ESSEC BUSINESS SCHOOL

PhD DISSERTATION • 2026

ESSEC BUSINESS SCHOOL

Contributions To The Theory Of

Bayesian Computation

Bias, Information, and Robustness

A dissertation submitted in partial fulfilment
of the requirements for the degrees of

PhD in Business Administration
and
Docteur en Mathématiques

de l'École Doctorale
« Économie, Management, Mathématiques,
Physique et Sciences informatiques » – ED 405
CY Cergy Paris Université

Presented and defended publicly
on the 21st of September 2026 by

El Mahdi Khribch

Sous la direction de

Pierre E. Jacob	Professeur, ESSEC Business School
Pierre Alquier	Professeur, ESSEC Business School

Composition du jury

Julyan Arbel	*Rapporteur*	Chargé de recherche, Inria, Université Grenoble Alpes
Gergely Neu	*Rapporteur*	Professeur de recherche, Universitat Pompeu Fabra & ICREA
Stéphanie Allassonnière	*Présidente du jury*	Professeure des universités, Université Paris Cité
Kamélia Daudel	*Examinatrice*	Professeure assistante, ESSEC Business School
Pierre E. Jacob	*Directeur de thèse*	Professeur, ESSEC Business School
Pierre Alquier	*Co-directeur de thèse*	Professeur, ESSEC Business School

© 2026 EL Mahdi Khribch. All rights reserved.

“Son of man,
You cannot say, or guess, for you know only
A heap of broken images, where the sun beats...”

— T. S. Eliot, *The Waste Land*

“The value of information does not survive the moment in which it was new. It lives only at that moment; it has to surrender to it completely and explain itself to it without losing any time.”

— Walter Benjamin, *Illuminations: Essays and Reflections*

“Hope cannot be said to exist, nor can it be said not to exist. It is just like roads across the earth. For actually the earth had no roads to begin with, but when many men pass one way, a road is made.”

— Lu Xun, *My Old Home*

"المخلوق جائز، الزمان محدث، والمكان محدث!
الزمان مفتقر، والمكان مفتقر "
من أوراد الشيخ الكامل المأخوذة من ورد الشيخ محمد بن سليم الجزولى : "سبحان الدائم لا يزول

Remerciements

Mes premiers remerciements vont à mes deux directeurs de thèse, Pierre E. Jacob et Pierre Alquier. Sans eux, cette thèse n'aurait été ni possible ni aussi heureuse. Je leur dois beaucoup : leur aide, leur franchise et leur générosité, la foi qu'ils m'ont toujours témoignée, et cette exigence bienveillante qui m'a sans cesse poussé à m'élever et à me dépasser.

Pierre J., tu es pour moi un modèle de chercheur. Je te suis reconnaissant de ta patience, de m'avoir inlassablement mis au défi et amené vers le meilleur de moi-même, sans jamais fermer ta porte ni cesser d'ouvrir mes horizons.

Pierre A., je songe à tes conseils et à nos réunions sur Zoom depuis Singapour, où nous passions du PAC-Bayes à Borges, et où tu accueillais volontiers mes excentricités ; tu m'as guidé avec douceur au fil de mes avancées, me laissant explorer librement les sujets qui me tenaient à cœur tout en veillant sur moi. À vous deux, pour votre érudition, l'appui sans faille et la confiance que vous avez placée en moi, ma gratitude sera sans fin ; puissé-je continuer de vous honorer et de me montrer digne de vous.

Mes remerciements vont également à Julyan Arbel et Gergely Neu, qui ont accepté d'être les rapporteurs de cette thèse, ainsi qu'à Stéphanie Allassonnière et Kamélia Daudel, qui me font l'honneur de siéger dans mon jury.

Je n'oublie pas ceux dont j'ai beaucoup appris. George, tu ne m'as jamais fait défaut, en ami autant qu'en guide. Guanyang, ton talent a tant apporté à notre travail ; ta finesse force l'admiration, et je forme le vœu que nos chemins ne cessent de se croiser. À Andi, je dois tant : tu m'as accueilli à Warwick et chez toi, auprès de Ludi et du petit « cho cho man ». Je suis reconnaissant à mes formidables amis de Warwick, Nicola et Rocco. Samy, mon ami de toujours, tu es pour moi un frère. Et à Simo et Badr, membres de mon comité de suivi de thèse et véritables grands frères, sur qui je sais pouvoir compter en toute circonstance.

Au merveilleux département de l'ESSEC : à Mohamed, El Houssein, Shiva et Alberto, et à tous les autres, trop nombreux pour être cités, mais qui se reconnaîtront sans peine.

Enfin, à ma famille. Voilà dix ans que la distance me sépare d'eux, et pourtant, tout au long de mes études, en France comme ailleurs, ils sont demeurés à mes côtés. C'est à vous que je dois d'avoir entrepris ce doctorat, moi qui suis le premier de la famille à le mener à son terme. À mon père, Rahhal, ma boussole : pour tes sacrifices et ton amour, pour cette foi que tu n'as jamais cessé de me porter, jusque dans les heures où je doutais de moi, j'espère t'avoir rendu fier. À ma mère, Samira, dont je porte le cœur en moi malgré l'éloignement : au fil de nos appels, tu étais là quoi qu'il advienne, que je fusse heureux ou triste, que je rie ou que je pleure, sans relâche à veiller sur moi. Je vous aime plus que les mots ne sauraient le dire. À vous deux, je voudrais rendre, ne fût-ce qu'un peu, tout ce que vous m'avez donné. À Ibrahim, mon frère, et à Aisha, ma sœur, pour l'amour et la lumière dont ils emplissent ma vie depuis toujours. Et à toi, Habiba : que serais-je sans toi qui vins à ma rencontre ? Tu as accompagné chacun de mes pas dans cette thèse, et sans toi, je n'aurais jamais trouvé la force d'aller jusqu'au bout.

Résumé substantiel

Titre : Contributions à la théorie du calcul bayésien : biais, information et robustesse.

Contexte et motivation. Soit $\{P_\theta : \theta \in \Theta\}$ un modèle statistique dominé par une mesure σ-finie μ, de densités $p_\theta = dP_\theta/d\mu$, soit π une loi a priori sur Θ, et soit $\mathcal{S} = (X_1, \dots, X_n)$ un échantillon i.i.d. de loi génératrice P^* inconnue.[1] À partir de la log-vraisemblance $\ell_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$, la règle de Bayes définit la loi a posteriori

$$\pi_n(d\theta) = \frac{e^{\ell_n(\theta)}\,\pi(d\theta)}{Z_n}, \qquad Z_n = \int_\Theta e^{\ell_n(\theta)}\,\pi(d\theta) \in (0, \infty),$$

où Z_n est la vraisemblance marginale. Toute quantité utile à l'inférence, à commencer par l'espérance a posteriori $\pi_n(h)$, s'écrit comme un rapport d'intégrales de la forme $\int_\Theta h(\theta)\, e^{\ell_n(\theta)}\, \pi(d\theta)$. Or, en dehors des familles conjuguées, aucune de ces intégrales n'admet de forme explicite, et le calcul bayésien repose sur deux grandes familles de méthodes, le Monte Carlo (Robert & Casella, 2004) et l'inférence variationnelle (Blei et al., 2017).

Problématique. Trois difficultés structurelles, une par chapitre, motivent les trois contributions de cette thèse. *(i) Biais computationnel des échantillonneurs.* Les échantillonneurs qui approchent π_n à budget fini sont biaisés. Les méthodes de Monte Carlo par chaînes de Markov (MCMC) de Metropolis et al. (1953); Hastings (1970) produisent une chaîne dont la loi marginale diffère de π_n à tout temps fini : ces méthodes itératives fournissent des estimateurs convergents lorsque le nombre d'itérations devient grand, mais potentiellement biaisés pour tout nombre d'itérations fixé, ce qui décourage l'exécution en parallèle de nombreuses chaînes courtes (Rosenthal, 2000). L'échantillonnage d'importance auto-normalisé (SNIS, *self-normalised importance sampling*), de poids $\omega = d\pi_n/dq$ contre une loi instrumentale q, emporte quant à lui un biais d'ordre $1/N$ (Agapiou et al., 2017), négligeable devant la variance mais dominant dans les procédures itératives, comme l'estimation des espérances emboîtées et des bornes variationnelles à poids d'importance. Le biais de Metropolis–Hastings indépendant (IMH, *independent Metropolis–Hastings*), qui repose sur la même loi de proposition globale q que le SNIS, est en revanche difficile à caractériser précisément dans le régime de poids non bornés, $\sup_x \omega(x) = \infty$: l'IMH y perd l'ergodicité géométrique qu'il possède

[1] L'appartenance ou non de P^* à $\{P_\theta\}$ (cadre bien spécifié ou mal spécifié) dépend du chapitre considéré ; le Chapitre 4 est dédié au régime mal spécifié.

sous poids bornés (Mengersen & Tweedie, 1996), et seules subsistent des vitesses polynomiales. *(ii) Certification des lois a posteriori généralisées.* Pour gagner en commodité de calcul ou en robustesse, on remplace souvent π_n par une loi a posteriori fractionnaire ou de Gibbs, ou par une approximation variationnelle. Leur contraction en échantillon fini est établie par des bornes PAC-bayésiennes (Bhattacharya et al., 2019; Alquier & Ridgway, 2020) gouvernées par une divergence de Kullback–Leibler à une loi a priori fixée (Catoni, 2007; Alquier, 2024) ; ce contrôle n'est pas assez fin et laisse subsister des facteurs $\log n$ superflus, que seule la théorie de l'information parvient à effacer. *(iii) Fragilité à la mauvaise spécification.* Sous mauvaise spécification ou contamination de Huber, de légères déviations de la loi génératrice P^* suffisent à concentrer π_n loin du meilleur paramètre (Grünwald & van Ommen, 2017; Owhadi et al., 2015). Deux courants y répondent sans se rejoindre : les lois a posteriori fractionnaires et variationnelles d'inspiration PAC-bayésienne (Alquier & Ridgway, 2020) sont calculables et certifiées en échantillon fini, mais n'offrent pas la robustesse d'un contraste borné ; les lois a posteriori ρ de Baraud & Birgé (2020), fondées sur un contraste borné de type Hellinger, sont robustes mais définies par un supremum sur l'espace des paramètres, difficile à approcher ou à calculer exactement en grande dimension, par MCMC comme par inférence variationnelle.

Contributions. Cette thèse établit trois résultats, un par chapitre : une analyse en échantillon fini de l'échantillonnage d'importance et de l'IMH en régime de poids à queues polynomiales, avec un échantillonnage d'importance sans biais par couplage (Coupling-UIS, Chapitre 2) ; une borne par information mutuelle certifiant la contraction de toute loi a posteriori dépendant des données, approximations variationnelles incluses (Chapitre 3) ; et une contrepartie variationnelle calculable de la loi a posteriori ρ de Baraud & Birgé (2020), robuste en distance de Hellinger (Chapitre 4).

Contribution 1 (Chapitre 2, avec G. Deligiannidis, P. E. Jacob et G. Wang; Deligiannidis et al., 2024). Nous analysons l'IMH et le SNIS dans le régime sous-géométrique $\sup_x \omega(x) = \infty$ avec $q(\omega^p) < \infty$ pour un certain $p > 1$, en dehors du régime de poids bornés sous lequel Mengersen & Tweedie (1996) établit l'ergodicité géométrique. *(a)* Nous prouvons que le couplage par nombres aléatoires communs (CRN) de deux chaînes IMH est maximal en variation totale à tout temps fini, seul exemple connu pour un algorithme MCMC. *(b)* Nous établissons que le biais en variation totale de l'IMH est $\mathcal{O}(N^{-(p-1)})$, plus faible que le biais $\mathcal{O}(N^{-1})$ du SNIS à budget égal dès que $p > 2$; l'intérêt se manifeste dans les procédures itératives (espérances emboîtées, bornes variationnelles, optimisation stochastique), où tout biais résiduel impose une erreur incompressible qu'aucune réduction de variance ne peut éliminer. *(c)* L'idée de débiaiser l'IMH en couplant deux chaînes décalées et en appliquant la correction télescopique de Jacob et al. (2020), dans la lignée de Glynn & Rhee (2014), est due à Middleton et al. (2019), pour l'IMH particulaire sous poids bornés. Nous observons que la même construction débiaise le SNIS, d'où une variante sans biais, le Coupling-UIS, dont nous établissons la variance finie dès que $p \geq 2$, en caractérisant la queue du temps de rencontre τ, objet central de l'analyse du

Monte Carlo par chaînes de Markov sans biais (UMCMC) (Douc et al., 2026; Atchadé & Jacob, 2026). Le caractère exactement sans biais du Coupling-UIS permet de l'associer à une large classe d'estimateurs robustes (médiane des moyennes, estimateurs sous-gaussiens) (Nemirovsky & Yudin, 1983; Minsker & Ndaoud, 2021; Lee & Valiant, 2022), contrairement au SNIS biaisé, dont l'analyse robuste de Dau (2022) se limite à la médiane des moyennes.

Contribution 2 (Chapitre 3, avec P. Alquier; Khribch & Alquier, 2024). Pour toute loi a posteriori $\hat{\rho}$ dépendant des données, nous bornons sa contraction en divergence de Rényi par un risque empirique et l'information mutuelle $\mathcal{I}(\theta; \mathcal{S})$, qui remplace la KL des bornes PAC-Bayes classiques (Catoni, 2007; Alquier, 2024) et lui est toujours inférieure (Russo & Zou, 2016; Xu & Raginsky, 2017). Combinée à la localisation de Catoni (2007), la borne donne, sous une condition usuelle de masse a priori, un taux d_π / n^{κ_π} (d_π, dimension effective du modèle) régi par un unique exposant $\kappa_\pi \in (0, 1]$: il couvre les régimes paramétrique ($\kappa_\pi = 1$) et non paramétrique ($\kappa_\pi < 1$) et efface, en $\kappa_\pi = 1$, le facteur $\log n$ de Bhattacharya et al. (2019); Alquier & Ridgway (2020), pour la loi a posteriori fractionnaire, l'approximation variationnelle et l'estimateur du maximum de vraisemblance.

Contribution 3 (Chapitre 4, avec P. Alquier; Khribch & Alquier, 2026). Nous proposons une variante calculable de la loi a posteriori ρ de Baraud & Birgé (2020) : le supremum qui la définit est remplacé par une agrégation à poids exponentiels à la Donsker–Varadhan, d'où une loi a posteriori $\tilde{\rho}$ de Gibbs se prêtant à l'analyse PAC-bayésienne. Nous en déduisons une inégalité oracle non asymptotique en distance de Hellinger, où le terme de mauvaise spécification $\mathcal{H}^2(P^*, P_{\theta_0})$ intervient de manière additive, borné par ε sous contamination de Huber. La formulation point-selle associée est non convexe–fortement concave de conditionnement $\mathcal{O}(1)$, et l'extragradient stochastique y atteint une complexité d'itération indépendante de n.

Portée et impact. Au fil des trois chapitres, la thèse établit de nouvelles garanties en échantillon fini, à la fois numériques et statistiques : un Monte Carlo sans biais au-delà de l'ergodicité géométrique, sous les seuls moments polynomiaux du poids, avec un estimateur débiaisé à coût équivalent (Chapitre 2) ; une unique borne en information mutuelle dont un même taux couvre les régimes paramétrique et non paramétrique des lois a posteriori variationnelle, fractionnaire et de Gibbs (Chapitre 3) ; et une loi a posteriori robuste et calculable, préservant l'inégalité oracle de Hellinger sous contamination (Chapitre 4).

Mots-clés : calcul bayésien ; échantillonnage d'importance ; MCMC sans biais ; PAC-Bayes ; information mutuelle ; inférence variationnelle robuste.

Abstract

Title: Contributions to the Theory of Bayesian Computation: Bias, Information, and Robustness.

Context and motivation. Let $\{P_\theta : \theta \in \Theta\}$ be a statistical model dominated by a measure μ with densities $p_\theta = dP_\theta/d\mu$, π a prior on Θ, and $\mathcal{S} = (X_1, \ldots, X_n)$ an i.i.d. sample from an unknown data-generating distribution P^*.[2] With log-likelihood $\ell_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$, Bayes' rule defines the posterior

$$\pi_n(d\theta) = \frac{e^{\ell_n(\theta)}\,\pi(d\theta)}{Z_n}, \qquad Z_n = \int_\Theta e^{\ell_n(\theta)}\,\pi(d\theta) \in (0, \infty),$$

where Z_n is the marginal likelihood. Every inferentially relevant quantity, beginning with the posterior expectation $\pi_n(h)$, is a ratio of integrals of the form $\int_\Theta h(\theta)\, e^{\ell_n(\theta)}\,\pi(d\theta)$. However, outside conjugate families none of these integrals admits a closed form, and Bayesian computation rests on two complementary toolboxes, Monte Carlo sampling (Robert & Casella, 2004) and variational inference (Blei et al., 2017).

Challenges. Three structural difficulties, one per chapter, motivate the contributions. *(i) Computational bias of samplers.* Outside conjugate families π_n has no closed form and finite-budget samplers are biased: MCMC chains differ from π_n at every finite time. These iterative methods yield estimators that are consistent as the number of iterations grows but potentially biased at any fixed iteration count, which discourages running many short chains in parallel (Rosenthal, 2000), while self-normalised importance sampling (SNIS) carries an $\mathcal{O}(1/N)$ bias (Agapiou et al., 2017) that is negligible against the variance but dominant inside iterative procedures such as nested expectations and importance-weighted variational bounds. By contrast, the bias of the independent Metropolis–Hastings (IMH) sampler, which uses the same global proposal q as SNIS, is harder to characterise precisely in the unbounded-weight regime $\sup_x \omega(x) = \infty$, where it loses geometric ergodicity (Mengersen & Tweedie, 1996). *(ii) Certification of generalised posteriors.* Fractional, Gibbs, and variational posteriors are certified through PAC-Bayes bounds (Bhattacharya et al., 2019; Alquier & Ridgway, 2020) governed by a Kullback–Leibler

[2] Whether P^* belongs to $\{P_\theta\}$ (the well-specified case) or not (misspecification) depends on the setting and is clarified chapter by chapter; Chapter 4 is dedicated to the misspecified regime.

divergence to a fixed prior (Catoni, 2007); this control is not tight and leaves spurious $\log n$ factors that only an information-theoretic perspective removes. *(iii) Fragility to misspecification.* Under misspecification or Huber contamination, slight deviations of P^* concentrate π_n far from the best parameter (Grünwald & van Ommen, 2017; Owhadi et al., 2015). Two strands of work respond, without ever coming together: the PAC-Bayes fractional and variational posteriors (Alquier & Ridgway, 2020) are computable and certified but lack the robustness of a bounded contrast, while the ρ-posteriors of Baraud & Birgé (2020), built on a bounded Hellinger contrast, are robust but defined by a supremum over the parameter space that is hard to approximate or compute exactly in high dimension.

This thesis contributes three results, one per chapter: a finite-sample analysis of importance sampling and independent Metropolis–Hastings under heavy-tailed weights, together with a coupling-based unbiased importance sampling estimator (Coupling-UIS) for self-normalised importance sampling (Chapter 2); a mutual-information bound certifying the contraction of any data-dependent posterior, variational families included (Chapter 3); and a tractable variational counterpart of the Hellinger-robust ρ-posterior of Baraud & Birgé (2020) (Chapter 4).
Contribution 1 (Deligiannidis et al., 2024, Chapter 2; joint with G. Deligiannidis, P. E. Jacob and G. Wang). We analyse the independent Metropolis–Hastings (IMH) sampler and the self-normalised importance sampling (SNIS) estimator in the subgeometric regime where the importance weight $\omega = d\pi_n/dq$ has $\sup_x \omega(x) = \infty$ with $q(\omega^p) < \infty$ for some $p > 1$; this regime lies outside the bounded-weight condition under which Mengersen & Tweedie (1996) establish geometric ergodicity of the IMH chain. Three results. *(a)* The common-random-numbers (CRN) coupling of two IMH chains saturates the coupling inequality at every finite time, the only known example of an all-time maximal coupling for a Markov chain Monte Carlo (MCMC) algorithm. *(b)* The IMH total-variation bias is $\mathcal{O}(N^{-(p-1)})$, smaller than the $\mathcal{O}(N^{-1})$ bias of SNIS at matched proposal budget once $p > 2$; the SNIS asymptotic variance is in turn smaller than that of IMH (Deligiannidis & Lee, 2018), so the benefit of IMH is realised inside iterative procedures (nested Monte Carlo, importance-weighted variational bounds, biased-gradient stochastic optimisation), where the $\mathcal{O}(N^{-1})$ SNIS bias compounds and dominates the mean-squared error. *(c)* The idea of debiasing independent Metropolis–Hastings by coupling two chains with a lag and applying the telescoping correction of Jacob et al. (2020), following Glynn & Rhee (2014), is due to Middleton et al. (2019), who introduced it for particle independent Metropolis–Hastings under bounded weights. We observe that the same construction debiases SNIS, and we establish finite variance beyond the bounded-weight regime: the resulting unbiased importance sampling (Coupling-UIS) estimator has finite variance under $q(\omega^p) < \infty$, $p \geq 2$, the proof converting the finite weight-moment $q(\omega^p)$ into a finite moment of the coupling meeting time τ, a quantity central to the analysis of unbiased MCMC (UMCMC) algorithms (Douc et al., 2026; Atchadé & Jacob, 2026). Because Coupling-UIS is exactly unbiased, it can be fed into any robust mean estimator (median-of-

means, sub-Gaussian) (Nemirovsky & Yudin, 1983; Minsker & Ndaoud, 2021; Lee & Valiant, 2022) to produce robust estimates of $\pi_n(h)$, a composition that biased SNIS does not admit (Dau, 2022).

Contribution 2 (Khribch & Alquier, 2024, Chapter 3; joint with P. Alquier). We establish a mutual-information bound on the contraction of any data-dependent distribution $\hat{\rho}$, controlling its Rényi risk by an empirical-risk term and the mutual information $\mathcal{I}(\theta; \mathcal{S})$, in the spirit of Russo & Zou (2016) and Xu & Raginsky (2017). The bound is strictly tighter than the Kullback–Leibler divergence appearing in classical PAC-Bayes (Catoni, 2007; Alquier, 2024), the gap quantifying exactly the price of committing to a fixed prior. Combined with the localisation technique of Catoni (2007), it yields, under standard prior-mass hypotheses, a contraction rate of order d_π / n^{κ_π} governed by a single prior-mass exponent $\kappa_\pi \in (0, 1]$, with d_π the effective model dimension, covering parametric ($\kappa_\pi = 1$) and nonparametric ($\kappa_\pi < 1$) Bayesian asymptotics; in the parametric case it attains the d_π / n rate, without the spurious $\log n$ factor carried by Bhattacharya et al. (2019) and Alquier & Ridgway (2020).

Contribution 3 (Khribch & Alquier, 2026, Chapter 4; joint with P. Alquier). The ρ-posteriors of Baraud & Birgé (2020) are robust under misspecification but hard to approximate in high dimension. We make them computable through tractable variational approximations that preserve the Huber-robustness guarantees: the supremum defining the ρ-posterior is replaced by a smooth Donsker–Varadhan exponential-weights aggregation (Donsker & Varadhan, 1976; Catoni, 2007), yielding a $\tilde{\rho}$-posterior that is an exponentially weighted Gibbs measure amenable to PAC-Bayesian analysis. We prove a high-probability oracle inequality in squared Hellinger distance whose model-misspecification term $\mathcal{H}^2(P^*, P_{\theta_0})$ enters additively rather than multiplying the complexity; under Huber's ε-contamination this term is bounded by ε, so the contamination penalty enters only additively. The resulting nonconvex–strongly-concave saddle-point problem has condition number $\mathcal{O}(1)$ uniformly in the temperature, so projected stochastic extragradient attains a sample-size-free iteration complexity.

Relevance and impact. Across the three chapters, the thesis establishes new finite-sample guarantees, computational and statistical: unbiased Monte Carlo beyond geometric ergodicity under polynomial weight moments, debiased at equal cost (Chapter 2); a mutual-information bound covering the parametric and nonparametric regimes of variational, fractional and Gibbs posteriors (Chapter 3); and a tractable robust posterior preserving the Hellinger oracle inequality under contamination (Chapter 4).

Keywords: Bayesian Computation; Importance Sampling; Unbiased Markov Chain Monte Carlo; PAC-Bayes and Mutual-Information Bounds; Robust Variational Inference.

To my family,

and to the memory of my grandmother

مي لالة الفايزة

Contents

Remerciements iv

Résumé substantiel v

Abstract viii

1 Introduction 2
- 1.1 Bayesian Inference and Posterior Approximation . . . 2
- 1.2 Importance Sampling and the IMH Sampler . . . 4
- 1.3 Two Finite-Sample Biases . . . 9
- 1.4 Couplings, Meeting Times, and Unbiased MCMC . . . 12
- 1.5 Variational Approximation of the Posterior . . . 15
- 1.6 PAC-Bayes, MI, and Posterior Certification . . . 17
- 1.7 Robust Posteriors under Misspecification . . . 22

2 Importance Sampling and IMH with Unbounded Weights 26
- 2.1 Introduction . . . 27
- 2.2 Bias and moments of importance sampling . . . 35
- 2.3 Optimality of coupling IMH with common draws . . . 37
- 2.4 Meeting times and polynomial convergence . . . 39
- 2.5 Bias removal for self-normalized importance sampling . . . 46
- 2.6 Applications . . . 53

3 Convergence of Statistical Estimators via MI Bounds 58
- 3.1 Introduction and Motivation . . . 59
- 3.2 Notations and Preliminaries . . . 61
- 3.3 Main Results . . . 64
- 3.4 Examples . . . 68

4 Robust Variational Inference via ρ-Posteriors 76
- 4.1 Introduction and Motivation . . . 77

4.2 Main Results . . . 82
4.3 Numerical Experiments . . . 93
4.4 Conclusion . . . 97

5 Conclusion and Perspectives 98
5.1 Sampling under global proposals . . . 98
5.2 Certifying and robustifying generalised posteriors . . . 101

Appendices 103

A Proofs of Chapter 2 104
A.1 Proofs of Section 2.1 . . . 104
A.2 Proofs of Section 2.2 . . . 105
A.3 Proofs of Section 2.3 . . . 111
A.4 Proofs of Section 2.4 . . . 112
A.5 Proofs of Section 2.5 . . . 127

B Proofs of Chapter 3 137
B.1 PAC-Bayes bounds . . . 137
B.2 Mutual information and the main-theorem proof . . . 139
B.3 Localized PAC-Bayes bounds and proofs . . . 140
B.4 Other proofs . . . 146

C Proofs of Chapter 4 156
C.1 Technical preliminaries . . . 156
C.2 Proof of the PAC-Bayes bound . . . 156
C.3 Proof of the corollary at an explicit inverse temperature . . . 159
C.4 Proof of the oracle inequality . . . 160
C.5 Proofs of the prior-mass condition . . . 161
C.6 Proofs for the regression application . . . 162
C.7 Proof of the variational oracle inequality . . . 167
C.8 Proof of the saddle-point equivalence and gap bound . . . 169
C.9 Proof of the nonconvex–strongly-concave geometry . . . 172

D Additional Results 177
D.1 Alternative unbiased importance sampling estimators . . . 177
D.2 Implementation Details . . . 181
D.3 Additional Numerical Results . . . 182
D.4 Leave-one-out SNIS and its asymptotic bias . . . 183
D.5 Robust Mean Estimation: Median-of-Means and Sub-Gaussian Rates . . . 194

D.6 Convergence Rates for Nested Expectations 195
D.7 Comparison with Hesterberg's Bias Formula 197

Bibliography **199**

List of Figures

1.1 Roadmap of the thesis. 5
1.2 IMH trace and importance-weight distribution. 9
1.3 L-lag coupled IMH and meeting-time histogram. 14
1.4 Variational zoom. 16
1.5 Variational fit $\widehat{\rho}_n$ against the target. 17
1.6 MIB–PAC-Bayes bridge. 21

2.1 Inefficiency of $1/Z$ estimators versus cost. 54
2.2 Mean-squared error of nested-expectation estimators. 56
2.3 Error quantiles for robust mean estimation. 57

4.1 Gaussian location model results. 95
4.2 Poisson intensity model results. 95
4.3 Uniform scale model results. 95
4.4 Fourier basis regression. 96
4.5 Correlated design regression. 96
4.6 Test residual distributions at 10% contamination. 97

D.1 Meeting time of the coupled IMH chains. 182
D.2 Variance and inefficiency of the Coupling-UIS estimator. 183

Chapter 1

Introduction

1.1 Bayesian Inference and Posterior Approximation

Throughout the chapter, $(\Theta, \mathcal{B})$ denotes a measurable parameter space carrying a prior probability measure π, and $\mathcal{S} = (X_1, \dots, X_n)$ are observations on a measurable space $(\mathbb{X}, \mathcal{X})$, conditionally i.i.d. given $\theta \in \Theta$ from a parametric family $\{P_\theta : \theta \in \Theta\}$ dominated by a σ-finite reference measure μ, with densities $p_\theta = dP_\theta/d\mu$ jointly measurable in (θ, x). The log-likelihood is $\ell_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$. The inversion rule formulated for binary events by Bayes (1763) and extended to densities by Laplace (1774) produces the *posterior*

$$\pi_n(d\theta) := \pi(d\theta \mid \mathcal{S}) = \frac{e^{\ell_n(\theta)}\,\pi(d\theta)}{Z_n}, \qquad Z_n := \int_\Theta e^{\ell_n(\vartheta)}\,\pi(d\vartheta) \in (0, \infty), \tag{1.1}$$

the marginal likelihood Z_n being finite by Fubini's theorem under the preceding measurability assumption. We write π_n for the posterior and reserve π for the prior.

Equation (1.1) is the mathematical content of Bayesian inference: a principled rule for updating beliefs about θ in the light of the sample $\mathcal{S}$. The prior π encodes the initial uncertainty; the posterior π_n encodes what remains after having observed $\mathcal{S}$. The framework treats uncertainty probabilistically from end to end, and Bayes' rule is an optimal information-processing rule (Zellner, 1988). That generality carries a computational price. Outside conjugate families, π_n is known only through the unnormalised numerator of (1.1), so inference must proceed by approximations whose finite-sample error has to be controlled.

Every quantity of inferential interest is a functional of π_n: the expectation $\pi_n(h) = \int h\, d\pi_n$ of a measurable test function h, the credible probability $\pi_n(A)$ of a measurable set $A \in \mathcal{B}$, the predictive density $p(x_{n+1} \mid \mathcal{S}) = \int p_\theta(x_{n+1})\, d\pi_n(\theta)$ of a future observation, and, under any proper loss, the Bayes action and Bayes risk of any decision rule. That every such func-

tional is determined by π_n alone, independently of any prior or sampling-distribution decomposition, is the foundational decision-theoretic statement of Bayesian inference (Bernardo & Smith, 1994; Robert, 2007); in practice, one therefore reports the posterior expectation $\pi_n(h)$ rather than the posterior law itself. Frequentist guarantees for inference based on those functionals are supplied by posterior contraction theory and, more sharply, by the Bernstein–von Mises theorem. Under appropriate prior-mass and testing conditions, π_n concentrates around the data-generating parameter: in the nonparametric case at the minimax-adaptive contraction rate (Ghosal & van der Vaart, 2017, Chs. 8–10) and (Castillo, 2024, Chs. 1–4), and in the parametric case at the rate $n^{-1/2}$, where the Bernstein–von Mises theorem (van der Vaart, 2000, Ch. 10) gives asymptotic normality of π_n around the maximum-likelihood estimator, with a misspecified counterpart due to Kleijn & van der Vaart (2012).

The choice of the prior π is itself a modelling problem, but for the purposes of this thesis it enters only through the prior-mass conditions of Ghosal & van der Vaart (2017), which can be checked directly and are the level at which the analyses of Chapters 3 and 4 operate.

The successes of Bayesian computation. The translation of (1.1) into routine inferential practice has been driven by two and a half centuries of methodological progress, recently traced in chronological detail by Martin et al. (2024). The story passes from the one-dimensional integrals first confronted by Bayes (1763) and Laplace (1774), through Laplace's approximation and its modern Bayesian treatment by Tierney & Kadane (1986), now equipped with computable finite-sample error bounds (Kasprzak et al., 2025), to the development of Markov chain Monte Carlo (MCMC) following Metropolis et al. (1953) and Hastings (1970), to the dissemination of the Gibbs sampler by Gelfand & Smith (1990) and the state-space analysis of Metropolis–Hastings due to Tierney (1994). The accompanying maturation of importance sampling (Geweke, 1989; Tokdar & Kass, 2010; Liu, 2008; Owen, 2013), sequential Monte Carlo (Del Moral et al., 2006; Chopin & Papaspiliopoulos, 2020) and variational inference (Jordan et al., 1999; Blei et al., 2017) has turned posterior expectation, credible-set construction and predictive evaluation into routine practice, at an ever-growing computational scale.

Challenges. The computational obstruction is intrinsic to (1.1): outside conjugate families, π_n is known only up to Z_n, so although the unnormalised density is evaluable pointwise, the integrals defining $\pi_n(h)$, $\pi_n(A)$ and Z_n admit no closed form. Inference proceeds by approximations whose error must be quantified (Robert & Casella, 2004, Ch. 1), and the difficulty compounds under hierarchical structure, nonparametric priors, large n and high-dimensional θ. The recent surveys of Winter et al. (2024) and Bhattacharya et al. (2024) identify two central open challenges: the control of approximation error, and the design of methods whose finite-sample behaviour can be *certified*. Computable, finite-sample bounds on posterior-approximation error are precisely the certificates these challenges call for, valid at the budgets

practitioners actually use rather than only in the limit.

Two complementary families of methods address the computational obstruction itself. *Sampling-based* methods, namely importance sampling and MCMC, target π_n via weighted particles or ergodic Markov-chain trajectories without ever evaluating Z_n. *Optimisation-based* methods replace π_n by the closest surrogate $\widehat{\rho}_n$ in a tractable family $\mathcal{F} \subset \mathcal{M}_1^+(\Theta)$ of distributions on Θ, minimising a divergence in which Z_n also cancels; $\mathcal{M}_1^+(\Theta)$ denotes the probability measures on $(\Theta, \mathcal{B})$. Sections 1.2 and 1.5 introduce the two families in turn. The thesis then contributes three results: Chapter 2 establishes finite-sample bias bounds for self-normalised importance sampling and independent Metropolis–Hastings under unbounded weights, and proposes an unbiased self-normalised estimator; Chapter 3 proves a mutual-information generalisation bound for variational and PAC-Bayesian posteriors; and Chapter 4 constructs a tractable, robust variational posterior for misspecified models. Figure 1.1 gives a visual overview of the thesis.

Guarantees matter most for the optimisation-based methods of Section 1.5, and the last two of these contributions supply them. Such methods do not target π_n directly: they search a tractable family $\mathcal{F}$ of distributions and return its member closest to π_n in Kullback–Leibler divergence. When π_n lies outside $\mathcal{F}$, this minimiser $\widehat{\rho}_n$ is the best approximation $\mathcal{F}$ can offer, not π_n itself, so a guarantee must be proved for $\widehat{\rho}_n$ directly rather than borrowed from π_n (Wang & Blei, 2019; Alquier et al., 2016; Pati et al., 2018). The PAC-Bayesian bounds of Section 1.6 provide exactly such guarantees, and they apply broadly: the generalised-Bayesian update of Bissiri et al. (2016), re-cast as optimisation-centric inference by Knoblauch et al. (2022), gathers Bayes' rule, fractional posteriors and Gibbs measures into a single decision-theoretic family built to tolerate misspecification. Chapter 3 sharpens the complexity term of these variational and PAC-Bayesian posteriors into a mutual-information bound; Chapter 4 builds on the same framework the first tractable variational counterpart of the ρ-posterior, provably robust under misspecification.

1.2 Importance Sampling and the IMH Sampler

Sampling-based methods generate random elements of Θ whose empirical distributions converge weakly to π_n, organised around two templates. Importance sampling, originating in neutron-transport Monte Carlo (Kahn, 1949; Hammersley & Handscomb, 1964) with the history reconstructed by Andral (2022), draws an i.i.d. sample $x_1, \ldots, x_N \sim q$ from a proposal q on Θ and reweights by the Radon–Nikodym derivative $\omega := d\pi_n/dq$ (formally defined in (1.6) below). Its self-normalised form, defined in (1.7) below, is a ratio estimator: the strong law of large numbers applied to numerator and denominator yields almost-sure consistency of their ratio, while a central limit theorem for the pair, combined with the delta method, supplies asymptotic normality (Owen, 2013, §9.2). Geweke (1989) imported the construction into Bayesian

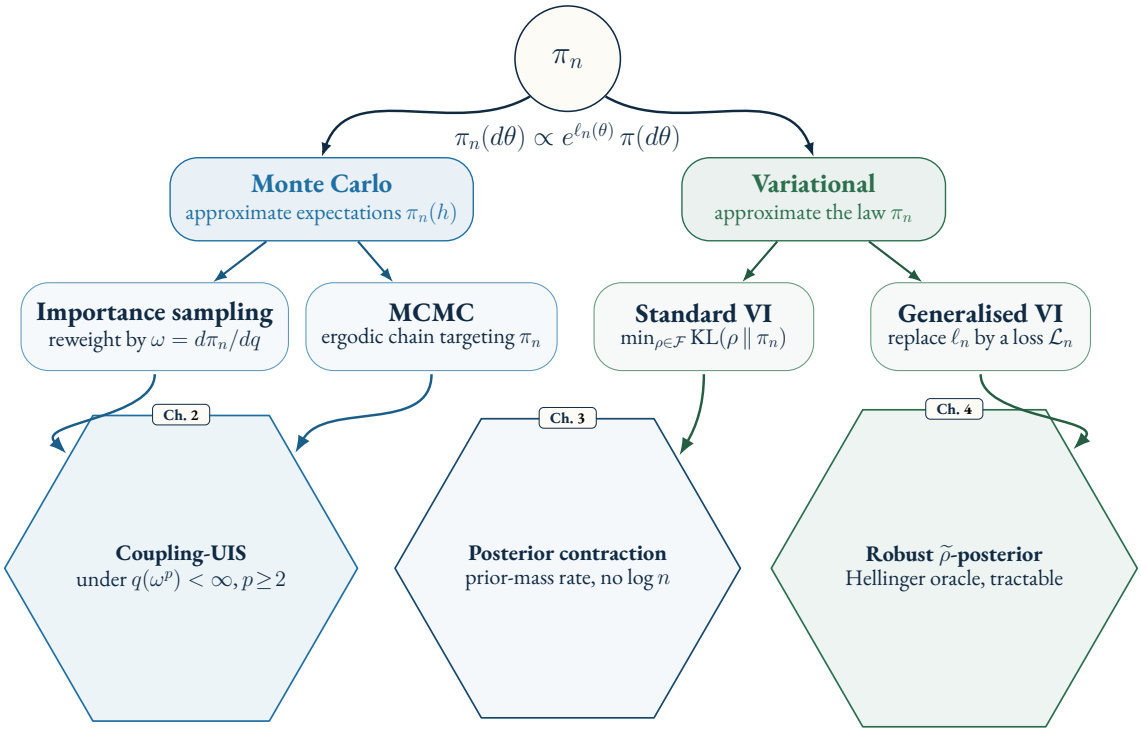


Figure 1.1: *Roadmap of the thesis.* Two complementary routes approach the intractable posterior π_n: Monte Carlo (importance sampling, MCMC) and variational inference (standard, generalised). Chapter 2 provides a coupling-based unbiased importance sampling estimator (Coupling-UIS) under finite weight moments; Chapter 3 gives a mutual-information contraction bound at the prior-mass rate; Chapter 4 builds a tractable, robust variational $\widetilde{\rho}$-posterior under misspecification.

computation under moment conditions $q(\omega^2) < \infty$ and $q(h^2\omega) < \infty$; Hesterberg (1988) derived the asymptotic bias and the SNIS-optimal proposal $q^\star \propto |h - \pi_n(h)|\,\pi_n$; Liu (2008, Ch. 2) and Owen (2013, Ch. 9) record how $q(\omega^2)$ controls the asymptotic variance and how higher moments $q(\omega^p)$, $p > 2$, control the higher-order terms in the estimator's bias and variance. This second moment also measures the discrepancy between target and proposal, through the identity

$$q(\omega^2) = 1 + \chi^2(\pi_n \,\|\, q), \tag{1.2}$$

with $\chi^2(\pi_n \,\|\, q)$ the chi-square divergence. It is this second moment that controls the variance of the importance-sampling estimator of Z_n (Orsak & Aazhang, 1991; Agapiou et al., 2017); the KL-based sample-size analysis of Chatterjee & Diaconis (2018) delineates the regime where χ^2 control fails but importance sampling remains informative. Minimising $\chi^2(\pi_n \,\|\, q)$ over $q \in \mathcal{F}$, or a Rényi-divergence extension (Dieng et al., 2017; Li & Turner, 2016), is therefore a function-free surrogate for the SNIS proposal problem. *Importance-weighted variational inference* (IWVI), introduced through an importance-weighted ELBO by Burda et al. (2016), provides a related SNIS-driven objective; Domke & Sheldon (2018) make the connection exact by recasting IWVI maximisation, for every finite N, as standard variational inference on the SNIS-

augmented joint over N proposals.

MCMC builds an ergodic Markov kernel P on Θ with stationary distribution π_n and approximates expectations by time averages along a single trajectory $(x_t)_{t\geq 0}$, following Metropolis et al. (1953) and Hastings (1970). A particular case is the *independent* Metropolis–Hastings (IMH) chain, whose proposal q does not depend on the current state. Like SNIS, its finite-sample properties are governed by the same importance weight $\omega = d\pi_n/dq$, that is by the discrepancy between target and proposal: the heavier the tail of ω, the harder both methods become, and in both the controlling quantity is the same family of weight moments $q(\omega^p)$. What differs is only how these moments enter. For SNIS, they bound the bias and variance of the estimator. For IMH, they set the convergence rate of the chain: a bounded weight, $\sup_x \omega(x) < \infty$, gives geometric ergodicity, while an unbounded weight with $q(\omega^p) < \infty$ gives a polynomial rate of order $p - 1$. The practical remedy is common to both: choose q with heavy enough tails that the weight is bounded, or at least light-tailed, which restores the favourable regime (Owen, 2013, Ch. 9). Chapter 2 treats the harder regime, where no such proposal is available and ω is genuinely unbounded, and there establishes results for both methods, the moments of SNIS and the bias and convergence of IMH, under the same moment conditions.

The Metropolis–Hastings algorithm. The Metropolis–Hastings (MH) algorithm of Metropolis et al. (1953) and Hastings (1970) builds such a kernel from a proposal $q(y \mid x)$: a candidate $y \sim q(\cdot \mid x)$ is accepted with probability

$$\alpha(x, y) = \min\Big(1,\ \frac{\pi_n(y)\, q(x \mid y)}{\pi_n(x)\, q(y \mid x)}\Big), \tag{1.3}$$

and is otherwise rejected. The target enters only through the ratio $\pi_n(y)/\pi_n(x)$, so the unknown normalising constant Z_n cancels and α is computable; the resulting kernel P is reversible with respect to π_n. Under irreducibility and aperiodicity the time averages $\widehat{\pi}_n^t(h) := t^{-1}\sum_{s=0}^{t-1} h(x_s)$ converge almost surely to $\pi_n(h)$ for every π_n-integrable h (Meyn & Tweedie, 2012, Thm. 17.0.1). The chain is *geometrically ergodic* when

$$\big|P^t(x, \cdot) - \pi_n\big|_{\mathrm{TV}} \leq C(x)\, \gamma^t, \qquad x \in \Theta,\ t \geq 0, \tag{1.4}$$

for some $\gamma \in (0, 1)$ and a π_n-integrable $C : \Theta \to (0, \infty)$, which implies a Markov-chain central limit theorem for $\sqrt{t}\,(\widehat{\pi}_n^t(h) - \pi_n(h))$ with asymptotic variance

$$v(P, h) := \mathbb{V}_{\pi_n}(h) + 2\sum_{k\geq 1} \mathbb{C}\mathrm{ov}_{\pi_n}\big(h(x_0), h(x_k)\big). \tag{1.5}$$

For the *independent* Metropolis–Hastings (IMH) specialisation used throughout, with state-independent proposal $q(\cdot \mid x) = q(\cdot)$, these conditions reduce to the single requirement $\pi_n \ll q$: the chain is then π_n-irreducible (Mengersen & Tweedie, 1996, Thm. 2.1) and aperiodic (Roberts & Rosenthal, 2004, §3.4), hence ergodic. Given its acceptance ratio (1.8), its convergence rate and asymptotic variance are developed in Section 2.4 of Chapter 2.

SNIS and IMH as global proposals. Both algorithms share a global-proposal template: a proposal q on Θ independent of the current state, an i.i.d. population $x_1, \dots, x_N \sim q$, and an importance weight converting q-draws into π_n-information. The same architecture underlies sampling–importance resampling (SIR, Rubin, 1988), which resamples one point proportional to ω, and the independent-proposal variant of multiple-try Metropolis (MTM-IS, Liu et al., 2000). Common to all of them is the importance weight: whenever $\pi_n \ll q$, the Radon–Nikodym derivative

$$\omega(\theta) := \frac{d\pi_n}{dq}(\theta) = \frac{\pi_n(\theta)}{q(\theta)}, \tag{1.6}$$

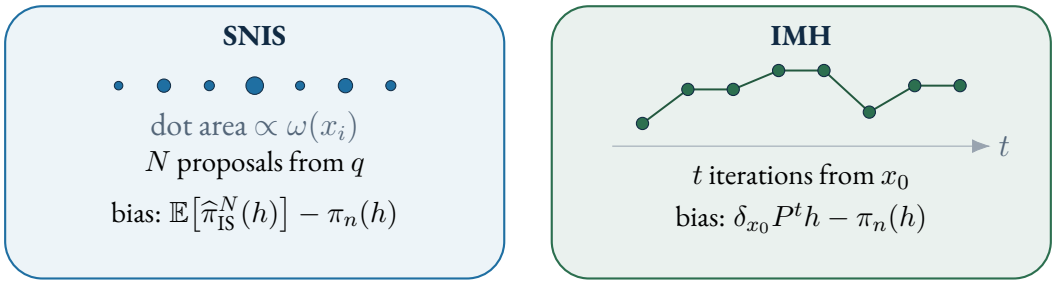
is well defined q-almost surely. Given i.i.d. draws $x_1, \dots, x_N \overset{\text{iid}}{\sim} q$, the *self-normalised importance-sampling* (SNIS) estimator of $\pi_n(h)$ is

$$\widehat{\pi}^N_{\text{IS}}(h) = \frac{\displaystyle\sum_{i=1}^{N} \omega(x_i)\, h(x_i)}{\displaystyle\sum_{i=1}^{N} \omega(x_i)}, \tag{1.7}$$

the ratio form mandated by the unavailability of Z_n. The *independent Metropolis–Hastings* (IMH) algorithm (Tierney, 1994) takes a state-independent proposal $q(\cdot \mid x) = q(\cdot)$, so that the acceptance probability simplifies to

$$\alpha(x, y) = \min\Big(1, \frac{\omega(y)}{\omega(x)}\Big). \tag{1.8}$$

SNIS and IMH are the global-proposal methods we consider in our analysis, and ones that also arise in practice, where a learned approximation of π_n (variational fit, Laplace, or normalising-flow surrogate; Rezende & Mohamed, 2015; Papamakarios et al., 2021) is plugged in as q. IMH can also serve as a building block inside a larger sampler, for instance as a step in Metropolis-within-Gibbs (Skare et al., 2003). The two methods place the bias differently: SNIS in its weighted average over N draws, IMH in the law of its chain after t steps.



Taking the supremum over bounded test functions turns the SNIS bias into the total-variation distance between the SIR resampling marginal and π_n, and the IMH bias into the total-variation distance of the IMH marginal to π_n, compared in (1.16) below. In practice no clear default exists between SNIS and IMH at matched proposal budget: SNIS has the lower asymptotic variance (Deligiannidis & Lee, 2018, Prop. 2), IMH the smaller total-variation bias once the weight has enough finite moments. Which of the two to prefer therefore turns on the tail of the weight ω, the trade-off taken up in Chapter 2.

Two regimes of the importance weight. Both SNIS (1.7) and the IMH time averages converge almost surely to $\pi_n(h)$ under the single condition $q(|h|\,\omega) = \pi_n(|h|) < \infty$ (Liu, 2008; Owen, 2013; Meyn & Tweedie, 2012, Thm. 17.0.1; Tierney, 1994, Cor. 2). Beyond consistency, the regime of convergence is determined by the boundedness of ω: Mengersen & Tweedie (1996, Thm. 2.1) show that IMH is uniformly ergodic if and only if $\sup_x \omega(x) < \infty$, in which case the geometric bound (1.4) holds, with the exact convergence rate known explicitly (Wang, 2022); the asymptotic variance of SNIS then exceeds that of direct Monte Carlo sampling from π_n by at most the factor $\omega^\star := \sup_x \omega(x)$ (Owen, 2013, Ch. 9, Ex. 9.8). When $\sup_x \omega(x) = \infty$, IMH is no longer geometrically ergodic; under moment assumptions on the weight, polynomial total-variation convergence rates follow from the general drift-condition machinery of Jarner & Roberts (2002), with quantitative non-geometric bounds in specific examples by Roberts & Rosenthal (2011). The same polynomial form is recovered by Andrieu et al. (2022, Rem. 29) through weak Poincaré inequalities under $q(\omega^p) < \infty$ for some $p > 2$. In both regimes the tail behaviour of ω, encoded in the moments $q(\omega^p)$, is the central object governing convergence (Figure 1.2). Whether this regime arises depends on the tail of the proposal relative to the posterior, and it is flagged in practice by the Pareto-$\hat{k}$ diagnostic of Vehtari et al. (2017), refined in Vehtari et al. (2024); finite-sample bias and variance bounds for SNIS and IMH in this regime, under only a finite weight moment $q(\omega^p)$, are the subject of Chapter 2.[1]

[1] Sequential Monte Carlo (SMC) gives a third reading of the two regimes: SMC samplers (Del Moral et al., 2006; Doucet & Johansen, 2009) interpolate from q to π_n along a schedule of bridging distributions, alternating SNIS-like reweighting with π_n-invariant MCMC mutation (Chopin, 2002); the schedule trades weight-moment control against mutation ergodicity. Extending the weight-moment analysis to coupled SMC samplers is a follow-up direction not pursued here.

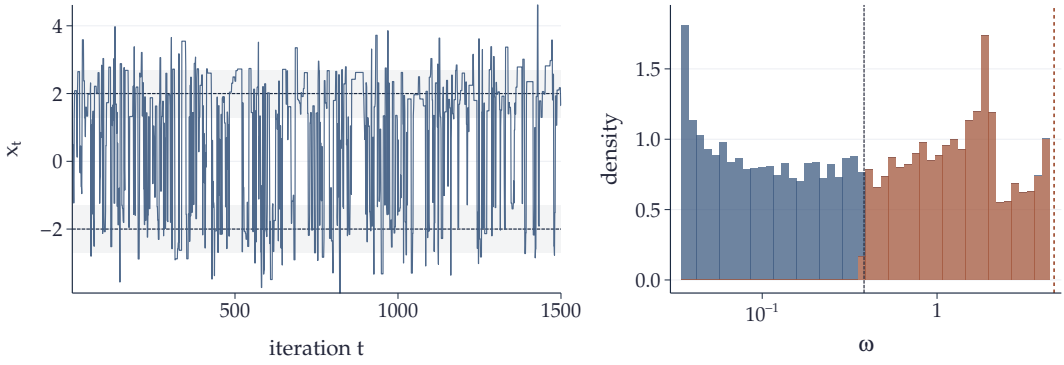


Figure 1.2: *Inside the importance mechanism. (a)* IMH sample path on the bimodal target. *(b)* Empirical distribution of $\omega(x) = \pi(x)/q(x)$ for $x \sim q$ (log axis); its right tail drives both the SNIS variance and the ergodicity of IMH.

1.3 Two Finite-Sample Biases

Although SNIS (1.7) and the IMH chain are both consistent as N and t go to infinity, each carries a finite-sample bias whose source and remedy differ qualitatively.

Initialisation bias of MCMC. At any finite time t, the marginal law $\mathcal{L}(x_t)$ of the IMH chain differs from the target π_n. Ergodicity guarantees that $|\mathcal{L}(x_t) - \pi_n|_{\mathrm{TV}} \to 0$ as $t \to \infty$ (Roberts & Rosenthal, 2004, Thm. 4), but the constants entering the rate cannot in general be evaluated in closed form; explicit upper bounds are available only for particular algorithms via minorisation–drift conditions (Rosenthal, 1995; Johnson, 1996; Jones & Hobert, 2001). The consequence is a systematic *initialisation bias*: for any initial law $\pi_0 \neq \pi_n$[2] and any π_n-integrable test function h,

$$\mathbb{E}_{\pi_0}\left[\frac{1}{t}\sum_{s=0}^{t-1} h(x_s)\right] - \pi_n(h) = \frac{1}{t}\sum_{s=0}^{t-1}\big(\pi_0 P^s(h) - \pi_n(h)\big) \qquad \text{for every finite } t, \tag{1.9}$$

and this quantity is non-zero whenever $\pi_0 \neq \pi_n$, and typically unknown (Atchadé & Jacob, 2026, §1.1). The bias vanishes as $t \to \infty$ and becomes a negligible part of the mean-squared error, which is dominated by the asymptotic variance $v(P, h)$ in the central limit theorem:

$$\sqrt{t}\left(\frac{1}{t}\sum_{s=0}^{t-1} h(x_s) - \pi_n(h)\right) \xrightarrow{d} \mathcal{N}\big(0,\, v(P,h)\big), \qquad t \to \infty, \tag{1.10}$$

where $v(P, h)$ is the asymptotic variance defined in Section 1.2 and the CLT holds under suitable ergodicity and moment conditions (Roberts & Rosenthal, 2004, §5). Despite vanishing

[2] The deterministic start $\pi_0 = \delta_{x_0}$ is a special case.

asymptotically, the initialisation bias matters at finite t and under parallel computation.

Parallelisation. Given M independent chains of length t each, define the cross-chain average

$$\bar{\pi}^{M,t}(h) := \frac{1}{Mt} \sum_{m=1}^{M} \sum_{s=0}^{t-1} h\big(x_s^{(m)}\big). \tag{1.11}$$

As $M \to \infty$ with t fixed,

$$\bar{\pi}^{M,t}(h) \longrightarrow \frac{1}{t} \sum_{s=0}^{t-1} \pi_0 P^s(h), \tag{1.12}$$

and this limit differs from $\pi_n(h)$ by precisely the right-hand side of (1.9): adding processors does not drive the bias to zero (Rosenthal, 2000; Atchadé & Jacob, 2026). To force the per-chain bias below $\varepsilon/2$ one must take $t = \mathcal{O}(\varepsilon^{-1})$, so the completion time of every parallel worker scales with the target accuracy rather than with M (Wang et al., 2024, §4.1). Discarding a burn-in segment of length b (Geyer, 1992; Gelman et al., 2014) reduces but does not remove the bias, and the burn-in length needed is application-dependent, ranging from dozens of iterations (Metropolis et al., 1953) to trillions (for sampling redistricting plans; McCartan & Imai, 2023). Bias removal at the estimator level changes the picture: if $\widehat{\pi}(h)$ is an unbiased and finite-variance estimator of $\pi_n(h)$, independent copies can be generated in parallel and averaged to give a consistent estimator at the standard Monte Carlo rate.

Composition. A residual bias of order N^{-1} is benign inside a single expectation but compounds under nested Monte Carlo. For a nested estimand

$$E = \mathbb{E}_1\big[\mathbb{E}_{2|1}[f(\Theta_1, \Theta_2) \mid \Theta_1]\big], \tag{1.13}$$

as in multiple imputation and experimental design (Rainforth et al., 2018a; Plummer, 2014), an inner estimator with bias $\mathcal{O}(N^{-\gamma})$ and variance $\mathcal{O}(N^{-1})$ attains, at the budget-optimal split, the mean-squared error

$$\mathcal{O}\big(C^{-2\gamma/(2\gamma+1)}\big)\ (\gamma < \infty), \qquad \mathcal{O}(C^{-1})\ (\gamma = \infty,\ \text{unbiased inner}). \tag{1.14}$$

For SNIS ($\gamma = 1$) this is $\mathcal{O}(C^{-2/3})$, strictly slower than $\mathcal{O}(C^{-1})$, so removing the inner bias improves the rate from $\mathcal{O}(C^{-2/3})$ to $\mathcal{O}(C^{-1})$, not merely the constant. Chapter 2 develops this nested-expectation application; the budget-allocation argument of Andradóttir & Glynn (2016) is given in Appendix D.6. The same point recurs for robust-mean estimators (Section 1.7), which deliver their finite-sample guarantees only on unbiased inputs.

Bias of self-normalised importance sampling. The SNIS estimator (1.7) is the ratio of two unbiased Monte Carlo averages, and the non-linearity introduces a finite-sample bias whose

leading-order asymptotic is

$$\lim_{N\to\infty} N \cdot \mathbb{E}\big[\widehat{\pi}_{\mathrm{IS}}^{N}(h) - \pi_n(h)\big] = -q\big(\omega^2(h - \pi_n(h))\big). \tag{1.15}$$

This formula appears in textbook form in Liu (2008, §2.5) but without a finite-sample proof; Chapter 2 (Section 2.2) supplies its first rigorous moment-level statement and proof, with the side-by-side comparison to the earlier derivations of Hesterberg (1988) deferred to Appendix D.7.

This same $\mathcal{O}(N^{-1})$ bias propagates wherever SNIS serves as a subroutine: it sets the nested-Monte-Carlo rate (1.14) and enters the gradient used to optimise importance-weighted variational bounds (Burda et al., 2016; Daudel et al., 2023).

Importance sampling versus independent Metropolis–Hastings. SNIS and IMH differ only in the accept–reject step. Write q_{SIR}^{N} for the law of the single point returned by sampling–importance resampling from N proposals $x_1, \dots, x_N \sim q$ (an index I drawn with probability proportional to $\omega(x_I)$), and qP^{N-1} for the law of the IMH chain after $N-1$ steps started from q. Comparing each to π_n in total variation, we show that, under $q(\omega^p) < \infty$, this accept–reject step lowers the finite-time bias by an order in N:

Independent Metropolis–Hastings reduces the importance-sampling bias

$$\big|q_{\mathrm{SIR}}^{N} - \pi_n\big|_{\mathrm{TV}} = \mathcal{O}\big(N^{-1}\big) \qquad \text{(SIR/SNIS bias)}, \tag{1.16}$$

$$\big|qP^{N-1} - \pi_n\big|_{\mathrm{TV}} = \mathcal{O}\big(N^{-(p-1)}\big) \qquad \text{(IMH bias)}. \tag{1.17}$$

Run for $t = N - 1$ iterations, IMH has bias $N^{-(p-1)}$: the rate $t^{-(p-1)}$ holds for $p > 1$ and is sharp up to logarithmic factors in an example (Chapter 2, Section 2.4). At equal budget, this bias decays strictly faster than the N^{-1} bias of SIR/SNIS once $p > 2$, a gain of $N^{-(p-2)}$. The hypothesis $p \geq 2$ is the standing assumption of the bias-removal results below. The idea of debiasing IMH by coupling two chains with a lag and applying the telescoping correction (Glynn & Rhee, 2014; Jacob et al., 2020) is due to Middleton et al. (2019), who introduced it for particle IMH under bounded weights; the same construction debiases SNIS and, via the common-random-numbers coupling of Section 1.4, extends to the unbounded-weight regime, where the resulting unbiased estimator has finite variance whenever $q(\omega^p) < \infty$ with $p \geq 2$.

Two questions for the analysis. The preceding analysis motivates two questions, taken up in Chapter 2:

Question 1. Best selection from a global proposal

Given N independent draws $x_1, \dots, x_N$ from q with $\pi_n \ll q$, what selection rule for one index $I \in \{1, \dots, N\}$ makes the chosen draw x_I as close to π_n as possible in total-variation distance?

Question 2. Unbiased self-normalised importance sampling

The self-normalised importance-sampling estimator is biased for the posterior expectation $\pi_n(h)$. Is there a finite-time procedure that returns an unbiased estimator of $\pi_n(h)$?

The IMH chain is a natural answer to Question 1, to which Chapter 2 gives a partial answer, while its coupling-based construction answers Question 2.

1.4 Couplings, Meeting Times, and Unbiased MCMC

The unbiased Markov-chain Monte Carlo (UMCMC) framework (Glynn & Rhee, 2014; Jacob et al., 2020; Atchadé & Jacob, 2026) removes the initialisation bias of Section 1.3 by coupling two copies of the chain so that the resulting estimator has expectation exactly $\pi_n(h)$, giving exact estimation without exact simulation from π_n (Glynn, 2016).

A *coupling* of two Markov chains $(x_t, y_t)_{t\geq 0}$ on Θ is a joint process driven by a transition kernel $\bar{P}$ whose marginals coincide with the prescribed kernel P and which, after an almost surely finite *meeting time* τ, takes values on the diagonal $\{x_t = y_t\}$ for all $t \geq \tau$. The textbook account is Lindvall (2002, Chapters I–II). The associated *coupling inequality*, due to Doeblin (1938) and recalled in Lindvall (2002, Ch. I), controls the total-variation distance to stationarity by the tail of τ:

$$|\mathcal{L}(x_t) - \pi_n|_{\mathrm{TV}} \leq \mathbb{P}(\tau > t), \qquad t \geq 0. \tag{1.18}$$

Coupling was first used to diagnose MCMC convergence by Johnson (1996, 1998); the toolkit now yields non-asymptotic total-variation diagnostics (Biswas et al., 2019) and, by using coupled chains to solve the Poisson equation, unbiased estimators of the asymptotic variance $v(P, h)$ (Douc et al., 2026); see Atchadé & Jacob (2026, §3) for a review.

Designing the coupling. Implementing UMCMC reduces to designing $\bar{P}$ with the correct marginals and a small meeting time τ (Atchadé & Jacob, 2026, §4). This thesis considers the *common-random-numbers (CRN) coupling of IMH*, which drives both chains from a single proposal $y \sim q$ and a single uniform. Chapter 2 shows that it saturates the coupling inequality (1.18) at every finite time, hence is *all-time maximal* (Griffeath, 1975; Pitman, 1976), the

only such example known for an MCMC algorithm, and bounds its meeting-time tail under moment conditions on ω.

The coupled-chains unbiased estimator. Glynn & Rhee (2014) showed that a telescoping series for a stationary expectation $\pi_n(h)$ may be debiased by *randomised truncation* at a level drawn independently of the chain. The variant used throughout this thesis avoids that external truncation: following Jacob et al. (2020, Prop. 1), run two chains (x_t) and (y_t) with a lag $L \geq 1$, coupled so that they meet at the time $\tau = \inf\{t > L : x_t = y_{t-L}\}$ and coincide thereafter; the meeting itself, rather than an independent draw, terminates the correction. For any burn-in index $k \geq 0$, the coupled-chain estimator averages $h(x_k)$ and the correction differences $h(x_\ell) - h(y_{\ell-L})$ up to meeting,

$$\widehat{\pi}_{n,\mathrm{u}}(h) = h(x_k) + \sum_{\ell=k+1}^{\tau-1} \big[\, h(x_\ell) - h(y_{\ell-L}) \,\big], \tag{1.19}$$

then satisfies

$$\mathbb{E}\big[\widehat{\pi}_{n,\mathrm{u}}(h)\big] = \pi_n(h), \tag{1.20}$$

exactly, without any burn-in correction and for any $k \geq 0$, $L \geq 1$, provided h is π_n-integrable and $\mathbb{E}[\tau] < \infty$ (Jacob et al., 2020, Prop. 1). The precise CRN coupling, the moment conditions securing $\mathbb{E}[\tau] < \infty$, and the unbiasedness argument are developed in Chapter 2 (Section 2.5). Jacob et al. (2020) turned this construction into a practical coupled-MCMC methodology (reviewed in Atchadé & Jacob, 2026, §2–4). Here the interest is bias elimination rather than the parallelisation that motivated unbiased MCMC in other settings, as SNIS is already parallel by construction. As discussed in Section 1.3, the construction removes the SNIS bias, which would otherwise propagate through nested expectations, and yields i.i.d. unbiased copies of $\pi_n(h)$ that compose directly with downstream procedures such as robust mean estimation.

The subgeometric regime and the role of ω. Under geometric ergodicity τ has exponential tails and $\widehat{\pi}_{n,\mathrm{u}}(h)$ has finite variance (Jacob et al., 2020, §3). In the subgeometric regime relevant to this thesis, an IMH chain with unbounded ω is at most polynomially ergodic (Mengersen & Tweedie, 1996, Thm. 2.1), and the meeting time of the coupled construction inherits the same polynomial decay (Figure 1.3), so that a finite weight moment $q(\omega^p)$ yields a corresponding moment bound on the meeting time τ, and, once $p \geq 2$, a finite-variance guarantee for $\widehat{\pi}_{n,\mathrm{u}}(h)$. Carrying out this transfer to derive non-asymptotic guarantees for SNIS and IMH under unbounded weights is the content of Chapter 2.

Unbiased MCMC and unbiased MLMC. The coupling construction belongs to a wider family of unbiased estimators built by randomised truncation of an infinite series; a parallel strand, randomised multilevel Monte Carlo (rMLMC) (McLeish, 2011; Rhee & Glynn, 2015;

Blanchet & Glynn, 2015; Vihola, 2018; Blanchet et al., 2019), debiases multilevel constructions (Giles, 2008, 2015) by the same mechanism. A shared caveat: naive parallelisation by retaining the first C of $M \gg C$ workers reintroduces an *impatience bias* (Atchadé & Jacob, 2026, §2.5); the cost–completion-time trade-off is analysed by Wang et al. (2024, §2).

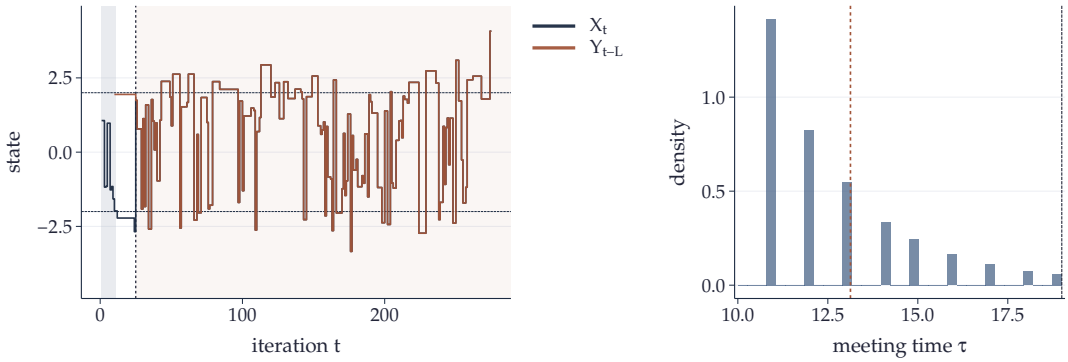


Figure 1.3: *L-lag coupled IMH for the target and proposal of Figure 1.2 with* $L = 10$. *(a)* Trajectories of two chains coupled by common random numbers, meeting at $\tau = \inf\{t > L : x_t = y_{t-L}\}$. *(b)* Histogram of τ over 5,000 runs; under $q(\omega^p) < \infty$, the tail decays polynomially and $\mathbb{E}[\tau] < \infty$ secures unbiasedness of $\widehat{\pi}_{n,\mathrm{u}}(h)$.

Maximal coupling and finite-time bias. Drawn with common random numbers, the coupling of independent Metropolis–Hastings is *all-time maximal*, the only such example known for an MCMC kernel: Chapter 2 shows that the coupling inequality (1.18) holds with equality at every finite time.

An all-time maximal coupling

$$|\mathcal{L}(x_t) - \pi_n|_{\mathrm{TV}} = \mathbb{P}(\tau > t), \qquad t \geq 1.$$

The meeting-time tail is therefore exactly the total-variation distance to π_n. Under a finite weight moment $q(\omega^p) < \infty$ this yields, for $p > 1$, the polynomial convergence rate $|\mathcal{L}(x_t) - \pi_n|_{\mathrm{TV}} \lesssim t^{-(p-1)}$, faster than the $\mathcal{O}(N^{-1})$ bias of self-normalised importance sampling, so independent Metropolis–Hastings has the smaller finite-time bias once $p > 2$. The same coupling makes the coupled-chain estimator $\widehat{\pi}_{n,\mathrm{u}}(h)$ of (1.19) unbiased with finite variance as soon as $p \geq 2$, at finite expected cost and with efficiency matching importance sampling, giving unbiased copies of $\pi_n(h)$ for nested expectations and stochastic-gradient pipelines. These results occupy Sections 2.3, 2.4, 2.2 and 2.5 of Chapter 2.

1.5 Variational Approximation of the Posterior

The methods of Sections 1.2–1.4 approximate π_n by sampling from it. A second family takes a different route: rather than target π_n directly, it returns the closest approximation to it within a tractable family of distributions. Variational inference (VI) approximates the posterior π_n by recasting Bayesian inference as an optimisation problem: a tractable family $\mathcal{F} \subset \mathcal{M}_1^+(\Theta)$ of candidate distributions is fixed, and the closest member $\widehat{\rho}_n \in \mathcal{F}$ in Kullback–Leibler divergence is selected,

$$\widehat{\rho}_n = \operatorname*{argmin}_{\rho \in \mathcal{F}} \mathrm{KL}(\rho \,\|\, \pi_n)\,. \tag{1.21}$$

The output $\widehat{\rho}_n$ is a deterministic parametric measure that can be reused at test time without further sampling. The framework is reviewed in Blei et al. (2017, §1), who position VI alongside Monte Carlo as "a valuable tool, alongside MCMC, in the statistician's toolbox": sampling and optimisation are complementary approaches to the same problem, with VI suited to large data sets and rapid model exploration and MCMC preferred when precise samples are needed at moderate scale. The framing has a classical origin: Bayes' rule itself solves a variational problem. The Donsker–Varadhan identity (Donsker & Varadhan, 1975; Csiszár, 1975), brought into the Bayesian fold by Zellner (1988), characterises π_n as the unconstrained minimiser of $-\mathbb{E}_\rho[\ell_n(\theta)] + \mathrm{KL}(\rho \,\|\, \pi)$ over all of $\mathcal{M}_1^+(\Theta)$, derived via Donsker–Varadhan in Section 1.6 (eq. (1.25)). The variational posterior (1.21) solves the same minimisation over a smaller domain, the tractable family $\mathcal{F} \subset \mathcal{M}_1^+(\Theta)$, so Bayes and VI differ only in the feasible set. Figure 1.4 summarises the specialisations of this unified variational template covered in the rest of the chapter.

Bayes' rule (1.1) yields the identity

$$\mathrm{KL}\left(\rho \,\|\, \pi_n\right) = \log Z_n - \mathrm{ELBO}(\rho), \tag{1.22}$$

where the *evidence lower bound* (ELBO) is

$$\mathrm{ELBO}(\rho) = \mathbb{E}_{\theta\sim\rho}\big[\ell_n(\theta)\big] - \mathrm{KL}(\rho \,\|\, \pi). \tag{1.23}$$

Since $\log Z_n$ does not depend on ρ, problem (1.21) is equivalent to maximising the ELBO (Blei et al., 2017, §2.2), and the unknown normalising constant Z_n cancels by construction; the resulting objective is amenable to stochastic optimisation (Hoffman et al., 2013, §3), with reparameterised (Kingma & Welling, 2014; Rezende et al., 2014) or score-function (Paisley et al., 2012; Ranganath et al., 2014) gradient estimators making the procedure applicable to large datasets and high-dimensional models under mini-batching.

The variational family $\mathcal{F}$ encodes the trade-off between expressiveness and tractability and ranges from the factorised *mean-field* class (Bishop, 2006, §10.1), through full-covariance

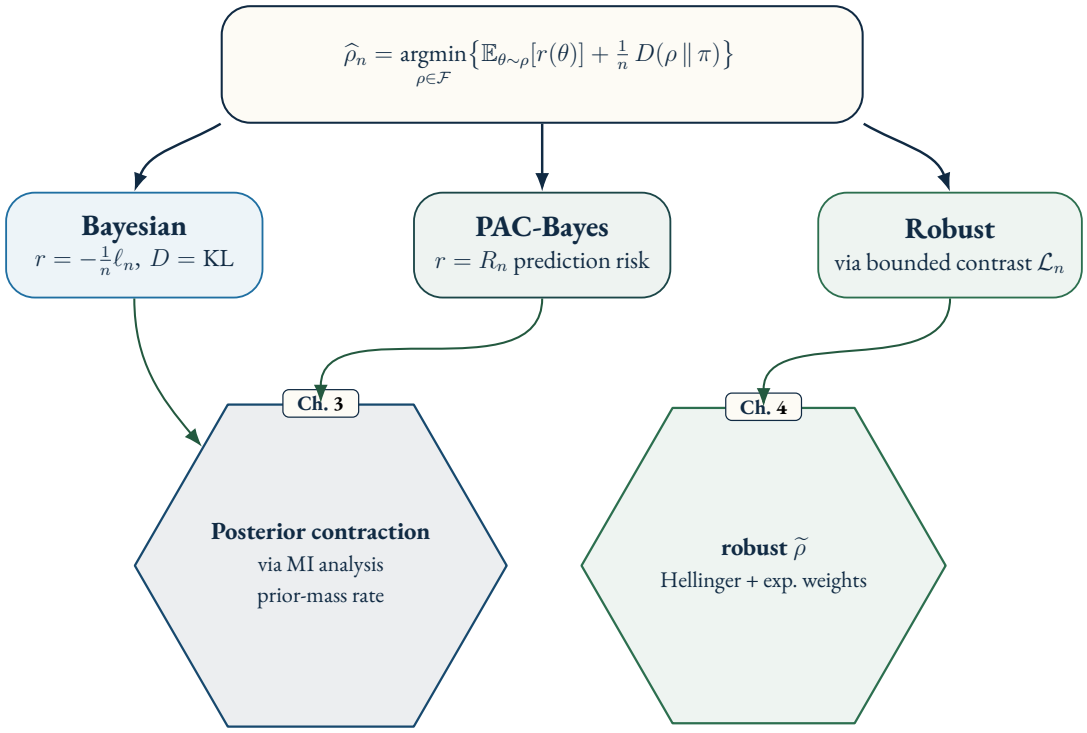


Figure 1.4: *Variational zoom.* One variational problem with three specialisations: Bayesian (likelihood-based), PAC-Bayes (prediction-risk-based), and robust (loss-based statistical contrast $\mathcal{L}_n$). Chapters 3–4 certify and robustify the variational optimum.

Gaussians, structured factorisations (Wainwright & Jordan, 2008, §5) and mixtures, to neural-network parametrisations: normalising flows (Rezende & Mohamed, 2015; Papamakarios et al., 2021, §3) and amortised inference networks (Kingma & Welling, 2014; Margossian & Blei, 2024, §2.3). Such families also drive the variational training of Bayesian neural networks, whose methodology and open questions are surveyed by Arbel et al. (2024). The variational gap $\inf_{\rho\in\mathcal{F}} \mathrm{KL}\,(\rho \,\|\, \pi_n)$ is the irreducible approximation error of restricting ρ to $\mathcal{F}$; enlarging $\mathcal{F}$ shrinks it at the cost of more parameters, and drives it to zero for families dense in $\mathcal{M}_1^+(\Theta)$ such as normalising flows. The generalisation error of $\widehat{\rho}_n$ is in turn controlled by the PAC-Bayes inequalities of Section 1.6.

Two sources of approximation error. Two distinct biases limit any variational fit. First, the family $\mathcal{F}$ is restricted: a mean-field Gaussian, for instance, cannot represent posterior correlations or multimodality. Second, the KL objective in (1.21) is asymmetric and *mode-seeking* (Bishop, 2006; Minka, 2005), penalising mass placed where π_n has none but not failure to cover its tails. The optimum thus concentrates on a single mode and underestimates posterior variance (Figure 1.5). The consequence for this thesis is direct: an overly concentrated $\widehat{\rho}_n$, deployed as the proposal q in an IMH sampler, induces an importance weight ω with heavy tails; the chain (x_t) loses uniform ergodicity, returning to the polynomially-ergodic regime of Chapter 2.

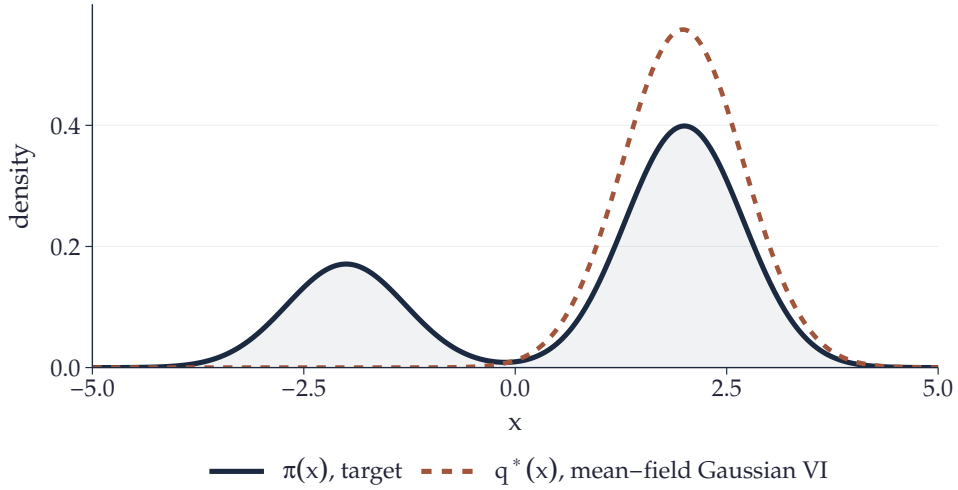


Figure 1.5: *Mean-field Gaussian VI on the bimodal target.* The optimum $\widehat{\rho}_n = \operatorname{argmin}_{\rho \in \mathcal{F}_{\text{Gauss}}} \mathrm{KL}(\rho \,\|\, \pi_n)$ (forest dashed) collapses onto the heavier mode, the mode-seeking bias addressed in Chapters 3 and 4.

Importance-weighted ELBOs and their gradient variance. Tighter lower bounds on $\log Z_n$ exist, most notably the importance-weighted bound of Burda et al. (2016), the logarithm of a K-sample importance estimator of Z_n that converges to $\log Z_n$ as $K \to \infty$. Its tightness improves with K, while the variance of its reparameterised gradient is governed by the moments of the importance weights ω (Rainforth et al., 2018b, §3). The unbounded-weight bias bounds for SNIS and IMH proved in Chapter 2 bear directly on the gradient quality of importance-weighted variational training.

Diagnostics and reliability. Diagnostics for black-box VI, such as the Pareto-$\hat{k}$ statistic of Yao et al. (2018); Vehtari et al. (2024) that estimates the tail index of ω under the proposal q, flag when $\widehat{\rho}_n$ is too far from π_n to be trusted but do not provide finite-sample concentration guarantees on the variational posterior. For the latter one turns to PAC-Bayesian theory.

1.6 PAC-Bayes, MI, and Posterior Certification

PAC-Bayesian theory is one of the analytical tools of this thesis. The variational posterior $\widehat{\rho}_n$ of (1.21), the posterior π_n, the fractional posterior $\pi_{n,\alpha} \propto e^{\alpha \ell_n} \pi$ ($\alpha \in (0, 1]$), the coherent generalised updates of Bissiri et al. (2016) (the loss-based *Gibbs posteriors* $\rho_n \propto e^{-\eta \mathcal{L}_n} \pi$, studied in PAC-Bayes and M-estimation; see Catoni, 2007; Jiang & Tanner, 2008), and the bounded-contrast posteriors of Section 1.7 are all data-dependent probability measures on Θ. PAC-Bayes bounds apply uniformly to every such measure: with high probability, the true risk of any data-dependent posterior is at most its empirical risk plus a complexity penalty, a divergence to the prior scaled by $1/n$. This section assembles such a bound and uses it to certify $\widehat{\rho}_n$.

A brief history. PAC-Bayes originated in supervised learning with Shawe-Taylor & Williamson (1997) and McAllester (1998, 1999); we refer to Alquier (2024) for its general theory. The localisation technique of Catoni (2007) underlies the rates of Chapter 3. Alquier et al. (2016) first analysed variational PAC-Bayes for possibly unbounded prediction risks, after which Bhattacharya et al. (2019) and Alquier & Ridgway (2020) extended the machinery to fractional posteriors. The information-theoretic bounds of Russo & Zou (2016); Xu & Raginsky (2017) replace the Kullback–Leibler complexity term by a mutual information, the object sharpened in Chapter 3.

Two features recur below: the prior π need not be the Bayesian prior but serves only as a complexity reference, so the bounds reach beyond Bayesian inference to the generalised updates of Bissiri et al. (2016) (Section 1.7); and all of it rests on one change-of-measure identity, recalled next.

The change-of-measure identity. For a probability measure π on $(\Theta, \mathcal{B})$ and a measurable $f : \Theta \to \mathbb{R}$ with $\int e^f \, d\pi < \infty$, the Donsker–Varadhan formula (Donsker & Varadhan, 1975) states

$$\log \int_\Theta e^{f(\theta)} \, \pi(d\theta) = \sup_{\rho \in \mathcal{M}_1^+(\Theta)} \Big\{ \mathbb{E}_\rho[f(\theta)] - \mathrm{KL}(\rho \,\|\, \pi) \Big\}, \tag{1.24}$$

attained at the Gibbs measure $\rho^\star(d\theta) \propto e^{f(\theta)} \, \pi(d\theta)$ (Catoni, 2007, Prop. 1.4.2). The PAC-Bayes bounds developed in this thesis are obtained through this route: a data-dependent $f = f(\theta; \mathcal{S})$ is chosen, and the log-MGF $\log \mathbb{E}_\pi[e^f]$ is controlled in probability over the sample $\mathcal{S}$. The Kullback–Leibler change of measure is, however, only one choice among several: other divergences yield related PAC-Bayes bounds, such as those based on the Rényi divergence (Bégin et al., 2016) or on general f-divergences (Alquier & Guedj, 2018), a freedom made explicit in the generalised-Bayes viewpoint of Knoblauch et al. (2022), where the divergence is a modelling choice.

The posterior and its variational approximation. Taking $f = \ell_n$ in (1.24) identifies the posterior with the minimiser of the Gibbs functional, the variational characterisation already introduced in Section 1.5,

$$\pi_n = \operatorname*{argmin}_{\rho \in \mathcal{M}_1^+(\Theta)} J(\rho), \qquad J(\rho) := -\mathbb{E}_\rho[\ell_n(\theta)] + \mathrm{KL}(\rho \,\|\, \pi), \tag{1.25}$$

with $J(\pi_n) = -\log Z_n$, and PAC-Bayes studies this problem uniformly in ρ (Germain et al., 2016, Thm. 3). Restricting J to a tractable family $\mathcal{F} \subset \mathcal{M}_1^+(\Theta)$ gives the variational posterior

$$\widehat{\rho}_n = \operatorname*{argmin}_{\rho \in \mathcal{F}} J(\rho), \tag{1.26}$$

which coincides with (1.21) after cancelling $\log Z_n$. Sampling targets the exact minimiser π_n, while variational inference settles for the closest member of the tractable family $\mathcal{F}$; the two solve the same optimisation and differ only in the set over which J is minimised. This shared optimisation is the thread linking the samplers of Chapter 2 to the variational methods of Chapters 3–4.

The fractional posterior. Applied to $\alpha\,\ell_n$ in place of ℓ_n, the same argument identifies the *fractional* (or *tempered*) *posterior*

$$\pi_{n,\alpha}(d\theta) \propto e^{\alpha\,\ell_n(\theta)}\,\pi(d\theta), \qquad \alpha \in (0,1], \tag{1.27}$$

introduced for consistency analysis by Walker & Hjort (2001), given an information-theoretic treatment by Zhang (2006a), and developed for posterior contraction by Bhattacharya et al. (2019). By the same Donsker–Varadhan argument,

$$\pi_{n,\alpha} = \operatorname*{argmin}_{\rho \in \mathcal{M}_1^+(\Theta)} \left\{ -\alpha\,\mathbb{E}_\rho[\ell_n(\theta)] + \mathrm{KL}(\rho \,\|\, \pi) \right\}, \tag{1.28}$$

and $\widetilde{\rho}_{n,\alpha}$ denotes the restriction of that minimisation to $\mathcal{F}$. The inverse temperature α rebalances fit against prior mass: for $\alpha < 1$, Bhattacharya et al. (2019) and Alquier & Ridgway (2020) establish prior-mass contraction for $\pi_{n,\alpha}$ and its variational restriction $\widetilde{\rho}_{n,\alpha}$ without a testing condition. At $\alpha = 1$, a parallel line outside the PAC-Bayes route studies the standard variational posterior under Ghosal-type prior-mass and testing assumptions (Zhang & Gao, 2020). The comparison of these rates with the bound obtained in this thesis is made in Chapter 3.

Setting and risks. A *randomised estimator* is a data-dependent probability measure $\widehat{\rho} = \widehat{\rho}(\mathcal{S})$ on Θ, a single draw $\widehat{\theta} \sim \widehat{\rho}$ being the estimator itself; we suppress the conditioning on $\mathcal{S}$ throughout this section. Fix a measurable loss $\ell : \Theta \times \mathbb{X} \to \mathbb{R}$, in general distinct from the log-likelihood ℓ_n. We treat the bounded case $\ell \in [0,1]$ first, where the bounds are cleanest, and turn to the unbounded log-likelihood-ratio loss below; the unbounded losses in between are covered under Hoeffding- or Bernstein-type exponential-moment conditions on the centred risk $R - R_n$ (Alquier et al., 2016, Defs. 2.3–2.4). For a fixed loss ℓ, write the empirical and population risks

$$R_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, X_i), \qquad R(\theta) := \mathbb{E}_{X\sim P}[\ell(\theta, X)]\,, \tag{1.29}$$

so that the generalisation gap is $\mathrm{gen}(\widehat{\theta}, \mathcal{S}) := R(\widehat{\theta}) - R_n(\widehat{\theta})$. We write $\mathbb{E}_{\mathcal{S}}$ for the expectation over $\mathcal{S} \sim P^{\otimes n}$ and $\mathbb{E}_{\widehat{\theta}\sim\widehat{\rho}}$ for the expectation over $\widehat{\theta}$ given $\mathcal{S}$.

The PAC-Bayes in-expectation bound. For a bounded loss $\ell : \Theta \times \mathbb{X} \to [0,1]$, every data-independent prior π on Θ, and every randomised estimator $\widehat{\rho} \ll \pi$,

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\widehat{\theta}\sim\widehat{\rho}}\Big[\operatorname{gen}(\widehat{\theta},\mathcal{S})\Big] \le \sqrt{\frac{\mathbb{E}_{\mathcal{S}}[\mathrm{KL}(\widehat{\rho}\,\|\,\pi)]}{2n}}, \tag{1.30}$$

where $\operatorname{gen}(\widehat{\theta},\mathcal{S}) := R(\widehat{\theta}) - R_n(\widehat{\theta})$ is the generalisation gap defined in (1.29); this is the Catoni in-expectation bound (Alquier, 2024, Thm. 2.1).

Mutual information and the PAC-Bayes bridge. The *mutual information* $\mathcal{I}(U;V) := \mathrm{KL}(Q_{U,V} \,\|\, Q_U \otimes Q_V)$ of two random variables, with joint law $Q_{U,V}$ and marginals Q_U, Q_V, quantifies their statistical dependence (Cover & Thomas, 2006, Ch. 2). Applied to the parameter–sample pair $(\widehat{\theta},\mathcal{S})$ with $\widehat{\theta} \mid \mathcal{S} \sim \widehat{\rho}$ and disintegrated along $\mathcal{S}$, it equals the expected divergence $\mathbb{E}_{\mathcal{S}}[\mathrm{KL}(\widehat{\rho}\,\|\,\mathbb{E}_{\mathcal{S}}[\widehat{\rho}])]$ between the data-dependent posterior $\widehat{\rho}$ and its sample-averaged marginal. For σ-sub-Gaussian losses, Russo & Zou (2016, Prop. 1) give

$$\Big|\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\widehat{\theta}\sim\widehat{\rho}}\Big[\operatorname{gen}(\widehat{\theta},\mathcal{S})\Big]\Big| \le \sigma\sqrt{\frac{2\,\mathcal{I}(\widehat{\theta};\mathcal{S})}{n}}; \tag{1.31}$$

for the bounded case $\ell \in [0,1]$, Hoeffding's lemma gives $\sigma = 1/2$ and recovers the familiar form $\sqrt{\mathcal{I}(\widehat{\theta};\mathcal{S})/(2n)}$. The same bound holds for general supervised learning (Xu & Raginsky, 2017, Thm. 1), with data-dependent refinements in Negrea et al. (2019).

The link to PAC-Bayes is exact. Write $\pi^\star = \mathbb{E}_{\mathcal{S}}[\widehat{\rho}]$. For any prior π the chain rule gives

$$\mathbb{E}_{\mathcal{S}}[\mathrm{KL}(\widehat{\rho}\,\|\,\pi)] = \mathcal{I}(\widehat{\theta};\mathcal{S}) + \mathrm{KL}(\pi^\star \,\|\, \pi),$$

so the expected KL penalty is minimised over π at the data-independent marginal $\pi = \pi^\star$, where it equals $\mathcal{I}(\widehat{\theta};\mathcal{S})$ (Catoni, 2007, pp. 14, 51). Hence the mutual-information bound (1.31) is the in-expectation PAC-Bayes bound (1.30) evaluated at the best (oracle) prior $\pi^\star$; any fixed π loosens it by $\mathrm{KL}(\pi^\star \,\|\, \pi) \ge 0$, the price of choosing the prior before seeing the sample (Figure 1.6). This identity is the starting point for the mutual-information bound of Chapter 3 (Section 3.3). Lugosi & Neu (2022, Thm. 1) recover both families from a single master bound via convex duality; Hellström et al. (2025) survey the lineage.

One analysis for sampling and variational inference. Both (1.30) and (1.31) hold for any data-dependent $\widehat{\rho} \ll \pi$, so the same inequality certifies the Gibbs and fractional posteriors π_n and $\pi_{n,\alpha}$ and, through the constrained-Gibbs identity (1.26), their variational approximation $\widehat{\rho}_n$, with no separate argument for each. Since the two differ only in the feasible set of (1.25)–(1.26), the certificate is blind to that difference: even the crude, mode-seeking variational fit

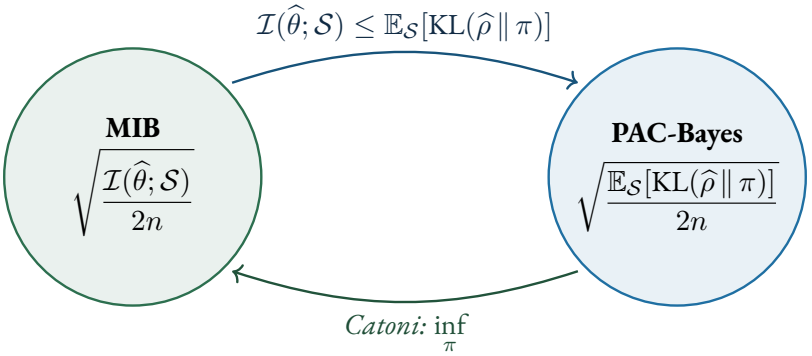


Figure 1.6: *MIB–PAC-Bayes bridge.* $\le$ holds for every fixed prior; $\inf_\pi$ in the opposite direction recovers the MIB at $\pi^\star = \mathbb{E}_{\mathcal{S}}[\widehat{\rho}]$. Adapted from P. Alquier, *PAC-Bayes: an introduction*, lecture slides.

of Figure 1.5 carries the same finite-sample generalisation guarantee as the posterior it approximates. The black-box diagnostics of Section 1.5 only signal when such a fit is unreliable; the PAC-Bayes bound certifies it.

From statistical learning to statistical inference. The in-expectation bound (1.30) is stated for a bounded loss $\ell \in [0,1]$; in statistical inference the natural loss is the per-observation negative log-likelihood ratio

$$\ell(\theta, X_i) := \log\!\left(\frac{p^*(X_i)}{p_\theta(X_i)}\right), \qquad R(\theta) = \mathrm{KL}(P^* \,\|\, P_\theta), \tag{1.32}$$

whose population risk $R(\theta) = \mathbb{E}_{X\sim P^*}[\ell(\theta, X)]$ is the Kullback–Leibler divergence (Alquier & Ridgway, 2020, §2); write $r_n(\theta) = \sum_{i=1}^n \ell(\theta, X_i)$ for the cumulative log-likelihood ratio. This loss is generally unbounded, so the Russo–Zou MIB (1.31) does not apply directly. Alquier & Ridgway (2020, Thm. 2.1) resolve the boundedness issue by replacing the population KL with the α-Rényi divergence $\mathcal{D}_\alpha$ for $\alpha \in (0,1)$: under a single prior-mass condition of the form $-\log \pi(B(\varepsilon_n)) \le n\,\varepsilon_n$, the fractional posterior $\pi_{n,\alpha}$ of (1.27) contracts at the prior-mass rate in $\mathcal{D}_\alpha$, with squared Hellinger recovered at $\alpha = 1/2$ (Alquier & Ridgway, 2020, Rem. 2.2).

A mutual-information bound beyond bounded losses. For the unbounded log-likelihood-ratio loss (1.32) the Russo–Zou bound (1.31) is unavailable. The benchmark guarantee is the Alquier & Ridgway (2020, Thm. 2.6) PAC-Bayes bound: for every data-dependent $\widehat{\rho} \ll \pi$ and every $\alpha \in (0,1)$,

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\widehat{\rho}}\big[\mathcal{D}_\alpha(P_\theta \,\|\, P^*)\big] - \frac{\alpha}{n(1-\alpha)}\,\mathbb{E}_{\theta\sim\widehat{\rho}}\big[r_n(\theta)\big]\right] \le \frac{\mathbb{E}_{\mathcal{S}}[\mathrm{KL}(\widehat{\rho}\,\|\,\pi)]}{n(1-\alpha)}, \tag{1.33}$$

its complexity term the divergence $\mathbb{E}_{\mathcal{S}}[\mathrm{KL}(\widehat{\rho}\,\|\,\pi)]$ to a prior π fixed before the sample is seen.

This is not a mutual-information bound: the sub-Gaussian argument behind (1.31) fails for the unbounded loss. Chapter 3 establishes the first mutual-information bound for statistical inference, minimising the complexity term over the prior, which replaces $\mathbb{E}_{\mathcal{S}}[\mathrm{KL}(\widehat{\rho} \,\|\, \pi)]$ by the mutual information $\mathcal{I}(\theta;\mathcal{S}) = \min_{\pi} \mathbb{E}_{\mathcal{S}}[\mathrm{KL}(\widehat{\rho} \,\|\, \pi)]$, attained at the oracle marginal $\pi^\star = \mathbb{E}_{\mathcal{S}}[\widehat{\rho}]$.

Mutual-information bound for statistical estimation

For every data-dependent posterior $\widehat{\rho} \ll \pi$ and every $\alpha \in (0,1)$,

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\widehat{\rho}}\big[\mathcal{D}_\alpha(P_\theta \,\|\, P^*)\big] - \frac{\alpha}{n(1-\alpha)}\,\mathbb{E}_{\theta\sim\widehat{\rho}}\big[r_n(\theta)\big]\right] \le \frac{\mathcal{I}(\theta;\mathcal{S})}{n(1-\alpha)}.$$

This is harder to prove than the bounded-loss bound (1.31), whose bounded generalisation gap Hoeffding's lemma controls at once. The log-likelihood ratio is unbounded, with possibly infinite moments, so that argument fails; what rescues the bound is that $\mathcal{I}(\theta;\mathcal{S})$ stays finite even when the population risk $R(\theta) = \mathrm{KL}(P^* \,\|\, P_\theta)$ does not, and controlling it is the more delicate analysis of Section 3.3.

Combined with the localisation device of Catoni (2007, Ch. 4.5), under a divergence-equivalence condition on the model and a prior-mass condition with exponent $\kappa_\pi \in (0,1]$ and constant d_π, the bound yields a common convergence theorem for the fractional posterior $\pi_{n,\alpha}$ and its variational restriction $\widetilde{\rho}_{n,\alpha}$:

Convergence rate for fractional and variational posteriors

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi_{n,\alpha}}\big[\mathrm{KL}(P^* \,\|\, P_\theta)\big] \lesssim \inf_{\tilde{\theta}\in\Theta} \mathrm{KL}(P^* \,\|\, P_{\tilde{\theta}}) + \frac{d_\pi}{n^{\kappa_\pi}}.$$

Minimising over the prior yields a sharper rate. Specialised to the fractional posterior, the complexity term of (1.33) is of order $d_\pi \log(n)/n$; this minimisation removes the logarithmic factor and leaves d_π/n when $\kappa_\pi = 1$, so the mutual-information bound improves the $d_\pi \log(n)/n$ rate of Bhattacharya et al. (2019); Alquier & Ridgway (2020) to the parametric d_π/n. The exponent $\kappa_\pi < 1$ matches the nonparametric prior-mass rates of Ghosal et al. (2000, Thm. 2.1), and the same bound governs the posterior, the fractional posterior, its variational restriction along the optimisation path, and the maximum-likelihood estimator.

1.7 Robust Posteriors under Misspecification

The guarantees of Section 1.6 presume a well-specified model. Under misspecification, i.e. when the data-generating distribution P^* lies outside $\{P_\theta : \theta \in \Theta\}$, the posterior π_n fails two stability properties: Grünwald & van Ommen (2017, §§1–2) exhibit models in which mild

misspecification drives π_n to concentrate on parameter values arbitrarily far from any sensible target as $n \to \infty$, and Owhadi et al. (2015, Thm. 1.1, Thm. 5.1) establish the complementary brittleness statement: arbitrarily small total-variation perturbations of the model produce arbitrarily large changes in posterior expectations. A mode-seeking variational fit (Section 1.5) inherits both defects. The fragility is structural: Bayes' rule minimises $\mathrm{KL}(P^* \,\|\, P_\theta)$, which is $+\infty$ as soon as P_θ vanishes on a P^*-positive event, whereas frequentist robustness asks only for continuity in total variation (Huber, 1981, Ch. 1, §§1.3, 4), a strictly weaker condition.

Generalised posteriors and bounded contrasts. Robustness can be achieved by several routes; the one developed in this thesis replaces the unbounded log-likelihood by a *bounded* statistical contrast and updates beliefs through a generalised posterior. Bissiri et al. (2016) derive such generalised posteriors axiomatically through an update map Ψ that takes a data-dependent loss ℓ on Θ and a prior π to a posterior $\Psi(\ell, \pi)$, the single requirement being a coherence property (Bissiri et al., 2016, eq. (3)): updating with two losses ℓ_1 and ℓ_2 in succession must agree with updating once by their sum,

$$\Psi\big(\ell_2,\ \Psi(\ell_1, \pi)\big) = \Psi\big(\ell_1 + \ell_2,\ \pi\big). \tag{1.34}$$

For an i.i.d. sample and a per-observation loss, iterating (1.34) forces the cumulative loss $\mathcal{L}_n(\theta) = \sum_{i=1}^n \ell(\theta, X_i)$ to be additive, and the unique coherent update at learning rate $\eta > 0$ is the Gibbs posterior $\rho_n \propto \exp(-\eta\, \mathcal{L}_n)\, \pi$ of Section 1.6 (Bissiri et al., 2016, eq. (8)), here taking ℓ a bounded statistical contrast rather than the negative log-likelihood. Taking $\mathcal{L}_n$ bounded by construction caps the influence of any single observation: the per-observation contrast $\mathcal{L}_n/n$ remains bounded under arbitrary contamination, so the PAC-Bayes apparatus of Section 1.6 applies through Hoeffding-type log-MGF bounds. The ρ-posteriors of Baraud & Birgé (2020), introduced in (1.36) below, build their bounded contrast from Hellinger affinities but lie outside this framework: their contrast is a supremum of a sum, not a sum, so (1.34) cannot be iterated across observations. The construction rests instead on the Hellinger-affinity oracle inequality of Baraud & Birgé (2018, 2020) and is the focus of Chapter 4.

Hellinger as the contamination metric. The robustness of a generalised posterior is inherited from the discrepancy its contrast measures, and not every discrepancy will do. The Kullback–Leibler divergence is unsuitable: $\mathrm{KL}(P^* \,\|\, P_\theta)$ is infinite as soon as the true law P^* charges a set on which P_θ vanishes, so a single contaminating outlier can drive the Bayes posterior arbitrarily far. The squared Hellinger distance

$$\mathcal{H}^2(P, Q) = \tfrac{1}{2} \int \big(\sqrt{dP/d\mu} - \sqrt{dQ/d\mu}\big)^2\, d\mu$$

avoids this: it is bounded by one, stays finite even when the supports of P and Q differ, and controls total variation through $\mathcal{H}^2 \le \mathrm{TV} \le \sqrt{2}\,\mathcal{H}$, the metric in which frequentist robustness is framed. Under Huber's ε-contamination model the contaminated law lies within ε of the clean model in this metric: for arbitrary Q,

$$\mathcal{H}^2\big((1-\varepsilon)P_{\theta_0} + \varepsilon Q,\ P_{\theta_0}\big) \le \varepsilon, \tag{1.35}$$

since $\mathcal{H}^2 \le \mathrm{TV}$ and $\mathrm{TV}\big((1-\varepsilon)P_{\theta_0} + \varepsilon Q, P_{\theta_0}\big) = \varepsilon\,\mathrm{TV}(Q, P_{\theta_0}) \le \varepsilon$ (Tsybakov, 2008, §2.4). The contrast $\widehat{R}_\psi$ below is therefore built from Hellinger affinities, and (1.35) lets contamination enter the oracle inequality of Chapter 4 only as an additive term of order ε.

The ρ-posterior. Baraud et al. (2017, Thm. 1) introduce ρ-estimation, Baraud & Birgé (2020, §3) extend it to a ρ-Bayes posterior, and Baraud & Birgé (2020, Thm. 1) establish oracle inequalities under ε-contamination; the Hellinger-comparison machinery is laid out in Baraud & Birgé (2018, Prop. 3). For each pair (θ, θ'), the pairwise contrast

$$\widehat{R}_\psi(\theta, \theta') = \frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{p_{\theta'}(X_i)}{p_\theta(X_i)}\right), \qquad \psi(x) = \frac{\sqrt{x}-1}{\sqrt{x}+1} \in [-1, 1], \tag{1.36}$$

measures relative fit through Hellinger affinities. Since ψ takes values in $[-1, 1]$, a single observation changes $\widehat{R}_\psi$ by at most $2/n$, so no single observation can dominate it, the bounded influence on which its robustness rests. The supremum contrast $\widehat{R}^\star_\psi(\theta) := \sup_{\theta' \in \Theta} \widehat{R}_\psi(\theta, \theta')$ yields the *ρ-estimator* $\widehat{\theta}_\rho = \operatorname{argmin}_\theta \widehat{R}^\star_\psi(\theta)$ and the Bayesian *ρ-posterior* $\hat{\rho}^\rho_n \propto \exp(-n\,\widehat{R}^\star_\psi)\,\pi$ of Baraud & Birgé (2020), both defined in full in Chapter 4. Under Huber's ε-contamination model $P^* = (1-\varepsilon)P_{\theta_0} + \varepsilon Q$ with arbitrary Q (Huber, 1992), Baraud & Birgé (2020, Thm. 1) establish that the ρ-posterior contracts around the best in-model approximation of P^* in squared Hellinger distance, the misspecification bias entering as an infimum of a Hellinger term, the estimation error governed by the metric dimension of the model, and the contamination level entering additively rather than multiplying the estimation term: informally,

$$\mathbb{E}_{\theta \sim \hat{\rho}^\rho_n}\big[\mathcal{H}^2(P^*, P_\theta)\big] \lesssim \inf_{\theta \in \Theta} \mathcal{H}^2(P^*, P_\theta) + v_n(\Theta) + \varepsilon, \tag{1.37}$$

where $v_n(\Theta)$ is the model's minimax rate, set by its metric dimension up to logarithmic factors, and the additive ε follows from (1.35). The rate thus degrades only gracefully under contamination.

Computational obstruction. Despite these guarantees, the ρ-posterior has received no implementation and remained an elegant theoretical device. The obstruction is structural: its supremum contrast $\widehat{R}^\star_\psi(\theta) = \sup_{\theta'} \widehat{R}_\psi(\theta, \theta')$ requires, for each θ, a worst-case comparison over the whole parameter space, which does not lend itself directly to MCMC or variational

inference.

A tractable surrogate via softmax aggregation. Chapter 4 removes this obstruction by replacing the intractable supremum with a softmax aggregation over competitor parameters drawn against an auxiliary prior π',

$$\Lambda_\lambda(\theta;\pi') = \frac{1}{\lambda}\log\int_\Theta e^{\lambda\,\widehat{R}_\psi(\theta,\theta')}\,\pi'(d\theta'),$$

which recovers $\widehat{R}^\star_\psi$ as $\lambda\to\infty$. The *generalised ρ-posterior* is the Gibbs minimiser

$$\hat\rho_\lambda = \operatorname*{argmin}_{\rho\in\mathcal{M}_1^+(\Theta)}\Big\{\mathbb{E}_{\theta\sim\rho}\big[\Lambda_\lambda(\theta;\pi')\big] + \frac{1}{\lambda}\,\mathrm{KL}(\rho\,\|\,\pi)\Big\}.$$

Restricting it to a tractable family $\mathcal{F}\subset\mathcal{M}_1^+(\Theta)$, as in Section 1.5, gives the *variational* counterpart $\widetilde\rho_\lambda$.

Oracle inequality for the generalised $\widetilde\rho$-posterior

At inverse temperature $\lambda = n/8$ and with high probability, the generalised ρ-posterior $\hat\rho_\lambda$ satisfies the squared-Hellinger oracle inequality, informally,

$$\mathbb{E}_{\theta\sim\hat\rho_\lambda}\big[\mathcal{H}^2(P^*,P_\theta)\big] \lesssim \inf_{\rho\in\mathcal{M}_1^+(\Theta)}\Big\{\mathbb{E}_{\theta\sim\rho}\big[\mathcal{H}^2(P^*,P_\theta)\big] + \frac{\mathrm{KL}(\rho\,\|\,\pi)}{n}\Big\} + \frac{\log(1/\delta)}{n}.$$

The same inequality holds for the variational $\widetilde\rho_\lambda$ with the infimum over $\mathcal{F}$, which thus pays nothing beyond the best fit attainable there. A prior-mass condition makes the rate explicit:

Concentration rate for the generalised ρ-posterior

Under a prior-mass condition with Catoni dimension d_π, with high probability

$$\mathbb{E}_{\theta\sim\hat\rho_\lambda}\big[\mathcal{H}^2(P^*,P_\theta)\big] \lesssim \inf_{\theta\in\Theta}\mathcal{H}^2(P^*,P_\theta) + \frac{d_\pi\log n}{n},$$

the best in-model fit plus a Catoni-dimension term, minimax-optimal up to a log factor.

These inequalities are proved in Section 4.2 of Chapter 4; the contributions previewed here are developed in full in the chapters that follow. Chapter 5 draws them together and closes with perspectives for future work.

Chapter 2

On Importance Sampling and Independent Metropolis–Hastings with an Unbounded Weight Function

This chapter reproduces, with minor typographic modifications, the paper:

G. Deligiannidis, P. E. Jacob, E. M. Khribch, G. Wang

On Importance Sampling and Independent Metropolis–Hastings with an Unbounded Weight Function

See Deligiannidis et al. (2024).

Importance sampling and independent Metropolis–Hastings (IMH) are among the fundamental building blocks of Monte Carlo methods. Both require a proposal distribution that globally approximates the target distribution, and pointwise evaluation of the Radon–Nikodym derivative of the target distribution relative to the proposal, also called the weight function. We study the bias of importance sampling and particle independent Metropolis–Hastings (PIMH), without assuming that the weight function is bounded. We show that the common random numbers coupling of PIMH is maximal. Using that coupling, we derive bounds on the total variation distance of a PIMH chain to its target distribution. We further consider bias removal techniques using couplings, and provide conditions under which the resulting unbiased estimators have finite moments. Experiments illustrate the performance of these unbiased estimators for the estimation of the inverse of a normalizing constant, for nested expectations, and in combination with median-of-means.

Keywords: Monte Carlo, importance sampling, Markov chain, robust mean estimation.

2.1 Introduction

2.1.1 Context and contributions

Monte Carlo with global proposals

Monte Carlo methods aim to approximate a target distribution π on a measurable space $(\mathbb{X}, \mathcal{X})$, for example $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. These techniques are useful when analytical computation of expectations under π is infeasible. The goal is to evaluate integrals of functions $f : \mathbb{X} \longrightarrow \mathbb{R}$ with respect to π:

$$\pi(f) := \mathbb{E}_\pi[f] = \int_{\mathcal{X}} f(x)\pi(x)\,\mathrm{d}x. \tag{2.1}$$

Two primary approaches are Markov chain Monte Carlo (MCMC) methods, that construct a Markov chain with π as its stationary distribution, and importance sampling (IS) methods, where the target distribution is approximated by weighted samples. Among MCMC methods, the independent Metropolis–Hastings (IMH) algorithm is a specialized form of the Metropolis–Rosenbluth–Teller–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), in which proposals are drawn from a proposal distribution q, independently of the current state of the chain. We assume that we can sample from q on $(\mathbb{X}, \mathcal{X})$, such that π is absolutely continuous with respect to q, and we can evaluate the Radon–Nikodym derivative $\omega(x) = \pi(x)/q(x)$ up to a multiplicative constant. We call ω the *weight* function. The same proposal q can be employed in importance sampling (Kahn, 1949), where draws from q are weighted by ω. Rejection sampling could be implemented if ω was bounded by a known constant, but in this chapter we let ω be unbounded.

Comparing MH with independent proposals and importance sampling

Independent MH and IS are two approaches to account for the discrepancy between a proposal q and a target π, and their comparison is a natural question. The asymptotic variance of self-normalized IS (SNIS) estimators is lower than that of IMH estimators (Deligiannidis & Lee, 2018), see (2.7) below. In both cases, the squared bias is negligible relative to the variance as the computing effort goes to infinity, and thus the bias can be ignored from the perspective of asymptotic mean squared error. Perhaps for this reason, the comparison of the bias of SNIS and IMH has received little attention, particularly in the general setting where the weight ω may be unbounded. However, when the resulting biased Monte Carlo is used for downstream tasks, the non-asymptotic bias matters and can even become the dominant issue. We review multiple such settings below.

The question of bias can be rephrased as follows, for bounded test functions. Among N independent draws $(X_n)_{n=1}^N$ from q, consider the selection of an index $I \in [N] = \{1, \dots, N\}$

such that X_I is as close as possible to π in total variation distance. Is it preferable to perform N steps of IMH and select the terminal state, or to draw I from a Categorical distribution with probabilities proportional to $\omega(X_1), \ldots, \omega(X_N)$, as in Sampling-Importance-Resampling (SIR, Rubin, 1988)? Our developments lead to the conclusion that, in some generality, IMH leads to a smaller bias than SIR; see Section 2.4.6 for a summary. We further explore ways of entirely removing such biases, following Middleton et al. (2019), and study their asymptotic efficiency in Section 2.5.

Effect of the bias of Monte Carlo estimators

We review scenarios where the bias of Monte Carlo estimators has noticeable and often detrimental effects on downstream tasks, motivating both its study and its removal.

Optimization with stochastic gradients Monte Carlo estimators are used in the approximation of gradients within stochastic gradient algorithms. The effect of the bias of gradient estimators in the encompassing optimization has received attention (e.g. Tadic & Doucet, 2017; Hu et al., 2021; Demidovich et al., 2023; Surendran et al., 2024). The literature shows that the bias can induce a non-vanishing error, or affect the convergence rate, or can have no significant effect, depending on its magnitude. The case of IS-based gradient estimators employed in training importance-weighted autoencoders (IWAEs) has received particular attention: the bias of gradient estimators is studied in Daudel & Roueff (2024, Theorem 1), and is shown to propagate into the encompassing optimization in Surendran et al. (2024, Corollary 4.5, Theorem 4.6 and Theorem B.1), and in Surendran et al. (2025, Theorem 3.7). A number of bias-reduction techniques have been proposed in the context of IWAEs, including jackknife correction (Nowozin, 2018, Proposition 3), randomized multilevel Monte Carlo (Shi & Cornish, 2021, Theorems 2 and 3), and the iterated SIR (Cardoso et al., 2022, Theorems 3–4), assuming bounded weights. Experiments in Shi & Cornish (2021); Cardoso et al. (2022); Surendran et al. (2024) show that, relative to SNIS gradient estimators, reducing or removing the bias can be overall beneficial at matched computational cost.

Nested expectations The bias of Monte Carlo estimators has a particularly clear effect in the context of nested expectations (Andradóttir & Glynn, 2016). The quantity of interest is of the form

$$\mathbb{E}_1 \left[\mathbb{E}_{2|1}[f(\Theta_1, \Theta_2)|\Theta_1] \right], \tag{2.2}$$

where $\mathbb{E}_{2|1}$ denotes a conditional expectation of Θ_2 given Θ_1, and f is a test function. We assume that we are able to sample Θ_1 from its marginal distribution, and that we can approximate $\mathbb{E}_{2|1}[f(\Theta_1, \Theta_2)|\Theta_1]$ for each Θ_1 by a numerical method, such as SNIS or IMH. Such nested expectations arise in modular Bayesian inference (Liu et al., 2009; Plummer, 2014), in-

cluding Bayesian inference after multiple imputation (Zhou & Reiter, 2010), experimental design (Rainforth et al., 2018a) or portfolio risk measurement (Gordy & Juneja, 2010).

Assume that, for each realization θ_1 of Θ_1, the conditional expectation $\mathbb{E}_{2|1}[f(\Theta_1, \Theta_2)|\Theta_1 = \theta_1]$ is approximated by an estimator $\hat{F}(\theta_1)$ with cost N, bias of order $N^{-\gamma}$ for some $\gamma \in (0, \infty)$, and variance of order N^{-1}. Consider the estimator $M^{-1}\sum_{m=1}^{M}\hat{F}(\Theta_{1,m})$, where $\Theta_{1,1}, \ldots, \Theta_{1,M}$ are i.i.d. draws from the marginal distribution of Θ_1. The total cost is $C = NM$, and optimizing the budget allocation leads to an overall mean squared error of order $C^{-2\gamma/(2\gamma+1)}$ (Remarks 3.5-3.6, Andradóttir & Glynn, 2016); a detailed proof of this rate is given in Appendix D.6 (Proposition D.8). Therefore, the magnitude of the bias of the inner estimator directly impacts the overall rate of convergence of the estimator of the nested expectation of interest, and therefore any reduction in bias brings concrete benefits in the overall mean squared error. Furthermore, unbiased estimators for the inner expectation recover the canonical Monte Carlo rate for the nested expectation. We illustrate this in Subsection 2.6.2.

Plug-and-play property of unbiased estimators In general, the effect of the bias of a Monte Carlo estimator in an encompassing method is unknown, and the magnitude of that bias is typically hard to bound explicitly. Removing the bias offers a plug-and-play solution. We illustrate this in Section 2.6.3 with the problem of robust mean estimation. Methods for robust mean estimation, such as median-of-means, typically assume access to variables that have as expectation the quantity of interest. Biased estimators can be used instead (see Dau (2022) for the use of SNIS within median-of-means), but require potential adjustments in both theory and implementation.

Our contributions

We consider self-normalized importance sampling, independent Metropolis–Hastings, and an unbiased estimator that relates to both. Our key assumption is that the *weight* function ω has p finite moments under q. We obtain results on the bias and higher moments of self-normalized importance sampling estimators in Section 2.2. For IMH we show that the common random numbers coupling is an "all time" maximal coupling in Section 2.3. Using that coupling, in Section 2.4 we show that the total variation distance between the IMH chain at iteration t and π decays as t^{p-1}. We obtain lower bounds that match up to logarithmic terms in an example, which suggests that our working assumption of p finite moments of ω under q is close to necessary for polynomial convergence rates. We further obtain explicit dependence in N for the particle IMH algorithm (Andrieu et al., 2010), a variant of IMH where N proposals are sampled at each iteration. To establish precise bounds that account for both t and N, we develop novel methods to analyze the average weight, $\hat{Z}_N = N^{-1}\sum_{n=1}^{N}\omega(X_n)$, and thereby control the rejection probability. We first use the *Paley–Zygmund inequality* (Petrov, 2007) (an anticoncen-

tration inequality) to provide a lower bound on the probability of $\hat{Z}_N$ being small, specifically controlling its behavior when $\hat{Z}_N \leq 2$. For larger values of $\hat{Z}_N$, we divide the range into an intermediate section $(2, 1+t)$ and a large section $[1+t, \infty)$. The *Markov inequality* directly handles the large range. For the intermediate range, we employ a *peeling argument*, breaking it down into a union of smaller intervals and applying the Markov inequality to each. Our multiscale analysis, combining both concentration and anticoncentration inequalities, allows us to control the t-th moment of the rejection probability. This eventually leads to a total-variation bound through the coupling inequality. In Section 2.5 we consider the bias removal technique of Glynn & Rhee (2014) applied by Middleton et al. (2019) to the particle IMH algorithm. We provide conditions under which these unbiased estimators have finite moments, and conditions under which their efficiency is asymptotically equivalent to that of self-normalized importance sampling. This distinguishes the bias removal technique of Middleton et al. (2019) from other debiasing strategies for IS (Blanchet & Glynn, 2015; Shi & Cornish, 2021; Chopin et al., 2025), which do not have this property. Section 2.6 presents numerical illustrations of these unbiased variants of importance sampling, for the estimation of the inverse of the normalizing constant of a distribution, for nested expectations, and in the context of robust mean estimation.

2.1.2 Importance sampling

Self-normalized importance sampling (IS) is described in Algorithm 1, see also Chapter 9.2 in Owen (2013). Central to IS is the weight function:

$$\omega : x \mapsto \frac{\pi(x)}{q(x)}, \quad \text{so that} \quad q(\omega) = 1. \tag{2.3}$$

Since multiplicative constants in ω have no effect on the IS estimator (2.5), it can be computed as long as the user can evaluate a function proportional to ω in (2.3). Unless specified otherwise, by IS we refer to the self-normalized version as in Algorithm 1 rather than the more basic estimator $N^{-1}\sum_{n=1}^{N}\omega(X_n)f(X_n)$ that depends on the multiplicative constant in ω.

For brevity we sometimes write $\mathbf{X}^N$ for $(X_1, \ldots, X_N)$, and $\hat{F}_N$ for $\hat{F}_N(\mathbf{X}^N)$ i.e. the IS estimator, and $\hat{Z}_N$ for $\hat{Z}_N(\mathbf{X}^N)$ i.e. the normalizing constant estimator, when the dependence on $\mathbf{X}^N$ is clear from the context. We make the following assumption throughout.

Assumption 2.1. For any measurable set $A \in \mathcal{X}$, if $q(A) = 0$, then $\pi(A) = 0$. Furthermore, $\omega(X)$ with $X \sim q$ is almost surely positive, and $q(\omega) = 1$.

Under Assumption 2.1, if $\pi(f)$ exists then $\hat{F}_N(X_1, \ldots, X_N) \to \pi(f)$ as $N \to \infty$ almost surely. The asymptotic variance of $\hat{F}_N$ can be directly computed from the delta method (Owen, 2013; Robert & Casella, 2004; Liu, 2008), assuming $q(\omega^2 \cdot f^2) < \infty$ and $q(\omega^2) < \infty$,

$$\lim_{N\to\infty} \mathbb{V}\left[\sqrt{N}(\hat{F}_N(X_1, \ldots, X_N) - \pi(f))\right] = q(\omega^2 \cdot (f - \pi(f))^2) =: \sigma^2_{\text{IS}}. \tag{2.6}$$

Algorithm 1. Self-normalized importance sampling

1. Sample N particles independently $X_1, \ldots, X_N$ from q.
2. Compute the importance weights $\omega(X_n) = \pi(X_n)/q(X_n)$ for $n \in [N] = \{1, \ldots, N\}$.
3. Compute
$$\hat{Z}_N(X_1, \ldots, X_N) = N^{-1} \sum_{n=1}^{N} \omega(X_n). \tag{2.4}$$
4. For any test function f, compute the IS estimator
$$\hat{F}_N(X_1, \ldots, X_N) = \frac{\sum_{n=1}^{N} \omega(X_n) f(X_n)}{\sum_{n=1}^{N} \omega(X_n)}. \tag{2.5}$$

Non-asymptotic bounds on the mean squared error and on the bias of $\hat{F}_N$ are derived in Agapiou et al. (2017). These bounds are inversely proportional to the number N of draws from q. The exact form of the asymptotic bias of the IS estimator $\hat{F}_N$ is well-known (e.g. Hesterberg, 1988; Liu, 2008), and we provide a formal statement in Section 2.2.

2.1.3 Independent Metropolis–Hastings

Independent Metropolis–Hastings (IMH) is an instance of the Metropolis–Rosenbluth–Teller–Hastings algorithm (Hastings, 1970, Section 2.5), described in Algorithm 2. The IMH chain is aperiodic and π-invariant by design, whereas under Assumption 2.1, it is also π-irreducible; thus for π-almost every x, $|P^t(x, \cdot) - \pi|_{\mathrm{TV}} \to 0$ as $t \to \infty$ (Theorem 4 in Roberts & Rosenthal, 2004), where P denotes the transition kernel of IMH, P^t denotes the t-steps transition kernel, and $|\mu - \nu|_{\mathrm{TV}} = \sup_{A \in \mathcal{X}} \mu(A) - \nu(A)$.

The asymptotic variance of the ergodic average $t^{-1} \sum_{s=0}^{t-1} f(X_s)$ generated by IMH, denoted by σ^2_{IMH}, is finite if and only if $\pi(f^2) < \infty$ and $q(\omega^2 \cdot f^2) < \infty$ (Theorem 2 in Deligiannidis & Lee, 2018). Furthermore, if σ^2_{IMH} is finite then Deligiannidis & Lee (2018, Proposition 2) provides a general comparison:
$$\sigma^2_{\mathrm{IS}} \leq \sigma^2_{\mathrm{IMH}}, \tag{2.7}$$
with σ^2_{IS} as in (2.6). Thus IS outperforms IMH in asymptotic variance and in asymptotic mean squared error. Since IMH defines a Markov transition, it can be used directly as a step within an encompassing Gibbs sampler (Skare et al., 2003), and it is commonly used within sequential Monte Carlo samplers (Chopin, 2002; South et al., 2019), and thus has its specific uses irrespective of the performance comparison with importance sampling.

Algorithm 2. One step of IMH from state X

1. Draw $X^\star \sim q$.
2. Compute the acceptance probability

$$\alpha(X, X^\star) = \min\left\{1, \frac{\omega(X^\star)}{\omega(X)}\right\}. \tag{2.8}$$

3. Draw U from a Uniform$(0,1)$ distribution, independently of $X^\star$.
4. If $U < \alpha(X, X^\star)$, return $X^\star$, otherwise return X.

When it comes to non-asymptotic behavior, for IMH there is an important distinction between two cases (Mengersen & Tweedie, 1996): either the weight is bounded, in which case the chain is geometrically ergodic and exact rates are obtained in Wang (2022), or the weight is unbounded, and the convergence cannot be geometric; in the latter case, various results are provided e.g. in Jarner & Roberts (2002); Douc et al. (2007); Roberts & Rosenthal (2011); Andrieu et al. (2022) and Douc et al. (2018, Chapter 17); see Section 2.4.5. In Section 2.4 we provide polynomial bounds on the total variation distance to stationarity under moment conditions on ω under q. Our results enable bias comparison of IS and IMH in Section 2.4.6, along with practical recommendations.

In the following we consider the particle generalization of IMH (PIMH), where N proposals are drawn at each iteration (Andrieu et al., 2010, Section 4.2); see Algorithm 3. We define the algorithm on the state space $\mathbb{X}^N$, use boldface to denote its elements, e.g. $\mathbf{X}^N = (X_1, \ldots, X_N) \in \mathbb{X}^N$, and denote the transition kernel by P. If $N = 1$ the algorithm corresponds to IMH, and our results apply for all $N \geq 1$. To view Algorithm 3 as a special case of IMH, define for any $N \geq 1$

$$\bar{\pi}(x_1, \ldots, x_N) = \sum_{k=1}^{N} \frac{\pi(x_k)}{N} \prod_{n \neq k} q(x_n) = \left(\frac{1}{N}\sum_{k=1}^{N} \omega(x_k)\right) \prod_{n=1}^{N} q(x_n), \tag{2.9}$$

$$\bar{q}(x_1, \ldots, x_N) = \prod_{n=1}^{N} q(x_n), \tag{2.10}$$

and, in the case $N = 1$, $\bar{\pi}(x_1) = \pi(x_1)$. From (2.4) and the above definitions, we define the Radon–Nikodym derivative of $\bar{\pi}$ with respect to $\bar{q}$ as:

$$\bar{\omega}(x_1, \ldots, x_N) = \frac{\bar{\pi}(x_1, \ldots, x_N)}{\bar{q}(x_1, \ldots, x_N)} = \frac{1}{N}\sum_{n=1}^{N} \omega(x_n) = \hat{Z}_N(x_1, \ldots, x_N). \tag{2.11}$$

Note that $\mathbb{E}_{\mathbf{X}^N \sim \bar{q}}[\bar{\omega}(\mathbf{X}^N)] = N^{-1}\sum_{n=1}^{N} q(\omega) = 1$ under Assumption 2.1, thus $\bar{\pi}$ is properly normalized. Hence, IMH as in Algorithm 2, with proposal $\bar{q}$ and target $\bar{\pi}$, is equivalent to

Algorithm 3.

Algorithm 3. One step of PIMH from state $\mathbf{X}^N$

1. Draw $\mathbf{X}^{N,\star} = (X_1^\star, \ldots, X_N^\star) \sim \bar{q}$.
2. Compute the acceptance probability

$$\alpha(\mathbf{X}^N, \mathbf{X}^{N,\star}) = \min\left\{1, \frac{\hat{Z}(\mathbf{X}^{N,\star})}{\hat{Z}(\mathbf{X}^N)}\right\}, \tag{2.12}$$

   where $\hat{Z}_N(\mathbf{X}^N) = N^{-1}\sum_{n=1}^N \omega(X_n)$.
3. Draw U from a Uniform$(0,1)$ distribution, independently of $\mathbf{X}^{N,\star}$.
4. If $U < \alpha(\mathbf{X}^N, \mathbf{X}^{N,\star})$, return $\mathbf{X}^{N,\star}$, otherwise return $\mathbf{X}^N$.

In order to estimate an expectation $\pi(f)$ from the PIMH output using all the generated particles, note that

$$\begin{aligned}
\mathbb{E}_{\bar{\pi}}[\hat{F}_N(\mathbf{X}^N)] &= \int \hat{F}_N(x_1,\ldots,x_N)\cdot\hat{Z}_N(x_1,\ldots,x_N)\cdot\bar{q}(x_1,\ldots,x_N)\mathrm{d}x_1\ldots\mathrm{d}x_N \\
&= \int\left\{\frac{1}{N}\sum_{n=1}^N \omega(x_n)f(x_n)\right\}\cdot\bar{q}(x_1,\ldots,x_N)\mathrm{d}x_1\ldots\mathrm{d}x_N \\
&= \int \omega(x_1)f(x_1)q(x_1)\mathrm{d}x_1 = \pi(f).
\end{aligned} \tag{2.13}$$

Thus, we can evaluate $\hat{F}_N$ at each state of the chain $(\mathbf{X}_t^N)_{t\geq 0}$ generated by the PIMH algorithm. The ergodic average $T^{-1}\sum_{t=0}^{T-1}\hat{F}_N(\mathbf{X}_t^N)$ converges to $\pi(f)$ under suitable conditions. With this notation, the IS estimator (2.5) is $\hat{F}_N(\mathbf{X}^N)$ with $\mathbf{X}^N \sim \bar{q}$.

2.1.4 Moment conditions on the weight

Our key assumption is the following.

Assumption 2.2. The weights have a finite p-th moment for $p > 1$: $q(\omega^p) < \infty$.

This is a weak assumption in the context of both self-normalized importance sampling and IMH. For bounded test functions, Assumption 2.2 with $p \geq 2$ is necessary for the asymptotic variance of both self-normalized importance sampling and IMH to be finite. The case where $p < 2$ is also of interest, and we include a number of results covering it.

Example 2.3 (Exponential distributions)**.** Let π be the Exponential(1) distribution and let q be the Exponential(k) distribution with $q(x) = ke^{-kx}$, both on $\mathbb{R}_+$. If $k \leq 1$, the weight $\omega(x)$ is upper bounded by k^{-1}, and Assumption 2.2 holds for all p. If $k > 1$, then $q(\omega^p) < \infty$

holds with any $p < k/(k-1)$. The example is considered in Jarner & Roberts (2007); Roberts & Rosenthal (2011); Andrieu et al. (2022); Orenstein (2022).

Example 2.4 (Normal distributions)**.** Let π be the Normal(0,1) distribution and let q be the Normal$(0, \sigma^2)$ distribution, both on $\mathbb{R}$. If $\sigma^2 \geq 1$, the weight $\omega(x)$ is upper bounded by σ, and Assumption 2.2 holds for all p. If $\sigma^2 < 1$, then $q(\omega^p) < \infty$ holds for $p < \sigma^{-2}/(\sigma^{-2} - 1)$. The example is considered in Roberts & Rosenthal (2011); Owen (2013).

Assumption 2.2 with $p \geq 2$ implies the following well-known behavior of the average of N independent and identically distributed random variables, obtained from the Marcinkiewicz–Zygmund inequalities. The proof is in Appendix A.1.

Proposition 2.5. *Let $\mathbf{X}^N = (X_n)_{n=1}^N$ be N independent random variables with distribution q. Under Assumptions 2.1–2.2, with $p \geq 2$, $\hat{Z}_N(\mathbf{X}^N) = N^{-1}\sum_{n=1}^N \omega(X_n)$ satisfies, for all $N \geq 1$, and $M(p) := 2p\, q(\omega^p)^{1/p}$,*

$$\mathbb{E}_{\bar{q}}[|\hat{Z}_N(\mathbf{X}^N) - 1|^p]^{1/p} \leq \frac{M(p)}{\sqrt{N}}, \tag{2.14}$$

$$\mathbb{E}_{\bar{q}}[\hat{Z}_N(\mathbf{X}^N)^p]^{1/p} \leq 1 + \frac{M(p)}{\sqrt{N}}, \tag{2.15}$$

$$\mathbb{P}_{\bar{q}}\left(\hat{Z}_N(\mathbf{X}^N) \geq 1 + t\right) \leq \left(\frac{M(p)}{t\sqrt{N}}\right)^p, \qquad t > 0. \tag{2.16}$$

Remark 2.6. Some of our results do not require $\hat{Z}_N(\mathbf{X}^N)$ to be an average of i.i.d. weights, but only that $\hat{Z}_N(\mathbf{X}^N)$ is non-negative, unbiased for the normalizing constant of the target distribution, and satisfies the moment bound (2.14), and thus (2.15) and (2.16) as well since they follow from (2.14). These properties may hold, for example, for the normalizing constant estimator generated by sequential Monte Carlo (SMC) samplers, under some conditions; see e.g. Del Moral (2013, Section 16.5) for related results. For example, the results on the coupling of particle IMH and the resulting meeting times in Sections 2.3-2.4 apply whenever $\hat{Z}_N(\mathbf{X}^N)$ satisfies these properties, as can be seen from the proofs. However, the results on importance sampling in Section 2.2 and bias removal in Section 2.5 rely on the i.i.d. structure of $\mathbf{X}^N$ in multiple places, and do not extend to the SMC setting without modification, which we leave as an open line of research.

Remark 2.7. If Assumption 2.2 holds for some $p \in (1, 2)$, the von Bahr–Esseen inequality (von Bahr & Esseen, 1965) can be used to obtain a bound similar to (2.14) but with $N^{-(p-1)/p}$ instead of $N^{-1/2}$ on the right-hand side. Then a bound similar to (2.15) can be obtained by the triangle inequality, a bound similar to (2.16) by Markov's inequality. In each case $\sqrt{N}$ is replaced by $N^{(p-1)/p}$.

2.2 Bias and moments of importance sampling

Our first result is a clean statement on the asymptotic bias of self-normalized importance sampling. Introductory material on importance sampling often notes that the basic importance sampling estimator $N^{-1}\sum_{n=1}^{N} f(X_n)\omega(X_n)$ is unbiased, but since ω can only be evaluated up to a multiplicative constant, users may need to resort to the self-normalized estimator $\hat{F}_N(\mathbf{X}^N)$ in (2.5), which is biased: $\mathbb{E}[\hat{F}_N(\mathbf{X}^N)] \neq \pi(f)$. The form of the asymptotic bias is well known, e.g. Section 2.5 in Liu (2008). Formal results with self-contained proofs are however hard to find. Theorem 2 in Skare et al. (2003) emphasizes on the pointwise relative error of the density of a particle selected from the IS approximation. Their Remark 1 translates this into a bound on the bias for bounded functions under the assumption of bounded weights. The asymptotic bias expression appears in Hesterberg (1988), equation (2.63), and is justified by the Edgeworth expansion of the cumulative distribution function of $\sqrt{N}(\hat{F}_N - \pi(f))$, using results from Bhattacharya & Ghosh (1978). However, as noted in Hesterberg (1988, p. 40), the expression is obtained as a first order term in an Edgeworth expansion which does not establish that the actual bias is well-defined or finite. Moreover, the approach requires third moments of ω and ωf under q, as well as Cramér's condition on the characteristic function of $(\omega, \omega^2, \omega f, \omega^2 f^2)$ (Hesterberg, 1988, p. 39). Theorem 2.8 below requires fewer positive moments on ω and ωf, but assumes the existence of an inverse moment for ω. The proof in Appendix A.2 is elementary.

Theorem 2.8. *Let f be such that $\pi(|f|) < \infty$. Let $\mathbf{X}^N = (X_1, \ldots, X_N) \sim \bar{q}$ as in* (2.10) *and let $\hat{F}_N(\mathbf{X}^N)$ be as in* (2.5). *Assume that $q(|\omega^2(f - \pi(f))|^{1+\epsilon}) < \infty$ for some $\epsilon > 0$ and $q(\omega^{-\eta}) < \infty$ for some $\eta > 0$. Then*

$$\lim_{N\to\infty} N \times \mathbb{E}_{\mathbf{X}^N\sim\bar{q}}\left[\hat{F}_N(\mathbf{X}^N) - \pi(f)\right] = -\int (f(x) - \pi(f))\,\omega^2(x) q(dx). \tag{2.17}$$

Theorem 2.8 relates to the question raised in Section 2.1.1. Indeed, consider a draw X_I obtained by Sampling-Importance Resampling (SIR), and use the representation of the total variation distance as a supremum of difference in expectations of test functions bounded by one. Theorem 2.8 then shows that the total variation distance between X_I and π is of order N^{-1} as $N \to \infty$. Instead of exact asymptotic order, Agapiou et al. (2017, Theorem 2.1) provides an upper bound on the bias under weaker assumptions. Agapiou et al. (2017, Theorem 2.3) provides upper bounds of order N^{-1} also for unbounded test functions, under moment conditions on f and on $f \cdot \omega$. Our Theorem 2.8 establishes that N^{-1} is the exact order of the asymptotic bias as a function of N, but requires additional conditions. Theorem 2.8 is similar to results proposed independently in Daudel & Roueff (2024). Finally, the bias of sequential Monte Carlo estimators is typically also of order N^{-1}, where N is the number of particles, as discussed in Del Moral et al. (2007), and in Del Moral (2013, Section 15.3).

We next provide a result on the s-th moments of the error in importance sampling for unbounded test functions. Theorem 2.9 generalizes the MSE part of Theorem 2.3 in Agapiou et al. (2017) to arbitrary orders $s \geq 2$, and its assumptions are weaker in the case $s = 2$, as discussed below. The proof is in Appendix A.2. These bounds are used to obtain the results of Section 2.5, and may be of independent interest.

Theorem 2.9. *Assume that there exist $p \in [2, \infty)$ and $r \in [2, \infty]$ such that $q(\omega^p) < \infty$ and $q(|f|^r) < \infty$, and $q(f^2 \cdot \omega^2) < \infty$, then for any $2 \leq s \leq pr/(p + r + 2)$ and any $N \geq 1$, we have:*

$$\mathbb{E}_{\mathbf{X}^N \sim \bar{q}} \left[\left| \hat{F}_N(\mathbf{X}^N) - \pi(f) \right|^s \right] \leq C N^{-s/2},$$

where the constant C depends on $r, p, s, q(|f|^r), q(\omega^p), q(f^2 \cdot \omega^2)$. When $r = \infty$, the statement holds for f such that $|f|_\infty < \infty$ and all $s \leq p$.

A few remarks are in order:

- The condition $s \leq pr/(p + r + 2)$ implies $s \leq \min\{p, r\}$.
- We have $q((|f|\omega)^{pr/(p+r)}) < \infty$ if $q(\omega^p) < \infty$ and $q(|f|^r) < \infty$. Indeed, when $r < \infty$, $q((|f|\omega)^{pr/(p+r)}) \leq q(|f|^r)^{p/(p+r)} q(\omega^p)^{r/(p+r)} < \infty$. When $r = \infty$, the claim remains correct (by understanding $pr/(p + r)$ as p), since $q((|f|\omega)^p) \leq |f|_\infty^p q(\omega^p)$. This observation leads to two facts: 1) If $pr/(p + r) \geq 2$ (e.g. $p = r = 4$ or $p = 2, r = \infty$), the assumption $q(f^2 \cdot \omega^2) < \infty$ in Theorem 2.9 can be derived from the assumptions $q(\omega^p) < \infty$ and $q(|f|^r) < \infty$. 2) The basic importance sampling estimator $N^{-1} \sum_{n=1}^N f(X_n)\omega(X_n)$ has a finite s-th moment under the same conditions, as it has a finite $pr/(p + r)$-th moment, and $s \leq pr/(p + r + 2) \leq pr/(p + r)$.
- We may be particularly interested in the mean-squared error (MSE) of IS, corresponding to $s = 2$. Theorem 2.9 implies that the MSE is of order $1/N$ as long as $2 \leq pr/(p + r + 2)$. This condition holds, for example, if $\min\{p, r\} \geq 2(1 + \sqrt{2}) \approx 4.828$, or if $p \geq 3$ and $r \geq 10$, or if $p = 2$ and $r = \infty$. The case $s = 2$ can be compared to the MSE part of Theorem 2.3 in Agapiou et al. (2017). In our notation, they require $q(|f \cdot \omega|^{2d}) < \infty$, $q(\omega^{2e}) < \infty$, $q(|f|^{2a}) < \infty$, $q(\omega^{2b(1+a^{-1})}) < \infty$, for $a, b, d, e > 1$ such that $a^{-1} + b^{-1} = 1$, $d^{-1} + e^{-1} = 1$. Their assumption implies ours, as can be seen by setting $r = 2a$ and $p = 2b(1 + a^{-1})$, since then

$$\begin{aligned} \frac{pr}{(p + r + 2)} &= \frac{4b(a + 1)}{2a + 2b + 2ba^{-1} + 2} \\ &= \frac{2(a + 1)}{a(1 - a^{-1}) + 1 + a^{-1} + (1 - a^{-1})} = \frac{2(a + 1)}{a + 1}, \end{aligned}$$

i.e. our theorem holds with $s = 2$ under their assumptions.

- Comparable results are given in Shi & Cornish (2021). For bounded test functions (Shi & Cornish, 2021, Supplementary material, Lemma 5), the two results are equivalent. For unbounded test functions (Shi & Cornish, 2021, Supplementary material, Lemma 6), Theorem 2.9 yields a bound of $\mathcal{O}(N^{-s/2})$ under strictly weaker conditions, whereas Lemma 6 of Shi & Cornish (2021) gives only a bound of $\mathcal{O}(N^{-s/2+s/r})$ under $q(\omega^p) < \infty$ and $q(|f|^r) < \infty$.
- Similar types of results have been obtained for sequential Monte Carlo estimators, e.g. in Del Moral (2013, Section 15.3) under strong conditions.

The following proposition extends Theorem 2.9 to the heavy-tailed regime $p \in (1, 2)$.

Proposition 2.10. *Assume that there exist $p \in (1, 2)$ and $r \in [1, \infty]$ such that $q(\omega^p) < \infty$ and $q(|f|^r) < \infty$, and such that $\alpha := pr/(p + r) > 1$. Then for any $1 \leq s \leq \alpha$ and any $N \geq 1$,*

$$\mathbb{E}_{\mathbf{X}^N \sim \bar{q}} \left[\left| \hat{F}_N(\mathbf{X}^N) - \pi(f) \right|^s \right] \leq C\, N^{-s(\alpha-1)/\alpha},$$

where the constant C depends on $r, p, s, q(|f|^r), q(\omega^p)$. When $r = \infty$, the statement holds for f such that $|f|_\infty < \infty$, giving $\alpha = p$ and the rate $N^{-s(p-1)/p}$ for all $1 \leq s \leq p$.

The proof is given in Appendix A.2.3.

2.3 Optimality of coupling IMH with common draws

With a view toward deriving upper bounds on the total variation distance of IMH to stationarity, we consider the common draws (or *common random numbers*) coupling of a generic IMH algorithm, described in Algorithm 4, and PIMH is retrieved as a special case. In this section we denote a state variable by X instead of the notation $\mathbf{X}^N$ employed in Algorithm 3, to simplify the notation, given that the results apply to both IMH and PIMH and for any N. We also write $\hat{Z}(X)$ instead of $\hat{Z}_N(X)$. The coupling is very simple and was considered in Liu (1996); Roberts & Rosenthal (2011). The pseudocode describes the transition kernel $\bar{P}((X, Y), \cdot)$ of the coupled chains, and we denote the transition of IMH by P. It was remarked around Lemma 1 in Wang et al. (2021) that this coupling is "one-step maximal": the meeting probability $\bar{P}((x, y), D)$, with $D = \{(x', y') : x' = y'\}$, is maximal over all couplings of $P(x, \cdot)$ and $P(y, \cdot)$, and hence equals $1 - |P(x, \cdot) - P(y, \cdot)|_{\mathrm{TV}}$. A direct computation gives the identity

$$\bar{P}((x, y), D) = 1 - |P(x, \cdot) - P(y, \cdot)|_{\mathrm{TV}} = \int \min \left\{ \frac{\hat{Z}(x^\star)}{\hat{Z}(x)}, \frac{\hat{Z}(x^\star)}{\hat{Z}(y)}, 1 \right\} \bar{q}(\mathrm{d}x^\star). \quad (2.18)$$

Algorithm 4. Common-draws coupling $\bar{P}$ of IMH at (X, Y)

1. Draw $X^\star \sim \bar{q}$.
2. Draw U from a Uniform$(0, 1)$ distribution, independently of $X^\star$.
3. If $U < \hat{Z}(X^\star)/\hat{Z}(X)$, set $X' = X^\star$, otherwise set $X' = X$.
4. If $U < \hat{Z}(X^\star)/\hat{Z}(Y)$, set $Y' = X^\star$, otherwise set $Y' = Y$.
5. Return (X', Y').

Let (X_t, Y_t) be a coupled chain started from (x, y) at time zero, and evolving according to $\bar{P}$. Denoting the meeting time by

$$\tau = \inf\{t \geq 1 : X_t = Y_t\}, \tag{2.19}$$

the coupling inequality states that, for $t \geq 1$,

$$|P^t(x, \cdot) - P^t(y, \cdot)|_{\text{TV}} \leq \mathbb{P}_{x,y}(\tau > t), \tag{2.20}$$

where the probability $\mathbb{P}_{x,y}$ is under the law of (X_t, Y_t) started from (x, y) at time zero. We will relate the probability $\mathbb{P}_{x,y}(\tau > t)$ to the rejection probabilities of IMH from x and y, and we define

$$r : x \mapsto \int_{x^\star \neq x} (1 - \alpha(x, x^\star))\, \bar{q}(\mathrm{d}x^\star), \tag{2.21}$$

where $\alpha(x, x^\star)$ is the acceptance probability defined in (2.12).

The meeting time τ is the first time at which both chains accept the proposal simultaneously, which corresponds to the first time at which the chain with the highest weight accepts the proposal. Indeed, if $\hat{Z}(x) \geq \hat{Z}(y)$, then $\alpha(x, y) \leq \alpha(y, x)$ for all x, y, and thus $u < \alpha(x, y)$ implies that $u < \alpha(y, x)$. Thus, conditionally on $x_0 = x, y_0 = y$, the meeting time τ follows a Geometric distribution with parameter $1-r(x)$, where $r(x)$ is defined in (2.21). Recall that the survival function of a Geometric variable T with parameter γ is given by: $\mathbb{P}(T > t) = (1-\gamma)^t$ for $t \in \mathbb{N}$. Still assuming $\hat{Z}(x) \geq \hat{Z}(y)$, we obtain, for $t \geq 1$,

$$|P^t(x, \cdot) - P^t(y, \cdot)|_{\text{TV}} \leq \mathbb{P}_{x,y}(\tau > t) = (r(x))^t. \tag{2.22}$$

The above upper bound is given in Roberts & Rosenthal (2011). In their remark following Theorem 5, they state that this is also a lower bound without providing a proof. We do so below, for both discrete and continuous state spaces; Roberts & Rosenthal (2011) focus on non-atomic spaces. First, we express

$$|P^t(x, \cdot) - P^t(y, \cdot)|_{\text{TV}} = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |P^t(x, A) - P^t(y, A)|, \tag{2.23}$$

and we select the set $A = \mathbb{R}^d \setminus \{x\}$ to obtain a lower bound. Then $P^t(y, A) = 1$ since $x \neq y$, assuming that $q(\{x\}) = 0$, while $P^t(x, A) = 1 - (r(x))^t$, i.e. the chain is in A at step t except if t proposals have been rejected. The situation is slightly more complicated if the proposal has non-zero mass on $\{x\}$ and $\{y\}$, i.e. in discrete state spaces, but the following result still holds. The proof is in Appendix A.3.

Theorem 2.11. *Let (X_t, Y_t) be a Markov chain evolving according to $\bar{P}$ in Algorithm 4 and starting from $X_0 = x$ and $Y_0 = y$. Let $\tau = \inf\{t \geq 1 : X_t = Y_t\}$, and let $r(x)$ be defined as in* (2.21)*. Then, under Assumption 2.1, for all $t \geq 1$,*

$$|P^t(x, \cdot) - P^t(y, \cdot)|_{\mathrm{TV}} = \mathbb{P}_{x,y}(\tau > t) = \max(r(x), r(y))^t. \tag{2.24}$$

Thus, the chain (X_t, Y_t) generated by the common draws coupling follows a *maximal coupling*, as in Pitman (1976): the coupling inequality is an equality at all times t. To the best of our knowledge, this is the only known case of "all time maximal" couplings of an MCMC algorithm. This peculiarity could prove useful when considering the problem of numerically solving for the optimal coupling between Markov chains, as in the literature on bicausal optimal transport (Moulos, 2021) and bisimulation metrics (Calo et al., 2024). The case of IMH provides a rare case of an explicit solution, namely the common draws coupling, which could serve as a benchmark for algorithmic approaches to the problem. Note that the upper bound in (2.24) decreases geometrically in t; the polynomial rates that come later come from the integration with respect to x or y.

2.4 Meeting times and polynomial convergence

2.4.1 Meeting times of lagged chains

We consider coupled IMH chains with a lag, as in Middleton et al. (2019). The construction is described in Algorithm 5. Compared to Section 2.3, note the redefinition of the meeting time τ, which now corresponds to $\inf\{t \geq 1 : \mathbf{X}_t^N = \mathbf{Y}_{t-1}^N\}$. The generated chains $(\mathbf{X}_t^N)_{t\geq 0}$ and $(\mathbf{Y}_t^N)_{t\geq 0}$ have the same marginal distribution, that of an IMH chain started from $\bar{q}$. Recall that we denote by P the transition kernel of the particle IMH chain in Algorithm 3, and by $\bar{P}$ the transition kernel of the coupled PIMH chain in Algorithm 4, both of which depend on the parameter N.

We relate the distribution of the meeting times generated by Algorithm 5 to the expected rejection probability in the following result, proved in Appendix A.4.1.

Proposition 2.12. *Consider τ generated by Algorithm 5. Under Assumption 2.1, for all $t \geq 1$,*

Algorithm 5. Coupled PIMH with a lag

1. Set $\tau = +\infty$ and $t = 1$.
2. Draw $\mathbf{X}_0^N \sim \bar{q}$ and $\mathbf{Y}_0^N \sim \bar{q}$ independently.
3. Draw U from a Uniform$(0,1)$ distribution, independently of $\mathbf{X}_0^N$ and $\mathbf{Y}_0^N$.
4. If $U < \hat{Z}_N(\mathbf{Y}_0^N)/\hat{Z}_N(\mathbf{X}_0^N)$, set $\mathbf{X}_1^N = \mathbf{Y}_0^N$, $\tau = 1$. Otherwise, set $\mathbf{X}_1^N = \mathbf{X}_0^N$.
5. While $\tau = +\infty$:
   (a) Sample $(\mathbf{X}_{t+1}^N, \mathbf{Y}_t^N) \sim \bar{P}((\mathbf{X}_t^N, \mathbf{Y}_{t-1}^N), \cdot)$, the common draws coupling of PIMH in Algorithm 4.
   (b) If $\mathbf{X}_{t+1}^N = \mathbf{Y}_t^N$, set $\tau = t + 1$.
   (c) Set $t = t + 1$.
6. Return τ and the generated chains up to time τ.

with r defined in (2.21)*, we have*

$$\mathbb{P}(\tau > t) \leq \mathbb{E}_{\bar{q}}\left[\left(r(\mathbf{X}^N)\right)^t\right]. \tag{2.25}$$

This connection between meeting times and expected rejection probability motivates our next result that bounds the latter. That rejection probability is known to be central in the study of IMH, e.g. Theorem 6 in Roberts & Rosenthal (2011). Our bounds are explicit functions of t and N. From this point onward we focus on $p \geq 2$, postponing the case $p \in (1,2)$ to Section 2.4.3.

Proposition 2.13. *Fix $p \geq 2$ and let*

$$\beta_p := 1 - \frac{1}{2^{\frac{3p-2}{p-1}} q(\omega^p)^{\frac{1}{p-1}}}. \tag{2.26}$$

Under Assumptions 2.1-2.2, $\beta_p \in (0,1)$ and there exist finite constants $A_p, C_p > 0$, depending only on p and $q(\omega^p)$, such that for all $N \geq 1$, for all $t \geq 1$, the following holds:

$$\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t\right] \leq \frac{A_p}{N^{(t\wedge p)/2}}\beta_p^t + \frac{C_p}{t^p N^{p/2}}. \tag{2.27}$$

Proposition 2.13, proved in Appendix A.4.2, holds for all $t \geq 1$ and all $N \geq 1$. The bounds decay to 0 as either N or t approaches infinity, polynomially with rate at most $N^{-1/2}$ with respect to N, and, for fixed N, polynomially with rate t^{-p}, where p is the number of assumed moments of ω under q. A direct consequence of the previous two propositions is a bound on the tails of the meeting times.

Proposition 2.14. *Consider τ generated by Algorithm 5. Under Assumptions 2.1-2.2, there*

exists a finite $C > 0$ *such that for all* $N \geq 1$ *and all* $t \geq 1$*, if* $p \geq 2$ *in Assumption 2.2,*

$$\mathbb{P}(\tau > t) \leq \frac{C}{\sqrt{N} t^p}. \tag{2.28}$$

As a consequence, we have $\mathbb{E}[\tau] \leq 1 + C'/\sqrt{N}$ *with* $C' = C \sum_{t \geq 1} t^{-p}$.

That bound retains the slowest rates in N and t from the previous result. Proposition 2.14 is consistent with Proposition 8 in Middleton et al. (2019), which showed that $\mathbb{P}(\tau = 1)$ approaches one as $N \to \infty$ under the assumption of bounded weights. However, our present assumptions are considerably weaker, and we provide explicit dependencies on both N and t.

Remark 2.15. We comment on the sharpness of the dependency on N in Proposition 2.14. For $t = 1$, the result reads $\mathbb{P}(\tau > 1) \leq C/\sqrt{N}$. The event $\{\tau > 1\}$ corresponds to the rejection of $\mathbf{X}^{N,\star}$ from a state $\mathbf{X}^N$, both $\mathbf{X}^N, \mathbf{X}^{N,\star}$ being independent draws from $\bar{q}$. Here we show that we cannot improve upon the rate $N^{-1/2}$ as a function of N. The central limit theorem implies $\sqrt{N}(\hat{Z}_N(\mathbf{X}^N) - 1) \to \text{Normal}(0, q(\omega^2) - 1)$ in distribution. Therefore, $\mathbb{P}(\hat{Z}_N(\mathbf{X}^N) \geq 1 + N^{-1/2}) \to p_0$ as $N \to \infty$, with p_0 depending on $q(\omega^2)$. The same argument shows $\mathbb{P}(\hat{Z}_N(\mathbf{X}^{N,\star}) \leq 1 - N^{-1/2}) \to p_1$ as $N \to \infty$, with p_1 depending on $q(\omega^2)$. Therefore, we can choose a large enough N that depends on $q(\omega^2)$ such that $\mathbb{P}(\hat{Z}_N(\mathbf{X}^N) \geq 1 + N^{-1/2}) \geq p_0/2$ and $\mathbb{P}(\hat{Z}_N(\mathbf{X}^{N,\star}) \leq 1 - N^{-1/2}) \geq p_1/2$. Thus, with a constant probability c, $\hat{Z}_N(\mathbf{X}^{N,\star}) \leq 1 - N^{-1/2}$ and $\hat{Z}_N(\mathbf{X}^N) \geq 1 + N^{-1/2}$ occur simultaneously, and thus the acceptance probability is at most $(1 - N^{-1/2})/(1 + N^{-1/2}) \leq 1 - N^{-1/2}$. In turn this means that the rejection probability is at least $cN^{-1/2}$.

2.4.2 Polynomial convergence rates

As discussed in Section 6 of Jacob et al. (2020) and in Biswas et al. (2019), lagged chains such as those generated by Algorithm 5 can be employed to bound the total variation distance between the chain at time t and its stationary distribution. We state upper bounds on the distance of IMH chains to stationarity that are explicit in their dependency on t and N. We present the following result.

Theorem 2.16. *Consider* τ *generated by Algorithm 5. Let* P *be the transition kernel of the PIMH chain as in Algorithm 3. Under Assumption 2.2 with* $p \geq 2$*, we have that for all* $t \geq 0$*,*

$$\left| \bar{q} P^t - \bar{\pi} \right|_{\text{TV}} \leq \mathbb{E}\left[\max\left(0, \tau - 1 - t\right) \right]. \tag{2.29}$$

Furthermore, still under Assumption 2.2 with $p \geq 2$*, there exists a constant* C*, independent of* t

and N, such that for all $N \geq 1$ and $t \geq 0$,

$$\left|\bar{q}P^t - \bar{\pi}\right|_{\mathrm{TV}} \leq \frac{C}{\sqrt{N}(1+t)^{p-1}}. \tag{2.30}$$

Remark 2.17. The case $t = 0$ states that $|\bar{q} - \bar{\pi}|_{\mathrm{TV}} \leq CN^{-1/2}$, despite both $\bar{\pi}$ and $\bar{q}$ being defined in (2.9)-(2.10) on spaces growing with N. With the density representation of the total variation distance and $\bar{\pi}(\mathbf{x}^N) = \hat{Z}_N(\mathbf{x}^N)\bar{q}(\mathbf{x}^N)$, we can directly compute

$$|\bar{q} - \bar{\pi}|_{\mathrm{TV}} = \frac{1}{2}\int |1 - \hat{Z}_N(\mathbf{x}^N)|\bar{q}(\mathrm{d}\mathbf{x}^N) \leq \frac{1}{2}\mathbb{E}_{\bar{q}}\left[|1 - \hat{Z}_N(\mathbf{X}^N)|^2\right]^{1/2} \leq \frac{1}{2}M(2)N^{-1/2}, \tag{2.31}$$

where the first inequality is Cauchy–Schwarz and the second uses Proposition 2.5 under Assumption 2.2 with $p \geq 2$, the function $M(\cdot)$ being defined in that proposition. Furthermore, in the large N asymptotics we expect $1 - \hat{Z}_N(\mathbf{X}^N)$ to behave as a Normal random variable with mean zero and standard deviation $\sqrt{q(\omega^2) - 1}/\sqrt{N}$, so that the expectation of its absolute value should indeed behave as $\sqrt{(2/\pi)(q(\omega^2) - 1)}/\sqrt{N}$.

We also state a bound for a PIMH chain started at any initial point $\mathbf{x}^N \in \mathbb{X}^N$.

Corollary 2.18. *Under Assumptions 2.1-2.2, with $p \geq 2$, there exists a constant $\tilde{C}$, independent of t and N, such that for all $N \geq 1$, $t \geq 1$, and any starting point $\mathbf{x}^N \in \mathbb{X}^N$,*

$$\left|P^t(\mathbf{x}^N, \cdot) - \bar{\pi}\right|_{\mathrm{TV}} \leq (r(\mathbf{x}^N))^t + \frac{\tilde{C}}{\sqrt{N}\, t^{p-1}}. \tag{2.32}$$

Here $r(\cdot) \in (0, 1)$ is the rejection probability defined in (2.21)*.*

Theorem 2.16 and Corollary 2.18, proven in Appendix A.4.3, provide explicit bounds on the convergence rate of the PIMH algorithm. Both results are interpretable in terms of the number of iterations t and the number of particles N, and apply to IMH as a special case when $N = 1$. The difference between these results lies in the starting distribution. Practitioners would typically start the algorithm from the proposal distribution, as it is the best available approximation of the target. Corollary 2.18 bounds the distance from a fixed starting point by a sum of two terms: an exponential term $r(\mathbf{x}^N)^t$, which decays in t but does not improve with N, and a polynomial term decaying in both t and N. Which of the two dominates depends on the starting point and on N. When $r(\mathbf{x}^N)$ is close to one, the exponential term controls the bound for small t and the polynomial term takes over as t grows; when $r(\mathbf{x}^N)$ is small, or when N is small enough that $\tilde{C}/\sqrt{N}$ already exceeds $r(\mathbf{x}^N)$, the polynomial term dominates from the outset.

Remark 2.19. The weight ω can be unbounded while having infinitely many moments under

q, i.e. $q(\omega^p) < \infty$ for all $p \geq 1$. For example, this happens when π is Gamma$(2,1)$ and q is Exponential(1), leading to $\omega(x) = x$. In that case Mengersen & Tweedie (1996) prove that the IMH chain cannot be geometrically ergodic, while Corollary 2.22 holds with any $p > 1$. Indeed the actual decay of $|P^t(x,\cdot) - \pi|_{\mathrm{TV}}$ could be between geometric and polynomial in t, for example of the form $\exp(-t^{1/2})$.

2.4.3 Case when the weight has less than two moments

We next study PIMH in the heavy-tail regime where $q(\omega^p) < \infty$ for $p \in (1,2)$; the case $p \geq 2$ was treated in the previous sections. The next two results show that the polynomial rate in t remains $t^{-(p-1)}$, but the dependence on the number of particles deteriorates from $N^{-1/2}$ to $N^{-(p-1)/p}$.

Theorem 2.20. *Consider τ generated by Algorithm 5. Let P be the transition kernel of the PIMH chain as in Algorithm 3. Under Assumption 2.1, and assuming that $q(\omega^p) < \infty$ for some $p \in (1,2)$, there exists a constant C, independent of t and N, such that for all $N \geq 1$ and $t \geq 0$,*

$$\left|\bar{q}P^t - \bar{\pi}\right|_{TV} \leq \frac{C}{N^{(p-1)/p}(1+t)^{p-1}}. \tag{2.33}$$

Corollary 2.21. *Under Assumption 2.1, and assuming that $q(\omega^p) < \infty$ for some $p \in (1,2)$, there exists a constant $\tilde{C}$, independent of t and N, such that for all $N \geq 1$, $t \geq 0$, and any starting point $\mathbf{x}^N \in \mathbb{X}^N$,*

$$\left|P^t(\mathbf{x}^N,\cdot) - \bar{\pi}\right|_{TV} \leq (r(\mathbf{x}^N))^t + \frac{\tilde{C}}{N^{(p-1)/p}(1+t)^{p-1}}. \tag{2.34}$$

Theorem 2.20 and Corollary 2.21 show that the same polynomial rate in t continues to hold in the heavy-tail regime $p \in (1,2)$, but the effect of increasing N is weaker. The proofs are in Appendix A.4.4.

The case $N = 1$ corresponds to standard IMH. Combining Corollaries 2.18 and 2.21, the rate $t^{-(p-1)}$ holds for every $p > 1$ as we summarize in the following corollary.

Corollary 2.22. *Consider IMH under Assumption 2.1, and Assumption 2.2 for any $p > 1$. There exists $D < \infty$ independent of t such that for all $t \geq 1$, and any starting point $x \in \mathbb{X}$,*

$$\left|P^t(x,\cdot) - \pi\right|_{TV} \leq (r(x))^t + \frac{D}{t^{p-1}}. \tag{2.35}$$

As above, $r(\cdot) \in (0,1)$ is defined in (2.21).

We provide a direct proof in Appendix A.4.4, which may be of independent interest. That proof is closer in spirit to existing analyses of IMH mentioned in Section 2.4.5.

2.4.4 Lower bounds on the convergence rate of IMH

The purpose of the following example is to demonstrate that the rate $t^{-(p-1)}$ in Corollary 2.22 cannot be improved beyond polylogarithmic factors, without further assumptions. The proof is in Section A.4.5.

Example 2.23. Fix any $p > 1$. Let $k \geq 1$ be such that $k > 1/(p-1)$. Consider the IMH algorithm targeting $\pi(x) := Z_\pi x^{-p}$ on $[2, \infty)$, with proposal distribution $q(x) := Z_q \log^k(x) x^{-(p+1)}$ on $[2, \infty)$, started from $x_0 = 3$. Then $q(\omega^p) < \infty$, and there exist $C < \infty$ and $t_0 \in \mathbb{N}$ such that, for all $t \geq t_0$,

$$\left| P^t(x_0, \cdot) - \pi \right|_{\text{TV}} \geq \frac{C}{t^{p-1} (\log t)^{(k+1)(p-1)}}. \tag{2.36}$$

2.4.5 Related results on IMH

The convergence of IMH has garnered significant interest over decades, and in particular the sub-geometric rates have been studied in several works including Jarner & Roberts (2002); Douc et al. (2007); Roberts & Rosenthal (2011); Andrieu et al. (2022). One approach utilizes drift and minorization techniques (Jarner & Roberts, 2002).

Theorem 2.24 (Theorem 5.3 in Jarner & Roberts (2002))**.** *Let P be the transition kernel of the IMH chain as in Algorithm 2. Assume that for some $r > 0$,*

$$\pi(A_\epsilon) = \mathcal{O}\left(\epsilon^{1/r}\right) \quad \textit{for} \quad \epsilon \to 0, \tag{2.37}$$

where $A_\epsilon = \{x \in \mathbb{X} : \omega(x) > 1/\epsilon\}$, for any $\epsilon > 0$. Then, for any $x \in \mathbb{X}$, and any $t \geq 1$, we have that

$$\lim_{t \to \infty} (1+t)^\beta \left| P^t(x, \cdot) - \pi \right|_{\text{TV}} = 0, \tag{2.38}$$

for any $0 \leq \beta \leq (s-r)/r$, with $r < s < r+1$.

The $\mathcal{O}$ notation here is such that if $f(x) = \mathcal{O}(g(x))$ then there exists a constant M such that $|f(x)| \leq M|g(x)|$ for all x in the domain of f. Theorem 2.24 provides a polynomial rate of convergence for the IMH chain in total variation of order $o\left(t^{-1/r+\kappa}\right)$ for any $\kappa > 0$ under the assumption that the tail weights satisfy the condition specified in equation (2.37). Under the assumption $q(\omega^p) < \infty$, the condition in (2.37) is satisfied for $r = 1/(p-1)$, using Markov's inequality. Our Corollary 2.22 differs slightly as our bounds are in $t^{-(p-1)}$ instead of $t^{-(p-1)+\kappa}$ for some arbitrarily small $\kappa > 0$. Similar results can be obtained using weak Poincaré inequalities as described in Remark 29 of Andrieu et al. (2022), under $\pi(\omega^p)$ with $p > 1$ which amounts to our Assumption 2.2 with $p > 2$.

Our bounds in Theorem 2.16 and Corollary 2.18 have the advantage of providing an explicit dependency on N in the case of PIMH, which is critical for the results on bias removal in Section 2.5. Our results in Section 2.4.3 cover the case where $p \in (1, 2)$, in which polynomial convergence of IMH does not seem to have been studied before.

2.4.6 Bias comparison and practical recommendations

Results in Sections 2.2 and 2.4 enable a comparison of the bias of IS and IMH under Assumptions 2.1-2.2, addressing the question raised in Section 2.1.1. Performing N steps of IMH to obtain a sample X_N, and considering a test function f bounded by one, Corollary 2.22 shows that the bias of $f(X_N)$ relative to $\pi(f)$ is at most $C\, N^{-(p-1)}$. In terms of variance, however, it is preferable to average over successive states of the chain. We recommend setting a burn-in period $T = \lceil cN^\eta \rceil$ with $\eta \in (0, 1)$ and $c > 0$, and averaging over $X_{T+1}, \dots, X_N$. We can then bound the bias of the average by

$$\frac{1}{N-T} \sum_{t=T+1}^{N} \left| qP^t f - \pi(f) \right| \leq \frac{C}{N-T} \sum_{t=T+1}^{\infty} t^{-(p-1)} = \mathcal{O}\left(N^{-1-\eta(p-2)}\right),$$

using $\sum_{t>T} t^{-(p-1)} = \mathcal{O}(T^{-(p-2)})$ as in (A.20) in the proof of Theorem 2.16 and $N - T \sim N$. For $\eta < 1$ the burn-in phase is negligible in the limit, and thus the ergodic average has the same asymptotic variance as if no burn-in were used. A value $\eta = 1$ makes the bias decrease to $N^{-(p-1)}$, but the asymptotic variance is inflated by $1/(1-c)$.

On the other hand, Theorem 2.8 shows that the bias of the SNIS estimator with N samples is upper bounded by $C' N^{-1}$ for some constant C', under an additional inverse weight assumption. Therefore, as soon as $q(\omega^p) < \infty$ with $p > 2$, the bias of IMH becomes smaller than that of SNIS for a similar budget of weight function evaluations. Our results also show that, in terms of bias and for a fixed total budget, the best choice for the number of proposals per iteration in PIMH is one. If the weights are bounded, MCMC methods such as PIMH or particle Gibbs (Andrieu et al., 2010) are geometrically ergodic (e.g. Wang, 2022; Lee et al., 2020; Cardoso et al., 2022), and thus the bias decays geometrically. The bias comparison is then clearly in favor of MCMC algorithms.

We therefore recommend considering IMH as a potential alternative to SNIS when low-bias estimators are desired, see e.g., Section 2.1.1 for examples of such settings such as gradient estimation or nested expectations. The tail behavior of the weight function is typically unknown in practice: the user does not know the largest value of p such that $q(\omega^p) < \infty$, and thus does not know the rate at which the bias of IMH decreases. This can be problematic: for example in the nested expectation setting of Andradóttir & Glynn (2016) recalled in Section 2.1.1, an optimal budget allocation requires knowledge of the bias of the inner expectation estimator. To address this, one approach would be to estimate the largest value of p such that $q(\omega^p) < \infty$. This is

a hard but well-known tail estimation problem, which has been considered in the context of importance sampling in (e.g. Monahan, 1993; Koopman et al., 2009; Vehtari et al., 2024). In Section 2.5, we pursue a different approach, which completely removes the bias without having to estimate the tail behavior of the weight. As we will see the bias removal technique (fully described in Algorithm 6) is generically applicable whenever SNIS is, and incurs a moderate increase in variance, which vanishes as the single tuning parameter N goes to infinity.

For completeness, we recall that various bias and variance reduction techniques are available for both IS and IMH. Bias reduction for IS includes the leave-one-out modification proposed in Skare et al. (2003), or the jackknife (Shao & Tu, 2012; Nowozin, 2018), both of which can reduce the bias to N^{-2}; we implement the former in Section 2.6.2, and we describe SNIS-LOO and show its bias is of order N^{-2} in Appendix D.4. Variance reduction for IS, beyond optimizing the choice of proposal distribution, includes Owen (2025); Henmi et al. (2007), and variance reduction for IMH includes Rao-Blackwellization of the acceptance step (Casella & Robert, 1996), averaging over different orderings of the proposed samples (Atchadé & Perron, 2005; Jacob et al., 2011), or, depending on the tractability of the proposal distribution, constructing control variates as in Neal & Pinto (2001); Alexopoulos et al. (2023).

2.5 Bias removal for self-normalized importance sampling

2.5.1 Motivation

We consider techniques to remove the bias of IS entirely. There exist at least three distinct methods to achieve this. We focus on the approach of Middleton et al. (2019), that follows the pioneering work of Glynn & Rhee (2014), and relies on couplings of PIMH chains as in Algorithm 5. We refer to this approach as Coupling-based Unbiased Importance Sampling (Coupling-UIS). Another approach is based on randomly-truncated multilevel Monte Carlo (MLMC) and was proposed in the context of SNIS by Blanchet & Glynn (2015); Shi & Cornish (2021). We refer to it as MLMC-UIS. Finally, any approach that provides unbiased estimators of $1/q(\omega)$ can be used to remove the bias of IS, since multiplying an unbiased estimator of $1/q(\omega)$ by an independent unbiased estimator of $q(\omega f)$ yields an unbiased estimator of $\pi(f)$; here we understand ω as being possibly unnormalized. In particular, an approach based on the Taylor expansion of the inverse function was proposed in Blanchet et al. (2015); Chopin et al. (2025). We refer to this approach as Taylor-UIS. Both MLMC-UIS and Taylor-UIS are described in Appendix D.1. Conversely, using the test function $f : x \mapsto \omega^{-1}(x)$, the Coupling-UIS estimator of $\pi(f)$ provides an unbiased estimator of $1/q(\omega)$; see Section 2.5.6 and Section 2.6.

We propose conditions under which Coupling-UIS has finite moments, we derive a Rao–Blackwellization of the scheme for increased efficiency, and we provide conditions under which

its asymptotic efficiency matches that of IS as $N \to \infty$ (see Proposition 2.32). The two other approaches, MLMC-UIS and Taylor-UIS, do not seem to match this asymptotic efficiency, as we observe in Section 2.6.1.

2.5.2 Construction

For bias removal, Middleton et al. (2019) employ common random numbers couplings of PIMH within the approach of Glynn & Rhee (2014). Upon running Algorithm 5 with $N \geq 1$, with $\tau = \inf\{t \geq 1 : \mathbf{X}_t^N = \mathbf{Y}_{t-1}^N\}$, one can compute the following unbiased estimator of $\pi(f)$:

$$\hat{F}_N^u = \hat{F}_N(\mathbf{X}_0^N) + \sum_{t=1}^{\tau-1} \{\hat{F}_N(\mathbf{X}_t^N) - \hat{F}_N(\mathbf{Y}_{t-1}^N)\}, \tag{2.39}$$

where $\hat{F}_N$ is as in (2.5). By convention the sum in (2.39) is zero in the event $\{\tau = 1\}$, and it is also equal to the infinite sum $\sum_{t=1}^{\infty} \{\hat{F}_N(\mathbf{X}_t^N) - \hat{F}_N(\mathbf{Y}_{t-1}^N)\}$ since $\hat{F}_N(\mathbf{X}_t^N) = \hat{F}_N(\mathbf{Y}_{t-1}^N)$ from time τ onward. The lack of bias can be seen via a telescopic sum argument, since $\mathbf{X}_t^N$ and $\mathbf{Y}_t^N$ have the same marginal distribution for all t, provided that limit and expectation can be swapped, and using (2.13). Since $\mathbf{X}_0^N \sim \bar{q}$, $\hat{F}(\mathbf{X}_0^N)$ is the (biased) SNIS estimator, so that the sum in (2.39) can be seen as a bias cancellation term, under some conditions. Middleton et al. (2019) consider the case where ω is uniformly upper bounded, and show that (2.39) can have a finite variance. Below we work under much weaker assumptions and bound the moments of these unbiased estimators, highlighting the effect of the tuning parameter N, and comparing its efficiency with that of SNIS.

Remark 2.25. The estimator $\hat{F}_N^u$ in (2.39) is an unbiased MCMC estimator whose efficiency is usually compared to that of the underlying MCMC algorithm (Jacob et al., 2020; Atchadé & Jacob, 2026), i.e. PIMH in the present case. However, it can also be compared to that of SNIS as $N \to \infty$, which is more compelling, as the asymptotic variance of SNIS is smaller than that of IMH (2.7).

2.5.3 Moments of Coupling-UIS

We subtract $\pi(f)$ from all terms in (2.39) to obtain

$$\hat{F}_N^u - \pi(f) = \hat{F}_N(\mathbf{X}_0^N) - \pi(f) + \sum_{t=1}^{\infty} \{\hat{F}_N(\mathbf{X}_t^N) - \hat{F}_N(\mathbf{Y}_{t-1}^N)\} \mathbb{1}(\tau > t). \tag{2.40}$$

We introduce the notation

$$\Delta_t = \hat{F}_N(\mathbf{X}_t^N) - \hat{F}_N(\mathbf{Y}_{t-1}^N), \quad \mathrm{BC} = \sum_{t=1}^{\infty} \Delta_t \mathbb{1}(\tau > t), \tag{2.41}$$

where BC stands for the bias cancellation term. Using Minkowski's inequality, the moments of the error of $\hat{F}_N^u$ can be bounded by the moments of the error of the IS estimator $\hat{F}_N(\mathbf{X}_0^N)$, as in Theorem 2.9, and the moments of BC.

A first result is that, for bounded test functions f, $\hat{F}_N^u$ has as many moments as the meeting time τ, which is up to p (non-included) under Assumption 2.2. The proof is in Appendix A.5.1.

Proposition 2.26. *Assume that $|f|_\infty \leq 1$ and let $s \geq 1$. If Assumptions 2.1-2.2 hold with $p > 1$ and $p > s$, then τ has s finite moments, and the unbiased self-normalized importance sampling (Coupling-UIS) estimator $\hat{F}_N^u$ in* (2.39) *has s finite moments for any $N \geq 1$.*

To deal with unbounded test functions, we need to control the moments of the terms Δ_t. For this we derive the following result about PIMH at any iteration t, under the same conditions as Theorem 2.9. The proof is in Appendix A.5.2. Here and in the rest of this section we focus on the case $p \geq 2$ for simplicity, but the result could be extended to $p > 1$ with a different dependency on N.

Proposition 2.27. *Assume that there exist $p \in [2, \infty)$ and $r \in [2, \infty]$ such that $q(\omega^p) < \infty$ and $q(|f|^r) < \infty$, and $q(f^2 \cdot \omega^2) < \infty$. Let $(\mathbf{X}_t^N)$ be the PIMH chain started from $\mathbf{X}_0^N \sim \bar{q}$. Then, for any $2 \leq s \leq pr/(p + r + 2)$ and any $N \geq 1$, there exists C such that for all $t \geq 0$:*

$$\mathbb{E}_{\mathbf{X}_0^N \sim \bar{q}} \left[\left| \hat{F}_N(\mathbf{X}_t^N) - \pi(f) \right|^s \right] \leq C N^{-s/2},$$

where the constant C depends on $r, p, s, q(|f|^r), q(\omega^p), q(f^2 \cdot \omega^2)$. When $r = \infty$, the statement holds for f such that $|f|_\infty < \infty$ and all $s \leq p$.

By Minkowski's inequality, under the conditions of Proposition 2.27, the moments of Δ_t have similar bounds. This can be used to obtain the following result, proven in Appendix A.5.3.

Proposition 2.28. *Assume that there exist $p \in (2, \infty)$ and $r \in [2, \infty]$ such that $q(\omega^p) < \infty$ and $q(|f|^r) < \infty$, and $q(f^2 \cdot \omega^2) < \infty$, then for any $2 \leq s < p$ such that $pr/(p + r + 2) > ps/(p - s)$, and for any $N \geq 1$, the unbiased importance sampling (Coupling-UIS) estimator satisfies:*

$$\mathbb{E} \left[\left| \hat{F}_N^u - \pi(f) \right|^s \right] \leq C N^{-s/2},$$

where the constant C depends on $r, p, s, q(|f|^r), q(\omega^p), q(f^2 \cdot \omega^2)$. When $r = \infty$, the statement holds for f such that $|f|_\infty < \infty$ and all $s < p$ such that $p > sp/(p - s)$.

Remark 2.29. The construction of Coupling-UIS in (2.39) can be generalized to the case of i.i.d. pairs of variables (Ω_n, F_n) that jointly follow a distribution p on $\mathbb{R}_+ \times \mathbb{R}$. In that case a similar construction allows the unbiased estimation of $\mathbb{E}[\Omega \cdot F]/\mathbb{E}[\Omega]$. As noted in Blanchet &

Glynn (2015, Section 4), this includes steady-state regenerative simulation.

2.5.4 The asymptotic price of bias removal

We consider the price of debiasing self-normalized importance sampling using Coupling-UIS in terms of inefficiency, defined by the mean squared error multiplied by the average cost (see e.g. Glynn & Whitt, 1992). We start by comparing the mean squared errors of unbiased and regular IS. From (2.40), squaring and using Cauchy–Schwarz, we obtain

$$\begin{aligned}&\left|\mathbb{E}\left[(\hat{F}_N^u - \pi(f))^2\right] - \mathbb{E}\left[(\hat{F}_N(\mathbf{X}_0^N) - \pi(f))^2\right]\right| \\ &\leq 2\sqrt{\mathbb{E}\left[(\hat{F}_N(\mathbf{X}_0^N) - \pi(f))^2\right] \cdot \mathbb{E}[\mathrm{BC}^2]} + \mathbb{E}[\mathrm{BC}^2],\end{aligned} \tag{2.42}$$

with $\mathrm{BC} = \sum_{t=1}^{\tau-1}\{\hat{F}_N(\mathbf{X}_t^N) - \hat{F}_N(\mathbf{Y}_{t-1}^N)\}$.

The mean squared error (MSE) of IS, which is the term $\mathbb{E}[(\hat{F}_N(\mathbf{X}_0^N) - \pi(f))^2]$, is of order N^{-1} under conditions stated in Theorem 2.9. If we can bound $\mathbb{E}[\mathrm{BC}^2]$ by a term that decreases faster than N^{-1}, then the MSE of Coupling-UIS would be asymptotically equivalent to that of IS. Intuitively, the bias cancellation term BC in (2.41) goes to zero for two reasons: first because τ goes to one as $N \to \infty$, since two independent draws of $\hat{Z}_N(\mathbf{X}^N)$ are more likely to take similar values as N increases. Furthermore, each term $\Delta_t = \hat{F}_N(\mathbf{X}_t^N) - \hat{F}_N(\mathbf{Y}_{t-1}^N)$ goes to 0 as $N \to \infty$, as described in Proposition 2.27. We obtain the following result, proven in Appendix A.5.4.

Proposition 2.30. *Let $\hat{F}_N^u$ be the Coupling-UIS estimator* (2.39) *and $\hat{F}_N(\mathbf{X}^N)$ with $\mathbf{X}^N \sim \bar{q}$ be the IS estimator. Suppose that the assumptions of Proposition 2.28 are satisfied with $s = 2$, that is: $p > 2$ and $r > 2$ such that $q(\omega^p) < \infty$, $q(|f|^r) < \infty$, and $q(f^2 \cdot \omega^2) < \infty$, with $2p + 4r + 4 < r \cdot p$. Then the mean squared errors of $\hat{F}_N^u$ and $\hat{F}_N(\mathbf{X}^N)$ are asymptotically equivalent:*

$$\lim_{N\to\infty} N \cdot \mathbb{E}\left[(\hat{F}_N^u - \pi(f))^2\right] = \lim_{N\to\infty} N \cdot \mathbb{E}_{\mathbf{X}^N\sim\bar{q}}\left[(\hat{F}_N(\mathbf{X}^N) - \pi(f))^2\right].$$

The assumption $2p + 4r + 4 < r \cdot p$ is for example satisfied if $r = \infty$ and $p = 4 + \epsilon$ with an arbitrary $\epsilon > 0$, or if $p = 5$ and $r = 15$. However it cannot be satisfied with $p \leq 4$.

The cost of Coupling-UIS is that of running Algorithm 5. It starts with two draws from $\bar{q}$, i.e. $2N$ draws from q, and as many evaluations of the weight function ω. Then either $\tau = 1$ or the algorithm enters its while loop up to the meeting time τ, drawing N new particles at each iterate of the loop. Counting the cost in units of number of evaluations of π, Coupling-UIS has an overall cost of $\mathcal{C} = 2N + N(\tau - 1)$. If Assumption 2.2 holds with $p \geq 2$, using Proposition 2.14 then $\mathbb{E}[\tau - 1] \leq CN^{-1/2}$ for a finite constant C. Thus, as $N \to \infty$, $\mathbb{E}[\mathcal{C}]$ is equivalent to $2N$. We can then compare the asymptotic inefficiencies of Coupling-UIS and IS

as in the next statement.

Proposition 2.31. *Denote the cost of the Coupling-UIS estimator $\hat{F}_N^u$ in* (2.39) *by $\mathcal{C} = 2N + N(\tau - 1)$. Under the conditions of Proposition 2.30, the inefficiency, defined as expected cost multiplied by mean squared error, of Coupling-UIS is twice that of IS as $N \to \infty$.*

The reason for the efficiency loss in Coupling-UIS is that, in Algorithm 5 the N particles in $\mathbf{Y}_0^N$ are required to determine τ but in the event $\{\tau = 1\}$, which is increasingly likely as $N \to \infty$, these N particles do not participate directly in the estimator $\hat{F}_N^u$.

2.5.5 An improved coupling-based unbiased importance sampling estimator

A simple trick provides a remedy, and cuts the asymptotic inefficiency by a half. We can view $\hat{F}_N^u$ as a deterministic function of initial states $\mathbf{X}_0^N$ and $\mathbf{Y}_0^N$ drawn from $\bar{q}$ independently, as well as additional variables: an arbitrary long sequence of proposals from $\bar{q}$, and a sequence of uniform random variables used to accept or reject proposals, of the same length. Denote these additional variables by ζ. Then $\hat{F}_N^u$ can be written as $\hat{F}_N^u = \mathcal{A}_u(\mathbf{X}_0^N, \mathbf{Y}_0^N, \zeta)$, where $\mathcal{A}_u$ is a deterministic function. Then we define the estimator that averages over swapping $\mathbf{X}_0^N$ and $\mathbf{Y}_0^N$:

$$\tilde{F}_N^u = \frac{1}{2}\left(\mathcal{A}_u(\mathbf{X}_0^N, \mathbf{Y}_0^N, \zeta) + \mathcal{A}_u(\mathbf{Y}_0^N, \mathbf{X}_0^N, \zeta)\right). \tag{2.43}$$

Computing (2.43) only requires simple modifications of Algorithm 5. Indeed, either $\hat{Z}_N(\mathbf{X}_0^N) \geq \hat{Z}_N(\mathbf{Y}_0^N)$ or $\hat{Z}_N(\mathbf{X}_0^N) < \hat{Z}_N(\mathbf{Y}_0^N)$. In the first case, we always have $\mathcal{A}_u(\mathbf{Y}_0^N, \mathbf{X}_0^N, \zeta) = \hat{F}_N(\mathbf{Y}_0^N)$, and $\mathcal{A}_u(\mathbf{X}_0^N, \mathbf{Y}_0^N, \zeta)$ can be computed following Algorithm 5 and (2.39). In the second case, we always have $\mathcal{A}_u(\mathbf{X}_0^N, \mathbf{Y}_0^N, \zeta) = \hat{F}_N(\mathbf{X}_0^N)$, and $\mathcal{A}_u(\mathbf{Y}_0^N, \mathbf{X}_0^N, \zeta)$ can be computed following Algorithm 5 and (2.39) with the role of $\mathbf{X}_0^N$ and $\mathbf{Y}_0^N$ swapped. The estimator (2.43) amounts to a *Rao–Blackwellized* version of $\hat{F}_N^u$ in (2.39) over the arbitrary specification of which of the two draws from $\bar{q}$ is used as $\mathbf{X}_0^N$ or as $\mathbf{Y}_0^N$. The following statement is a mild variation of the previous results and is stated without a proof.

Proposition 2.32. *Consider the Coupling-UIS estimator $\tilde{F}_N^u$ in* (2.43)*, with cost $\tilde{\mathcal{C}}$. Suppose that the conditions of Proposition 2.30 are satisfied. Then $\tilde{F}_N^u$ is an unbiased estimator of $\pi(f)$ for any $N \geq 1$, with finite expected cost and finite variance, and its inefficiency is equivalent to that of IS as $N \to \infty$.*

The result supports the intuition that $\tilde{F}_N^u$ should be preferred to $\hat{F}_N^u$ in practice, and from now on we refer to $\tilde{F}_N^u$ as the Coupling-UIS estimator. Note however that the cost of the symmetrized estimator (2.43) is at least as large as that of (2.39).

Algorithm 6 provides pseudo-code for the Coupling-UIS estimator, which can be imple-

mented under the same conditions as Algorithm 1 (SNIS) or Algorithm 3 (PIMH). Its only tuning parameter is N. The pseudo-code implements a Rao–Blackwellization argument (Casella & Robert, 1996) that replaces each $\hat{F}_N(\mathbf{X}_t^N)$ by

$$\hat{F}_N^{\mathrm{RB}}(\mathbf{X}_{t-1}^N, \mathbf{X}^{N,\star}) = \big(1 - \alpha(\mathbf{X}_{t-1}^N, \mathbf{X}^{N,\star})\big)\, \hat{F}_N(\mathbf{X}_{t-1}^N) + \alpha(\mathbf{X}_{t-1}^N, \mathbf{X}^{N,\star}) \hat{F}_N(\mathbf{X}^{N,\star}), \quad (2.44)$$

where $\alpha(\mathbf{X}_{t-1}^N, \mathbf{X}^{N,\star}) = \min(1, \hat{Z}_N(\mathbf{X}^{N,\star})/\hat{Z}_N(\mathbf{X}_{t-1}^N))$ is the acceptance probability of the PIMH step, as in (2.12).

Algorithm 6. Coupling-based unbiased importance sampling (Coupling-UIS)

1. Sample $\mathbf{X}^N = (X_1, \ldots, X_N) \sim \bar{q}$ and $\mathbf{Y}^N = (Y_1, \ldots, Y_N) \sim \bar{q}$ independently.
2. If $\hat{Z}_N(\mathbf{X}^N) < \hat{Z}_N(\mathbf{Y}^N)$, swap $\mathbf{X}^N$ and $\mathbf{Y}^N$, so that $\hat{Z}_N(\mathbf{X}^N) \geq \hat{Z}_N(\mathbf{Y}^N)$ henceforth.
3. Set $\tau \leftarrow 1$ and initialize
$$\tilde{F}_N^u \leftarrow \tfrac{1}{2}\hat{F}_N(\mathbf{X}^N) + \tfrac{1}{2}\hat{F}_N(\mathbf{Y}^N).$$
4. Update the estimate, using Rao–Blackwellization as in (2.44):
$$\tilde{F}_N^u \leftarrow \tilde{F}_N^u + \tfrac{1}{2}\big(1 - \alpha(\mathbf{X}^N, \mathbf{Y}^N)\big)\Big(\hat{F}_N(\mathbf{X}^N) - \hat{F}_N(\mathbf{Y}^N)\Big).$$
Draw $U \sim \mathrm{Uniform}(0,1)$. If $U < \hat{Z}_N(\mathbf{Y}^N)/\hat{Z}_N(\mathbf{X}^N)$, go to step 6.
5. Iterate until both conditions in (e) and (f) are satisfied:
   (a) $\tau \leftarrow \tau + 1$.
   (b) Sample $\mathbf{X}^{N,\star} = (X_1^\star, \ldots, X_N^\star) \sim \bar{q}$.
   (c) Update the estimate:
$$\begin{aligned}\tilde{F}_N^u \leftarrow \tilde{F}_N^u + \tfrac{1}{2}\Big(&(1 - \alpha(\mathbf{X}^N, \mathbf{X}^{N,\star}))\hat{F}_N(\mathbf{X}^N) \\ &- (1 - \alpha(\mathbf{Y}^N, \mathbf{X}^{N,\star}))\hat{F}_N(\mathbf{Y}^N) \\ &+ \big(\alpha(\mathbf{X}^N, \mathbf{X}^{N,\star}) - \alpha(\mathbf{Y}^N, \mathbf{X}^{N,\star})\big)\, \hat{F}_N(\mathbf{X}^{N,\star})\Big).\end{aligned}$$
   (d) Draw $U \sim \mathrm{Uniform}(0,1)$.
   (e) If $U < \hat{Z}_N(\mathbf{X}^{N,\star})/\hat{Z}_N(\mathbf{X}^N)$, set $\mathbf{X}^N \leftarrow \mathbf{X}^{N,\star}$.
   (f) If $U < \hat{Z}_N(\mathbf{X}^{N,\star})/\hat{Z}_N(\mathbf{Y}^N)$, set $\mathbf{Y}^N \leftarrow \mathbf{X}^{N,\star}$.
6. Set cost $\leftarrow N(\tau + 1)$.
7. Return $\tilde{F}_N^u$.

2.5.6 Unbiased estimation of the inverse of the normalizing constant

A particular case is the unbiased estimation of $1/Z$, where $Z = q(\omega_u)$ for an unnormalized weight function ω_u proportional to ω. This problem was part of the motivation for Wang & Wang (2023) and for Chopin et al. (2025), where such unbiased estimators are used for inference in doubly-intractable models. Coupling-UIS can be used for this purpose by setting $f = 1/\omega_u$. In this case $\hat{F}_N(\mathbf{X}^N) = 1/\hat{Z}_N(\mathbf{X}^N)$, the natural but biased importance sampling estimator of $1/Z$; here we redefine $\hat{Z}_N(\mathbf{X}^N)$ as $N^{-1}\sum_{n=1}^N \omega_u(X_n)$, so that its mean is Z rather than one. We are interested in the comparison of the efficiencies of Coupling-UIS and SNIS. For Coupling-UIS, we could directly apply the previous results with the choice of test function $f = 1/\omega_u$, but we obtain weaker conditions by leveraging the specific structure of the problem, in particular the identity $\omega_u \cdot f = 1$. Thus we rederive Propositions 2.26, 2.28 and 2.30, for the case $f : x \mapsto 1/\omega_u(x)$, and state the results in Proposition 2.33, with the proof in Appendix A.5.5.

Proposition 2.33 (Properties of Coupling-UIS for $1/Z$)**.** *Denote by $\hat{I}_N$ the Coupling-UIS estimator* (2.39) *with $f = 1/\omega_u$, for which $\hat{F}_N(\mathbf{X}^N) = 1/\hat{Z}_N(\mathbf{X}^N)$. Assume that $q(\omega^p) < \infty$ for some $p \geq 2$ and $q(\omega^{-\eta}) < \infty$ for some $\eta > 0$. Then for any $1 \leq s < p$ and $N \geq sp/((p-s)\eta)$, $\hat{I}_N$ has s finite moments:*

$$\mathbb{E}\left[\left|\hat{I}_N - \frac{1}{Z}\right|^s\right] < \infty. \tag{2.45}$$

If furthermore $p > 2s$, then for $N \geq sp/((p-s)\eta)$ there exists $C > 0$ such that

$$\mathbb{E}\left[\left|\hat{I}_N - \frac{1}{Z}\right|^s\right] \leq C\, N^{-s/2}. \tag{2.46}$$

In particular, with $s = 2$: $\hat{I}_N$ has finite variance for $p > 2$ and $N > 2p/((p-2)\eta)$; if $p > 4$, then

$$\lim_{N\to\infty} N \cdot \mathbb{E}\left[\left(\hat{I}_N - \frac{1}{Z}\right)^2\right] = \lim_{N\to\infty} N \cdot \mathbb{E}_{\mathbf{X}^N\sim\bar{q}}\left[\left(\frac{1}{\hat{Z}_N(\mathbf{X}^N)} - \frac{1}{Z}\right)^2\right].$$

The distinctive feature of Coupling-UIS is the last item of the proposition, i.e. the asymptotic equivalence of the MSE of $\hat{I}_N$ and that of the SNIS estimator $1/\hat{Z}_N(\mathbf{X}^N)$, which is the natural estimator of $1/Z$. We will see in experiments that the condition $p > 4$ is not necessary.

Remark 2.34. The general Coupling-UIS moment bound (Proposition 2.28) specialized to $f = 1/\omega_u$ requires $q(\omega^{-r}) < \infty$ for some $r \geq 2$ together with the compatibility condition $pr/(p+r+2) > ps/(p-s)$. Proposition 2.33 replaces this by $q(\omega^{-\eta}) < \infty$ for any $\eta > 0$

and $p > 2s$, with no compatibility condition between η and p. When $s = 2$, the general compatibility condition reduces to $2p + 4r + 4 < rp$, i.e. $r > (2p+4)/(p-4)$, which forces $p > 4$. Both results thus require $p > 4$, but they differ on the assumption required on ω^{-1}. The general bound demands $q(\omega^{-r}) < \infty$ with $r > (2p+4)/(p-4)$, which is very restrictive for moderate p: for instance, $p = 5$ requires $r > 14$, $p = 6$ requires $r > 8$. Proposition 2.33, on the other hand, requires only $p > 4$ and $q(\omega^{-\eta}) < \infty$ for any $\eta > 0$.

Remark 2.35. The strict inequality $s < p$ can be relaxed to $s = p$ for the SNIS and PIMH pointwise estimators of $1/\hat{Z}_N$ (Propositions A.15 and A.16), at the cost of a slower convergence rate. A truncation argument splitting at the event $\{\hat{Z}_N \geq 1/2\}$ shows that, under the same assumptions with $p \geq 2$, for every $\alpha \in (0, p/2)$ there exists $C_\alpha > 0$ such that

$$\mathbb{E}_{\bar{q}}\left[\left|\frac{1}{\hat{Z}_N(\mathbf{X}^N)} - \frac{1}{Z}\right|^p\right] \leq C_\alpha \, N^{-\alpha},$$

for N large enough, and the same bound extends to the PIMH chain uniformly in t.

This endpoint relaxation does *not* extend to the Coupling-UIS estimator $\hat{I}_N$: the condition $s < p$ in (2.45) is sharp. The Minkowski–Hölder framework underlying the proof requires the series $\sum_t \mathbb{P}(\tau > t)^{(\kappa-1)/(s\kappa)}$ to converge, and at $s = p$ this fails for every finite κ due to the polynomial tail of the meeting time. See Appendix A.5.5 for details.

2.6 Applications

2.6.1 Unbiased estimation of $1/Z$

We consider the unbiased estimation of $1/Z$ described in Subsection 2.5.6. As described in Chopin et al. (2025, Section 4.3) the problem arises for example in likelihood-based inference for exponential random graphs. We compare Coupling-UIS with two existing unbiased estimators for $1/Z$, MLMC-UIS (Blanchet & Glynn, 2015; Shi & Cornish, 2021) and Taylor-UIS (Blanchet et al., 2015; Chopin et al., 2025), described in Appendix D.1.

We consider Example 2.3, with $p = 3$, so that the importance weight has strictly less than 3 finite moments. Here the actual value of Z^{-1} is 1. For SNIS and Coupling-UIS, the unique tuning parameter is the number of samples N, which we vary from 4 to 128. For MLMC-UIS, we select a Geometric parameter based on the number of moments of ω under q as described in Appendix D.1. We vary the initial level of the multi-level construction from 1 to 5, and we report both the "single sample" (SS) and the "Russian roulette" (RR) estimates (Shi & Cornish, 2021). For Taylor-UIS, the input estimators of Z are based on N samples, with N ranging from 4 to 128. For each value of N, we tune Taylor-UIS following the guidelines of Chopin et al. (2025) and we do not count the computational cost of the tuning step, which is negligible. The

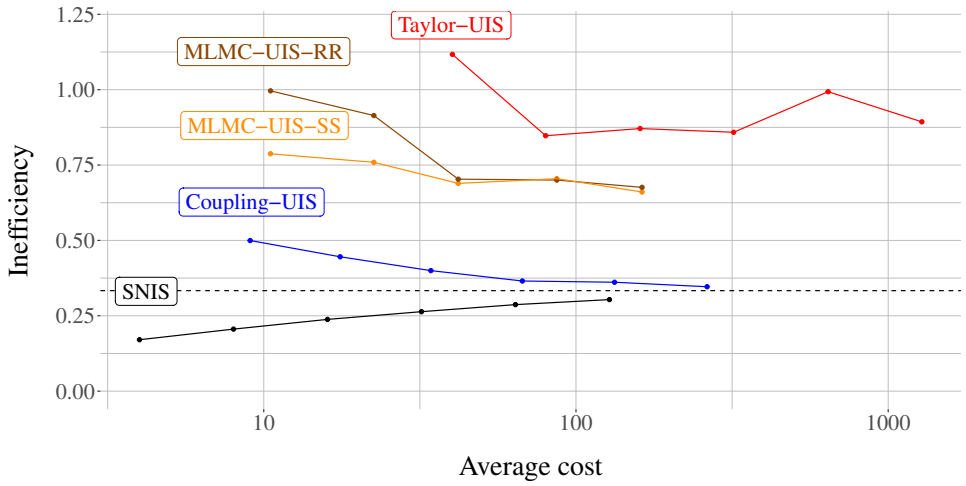


Figure 2.1: Inefficiency of different estimators of $1/Z$, against average cost per estimator. For each method, we vary a tuning parameter that impacts both the inefficiency and the cost. For SNIS and Coupling-UIS the tuning parameter is the number of samples N; see Section 2.6.1. The dashed line represents the asymptotic inefficiency of SNIS.

results are shown in Figure 2.1. The inefficiency is defined as the product of the variance of the estimator and its expected cost, where the cost is measured in number of evaluations of ω per estimator. The results are based on $250,000$ independent repeats of each method.

Firstly, we note that inefficiencies of Coupling-UIS and SNIS converge to the same limit as N increases, as predicted by Propositions 2.32 and 2.33, even though the setting violates the assumption $p > 4$ in these propositions. Secondly, Coupling-UIS provides a smaller inefficiency than the other unbiased estimators of $1/Z$, for all tuning parameters. This, combined with the fact that Coupling-UIS only requires a choice of N as tuning parameter, suggests that Coupling-UIS is a competitive method for unbiased estimation of $1/Z$.

2.6.2 Nested expectations

We consider a nested expectation problem, as defined in Subsection 2.1.1. The setting is that of Bayesian inference after multiple imputation of missing data (e.g. Zhou & Reiter, 2010). Specifically, we consider a logistic regression where the covariates X are continuous and contain missing entries. We generate the data by sampling $X \in \mathbb{R}^{n\times d}$ independently from a Normal$(0,1)$ distribution, with $n = 100$ and $d = 10$. The response Y is generated given X from a logistic regression with parameters $\beta^\star$ defined as $\beta_i^\star = i/10$ for $i = 1,\dots,5$, and $\beta_i^\star = 0$ for $i = 6,\dots,10$. We obtain $\sum_{i=1}^n Y_i = 52$. We then remove some entries from X, with X_{ij} becoming missing with probability 0.4 if $Y_i = 1$ and 0.2 if $Y_i = 0$, independently across i and j. As a result, 285 entries are missing out of 1000. We use the correct Normal$(0,1)$ distribution as the imputation model for each missing entry.

For each completed data set, we approximate the posterior distribution in the logistic regression model, using the Normal$(0, 10\cdot I_d)$ distribution as prior on the coefficients. We use

importance sampling using as proposal a multivariate Student distribution with 10 degrees of freedom, centered at the MLE, and with covariance matrix equal to the inverse of minus the Hessian of the log-posterior density at the MLE. We consider three variants of importance sampling: SNIS, with bias of order N^{-1}, the variant of Skare et al. (2003), referred to as SNIS-LOO, where each importance weight is divided by the sum of the other $N-1$ importance weights, with bias of order N^{-2} as shown in Appendix D.4, and finally the Coupling-UIS estimator from Section 2.5.5, with no bias. We consider the test function that maps the parameter to its first component, so that we are estimating the posterior mean of the first coefficient, averaged over the distribution of the imputed data. For a total budget C, we need to allocate the budget between the number of imputed data sets M and the number of samples N per estimator for the inner expectation, so that $C = MN$. We follow the optimal allocation from Andradóttir & Glynn (2016): for SNIS we use $M = C^{2/3}$ and $N = C^{1/3}$. For SNIS-LOO, we use $M = C^{4/5}$ and $N = C^{1/5}$. For Coupling-UIS, we use a fixed $N = 5$, and $M = C/5$. In the latter case the cost is random so we report its average.

Figure 2.2 displays the MSE of the three methods as a function of the computing budget C. The results are based on 200 independent repeats. The MSE is computed using an approximate ground truth set to the average of the 200 estimators obtained with Coupling-UIS at the highest budget of $M = 32000$ and $N = 5$. From those 200 estimators we compute a 95% confidence interval for the quantity of interest and obtain $[0.2286, 0.2289]$. The figure shows regression lines corresponding to log-mean squared error over log-average cost. We obtain a slope of -0.65 for SNIS, -0.83 for SNIS-LOO and -1.06 for Coupling-UIS, matching closely the theoretical rates of $-2/3$, $-4/5$ and -1, respectively. The figure also shows that, for a total budget between 10^4 and 3×10^5, nested expectation estimators using Coupling-UIS outperform those using SNIS throughout the range, and start outperforming those using SNIS-LOO for the larger budgets.

2.6.3 Robust mean estimation

A motivation for eliminating the bias is that one can then use any statistical technique that requires unbiased inputs. An example is robust mean estimation (Lugosi & Mendelson, 2019). The goal is to estimate the mean μ of a random variable with finite variance σ^2 from n i.i.d. copies $X_1, \ldots, X_n$: for a confidence level $\delta \in (0,1)$, we seek $\hat{\mu}_n = \hat{\mu}_n(X_1, \ldots, X_n, \delta)$ such that the inequality

$$\mathbb{P}(|\hat{\mu}_n - \mu| > \epsilon) \leq \delta \tag{2.47}$$

holds for the smallest possible $\epsilon = \epsilon(n, \delta)$. The empirical average satisfies (2.47) asymptotically with $\epsilon = \sigma\sqrt{2\log(2/\delta)/n}$ via the CLT. On the other hand, Chebyshev's inequality only yields $\epsilon = \sigma\sqrt{1/(n\delta)}$, with a δ-dependence that cannot be improved under finite variance alone (Lugosi & Mendelson, 2019, Section 2). Remarkably, implementable alternatives recover the

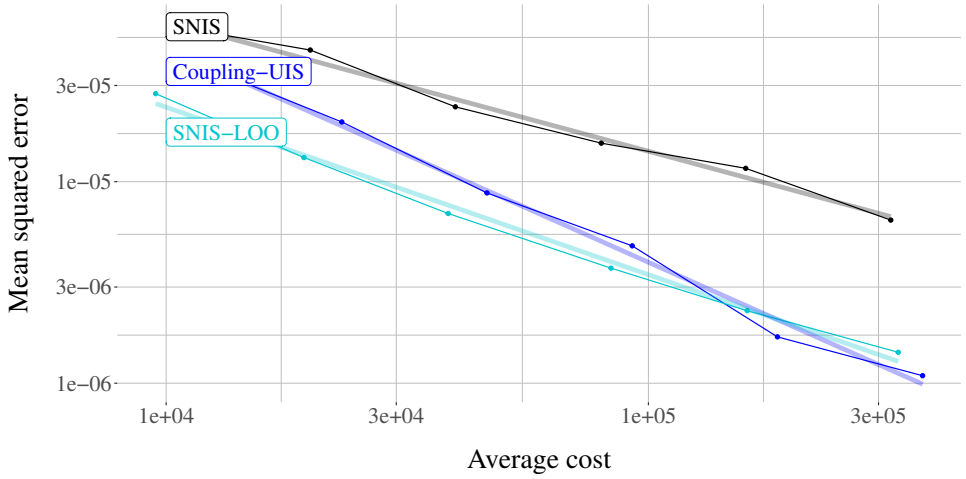


Figure 2.2: Mean-squared error of nested expectation estimators, in the context of Bayesian inference after multiple imputation, using either SNIS, SNIS-LOO or Coupling-UIS for the inner expectation, as a function of the computing budget C. The dashed lines are regression lines fitted on log-MSE $\sim \log C$.

sub-Gaussian rate, that is, the $\sqrt{\log(1/\delta)/n}$ dependence of the central limit theorem, but at every finite n. The Median-of-Means (MoM, Nemirovsky & Yudin (1983)) estimator provides a prime example, where the $X_1, \ldots, X_n$ are partitioned into K blocks of size m with $n = mK$, and the estimator is obtained as the median of the blockwise means. The following result shows that the MoM estimator achieves sub-Gaussian performance. We recall this guarantee in Appendix D.5 (Theorem D.7): with $K = \lceil 8 \log(1/\delta) \rceil$ blocks, the MoM estimator $\hat{\mu}_{\mathrm{MoM},n}$ satisfies $|\hat{\mu}_{\mathrm{MoM},n} - \mu| \leq \sigma \sqrt{32 \log(1/\delta)/n}$ with probability at least $1 - \delta$.

The combination of MoM with self-normalized importance sampling (SNIS) was explored by Dau (2022). The proposed estimator, termed MoM-SNIS, is the median of K SNIS estimators obtained with M draws each, so that the cost is equivalent to SNIS with $N = M \times K$ draws. The above result on MoM does not directly apply because SNIS is biased for the quantity of interest. Dau (2022, Proposition 2) shows that MoM-SNIS achieves sub-Gaussian performance under minimal assumptions on the weight and test function, but only once N exceeds a problem-dependent threshold involving the typically unknown variance of the weights. Furthermore, as shown in Dau (2022, Proposition 3), it is not possible to obtain a sub-Gaussian concentration inequality for MoM-SNIS that would hold for all N larger than a threshold that would depend on δ only.

In contrast, since Coupling-UIS estimators are unbiased, one can directly use MoM and obtain sub-Gaussian performance for all n larger than $\lceil 8 \log(1/\delta) \rceil$ as in Theorem D.7, where n is the number of independent copies of Coupling-UIS. We could also use any other robust mean estimator, such as those proposed in Minsker & Ndaoud (2021) or Lee & Valiant (2022). Drawbacks relative to MoM-SNIS (Dau, 2022) include the higher variance of Coupling-UIS relative to SNIS for matching costs, and the random nature of the computing cost.

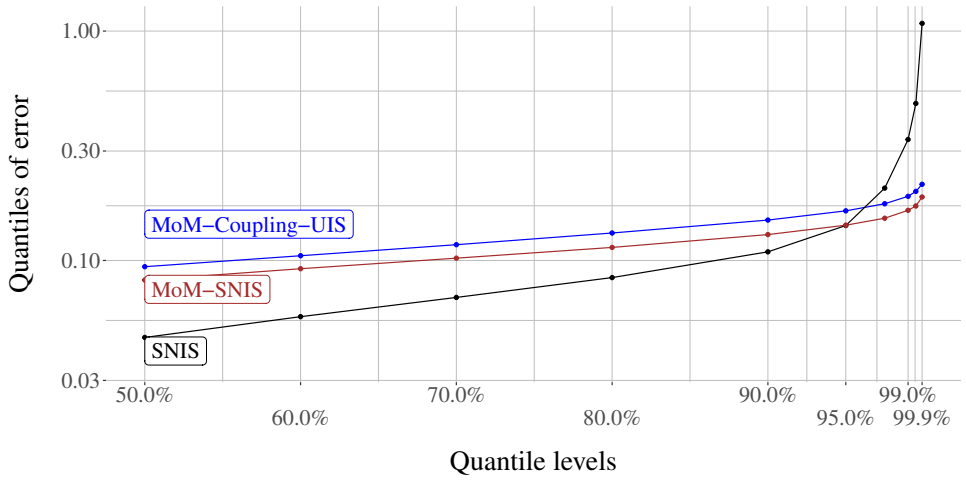


Figure 2.3: Quantiles of the absolute error for SNIS, MoM-SNIS, MoM-Coupling-UIS for the Exponential example with $p = 2.1$.

We revisit Example 2.3, with $p = 2.1$, and the test function $f(x) = x$, so that $\pi(f) = 1$. We compare the distribution of the errors of self-normalized importance sampling (SNIS) with $N = 4000$, with the MoM-SNIS method of Dau (2022), tuned with $\delta = 0.05$, leading to $K = \lceil 8 \log(1/\delta) \rceil = 24$ and $M = 192$ so that $K \times M = 4608$ is comparable to N. For MoM-Coupling-UIS, we use 24 samples for each unbiased estimator, and then we compute the median of $K = 24$ averages, each over $M = 3$ unbiased estimators per block. This leads to an average cost of 3856 target density evaluations per MoM-Coupling-UIS estimator. We generate 10^6 independent runs of SNIS with 4000 samples, and then for MoM, of SNIS with 192 samples and Coupling-UIS with 24 samples. This allows us to collect 41,660 MoM-SNIS and 13,880 MoM-Coupling-UIS estimates.

Figure 2.3 displays the quantiles of the absolute value of the difference between the estimates and $\pi(f)$, for SNIS, MoM-SNIS, and MoM-Coupling-UIS. The median absolute error is smaller for SNIS than for the robust estimates. However, for quantile of very high order, the error of SNIS increases faster than the error of the robust methods. From the experiments we observe a small advantage of MoM-SNIS over MoM-Coupling-UIS at comparable costs, so the appeal of the latter appears mostly theoretical.

Chapter 3

Convergence of Statistical Estimators via Mutual Information Bounds

This chapter reproduces, with minor typographic modifications, the paper:

E. M. Khribch, P. Alquier

Convergence of Statistical Estimators via Mutual Information Bounds

See Khribch & Alquier (2024).

This chapter introduces a mutual information (MI) bound for statistical models. The bound yields improved contraction rates for fractional posteriors in Bayesian nonparametrics, and applies to other estimation methods, including variational inference and the maximum likelihood estimator (MLE). It also applies under model misspecification, where it takes the form of oracle inequalities relevant, for instance, in model selection. The results connect the analysis of statistical estimators with information-theoretic tools and PAC-Bayesian theory.

Keywords: PAC-Bayes bounds, information bounds, variational approximations.

3.1 Introduction and Motivation

3.1.1 Mutual Information Bounds in Machine Learning

Mutual information (MI) bounds are a tool for analyzing generalization and learning in machine learning. This approach provides an information-theoretic perspective on the generalization of learning algorithms.

The use of information-theoretic approaches in machine learning goes back to Rissanen's Minimum Description Length (MDL) principle (Rissanen, 1978); see Grünwald (2007) for a more recent view on the topic. A related approach is PAC-Bayes bounds (McAllester, 1998, 1999). These bounds were initially developed as Bayesian-inspired generalization bounds in machine learning, but the connection to information theory was later highlighted by Zhang (2006b) and Catoni (2007). The Kullback-Leibler divergence terms in the PAC-Bayes bound can be interpreted in terms of length of the description of a model, and more generally as a complexity measure. In particular, Catoni (2007) proved that an optimal prior choice in the PAC-Bayes bound makes the Kullback-Leibler divergence term equal to the mutual information between the sample and the parameter. This fact, rediscovered independently by Russo & Zou (2016) under the name "mutual information bound", has received considerable attention in the past years (Xu & Raginsky, 2017; Negrea et al., 2019; Lugosi & Neu, 2022).

In parallel, it became clear that PAC-Bayes bounds can be used to analyze various kinds of machine learning algorithms, including empirical risk minimization (Catoni & Giulini, 2017) as well as various kinds of robust estimators (Catoni, 2012), generalized posteriors and their variational approximations (Alquier et al., 2016). We refer the reader to Alquier (2024) for a recent overview on PAC-Bayes, and to Hellström et al. (2025) for a global presentation of all information bounds.

PAC-Bayes bounds have also been used to analyze estimators in a more statistical context where the objective is inference on parameters: Bhattacharya et al. (2019) analyzed Bayesian fractional posteriors with related tools, and this was extended to variational approximations by Alquier & Ridgway (2020) and to (non-tempered) posteriors (Yang et al., 2020; Zhang & Gao, 2020); see also Cherief-Abdellatif (2019); Ohn & Lin (2024) for results on model selection. Since mutual information bounds in machine learning are optimized PAC-Bayes bounds, a natural question arises: Can we similarly optimize these statistical bounds to achieve tighter results? In this chapter, we prove a mutual information version of the PAC-Bayes bounds of Bhattacharya et al. (2019) and Alquier & Ridgway (2020). We show that it can be used to prove convergence rates for tempered posteriors, variational approximations, and the maximum likelihood estimator (MLE). When applied to Bayesian-style estimators, this result leads to improved rates compared to Bhattacharya et al. (2019); Alquier & Ridgway (2020); Yang et al.

(2020); Zhang & Gao (2020). Our bound thus extends the existing literature on MI bounds, providing tools to study a range of statistical methods.

3.1.2 Bayesian Nonparametrics and Posterior Contraction

The study of frequentist properties of Bayesian procedures, particularly in nonparametric settings, has received much attention in recent years. Two main approaches have emerged: the analysis of posterior concentration rates in the Bayesian nonparametric literature, and the PAC-Bayesian framework from statistical learning theory. These approaches have largely developed independently, but recent work suggests that connecting them can lead to new results and improved theoretical guarantees.

The study of posterior concentration rates has a long history in Bayesian nonparametrics. Early work by Ghosal et al. (2000) and Shen & Wasserman (2001) established general conditions for posterior consistency and convergence rates. These results were later extended to more general settings by Ghosal & van der Vaart (2007) and van der Vaart & van Zanten (2008). An overview of these developments can be found in Rousseau (2016); Ghosal & van der Vaart (2017) and Castillo (2024).

In parallel, the PAC-Bayesian approach, introduced by McAllester (1999), has become a flexible tool for deriving generalization bounds in statistical learning. It has recently gained traction within the Bayesian community, offering a new perspective on the analysis of Bayesian methods.

Some progress has been made in connecting PAC-Bayesian theory and Bayesian nonparametrics. Notably, Grünwald & Mehta (2018) provided a unified analysis of excess risk bounds for general loss functions, bridging the gap between empirical risk minimization and generalized Bayesian approaches. Additionally, Bhattacharya et al. (2019); Alquier & Ridgway (2020); Banerjee & Montufar (2021) used PAC-Bayesian techniques to prove the contraction of tempered posteriors under essentially only a prior mass condition.

3.1.3 Our Contributions

We derive a new MI bound tailored for statistical estimation. This bound provides guarantees in terms of statistical inference, rather than on the generalization risk as in Russo & Zou (2016); Xu & Raginsky (2017). From a technical perspective, we obtain this bound by applying the localization technique of Catoni (2007) to PAC-Bayes bounds for density estimation. Formally, we use bounds proven by Bhattacharya et al. (2019); Alquier & Ridgway (2020), but similar bounds can be traced back to Zhang (2006a). Our bounds also apply when the model is misspecified, where they take the form of oracle inequalities, following the misspecified analysis of Alquier & Ridgway (2020, Section 2.4); the well-specified contraction results are recovered

when the data-generating distribution belongs to the model. Under the divergence-equivalence condition of Assumption 3.3, these oracle inequalities involve the Kullback-Leibler divergence on both sides, both in expectation and with high probability.

This chapter is organized as follows. In Section 3.2, we introduce our notation and state our MI bound in its most general form: Theorem 3.2. In Section 3.3, we specify this bound in various settings. Under specific assumptions on the statistical model, we derive convergence rates for tempered posteriors and for variational inference. Our rates are tighter than those in Bhattacharya et al. (2019); Alquier & Ridgway (2020); Yang et al. (2020); Zhang & Gao (2020), as we eliminate suboptimal logarithmic terms. Finally, in Section 3.4, we work through concrete examples: we verify our assumptions in classical models and illustrate an application of Theorem 3.2 to prove the convergence of the MLE.

All proofs are provided in Appendix B, where we also clarify the connection between our MI bound and PAC-Bayes bounds.

3.2 Notations and Preliminaries

Consider a collection of n i.i.d. random variables $(X_1, \dots, X_n) =: \mathcal{S}$ in a measured sample space $(\mathbb{X}, \mathcal{X}, P^*)$, where P^* denotes the (unknown) data-generating distribution. Let $\{P_\theta, \theta \in \Theta\}$ be a statistical model, that is, a collection of probability distributions on $(\mathbb{X}, \mathcal{X})$ indexed by a parameter set Θ. Let $\mathcal{M}_1^+$ be the set of probability measures on Θ equipped with a suitable σ-algebra $\mathcal{B}$.

In this chapter, we do *not* assume that the model is well-specified: the data-generating distribution P^* need not belong to $\{P_\theta : \theta \in \Theta\}$. In other words, we allow the model to be *misspecified*. This is the relevant setting, for instance, in model selection, where all the candidate models but (at most) one are misspecified. The well-specified case, in which there is a $\theta_0 \in \Theta$ with $P_{\theta_0} = P^*$, is recovered as a special case; whenever we restrict to it we write θ_0 for the true parameter.

We assume that the model is dominated: all the P_θ's have a density p_θ with respect to a specified measure μ (for example, the Lebesgue measure or the counting measure). We also assume that the data-generating distribution is dominated, $P^* \ll \mu$, with density $p^* = \mathrm{d}P^*/\mathrm{d}\mu$. Let $\mathcal{L}_n(\theta)$ denote the likelihood function, and $r_n(\theta)$ the negative log-likelihood ratio with respect to the truth. That is:

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i)$$
$$r_n(\theta) = \sum_{i=1}^{n} \log \frac{p^*(X_i)}{p_\theta(X_i)},$$

with the convention $\log(p^*(X_i)/0) = +\infty$ if $p^*(X_i) > 0$ and $\log 0/0 = 0$. The (generalized) posteriors defined below do not depend on p^*, since $e^{-\alpha r_n(\theta)} \propto \mathcal{L}_n^\alpha(\theta)$; in the well-specified case $P^* = P_{\theta_0}$ we recover the usual $r_n(\theta) = r_n(\theta, \theta_0) = \sum_{i=1}^n \log\big(p_{\theta_0}(X_i)/p_\theta(X_i)\big)$.

We recall the definitions of the α-Rényi divergence $\mathcal{D}_\alpha$, for $\alpha \in (0,1)$, and of the Kullback-Leibler (KL) divergence, KL. Conventions differ across the literature, with or without the factor $1/\alpha$ and with or without the logarithm; we use the logarithmic form below throughout, and refer to van Erven & Harremoës (2014) for a systematic comparison.

Given two probability distributions P and Q defined on the same measurable space, and μ a reference measure such that both P and Q are absolutely continuous with respect to μ,

$$\mathcal{D}_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \left(\int \left(\frac{\mathrm{d}P}{\mathrm{d}\mu}(x) \right)^\alpha \left(\frac{\mathrm{d}Q}{\mathrm{d}\mu}(x) \right)^{1-\alpha} \mu(\mathrm{d}x) \right),$$
$$\mathrm{KL}(P\|Q) = \int \log \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x) \right) P(\mathrm{d}x)$$

with the convention that $\mathrm{KL}(P\|Q) = +\infty$ if P is not absolutely continuous with respect to Q. The divergence $\mathcal{D}_\alpha(P\|Q)$ does not depend on the choice of the reference measure μ.

Definition 3.1. Let U and V be two random variables, $P_{U,V}$ the joint distribution of U and V, and P_U the marginal distribution of U and P_V the marginal distribution of V. The mutual information (MI) between U and V is defined as:

$$\mathcal{I}(U;V) := \mathrm{KL}(P_{U,V}\|P_U \otimes P_V).$$

We now present our main result, which establishes a mutual information bound for any randomized procedure, that is, any probability distribution $\widehat{\rho}$ learned from the data, where $\widehat{\rho} = \widehat{\rho}(\mathcal{S})$. Examples of such procedures include the posterior, tempered posteriors, and variational approximations.

By analogy with the Bayesian framework, we will sometimes call $\widehat{\rho}$ a (generalized) posterior.

To clarify the notation in this setting, consider that $V = \mathcal{S} = (X_1, \ldots, X_n) \sim P^{*\otimes n}$ and $U = \theta$ is sampled conditionally on the data: $U|V \sim \widehat{\rho}$. This describes a joint distribution on the data and the parameter: $P_{U,V}$, and thus a mutual information $\mathcal{I}(U;V) = \mathcal{I}(\theta;\mathcal{S})$.

Theorem 3.2. *Given a generalized posterior $\widehat{\rho}$, we have, for any $\alpha \in (0,1)$:*

$$\mathbb{E}_{\mathcal{S}} \left[\mathbb{E}_{\theta\sim\widehat{\rho}}[\mathcal{D}_\alpha(P_\theta\|P^*)] - \frac{\alpha}{n(1-\alpha)} \mathbb{E}_{\theta\sim\widehat{\rho}}[r_n(\theta)] \right] \leq \frac{\mathcal{I}(\theta;\mathcal{S})}{n(1-\alpha)}. \tag{3.1}$$

The proof of Theorem 3.2 is provided in Appendix B.

Intuitively, it is natural to sample from generalized posteriors $\widehat{\rho}$ concentrated on θ with a large likelihood, that is, for which $r_n(\theta)$ is small. We expect this to make P_θ close to P^* (more precisely, to the best in-model approximation of P^*). Theorem 3.2 states that this is provably the case, in the sense that $\mathcal{D}_\alpha(P_\theta \| P^*)$ will be small, provided that $\mathcal{I}(\theta; \mathcal{S})/n$ also remains small.

We will see in Section 3.3 that generalized Bayesian methods such as fractional posteriors and their variational approximations can be tuned to ensure that $\mathcal{I}(\theta; \mathcal{S})/n$ is small. In this case, we will derive tight rates of convergence from Theorem 3.2. Theorem 3.2 is quite general: in this form, there are no regularity assumptions on the densities p_θ (even though such assumptions can be useful to make the bound tight). Moreover, there is no restriction on $\widehat{\rho}$. While the bound will give tight results when applied to Bayesian methods, it can also be used to prove convergence of the Maximum Likelihood Estimator (MLE). This is not entirely straightforward: one might be tempted to take $\widehat{\rho}$ as the Dirac mass on the MLE, but this will often cause $\mathcal{I}(\theta; \mathcal{S})$ to be infinite. However, by taking $\widehat{\rho}$ with support on a small neighborhood of the MLE instead (a technique developed by Catoni (2004) for PAC-Bayes bounds), we can analyze the MLE.

For completeness, we recall the definition of the MLE:

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta} \mathcal{L}_n(\theta) = \operatorname*{argmin}_{\theta \in \Theta} r_n(\theta)$$

when this argmin exists and is unique. Let π be a prior distribution on $(\Theta, \mathcal{B})$. The fractional posterior for $\alpha \in (0, 1]$ is defined as:

$$\pi_{n,\alpha}(\theta) := \frac{e^{-\alpha r_n(\theta)} \pi(\theta)}{\mathbb{E}_\pi\left[e^{-\alpha r_n(\theta)}\right]} \propto \mathcal{L}_n^\alpha(\theta)\pi(\theta).$$

Here π and $\pi_{n,\alpha}$ denote densities with respect to μ. It reduces to the standard posterior when $\alpha = 1$. For $\mathcal{F} \subset \mathcal{M}_1^+$, we define the variational approximation:

$$\widetilde{\rho}_{n,\alpha} = \operatorname*{argmin}_{\rho \in \mathcal{F}} \left[\frac{\alpha}{n(1-\alpha)} \mathbb{E}_{\theta \sim \rho}\left[r_n(\theta)\right] + \frac{\mathrm{KL}(\rho \| \pi)}{n(1-\alpha)} \right].$$

When $\mathcal{F} = \mathcal{M}_1^+$, we have $\widetilde{\rho}_{n,\alpha} = \pi_{n,\alpha}$. In practice, variational approximations are useful when $\mathcal{F}$ leads to feasible algorithms; see Bishop (2006) for an introduction.

Many other methods could be analyzed with this theorem. For example, the original MI bound was applied by Neu et al. (2021) to analyze the output of finite steps of stochastic gradient descent on the empirical risk. A similar approach could be used with our bound to analyze stochastic optimization algorithms on the log-likelihood.

In the next section, we formulate assumptions on the model (and the prior π) that will lead to explicit rates of convergence for the fractional posteriors and their variational approxima-

tions. In this setting, Theorem 3.2 leads to better rates than the bounds in Bhattacharya et al. (2019); Alquier et al. (2016), as it eliminates the extra $\log(n)$ factor. The MLE will be discussed later, in Section 3.4.

3.3 Main Results

Our results rely on two sets of assumptions. Assumptions 3.4, 3.5 and 3.6 are assumptions on the model complexity. Similar assumptions appear in Catoni (2007); Bhattacharya et al. (2019); Alquier et al. (2016). While these assumptions alone could yield the same rates as in these works, an additional assumption on the model, Assumption 3.3, allows us to obtain improved rates from Theorem 3.2.

Throughout this section, θ_0 no longer plays a special role. The natural benchmark is the best Kullback-Leibler approximation of P^* within the model, that is,

$$\inf_{\tilde{\theta}\in\Theta} \mathrm{KL}(P^* \| P_{\tilde{\theta}}).$$

This quantity is zero if and only if the model is well-specified, and is otherwise the irreducible approximation error. When the infimum is attained, we write $\theta^\star \in \operatorname{argmin}_{\tilde{\theta}\in\Theta} \mathrm{KL}(P^* \| P_{\tilde{\theta}})$ for a Kullback-Leibler projection of P^* onto the model. Following the misspecified analysis of Alquier & Ridgway (2020, Section 2.4), the complexity assumptions below are stated relative to this benchmark.

Assumption 3.3. For any $\alpha \in (0,1)$, there is a $c(\alpha) > 0$ such that

$$\forall \theta \in \Theta, \quad \mathrm{KL}(P^* \| P_\theta) \le c(\alpha) \mathcal{D}_\alpha(P_\theta \| P^*). \tag{3.2}$$

To understand this property, we recall the following facts (van Erven & Harremoës, 2014): for any $\alpha < 1$, we have

$$\mathcal{D}_\alpha(q\|p) = \frac{\alpha}{1-\alpha}\mathcal{D}_{1-\alpha}(p\|q) \le \frac{\alpha}{1-\alpha}\mathrm{KL}(p\|q).$$

The monotonicity of Rényi divergence in α follows from Jensen's inequality, which directly gives the upper bound on the KL divergence. Applying this identity with $p = P^*$ and $q = P_\theta$, Assumption 3.3 is equivalent to $\mathrm{KL}(P^*\|P_\theta) \le c(\alpha)\,\alpha/(1-\alpha)\,\mathcal{D}_{1-\alpha}(P^*\|P_\theta)$, where both divergences compare P^* and P_θ in the same order. Chaining the two displays gives $\mathrm{KL}(P^*\|P_\theta) \le c(\alpha)\,\alpha/(1-\alpha)\ \mathrm{KL}(P^*\|P_\theta)$, so that

$$c(\alpha) \ge \frac{1-\alpha}{\alpha} \tag{3.3}$$

as soon as $\mathrm{KL}(P^*\|P_\theta) > 0$ for some $\theta \in \Theta$: the constant of Assumption 3.3 cannot be smaller than $(1-\alpha)/\alpha$. In particular $n(1-\alpha)/c(\alpha) \le \alpha n$, so the range of β in Theorem 3.7 below is contained in $(0, \alpha n)$. Moreover,

$$\mathcal{D}_\alpha(p\|q) \xrightarrow[\alpha\to 1^-]{} \mathrm{KL}(p\|q).$$

Thus, Assumption 3.3 states that the Rényi divergence and the KL divergence are equivalent on our statistical model. Similar assumptions have appeared in various forms in the literature. Notably, Grünwald & Mehta (2018, Lemma 13) relates the KL divergence to the Hellinger distance. A similar result was previously established by Wong & Shen (1995, Theorem 5), allowing one to bound the KL divergence in terms of the Hellinger distance. In particular,

$$\mathrm{KL}(P^*\|P_\theta) \le (2 + \log V)\, H^2(P^*, P_\theta), \tag{3.4}$$

where V is a bound on the density ratio p^*/p_θ, which makes the equivalence postulated in Assumption 3.3 explicit. The Hellinger distance $H(P,Q)$ is also related to the Rényi divergence, specifically for $\alpha = 1/2$:

$$H^2(P,Q) = 2\left(1 - \exp\left(-\frac{1}{2}\mathcal{D}_{1/2}(P\|Q)\right)\right) \le \mathcal{D}_{1/2}(P\|Q), \tag{3.5}$$

the inequality holding because $2(1-e^{-x/2}) \le x$ for $x \ge 0$. Models where $H^2(P,Q)$ is locally equivalent to KL have been extensively studied in asymptotic statistics theory. We examine several examples of models satisfying Assumption 3.3 in Section 3.4.

In the well-specified case, Assumption 3.3 is a property of the model alone. Under misspecification it becomes a joint condition on the model and the truth P^*, requiring the Kullback-Leibler and Rényi divergences to the truth to be equivalent. It remains satisfiable in genuinely misspecified models. For instance, if $P^* = \mathcal{N}(\mu_0, s^2)$ and $P_\theta = \mathcal{N}(\theta, v^2)$ with $s^2 \ne v^2$, then both $\mathrm{KL}(P^*\|P_\theta)$ and $\mathcal{D}_\alpha(P_\theta\|P^*)$ are quadratic in θ with strictly positive leading coefficients, so the inequality holds for a suitable $c(\alpha)$. This equivalence to the truth is what lets the Kullback-Leibler divergence, rather than the Rényi divergence, appear on the left-hand side of the oracle inequalities below. The bound of Alquier & Ridgway (2020, Theorem 2.7) instead controls a Rényi divergence by a Kullback-Leibler divergence.

Assumption 3.4. For $\beta > 0$, we define the following localized prior:

$$\pi_{-\beta}(d\theta) = \frac{e^{-\beta\,\mathrm{KL}(P^*\|P_\theta)}\pi(d\theta)}{Z_\beta}$$

where

$$Z_\beta = \int e^{-\beta \,\mathrm{KL}(P^* \| P_\theta)} \pi(d\theta).$$

We assume that there is a $\kappa_\pi \in (0, 1]$ and a $d_\pi > 0$ such that

$$\sup_{\beta > 0} \beta^{\kappa_\pi} \left\{ \mathbb{E}_{\theta \sim \pi_{-\beta}} \left[\mathrm{KL}(P^* \| P_\theta) \right] - \inf_{\tilde{\theta} \in \Theta} \mathrm{KL}(P^* \| P_{\tilde{\theta}}) \right\} \leq d_\pi.$$

The bracketed term is the average over $\pi_{-\beta}$ of the excess Kullback-Leibler risk of P_θ over the best in-model approximation. This is the misspecified complexity measure used by Alquier & Ridgway (2020, Section 2.4); here it is required to decay at rate $\beta^{-\kappa_\pi}$ as the localization parameter β grows. In the well-specified case $\inf_{\tilde{\theta} \in \Theta} \mathrm{KL}(P^* \| P_{\tilde{\theta}}) = 0$ and Assumption 3.4 reduces to the dimension condition of Catoni (2007).

When $\kappa_\pi = 1$, Assumption 3.4 is a standard assumption for obtaining parametric rates of convergence, with d_π interpreted as a notion of model complexity or dimension (Catoni, 2007). Similar assumptions are discussed by Baraud (2024); Alquier (2024). The case $\kappa_\pi < 1$ leads to nonparametric rates. We provide numerous examples where this assumption holds in Section 3.4. Lemma 3.25 below justifies the interpretation of d_π as a dimension.

While Assumptions 3.3 and 3.4 suffice to prove the convergence of the tempered posterior in expectation, stronger contraction results require a stronger version of Assumption 3.4: Assumption 3.5.

Assumption 3.5. Define

$$\mathcal{V}(\theta) = \mathbb{E}_{X \sim P^*} \left[\left(\log \frac{p^*(X)}{p_\theta(X)} \right)^2 \right].$$

We assume that the infimum $\inf_{\tilde{\theta} \in \Theta} \mathrm{KL}(P^* \| P_{\tilde{\theta}})$ is attained at some projection $\theta^\star$, and that there are $\kappa_\pi \in (0, 1]$, $d_\pi > 0$ and $d'_\pi > 0$ such that

$$\sup_{\beta > 0} \beta^{\kappa_\pi} \left\{ \mathbb{E}_{\theta \sim \pi_{-\beta}} \left[\mathrm{KL}(P^* \| P_\theta) \right] - \inf_{\tilde{\theta} \in \Theta} \mathrm{KL}(P^* \| P_{\tilde{\theta}}) \right\} \leq d_\pi,$$
$$\sup_{\beta > 0} \beta^{\kappa_\pi} \left\{ \mathbb{E}_{\theta \sim \pi_{-\beta}} \left[\mathcal{V}(\theta) \right] - \mathcal{V}(\theta^\star) \right\} \leq d'_\pi.$$

Here $\mathcal{V}(\theta^\star)$ is the residual fluctuation of the log-likelihood ratio at the projection $\theta^\star$. It vanishes in the well-specified case, where $\theta^\star = \theta_0$ and the log-ratio is identically zero, but is generally positive under misspecification.

We need a slightly different assumption to study the variational approximation $\widetilde{\rho}_{n,\alpha}$, fol-

lowing Alquier & Ridgway (2020). Assumption 3.6 extends Assumption 3.4 to this setting.

Assumption 3.6. We assume that there exist $\kappa_\pi \in (0, 1]$ and $d_\pi > 0$ such that, for any $\beta > 0$, there exists a $\rho \in \mathcal{F}$ such that

$$\mathbb{E}_{\theta\sim\rho}[\mathrm{KL}(P^*\|P_\theta)] - \inf_{\tilde\theta\in\Theta} \mathrm{KL}(P^*\|P_{\tilde\theta}) + \frac{\mathrm{KL}(\rho\|\pi_{-\beta})}{n} \le \frac{d_\pi}{\beta^{\kappa_\pi}}.$$

Theorem 3.7. *Under Assumptions 3.3 and 3.4, we have the following bound for any $\alpha \in (0, 1)$ and $\beta < n(1-\alpha)/c(\alpha)$, a range contained in $(0, \alpha n)$ by* (3.3)*:*

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi_{n,\alpha}}[\mathrm{KL}(P^*\|P_\theta)] \le \frac{c(\alpha)\alpha n - \beta c(\alpha)}{n(1-\alpha) - \beta c(\alpha)} \left(\inf_{\tilde\theta\in\Theta} \mathrm{KL}(P^*\|P_{\tilde\theta}) + \frac{d_\pi}{\beta^{\kappa_\pi}} \right).$$

Moreover, if Assumption 3.6 holds,

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\tilde\rho_{n,\alpha}}[\mathrm{KL}(P^*\|P_\theta)] \le \frac{c(\alpha)\alpha n - \beta c(\alpha)}{n(1-\alpha) - \beta c(\alpha)} \left(\inf_{\tilde\theta\in\Theta} \mathrm{KL}(P^*\|P_{\tilde\theta}) + \frac{d_\pi}{\beta^{\kappa_\pi}} \right).$$

In particular, for $\beta = n(1-\alpha)/(2c(\alpha))$, we have

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi_{n,\alpha}}[\mathrm{KL}(P^*\|P_\theta)] \le \alpha \frac{2c(\alpha)}{1-\alpha} \inf_{\tilde\theta\in\Theta} \mathrm{KL}(P^*\|P_{\tilde\theta}) + \alpha \left(\frac{2c(\alpha)}{1-\alpha} \right)^{1+\kappa_\pi} \frac{d_\pi}{n^{\kappa_\pi}}.$$

Remark 3.8. Theorem 3.7 is an oracle inequality of the same form as the misspecified bound of Alquier & Ridgway (2020, Theorem 2.7): the Kullback-Leibler risk of the fractional posterior is bounded by the best achievable in-model risk $\inf_{\tilde\theta\in\Theta} \mathrm{KL}(P^*\|P_{\tilde\theta})$, called the *oracle term* because it is the risk of an oracle that knows the best member of the model, plus a complexity term of order d_π/n^{κ_π}. When $\kappa_\pi = 1$, which holds in many examples below, this complexity term is of order $1/n$, against $\log(n)/n$ in Bhattacharya et al. (2019); Alquier & Ridgway (2020). In the well-specified case the oracle term vanishes and the bound is a contraction rate. The bound of Alquier & Ridgway (2020) controls a Rényi divergence by a Kullback-Leibler divergence; under Assumption 3.3, both sides here are Kullback-Leibler divergences.

Remark 3.9. The proof of Theorem 3.7 appears in Appendix B. While it can be derived as a corollary of Theorem 3.2, we follow a slightly different approach in Appendix B: we begin with PAC-Bayes bounds in expectation based on Bhattacharya et al. (2019); Alquier & Ridgway (2020) and demonstrate that optimization with respect to the prior leads to Theorem 3.2. The "optimal" prior, however, does not yield very explicit bounds. Catoni's localized priors (Catoni, 2007) were developed as approximations of the optimal priors and provide more explicit rates as shown in Theorem 3.7.

We also state a variant that holds with large probability rather than in expectation.

Theorem 3.10. *Under Assumptions (3.3,3.4,3.5), we have, for any* $\alpha, \delta, \eta \in (0,1)$ *and* $\beta = n(1-\alpha)/(2c(\alpha))$*, with probability at least* $1-\delta-\eta$ *on the sample* $\mathcal{S}$*,*

$$\mathbb{E}_{\theta\sim\pi_{n,\alpha}}[\mathrm{KL}(P^*\|P_\theta)] \le \alpha\,\frac{2c(\alpha)}{1-\alpha}\inf_{\tilde\theta\in\Theta}\mathrm{KL}(P^*\|P_{\tilde\theta}) + \alpha\left(\frac{2c(\alpha)}{1-\alpha}\right)^{1+\kappa_\pi}\frac{d_\pi}{n^{\kappa_\pi}} + \alpha\,\frac{2c(\alpha)}{1-\alpha}\sqrt{\frac{\mathcal{V}(\theta^\star)+d'_\pi/\beta^{\kappa_\pi}}{n\eta}} + \frac{2c(\alpha)\log\frac{1}{\delta}}{n(1-\alpha)}.$$

The square-root term is an irreducible $O(1/\sqrt{n})$ fluctuation governed by the residual variance $\mathcal{V}(\theta^\star)$ at the projection. In the well-specified case $\mathcal{V}(\theta^\star) = 0$ and this term is of order $n^{-(\kappa_\pi+1)/2}$, of the same nature as the fluctuation term in Alquier & Ridgway (2020). The misspecified analysis of Alquier & Ridgway (2020) is carried out in expectation; the high-probability statement here uses the same Chebyshev argument, now retaining the residual fluctuation at $\theta^\star$.

3.4 Examples

In this section, we consider various examples of statistical models. We discuss the validity of Assumption 3.3 and Assumption 3.4 (or its variants), and apply Theorem 3.7.

Throughout this section we place ourselves in the well-specified case, that is, $P^* = P_{\theta_0}$ for some $\theta_0 \in \Theta$. Then $\inf_{\tilde\theta\in\Theta}\mathrm{KL}(P^*\|P_{\tilde\theta}) = 0$ and $\mathcal{V}(\theta^\star) = 0$, so the oracle terms in Theorems 3.7 and 3.10 vanish and these results reduce to contraction bounds; we revert to writing θ_0 for the true parameter and $r_n(\theta,\theta_0)$ for the log-likelihood ratio. The misspecified statements specialize to this case, so verifying the assumptions here also illustrates the general theorems.

3.4.1 Gaussian Model

Gaussian mean estimation

Consider the case where the model is a d dimensional Gaussian $\mathcal{N}(\theta, v^2 I)$ with known covariance matrix $v^2 I$ and unknown mean θ, and a Gaussian prior on $\Theta = \mathbb{R}^d$:

$$\pi(\theta) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{\theta^T\Sigma^{-1}\theta}{2}\right), \tag{3.6}$$

$$P_\theta(x) = \frac{1}{(2\pi)^{d/2}v^d}\exp\left(-\frac{(x-\theta)^T(x-\theta)}{2v^2}\right). \tag{3.7}$$

Lemma 3.11. *When the model is given by* (3.7)*,*

$$\mathrm{KL}(P_{\theta_0}\|P_\theta) = \frac{1}{2v^2}\|\theta-\theta_0\|^2 = \frac{\frac{1}{\alpha}\alpha}{2v^2}\|\theta-\theta_0\|^2 = \frac{1}{\alpha}\mathcal{D}_\alpha(P_\theta\|P_{\theta_0}).$$

In other words, Assumption 3.3 is satisfied with $c(\alpha) = 1/\alpha$*.*

Lemma 3.12. *When the model and the prior are given by* (3.6) *and* (3.7)*,*

$$\beta\mathbb{E}_{\pi_{-\beta}}\left[\mathrm{KL}(P_{\theta_0}\|P_\theta)\right] = \frac{\beta}{2v^2}\left\|\left(\Sigma^{-1}+\frac{\beta}{v^2}I_d\right)^{-1}\Sigma^{-1}\theta_0\right\|^2 + \frac{\beta}{2v^2}\mathrm{Tr}\left[\left(\Sigma^{-1}+\frac{\beta}{v^2}I_d\right)^{-1}\right].$$

In particular, when $\Sigma = \sigma^2 I_d$*,*

$$\beta\mathbb{E}_{\pi_{-\beta}}\left[\mathrm{KL}(P_{\theta_0}\|P_\theta)\right] = \frac{\frac{\beta}{v^2}\|\theta_0\|^2}{2\sigma^4\left(\frac{\beta}{v^2}+\frac{1}{\sigma^2}\right)^2} + \frac{\frac{\beta}{v^2}d}{2\left(\frac{\beta}{v^2}+\frac{1}{\sigma^2}\right)} \le \frac{d}{2} + \frac{\|\theta_0\|^2}{8\sigma^2}$$

and thus Assumption 3.4 is satisfied with $\kappa_\pi = 1$ *and* $d_\pi = d/2 + \|\theta_0\|^2/(8\sigma^2)$*.*

Plugging Lemmas 3.11 and 3.12 into Theorem 3.7, we obtain the following result.

Corollary 3.13. *In the Gaussian model given by* (3.6) *and* (3.7) *and* $\Sigma = \sigma^2 I$*,*

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi_{n,\alpha}}\left[\mathrm{KL}(P_{\theta_0}\|P_\theta)\right] \le \frac{2d + \|\theta_0\|^2/(2\sigma^2)}{\alpha(1-\alpha)^2 n}.$$

In other words,

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi_{n,\alpha}}\left[\|\theta-\theta_0\|^2\right] \le \frac{v^2\left(4d + \|\theta_0\|^2/\sigma^2\right)}{\alpha(1-\alpha)^2 n}.$$

In this case the fractional posterior has an explicit form, so the expected risk could be derived directly. Still, this simple example illustrates an application of Theorem 3.7, and even in such a simple model the results of Bhattacharya et al. (2019); Alquier et al. (2016) would yield rates in $\log(n)/n$ rather than $1/n$.

We can also verify Assumption 3.5 in this setting.

Lemma 3.14. *In the Gaussian model given by* (3.6) *and* (3.7) *and* $\Sigma = \sigma^2 I$*, Assumption 3.5 holds with* $\kappa_\pi = 1$*,* $d_\pi = d/2 + \|\theta_0\|^2/(8\sigma^2)$ *and a finite constant* d'_π*.*

Gaussian sequence model

We now examine the Gaussian sequence model. While this model can be viewed as a variant of the previous one and remains simple, it usefully illustrates how Theorem 3.7 can lead to

nonparametric rates.

Following the notation of Castillo (2024): let $X_{i,j} = \theta_j + \varepsilon_{i,j}$ where $i = 1, 2, \dots, n$ and $j = 1, 2, \dots,$ the $\varepsilon_{i,j}$ are i.i.d. $\mathcal{N}(0,1)$, and

$$\theta_0 \in S_b(L) = \left\{ \theta : \sum_{\ell=1}^{\infty} \ell^{2b} \theta_\ell^2 \le L \right\}$$

for some $L, b > 0$. While this is infinite dimensional, we can truncate for $j \ge n$. That is, we consider p_θ as the density of $\mathcal{N}((\theta_1, \dots, \theta_n), 1) = \mathcal{N}((\theta_1, \dots, \theta_n), v^2)$ with $v = 1$. Using Lemma 3.11, we immediately obtain Assumption 3.3 with $c(\alpha) = 1/\alpha$.

Lemma 3.15. *Let $\pi = \mathcal{N}(0, \Sigma)$ with Σ a diagonal matrix, and let σ_i denote its diagonal elements. Put $\sigma_i = i^{-1-2b}$ (Castillo, 2024). Then*

$$\sup_\beta \beta^{\frac{2b}{1+2b}} \mathbb{E}_{\pi_{-\beta}} \left[\mathrm{KL}(P_{\theta_0} \| P_\theta) \right] \le \left(\frac{3L}{2} + \frac{1}{2} + \frac{1}{4b} \right).$$

In other words, Assumption 3.4 is satisfied with $d_\pi = 3L/2 + 1/2 + 1/(4b)$ and $\kappa_\pi = 2b/(1+2b)$.

Corollary 3.16. *In the Gaussian sequence model $p_\theta = \mathcal{N}((\theta_1, \dots, \theta_n), 1)$ with $(\theta_1, \theta_2, \dots) \in S_b(L)$ and π as in Lemma 3.15,*

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\theta \sim \pi_{n,\alpha}} [\mathrm{KL}(P_{\theta_0} \| P_\theta)] \le \alpha \left(\frac{2c(\alpha)}{1-\alpha} \right)^{1 + \frac{2b}{2b+1}} \frac{\frac{3L}{2} + \frac{1}{2} + \frac{1}{4b}}{n^{\frac{2b}{2b+1}}}.$$

3.4.2 Exponential Family

We now turn to more general statistical models. Assumption 3.3 can be verified in a broad class of models in the exponential family.

Definition 3.17. We say that $(P_\theta, \theta \in \Theta)$ is an exponential family with respect to a measure μ on $(\mathbb{X}, \mathcal{X})$ if there exist measurable functions $h : \mathbb{X} \to \mathbb{R}$ and $T : \mathbb{X} \to \mathbb{R}^d$ such that:

$$p_\theta(x) = \frac{dP_\theta}{d\mu}(x) = \exp\left(h(x) + \langle T(x), \theta \rangle - \psi(\theta) \right),$$

where $\psi(\theta) = \log\left(\int_{\mathbb{X}} \exp\left(h(x) + \langle T(x), \theta \rangle \right) \mu(dx) \right)$ is the partition function.

Lemma 3.18. *In the case $\mathbb{X} = \mathbb{R}^d$ and μ is the Lebesgue measure, assume that ψ is m-strongly convex and that its gradient is Lipschitz with constant L. Define the condition number of ψ as*

$\kappa = L/m$. *For any* $\alpha \in (0, 1)$,

$$\mathrm{KL}(P_{\theta_0} \| P_\theta) \leq \frac{L}{2}\|\theta_0 - \theta\|^2 \leq \frac{\kappa}{\alpha}\mathcal{D}_\alpha(P_\theta \| P_{\theta_0})$$

and thus Assumption 3.3 is satisfied with $c(\alpha) = \kappa/\alpha$.

Next, we provide examples of exponential families for which Assumption 3.3 holds.

Example 3.19 (Poisson Distribution)**.** Consider the Poisson distribution with probability mass function $\lambda^x/x!\, e^{-\lambda}$. The canonical decomposition yields $\theta = \log \lambda$ and $T(x) = x$. The partition function is given by:

$$\psi(\theta) = \lambda = e^\theta.$$

For $\Theta \subseteq [a, b]$, we have $m = e^a$ (strong convexity constant) and $L = e^b$ (Lipschitz constant), giving condition number $\kappa = L/m = e^{b-a}$. Thus, Assumption 3.3 is satisfied for the Poisson distribution when the parameter set Θ is compact.

Example 3.20 (Bernoulli Distribution)**.** Consider the Bernoulli distribution with probability mass function $p^x(1-p)^{1-x}$. The canonical decomposition yields $\theta = \log(p/(1-p))$ and $T(x) = x$. The partition function is given by:

$$\psi(\theta) = \log\left(1 + e^\theta\right).$$

The second derivative is:

$$\psi''(\theta) = \frac{e^\theta}{(1+e^\theta)^2}.$$

For $\Theta \subseteq [a, b]$, we have $m = \min\{\psi''(a), \psi''(b)\}$ (strong convexity constant), $L = 1/4$ (Lipschitz constant), and condition number $\kappa = L/m = 1/4m$. Thus, Assumption 3.3 is satisfied for the Bernoulli distribution on compact sets.

A discussion of Assumption 3.4 in these two examples will be included in Subsection 3.4.4.

3.4.3 Smooth models

The relationships between Fisher information and various divergences are closely tied to Le Cam's quadratic mean differentiability (QMD) (Le Cam, 2012) in asymptotic statistics. Following Polyanskiy & Wu (2025, Section 7.11), we present calculations that verify Assumption 3.3 holds in a local neighborhood of θ_0. These calculations build upon previous results on the local behavior of the Hellinger distance (Baraud et al., 2017) and KL divergence (van der

Vaart, 2000, Ch. 7.1), unified under the f-divergence framework introduced by Csiszár (1967). We first recall the Fisher information, which sets the scale of all three expansions below.

Definition 3.21. The Fisher information at θ is

$$I(\theta) = \mathbb{E}_{P_\theta}\left[\left(\frac{\partial}{\partial\theta}\log p_\theta(X)\right)^2\right].$$

We work under the following regularity assumption, in which $\alpha \in (0,1)$ is the order at which Assumption 3.3 is to be verified.

Assumption 3.22. Fix $\alpha \in (0,1)$. There is an open interval U containing θ_0 such that

(i) the densities p_θ, $\theta \in U$, have common support;

(ii) for μ-almost every x, the map $\theta \mapsto \log p_\theta(x)$ is three times continuously differentiable on U;

(iii) for every compact $V \subset U$, the derivatives in θ through order three of

$$p_\theta(x), \qquad p_{\theta_0}(x)\left(\log p_{\theta_0}(x) - \log p_\theta(x)\right), \qquad p_\theta(x)^\gamma p_{\theta_0}(x)^{1-\gamma} \quad (\gamma \in \{\alpha, 1/2\})$$

are bounded in absolute value, uniformly in $\theta \in V$, by μ-integrable functions of x;

(iv) $0 < I(\theta_0) < \infty$.

The envelopes in (iii) allow differentiation under the integral sign, by dominated convergence, of $\int p_\theta\, d\mu$, of $\theta \mapsto \mathrm{KL}(P_{\theta_0}\|P_\theta)$ and of the affinities $A_\gamma(\theta) := \int p_\theta^\gamma p_{\theta_0}^{1-\gamma} d\mu$ for $\gamma \in \{\alpha, 1/2\}$, and make the resulting third derivatives continuous on U. The order α is fixed; a statement uniform over all orders would require uniform control of the envelopes.

The three divergences are then quadratic at θ_0, with a cubic remainder controlled in absolute value.

Theorem 3.23. *Let Assumption 3.22 hold. There exist $r > 0$ and finite constants $C_H, C_K, C_R \geq 0$ such that, for every θ with $|\theta - \theta_0| \leq r$,*

$$\left| H^2(P_\theta, P_{\theta_0}) - \frac{I(\theta_0)}{4}(\theta-\theta_0)^2 \right| \leq C_H\,|\theta-\theta_0|^3, \tag{3.8}$$

$$\left| \mathrm{KL}(P_{\theta_0}\|P_\theta) - \frac{I(\theta_0)}{2}(\theta-\theta_0)^2 \right| \leq C_K\,|\theta-\theta_0|^3, \tag{3.9}$$

$$\left| \mathcal{D}_\alpha(P_\theta\|P_{\theta_0}) - \frac{\alpha\, I(\theta_0)}{2}(\theta-\theta_0)^2 \right| \leq C_R\,|\theta-\theta_0|^3. \tag{3.10}$$

One may take $C_j = \frac{1}{6}\sup_{|t-\theta_0|\leq r}|\Delta_j'''(t)|$, where Δ_H, Δ_K and Δ_R denote the three divergences read as functions of θ.

The leading coefficients are those of the classical expansions; only the remainders are stated differently, as bounds rather than as identities. A bound on the third derivative is all a cubic remainder needs, so no derivative of the Fisher information enters the statement. The radius r and the constant C_R may depend on α. Expansions of this kind are classical; van der Vaart (2000, Section 7.2) obtains the Hellinger expansion from differentiability in quadratic mean.

Prescribing a margin η turns the expansions into two-sided bounds on an explicit neighborhood, and the comparison required by Assumption 3.3 follows by division.

Corollary 3.24. *Let Assumption 3.22 hold, let r and C_H, C_K, C_R be as in Theorem 3.23, and fix $\eta \in (0,1)$. Set*

$$b_H := \frac{I(\theta_0)}{4}, \qquad b_K := \frac{I(\theta_0)}{2}, \qquad b_R := \frac{\alpha I(\theta_0)}{2},$$
$$r_{\alpha,\eta} := \min\left\{ r,\ \min_{j\in\{H,K,R\}:\, C_j>0} \frac{\eta\, b_j}{C_j} \right\}, \tag{3.11}$$

the inner minimum being omitted when every C_j vanishes. Then $r_{\alpha,\eta} \in (0, r]$ and, for every θ with $|\theta - \theta_0| \le r_{\alpha,\eta}$,

$$(1-\eta)\frac{I(\theta_0)}{4}(\theta-\theta_0)^2 \le H^2(P_\theta, P_{\theta_0}) \le (1+\eta)\frac{I(\theta_0)}{4}(\theta-\theta_0)^2 \tag{3.12}$$
$$(1-\eta)\frac{I(\theta_0)}{2}(\theta-\theta_0)^2 \le \mathrm{KL}(P_{\theta_0}\|P_\theta) \le (1+\eta)\frac{I(\theta_0)}{2}(\theta-\theta_0)^2 \tag{3.13}$$
$$(1-\eta)\frac{\alpha I(\theta_0)}{2}(\theta-\theta_0)^2 \le \mathcal{D}_\alpha(P_\theta\|P_{\theta_0}) \le (1+\eta)\frac{\alpha I(\theta_0)}{2}(\theta-\theta_0)^2. \tag{3.14}$$

Consequently, on the same neighborhood,

$$\mathrm{KL}(P_{\theta_0}\|P_\theta) \ \le\ \frac{1+\eta}{\alpha(1-\eta)}\, \mathcal{D}_\alpha(P_\theta\|P_{\theta_0}). \tag{3.15}$$

Thus, if the model is smooth enough, then Assumption 3.3 is satisfied at the prescribed order α, on the neighborhood of θ_0 of radius $r_{\alpha,\eta}$, with $c(\alpha) = (1+\eta)/(\alpha(1-\eta))$ by (3.15). Both the radius and the constant depend on α; a single neighborhood serving every order would require a bound on C_R/α uniform in α. This expansion is stated for a one-dimensional parameter; the extension to multivariate product models follows from Lemma 3.25.

3.4.4 Dimension Assumption

We begin with a lemma that explains why we refer to d_π as a dimension.

Lemma 3.25. *Assume that Θ, equipped with the prior π, satisfies Assumption 3.4 uniformly, in*

other words, for any θ_0,

$$\sup_{\beta>0} \beta^{\kappa_\pi} \mathbb{E}_{\theta\sim\pi_{-\beta}} [\mathrm{KL}(P_{\theta_0} \| P_\theta)] \le d_\pi$$

where κ_π *and* d_π *do not depend on* θ_0. *Define a new statistical model* $\Theta_p := \Theta^p$ *that parametrizes product distributions: for any* $\theta = (\theta_1, \dots, \theta_p) \in \Theta_p$, $P_\theta := P_{\theta_1} \otimes \cdots \otimes P_{\theta_p}$, *and equip* Θ_p *with the product prior* $\pi^{(p)} := \pi^{\otimes p}$. *Then, for any* $\theta_0 \in \Theta_p$,

$$\sup_{\beta>0} \beta^{\kappa_\pi} \mathbb{E}_{\theta\sim\pi^{(p)}_{-\beta}} [\mathrm{KL}(P_{\theta_0} \| P_\theta)] \le d_\pi \, p.$$

Proof. Write $V_j(\theta_j) := \mathrm{KL}(P_{\theta_{0,j}} \| P_{\theta_j})$, so that additivity of the Kullback-Leibler divergence over product measures gives $\mathrm{KL}(P_{\theta_0} \| P_\theta) = \sum_{j=1}^p V_j(\theta_j)$. With the product prior $\pi^{(p)} = \pi^{\otimes p}$, the localized prior factorizes:

$$\pi^{(p)}_{-\beta}(d\theta) \propto \exp\left(-\beta \sum_{j=1}^p V_j(\theta_j)\right) \prod_{j=1}^p \pi(d\theta_j) = \prod_{j=1}^p e^{-\beta V_j(\theta_j)} \pi(d\theta_j),$$

that is, $\pi^{(p)}_{-\beta} = \bigotimes_{j=1}^p \pi^{(j)}_{-\beta}$ with one-coordinate marginals $\pi^{(j)}_{-\beta}(d\theta_j) \propto e^{-\beta V_j(\theta_j)} \pi(d\theta_j)$, which differ from one another when the $\theta_{0,j}$ do. Taking expectations coordinate by coordinate,

$$\sup_{\beta>0} \beta^{\kappa_\pi} \mathbb{E}_{\theta\sim\pi^{(p)}_{-\beta}} [\mathrm{KL}(P_{\theta_0} \| P_\theta)] \le \sum_{j=1}^p \sup_{\beta>0} \beta^{\kappa_\pi} \mathbb{E}_{\theta_j\sim\pi^{(j)}_{-\beta}} [V_j(\theta_j)] \le p \, d_\pi,$$

each term being bounded by d_π through the hypothesis applied in the one-dimensional model at $\theta_{0,j}$. □

We now verify that under suitable conditions on the KL divergence, Assumption 3.4 holds for one-dimensional compact models such as those considered in Examples 3.19 and 3.20 above. Thanks to Lemma 3.25, this can be extended to multidimensional examples. Motivated by the discussion in the exponential family, we consider models where the KL divergence satisfies

$$\frac{m}{2} \|\theta - \theta_0\|^2 \le \mathrm{KL}(P_{\theta_0} \| P_\theta) \le \frac{L}{2} \|\theta - \theta_0\|^2 \tag{3.16}$$

for some $0 < m < L < +\infty$.

Lemma 3.26. *When the prior is uniform on* $\Theta = (-M, M)$ *and* $\theta_0 \in (-M, M)$, *under the assumption that* (3.16) *holds, with* $\kappa = L/m$,

$$\sup_{\beta} \beta \mathbb{E}_{\theta\sim\pi_{-\beta}} [\mathrm{KL}(P_{\theta_0} \| P_\theta)] \le \frac{\kappa^{\frac{3}{2}}}{2},$$

and thus Assumption 3.4 is satisfied with $\kappa_\pi = 1$ *and* $d_\pi = \kappa^{3/2}/2$.

3.4.5 Application to the MLE

The previous subsections provided many examples where $\mathcal{D}_\alpha(P_\theta \| P_{\theta_0}) \geq m\|\theta - \theta_0\|^2$ for some constant $m > 0$. In this setting, and under additional smoothness assumptions on the likelihood, we can use Theorem 3.2 to study the MLE. Specifically, we will assume that $r_n(\theta, \theta_0)/n$ is Lipschitz with respect to θ.

Corollary 3.27. *Fix* $\alpha \in (0, 1)$. *Assume* Θ *is a compact set in* $\mathbb{R}^d$. *Assume there exists a constant* $m > 0$ *such that, for any* $\theta \in \Theta$, $\mathcal{D}_\alpha(P_\theta \| P_{\theta_0}) \geq m\|\theta - \theta_0\|^2$. *Assume moreover that, almost surely with respect to the sample,* $\theta \mapsto r_n(\theta, \theta_0)/n$ *is L-Lipschitz. Let* $\mathcal{N}(\Theta, \varepsilon)$ *denote the minimal number of balls of radius* $\varepsilon > 0$ *(with respect to the Euclidean norm) required to cover* Θ. *Then*

$$\mathbb{E}_{\mathcal{S}}\left[\|\hat{\theta}_n - \theta_0\|^2\right] \leq \frac{1}{n}\left[\frac{2\alpha L + 2\log \mathcal{N}(\Theta, 1/n)}{m(1-\alpha)} + \frac{1}{n}\right]. \tag{3.17}$$

While the analysis of the MLE is not new, the point here is that this result follows as a corollary of Theorem 3.2. This shows that Theorem 3.2 extends beyond the analysis of Bayesian estimators.

Since Θ is a compact set in $\mathbb{R}^d$, we expect $\mathcal{N}(\Theta, \varepsilon) = \mathcal{O}\left((1/\varepsilon)^d\right)$. In this case, Corollary 3.27 gives

$$\mathbb{E}_{\mathcal{S}}\left[\|\hat{\theta}_n - \theta_0\|^2\right] = \mathcal{O}\left(\frac{2d\log(n)}{nm(1-\alpha)}\right).$$

Compared to the Bayesian estimators, we lose a $\log(n)$ term. Whether this $\log(n)$ term can be eliminated in a general setting while still using an MI bound remains, to our knowledge, an open question.

Chapter 4

Robust Bayesian Inference via Variational Approximations of Generalized ρ-Posteriors

This chapter reproduces, with minor typographic modifications, the paper:

E. M. Khribch, P. Alquier

Robust Bayesian Inference via Variational Approximations of Generalized ρ-Posteriors

See Khribch & Alquier (2026).

We introduce the $\tilde{\rho}$-posterior, a modified version of the ρ-posterior of Baraud & Birgé (2020), obtained by replacing the supremum over competitor parameters with a softmax aggregation. This modification allows a PAC-Bayesian analysis of the $\tilde{\rho}$-posterior. This yields finite-sample oracle inequalities with explicit convergence rates that inherit the robustness properties of the original framework, in particular graceful degradation under model misspecification and data contamination: the guarantee is not lost when the model is wrong, but weakened by an additive term that vanishes with the size of the perturbation. The PAC-Bayesian oracle inequalities extend to variational approximations of the $\tilde{\rho}$-posterior, providing theoretical guarantees for tractable inference. Numerical experiments on exponential families, regression, and real-world datasets confirm that the resulting variational procedures achieve robustness competitive with theoretical predictions at computational cost comparable to standard variational Bayes.

Keywords: Robust Bayesian inference; PAC-Bayes; Variational inference; ρ-posterior; Hellinger distance; Generalized Bayesian inference.

4.1 Introduction and Motivation

4.1.1 Universal Estimation and the Robustness Problem

A fundamental challenge in statistical inference is the construction of universal estimation procedures: given an observed sample $\mathcal{S} = (X_1, \dots, X_n)$ from unknown distributions $P_*^1, \dots, P_*^n$, one seeks methods that achieve minimax-optimal rates when the model is well-specified, yet degrade gracefully under misspecification (Bickel et al., 1976; Birgé, 2006). Most classical procedures, including maximum likelihood estimation, method of moments, and standard Bayesian inference, fail to satisfy both requirements simultaneously.

The non-universality of maximum likelihood is well illustrated by an example due to Birgé (2006). Consider n independent observations from the mixture

$$P_* = \left(1 - \frac{2}{n}\right) \mathcal{U}\left(\left[0, \frac{1}{10}\right]\right) + \frac{2}{n} \mathcal{U}\left(\left[\frac{1}{10}, \frac{9}{10}\right]\right),$$

modeled by $\{P_\theta = \mathcal{U}([0,\theta]) : \theta \in [0,1)\}$. Since $\mathcal{H}^2(P_*, P_{1/10}) < 5/(4n)$ for $n \geq 4$, the parameter $\theta = 1/10$ provides an excellent approximation in squared Hellinger risk. Yet the MLE $\hat{\theta}_{\text{MLE}} = X_{(n)}$ satisfies $\mathbb{E}[\mathcal{H}^2(P_*, P_{\hat{\theta}_{\text{MLE}}})] > 0.38$, failing even to be consistent. Although maximum likelihood achieves optimality under classical regularity conditions (LeCam, 1970; van der Vaart, 2000), it can degrade catastrophically under misspecification.

Standard Bayesian inference suffers from analogous brittleness. The posterior

$$\pi_n(\theta) \propto \pi(\theta) \prod_{i=1}^{n} p_\theta(X_i)$$

concentrates around θ_0 at the optimal rate in the well-specified setting $P_*^i = P_{\theta_0}$, and credible sets are asymptotically valid confidence sets by the Bernstein–von Mises theorem (van der Vaart, 2000; Kleijn & van der Vaart, 2012). Under misspecification, however, a single corrupted observation can cause the posterior to concentrate arbitrarily far from the truth (Barron et al., 1999; Grünwald & van Ommen, 2017; Owhadi et al., 2015; Baraud et al., 2017). The source of this fragility is the implicit commitment of Bayes' rule to minimizing $\mathrm{KL}(P_* \| P_\theta)$, a divergence that is unbounded whenever the model assigns zero density to events of positive probability (Hampel et al., 2009).

The connection between universality and robustness is made precise through the contamination model of Huber (1992): a fraction ε of observations is adversarially corrupted, so that $P_* = (1-\varepsilon)P_{\theta_0} + \varepsilon Q$ for some $\theta_0 \in \Theta$ and arbitrary Q. Frequentist robustness requires continuity with respect to total variation (Hampel et al., 2009), while Bayesian posterior consistency relies on Kullback–Leibler divergence (Kleijn & van der Vaart, 2012), a substantially stronger condition. As Owhadi et al. (2015) observe, this mismatch creates fundamental ob-

stacles for robust Bayesian inference. The Hellinger distance bridges the two perspectives: it satisfies $\mathcal{H}^2(P,Q) \leq \mathrm{TV}(P,Q) \leq \sqrt{2}\,\mathcal{H}(P,Q)$ (Tsybakov, 2008) and, unlike KL divergence, remains bounded when the supports of P and Q differ.

4.1.2 Robust Bayesian Inference via Generalized Posteriors

Generalized Bayesian inference (Chernozhukov & Hong, 2003; Bissiri et al., 2016) provides a principled route to robustness by replacing the log-likelihood with an alternative loss $\ell_n(\theta)$:

$$\pi_n^{\ell}(\theta) \propto \pi(\theta)\exp(-\beta n\,\ell_n(\theta))\,.$$

Equivalently, the generalized posterior minimizes

$$\pi_n^{\ell} = \operatorname*{argmin}_{\rho\in\mathcal{P}(\Theta)} \left\{\beta n\,\mathbb{E}_{\theta\sim\rho}[\ell_n(\theta)] + \mathrm{KL}(\rho\|\pi)\right\},$$

an optimization-centric viewpoint that unifies Bayesian updating, variational inference, and generalized posteriors (Knoblauch et al., 2022), and reveals that standard Bayesian inference implicitly minimizes $\mathrm{KL}(P_*\|P_\theta)$.

An important special case is that of *tempered* or *fractional* posteriors, obtained by raising the likelihood to a power $\beta \in (0,1)$. Non-asymptotic concentration inequalities for such posteriors and their variational approximations were established by Bhattacharya et al. (2019); Alquier & Ridgway (2020); Khribch & Alquier (2024) subsequently sharpened these rates through an information-theoretic framework based on mutual information bounds.

More broadly, the robustness properties of a generalized posterior are inherited from the underlying loss function. Replacing KL divergence with Hellinger distance yields posteriors insensitive to tail behavior (Hooker & Vidyashankar, 2014); this principle extends to α- and β-divergences (Ghosh & Basu, 2016), γ-divergences (Nakagawa & Hashimoto, 2020), and was systematically formalized by Jewson et al. (2018), who showed that boundedness and total variation continuity of the divergence translate directly to posterior robustness. Related work includes kernel Stein discrepancy methods (Matsubara et al., 2022), scoring-rule inference (Giummolè et al., 2019; Pacchiardi & Dutta, 2021), power posteriors (Holmes & Walker, 2017), coarsened posteriors (Miller & Dunson, 2019), bagged posteriors (Huggins & Miller, 2019), the safe Bayes framework (Grünwald, 2011; Grünwald & van Ommen, 2017; Heide et al., 2020), and MMD-Bayes (Chérief-Abdellatif & Alquier, 2020; Dellaporta et al., 2022), which achieves minimax-optimal rates via kernel embeddings.

An alternative strategy constructs log-likelihood proxies through the median-of-means principle (Lecué & Lerasle, 2020; Minsker & Yao, 2025). A fundamentally different approach, due to Catoni (2012), Baraud et al. (2017), and Baraud & Birgé (2020), replaces the log-likelihood with bounded contrast functions, yielding the ρ-estimator (frequentist) and the ρ-

posterior (Bayesian). The resulting competitor-based risk decomposition separates estimation error from model misspecification, achieving universal estimation in both frameworks.

4.1.3 The Rho-Posterior Framework

The theoretical foundations for robust testing-based estimation originate in the work of LeCam (1973); Le Cam (1975) on asymptotic decision theory, subsequently developed by Birgé (1983, 2006) into T-estimation, a framework that constructs estimators through robust tests comparing the fitted model against reference distributions. While T-estimators achieve consistency and robustness, they require compactness of the parameter space.

To remove this restriction, Baraud & Birgé (2018) and Baraud et al. (2017) introduced ρ-estimation. The method aggregates robust pairwise tests between models into an empirical supremum risk, and defines the ρ-estimator as the parameter minimizing the worst-case discrepancy against all competitors. Unlike T-estimation, ρ-estimation extends to noncompact models, including linear regression under various error distributions, and recovers the MLE asymptotically when the model is well-specified and regular.

Baraud & Birgé (2020) introduced the ρ-posterior, a Bayesian counterpart that replaces the likelihood with the same robust testing-based criterion. The resulting posterior satisfies several theoretical properties: explicit contamination rates degrading gracefully with the fraction of corrupted observations, minimax-optimal convergence, and versions of the Bernstein–von Mises theorem valid under outlier contamination. These guarantees place the ρ-posterior among the most theoretically developed robust Bayesian frameworks.

The ρ-posterior has, however, remained a theoretical construct. The computational obstruction is structural: its definition requires, for each parameter value, a worst-case comparison over the entire parameter space, a structure that admits neither MCMC sampling nor standard variational inference. The present work addresses this gap.

Our Contributions

We define a modified version of the ρ-posteriors of Baraud & Birgé (2020), which we call *generalized ρ-posteriors* (or $\tilde{\rho}$-posteriors). While the original ρ-posterior is constructed from a supremum contrast that resists both PAC-Bayesian analysis and tractable computation, the $\tilde{\rho}$-posterior replaces this supremum with a softmax aggregation over competitor parameters, yielding a Gibbs posterior amenable to the PAC-Bayesian machinery of McAllester (1999); Catoni (2007); Alquier (2024).

We show that, unlike the ρ-posteriors of Baraud & Birgé (2020), the $\tilde{\rho}$-posteriors can be analyzed via PAC-Bayesian bounds. In Theorem 4.5, we establish finite-sample bounds controlling the expected Hellinger risk of the $\tilde{\rho}$-posterior by the oracle approximation error plus a complexity penalty, under both i.i.d. and independent non-identically distributed observa-

tions. Through softmax aggregation, Theorem 4.10 yields oracle inequalities with dimension-dependent rates matching the minimax-optimal rates obtained by Baraud & Birgé (2020) for the original ρ-posterior. As an application, we derive corresponding results for fixed-design regression (Theorem 4.17).

An advantage of the PAC-Bayesian perspective is that the resulting oracle inequalities extend naturally to variational approximations of the $\tilde{\rho}$-posterior. Theorem 4.23 establishes that variational approximations inherit the robustness properties of the exact $\tilde{\rho}$-posterior with explicit control on the approximation quality. For exponential families with mean-field Gaussian variational families, the resulting saddle-point problem admits favorable nonconvex–strongly concave geometry, ensuring convergence of first-order methods uniformly in the temperature parameter (Theorem 4.33).

Finally, numerical experiments on exponential families under contamination, regression with correlated designs, and real-world datasets confirm that the variational $\tilde{\rho}$-posteriors achieve robustness competitive with theoretical predictions, at computational cost comparable to standard variational Bayes.

Setup and Notation

Let $\mathcal{S} = (X_1, \dots, X_n)$ denote n independent observations taking values in a measurable space $(\mathcal{X}, \mathcal{A})$ equipped with a σ-finite reference measure μ, and let $\{P_\theta : \theta \in \Theta\}$ be a parametric model. We impose two standing assumptions.

Assumption 4.1 (Dominated model)**.** For each $\theta \in \Theta$ and $i \in [n]$, the model distribution P_θ^i for the i-th observation is dominated by the σ-finite measure μ, and admits a density $p_\theta^i = dP_\theta^i/d\mu$. The product model distribution is $P_\theta^{(n)} := \bigotimes_{i=1}^n P_\theta^i$.

Assumption 4.2 (Data-generating distribution)**.** The observed sample $\mathcal{S} = (X_1, \dots, X_n)$ consists of n independent observations, where X_i is drawn from an unknown distribution P_*^i with density p_*^i with respect to μ for each $i \in [n]$. The product distribution is $P_*^{(n)} := \bigotimes_{i=1}^n P_*^i$. We write $\mathbb{E}_\mathcal{S}$ and $\mathbb{P}_\mathcal{S}$ for expectations and probabilities under $P_*^{(n)}$.

Remark 4.3 (i.i.d. observations as a special case)**.** When $p_*^1 = \cdots = p_*^n =: p_*$ and $p_\theta^1 = \cdots = p_\theta^n =: p_\theta$ for all $\theta \in \Theta$, Assumptions 4.1–4.2 reduce to the standard i.i.d. framework with $P_* = p_* \cdot \mu$ and model $\{P_\theta = p_\theta \cdot \mu : \theta \in \Theta\}$. All subsequent results specialize accordingly.

Sample Hellinger distance The natural metric for independent observations is the *coordinate-averaged squared Hellinger distance*:

$$\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) := \frac{1}{n}\sum_{i=1}^n \mathcal{H}^2(P_*^i, P_\theta^i) = \frac{1}{n}\sum_{i=1}^n \mathcal{H}^2(p_*^i, p_\theta^i), \tag{4.1}$$

where $\mathcal{H}^2(P,Q) = \frac{1}{2}\int(\sqrt{dP} - \sqrt{dQ})^2$ denotes the squared Hellinger distance. In the i.i.d. case (Remark 4.3), this reduces to $\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) = \mathcal{H}^2(p_*, p_\theta) =: \mathcal{H}^2(P_*, P_\theta)$.

The contrast function ψ Following Baraud et al. (2017), the unbounded log-likelihood is replaced by a bounded contrast $\psi : \mathbb{R}_+ \cup \{+\infty\} \to [-1, 1]$,

$$\psi(x) = \begin{cases} \dfrac{\sqrt{x}-1}{\sqrt{x}+1} & \text{for } 0 \le x < +\infty, \\ 1 & \text{for } x = +\infty, \end{cases} \tag{4.2}$$

with conventions $0/0 = 1$ and $a/0 = +\infty$ for $a > 0$. The function ψ is strictly increasing, antisymmetric ($\psi(1/x) = -\psi(x)$), bounded in $[-1,1]$, and Lipschitz continuous with constant 2 as a function of $\sqrt{x}$. Setting $\phi = 4\psi$, one has $|\phi(x) - \log x| \le 0.055|x-1|^3$ for $x \in [1/2, 2]$, so that ψ approximates the log-likelihood locally while remaining globally bounded.

Pairwise and supremum contrasts For $\theta, \theta' \in \Theta$ and observation X_i, the *coordinate-wise contrast* is defined by

$$\ell_\psi(x_i; \theta, \theta') := \psi\left(\frac{p_{\theta'}^i(x_i)}{p_\theta^i(x_i)}\right), \tag{4.3}$$

with population and empirical versions given by coordinate-averaging:

$$R_\psi(\theta, \theta') := \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{X_i \sim P_*^i}[\ell_\psi(X_i; \theta, \theta')] = \frac{1}{n}\sum_{i=1}^n \int \psi\left(\frac{p_{\theta'}^i}{p_\theta^i}\right) dP_*^i, \tag{4.4}$$

$$\hat{R}_\psi(\theta, \theta') := \frac{1}{n}\sum_{i=1}^n \ell_\psi(X_i; \theta, \theta'). \tag{4.5}$$

We also define the variance

$$V_\psi(\theta, \theta') := \mathbb{V}_{\mathcal{S}}[\hat{R}_\psi(\theta, \theta')] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}_{X_i \sim P_*^i}[\ell_\psi(X_i; \theta, \theta')], \tag{4.6}$$

by independence. The *supremum contrasts* are

$$R_\psi^*(\theta) := \sup_{\theta' \in \Theta} R_\psi(\theta, \theta'), \quad \hat{R}_\psi^*(\theta) := \sup_{\theta' \in \Theta} \hat{R}_\psi(\theta, \theta'), \tag{4.7}$$

and the ρ-estimator is $\hat{\theta}_{n,\psi} \in \operatorname{argmin}_{\theta\in\Theta} \hat{R}^*_\psi(\theta)$. In the i.i.d. case (Remark 4.3), these reduce to $R_\psi(\theta,\theta') = \mathbb{E}_{X\sim P_*}[\ell_\psi(X;\theta,\theta')]$ and $V_\psi(\theta,\theta') = n^{-1}\mathbb{V}_{X\sim P_*}[\ell_\psi(X;\theta,\theta')]$.

The following lemma, due to Baraud & Birgé (2018), relates the ψ-contrast to the Hellinger geometry of the product laws $P_\theta^{(n)} := \bigotimes_{i=1}^n P_\theta^i$.

Lemma 4.4 (Proposition 3 in Baraud & Birgé (2018)). *For $\psi(x) = (\sqrt{x}-1)/(\sqrt{x}+1)$, there exist universal constants $a_0 = 4$, $a_1 = 3/8$, $a_2^2 = 3\sqrt{2}$ such that for all $(\theta,\theta') \in \Theta^2$,*

$$\begin{aligned} a_1\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) - a_0\mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)}) &\le R_\psi(\theta,\theta') \\ &\le a_0\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) - a_1\mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)}), \quad (4.8) \\ V_\psi(\theta,\theta') &\le \frac{a_2^2}{n}\left(\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) + \mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)})\right). \quad (4.9) \end{aligned}$$

In the i.i.d. case (Remark 4.3), these bounds hold with $\mathcal{H}_n^2$ replaced by $\mathcal{H}^2$.

4.2 Main Results

Our main result is a finite-sample PAC-Bayes bound for a Gibbs posterior constructed from the ψ-contrast.

Theorem 4.5 (PAC-Bayes bound for independent observations). *Fix $\delta \in (0,1)$, $\lambda > 0$, and priors $\pi, \pi' \in \mathcal{P}(\Theta)$. Define the softmax competitor functional*

$$\Lambda_\lambda(\theta;\pi') := \frac{1}{\lambda}\log\left(\int_\Theta e^{\lambda\,\hat{R}_\psi(\theta,\theta')}\pi'(d\theta')\right), \quad (4.10)$$

and let $\hat{\rho}_\lambda$ be any minimizer of

$$\rho \mapsto \mathbb{E}_{\theta\sim\rho}\left[\Lambda_\lambda(\theta;\pi')\right] + \frac{1}{\lambda}\mathrm{KL}(\rho\|\pi). \quad (4.11)$$

Then with probability at least $1-\delta$ over $\mathcal{S} = (X_1,\dots,X_n)$,

$$\begin{aligned} &\left(a_1 - \beta_{n,\lambda}a_2^2\right)\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}\left[\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)})\right] \\ &\le \inf_{\rho\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda(\theta;\pi')] + \frac{1}{\lambda}\mathrm{KL}(\rho\|\pi)\right\} \\ &\quad + \inf_{\rho'\in\mathcal{P}(\Theta)}\left\{\left(a_0 + \beta_{n,\lambda}a_2^2\right)\mathbb{E}_{\theta'\sim\rho'}\left[\mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)})\right] + \frac{2}{\lambda}\mathrm{KL}(\rho'\|\pi')\right\} \\ &\quad + \frac{\log(1/\delta)}{\lambda}, \end{aligned} \quad (4.12)$$

where $a_0 = 4$, $a_1 = 3/8$, $a_2^2 = 3\sqrt{2}$, *and*

$$\beta_{n,\lambda} = g\left(\frac{2\lambda}{n}\right)\frac{\lambda}{n}, \quad g(x) = \begin{cases} (e^x - 1 - x)/x^2 & x \neq 0 \\ 1/2 & x = 0. \end{cases} \tag{4.13}$$

The proof is given in Appendix C.

An explicit choice of inverse temperature yields a simplified bound with interpretable constants.

Corollary 4.6. *Under the setting of Theorem 4.5, take* $\lambda = n/8$. *Then with probability at least* $1 - \delta$,

$$\begin{aligned} \frac{1}{12}\,\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}\left[\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)})\right] \le \inf_{\rho\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda(\theta;\pi')] + \frac{8}{n}\mathrm{KL}(\rho\|\pi)\right\} \\ + \inf_{\rho'\in\mathcal{P}(\Theta)}\left\{\frac{13}{3}\,\mathbb{E}_{\theta'\sim\rho'}\left[\mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)})\right] + \frac{16}{n}\mathrm{KL}(\rho'\|\pi')\right\} \\ + \frac{8\log(1/\delta)}{n}. \end{aligned} \tag{4.14}$$

The remainder of this section specializes to the i.i.d. case (Remark 4.3); all results extend to independent observations upon replacing $\mathcal{H}^2(P_*, P_\theta)$ by $\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)})$.

Corollary 4.7 (i.i.d. case)**.** *Suppose* $X_1, \ldots, X_n$ *are i.i.d. with* $X_i \sim P_* := p_* \cdot \mu$ *for all* i, *and the model densities satisfy* $p_\theta^i = p_\theta$ *for all* i *and* $\theta \in \Theta$. *Then Theorem 4.5 holds with* $\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)})$ *replaced by* $\mathcal{H}^2(P_*, P_\theta)$ *throughout. Specifically, with probability at least* $1 - \delta$,

$$\begin{aligned} &\left(a_1 - \beta_{n,\lambda}a_2^2\right)\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}\left[\mathcal{H}^2(P_*, P_\theta)\right] \\ &\le \inf_{\rho\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda(\theta;\pi')] + \frac{1}{\lambda}\mathrm{KL}(\rho\|\pi)\right\} \\ &\quad + \inf_{\rho'\in\mathcal{P}(\Theta)}\left\{\left(a_0 + \beta_{n,\lambda}a_2^2\right)\mathbb{E}_{\theta'\sim\rho'}\left[\mathcal{H}^2(P_*, P_{\theta'})\right] + \frac{2}{\lambda}\mathrm{KL}(\rho'\|\pi')\right\} \\ &\quad + \frac{\log(1/\delta)}{\lambda}. \end{aligned} \tag{4.15}$$

The proof is given in Appendix C.

Setting $\lambda = n/8$ yields explicit constants.

Corollary 4.8 (Explicit inverse temperature, i.i.d. case)**.** *Under the setting of Corollary 4.7, take*

$\lambda = n/8$. *Then with probability at least* $1 - \delta$,

$$\frac{1}{12}\,\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[\mathcal{H}^2(P^*, P_\theta)] \le \inf_{\rho\in\mathcal{P}(\Theta)} \left\{ \mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda(\theta;\pi')] + \frac{8\,\mathrm{KL}(\rho\|\pi)}{n} \right\} + \inf_{\rho'\in\mathcal{P}(\Theta)} \left\{ \frac{13}{3}\,\mathbb{E}_{\theta'\sim\rho'}\left[\mathcal{H}^2(P^*, P_{\theta'})\right] + \frac{16\,\mathrm{KL}(\rho'\|\pi')}{n} \right\} + \frac{8\log(1/\delta)}{n}. \tag{4.16}$$

Remark 4.9 (Choice of competitor prior)**.** Theorem 4.5 and Corollaries 4.6–4.8 allow the target prior π and the competitor prior π' to differ. In practice, there is seldom reason to choose them differently: a single prior π encoding the available information about Θ serves both roles equally well. Accordingly, all subsequent results are stated under the simplifying assumption $\pi' = \pi$. The proofs are carried out for general $\pi' \neq \pi$ and specialize immediately.

Bounding Λ_λ in terms of Hellinger distance yields a purely geometric oracle inequality.

Theorem 4.10 (Oracle inequality, i.i.d. case)**.** *Fix* $\delta \in (0, 1/2)$ *and* $\lambda = n/8$. *With probability at least* $1 - 2\delta$ *over* $\mathcal{S}$,

$$\frac{1}{12}\,\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}[\mathcal{H}^2(P^*, P_\theta)] \le \inf_{\rho\in\mathcal{P}(\Theta)} \left\{ \frac{26}{3}\,\mathbb{E}_{\theta\sim\rho}\left[\mathcal{H}^2(P^*, P_\theta)\right] + \frac{32\,\mathrm{KL}(\rho\|\pi)}{n} \right\} + \frac{16\log(1/\delta)}{n}. \tag{4.17}$$

The proof is given in Appendix C.

Remark 4.11 (Inverse temperature choice)**.** The condition $\beta_{n,\lambda} < a_1/a_2^2 = \sqrt{2}/16$ ensures positivity of $a_1 - \beta_{n,\lambda} a_2^2$. With $\lambda = n/8$, and $g(\cdot)$ as defined in (4.13), one verifies $\beta_{n,\lambda} = g(1/4)/8 \approx 0.068$, giving $a_1 - \beta_{n,\lambda} a_2^2 \approx 0.086 \ge 1/12$ and $a_0 + \beta_{n,\lambda} a_2^2 \approx 4.29 \le 13/3$. Any λ satisfying $(\lambda/n) g(2\lambda/n) < \sqrt{2}/16$ is admissible.

Under a prior mass condition, the oracle inequality yields explicit convergence rates. The following definition introduces the relevant condition, taken from Alquier (2024, Section 6.1).

Definition 4.12 (Prior mass condition and Catoni dimension)**.** Let $\theta_0 \in \operatorname{argmin}_{\theta\in\Theta} \mathcal{H}^2(P^*, P_\theta)$. We say that the *prior mass condition* is satisfied with constants $c > 0$ and $d_\pi \ge 0$ if there exists $r_0 > 0$ such that, for any $r \le r_0$,

$$\pi\big(\{\theta \in \Theta : \mathcal{H}^2(P^*, P_\theta) \le \mathcal{H}^2(P^*, P_{\theta_0}) + r\}\big) \ge \left(\frac{r}{c}\right)^{d_\pi}. \tag{4.18}$$

The exponent d_π is called the *Catoni dimension* of the model with respect to the prior π. This type of condition is classical in the analysis of posterior contraction rates in Bayesian statistics (Ghosal & van der Vaart, 2017).

The following lemma bounds the PAC-Bayesian trade-off under the prior mass condition.

Lemma 4.13. *Under the prior mass condition (Definition 4.12) with constants c, d_π, and r_0, for any $a > 0$ and $\beta > 0$ satisfying $a\beta \geq d_\pi / r_0$,*

$$\inf_{\rho \in \mathcal{P}(\Theta)} \left\{ a\, \mathbb{E}_{\theta\sim\rho}\big[\mathcal{H}^2(P^*, P_\theta)\big] + \frac{\mathrm{KL}(\rho\|\pi)}{\beta} \right\} \leq a\, \mathcal{H}^2(P^*, P_{\theta_0}) + \frac{d_\pi}{\beta} \log\left(\frac{ec\, a\beta}{d_\pi}\right). \quad (4.19)$$

The proof is given in Appendix C.

Combining Theorem 4.10 with Lemma 4.13 yields the following concentration result.

Corollary 4.14 (Concentration under prior mass condition)**.** *Fix $\delta \in (0, 1/2)$ and $\lambda = n/8$. Suppose the prior mass condition (Definition 4.12) holds with constants c, d_π, and r_0 satisfying $13n \geq 48\, d_\pi / r_0$. Then with probability at least $1 - 2\delta$,*

$$\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}\big[\mathcal{H}^2(P^*, P_\theta)\big] \leq 104\, \mathcal{H}^2(P^*, P_{\theta_0}) + \frac{384\, d_\pi}{n} \log\left(\frac{13ecn}{48 d_\pi}\right) + \frac{192 \log(1/\delta)}{n}. \quad (4.20)$$

In particular, $\mathcal{H}^2(P^*, P_{\theta_0}) + d_\pi \log(n)/n$ is a concentration rate for the $\tilde{\rho}$-posterior.

The proof is given in Appendix C.

4.2.1 Application to Fixed-Design Regression

We illustrate Theorem 4.5 in the context of fixed-design regression, recovering the rates of Baraud & Birgé (2020).

Model setup Following Baraud & Birgé (2020), consider

$$Y_i = f^\star(w_i) + \varepsilon_i, \quad i \in [n], \quad (4.21)$$

where $w_1, \ldots, w_n \in \mathcal{W}$ are fixed design points, ε_i are i.i.d. errors with unknown density p with respect to Lebesgue measure λ, and $f^\star : \mathcal{W} \to \mathbb{R}$ satisfies $\|f^\star\|_\infty \leq B$.

Embedding in the independent framework Setting $X_i = (w_i, Y_i)$ with reference measure $\mu_i = \delta_{w_i} \otimes \lambda$, the true density is

$$p^i_*(w, y) = p(y - f^\star(w)) \cdot \mathbb{1}(w = w_i). \quad (4.22)$$

The parameter space is a function class $\Theta = \mathcal{F}$ with $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}$. Fixing a candidate noise density q, which may differ from the true p, define

$$p^i_f(w, y) = q(y - f(w)) \cdot \mathbb{1}(w = w_i), \qquad f \in \mathcal{F}. \quad (4.23)$$

This defines an independent (non-i.i.d.) model to which Theorem 4.5 applies.

Structural assumption The requirement is that translations of q are controlled in Hellinger distance.

Assumption 4.15 (Order-α candidate density)**.** The candidate density q is unimodal and of order $\alpha \in (-1, 1]$: there exist constants $0 < c_q \le C_q$ such that

$$c_q\big(|\delta|^{1+\alpha} \wedge C_q^{-1}\big) \le \mathcal{H}^2(q_\delta, q) \le C_q\big(|\delta|^{1+\alpha} \wedge C_q^{-1}\big) \quad \forall\, \delta \in \mathbb{R}, \tag{4.24}$$

where $q_\delta(\cdot) = q(\cdot - \delta)$ denotes translation by δ.

This is Definition 26 of Baraud & Birgé (2018); examples include uniform ($\alpha = 0$) and Gaussian ($\alpha = 1$) densities, and more generally any regular translation model (Ibragimov & Has' Minskii, 2013, Chapter VI).

Regression loss Following Baraud & Birgé (2018), define the empirical $(1+\alpha)$-loss

$$d_{1+\alpha}(f, g) := \sum_{i=1}^{n} \big(|f(w_i) - g(w_i)|^{1+\alpha} \wedge C_q^{-1}\big)\,. \tag{4.25}$$

When $\|f\|_\infty, \|g\|_\infty \le B$ and $2B \le C_q^{-1/(1+\alpha)}$, the truncation is inactive, giving $d_{1+\alpha}(f, g) = n\|f - g\|_{n,1+\alpha}^{1+\alpha}$, where

$$\|f - g\|_{n,1+\alpha}^{1+\alpha} := \frac{1}{n}\sum_{i=1}^{n} |f(w_i) - g(w_i)|^{1+\alpha} \tag{4.26}$$

is the empirical $(1+\alpha)$-norm. For comparison with Baraud & Birgé (2020), define also the population $(1+\alpha)$-norm

$$\|f - g\|_{1+\alpha} := \left(\int_{\mathcal{W}} |f(w) - g(w)|^{1+\alpha}\, dP_W(w)\right)^{1/(1+\alpha)}. \tag{4.27}$$

All results below are stated in terms of $\|\cdot\|_{n,1+\alpha}$; under standard design conditions, these translate to bounds on the population norm.

The following lemma connects the sample Hellinger distance to the regression loss.

Lemma 4.16. *Under the regression setup above and Assumption 4.15, assume further that*

$$2B \le C_q^{-1/(1+\alpha)}. \tag{4.28}$$

Then there exist constants $c_\alpha, C_\alpha > 0$ depending only on (c_q, C_q, B, α) such that for all $f \in \mathcal{F}$,

$$c_\alpha \|f - f^\star\|_{n,1+\alpha}^{1+\alpha} - \mathcal{H}^2(p,q) \le \mathcal{H}_n^2(P_*^{(n)}, P_f^{(n)}) \le C_\alpha \|f - f^\star\|_{n,1+\alpha}^{1+\alpha} + 2\mathcal{H}^2(p,q). \quad (4.29)$$

The proof is given in Appendix C. The additive decomposition into regression error and noise misspecification $\mathcal{H}^2(p,q)$ is the source of robustness: misspecification of q contributes a bias term rather than invalidating the procedure.

The main PAC-Bayes bound for regression requires no covering or entropy condition.

Theorem 4.17 (PAC-Bayes bound for fixed-design regression)**.** *Assume the fixed-design regression model* (4.21)*, Assumption 4.15, and the boundedness condition $2B \le C_q^{-1/(1+\alpha)}$. Fix $\delta \in (0, 1/2)$ and a prior $\pi \in \mathcal{P}(\mathcal{F})$. Set $\lambda = n/8$ and let $\hat{\rho}_\lambda$ be the Gibbs posterior defined by* (4.11)*. Then with probability at least $1 - 2\delta$,*

$$\begin{aligned}
&\mathbb{E}_{f\sim\hat{\rho}_\lambda}\left[\|f - f^\star\|_{n,1+\alpha}^{1+\alpha}\right] \\
&\le \frac{12}{c_\alpha}\Bigg[\inf_{\rho\in\mathcal{P}(\mathcal{F})}\left\{\frac{13C_\alpha}{3}\,\mathbb{E}_{f\sim\rho}\left[\|f - f^\star\|_{n,1+\alpha}^{1+\alpha}\right] + \frac{8\,\mathrm{KL}(\rho\|\pi)}{n}\right\} \\
&\quad + \inf_{\rho'\in\mathcal{P}(\mathcal{F})}\left\{\frac{13C_\alpha}{3}\,\mathbb{E}_{f'\sim\rho'}\left[\|f' - f^\star\|_{n,1+\alpha}^{1+\alpha}\right] + \frac{16\,\mathrm{KL}(\rho'\|\pi)}{n}\right\} \\
&\quad + \frac{209}{12}\,\mathcal{H}^2(p,q) + \frac{16\log(1/\delta)}{n}\Bigg], \qquad (4.30)
\end{aligned}$$

where $c_\alpha = c_q/2$ and $C_\alpha = 2C_q$ are as in Lemma 4.16.

The proof is given in Appendix C. The bound holds for arbitrary priors on $\mathcal{F}$, with noise misspecification entering only through the additive term $\mathcal{H}^2(p,q)$, the characteristic feature of ρ-estimation.

To obtain explicit rates, we specialize to priors on finite ε-nets.

Assumption 4.18 (Metric entropy of function class)**.** There exists a non-increasing function $H : (0,\infty) \to \mathbb{R}_+$ such that for every $\varepsilon > 0$, one can find a subset $\mathcal{F}_\varepsilon \subset \mathcal{F}$ with $|\mathcal{F}_\varepsilon| \le \exp(H(\varepsilon))$ satisfying

$$\inf_{g\in\mathcal{F}_\varepsilon} \|f - g\|_\infty \le \varepsilon \quad \text{for all } f \in \mathcal{F}.$$

Corollary 4.19 (Entropy bound via ε-net prior)**.** *Assume the setting of Theorem 4.17 and additionally Assumption 4.18. Fix $\varepsilon > 0$ and let π be the uniform distribution on an ε-net $\mathcal{F}_\varepsilon$ with*

$|\mathcal{F}_\varepsilon| \le e^{H(\varepsilon)}$. *Then with probability at least* $1-2\delta$,

$$\mathbb{E}_{f\sim\hat{\rho}_\lambda}\left[\|f-f^\star\|_{n,1+\alpha}^{1+\alpha}\right] \le K_\alpha\left[\mathcal{H}^2(p,q)+\inf_{f\in\mathcal{F}}\|f-f^\star\|_\infty^{1+\alpha}+\varepsilon^{1+\alpha}+\frac{H(\varepsilon)}{n}+\frac{\log(1/\delta)}{n}\right], \quad (4.31)$$

where $K_\alpha > 0$ *depends only on* (c_q, C_q, B, α).

The three error sources, noise misspecification $\mathcal{H}^2(p,q)$, model misspecification $\inf_{f\in\mathcal{F}}\|f-f^\star\|_\infty^{1+\alpha}$, and statistical complexity $\varepsilon^{1+\alpha}+H(\varepsilon)/n$, contribute additively.

Optimizing ε yields the minimax rate.

Corollary 4.20 (Optimal rate under polynomial entropy)**.** *Assume the setting of Corollary 4.19 with* $\alpha \ge 0$*, and suppose* $H(\varepsilon) \le M\varepsilon^{-d}$ *for constants* $M, d > 0$*. Choose*

$$\varepsilon_n = \left(\frac{M}{n}\right)^{1/(d+1+\alpha)}. \quad (4.32)$$

Then with probability at least $1-2\delta$,

$$\mathbb{E}_{f\sim\hat{\rho}_\lambda}\left[\|f-f^\star\|_{n,1+\alpha}\right] \le K'_\alpha\left[\mathcal{H}^2(p,q)+\inf_{f\in\mathcal{F}}\|f-f^\star\|_\infty^{1+\alpha}+n^{-\frac{1+\alpha}{d+1+\alpha}}+\frac{\log(1/\delta)}{n}\right]^{1/(1+\alpha)}, \quad (4.33)$$

where $K'_\alpha > 0$ *depends on* $(c_q, C_q, B, \alpha, M, d)$.

Remark 4.21. The ε-net prior is illustrative; Theorem 4.17 accommodates sparsity-inducing priors Dalalyan & Tsybakov (2012); Alquier & Lounici (2011), Gaussian processes Reeb et al. (2018), and priors on deep neural networks Chérief-Abdellatif (2020), among others.

The rate $n^{-(1+\alpha)/(d+1+\alpha)}$ is minimax-optimal for nonparametric regression with metric entropy of order ε^{-d}; setting $d = V/(1+\alpha)$ recovers the classical rate $n^{-(1+\alpha)/(V+1+\alpha)}$. The proofs of Corollaries 4.19 and 4.20 are given in Appendix C.

4.2.2 Variational Approximation

The oracle inequalities above optimize over all of $\mathcal{P}(\Theta)$, yielding Gibbs posteriors with intractable partition functions. We now restrict both ρ and ρ' to tractable variational families $\mathcal{F}, \mathcal{F}' \subset \mathcal{P}(\Theta)$.

By the Donsker–Varadhan formula, the softmax competitor (4.10) admits the representa-

tion

$$\Lambda_\lambda(\theta;\pi) = \frac{1}{\lambda}\log\left(\int_\Theta e^{\lambda\,\hat{R}_\psi(\theta,\theta')}\pi(d\theta')\right) = \sup_{\rho'\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta'\sim\rho'}[\hat{R}_\psi(\theta,\theta')] - \frac{1}{\lambda}\mathrm{KL}(\rho'\|\pi)\right\}.$$

Restricting to $\mathcal{F}' \subset \mathcal{P}(\Theta)$ yields a computable lower bound.

Definition 4.22 (Variational softmax)**.** Let $\mathcal{F}' \subset \mathcal{P}(\Theta)$ be a variational family. For each $\theta \in \Theta$, define the *variational softmax*

$$\Lambda_\lambda^{\mathcal{F}'}(\theta;\pi) := \sup_{\rho'\in\mathcal{F}'}\left\{\mathbb{E}_{\theta'\sim\rho'}[\hat{R}_\psi(\theta,\theta')] - \frac{1}{\lambda}\,\mathrm{KL}(\rho'\|\pi)\right\}.$$

By definition, $\Lambda_\lambda^{\mathcal{F}'}(\theta;\pi) \le \Lambda_\lambda(\theta;\pi)$ for all $\theta \in \Theta$.

Similarly, restricting the target posterior to $\mathcal{F} \subset \mathcal{P}(\Theta)$ gives the *variational target posterior*

$$\tilde{\rho}_\lambda \in \operatorname*{argmin}_{\rho\in\mathcal{F}}\left\{\mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda^{\mathcal{F}'}(\theta;\pi)] + \frac{1}{\lambda}\mathrm{KL}(\rho\|\pi)\right\}. \tag{4.34}$$

Theorem 4.23 (Variational oracle inequality)**.** *Fix $\delta \in (0,1/2)$, $\lambda = n/8$, and $\mathcal{F} = \mathcal{F}'$. Let $\tilde{\rho}_\lambda$ be the variational target posterior defined in* (4.34)*. With probability at least $1-2\delta$,*

$$\frac{1}{12}\,\mathbb{E}_{\theta\sim\tilde{\rho}_\lambda}[\mathcal{H}^2(P^*,P_\theta)] \le \inf_{\rho\in\mathcal{F}}\left\{\frac{26}{3}\,\mathbb{E}_{\theta\sim\rho}[\mathcal{H}^2(P^*,P_\theta)] + \frac{32}{n}\mathrm{KL}(\rho\|\pi)\right\} + \frac{16\log(1/\delta)}{n}. \tag{4.35}$$

The proof is given in Appendix C.

As in Corollary 4.14, the oracle inequality simplifies under a prior mass condition adapted to the variational family, taken from Alquier (2024, Section 6.1).

Corollary 4.24 (Variational concentration under prior mass condition)**.** *Fix $\delta \in (0,1/2)$ and $\lambda = n/8$. Assume that there exists $\rho_n \in \mathcal{F}$ satisfying*

$$\mathbb{E}_{\theta\sim\rho_n}\left[\mathcal{H}^2(P^*,P_\theta)\right] \le \frac{1}{n} \quad \textit{and} \quad \mathrm{KL}(\rho_n\|\pi) \le d_\pi \log n, \tag{4.36}$$

where d_π is the Catoni dimension (Definition 4.12). Then with probability at least $1-2\delta$,

$$\mathbb{E}_{\theta\sim\tilde{\rho}_\lambda}\left[\mathcal{H}^2(P^*,P_\theta)\right] \le \frac{384\,d_\pi\log n + 104}{n} + \frac{192\log(1/\delta)}{n}. \tag{4.37}$$

In particular, the variational $\tilde{\rho}$-posterior concentrates at the rate $\mathcal{O}(d_\pi\log(n)/n)$.

The proof is given in Appendix C.

Remark 4.25 (Comparison with the unrestricted case)**.** When $\mathcal{F} = \mathcal{P}(\Theta)$, condition (4.36) is implied by the prior mass condition (Definition 4.12) by choosing $\rho_n = \pi(\cdot \mid B(1/n))$, which gives $\mathrm{KL}(\rho_n \| \pi) = -\log \pi(B(1/n)) \leq d_\pi \log(cn)$. For strict variational subfamilies $\mathcal{F} \subsetneq \mathcal{P}(\Theta)$, condition (4.36) is a genuine requirement on the approximation capacity of $\mathcal{F}$: it asks that $\mathcal{F}$ contains a distribution that simultaneously concentrates near P^* in Hellinger distance and remains close to the prior in KL divergence. The rate $\mathcal{O}(d_\pi \log(n)/n)$ is suboptimal by a logarithmic factor compared to the minimax rate $\mathcal{O}(d_\pi/n)$.

4.2.3 Computational Analysis

Computing the variational posterior (4.34) requires solving an optimization problem that we now reformulate as a saddle-point problem admitting efficient first-order methods. We develop this for canonical exponential families with Gaussian variational approximations, establishing convergence guarantees uniform in λ.

Variational objectives

Let $\Phi \subset \mathbb{R}^{d_\phi}$ and $\mathcal{N} \subset \mathbb{R}^{d_\nu}$ parameterize $\mathcal{F}$ and $\mathcal{F}'$ surjectively, writing $\rho_\phi \in \mathcal{F}$ and $\rho'_\nu \in \mathcal{F}'$. Two natural objectives arise depending on the order of optimization over the competitor ρ' and integration over $\theta \sim \rho_\phi$.

Definition 4.26 (Pointwise and joint variational objectives)**.**

1. The *pointwise variational objective* $\mathcal{J} : \Phi \to \mathbb{R}$ is

$$\mathcal{J}(\phi) := \mathbb{E}_{\theta \sim \rho_\phi}\left[\Lambda_\lambda^{\mathcal{F}'}(\theta; \pi)\right] + \frac{1}{\lambda}\mathrm{KL}(\rho_\phi \| \pi), \tag{4.38}$$

   where the supremum over $\rho' \in \mathcal{F}'$ implicit in $\Lambda_\lambda^{\mathcal{F}'}$ is taken separately for each θ.

2. The *joint variational objective* $\tilde{\mathcal{J}} : \Phi \to \mathbb{R}$ is

$$\tilde{\mathcal{J}}(\phi) := \sup_{\rho' \in \mathcal{F}'} \left\{ \mathbb{E}_{(\theta,\theta') \sim \rho_\phi \otimes \rho'}\left[\hat{R}_\psi(\theta, \theta')\right] - \frac{1}{\lambda}\mathrm{KL}(\rho' \| \pi) \right\} + \frac{1}{\lambda}\mathrm{KL}(\rho_\phi \| \pi), \tag{4.39}$$

   where a single ρ' is chosen to be optimal on average over $\theta \sim \rho_\phi$.

Jensen's inequality gives $\tilde{\mathcal{J}}(\phi) \leq \mathcal{J}(\phi)$, with equality when the pointwise maximizer $\rho'^\star(\theta)$ does not depend on θ. The gap $\Delta(\phi) := \mathcal{J}(\phi) - \tilde{\mathcal{J}}(\phi) \geq 0$ is bounded in Proposition 4.32.

Saddle-point reformulation

Definition 4.27 (Primal–dual objective)**.** Define $\mathcal{L}_n : \Phi \times \mathcal{N} \to \mathbb{R}$ by

$$\mathcal{L}_n(\phi, \nu) := \mathbb{E}_{\theta \sim \rho_\phi} \mathbb{E}_{\theta' \sim \rho'_\nu} \left[\hat{R}_\psi(\theta, \theta') \right] + \frac{1}{\lambda} \mathrm{KL}(\rho_\phi \| \pi) - \frac{1}{\lambda} \mathrm{KL}(\rho'_\nu \| \pi). \tag{4.40}$$

Theorem 4.28 (Saddle-point equivalence)**.** *Under the surjectivity assumption, the following hold.*

(i) *For every $\phi \in \Phi$, the supremum over ν is attained and*

$$\sup_{\nu \in \mathcal{N}} \mathcal{L}_n(\phi, \nu) = \tilde{\mathcal{J}}(\phi). \tag{4.41}$$

(ii) *For every $\phi \in \Phi$,*

$$\tilde{\mathcal{J}}(\phi) \le \mathcal{J}(\phi) \le \tilde{\mathcal{J}}(\phi) + \Delta(\phi), \tag{4.42}$$

where the gap satisfies

$$\Delta(\phi) \le \frac{\bar{G}^2}{16} \operatorname{tr}(\Sigma_\phi), \tag{4.43}$$

with Σ_ϕ the covariance of ρ_ϕ and $\bar{G}$ the score bound from Assumption 4.31(3).

(iii) *The infimum satisfies*

$$\inf_{\phi \in \Phi} \sup_{\nu \in \mathcal{N}} \mathcal{L}_n(\phi, \nu) = \inf_{\phi \in \Phi} \tilde{\mathcal{J}}(\phi) \le \inf_{\phi \in \Phi} \mathcal{J}(\phi). \tag{4.44}$$

Moreover, $\phi^\star$ is a stationary point of $\tilde{\mathcal{J}}$ if and only if there exists $\nu^\star \in \mathcal{N}$ such that $(\phi^\star, \nu^\star)$ is a first-order stationary point of $\mathcal{L}_n$.

(iv) *In the PAC-Bayes regime $\lambda = \Theta(n)$, if $\operatorname{tr}(\Sigma_{\phi^\star}) = \mathcal{O}(d/n)$, then $\Delta(\phi^\star) = \mathcal{O}(d\bar{G}^2/n) \to 0$, so $\tilde{\mathcal{J}}(\phi^\star) = \mathcal{J}(\phi^\star)$ asymptotically.*

The proof is given in Appendix C.

Convergence guarantees for exponential families

Definition 4.29 (Canonical exponential family)**.** Let $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ be a canonical exponential family on $(\mathcal{X}, \mathcal{A})$:

$$p_\theta(x) = h(x) \exp\big\{ \langle \theta, T(x) \rangle - A(\theta) \big\},$$

where $T : \mathcal{X} \to \mathbb{R}^d$ is the sufficient statistic, $A : \Theta \to \mathbb{R}$ the log-partition function, and $\mu(\theta) := \nabla A(\theta)$, $I(\theta) := \nabla^2 A(\theta)$ the mean map and Fisher information matrix.

Definition 4.30 (Variational families and priors)**.** We consider mean-field Gaussian variational families:

- *Target posterior:* $\rho_\phi = \mathcal{N}(m, \Sigma), \phi = (m, \Sigma) \in \mathbb{R}^d \times \mathbb{S}^d_{++}$.
- *Competitor posterior:* $\rho'_\nu = \mathcal{N}(m', \mathrm{diag}(\sigma'^2)), \nu = (m', s) \in \mathbb{R}^d \times \mathbb{R}^d, \sigma'^2_i = e^{s_i}$.
- *Prior:* $\pi = \mathcal{N}(m_\pi, \Sigma_\pi)$.

Assumption 4.31 (Regularity conditions)**.** The following hold uniformly on a compact region $\bar{\Theta} \subset \Theta$ containing the supports of ρ_ϕ and ρ'_ν:

1. *Fisher lower bound:* $I(\theta) \succeq \sigma_0^2 \mathbb{I}_d$ for some $\sigma_0 > 0$.
2. *Bounded second moment:* $\mathbb{E}_{X \sim P^*}\left[\|T(X) - \mu(\theta')\|^2\right] \leq B^2 < \infty$ uniformly in $\theta' \in \bar{\Theta}$.
3. *Bounded scores:* $\bar{G} := \sup_{\theta \in \bar{\Theta}} \sup_{x \in \mathcal{X}} \|\nabla_\theta \log p_\theta(x)\| < \infty$.
4. *Bounded log-likelihood ratios:* there exists $M < \infty$ such that $|\log p_{\theta'}(x) - \log p_\theta(x)| \leq M$ for all $\theta, \theta' \in \bar{\Theta}$ and P^*-a.e. x. This is implied by (3) together with compactness of $\bar{\Theta}$.
5. *Curvature margin:* $\mu_0 := c_M \sigma_0^2 - B^2/4 > 0$, where $c_M := \min_{t \in [e^{-M/2}, e^{M/2}]} t/(t+1)^2 > 0$.

Here c_M is the minimum of $\varphi'(u)$ on $\{|u| \leq M\}$ with $\varphi(u) := \psi(e^u)$. The curvature margin $\mu_0 > 0$ requires that the weighted Fisher information dominates the variance of the sufficient statistic.

Proposition 4.32 (Gap bound)**.** *Under Assumption 4.31(3), the gap between the pointwise and joint variational objectives satisfies, for every* $\phi \in \Phi$*,*

$$0 \leq \Delta(\phi) := \mathcal{J}(\phi) - \tilde{\mathcal{J}}(\phi) \leq \frac{\bar{G}^2}{16} \mathrm{tr}(\Sigma_\phi), \tag{4.45}$$

where Σ_ϕ *is the covariance matrix of* ρ_ϕ *and* $\bar{G}$ *is the score bound from Assumption 4.31(3). In particular, in the PAC-Bayes regime* $\lambda = \Theta(n)$*, if* $\mathrm{tr}(\Sigma_{\phi^\star}) = \mathcal{O}(d/n)$ *then* $\Delta(\phi^\star) = \mathcal{O}(d\bar{G}^2/n) \to 0$*, so the two objectives coincide asymptotically.*

The proof is given in Appendix C.

Theorem 4.33 (NC-SC geometry)**.** *Under Assumption 4.31:*

1. $\mathcal{L}_n(\phi, \nu)$ *is* L*-smooth in* (ϕ, ν) *with* $L = L_\psi + L_{\mathrm{KL}}/\lambda$*, where* L_ψ *and* L_{KL} *are independent of* λ*.*

2. *For any fixed ϕ and any data realization $(X_1, \dots, X_n)$ in the support of P^{*n}, the map $\nu \mapsto \mathcal{L}_n(\phi, \nu)$ is μ-strongly concave with $\mu \geq \mu_0 > 0$ independent of both λ and the data.*

The NC-SC condition number $\kappa := L/\mu = \mathcal{O}(1)$ uniformly in $\lambda > 0$.

The proof is given in Appendix C.

Corollary 4.34 (Optimization guarantee)**.** *Assume that $\mathcal{L}_n$ satisfies the NC–SC geometry of Theorem 4.33. Let $(x_t)_{t\geq 0}$ be the iterates of a projected stochastic extragradient method applied to*

$$\min_{\phi\in\Phi} \max_{\nu\in\mathcal{N}} \mathcal{L}_n(\phi, \nu),$$

with stepsize $\eta \asymp 1/L$ and unbiased stochastic gradients with variance bounded by σ^2. Then, for any $T \geq 1$,

$$\min_{0\leq t<T} \mathbb{E}\big[\|\mathcal{G}_\eta(x_t)\|^2\big] = \mathcal{O}\left(\frac{\kappa\,\Delta}{T} + \frac{\kappa\,\sigma^2}{\sqrt{T}}\right),$$

where $\mathcal{G}_\eta$ denotes the gradient mapping, $\Delta := \mathcal{L}_n(x_0) - \inf_\phi \sup_\nu \mathcal{L}_n(\phi, \nu)$ is the initial optimality gap, and $\kappa := L/\mu$ is the condition number.

Since $\kappa = \mathcal{O}(1)$ uniformly in λ (Theorem 4.33), this guarantee holds for any $\lambda > 0$, including $\lambda = \Theta(n)$; see Juditsky et al. (2011); Lin et al. (2020).

4.3 Numerical Experiments

We assess the variational $\tilde{\rho}$-posterior on one-dimensional exponential families under ε-contamination. In each case the model is correctly specified, but the data-generating distribution is a contaminated mixture.

Gaussian location model Consider

$$P_\theta = \mathcal{N}(\theta, 1), \qquad \theta \in \mathbb{R},$$

with true parameter $\theta^\star = 0$ and contaminated distribution

$$P_\varepsilon^\star = (1-\varepsilon)\,\mathcal{N}(0,1) + \varepsilon\,\mathcal{N}(8,1).$$

A Gaussian prior $\pi = \mathcal{N}(0, 4)$ is placed on θ, and the $\tilde{\rho}$-posterior is approximated by a Gaussian variational family $q_\phi(\theta) = \mathcal{N}(m, s^2)$, optimized via the saddle-point objective $\mathcal{L}_n(\phi, \nu)$ with inverse temperature $\lambda = \tau n$ ($\tau = 0.5$), using Adam with 200 iterations and the reparameterization trick. The $\tilde{\rho}$-estimator is the variational posterior mean $\hat{\theta}_\rho = m$. We compare with the MLE $\hat{\theta}_{\mathrm{MLE}} = \bar{X}$ and the conjugate Bayes posterior mean $\hat{\theta}_B$.

Poisson intensity model Consider

$$X_i \mid \mu \sim \text{Pois}(\mu), \qquad \mu > 0,$$

with true intensity $\mu^\star = 3$ and contaminated distribution

$$P_\varepsilon^\star = (1-\varepsilon)\,\text{Pois}(3) + \varepsilon\,\text{Pois}(30).$$

The Bayesian estimator uses a Gamma$(1,1)$ prior with conjugate posterior mean $\hat{\mu}_B$. For the $\tilde{\rho}$-posterior, we reparameterize $\eta = \log\mu$, place a Gaussian prior on η, and optimize a Gaussian variational family $q_\psi(\eta) = \mathcal{N}(m_\rho, s_\rho^2)$ as above. The $\tilde{\rho}$-estimator is

$$\hat{\mu}_\rho = \mathbb{E}_{q_\psi}[e^\eta] = \exp\left(m_\rho + \tfrac{1}{2}s_\rho^2\right).$$

We compare with $\hat{\mu}_{\text{MLE}} = \bar{X}$ and $\hat{\mu}_B$.

Uniform scale model Consider

$$X_i \mid \theta \sim \text{Uniform}(0,\theta), \qquad \theta > 0,$$

with true parameter $\theta^\star = 1$ and contaminated distribution

$$P_\varepsilon^\star = (1-\varepsilon)\,\text{Uniform}(0,1) + \varepsilon\,\text{Uniform}(101,102),$$

where contaminated observations lie far outside the support of the clean distribution. The MLE $\hat{\theta}_{\text{MLE}} = \max_i X_i$ is highly sensitive to outliers. The Bayesian estimator uses the prior $\pi(\theta) \propto \theta^{-\alpha}\mathbb{1}_{\theta \geq a}$ with $\alpha = 2$ and $a = 0.5$, giving Pareto posterior mean

$$\hat{\theta}_B = \frac{n+\alpha-1}{n+\alpha-2} \cdot (a \vee X_{(n)}), \qquad X_{(n)} = \max_i X_i.$$

For the $\tilde{\rho}$-posterior, we reparameterize $u = \log\theta$ and use a Gaussian variational approximation $q_\psi(u) = \mathcal{N}(m_\rho, s_\rho^2)$, giving $\tilde{\rho}$-estimator

$$\hat{\theta}_\rho = \exp\left(m_\rho + \tfrac{1}{2}s_\rho^2\right).$$

Experimental setup For each configuration (n, τ, ε), $T = 1000$ independent datasets are generated from $P_\varepsilon^\star$ with $n = 200$, $\tau = 0.5$, and $\varepsilon \in \{0, 0.05, 0.08, 0.10\}$. Performance is measured by the empirical risk

$$\widehat{R}(\delta;\theta^\star) = \frac{1}{T}\sum_{k=1}^{T}\left(\delta(X^{(k)}) - \theta^\star\right)^2.$$

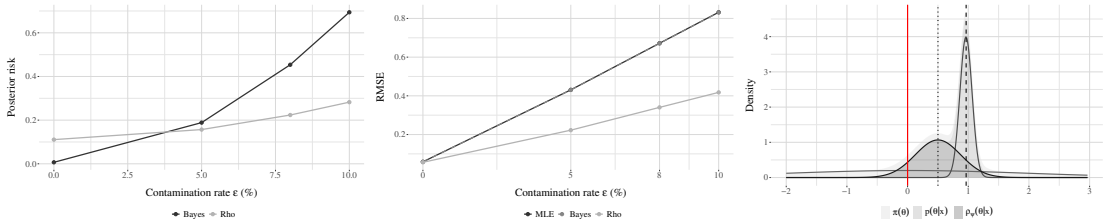


Figure 4.1: Gaussian location model results for $n = 200$ and $\tau = 0.5$. *Left:* Posterior risk vs. contamination rate ε. *Middle:* RMSE vs. contamination rate. *Right:* Posterior densities for a single dataset at $\varepsilon = 10\%$.

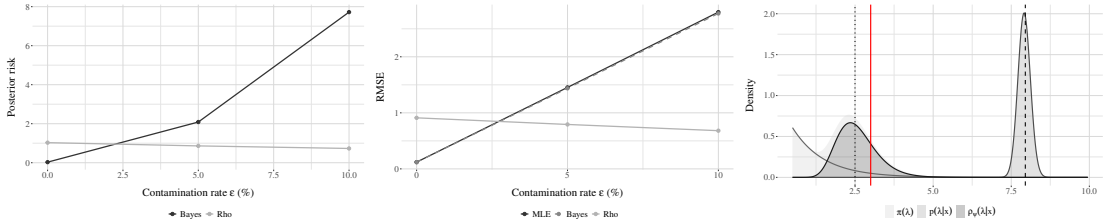


Figure 4.2: Poisson intensity model results for $n = 200$ and $\tau = 0.5$. *Left:* Posterior risk vs. contamination rate ε. *Middle:* RMSE vs. contamination rate. *Right:* Posterior densities for a single dataset at $\varepsilon = 10\%$.

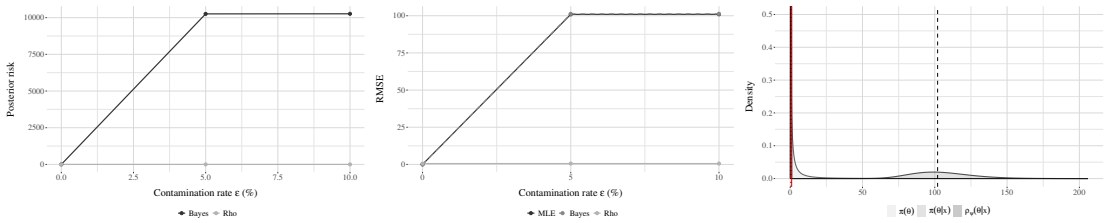


Figure 4.3: Uniform scale model results for $n = 200$ and $\tau = 0.5$. *Left:* Posterior risk vs. contamination rate ε. *Middle:* RMSE vs. contamination rate. *Right:* Posterior densities for a single dataset at $\varepsilon = 10\%$.

Results Figures 4.1–4.3 report posterior risk (left), RMSE (middle), and posterior densities at $\varepsilon = 10\%$ (right) for each model. Across all three settings, the $\tilde{\rho}$-posterior maintains stability under contamination while the MLE and Bayes estimators deteriorate rapidly. The effect is most pronounced in the uniform model, where a single outlier can inflate the MLE to ≈ 102; the $\tilde{\rho}$-posterior remains concentrated near $\theta^\star = 1$.

We next evaluate the $\tilde{\rho}$-posterior on three increasingly complex regression settings, each designed to test different aspects of robustness under heavy-tailed contamination.

Fourier basis regression with smooth target function We first consider a regression problem with known smooth structure. The model takes the form $Y_i = f^*(w_i) + \xi_i$ for $i = 1, \ldots, n$, where w_i are equally spaced points in $[0, 1]$ and the true regression function is

$$f^*(w) = \sin(2\pi w) + 0.3\cos(6\pi w).$$

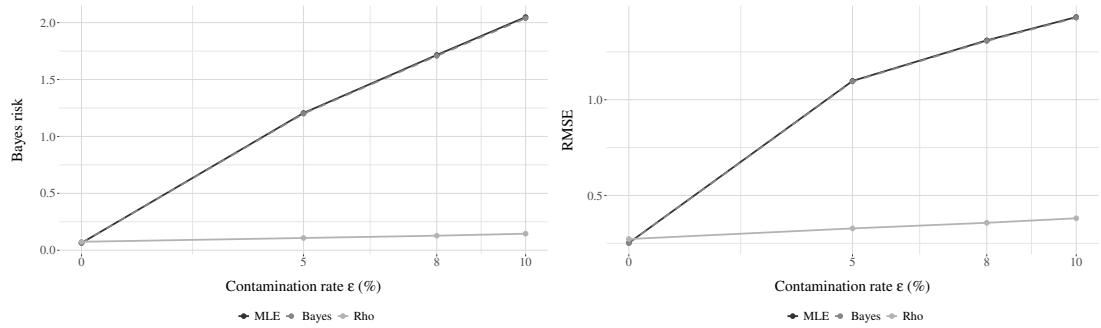


Figure 4.4: Fourier basis regression ($n = 200$, $K = 6$). *Left:* Posterior risk. *Right:* RMSE. The $\tilde{\rho}$-posterior maintains stability under contamination.

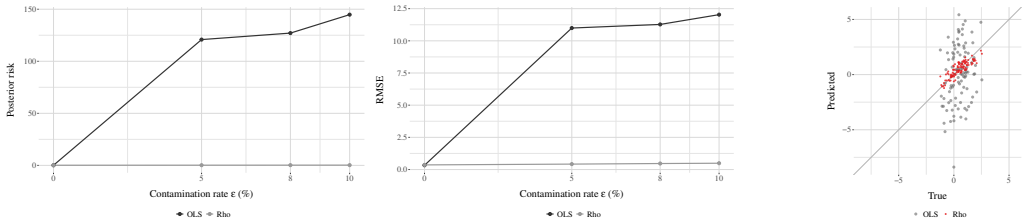


Figure 4.5: Correlated design regression ($n = 100$, $p = 11$). *Left:* Posterior risk. *Middle:* RMSE. *Right:* Predicted vs. true values at $\varepsilon = 10\%$.

The design matrix $\Phi \in \mathbb{R}^{n \times p}$ consists of Fourier basis functions with $K = 6$ frequency components, yielding $p = 13$ features including the intercept. The noise follows an ε-contaminated mixture

$$\xi_i \sim (1 - \varepsilon)\mathcal{N}(0, 1) + \varepsilon \cdot \text{Pareto}_{\text{two-sided}}(6, 2),$$

where the two-sided Pareto generates symmetric outliers with minimum magnitude 6. We set $n = 200$ and vary $\varepsilon \in \{0, 0.05, 0.08, 0.10\}$.

For this setting, we compared three estimators: MLE (ordinary least squares), standard Bayes posterior mean with conjugate Gaussian prior $\beta \sim \mathcal{N}(0, 4I_p)$, and the $\tilde{\rho}$-posterior approximated via variational inference with temperature $\lambda = 0.5n$.

Figure 4.4 reports results averaged over 1000 replications. Both MLE and Bayes posterior risk grow from near zero to over 2 at $\varepsilon = 10\%$; the $\tilde{\rho}$-posterior risk increases only from 0.07 to 0.14. RMSE exhibits the same pattern.

Correlated design with sparse parameters The second setting uses $n = 100$ observations, $d = 10$ correlated features with Toeplitz covariance $\Sigma_{jk} = 0.7^{|j-k|}$, and a sparse true parameter β^* with five nonzero entries. Contamination uses two-sided Pareto noise with minimum magnitude 10 and shape 1.5. Figure 4.5 shows OLS risk growing to over 150 at $\varepsilon = 10\%$, while the $\tilde{\rho}$-posterior remains below 4.

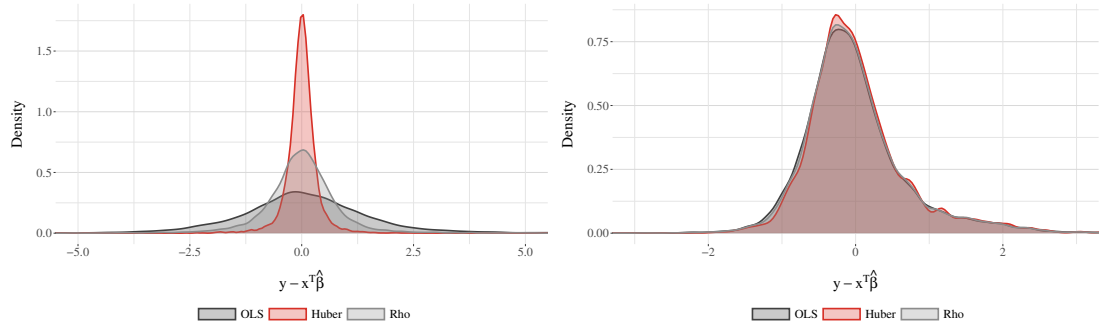


Figure 4.6: Test residual distributions at $\varepsilon = 10\%$ contamination. *Left:* Ames Housing. *Right:* Abalone.

Real-world datasets We evaluate on the Ames Housing dataset (De Cock, 2011) ($n = 2{,}930$, 79 features) and the Abalone dataset (Nash et al., 1994) ($n = 4{,}177$, 8 features). Training labels are contaminated by adding $\pm 15 \times \mathrm{MAD}(Y)$ to a random ε-fraction; test data remains clean. We compare OLS, Huber regression (Hammouda et al., 2024), and the variational $\tilde{\rho}$-posterior. Figure 4.6 shows test residual distributions at $\varepsilon = 10\%$ over 1000 trials: the $\tilde{\rho}$-posterior achieves concentration competitive with Huber regression.

These experiments confirm that the robustness established in Theorems 4.5–4.23 extends to practical settings with varying dimensionality and contamination patterns.

4.4 Conclusion

We introduced the $\tilde{\rho}$-posterior, a softmax relaxation of the ρ-posterior of Baraud & Birgé (2020): replacing the intractable supremum contrast by a softmax aggregation over competitor parameters yields a Gibbs posterior amenable to the PAC-Bayesian machinery, while preserving the competitor-based risk decomposition central to Baraud et al. (2017); Baraud & Birgé (2020). Finite-sample oracle inequalities (Theorems 4.5 and 4.10) give concentration rates under Hellinger risk; these are inherited by variational approximations (Theorem 4.23), whose saddle-point optimization is solved efficiently by first-order methods (Theorem 4.33).

Several directions remain open. The softmax principle applies to any bounded contrast ψ; the present choice is motivated by its link to the Hellinger distance and the original ρ-estimator, and by the favorable landscape it induces, a nonconvex–strongly concave saddle-point problem on which standard first-order methods converge. Whether other bounded contrasts, tied to different divergences between probability measures, can simultaneously deliver strong robustness to contamination, sharp concentration, and a tractable optimization geometry is left to future work.

Chapter 5

Conclusion and Perspectives

Each contribution of this thesis is a certificate: a finite-sample guarantee on the error of an approximation, of $\pi_n(h)$ by importance sampling and Metropolis–Hastings in Chapter 2, of the risk of a data-dependent posterior in Chapter 3, and of a misspecified model by a robust surrogate in Chapter 4. Each such guarantee isolates the object that governs the error and points to the same next step, optimising it. On the Monte Carlo front, the bias and variance of self-normalised importance sampling and of independent Metropolis–Hastings, and the trade-off between them, are what the analysis must pin down; doing so is what opens the way to optimising the proposal q and to evaluating alternative schemes. The prior of the mutual-information bound (Chapter 3) and the loss-based contrast of the ρ-posterior (Chapter 4) are the analogous levers on the optimisation-based side. We sketch these directions below.

5.1 Sampling under global proposals

For the global Monte Carlo schemes of Chapter 2, self-normalised importance sampling and the independent Metropolis–Hastings sampler, the proposal q is the quantity that fixes both the discrepancy with π_n and the behaviour of the weight $\omega = d\pi_n/dq$, whose moments $q(\omega^p)$ control every bound of the chapter. Recent methodological advances, and the computing power now available, have made q both more accurate and readily usable within Monte Carlo: normalising flows (Papamakarios et al., 2021) and the integrated nested Laplace approximation (Rue et al., 2009), the latter now run inside Markov chain Monte Carlo (Gómez-Rubio & Rue, 2018) and importance sampling (Berild et al., 2022), with their finite-sample accuracy now quantified (Kasprzak et al., 2025; Katsevich & Rigollet, 2024).

With the proposal fixed, the finite-sample performance is decided by the scheme that uses it. In the unbounded-weight regime of this thesis, where q admits only finitely many weight

moments, the behaviour of a scheme is governed by one index, the largest order

$$p^\star(q) = \sup\{\, p \geq 1 : q(\omega^p) < \infty \,\}, \tag{5.1}$$

at which a moment of ω is finite, the moment index of the unbounded-weight regime of Chapter 2. That chapter established the role of $p^\star(q)$ for two schemes. Self-normalised importance sampling carries a self-normalisation bias of order N^{-1} (Theorem 2.8), while coupling two independent Metropolis–Hastings chains by common random numbers turns a weight moment into a meeting-time moment and gives a total-variation bias that, under $q(\omega^p) < \infty$, decays as $t^{-(p-1)}$ for every $p < p^\star(q)$ (Theorem 2.16), so IMH carries the smaller bias once $p^\star(q) > 2$.

The same tail behaviour of the weight, encoded in its moments $q(\omega^p)$, should govern the wider family of global-proposal MCMC samplers. The multiple-try Metropolis independent sampler (Liu et al., 2000; Yang & Liu, 2023) and the iterated SIR sampler (Andrieu et al., 2010, 2018) are, under a bounded weight $\omega^\star := \operatorname{ess\,sup} \omega < \infty$, uniformly ergodic at a rate that sharpens with the number N of proposals. From any starting point x, the iterated SIR kernel obeys

$$\left|\delta_x \left(P_N^{\text{i-SIR}}\right)^t - \pi_n\right|_{\text{TV}} \leq \left(1 - \frac{N-1}{2\omega^\star + N - 2}\right)^t, \tag{5.2}$$

a geometric rate that contracts to 0 as the number of proposals N grows; the exact rate of MTM-IS under bounded weights has likewise been characterised (Yang & Liu, 2023),

$$\begin{aligned} \left|\delta_x \left(P_N^{\text{MTM-IS}}\right)^t - \pi_n\right|_{\text{TV}} &\leq \left(1 - H_N(\omega^\star)\right)^t, \\ H_N(\omega^\star) &= \mathbb{E}_q\left[\frac{N}{\omega^\star + \sum_{i=1}^{N-1} \omega(X_i)}\right], \end{aligned} \tag{5.3}$$

the multiple-try rate reducing to the independent-Metropolis rate $(1 - 1/\omega^\star)^t$ (Mengersen & Tweedie, 1996; Wang, 2022) at a single try. Under unbounded weights, geometric ergodicity fails (Mengersen & Tweedie, 1996), and the bias of both samplers is unknown. Bounds there, explicit in t and N under a moment assumption $q(\omega^p) < \infty$, would tell a practitioner which scheme best reduces the bias at a given budget and application. These samplers are also elementary building blocks of larger constructions, among them particle Gibbs and sequential Monte Carlo. Coupling their kernels maximally, as Chapter 2 does for IMH, would debias the samplers and the schemes assembled from them, making both composable inside iterated procedures such as nested expectations and IWAE gradient estimation.

The order $p^\star(q)$ that controls these unbounded-weight rates is a property of π_n through the weight ω, and for a learned variational or normalising-flow proposal q it is not known in advance. Knowing it fixes the order of the bias, and hence the budget-optimal allocation between the inner and outer stages of a nested expectation, following Andradóttir & Glynn (2016), as recalled in Section 1.3.

Estimating $p^\star(q)$ from the realised weights is therefore essential. Chapter 2 already flags this as a hard tail-estimation problem; extreme-value theory is its natural apparatus, the tail-index estimator of Hill (1975) or a generalised-Pareto fit to the largest weights (Pickands, 1975). As a diagnostic for importance sampling, an estimate of $p^\star(q)$ would certify finite variance ($p^\star(q) \geq 2$) and predict the bias rate. It would also flag when the data justify the added cost of debiasing ($p^\star(q) > 2$).

On the variance side the two schemes are already ordered, by the asymptotic-variance comparison of Deligiannidis & Lee (2018, Proposition 2): reweighting has the smaller variance and accept–reject the smaller bias. Whether the same tools rank the asymptotic variances of MTM-IS and i-SIR remains open; settling it would order the global proposals along both axes.

One finding of Chapter 2 is that an exactly unbiased estimator makes robust mean estimation available: its R i.i.d. replicates feed a median-of-means, or any estimator built for unbiased inputs, sub-Gaussian under the single finite variance of one replicate where a biased estimator would leave a floor, and the same composition extends to any other unbiased-MCMC scheme (Atchadé & Jacob, 2026).

Finally, the analysis assumed i.i.d. proposal draws $X_1, \dots, X_N \sim q$. Using importance sampling within MCMC is an old idea, raised already by Hastings (1970, Section 5): one runs a q-invariant chain (X_t) and reweights its trajectory,

$$\widehat{\pi}_N(f) = \frac{\sum_{t=1}^N f(X_t)\,\omega(X_t)}{\sum_{t=1}^N \omega(X_t)}, \tag{5.4}$$

the self-normalised estimator of Section 1.2 run along a chain, as in the importance Markov chain of Andral et al. (2024). Carrying the guarantees of Chapter 2 into this setting would be valuable twice over: a well-mixing chain can reduce variance below i.i.d. sampling, and the scheme reaches targets that admit no direct sampler from q. What must transfer is the $N^{-s/2}$ moment bound of Theorem 2.9, and the bias analysis behind it, once the summands become dependent.

In the i.i.d. case the bound rests on the Rosenthal inequality

$$\mathbb{E}\Big|\textstyle\sum_{t=1}^N Y_t\Big|^s \leq C_s\Big\{N^{s/2}\,(\mathbb{E}Y_1^2)^{s/2} + N\,\mathbb{E}|Y_1|^s\Big\}, \qquad s \geq 2, \tag{5.5}$$

which sharpens Marcinkiewicz–Zygmund (Rosenthal, 1970). Its chain analogue runs through the martingale decomposition: with $\widehat{g}$ solving the Poisson equation $(I - P)\widehat{g} = f - \pi(f)$,

$$\sum_{t=1}^N \big(f(X_t) - \pi(f)\big) = \underbrace{\sum_{t=1}^N \big(\widehat{g}(X_t) - P\widehat{g}(X_{t-1})\big)}_{\text{martingale } M_N} + \underbrace{P\widehat{g}(X_0) - P\widehat{g}(X_N)}_{\text{boundary}}, \tag{5.6}$$

a martingale plus boundary terms: Burkholder's inequality bounds M_N at the $N^{s/2}$ rate, the

boundary terms are $\mathcal{O}(1)$, and the scaling survives with a constant set by the mixing (Jones, 2004; Douc et al., 2018). Under uniform geometric ergodicity $\widehat{g}$ is bounded and the argument is direct; under geometric ergodicity it needs V-norm control of $\widehat{g}$ (Mijatović & Vogrinc, 2018). Carrying the SNIS moment and bias analysis through both, with the mixing penalty explicit, and settling whether a well-mixing chain then improves on independent draws, is left to future work.

5.2 Certifying and robustifying generalised posteriors

The mutual-information bound of Chapter 3 (Theorem 3.2) is one-sided: for the Rényi divergence $\mathcal{D}_\alpha$, $\alpha \in (0,1)$, it controls the expected divergence of the generalised posterior, fractional or its variational projection, and yields the oracle inequality of Theorem 3.7, where the in-model risk is bounded by the best achievable in-model risk plus a complexity term of order d_π/n^{κ_π}, with κ_π the prior-mass exponent. Whether it can be made two-sided, controlling the absolute deviation between the divergence and its empirical counterpart rather than the signed gap, is open. The obstacle is intrinsic: $\mathcal{D}_\alpha$ only lower-bounds the Kullback–Leibler divergence, so a genuine two-sided statement must control the difference $\mathrm{KL} - \mathcal{D}_\alpha$, presumably at a slower rate. Such concentration would furnish confidence statements, not contraction alone. The complexity term, too, ties the bound to the wider information-theoretic theory of generalisation (Hellström et al., 2025): $\mathcal{I}(\widehat{\theta}; \mathcal{S})$ can be infinite for a continuous model, whereas the conditional mutual information of Steinke & Zakynthinou (2020), built on a supersample, never exceeds $n \log 2$, and a bound in it, transductive by nature (Bégin et al., 2014), would stay finite where the unconditional one diverges. Recasting the bound in these sharper complexities, and so connecting it to the information-theoretic analysis of statistical inference, is left for future work.

Two-sided control is what adaptive model selection asks for. Lepski's method (Lepskiĭ, 1991) compares estimators across scales and keeps the least regularised one still consistent with the rest, a comparison the one-sided bound cannot support. Alquier (2008) carried this into PAC-Bayes for general losses, where a symmetric bound and a margin condition yield a Lepski-type rule at the oracle rate. Whether the mutual-information bound of Chapter 3 can be turned the same way, localising each candidate model's prior and aggregating them into one mixture to select at the oracle rate without the spurious factors of discrete selection, is unresolved. Localisation also serves beyond model selection, at the meta level, where a prior learned from T tasks is itself estimated at the fast rate $\mathcal{O}(1/T)$ (Riou et al., 2025).

The mutual-information bound holds for the variational projection, not only the exact posterior, so it certifies the object one actually computes. A variational approximation of a tempered posterior over a sparse deep network contracts at the near-minimax rate $\mathcal{O}(n^{-2\beta/(2\beta+d)} \log^2 n)$ for a β-Hölder regression function (Chérief-Abdellatif, 2020), and

high-dimensional spike-and-slab priors both contract and select at the minimax rate (Castillo et al., 2015; Castillo, 2024); the variational and fractional-posterior bounds of Alquier & Ridgway (2020) and Bhattacharya et al. (2019) are the PAC-Bayes counterpart, reaching the same rates from a prior-mass condition. A localised mutual-information bound would certify these rates at finite n, the sharper complexity shedding the residual log factors; across a family of architectures, the same model-selection step would then choose depth and sparsity adaptively. A two-sided bound would add what contraction alone cannot: lower-bounding the divergence as well would certify that the variational posterior is concentrated at the right scale, neither too loosely nor too tightly, the property that frequentist coverage of its credible sets would rest on.

Robustness asks the same of the loss. The $\widetilde{\rho}_\lambda$-posterior of Chapter 4 smooths the intractable supremum contrast of the ρ-posterior by the log-sum-exp surrogate Λ_λ ((4.10)), which recovers the supremum as $\lambda \to \infty$ and yields a Gibbs posterior ((4.34)). The bounded contrast is one of several in robust statistics — redescending M-estimation scores (Hampel, 1974), density-power divergences (Basu et al., 1998), the kernel discrepancies of MMD-Bayes (Chérief-Abdellatif & Alquier, 2020) — each trading the penalisation of outliers against the convexity of the resulting optimisation. Choosing among them is the loss-side analogue of choosing the prior. The contrast of Chapter 4 is built on the Hellinger distance, matching the log-likelihood near the model yet staying bounded under contamination, and its smoothed objective, though nonconvex, is strongly concave and well-conditioned uniformly in λ. A two-step design would take both ends: a convex contrast first learns a data-driven prior, on which the $\widetilde{\rho}_\lambda$-posterior is run for the guarantee, the first step buying optimisation and the second the robustness certificate.

Appendices

Appendix A

Proofs of Chapter 2

This appendix collects the proofs of the results stated in Chapter 2.

A.1 Proofs of Section 2.1

Proof of Proposition 2.5. It suffices to establish (2.14), since (2.15) and (2.16) then follow by the triangle inequality in L^p and Markov's inequality, respectively. By the Marcinkiewicz–Zygmund-type inequality of Petrov (1975) (Section III.5, item 16, p. 60), for $U_1, \dots, U_N$ independent, mean-zero random variables with finite p-th moment, $p \geq 2$,

$$\mathbb{E}\left[\left|\sum_{n=1}^{N} U_n\right|^p\right] \leq m(p)\, N^{p/2-1} \sum_{n=1}^{N} \mathbb{E}[|U_n|^p],$$

where $m(p)$ depends only on p and satisfies $(m(p))^{1/p} \leq p - 1$ (Ren & Liang, 2001). Setting $U_n = \omega(X_n) - 1$, which has zero mean since $q(\omega) = 1$, and passing to the average $\hat{Z}_N(\mathbf{X}^N) - 1 = N^{-1} \sum_{n=1}^{N} U_n$, we obtain

$$\mathbb{E}_{\bar{q}}\left[|\hat{Z}_N(\mathbf{X}^N) - 1|^p\right] \leq (p-1)^p\, N^{-p/2}\, q(|\omega - 1|^p). \tag{A.1}$$

The C_p-inequality gives $q(|\omega - 1|^p) \leq 2^{p-1}(1 + q(\omega^p))$, and since $q(\omega^p) \geq q(\omega)^p = 1$ by Jensen's inequality, we have $1 + q(\omega^p) \leq 2\, q(\omega^p)$, so $q(|\omega - 1|^p) \leq 2^p\, q(\omega^p)$. Substituting into (A.1) and using $p - 1 \leq p$ yields

$$\mathbb{E}_{\bar{q}}\left[|\hat{Z}_N(\mathbf{X}^N) - 1|^p\right] \;\leq\; (2p)^p\, q(\omega^p)\, N^{-p/2} \;=\; M(p)^p\, N^{-p/2},$$

and taking p-th roots gives (2.14). □

A.2 Proofs of Section 2.2

A.2.1 Proof of Theorem 2.8

We start with a technical result on the inverse moments of averages.

Proposition A.1. *Let $r \geq 1$ and let $\mathbf{X}^N = (X_1, \dots, X_N)$ be i.i.d. from q on $\mathbb{X}$. Suppose $\omega : \mathbb{X} \to (0, \infty)$ satisfies $q(\omega^{-\eta}) < \infty$ for some $\eta > 0$. Then, for all integers $N \geq r/\eta$,*

$$\mathbb{E}_{\mathbf{X}^N \sim \bar{q}}\left[\hat{Z}_N(\mathbf{X}^N)^{-r}\right] \;\leq\; q(\omega^{-\eta})^{r/\eta} \;<\; \infty. \tag{A.2}$$

The bound on the right-hand side is independent of N.

Proof. Write $\Omega_n = \omega(X_n)$ and recall $\hat{Z}_N(\mathbf{X}^N) = N^{-1}\sum_{n=1}^N \Omega_n$. By the AM–GM inequality applied to $\Omega_1, \dots, \Omega_N$,

$$\hat{Z}_N(\mathbf{X}^N) = \frac{\Omega_1 + \cdots + \Omega_N}{N} \;\geq\; (\Omega_1 \cdots \Omega_N)^{1/N},$$

so, taking reciprocals and raising to the power r,

$$\hat{Z}_N(\mathbf{X}^N)^{-r} \;\leq\; (\Omega_1 \cdots \Omega_N)^{-r/N} = \prod_{n=1}^N \Omega_n^{-r/N}.$$

Taking expectations and using independence of $X_1, \dots, X_N$ under $\bar{q}$,

$$\mathbb{E}_{\mathbf{X}^N \sim \bar{q}}\left[\hat{Z}_N(\mathbf{X}^N)^{-r}\right] \;\leq\; \prod_{n=1}^N q(\omega^{-r/N}) \;=\; q(\omega^{-r/N})^N.$$

Since $N \geq r/\eta$, we have $r/N \leq \eta$, so the map $t \mapsto t^{r/(N\eta)}$ is concave on $(0, \infty)$. Applying Jensen's inequality to the random variable $\omega(x)^{-\eta}$ with $x \sim q$,

$$q(\omega^{-r/N}) \;=\; q\left[(\omega^{-\eta})^{r/(N\eta)}\right] \;\leq\; q(\omega^{-\eta})^{r/(N\eta)}.$$

Substituting back,

$$q(\omega^{-r/N})^N \;\leq\; q(\omega^{-\eta})^{r/\eta},$$

which gives the claimed bound. □

Proof of Theorem 2.8. We first write the rescaled bias of the SNIS estimator as,

$$N \times \mathbb{E}_{\mathbf{X}^N \sim \bar{q}}\left[\hat{F}_N(\mathbf{X}^N) - \pi(f)\right] = \mathbb{E}\left[\frac{\sum_{n=1}^N \omega(X_n)(f(X_n) - \pi(f))}{\sum_{n=1}^N \omega(X_n)/N}\right]$$

$$= N\mathbb{E}\left[\frac{\omega(X_1)(f(X_1)-\pi(f))}{\sum_{n=1}^N \omega(X_n)/N}\right]$$

by identity in distribution

$$= N\mathbb{E}\left[\frac{\omega(X_1)(f(X_1)-\pi(f))}{\sum_{n=2}^N \omega(X_n)/N}\right]$$

$$+ N\mathbb{E}\Big[\omega(X_1)(f(X_1)-\pi(f)) \Big\{\frac{1}{\sum_{n=1}^N \omega(X_n)/N} - \frac{1}{\sum_{n=2}^N \omega(X_n)/N}\Big\}\Big].$$

Since X_1 is independent of $(X_2,\ldots,X_N)$ and $q(\omega(f-\pi(f))) = q(\omega f) - \pi(f)q(\omega) = 0$, the first term vanishes. For the second term,

$$\frac{1}{\sum_{n=1}^N \omega(X_n)/N} - \frac{1}{\sum_{n=2}^N \omega(X_n)/N} = \frac{-\omega(X_1)/N}{(\sum_{n=1}^N \omega(X_n)/N)(\sum_{n=2}^N \omega(X_n)/N)}. \quad \text{(A.3)}$$

Substituting, we obtain

$$N \times \mathbb{E}_{\mathbf{X}^N\sim\bar{q}}\left[\hat{F}_N(\mathbf{X}^N) - \pi(f)\right] = -\mathbb{E}\left[\frac{\omega(X_1)^2(f(X_1)-\pi(f))}{(\sum_{n=1}^N \omega(X_n)/N)(\sum_{n=2}^N \omega(X_n)/N)}\right].$$

Define

$$A_N := \frac{\omega(X_1)^2(f(X_1)-\pi(f))}{(\sum_{n=1}^N \omega(X_n)/N)(\sum_{n=2}^N \omega(X_n)/N)}.$$

By the strong law of large numbers, $\sum_{n=1}^N \omega(X_n)/N \xrightarrow{a.s.} 1$ and $\sum_{n=2}^N \omega(X_n)/N \xrightarrow{a.s.} 1$ as $N\to\infty$, so

$$A_N \xrightarrow{a.s.} \omega(X_1)^2(f(X_1)-\pi(f)) \quad \text{as } N\to\infty.$$

If $\{A_N\}_{N\geq 1}$ is uniformly integrable, we may exchange expectation and limit to get

$$\lim_{N\to\infty} N \times \mathbb{E}_{\mathbf{X}^N\sim\bar{q}}\left[\hat{F}_N(\mathbf{X}^N) - \pi(f)\right] = -\mathbb{E}[\lim_{N\to\infty} A_N] = -q\left(\omega^2(f-\pi(f))\right),$$

which is the claimed result.

It remains to establish uniform integrability of $\{A_N\}_{N\geq 1}$. By Billingsley (1999), Equation (3.18), it suffices to find $\epsilon > 0$ such that

$$\sup_N \mathbb{E}[|A_N|^{1+\epsilon}] < \infty.$$

Since $\omega > 0$ and $\sum_{n=1}^N \omega(X_n)/N \geq \sum_{n=2}^N \omega(X_n)/N$ a.s., we can a.s. bound

$$|A_N| \leq \frac{|\omega(X_1)^2(f(X_1) - \pi(f))|}{(\sum_{n=2}^N \omega(X_n)/N)^2}.$$

Using independence of X_1 and $(X_2, \ldots, X_N)$,

$$\mathbb{E}[|A_N|^{1+\epsilon}] \leq \mathbb{E}\left[\left|\omega(X_1)^2(f(X_1) - \pi(f))\right|^{1+\epsilon}\right] \mathbb{E}\left[\left(\frac{N}{\sum_{n=2}^N \omega(X_n)}\right)^{2(1+\epsilon)}\right].$$

Write $\hat{Z}_{N\setminus 1} = (N-1)^{-1} \sum_{n=2}^N \omega(X_n)$ for the sample average of the last $N-1$ weights. Since $N \leq 2(N-1)$ for $N \geq 2$, we have $N^{-1} \sum_{n=2}^N \omega(X_n) = \frac{N-1}{N} \hat{Z}_{N\setminus 1} \geq \frac{1}{2} \hat{Z}_{N\setminus 1}$, so

$$\left(\frac{N}{\sum_{n=2}^N \omega(X_n)}\right)^{2(1+\epsilon)} \leq 2^{2(1+\epsilon)}\, \hat{Z}_{N\setminus 1}^{-2(1+\epsilon)}.$$

Applying Proposition A.1 with $r = 2(1+\epsilon)$ to the $N-1$ i.i.d. variables $X_2, \ldots, X_N$, we get $\mathbb{E}[\hat{Z}_{N\setminus 1}^{-2(1+\epsilon)}] \leq q(\omega^{-\eta})^{2(1+\epsilon)/\eta}$ for $N - 1 \geq 2(1+\epsilon)/\eta$. The first factor in the bound on $\mathbb{E}[|A_N|^{1+\epsilon}]$ is finite by assumption. Hence $\sup_N \mathbb{E}[|A_N|^{1+\epsilon}] < \infty$, and uniform integrability follows by Billingsley (1999, (3.18)). □

A.2.2 Proof of Theorem 2.9

Proof of Theorem 2.9. For a sample $X_1, \ldots, X_N$ i.i.d. from q, we write their empirical distribution q^N so that, for example, $q^N(\omega) := \sum_{n=1}^N \omega(X_n)/N$. The IS estimator is thus $q^N(f\omega)/q^N(\omega)$. First, it is enough to consider the case where the test function f is non-negative. Indeed, for a general function f we write $f = f_+ - f_-$ where $f_+(x) := \max\{f(x), 0\}$ and $f_-(x) := -\min\{f(x), 0\}$. Then

$$\begin{aligned}\left|\frac{q^N(f\omega)}{q^N(\omega)} - \pi(f)\right| &= \left|\frac{q^N(f_+\omega) - q^N(f_-\omega)}{q^N(\omega)} - (\pi(f_+) - \pi(f_-))\right| \\ &\leq \left|\frac{q^N(f_+\omega)}{q^N(\omega)} - \pi(f_+)\right| + \left|\frac{q^N(f_-\omega)}{q^N(\omega)} - \pi(f_-)\right|.\end{aligned}$$

Using $(a+b)^s \leq 2^{s-1}(a^s + b^s)$, and applying the result for non-negative functions f_+ and f_- separately, we obtain the result for general f. We thus assume that f is non-negative.

We write the absolute error between the IS estimator with the target $\pi(f) = q(f\omega)$ in two different ways. The first is:

$$\left|\frac{q^N(f\omega)}{q^N(\omega)} - q(f\omega)\right| \leq \max_{1\leq i\leq N} f(x_i) + q(f\omega). \tag{A.4}$$

The second is:

$$\left|\frac{q^N(f\omega)}{q^N(\omega)} - q(f\omega)\right| \leq \left|\frac{q^N(f\omega)}{q^N(\omega)} - \frac{q(f\omega)}{q^N(\omega)}\right| + q(f\omega)\left|\frac{1}{q^N(\omega)} - 1\right|.$$

Now we consider two cases: 1) $|q^N(\omega) - 1| > \frac{1}{2}$, 2) $|q^N(\omega) - 1| \leq \frac{1}{2}$. We will separately bound the expected error under the two cases using the two inequalities above.

We start with the first case, and we assume $r < \infty$. First, we use $(a+b)^s \leq 2^{s-1}(a^s + b^s)$ to write

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{q^N(f\omega)}{q^N(\omega)} - q(f\omega)\right|^s \mathbb{1}(|q^N(\omega) - 1| > \tfrac{1}{2})\right] \\
&\leq \mathbb{E}\left[\left(\max_{1\leq i\leq N} f(X_i) + q(f\omega)\right)^s \mathbb{1}(|q^N(\omega) - 1| > \tfrac{1}{2})\right] \\
&\leq 2^{s-1}\mathbb{E}\left[\left(\max_{1\leq i\leq N} f(X_i)\right)^s \mathbb{1}(|q^N(\omega) - 1| > \tfrac{1}{2})\right] \\
&\quad + 2^{s-1}\left(q(f\omega)\right)^s \mathbb{P}(|q^N(\omega) - 1| > \tfrac{1}{2}).
\end{aligned}$$

The second term leads to a bound in $N^{-s/2}$ using Markov's inequality as in Proposition 2.5, since $q(\omega^s) < \infty$ under the assumptions. The first term is dealt with first using Hölder's inequality with exponents r/s and $(1 - s/r)^{-1}$,

$$\begin{aligned}
&\mathbb{E}\left[\left(\max_{1\leq i\leq N} f(X_i)\right)^s \mathbb{1}(|q^N(\omega) - 1| > \tfrac{1}{2})\right] \\
&\leq \mathbb{E}\left[\left(\max_{1\leq i\leq N} f(X_i)\right)^r\right]^{s/r} \times \mathbb{P}(|q^N(\omega) - 1| > \tfrac{1}{2})^{1-s/r} \\
&\leq q(f^r)^{s/r} N^{s/r} \cdot C \cdot N^{-\frac{1}{2}p(1-s/r)},
\end{aligned}$$

for a constant C. The last inequality uses the fact that $\mathbb{E}[(\max_{1\leq i\leq N} f(X_i))^r] \leq N\mathbb{E}[f(X_1)^r]$, and Markov's inequality using $q(\omega^p) < \infty$. Given $s \leq pr/(p+r+2)$, the exponent of N satisfies

$$\frac{s}{r} - \frac{p(r-s)}{2r} = \frac{2s + ps - pr}{2r} \leq \frac{-s}{2},$$

using $s \leq pr/(p+r+2) \Leftrightarrow -pr \leq -s(p+r+2)$. Altogether we arrive at

$$\mathbb{E}\left[\left|\frac{q^N(f\omega)}{q^N(\omega)} - q(f\omega)\right|^s \mathbb{1}(|q^N(\omega) - 1| > \tfrac{1}{2})\right] \leq CN^{-s/2},$$

for another constant C.

In the case $r = \infty$, we can directly write

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{q^N(f\omega)}{q^N(\omega)} - q(f\omega)\right|^s \mathbb{1}(|q^N(\omega) - 1| > \tfrac{1}{2})\right] \\
&\leq \mathbb{E}\left[\left(\max_{1\leq i\leq N} f(X_i) + q(f\omega)\right)^s \mathbb{1}(|q^N(\omega) - 1| > \tfrac{1}{2})\right] \\
&\leq 2^s |f|_\infty^s \mathbb{P}(|q^N(\omega) - 1| > \tfrac{1}{2}) \\
&\leq 2^s C |f|_\infty^s N^{-\frac{1}{2}p} \leq C N^{-\frac{1}{2}s},
\end{aligned}$$

using $s \leq \min\{p, r\} \leq p$ in the last line, and changing the value of C between inequalities.

For the case $|q^N(\omega) - 1| \leq \frac{1}{2}$,

$$\begin{aligned}
\left|\frac{q^N(f\omega)}{q^N(\omega)} - \pi(f)\right| \mathbb{1}(|q^N(\omega) - 1| \leq 0.5) &\leq 2|q^N(f\omega) - \pi(f)| + \pi(f)\left|\frac{q^N(\omega) - 1}{q^N(\omega)}\right| \\
&\leq 2|q^N(f\omega) - \pi(f)| + 2\pi(f)\left|q^N(\omega) - 1\right|.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{q^N(f\omega)}{q^N(\omega)} - \pi(f)\right|^s \mathbb{1}(|q^N(\omega) - 1| \leq 0.5)\right] \\
&\leq C\left(\mathbb{E}[|q^N(f\omega) - \pi(f)|^s] + \mathbb{E}[\left|q^N(\omega) - 1\right|^s]\right) \\
&\leq C N^{-s/2},
\end{aligned}$$

for some constant C that changes at each line. The first term is $O(N^{-s/2})$ with a reasoning similar to that in the proof of Proposition 2.5, since $q^N(f\omega)$ is the average of N i.i.d. random variables with mean $q(f\omega)$ and s finite moments, since $s \leq pr/(p + r + 2) \leq pr/(p + r)$. Putting everything together gives

$$\mathbb{E}\left[\left|\frac{q^N(f\omega)}{q^N(\omega)} - \pi(f)\right|^s\right] \leq C N^{-s/2}.$$

□

A.2.3 Proof of Proposition 2.10

Proof of Proposition 2.10. Set $\alpha := pr/(p + r)$, so that $1/\alpha = 1/p + 1/r$ and $1 < \alpha \leq 2$ under the assumptions of Proposition 2.10. By Hölder's inequality with conjugate exponents p/α and r/α,

$$q(|f\omega|^\alpha) \leq [q(\omega^p)]^{\alpha/p}\ [q(|f|^r)]^{\alpha/r} < \infty, \tag{A.5}$$

so the summands $\omega(X_n)f(X_n)$ have a finite absolute moment of order α. The assumption $\alpha > 1$, equivalent to $1/p + 1/r < 1$, is exactly the integrability threshold beyond which the von Bahr–Esseen inequality of Lemma A.8 sharpens the trivial $\mathcal{O}(1)$ bound on centred sums into the decaying rate $N^{-(\alpha-1)/\alpha}$.

Proposition A.9 controls the centred sum of weights:

$$\mathbb{E}\big[\,\big|q^N(\omega) - 1\big|^p\,\big] \le \frac{C}{N^{p-1}}, \qquad \mathbb{P}\big(\,\big|q^N(\omega) - 1\big| > \tfrac{1}{2}\,\big) \le \frac{C}{N^{p-1}}. \tag{A.6}$$

For the centred numerator, set $Y_n := \omega(X_n)f(X_n) - q(f\omega)$ and apply Lemma A.8 at order α to $\sum_n Y_n$, then Lyapunov's inequality at $1 \le s \le \alpha$:

$$\mathbb{E}\big[\,\big|q^N(f\omega) - q(f\omega)\big|^s\,\big] \le C\, N^{-s(\alpha-1)/\alpha}. \tag{A.7}$$

Specializing to $f \equiv 1$ gives the analogous bound

$$\mathbb{E}\big[\,\big|q^N(\omega) - 1\big|^s\,\big] \le C\, N^{-s(p-1)/p}, \qquad 1 \le s \le p. \tag{A.8}$$

From here the argument follows that of Theorem 2.9, the only change being that its three Marcinkiewicz–Zygmund rates are replaced by the heavy-tail bounds derived above: on the event $\{|q^N(\omega) - 1| > 1/2\}$ the tail of $q^N(\omega) - 1$ becomes $N^{-(p-1)}$ via (A.6), while on its complement the L^s-norms of $q^N(f\omega) - q(f\omega)$ and of $q^N(\omega) - 1$ become $N^{-s(\alpha-1)/\alpha}$ and $N^{-s(p-1)/p}$ via (A.7) and (A.8). The reduction to $f \ge 0$, the deterministic bounds on each event, and the Hölder split with conjugate exponents r/s and $(1 - s/r)^{-1}$ (and its simpler $r = \infty$ counterpart) are unchanged.

On $\{|q^N(\omega) - 1| > 1/2\}$ with $r < \infty$, the Hölder split, together with $\mathbb{E}[(\max_i f(X_i))^r] \le N\, q(|f|^r)$ and (A.6), contributes a term of order $N^{\,sp/r-(p-1)}$, while on the complementary event the slower of the two heavy-tail rates is $N^{-s(\alpha-1)/\alpha}$, since $\alpha \le p$. The former is dominated by the latter precisely when

$$\frac{sp}{r} - (p-1) \le -\frac{s(\alpha-1)}{\alpha}, \tag{A.9}$$

which, using $1/\alpha = 1/p + 1/r$, rearranges to

$$s\,\frac{(p-1)(p+r)}{pr} \le p - 1, \qquad \text{i.e.} \qquad s \le \alpha, \tag{A.10}$$

the admissible range of Proposition 2.10. Combining the two events,

$$\mathbb{E}_{\bar{q}}\Big[\,\Big|\hat{F}(\mathbf{X}^N) - \pi(f)\Big|^s\,\Big] \le C\, N^{-s(\alpha-1)/\alpha}, \qquad 1 \le s \le \alpha.$$

For bounded f, $r = \infty$ gives $\alpha = p$ and the rate $N^{-s(p-1)/p}$ for $1 \leq s \leq p$. As $p \uparrow 2$, $\alpha \uparrow 2$ and $s(\alpha - 1)/\alpha \to s/2$, so the bound is continuous with Theorem 2.9 at the boundary. □

A.3 Proofs of Section 2.3

We assume that both target and proposal distributions admit densities with respect to a measure λ. Although we will express all subsequent notations using integration, this should be interpreted as summation when the space is discrete and λ represents the counting measure. The rejection probability at x is denoted by

$$r(x) = \int_{x^\star \neq x} \left(1 - \min\left(1, \frac{\hat{Z}(x^\star)}{\hat{Z}(x)}\right)\right) \bar{q}(x^\star)\, \lambda(\mathrm{d}x^\star).$$

That definition only considers the probability of moves to states different than x that are rejected. We will use the following fact: at every iteration, for each chain one of the following three events occurs: 1) a proposal to a different state is accepted, 2) a proposal to a different state is rejected, 3) a proposal is made to the current state (and systematically accepted). In a continuous state space with an atomless measure λ, the last event occurs with probability zero. We assume Assumption 2.1 throughout so that $\mathbb{P}(\hat{Z}(x) = 0) = 0$ under q.

We first prove a lemma that describes the coupling time τ.

Lemma A.2. *Assuming $\omega(x) \geq \omega(y)$, we have the following facts:*

- *Let τ_0 be the first time when the X-chain moves to a different state. Then $\tau \leq \tau_0$, i.e. the chains meet at τ_0 or earlier.*
- *Let τ_1 be the first time when a common proposal is x. Then $\tau \leq \tau_1$, i.e. the chains meet at τ_1 or earlier.*
- *The meeting time satisfies $\tau = \min\{\tau_0, \tau_1\}$.*

Proof of Lemma A.2. The first two observations can be proven by induction, once we recognize that the common draws coupling of Algorithm 4 implies $\omega(X_t) \geq \omega(Y_t)$ for all $t \geq 0$. Regarding the last observation, for every $t < \min\{\tau_0, \tau_1\}$, the X-chain must have rejected moves to a different state than x at each iteration up to t. In that situation, the X-chain is still at x, while the Y-chain never proposed a move to x and thus $X_t = x \neq Y_t$, as claimed. □

Now we calculate the tail probability of τ.

Lemma A.3. *For all $t \geq 1$, $|P^t(x, \cdot) - P^t(y, \cdot)|_{TV} \leq \mathbb{P}_{x,y}(\tau > t) = \max(r(x), r(y))^t$.*

Proof of Lemma A.3. The inequality in the statement is the celebrated coupling inequality. For the equality, we assume $\omega(x) \geq \omega(y)$ without loss of generality, which implies $r(x) \geq r(y)$.

By Lemma A.2, the event $\{\tau > t\}$ is equivalent to $\{\min\{\tau_0, \tau_1\} > t\}$. The latter event corresponds to the event: "the X-chain proposes to move to a different state but gets rejected at each of the first t iterations". Then its probability is $r(x)^t$, since $r(x)$ is the probability of a failed attempt to move to a different state. □

It remains to show the following lower bound.

Lemma A.4. *For all* $t \geq 1$, $|P^t(x, \cdot) - P^t(y, \cdot)|_{TV} \geq \mathbb{P}_{x,y}(\tau > t)$.

Proof of Lemma A.4. Again, we assume $\omega(x) \geq \omega(y)$ without loss of generality. The definition of total variation distance as a supremum over measurable sets implies $|P^t(x, \cdot) - P^t(y, \cdot)|_{\text{TV}} \geq P^t(x, \{x\}) - P^t(y, \{x\})$, considering the set $\{x\}$.

Under the distribution of the coupled chains, we can write $P^t(x, \{x\}) - P^t(y, \{x\})$ as $\mathbb{P}(X_t = x) - \mathbb{P}(Y_t = x)$. Now we decompose each probability according to τ being greater or less than t, for any $t \geq 1$:

$$\begin{aligned}\mathbb{P}(X_t = x) - \mathbb{P}(Y_t = x) = \mathbb{P}(X_t = x; \tau > t) + \mathbb{P}(X_t = x; \tau \leq t) \\ - \mathbb{P}(Y_t = x; \tau > t) - \mathbb{P}(Y_t = x; \tau \leq t).\end{aligned}$$

We simplify with the following observations.

- Under the event $\tau > t$: we have $X_t = x$; otherwise, the X-chain would have successfully moved to a new state jointly with the Y-chain implying $\tau \leq t$ by Lemma A.2. Therefore,

  $$\mathbb{P}(X_t = x; \tau > t) = \mathbb{P}(\tau > t)\,\mathbb{P}(X_t = x \mid \tau > t) = \mathbb{P}(\tau > t).$$

  Meanwhile, under that event we have $Y_t \neq x$; otherwise, the Y-chain must have proposed a move to x at or before time t, and that would have resulted in a meeting by Lemma A.2. Therefore,

  $$\mathbb{P}(Y_t = x; \tau > t) = 0.$$

- Under the event $\tau \leq t$: we have $X_t = Y_t$, therefore $\mathbb{P}(X_t = x; \tau \leq t) = \mathbb{P}(Y_t = x; \tau \leq t)$.

Putting these together, we conclude that $\mathbb{P}(X_t = x) - \mathbb{P}(Y_t = x) = \mathbb{P}(\tau > t)$. □

Theorem 2.11 is obtained by combining Lemmas A.3 and A.4.

A.4 Proofs of Section 2.4

A.4.1 Proof of Proposition 2.12

Let $t \geq 1$. The event $\{\tau > t\}$ only occurs when Algorithm 5 enters its while loop, in which case we must have that 1) $\mathbf{X}_1^N = \mathbf{X}_0^N$, 2) $\hat{Z}(\mathbf{X}_0^N) > \hat{Z}(\mathbf{Y}_0^N)$, and 3) the first generated Uniform variable was greater than $\hat{Z}(\mathbf{Y}_0^N)/\hat{Z}(\mathbf{X}_0^N)$. Thus,

$$\mathbb{P}(\tau > t) = \iint \mathbb{P}_{\mathbf{x}_0^N, \mathbf{y}_0^N}(\tau > t) \left(1 - \min\left\{1, \frac{\hat{Z}(\mathbf{y}_0^N)}{\hat{Z}(\mathbf{x}_0^N)}\right\}\right) \tag{A.11}$$

$$\times \mathbb{1}\left(\hat{Z}(\mathbf{y}_0^N) < \hat{Z}(\mathbf{x}_0^N)\right) \mathbb{1}(\mathbf{y}_0^N \neq \mathbf{x}_0^N) \bar{q}(\mathrm{d}\mathbf{x}_0^N) \bar{q}(\mathrm{d}\mathbf{y}_0^N). \tag{A.12}$$

The quantity $\mathbb{P}_{\mathbf{x}_0^N, \mathbf{y}_0^N}(\tau > t)$ in the event $\hat{Z}(\mathbf{y}_0^N) < \hat{Z}(\mathbf{x}_0^N)$ is equal to $r(\mathbf{x}_0^N)^t$, as in Theorem 2.11. By upper-bounding the other terms by one and integrating with respect to $\bar{q}(\mathrm{d}\mathbf{y}_0^N)$, we obtain the upper bound

$$\mathbb{P}(\tau > t) \leq \int (r(\mathbf{x}_0^N))^t \bar{q}(\mathrm{d}\mathbf{x}_0^N) = \mathbb{E}_{\bar{q}}\left[(r(\mathbf{X}_0^N))^t\right]. \tag{A.13}$$

A.4.2 Proof of Proposition 2.13

We prove Proposition 2.13 by first splitting the expectation according to whether $\hat{Z}_N(\mathbf{X}^N)$ is less than or greater than 2:

$$\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t\right] = \mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) \leq 2)\right] + \mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) > 2)\right]. \tag{A.14}$$

We then proceed through a series of lemmas to bound each term. The following lemmas are used to handle the case where $\hat{Z}_N(\mathbf{X}^N) > 2$.

Lemma A.5. *Under Assumption 2.1 and $q(\omega^p) < \infty$ for any $p > 1$, the rejection probability* (2.21) *is upper bounded as follows, for any $\theta \in [0,1]$ and any $\mathbf{x}^N \in \mathbb{X}^N$:*

$$r(\mathbf{x}^N) \leq 1 - \min\left\{1, \frac{\theta}{\hat{Z}_N(\mathbf{x}^N)}\right\} c_p(\theta), \quad \textit{with} \quad c_p(\theta) = \frac{(1-\theta)^{p/(p-1)}}{q(\omega^p)^{1/(p-1)}} \in [0,1]. \tag{A.15}$$

Proof. Let $\theta \in [0,1]$. We start with an L^p-version of the Paley–Zygmund inequality, as on page 2705, equation (12) of Petrov (2007) with $r = 1$. If W is a non-negative random variable and $p > 1$, then

$$\mathbb{P}\left(W > \theta \mathbb{E}[W]\right) \geq \frac{(1-\theta)^{p/(p-1)} \left(\mathbb{E}[W]\right)^{p/(p-1)}}{\left(\mathbb{E}[W^p]\right)^{1/(p-1)}}. \tag{A.16}$$

Indeed, for any $b > 0$, Hölder's inequality implies

$$\begin{aligned}\mathbb{E}[W] &= \mathbb{E}[W\mathbb{1}(W > b)] + \mathbb{E}[W\mathbb{1}(W \leq b)] \\ &\leq \mathbb{P}(W > b)^{(1-1/p)}\mathbb{E}[W^p]^{1/p} + b.\end{aligned}$$

Rearranging with $b = \theta\mathbb{E}[W]$ implies (A.16). We apply this to $\hat{Z}_N(\mathbf{X}^N)$, under Assumption 2.1:

$$\mathbb{P}_{\bar{q}}\left(\hat{Z}_N(\mathbf{X}^N) > \theta\right) \geq \frac{(1-\theta)^{p/(p-1)}}{\left(\mathbb{E}_{\bar{q}}\left[\left(\hat{Z}_N(\mathbf{X}^N)\right)^p\right]\right)^{1/(p-1)}} \geq \frac{(1-\theta)^{p/(p-1)}}{q(\omega^p)^{1/(p-1)}}. \tag{A.17}$$

The latter inequality comes from Jensen's inequality, since $z \mapsto z^p$ with $p > 1$ is convex:

$$\mathbb{E}_{\bar{q}}\left[\left(\hat{Z}_N(\mathbf{X}^N)\right)^p\right] \leq \mathbb{E}_{\bar{q}}\left[\frac{1}{N}\sum_{n=1}^{N}\omega(X_n)^p\right] = q(\omega^p). \tag{A.18}$$

Inequality (A.17) implies that

$$\begin{aligned}\int \min\left\{1, \frac{\hat{Z}_N(\mathbf{x}^{N,\star})}{\hat{Z}_N(\mathbf{x}^N)}\right\}\bar{q}(\mathrm{d}\mathbf{x}^{N,\star}) &= \int_{\{\mathbf{x}^{N,\star}:\hat{Z}_N(\mathbf{x}^{N,\star})\leq\theta\}} \min\left\{1, \frac{\hat{Z}_N(\mathbf{x}^{N,\star})}{\hat{Z}_N(\mathbf{x}^N)}\right\}\bar{q}(\mathrm{d}\mathbf{x}^{N,\star}) \\ &\quad + \int_{\{\mathbf{x}^{N,\star}:\hat{Z}_N(\mathbf{x}^{N,\star})>\theta\}} \min\left\{1, \frac{\hat{Z}_N(\mathbf{x}^{N,\star})}{\hat{Z}_N(\mathbf{x}^N)}\right\}\bar{q}(\mathrm{d}\mathbf{x}^{N,\star}) \\ &\geq 0 + \min\left\{1, \frac{\theta}{\hat{Z}_N(\mathbf{x}^N)}\right\}\mathbb{P}_{\bar{q}}\left(\hat{Z}_N(\mathbf{x}^{N,\star}) > \theta\right) \\ &\geq \min\left\{1, \frac{\theta}{\hat{Z}_N(\mathbf{x}^N)}\right\}\frac{(1-\theta)^{p/(p-1)}}{q(\omega^p)^{1/(p-1)}}.\end{aligned}$$

This yields the desired result. □

Lemma A.6. *Under Assumptions 2.1-2.2 with $p \geq 2$, there exists a constant $C > 0$ such that for all $t \geq 1$, $N \geq 1$,*

$$\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t\mathbb{1}(\hat{Z}_N(\mathbf{X}^N) > 2)\right] \leq \frac{C}{N^{p/2}t^p}.$$

Proof. We split the expectation into two parts:

$$\begin{aligned}\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t\mathbb{1}\left(\hat{Z}_N(\mathbf{X}^N) > 2\right)\right] &= \mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t\mathbb{1}\left(\hat{Z}_N(\mathbf{X}^N) \in (2, 1+t)\right)\right] \\ &\quad + \mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t\mathbb{1}\left(\hat{Z}_N(\mathbf{X}^N) \geq 1+t\right)\right].\end{aligned}$$

For $\{\hat{Z}_N(\mathbf{X}^N) \geq 1+t\}$, we directly apply Proposition 2.5:

$$\begin{aligned}\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) \geq 1+t)\right] &\leq \mathbb{P}_{\bar{q}}\left(\hat{Z}_N(\mathbf{X}^N) \geq 1+t\right) \\ &\leq \frac{M(p)^p}{N^{p/2}t^p}.\end{aligned}$$

For $\{\hat{Z}_N(\mathbf{X}^N) \in (2, 1+t)\}$, we use Lemma A.5 with $\theta = 1/2$:

$$r(\mathbf{x}^N) \leq 1 - \frac{c_p(1/2)}{2\hat{Z}_N(\mathbf{x}^N)} \leq \exp\left(-\frac{c_p(1/2)}{2\hat{Z}_N(\mathbf{x}^N)}\right).$$

Let $c = c_p(1/2)/4$. Then using the fact that $\hat{Z}_N(\mathbf{x}^N) > 2$ implies that $\hat{Z}_N(\mathbf{x}^N) \leq 2(\hat{Z}_N(\mathbf{x}^N) - 1)$, we have:

$$r(\mathbf{x}^N)^t \leq \exp\left(-\frac{2ct}{\hat{Z}_N(\mathbf{x}^N)}\right) \leq \exp\left(-\frac{ct}{\hat{Z}_N(\mathbf{x}^N) - 1}\right).$$

We introduce the sets $A_k = [t/(k+1), t/k]$ for $k \geq 1$, so that $\cup_{k=1}^{\infty} A_k = (0, t]$ which contains $(1, t)$. Using the result of Proposition 2.5, we obtain the bound:

$$\begin{aligned}&\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) \in (2, 1+t))\right] \\ &\leq \mathbb{E}_{\bar{q}}\left[\exp\left(-\frac{ct}{\hat{Z}_N(\mathbf{X}^N) - 1}\right) \mathbb{1}\left(\hat{Z}_N(\mathbf{X}^N) - 1 \in (1, t)\right)\right] \\ &\leq \sum_{k=1}^{\infty} \mathbb{E}_{\bar{q}}\left[\exp\left(-\frac{ct}{\hat{Z}_N(\mathbf{X}^N) - 1}\right) \mathbb{1}\left(\hat{Z}_N(\mathbf{X}^N) - 1 \in A_k\right)\right] \\ &\leq \sum_{k=1}^{\infty} \exp(-ck) \mathbb{P}_{\bar{q}}\left(\hat{Z}_N(\mathbf{X}^N) \geq 1 + t/(k+1)\right) \\ &\leq \sum_{k=1}^{\infty} \exp(-ck) \frac{M(p)^p}{N^{p/2}} \left(\frac{k+1}{t}\right)^p.\end{aligned}$$

Let $S_p = \sum_{k=1}^{\infty} \exp(-ck)(k+1)^p$, which is finite. Then:

$$\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}\left(\hat{Z}_N(\mathbf{X}^N) \in (2, 1+t)\right)\right] \leq \frac{M(p)^p S_p}{N^{p/2}t^p}.$$

Combining the bounds for both parts, we get:

$$\begin{aligned}\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}\left(\hat{Z}_N(\mathbf{X}^N) > 2\right)\right] &\leq \frac{M(p)^p}{N^{p/2}t^p} + \frac{M(p)^p S_p}{N^{p/2}t^p} \\ &\leq \frac{M(p)^p (1 + S_p)}{N^{p/2}t^p}.\end{aligned}$$

Setting the new constant $C := M(p)^p\,(1 + S_p)$ completes the proof. □

Now, we turn our attention to controlling the expectation when $\hat{Z}_N(\mathbf{X}^N) \leq 2$.

Lemma A.7. *Fix $p \geq 2$ and let β_p be defined as in* (2.26)*. There exist constants $A_p, B_p > 0$, depending only on p and $q(\omega^p)$, such that for all $N \geq 1$, for all $t \geq 1$, the following holds:*

$$\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}\left(\hat{Z}_N(\mathbf{X}^N) \leq 2\right)\right] \leq \left[\frac{A_p}{N^{\frac{t\wedge p}{2}}} + \frac{B_p}{N^{p/2}}\right]\beta_p^t.$$

Proof. We abuse notation to write r as a function of the value z taken by $\hat{Z}_N(\mathbf{X}^N)$, instead of a function of $\mathbf{X}^N$, in various places in this proof. First notice that $r(z)$ is increasing in z. We thus have that for $t \geq 1$ that

$$\begin{aligned}&\mathbb{E}_{\bar{q}}\left[r(\hat{Z}_N(\mathbf{X}^N))^t \mathbb{1}\left(\hat{Z}_N(\mathbf{X}^N) \leq 2\right)\right]\\ &\leq r(2)^{t-t\wedge p}\mathbb{E}_{\bar{q}}\left[r(\hat{Z}_N(\mathbf{X}^N))^{t\wedge p}\mathbb{1}\left(\hat{Z}_N(\mathbf{X}^N) \leq 2\right)\right].\end{aligned}$$

We first consider the second factor. We have for any $\alpha \in (0,1)$,

$$\begin{aligned}&\mathbb{E}_{\bar{q}}\left[r(\hat{Z}_N(\mathbf{X}^N))^{t\wedge p}\mathbb{1}\left(\hat{Z}_N(\mathbf{X}^N) \leq 2\right)\right]\\ &\leq \mathbb{E}_{\bar{q}}\left[r(\hat{Z}_N(\mathbf{X}^N))^{t\wedge p}\mathbb{1}\left(1-\alpha \leq \hat{Z}_N(\mathbf{X}^N) \leq 2\right)\right] + r(2)^{t\wedge p}\bar{q}\left\{|\hat{Z}_N(\mathbf{X}^N) - 1| \geq \alpha\right\}.\end{aligned} \tag{A.19}$$

That is because $\{\hat{Z}_N(\mathbf{X}^N) \leq 1-\alpha\} \subset \{|\hat{Z}_N(\mathbf{X}^N) - 1| \geq \alpha\}$, and $r(z) \leq r(2)$ for $z \leq 2$.

At this point, notice that by Lemma A.5 with $\theta = 1/2$ we have that $r(2) \leq 1 - c_p(1/2)/4 = \beta_p$, where β_p is defined in (2.26). Writing ν for the law of $\hat{Z}_N$ under $\bar{q}$, we have that

$$\begin{aligned}r(z) &= 1 - \int \min\left\{1, \frac{z^*}{z}\right\}\nu(\mathrm{d}z^*) = \int_{z^*=0}^{\infty}\nu(\mathrm{d}z^*) - \int_{z^*=0}^{z}\frac{z^*}{z}\nu(\mathrm{d}z^*) - \int_{z^*=z}^{\infty}\nu(\mathrm{d}z^*)\\ &= \int_{z^*=0}^{z}\nu(\mathrm{d}z^*) - \int_{z^*=0}^{z}\frac{z^*}{z}\nu(\mathrm{d}z^*) = \int_{z^*=0}^{z}\left(\frac{z-z^*}{z}\right)\nu(\mathrm{d}z^*).\end{aligned}$$

Returning to our calculation regarding the first term in (A.19),

$$\begin{aligned}&\mathbb{E}_{\bar{q}}\left[r(\hat{Z}_N(\mathbf{X}^N))^{t\wedge p}\mathbb{1}\left(1-\alpha \leq \hat{Z}_N(\mathbf{X}^N) \leq 2\right)\right]\\ &= \mathbb{E}_{\bar{q}}\left[\left(\frac{1}{\hat{Z}_N(\mathbf{X}^N)}\int_{z^*=0}^{\hat{Z}_N(\mathbf{X}^N)}(\hat{Z}_N(\mathbf{X}^N) - z^*)\nu(\mathrm{d}z^*)\right)^{t\wedge p}\mathbb{1}\left(1-\alpha \leq \hat{Z}_N(\mathbf{X}^N) \leq 2\right)\right]\\ &\leq \frac{\nu\{[0, \hat{Z}_N(\mathbf{X}^N)]\}^{t\wedge p}}{(1-\alpha)^{t\wedge p}}\end{aligned}$$

$$
\begin{aligned}
&\times \mathbb{E}_{\bar{q}}\left[\left(\int_{z^*=0}^{\hat{Z}_N(\mathbf{X}^N)} (\hat{Z}_N(\mathbf{X}^N) - z^*)\frac{\nu(\mathrm{d}z^*)}{\nu\{[0, \hat{Z}_N(\mathbf{X}^N)]\}}\right)^{t\wedge p} \mathbb{1}\left(1-\alpha \leq \hat{Z}_N(\mathbf{X}^N) \leq 2\right)\right] \\
&\leq \frac{\nu\{[0, \hat{Z}_N(\mathbf{X}^N)]\}^{t\wedge p-1}}{(1-\alpha)^{t\wedge p}} \\
&\times \mathbb{E}_{\bar{q}}\left[\int_{z^*=0}^{\hat{Z}_N(\mathbf{X}^N)} (\hat{Z}_N(\mathbf{X}^N) - z^*)^{t\wedge p}\nu(\mathrm{d}z^*) \cdot \mathbb{1}\left(1-\alpha \leq \hat{Z}_N(\mathbf{X}^N) \leq 2\right)\right] \\
&\leq \frac{1}{(1-\alpha)^{t\wedge p}}\mathbb{E}_{\bar{q}}\left[\int_{z^*=0}^{\hat{Z}_N(\mathbf{X}^N)} |\hat{Z}_N(\mathbf{X}^N) - z^*|^{t\wedge p}\nu(\mathrm{d}z^*)\right] \\
&\leq \frac{1}{(1-\alpha)^{t\wedge p}}\mathbb{E}_{(\mathbf{X}^N,\mathbf{X}^{N,\prime})\sim\bar{q}\otimes\bar{q}}\left[|\hat{Z}_N(\mathbf{X}^N) - \hat{Z}_N(\mathbf{X}^{N,\prime})|^{t\wedge p}\right] \leq \frac{1}{(1-\alpha)^{t\wedge p}}\frac{A_p q(\omega^p)}{N^{\frac{t\wedge p}{2}}},
\end{aligned}
$$

for a constant A_p depending only on p. The first inequality comes from $\hat{Z}_N(\mathbf{X}^N)^{-1} \leq (1-\alpha)^{-1}$ on the event of interest, the second inequality is from Jensen's since the function $u \mapsto u^{t\wedge p}$ is convex, the third inequality is from $\bar{q}(A) \leq 1$ and the indicator being smaller than one, the fourth is obtained by completing the integral over all $z^* \in (0, \infty)$, and the last is from a reasoning similar to the proof of Proposition 2.5, or by direct application of Minkowski's inequality and Proposition 2.5.

Overall, choosing $\alpha = 1/2$ we have that

$$
\begin{aligned}
\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^{t\wedge p}\mathbb{1}\left(\hat{Z}_N(\mathbf{X}^N) \leq 2\right)\right] &\leq 2^{t\wedge p}A_p q(\omega^p)N^{-t\wedge p/2} \\
&\quad + r(2)^{t\wedge p}\nu\left\{|\hat{Z}_N(\mathbf{X}^N) - 1| \geq \alpha\right\} \\
&\leq 2^{t\wedge p}A_p q(\omega^p)N^{-t\wedge p/2} + \beta_p^{t\wedge p}C_p N^{-p/2}2^p,
\end{aligned}
$$

using Markov's inequality as in Proposition 2.5. Finally, multiplying by $r(2)^{t-t\wedge p}$ we obtain

$$
\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t\mathbb{1}\left(\hat{Z}_N(\mathbf{X}^N) \leq 2\right)\right] \leq \frac{\beta_p^t\beta_p^{-t\wedge p}2^{t\wedge p}A_p q(\omega^p)}{N^{\frac{t\wedge p}{2}}} + \frac{2^p C_p \beta_p^t}{N^{p/2}},
$$

and we note that, since $\beta_p \leq 1$, we have $\beta_p^{-t\wedge p}2^{t\wedge p} \leq \beta_p^{-p}2^p$, and thus we can define A_p and B_p to obtain Lemma A.7. □

Proof of Proposition 2.13. We combine the bounds from Lemmas A.6 and A.7, and note that the two terms in the bound of Lemma A.7 can be bounded by $A_p\beta_p^t N^{-(t\wedge p)/2}$ for some constant A_p, which is not the same A_p as in the statement of Lemma A.7. □

A.4.3 Proofs of Theorem 2.16 and Corollary 2.18

Proof of Theorem 2.16. Under Assumption 2.1, the PIMH chain is $\bar{\pi}$-irreducible, and by construction it is aperiodic and $\bar{\pi}$-invariant, therefore $|\bar{q}P^t - \bar{\pi}|_{\mathrm{TV}} \to 0$ as $t \to \infty$ (Theorem 4 in

Roberts & Rosenthal, 2004). Thus, for any $t \geq 0$, by the triangle inequality,

$$|\bar{q}P^t - \bar{\pi}|_{\text{TV}} \leq \sum_{j=1}^{\infty} |\bar{q}P^{t+j} - \bar{q}P^{t+j-1}|_{\text{TV}}.$$

By the coupling representation of the TV distance, for any $t \geq 0$, $j \geq 1$,

$$|\bar{q}P^{t+j} - \bar{q}P^{t+j-1}|_{\text{TV}} \leq \mathbb{E}[\mathbb{1}(\mathbf{X}_{t+j}^N \neq \mathbf{Y}_{t+j-1}^N)] = \mathbb{P}(\tau > t + j),$$

where $(\mathbf{X}_t^N)$ and $(\mathbf{Y}_t^N)$ are jointly generated by Algorithm 5. Under Assumption 2.2, Proposition 2.14 applies and thus the series $\sum_{j=1}^{\infty} \mathbb{P}(\tau > t + j)$ converges. Thus, by the dominated convergence theorem we may swap expectation and limit to write

$$|\bar{q}P^t - \bar{\pi}|_{\text{TV}} \leq \mathbb{E}\left[\sum_{j=1}^{\infty} \mathbb{1}(\mathbf{X}_{t+j}^N \neq \mathbf{Y}_{t+j-1}^N)\right] = \mathbb{E}\left[\max(0, \tau - t - 1)\right],$$

for all $t \geq 0$. This is (2.29).

We may express the expectation of a non-negative variable as a series of survival probabilities:

$$\mathbb{E}\left[\max\left(0, \tau - t - 1\right)\right] = \sum_{s=1}^{\infty} \mathbb{P}\left(\max\left(0, \tau - t - 1\right) \geq s\right).$$

For any $t \geq 0$, $s \geq 1$, $\max(0, \tau - 1 - t) \geq s$ if and only if $\tau > s + t$. Under Assumption 2.2, Proposition 2.14 obtains

$$\mathbb{P}\left(\tau > s + t\right) \leq CN^{-1/2}(s + t)^{-p}.$$

The series $\sum_{s=1}^{\infty}(s + t)^{-p}$ can be bounded as follows:

$$\begin{aligned}\sum_{s=1}^{\infty}(s + t)^{-p} &\leq (1 + t)^{-p} + \int_{1+t}^{\infty} x^{-p}dx \\ &= (1 + t)^{-p+1}\left(\frac{(1 + t/p)p}{(1 + t)(p - 1)}\right) \\ &\leq \frac{p}{(p - 1)(1 + t)^{p-1}},\end{aligned} \tag{A.20}$$

using the fact that $f(k) \leq \int_{k-1}^{k} f(x)dx$ for any decreasing function f. Thus, for $t \geq 0$,

$$|\bar{q}P^t - \bar{\pi}|_{\text{TV}} \leq \frac{Cp}{\sqrt{N}(p - 1)(1 + t)^{p-1}},$$

which completes the proof. □

Proof of Corollary 2.18. The proof starts with multiple applications of the triangle inequality, Theorem 2.11, $\max(a,b) \le a+b$ for $a,b \ge 0$:

$$\begin{aligned}
\left|P^t(\mathbf{X}^N,\cdot) - \bar{\pi}\right|_{\mathrm{TV}} &\le \left|P^t(\mathbf{X}^N,\cdot) - \bar{q}P^t\right|_{\mathrm{TV}} + \left|\bar{q}P^t - \bar{\pi}\right|_{\mathrm{TV}} \\
&\le \int \left|P^t(\mathbf{X}^N,\cdot) - P^t(\mathbf{y}^N,\cdot)\right|_{\mathrm{TV}} \bar{q}(\mathrm{d}\mathbf{y}^N) + \left|\bar{q}P^t - \bar{\pi}\right|_{\mathrm{TV}} \\
&= \int \max\left(r(\mathbf{X}^N), r(\mathbf{y}^N)\right)^t \bar{q}(\mathrm{d}\mathbf{y}^N) + \left|\bar{q}P^t - \bar{\pi}\right|_{\mathrm{TV}} \\
&\le \left(r(\mathbf{X}^N)\right)^t + \mathbb{E}_{\bar{q}}[\left(r(\mathbf{Y}^N)\right)^t] + \left|\bar{q}P^t - \bar{\pi}\right|_{\mathrm{TV}}.
\end{aligned}$$

The result then follows from Proposition 2.13 and Theorem 2.16. □

A.4.4 Proofs of Theorem 2.20, Corollary 2.21 and Corollary 2.22

We first treat PIMH in the regime $p \in (1,2)$, and then prove the IMH specialization in Corollary 2.22.

Lemma A.8 (von Bahr–Esseen inequality (von Bahr & Esseen, 1965))**.** *Let $X_1, \dots, X_N$ be independent random variables with $\mathbb{E}[X_i] = 0$, and let $p \in [1,2]$. Then*

$$\mathbb{E}\left[\left|\sum_{n=1}^N X_n\right|^p\right] \le 2\sum_{n=1}^N \mathbb{E}\left[|X_n|^p\right].$$

Alternatively we can use the Marcinkiewicz–Zygmund inequality, followed by the subadditivity of the function $x \mapsto x^{p/2}$ for $p \in (1,2)$, to obtain a similar bound with a multiplicative constant that depends on p.

Proposition A.9. *Assume Assumption 2.1, and assume that $q(\omega^p) < \infty$ for some $p \in (1,2)$. Then, for all $N \ge 1$,*

$$\mathbb{E}_{\bar{q}}\left[\left|\hat{Z}_N(\mathbf{X}^N) - 1\right|^p\right] \le \frac{2q\left(|\omega-1|^p\right)}{N^{p-1}}, \tag{A.21}$$

$$\mathbb{P}_{\bar{q}}\left(\hat{Z}_N(\mathbf{X}^N) \ge 1+t\right) \le \frac{2q\left(|\omega-1|^p\right)}{N^{p-1}t^p}, \qquad t > 0. \tag{A.22}$$

If $\hat{Z}_N(\mathbf{X}^{N,\prime})$ is an independent copy of $\hat{Z}_N(\mathbf{X}^N)$ under $\bar{q}$, then

$$\mathbb{E}_{\bar{q}\otimes\bar{q}}\left[\left|\hat{Z}_N(\mathbf{X}^N) - \hat{Z}_N(\mathbf{X}^{N,\prime})\right|^p\right] \le \frac{2^{p+1}q(\omega^p)}{N^{p-1}}, \tag{A.23}$$

$$\mathbb{E}_{\bar{q}\otimes\bar{q}}\left[\left|\hat{Z}_N(\mathbf{X}^N) - \hat{Z}_N(\mathbf{X}^{N,\prime})\right|\right] \le 2^{(p+1)/p} q(\omega^p)^{1/p} N^{-(p-1)/p}. \tag{A.24}$$

Proof. Let $\Omega_1, \dots, \Omega_N$ be i.i.d. with law q, and write $\hat{Z}_N = N^{-1}\sum_{n=1}^N \Omega_n$. Set $X_n := \Omega_n - 1$,

so that $\mathbb{E}[X_n] = 0$. Lemma A.8 gives

$$\mathbb{E}\left[\left|\sum_{n=1}^{N}(\Omega_n - 1)\right|^p\right] \leq 2\sum_{n=1}^{N}\mathbb{E}\left[|\Omega_n - 1|^p\right] = 2Nq\left(|\omega - 1|^p\right).$$

Dividing by N^p yields (A.21). Markov's inequality then gives (A.22).

For (A.23), let $\Omega'_1, \ldots, \Omega'_N$ be an independent copy and set $\hat{Z}'_N = N^{-1}\sum_{n=1}^{N}\Omega'_n$. Then

$$\hat{Z}_N - \hat{Z}'_N = \frac{1}{N}\sum_{n=1}^{N}(\Omega_n - \Omega'_n),$$

and the summands are independent and mean-zero. Applying Lemma A.8 again,

$$\mathbb{E}\left[|\hat{Z}_N - \hat{Z}'_N|^p\right] \leq \frac{2}{N^{p-1}}\mathbb{E}\left[|\Omega - \Omega'|^p\right].$$

Since $|a-b|^p \leq 2^{p-1}(a^p + b^p)$ for $a, b \geq 0$, we obtain $\mathbb{E}[|\Omega - \Omega'|^p] \leq 2^p q(\omega^p)$, which gives (A.23). Finally, (A.24) follows from Lyapunov's inequality. □

The following is an adaptation of Lemma A.6 in the case $p \in (1, 2)$.

Lemma A.10. *Under Assumptions 2.1-2.2 with $p \in (1, 2)$, there exists a constant $B > 0$ such that for all $t \geq 1$ and $N \geq 1$,*

$$\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) > 2)\right] \leq \frac{B}{N^{p-1}t^p}.$$

Proof. We split the expectation into two parts:

$$\begin{aligned}\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) > 2)\right] &= \mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) \in (2, 1+t))\right] \\ &\quad + \mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) \geq 1+t)\right].\end{aligned}$$

For $\{\hat{Z}_N(\mathbf{X}^N) \geq 1 + t\}$, we directly apply Proposition A.9:

$$\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) \geq 1+t)\right] \leq \mathbb{P}_{\bar{q}}\left(\hat{Z}_N(\mathbf{X}^N) \geq 1+t\right) \leq \frac{2q(|\omega-1|^p)}{N^{p-1}t^p}.$$

For $\{\hat{Z}_N(\mathbf{X}^N) \in (2, 1+t)\}$, we use Lemma A.5 with $\theta = 1/2$:

$$r(\mathbf{x}^N) \leq 1 - \frac{c_p(1/2)}{2\hat{Z}_N(\mathbf{x}^N)} \leq \exp\left(-\frac{c_p(1/2)}{2\hat{Z}_N(\mathbf{x}^N)}\right).$$

Let $c = c_p(1/2)/4$. Since $\hat{Z}_N(\mathbf{x}^N) > 2$ implies $\hat{Z}_N(\mathbf{x}^N) \leq 2(\hat{Z}_N(\mathbf{x}^N) - 1)$, we have

$$r(\mathbf{x}^N)^t \leq \exp\left(-\frac{ct}{\hat{Z}_N(\mathbf{x}^N) - 1}\right).$$

Introduce the sets $A_k = [t/(k+1), t/k]$ for $k \geq 1$, so that $\cup_{k=1}^{\infty} A_k = (0, t]$ and in particular covers $(1, t)$. Then

$$\begin{aligned}
&\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) \in (2, 1+t))\right] \\
&\leq \mathbb{E}_{\bar{q}}\left[\exp\left(-\frac{ct}{\hat{Z}_N(\mathbf{X}^N) - 1}\right) \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) - 1 \in (1, t))\right] \\
&\leq \sum_{k=1}^{\infty} \exp(-ck)\, \mathbb{P}_{\bar{q}}\left(\hat{Z}_N(\mathbf{X}^N) \geq 1 + \frac{t}{k+1}\right) \\
&\leq \sum_{k=1}^{\infty} \exp(-ck) \frac{2q(|\omega - 1|^p)}{N^{p-1}} \left(\frac{k+1}{t}\right)^p,
\end{aligned}$$

where the last inequality uses Proposition A.9. Since $\sum_{k=1}^{\infty} e^{-ck}(k+1)^p < \infty$, this term is also bounded by a constant multiple of $(N^{p-1}t^p)^{-1}$. Combining both parts completes the proof. □

The next Lemma is similar to Lemma A.7 but adapted for the case $p \in (1, 2)$.

Lemma A.11. *Under Assumptions 2.1-2.2 with $p \in (1, 2)$, define*

$$\beta_p := 1 - \frac{1}{2^{\frac{3p-2}{p-1}} q(\omega^p)^{\frac{1}{p-1}}}. \tag{A.25}$$

Then $\beta_p \in (0, 1)$, and there exists a constant $A > 0$ such that for all $t \geq 1$ and $N \geq 1$,

$$\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) \leq 2)\right] \leq \frac{A}{N^{(p-1)/p}} \beta_p^t.$$

Proof. We abuse notation slightly and write $r(z)$ for the rejection probability as a function of the value $z = \hat{Z}_N(\mathbf{X}^N)$. Since $r(z)$ is non-decreasing in z, on $\{\hat{Z}_N(\mathbf{X}^N) \leq 2\}$ we have

$$r(\mathbf{X}^N)^t \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) \leq 2) \leq r(2)^{t-1} r(\mathbf{X}^N) \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) \leq 2).$$

By Lemma A.5 with $\theta = 1/2$, we have

$$r(2) \leq 1 - \frac{c_p(1/2)}{4} = \beta_p.$$

It remains to bound $\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N) \mathbb{1}(\hat{Z}_N(\mathbf{X}^N) \leq 2)\right]$.

Let $\hat{Z}_N(\mathbf{X}^{N,\prime})$ be an independent copy of $\hat{Z}_N(\mathbf{X}^N)$ under $\bar{q}$. Writing ν for the law of $\hat{Z}_N(\mathbf{X}^{N,\prime})$, for any $z > 0$,

$$r(z) = 1 - \int \min\left\{1, \frac{z^*}{z}\right\} \nu(\mathrm{d}z^*) = \int_0^z \frac{z - z^*}{z} \nu(\mathrm{d}z^*).$$

Therefore,

$$r(\hat{Z}_N(\mathbf{X}^N)) = \frac{1}{\hat{Z}_N(\mathbf{X}^N)} \mathbb{E}_{\bar{q}}\left[\left(\hat{Z}_N(\mathbf{X}^N) - \hat{Z}_N(\mathbf{X}^{N,\prime})\right)_+ \middle| \hat{Z}_N(\mathbf{X}^N)\right].$$

We split the event $\{\hat{Z}_N(\mathbf{X}^N) \le 2\}$ into $\{\hat{Z}_N(\mathbf{X}^N) < 1/2\} \cup \{1/2 \le \hat{Z}_N(\mathbf{X}^N) \le 2\}$.

On $\{1/2 \le \hat{Z}_N(\mathbf{X}^N) \le 2\}$ we have $1/\hat{Z}_N(\mathbf{X}^N) \le 2$, hence

$$r(\hat{Z}_N(\mathbf{X}^N))\mathbb{1}(1/2 \le \hat{Z}_N(\mathbf{X}^N) \le 2) \le 2\mathbb{E}_{\bar{q}}\left[\left|\hat{Z}_N(\mathbf{X}^N) - \hat{Z}_N(\mathbf{X}^{N,\prime})\right| \middle| \hat{Z}_N(\mathbf{X}^N)\right].$$

Taking expectations and using Proposition A.9,

$$\begin{aligned}\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)\mathbb{1}(1/2 \le \hat{Z}_N(\mathbf{X}^N) \le 2)\right] &\le 2\mathbb{E}_{\bar{q}\otimes\bar{q}}\left[\left|\hat{Z}_N(\mathbf{X}^N) - \hat{Z}_N(\mathbf{X}^{N,\prime})\right|\right] \\ &\le 4 \cdot 2^{1/p} \cdot q(\omega^p)^{1/p} N^{-(p-1)/p}.\end{aligned}$$

On $\{\hat{Z}_N(\mathbf{X}^N) < 1/2\}$, we use the crude bound $r \le 1$ and Proposition A.9:

$$\begin{aligned}\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)\mathbb{1}(\hat{Z}_N(\mathbf{X}^N) < 1/2)\right] &\le \mathbb{P}_{\bar{q}}\left(\hat{Z}_N(\mathbf{X}^N) < 1/2\right) \le \mathbb{P}_{\bar{q}}\left(|\hat{Z}_N(\mathbf{X}^N) - 1| \ge 1/2\right) \\ &\le \frac{2^{p+1} q(|\omega - 1|^p)}{N^{p-1}}.\end{aligned}$$

Since $p - 1 > (p-1)/p$ and $N \ge 1$, the last term is $O(N^{-(p-1)/p})$ and can be absorbed into the previous bound. Thus there exists $K > 0$, depending only on p and $q(\omega^p)$, such that

$$\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)\mathbb{1}(\hat{Z}_N(\mathbf{X}^N) \le 2)\right] \le \frac{K}{N^{(p-1)/p}}.$$

Combining this with $r(2) \le \beta_p$ yields

$$\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t\mathbb{1}(\hat{Z}_N(\mathbf{X}^N) \le 2)\right] \le \frac{K}{\beta_p} N^{-(p-1)/p} \beta_p^t,$$

which proves the claim. □

The following result is analogous to Proposition 2.13 but adapted for the case $p \in (1, 2)$.

Proposition A.12. *Under Assumptions 2.1-2.2 with $p \in (1, 2)$, there exists a constant $C > 0$*

such that for all $N \geq 1$ *and all* $t \geq 1$,

$$\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t\right] \leq \frac{C}{N^{(p-1)/p} t^p}.$$

Proof. We combine the bounds from Lemmas A.10 and A.11 and get

$$\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t\right] \leq \frac{A}{N^{(p-1)/p}}\beta_p^t + \frac{B}{N^{p-1} t^p}.$$

Since β_p defined in (A.25) is in $(0,1)$, the quantity $\sup_{t\geq 1} t^p \beta_p^t$ is finite; call it $K_{\beta,p}$. Hence $\beta_p^t \leq K_{\beta,p}/t^p$ for all $t \geq 1$. Also $N^{-(p-1)} \leq N^{-(p-1)/p}$ for $N \geq 1$ because $p-1 > (p-1)/p$. Therefore we can pick $C = AK_{\beta,p} + B$ and conclude

$$\mathbb{E}_{\bar{q}}\left[r(\mathbf{X}^N)^t\right] \leq \frac{C}{N^{(p-1)/p} t^p}.$$

□

The next result is analogous to Proposition 2.14 but adapted for the case $p \in (1,2)$.

Proposition A.13. *Consider* τ *generated by Algorithm 5. Under Assumptions 2.1-2.2 with* $p \in (1,2)$, *there exists a finite* $C > 0$ *such that for all* $N \geq 1$ *and all* $t \geq 1$,

$$\mathbb{P}(\tau > t) \leq \frac{C}{N^{(p-1)/p} t^p}.$$

Proof. Combining Proposition 2.12 and Proposition A.12 immediately yields the claim. □

Proof of Theorem 2.20. Under Assumption 2.1, the PIMH chain is $\bar{\pi}$-irreducible, and by construction it is aperiodic and $\bar{\pi}$-invariant, therefore $|\bar{q}P^t - \bar{\pi}|_{\mathrm{TV}} \to 0$ as $t \to \infty$ (Theorem 4 in Roberts & Rosenthal, 2004). Thus, for any $t \geq 0$, by the triangle inequality,

$$|\bar{q}P^t - \bar{\pi}|_{\mathrm{TV}} \leq \sum_{j=1}^{\infty} |\bar{q}P^{t+j} - \bar{q}P^{t+j-1}|_{\mathrm{TV}}.$$

By the coupling representation of the total variation distance, for any $t \geq 0$ and $j \geq 1$,

$$|\bar{q}P^{t+j} - \bar{q}P^{t+j-1}|_{\mathrm{TV}} \leq \mathbb{E}\left[\mathbb{1}(\mathbf{X}_{t+j}^N \neq \mathbf{Y}_{t+j-1}^N)\right] = \mathbb{P}(\tau > t+j),$$

where $(\mathbf{X}_t^N)$ and $(\mathbf{Y}_t^N)$ are jointly generated by Algorithm 5. Under the present assumptions, Proposition A.13 applies, and therefore

$$|\bar{q}P^t - \bar{\pi}|_{\mathrm{TV}} \leq \sum_{j=1}^{\infty} \mathbb{P}(\tau > t+j) \leq \frac{C}{N^{(p-1)/p}} \sum_{j=1}^{\infty} (t+j)^{-p}.$$

Finally, as in (A.20) in the proof of Theorem 2.16, we can bound

$$\sum_{j=1}^{\infty}(t+j)^{-p} \le \frac{p}{(p-1)(1+t)^{p-1}},$$

which gives the result. □

Proof of Corollary 2.21. Similarly to the proof of Corollary 2.18, we start from

$$\begin{aligned}
\left|P^t(\mathbf{X}^N,\cdot)-\bar{\pi}\right|_{\mathrm{TV}} &\le \left|P^t(\mathbf{X}^N,\cdot)-\bar{q}P^t\right|_{\mathrm{TV}} + \left|\bar{q}P^t-\bar{\pi}\right|_{\mathrm{TV}} \\
&\le \int \left|P^t(\mathbf{X}^N,\cdot)-P^t(\mathbf{Y}^N,\cdot)\right| \bar{q}(\mathrm{d}\mathbf{Y}^N) + \left|\bar{q}P^t-\bar{\pi}\right|_{\mathrm{TV}} \\
&= \int \max\left(r(\mathbf{X}^N), r(\mathbf{Y}^N)\right)^t \bar{q}(\mathrm{d}\mathbf{Y}^N) + \left|\bar{q}P^t-\bar{\pi}\right|_{\mathrm{TV}} \\
&\le \left(r(\mathbf{X}^N)\right)^t + \mathbb{E}_{\bar{q}}\left[\left(r(\mathbf{Y}^N)\right)^t\right] + \left|\bar{q}P^t-\bar{\pi}\right|_{\mathrm{TV}}.
\end{aligned}$$

The result then follows from Proposition A.12 and Theorem 2.20. □

For $N = 1$ the rejection probability can be controlled directly, without the von Bahr–Esseen inequality, which gives the shorter argument below.

Proof of Corollary 2.22. Similarly to the proof of Corollary 2.18, we start from

$$|P^t(x,\cdot)-\pi|_{\mathrm{TV}} = |P^t(x,\cdot)-\pi P^t|_{\mathrm{TV}} \le r(x)^t + \mathbb{E}_{Y\sim\pi}[r(Y)^t].$$

Let $p > 1$. We can apply Lemma A.5 to obtain

$$r(Y) \le 1 - \min\left\{1, \frac{\theta}{\omega(Y)}\right\}\frac{(1-\theta)^{p/(p-1)}}{q(\omega^p)^{1/(p-1)}},$$

and we set $\theta = 1/2$, and $c = (1-\theta)^{p/(p-1)}/q(\omega^p)^{1/(p-1)}$. Note that $c \le 1$ as $q(\omega^p) \ge q(\omega)^p = 1$. We next bound the expected rejection probability as follows

$$\begin{aligned}
\mathbb{E}_{Y\sim\pi}\left[r(Y)^t\right] &\le \mathbb{E}_{Y\sim\pi}\left[\left(1-\min\left\{\frac{1}{2\omega(Y)}, 1\right\}c\right)^t\right] \\
&\le \mathbb{E}_{Y\sim\pi}\left[\left(1-\frac{c}{2}\right)^t \mathbb{1}(\omega(Y)\le 1)\right] \\
&\quad + \mathbb{E}_{Y\sim\pi}\left[\left(1-\frac{c}{2\omega(Y)}\right)^t \mathbb{1}(\omega(Y)\in[1,t])\right] + \mathbb{P}\left(\omega(Y)\ge t\right) \\
&\le \left(1-\frac{c}{2}\right)^t + \mathbb{E}_{Y\sim\pi}\left[\left(1-\frac{c}{2\omega(Y)}\right)^t \mathbb{1}(\omega(Y)\in[1,t])\right] + \frac{\tilde{C}}{t^{p-1}} \\
&\le \left(1-\frac{c}{2}\right)^t + \mathbb{E}_{Y\sim\pi}\left[\exp\{-Ct/\omega(Y)\}\mathbb{1}(\omega(Y)\in[1,t])\right] + \frac{\tilde{C}}{t^{p-1}}.
\end{aligned}$$

The second inequality follows by splitting the weight into $\omega \leq 1, \omega \in [1,t]$ and $\omega > t$. The third inequality employs Markov's inequality and the assumption that $p > 1$. The last inequality uses $\log(1+x) \leq x$ with $x = -c/(2\omega(Y))$ and $C = c/2$. Consider the three terms on the last line. The first term decays exponentially fast with t, the third term decays at the rate of $t^{-(p-1)}$. It remains to bound the second term.

Define $A_k := [t/(k+1), t/k]$, then clearly $\cup_{k=1}^{\infty} A_k = (0,t]$. We bound the second term as follows:

$$\begin{aligned}
\mathbb{E}_{Y\sim\pi}\left[\exp\{-Ct/\omega(Y)\}\mathbb{1}(\omega(Y)\in[1,t])\right] &\leq \mathbb{E}_{Y\sim\pi}\left[\exp\{-Ct/\omega(Y)\}\mathbb{1}(\omega(Y)\in[0,t])\right] \\
&= \sum_{k=1}^{\infty} \mathbb{E}_{Y\sim\pi}\left[\exp\{-Ct/\omega(Y)\}\mathbb{1}(\omega(Y)\in A_k)\right] \\
&\leq \sum_{k=1}^{\infty} \exp\{-Ct/(t/k)\}\mathbb{P}(\omega(Y)\geq t/(k+1)) \\
&\leq \sum_{k=1}^{\infty} \exp\{-Ck\}\frac{C'(k+1)^{p-1}}{t^{p-1}} \\
&= \frac{C''}{t^{p-1}}\sum_{k=1}^{\infty}\exp\{-Ck\}(k+1)^{p-1} \\
&\leq \frac{C'''}{t^{p-1}}.
\end{aligned}$$

The last inequality holds as $\sum_{k=1}^{\infty}\exp\{-Ck\}(k+1)^{p-1} < \infty$ (the terms inside the summation decay exponentially fast). This concludes the proof. □

A.4.5 Proof of the result in Example 2.23

Given

$$\pi(x) := Z_\pi x^{-p}, \qquad q(x) := Z_q \log^k(x) x^{-(p+1)}, \qquad x \in [2,\infty),$$

with $p > 1$ and $k > 1/(p-1)$. We have

$$\omega(x) = \frac{\pi(x)}{q(x)} = \frac{Z_\pi}{Z_q}\frac{x}{\log^k(x)}.$$

We first verify that $q(\omega^p) < \infty$. Since

$$q(\omega^p) = Z_q\left(\frac{Z_\pi}{Z_q}\right)^p \int_2^\infty \frac{x^p}{\log^{kp}(x)}\frac{\log^k(x)}{x^{p+1}}\mathrm{d}x = Z_q\left(\frac{Z_\pi}{Z_q}\right)^p \int_2^\infty \frac{1}{x\log^{k(p-1)}(x)}\mathrm{d}x,$$

the change of variable $u = \log x$ gives

$$q(\omega^p) = Z_q\left(\frac{Z_\pi}{Z_q}\right)^p \int_{\log 2}^\infty u^{-k(p-1)}\mathrm{d}u < \infty,$$

since $k(p-1) > 1$.

Now we estimate $\mathbb{P}_{X\sim\pi}(X > s)$ and $\mathbb{P}_{X\sim q}(X > s)$ for any $s > 2$. For the former,

$$\mathbb{P}_{X\sim\pi}(X > s) = Z_\pi \int_s^\infty x^{-p}\mathrm{d}x = \frac{C_1}{s^{p-1}}, \qquad C_1 := \frac{Z_\pi}{p-1}.$$

For the latter,

$$\begin{aligned}\mathbb{P}_{X\sim q}(X > s) &= Z_q \int_s^\infty \frac{\log^k(x)}{x^{p+1}}\mathrm{d}x = Z_q s^{-p}\int_1^\infty \log^k(su)u^{-(p+1)}\mathrm{d}u \\ &\le Z_q 2^{k-1}s^{-p}\int_1^\infty \frac{\log^k(s)+\log^k(u)}{u^{p+1}}\mathrm{d}u \\ &\le Z_q 2^{k-1}s^{-p}\left(\frac{\log^k(s)}{p} + \kappa_{p,k}\right) \le \frac{C_2\log^k(s)}{s^p},\end{aligned}$$

where

$$\kappa_{p,k} := \int_1^\infty \log^k(u)u^{-(p+1)}\mathrm{d}u < \infty, \qquad C_2 := Z_q 2^{k-1}\left(\frac{1}{p} + \frac{\kappa_{p,k}}{\log^k(2)}\right),$$

and we used $\log^k(s) \ge \log^k(2)$ for all $s \ge 2$.

Consider an IMH chain $(X_t)_{t\ge 0}$ targeting π with proposal q, started from $x_0 = 3$. Fix any $t \ge 2$, define

$$s_t := t(\log t)^{k+1}, \qquad A_t := (s_t, \infty).$$

Then the probability of A_t under π is

$$\mathbb{P}_{X\sim\pi}(X \in A_t) = \frac{C_1}{t^{p-1}(\log t)^{(k+1)(p-1)}}.$$

On the other hand, $X_t \in A_t$ implies that at least one of $x_0, Y_1, \ldots, Y_t$ falls in A_t, where $(Y_j)_{j\ge 1}$ are the proposals. For $t \ge 3$, we have $x_0 \notin A_t$, thus, by the union bound,

$$\mathbb{P}(X_t \in A_t) \le t\mathbb{P}_{Y\sim q}(Y \in A_t) \le t\frac{C_2\log^k(s_t)}{s_t^p}.$$

Let $t_1 = t_1(k)$ be large enough that $t_1 \ge 3$ and $(k+1)\log\log t \le \log t$ for all $t \ge t_1$. Then, for $t \ge t_1$,

$$\log(s_t) = \log t + (k+1)\log\log t \le 2\log t,$$

so that $\log^k(s_t) \le 2^k(\log t)^k$. Therefore, for all $t \ge t_1$,

$$\mathbb{P}(X_t \in A_t) \le t \cdot \frac{C_2 2^k(\log t)^k}{t^p(\log t)^{p(k+1)}} = \frac{2^k C_2}{t^{p-1}(\log t)^{p(k+1)-k}}.$$

Since

$$p(k+1) - k - (k+1)(p-1) = 1,$$

we have

$$\mathbb{P}(X_t \in A_t) \leq \frac{2^k C_2}{C_1 \log t} \mathbb{P}_{X\sim\pi}(X \in A_t).$$

Define

$$t_0 := \left\lceil \exp\left(\frac{2^{k+1}C_2}{C_1}\right)\right\rceil \vee t_1. \tag{A.26}$$

Then for all $t \geq t_0$,

$$\mathbb{P}(X_t \in A_t) \leq \frac{1}{2}\mathbb{P}_{X\sim\pi}(X \in A_t).$$

Therefore, for all $t \geq t_0$,

$$\begin{aligned}\left|P^t(x_0,\cdot) - \pi\right|_{\mathrm{TV}} &\geq \mathbb{P}_{X\sim\pi}(X \in A_t) - \mathbb{P}(X_t \in A_t) \\ &\geq \frac{1}{2}\mathbb{P}_{X\sim\pi}(X \in A_t) = \frac{C_1/2}{t^{p-1}(\log t)^{(k+1)(p-1)}}.\end{aligned}$$

This proves the claim.

A.5 Proofs of Section 2.5

A.5.1 Proof of Proposition 2.26

Proof. Note that $\hat{F}_N^u$ is not bounded even if $|f|_\infty \leq 1$, because the sum in (2.39) can be arbitrarily large. By Minkowski's inequality, for any $s \geq 1$,

$$\mathbb{E}\left[|\hat{F}_N^u|^s\right]^{1/s} \leq \mathbb{E}\left[|\hat{F}(\mathbf{x}_0)|^s\right]^{1/s} + \mathbb{E}\left[\left|\sum_{t=1}^{\tau-1}\{\hat{F}(\mathbf{x}_t) - \hat{F}(\mathbf{y}_{t-1})\}\right|^s\right]^{1/s}.$$

Furthermore, if $|f|_\infty \leq 1$ then $|\hat{F}(\mathbf{x})| \leq 1$ for all $\mathbf{x}$, thus

$$\mathbb{E}\left[|\hat{F}_N^u|^s\right]^{1/s} \leq \mathbb{E}\left[|\hat{F}(\mathbf{x}_0)|^s\right]^{1/s} + 2\mathbb{E}\left[\mathbb{1}(\tau > 1)\,|\tau - 1|^s\right]^{1/s}.$$

Since $\hat{F}(\mathbf{x}_0) \leq 1$ almost surely, $\mathbb{E}[|\hat{F}(\mathbf{x}_0)|^s]^{1/s}$ is finite for all $s \geq 1$. The latter expectation is smaller than $\mathbb{E}[|\tau|^s]^{1/s}$. Thus, $\hat{F}_N^u$ has s finite moments if τ has s finite moments. Note that $\hat{F}_N^u$ can have higher moments as well: for example, if f is constant, then $\hat{F}_N^u$ is constant.

Next, in order for τ to have $s \geq 1$ moments, we can resort to Proposition 2.14. If Assumption 2.2 holds with $p > s$, then $\mathbb{P}(\tau > t) \leq CN^{-1/2}t^{-p}$. We can then follow the proof of

Proposition 8 in Douc et al. (2026), using Tonelli's theorem:

$$\mathbb{E}\left[\tau^s\right] = \mathbb{E}\left[\int_0^\infty \mathbb{1}(u<\tau) s u^{s-1} \mathrm{d}u\right] = \int_0^\infty s u^{s-1} \mathbb{P}(\tau > u) \mathrm{d}u$$
$$= \sum_{i=0}^{\infty} \mathbb{P}(\tau > i) \int_i^{i+1} s u^{s-1} \mathrm{d}u \;\leq\; \sum_{i=0}^{\infty} \mathbb{P}(\tau > i) s (i+1)^{s-1}.$$

The sum is finite under the assumption $p > s$. □

A.5.2 Proof of Proposition 2.27

Proof of Proposition 2.27. We now consider the PIMH chain $(\mathbf{X}_t^N)_{t\geq 0}$, started from $\bar{q}$. The case $t = 0$ corresponds to Theorem 2.9. Let $t \geq 1$. We can assume that f is non-negative, using the same separate treatment of f_+ and f_- as in the beginning of the proof of Theorem 2.9.

We write

$$\hat{F}^\circ : \mathbf{X}^N \mapsto \hat{F}(\mathbf{X}^N) - \pi(f) = \frac{\sum_{n=1}^N \omega(x_n)\{f(x_n) - \pi(f)\}}{\sum_{m=1}^N \omega(x_m)}.$$

We can write

$$\mathbb{E}_{\mathbf{X}_0^N \sim \bar{q}}\left[|\hat{F}(\mathbf{x}_t) - \pi(f)|^s\right] = \int \bar{q}(\mathrm{d}\mathbf{X}_0^N) P(\mathbf{X}_0^N, \mathrm{d}\mathbf{X}_1^N) \ldots P(\mathbf{X}_{t-1}^N, \mathrm{d}\mathbf{X}_t^N) |\hat{F}^\circ(\mathbf{X}_t^N)|^s$$
$$= \int \bar{q}(\mathrm{d}\mathbf{X}_0^N) P(\mathbf{X}_0^N, \mathrm{d}\mathbf{X}_1^N) \ldots P(\mathbf{X}_{t-1}^N, \mathrm{d}\mathbf{X}_t^N)$$
$$|\hat{F}^\circ(\mathbf{X}_t^N)|^s \{\mathbb{1}(A_t) + \mathbb{1}(A_t^c)\},$$

where the event A_t represents "there was an acceptance in the first t steps".

In the event A_t^c, $\mathbf{X}_t^N = \mathbf{X}_0^N$ so

$$\int \bar{q}(\mathrm{d}\mathbf{X}_0^N) P(\mathbf{X}_0^N, \mathrm{d}\mathbf{X}_1^N) \ldots P(\mathbf{X}_{t-1}^N, \mathrm{d}\mathbf{X}_t^N) |\hat{F}^\circ(\mathbf{X}_t^N)|^s \cdot \mathbb{1}(A_t^c)$$
$$= \int \bar{q}(\mathrm{d}\mathbf{X}_0^N) P(\mathbf{X}_0^N, \mathrm{d}\mathbf{X}_1^N) \ldots P(\mathbf{X}_{t-1}^N, \mathrm{d}\mathbf{X}_t^N) |\hat{F}^\circ(\mathbf{X}_0^N)|^s \cdot \mathbb{1}(A_t^c)$$
$$\leq \int \bar{q}(\mathrm{d}\mathbf{X}_0^N) |\hat{F}^\circ(\mathbf{X}_0^N)|^s,$$

by bounding the indicator by one, and we can use Theorem 2.9 to obtain a bound in $N^{-s/2}$.

Now we consider the case A_t. For $1 \leq j \leq t$ define the events

$$A_{j,t} := \{\mathbf{X}_{j-1}^N \neq \mathbf{X}_j^N = \mathbf{X}_{j+1}^N = \cdots = \mathbf{X}_t^N\},$$

where $A_{j,t}$ is the event that there is a jump at time j and no jump after that. Then $A_{j,t} \cap A_{j',t} =$

$\emptyset$ for $j \neq j'$ and $A_t = \cup_{j=1}^t A_{j,t}$. We can decompose $\mathbb{1}(A_t)$ into $\sum_{j=1}^t \mathbb{1}(A_{j,t})$ to get

$$\begin{aligned}
&\int \bar{q}(\mathrm{d}\mathbf{X}_0^N) P(\mathbf{X}_0^N, \mathrm{d}\mathbf{X}_1^N) \dots P(\mathbf{X}_{t-1}^N, \mathrm{d}\mathbf{X}_t^N) |\hat{F}^\circ(\mathbf{X}_t^N)|^s \mathbb{1}(A_t) \\
&= \sum_{j=1}^t \mathbb{E}_{\mathbf{X}_0^N \sim \bar{q}} \left[|\hat{F}^\circ(\mathbf{X}_t^N)|^s \mathbb{1}(A_{j,t}) \right] = \sum_{j=1}^t \mathbb{E}_{\mathbf{X}_0^N \sim \bar{q}} \left[\mathbb{E} \left\{ |\hat{F}^\circ(\mathbf{X}_j^N)|^s \mathbb{1}(A_{j,t}) \middle| \mathbf{X}_{j-1}^N \right\} \right].
\end{aligned}$$

Conditional on $\mathbf{X}_{j-1}^N$,

$$\begin{aligned}
&\int P(\mathbf{X}_{j-1}^N, d\mathbf{X}_j^N) P(\mathbf{X}_j^N, d\mathbf{X}_{j+1}^N) \cdots P(\mathbf{X}_{t-1}^N, d\mathbf{X}_t^N) \\
&\qquad |\hat{F}^\circ(\mathbf{X}_j^N)|^s \mathbb{1}\{\mathbf{X}_{j-1}^N \neq \mathbf{X}_j^N = \cdots = \mathbf{X}_t^N\} \\
&= \int P(\mathbf{X}_{j-1}^N, d\mathbf{X}_j^N) |\hat{F}^\circ(\mathbf{X}_j^N)|^s \mathbb{1}\{\mathbf{X}_{j-1}^N \neq \mathbf{X}_j^N\} \\
&\qquad \int P(\mathbf{X}_j^N, d\mathbf{X}_{j+1}^N) \cdots P(\mathbf{X}_{t-1}^N, d\mathbf{X}_t^N) \mathbb{1}\{\mathbf{X}_j^N = \cdots = \mathbf{X}_t^N\} \\
&= \int P(\mathbf{X}_{j-1}^N, d\mathbf{X}_j^N) |\hat{F}^\circ(\mathbf{X}_j^N)|^s \mathbb{1}\{\mathbf{X}_{j-1}^N \neq \mathbf{X}_j^N\} r(\mathbf{X}_j^N)^{t-j} \\
&= \int \bar{q}(d\zeta) \alpha(\mathbf{X}_{j-1}^N, \zeta) |\hat{F}^\circ(\zeta)|^s r(\zeta)^{t-j}.
\end{aligned}$$

We can then upper bound α by one, and upper bound $\sum_{j=1}^t r(\zeta)^{t-j}$ by $(1-r(\zeta))^{-1}$ to obtain

$$\begin{aligned}
&\sum_{j=1}^t \mathbb{E}_{\mathbf{X}_0^N \sim \bar{q}, \zeta \sim \bar{q}} \left[\alpha(\mathbf{X}_{j-1}^N, \zeta) |\hat{F}^\circ(\zeta)|^s r(\zeta)^{t-j} \right] \\
&\leq \mathbb{E}_{\zeta \sim \bar{q}} \left[|\hat{F}^\circ(\zeta)|^s \sum_{j=1}^t r(\zeta)^{t-j} \right] \\
&\leq \mathbb{E}_{\zeta \sim \bar{q}} \left[|\hat{F}^\circ(\zeta)|^s \frac{1}{1-r(\zeta)} \right].
\end{aligned}$$

Next, split the expectation into the cases $\hat{Z}(\zeta) > 2$ and $\hat{Z}(\zeta) \leq 2$:

$$\begin{aligned}
\mathbb{E}_{\zeta \sim \bar{q}} \left[|\hat{F}^\circ(\zeta)|^s \frac{1}{1-r(\zeta)} \right] &= \mathbb{E}_{\zeta \sim \bar{q}} \left[|\hat{F}^\circ(\zeta)|^s \frac{1}{1-r(\zeta)} \mathbb{1}(\hat{Z}(\zeta) \leq 2) \right] \\
&\quad + \mathbb{E}_{\zeta \sim \bar{q}} \left[|\hat{F}^\circ(\zeta)|^s \frac{1}{1-r(\zeta)} \mathbb{1}(\hat{Z}(\zeta) > 2) \right].
\end{aligned}$$

When $\hat{Z}(\zeta) \leq 2$, since r is increasing with $\hat{Z}$, we have $r(\zeta) \leq r(2)$ and thus $(1-r(\zeta))^{-1} \leq (1-r(2))^{-1}$. This yields:

$$\mathbb{E}_{\zeta \sim \bar{q}} \left[|\hat{F}^\circ(\zeta)|^s \frac{1}{1-r(\zeta)} \mathbb{1}(\hat{Z}(\zeta) \leq 2) \right] \leq \frac{1}{1-r(2)} \int \bar{q}(\mathrm{d}\zeta) |\hat{F}^\circ(\zeta)|^s \mathbb{1}(\hat{Z}(\zeta) \leq 2)$$

$$\leq \frac{1}{1-r(2)}\mathbb{E}_{\mathbf{x}_0\sim\bar{q}}\left[|\hat{F}^\circ(\mathbf{x}_0)|^s\right].$$

We obtain a bound in $N^{-s/2}$ using Theorem 2.9.

When $\hat{Z}(\zeta) > 2$, from Lemma A.5, we have $r(\zeta) \leq 1 - c_p(1/2)/(2\hat{Z}(\zeta))$. Thus $1 - r(\zeta) \geq c_p(1/2)/(2\hat{Z}(\zeta))$, and $(1-r(\zeta))^{-1} \leq (2/c_p(1/2))\hat{Z}(\zeta)$. This yields:

$$\mathbb{E}_{\zeta\sim\bar{q}}\left[|\hat{F}^\circ(\zeta)|^s\frac{1}{1-r(\zeta)}\mathbb{1}(\hat{Z}(\zeta)>2)\right] \leq \frac{2}{c_p(1/2)}\int \bar{q}(\mathrm{d}\zeta)|\hat{F}^\circ(\zeta)|^s\hat{Z}(\zeta)\mathbb{1}(\hat{Z}(\zeta)>2).$$

Since we assume $f \geq 0$, we can use the inequality (A.4):

$$|\hat{F}^\circ(\zeta)|^s \leq \left(\max_{1\leq i\leq N} f(\zeta_i) + q(\omega f)\right)^s,$$

from which we obtain

$$\begin{aligned}
&\mathbb{E}_{\zeta\sim\bar{q}}\left[|\hat{F}^\circ(\zeta)|^s\frac{1}{1-r(\zeta)}\mathbb{1}(\hat{Z}(\zeta)>2)\right] \\
&\quad\leq \frac{2}{c_p(1/2)}\mathbb{E}_{\zeta\sim\bar{q}}\left[\left(\max_{1\leq i\leq N} f(\zeta_i) + q(\omega f)\right)^s\cdot\hat{Z}(\zeta)\mathbb{1}(\hat{Z}(\zeta)>2)\right] \\
&\quad\leq \frac{2^s}{c_p(1/2)}\Big(\mathbb{E}_{\zeta\sim\bar{q}}\left[\left(\max_{1\leq i\leq N} f(\zeta_i)\right)^s\cdot\hat{Z}(\zeta)\mathbb{1}(\hat{Z}(\zeta)>2)\right] \\
&\qquad + q(\omega f)^s\cdot\mathbb{E}_{\zeta\sim\bar{q}}\left[\hat{Z}(\zeta)\mathbb{1}(\hat{Z}(\zeta)>2)\right]\Big).
\end{aligned}$$

Using the facts that $\hat{Z}(\zeta) \leq 2(\hat{Z}(\zeta)-1)$ when $\hat{Z}(\zeta) > 2$ and $\mathbb{1}(\hat{Z}(\zeta)\geq 2) \leq |\hat{Z}(\zeta)-1|^{p-1}$, we obtain via Proposition 2.5:

$$\begin{aligned}
q(\omega f)^s\cdot\mathbb{E}_{\zeta\sim\bar{q}}\left[\hat{Z}(\zeta)\mathbb{1}(\hat{Z}(\zeta)>2)\right] &\leq 2q(\omega f)^s\mathbb{E}_{\zeta\sim\bar{q}}\left[|\hat{Z}(\zeta)-1|^p\right] \\
&\leq \frac{2M(p)^p}{N^{p/2}}q(\omega f)^s.
\end{aligned}$$

For the remaining term, using Hölder's inequality yields:

$$\begin{aligned}
&\mathbb{E}_{\zeta\sim\bar{q}}\left[\left(\max_{1\leq i\leq N} f(\zeta_i)\right)^s\cdot\hat{Z}(\zeta)\mathbb{1}(\hat{Z}(\zeta)>2)\right] \\
&\quad\leq 2\mathbb{E}_{\zeta\sim\bar{q}}\left[\left(\max_{1\leq i\leq N} f(\zeta_i)\right)^r\right]^{s/r}\cdot\mathbb{E}_{\zeta\sim\bar{q}}\left[\left|\hat{Z}(\zeta)-1\right|^{\frac{r}{r-s}}\mathbb{1}(\hat{Z}(\zeta)>2)\right]^{1-s/r}.
\end{aligned}$$

Under the assumptions, with $s \leq \frac{pr}{p+r+2}$, we have:

$$\frac{r}{r-s} \leq \frac{p+r+2}{r+2} = 1 + \frac{p}{r+2} \leq p,$$

where the inequality holds since $r \geq 2$ by assumption. This gives us:

$$\mathbb{E}_{\zeta\sim\bar{q}}\left[\left|\hat{Z}(\zeta)-1\right|^{\frac{r}{r-s}}\mathbb{1}(\hat{Z}(\zeta)>2)\right]^{1-s/r} \leq \left(\frac{M(p)^p}{N^{p/2}}\right)^{1-s/r}.$$

Finally we use the fact that $\max\{a_1,\ldots,a_n\} \leq a_1+\cdots+a_n$ for non-negative a_i to derive

$$\begin{aligned}&\mathbb{E}_{\zeta\sim\bar{q}}\left[\left(\max_{1\leq i\leq N} f(\zeta_i)\right)^r\right]\\ &= \mathbb{E}_{\zeta\sim\bar{q}}\left[\max_{1\leq i\leq N} f(\zeta_i)^r\right] \leq \mathbb{E}_{x\sim q}\left[f(x)^r N\right],\end{aligned}$$

so that

$$\begin{aligned}&\mathbb{E}_{\zeta\sim\bar{q}}\left[\left(\max_{1\leq i\leq N} f(\zeta_i)\right)^s \cdot \hat{Z}(\zeta)\mathbb{1}(\hat{Z}(\zeta)>2)\right]\\ &\leq 2\left(q(f^r)N\right)^{s/r}\cdot\left(\frac{M(p)^p}{N^{p/2}}\right)^{1-s/r}.\end{aligned}$$

We end up with an exponent of N equal to $s/r-(p/2)(r-s)/r$, which under the assumptions is less than $-s/2$, as detailed in the proof of Theorem 2.9. Therefore, we obtain an upper bound in $N^{-s/2}$ on all terms.

□

Remark A.14. Under Assumption 2.1, PIMH converges in total variation. Thus, $(\mathbf{X}_t^N)$ converges weakly to $\bar{\pi}$. We consider the transformation $\mathbf{X}^N \mapsto |\hat{F}^\circ(\mathbf{X}^N)|^s$ and Fatou's lemma as in Theorem 3.4 of Billingsley (1999), to obtain

$$\mathbb{E}_{\bar{\pi}}[|\hat{F}^\circ(\mathbf{X}^N)|^s] \leq \liminf_t \mathbb{E}[|\hat{F}^\circ(\mathbf{X}_t^N)|^s].$$

Thus, the bound of Proposition 2.27, valid for all $t \geq 0$, applies also to the s-th moment of $\hat{F}(\mathbf{X}^N)-\pi(f)$ under $\bar{\pi}$.

A.5.3 Proof of Proposition 2.28

Proof of Proposition 2.28. We start as in the proof of Proposition 2.26 in Appendix A.5.1, and employ Theorem 2.9 for the moments of the error of IS with unbounded functions. Regarding the bias cancellation term,

$$\mathrm{BC} = \sum_{t=1}^{\infty}\Delta_t\mathbb{1}(\tau>t),$$

we use Minkowski with exponent $s \geq 1$:

$$\mathbb{E}\left[|\mathrm{BC}|^s\right]^{1/s} \leq \sum_{t=1}^{\infty} \mathbb{E}\left[|\Delta_t|^s \mathbb{1}(\tau > t)\right]^{1/s}. \tag{A.27}$$

Next, for each time t, using Hölder's inequality with an arbitrary $\kappa > 1$,

$$\mathbb{E}\left[|\Delta_t|^s \mathbb{1}(\tau > t)\right] \leq \mathbb{E}\left[|\Delta_t|^{s\kappa}\right]^{1/\kappa} \mathbb{P}(\tau > t)^{(\kappa-1)/\kappa}.$$

For the sum over t in (A.27) to be finite, and using Proposition 2.14 to bound $\mathbb{P}(\tau > t)$, we have the condition on κ and s,

$$-\frac{p(\kappa-1)}{s\kappa} < -1 \quad \Leftrightarrow \quad \kappa > p/(p-s).$$

To establish the finiteness of $\mathbb{E}\left[|\Delta_t|^{s\kappa}\right]$ we can resort to Proposition 2.27 if $s\kappa$ satisfies the condition

$$s\kappa \leq \frac{pr}{p+r+2}.$$

We can find such κ if

$$\frac{ps}{p-s} < \frac{pr}{p+r+2}.$$

□

A.5.4 Proof of Proposition 2.30

Proof of Proposition 2.30. We follow the proof of Proposition 2.28, with $s = 2$. We thus have an exponent $\kappa > 1$ that must satisfy $\kappa > p/(p-2)$, and $2\kappa \leq pr/(p+r+2)$. We choose any number κ strictly between $p/(p-2)$ and $pr/(2p+2r+4)$, which is possible by assumption, since

$$1 < \frac{p}{p-2} < \frac{pr}{2p+2r+4} \Leftrightarrow 2p+4r+4 < rp.$$

For that κ, we can apply Proposition 2.27 to bound $\mathbb{E}[|\Delta_t|^{2\kappa}]^{1/\kappa}$ by a constant times N^{-1}. Meanwhile, the sum $\sum_{t=1}^{\infty} \mathbb{P}(\tau > t)^{(\kappa-1)/(\kappa s)}$ is finite using Proposition 2.14, and is of the form CN^{-a} for some positive a, namely $a = (\kappa-1)/(2\kappa s)$. Thus, $\mathbb{E}|\mathrm{BC}|^2$ can be bounded by a constant times N^{-1-a} for some positive a, and is negligible compared to the MSE of IS as $N \to \infty$. □

A.5.5 Proof of Proposition 2.33 on the unbiased estimation of $1/Z$

The proof of Proposition 2.33 relies on the inverse-moment bound of Proposition A.1, which provides a finite constant bound on $\mathbb{E}[\hat{Z}_N(\mathbf{X}^N)^{-r}]$ for all $r \geq 1$ with no constraint on r relative to p. We first establish sharper SNIS and PIMH moment bounds for $f = 1/\omega_u$.

Proposition A.15 (SNIS moments of $1/\hat{Z}_N$)**.** *Assume that $q(\omega^p) < \infty$ for some $p \geq 2$ and $q(\omega^{-\eta}) < \infty$ for some $\eta > 0$. Then for any $1 \leq s < p$ and $N \geq sp/((p-s)\eta)$, there exists $C > 0$ such that*

$$\mathbb{E}_{\bar{q}}\left[\left|\frac{1}{\hat{Z}_N(\mathbf{X}^N)} - \frac{1}{Z}\right|^s\right] \leq C\, N^{-s/2}. \tag{A.28}$$

Proof. Using $|1/\hat{Z}_N - 1/Z|^s = |\hat{Z}_N - Z|^s/(Z^s\, \hat{Z}_N^s)$ and Hölder's inequality with exponents $\kappa = p/s$ and $\kappa' = p/(p-s)$:

$$\mathbb{E}\left[\frac{|\hat{Z}_N - Z|^s}{Z^s\, \hat{Z}_N^s}\right] \leq Z^{-s}\, \mathbb{E}\Big[|\hat{Z}_N - Z|^p\Big]^{s/p} \cdot \mathbb{E}\Big[\hat{Z}_N^{-sp/(p-s)}\Big]^{(p-s)/p}.$$

The first factor is $O(N^{-s/2})$ by Proposition 2.5. The second is bounded by a constant via Proposition A.1 with $r = sp/(p-s)$, for $N \geq sp/((p-s)\eta)$. □

Proposition A.16 (PIMH moments of $1/\hat{Z}_N$)**.** *Under the same conditions as Proposition A.15, there exists $C > 0$ such that for all $t \geq 0$:*

$$\mathbb{E}_{\mathbf{X}_0^N \sim \bar{q}}\left[\left|\frac{1}{\hat{Z}_N(\mathbf{X}_t^N)} - \frac{1}{Z}\right|^s\right] \leq C\, N^{-s/2}, \tag{A.29}$$

where $(\mathbf{X}_t^N)_{t\geq 0}$ is the PIMH chain initialized at $\mathbf{X}_0^N \sim \bar{q}$.

Proof of Proposition A.16. The case $t = 0$ is Proposition A.15. For $t \geq 1$, we follow the proof of Proposition 2.27 in Appendix A.5.2. The decomposition into the rejection event (A_t^c), small-weight regime ($\hat{Z}_N(\zeta) \leq 2$), and large-weight regime ($\hat{Z}_N(\zeta) > 2$) carries over. The A_t^c and small-weight contributions are each bounded by a constant times $\mathbb{E}_{\bar{q}}[|1/\hat{Z}_N - 1/Z|^s] \leq CN^{-s/2}$ by Proposition A.15.

The key simplification occurs in the large-weight regime. For $f = 1/\omega_u$, we have $|\hat{F}_N^\circ(\zeta)|^s = |\hat{Z}_N(\zeta) - Z|^s/(Z^s\, \hat{Z}_N(\zeta)^s)$, and using Lemma A.5 with $\theta = 1/2$, $(1-r(\zeta))^{-1} \leq 2\hat{Z}_N(\zeta)/c_p(1/2)$ when $\hat{Z}_N(\zeta) > 2$. The product becomes

$$\frac{|\hat{Z}_N(\zeta) - Z|^s}{Z^s\, \hat{Z}_N(\zeta)^s} \cdot \frac{2\hat{Z}_N(\zeta)}{c_p(1/2)} = \frac{2}{Z^s\, c_p(1/2)} \cdot \frac{|\hat{Z}_N(\zeta) - Z|^s}{\hat{Z}_N(\zeta)^{s-1}},$$

where one power of $\hat{Z}_N$ cancels. This is the decisive simplification afforded by the inverse weight structure: for a general test function f, the analogous step requires Hölder's inequality to separate $|\hat{F}_N^\circ|^s$ from $\hat{Z}_N \cdot \mathbb{1}(\hat{Z}_N > 2)$, introducing the test-function moment $q(|f|^r)$ and the constraint $s \leq pr/(p+r+2)$. For $f = 1/\omega_u$, the partial cancellation leaves only inverse moments of $\hat{Z}_N$, controlled by Proposition A.1.

Applying Hölder's inequality with exponents p/s and $p/(p-s)$:

$$\mathbb{E}\left[\frac{|\hat{Z}_N - Z|^s}{\hat{Z}_N^{s-1}}\right] \leq \mathbb{E}[|\hat{Z}_N - Z|^p]^{s/p} \cdot \mathbb{E}[\hat{Z}_N^{-(s-1)p/(p-s)}]^{(p-s)/p}.$$

The first factor is $O(N^{-s/2})$ by Proposition 2.5, and the second is bounded by a constant via Proposition A.1 with $r = (s-1)p/(p-s)$. Combining the three contributions (A_t^c, small weight, large weight) yields the bound. □

Endpoint moments at slower rate The Hölder decomposition in the proofs of Propositions A.15 and A.16 requires $s < p$: at $s = p$ the exponent $\kappa = p/s = 1$ no longer separates the numerator from the denominator. A truncation argument recovers the endpoint. Split the expectation at the event $\{\hat{Z}_N \geq 1/2\}$. On $\{\hat{Z}_N \geq 1/2\}$, $\hat{Z}_N^{-p} \leq 2^p$, so this contribution is bounded by $2^p Z^{-p}\,\mathbb{E}[|\hat{Z}_N - Z|^p] \leq C_1 N^{-p/2}$ by Proposition 2.5. On $\{\hat{Z}_N < 1/2\}$, $|\hat{Z}_N - Z|$ is bounded by a constant so $|\hat{Z}_N - Z|^p/\hat{Z}_N^p \leq C\,\hat{Z}_N^{-p}$. Applying Hölder with exponents $1+\delta$ and $(1+\delta)/\delta$ for $\delta = 2\alpha/(p-2\alpha)$, the inverse moment factor is controlled by Proposition A.1 and the probability factor by Markov's inequality. This yields: for every $\alpha \in (0, p/2)$, there exists $C_\alpha > 0$ such that for N large enough,

$$\mathbb{E}_{\bar{q}}\left[\left|\frac{1}{\hat{Z}_N(\mathbf{X}^N)} - \frac{1}{Z}\right|^p\right] \leq C_\alpha\, N^{-\alpha}.$$

The same argument extends to the PIMH chain uniformly in t, since the A_t^c and small-weight contributions are bounded by $\mathbb{E}_{\bar{q}}[|1/\hat{Z}_N - \frac{1}{Z}|^p] \leq C_\alpha N^{-\alpha}$, and the large-weight contribution with $s = p$ involves $|\hat{Z}_N - Z|^p/\hat{Z}_N^{p-1}$, which is handled by the same truncation.

Proof of Proposition 2.33. Finite s-th moments under $p > s$. We follow the proof of Proposition 2.28 in Appendix A.5.3 with $f = 1/\omega$. By Minkowski's inequality, it suffices to bound $\mathbb{E}[|\mathrm{BC}|^s]^{1/s}$, where $\mathrm{BC} = \sum_{t=1}^{\infty} \Delta_t \mathbb{1}(\tau > t)$. Applying Hölder's inequality with exponent $\kappa > 1$ and conjugate $\kappa' = \kappa/(\kappa-1)$:

$$\mathbb{E}[|\Delta_t|^s \mathbb{1}(\tau > t)] \leq \mathbb{E}[|\Delta_t|^{s\kappa}]^{1/\kappa} \cdot \mathbb{P}(\tau > t)^{(\kappa-1)/\kappa}. \tag{A.30}$$

For general f, bounding $\mathbb{E}[|\Delta_t|^{s\kappa}]$ requires $s\kappa \leq pr/(p+r+2)$, imposing an upper bound on κ. For $f = 1/\omega$, we show that a constant bound on $\mathbb{E}[|\Delta_t|^{s\kappa}]$ holds for *all* κ, so that κ can be chosen freely.

By the triangle inequality,

$$|\Delta_t| = \left|\frac{1}{\hat{Z}_N(\mathbf{X}_t^N)} - \frac{1}{\hat{Z}_N(\mathbf{Y}_{t-1}^N)}\right| \leq \frac{1}{\hat{Z}_N(\mathbf{X}_t^N)} + \frac{1}{\hat{Z}_N(\mathbf{Y}_{t-1}^N)}, \tag{A.31}$$

so by Minkowski's inequality in $L^{s\kappa}$,

$$\mathbb{E}[|\Delta_t|^{s\kappa}]^{1/(s\kappa)} \leq \mathbb{E}\Big[\hat{Z}_N(\mathbf{X}_t^N)^{-s\kappa}\Big]^{1/(s\kappa)} + \mathbb{E}\Big[\hat{Z}_N(\mathbf{Y}_{t-1}^N)^{-s\kappa}\Big]^{1/(s\kappa)}. \tag{A.32}$$

We now show that $\mathbb{E}[\hat{Z}_N(\mathbf{X}_t^N)^{-s\kappa}] \leq \mathbb{E}_{\bar{q}}[\hat{Z}_N^{-s\kappa}]$ for all $t \geq 0$, by induction on the PIMH chain. The base case $t = 0$ is immediate since $\mathbf{X}_0^N \sim \bar{q}$. For the inductive step, recall that the Metropolis–Hastings update at time t draws a proposal $\boldsymbol{\zeta}_t \sim \bar{q}$ independently and accepts it with probability $\min(1, \hat{Z}_N(\boldsymbol{\zeta}_t)/\hat{Z}_N(\mathbf{X}_{t-1}^N))$. Conditioning on $\mathbf{X}_{t-1}^N$ and $\boldsymbol{\zeta}_t$:

$$\begin{aligned}\mathbb{E}\Big[\hat{Z}_N(\mathbf{X}_t^N)^{-s\kappa} \mid \mathbf{X}_{t-1}^N, \boldsymbol{\zeta}_t\Big] &= \min\left(1, \frac{\hat{Z}_N(\boldsymbol{\zeta}_t)}{\hat{Z}_N(\mathbf{X}_{t-1}^N)}\right) \hat{Z}_N(\boldsymbol{\zeta}_t)^{-s\kappa} \\ &\quad + \left(1 - \min\left(1, \frac{\hat{Z}_N(\boldsymbol{\zeta}_t)}{\hat{Z}_N(\mathbf{X}_{t-1}^N)}\right)\right) \hat{Z}_N(\mathbf{X}_{t-1}^N)^{-s\kappa}. \end{aligned} \tag{A.33}$$

This is a convex combination. If $\hat{Z}_N(\boldsymbol{\zeta}_t) \geq \hat{Z}_N(\mathbf{X}_{t-1}^N)$, the proposal is accepted with certainty and the expression equals $\hat{Z}_N(\boldsymbol{\zeta}_t)^{-s\kappa}$. If $\hat{Z}_N(\boldsymbol{\zeta}_t) < \hat{Z}_N(\mathbf{X}_{t-1}^N)$, then $\hat{Z}_N(\mathbf{X}_{t-1}^N)^{-s\kappa} \leq \hat{Z}_N(\boldsymbol{\zeta}_t)^{-s\kappa}$, so the convex combination is also bounded by $\hat{Z}_N(\boldsymbol{\zeta}_t)^{-s\kappa}$. In both cases,

$$\mathbb{E}\Big[\hat{Z}_N(\mathbf{X}_t^N)^{-s\kappa} \mid \mathbf{X}_{t-1}^N, \boldsymbol{\zeta}_t\Big] \leq \hat{Z}_N(\boldsymbol{\zeta}_t)^{-s\kappa}. \tag{A.34}$$

Since $\boldsymbol{\zeta}_t \sim \bar{q}$ independently of $\mathbf{X}_{t-1}^N$, taking expectations yields

$$\mathbb{E}\Big[\hat{Z}_N(\mathbf{X}_t^N)^{-s\kappa}\Big] \leq \mathbb{E}_{\bar{q}}\Big[\hat{Z}_N^{-s\kappa}\Big] \leq q(\omega^{-\eta})^{s\kappa/\eta}, \tag{A.35}$$

where the last inequality is Proposition A.1 with $r = s\kappa$, for $N \geq s\kappa/\eta$. The same bound holds for $\mathbf{Y}_{t-1}^N$. Combining (A.32) and (A.35):

$$\mathbb{E}[|\Delta_t|^{s\kappa}]^{1/(s\kappa)} \leq 2\, q(\omega^{-\eta})^{1/\eta} =: \bar{C}_\eta < \infty, \tag{A.36}$$

for all κ with $N \geq s\kappa/\eta$, with *no constraint* on $s\kappa$ relative to p.

Substituting (A.36) and the meeting-time bound $\mathbb{P}(\tau > t) \leq C/(N^{1/2}t^p)$ of Proposition 2.14 into (A.30), and summing over t:

$$\sum_{t=1}^{\infty} \mathbb{E}[|\Delta_t|^s \mathbb{1}(\tau > t)]^{1/s} \leq \bar{C}_\eta \sum_{t=1}^{\infty} \left(\frac{C}{N^{1/2}t^p}\right)^{(\kappa-1)/(s\kappa)}.$$

The series converges if and only if $p(\kappa-1)/(s\kappa) > 1$, i.e. $\kappa > p/(p-s)$. Since $p > s$, we have $p/(p-s) < \infty$, and since (A.36) imposes no upper bound on κ, such a choice exists. This gives $\mathbb{E}[|\mathrm{BC}|^s] < \infty$.

Rate $N^{-s/2}$ under $p > 2s$. To upgrade from finiteness to the rate $N^{-s/2}$, replace the con-

stant bound (A.36) by the PIMH rate of Proposition A.16: for $s\kappa < p$,

$$\mathbb{E}[|\Delta_t|^{s\kappa}]^{1/(s\kappa)} \leq C_1 N^{-1/2}. \tag{A.37}$$

Two constraints on κ must now hold simultaneously:

$$\frac{p}{p-s} < \kappa < \frac{p}{s}. \tag{A.38}$$

The first ensures series convergence; the second ensures the PIMH bound applies. This interval is non-empty if and only if $p/(p-s) < p/s$, i.e. $p > 2s$. Choosing such κ and separating the N-dependent factors from the convergent t-series gives $\mathbb{E}[|\mathrm{BC}|^s] = o(N^{-s/2})$, and the MSE equivalence for $s = 2$ follows from (2.42). □

Appendix B

Proofs of Chapter 3

This appendix collects the proofs of the results stated in Chapter 3.

This section proves all the theorems. Section B.1 reviews the PAC-Bayes bounds of Bhattacharya et al. (2019); Alquier & Ridgway (2020) and gives variants in expectation. Section B.2 uses these to prove Theorem 3.2. Section B.3 then proves the results of Section 3.3, including Theorem 3.7; although these follow from Theorem 3.2, we derive them from a localized PAC-Bayes bound, following Catoni (2007). Finally, Section B.4 gathers the proofs of the results in Section 3.4.

B.1 PAC-Bayes bounds

We first state a PAC-Bayes bound in expectation for the fractional posterior. In this form, the bound is due to Alquier & Ridgway (2020), who obtained it as a variant of a result of Bhattacharya et al. (2019) using the same proof technique. Similar results can be found in Zhang (2006a). We recall the proof for completeness.

Theorem B.1 (Theorem 2.6 in Alquier & Ridgway (2020))**.** *Given a data-dependent posterior $\widehat{\rho}$, a prior π, and observations drawn from a data-generating distribution P^* with density p^* with respect to μ, we have the following bound, for any $\alpha \in (0,1)$:*

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widehat{\rho}}[\mathcal{D}_\alpha(P_\theta\|P^*)] \leq \mathbb{E}_{\mathcal{S}}\left[\frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\theta\sim\widehat{\rho}}[r_n(\theta)]\right] + \frac{1}{n(1-\alpha)}\mathbb{E}_{\mathcal{S}}[\mathrm{KL}(\widehat{\rho}\|\pi)]$$

Proof. Fix $\alpha \in (0,1)$ and $\theta \in \Theta$. Since $P^* \ll \mu$, the Rényi divergence $\mathcal{D}_\alpha(P_\theta\|P^*)$ to the

truth is well defined; since the X_i are i.i.d. with law P^*,

$$\mathbb{E}_{\mathcal{S}}\left[\exp\left(-\alpha r_n(\theta)\right)\right] = \mathbb{E}_{\mathcal{S}}\left[\prod_{i=1}^{n}\left(\frac{p_\theta(X_i)}{p^*(X_i)}\right)^{\alpha}\right] = \mathbb{E}_{\mathcal{S}}\left[\left(\frac{p_\theta(X_1)}{p^*(X_1)}\right)^{\alpha}\right]^n$$
$$= \exp\left(-n(1-\alpha)\mathcal{D}_\alpha(P_\theta\|P^*)\right).$$

This can be rewritten as follows:

$$\mathbb{E}_{\mathcal{S}}\left[\exp\left(-\alpha r_n(\theta) + n(1-\alpha)\mathcal{D}_\alpha(P_\theta\|P^*)\right)\right] = 1.$$

Integrating with respect to π and using Fubini's theorem,

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\pi}\left[\exp\left(-\alpha r_n(\theta) + n(1-\alpha)\mathcal{D}_\alpha(P_\theta\|P^*)\right)\right]\right] = 1.$$

The main step is the Donsker–Varadhan variational formula (1.24), recalled in the introduction. Applied to the function $h(\theta) = -\alpha r_n(\theta) + n(1-\alpha)\mathcal{D}_\alpha(P_\theta\|P^*)$, for which $\log\int e^h\,\pi = \sup_{\rho\in\mathcal{M}_1^+}\left[\int h\rho - \mathrm{KL}(\rho\|\pi)\right]$, we have that:

$$\mathbb{E}_{\mathcal{S}}\left[\exp\left[\sup_{\rho\in\mathcal{M}_1^+}\left[\int h\rho - \mathrm{KL}(\rho\|\pi)\right]\right]\right] = 1.$$

Using Jensen's inequality, we have that:

$$\mathbb{E}_{\mathcal{S}}\left[\sup_{\rho\in\mathcal{M}_1^+}\left[\int h\rho - \mathrm{KL}(\rho\|\pi)\right]\right] \leq 0.$$

In particular, for any data-dependent posterior $\widehat{\rho}$, we have that:

$$\mathbb{E}_{\mathcal{S}}\left[\int h\widehat{\rho} - \mathrm{KL}(\widehat{\rho}\|\pi)\right] \leq 0.$$

Using the definition of h, we have that:

$$\mathbb{E}_{\mathcal{S}}\left[\int -\alpha r_n(\theta) + n(1-\alpha)\mathcal{D}_\alpha(P_\theta\|P^*)\widehat{\rho}(d\theta) - \mathrm{KL}(\widehat{\rho}\|\pi)\right] \leq 0.$$

Rearranging the terms, we have that:

$$n(1-\alpha)\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widehat{\rho}}\left[\mathcal{D}_\alpha(P_\theta\|P^*)\right] \leq \alpha\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widehat{\rho}}\left[r_n(\theta)\right] + \mathbb{E}_{\mathcal{S}}\left[\mathrm{KL}(\widehat{\rho}\|\pi)\right].$$

Dividing by $n(1-\alpha)$, we get the desired result. □

Consider the special case $\widehat{\rho} = \pi_{n,\alpha}$ and take the infimum over a family $\mathcal{F} \subset \mathcal{M}_1^+$ on the

right-hand side. This gives the following corollary, which is the in-expectation bound of Alquier & Ridgway (2020, Theorem 2.6).

Corollary B.2. *Fix $\mathcal{F} \subset \mathcal{M}_1^+$, and $\alpha \in (0,1)$, we have the following bound:*

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi_{n,\alpha}}\left[\mathcal{D}_\alpha(P_\theta\|P^*)\right] \leq \inf_{\rho\in\mathcal{F}}\left[\frac{\alpha}{1-\alpha}\mathbb{E}_{\theta\sim\rho}\left[\mathrm{KL}(P^*\|P_\theta)\right] + \frac{\mathrm{KL}(\rho\|\pi)}{n(1-\alpha)}\right].$$

Under Assumption 3.4, the bound of Corollary B.2 is of order $\mathcal{O}\left(d_\pi \log(n)/n\right)$, as shown in Alquier & Ridgway (2020). The bound holds for any prior π, and can therefore be optimized over π. We carry out this optimization using mutual information, which is the route to Theorem 3.2.

We also state the variant in probability.

Theorem B.3 (Theorem 2.1 and Corollary 2.3 in Alquier & Ridgway (2020))**.** *Given a data-dependent posterior $\widehat{\rho}$ and observations drawn from a data-generating distribution P^* with density p^* with respect to μ, we have, for any $\alpha \in (0,1)$, with probability at least $1-\delta$ over the sample $\mathcal{S}$,*

$$\mathbb{E}_{\theta\sim\widehat{\rho}}[\mathcal{D}_\alpha(P_\theta\|P^*)] \leq \frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\theta\sim\widehat{\rho}}[r_n(\theta)] + \frac{\mathrm{KL}(\widehat{\rho}\|\pi) + \log\frac{1}{\delta}}{n(1-\alpha)}.$$

In particular,

$$\mathbb{E}_{\theta\sim\widetilde{\rho}_{n,\alpha}}[\mathcal{D}_\alpha(P_\theta\|P^*)] \leq \inf_{\rho\in\mathcal{F}}\left[\frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\theta\sim\rho}[r_n(\theta)] + \frac{\mathrm{KL}(\rho\|\pi) + \log\frac{1}{\delta}}{n(1-\alpha)}\right].$$

B.2 Mutual information and the main-theorem proof

We are now ready to prove Theorem 3.2, based on ideas from page 51 of Catoni (2007).

Proof of Theorem 3.2. Starting from Theorem B.1, we have that:

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widehat{\rho}}[\mathcal{D}_\alpha(P_\theta\|P^*)] - \frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widehat{\rho}}[r_n(\theta)] \leq \frac{1}{n(1-\alpha)}\mathbb{E}_{\mathcal{S}}[\mathrm{KL}(\widehat{\rho}\|\pi)].$$

Let $\widehat{\rho}$ be any data-dependent posterior that is absolutely continuous with respect to π. Following Catoni (2007), $\mathbb{E}_{\mathcal{S}}[\widehat{\rho}]$ is the probability measure defined by:

$$\frac{d\mathbb{E}_{\mathcal{S}}[\widehat{\rho}]}{d\pi}(\theta) = \mathbb{E}_{\mathcal{S}}\left[\frac{d\widehat{\rho}}{d\pi}(\theta)\right].$$

Direct calculations show that:

$$\mathbb{E}_{\mathcal{S}}[\mathrm{KL}(\widehat{\rho}\|\pi)] = \mathbb{E}_{\mathcal{S}}\left[\mathrm{KL}(\widehat{\rho}\|\mathbb{E}_{\mathcal{S}}\left[\widehat{\rho}\right])\right] + \mathrm{KL}(\mathbb{E}_{\mathcal{S}}[\widehat{\rho}]\|\pi) = \mathcal{I}(\theta;\mathcal{S}) + \mathrm{KL}(\mathbb{E}_{\mathcal{S}}[\widehat{\rho}]\|\pi).$$

Using the above equality, we have that:

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widehat{\rho}}[\mathcal{D}_\alpha(P_\theta\|P^*)] - \frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widehat{\rho}}[r_n(\theta)] \leq \frac{1}{n(1-\alpha)}\left(\mathcal{I}(\theta;\mathcal{S}) + \mathrm{KL}(\mathbb{E}_{\mathcal{S}}[\widehat{\rho}]\|\pi)\right).$$

Choosing $\pi = \mathbb{E}_{\mathcal{S}}[\widehat{\rho}]$ sets $\mathrm{KL}(\mathbb{E}_{\mathcal{S}}[\widehat{\rho}]\|\pi) = 0$, the smallest possible value of this nonnegative term, and yields the bound of Theorem 3.2. □

In other words, MI bounds can be viewed as KL bounds optimized with respect to the choice of prior. This reflects the well-established relationship between relative entropy and mutual information (see, e.g., Melbourne et al. (2022, Proposition 2)).

The MI bound cannot be computed in practice, so the above bound is not directly computable. However, one can settle for a suboptimal prior π, said to be localized (Catoni, 2007; Alquier, 2024), by selecting a prior $\pi_{-\beta\mathcal{D}_\alpha}$ that approximates $\mathbb{E}_{\mathcal{S}}[\pi_{n,\alpha}]$.

B.3 Localized PAC-Bayes bounds and proofs

In our context, Catoni's localization approach can be motivated by considering the localized prior and then upper bounding $\mathrm{KL}(\widehat{\rho}\|\pi_{-\beta})$ via empirical bounds. The following theorem provides a bound on the localized prior under Assumption 3.3.

Theorem B.4. *Let $\alpha \in (0,1)$, and $\mathcal{F} \subset \mathcal{M}_1^+$, define the following posterior:*

$$\widetilde{\rho}_{n,\alpha} = \operatorname*{argmin}_{\rho\in\mathcal{F}}\left[\frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\theta\sim\rho}\left[r_n(\theta)\right] + \frac{\mathrm{KL}(\rho\|\pi)}{n(1-\alpha)}\right],$$

and the localized prior defined as in Assumption 3.4. If Assumption 3.3 is satisfied, then we have the following bound, for any $\beta < n(1-\alpha)/c(\alpha)$:

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widetilde{\rho}_{n,\alpha}}[\mathrm{KL}(P^*\|P_\theta)] \leq \frac{c(\alpha)n}{n(1-\alpha)-\beta c(\alpha)} \times \inf_{\rho\in\mathcal{F}}\left[\left(\alpha - \frac{\beta}{n}\right)\mathbb{E}_{\theta\sim\rho}\left[\mathrm{KL}(P^*\|P_\theta)\right] + \frac{\mathrm{KL}(\rho\|\pi_{-\beta})}{n}\right].$$

Proof. Consider the bound given in Theorem B.1, for the choice of $\pi = \pi_{-\beta}$, and for any $\rho \in \mathcal{F}$, we have that:

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widehat{\rho}}[\mathcal{D}_\alpha(P_\theta\|P^*)] \leq \mathbb{E}_{\mathcal{S}}\left[\frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\theta\sim\widehat{\rho}}[r_n(\theta)]\right] + \frac{1}{n(1-\alpha)}\mathbb{E}_{\mathcal{S}}[\mathrm{KL}(\widehat{\rho}\|\pi_{-\beta})].$$

Now consider the following decomposition:

$$\begin{aligned}
\mathrm{KL}(\widehat{\rho}\|\pi_{-\beta}) &= \mathbb{E}_{\theta\sim\widehat{\rho}}\left[\log\frac{\widehat{\rho}(\theta)}{\pi_{-\beta}(\theta)}\right] && \text{(B.1)}\\
&= \mathbb{E}_{\theta\sim\widehat{\rho}}\left[\log\frac{\widehat{\rho}(\theta)}{\pi(\theta)}\right] + \mathbb{E}_{\theta\sim\widehat{\rho}}\left[\log\frac{\pi(\theta)}{\pi_{-\beta}(\theta)}\right] && \text{(B.2)}\\
&= \mathrm{KL}(\widehat{\rho}\|\pi) + \mathbb{E}_{\theta\sim\widehat{\rho}}\left[\log\frac{\pi(\theta)}{\pi_{-\beta}(\theta)}\right] && \text{(B.3)}\\
&= \mathrm{KL}(\widehat{\rho}\|\pi) + \mathbb{E}_{\theta\sim\widehat{\rho}}\left[\log\left(\frac{\mathbb{E}_{\pi}[e^{-\beta\,\mathrm{KL}(P^*\|P_\theta)}]}{e^{-\beta\,\mathrm{KL}(P^*\|P_\theta)}}\right)\right] && \text{(B.4)}\\
&= \mathrm{KL}(\widehat{\rho}\|\pi) + \beta\mathbb{E}_{\theta\sim\widehat{\rho}}\left[\mathrm{KL}(P^*\|P_\theta)\right] + \log\mathbb{E}_{\pi}[e^{-\beta\,\mathrm{KL}(P^*\|P_\theta)}]. && \text{(B.5)}
\end{aligned}$$

Using the above decomposition, and using Assumption 3.3, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widehat{\rho}}[\mathcal{D}_\alpha(P_\theta\|P^*)]\\
&\leq \mathbb{E}_{\mathcal{S}}\left[\frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\theta\sim\widehat{\rho}}[r_n(\theta)]\right] + \frac{1}{n(1-\alpha)}\mathbb{E}_{\mathcal{S}}\Big[\mathrm{KL}(\widehat{\rho}\|\pi) + \beta\mathbb{E}_{\theta\sim\widehat{\rho}}\big[\mathrm{KL}(P^*\|P_\theta)\big]\\
&\qquad + \log\mathbb{E}_{\pi}[e^{-\beta\,\mathrm{KL}(P^*\|P_\theta)}]\Big]\\
&\leq \mathbb{E}_{\mathcal{S}}\left[\frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\theta\sim\widehat{\rho}}[r_n(\theta)]\right] + \frac{1}{n(1-\alpha)}\mathbb{E}_{\mathcal{S}}\Big[\mathrm{KL}(\widehat{\rho}\|\pi) + \beta c(\alpha)\,\mathbb{E}_{\theta\sim\widehat{\rho}}\big[\mathcal{D}_\alpha(P_\theta\|P^*)\big]\\
&\qquad + \log\mathbb{E}_{\pi}[e^{-\beta\,\mathrm{KL}(P^*\|P_\theta)}]\Big],
\end{aligned}$$

Rearranging the terms, we have that:

$$\begin{aligned}
&\left(1 - \frac{\beta c(\alpha)}{n(1-\alpha)}\right)\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widehat{\rho}}[\mathcal{D}_\alpha(P_\theta\|P^*)]\\
&\leq \frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\widehat{\rho}}[r_n(\theta)]\right] + \frac{1}{n(1-\alpha)}\mathbb{E}_{\mathcal{S}}\left[\mathrm{KL}(\widehat{\rho}\|\pi) + \log\mathbb{E}_{\pi}\left[e^{-\beta\,\mathrm{KL}(P^*\|P_\theta)}\right]\right].
\end{aligned}$$

Next, we use the definition of $\widetilde{\rho}_{n,\alpha}$, we have that:

$$\begin{aligned}
&\left(1 - \frac{\beta c(\alpha)}{n(1-\alpha)}\right)\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widetilde{\rho}_{n,\alpha}}[\mathcal{D}_\alpha(P_\theta\|P^*)] \leq\\
&\inf_{\rho\in\mathcal{F}}\Big[\frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\rho}\left[r_n(\theta)\right] + \frac{1}{n(1-\alpha)}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\rho}\big[\mathrm{KL}(\rho\|\pi)\\
&\qquad + \log\mathbb{E}_{\pi}\left[e^{-\beta\,\mathrm{KL}(P^*\|P_\theta)}\right]\big]\Big],
\end{aligned}$$

using Fubini and inverting the inf and the expectation with respect to $\mathcal{S}$, we have that:

$$\left(1 - \frac{\beta c(\alpha)}{n(1-\alpha)}\right)\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\widetilde{\rho}_{n,\alpha}}[\mathcal{D}_\alpha(P_\theta\|P^*)]$$

$$\leq \inf_{\rho \in \mathcal{F}} \Big[\frac{\alpha}{(1-\alpha)} \mathbb{E}_{\theta \sim \rho} \left[\mathrm{KL}(P^* \| P_\theta) \right] + \frac{1}{n(1-\alpha)} \mathbb{E}_{\theta \sim \rho} \left[\mathrm{KL}(\rho \| \pi) + \log \mathbb{E}_\pi \left[e^{-\beta \, \mathrm{KL}(P^* \| P_\theta)} \right] \right] \Big].$$

Using the decomposition of equation B.5, we have that:

$$\begin{aligned} &\left(1 - \frac{\beta c(\alpha)}{n(1-\alpha)} \right) \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\theta \sim \widetilde{\rho}_{n,\alpha}} [\mathcal{D}_\alpha(P_\theta \| P^*)] \\ &\leq \inf_{\rho \in \mathcal{F}} \Big[\frac{\alpha}{(1-\alpha)} \mathbb{E}_{\theta \sim \rho} \left[\mathrm{KL}(P^* \| P_\theta) \right] \\ &\quad + \frac{1}{n(1-\alpha)} \mathbb{E}_{\theta \sim \rho} \left[\mathrm{KL}(\rho \| \pi_{-\beta}) \right] - \beta \mathbb{E}_{\theta \sim \rho} \left[\mathrm{KL}(P^* \| P_\theta) \right] \Big]. \end{aligned}$$

Rearranging the terms, we have that:

$$\begin{aligned} &\left(1 - \frac{\beta c(\alpha)}{n(1-\alpha)} \right) \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\theta \sim \widetilde{\rho}_{n,\alpha}} [\mathcal{D}_\alpha(P_\theta \| P^*)] \\ &\qquad \leq \inf_{\rho \in \mathcal{F}} \Big[\left(\frac{\alpha - \frac{\beta}{n}}{1-\alpha} \right) \mathbb{E}_{\theta \sim \rho} \left[\mathrm{KL}(P^* \| P_\theta) \right] \\ &\qquad\quad + \frac{\mathrm{KL}(\rho \| \pi_{-\beta})}{n(1-\alpha)} \Big]. \end{aligned}$$

Using Assumption 3.3 again,

$$\begin{aligned} &\left(\frac{1}{c(\alpha)} - \frac{\beta}{n(1-\alpha)} \right) \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\theta \sim \widetilde{\rho}_{n,\alpha}} [\mathrm{KL}(P^* \| P_\theta)] \\ &\leq \inf_{\rho \in \mathcal{F}} \Big[\left(\frac{\alpha - \frac{\beta}{n}}{1-\alpha} \right) \mathbb{E}_{\theta \sim \rho} \left[\mathrm{KL}(P^* \| P_\theta) \right] \\ &\quad + \frac{\mathrm{KL}(\rho \| \pi_{-\beta})}{n(1-\alpha)} \Big]. \end{aligned}$$

We can divide both sides by $1/c(\alpha) - \beta/[n(1-\alpha)]$ on the condition that this quantity is > 0. □

Let us investigate the above bound. First, we note that when $\beta = 0$, we recover the bound of Corollary B.2. Thus, Theorem B.4 generalizes the bound of Corollary B.2.

We are now ready to prove Theorem 3.7.

Proof of Theorem 3.7. Apply Theorem B.4. If Assumption 3.4 is satisfied and $\mathcal{F} = \mathcal{M}_1^+$ (so that $\widetilde{\rho}_{n,\alpha} = \pi_{n,\alpha}$), we take $\rho = \pi_{-\beta} \in \mathcal{F}$, for which $\mathrm{KL}(\rho\|\pi_{-\beta}) = 0$. As β grows, the localized prior $\pi_{-\beta}$ concentrates on the Kullback-Leibler projection of P^* onto the model, so its average risk $\mathbb{E}_{\pi_{-\beta}}[\mathrm{KL}(P^*\|P_\theta)]$ cannot fall below $\inf_{\tilde{\theta}\in\Theta} \mathrm{KL}(P^*\|P_{\tilde{\theta}})$. Assumption 3.4 bounds the excess above this value, namely

$$\mathbb{E}_{\pi_{-\beta}}[\mathrm{KL}(P^*\|P_\theta)] \leq \inf_{\tilde{\theta}\in\Theta} \mathrm{KL}(P^*\|P_{\tilde{\theta}}) + \frac{d_\pi}{\beta^{\kappa_\pi}}.$$

Substituting into Theorem B.4 gives

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\pi_{n,\alpha}}[\mathrm{KL}(P^*\|P_\theta)] \leq \frac{c(\alpha)\alpha n - \beta c(\alpha)}{n(1-\alpha) - \beta c(\alpha)} \left(\inf_{\tilde{\theta}\in\Theta} \mathrm{KL}(P^*\|P_{\tilde{\theta}}) + \frac{d_\pi}{\beta^{\kappa_\pi}} \right).$$

For $\beta = n(1-\alpha)/(2c(\alpha))$, the prefactor equals $2c(\alpha)\alpha/(1-\alpha) - 1 \leq \alpha\, 2c(\alpha)/(1-\alpha)$, while $d_\pi/\beta^{\kappa_\pi} = d_\pi\big(2c(\alpha)/(1-\alpha)\big)^{\kappa_\pi} n^{-\kappa_\pi}$; this yields the closed-form bound of the theorem. If Assumption 3.6 holds instead, take the $\rho \in \mathcal{F}$ it provides, which satisfies $\mathbb{E}_\rho[\mathrm{KL}(P^*\|P_\theta)] - \inf_{\tilde{\theta}\in\Theta} \mathrm{KL}(P^*\|P_{\tilde{\theta}}) + \mathrm{KL}(\rho\|\pi_{-\beta})/n \leq d_\pi/\beta^{\kappa_\pi}$, and the same bound follows for $\widetilde{\rho}_{n,\alpha}$. □

Remark B.5. We now locate where the cost of misspecification enters. The mutual information bound of Theorem 3.2 carries no such cost: it relates the Rényi risk $\mathcal{D}_\alpha(P_\theta\|P^*)$ to the empirical log-likelihood ratio $r_n(\theta)$ and to the mutual information $\mathcal{I}(\theta;\mathcal{S})$, and its proof carries over, with P^* in place of P_{θ_0}, as soon as $P^* \ll \mu$. The two ingredients specific to Theorem 3.7 are responsible for the oracle term. First, the localized prior $\pi_{-\beta}$ can concentrate only on the best in-model approximation, not on P^* itself, so its risk floor is $\inf_{\tilde{\theta}\in\Theta} \mathrm{KL}(P^*\|P_{\tilde{\theta}})$ rather than zero. Second, Assumption 3.3 turns the Rényi risk into the Kullback-Leibler risk on both sides. In the well-specified case $P^* = P_{\theta_0}$ the infimum equals zero, attained at θ_0, and Theorem 3.7 reduces to a contraction bound of the type proved by Bhattacharya et al. (2019); Alquier & Ridgway (2020), without their extra $\log(n)$ factor.

Let us now state a version of Theorem B.4 that holds with large probability.

Theorem B.6. *Let $\alpha \in (0,1)$, $\delta, \eta \in (0,1)$, and $\mathcal{F} \subset \mathcal{M}_1^+$. If Assumption 3.3 is satisfied, then we have, for any $\beta < n(1-\alpha)/c(\alpha)$, with probability at least $1-\delta-\eta$ on the sample $\mathcal{S}$,*

$$\begin{aligned}\mathbb{E}_{\theta\sim\widetilde{\rho}_{n,\alpha}}[\mathrm{KL}(P^*\|P_\theta)] \\ \leq \frac{c(\alpha)n}{n(1-\alpha)-\beta c(\alpha)} \inf_{\rho\in\mathcal{F}} \Bigg[\left(\alpha - \frac{\beta}{n}\right) \mathbb{E}_{\theta\sim\rho}\left[\mathrm{KL}(P^*\|P_\theta)\right] \\ + \alpha\sqrt{\frac{\mathbb{E}_{\theta\sim\rho}\left[\mathcal{V}(\theta)\right]}{n\eta}} + \frac{\mathrm{KL}(\rho\|\pi_{-\beta}) + \log\frac{1}{\delta}}{n} \Bigg].\end{aligned}$$

Proof. Consider the bound given in Theorem B.3, for the choice of $\pi = \pi_{-\beta}$, and for any

$\rho \in \mathcal{F}$, with probability at least $1-\delta$ over $\mathcal{S}$,

$$\mathbb{E}_{\theta\sim\widehat{\rho}}[\mathcal{D}_\alpha(P_\theta\|P^*)] \leq \frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\theta\sim\widehat{\rho}}[r_n(\theta)] + \frac{\mathrm{KL}(\widehat{\rho}\|\pi_{-\beta}) + \log\frac{1}{\delta}}{n(1-\alpha)}.$$

Using the decomposition from the previous proof, and using Assumption 3.3, we have

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\widehat{\rho}}[\mathcal{D}_\alpha(P_\theta\|P^*)] \\
&\leq \frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\theta\sim\widehat{\rho}}[r_n(\theta)] + \frac{1}{n(1-\alpha)}\Big[\,\mathrm{KL}(\widehat{\rho}\|\pi) + \beta\mathbb{E}_{\theta\sim\widehat{\rho}}\big[\,\mathrm{KL}(P^*\|P_\theta)\big] \\
&\qquad + \log\mathbb{E}_\pi[e^{-\beta\,\mathrm{KL}(P^*\|P_\theta)}] + \log\frac{1}{\delta}\Big] \\
&\leq \frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\theta\sim\widehat{\rho}}[r_n(\theta)] + \frac{1}{n(1-\alpha)}\Big[\,\mathrm{KL}(\widehat{\rho}\|\pi) + \beta c(\alpha)\,\mathbb{E}_{\theta\sim\widehat{\rho}}\big[\mathcal{D}_\alpha(P_\theta\|P^*)\big] \\
&\qquad + \log\mathbb{E}_\pi[e^{-\beta\,\mathrm{KL}(P^*\|P_\theta)}] + \log\frac{1}{\delta}\Big],
\end{aligned}$$

Rearranging the terms, we have that:

$$\begin{aligned}
&\left(1 - \frac{\beta c(\alpha)}{n(1-\alpha)}\right)\mathbb{E}_{\theta\sim\widehat{\rho}}[\mathcal{D}_\alpha(P_\theta\|P^*)] \\
&\leq \frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\theta\sim\widehat{\rho}}[r_n(\theta)] + \frac{1}{n(1-\alpha)}\left[\mathrm{KL}(\widehat{\rho}\|\pi) + \log\mathbb{E}_\pi\left[e^{-\beta\,\mathrm{KL}(P^*\|P_\theta)}\right] + \log\frac{1}{\delta}\right].
\end{aligned}$$

Next, we use the definition of $\widetilde{\rho}_{n,\alpha}$, we have that:

$$\begin{aligned}
&\left(1 - \frac{\beta c(\alpha)}{n(1-\alpha)}\right)\mathbb{E}_{\theta\sim\widetilde{\rho}_{n,\alpha}}[\mathcal{D}_\alpha(P_\theta\|P^*)] \\
&\leq \inf_{\rho\in\mathcal{F}}\Big[\frac{\alpha}{n(1-\alpha)}\mathbb{E}_{\theta\sim\rho}\left[r_n(\theta)\right] + \frac{1}{n(1-\alpha)}\mathbb{E}_{\theta\sim\rho}\big[\,\mathrm{KL}(\rho\|\pi) \\
&\qquad\qquad + \log\mathbb{E}_\pi\left[e^{-\beta\,\mathrm{KL}(P^*\|P_\theta)}\right]\Big] + \log\frac{1}{\delta}\Big].
\end{aligned}$$

By Chebyshev's inequality and Fubini's theorem,

$$\mathbb{P}_{\mathcal{S}}\left[\mathbb{E}_{\theta\sim\rho}\left[r_n(\theta) - n\ \mathrm{KL}(P^*\|P_\theta)\right] \geq t\right] \leq \frac{\mathbb{V}_{\mathcal{S}}(\mathbb{E}_{\theta\sim\rho}[r_n(\theta)])}{t^2} \leq \frac{n\,\mathbb{E}_{\theta\sim\rho}[\mathcal{V}(\theta)]}{t^2},$$

where the last step uses, by Jensen's inequality and the bound of the variance by the second moment,

$$\mathbb{V}_{\mathcal{S}}(\mathbb{E}_{\theta\sim\rho}[r_n(\theta)]) \leq \mathbb{E}_{\theta\sim\rho}[\mathbb{V}_{\mathcal{S}}(r_n(\theta))] = n\,\mathbb{E}_{\theta\sim\rho}\left[\mathbb{V}_{P^*}\left(\log\frac{p^*}{p_\theta}\right)\right] \leq n\,\mathbb{E}_{\theta\sim\rho}[\mathcal{V}(\theta)].$$

Thus, with probability at least $1 - \eta$ on the sample $\mathcal{S}$,

$$\mathbb{E}_{\theta\sim\rho}\left[r_n(\theta) - n\,\mathrm{KL}(P^*\|P_\theta)\right] \leq \sqrt{n\frac{\mathbb{E}_{\theta\sim\rho}[\mathcal{V}(\theta)]}{\eta}}.$$

Thus, with probability at least $1 - \delta - \eta$ on $\mathcal{S}$,

$$\begin{aligned}\left(1 - \frac{\beta c(\alpha)}{n(1-\alpha)}\right)&\mathbb{E}_{\theta\sim\widetilde{\rho}_{n,\alpha}}[\mathcal{D}_\alpha(P_\theta\|P^*)] \\ &\leq \inf_{\rho\in\mathcal{F}}\left[\frac{\alpha}{1-\alpha}\left[\mathbb{E}_{\theta\sim\rho}\left[\mathrm{KL}(P^*\|P_\theta)\right] + \sqrt{\frac{\mathbb{E}_{\theta\sim\rho}[\mathcal{V}(\theta)]}{n\eta}}\right]\right. \\ &\quad\left. + \frac{1}{n(1-\alpha)}\mathbb{E}_{\theta\sim\rho}\left[\mathrm{KL}(\rho\|\pi) + \log\mathbb{E}_\pi\left[e^{-\beta\,\mathrm{KL}(P^*\|P_\theta)}\right]\right] + \log\frac{1}{\delta}\right].\end{aligned}$$

Using the decomposition of equation B.5, we have that:

$$\begin{aligned}\left(1 - \frac{\beta c(\alpha)}{n(1-\alpha)}\right)&\mathbb{E}_{\theta\sim\widetilde{\rho}_{n,\alpha}}[\mathcal{D}_\alpha(P_\theta\|P^*)] \\ &\leq \inf_{\rho\in\mathcal{F}}\left[\frac{\alpha}{1-\alpha}\left[\mathbb{E}_{\theta\sim\rho}\left[\mathrm{KL}(P^*\|P_\theta)\right] + \sqrt{\frac{\mathbb{E}_{\theta\sim\rho}[\mathcal{V}(\theta)]}{n\eta}}\right]\right. \\ &\quad\left. + \frac{1}{n(1-\alpha)}\mathbb{E}_{\theta\sim\rho}\left[\mathrm{KL}(\rho\|\pi_{-\beta})\right] - \beta\mathbb{E}_{\theta\sim\rho}\left[\mathrm{KL}(P^*\|P_\theta)\right] + \log\frac{1}{\delta}\right].\end{aligned}$$

Rearranging the terms, we have that:

$$\begin{aligned}\left(1 - \frac{\beta c(\alpha)}{n(1-\alpha)}\right)&\mathbb{E}_{\theta\sim\widetilde{\rho}_{n,\alpha}}[\mathcal{D}_\alpha(P_\theta\|P^*)] \\ &\leq \inf_{\rho\in\mathcal{F}}\left[\left(\frac{\alpha - \frac{\beta}{n}}{1-\alpha}\right)\mathbb{E}_{\theta\sim\rho}\left[\mathrm{KL}(P^*\|P_\theta)\right]\right. \\ &\quad + \frac{\alpha}{1-\alpha}\sqrt{\frac{\mathbb{E}_{\theta\sim\rho}[\mathcal{V}(\theta)]}{n\eta}} \\ &\quad\left. + \frac{\mathrm{KL}(\rho\|\pi_{-\beta}) + \log\frac{1}{\delta}}{n(1-\alpha)}\right].\end{aligned}$$

Using Assumption 3.3 again,

$$\begin{aligned}\left(\frac{1}{c(\alpha)} - \frac{\beta}{n(1-\alpha)}\right)&\mathbb{E}_{\theta\sim\widetilde{\rho}_{n,\alpha}}[\mathrm{KL}(P^*\|P_\theta)] \\ &\leq \inf_{\rho\in\mathcal{F}}\left[\left(\frac{\alpha - \frac{\beta}{n}}{1-\alpha}\right)\mathbb{E}_{\theta\sim\rho}\left[\mathrm{KL}(P^*\|P_\theta)\right]\right.\end{aligned}$$

$$+\frac{\alpha}{1-\alpha}\sqrt{\frac{\mathbb{E}_{\theta\sim\rho}[\mathcal{V}(\theta)]}{n\eta}}\\ +\frac{\mathrm{KL}(\rho\|\pi_{-\beta})+\log\frac{1}{\delta}}{n(1-\alpha)}\Big].$$

We can divide both sides by $1/c(\alpha)-\beta/[n(1-\alpha)]$ on the condition that this quantity is >0. □

We can now prove Theorem 3.10.

Proof of Theorem 3.10. Apply Theorem B.6 with $\rho=\pi_{-\beta}$, so that $\mathrm{KL}(\rho\|\pi_{-\beta})=0$. By Assumption 3.5,

$$\mathbb{E}_{\pi_{-\beta}}[\mathrm{KL}(P^*\|P_\theta)]\le\inf_{\tilde\theta\in\Theta}\mathrm{KL}(P^*\|P_{\tilde\theta})+\frac{d_\pi}{\beta^{\kappa_\pi}}\qquad\text{and}\qquad\mathbb{E}_{\pi_{-\beta}}[\mathcal{V}(\theta)]\le\mathcal{V}(\theta^\star)+\frac{d'_\pi}{\beta^{\kappa_\pi}}.$$

Bounding $(\alpha-\beta/n)/(1-\alpha)\le\alpha/(1-\alpha)$ and taking $\beta=n(1-\alpha)/(2c(\alpha))$, so that the prefactor $1/c(\alpha)-\beta/[n(1-\alpha)]$ equals $1/[2c(\alpha)]$, gives the announced bound. We keep the square-root term as is: the linearization

$$\sqrt{\mathbb{E}_{\pi_{-\beta}}[\mathcal{V}(\theta)]/(n\eta)}\le 2\,\mathbb{E}_{\pi_{-\beta}}[\mathcal{V}(\theta)]+\frac{1}{8n\eta},$$

used in the well-specified case, would replace the fluctuation by a term proportional to $\mathbb{E}_{\pi_{-\beta}}[\mathcal{V}(\theta)]$, which under misspecification does not vanish but tends to $\mathcal{V}(\theta^\star)>0$. □

B.4 Other proofs

Proof of Lemma 3.11. In order to calculate the α-divergence, we use the formulas in VanDerwerken & Schmidler (2013); Liese & Vajda (1987). In our setting where the variance matrices are equal, they give

$$\mathcal{D}_\alpha(P_\theta\|P_{\theta_0})=\frac{\alpha}{2v^2}\left\|\theta-\theta_0\right\|^2.$$

Moreover,

$$\mathrm{KL}(P_{\theta_0}\|P_\theta)=\frac{1}{2v^2}\left\|\theta-\theta_0\right\|^2.$$

This shows that

$$\mathrm{KL}(P_{\theta_0}\|P_\theta)=\frac{1}{2v^2}\left\|\theta-\theta_0\right\|^2=\frac{\frac{1}{\alpha}\alpha}{2v^2}\left\|\theta-\theta_0\right\|^2=\frac{1}{\alpha}\mathcal{D}_\alpha(P_\theta\|P_{\theta_0}).$$

□

Proof of Lemma 3.12. From the Proof of Lemma 3.11,

$$\mathrm{KL}(P_{\theta_0} \| P_\theta) = \frac{1}{2v^2} \|\theta - \theta_0\|^2 .$$

Our prior is $\pi = \mathcal{N}(0, \Sigma)$ for some positive definite matrix Σ. For any $\beta > 0$, we have that:

$$\begin{aligned}
\pi_{-\beta}(\theta) &\propto \exp\left(-\frac{\theta'\Sigma^{-1}\theta}{2} - \frac{\beta}{2v^2}\|\theta - \theta_0\|^2\right) \\
&\propto \exp\left(-\frac{\theta'\Sigma^{-1}\theta}{2} - \frac{\beta}{2v^2}\|\theta\|^2 + \frac{\beta}{v^2}\langle\theta, \theta_0\rangle\right) \\
&\propto \exp\left(-\frac{\theta'\left(\Sigma^{-1} + \frac{\beta}{v^2}I_d\right)\theta}{2} + \frac{\beta}{v^2}\langle\theta, \theta_0\rangle\right) \\
&\propto \exp\left(-\frac{\theta'\left(\Sigma^{-1} + \frac{\beta}{v^2}I_d\right)\theta}{2} + \frac{\beta}{v^2}\theta'\left(\Sigma^{-1} + \frac{\beta}{v^2}I_d\right)\left[\left(\Sigma^{-1} + \frac{\beta}{v^2}I_d\right)^{-1}\theta_0\right]\right).
\end{aligned}$$

Thus, we have that:

$$\pi_{-\beta}(\theta) \sim \mathcal{N}\left(\frac{\beta}{v^2}\left(\Sigma^{-1} + \frac{\beta}{v^2}I_d\right)^{-1}\theta_0, \left(\Sigma^{-1} + \frac{\beta}{v^2}I_d\right)^{-1}\right).$$

Thus,

$$\begin{aligned}
\beta\mathbb{E}_{\pi_{-\beta}}\left[\mathrm{KL}(P_{\theta_0}\|P_\theta)\right] &= \frac{\beta}{2v^2}\mathbb{E}_{\pi_{-\beta}}\left[\|\theta - \theta_0\|^2\right] \\
&= \frac{\beta}{2v^2}\left\|\left(\Sigma^{-1} + \frac{\beta}{v^2}I_d\right)^{-1}\Sigma^{-1}\theta_0\right\|^2 + \frac{\beta}{2v^2}\mathrm{Tr}\left[\left(\Sigma^{-1} + \frac{\beta}{v^2}I_d\right)^{-1}\right].
\end{aligned}$$

In particular, when $\Sigma = \sigma^2 I$,

$$\beta\mathbb{E}_{\pi_{-\beta}}\left[\mathrm{KL}(P_{\theta_0}\|P_\theta)\right] = \frac{\frac{\beta}{v^2}\|\theta_0\|^2}{2\sigma^4\left(\frac{\beta}{v^2} + \frac{1}{\sigma^2}\right)^2} + \frac{\frac{\beta}{v^2}d}{2\left(\frac{\beta}{v^2} + \frac{1}{\sigma^2}\right)} \leq \frac{d}{2} + \frac{\|\theta_0\|^2}{8\sigma^2}.$$

We finally check the last inequality in the lemma. In order to do so, check that the function $f(\beta) = \frac{\frac{\beta}{v^2}}{2\sigma^4\left(\frac{\beta}{v^2} + \frac{1}{\sigma^2}\right)^2}$ is maximized at $\beta_0 = \frac{v^2}{\sigma^2}$, and thus

$$\beta\mathbb{E}_{\pi_{-\beta}}\left[\mathrm{KL}(P_{\theta_0}\|P_\theta)\right] \leq \frac{d}{2} + \frac{\|\theta_0\|^2}{8\sigma^2}.$$

Thus, Assumption 3.4 is satisfied with $\kappa_\pi = 1$ and $d_\pi = d/2 + \|\theta_0\|^2/(8\sigma^2)$. □

Proof of Lemma 3.14. Write $u = \theta - \theta_0$ and $X = \theta_0 + v\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$, so that

$$\log \frac{p_{\theta_0}(X)}{p_\theta(X)} = \frac{\|X-\theta\|^2 - \|X-\theta_0\|^2}{2v^2} = \frac{\langle u,\, u - 2v\varepsilon\rangle}{2v^2} = \frac{\|u\|^2 - 2v\langle u, \varepsilon\rangle}{2v^2}.$$

Using $\mathbb{E}\langle u, \varepsilon\rangle = 0$ and $\mathbb{E}\langle u, \varepsilon\rangle^2 = \|u\|^2$,

$$\mathcal{V}(\theta) = \mathbb{E}\left[\log^2 \frac{p_{\theta_0}(X)}{p_\theta(X)}\right] = \frac{\|u\|^4 + 4v^2\|u\|^2}{4v^4} = \frac{\|\theta-\theta_0\|^2\big(\|\theta-\theta_0\|^2 + 4v^2\big)}{4v^4},$$

so that $\mathcal{V}(\theta^\star) = \mathcal{V}(\theta_0) = 0$. Hence

$$\beta\, \mathbb{E}_{\pi_{-\beta}}[\mathcal{V}(\theta)] = \frac{\beta\, \mathbb{E}_{\pi_{-\beta}}\big[\|u\|^4\big]}{4v^4} + \frac{\beta\, \mathbb{E}_{\pi_{-\beta}}\big[\|u\|^2\big]}{v^2}.$$

By Lemma 3.12 the second term equals $2\beta\, \mathbb{E}_{\pi_{-\beta}}[\mathrm{KL}(P_{\theta_0}\|P_\theta)] \le 2d_\pi$. For the first term, $\pi_{-\beta} = \mathcal{N}(m_\beta, s_\beta^2 I)$ with $s_\beta^2 = (\sigma^{-2} + \beta/v^2)^{-1}$ and $m_\beta \to \theta_0$, so $\mathbb{E}_{\pi_{-\beta}}[\|u\|^4] = \mathcal{O}(\beta^{-2})$ as $\beta \to \infty$ and stays bounded for β small; thus $\beta\, \mathbb{E}_{\pi_{-\beta}}[\|u\|^4]$ is bounded uniformly in β. Therefore $\sup_{\beta \ge 0} \beta\, \mathbb{E}_{\pi_{-\beta}}[\mathcal{V}(\theta)] \le d'_\pi$ for a finite constant d'_π, and Assumption 3.5 holds with $\kappa_\pi = 1$. □

Proof of Lemma 3.15. Lemma 3.12 gives

$$\beta \mathbb{E}_{\pi_{-\beta}}\left[\mathrm{KL}(P_{\theta_0}\|P_\theta)\right] = \frac{\beta}{2}\left\|\left(\Sigma^{-1} + \beta I_n\right)^{-1}\Sigma^{-1}\theta_0\right\|^2 + \frac{\beta}{2}\mathrm{Tr}\left[\left(\Sigma^{-1} + \beta I_n\right)^{-1}\right].$$

We now use the diagonal form of Σ to get

$$\begin{aligned}
\beta \mathbb{E}_{\pi_{-\beta}}\left[\mathrm{KL}(P_{\theta_0}\|P_\theta)\right] &= \sum_{i=1}^n \left\{ \frac{\frac{\theta_{0,i}^2}{\sigma_i^2}\beta}{2(\frac{1}{\sigma_i}+\beta)^2} + \frac{\beta}{2\left(\frac{1}{\sigma_i}+\beta\right)} \right\} \\
&= \sum_{\frac{1}{\sigma_i} \le \beta} \left\{ \frac{\frac{\theta_{0,i}^2}{\sigma_i^2}\beta}{2(\frac{1}{\sigma_i}+\beta)^2} + \frac{\beta}{2\left(\frac{1}{\sigma_i}+\beta\right)} \right\} \\
&\quad + \sum_{\frac{1}{\sigma_i} > \beta} \left\{ \frac{\frac{\theta_{0,i}^2}{\sigma_i^2}\beta}{2(\frac{1}{\sigma_i}+\beta)^2} + \frac{\beta}{2\left(\frac{1}{\sigma_i}+\beta\right)} \right\} \\
&\le \sum_{\frac{1}{\sigma_i} \le \beta} \left\{ \frac{\theta_{0,i}^2}{\sigma_i^2\beta} + \frac{1}{2} \right\} + \sum_{\frac{1}{\sigma_i} > \beta} \left\{ \frac{\beta\theta_{0,i}^2}{2} + \frac{\beta\sigma_i}{2} \right\}.
\end{aligned}$$

We now use explicitly $\sigma_i = i^{-1-2b}$. We have $1/\sigma_i \leq \beta \Leftrightarrow i^{1+2b} \leq \beta \Leftrightarrow i \leq \beta^{1/(1+2b)}$. Thus,

$$\begin{aligned}\sum_{\frac{1}{\sigma_i}\leq\beta} \frac{\theta_{0,i}^2}{\sigma_i^2\beta} &= \frac{1}{\beta}\sum_{1\leq i\leq\beta^{1/(1+2b)}} \theta_{0,i}^2 i^{2(1+2b)} \\ &\leq \frac{1}{\beta}\sum_{1\leq i\leq\beta^{1/(1+2b)}} \theta_{0,i}^2 i^{2b}\beta^{(2+2b)/(1+2b)} \\ &\leq \beta^{1/(1+2b)}\sum_{i=1}^{\infty}\theta_{0,i}^2 i^{2b} \\ &\leq L\beta^{1/(1+2b)}.\end{aligned}$$

Then,

$$\sum_{\frac{1}{\sigma_i}\leq\beta}\frac{1}{2} = \sum_{1\leq i\leq\beta^{1/(1+2b)}}\frac{1}{2} \leq \frac{1}{2}\beta^{1/(1+2b)}.$$

Then,

$$\sum_{\frac{1}{\sigma_i}>\beta}\frac{\beta\theta_{0,i}^2}{2} = \frac{\beta}{2}\sum_{\beta^{1/(1+2b)}<i\leq n}\theta_{0,i}^2 \leq \frac{\beta}{2}\sum_{i>\beta^{1/(1+2b)}}\frac{\theta_{0,i}^2 i^{2b}}{\beta^{2b/(1+2b)}} \leq \frac{L}{2}\beta^{1/(1+2b)}.$$

Finally, using a sum-integral comparison inequality,

$$\sum_{\frac{1}{\sigma_i}>\beta}\frac{\beta\sigma_i}{2} \leq \frac{\beta}{2}\sum_{i>\beta^{1/(1+2b)}}\frac{1}{i^{1+2b}} \leq \frac{\beta}{4b}\frac{1}{\beta^{2b/(1+2b)}} = \frac{1}{4b}\beta^{1/(1+2b)}.$$

So we can conclude that

$$\beta\mathbb{E}_{\pi_{-\beta}}\left[\mathrm{KL}(P_\theta\|P_{\theta_0})\right] \leq \left(\frac{3L}{2}+\frac{1}{2}+\frac{1}{4b}\right)\beta^{1/(1+2b)}.$$

In other words,

$$\sup_\beta \beta^{\frac{2b}{1+2b}}\mathbb{E}_{\pi_{-\beta}}\left[\mathrm{KL}(P_\theta\|P_{\theta_0})\right] \leq \left(\frac{3L}{2}+\frac{1}{2}+\frac{1}{4b}\right).$$

□

Proof of Lemma 3.18. The α-divergence can be calculated for any member of the exponential family. Consider the case where $\mathcal{X} = \mathbb{R}^d$, and μ is the Lebesgue measure on $\mathbb{R}^d$. In this case, we have that:

$$\psi(\theta) = \log\left(\int_{\mathbb{R}^d}\exp\left(h(x)+\langle T(x),\theta\rangle\right)dx\right).$$

From (Nielsen & Nock, 2011), the following holds:

$$\mathcal{D}_\alpha(P_\theta \| P_{\theta_0}) = \frac{1}{1-\alpha}\left(\alpha\psi(\theta) + (1-\alpha)\psi(\theta_0) - \psi(\alpha\theta + (1-\alpha)\theta_0)\right),$$
$$\mathrm{KL}(P_{\theta_0} \| P_\theta) = \psi(\theta_0) - \psi(\theta) - \langle \theta_0 - \theta, \nabla\psi(\theta)\rangle.$$

As ψ is m-strongly convex, we have that:

$$\psi(\alpha\theta + (1-\alpha)\theta_0) \leq \alpha\psi(\theta) + (1-\alpha)\psi(\theta_0) - \frac{m}{2}\alpha(1-\alpha)\left\|\theta_0 - \theta\right\|^2.$$

This yields that for any $\alpha \in (0,1)$, we have that:

$$\mathcal{D}_\alpha(P_\theta \| P_{\theta_0}) \geq \frac{m\alpha}{2}\left\|\theta_0 - \theta\right\|^2.$$

Similarly, if ψ has a Lipschitz gradient with constant L, we have that:

$$\left\|\nabla\psi(\theta_0) - \nabla\psi(\theta)\right\| \leq L\left\|\theta_0 - \theta\right\|,$$

which implies

$$\psi(\theta_0) - \psi(\theta) - \langle \theta_0 - \theta, \nabla\psi(\theta)\rangle \leq \frac{L}{2}\left\|\theta_0 - \theta\right\|^2.$$

This in turn yields:

$$\mathrm{KL}(P_{\theta_0} \| P_\theta) \leq \frac{L}{2}\left\|\theta_0 - \theta\right\|^2.$$

We obtain finally:

$$\mathrm{KL}(P_{\theta_0} \| P_\theta) \leq \frac{L}{2}\left\|\theta_0 - \theta\right\|^2 = \frac{m\kappa}{2}\left\|\theta_0 - \theta\right\|^2 = \frac{m\frac{\kappa}{\alpha}\alpha}{2}\left\|\theta_0 - \theta\right\|^2 \leq \frac{\kappa}{\alpha}\mathcal{D}_\alpha(P_\theta \| P_{\theta_0}).$$

□

Proof of Theorem 3.23. Write $\ell_\theta := \log p_\theta$, let ∂_θ denote differentiation of the log-density in θ and primes differentiation of functions of θ alone, and set $I_0 := I(\theta_0)$. For fixed θ_0 and α, the three divergences are functions of θ:

$$\Delta_H(\theta) := H^2(P_\theta, P_{\theta_0}), \qquad \Delta_K(\theta) := \mathrm{KL}(P_{\theta_0} \| P_\theta), \qquad \Delta_R(\theta) := \mathcal{D}_\alpha(P_\theta \| P_{\theta_0}). \tag{B.6}$$

The Kullback-Leibler divergence can be written as $\Delta_K(\theta) = \mathbb{E}_{P_{\theta_0}}\left[\ell_{\theta_0}(X) - \ell_\theta(X)\right]$. The two

others can be expressed as a function of the affinity of order $\gamma \in (0,1)$,

$$A_\gamma(\theta) := \int p_\theta^\gamma p_{\theta_0}^{1-\gamma}\, d\mu = \mathbb{E}_{P_{\theta_0}}\left[\exp\left(\gamma\left(\ell_\theta(X) - \ell_{\theta_0}(X)\right)\right)\right], \tag{B.7}$$

through

$$\Delta_H = 2\left(1 - A_{1/2}\right), \qquad \Delta_R = \frac{1}{\alpha - 1}\log A_\alpha. \tag{B.8}$$

Both Δ_K and A_γ are expectations under P_{θ_0} of explicit functions of ℓ_θ, so their derivatives at θ_0 follow from the score identities alone.

Differentiating $\int p_\theta\, d\mu = 1$ under the integral sign, which Assumption 3.22(iii) permits by dominated convergence, and using $\partial_\theta p_\theta = p_\theta\, \partial_\theta \ell_\theta$, gives the score identities

$$\mathbb{E}_{P_\theta}\left[\partial_\theta \ell_\theta(X)\right] = 0, \qquad \mathbb{E}_{P_\theta}\left[\left(\partial_\theta \ell_\theta(X)\right)^2 + \partial_\theta^2 \ell_\theta(X)\right] = 0, \tag{B.9}$$

for $\theta \in U$, and at θ_0, by Definition 3.21,

$$\mathbb{E}_{P_{\theta_0}}\left[\partial_\theta \ell_{\theta_0}(X)\right] = 0, \qquad \mathbb{E}_{P_{\theta_0}}\left[\partial_\theta^2 \ell_{\theta_0}(X)\right] = -I_0. \tag{B.10}$$

Fix $\gamma \in \{\alpha, 1/2\}$ and write $g_\theta(x) := \exp\left(\gamma\left(\ell_\theta(x) - \ell_{\theta_0}(x)\right)\right)$, so that $A_\gamma(\theta) = \mathbb{E}_{P_{\theta_0}}\left[g_\theta(X)\right]$ by (B.7), with $g_{\theta_0} = 1$. Differentiating in θ gives

$$\partial_\theta g_\theta = \gamma\, g_\theta\, \partial_\theta \ell_\theta, \qquad \partial_\theta^2 g_\theta = \gamma\, g_\theta\left(\partial_\theta^2 \ell_\theta + \gamma\left(\partial_\theta \ell_\theta\right)^2\right). \tag{B.11}$$

Taking expectations under P_{θ_0} and evaluating at θ_0, where $g_{\theta_0} = 1$, (B.10) yields

$$\begin{aligned} A_\gamma'(\theta_0) &= \gamma\, \mathbb{E}_{P_{\theta_0}}\left[\partial_\theta \ell_{\theta_0}(X)\right] = 0, \\ A_\gamma''(\theta_0) &= \gamma\left(-I_0 + \gamma I_0\right) = -\gamma(1-\gamma) I_0, \end{aligned} \tag{B.12}$$

and $A_\gamma(\theta_0) = 1$ follows from (B.7).

Next, notice from $\Delta_K(\theta) = \mathbb{E}_{P_{\theta_0}}\left[\ell_{\theta_0}(X) - \ell_\theta(X)\right]$ that

$$\Delta_K'(\theta) = -\mathbb{E}_{P_{\theta_0}}\left[\partial_\theta \ell_\theta(X)\right], \qquad \Delta_K''(\theta) = -\mathbb{E}_{P_{\theta_0}}\left[\partial_\theta^2 \ell_\theta(X)\right], \tag{B.13}$$

and (B.10) gives $\Delta_K'(\theta_0) = 0$ and $\Delta_K''(\theta_0) = I_0$; the value $\Delta_K(\theta_0) = 0$ is immediate.

Finally, by (B.8) and (B.12) read at $\gamma = 1/2$, we have that $\Delta_H(\theta_0) = \Delta_H'(\theta_0) = 0$ and

$$\Delta_H''(\theta_0) = -2A_{1/2}''(\theta_0) = 2 \cdot \tfrac{1}{2} \cdot \tfrac{1}{2}\, I_0 = \frac{I_0}{2}. \tag{B.14}$$

The Rényi curve is defined as $\Delta_R = (\alpha - 1)^{-1} \log A_\alpha$ where $A_\alpha > 0$, which holds on a certain

neighborhood of θ_0 by continuity, and there

$$\Delta_R' = \frac{1}{\alpha-1}\cdot\frac{A_\alpha'}{A_\alpha}, \qquad \Delta_R'' = \frac{1}{\alpha-1}\cdot\frac{A_\alpha'' A_\alpha - (A_\alpha')^2}{A_\alpha^2}. \tag{B.15}$$

Inserting (B.12) at $\gamma = \alpha$ gives $\Delta_R(\theta_0) = \Delta_R'(\theta_0) = 0$ and

$$\Delta_R''(\theta_0) = \frac{-\alpha(1-\alpha)I_0}{\alpha-1} = \alpha\, I_0. \tag{B.16}$$

Thus $\Delta_j(\theta_0) = \Delta_j'(\theta_0) = 0$ and $\Delta_j''(\theta_0) = 2b_j$ for $j \in \{H, K, R\}$, with $b_H = I_0/4$, $b_K = I_0/2$ and $b_R = \alpha I_0/2$.

Choose $r > 0$ with $U_0 := [\theta_0 - r, \theta_0 + r] \subset U$ and $A_\alpha > 0$ on U_0, which is valid since $A_\alpha(\theta_0) = 1$ and A_α is continuous. On U_0 each curve Δ_j is three times continuously differentiable, by Assumption 3.22(iii) and (B.8), and we set

$$C_j := \frac{1}{6}\sup_{t\in U_0}\left|\Delta_j'''(t)\right|, \qquad j \in \{H, K, R\}, \tag{B.17}$$

which is finite since Δ_j''' is continuous on the compact U_0. Fix an index j and $\theta \in U_0$. Since $\Delta_j(\theta_0) = \Delta_j'(\theta_0) = 0$ and $\Delta_j''(\theta_0) = 2b_j$, Taylor's formula with integral remainder reads

$$\Delta_j(\theta) = b_j(\theta-\theta_0)^2 + R_j(\theta), \qquad R_j(\theta) := \frac{1}{2}\int_{\theta_0}^{\theta}(\theta - t)^2\Delta_j'''(t)\,dt, \tag{B.18}$$

and the remainder is bounded by

$$|R_j(\theta)| \le \frac{1}{2}\sup_{t\in U_0}\left|\Delta_j'''(t)\right|\cdot\left|\int_{\theta_0}^{\theta}(\theta-t)^2 dt\right| = \frac{1}{2}\sup_{t\in U_0}\left|\Delta_j'''(t)\right|\cdot\frac{|\theta-\theta_0|^3}{3} = C_j|\theta-\theta_0|^3. \tag{B.19}$$

For $j = H, K, R$ this is the statement of the theorem. The driving force behind (B.17) is that we need only a bound on the third derivative to control the cubic remainder. □

Proof of Corollary 3.24. With b_H, b_K and b_R as in (3.11), the three displays of the corollary read

$$(1-\eta)\,b_j(\theta-\theta_0)^2 \;\le\; \Delta_j(\theta) \;\le\; (1+\eta)\,b_j(\theta-\theta_0)^2, \qquad j \in \{H, K, R\}, \tag{B.20}$$

with Δ_j as in (B.6). Each b_j is positive by Assumption 3.22(iv), so $r_{\alpha,\eta} \in (0, r]$. Fix j and θ with $|\theta - \theta_0| \le r_{\alpha,\eta}$. If $C_j = 0$, then $R_j(\theta) = 0$ by (B.19), and (B.20) holds with equality on both sides. If $C_j > 0$, then $r_{\alpha,\eta} \le \eta b_j/C_j$, and (B.19) gives

$$|R_j(\theta)| \le C_j|\theta-\theta_0|^3 = C_j|\theta-\theta_0|\cdot(\theta-\theta_0)^2 \le C_j\, r_{\alpha,\eta}(\theta-\theta_0)^2 \le \eta\, b_j(\theta-\theta_0)^2, \tag{B.21}$$

which yields (B.20), since $\Delta_j = b_j(\theta-\theta_0)^2 + R_j$ by (B.18). For the last display of the corollary, we take $\theta \neq \theta_0$ and divide the upper bound of (B.20) at $j = K$ by its lower bound at $j = R$:

$$\frac{\mathrm{KL}(P_{\theta_0}\|P_\theta)}{\mathcal{D}_\alpha(P_\theta\|P_{\theta_0})} \leq \frac{(1+\eta)b_K}{(1-\eta)b_R} = \frac{1+\eta}{1-\eta}\cdot\frac{I(\theta_0)/2}{\alpha I(\theta_0)/2} = \frac{1+\eta}{\alpha(1-\eta)}. \tag{B.22}$$

At $\theta = \theta_0$ both sides vanish. The mechanism is only that $|\theta-\theta_0|^3/(\theta-\theta_0)^2 = |\theta-\theta_0|$ tends to zero. □

Proof of Lemma 3.26. Fix $\beta > 0$ and write $F(\beta) := \mathbb{E}_{\theta\sim\pi_{-\beta}}[\mathrm{KL}(P_{\theta_0}\|P_\theta)]$. Since π is uniform and $\Delta_K(\theta) = \mathrm{KL}(P_{\theta_0}\|P_\theta)$,

$$F(\beta) = \frac{\displaystyle\int_{-M}^{M} \Delta_K(\theta)e^{-\beta\Delta_K(\theta)}d\theta}{\displaystyle\int_{-M}^{M} e^{-\beta\Delta_K(\theta)}d\theta}. \tag{B.23}$$

Use the change of variable $x = \theta - \theta_0$, whose range is $[a, b]$ with $a := -M - \theta_0 < 0$ and $b := M - \theta_0 > 0$.

We use (3.16) in three separate instances, each time in the direction that increases the ratio (B.23): the upper bound $\Delta_K \leq Lx^2/2$ on the factor Δ_K in the numerator, the lower bound $\Delta_K \geq mx^2/2$ in the exponent of the numerator, and the upper bound $\Delta_K \leq Lx^2/2$ in the exponent of the denominator. Together they yield

$$F(\beta) \;\leq\; \frac{L}{2}\cdot\frac{\displaystyle\int_a^b x^2 e^{-\beta m x^2/2}dx}{\displaystyle\int_a^b e^{-\beta L x^2/2}dx}. \tag{B.24}$$

Writing $x^2e^{-\beta mx^2/2} = x\cdot\left(xe^{-\beta mx^2/2}\right)$ and using

$$\frac{d}{dx}e^{-\beta mx^2/2} = -\beta m\, x\, e^{-\beta mx^2/2}, \tag{B.25}$$

integration by parts gives

$$\int_a^b x^2 e^{-\beta mx^2/2}dx = -\frac{1}{\beta m}\left[x\, e^{-\beta mx^2/2}\right]_a^b + \frac{1}{\beta m}\int_a^b e^{-\beta mx^2/2}dx. \tag{B.26}$$

The bracket equals $be^{-\beta mb^2/2} - ae^{-\beta ma^2/2}$, which is positive because $a < 0 < b$, so the first term on the right of (B.26) is negative and

$$\int_a^b x^2 e^{-\beta mx^2/2}dx \;\leq\; \frac{1}{\beta m}\int_a^b e^{-\beta mx^2/2}dx. \tag{B.27}$$

Substituting $u = \sqrt{\beta m}\, x$ in the first integral and $u = \sqrt{\beta L}\, x$ in the second,

$$\int_a^b e^{-\beta m x^2/2} dx = \frac{1}{\sqrt{\beta m}} \int_{\sqrt{\beta m}\, a}^{\sqrt{\beta m}\, b} e^{-u^2/2} du, \qquad \int_a^b e^{-\beta L x^2/2} dx = \frac{1}{\sqrt{\beta L}} \int_{\sqrt{\beta L}\, a}^{\sqrt{\beta L}\, b} e^{-u^2/2} du. \tag{B.28}$$

Since $a < 0 < b$ and $m \le L$, we have $\left[\sqrt{\beta m}\, a, \sqrt{\beta m}\, b\right] \subseteq \left[\sqrt{\beta L}\, a, \sqrt{\beta L}\, b\right]$, and hence, since $e^{-u^2/2} > 0$,

$$\int_{\sqrt{\beta m}\, a}^{\sqrt{\beta m}\, b} e^{-u^2/2} du \;\le\; \int_{\sqrt{\beta L}\, a}^{\sqrt{\beta L}\, b} e^{-u^2/2} du. \tag{B.29}$$

Dividing,

$$\frac{\displaystyle\int_a^b e^{-\beta m x^2/2} dx}{\displaystyle\int_a^b e^{-\beta L x^2/2} dx} \;\le\; \sqrt{\frac{\beta L}{\beta m}} \;=\; \sqrt{\kappa}. \tag{B.30}$$

By (B.24), (B.27) and (B.30),

$$F(\beta) \;\le\; \frac{L}{2} \cdot \frac{1}{\beta m} \cdot \sqrt{\kappa} \;=\; \frac{\kappa^{3/2}}{2\beta}. \tag{B.31}$$

Multiplying by β and taking the supremum over $\beta > 0$ gives the claim, so Assumption 3.4 holds with $\kappa_\pi = 1$ and $d_\pi = \kappa^{3/2}/2$, as stated; only the proof changes. No crossover point β_0 and no monotonicity of $\beta \mapsto \beta F(\beta)$ is used. □

Proof of Corollary 3.27. Fix $\varepsilon > 0$, put $\mathcal{N}_\varepsilon = \mathcal{N}(\Theta, \varepsilon)$ for short and let $\theta_1, \dots, \theta_{\mathcal{N}_\varepsilon} \in \Theta$ be such that $\Theta \subset B(\theta_1, \varepsilon) \cup \cdots \cup B(\theta_{\mathcal{N}_\varepsilon}, \varepsilon)$. We define

$$\pi = \frac{1}{\mathcal{N}_\varepsilon} \sum_{i=1}^{\mathcal{N}_\varepsilon} \delta_{\theta_i}$$

and $\widehat{\rho} = \delta_{\theta_{\hat{i}}}$ where

$$\hat{i} = \operatorname*{argmin}_{1 \le i \le \mathcal{N}_\varepsilon} \|\theta_i - \hat{\theta}_n\|,$$

in other words, $\widehat{\rho}$ is a Dirac mass on the closest parameter to the MLE among the $\theta_1, \dots, \theta_{\mathcal{N}_\varepsilon}$. Theorem 3.2 states

$$\mathbb{E}_{\mathcal{S}} \left[\mathbb{E}_{\theta \sim \widehat{\rho}}[\mathcal{D}_\alpha(P_\theta \| P_{\theta_0})] - \frac{\alpha}{n(1-\alpha)} \mathbb{E}_{\theta \sim \widehat{\rho}}[r_n(\theta, \theta_0)] \right] \le \frac{\mathcal{I}(\theta; \mathcal{S})}{n(1-\alpha)},$$

and thus, in our setting, this gives

$$\mathbb{E}_{\mathcal{S}}\left[\mathcal{D}_\alpha(P_{\theta_{\hat{i}}}\|P_{\theta_0}) - \frac{\alpha}{n(1-\alpha)} r_n(\theta_{\hat{i}}, \theta_0)\right] \leq \frac{\log \mathcal{N}_\varepsilon}{n(1-\alpha)}$$

and thus

$$\mathbb{E}_{\mathcal{S}}\left[\mathcal{D}_\alpha(P_{\theta_{\hat{i}}}\|P_{\theta_0})\right] \leq \mathbb{E}_{\mathcal{S}}\left[\frac{\alpha}{n(1-\alpha)} r_n(\theta_{\hat{i}}, \theta_0)\right] + \frac{\log \mathcal{N}_\varepsilon}{n(1-\alpha)}.$$

By the Lipschitz assumption,

$$\frac{\alpha}{n(1-\alpha)} r_n(\theta_{\hat{i}}, \theta_0) \leq \frac{\alpha}{n(1-\alpha)} r_n(\hat{\theta}_n, \theta_0) + L\|\theta_{\hat{i}} - \hat{\theta}_n\| \leq \frac{\alpha}{n(1-\alpha)} r_n(\hat{\theta}_n, \theta_0) + L\varepsilon.$$

Moreover, by definition of $\hat{\theta}_n$, $r_n(\hat{\theta}_n, \theta_0) \leq r_n(\theta_0, \theta_0) = 0$. Thus,

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\left[\mathcal{D}_\alpha(P_{\theta_{\hat{i}}}\|P_{\theta_0})\right] &\leq \mathbb{E}_{\mathcal{S}}\left[\frac{\alpha}{n(1-\alpha)} r_n(\hat{\theta}_n, \theta_0)\right] + \frac{L\varepsilon\alpha}{(1-\alpha)} + \frac{\log \mathcal{N}_\varepsilon}{n(1-\alpha)} \\ &\leq \frac{L\varepsilon\alpha}{(1-\alpha)} + \frac{\log \mathcal{N}_\varepsilon}{n(1-\alpha)}.\end{aligned}$$

Finally, note that

$$\|\hat{\theta}_n - \theta_0\|^2 \leq 2\|\theta_{\hat{i}} - \theta_0\|^2 + 2\|\theta_{\hat{i}} - \hat{\theta}_n\|^2 \leq 2\|\theta_{\hat{i}} - \theta_0\|^2 + 2\varepsilon^2$$

and as, by assumption, $\mathcal{D}_\alpha(P_{\theta_{\hat{i}}}\|P_{\theta_0}) \geq m\|\theta_{\hat{i}} - \theta_0\|^2$, we obtain:

$$\mathcal{D}_\alpha(P_{\theta_{\hat{i}}}\|P_{\theta_0}) \geq m\left(\|\hat{\theta}_n - \theta_0\|^2 - \varepsilon^2\right).$$

Putting everything together gives:

$$\mathbb{E}_{\mathcal{S}}\left[\|\hat{\theta}_n - \theta_0\|^2\right] \leq \varepsilon^2 + \frac{2L\varepsilon\alpha}{m(1-\alpha)} + \frac{2\log \mathcal{N}_\varepsilon}{nm(1-\alpha)}.$$

Take $\varepsilon = 1/n$ to end the proof. □

Appendix C

Proofs of Chapter 4

This appendix collects the proofs of the results stated in Chapter 4. Notation follows the chapter; in particular, the sample Hellinger distance (4.1), the variance functional (4.6), and the Hellinger comparison of Lemma 4.4 are used throughout.

C.1 Technical preliminaries

Lemma C.1 (Bernstein inequality for independent bounded random variables)**.** *Let $Y_1, \dots, Y_n$ be independent real-valued random variables with $\mathbb{E}[Y_i] = 0$ for all i, and suppose $|Y_i| \le C$ almost surely for some $C > 0$. Define the total variance $\sigma^2 := \sum_{i=1}^n \mathbb{V}(Y_i)$ and the function*

$$g(x) := \begin{cases} \dfrac{e^x - 1 - x}{x^2} & x \neq 0, \\ \dfrac{1}{2} & x = 0. \end{cases} \tag{C.1}$$

Then for all $t \in \mathbb{R}$,

$$\log \mathbb{E}\left[\exp\left(t \sum_{i=1}^n Y_i\right)\right] \le \sigma^2 t^2 g(C|t|). \tag{C.2}$$

Proof. See Boucheron et al. (2013). The moment generating function factorizes by independence, and each factor is bounded via Bernstein's lemma for a single bounded variable. □

C.2 Proof of the PAC-Bayes bound

Proof of Theorem 4.5. Fix $(\theta, \theta') \in \Theta^2$. For each $i \in \{1, \dots, n\}$, define the random variable

$$U_i(\theta, \theta') := \ell_\psi(X_i; \theta, \theta') = \psi\left(\frac{p^i_{\theta'}(X_i)}{p^i_\theta(X_i)}\right) \in [-1, 1], \tag{C.3}$$

and its centered version

$$Y_i(\theta,\theta') := U_i(\theta,\theta') - \mathbb{E}_{X_i\sim P_*^i}[U_i(\theta,\theta')]. \tag{C.4}$$

Since $|U_i| \le 1$, we have $|Y_i| \le 2$ almost surely. The variables $\{Y_i\}_{i=1}^n$ are independent with $\mathbb{E}[Y_i] = 0$, and

$$\sum_{i=1}^n Y_i(\theta,\theta') = n\big(\widehat{R}_\psi(\theta,\theta') - R_\psi(\theta,\theta')\big).$$

By independence,

$$\mathbb{V}\left[\sum_{i=1}^n Y_i(\theta,\theta')\right] = \sum_{i=1}^n \mathbb{V}[Y_i(\theta,\theta')] = n^2\, V_\psi(\theta,\theta'),$$

where the last equality uses the definition (4.6). The individual variances $\mathbb{V}_{X_i\sim P_*^i}[\ell_\psi(X_i;\theta,\theta')]$ may differ across coordinates; in the i.i.d. case, the sum reduces to n times a common variance.

Applying Lemma C.1 with $C = 2$ and $t = \lambda/n$,

$$\log \mathbb{E}_{\mathcal{S}} \exp\Big(\lambda(\widehat{R}_\psi - R_\psi)\Big) \le \lambda^2 g\left(\frac{2\lambda}{n}\right) V_\psi(\theta,\theta'). \tag{C.5}$$

Applying the same bound to $-Y_i$,

$$\log \mathbb{E}_{\mathcal{S}} \exp\Big(\lambda(R_\psi - \widehat{R}_\psi)\Big) \le \lambda^2 g\left(\frac{2\lambda}{n}\right) V_\psi(\theta,\theta'). \tag{C.6}$$

Equivalently,

$$\mathbb{E}_{\mathcal{S}} \exp\left(\lambda(R_\psi - \widehat{R}_\psi) - \lambda^2 g\left(\frac{2\lambda}{n}\right) V_\psi\right) \le 1. \tag{C.7}$$

Integrating (C.7) over $(\theta,\theta') \sim \pi\otimes\pi'$ and applying Markov's inequality, with probability at least $1-\delta$,

$$\int \exp\left(\lambda(R_\psi - \widehat{R}_\psi) - \lambda^2 g\left(\frac{2\lambda}{n}\right) V_\psi\right) d(\pi\otimes\pi') \le \frac{1}{\delta}.$$

Taking logarithms and applying the Donsker–Varadhan variational formula (1.24) with $h(\theta,\theta') := \lambda(R_\psi - \widehat{R}_\psi) - \lambda^2 g(2\lambda/n) V_\psi$ and prior $\pi\otimes\pi'$, for any $\mu = \rho\otimes\rho'$ with $\rho \ll \pi$ and $\rho' \ll \pi'$,

$$\int h(\theta,\theta')\,\mu(d\theta, d\theta') - \mathrm{KL}(\mu\|\pi\otimes\pi') \le \log(1/\delta). \tag{C.8}$$

The KL divergence decomposes as $\mathrm{KL}(\rho\otimes\rho'\|\pi\otimes\pi') = \mathrm{KL}(\rho\|\pi) + \mathrm{KL}(\rho'\|\pi')$. Substituting into (C.8) and dividing by λ,

$$\mathbb{E}_{\rho\otimes\rho'}[R_\psi(\theta,\theta')] - \lambda g\left(\frac{2\lambda}{n}\right) \mathbb{E}_{\rho\otimes\rho'}[V_\psi(\theta,\theta')]$$

$$\leq \mathbb{E}_{\rho\otimes\rho'}\Big[\widehat{R}_\psi(\theta,\theta')\Big] + \frac{1}{\lambda}\,\mathrm{KL}(\rho\|\pi) + \frac{1}{\lambda}\,\mathrm{KL}(\rho'\|\pi') + \frac{\log(1/\delta)}{\lambda}. \tag{C.9}$$

By Lemma 4.4,

$$V_\psi(\theta,\theta') \leq \frac{a_2^2}{n}\Big(\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) + \mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)})\Big),$$

hence

$$\lambda g\left(\frac{2\lambda}{n}\right) V_\psi(\theta,\theta') \leq \beta_{n,\lambda} a_2^2\Big(\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) + \mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)})\Big),$$

with $\beta_{n,\lambda} = g(2\lambda/n)\lambda/n$. Substituting this into (C.9) yields that, with probability at least $1-\delta$,

$$\begin{aligned}\mathbb{E}_{\rho\otimes\rho'}[R_\psi] \leq \mathbb{E}_{\rho\otimes\rho'}[\widehat{R}_\psi] + \beta_{n,\lambda} a_2^2 \Big(\mathbb{E}_\rho[\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)})] + \mathbb{E}_{\rho'}[\mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)})]\Big) \\ + \frac{1}{\lambda}\,\mathrm{KL}(\rho\|\pi) + \frac{1}{\lambda}\,\mathrm{KL}(\rho'\|\pi') + \frac{\log(1/\delta)}{\lambda}.\end{aligned} \tag{C.10}$$

By Lemma 4.4,

$$R_\psi(\theta,\theta') \geq a_1 \mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) - a_0 \mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)}).$$

Taking $\rho\otimes\rho'$-expectations and combining with (C.10) gives

$$\begin{aligned}(a_1 - \beta_{n,\lambda} a_2^2)\,\mathbb{E}_\rho\Big[\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)})\Big] \leq \mathbb{E}_{\rho\otimes\rho'}[\widehat{R}_\psi(\theta,\theta')] \\ + \big(a_0 + \beta_{n,\lambda} a_2^2\big)\,\mathbb{E}_{\rho'}\Big[\mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)})\Big] \\ + \frac{1}{\lambda}\,\mathrm{KL}(\rho\|\pi) + \frac{1}{\lambda}\,\mathrm{KL}(\rho'\|\pi') + \frac{\log(1/\delta)}{\lambda}.\end{aligned} \tag{C.11}$$

Fix $\theta \in \Theta$ and apply the Donsker–Varadhan variational formula (1.24) with $h(\theta') = \lambda\widehat{R}_\psi(\theta,\theta')$ and prior π'. For any $\rho' \ll \pi'$,

$$\mathbb{E}_{\theta'\sim\rho'}[\widehat{R}_\psi(\theta,\theta')] - \frac{1}{\lambda}\,\mathrm{KL}(\rho'\|\pi') \leq \frac{1}{\lambda}\log\left(\int_\Theta e^{\lambda\widehat{R}_\psi(\theta,\theta')}\pi'(d\theta')\right) =: \Lambda_\lambda(\theta;\pi'). \tag{C.12}$$

Rearranging (C.12),

$$\mathbb{E}_{\theta'\sim\rho'}[\widehat{R}_\psi(\theta,\theta')] \leq \Lambda_\lambda(\theta;\pi') + \frac{1}{\lambda}\,\mathrm{KL}(\rho'\|\pi'). \tag{C.13}$$

Averaging over $\theta\sim\rho$ yields

$$\mathbb{E}_{\rho\otimes\rho'}[\widehat{R}_\psi(\theta,\theta')] \leq \mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda(\theta;\pi')] + \frac{1}{\lambda}\,\mathrm{KL}(\rho'\|\pi').$$

Plugging into (C.11) gives

$$(a_1 - \beta_{n,\lambda} a_2^2)\, \mathbb{E}_\rho\left[\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)})\right] \le \mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda(\theta;\pi')] + \left(a_0 + \beta_{n,\lambda} a_2^2\right) \mathbb{E}_{\rho'}\left[\mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)})\right] + \frac{1}{\lambda}\,\mathrm{KL}(\rho\|\pi) + \frac{2}{\lambda}\,\mathrm{KL}(\rho'\|\pi') + \frac{\log(1/\delta)}{\lambda}. \quad \text{(C.14)}$$

Taking the infimum over $\rho' \in \mathcal{P}(\Theta)$ and evaluating at $\rho = \hat{\rho}_\lambda$, the minimizer of $\rho \mapsto \mathbb{E}_\rho[\Lambda_\lambda(\theta;\pi')] + \lambda^{-1}\,\mathrm{KL}(\rho\|\pi)$, yields (4.12). □

C.3 Proof of the corollary at an explicit inverse temperature

Proof of Corollary 4.6. Set $\lambda = n/8$. Then

$$\beta_{n,\lambda} = g\left(\frac{2\lambda}{n}\right)\frac{\lambda}{n} = \frac{1}{8}\,g\left(\frac{1}{4}\right),$$

where $g(x) = (e^x - 1 - x)/x^2$. Evaluating,

$$g\left(\frac{1}{4}\right) = 16\left(e^{1/4} - \frac{5}{4}\right) \approx 0.5444,$$

hence $\beta_{n,\lambda} \approx 0.068$. For the leading coefficient,

$$a_1 - \beta_{n,\lambda} a_2^2 = \frac{3}{8} - \beta_{n,\lambda}\cdot 3\sqrt{2} \approx 0.375 - 0.068 \times 4.242 \approx 0.086 \ge \frac{1}{12}.$$

For the competitor coefficient,

$$a_0 + \beta_{n,\lambda} a_2^2 = 4 + \beta_{n,\lambda}\cdot 3\sqrt{2} \approx 4 + 0.289 \approx 4.289 \le \frac{13}{3}.$$

Substituting these bounds into Theorem 4.5 yields (4.14). □

Proof of Corollary 4.7. In the i.i.d. case $p_\theta^i = p_\theta$ and $P_*^i = P_*$ for every i, so each coordinate Hellinger distance equals $\mathcal{H}^2(P_*, P_\theta)$ and the sample Hellinger distance (4.1) reduces to

$$\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{H}^2(P_*, P_\theta) = \mathcal{H}^2(P_*, P_\theta).$$

Substituting this identity throughout Theorem 4.5 gives (4.15). □

Proof of Corollary 4.8. Combine Corollary 4.7 with the explicit constants of Corollary 4.6: at

$\lambda = n/8$ one has $a_1 - \beta_{n,\lambda} a_2^2 \geq 1/12$ and $a_0 + \beta_{n,\lambda} a_2^2 \leq 13/3$, and $1/\lambda = 8/n, 2/\lambda = 16/n$. Substituting these values into (4.15) yields (4.16). □

C.4 Proof of the oracle inequality

The proof combines Theorem 4.5 with an upper bound on Λ_λ that holds with high probability.

Lemma C.2 (Upper bound on Λ_λ). *Fix $\delta \in (0,1)$ and $\lambda > 0$ such that $\beta_{n,\lambda} < a_1/a_2^2$. Then for any fixed $\rho \in \mathcal{P}(\Theta)$, with probability at least $1-\delta$,*

$$\mathbb{E}_{\theta\sim\rho}\big[\Lambda_\lambda(\theta;\pi')\big] \leq (a_0 + \beta_{n,\lambda} a_2^2)\, \mathbb{E}_{\theta\sim\rho}\Big[\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)})\Big] + \frac{\log(1/\delta)}{\lambda}. \tag{C.15}$$

Proof. Fix $\theta \in \Theta$. Apply the Bernstein moment generating function bound exactly as in the proof of Theorem 4.5 with $t = \lambda/n$, to obtain for each $\theta' \in \Theta$,

$$\mathbb{E}_{\mathcal{S}} \exp\left(\lambda(\widehat{R}_\psi(\theta,\theta') - R_\psi(\theta,\theta')) - \lambda^2 g\left(\frac{2\lambda}{n}\right) V_\psi(\theta,\theta')\right) \leq 1.$$

Integrating over $\theta' \sim \pi'$ and applying Markov's inequality, with probability at least $1-\delta$,

$$\int_\Theta \exp\left(\lambda(\widehat{R}_\psi(\theta,\theta') - R_\psi(\theta,\theta')) - \lambda^2 g\left(\frac{2\lambda}{n}\right) V_\psi(\theta,\theta')\right) \pi'(d\theta') \leq \frac{1}{\delta}. \tag{C.16}$$

Taking the logarithm and applying the Donsker–Varadhan variational formula (1.24) yields that, for any $\rho' \ll \pi'$,

$$\mathbb{E}_{\theta'\sim\rho'}[\widehat{R}_\psi(\theta,\theta')] \leq \mathbb{E}_{\theta'\sim\rho'}\left[R_\psi(\theta,\theta') + \lambda g\left(\frac{2\lambda}{n}\right) V_\psi(\theta,\theta')\right] + \frac{\mathrm{KL}(\rho'\|\pi') + \log(1/\delta)}{\lambda}.$$

Choose $\rho' = \rho'_\lambda(\theta)$, the Gibbs distribution proportional to $\exp(\lambda \widehat{R}_\psi(\theta,\theta'))\pi'(d\theta')$. Then

$$\Lambda_\lambda(\theta;\pi') = \mathbb{E}_{\theta'\sim\rho'_\lambda}[\widehat{R}_\psi(\theta,\theta')] - \frac{1}{\lambda}\,\mathrm{KL}(\rho'_\lambda\|\pi'),$$

and substituting gives

$$\Lambda_\lambda(\theta;\pi') \leq \mathbb{E}_{\theta'\sim\rho'_\lambda}\left[R_\psi(\theta,\theta') + \lambda g\left(\frac{2\lambda}{n}\right) V_\psi(\theta,\theta')\right] + \frac{\log(1/\delta)}{\lambda}.$$

Using the upper risk bound and the variance bound from Lemma 4.4,

$$R_\psi(\theta,\theta') \leq a_0 \mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) - a_1 \mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)}),$$

and

$$\lambda g\left(\frac{2\lambda}{n}\right) V_\psi(\theta,\theta') \le \beta_{n,\lambda} a_2^2 \Big(\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) + \mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)}) \Big),$$

we obtain

$$\Lambda_\lambda(\theta;\pi') \le (a_0 + \beta_{n,\lambda} a_2^2)\, \mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) + (\beta_{n,\lambda} a_2^2 - a_1)\, \mathbb{E}_{\theta' \sim \rho'_\lambda}\big[\mathcal{H}_n^2(P_*^{(n)}, P_{\theta'}^{(n)})\big] + \frac{\log(1/\delta)}{\lambda}.$$

Since $\beta_{n,\lambda} < a_1/a_2^2$, the coefficient $\beta_{n,\lambda} a_2^2 - a_1$ is negative, and dropping the corresponding nonpositive term yields

$$\Lambda_\lambda(\theta;\pi') \le (a_0 + \beta_{n,\lambda} a_2^2)\, \mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) + \frac{\log(1/\delta)}{\lambda}.$$

Taking the expectation over $\theta \sim \rho$ gives (C.15). □

Proof of Theorem 4.10. We work in the i.i.d. setting, so that $\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) = \mathcal{H}^2(P^*, P_\theta)$ throughout. By Theorem 4.5 with $\lambda = n/8$, with probability at least $1-\delta$,

$$\begin{aligned}
&\big(a_1 - \beta_{n,\lambda} a_2^2\big)\, \mathbb{E}_{\theta\sim\hat\rho_\lambda}[\mathcal{H}^2(P^*, P_\theta)] \\
&\le \inf_{\rho\in\mathcal{P}(\Theta)} \left\{ \mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda(\theta;\pi')] + \frac{1}{\lambda}\,\mathrm{KL}(\rho\|\pi) \right\} \\
&\quad + \inf_{\rho'\in\mathcal{P}(\Theta)} \left\{ (a_0 + \beta_{n,\lambda} a_2^2)\, \mathbb{E}_{\theta'\sim\rho'} \left[\mathcal{H}^2(P^*, P_{\theta'})\right] + \frac{2}{\lambda}\,\mathrm{KL}(\rho'\|\pi') \right\} + \frac{\log(1/\delta)}{\lambda}.
\end{aligned}$$

By Lemma C.2, with probability at least $1-\delta$, for any $\rho \in \mathcal{P}(\Theta)$,

$$\mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda(\theta;\pi')] \le (a_0 + \beta_{n,\lambda} a_2^2)\, \mathbb{E}_{\theta\sim\rho}[\mathcal{H}^2(P^*, P_\theta)] + \frac{\log(1/\delta)}{\lambda}.$$

A union bound over the two events gives probability at least $1-2\delta$. With $\lambda = n/8$, the explicit constants of Corollary 4.6 give $a_1 - \beta_{n,\lambda} a_2^2 \ge 1/12$ and $a_0 + \beta_{n,\lambda} a_2^2 \le 13/3$; substituting the upper bound on Λ_λ into the first infimum and collecting the resulting Hellinger terms (the competitor and the localization each contribute a factor $13/3$, summing to $26/3$) yields (4.17). □

C.5 Proofs of the prior-mass condition

Proof of Lemma 4.13. Let $\theta_0 \in \operatorname{argmin}_{\theta\in\Theta} \mathcal{H}^2(P^*, P_\theta)$ and, for $r \le r_0$, write $B(r) := \{\theta \in \Theta : \mathcal{H}^2(P^*, P_\theta) \le \mathcal{H}^2(P^*, P_{\theta_0}) + r\}$. Restrict the prior to $B(r)$, that is, take $\rho = \pi(\cdot \mid B(r))$. Then $\mathrm{KL}(\rho\|\pi) = -\log \pi(B(r)) \le d_\pi \log(c/r)$ by the prior mass condition (4.18), while

$\mathbb{E}_{\theta\sim\rho}[\mathcal{H}^2(P^*, P_\theta)] \le \mathcal{H}^2(P^*, P_{\theta_0}) + r$ since ρ is supported on $B(r)$. Hence

$$\inf_{\rho\in\mathcal{P}(\Theta)} \left\{ a\,\mathbb{E}_{\theta\sim\rho}[\mathcal{H}^2(P^*, P_\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\beta} \right\} \le a\,\mathcal{H}^2(P^*, P_{\theta_0}) + a\,r + \frac{d_\pi}{\beta}\log\!\left(\frac{c}{r}\right).$$

Minimizing the right-hand side over $r \in (0, r_0]$, the unconstrained minimizer of $r \mapsto ar + (d_\pi/\beta)\log(c/r)$ is $r^\star = d_\pi/(a\beta)$, which satisfies $r^\star \le r_0$ under the hypothesis $a\beta \ge d_\pi/r_0$. Substituting $r = r^\star$ gives

$$a\,r^\star + \frac{d_\pi}{\beta}\log\!\left(\frac{c}{r^\star}\right) = \frac{d_\pi}{\beta} + \frac{d_\pi}{\beta}\log\!\left(\frac{c\,a\beta}{d_\pi}\right) = \frac{d_\pi}{\beta}\log\!\left(\frac{ec\,a\beta}{d_\pi}\right),$$

which is (4.19). □

Proof of Corollary 4.14. With $\lambda = n/8$, Theorem 4.10 gives, with probability at least $1 - 2\delta$,

$$\frac{1}{12}\,\mathbb{E}_{\theta\sim\hat\rho_\lambda}[\mathcal{H}^2(P^*, P_\theta)] \le \inf_{\rho\in\mathcal{P}(\Theta)} \left\{ \frac{26}{3}\,\mathbb{E}_{\theta\sim\rho}[\mathcal{H}^2(P^*, P_\theta)] + \frac{32}{n}\,\mathrm{KL}(\rho\|\pi) \right\} + \frac{16\log(1/\delta)}{n}.$$

Apply Lemma 4.13 to the infimum with $a = 26/3$ and $\beta = n/32$, so that $a\beta = 13n/48$; the condition $a\beta \ge d_\pi/r_0$ is exactly $13n \ge 48\,d_\pi/r_0$. This yields

$$\inf_{\rho\in\mathcal{P}(\Theta)} \left\{ \frac{26}{3}\,\mathbb{E}_{\theta\sim\rho}[\mathcal{H}^2(P^*, P_\theta)] + \frac{32}{n}\,\mathrm{KL}(\rho\|\pi) \right\} \le \frac{26}{3}\,\mathcal{H}^2(P^*, P_{\theta_0}) + \frac{32\,d_\pi}{n}\log\!\left(\frac{13ecn}{48 d_\pi}\right).$$

Multiplying through by 12 gives

$$\mathbb{E}_{\theta\sim\hat\rho_\lambda}[\mathcal{H}^2(P^*, P_\theta)] \le 104\,\mathcal{H}^2(P^*, P_{\theta_0}) + \frac{384\,d_\pi}{n}\log\!\left(\frac{13ecn}{48 d_\pi}\right) + \frac{192\log(1/\delta)}{n},$$

which is (4.20). □

C.6 Proofs for the regression application

Proof of Lemma 4.16. We adapt Baraud & Birgé (2020, Proposition 3). By (4.1),

$$\mathcal{H}_n^2(P_*^{(n)}, P_f^{(n)}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{H}^2(p_*^i, p_f^i). \tag{C.17}$$

Since w_i is deterministic, each coordinate Hellinger distance reduces to

$$\mathcal{H}^2(p_*^i, p_f^i) = \mathcal{H}^2(p_{f^\star(w_i)}, q_{f(w_i)}), \tag{C.18}$$

where $p_\delta(\cdot) = p(\cdot - \delta)$ and $q_\delta(\cdot) = q(\cdot - \delta)$. By translation invariance, $\mathcal{H}^2(p_{f^\star(w_i)}, q_{f^\star(w_i)}) = \mathcal{H}^2(p, q)$.

Upper bound. The triangle inequality for Hellinger distance gives

$$\mathcal{H}^2(u, w) \le 2\mathcal{H}^2(u, v) + 2\mathcal{H}^2(v, w). \tag{C.19}$$

Applying (C.19) with the triple $(u, v, w) = (p_{f^\star(w_i)}, q_{f^\star(w_i)}, q_{f(w_i)})$,

$$\begin{aligned}\mathcal{H}^2(p_{f^\star(w_i)}, q_{f(w_i)}) &\le 2\mathcal{H}^2(p_{f^\star(w_i)}, q_{f^\star(w_i)}) + 2\mathcal{H}^2(q_{f^\star(w_i)}, q_{f(w_i)}) \\ &= 2\mathcal{H}^2(p, q) + 2\mathcal{H}^2(q_{f^\star(w_i)-f(w_i)}, q),\end{aligned} \tag{C.20}$$

by translation invariance. By Assumption 4.15 and the boundedness condition (4.28), $|f(w_i) - f^\star(w_i)|^{1+\alpha} \le (2B)^{1+\alpha} \le C_q^{-1}$, so the truncation is inactive and

$$\mathcal{H}^2(q_{f^\star(w_i)-f(w_i)}, q) \le C_q |f(w_i) - f^\star(w_i)|^{1+\alpha}. \tag{C.21}$$

Substituting (C.21) into (C.20) and averaging over $i \in [n]$,

$$\begin{aligned}\mathcal{H}_n^2(P_*^{(n)}, P_f^{(n)}) &= \frac{1}{n}\sum_{i=1}^n \mathcal{H}^2(p_{f^\star(w_i)}, q_{f(w_i)}) \\ &\le \frac{1}{n}\sum_{i=1}^n \left[2\mathcal{H}^2(p, q) + 2C_q |f(w_i) - f^\star(w_i)|^{1+\alpha}\right] \\ &= 2\mathcal{H}^2(p, q) + 2C_q \|f - f^\star\|_{n,1+\alpha}^{1+\alpha},\end{aligned} \tag{C.22}$$

by (4.26).

Lower bound. The reverse triangle inequality gives

$$\mathcal{H}^2(u, w) \ge \frac{1}{2}\mathcal{H}^2(v, w) - \mathcal{H}^2(u, v), \tag{C.23}$$

since $(a - b)^2 \ge \frac{1}{2}a^2 - b^2$. Applying with $(u, v, w) = (p_{f^\star(w_i)}, q_{f^\star(w_i)}, q_{f(w_i)})$,

$$\begin{aligned}\mathcal{H}^2(p_{f^\star(w_i)}, q_{f(w_i)}) &\ge \frac{1}{2}\mathcal{H}^2(q_{f^\star(w_i)}, q_{f(w_i)}) - \mathcal{H}^2(p_{f^\star(w_i)}, q_{f^\star(w_i)}) \\ &= \frac{1}{2}\mathcal{H}^2(q_{f^\star(w_i)-f(w_i)}, q) - \mathcal{H}^2(p, q).\end{aligned} \tag{C.24}$$

By the lower bound in Assumption 4.15 and condition (4.28),

$$\mathcal{H}^2(q_{f^\star(w_i)-f(w_i)}, q) \ge c_q |f(w_i) - f^\star(w_i)|^{1+\alpha}. \tag{C.25}$$

Substituting (C.25) into (C.24) and averaging over $i \in [n]$,

$$\mathcal{H}_n^2(P_*^{(n)}, P_f^{(n)}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{H}^2(p_{f^\star(w_i)}, q_{f(w_i)}) \geq \frac{c_q}{2}\|f - f^\star\|_{n,1+\alpha}^{1+\alpha} - \mathcal{H}^2(p,q). \tag{C.26}$$

Combining (C.22) and (C.26) gives (4.29) with $C_\alpha = 2C_q$ and $c_\alpha = c_q/2$. □

Lemma C.3 (Upper bound on Λ_λ for regression)**.** *Fix $\delta \in (0,1)$ and set $\lambda = n/8$. With probability at least $1-\delta$ over the data $\mathcal{S}$, for all $\rho \in \mathcal{P}(\mathcal{F})$ and any prior $\pi' \in \mathcal{P}(\mathcal{F})$,*

$$\mathbb{E}_{f\sim\rho}[\Lambda_\lambda(f;\pi')] \leq \frac{13}{3}\,\mathbb{E}_{f\sim\rho}\big[\mathcal{H}_n^2(P_*^{(n)}, P_f^{(n)})\big] + \frac{8\log(1/\delta)}{n}. \tag{C.27}$$

Proof. Identical to Lemma C.2 with Θ replaced by $\mathcal{F}$ and $\lambda = n/8$, using $a_0 + \beta_{n,\lambda} a_2^2 \leq 13/3$ and $1/\lambda = 8/n$. □

Proof of Theorem 4.17. Set $\lambda = n/8$. By Corollary 4.6, with probability at least $1-\delta$,

$$\begin{aligned}\frac{1}{12}\,\mathbb{E}_{\hat\rho_\lambda}\big[\mathcal{H}_n^2(P_*^{(n)}, P_f^{(n)})\big] \leq{}& \inf_{\rho\in\mathcal{P}(\mathcal{F})}\left\{\mathbb{E}_\rho[\Lambda_\lambda(f;\pi')] + \frac{8\ \mathrm{KL}(\rho\|\pi)}{n}\right\} \\ &+ \inf_{\rho'\in\mathcal{P}(\mathcal{F})}\left\{\frac{13}{3}\,\mathbb{E}_{\rho'}\big[\mathcal{H}_n^2(P_*^{(n)}, P_{f'}^{(n)})\big] + \frac{16\ \mathrm{KL}(\rho'\|\pi')}{n}\right\} \\ &+ \frac{8\log(1/\delta)}{n}.\end{aligned} \tag{C.28}$$

By Lemma C.3, with probability at least $1-\delta$,

$$\mathbb{E}_{f\sim\rho}[\Lambda_\lambda(f;\pi')] \leq \frac{13}{3}\,\mathbb{E}_{f\sim\rho}\big[\mathcal{H}_n^2(P_*^{(n)}, P_f^{(n)})\big] + \frac{8\log(1/\delta)}{n}. \tag{C.29}$$

A union bound gives probability at least $1-2\delta$. By Lemma 4.16,

$$c_\alpha\|f - f^\star\|_{n,1+\alpha}^{1+\alpha} - \mathcal{H}^2(p,q) \leq \mathcal{H}_n^2(P_*^{(n)}, P_f^{(n)}) \leq C_\alpha\|f - f^\star\|_{n,1+\alpha}^{1+\alpha} + 2\mathcal{H}^2(p,q). \tag{C.30}$$

The lower bound applied to the left side of (C.28) gives

$$\mathbb{E}_{\hat\rho_\lambda}\big[\mathcal{H}_n^2(P_*^{(n)}, P_f^{(n)})\big] \geq c_\alpha\,\mathbb{E}_{\hat\rho_\lambda}\big[\|f - f^\star\|_{n,1+\alpha}^{1+\alpha}\big] - \mathcal{H}^2(p,q).$$

Applying the upper bound to (C.29),

$$\mathbb{E}_{f\sim\rho}[\Lambda_\lambda(f;\pi')] \leq \frac{13C_\alpha}{3}\,\mathbb{E}_{f\sim\rho}\big[\|f - f^\star\|_{n,1+\alpha}^{1+\alpha}\big] + \frac{26}{3}\mathcal{H}^2(p,q) + \frac{8\log(1/\delta)}{n}.$$

Similarly, for the second infimum in (C.28),

$$\inf_{\rho' \in \mathcal{P}(\mathcal{F})} \left\{ \frac{13}{3} \mathbb{E}_{\rho'} \left[\mathcal{H}_n^2(P_*^{(n)}, P_{f'}^{(n)}) \right] + \frac{16 \ \mathrm{KL}(\rho' \| \pi')}{n} \right\}$$
$$\leq \inf_{\rho' \in \mathcal{P}(\mathcal{F})} \left\{ \frac{13 C_\alpha}{3} \mathbb{E}_{\rho'} \left[\| f' - f^\star \|_{n,1+\alpha}^{1+\alpha} \right] + \frac{16 \ \mathrm{KL}(\rho' \| \pi')}{n} \right\} + \frac{26}{3} \mathcal{H}^2(p, q).$$

Substituting into (C.28),

$$\begin{aligned} &\frac{1}{12} \left(c_\alpha \, \mathbb{E}_{\hat{\rho}_\lambda} \left[\| f - f^\star \|_{n,1+\alpha}^{1+\alpha} \right] - \mathcal{H}^2(p, q) \right) \\ &\leq \inf_{\rho \in \mathcal{P}(\mathcal{F})} \left\{ \frac{13 C_\alpha}{3} \mathbb{E}_{\rho} \left[\| f - f^\star \|_{n,1+\alpha}^{1+\alpha} \right] + \frac{8 \ \mathrm{KL}(\rho \| \pi)}{n} \right\} \\ &\quad + \inf_{\rho' \in \mathcal{P}(\mathcal{F})} \left\{ \frac{13 C_\alpha}{3} \mathbb{E}_{\rho'} \left[\| f' - f^\star \|_{n,1+\alpha}^{1+\alpha} \right] + \frac{16 \ \mathrm{KL}(\rho' \| \pi')}{n} \right\} \\ &\quad + \left(\frac{26}{3} + \frac{26}{3} \right) \mathcal{H}^2(p, q) + \frac{16 \log(1/\delta)}{n}. \end{aligned}$$

The coefficient of $\mathcal{H}^2(p, q)$ is $1/12 + 52/3 = 209/12$. Multiplying through by $12/c_\alpha$ yields (4.30). □

Proof of Corollary 4.19. Set $\pi = \pi'$ equal to the uniform distribution on the ε-net $\mathcal{F}_\varepsilon$ from Assumption 4.18. For arbitrary $f_0 \in \mathcal{F}$, there exists $f_\varepsilon \in \mathcal{F}_\varepsilon$ with $\| f_\varepsilon - f_0 \|_\infty \leq \varepsilon$. By the triangle inequality,

$$\| f_\varepsilon - f^\star \|_\infty \leq \varepsilon + \| f_0 - f^\star \|_\infty. \tag{C.31}$$

Since the empirical norm is dominated by the supremum norm,

$$\| f_\varepsilon - f^\star \|_{n,1+\alpha}^{1+\alpha} = \frac{1}{n} \sum_{i=1}^{n} | f_\varepsilon(w_i) - f^\star(w_i) |^{1+\alpha} \leq \| f_\varepsilon - f^\star \|_\infty^{1+\alpha}. \tag{C.32}$$

We bound $\| f_\varepsilon - f^\star \|_\infty^{1+\alpha}$ via (C.31), distinguishing two cases.

Case 1: $\alpha \geq 0$. For $1 + \alpha \geq 1$, the function $x \mapsto x^{1+\alpha}$ is convex. By the convexity inequality $(a + b)^p \leq 2^{p-1}(a^p + b^p)$ for $p \geq 1$,

$$\| f_\varepsilon - f^\star \|_{n,1+\alpha}^{1+\alpha} \leq \left(\varepsilon + \| f_0 - f^\star \|_\infty \right)^{1+\alpha} \leq 2^\alpha \left(\varepsilon^{1+\alpha} + \| f_0 - f^\star \|_\infty^{1+\alpha} \right). \tag{C.33}$$

Case 2: $-1 < \alpha < 0$. For $0 < 1 + \alpha < 1$, the function $x \mapsto x^{1+\alpha}$ is concave. By the subadditivity property $(a + b)^p \leq a^p + b^p$ for $0 < p < 1$,

$$\| f_\varepsilon - f^\star \|_{n,1+\alpha}^{1+\alpha} \leq \left(\varepsilon + \| f_0 - f^\star \|_\infty \right)^{1+\alpha} \leq \varepsilon^{1+\alpha} + \| f_0 - f^\star \|_\infty^{1+\alpha}. \tag{C.34}$$

In both cases,

$$\|f_\varepsilon - f^\star\|_{n,1+\alpha}^{1+\alpha} \le \tilde{C}_\alpha\big(\varepsilon^{1+\alpha} + \|f_0 - f^\star\|_\infty^{1+\alpha}\big), \tag{C.35}$$

where $\tilde{C}_\alpha = 2^\alpha$ if $\alpha \ge 0$ and $\tilde{C}_\alpha = 1$ if $-1 < \alpha < 0$.

Choose $\rho = \rho' = \delta_{f_\varepsilon}$. Since π is uniform on $\mathcal{F}_\varepsilon$ with $|\mathcal{F}_\varepsilon| \le e^{H(\varepsilon)}$,

$$\mathrm{KL}(\delta_{f_\varepsilon}\|\pi) = \log|\mathcal{F}_\varepsilon| \le H(\varepsilon). \tag{C.36}$$

By Theorem 4.17, with probability at least $1 - 2\delta$,

$$\begin{aligned}
&\mathbb{E}_{f\sim\hat{\rho}_\lambda}\big[\|f - f^\star\|_{n,1+\alpha}^{1+\alpha}\big] \\
&\le \frac{12}{c_\alpha}\Bigg[\inf_{\rho\in\mathcal{P}(\mathcal{F})}\left\{\frac{13C_\alpha}{3}\,\mathbb{E}_{f\sim\rho}\big[\|f - f^\star\|_{n,1+\alpha}^{1+\alpha}\big] + \frac{8\ \mathrm{KL}(\rho\|\pi)}{n}\right\} \\
&\qquad + \inf_{\rho'\in\mathcal{P}(\mathcal{F})}\left\{\frac{13C_\alpha}{3}\,\mathbb{E}_{f'\sim\rho'}\big[\|f' - f^\star\|_{n,1+\alpha}^{1+\alpha}\big] + \frac{16\ \mathrm{KL}(\rho'\|\pi')}{n}\right\} \\
&\qquad + \frac{209}{12}\,\mathcal{H}^2(p,q) + \frac{16\log(1/\delta)}{n}\Bigg].
\end{aligned} \tag{C.37}$$

Substituting $\rho = \rho' = \delta_{f_\varepsilon}$ and using (C.35) and (C.36),

$$\begin{aligned}
&\mathbb{E}_{f\sim\hat{\rho}_\lambda}\big[\|f - f^\star\|_{n,1+\alpha}^{1+\alpha}\big] \\
&\le \frac{12}{c_\alpha}\left[\frac{26C_\alpha}{3}\cdot\tilde{C}_\alpha\big(\varepsilon^{1+\alpha} + \|f_0 - f^\star\|_\infty^{1+\alpha}\big) + \frac{24H(\varepsilon)}{n} + \frac{209}{12}\,\mathcal{H}^2(p,q) + \frac{16\log(1/\delta)}{n}\right].
\end{aligned} \tag{C.38}$$

Since this holds for arbitrary $f_0 \in \mathcal{F}$, taking the infimum over f_0 yields

$$\begin{aligned}
&\mathbb{E}_{f\sim\hat{\rho}_\lambda}\big[\|f - f^\star\|_{n,1+\alpha}^{1+\alpha}\big] \\
&\le \frac{12}{c_\alpha}\left[\frac{26C_\alpha\tilde{C}_\alpha}{3}\varepsilon^{1+\alpha} + \frac{26C_\alpha\tilde{C}_\alpha}{3}\inf_{f\in\mathcal{F}}\|f - f^\star\|_\infty^{1+\alpha}\right. \\
&\qquad \left. + \frac{24H(\varepsilon)}{n} + \frac{209}{12}\,\mathcal{H}^2(p,q) + \frac{16\log(1/\delta)}{n}\right].
\end{aligned} \tag{C.39}$$

Defining

$$K_\alpha := \frac{12}{c_\alpha}\max\left\{\frac{26C_\alpha\tilde{C}_\alpha}{3},\ 24,\ \frac{209}{12},\ 16\right\}, \qquad \tilde{C}_\alpha = \max(1, 2^\alpha),$$

we obtain (4.31). □

Proof of Corollary 4.20. We optimize (4.31) over ε. Under $H(\varepsilon) \le M\varepsilon^{-d}$, it suffices to mini-

mize

$$\varepsilon^{1+\alpha} + \frac{H(\varepsilon)}{n} \le \varepsilon^{1+\alpha} + \frac{M\varepsilon^{-d}}{n}.$$

Setting the derivative to zero,

$$(1+\alpha)\varepsilon^{\alpha} - \frac{Md\varepsilon^{-d-1}}{n} = 0 \quad \Longrightarrow \quad \varepsilon^{d+1+\alpha} = \frac{Md}{n(1+\alpha)}.$$

Up to constants, the optimal choice is $\varepsilon_n \asymp (M/n)^{1/(d+1+\alpha)}$, giving

$$\varepsilon_n^{1+\alpha} + \frac{H(\varepsilon_n)}{n} \asymp n^{-\frac{1+\alpha}{d+1+\alpha}}.$$

Substituting into (4.31) and applying Jensen's inequality (valid since $1+\alpha \ge 1$),

$$\begin{aligned}\mathbb{E}_{f\sim\hat{\rho}_\lambda}\left[\|f-f^\star\|_{n,1+\alpha}\right] &\le \left(\mathbb{E}_{f\sim\hat{\rho}_\lambda}\left[\|f-f^\star\|_{n,1+\alpha}^{1+\alpha}\right]\right)^{1/(1+\alpha)} \\ &\le K_\alpha^{1/(1+\alpha)}\Bigg[\mathcal{H}^2(p,q) + \inf_{f\in\mathcal{F}}\|f-f^\star\|_\infty^{1+\alpha} \\ &\qquad + n^{-\frac{1+\alpha}{d+1+\alpha}} + \frac{\log(1/\delta)}{n}\Bigg]^{1/(1+\alpha)}.\end{aligned}$$

Setting $K'_\alpha = K_\alpha^{1/(1+\alpha)}$ yields (4.33). □

C.7 Proof of the variational oracle inequality

The proof adapts the oracle inequality to variational families, exploiting the fact that the Donsker–Varadhan upper bound holds for any member of the variational family even though the supremum need not be attained.

Lemma C.4 (Variational Donsker–Varadhan bound). *For any $\rho' \in \mathcal{F}'$ and any $\theta \in \Theta$,*

$$\mathbb{E}_{\theta'\sim\rho'}[\widehat{R}_\psi(\theta,\theta')] \le \Lambda_\lambda^{\mathcal{F}'}(\theta;\pi') + \frac{1}{\lambda}\,\mathrm{KL}(\rho'\|\pi').$$

Proof. By Definition 4.22,

$$\Lambda_\lambda^{\mathcal{F}'}(\theta;\pi') = \sup_{\nu\in\mathcal{F}'}\left\{\mathbb{E}_{\theta'\sim\nu}[\widehat{R}_\psi(\theta,\theta')] - \frac{1}{\lambda}\,\mathrm{KL}(\nu\|\pi')\right\} \ge \mathbb{E}_{\theta'\sim\rho'}[\widehat{R}_\psi(\theta,\theta')] - \frac{1}{\lambda}\,\mathrm{KL}(\rho'\|\pi'),$$

where the inequality holds because $\rho' \in \mathcal{F}'$. □

Proposition C.5 (PAC-Bayes bound with variational softmax). *Fix $\delta \in (0,1)$ and $\lambda > 0$.*

With probability at least $1-\delta$*, for any* $\rho \in \mathcal{P}(\Theta)$ *and any* $\rho' \in \mathcal{F}'$*,*

$$\begin{aligned}(a_1 - \beta_{n,\lambda} a_2^2)\, \mathbb{E}_{\theta\sim\rho}[\mathcal{H}^2(P^*, P_\theta)] \le \mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda^{\mathcal{F}'}(\theta;\pi')] \\ + (a_0 + \beta_{n,\lambda} a_2^2)\, \mathbb{E}_{\theta'\sim\rho'}\left[\mathcal{H}^2(P^*, P_{\theta'})\right] \\ + \frac{1}{\lambda}\,\mathrm{KL}(\rho\|\pi) + \frac{2}{\lambda}\,\mathrm{KL}(\rho'\|\pi') + \frac{\log(1/\delta)}{\lambda}. \quad \text{(C.40)}\end{aligned}$$

Proof. Starting from the intermediate bound (C.11), specialized to the i.i.d. case so that $\mathcal{H}_n^2(P_*^{(n)}, P_\theta^{(n)}) = \mathcal{H}^2(P^*, P_\theta)$, which holds with probability at least $1-\delta$ for any $\rho \in \mathcal{P}(\Theta)$ and any $\rho' \in \mathcal{P}(\Theta)$, we have

$$\begin{aligned}(a_1 - \beta_{n,\lambda} a_2^2)\mathbb{E}_{\theta\sim\rho}[\mathcal{H}^2(P^*, P_\theta)] \le \mathbb{E}_{\theta\sim\rho}\mathbb{E}_{\theta'\sim\rho'}[\widehat{R}_\psi(\theta,\theta')] \\ + (a_0 + \beta_{n,\lambda} a_2^2)\mathbb{E}_{\theta'\sim\rho'}\left[\mathcal{H}^2(P^*, P_{\theta'})\right] \\ + \frac{1}{\lambda}\,\mathrm{KL}(\rho\|\pi) + \frac{1}{\lambda}\,\mathrm{KL}(\rho'\|\pi') + \frac{\log(1/\delta)}{\lambda}.\end{aligned}$$

For any fixed $\rho' \in \mathcal{F}'$, apply Lemma C.4 and take the expectation over $\theta \sim \rho$,

$$\mathbb{E}_{\theta\sim\rho}\mathbb{E}_{\theta'\sim\rho'}[\widehat{R}_\psi(\theta,\theta')] \le \mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda^{\mathcal{F}'}(\theta;\pi')] + \frac{1}{\lambda}\,\mathrm{KL}(\rho'\|\pi').$$

Substituting this bound and collecting the KL terms involving ρ' yields (C.40). □

Lemma C.6 (Upper bound on $\Lambda_\lambda^{\mathcal{F}'}$)**.** *Fix* $\varepsilon \in (0,1)$ *and* $\lambda > 0$ *such that* $\beta_{n,\lambda} < a_1/a_2^2$*. With probability at least* $1-\varepsilon$*, for any* $\rho \in \mathcal{P}(\Theta)$*,*

$$\mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda^{\mathcal{F}'}(\theta;\pi')] \le (a_0 + \beta_{n,\lambda} a_2^2)\, \mathbb{E}_{\theta\sim\rho}[\mathcal{H}^2(P^*, P_\theta)] + \frac{\log(1/\varepsilon)}{\lambda}.$$

Proof. By Definition 4.22, $\Lambda_\lambda^{\mathcal{F}'}(\theta;\pi') \le \Lambda_\lambda(\theta;\pi')$ for all $\theta \in \Theta$, since the supremum over $\mathcal{F}'$ is at most the supremum over $\mathcal{P}(\Theta)$. Therefore, for any $\rho \in \mathcal{P}(\Theta)$,

$$\mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda^{\mathcal{F}'}(\theta;\pi')] \le \mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda(\theta;\pi')].$$

Applying Lemma C.2 to the right-hand side with confidence parameter ε yields the claim. □

Proof of Theorem 4.23. We combine the two high-probability events via a union bound. By Proposition C.5, (C.40) holds with probability at least $1-\delta$. Taking the infimum over $\rho' \in \mathcal{F}'$ and then over $\rho \in \mathcal{F}$, the minimizer of the latter being $\tilde{\rho}_\lambda$ by definition (4.34),

$$\begin{aligned}&(a_1 - \beta_{n,\lambda} a_2^2)\mathbb{E}_{\theta\sim\tilde{\rho}_\lambda}[\mathcal{H}^2(P^*, P_\theta)] \\ &\le \inf_{\rho\in\mathcal{F}}\left\{\mathbb{E}_{\theta\sim\rho}[\Lambda_\lambda^{\mathcal{F}'}(\theta;\pi')] + \frac{1}{\lambda}\,\mathrm{KL}(\rho\|\pi)\right\}\end{aligned}$$

$$+ \inf_{\rho' \in \mathcal{F}'} \left\{ (a_0 + \beta_{n,\lambda} a_2^2) \mathbb{E}_{\theta' \sim \rho'} \left[\mathcal{H}^2(P^*, P_{\theta'}) \right] + \frac{2}{\lambda} \mathrm{KL}(\rho' \| \pi') \right\} + \frac{\log(1/\delta)}{\lambda}.$$

By Lemma C.6, with probability at least $1 - \varepsilon$, for any $\rho \in \mathcal{F}$,

$$\mathbb{E}_{\theta \sim \rho}[\Lambda_\lambda^{\mathcal{F}'}(\theta; \pi')] \le (a_0 + \beta_{n,\lambda} a_2^2) \mathbb{E}_{\theta \sim \rho}[\mathcal{H}^2(P^*, P_\theta)] + \frac{\log(1/\varepsilon)}{\lambda},$$

hence

$$\begin{aligned} &\inf_{\rho \in \mathcal{F}} \left\{ \mathbb{E}_{\theta \sim \rho}[\Lambda_\lambda^{\mathcal{F}'}(\theta; \pi')] + \frac{1}{\lambda} \mathrm{KL}(\rho \| \pi) \right\} \\ &\le \inf_{\rho \in \mathcal{F}} \left\{ (a_0 + \beta_{n,\lambda} a_2^2) \mathbb{E}_{\theta \sim \rho}[\mathcal{H}^2(P^*, P_\theta)] + \frac{1}{\lambda} \mathrm{KL}(\rho \| \pi) \right\} + \frac{\log(1/\varepsilon)}{\lambda}. \end{aligned}$$

By a union bound, both events hold simultaneously with probability at least $1 - \delta - \varepsilon$. Taking $\varepsilon = \delta$, $\lambda = n/8$ and using the explicit constants of Corollary 4.6 ($a_1 - \beta_{n,\lambda} a_2^2 \ge 1/12$, $a_0 + \beta_{n,\lambda} a_2^2 \le 13/3$, the competitor and localization terms summing to $26/3$) yields (4.35). □

Proof of Corollary 4.24. By Theorem 4.23, with probability at least $1 - 2\delta$,

$$\frac{1}{12} \mathbb{E}_{\theta \sim \tilde{\rho}_\lambda}[\mathcal{H}^2(P^*, P_\theta)] \le \inf_{\rho \in \mathcal{F}} \left\{ \frac{26}{3} \mathbb{E}_{\theta \sim \rho}[\mathcal{H}^2(P^*, P_\theta)] + \frac{32}{n} \mathrm{KL}(\rho \| \pi) \right\} + \frac{16 \log(1/\delta)}{n}.$$

Evaluate the infimum at $\rho = \rho_n \in \mathcal{F}$. By the hypotheses (4.36), $\mathbb{E}_{\theta \sim \rho_n}[\mathcal{H}^2(P^*, P_\theta)] \le 1/n$ and $\mathrm{KL}(\rho_n \| \pi) \le d_\pi \log n$, so

$$\inf_{\rho \in \mathcal{F}} \left\{ \frac{26}{3} \mathbb{E}_{\theta \sim \rho}[\mathcal{H}^2(P^*, P_\theta)] + \frac{32}{n} \mathrm{KL}(\rho \| \pi) \right\} \le \frac{26}{3n} + \frac{32\, d_\pi \log n}{n}.$$

Multiplying through by 12,

$$\mathbb{E}_{\theta \sim \tilde{\rho}_\lambda}[\mathcal{H}^2(P^*, P_\theta)] \le \frac{104}{n} + \frac{384\, d_\pi \log n}{n} + \frac{192 \log(1/\delta)}{n},$$

which is (4.37). □

C.8 Proof of the saddle-point equivalence and gap bound

Proof of Theorem 4.28. We establish each part in turn.

Part (i). Fix $\phi \in \Phi$. Since the KL term $\frac{1}{\lambda} \mathrm{KL}(\rho_\phi \| \pi)$ does not depend on ν, we may separate it from the supremum:

$$\sup_{\nu \in \mathcal{N}} \mathcal{L}_n(\phi, \nu) = \sup_{\nu \in \mathcal{N}} \left\{ \mathbb{E}_{(\theta, \theta') \sim \rho_\phi \otimes \rho'_\nu}[\widehat{R}_\psi(\theta, \theta')] - \frac{1}{\lambda} \mathrm{KL}(\rho'_\nu \| \pi') \right\} + \frac{1}{\lambda} \mathrm{KL}(\rho_\phi \| \pi). \quad \text{(C.41)}$$

By the surjectivity assumption, the parameterization $\nu \mapsto \rho'_\nu$ covers the entire variational family $\mathcal{F}'$. Therefore, we may replace $\sup_{\nu\in\mathcal{N}}$ by $\sup_{\rho'\in\mathcal{F}'}$, and the right-hand side becomes exactly $\tilde{\mathcal{J}}(\phi)$ as defined in (4.39), establishing (4.41).

It remains to verify that the supremum is attained. Define

$$f(\nu) := \mathbb{E}_{(\theta,\theta')\sim\rho_\phi\otimes\rho'_\nu}[\widehat{R}_\psi(\theta,\theta')] - \frac{1}{\lambda}\,\mathrm{KL}(\rho'_\nu\|\pi').$$

The map $\nu \mapsto f(\nu)$ is continuous on $\overline{\mathcal{N}}$: expectations of bounded functions of Gaussian parameters are continuous, and the Gaussian KL divergence is smooth. Since $\overline{\mathcal{N}}$ is compact, the extreme value theorem guarantees a maximizer $\nu^\star(\phi)$. By Proposition C.9, this maximizer is unique and lies in the interior of $\mathcal{N}$.

Part (ii). Define

$$h(\theta,\nu) := \mathbb{E}_{\theta'\sim\rho'_\nu}[\widehat{R}_\psi(\theta,\theta')] - \frac{1}{\lambda}\,\mathrm{KL}(\rho'_\nu\|\pi').$$

Since the supremum over ν and the expectation over θ do not commute in general, Jensen's inequality gives

$$\tilde{\mathcal{J}}(\phi) = \sup_\nu\, \mathbb{E}_{\theta\sim\rho_\phi}[h(\theta,\nu)] \le \mathbb{E}_{\theta\sim\rho_\phi}\big[\sup_\nu\, h(\theta,\nu)\big] = \mathcal{J}(\phi),$$

where the last equality uses the definition of $\mathcal{J}$ in (4.38). The upper bound $\mathcal{J}(\phi) \le \tilde{\mathcal{J}}(\phi) + \Delta(\phi)$ is established in the proof of Proposition 4.32 below, which also gives (4.43). Together these yield (4.42).

Part (iii). Suppose $\phi^\star$ is a stationary point of $\tilde{\mathcal{J}}$, and let $\nu^\star = \nu^\star(\phi^\star)$ be the unique interior maximizer from Part (i). By first-order optimality in ν, $\nabla_\nu\mathcal{L}_n(\phi^\star,\nu^\star) = 0$. Since the Hessian satisfies $\nabla^2_{\nu\nu}\mathcal{L}_n(\phi,\nu^\star) \preceq -\mu_0\mathbb{I}$ (Proposition C.9), it is nonsingular, and the implicit function theorem implies that $\nu^\star(\phi)$ is smooth in ϕ. By the envelope theorem,

$$\nabla_\phi\tilde{\mathcal{J}}(\phi) = \nabla_\phi\mathcal{L}_n\big(\phi,\nu^\star(\phi)\big) + \underbrace{\nabla_\nu\mathcal{L}_n\big(\phi,\nu^\star(\phi)\big)}_{=0}\cdot\nabla_\phi\nu^\star(\phi) = \nabla_\phi\mathcal{L}_n\big(\phi,\nu^\star(\phi)\big).$$

Therefore $\nabla_\phi\tilde{\mathcal{J}}(\phi^\star) = 0$ implies $\nabla_\phi\mathcal{L}_n(\phi^\star,\nu^\star) = 0$, so $(\phi^\star,\nu^\star)$ is a first-order stationary point of $\mathcal{L}_n$. Conversely, if $(\phi^\star,\nu^\star)$ is a first-order stationary point of $\mathcal{L}_n$, then $\nabla_\nu\mathcal{L}_n(\phi^\star,\nu^\star) = 0$, which by uniqueness of the maximizer implies $\nu^\star = \nu^\star(\phi^\star)$, and the envelope theorem yields $\nabla_\phi\tilde{\mathcal{J}}(\phi^\star) = \nabla_\phi\mathcal{L}_n(\phi^\star,\nu^\star) = 0$. For the infimum equivalence, Part (i) and the inequality $\tilde{\mathcal{J}} \le \mathcal{J}$ from Part (ii) give (4.44).

Part (iv). In the PAC-Bayes regime $\lambda = \Theta(n)$, posterior concentration (Theorem 4.23) gives

$\mathrm{tr}(\Sigma_{\phi^\star}) \le Cd/n$. Applying Proposition 4.32,

$$\Delta(\phi^\star) \le \frac{\bar{G}^2}{16}\,\mathrm{tr}(\Sigma_{\phi^\star}) \le \frac{C\bar{G}^2\,d}{16\,n} \to 0 \qquad \text{as } n \to \infty.$$

Hence $\tilde{\mathcal{J}}(\phi^\star) = \mathcal{J}(\phi^\star)$ asymptotically. □

Proof of Proposition 4.32. Write $V(\theta) := \Lambda_\lambda^{\mathcal{F}'}(\theta;\pi')$ for the pointwise optimal value function. By the definition of $\mathcal{J}$ in (4.38), $\mathcal{J}(\phi) = \mathbb{E}_{\theta\sim\rho_\phi}[V(\theta)] + \frac{1}{\lambda}\,\mathrm{KL}(\rho_\phi\|\pi)$. Since the KL term appears identically in $\mathcal{J}(\phi)$ and $\tilde{\mathcal{J}}(\phi)$, it cancels in the gap:

$$\Delta(\phi) = \mathbb{E}_{\theta\sim\rho_\phi}[V(\theta)] - \sup_{\rho'\in\mathcal{F}'}\Big\{\mathbb{E}_{(\theta,\theta')\sim\rho_\phi\otimes\rho'}[\widehat{R}_\psi(\theta,\theta')] - \frac{1}{\lambda}\,\mathrm{KL}(\rho'\|\pi')\Big\}.$$

We expand V around the mean $m := \mathbb{E}_{\rho_\phi}[\theta]$. By Taylor's theorem with integral remainder, for each θ there is an intermediate point ξ_θ on the segment $[m,\theta]$ with

$$V(\theta) = V(m) + \nabla_\theta V(m)^\top(\theta - m) + \frac{1}{2}(\theta-m)^\top\nabla_\theta^2 V(\xi_\theta)\,(\theta-m).$$

Taking expectations under ρ_ϕ and using $\mathbb{E}_{\rho_\phi}[\theta - m] = 0$, the linear term vanishes:

$$\mathbb{E}_{\theta\sim\rho_\phi}[V(\theta)] = V(m) + \frac{1}{2}\,\mathbb{E}_{\theta\sim\rho_\phi}\big[(\theta-m)^\top\nabla_\theta^2 V(\xi_\theta)\,(\theta-m)\big].$$

By the envelope theorem applied to the pointwise supremum defining $V(\theta)$, and the derivative bounds of Lemma C.7, the Hessian of V satisfies $\|\nabla_\theta^2 V(\theta)\|_{\mathrm{op}} \le \bar{G}^2/8$ for all $\theta \in \bar{\Theta}$. Substituting,

$$\mathbb{E}_{\theta\sim\rho_\phi}[V(\theta)] - V(m) \le \frac{1}{2}\cdot\frac{\bar{G}^2}{8}\,\mathbb{E}_{\theta\sim\rho_\phi}\big[\|\theta-m\|^2\big] = \frac{\bar{G}^2}{16}\,\mathrm{tr}(\Sigma_\phi).$$

Next, choosing $\rho' = \rho'^\star(m)$, the pointwise optimal competitor at $\theta = m$, in the definition of $\tilde{\mathcal{J}}$ gives the lower bound

$$\tilde{\mathcal{J}}(\phi) \ge \mathbb{E}_{\theta\sim\rho_\phi}\mathbb{E}_{\theta'\sim\rho'^\star(m)}[\widehat{R}_\psi(\theta,\theta')] - \frac{1}{\lambda}\,\mathrm{KL}(\rho'^\star(m)\|\pi') + \frac{1}{\lambda}\,\mathrm{KL}(\rho_\phi\|\pi)$$
$$\ge V(m) + \frac{1}{\lambda}\,\mathrm{KL}(\rho_\phi\|\pi).$$

Combining the gap expression, the lower bound on $\tilde{\mathcal{J}}$, and the Hessian bound,

$$\Delta(\phi) = \mathcal{J}(\phi) - \tilde{\mathcal{J}}(\phi) \le \mathbb{E}_{\theta\sim\rho_\phi}[V(\theta)] - V(m) \le \frac{\bar{G}^2}{16}\,\mathrm{tr}(\Sigma_\phi),$$

which is (4.45). □

C.9 Proof of the nonconvex–strongly-concave geometry

We write $\varphi(u) := \psi(e^u)$ and define the contrast

$$\ell_\psi(x;\theta,\theta') := \varphi\big(\log p_{\theta'}(x) - \log p_\theta(x)\big) \in [-1,1].$$

Lemma C.7 (Derivative bounds). *Set $t := e^{u/2}$. Then $\varphi'(u) = t/(t+1)^2$ and $\varphi''(u) = t(1-t)/[2(t+1)^3]$. In particular,*

$$\varphi'(u) = \frac{e^{u/2}}{(e^{u/2}+1)^2}, \tag{C.42}$$

$$\varphi''(u) = \frac{e^{u/2}(1-e^{u/2})}{2(e^{u/2}+1)^3}. \tag{C.43}$$

For all $u \in \mathbb{R}$, $0 < \varphi'(u) \le 1/4$ and $|\varphi''(u)| \le 1/8$. Moreover, on $|u| \le M$, $\varphi'(u) \ge c_M := \min_{t\in[e^{-M/2},\, e^{M/2}]} t/(t+1)^2 > 0$.

Proof. Recall $\psi(r) = (\sqrt{r}-1)/(\sqrt{r}+1)$. Differentiating with respect to r,

$$\psi'(r) = \frac{\frac{1}{2}r^{-1/2}(\sqrt{r}+1) - (\sqrt{r}-1)\cdot\frac{1}{2}r^{-1/2}}{(\sqrt{r}+1)^2} = \frac{r^{-1/2}}{(\sqrt{r}+1)^2}.$$

Since $\varphi(u) = \psi(e^u)$, the chain rule gives

$$\varphi'(u) = \psi'(e^u)\cdot e^u = \frac{e^{-u/2}}{(e^{u/2}+1)^2}\cdot e^u = \frac{e^{u/2}}{(e^{u/2}+1)^2}.$$

With $t := e^{u/2} > 0$, this is $\varphi'(u) = t/(t+1)^2$, establishing (C.42). Define $f(t) := t/(t+1)^2$ for $t > 0$. Differentiating,

$$f'(t) = \frac{(t+1)^2 - t\cdot 2(t+1)}{(t+1)^4} = \frac{1-t}{(t+1)^3}.$$

Since $f'(t) > 0$ for $t \in (0,1)$ and $f'(t) < 0$ for $t > 1$, the function f attains its unique global maximum at $t = 1$, where $f(1) = 1/4$. Moreover $f(t) > 0$ for all $t > 0$ and $f(t) \to 0$ as $t \to 0^+$ and as $t \to +\infty$. Therefore $0 < \varphi'(u) \le 1/4$ for all $u \in \mathbb{R}$.

For the second derivative, differentiate $\varphi'(u) = f(t)$ with $t = e^{u/2}$:

$$\varphi''(u) = f'(t)\cdot\frac{dt}{du} = \frac{1-t}{(t+1)^3}\cdot\frac{t}{2} = \frac{t(1-t)}{2(t+1)^3},$$

which establishes (C.43). Define $g(t) := t(1-t)/[2(t+1)^3]$ for $t > 0$. Then

$$g'(t) = \frac{(1-2t)\cdot 2(t+1)^3 - t(1-t)\cdot 6(t+1)^2}{4(t+1)^6} = \frac{1-4t+t^2}{2(t+1)^4}.$$

Setting $g'(t) = 0$ gives $t^2 - 4t + 1 = 0$, with positive roots $t_\pm = 2 \pm \sqrt{3}$. Evaluating,

$$g(t_-) = \frac{(2-\sqrt{3})(\sqrt{3}-1)}{2(3-\sqrt{3})^3} \approx 0.048,$$

and by symmetry $g(t_+) \approx -0.048$. Since $g(t) \to 0$ as $t \to 0^+$ and $t \to +\infty$, the global extrema of $|g|$ are attained at $t_\pm$, giving $|\varphi''(u)| = |g(t)| \le 0.048 < 1/8$ for all $u \in \mathbb{R}$.

Finally, on $|u| \le M$ the substitution $t = e^{u/2}$ ranges over the compact set $[e^{-M/2}, e^{M/2}]$. Since $f(t) = t/(t+1)^2$ is continuous and strictly positive on $(0, \infty)$, its minimum on this interval is attained and positive:

$$c_M := \min_{t \in [e^{-M/2}, e^{M/2}]} \frac{t}{(t+1)^2} > 0,$$

hence $\varphi'(u) \ge c_M$ for all $|u| \le M$. □

Proposition C.8 (Global L-smoothness)**.** *Under Assumption 4.31, $\mathcal{L}_n$ has Lipschitz gradient on the compact feasible set $\mathcal{B}$ with constant $L = L_\psi + L_{\mathrm{KL}}/\lambda$, where L_ψ and L_{KL} are independent of λ.*

Proof. Decompose $\mathcal{L}_n(\phi, \nu) = F(\phi, \nu) + \frac{1}{\lambda} K(\phi, \nu)$, where

$$F(\phi, \nu) := \mathbb{E}_{\theta \sim \rho_\phi} \mathbb{E}_{\theta' \sim \rho'_\nu}[\widehat{R}_\psi(\theta, \theta')], \qquad K(\phi, \nu) := \mathrm{KL}(\rho_\phi \| \pi) - \mathrm{KL}(\rho'_\nu \| \pi').$$

We bound the Hessian of each term separately. For a single observation x, write $u := \log p_{\theta'}(x) - \log p_\theta(x)$ so that $\ell_\psi(x; \theta, \theta') = \varphi(u)$. The exponential-family structure gives $\nabla_{\theta'} u = T(x) - \mu(\theta')$, using $\nabla_{\theta'} \log p_{\theta'}(x) = T(x) - \nabla A(\theta') = T(x) - \mu(\theta')$. By the chain rule,

$$\nabla_{\theta'} \ell_\psi(x; \theta, \theta') = \varphi'(u)\big(T(x) - \mu(\theta')\big),$$

and differentiating again,

$$\nabla^2_{\theta'\theta'} \ell_\psi(x; \theta, \theta') = \varphi''(u)\big(T(x) - \mu(\theta')\big)\big(T(x) - \mu(\theta')\big)^\top - \varphi'(u)\, I(\theta'),$$

where $I(\theta') = \nabla^2 A(\theta')$. Applying the triangle inequality and the bounds $|\varphi''(u)| \le 1/8$ and $0 < \varphi'(u) \le 1/4$ from Lemma C.7,

$$\|\nabla^2_{\theta'\theta'} \ell_\psi\|_{\mathrm{op}} \le \frac{1}{8}\|T(x) - \mu(\theta')\|^2 + \frac{1}{4}\|I(\theta')\|_{\mathrm{op}},$$

and analogously for the θ-block.

For the mean block of F, reparameterize $\theta' = m' + D(s)^{1/2}\varepsilon'$ with $\varepsilon' \sim \mathcal{N}(0, I_d)$ and $D(s) = \mathrm{diag}(e^{s_1}, \dots, e^{s_d})$. Since $\partial_{m'}\theta' = I_d$,

$$\nabla^2_{m'm'} F = \mathbb{E}_{\varepsilon'}\big[\nabla^2_{\theta'\theta'}\hat{g}(\theta')\big], \qquad \hat{g}(\theta') = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta\sim\rho_\phi}[\ell_\psi(X_i;\theta,\theta')].$$

Taking the operator norm and applying the Hessian bound with Assumption 4.31(2),

$$\|\nabla^2_{m'm'} F\|_{\mathrm{op}} \le \frac{B^2}{8} + \frac{\bar{I}}{4} =: L_{m'}, \qquad \bar{I} := \sup_{\theta\in\bar{\Theta}} \|I(\theta)\|_{\mathrm{op}} < \infty.$$

For the log-variance parameters, $\partial_{s_i}\theta' = \frac{1}{2}e^{s_i/2}\varepsilon'_i\, e_i$, so

$$\partial^2_{s_i s_i} F = \frac{e^{s_i}}{4}\,\mathbb{E}\big[(\varepsilon'_i)^2\, e_i^\top \nabla^2_{\theta'\theta'}\hat{g}\, e_i\big] + \frac{e^{s_i/2}}{4}\,\mathbb{E}\big[\varepsilon'_i\, e_i^\top \nabla_{\theta'}\hat{g}\big].$$

Using $\mathbb{E}[(\varepsilon'_i)^2] = 1$, $\mathbb{E}[|\varepsilon'_i|] = \sqrt{2/\pi}$, the Hessian bound, and $\|\nabla_{\theta'}\hat{g}\| \le \bar{G}/4$ (since $\varphi' \le 1/4$), with $\bar{s} := \sup_{s\in\mathcal{B}} \max_i |s_i|$,

$$|\partial^2_{s_i s_i} F| \le \frac{e^{\bar{s}}}{4}\left(\frac{B^2}{8} + \frac{\bar{I}}{4}\right) + \frac{e^{\bar{s}/2}\,\bar{G}}{16}\sqrt{\frac{2}{\pi}} =: L_{s_i}.$$

The cross-block Hessians are bounded on the compact set $\mathcal{B}$ by the Cauchy–Schwarz inequality, e.g. $\|\nabla^2_{\phi\nu} F\|_{\mathrm{op}} \le \sqrt{\|\nabla^2_{\phi\phi} F\|_{\mathrm{op}}\, \|\nabla^2_{\nu\nu} F\|_{\mathrm{op}}} < \infty$. Combining all blocks, the Lipschitz constant of ∇F satisfies $L_\psi = \mathcal{O}(B^2 + \bar{I} + \bar{G}\, e^{\bar{s}/2})$.

For the KL contribution, the Gaussian KL divergences have closed-form Hessians bounded by

$$\|\nabla^2_{\phi\phi}\,\mathrm{KL}(\rho_\phi\|\pi)\|_{\mathrm{op}} \le \|\Sigma_\pi^{-1}\|_{\mathrm{op}}, \qquad \|\nabla^2_{\nu\nu}\,\mathrm{KL}(\rho'_\nu\|\pi')\|_{\mathrm{op}} \le \frac{e^{\bar{s}}}{2\,\lambda_{\min}(\Sigma'_\pi)}.$$

Since both KL terms carry the prefactor $1/\lambda$, the KL contribution to smoothness is L_{KL}/λ, with

$$L_{\mathrm{KL}} := \max\left\{\|\Sigma_\pi^{-1}\|_{\mathrm{op}},\ \frac{e^{\bar{s}}}{2\,\lambda_{\min}(\Sigma'_\pi)}\right\}$$

independent of λ. Hence $\nabla\mathcal{L}_n = \nabla F + \nabla K/\lambda$ is Lipschitz with constant $L = L_\psi + L_{\mathrm{KL}}/\lambda$. □

Proposition C.9 (Strong concavity in ν)**.** *Under Assumption 4.31, for any fixed ϕ and any data realization $(X_1, \dots, X_n)$ in the support of P^{*n}, the map $\nu \mapsto \mathcal{L}_n(\phi, \nu)$ is μ-strongly concave with $\mu \ge \mu_0 > 0$ independent of both λ and the data.*

Proof. Fix a data realization $(X_1, \dots, X_n)$ in the support of P^{*n} and fix $\phi \in \Phi$. Define the

averaged contrast

$$\hat{g}(\theta') := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta\sim\rho_\phi}\big[\ell_\psi(X_i;\theta,\theta')\big]. \tag{C.44}$$

Since $\mathcal{L}_n(\phi,\nu) = G(\nu) + \frac{1}{\lambda}\,\mathrm{KL}(\rho_\phi\|\pi) - \frac{1}{\lambda}\,\mathrm{KL}(\rho'_\nu\|\pi')$ with $G(\nu) := \mathbb{E}_{\theta'\sim\rho'_\nu}[\hat{g}(\theta')]$, the term in ϕ is constant in ν while $-\frac{1}{\lambda}\,\mathrm{KL}(\rho'_\nu\|\pi')$ is concave in ν, it suffices to show that G is μ_G-strongly concave for some $\mu_G > 0$ independent of λ and the data. We first establish strong concavity of $\hat{g}$ in θ', then propagate it.

For each i, set $u_i := \log p_{\theta'}(X_i) - \log p_\theta(X_i)$. By Assumption 4.31(4), $|u_i| \le M$ for every $\theta,\theta' \in \bar{\Theta}$. The Hessian of $\ell_\psi(X_i;\theta,\theta')$ in θ' was computed in the proof of Proposition C.8:

$$\nabla^2_{\theta'\theta'}\ell_\psi(X_i;\theta,\theta') = \varphi''(u_i)\big(T(X_i)-\mu(\theta')\big)\big(T(X_i)-\mu(\theta')\big)^\top - \varphi'(u_i)\,I(\theta'). \tag{C.45}$$

For the first term, $|\varphi''(u_i)| \le 1/8$ (Lemma C.7) and the outer product has operator norm $\|T(X_i)-\mu(\theta')\|^2$, so

$$\varphi''(u_i)\big(T(X_i)-\mu(\theta')\big)\big(T(X_i)-\mu(\theta')\big)^\top \preceq \frac{1}{8}\|T(X_i)-\mu(\theta')\|^2\,\mathbb{I}_d.$$

For the second term, $|u_i| \le M$ gives $\varphi'(u_i) \ge c_M > 0$, and combined with Assumption 4.31(1), $I(\theta') \succeq \sigma_0^2\,\mathbb{I}_d$, so $-\varphi'(u_i)\,I(\theta') \preceq -c_M\,\sigma_0^2\,\mathbb{I}_d$. Combining,

$$\nabla^2_{\theta'\theta'}\ell_\psi(X_i;\theta,\theta') \preceq \left(\frac{1}{8}\|T(X_i)-\mu(\theta')\|^2 - c_M\,\sigma_0^2\right)\mathbb{I}_d. \tag{C.46}$$

Taking the expectation over $\theta\sim\rho_\phi$, averaging over i, and applying Assumption 4.31(2), so that $\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\theta\sim\rho_\phi}[\|T(X_i)-\mu(\theta')\|^2] \le B^2$,

$$\nabla^2_{\theta'\theta'}\hat{g}(\theta') \preceq \left(\frac{B^2}{8} - c_M\,\sigma_0^2\right)\mathbb{I}_d. \tag{C.47}$$

By Assumption 4.31(5), the curvature margin $\mu_0 := c_M\,\sigma_0^2 - B^2/4 > 0$, and since $B^2/8 < B^2/4$, we have $c_M\sigma_0^2 - B^2/8 > \mu_0 > 0$, so $\nabla^2_{\theta'\theta'}\hat{g}(\theta') \preceq -\mu_0\,\mathbb{I}_d$ pointwise in the data.

We propagate this through the reparameterization $\theta' = m' + D(s)^{1/2}\varepsilon'$. For the mean block, since the Loewner order is preserved under expectation,

$$\nabla^2_{m'm'}G = \mathbb{E}_{\varepsilon'}\big[\nabla^2_{\theta'\theta'}\hat{g}(m' + D(s)^{1/2}\varepsilon')\big] \preceq -\mu_0\,\mathbb{I}_d.$$

For the log-variance parameters, $\partial_{s_i}\theta' = \frac{1}{2}e^{s_i/2}\varepsilon'_i\,e_i$ and

$$\partial^2_{s_is_i}G = \frac{e^{s_i}}{4}\,\mathbb{E}\big[(\varepsilon'_i)^2\,e_i^\top\nabla^2_{\theta'\theta'}\hat{g}\,e_i\big] + \frac{e^{s_i/2}}{4}\,\mathbb{E}\big[\varepsilon'_i\,e_i^\top\nabla_{\theta'}\hat{g}\big]. \tag{C.48}$$

The first term is at most $-\mu_0 e^{\underline{s}}/4$ with $\underline{s} := \inf_{s\in\mathcal{B}}\min_i s_i > -\infty$, using $\mathbb{E}[(\varepsilon'_i)^2] = 1$

and (C.47); the second is bounded in absolute value by $e^{\bar{s}/2}\bar{G}\sqrt{2/\pi}/16$ using $\|\nabla_{\theta'}\hat{g}\| \le \bar{G}/4$ and $\mathbb{E}[|\varepsilon_i'|] = \sqrt{2/\pi}$. Hence

$$\partial^2_{s_i s_i} G \le -\frac{\mu_0\, e^{\underline{s}}}{4} + \frac{e^{\bar{s}/2}\,\bar{G}}{16}\sqrt{\frac{2}{\pi}},$$

which Assumption 4.31(5) renders strictly negative. The cross-block Hessians are bounded on $\mathcal{B}$ by Cauchy–Schwarz, so the smallest eigenvalue of $\nabla^2_{\nu\nu} G$ satisfies

$$\mu_G := \min\left\{\mu_0,\ \min_{i\in[d]}\left(\frac{\mu_0\, e^{\underline{s}}}{4} - \frac{\bar{G}\, e^{\bar{s}/2}}{16}\sqrt{\frac{2}{\pi}}\right)\right\} > 0. \tag{C.49}$$

Finally, since $\mathrm{KL}(\rho'_\nu\|\pi')$ is convex in ν, its negative is concave, so $-\frac{1}{\lambda}\nabla^2_{\nu\nu}\,\mathrm{KL}(\rho'_\nu\|\pi') \preceq 0$. Combining,

$$\nabla^2_{\nu\nu}\mathcal{L}_n(\phi,\nu) = \nabla^2_{\nu\nu} G(\nu) - \frac{1}{\lambda}\nabla^2_{\nu\nu}\,\mathrm{KL}(\rho'_\nu\|\pi') \preceq -\mu_G\,\mathbb{I}.$$

In particular $\mu \ge \mu_G \ge \mu_0 > 0$, independent of both λ and the data realization. □

Proof of Theorem 4.33. By Proposition C.8, $\mathcal{L}_n$ has Lipschitz gradient with constant $L = L_\psi + L_{\mathrm{KL}}/\lambda$, where L_ψ and L_{KL} are independent of λ. By Proposition C.9, $\nu \mapsto \mathcal{L}_n(\phi,\nu)$ is μ-strongly concave with $\mu \ge \mu_0 > 0$ independent of both λ and the data; more precisely, the KL contribution adds a concavity constant $\mu_{\mathrm{KL}}/\lambda \ge 0$, so $\mu \ge \mu_G + \mu_{\mathrm{KL}}/\lambda \ge \mu_0 > 0$. Therefore

$$\kappa = \frac{L}{\mu} = \frac{L_\psi + L_{\mathrm{KL}}/\lambda}{\mu_G + \mu_{\mathrm{KL}}/\lambda}.$$

As $\lambda \to \infty$, $\kappa \to L_\psi/\mu_G < \infty$; as $\lambda \to 0^+$, $\kappa \to L_{\mathrm{KL}}/\mu_{\mathrm{KL}} < \infty$. Since the numerator and denominator are continuous in $1/\lambda$ and the denominator is bounded away from zero,

$$\kappa \le \max\left\{\frac{L_\psi}{\mu_G}, \frac{L_{\mathrm{KL}}}{\mu_{\mathrm{KL}}}\right\} = \mathcal{O}(1) \qquad \text{uniformly in } \lambda > 0.$$

□

Proof of Corollary 4.34. Since $\mathcal{L}_n$ is L-smooth (Proposition C.8) and μ-strongly concave in ν (Proposition C.9), the saddle-point problem $\min_\phi \max_\nu \mathcal{L}_n(\phi,\nu)$ has nonconvex–strongly-concave structure. The convergence rate of the projected stochastic extragradient method then follows from the theory of Juditsky et al. (2011) and Lin et al. (2020). Since the condition number $\kappa = \mathcal{O}(1)$ uniformly in λ (Theorem 4.33), the guarantee holds for any $\lambda > 0$, including the statistically optimal choice $\lambda = \Theta(n)$. □

Appendix D

Additional Results

D.1 Alternative unbiased importance sampling estimators

Here we recall two other approaches to unbiased importance sampling mentioned in the beginning of Section 2.5, which are not based on coupling PIMH chains. These are implemented in the experiments of Section 2.6.

D.1.1 MLMC-UIS

One approach to removing the bias from a sequence of consistent estimates such as SNIS estimates as $N \to \infty$, is based on multilevel Monte Carlo and randomization; we refer to it as MLMC-UIS. This is the approach described in Blanchet & Glynn (2015); Wang & Wang (2023), and specifically applied to SNIS in Shi & Cornish (2021, Section 4). Let ρ in $(0, 1)$ be the parameter of a Geometric distribution, with $\mathbb{P}(R = k) = (1-\rho)^k \rho$ for $k \geq 0$, and $\mathbb{P}(R \geq k) = (1-\rho)^k$. Introduce the following notation:

- For a sample $x_1, \ldots, x_n$, with n even, we write $E(x_1, \ldots, x_n)$ for the sample $(x_{2i})_{i=1}^{n/2}$ of the even-indexed variables, and $O(x_1, \ldots, x_n)$ for the sample $(x_{2i-1})_{i=1}^{n/2}$ of the odd-indexed variables.

- We write $\hat{F}(x_1, \ldots, x_n)$ for the SNIS estimator of $\pi(f)$ based on the sample $x_1, \ldots, x_n$, and $\hat{F} \circ E(x_1, \ldots, x_n)$ for the SNIS estimator of $\pi(f)$ based on the even-indexed variables, and $\hat{F} \circ O(x_1, \ldots, x_n)$ for the SNIS estimator of $\pi(f)$ based on the odd-indexed variables.

The MLMC-UIS estimator is described in Algorithm 7.

Algorithm 7. MLMC-UIS

1. Choose an initial level $\ell \geq 0$. You can e.g. take $\ell = 1$.
2. Draw $R \sim \text{Geometric}(\rho)$.
3. Draw $x_1, \ldots, x_{2^{\ell+R+1}}$ i.i.d. samples from q. You may compute once and for all $\omega(x_i)$ and $f(x_i)$ for $i = 1, \ldots, 2^{\ell+R+1}$, as these will be used in the different SNIS estimators computed below.
4. Compute

$$S_R = \hat{F}(x_1, \ldots, x_{2^{\ell+R+1}}), \tag{D.1}$$
$$S_R^O = \hat{F} \circ O(x_1, \ldots, x_{2^{\ell+R+1}}), \tag{D.2}$$
$$S_R^E = \hat{F} \circ E(x_1, \ldots, x_{2^{\ell+R+1}}). \tag{D.3}$$

5. Compute

$$\Delta_R = S_R - \frac{S_R^O + S_R^E}{2}, \tag{D.4}$$

whose expectation equals the bias difference between SNIS estimators with $2^{\ell+R+1}$ and $2^{\ell+R}$ samples.

6. Return either the "single sample" or the "Russian roulette" estimator, defined as follows:

$$\text{SS} = \frac{1}{2^{R+1}} \sum_{n=1}^{2^{R+1}} \hat{F}(x_{(n-1)2^\ell+1}, \ldots, x_{n2^\ell}) + \frac{\Delta_R}{(1-\rho)^R \rho}, \tag{D.5}$$
$$\text{RR} = \frac{1}{2^{R+1}} \sum_{n=1}^{2^{R+1}} \hat{F}(x_{(n-1)2^\ell+1}, \ldots, x_{n2^\ell}) + \sum_{r=0}^{R} \frac{\Delta_r}{(1-\rho)^r}. \tag{D.6}$$

The first term in both the SS and RR estimators represents the average of many SNIS estimators of $\pi(f)$, each based on 2^ℓ samples. This is biased for $\pi(f)$. The second terms in SS and RR are bias correction terms.

Remark D.1. If the test function f is chosen to be $x \mapsto \omega_u(x)^{-1}$, where ω_u is the unnormalized weight function and $Z = q(\omega_u)$, then the unbiased MLMC estimators are unbiased estimators of $1/Z$.

The choice of the level ℓ and the Geometric parameter ρ, together with the finite-variance and finite-cost conditions of Shi & Cornish (2021) and the value used in the experiments, is deferred to Appendix D.2.

D.1.2 Taylor-UIS

Another approach to removing the bias of SNIS is based on Taylor expansions and randomization, and was proposed in Blanchet et al. (2015), with useful improvements and analyses in Chopin et al. (2025). In these articles, the goal is to estimate a smooth function of the expectation of a random variable that we can sample from. In the context of importance sampling, the smooth function is the inverse function, and we apply the method of Chopin et al. (2025) to estimate $1/Z$ without bias. We can then multiply the resulting estimator by an unbiased estimator of $q(f\omega)$ to obtain an unbiased estimator of $\pi(f)$. We refer to this approach as Taylor-UIS.

From $g : m \mapsto 1/m$, the starting point is the identity

$$g(m) = \sum_{k=0}^{\infty} \gamma_k \left(\frac{m}{x_0} - 1 \right)^k , \tag{D.7}$$

where x_0 is a tuning parameter, and the coefficients γ_k are given by

$$\forall k \geq 1 \quad \gamma_k = \frac{g^{(k)}(x_0)}{k!} x_0^k = \frac{(-1)^k}{x_0}. \tag{D.8}$$

The basic idea is that:

1. The infinite sum in (D.7) can be estimated without bias with a truncated sum of reweighted terms. Introduce $R \sim \text{Geometric}(\rho)$, such that $\mathbb{P}(R = k) = (1-\rho)^k \rho$ for $k \geq 0$, where $\rho \in (0, 1)$. Note that $\mathbb{P}(R \geq k) = (1-\rho)^k$.

2. Each term $(m/x_0 - 1)^k$ in (D.7) can be unbiasedly estimated using draws from Y with expectation m. The most natural estimator is (Blanchet et al., 2015):

$$U_{r,k} = \prod_{i=1}^{k} \left(\frac{Y_i}{x_0} - 1 \right). \tag{D.9}$$

   However, a better estimator is as follows: if one has access to r random variables $Y_1, \ldots, Y_r$ with $r > k$, then one can average over different possible products of k distinct variables. Chopin et al. (2025) consider averaging over r circular shifts of the variables:

$$U_{r,k}^C = \frac{1}{r} \left\{ \sum_{j=0}^{r-1} \prod_{i=1}^{k} \left(\frac{Y_{j+i}}{x_0} - 1 \right) \right\}, \tag{D.10}$$

   where the index $j + i$ is taken modulo r, i.e. $Y_{r+1} = Y_1$, $Y_{r+2} = Y_2$, etc.

Combining the above considerations, an unbiased estimator of $1/m$ is

$$R \sim \text{Geometric}(\rho), \quad Y_1, \dots, Y_R \sim \mathcal{L}(m, \sigma^2), \quad \text{CCS} = \sum_{k=0}^{R} \frac{(-1)^k U_{R,k}^C}{x_0 \mathbb{P}(R \geq k)}. \tag{D.11}$$

where $\mathcal{L}(m, \sigma^2)$ is the distribution of Y_i with expectation m and variance σ^2. In our setting each Y_i corresponds to $\hat{Z} = N^{-1} \sum_{n=1}^{N} \omega(x_n)$, where x_n are i.i.d. samples from q, and thus we need to choose a tuning parameter N. The two additional parameters are x_0 and ρ. The resulting estimator is described in Algorithm 8.

Algorithm 8. Taylor-UIS

1. Choose the tuning parameters x_0 and ρ from a pilot sample of Y_i's, as described in Appendix D.2.
2. Draw $R \sim \text{Geometric}(\rho)$.
3. Draw R i.i.d. samples $Y_1, \dots, Y_R$, each $Y_i = \hat{Z} = N^{-1} \sum_{n=1}^{N} \omega(x_n)$ with x_n i.i.d. from q.
4. Compute $U_{R,k}^C$ for $k = 1, \dots, R$ using (D.10).
5. Return the estimator CCS of (D.11), an unbiased estimator of $1/m$.

The pilot-sample choice of the tuning parameters x_0 and ρ, following Chopin et al. (2025), is deferred to Appendix D.2. Having obtained ρ and x_0, we run the CCS estimator of Algorithm 8, which expands to

$$\text{CCS} = x_0^{-1} + \sum_{k=1}^{R} \frac{(-1)^k U_{R,k}^C}{x_0 (1-\rho)^k}.$$

A natural estimator of $q(\omega \cdot f)$ is given by $M^{-1} \sum_{n=1}^{M} \omega(x'_n) f(x'_n)$, where $x'_1, \dots, x'_M$ are i.i.d. samples from q, independent of all the samples used to obtain the unbiased estimator of Z^{-1}. Finally we obtain

$$\hat{\pi}_{\text{CCS}}(f) = \text{CCS} \times \left(\frac{1}{M} \sum_{n=1}^{M} \omega(x'_n) f(x'_n) \right).$$

One would naturally need to tune M relative to the tuning of each Y_i (which itself is a $\hat{Z}_N$ estimator). In the application of Section 2.6.1, our interest is in estimating $\pi(f)$ for $f(x) = \omega(x)^{-1}$, which is equivalent to estimating $1/Z$, so we do not need to estimate $q(\omega \cdot f)$, and we can simply use the CCS estimator of $1/Z$. We leave the question of how to optimally split the budget between estimating Z^{-1} and estimating $q(\omega \cdot f)$ for future work.

D.2 Implementation Details

This section collects the choice of tuning parameters for the unbiased importance-sampling estimators of Chapter 2, namely MLMC-UIS (Algorithm 7) and Taylor-UIS (Algorithm 8).

D.2.1 Tuning of MLMC-UIS

The tuning parameters are ℓ, which can be set to e.g. one, and the parameter ρ of the Geometric distribution, which controls the tail of the distribution of R, and thus both the cost and variance of the estimator. Theorem 3 of Shi & Cornish (2021) provides conditions under which the variance of the SS and RR estimators is finite, and the expected cost is finite. There are two sets of conditions:

- Assume that $q(\omega^a + |f|^b) < \infty$ with $a > 4$ and $b > 2a/(a-4)$. Then set $\alpha = 1 - 2/b$.
- Assume that $q(\omega^a + |f|^b) < \infty$ with $2 < a \leq 4$ and $b > (2a+4)/(a-2)$, or $a > 4, b \in ((2a+4)/(a-2), 2a/(a-4))$. Then set $\alpha = a/2 - (a+2)/b - 1$.

Then the estimator has finite expected cost and finite variance for $\rho \in (2^{-1}, 1 - 2^{-1-\alpha})$. The larger the α, the larger the range of ρ for which the estimator has finite variance and finite expected cost. In the Exponential example considered in Section 2.6.1, with $p = 3$, we are in the case $2 < a \leq 4$. Considering the function $x \mapsto \omega(x)^{-1}$, we can take b arbitrarily large. Therefore, we select $\alpha = 1/2$, define $\rho_{\max} = 1 - 2^{-1-\alpha} = 1 - 2^{-3/2} \approx 0.6464$, and choose $\rho = 0.95\, \rho_{\max}$ in the experiments to reflect a slightly conservative but well-informed choice.

D.2.2 Tuning of Taylor-UIS

Chopin et al. (2025) provide guidance on tuning parameters in their Section 3. Introduce the following notation:

$$\beta_0 = |m/x_0 - 1|, \sigma^2 = \mathbb{V}[Y_i],$$
$$\hat{m} = \frac{1}{n_0}\sum_{i=1}^{n_0} Y_i, \quad \hat{\sigma}^2 = \frac{1}{n_0 - 1}\sum_{i=1}^{n_0}(Y_i - \hat{m})^2,$$

where n_0 is the length of a pilot sample of Y_i's, used to estimate m and σ^2. Chopin et al. (2025) choose $n_0 = 10$. Then define

$$\hat{x_0^\star} = (\hat{m}^2 + \hat{\sigma}^2)/\hat{m},$$
$$x_{0,\text{ub}} = \text{bootstrap percentile at level } (1-\alpha) \text{ of } (\hat{m}^2 + \hat{\sigma}^2)/(2\hat{m}),$$
$$x_0 = \max(\hat{x_0^\star}, x_{0,\text{ub}}),$$

where the bootstrap is performed by resampling with replacement from the pilot sample of Y_i's. By default we follow Chopin et al. (2025) and use 100 bootstrap samples and $\alpha = 0.01$. Having defined x_0, we can define

$$\hat{\beta}^2 = \frac{\hat{\sigma}^2}{x_0^2} + \left(\frac{\hat{m}}{x_0} - 1\right)^2,$$

and

$$\rho = \min(1 - \hat{\beta}^2, 1/(n_0 + 1)).$$

See Section 3.2 of Chopin et al. (2025) for details on the choice of ρ.

D.3 Additional Numerical Results

This section reports supplementary numerical illustrations for the importance-sampling and coupling results of Chapter 2. Throughout, we take the Exponential setting of Example 2.3: the target is the Exponential(1) distribution and the proposal the Exponential(k) distribution on $\mathbb{R}_+$, whose weight $\omega = \pi/q$ satisfies $q(\omega^p) < \infty$ for every $p < k/(k-1)$; the two reported moment orders $p \in \{3, 5\}$ correspond to $k = 3/2$ and $k = 5/4$. Figure D.1 illustrates the meeting time of the coupled IMH chains and its scaling, and Figure D.2 the variance and inefficiency of the Coupling-UIS estimator, both as functions of the number of proposals N and of the moment order p.

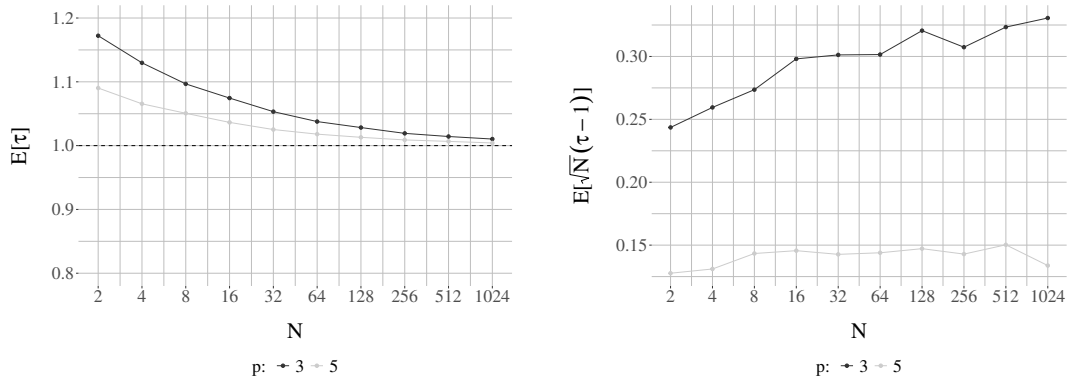


Figure D.1: Meeting time τ of the common-draws coupling of IMH (Algorithm 4, shown maximal in Theorem 2.11), whose finite moments under $q(\omega^p) < \infty$ drive the polynomial convergence rate of Theorem 2.16. Left: the average meeting time $\mathbb{E}[\tau]$, against the number of proposals N, for $p \in \{3, 5\}$; it decreases towards 1 as N grows. Right: the centred and rescaled meeting time $\mathbb{E}[\sqrt{N}\,(\tau - 1)]$, which stabilises as $N \to \infty$, consistent with $\tau = 1 + \mathcal{O}_p(N^{-1/2})$.

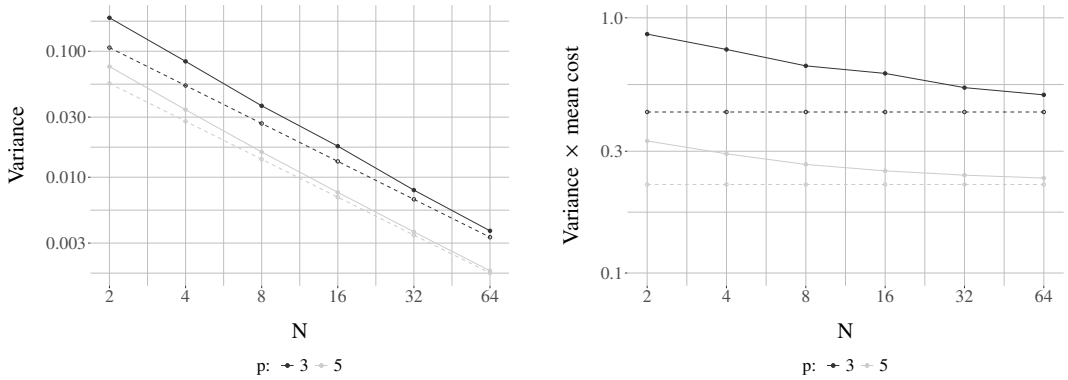


Figure D.2: Efficiency of the Coupling-UIS estimator, illustrating the asymptotic price of bias removal (Propositions 2.30 and 2.31). Left: its variance (solid) matches the asymptotic SNIS variance $\sigma_{\mathrm{IS}}^2/(2N)$ (dashed), for $p \in \{3, 5\}$. Right: its inefficiency (variance $\times$ mean cost) approaches twice the SNIS asymptote (dashed) as N grows, the price paid for exact unbiasedness.

D.4 Leave-one-out SNIS and its asymptotic bias

This section characterizes the asymptotic bias of the leave-one-out self-normalized importance sampling estimator (SNIS-LOO) of Skare et al. (2003), used in the nested-expectation experiment of Chapter 2. Under moment assumptions on the importance weights, SNIS-LOO achieves an asymptotic bias of order N^{-2}, an improvement over the order N^{-1} bias of standard self-normalized importance sampling. To our knowledge, this is the first proof of this result, and it may be of independent interest.

We adopt the notation of Chapter 2 throughout: the particle vector is $\mathbf{X}^N = (X_1, \dots, X_N)$ with $X_n \sim q$ independently, the importance weight is $\omega(X) = \pi(X)/q(X)$ with $q(\omega) = 1$, and $[N] = \{1, \dots, N\}$. Algorithm 9 describes the leave-one-out self-normalized importance sampling (SNIS-LOO) estimator introduced in Skare et al. (2003). The name reflects the defining feature of the method: the weight assigned to each particle is self-normalized using the sum of the weights of all other particles, that is, leaving that particle out.

Theorem D.2. *With the notation of Chapter 2, let $\hat{F}_{\text{SNIS-LOO}}(\mathbf{X}^N)$ be the SNIS-LOO estimator of Algorithm 9 for a test function f, with leave-one-out weights as in* (D.13)*. Assume that*

$$q(\omega\,|f - \pi(f)|) < \infty, \quad q(\omega^7\,|f - \pi(f)|) < \infty, \quad q(\omega^8) < \infty, \\ \text{and} \quad q(\omega^{-\eta}) < \infty \text{ for some } \eta > 0. \tag{D.15}$$

Then

$$\lim_{N\to\infty} N^2 \times \mathbb{E}_{\mathbf{X}^N \sim \bar{q}}\left[\hat{F}_{\text{SNIS-LOO}}(\mathbf{X}^N) - \pi(f)\right] = -q(\omega^3 \cdot (f - \pi(f))) \\ + 2q(\omega^2) \cdot q(\omega^2 \cdot (f - \pi(f))). \tag{D.16}$$

Algorithm 9. Leave-one-out self-normalized importance sampling (SNIS-LOO)

1. Sample N particles independently $\mathbf{X}^N = (X_1, \dots, X_N)$ from $\bar{q} = \bigotimes_{n=1}^N q$.
2. Compute the importance weights $\omega(X_n) = \pi(X_n)/q(X_n)$ for $n \in [N]$.
3. Compute the total sum of weights:

$$S_N(\mathbf{X}^N) = \sum_{j=1}^N \omega(X_j). \tag{D.12}$$

4. Define the *leave-one-out* weights

$$\tilde{\omega}_n(\mathbf{X}^N) = \frac{\dfrac{\omega(X_n)}{S_N(\mathbf{X}^N) - \omega(X_n)}}{\displaystyle\sum_{k=1}^N \frac{\omega(X_k)}{S_N(\mathbf{X}^N) - \omega(X_k)}}, \quad n \in [N]. \tag{D.13}$$

5. Return the estimate

$$\hat{F}_{\text{SNIS-LOO}}(\mathbf{X}^N) = \sum_{n=1}^N \tilde{\omega}_n(\mathbf{X}^N) f(X_n). \tag{D.14}$$

The proof below uses the weighted moments $q(\omega^r|f - \pi(f)|)$ for every $r \in \{1, 2, 3, 4, 5, 7\}$. As the following lemma shows, it is enough to control the two endpoints $r = 1$ and $r = 7$, which is why only these appear in (D.15).

Lemma D.3. *Let $g \geq 0$ be measurable. If $q(\omega\, g) < \infty$ and $q(\omega^7\, g) < \infty$, then $q(\omega^r\, g) < \infty$ for every $r \in [1, 7]$.*

Proof. For $r \in [1, 7]$ and $\omega > 0$ we have the pointwise bound

$$\omega^r \leq \mathbb{1}(\omega > 1)\, \omega^7 + \mathbb{1}(\omega \leq 1)\, \omega \leq \omega^7 + \omega,$$

since $\omega^r \leq \omega^7$ when $\omega > 1$ and $\omega^r \leq \omega$ when $\omega \leq 1$. Multiplying by $g \geq 0$ and integrating against q gives $q(\omega^r g) \leq q(\omega^7 g) + q(\omega\, g) < \infty$. □

Theorem D.2 assumes a finite inverse moment of the weight, and for bounded f it requires $q(\omega^8) < \infty$. The asymptotic bias is of order N^{-2}, which is a significant improvement over the order N^{-1} asymptotic bias of self-normalized importance sampling established in Theorem 2.8 of Chapter 2. The proof relies on the inverse-moment bound of Proposition A.1 of Chapter 2, together with the two auxiliary results stated next.

Lemma D.4 (Zero-mean identity). *For any measurable g with $q(\omega\,|g|) < \infty$,*

$\mathbb{E}_{X\sim q}[\omega(X)\, g(X)] = \pi(g)$. *In particular, if* $q(\omega\,|f|) < \infty$, *then*

$$\mathbb{E}_{X\sim q}[\omega(X)\,(f(X) - \pi(f))] = 0.$$

Proof. Since $\omega = \pi/q$, we have $\mathbb{E}_{X\sim q}[\omega(X)g(X)] = \int(\pi(X)/q(X))\, g(X)\, q(X)\,\mathrm{d}X = \int g(X)\,\pi(X)\,\mathrm{d}X = \pi(g)$. Taking $g = f - \pi(f)$ gives $\pi(f) - \pi(f) = 0$. □

Lemma D.5 (Convergence of normalized power sums). *Suppose* $q(\omega^{-\eta}) < \infty$ *for some* $\eta > 0$, *and recall* $\bar{T}_N^{(k)} = N^{-1}\sum_{j=2}^N \omega^k(X_j)$ *and* $\bar{T}_N = \bar{T}_N^{(1)}$. *Then* $\bar{T}_N \to 1$ *almost surely, and* $\bar{T}_N^{(k)} \to q(\omega^k)$ *almost surely for every integer* $k \geq 1$ *with* $q(\omega^k) < \infty$. *Moreover, for integers* $k \geq 0$ *and* $m \geq 1$ *such that* $q(\omega^{k+\epsilon}) < \infty$ *for some* $\epsilon > 0$, *the family* $\{\bar{T}_N^{(k)}/\bar{T}_N^m\}_N$ *is uniformly integrable and*

$$\lim_{N\to\infty} \mathbb{E}\left[\frac{\bar{T}_N^{(k)}}{\bar{T}_N^m}\right] = q(\omega^k),$$

with the conventions $\bar{T}_N^{(0)} = (N-1)/N$ *and* $q(\omega^0) = 1$.

Proof. The almost sure limits follow from the strong law of large numbers, using $q(\omega) = \pi(\mathbb{X}) = 1$ for $\bar{T}_N$. Since $\bar{T}_N^{(k)}/\bar{T}_N^m \to q(\omega^k)$ almost surely, it suffices to upgrade this to convergence in L^1, for which uniform integrability is enough. By Hölder's inequality, for some $\delta > 0$, $p > 1$ close to 1, and its conjugate exponent p',

$$\mathbb{E}\left[\left(\frac{\bar{T}_N^{(k)}}{\bar{T}_N^m}\right)^{1+\delta}\right] \leq \left(\mathbb{E}\left[(\bar{T}_N^{(k)})^{(1+\delta)p}\right]\right)^{1/p}\left(\mathbb{E}\left[\bar{T}_N^{-(1+\delta)mp'}\right]\right)^{1/p'}.$$

The first factor is bounded uniformly in N when $q(\omega^{k+\epsilon}) < \infty$ for some $\epsilon > 0$ (choosing δ and $p-1$ small), and the second is bounded uniformly in N by Proposition A.1 for N large enough. Hence $\sup_N \mathbb{E}[(\bar{T}_N^{(k)}/\bar{T}_N^m)^{1+\delta}] < \infty$, which yields uniform integrability (Billingsley, 1999, Eq. (3.18)) and therefore convergence of the expectations. □

Proof of Theorem D.2. Using $\sum_{n=1}^N \tilde{\omega}_n(\mathbf{X}^N) = 1$ and that the X_n are identically distributed, the rescaled bias of SNIS-LOO is

$$N^2 \times \mathbb{E}_{\mathbf{X}^N\sim\bar{q}}\left[\hat{F}_{\text{SNIS-LOO}}(\mathbf{X}^N) - \pi(f)\right] = N^2\mathbb{E}\left[\sum_{n=1}^N \tilde{\omega}_n(\mathbf{X}^N)(f(X_n) - \pi(f))\right] \quad \text{(D.17)}$$

$$= N^3\mathbb{E}_{\mathbf{X}^N\sim\bar{q}}\left[\tilde{\omega}_1(\mathbf{X}^N)(f(X_1) - \pi(f))\right]. \quad \text{(D.18)}$$

Define for $k \geq 1$

$$T_N^{(k)} = \sum_{j=2}^N \omega^k(X_j), \quad T_N = T_N^{(1)}, \quad \bar{T}_N = \frac{T_N}{N}, \quad \bar{T}_N^{(k)} = \frac{T_N^{(k)}}{N}.$$

Note that $T_N = S_N(\mathbf{X}^N) - \omega(X_1)$. From the definition of the leave-one-out weights in (D.13), we have

$$\begin{aligned}\tilde{\omega}_1(\mathbf{X}^N) &= \frac{\dfrac{\omega(X_1)}{T_N}}{\displaystyle\sum_{k=1}^{N} \frac{\omega(X_k)}{T_N + \omega(X_1) - \omega(X_k)}} \\ &= \frac{\omega(X_1)}{T_N \left(\dfrac{\omega(X_1)}{T_N + \omega(X_1) - \omega(X_1)} + \displaystyle\sum_{k=2}^{N} \frac{\omega(X_k)}{T_N + \omega(X_1) - \omega(X_k)} \right)}.\end{aligned}$$

Thus,

$$\tilde{\omega}_1(\mathbf{X}^N) = \frac{\omega(X_1)}{\omega(X_1) + \displaystyle\sum_{k=2}^{N} \frac{\omega(X_k)}{1 + \dfrac{\omega(X_1) - \omega(X_k)}{N\bar{T}_N}}}. \tag{D.19}$$

Next, we expand the denominator of (D.19), piece by piece. We start by defining for each $k \geq 2$

$$u_k = \frac{\omega(X_1) - \omega(X_k)}{N\bar{T}_N}, \tag{D.20}$$

and use the following expansion of $1/(1+u_k)$ to second order in u_k

$$\frac{1}{1+u_k} = 1 - u_k + u_k^2 - \frac{u_k^3}{1+u_k} = 1 - u_k + u_k^2 + R_k^u. \tag{D.21}$$

Thus, we obtain

$$\begin{aligned}\frac{\omega(X_k)}{1+u_k} &= \omega(X_k)(1 - u_k + u_k^2 + R_k^u) \\ &= \omega(X_k) - \omega(X_k)u_k + \omega(X_k)u_k^2 + \omega(X_k)R_k^u.\end{aligned}$$

Substituting u_k with its definition in (D.20), we obtain

$$\begin{aligned}\frac{\omega(X_k)}{1+u_k} = \omega(X_k) &- \frac{\omega(X_k)\omega(X_1)}{N\bar{T}_N} + \frac{\omega(X_k)^2}{N\bar{T}_N} + \frac{\omega(X_k)\omega(X_1)^2}{N^2\bar{T}_N^2} \\ &- \frac{2\omega(X_k)^2\omega(X_1)}{N^2\bar{T}_N^2} + \frac{\omega(X_k)^3}{N^2\bar{T}_N^2} + \omega(X_k)R_k^u,\end{aligned}$$

where the remainder term is given by

$$\omega(X_k)R_k^u = \omega(X_k) \cdot \frac{-u_k^3}{1+u_k} = \frac{-\omega(X_k)(\omega(X_1) - \omega(X_k))^3}{N^3\bar{T}_N^3(1+u_k)}.$$

Summing over $k = 2, \ldots, N$, we obtain

$$\sum_{k=2}^{N} \frac{\omega(X_k)}{1+u_k} = N\bar{T}_N - \omega(X_1) + \frac{\bar{T}_N^{(2)}}{\bar{T}_N} + \frac{\omega(X_1)^2}{N\bar{T}_N} - \frac{2\omega(X_1)\bar{T}_N^{(2)}}{N\bar{T}_N^2} + \frac{\bar{T}_N^{(3)}}{N\bar{T}_N^2} + \tilde{R}_N^u, \tag{D.22}$$

where the remainder term is given by

$$\tilde{R}_N^u = \sum_{k=2}^{N} \omega(X_k) R_k^u = \sum_{k=2}^{N} \frac{-\omega(X_k)(\omega(X_1) - \omega(X_k))^3}{N^3 \bar{T}_N^3 (1+u_k)}. \tag{D.23}$$

Observe that $\tilde{R}_N^u = \mathcal{O}_p(N^{-2})$: it is a sum of $N-1$ terms, each of order $\mathcal{O}_p(N^{-3})$.

The denominator of $\tilde{\omega}_1(\mathbf{X}^N)$ in (D.19) becomes

$$\omega(X_1) + \sum_{k=2}^{N} \frac{\omega(X_k)}{1+u_k} = N\bar{T}_N + \frac{\bar{T}_N^{(2)}}{\bar{T}_N} + \frac{\omega(X_1)^2}{N\bar{T}_N} - \frac{2\omega(X_1)\bar{T}_N^{(2)}}{N\bar{T}_N^2} + \frac{\bar{T}_N^{(3)}}{N\bar{T}_N^2} + \tilde{R}_N^u. \tag{D.24}$$

Factoring out $N\bar{T}_N$ from the denominator, we obtain

$$\tilde{\omega}_1(\mathbf{X}^N) = \frac{\omega(X_1)}{N\bar{T}_N} \cdot \frac{1}{1 + \frac{\bar{T}_N^{(2)}}{N\bar{T}_N^2} + \frac{\omega(X_1)^2}{N^2\bar{T}_N^2} - \frac{2\omega(X_1)\bar{T}_N^{(2)}}{N^2\bar{T}_N^3} + \frac{\bar{T}_N^{(3)}}{N^2\bar{T}_N^3} + \frac{\tilde{R}_N^u}{N\bar{T}_N}}. \tag{D.25}$$

Next, let

$$v_N(\mathbf{X}^N) = \frac{\bar{T}_N^{(2)}}{N\bar{T}_N^2} + \frac{\omega(X_1)^2}{N^2\bar{T}_N^2} - \frac{2\omega(X_1)\bar{T}_N^{(2)}}{N^2\bar{T}_N^3} + \frac{\bar{T}_N^{(3)}}{N^2\bar{T}_N^3} + \frac{\tilde{R}_N^u}{N\bar{T}_N}. \tag{D.26}$$

Recalling that $\tilde{R}_N^u = \mathcal{O}_p(N^{-2})$, we see that $v_N(\mathbf{X}^N) = \mathcal{O}_p(N^{-1})$, its first term being the leading one: the next three terms are $\mathcal{O}_p(N^{-2})$ and the last is $\mathcal{O}_p(N^{-3})$, using that $\bar{T}_N^{(k)} = \mathcal{O}_p(1)$ for $k = 1, 2, 3$ by Lemma D.5. Note that from (D.25) and (D.13), we have

$$\frac{1}{1+v_N(\mathbf{X}^N)} = \frac{1}{\sum_{k=1}^{N} \frac{\omega(X_k)}{S_N(\mathbf{X}^N) - \omega(X_k)}}. \tag{D.27}$$

Using a first-order expansion of $(1+v_N(\mathbf{X}^N))^{-1}$, we have

$$\frac{1}{1+v_N(\mathbf{X}^N)} = 1 - v_N(\mathbf{X}^N) + \frac{v_N(\mathbf{X}^N)^2}{1+v_N(\mathbf{X}^N)},$$

and substituting the first-order expansion back into (D.25), we obtain

$$\tilde{\omega}_1(\mathbf{X}^N) = \frac{\omega(X_1)}{N\bar{T}_N} \cdot \frac{1}{1+v_N(\mathbf{X}^N)} = \frac{\omega(X_1)}{N\bar{T}_N}\left(1 - v_N(\mathbf{X}^N) + \frac{v_N(\mathbf{X}^N)^2}{1+v_N(\mathbf{X}^N)}\right). \quad \text{(D.28)}$$

Next, we expand term by term using the definition of $v_N(\mathbf{X}^N)$, and group by order of N

$$\begin{aligned}\tilde{\omega}_1(\mathbf{X}^N) = \frac{\omega(X_1)}{N\bar{T}_N} - \frac{\omega(X_1)\bar{T}_N^{(2)}}{N^2\bar{T}_N^3} - \frac{\omega(X_1)^3}{N^3\bar{T}_N^3} + \frac{2\omega(X_1)^2\bar{T}_N^{(2)}}{N^3\bar{T}_N^4} - \frac{\omega(X_1)\bar{T}_N^{(3)}}{N^3\bar{T}_N^4} \\ - \frac{\omega(X_1)\tilde{R}_N^u}{N^2\bar{T}_N^2} + \frac{\omega(X_1)v_N(\mathbf{X}^N)^2}{N\bar{T}_N(1+v_N(\mathbf{X}^N))}.\end{aligned} \quad \text{(D.29)}$$

Then, we take expectations of the terms in (D.29), multiplied by $(f(X_1) - \pi(f))$. Recall from (D.18) that we are interested in identifying the terms that lead to non-zero expectations of order N^{-3}.

1. Since X_1 is independent of $(X_2, \ldots, X_N)$, the expectation of the first term factorizes, and Lemma D.4 makes it vanish:

$$\mathbb{E}\left[\frac{\omega(X_1)(f(X_1)-\pi(f))}{N\bar{T}_N}\right] = \frac{1}{N}\,\mathbb{E}\left[\frac{1}{\bar{T}_N}\right]\,\mathbb{E}[\omega(X_1)(f(X_1)-\pi(f))] = 0, \quad \text{(D.30)}$$

   where $\mathbb{E}[1/\bar{T}_N] < \infty$ for N large by Proposition A.1.

2. By the same factorization, the second term also vanishes through Lemma D.4:

$$\mathbb{E}\left[\frac{\omega(X_1)\bar{T}_N^{(2)}(f(X_1)-\pi(f))}{N^2\bar{T}_N^3}\right] = \mathbb{E}[\omega(X_1)(f(X_1)-\pi(f))] \times \mathbb{E}\left[\frac{\bar{T}_N^{(2)}}{N^2\bar{T}_N^3}\right] = 0,$$

   where the second expectation is finite for N large by Lemma D.5 under $q(\omega^{2+\epsilon}) < \infty$.

3. For the third term, independence gives

$$\mathbb{E}\left[-\frac{\omega(X_1)^3(f(X_1)-\pi(f))}{N^3\bar{T}_N^3}\right] = -N^{-3}\,\mathbb{E}\left[\frac{1}{\bar{T}_N^3}\right]\,\mathbb{E}\left[\omega(X_1)^3(f(X_1)-\pi(f))\right].$$

   By Lemma D.5, $\mathbb{E}[\bar{T}_N^{-3}] \to 1$ as $N \to \infty$, so this term contributes $-q(\omega^3(f-\pi(f)))$, the first term on the right-hand side of (D.16).

4. For the fourth term, independence gives

$$\mathbb{E}\left[\frac{2\omega(X_1)^2\bar{T}_N^{(2)}(f(X_1)-\pi(f))}{N^3\bar{T}_N^4}\right] = \frac{2}{N^3}\,\mathbb{E}\left[\frac{\bar{T}_N^{(2)}}{\bar{T}_N^4}\right] \cdot q(\omega^2(f-\pi(f))). \quad \text{(D.31)}$$

By Lemma D.5 (with $q(\omega^{2+\epsilon}) < \infty$), $\mathbb{E}[\bar{T}_N^{(2)}/\bar{T}_N^4] \to q(\omega^2)$, so this term contributes $2q(\omega^2)\, q(\omega^2(f - \pi(f)))$, the second term on the right-hand side of (D.16).

5. By independence and Lemma D.4, the fifth term vanishes:

$$\mathbb{E}\left[\frac{\omega(X_1)\bar{T}_N^{(3)}(f(X_1) - \pi(f))}{N^3\bar{T}_N^4}\right] = \frac{1}{N^3}\,\mathbb{E}[\omega(X_1)(f(X_1) - \pi(f))]\;\mathbb{E}\left[\frac{\bar{T}_N^{(3)}}{\bar{T}_N^4}\right] = 0, \tag{D.32}$$

where $\mathbb{E}[\bar{T}_N^{(3)}/\bar{T}_N^4] < \infty$ for N large by Lemma D.5 under $q(\omega^{3+\epsilon}) < \infty$.

6. For the sixth term, we can write it as

$$\sum_{k=2}^{N} \frac{-\omega(X_1)\omega(X_k)(\omega(X_1) - \omega(X_k))^3(f(X_1) - \pi(f))}{N^5\bar{T}_N^5(1 + (\omega(X_1) - \omega(X_k))/(N\bar{T}_N))}. \tag{D.33}$$

We can view the denominator as

$$N^5\bar{T}_N^5(1 + (\omega(X_1) - \omega(X_k))/(N\bar{T}_N)) = N^4\bar{T}_N^4 \cdot (\sum_{j\neq k}\omega(X_j)).$$

For the lower bound, we employ a leave-two-out technique by defining

$$\widetilde{T}_{N,k} = \frac{1}{N-2}\sum_{j\notin\{1,k\}}\omega(X_j),$$

which excludes both particles 1 and k. Since all weights are non-negative

$$\sum_{j\neq k}\omega(X_j) \geq (N-2)\widetilde{T}_{N,k}, \quad \text{and} \quad \bar{T}_N \geq \frac{N-2}{N}\widetilde{T}_{N,k}. \tag{D.34}$$

This implies that the denominator can be bounded below by $(N-2)^5\widetilde{T}_{N,k}^5$, which grows as N^5 times the fifth power of the leave-two-out average. The numerator expands to 4 terms

$$\begin{aligned}&-\omega(X_1)^4\omega(X_k)(f(X_1) - \pi(f)) + 3\omega(X_1)^3\omega(X_k)^2(f(X_1) - \pi(f))\\ &\quad - 3\omega(X_1)^2\omega(X_k)^3(f(X_1) - \pi(f)) + \omega(X_1)\omega(X_k)^4(f(X_1) - \pi(f)).\end{aligned}$$

When taking the expectation corresponding to any of the terms in the numerator

$$\left|\mathbb{E}\left[\frac{\pm\omega(X_1)^a\omega(X_k)^b(f(X_1) - \pi(f))}{N^4\bar{T}_N^4 \cdot (\sum_{j\neq k}\omega(X_j))}\right]\right|$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\frac{\omega(X_1)^a\omega(X_k)^b|f(X_1)-\pi(f)|}{N^4\bar{T}_N^4\cdot(N-2)\widetilde{T}_{N,k}}\right]\\
&\leq \mathbb{E}\left[\frac{\omega(X_1)^a\omega(X_k)^b|f(X_1)-\pi(f)|}{N^4\left(\frac{N-2}{N}\widetilde{T}_{N,k}\right)^4\cdot(N-2)\widetilde{T}_{N,k}}\right]\\
&= \frac{1}{(N-2)^5}\mathbb{E}\left[\omega(X_1)^a|f(X_1)-\pi(f)|\right]\mathbb{E}\left[\omega(X_k)^b\right]\mathbb{E}\left[\widetilde{T}_{N,k}^{-5}\right].
\end{aligned}
$$

where $(a,b)\in\{(4,1),(3,2),(2,3),(1,4)\}$ and the last equality uses the independence of $\widetilde{T}_{N,k}$ from both X_1 and X_k. Since $\mathbb{E}[\widetilde{T}_{N,k}^{-5}]<\infty$ for N large by Proposition A.1 (applied to the $N-2$ weights defining $\widetilde{T}_{N,k}$), each such expectation is $\mathcal{O}(N^{-5})$ provided $q(\omega^4)<\infty$ and $q(\omega^4|f-\pi(f)|)<\infty$. Hence the expectation of (D.33) is $\mathcal{O}(N^{-4})$, being a sum of $N-1$ terms of order N^{-5}.

7. For the last term in (D.29), recall that

$$
\begin{aligned}
&\frac{\omega(X_1)(f(X_1)-\pi(f))}{N\bar{T}_N(1+v_N(\mathbf{X}^N))}v_N(\mathbf{X}^N)^2\\
&=\frac{\omega(X_1)(f(X_1)-\pi(f))}{N\bar{T}_N(1+v_N(\mathbf{X}^N))}\\
&\quad\times\left(\frac{\bar{T}_N^{(2)}}{N\bar{T}_N^2}+\frac{\omega(X_1)^2}{N^2\bar{T}_N^2}-\frac{2\omega(X_1)\bar{T}_N^{(2)}}{N^2\bar{T}_N^3}+\frac{\bar{T}_N^{(3)}}{N^2\bar{T}_N^3}+\frac{\tilde{R}_N^u}{N\bar{T}_N}\right)^2,
\end{aligned}
$$

we will show that this expression is $o(N^{-3})$, under the assumptions of Theorem D.2.

Using the inequality $\left(\sum_{i=1}^5 a_i\right)^2\leq 5\left(\sum_{i=1}^5 a_i^2\right)$, we bound

$$
\begin{aligned}
v_N(\mathbf{X}^N)^2\leq 5\Bigg[&\left(\frac{\bar{T}_N^{(2)}}{N\bar{T}_N^2}\right)^2+\left(\frac{\omega(X_1)^2}{N^2\bar{T}_N^2}\right)^2+\left(\frac{2\omega(X_1)\bar{T}_N^{(2)}}{N^2\bar{T}_N^3}\right)^2\\
&+\left(\frac{\bar{T}_N^{(3)}}{N^2\bar{T}_N^3}\right)^2+\left(\frac{\tilde{R}_N^u}{N\bar{T}_N}\right)^2\Bigg].
\end{aligned}
\tag{D.35}
$$

Note that by (D.27), we have

$$
\frac{1}{1+v_N(\mathbf{X}^N)}=\frac{1}{\sum_{k=1}^N\frac{\omega(X_k)}{S_N(\mathbf{X}^N)-\omega(X_k)}}\leq\frac{1}{\sum_{k=1}^N\frac{\omega(X_k)}{S_N(\mathbf{X}^N)}}=1,
\tag{D.36}
$$

since $S_N(\mathbf{X}^N)-\omega(X_k)\leq S_N(\mathbf{X}^N)$ for all $k\in[N]$.

Thus, since $(1+v_N(\mathbf{X}^N))^{-1}\leq 1$, it suffices to show that each squared component

contributes $o(N^{-3})$ to the expectation

$$\mathbb{E}\left[\frac{\omega(X_1)|f(X_1)-\pi(f)|}{N\bar{T}_N}\cdot(\text{squared component})\right]. \tag{D.37}$$

The analysis requires different techniques depending on the order of each term. For the leading term in N corresponding to the first squared component, the random variable inside the expectation converges almost surely to a limit with zero expectation. A uniform integrability argument is then used to show that the expectation of the sequence also converges to zero. The remaining terms obtained by expanding the square are of order N^{-4} or smaller and have finite expectations that decay faster than N^{-3}, so they do not contribute to the asymptotic limit and can be handled more directly. We now analyze each term separately.

First squared term. The leading term in N corresponds to the first component in the square expansion

$$N^{-3}\mathbb{E}\left[\frac{\omega(X_1)(f(X_1)-\pi(f))\{\bar{T}_N^{(2)}\}^2}{\bar{T}_N^5(1+v_N(\mathbf{X}^N))}\right]. \tag{D.38}$$

Let A_N be the random variable inside the expectation. As $N\to\infty$, the numerator converges a.s. to $q(\omega^2)^2\omega(X_1)(f(X_1)-\pi(f))$, and the denominator converges to 1, so $A_N\to q(\omega^2)^2\omega(X_1)(f(X_1)-\pi(f))$ a.s. Thus, if A_N is uniformly integrable, we can swap expectation and limit, and obtain that $\mathbb{E}[A_N]$ goes to zero.

To show that A_N is uniformly integrable, we use (D.27) to write $(1+v_N(\mathbf{X}^N))^{-1}\leq 1$. Thus, upon taking absolute values, and by independence between X_1 and the other particles, we have, for any $\epsilon>0$,

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{\omega(X_1)(f(X_1)-\pi(f))\{\bar{T}_N^{(2)}\}^2}{\bar{T}_N^5(1+v_N(\mathbf{X}^N))}\right|^{1+\epsilon}\right] \\
&\leq \mathbb{E}\left[\left|\frac{\omega(X_1)(f(X_1)-\pi(f))\{\bar{T}_N^{(2)}\}^2}{\bar{T}_N^5}\right|^{1+\epsilon}\right] \\
&\leq \mathbb{E}\left[\frac{\{\bar{T}_N^{(2)}\}^{2(1+\epsilon)}}{\bar{T}_N^{5(1+\epsilon)}}\right]\cdot\mathbb{E}\left[|\omega(X_1)(f(X_1)-\pi(f))|^{1+\epsilon}\right].
\end{aligned} \tag{D.39}$$

The supremum over N is finite under the assumption that $q(\omega^{4+\epsilon'})<\infty$ for some $\epsilon'>0$, and thus $\{A_N\}$ is uniformly integrable by Billingsley (1999, Eq. (3.18)). This establishes that the first squared term contributes $o(N^{-3})$.

Second squared term. For the second squared component, we have

$$\mathbb{E}\left[\frac{\omega(X_1)|f(X_1)-\pi(f)|}{N\bar{T}_N}\cdot\frac{\omega^4(X_1)}{N^4\bar{T}_N^4}\right]=\mathbb{E}\left[\frac{\omega^5(X_1)|f(X_1)-\pi(f)|}{N^5\bar{T}_N^5}\right] \\ =\frac{q\left(\omega^5|f-\pi(f)|\right)}{N^5}\cdot\mathbb{E}\left[\frac{1}{\bar{T}_N^5}\right]. \tag{D.40}$$

This term is finite and of order $N^{-5}=o(N^{-3})$, provided that $q\left(\omega^5|f-\pi(f)|\right)<\infty$ and $\mathbb{E}[\bar{T}_N^{-5}]<\infty$. The latter condition holds for sufficiently large N under the moment assumptions of Proposition A.1.

Third squared term. For the third squared component, we have

$$\mathbb{E}\left[\frac{\omega(X_1)|f(X_1)-\pi(f)|}{N\bar{T}_N}\cdot\frac{4\omega^2(X_1)(\bar{T}_N^{(2)})^2}{N^4\bar{T}_N^6}\right] \\ =\frac{4}{N^5}q\left(\omega^3|f-\pi(f)|\right)\mathbb{E}\left[\frac{(\bar{T}_N^{(2)})^2}{\bar{T}_N^7}\right], \tag{D.41}$$

where the equality uses independence between X_1 and the remaining particles. This term is of order $N^{-5}=o(N^{-3})$, provided that $q\left(\omega^3|f-\pi(f)|\right)<\infty$ and $\mathbb{E}[(\bar{T}_N^{(2)})^2/\bar{T}_N^7]<\infty$. The latter condition holds for sufficiently large N under the moment assumption $q(\omega^{4+\epsilon})<\infty$ for some $\epsilon>0$, which can be verified using Hölder's inequality and Proposition A.1.

Fourth squared term. For the fourth squared component, we have

$$\mathbb{E}\left[\frac{\omega(X_1)|f(X_1)-\pi(f)|}{N\bar{T}_N}\cdot\frac{(\bar{T}_N^{(3)})^2}{N^4\bar{T}_N^6}\right]=\frac{1}{N^5}q\left(\omega|f-\pi(f)|\right)\cdot\mathbb{E}\left[\frac{(\bar{T}_N^{(3)})^2}{\bar{T}_N^7}\right], \tag{D.42}$$

where the equality uses independence between X_1 and the remaining particles. This term is of order $N^{-5}=o(N^{-3})$, provided that $q\left(\omega|f-\pi(f)|\right)<\infty$ and $\mathbb{E}[(\bar{T}_N^{(3)})^2/\bar{T}_N^7]<\infty$. The latter condition holds for sufficiently large N under the moment assumption $q(\omega^{6+\epsilon})<\infty$ for some $\epsilon>0$, which can be verified using Hölder's inequality and Proposition A.1.

Fifth squared term. For the final squared component, we proceed with more care. Recall that

$$\tilde{R}_N^u=\sum_{k=2}^{N}\frac{-\omega(X_k)(\omega(X_1)-\omega(X_k))^3}{N^3\bar{T}_N^3(1+u_k)}.$$

Define

$$b_k=\frac{\omega(X_k)(\omega(X_1)-\omega(X_k))^3}{N^4\bar{T}_N^4(1+u_k)},$$

so that $\tilde{R}_N^u / N\bar{T}_N = -\sum_{k=2}^N b_k$. Using the inequality $\left(\sum_{k=2}^N b_k\right)^2 \leq (N-1)\sum_{k=2}^N b_k^2$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\omega(X_1)|f(X_1)-\pi(f)|}{N\bar{T}_N}\left(\frac{\tilde{R}_N^u}{N\bar{T}_N}\right)^2\right] \\
&\leq (N-1)\sum_{k=2}^N \mathbb{E}\left[\frac{\omega(X_1)|f(X_1)-\pi(f)|b_k^2}{N\bar{T}_N}\right] \\
&= (N-1)\sum_{k=2}^N \mathbb{E}\left[\frac{\omega(X_1)|f(X_1)-\pi(f)|\omega^2(X_k)(\omega(X_1)-\omega(X_k))^6}{N^9\bar{T}_N^9(1+u_k)^2}\right].
\end{aligned} \tag{D.43}$$

Using the key identity

$$1+u_k = \frac{\sum_{j\neq k}\omega(X_j)}{N\bar{T}_N},$$

we can rewrite the denominator as

$$N^9\bar{T}_N^9(1+u_k)^2 = (N\bar{T}_N)^7\cdot\left(\sum_{j\neq k}\omega(X_j)\right)^2. \tag{D.44}$$

Recalling our leave-two-out technique with $\widetilde{T}_{N,k} = (N-2)^{-1}\sum_{j\notin\{1,k\}}\omega(X_j)$, we have the inequalities

$$\sum_{j\neq k}\omega(X_j) \geq (N-2)\widetilde{T}_{N,k}, \quad \text{and} \quad \bar{T}_N \geq \frac{N-2}{N}\widetilde{T}_{N,k}. \tag{D.45}$$

This gives us the lower bound for each term in the sum

$$(N\bar{T}_N)^7\cdot\left(\sum_{j\neq k}\omega(X_j)\right)^2 \geq (N-2)^9\widetilde{T}_{N,k}^9.$$

Using the inequality $|a-b|^6 \leq 32(|a|^6+|b|^6)$ and independence of $\widetilde{T}_{N,k}$ from both X_1 and X_k, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\frac{\omega(X_1)|f(X_1)-\pi(f)|}{N\bar{T}_N}\left(\frac{\tilde{R}_N^u}{N\bar{T}_N}\right)^2\right] \\
&\leq (N-1)\sum_{k=2}^N \frac{32\cdot\mathbb{E}[\widetilde{T}_{N,k}^{-9}]}{(N-2)^9}\left(q\left(\omega^7|f-\pi(f)|\right)q(\omega^2)+q\left(\omega|f-\pi(f)|\right)q(\omega^8)\right) \\
&= \frac{32(N-1)^2\cdot\max_{2\leq k\leq N}\mathbb{E}[\widetilde{T}_{N,k}^{-9}]}{(N-2)^9}
\end{aligned} \tag{D.46}$$

$$\times \left(q\left(\omega^7|f-\pi(f)|\right)q(\omega^2)+q\left(\omega|f-\pi(f)|\right)q(\omega^8)\right).$$

This term is of order $\mathcal{O}(N^{-7}) = o(N^{-3})$, provided that $q(\omega^8) < \infty$, $q(\omega^7|f - \pi(f)|) < \infty$, $q(\omega|f - \pi(f)|) < \infty$, and $\max_{2\le k\le N} \mathbb{E}[\widetilde{T}_{N,k}^{-9}] < \infty$. The latter condition holds for sufficiently large N under the assumption $q(\omega^{-\eta}) < \infty$ for some $\eta > 0$, which can be verified using Proposition A.1.

Conclusion. Under the moment conditions stated in Theorem D.2, all terms in $v_N(\mathbf{X}^N)^2$ contribute $o(N^{-3})$, confirming that the last term in (D.29) does not affect the asymptotic bias.

□

D.5 Robust Mean Estimation: Median-of-Means and Sub-Gaussian Rates

This appendix records the median-of-means (MoM) sub-Gaussian deviation bound invoked in Section 1.7. The construction is due to Nemirovsky & Yudin (1983); the modern finite-sample form is Devroye et al. (2016, Thm. 2), surveyed in Lugosi & Mendelson (2019). The proof we give is the standard Chebyshev–Hoeffding argument; we adopt the notation of Section 1.7.

Setting. Let $X_1, \dots, X_n$ be independent and identically distributed under a law P^* with mean $\mu := \mathbb{E}_{P^*}[X_1]$ and finite variance $\sigma^2 := \mathbb{V}_{P^*}(X_1) < \infty$. Fix $K \in \{1, \dots, n\}$, set $m := \lfloor n/K \rfloor$ and partition $\{1, \dots, n\}$ into K disjoint blocks $B_1, \dots, B_K$ with $|B_k| = m$ for each k. Define the block means

$$\bar{X}^{(k)} := \frac{1}{m}\sum_{i\in B_k} X_i, \qquad k = 1, \dots, K, \tag{D.47}$$

and the median-of-means estimator

$$\widehat{\mu}_K := \operatorname{median}\left(\bar{X}^{(1)}, \dots, \bar{X}^{(K)}\right). \tag{D.48}$$

D.5.1 Block-mean moments under the i.i.d. assumption

Lemma D.6 (Block-mean unbiasedness and variance). *For every $k \in \{1, \dots, K\}$, $\mathbb{E}_{P^*}[\bar{X}^{(k)}] = \mu$ and $\mathbb{V}_{P^*}(\bar{X}^{(k)}) = \sigma^2/m$.*

Proof. By the i.i.d. assumption, linearity of expectation gives $\mathbb{E}_{P^*}[\bar{X}^{(k)}] = m^{-1}\sum_{i\in B_k} \mathbb{E}_{P^*}[X_i] = \mu$. Independence of the summands within each block yields $\mathbb{V}_{P^*}(\bar{X}^{(k)}) = m^{-2}\sum_{i\in B_k} \mathbb{V}_{P^*}(X_i) = \sigma^2/m$. □

D.5.2 Sub-Gaussian deviation

Theorem D.7 (Sub-Gaussian deviation of MoM). *For every $\delta \in (0, 1/2]$, choose $K = \lceil 8 \log(1/\delta) \rceil$ and assume $K \leq n/2$. Then*

$$\mathbb{P}_{P^*}\left(\left| \widehat{\mu}_K - \mu \right| > \sigma \sqrt{\frac{32 \log(1/\delta)}{n}} \right) \leq \delta. \tag{D.49}$$

Proof. Set $r := 2\sigma/\sqrt{m}$. Chebyshev's inequality applied to each block mean using Lemma D.6 gives

$$\mathbb{P}_{P^*}\left(\bar{X}^{(k)} - \mu > r \right) \leq \frac{\mathbb{V}_{P^*}(\bar{X}^{(k)})}{r^2} = \frac{1}{4}. \tag{D.50}$$

Define $Z_k := \mathbb{1}\left(\bar{X}^{(k)} - \mu > r \right)$ for $k = 1, \ldots, K$. The (Z_k) are independent Bernoullis, since the blocks are disjoint and the sample is i.i.d., with $p := \mathbb{E}_{P^*}[Z_k] \leq 1/4$.

The event $\{\widehat{\mu}_K - \mu > r\}$ implies that more than half of the block means exceed $\mu + r$, hence $\sum_{k=1}^{K} Z_k \geq K/2$. Hoeffding's inequality (Boucheron et al., 2013, Thm. 2.8) applied to the centred sum $\sum_{k=1}^{K}(Z_k - p)$ yields, for every $t > 0$,

$$\mathbb{P}_{P^*}\left(\sum_{k=1}^{K} Z_k \geq Kp + Kt \right) \leq \exp\left(-2Kt^2\right). \tag{D.51}$$

Taking $t = 1/2 - p \geq 1/4$ gives $\mathbb{P}_{P^*}\left(\sum_{k=1}^{K} Z_k \geq K/2 \right) \leq \exp(-K/8)$. The same bound applied to the lower-tail event $\{\bar{X}^{(k)} - \mu < -r\}$ controls the lower deviation, and a union bound delivers

$$\mathbb{P}_{P^*}\left(\left| \widehat{\mu}_K - \mu \right| > r \right) \leq 2 \exp(-K/8). \tag{D.52}$$

For the choice $K = \lceil 8 \log(1/\delta) \rceil$ with $\delta \leq 1/2$, the right-hand side is at most δ. The block size is $m \geq n/K - 1 \geq n/(2K)$ for $K \leq n/2$, so $r = 2\sigma/\sqrt{m} \leq 2\sigma\sqrt{2K/n} \leq 4\sigma\sqrt{2\log(1/\delta)/n}$, which gives (D.49). □

D.6 Convergence Rates for Nested Expectations

This appendix reproduces, in the notation of Chapter 2, the budget-allocation argument of Andradóttir & Glynn (2016) invoked in the Composition paragraph of Section 1.3. The conclusion is a strict-improvement statement: an inner estimator with $\mathcal{O}(N^{-\gamma})$ bias and $\mathcal{O}(N^{-1})$ variance attains mean-squared-error rate $\mathcal{O}\left(C^{-2\gamma/(2\gamma+1)}\right)$ at total budget C, strictly slower than the canonical Monte Carlo rate $\mathcal{O}(C^{-1})$ for every finite γ; the canonical rate is recovered only by an exactly unbiased inner estimator ($\gamma = \infty$).

Setting. Let Θ_1 be an outer variable sampled from its marginal distribution, and let $f(\Theta_1, \Theta_2)$ be a test function whose conditional expectation given Θ_1 exists almost surely. The nested expectation of interest is

$$E := \mathbb{E}_1\big[g(\Theta_1)\big], \qquad g(\theta_1) := \mathbb{E}_{2|1}[f(\Theta_1, \Theta_2) \mid \Theta_1 = \theta_1]. \tag{D.53}$$

Such quantities arise in modular Bayesian inference (Liu et al., 2009; Plummer, 2014), multiple imputation (Zhou & Reiter, 2010), expected information gain in experimental design (Rainforth et al., 2018a) and portfolio risk measurement (Gordy & Juneja, 2010).

Two-stage estimator. Fix a bias exponent $\gamma \in (0, \infty]$ and assume that for every θ_1 an estimator $\hat{F}(\theta_1)$ of $g(\theta_1)$ is available at cost proportional to N, with the bias–variance decomposition

$$\hat{F}(\theta_1) = g(\theta_1) + \frac{b(\theta_1)}{N^\gamma} + \frac{\sigma(\theta_1)}{\sqrt{N}}\, Z(\theta_1), \qquad Z(\theta_1) \sim \mathcal{N}(0,1), \tag{D.54}$$

where b, σ are square-integrable under the marginal of Θ_1. SNIS corresponds to $\gamma = 1$; an exactly unbiased estimator to $\gamma = \infty$ ($b \equiv 0$). Drawing $\Theta_{1,1}, \ldots, \Theta_{1,M}$ i.i.d. from the marginal of Θ_1, independently of the inner randomness, the outer plug-in is

$$\hat{F}_{M,N} := \frac{1}{M} \sum_{m=1}^{M} \hat{F}(\Theta_{1,m}), \qquad C := MN. \tag{D.55}$$

Error decomposition. Write $B := \mathbb{E}_1[b(\Theta_1)]$, $V := \mathbb{E}_1[\sigma^2(\Theta_1)]$, and $v := \mathbb{V}_1\big(g(\Theta_1)\big) +$ l.o.t.. Substituting (D.54) into (D.55) and applying the central limit theorem in the outer variable yields

$$\hat{F}_{M,N} - E = \underbrace{\frac{B}{N^\gamma}}_{\text{inner bias}} + \underbrace{\frac{\sqrt{V}}{\sqrt{MN}}\, G}_{\text{inner variance}} + \underbrace{\frac{\sqrt{v}}{\sqrt{M}}\, \xi}_{\text{outer variance}} + \text{l.o.t.}, \qquad G, \xi \sim \mathcal{N}(0,1). \tag{D.56}$$

Proposition D.8 (Nested-expectation rates, Andradóttir & Glynn, 2016)**.** *Under* (D.54) *with* $C = MN$ *fixed, splitting the budget as* $M = C^\alpha$, $N = C^{1-\alpha}$ *with* $\alpha \in (0, 1]$*:*

(i) *For finite* $\gamma > 0$*, the optimal split is* $\alpha^\star = 2\gamma/(2\gamma + 1)$*, giving*

$$\mathbb{E}\big[(\hat{F}_{M,N} - E)^2\big] = \mathcal{O}\Big(C^{-2\gamma/(2\gamma+1)}\Big). \tag{D.57}$$

The SNIS case $\gamma = 1$ *gives* $\mathcal{O}(C^{-2/3})$*.*

(ii) *For* $\gamma = \infty$ *(unbiased inner), the bias term in* (D.56) *vanishes, the optimum is attained at* $\alpha = 1$*, and*

$$\mathbb{E}\big[(\hat{F}_{M,N} - E)^2\big] = \mathcal{O}(C^{-1}). \tag{D.58}$$

Proof. Writing $M = C^{\alpha}$, $N = C^{1-\alpha}$, the three terms in (D.56) have orders $C^{-\gamma(1-\alpha)}$, $C^{-1/2}$ and $C^{-\alpha/2}$ respectively. The dominant standard-deviation order is $C^{-\beta(\alpha)}$ with

$$\beta(\alpha) = \min\Big\{\gamma(1-\alpha),\ \tfrac{1}{2},\ \tfrac{\alpha}{2}\Big\}.$$

For finite γ, β is monotone increasing on $[0, \alpha^\star]$ and monotone decreasing on $[\alpha^\star, 1]$, with maximum where the inner bias and outer variance balance, $\gamma(1-\alpha^\star) = \alpha^\star/2$, giving $\alpha^\star = 2\gamma/(2\gamma+1)$ and $\beta^\star = \gamma/(2\gamma+1)$. Squaring yields (D.57). For $\gamma = \infty$ the inner-bias term is identically zero; the remaining terms $C^{-1/2}$ and $C^{-\alpha/2}$ are jointly maximised at $\alpha = 1$, where both equal $C^{-1/2}$, and squaring gives (D.58). □

Discussion. Proposition D.8 makes precise the point of Section 2.1.1: the inner bias is the binding constraint, since variance reduction moves only the constant in $C^{-2\gamma/(2\gamma+1)}$ whereas only exact debiasing ($\gamma = \infty$) lifts the rate to C^{-1}, at $\gamma = 1$ the ratio of mean-squared errors growing as $C^{1/3}$. The argument is independent of the inner algorithm, covering the coupled-IMH estimator of Chapter 2 and randomised multilevel constructions (McLeish, 2011; Rhee & Glynn, 2015); it assumes deterministic per-call cost, the random-cost extension being treated in Wang et al. (2024, §2).

D.7 Comparison with Hesterberg's Bias Formula

The leading-order asymptotic-bias formula (1.15) first appears in Hesterberg (1988), whose relationship to the moment-level statement of Chapter 2 we record here.

Hesterberg's derivation. Hesterberg (1988) obtains the formula by two complementary arguments. The first is a second-order Taylor expansion of $\hat{\pi}^N_{\mathrm{IS}}(h) = \bar{Y}/\bar{W}$ around its mean (Hesterberg, 1988, Appendix 3, pp. 207–212): writing

$$C := \frac{1}{N}\sum_{n=1}^{N} \omega(x_n)\, h(x_n) - \pi_n(h), \qquad D := \frac{1}{N}\sum_{n=1}^{N} \omega(x_n) - 1,$$

and expanding $\hat{\pi}^N_{\mathrm{IS}}(h) = (\pi_n(h)+C)/(1+D) \approx \pi_n(h)+C-\pi_n(h)\,D-CD+\pi_n(h)\,D^2+\cdots$, taking expectations and using $\mathbb{E}[C] = \mathbb{E}[D] = 0$ gives

$$\mathbb{E}\big[\hat{\pi}^N_{\mathrm{IS}}(h)\big] \approx \pi_n(h) - \frac{1}{N}\big(\mathbb{C}\mathrm{ov}_q(\omega h, \omega) - \pi_n(h)\,\mathbb{V}_q(\omega)\big) = \pi_n(h) - \frac{1}{N}\,q\big(\omega^2(h-\pi_n(h))\big), \tag{D.59}$$

matching (1.15). The derivation truncates the Taylor series and interchanges expansion with expectation without controlling the remainder.

The second argument is the Edgeworth expansion of Bhattacharya & Ghosh (1978), which

gives

$$\mathbb{P}\big(\sqrt{N}(\hat{\pi}^N_{\mathrm{IS}}(h) - \pi_n(h)) \le z\,\sigma\big) = \Phi(z) + \frac{1}{\sqrt{N}}\left(-\frac{b}{\sigma} + \frac{\kappa_3}{6\sigma^3}(z^2-1)\right)\phi(z) + o\big(N^{-1/2}\big), \tag{D.60}$$

with $b = -q(\omega^2(h - \pi_n(h)))$ identified as the mean-shift term of the distributional approximation.

Why the Edgeworth identification does not give the moment formula. Equation (D.60) is a pointwise (and even uniform) approximation of the cumulative distribution function of $\sqrt{N}\,(\hat{\pi}^N_{\mathrm{IS}}(h) - \pi_n(h))$; it characterises the distribution of the estimator, not its moments. Convergence in distribution does not imply convergence of expectations without *uniform integrability* (Billingsley, 1999, Theorem 3.5), and a uniform CDF error does not by itself imply uniform tail control. Hesterberg (1988, p. 40) acknowledges this explicitly, noting that the actual biases "may be infinite or undefined".

Comparison of conditions. The Edgeworth route of Bhattacharya & Ghosh (1978) requires $\pi_n \ll q$, finite third moments $q(\omega^3) < \infty$ and $q(\omega^3|h|^3) < \infty$, and Cramér's condition on the joint characteristic function of $(\omega, \omega^2, \omega h, \omega^2 h^2)$. The moment-level result proved in Chapter 2 requires only $\pi_n(|h|) < \infty$, $q(|\omega^2(h - \pi_n(h))|^{1+\varepsilon}) < \infty$ for some $\varepsilon > 0$, and an inverse-moment condition $q(\omega^{-\eta}) < \infty$ for some $\eta > 0$. For bounded h the second condition reduces to $q(\omega^{2+2\varepsilon}) < \infty$, strictly weaker than $q(\omega^3) < \infty$, and no smoothness or non-lattice (Cramér) condition is imposed. The independently obtained asymptotic-bias formula of Daudel & Roueff (2024, Proposition 6) requires separate moment conditions on ω and on $\omega(h - \pi_n(h))$ that are also strictly stronger than ours.

Nature of the proof. The argument of Chapter 2 establishes the moment convergence (1.15) directly. An exchangeability rewrite of $N \cdot \mathbb{E}[\hat{\pi}^N_{\mathrm{IS}}(h) - \pi_n(h)]$ as $-\mathbb{E}[A_N]$ for an auxiliary sequence A_N that converges almost surely to $\omega(x_1)^2(h(x_1) - \pi_n(h))$, combined with uniform integrability obtained from the $(1+\varepsilon)$-moment condition on the numerator and the inverse-moment condition on the denominator, allows the interchange of limit and expectation via Vitali's convergence theorem. The full proof and the numerical illustration are in Chapter 2; Daudel & Roueff (2024) obtain similar results.

Bibliography

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., & Stuart, A. M. (2017). Importance sampling: intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431.

Alexopoulos, A., Dellaportas, P., & Titsias, M. K. (2023). Variance reduction for Metropolis–Hastings samplers. *Statistics and Computing*, 33(1):6.

Alquier, P. (2008). PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304.

Alquier, P. (2024). User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303.

Alquier, P. & Guedj, B. (2018). Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902.

Alquier, P. & Lounici, K. (2011). Pac-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145.

Alquier, P. & Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475–1497.

Alquier, P., Ridgway, J., & Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(236):1–41.

Andradóttir, S. & Glynn, P. W. (2016). Computing Bayesian means using simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 26(2):1–26.

Andral, C. (2022). An attempt to trace the birth of importance sampling.

Andral, C., Douc, R., Marival, H., & Robert, C. P. (2024). The importance Markov chain. *Stochastic Processes and their Applications*, 171:104316.

Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Andrieu, C., Lee, A., Power, S., & Wang, A. Q. (2022). Comparison of Markov chains via weak Poincaré inequalities with application to pseudo-marginal MCMC. *The Annals of Statistics*, 50(6):3592–3618.

Andrieu, C., Lee, A., & Vihola, M. (2018). Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24(2):842–872.

Arbel, J., Pitas, K., Vladimirova, M., & Fortuin, V. (2024). A primer on Bayesian neural networks: review and debates. *Statistical Science*.

Atchadé, Y. F. & Jacob, P. E. (2026). Unbiased MCMC. In Craiu, R. V., Vats, D., Jones, G. L., Brooks, S., Gelman, A., & Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*, chapter 17. Chapman and Hall/CRC, 2nd edition.

Atchadé, Y. F. & Perron, F. (2005). Improving on the independent Metropolis–Hastings algorithm. *Statistica Sinica*, pages 3–18.

Banerjee, P. K. & Montufar, G. (2021). Information Complexity and Generalization Bounds. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1709–1714. IEEE.

Baraud, Y. (2024). From robust tests to Bayes-like posterior distributions. *Probability Theory and Related Fields*, 188(1):159–234.

Baraud, Y. & Birgé, L. (2018). Rho-estimators revisited: general theory and applications. *The Annals of Statistics*, 46(6B):3767–3804.

Baraud, Y. & Birgé, L. (2020). Robust Bayes-like estimation: Rho-Bayes estimation. *The Annals of Statistics*, 48(6):3699–3720.

Baraud, Y., Birgé, L., & Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Inventiones mathematicae*, 207(2):425–517.

Barron, A., Schervish, M. J., & Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561.

Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418.

Bégin, L., Germain, P., Laviolette, F., & Roy, J.-F. (2014). PAC-Bayesian theory for transductive learning. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 33 of *PMLR*, pages 105–113.

Bégin, L., Germain, P., Laviolette, F., & Roy, J.-F. (2016). PAC-Bayesian bounds based on the Rényi divergence. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51 of *PMLR*, pages 435–444.

Berild, M. O., Martino, S., Gómez-Rubio, V., & Rue, H. (2022). Importance sampling with the integrated nested Laplace approximation. *Journal of Computational and Graphical Statistics*, 31(4):1225–1237.

Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons.

Bhattacharya, A., Linero, A., & Oates, C. J. (2024). Grand challenges in Bayesian computation. *Bulletin of the International Society for Bayesian Analysis*, 31(3):1–6.

Bhattacharya, A., Pati, D., & Yang, Y. (2019). Bayesian fractional posteriors. *The Annals of Statistics*, 47(1):39–66.

Bhattacharya, R. N. & Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion. *The Annals of Statistics*, 6(2):434–451.

Bickel, P. J., Holm, S., Rosén, B., Spjøtvoll, E., Lauritzen, S., Johansen, S., & Barndorff-Nielsen, O. (1976). Another look at robustness: a review of reviews and some new developments [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 145–168.

Billingsley, P. (1999). *Convergence of probability measures*. John Wiley & Sons, second edition.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65(2):181–237.

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. In *Annales de l'IHP Probabilités et statistiques*, volume 42, pages 273–325.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bissiri, P. G., Holmes, C. C., & Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78(5):1103–1130.

Biswas, N., Jacob, P. E., & Vanetti, P. (2019). Estimating convergence of Markov chains with L-lag couplings. In *Advances in Neural Information Processing Systems*, pages 7389–7399.

Blanchet, J. H., Chen, N., & Glynn, P. W. (2015). Unbiased Monte Carlo computation of smooth functions of expectations via taylor expansions. In *2015 Winter Simulation Conference (WSC)*, pages 360–367. IEEE.

Blanchet, J. H. & Glynn, P. W. (2015). Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization. In *2015 Winter Simulation Conference (WSC)*, pages 3656–3667. IEEE.

Blanchet, J. H., Glynn, P. W., & Pei, Y. (2019). Unbiased multilevel Monte Carlo: stochastic optimization, steady-state simulation, quantiles, and other applications. *arXiv preprint arXiv:1904.09929*.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Burda, Y., Grosse, R., & Salakhutdinov, R. (2016). Importance weighted autoencoders. In *4th International Conference on Learning Representations (ICLR)*. arXiv:1509.00519.

Calo, S., Jonsson, A., Neu, G., Schwartz, L., & Segovia-Aguas, J. (2024). Bisimulation metrics are optimal transport distances, and can be computed efficiently. *Advances in Neural Information Processing Systems*, 37:134345–134386.

Cardoso, G., Samsonov, S., Thin, A., Moulines, E., & Olsson, J. (2022). BR-SNIS: bias reduced self-normalized importance sampling. *Advances in Neural Information Processing Systems*, 35:716–729.

Casella, G. & Robert, C. (1996). Rao-Blackwellisation of sampling schemes. 83(1):81–94.

Castillo, I. (2024). Bayesian nonparametric statistics, St-Flour lecture notes. *arXiv preprint arXiv:2402.16422*.

Castillo, I., Schmidt-Hieber, J., & van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018.

Catoni, O. (2004). *Statistical learning theory and stochastic optimization*. Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics. Springer.

Catoni, O. (2007). *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Institute of Mathematical Statistics Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Beachwood, OH.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185.

Catoni, O. & Giulini, I. (2017). Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*.

Chatterjee, S. & Diaconis, P. (2018). The sample size required in importance sampling. *Annals of Applied Probability*, 28(2):1099–1135.

Cherief-Abdellatif, B.-E. (2019). Consistency of ELBO maximization for model selection. In Ruiz, F., Zhang, C., Liang, D., & Bui, T., editors, *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*, volume 96 of *Proceedings of Machine Learning Research*, pages 11–31. PMLR.

Chérief-Abdellatif, B.-E. (2020). Convergence rates of variational inference in sparse deep learning. In III, H. D. & Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1831–1842. PMLR.

Chérief-Abdellatif, B.-E. & Alquier, P. (2020). MMD-Bayes: robust Bayesian estimation via maximum mean discrepancy. In *Proceedings of the 2nd Symposium on Advances in Approximate Bayesian Inference (AABI)*, volume 118 of *Proceedings of Machine Learning Research*, pages 1–21.

Chernozhukov, V. & Hong, H. (2003). An mcmc approach to classical estimation. *Journal of econometrics*, 115(2):293–346.

Chopin, N. (2002). A sequential particle filter method for static models. 89:539–552.

Chopin, N., Crucinio, F. R., & Singh, S. S. (2025). Towards a turnkey approach for unbiased Monte Carlo estimation of smooth functions of expectations. *Biometrika*, 112(3):asaf030.

Chopin, N. & Papaspiliopoulos, O. (2020). *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Springer.

Cover, T. M. & Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience, 2nd edition.

Csiszár, I. (1967). On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158.

Dalalyan, A. S. & Tsybakov, A. B. (2012). Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443.

Dau, H. D. (2022). *Sequential Bayesian Computation*. PhD thesis, Institut Polytechnique de Paris.

Daudel, K., Benton, J., Shi, Y., & Doucet, A. (2023). Alpha-divergence variational inference meets importance weighted auto-encoders: methodology and asymptotics. *Journal of Machine Learning Research*, 24(243):1–83.

Daudel, K. & Roueff, F. (2024). Learning with importance weighted variational inference: Asymptotics for gradient estimators of the vr-iwae bound. *arXiv preprint arXiv:2410.12035*.

De Cock, D. (2011). Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3).

Del Moral, P. (2013). Mean field simulation for Monte Carlo integration. *Monographs on Statistics and Applied Probability*, 126:26.

Del Moral, P., Doucet, A., & Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3):411–436.

Del Moral, P., Doucet, A., & Peters, G. W. (2007). Sharp propagation of chaos estimates for feynman–kac particle models. *Theory of Probability & Its Applications*, 51(3):459–485.

Deligiannidis, G., Jacob, P. E., Khribch, E. M., & Wang, G. (2024). On importance sampling and independent Metropolis–Hastings with an unbounded weight function. *arXiv preprint arXiv:2411.09514*.

Deligiannidis, G. & Lee, A. (2018). Which ergodic averages have finite asymptotic variance? *The Annals of Applied Probability*, 28(4):2309–2334.

Dellaporta, C., Knoblauch, J., Damoulas, T., & Briol, F.-X. (2022). Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics*, pages 943–970. PMLR.

Demidovich, Y., Malinovsky, G., Sokolov, I., & Richtárik, P. (2023). A guide through the zoo of biased SGD. In *Advances in Neural Information Processing Systems*, volume 36.

Devroye, L., Lerasle, M., Lugosi, G., & Oliveira, R. I. (2016). Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., & Blei, D. M. (2017). Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*.

Doeblin, W. (1938). Exposé de la théorie des chaînes simples constantes de Markov à un nombre fini d'états. *Revue Mathématique de l'Union Interbalkanique*, 2:77–105.

Domke, J. & Sheldon, D. R. (2018). Importance weighting and variational inference. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 4470–4479.

Donsker, M. D. & Varadhan, S. R. S. (1976). Asymptotic evaluation of certain Markov process expectations for large time—III. *Communications on Pure and Applied Mathematics*, 29(4):389–461.

Donsker, M. D. & Varadhan, S. S. (1975). Asymptotic evaluation of certain Markov process expectations for large time, i. *Communications on pure and applied mathematics*, 28(1):1–47.

Douc, R., Jacob, P. E., Lee, A., & Vats, D. (2026). Solving the Poisson equation using coupled Markov chains. *The Annals of Statistics*, 54(1):201–225.

Douc, R., Moulines, E., Priouret, P., & Soulier, P. (2018). *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer.

Douc, R., Moulines, E., & Soulier, P. (2007). Computable convergence rates for sub-geometric ergodic Markov chains.

Doucet, A. & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: fifteen years later. *Handbook of Nonlinear Filtering*, 12:656–704.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis*. CRC Press, 3rd edition.

Germain, P., Bach, F., Lacoste, A., & Lacoste-Julien, S. (2016). PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pages 1884–1892.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–483.

Ghosal, S., Ghosh, J. K., & van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531.

Ghosal, S. & van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

Ghosal, S. & van der Vaart, A. W. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *The Annals of Statistics*, 35(1):192–223.

Ghosh, A. & Basu, A. (2016). Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437.

Giles, M. B. (2008). Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617.

Giles, M. B. (2015). Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328.

Giummolè, F., Mameli, V., Ruli, E., & Ventura, L. (2019). Objective Bayesian inference with proper scoring rules. *TEST*, 28(3):728–755.

Glynn, P. W. (2016). Exact simulation vs exact estimation. In *2016 Winter Simulation Conference (WSC)*, pages 193–205. IEEE.

Glynn, P. W. & Rhee, C.-H. (2014). Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51A:377–389.

Glynn, P. W. & Whitt, W. (1992). The asymptotic efficiency of simulation estimators. *Operations Research*, 40(3):505–520.

Gómez-Rubio, V. & Rue, H. (2018). Markov chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing*, 28(5):1033–1051.

Gordy, M. B. & Juneja, S. (2010). Nested simulation in portfolio risk measurement. *Management Science*, 56(10):1833–1848.

Griffeath, D. (1975). A maximal coupling for Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 31:95–106.

Grünwald, P. (2011). Safe learning: bridging the gap between Bayes, mdl and statistical learning theory via empirical convexity. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 397–420. JMLR Workshop and Conference Proceedings.

Grünwald, P. & van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.

Grünwald, P. D. & Mehta, N. A. (2018). Fast rates for general unbounded loss functions: From ERM to generalized Bayes. *Journal of Machine Learning Research*, 19(1):2317–2380.

Hammersley, J. M. & Handscomb, D. C. (1964). *Monte Carlo Methods*. Methuen, London.

Hammouda, I., Ndaoud, M., & Seghouane, A.-K. (2024). Outlier-bias removal with alpha divergence: A robust non-convex estimator for linear regression. *arXiv preprint arXiv:2412.19183*.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (2009). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Heide, R., Kirichenko, A., Grunwald, P., & Mehta, N. (2020). Safe-Bayesian generalized linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 2623–2633. PMLR.

Hellström, F., Durisi, G., Guedj, B., & Raginsky, M. (2025). Generalization bounds: perspectives from information theory and PAC-Bayes. *Foundations and Trends in Machine Learning*, 18(1):1–223. arXiv:2309.04381.

Henmi, M., Yoshida, R., & Eguchi, S. (2007). Importance sampling via the estimated sampler. *Biometrika*, 94(4):985–991.

Hesterberg, T. C. (1988). *Advances in Importance Sampling*. PhD dissertation, Stanford University. Department of Statistics, Technical Report No. 300.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347.

Holmes, C. C. & Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503.

Hooker, G. & Vidyashankar, A. N. (2014). Bayesian model robustness via disparities. *TEST*, 23(3):556–584.

Hu, Y., Chen, X., & He, N. (2021). On the bias-variance-cost tradeoff of stochastic optimization. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., & Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22119–22131. Curran Associates, Inc.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons.

Huber, P. J. (1992). Robust estimation of a location parameter. In Kotz, S. & Johnson, N. L., editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 492–518. Springer.

Huggins, J. H. & Miller, J. W. (2019). Robust inference and model criticism using bagged posteriors. *arXiv preprint arXiv:1912.07104*.

Ibragimov, I. A. & Has' Minskii, R. Z. (2013). *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media.

Jacob, P. E., O'Leary, J., & Atchadé, Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B*, 82(3):543–600.

Jacob, P. E., Robert, C. P., & Smith, M. H. (2011). Using parallel computation to improve independent Metropolis–Hastings based estimation. *Journal of Computational and Graphical Statistics*, 20(3):616–635.

Jarner, S. F. & Roberts, G. O. (2002). Polynomial convergence rates of Markov chains. *The Annals of Applied Probability*, 12(1):224–247.

Jarner, S. F. & Roberts, G. O. (2007). Convergence of heavy-tailed Monte Carlo Markov chain algorithms. *Scandinavian Journal of Statistics*, 34(4):781–815.

Jewson, J., Smith, J. Q., & Holmes, C. (2018). Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442.

Jiang, W. & Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, 36(5):2207–2231.

Johnson, V. E. (1996). Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. *Journal of the American Statistical Association*, 91(433):154–166.

Johnson, V. E. (1998). A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *Journal of the American Statistical Association*, 93(441):238–248.

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320.

Jones, G. L. & Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, 16(4):312–334.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Juditsky, A., Nemirovski, A., & Tauvel, C. (2011). Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58.

Kahn, H. (1949). Stochastic (Monte Carlo) attenuation analysis. Technical Report P-88, RAND Corporation.

Kasprzak, M. J., Giordano, R., & Broderick, T. (2025). How good is your Laplace approximation of the Bayesian posterior? finite-sample computable error bounds for a variety of useful divergences. *Journal of Machine Learning Research*, 26(87):1–81.

Katsevich, A. & Rigollet, P. (2024). On the approximation accuracy of Gaussian variational inference. *The Annals of Statistics*, 52(4):1384–1409.

Khribch, E. M. & Alquier, P. (2024). Convergence of statistical estimators via mutual information bounds. *arXiv preprint arXiv:2412.18539*.

Khribch, E. M. & Alquier, P. (2026). Robust Bayesian inference via variational approximations of generalized rho-posteriors. *arXiv preprint arXiv:2601.07325*.

Kingma, D. P. & Welling, M. (2014). Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Kleijn, B. J. K. & van der Vaart, A. W. (2012). The Bernstein–Von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Knoblauch, J., Jewson, J., & Damoulas, T. (2022). An optimization-centric view on Bayes' rule: reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109.

Koopman, S. J., Shephard, N., & Creal, D. (2009). Testing the assumptions behind importance sampling. *Journal of Econometrics*, 149(1):2–11.

Laplace, P. S. d. (1774). Mémoire sur la probabilité des causes par les événements. *Mémoires de Mathématique et de Physique, Académie Royale des Sciences*, pages 621–656.

Le Cam, L. (1975). On local and global properties in the theory of asymptotic normality of experiments. *Stochastic processes and related topics*, 1:13–54.

Le Cam, L. (2012). *Asymptotic Methods in Statistical Decision Theory*. Springer Science & Business Media.

LeCam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, 41(3):802–828.

LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53.

Lecué, G. & Lerasle, M. (2020). Robust machine learning by median-of-means: theory and practice. *The Annals of Statistics*, 48(2):906–931.

Lee, A., Singh, S. S., & Vihola, M. (2020). Coupled conditional backward sampling particle filter. *The Annals of Statistics*, 48(5):3066 – 3089.

Lee, J. C. & Valiant, P. (2022). Optimal sub-Gaussian mean estimation in $\mathbb{R}$. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 672–683. IEEE.

Lepskiĭ, O. V. (1991). Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates. *Theory of Probability and Its Applications*, 36(4):682–697.

Li, Y. & Turner, R. E. (2016). Rényi divergence variational inference. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*.

Liese, F. & Vajda, I. (1987). *Convex Statistical Distances*. Teubner-Texte zur Mathematik. Teubner.

Lin, T., Jin, C., & Jordan, M. (2020). On gradient descent ascent for nonconvex-concave minimax problems. In *International conference on machine learning*, pages 6083–6093. PMLR.

Lindvall, T. (2002). *Lectures on the Coupling Method*. Dover Publications.

Liu, F., Bayarri, M., & Berger, J. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150.

Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and computing*, 6:113–119.

Liu, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer.

Liu, J. S., Liang, F., & Wong, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95(449):121–134.

Lugosi, G. & Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190.

Lugosi, G. & Neu, G. (2022). Generalization bounds via convex analysis. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, PMLR, pages 3524–3546.

Margossian, C. C. & Blei, D. M. (2024). Amortized variational inference: when and why? In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 244 of *PMLR*, pages 2434–2449. arXiv:2307.11018.

Martin, G. M., Frazier, D. T., & Robert, C. P. (2024). Computing Bayes: From then 'til now. *Statistical Science*, 39(1):3–19.

Matsubara, T., Knoblauch, J., Briol, F.-X., & Oates, C. J. (2022). Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):997–1022.

McAllester, D. A. (1998). Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT)*, pages 230–234.

McAllester, D. A. (1999). PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT)*, pages 164–170. ACM.

McCartan, C. & Imai, K. (2023). Sequential Monte Carlo for sampling balanced and compact redistricting plans. *Annals of Applied Statistics*, 17(4):3236–3267.

McLeish, D. (2011). A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods and Applications*, 17(4):301–315.

Melbourne, J., Talukdar, S., Bhaban, S., Madiman, M., & Salapaka, M. V. (2022). The differential entropy of mixtures: New bounds and applications. *IEEE Transactions on Information Theory*, 68(4):2123–2146.

Mengersen, K. L. & Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101–121.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Meyn, S. P. & Tweedie, R. L. (2012). *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition.

Middleton, L., Deligiannidis, G., Doucet, A., & Jacob, P. E. (2019). Unbiased smoothing using particle independent Metropolis–Hastings. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *PMLR*, pages 2378–2387.

Mijatović, A. & Vogrinc, J. (2018). On the Poisson equation for Metropolis–Hastings chains. *Bernoulli*, 24(3):2401–2428.

Miller, J. W. & Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*.

Minka, T. (2005). Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research.

Minsker, S. & Ndaoud, M. (2021). Robust and efficient mean estimation: an approach based on the properties of self-normalized sums. *Electronic Journal of Statistics*, 15(2):6036–6070.

Minsker, S. & Yao, S. (2025). Generalized median of means principle for Bayesian inference. *Machine Learning*, 114(4).

Monahan, J. F. (1993). Testing the behavior of importance sampling weights. In *Computing Science and Statistics: Proceedings of the 25th Symposium on the Interface*, pages 112–117.

Moulos, V. (2021). Bicausal optimal transport for Markov chains via dynamic programming. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1688–1693. IEEE.

Nakagawa, T. & Hashimoto, S. (2020). Robust Bayesian inference via γ-divergence. *Communications in Statistics — Theory and Methods*, 49(2):343–360.

Nash, W. J., Sellers, T. L., Talbot, S. R., Cawthorn, A. J., & Ford, W. B. (1994). The population biology of abalone (haliotis species) in tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait. *Sea Fisheries Division, Technical Report*, 48:p411.

Neal, R. M. & Pinto, R. L. (2001). Improving Markov chain Monte Carlo estimators by coupling to an approximating chain.

Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A., & Roy, D. M. (2019). Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *Advances in Neural Information Processing Systems*, volume 32.

Nemirovsky, A. S. & Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, New York.

Neu, G., Dziugaite, G. K., Haghifam, M., & Roy, D. M. (2021). Information-Theoretic Generalization Bounds for Stochastic Gradient Descent. In Belkin, M. & Kpotufe, S., editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3526–3545. PMLR.

Nielsen, F. & Nock, R. (2011). On Rényi and Tsallis Entropies and Divergences for Exponential Families. *arXiv preprint arXiv:1105.3259*.

Nowozin, S. (2018). Debiasing evidence approximations: On importance-weighted autoencoders and jackknife variational inference. In *International conference on learning representations*.

Ohn, I. & Lin, L. (2024). Adaptive variational Bayes: Optimality, computation and applications. *The Annals of Statistics*, 52(1):335–363.

Orenstein, P. (2022). Robust importance sampling with adaptive winsorization. *Bernoulli*, 28(4):2862–2873.

Orsak, G. C. & Aazhang, B. (1991). Constrained solutions in importance sampling via robust statistics. *IEEE Transactions on Information Theory*, 37(2):307–316.

Owen, A. B. (2013). Monte Carlo theory, methods and examples.

Owen, A. B. (2025). Zero variance self-normalized importance sampling via estimating equations. *arXiv preprint arXiv:2510.00389*.

Owhadi, H., Scovel, C., & Sullivan, T. (2015). Brittleness of Bayesian inference under finite information in a continuous world. *Electronic Journal of Statistics*, 9(1):1–79.

Pacchiardi, L. & Dutta, R. (2021). Generalized Bayesian likelihood-free inference using scoring rules estimators. *arXiv preprint arXiv:2104.03889*, 2(8).

Paisley, J., Blei, D. M., & Jordan, M. I. (2012). Variational Bayesian inference with stochastic search. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1363–1370.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64.

Pati, D., Bhattacharya, A., & Yang, Y. (2018). On statistical optimality of variational Bayes. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of *PMLR*, pages 1579–1588.

Petrov, V. V. (1975). Sums of independent random variables. In *Sums of Independent Random Variables*. De Gruyter.

Petrov, V. V. (2007). On lower bounds for tail probabilities. *Journal of statistical planning and inference*, 137(8):2703–2705.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131.

Pitman, J. W. (1976). On coupling of Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 35(4):315–322.

Plummer, M. (2014). Cuts in Bayesian graphical models. *Statistics and Computing*, pages 1–7.

Polyanskiy, Y. & Wu, Y. (2025). *Information Theory: From Coding to Learning*. Cambridge University Press, Cambridge.

Rainforth, T., Cornish, R., Yang, H., Warrington, A., & Wood, F. (2018a). On nesting Monte Carlo estimators. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pages 4267–4276.

Rainforth, T., Kosiorek, A. R., Le, T. A., Maddison, C. J., Igl, M., Wood, F., & Teh, Y. W. (2018b). Tighter variational bounds are not necessarily better. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pages 4277–4285.

Ranganath, R., Gerrish, S., & Blei, D. M. (2014). Black box variational inference. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 814–822.

Reeb, D., Doerr, A., Gerwinn, S., & Rakitsch, B. (2018). Learning Gaussian processes by minimizing pac-Bayesian generalization bounds. *Advances in Neural Information Processing Systems*, 31.

Ren, Y.-F. & Liang, H.-Y. (2001). On the best constant in Marcinkiewicz-Zygmund inequality. *Statistics & Probability Letters*, 53(3):227–233.

Rezende, D. J. & Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1278–1286.

Rhee, C.-h. & Glynn, P. W. (2015). Unbiased estimation with square root convergence for SDE models. *Operations Research*, 63(5):1026–1043.

Riou, C., Alquier, P., & Chérief-Abdellatif, B.-E. (2025). Bayes meets Bernstein at the meta level: An analysis of fast rates in meta-learning with PAC-Bayes. *Journal of Machine Learning Research*, 26:1–60.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

Robert, C. P. (2007). *The Bayesian Choice*. Springer, 2nd edition.

Robert, C. P. & Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, 2nd edition.

Roberts, G. O. & Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71.

Roberts, G. O. & Rosenthal, J. S. (2011). Quantitative non-geometric convergence bounds for independence samplers. *Methodology and Computing in Applied Probability*, 13(2):391–403.

Rosenthal, H. P. (1970). On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8:273–303.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566.

Rosenthal, J. S. (2000). Parallel computing and Monte Carlo algorithms. *Far East Journal of Theoretical Statistics*, 4:207–236.

Rousseau, J. (2016). On the frequentist properties of Bayesian nonparametric methods. *Annual Review of Statistics and Its Application*, 3:211–231.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., & Smith, A. F. M., editors, *Bayesian Statistics 3*, pages 395–402. Oxford University Press.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2):319–392.

Russo, D. & Zou, J. (2016). Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51 of *Proceedings of Machine Learning Research*, pages 1232–1240.

Shao, J. & Tu, D. (2012). *The jackknife and bootstrap*. Springer Science & Business Media.

Shawe-Taylor, J. & Williamson, R. C. (1997). A PAC analysis of a Bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory (COLT)*, pages 2–9.

Shen, X. & Wasserman, L. (2001). Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714.

Shi, Y. & Cornish, R. (2021). On multilevel Monte Carlo unbiased gradient estimation for deep latent variable models. In Banerjee, A. & Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3925–3933. PMLR.

Skare, Ø., Bølviken, E., & Holden, L. (2003). Improved sampling-importance resampling and reduced bias importance sampling. *Scandinavian Journal of Statistics*, 30(4):719–737.

South, L. F., Pettitt, A. N., & Drovandi, C. C. (2019). Sequential Monte Carlo Samplers with Independent Markov Chain Monte Carlo Proposals. *Bayesian Analysis*, 14(3):753 – 776.

Steinke, T. & Zakynthinou, L. (2020). Reasoning about generalization via conditional mutual information. In *Proceedings of the 33rd Conference on Learning Theory (COLT)*, volume 125 of *PMLR*, pages 3437–3452.

Surendran, S., Fermanian, A., Godichon-Baggioni, A., & Le Corff, S. (2024). Non-asymptotic analysis of biased adaptive stochastic approximation. In *Advances in Neural Information Processing Systems*, volume 37.

Surendran, S., Godichon-Baggioni, A., & Le Corff, S. (2025). Theoretical convergence guarantees for variational autoencoders. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*. PMLR.

Tadic, V. B. & Doucet, A. (2017). Asymptotic bias of stochastic gradient search.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1762.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Tokdar, S. T. & Kass, R. E. (2010). Importance sampling: a review. *WIREs Computational Statistics*, 2(1):54–60.

Tsybakov, A. B. (2008). Nonparametric estimators. In *Introduction to Nonparametric Estimation*, pages 1–76. Springer.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. & van Zanten, J. H. (2008). Rates of convergence for posterior distributions. 36(1):28–59.

van Erven, T. & Harremoës, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

VanDerwerken, D. N. & Schmidler, S. C. (2013). Parallel Markov chain Monte Carlo. *arXiv preprint arXiv:1312.7479*.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2024). Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72):1–58. arXiv:1507.02646.

Vihola, M. (2018). Unbiased estimators and multilevel Monte Carlo. *Operations Research*, 66(2):448–462.

von Bahr, B. & Esseen, C.-G. (1965). Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303.

Wainwright, M. J. & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Walker, S. G. & Hjort, N. L. (2001). On Bayesian consistency. *Journal of the Royal Statistical Society: Series B*, 63(4):811–821.

Wang, G. (2022). Exact convergence analysis of the independent Metropolis–Hastings algorithms. *Bernoulli*, 28(3):2012–2033.

Wang, G., Blanchet, J., & Glynn, P. W. (2024). When are unbiased Monte Carlo estimators more preferable than biased ones? *arXiv preprint arXiv:2404.01431*.

Wang, G., O'Leary, J., & Jacob, P. (2021). Maximal Couplings of the Metropolis-Hastings Algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 1225–1233. PMLR.

Wang, T. & Wang, G. (2023). Unbiased multilevel Monte Carlo methods for intractable distributions: Mlmc meets mcmc. *Journal of Machine Learning Research*, 24(249):1–40.

Wang, Y. & Blei, D. M. (2019). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Winter, S., Campbell, T., Lin, L., Srivastava, S., & Dunson, D. B. (2024). Emerging directions in Bayesian computation. *Statistical Science*, 39(1):62–89.

Wong, W. H. & Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. 23(2):339–362.

Xu, A. & Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30.

Yang, X. & Liu, J. S. (2023). Convergence rate of multiple-try Metropolis independent sampler. *Statistics and Computing*, 33(5):117. arXiv:2111.15084.

Yang, Y., Pati, D., & Bhattacharya, A. (2020). α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Yes, but did it work?: evaluating variational inference. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pages 5581–5590.

Zellner, A. (1988). Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4):278–280.

Zhang, F. & Gao, C. (2020). Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180–2207.

Zhang, T. (2006a). From ε-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210.

Zhang, T. (2006b). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321.

Zhou, X. & Reiter, J. P. (2010). A note on Bayesian inference with missing data. *The American Statistician*, 64(2):159–163.