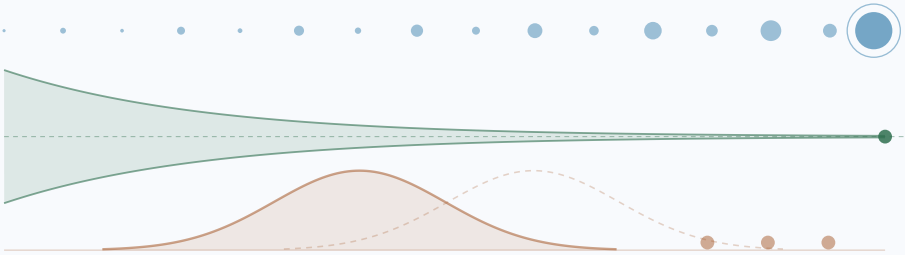
Contributions to the Theory of Bayesian Computation

El Mahdi Khribch

PhD Defense · ESSEC Business School · 21 September 2026

Supervised by Pierre E. Jacob · Pierre Alquier



Bias

$\omega = \pi_n / q$

Information

$\mathcal{I}(\theta; \mathcal{S})$

Robustness

$\mathcal{H}^2(P_*, P_\theta)$

Outline

Part 1



Bias

1. Importance sampling and particle Metropolis–Hastings
2. The bias each one carries under unbounded weights
3. Coupling, and unbiased estimation

Part 2



Information

1. From the price of a prior to mutual information
2. Tempered posteriors and their variational approximations
3. Rates without the logarithmic factor

Part 3



Robustness

1. Hellinger in place of the log-likelihood
2. Oracle inequalities under contamination
3. A saddle point we can solve

The posterior and its functionals

THE MODEL

a family $\{P_\theta\}_{\theta \in \Theta}$ with densities p_θ, a prior π on Θ, and a sample $\mathcal{S} = (X_1, \dots, X_n)$ from an unknown P_*

The posterior and its functionals

THE MODEL

a family $\{P_\theta\}_{\theta\in\Theta}$ with densities p_θ, a prior π on Θ, and a sample $\mathcal{S} = (X_1, \dots, X_n)$ from an unknown P_*

THE POSTERIOR

$$\pi_n(\mathrm{d}\theta) := \pi(\mathrm{d}\theta \mid \mathcal{S}) = \frac{e^{\ell_n(\theta)}\pi(\mathrm{d}\theta)}{Z_n}$$

$$\ell_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i), \qquad Z_n = \int_\Theta e^{\ell_n(\theta)}\pi(\mathrm{d}\theta)$$

The posterior and its functionals

THE MODEL

a family $\{P_\theta\}_{\theta\in\Theta}$ with densities p_θ, a prior π on Θ, and a sample $\mathcal{S} = (X_1, \dots, X_n)$ from an unknown P_*

THE POSTERIOR

$$\pi_n(\mathrm{d}\theta) := \pi(\mathrm{d}\theta \mid \mathcal{S}) = \frac{e^{\ell_n(\theta)}\pi(\mathrm{d}\theta)}{Z_n}$$

$\ell_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$, $\quad Z_n = \int_\Theta e^{\ell_n(\theta)}\pi(\mathrm{d}\theta)$

WHAT WE WANT TO COMPUTE

$$\pi_n(f) = \int_\Theta f(\theta)\,\pi_n(\mathrm{d}\theta)$$

for a test function $f : \Theta \to \mathbb{R}$

The posterior and its functionals

THE MODEL

a family $\{P_\theta\}_{\theta\in\Theta}$ with densities p_θ, a prior π on Θ, and a sample $\mathcal{S} = (X_1, \dots, X_n)$ from an unknown P_*

THE POSTERIOR

$$\pi_n(\mathrm{d}\theta) := \pi(\mathrm{d}\theta \mid \mathcal{S}) = \frac{e^{\ell_n(\theta)}\pi(\mathrm{d}\theta)}{Z_n}$$

$$\ell_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i), \qquad Z_n = \int_\Theta e^{\ell_n(\theta)}\pi(\mathrm{d}\theta)$$

WHAT WE WANT TO COMPUTE

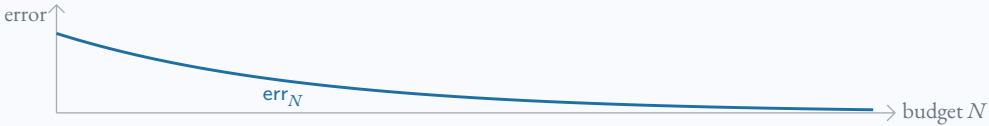
$$\pi_n(f) = \int_\Theta f(\theta)\,\pi_n(\mathrm{d}\theta)$$

for a test function $f : \Theta \to \mathbb{R}$

We evaluate $e^{\ell_n(\theta)}\pi(\theta)$ at any θ, and neither Z_n nor $\pi_n(f)$ in closed form.

So we approximate them, and every approximation carries an error.

Non-asymptotic bounds on the error

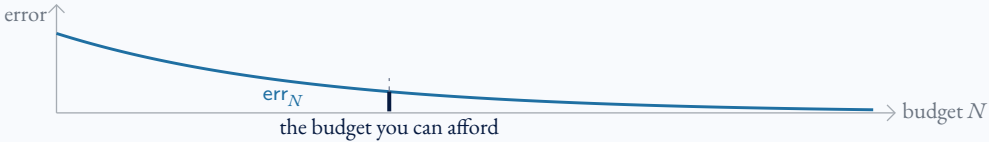


THE APPROXIMATION

$$\hat{\pi} \in \mathscr{P}(\Theta), \qquad \hat{\pi}(f) \approx \pi_n(f)$$

$\mathscr{P}(\Theta)$: the probability measures on Θ

Non-asymptotic bounds on the error



THE APPROXIMATION

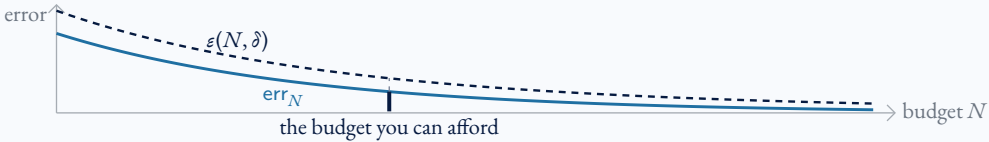
$$\hat{\pi} \in \mathscr{P}(\Theta), \qquad \hat{\pi}(f) \approx \pi_n(f)$$

$\mathscr{P}(\Theta)$: the probability measures on Θ

ITS ERROR, SAY IN TOTAL VARIATION

$$\mathrm{err}_N = \left|\hat{\pi} - \pi_n\right|_{\mathrm{TV}} = \sup_{|f|\leq 1} \left|\hat{\pi}(f) - \pi_n(f)\right|$$

Non-asymptotic bounds on the error



THE APPROXIMATION

$$\hat{\pi} \in \mathscr{P}(\Theta), \qquad \hat{\pi}(f) \approx \pi_n(f)$$

$\mathscr{P}(\Theta)$: the probability measures on Θ

ITS ERROR, SAY IN TOTAL VARIATION

$$\mathrm{err}_N = \left|\hat{\pi} - \pi_n\right|_{\mathrm{TV}} = \sup_{|f|\leq 1} \left|\hat{\pi}(f) - \pi_n(f)\right|$$

WHAT WE ASK OF IT

$$\mathbb{P}\Big(\mathrm{err}_N \leq \varepsilon(N,\delta)\Big) \geq 1 - \delta$$



Part 1 · Chapter 2

Bias under unbounded weights



$$\omega = \pi_n / q$$

Outline

Part 1



Bias

1 Importance sampling and particle Metropolis–Hastings
2 The bias each one carries under unbounded weights
3 Coupling, and unbiased estimation

Part 2

Information

Part 3

Robustness

Self-normalised importance sampling

SNIS *Kahn, 1949*

A weight on each draw from a proposal q, computable up to Z_n alone:

$$\omega = \frac{\pi_n}{q} = \frac{e^{\ell_n}\pi}{Z_n q}$$

Self-normalised importance sampling

SNIS *Kahn, 1949*

A weight on each draw from a proposal q, computable up to Z_n alone:

$$\omega = \frac{\pi_n}{q} = \frac{e^{\ell_n}\pi}{Z_n q}$$

INDEPENDENT DRAWS $\theta^{1:N} = (\theta_1, \dots, \theta_N)$ FROM q, AND THEIR TWO AVERAGES

$$\widehat{Z}_N(\theta^{1:N}) = N^{-1} \sum_{j=1}^{N} \omega(\theta_j), \qquad \widehat{F}_N(\theta^{1:N}) = \frac{\sum_{j=1}^{N} \omega(\theta_j) f(\theta_j)}{\sum_{j=1}^{N} \omega(\theta_j)}$$

Self-normalised importance sampling

SNIS *Kahn, 1949*

A weight on each draw from a proposal q, computable up to Z_n alone:

$$\omega = \frac{\pi_n}{q} = \frac{e^{\ell_n}\pi}{Z_n q}$$

INDEPENDENT DRAWS $\theta^{1:N} = (\theta_1, \dots, \theta_N)$ FROM q, AND THEIR TWO AVERAGES

$$\widehat{Z}_N(\theta^{1:N}) = N^{-1}\sum_{j=1}^{N}\omega(\theta_j), \qquad \hat{F}_N(\theta^{1:N}) = \frac{\sum_{j=1}^{N}\omega(\theta_j)f(\theta_j)}{\sum_{j=1}^{N}\omega(\theta_j)}$$

Z_n cancels in $\hat{F}_N$, $\widehat{Z}_N$ has mean one, and $\hat{F}_N \to \pi_n(f)$ almost surely.

$\hat{F}_N$ is a ratio of two averages, hence biased at every finite N. By how much?

Particle independent Metropolis–Hastings

PIMH *Andrieu et al., 2010*

A chain on Θ^N whose state is a whole particle set $\theta^{1:N}$:

$$\bar{q}(\mathrm{d}\theta^{1:N}) = \prod_{j=1}^{N} q(\mathrm{d}\theta_j), \qquad \bar{\pi}(\mathrm{d}\theta^{1:N}) = \widehat{Z}_N(\theta^{1:N})\, \bar{q}(\mathrm{d}\theta^{1:N})$$

Particle independent Metropolis–Hastings

PIMH *Andrieu et al., 2010*

A chain on Θ^N whose state is a whole particle set $\theta^{1:N}$:

$$\bar{q}(\mathrm{d}\theta^{1:N}) = \prod_{j=1}^{N} q(\mathrm{d}\theta_j), \qquad \bar{\pi}(\mathrm{d}\theta^{1:N}) = \widehat{Z}_N(\theta^{1:N})\, \bar{q}(\mathrm{d}\theta^{1:N})$$

ONE STEP: PROPOSE $\theta^{1:N,\star}$, ACCEPT IT WITH PROBABILITY α

$$\theta^{1:N,\star} \sim \bar{q}, \qquad \alpha = \min\{1,\ \widehat{Z}_N(\theta^{1:N,\star}) / \widehat{Z}_N(\theta^{1:N})\}$$

Particle independent Metropolis–Hastings

PIMH *Andrieu et al., 2010*

A chain on Θ^N whose state is a whole particle set $\theta^{1:N}$:

$$\bar{q}(\mathrm{d}\theta^{1:N}) = \prod_{j=1}^{N} q(\mathrm{d}\theta_j), \qquad \bar{\pi}(\mathrm{d}\theta^{1:N}) = \widehat{Z}_N(\theta^{1:N})\, \bar{q}(\mathrm{d}\theta^{1:N})$$

ONE STEP: PROPOSE $\theta^{1:N,\star}$, ACCEPT IT WITH PROBABILITY α

$$\theta^{1:N,\star} \sim \bar{q}, \qquad \alpha = \min\{1,\ \widehat{Z}_N(\theta^{1:N,\star}) / \widehat{Z}_N(\theta^{1:N})\}$$

$N = 1$ recovers independent Metropolis–Hastings *Hastings, 1970*

$$\theta^\star \sim q, \qquad \alpha = \min\{1,\ \omega(\theta^\star)/\omega(\theta)\}$$

Particle independent Metropolis–Hastings

PIMH *Andrieu et al., 2010*

A chain on Θ^N whose state is a whole particle set $\theta^{1:N}$:

$$\bar{q}(\mathrm{d}\theta^{1:N}) = \prod_{j=1}^{N} q(\mathrm{d}\theta_j), \qquad \bar{\pi}(\mathrm{d}\theta^{1:N}) = \widehat{Z}_N(\theta^{1:N})\, \bar{q}(\mathrm{d}\theta^{1:N})$$

ONE STEP: PROPOSE $\theta^{1:N,\star}$, ACCEPT IT WITH PROBABILITY α

$$\theta^{1:N,\star} \sim \bar{q}, \qquad \alpha = \min\{1,\ \widehat{Z}_N(\theta^{1:N,\star}) / \widehat{Z}_N(\theta^{1:N})\}$$

$N = 1$ recovers independent Metropolis–Hastings *Hastings, 1970*

$$\theta^\star \sim q, \qquad \alpha = \min\{1,\ \omega(\theta^\star)/\omega(\theta)\}$$

THE ESTIMATOR, ALONG T STEPS OF THE IMH CHAIN

$$\hat{\pi}(f) = T^{-1} \sum_{t=0}^{T-1} f(\theta_t)$$

The variance is settled, the bias is not

The asymptotic variance σ^2_{IMH} of the IMH ergodic average is finite if and only if $\pi_n(f^2) < \infty$ and $q(\omega^2 f^2) < \infty$ *Deligiannidis & Lee, 2018*. Whenever both hold:

$$\sigma^2_{\text{IS}} \leq \sigma^2_{\text{IMH}}$$

$\sigma^2_{\text{IS}} = q\big(\omega^2 (f - \pi_n(f))^2\big)$ is the asymptotic variance of $\sqrt{N}(\widehat{F}_N - \pi_n(f))$

The variance is settled, the bias is not

The asymptotic variance σ^2_{IMH} of the IMH ergodic average is finite if and only if $\pi_n(f^2) < \infty$ and $q(\omega^2 f^2) < \infty$ *Deligiannidis & Lee, 2018*. Whenever both hold:

$$\sigma^2_{\text{IS}} \;\leq\; \sigma^2_{\text{IMH}}$$

$\sigma^2_{\text{IS}} = q\big(\omega^2 (f - \pi_n(f))^2\big)$ is the asymptotic variance of $\sqrt{N}(\widehat{F}_N - \pi_n(f))$

At finite N both are biased: $\widehat{F}_N$ through the ratio of two averages, the chain through its initial law.

The variance is settled, the bias is not

The asymptotic variance σ^2_{IMH} of the IMH ergodic average is finite if and only if $\pi_n(f^2) < \infty$ and $q(\omega^2 f^2) < \infty$ *Deligiannidis & Lee, 2018*. Whenever both hold:

$$\sigma^2_{\text{IS}} \;\leq\; \sigma^2_{\text{IMH}}$$

$\sigma^2_{\text{IS}} = q\big(\omega^2 (f - \pi_n(f))^2\big)$ is the asymptotic variance of $\sqrt{N}(\widehat{F}_N - \pi_n(f))$

At finite N both are biased: $\widehat{F}_N$ through the ratio of two averages, the chain through its initial law.

The ordering of the variances is known. The ordering of the biases is not.

What a biased estimator costs downstream

THREE PLACES WHERE THE BIAS IS PAID FOR

- stochastic gradients *Surendran et al., 2025*
- importance-weighted autoencoders *Daudel et al., 2024*
- cut models, experimental design *Plummer, 2014; Rainforth et al., 2018*

What a biased estimator costs downstream

THREE PLACES WHERE THE BIAS IS PAID FOR

- stochastic gradients *Surendran et al., 2025*
- importance-weighted autoencoders *Daudel et al., 2024*
- cut models, experimental design *Plummer, 2014; Rainforth et al., 2018*

A MOTIVATING APPLICATION: A NESTED EXPECTATION OVER (Θ_1, Θ_2)

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

What a biased estimator costs downstream

THREE PLACES WHERE THE BIAS IS PAID FOR

- stochastic gradients *Surendran et al., 2025*
- importance-weighted autoencoders *Daudel et al., 2024*
- cut models, experimental design *Plummer, 2014; Rainforth et al., 2018*

A MOTIVATING APPLICATION: A NESTED EXPECTATION OVER (Θ_1, Θ_2)

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

INNER $\hat{F}(\theta_1)$ approximates $\mathbb{E}_{2|1}[f \mid \Theta_1 = \theta_1]$ at cost N, with bias $N^{-\gamma}$

What a biased estimator costs downstream

THREE PLACES WHERE THE BIAS IS PAID FOR

- stochastic gradients *Surendran et al., 2025*
- importance-weighted autoencoders *Daudel et al., 2024*
- cut models, experimental design *Plummer, 2014; Rainforth et al., 2018*

A MOTIVATING APPLICATION: A NESTED EXPECTATION OVER (Θ_1, Θ_2)

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

INNER $\hat{F}(\theta_1)$ approximates $\mathbb{E}_{2|1}[f \mid \Theta_1 = \theta_1]$ at cost N, with bias $N^{-\gamma}$

OUTER $\hat{F}_{M,N} = M^{-1} \sum_{m=1}^{M} \hat{F}(\Theta_{1,m})$ over M draws, at total cost $C = MN$

What a biased estimator costs downstream

THREE PLACES WHERE THE BIAS IS PAID FOR

- stochastic gradients *Surendran et al., 2025*
- importance-weighted autoencoders *Daudel et al., 2024*
- cut models, experimental design *Plummer, 2014; Rainforth et al., 2018*

A MOTIVATING APPLICATION: A NESTED EXPECTATION OVER (Θ_1, Θ_2)

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

INNER $\hat{F}(\theta_1)$ approximates $\mathbb{E}_{2|1}[f \mid \Theta_1 = \theta_1]$ at cost N, with bias $N^{-\gamma}$

OUTER $\hat{F}_{M,N} = M^{-1} \sum_{m=1}^{M} \hat{F}(\Theta_{1,m})$ over M draws, at total cost $C = MN$

$\mathrm{MSE} = \mathcal{O}\big(C^{-2\gamma/(2\gamma+1)}\big)$ at the optimal split of C *Andradóttir & Glynn, 2016*

Reducing the bias improves the rate; removing it attains the canonical rate C^{-1}.

The asymptotic bias, proved

Chapter 2, Theorem 2.8

hypotheses: $\pi_n(|f|) < \infty$; $q\big(|\omega^2(f - \pi_n(f))|^{1+\varepsilon}\big) < \infty$ for some $\varepsilon > 0$; $q(\omega^{-\eta}) < \infty$ for some $\eta > 0$; $\theta^{1:N} \sim \tilde{q}$

$$\lim_{N\to\infty} N\mathbb{E}\big[\hat{F}_N(\theta^{1:N}) - \pi_n(f)\big] = -\int \big(f - \pi_n(f)\big)\omega^2 \mathrm{d}q$$

The asymptotic bias, proved

Chapter 2, Theorem 2.8

hypotheses: $\pi_n(|f|) < \infty$; $q\big(|\omega^2(f - \pi_n(f))|^{1+\varepsilon}\big) < \infty$ for some $\varepsilon > 0$; $q(\omega^{-\eta}) < \infty$ for some $\eta > 0$; $\theta^{1:N} \sim \tilde{q}$

$$\lim_{N\to\infty} N\mathbb{E}\big[\hat{F}_N(\theta^{1:N}) - \pi_n(f)\big] = -\int \big(f - \pi_n(f)\big)\omega^2 \mathrm{d}q$$

Long known *Hesterberg, 1988; Liu, 2008*, but never proved: we give the first proof, by an elementary argument. Another proof, independent and under weaker conditions, is due to *Daudel & Roueff, 2024*.

The asymptotic bias, proved

Chapter 2, Theorem 2.8

hypotheses: $\pi_n(|f|) < \infty$; $q\big(|\omega^2(f - \pi_n(f))|^{1+\varepsilon}\big) < \infty$ for some $\varepsilon > 0$; $q(\omega^{-\eta}) < \infty$ for some $\eta > 0$; $\theta^{1:N} \sim \tilde{q}$

$$\lim_{N\to\infty} N\mathbb{E}\big[\hat{F}_N(\theta^{1:N}) - \pi_n(f)\big] = -\int \big(f - \pi_n(f)\big)\omega^2 \mathrm{d}q$$

Long known *Hesterberg, 1988; Liu, 2008*, but never proved: we give the first proof, by an elementary argument. Another proof, independent and under weaker conditions, is due to *Daudel & Roueff, 2024*.

at $q(\omega^2) < \infty$, a bound at every N

$$\sup_{|f|\le 1} \big|\mathbb{E}[\hat{F}_N] - \pi_n(f)\big| \;\le\; \frac{12q(\omega^2)}{N}$$

Agapiou et al., 2017

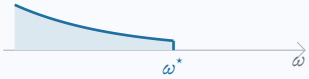
When the weight is unbounded

For IMH the distinction is between two cases *Mengersen & Tweedie, 1996*

When the weight is unbounded

For IMH the distinction is between two cases *Mengersen & Tweedie, 1996*

BOUNDED WEIGHT



uniformly ergodic if and only if $\omega^\star = \operatorname{ess\,sup} \omega < \infty$, and then, over starting points θ in the support of π_n,

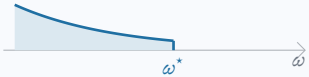
$$\sup_\theta |P^t(\theta, \cdot) - \pi_n|_{\mathrm{TV}} = (1 - 1/\omega^\star)^t$$

Wang, 2022

When the weight is unbounded

For IMH the distinction is between two cases *Mengersen & Tweedie, 1996*

BOUNDED WEIGHT



uniformly ergodic if and only if $\omega^\star = \operatorname{ess\,sup} \omega < \infty$, and then, over starting points θ in the support of π_n,

$$\sup_\theta \left| P^t(\theta, \cdot) - \pi_n \right|_{\mathrm{TV}} = (1 - 1/\omega^\star)^t$$

Wang, 2022

UNBOUNDED WEIGHT



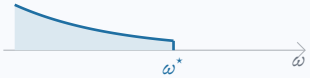
if $\omega^\star = \infty$, no geometric rate, whatever the moments of ω:

$$\sup_{t \geq 0} \rho^{-t} \left| P^t(\theta, \cdot) - \pi_n \right|_{\mathrm{TV}} = \infty$$

When the weight is unbounded

For IMH the distinction is between two cases *Mengersen & Tweedie, 1996*

BOUNDED WEIGHT



uniformly ergodic if and only if $\omega^\star = \operatorname{ess\,sup} \omega < \infty$, and then, over starting points θ in the support of π_n,

$$\sup_\theta \left| P^t(\theta, \cdot) - \pi_n \right|_{\mathrm{TV}} = (1 - 1/\omega^\star)^t$$

Wang, 2022

UNBOUNDED WEIGHT



if $\omega^\star = \infty$, no geometric rate, whatever the moments of ω:

$$\sup_{t \geq 0} \rho^{-t} \left| P^t(\theta, \cdot) - \pi_n \right|_{\mathrm{TV}} = \infty$$

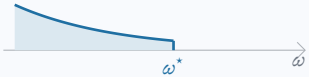
We assume a moment on the weight,

$$q(\omega^p) < \infty \quad \text{for some } p > 1,$$

When the weight is unbounded

For IMH the distinction is between two cases *Mengersen & Tweedie, 1996*

BOUNDED WEIGHT



uniformly ergodic if and only if $\omega^\star = \operatorname{ess\,sup} \omega < \infty$, and then, over starting points θ in the support of π_n,

$$\sup_\theta \left| P^t(\theta, \cdot) - \pi_n \right|_{\mathrm{TV}} = (1 - 1/\omega^\star)^t$$

Wang, 2022

UNBOUNDED WEIGHT



if $\omega^\star = \infty$, no geometric rate, whatever the moments of ω:

$$\sup_{t \geq 0} \rho^{-t} \left| P^t(\theta, \cdot) - \pi_n \right|_{\mathrm{TV}} = \infty$$

We assume a moment on the weight,

$$q(\omega^p) < \infty \quad \text{for some } p > 1,$$

and ask what rates in t and in N we can obtain.

A coupling that is maximal at all times

Run two chains (θ_t) and (ϑ_t) from the same proposals and the same uniforms. With $\tau = \inf\{t \geq 1 : \theta_t = \vartheta_t\}$, the coupling is faithful: $\theta_t = \vartheta_t$ for every $t \geq \tau$.

A coupling that is maximal at all times

Run two chains (θ_t) and (ϑ_t) from the same proposals and the same uniforms. With $\tau = \inf\{t \geq 1 : \theta_t = \vartheta_t\}$, the coupling is faithful: $\theta_t = \vartheta_t$ for every $t \geq \tau$.

Chapter 2, Theorem 2.11

$$\left| P^t(\theta, \cdot) - P^t(\vartheta, \cdot) \right|_{\mathrm{TV}} = \mathbb{P}_{\theta,\vartheta}(\tau > t)$$

maximal, in the sense of *Pitman, 1976*

A coupling that is maximal at all times

Run two chains (θ_t) and (ϑ_t) from the same proposals and the same uniforms. With $\tau = \inf\{t \geq 1 : \theta_t = \vartheta_t\}$, the coupling is faithful: $\theta_t = \vartheta_t$ for every $t \geq \tau$.

Chapter 2, Theorem 2.11

$$\left| P^t(\theta, \cdot) - P^t(\vartheta, \cdot) \right|_{\mathrm{TV}} = \mathbb{P}_{\theta,\vartheta}(\tau > t)$$

maximal, in the sense of *Pitman, 1976*

meeting times, $p \geq 2$

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

Prop. 2.14

A coupling that is maximal at all times

Run two chains (θ_t) and (ϑ_t) from the same proposals and the same uniforms. With $\tau = \inf\{t \geq 1 : \theta_t = \vartheta_t\}$, the coupling is faithful: $\theta_t = \vartheta_t$ for every $t \geq \tau$.

Chapter 2, Theorem 2.11

$$\left| P^t(\theta, \cdot) - P^t(\vartheta, \cdot) \right|_{\mathrm{TV}} = \mathbb{P}_{\theta,\vartheta}(\tau > t)$$

maximal, in the sense of *Pitman, 1976*

meeting times, $p \geq 2$

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

Prop. 2.14

The first maximal Markovian coupling known for an MCMC algorithm.

Polynomial rates, and their sharpness

Chapter 2, Theorem 2.16

PIMH started from $\bar{q}$, under $q(\omega^p) < \infty$ with $p \geq 2$

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

Polynomial rates, and their sharpness

Chapter 2, Theorem 2.16

PIMH started from $\bar{q}$, under $q(\omega^p) < \infty$ with $p \geq 2$

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

for $1 < p < 2$, the same bound with $N^{(p-1)/p}$ in place of $\sqrt{N}$ Thm. 2.20

Polynomial rates, and their sharpness

Chapter 2, Theorem 2.16

PIMH started from $\tilde{q}$, under $q(\omega^p) < \infty$ with $p \geq 2$

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

for $1 < p < 2$, the same bound with $N^{(p-1)/p}$ in place of $\sqrt{N}$ — Thm. 2.20

$$\left|P^t(\theta,\cdot) - \pi_n\right|_{\mathrm{TV}} \leq r(\theta)^t + \frac{D}{t^{p-1}} \qquad N = 1,\ p > 1 \qquad \text{Cor. 2.18}$$

sharp: for some target and proposal, $\left|P^t(\theta,\cdot) - \pi_n\right|_{\mathrm{TV}} \geq c\, t^{-(p-1)} (\log t)^{-k}$ — Ex. 2.23

Polynomial rates, and their sharpness

Chapter 2, Theorem 2.16

PIMH started from $\tilde{q}$, under $q(\omega^p) < \infty$ with $p \geq 2$

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

for $1 < p < 2$, the same bound with $N^{(p-1)/p}$ in place of $\sqrt{N}$ — Thm. 2.20

$$\left|P^t(\theta,\cdot) - \pi_n\right|_{\mathrm{TV}} \leq r(\theta)^t + \frac{D}{t^{p-1}} \qquad N = 1,\ p > 1 \qquad \text{Cor. 2.18}$$

sharp: for some target and proposal, $\left|P^t(\theta,\cdot) - \pi_n\right|_{\mathrm{TV}} \geq c\, t^{-(p-1)} (\log t)^{-k}$ — Ex. 2.23

the two biases, at a matched budget of N weight evaluations

$$\sup_{|f|\leq 1} \left|\mathbb{E}[\hat{F}_N] - \pi_n(f)\right| \leq 12 q(\omega^2) N^{-1} \qquad \left|qP^N - \pi_n\right|_{\mathrm{TV}} \leq C N^{-(p-1)}$$

Polynomial rates, and their sharpness

Chapter 2, Theorem 2.16

PIMH started from $\bar{q}$, under $q(\omega^p) < \infty$ with $p \geq 2$

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

for $1 < p < 2$, the same bound with $N^{(p-1)/p}$ in place of $\sqrt{N}$ — Thm. 2.20

$$\left|P^t(\theta,\cdot) - \pi_n\right|_{\mathrm{TV}} \leq r(\theta)^t + \frac{D}{t^{p-1}} \qquad N = 1,\ p > 1 \qquad \text{Cor. 2.18}$$

sharp: for some target and proposal, $\left|P^t(\theta,\cdot) - \pi_n\right|_{\mathrm{TV}} \geq c\, t^{-(p-1)} (\log t)^{-k}$ — Ex. 2.23

THE TWO BIASES, AT A MATCHED BUDGET OF N WEIGHT EVALUATIONS

$$\sup_{|f|\leq 1} \left|\mathbb{E}[\hat{F}_N] - \pi_n(f)\right| \leq 12 q(\omega^2) N^{-1} \qquad \left|qP^N - \pi_n\right|_{\mathrm{TV}} \leq C N^{-(p-1)}$$

Once $p > 2$, the state of the IMH chain at time N carries the smaller bias.

Removing the bias, at the inefficiency of SNIS

COUPLING-UIS *Glynn & Rhee, 2014; Middleton et al., 2019* SEC. 2.5

$$\hat{F}_N^u = \hat{F}_N(\theta_0^{1:N}) + \sum_{t=1}^{\tau-1} \left\{ \hat{F}_N(\theta_t^{1:N}) - \hat{F}_N(\vartheta_{t-1}^{1:N}) \right\}$$

Removing the bias, at the inefficiency of SNIS

COUPLING-UIS *Glynn & Rhee, 2014; Middleton et al., 2019* SEC. 2.5

$$\hat{F}_N^u = \hat{F}_N(\theta_0^{1:N}) + \sum_{t=1}^{\tau-1} \left\{ \hat{F}_N(\theta_t^{1:N}) - \hat{F}_N(\vartheta_{t-1}^{1:N}) \right\}$$

unbiased for $\pi_n(f)$ at every $N \geq 1$, at cost $\mathscr{C} = N(\tau + 1)$ with $\mathbb{E}[\mathscr{C}] \sim 2N$

Removing the bias, at the inefficiency of SNIS

COUPLING-UIS *Glynn & Rhee, 2014; Middleton et al., 2019* SEC. 2.5

$$\hat{F}_N^u = \hat{F}_N(\theta_0^{1:N}) + \sum_{t=1}^{\tau-1} \left\{ \hat{F}_N(\theta_t^{1:N}) - \hat{F}_N(\vartheta_{t-1}^{1:N}) \right\}$$

unbiased for $\pi_n(f)$ at every $N \geq 1$, at cost $\mathscr{C} = N(\tau + 1)$ with $\mathbb{E}[\mathscr{C}] \sim 2N$

CHAPTER 2, PROPOSITION 2.30

$$\lim_{N \to \infty} N \, \mathbb{E}\left[(\hat{F}_N^u - \pi_n(f))^2 \right] = \sigma_{\text{IS}}^2$$

the mean squared error of SNIS itself, under a few more moments

Removing the bias, at the inefficiency of SNIS

COUPLING-UIS *Glynn & Rhee, 2014; Middleton et al., 2019* SEC. 2.5

$$\hat{F}_N^u = \hat{F}_N(\theta_0^{1:N}) + \sum_{t=1}^{\tau-1} \left\{ \hat{F}_N(\theta_t^{1:N}) - \hat{F}_N(\vartheta_{t-1}^{1:N}) \right\}$$

unbiased for $\pi_n(f)$ at every $N \geq 1$, at cost $\mathscr{C} = N(\tau + 1)$ with $\mathbb{E}[\mathscr{C}] \sim 2N$

CHAPTER 2, PROPOSITION 2.30

$$\lim_{N \to \infty} N \, \mathbb{E}\left[(\hat{F}_N^u - \pi_n(f))^2 \right] = \sigma_{\text{IS}}^2$$

the mean squared error of SNIS itself, under a few more moments

Its inefficiency, expected cost times mean squared error, is twice that of SNIS, and equal to it once we average over the two initial draws. PROP. 2.32

Removing the bias, at the inefficiency of SNIS

COUPLING-UIS *Glynn & Rhee, 2014; Middleton et al., 2019* SEC. 2.5

$$\widehat{F}_N^u = \widehat{F}_N(\theta_0^{1:N}) + \sum_{t=1}^{\tau-1} \left\{ \widehat{F}_N(\theta_t^{1:N}) - \widehat{F}_N(\vartheta_{t-1}^{1:N}) \right\}$$

unbiased for $\pi_n(f)$ at every $N \geq 1$, at cost $\mathscr{C} = N(\tau + 1)$ with $\mathbb{E}[\mathscr{C}] \sim 2N$

CHAPTER 2, PROPOSITION 2.30

$$\lim_{N \to \infty} N \, \mathbb{E}\left[\left(\widehat{F}_N^u - \pi_n(f) \right)^2 \right] = \sigma_{\text{IS}}^2$$

the mean squared error of SNIS itself, under a few more moments

Its inefficiency, expected cost times mean squared error, is twice that of SNIS, and equal to it once we average over the two initial draws. PROP. 2.32

Unbiased at every N, at the asymptotic inefficiency of SNIS.

Part 1 in one page

what Chapter 2 establishes

- bias and moments of importance sampling, unbounded test functions included Thm. 2.8, 2.9
- the first maximal Markovian coupling known for MCMC, and the rates it gives Thm. 2.11, 2.16
- several ways of removing the bias, and what each costs in efficiency Prop. 2.26, 2.32
- unbiased estimation of $1/Z$, nested expectations, robust mean estimation Sec. 2.6

Part 1 in one page

what Chapter 2 establishes

- bias and moments of importance sampling, unbounded test functions included Thm. 2.8, 2.9
- the first maximal Markovian coupling known for MCMC, and the rates it gives Thm. 2.11, 2.16
- several ways of removing the bias, and what each costs in efficiency Prop. 2.26, 2.32
- unbiased estimation of $1/Z$, nested expectations, robust mean estimation Sec. 2.6

the same two ingredients; they differ only in what they do with them

$$\theta_1, \theta_2, \ldots \overset{\text{i.i.d.}}{\sim} q, \qquad \omega(\theta_1), \omega(\theta_2), \ldots$$

Part 1 in one page

what Chapter 2 establishes

- bias and moments of importance sampling, unbounded test functions included Thm. 2.8, 2.9
- the first maximal Markovian coupling known for MCMC, and the rates it gives Thm. 2.11, 2.16
- several ways of removing the bias, and what each costs in efficiency Prop. 2.26, 2.32
- unbiased estimation of $1/Z$, nested expectations, robust mean estimation Sec. 2.6

the same two ingredients; they differ only in what they do with them

$$\theta_1, \theta_2, \ldots \overset{\text{i.i.d.}}{\sim} q, \qquad \omega(\theta_1), \omega(\theta_2), \ldots$$

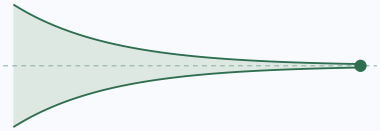
in progress

iterated sampling-importance resampling, and multiple-try Metropolis with an independent sampler, under unbounded weights *Bertholom et al., 2026*

Deligiannidis, Jacob, Khribch & Wang, *Annals of Statistics*, accepted

Part 2 · Chapter 3

Mutual information bounds



$\mathcal{I}(\theta; \mathcal{S})$

Outline

Part 1

Bias

Part 2



Information

1. From the price of a prior to mutual information
2. Tempered posteriors and their variational approximations
3. Rates without the logarithmic factor

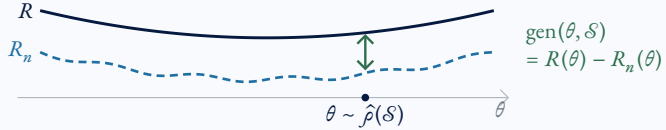
Part 3

Robustness

PAC-Bayes bounds and the price of a prior

A LOSS ℓ, ITS RISK UNDER THE LAW P OF THE DATA, AND ITS EMPIRICAL VERSION

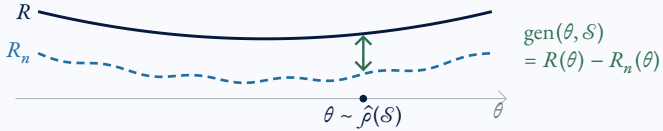
$$R(\theta) = \mathbb{E}_{X\sim P}\big[\ell(\theta, X)\big] \qquad R_n(\theta) = n^{-1}\sum_{i=1}^{n} \ell(\theta, X_i)$$



PAC-Bayes bounds and the price of a prior

A LOSS ℓ, ITS RISK UNDER THE LAW P OF THE DATA, AND ITS EMPIRICAL VERSION

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



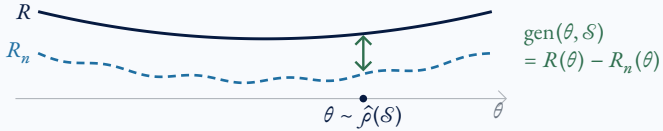
$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}}\big[\mathrm{gen}(\theta, \mathcal{S})\big] \le \sqrt{\frac{\mathbb{E}_{\mathcal{S}}\big[\mathrm{KL}(\hat{\rho}\|\pi)\big]}{2n}}$$

$\hat{\rho}$ is any law on Θ built from the sample, a posterior included; $0 \le \ell \le 1$, and π is fixed before it *Catoni, 2007*

PAC-Bayes bounds and the price of a prior

A LOSS ℓ, ITS RISK UNDER THE LAW P OF THE DATA, AND ITS EMPIRICAL VERSION

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



$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}}\big[\mathrm{gen}(\theta, \mathcal{S})\big] \le \sqrt{\frac{\mathbb{E}_{\mathcal{S}}\big[\mathrm{KL}(\hat{\rho}\|\pi)\big]}{2n}}$$

$\hat{\rho}$ is any law on Θ built from the sample, a posterior included; $0 \le \ell \le 1$, and π is fixed before it *Catoni, 2007*

The bound holds for every prior π, and π is a free choice in it.

At the best prior, the price is the mutual information

THE SAME GAP, WITH THE PRIOR OPTIMISED AWAY

$$\mathcal{J}(\theta;\mathcal{S}) = \mathrm{KL}\big(\mathbb{P}_{\theta,\mathcal{S}} \| \mathbb{P}_\theta \otimes \mathbb{P}_{\mathcal{S}}\big)$$

At the best prior, the price is the mutual information

THE SAME GAP, WITH THE PRIOR OPTIMISED AWAY

$$\mathcal{I}(\theta;\mathcal{S}) = \mathrm{KL}\big(\mathbb{P}_{\theta,\mathcal{S}} \,\|\, \mathbb{P}_\theta \otimes \mathbb{P}_{\mathcal{S}}\big)$$

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}}\big[\mathrm{gen}(\theta,\mathcal{S})\big] \le \sqrt{\frac{\mathcal{I}(\theta;\mathcal{S})}{2n}}$$

for a **bounded loss**, $0 \le \ell \le 1$ *Russo & Zou, 2016*

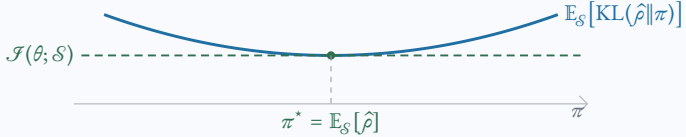
At the best prior, the price is the mutual information

THE SAME GAP, WITH THE PRIOR OPTIMISED AWAY

$$\mathcal{J}(\theta;\mathcal{S}) = \mathrm{KL}\big(\mathbb{P}_{\theta,\mathcal{S}} \,\|\, \mathbb{P}_\theta \otimes \mathbb{P}_{\mathcal{S}}\big)$$

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}}\big[\mathrm{gen}(\theta,\mathcal{S})\big] \le \sqrt{\frac{\mathcal{J}(\theta;\mathcal{S})}{2n}}$$

for a **bounded loss**, $0 \le \ell \le 1$ *Russo & Zou, 2016*



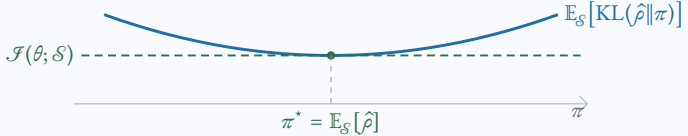
At the best prior, the price is the mutual information

THE SAME GAP, WITH THE PRIOR OPTIMISED AWAY

$$\mathcal{J}(\theta;\mathcal{S}) = \mathrm{KL}\big(\mathbb{P}_{\theta,\mathcal{S}} \| \mathbb{P}_\theta \otimes \mathbb{P}_{\mathcal{S}}\big)$$

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}}\big[\mathrm{gen}(\theta,\mathcal{S})\big] \le \sqrt{\frac{\mathcal{J}(\theta;\mathcal{S})}{2n}}$$

for a **bounded loss**, $0 \le \ell \le 1$ *Russo & Zou, 2016*



$$\mathbb{E}_{\mathcal{S}}\big[\mathrm{KL}(\hat{\rho}\|\pi)\big] = \mathcal{J}(\theta;\mathcal{S}) + \mathrm{KL}(\pi^\star\|\pi)$$ *Catoni, 2007*

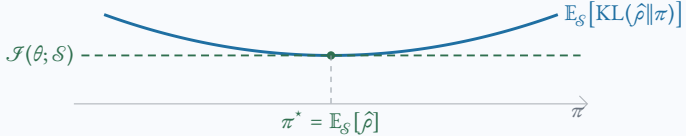
At the best prior, the price is the mutual information

THE SAME GAP, WITH THE PRIOR OPTIMISED AWAY

$$\mathcal{J}(\theta;\mathcal{S}) = \mathrm{KL}\big(\mathbb{P}_{\theta,\mathcal{S}} \,\|\, \mathbb{P}_\theta \otimes \mathbb{P}_{\mathcal{S}}\big)$$

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}}\big[\mathrm{gen}(\theta,\mathcal{S})\big] \le \sqrt{\frac{\mathcal{J}(\theta;\mathcal{S})}{2n}}$$

for a **bounded loss**, $0 \le \ell \le 1$ *Russo & Zou, 2016*



$$\mathbb{E}_{\mathcal{S}}\big[\mathrm{KL}(\hat{\rho}\|\pi)\big] = \mathcal{J}(\theta;\mathcal{S}) + \mathrm{KL}(\pi^\star\|\pi)$$ *Catoni, 2007*

The mutual information is the smallest complexity term over all priors.

Unbounded losses, and a Rényi divergence

STATISTICAL INFERENCE: THE LOSS IS A LOG-RATIO

$$\mathrm{LR}_n(\theta_0, \theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)} \longrightarrow \mathrm{KL}(P_{\theta_0} \| P_\theta)$$

Unbounded losses, and a Rényi divergence

STATISTICAL INFERENCE: THE LOSS IS A LOG-RATIO

$$\mathrm{LR}_n(\theta_0, \theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)} \longrightarrow \mathrm{KL}(P_{\theta_0} \| P_\theta)$$

$0 \le \ell \le 1$ fails here, and with it the argument behind both bounds above.

Unbounded losses, and a Rényi divergence

STATISTICAL INFERENCE: THE LOSS IS A LOG-RATIO

$$\mathrm{LR}_n(\theta_0, \theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)} \longrightarrow \mathrm{KL}(P_{\theta_0} \| P_\theta)$$

$0 \le \ell \le 1$ fails here, and with it the argument behind both bounds above.

ONE SURVIVES, WITH A RÉNYI DIVERGENCE ON THE LEFT

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\theta \sim \hat{\rho}} \left[\mathscr{D}_\alpha(P_\theta \| P_{\theta_0}) - \frac{\alpha}{1-\alpha} \mathrm{LR}_n(\theta_0, \theta) \right] \le \frac{\mathbb{E}_{\mathcal{S}}\left[\mathrm{KL}(\hat{\rho} \| \pi)\right]}{n(1-\alpha)}$$

Alquier & Ridgway, 2020; after Bhattacharya et al., 2019

Unbounded losses, and a Rényi divergence

STATISTICAL INFERENCE: THE LOSS IS A LOG-RATIO

$$\mathrm{LR}_n(\theta_0, \theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)} \longrightarrow \mathrm{KL}(P_{\theta_0} \| P_\theta)$$

$0 \le \ell \le 1$ fails here, and with it the argument behind both bounds above.

ONE SURVIVES, WITH A RÉNYI DIVERGENCE ON THE LEFT

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\theta \sim \hat{\rho}} \left[\mathscr{D}_\alpha(P_\theta \| P_{\theta_0}) - \frac{\alpha}{1-\alpha} \mathrm{LR}_n(\theta_0, \theta) \right] \le \frac{\mathbb{E}_{\mathcal{S}}\left[\mathrm{KL}(\hat{\rho} \| \pi)\right]}{n(1-\alpha)}$$

Alquier & Ridgway, 2020; after Bhattacharya et al., 2019

We can optimise the prior on the right-hand side again.

A bound with no prior in it

Chapter 3, Theorem 3.2

for every $\hat{\rho}$, every $\alpha \in (0,1)$, and no prior on the right

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}}\left[\mathscr{D}_{\alpha}(P_{\theta}\|P_{\theta_0}) - \frac{\alpha}{1-\alpha}\mathrm{LR}_n(\theta_0,\theta)\right] \leq \frac{\mathcal{I}(\theta;\mathcal{S})}{n(1-\alpha)}$$

A bound with no prior in it

Chapter 3, Theorem 3.2

for every $\hat{\rho}$, every $\alpha \in (0,1)$, and no prior on the right

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}}\left[\mathscr{D}_{\alpha}(P_{\theta}\|P_{\theta_0}) - \frac{\alpha}{1-\alpha}\mathrm{LR}_n(\theta_0,\theta)\right] \le \frac{\mathcal{I}(\theta;\mathcal{S})}{n(1-\alpha)}$$

at $\alpha = 1/2$, weakened by $\mathscr{H}^2 \le \mathscr{D}_{1/2}$

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}}\left[\mathscr{H}^2(P_{\theta},P_{\theta_0}) - \mathrm{LR}_n(\theta_0,\theta)\right] \le \frac{2\mathcal{I}(\theta;\mathcal{S})}{n}$$

A bound with no prior in it

Chapter 3, Theorem 3.2

for every $\hat{\rho}$, every $\alpha \in (0,1)$, and no prior on the right

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}}\left[\mathscr{D}_{\alpha}(P_{\theta}\|P_{\theta_0}) - \frac{\alpha}{1-\alpha}\mathrm{LR}_n(\theta_0,\theta)\right] \le \frac{\mathcal{I}(\theta;\mathcal{S})}{n(1-\alpha)}$$

at $\alpha = 1/2$, weakened by $\mathscr{H}^2 \le \mathscr{D}_{1/2}$

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\hat{\rho}}\left[\mathscr{H}^2(P_{\theta},P_{\theta_0}) - \mathrm{LR}_n(\theta_0,\theta)\right] \le \frac{2\mathcal{I}(\theta;\mathcal{S})}{n}$$

No regularity on the densities p_θ, and no restriction on $\hat{\rho}$.

the same inequality covers an exact minimiser and the output of a gradient method

Tempered posteriors, and what approximates them

MINIMISE THE FIXED-PRIOR BOUND OVER $\hat{\rho}$ *Donsker & Varadhan, 1976*

$$\pi_{n,\alpha}(\mathrm{d}\theta) \;\propto\; \Big(\prod_{i=1}^{n} p_\theta(X_i)\Big)^{\alpha} \pi(\mathrm{d}\theta)$$

it contracts under a prior mass condition alone *Bhattacharya et al., 2019*

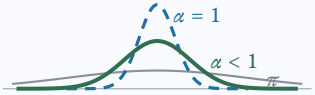
Tempered posteriors, and what approximates them

MINIMISE THE FIXED-PRIOR BOUND OVER $\hat{\rho}$ *Donsker & Varadhan, 1976*

$$\pi_{n,\alpha}(\mathrm{d}\theta) \;\propto\; \Big(\prod_{i=1}^{n} p_\theta(X_i)\Big)^{\alpha} \pi(\mathrm{d}\theta)$$

it contracts under a prior mass condition alone *Bhattacharya et al., 2019*

THE TEMPERATURE



$\alpha = 1$ out of reach: $1/(1-\alpha)$

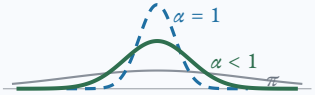
Tempered posteriors, and what approximates them

MINIMISE THE FIXED-PRIOR BOUND OVER $\hat{\rho}$ *Donsker & Varadhan, 1976*

$$\pi_{n,\alpha}(\mathrm{d}\theta) \;\propto\; \Big(\prod_{i=1}^{n} p_\theta(X_i)\Big)^{\alpha} \pi(\mathrm{d}\theta)$$

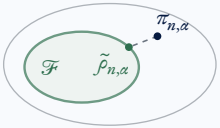
it contracts under a prior mass condition alone *Bhattacharya et al., 2019*

THE TEMPERATURE



$\alpha = 1$ out of reach: $1/(1-\alpha)$

AND THE VARIATIONAL FAMILY $\mathcal{F}$



the same minimisation, restricted to $\rho \in \mathcal{F}$

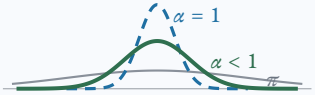
Tempered posteriors, and what approximates them

MINIMISE THE FIXED-PRIOR BOUND OVER $\hat{\rho}$ *Donsker & Varadhan, 1976*

$$\pi_{n,\alpha}(\mathrm{d}\theta) \;\propto\; \Big(\prod_{i=1}^{n} p_\theta(X_i)\Big)^{\alpha} \pi(\mathrm{d}\theta)$$

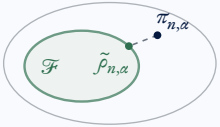
it contracts under a prior mass condition alone *Bhattacharya et al., 2019*

THE TEMPERATURE



$\alpha = 1$ out of reach: $1/(1-\alpha)$

AND THE VARIATIONAL FAMILY $\mathcal{F}$



the same minimisation, restricted to $\rho \in \mathcal{F}$

The same bound covers the tempered posterior and its variational approximation.

Localised priors and the dimension they carry

ONE CONDITION ON THE MODEL

$$\mathrm{KL}(P_{\theta_0} \| P_\theta) \le c(\alpha) \mathcal{D}_\alpha(P_\theta \| P_{\theta_0}) \qquad \text{for all } \theta \in \Theta$$

ASM. 3.3

it reverses the free direction, and we verify it on the models of the paper

Localised priors and the dimension they carry

ONE CONDITION ON THE MODEL

$$\mathrm{KL}(P_{\theta_0} \| P_\theta) \le c(\alpha) \mathcal{D}_\alpha(P_\theta \| P_{\theta_0}) \qquad \text{for all } \theta \in \Theta$$ Asm. 3.3

it reverses the free direction, and we verify it on the models of the paper

AND ONE ON THE TILT

$$\pi_{-\beta} \propto e^{-\beta \,\mathrm{KL}(P_{\theta_0} \| P_\theta)} \pi, \qquad \sup_{\beta \ge 0} \beta^{\kappa_\pi} \mathbb{E}_{\pi_{-\beta}} \big[\mathrm{KL}(P_{\theta_0} \| P_\theta) \big] = d_\pi < \infty$$ Asm. 3.4

Catoni's localisation, and $\kappa_\pi \in (0, 1]$ *Catoni, 2007*

Localised priors and the dimension they carry

ONE CONDITION ON THE MODEL

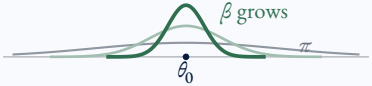
$$\mathrm{KL}(P_{\theta_0} \| P_\theta) \le c(\alpha) \mathcal{D}_\alpha(P_\theta \| P_{\theta_0}) \quad \text{for all } \theta \in \Theta$$ Asm. 3.3

it reverses the free direction, and we verify it on the models of the paper

AND ONE ON THE TILT

$$\pi_{-\beta} \propto e^{-\beta \,\mathrm{KL}(P_{\theta_0} \| P_\theta)} \pi, \quad \sup_{\beta \ge 0} \beta^{\kappa_\pi} \mathbb{E}_{\pi_{-\beta}} \big[\mathrm{KL}(P_{\theta_0} \| P_\theta) \big] = d_\pi < \infty$$ Asm. 3.4

Catoni's localisation, and $\kappa_\pi \in (0, 1]$ *Catoni, 2007*



Localised priors and the dimension they carry

ONE CONDITION ON THE MODEL

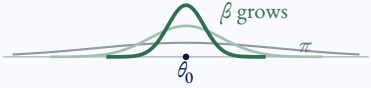
$$\mathrm{KL}(P_{\theta_0} \| P_\theta) \le c(\alpha) \mathcal{D}_\alpha(P_\theta \| P_{\theta_0}) \quad \text{for all } \theta \in \Theta$$ Asm. 3.3

it reverses the free direction, and we verify it on the models of the paper

AND ONE ON THE TILT

$$\pi_{-\beta} \propto e^{-\beta \,\mathrm{KL}(P_{\theta_0} \| P_\theta)} \pi, \qquad \sup_{\beta \ge 0} \beta^{\kappa_\pi} \mathbb{E}_{\pi_{-\beta}} \big[\mathrm{KL}(P_{\theta_0} \| P_\theta) \big] = d_\pi < \infty$$ Asm. 3.4

Catoni's localisation, and $\kappa_\pi \in (0, 1]$ *Catoni, 2007*



$\mathcal{J}(\theta; \mathcal{S}) = \inf_\pi \mathbb{E}_{\mathcal{S}}[\mathrm{KL}(\hat{\rho} \| \pi)]$ is not evaluable; evaluating at $\pi_{-\beta}$ gives $d_\pi / \beta^{\kappa_\pi}$ instead.

Rates without the logarithmic factor

Chapter 3, Theorem 3.7

trading β against n, at $\beta = n(1-\alpha)/(2c(\alpha))$

$$\mathbb{E}_S \mathbb{E}_{\theta \sim \pi_{n,\alpha}} \left[\mathrm{KL}(P_{\theta_0} \| P_\theta) \right] \;\leq\; \alpha \left(\frac{2c(\alpha)}{1-\alpha} \right)^{1+\kappa_\pi} \frac{d_\pi}{n^{\kappa_\pi}}$$

Rates without the logarithmic factor

Chapter 3, Theorem 3.7

trading β against n, at $\beta = n(1-\alpha)/(2c(\alpha))$

$$\mathbb{E}_S \mathbb{E}_{\theta \sim \pi_{n,\alpha}} \left[\mathrm{KL}(P_{\theta_0} \| P_\theta) \right] \;\leq\; \alpha \left(\frac{2c(\alpha)}{1-\alpha} \right)^{1+\kappa_\pi} \frac{d_\pi}{n^{\kappa_\pi}}$$

The same bound for $\tilde{\rho}_{n,\alpha}$, when $\mathcal{F}$ contains a competitor for $\pi_{-\beta}$.

Rates without the logarithmic factor

Chapter 3, Theorem 3.7

trading β against n, at $\beta = n(1-\alpha)/(2c(\alpha))$

$$\mathbb{E}_S \mathbb{E}_{\theta \sim \pi_{n,\alpha}} \left[\mathrm{KL}(P_{\theta_0} \| P_\theta) \right] \;\leq\; \alpha \left(\frac{2c(\alpha)}{1-\alpha} \right)^{1+\kappa_\pi} \frac{d_\pi}{n^{\kappa_\pi}}$$

The same bound for $\tilde{\rho}_{n,\alpha}$, when $\mathcal{F}$ contains a competitor for $\pi_{-\beta}$.

$$\mathcal{O}(d_\pi \log(n)/n) \quad \longrightarrow \quad \mathcal{O}(d_\pi/n)$$

at a prior fixed in advance — at the localised one, $\kappa_\pi = 1$

Bhattacharya et al., 2019; Alquier & Ridgway, 2020

Rates without the logarithmic factor

Chapter 3, Theorem 3.7

trading β against n, at $\beta = n(1-\alpha)/(2c(\alpha))$

$$\mathbb{E}_S \mathbb{E}_{\theta \sim \pi_{n,\alpha}} \left[\mathrm{KL}(P_{\theta_0} \| P_\theta) \right] \;\le\; \alpha \left(\frac{2c(\alpha)}{1-\alpha} \right)^{1+\kappa_\pi} \frac{d_\pi}{n^{\kappa_\pi}}$$

The same bound for $\tilde{\rho}_{n,\alpha}$, when $\mathcal{F}$ contains a competitor for $\pi_{-\beta}$.

$$\mathcal{O}(d_\pi \log(n)/n) \quad \longrightarrow \quad \mathcal{O}(d_\pi/n)$$

at a prior fixed in advance — at the localised one, $\kappa_\pi = 1$

Bhattacharya et al., 2019; Alquier & Ridgway, 2020

At $\kappa_\pi = 1$ the logarithm disappears; for $\kappa_\pi < 1$ the rate is $n^{-\kappa_\pi}$.

The Gaussian sequence model, at the minimax rate

Gaussian sequence model

$$\theta_0 \in S_b(L) = \left\{\theta : \sum_\ell \ell^{2b}\theta_\ell^2 \le L\right\}, \qquad b, L > 0$$

The Gaussian sequence model, at the minimax rate

Gaussian sequence model

$$\theta_0 \in S_b(L) = \left\{\theta : \sum_\ell \ell^{2b}\theta_\ell^2 \le L\right\}, \qquad b, L > 0$$

a prior with one free scale, set to $\tau_\star$

$$\pi = \bigotimes_\ell \mathcal{N}\left(0, \tau_\star^2 \ell^{-1-2b}\right), \qquad \tau_\star^2 = 12b^2L/(2b+1)$$

The Gaussian sequence model, at the minimax rate

Gaussian sequence model

$$\theta_0 \in S_b(L) = \left\{\theta : \textstyle\sum_\ell \ell^{2b}\theta_\ell^2 \le L\right\}, \qquad b, L > 0$$

a prior with one free scale, set to $\tau_\star$

$$\pi = \textstyle\bigotimes_\ell \mathcal{N}\left(0, \tau_\star^2 \ell^{-1-2b}\right), \qquad \tau_\star^2 = 12b^2L/(2b+1)$$

$$\mathbb{E}_S \mathbb{E}_{\theta\sim\pi_{n,\alpha}}\left[\mathrm{KL}(P_{\theta_0}\|P_\theta)\right] \lesssim \frac{L^{1/(2b+1)}}{(n\alpha)^{2b/(2b+1)}}$$

Cor. 3.16

The Gaussian sequence model, at the minimax rate

GAUSSIAN SEQUENCE MODEL

$$\theta_0 \in S_b(L) = \left\{\theta : \textstyle\sum_\ell \ell^{2b}\theta_\ell^2 \le L\right\}, \qquad b, L > 0$$

A PRIOR WITH ONE FREE SCALE, SET TO $\tau_\star$

$$\pi = \textstyle\bigotimes_\ell \mathcal{N}\left(0, \tau_\star^2 \ell^{-1-2b}\right), \qquad \tau_\star^2 = 12b^2L/(2b+1)$$

$$\mathbb{E}_S \mathbb{E}_{\theta\sim\pi_{n,\alpha}}\left[\mathrm{KL}(P_{\theta_0}\|P_\theta)\right] \lesssim \frac{L^{1/(2b+1)}}{(n\alpha)^{2b/(2b+1)}}$$ COR. 3.16

Minimax over $S_b(L)$ in both exponents: the radius, and $n\alpha$ for the sample size.

in the exponents alone *Giné & Nickl, 2016*

The Gaussian sequence model, at the minimax rate

GAUSSIAN SEQUENCE MODEL

$$\theta_0 \in S_b(L) = \left\{\theta : \textstyle\sum_\ell \ell^{2b}\theta_\ell^2 \le L\right\}, \qquad b, L > 0$$

A PRIOR WITH ONE FREE SCALE, SET TO $\tau_\star$

$$\pi = \textstyle\bigotimes_\ell \mathcal{N}\left(0, \tau_\star^2 \ell^{-1-2b}\right), \qquad \tau_\star^2 = 12b^2L/(2b+1)$$

$$\mathbb{E}_S \mathbb{E}_{\theta\sim\pi_{n,\alpha}}\left[\mathrm{KL}(P_{\theta_0}\|P_\theta)\right] \lesssim \frac{L^{1/(2b+1)}}{(n\alpha)^{2b/(2b+1)}} \qquad \text{COR. 3.16}$$

Minimax over $S_b(L)$ in both exponents: the radius, and $n\alpha$ for the sample size.

in the exponents alone *Giné & Nickl, 2016*

We recover the classical minimax rate, at the effective sample size $n\alpha$. *L'Huillier et al., 2023*

When the model is wrong

THE BENCHMARK THE BOUND COMPARES AGAINST



When the model is wrong

THE BENCHMARK THE BOUND COMPARES AGAINST



CHAPTER 3, THEOREM 3.7

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

When the model is wrong

THE BENCHMARK THE BOUND COMPARES AGAINST



CHAPTER 3, THEOREM 3.7

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

The benchmark is a KL projection: infinite wherever p_θ vanishes and the truth does not.

Part 2 in one page

what Chapter 3 establishes

- a bound in the mutual information $\mathcal{I}(\theta; \mathcal{S})$, with the prior optimised away Thm. 3.2
- and, from it, the sharp rate d_π / n^{κ_π}, minimax over a Sobolev ball Thm. 3.7
- the same bound for the maximum likelihood estimator Thm. 3.27, Sec. 3.4

Part 2 in one page

what Chapter 3 establishes

- a bound in the mutual information $\mathcal{I}(\theta; \mathcal{S})$, with the prior optimised away Thm. 3.2
- and, from it, the sharp rate d_π / n^{κ_π}, minimax over a Sobolev ball Thm. 3.7
- the same bound for the maximum likelihood estimator Thm. 3.27, Sec. 3.4

The same bound certifies the tempered posterior and the variational approximation, at the same rate and with the same constants.

Part 2 in one page

what Chapter 3 establishes

- a bound in the mutual information $\mathcal{I}(\theta; \mathcal{S})$, with the prior optimised away Thm. 3.2
- and, from it, the sharp rate d_π / n^{κ_π}, minimax over a Sobolev ball Thm. 3.7
- the same bound for the maximum likelihood estimator Thm. 3.27, Sec. 3.4

The same bound certifies the tempered posterior and the variational approximation, at the same rate and with the same constants.

in progress

other information-theoretic tools, such as the conditional mutual information

Steinke & Zakynthinou, 2020



Part 3 · Chapter 4

Robust generalised posteriors



$$\mathscr{H}^2(P_*, P_\theta)$$

Outline

Part 1

Bias

Part 2

Information

Part 3



Robustness

1. Hellinger in place of the log-likelihood
2. Oracle inequalities under contamination
3. A saddle point we can solve

Misspecification, and why the likelihood fails

FITTED WITH THE UNIFORMS $\mathcal{U}([0,\theta])$, $\theta \in [0,1)$ *Birgé, 2006*

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



Misspecification, and why the likelihood fails

FITTED WITH THE UNIFORMS $\mathcal{U}([0,\theta])$, $\theta \in [0,1)$ *Birgé, 2006*

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



$$\mathscr{H}^2(P_*, P_{1/10}) < 5/(4n)$$

$$\mathrm{E}\big[\mathscr{H}^2(P_*, P_{\hat{\theta}})\big] > 0.38$$

$\hat{\theta} = \max_i X_i$ is the maximum likelihood estimator: any smaller θ has likelihood zero, and $\mathrm{KL}(P_* \| P_{1/10}) = +\infty$

Misspecification, and why the likelihood fails

FITTED WITH THE UNIFORMS $\mathcal{U}([0,\theta])$, $\theta \in [0,1)$ *Birgé, 2006*

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



$\mathscr{H}^2(P_*, P_{1/10}) < 5/(4n)$

$\mathrm{E}\big[\mathscr{H}^2(P_*, P_{\hat{\theta}})\big] > 0.38$

$\hat{\theta} = \max_i X_i$ is the maximum likelihood estimator: any smaller θ has likelihood zero, and $\mathrm{KL}(P_* \| P_{1/10}) = +\infty$

The model contains a near-perfect member, and maximum likelihood misses it. Bayes commits to the same KL; $\mathscr{H}^2$ does not.

Generalised Bayes with a Hellinger loss

GENERALISED BAYES: THE LOSS DEFINES THE POSTERIOR, AND $-\log p_\theta$ GIVES THE USUAL ONE

$$\pi_n^\ell(\theta) \;\propto\; \pi(\theta)\exp\bigl(-\beta n\, R_n(\theta)\bigr) \qquad \textit{Bissiri et al., 2016}$$

Generalised Bayes with a Hellinger loss

GENERALISED BAYES: THE LOSS DEFINES THE POSTERIOR, AND $-\log p_\theta$ GIVES THE USUAL ONE

$$\pi_n^{\ell}(\theta) \;\propto\; \pi(\theta)\exp\bigl(-\beta n\, R_n(\theta)\bigr) \qquad \textit{Bissiri et al., 2016}$$

Part 2 moved α; here we move ℓ. Many losses have been proposed for robustness, and the criterion is that the loss be bounded and continuous in total variation *Jewson et al., 2018*.

Generalised Bayes with a Hellinger loss

GENERALISED BAYES: THE LOSS DEFINES THE POSTERIOR, AND $-\log p_\theta$ GIVES THE USUAL ONE

$$\pi_n^{\ell}(\theta) \;\propto\; \pi(\theta)\exp\!\big(-\beta n\, R_n(\theta)\big) \qquad \textit{Bissiri et al., 2016}$$

Part 2 moved α; here we move ℓ. Many losses have been proposed for robustness, and the criterion is that the loss be bounded and continuous in total variation *Jewson et al., 2018*.

$\mathscr{H}^2$ HAS BOTH; WHAT IT BUYS UNDER CONTAMINATION

$$\mathscr{H}^2\big((1-\varepsilon)P_{\theta_0} + \varepsilon Q,\; P_{\theta_0}\big) \;\le\; \varepsilon \qquad \textit{Huber, 1964}$$

uniformly in Q: no moment, no tail, no support condition on the contaminating part

Generalised Bayes with a Hellinger loss

GENERALISED BAYES: THE LOSS DEFINES THE POSTERIOR, AND $-\log p_\theta$ GIVES THE USUAL ONE

$$\pi_n^{\ell}(\theta) \;\propto\; \pi(\theta)\exp\!\big(-\beta n\, R_n(\theta)\big) \qquad \textit{Bissiri et al., 2016}$$

Part 2 moved α; here we move ℓ. Many losses have been proposed for robustness, and the criterion is that the loss be bounded and continuous in total variation *Jewson et al., 2018*.

$\mathscr{H}^2$ HAS BOTH; WHAT IT BUYS UNDER CONTAMINATION

$$\mathscr{H}^2\big((1-\varepsilon)P_{\theta_0} + \varepsilon Q,\; P_{\theta_0}\big) \;\le\; \varepsilon \qquad \textit{Huber, 1964}$$

uniformly in Q: no moment, no tail, no support condition on the contaminating part

$\mathscr{H}^2$ is bounded. We now need an empirical criterion that estimates it and is bounded too.

A contrast no observation can dominate

A BOUNDED CONTRAST IN PLACE OF THE LOG-LIKELIHOOD *Baraud, Birgé & Sart, 2017*

$$\widehat{R}_\psi(\theta,\theta') = n^{-1}\sum_i \psi\big(p_{\theta'}(X_i)/p_\theta(X_i)\big), \qquad \psi(x) = \frac{\sqrt{x}-1}{\sqrt{x}+1}$$

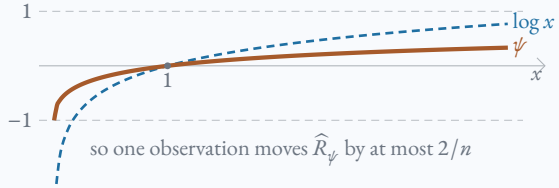
$X_1, \dots, X_n$ drawn independently from $P_\star$, in or out of the model

A contrast no observation can dominate

A BOUNDED CONTRAST IN PLACE OF THE LOG-LIKELIHOOD *Baraud, Birgé & Sart, 2017*

$$\widehat{R}_\psi(\theta,\theta') = n^{-1}\sum_i \psi\big(p_{\theta'}(X_i)/p_\theta(X_i)\big), \qquad \psi(x) = \frac{\sqrt{x}-1}{\sqrt{x}+1}$$

$X_1, \dots, X_n$ drawn independently from $P_\star$, in or out of the model



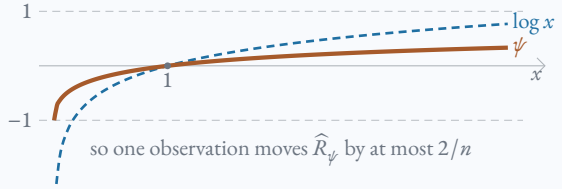
so one observation moves $\widehat{R}_\psi$ by at most $2/n$

A contrast no observation can dominate

A BOUNDED CONTRAST IN PLACE OF THE LOG-LIKELIHOOD *Baraud, Birgé & Sart, 2017*

$$\widehat{R}_\psi(\theta,\theta') = n^{-1}\sum_i \psi\big(p_{\theta'}(X_i)/p_\theta(X_i)\big), \qquad \psi(x) = \frac{\sqrt{x}-1}{\sqrt{x}+1}$$

$X_1, \dots, X_n$ drawn independently from P_*, in or out of the model



IN EXPECTATION, $R_\psi = \mathrm{E}_{\mathcal{S}}[\widehat{R}_\psi]$: HELLINGER RISK ON BOTH SIDES

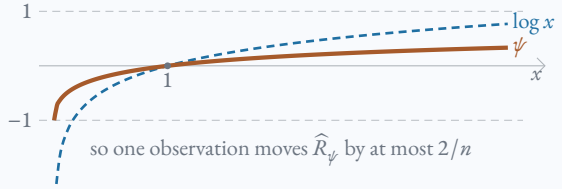
$$\tfrac{3}{8}\mathscr{H}^2(P_*,P_\theta) - 4\mathscr{H}^2(P_*,P_{\theta'}) \;\le\; R_\psi(\theta,\theta') \;\le\; 4\mathscr{H}^2(P_*,P_\theta) - \tfrac{3}{8}\mathscr{H}^2(P_*,P_{\theta'})$$

A contrast no observation can dominate

A BOUNDED CONTRAST IN PLACE OF THE LOG-LIKELIHOOD *Baraud, Birgé & Sart, 2017*

$$\widehat{R}_\psi(\theta,\theta') = n^{-1}\sum_i \psi\big(p_{\theta'}(X_i)/p_\theta(X_i)\big), \qquad \psi(x) = \frac{\sqrt{x}-1}{\sqrt{x}+1}$$

$X_1, \dots, X_n$ drawn independently from P_*, in or out of the model



IN EXPECTATION, $R_\psi = \mathrm{E}_\mathcal{S}[\widehat{R}_\psi]$: HELLINGER RISK ON BOTH SIDES

$$\tfrac{3}{8}\mathscr{H}^2(P_*,P_\theta) - 4\mathscr{H}^2(P_*,P_{\theta'}) \;\le\; R_\psi(\theta,\theta') \;\le\; 4\mathscr{H}^2(P_*,P_\theta) - \tfrac{3}{8}\mathscr{H}^2(P_*,P_{\theta'})$$

Minimising the worst case over θ' gives the ρ-estimator; a posterior needs that criterion at every θ.

From a worst case to a soft maximum

THE ρ-POSTERIOR, AT TEMPERATURE $\lambda = \beta n$

$$\pi_n^\rho(\mathrm{d}\theta) \;\propto\; \exp\!\big(-\lambda \sup_{\theta' \in \Theta} \widehat{R}_\psi(\theta, \theta')\big)\, \pi(\mathrm{d}\theta)$$

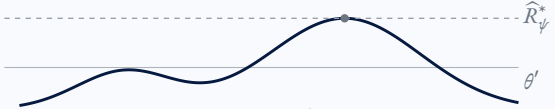
minimax rates, and Bernstein–von Mises under outliers *Baraud et al., 2017; Baraud & Birgé, 2020*

From a worst case to a soft maximum

THE ρ-POSTERIOR, AT TEMPERATURE $\lambda = \beta n$

$$\pi_n^{\rho}(\mathrm{d}\theta) \;\propto\; \exp\!\left(-\lambda \sup_{\theta' \in \Theta} \widehat{R}_{\psi}(\theta, \theta')\right) \pi(\mathrm{d}\theta)$$

minimax rates, and Bernstein–von Mises under outliers *Baraud et al., 2017; Baraud & Birgé, 2020*



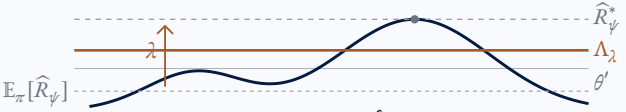
Every density evaluation carries a supremum over Θ, so π_n^{ρ} can be neither sampled nor optimised in practice.

From a worst case to a soft maximum

THE ρ-POSTERIOR, AT TEMPERATURE $\lambda = \beta n$

$$\pi_n^{\rho}(\mathrm{d}\theta) \;\propto\; \exp\!\big(-\lambda \sup_{\theta' \in \Theta} \widehat{R}_{\psi}(\theta, \theta')\big)\, \pi(\mathrm{d}\theta)$$

minimax rates, and Bernstein–von Mises under outliers *Baraud et al., 2017; Baraud & Birgé, 2020*



Every density evaluation carries a supremum over Θ, so π_n^{ρ} can be neither sampled nor optimised in practice.

CHAPTER 4, THEOREM 4.5

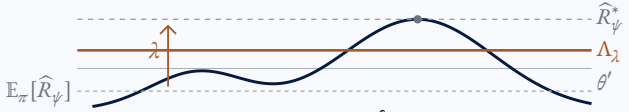
$$\Lambda_\lambda(\theta;\pi) = \frac{1}{\lambda}\log\int_{\Theta} e^{\lambda \widehat{R}_{\psi}(\theta,\theta')}\pi(d\theta') \;\nearrow\; \widehat{R}^*_{\psi}(\theta)$$

From a worst case to a soft maximum

THE ρ-POSTERIOR, AT TEMPERATURE $\lambda = \beta n$

$$\pi_n^{\rho}(\mathrm{d}\theta) \;\propto\; \exp\!\big(-\lambda \sup_{\theta' \in \Theta} \widehat{R}_{\psi}(\theta, \theta')\big)\, \pi(\mathrm{d}\theta)$$

minimax rates, and Bernstein–von Mises under outliers *Baraud et al., 2017; Baraud & Birgé, 2020*



Every density evaluation carries a supremum over Θ, so π_n^{ρ} can be neither sampled nor optimised in practice.

CHAPTER 4, THEOREM 4.5

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

Replacing the supremum by a soft maximum returns a Gibbs posterior.

An oracle inequality, and what contamination costs

Chapter 4, Theorem 4.10

with probability $1 - 2\delta$, at $\lambda = n/8$

$$\mathbb{E}_{\theta \sim \hat{\rho}_\lambda}\left[\mathscr{H}^2(P_*, P_\theta)\right] \;\le\; \inf_\rho \left\{ 104\, \mathbb{E}_{\theta\sim\rho}\left[\mathscr{H}^2(P_*, P_\theta)\right] + 384\, \mathrm{KL}(\rho\|\pi)/n \right\} + 192 \log(1/\delta)/n$$

ρ ranges over $\mathscr{P}(\Theta)$: no entropy and no covering condition, any prior, and absolute constants

An oracle inequality, and what contamination costs

Chapter 4, Theorem 4.10

with probability $1 - 2\delta$, at $\lambda = n/8$

$$\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}\big[\mathscr{H}^2(P_*, P_\theta)\big] \;\le\; \inf_\rho\Big\{104\,\mathbb{E}_{\theta\sim\rho}\big[\mathscr{H}^2(P_*, P_\theta)\big] + 384\,\mathrm{KL}(\rho\|\pi)/n\Big\} + 192\log(1/\delta)/n$$

ρ ranges over $\mathscr{P}(\Theta)$: no entropy and no covering condition, any prior, and absolute constants

UNDER A PRIOR MASS CONDITION

$$\pi\Big(\big\{\theta : \mathscr{H}^2(P_*, P_\theta) \le \mathscr{H}^2(P_*, P_{\theta_0}) + r\big\}\Big) \;\ge\; (r/c)^{d_\pi} \qquad \text{for every } r \le r_0$$

$$\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}\big[\mathscr{H}^2(P_*, P_\theta)\big] \;\le\; 104\,\mathscr{H}^2(P_*, P_{\theta_0}) + \mathcal{O}\big(d_\pi \log(n)/n\big)$$ Cor. 4.14

$\theta_0 \in \operatorname{argmin}_\theta \mathscr{H}^2(P_*, P_\theta)$, and d_π the Catoni dimension at π: minimax up to the $\log n$

An oracle inequality, and what contamination costs

Chapter 4, Theorem 4.10

with probability $1 - 2\delta$, at $\lambda = n/8$

$$\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}\big[\mathscr{H}^2(P_*, P_\theta)\big] \;\le\; \inf_\rho\Big\{104\mathbb{E}_{\theta\sim\rho}\big[\mathscr{H}^2(P_*, P_\theta)\big] + 384\,\mathrm{KL}(\rho\|\pi)/n\Big\} + 192\log(1/\delta)/n$$

ρ ranges over $\mathscr{P}(\Theta)$: no entropy and no covering condition, any prior, and absolute constants

UNDER A PRIOR MASS CONDITION

$$\pi\Big(\big\{\theta : \mathscr{H}^2(P_*, P_\theta) \le \mathscr{H}^2(P_*, P_{\theta_0}) + r\big\}\Big) \;\ge\; (r/c)^{d_\pi} \qquad \text{for every } r \le r_0$$

$$\mathbb{E}_{\theta\sim\hat{\rho}_\lambda}\big[\mathscr{H}^2(P_*, P_\theta)\big] \;\le\; 104\mathscr{H}^2(P_*, P_{\theta_0}) + \mathcal{O}\big(d_\pi \log(n)/n\big)$$ Cor. 4.14

$\theta_0 \in \operatorname{argmin}_\theta \mathscr{H}^2(P_*, P_\theta)$, and d_π the Catoni dimension at π: minimax up to the $\log n$

WHAT CONTAMINATION COSTS

$$P_* = (1-\varepsilon)P_{\theta_0} + \varepsilon Q \implies \mathbb{E}_{\theta\sim\hat{\rho}_\lambda}\big[\mathscr{H}^2(P_*, P_\theta)\big] \;\le\; 104\varepsilon + \mathcal{O}\big(d_\pi \log(n)/n\big)$$

An oracle inequality, and what contamination costs

Chapter 4, Theorem 4.10

with probability $1-2\delta$, at $\lambda = n/8$

$$\mathbb{E}_{\theta\sim\hat{p}_\lambda}\big[\mathscr{H}^2(P_*,P_\theta)\big] \;\le\; \inf_\rho\Big\{104\,\mathbb{E}_{\theta\sim\rho}\big[\mathscr{H}^2(P_*,P_\theta)\big] + 384\,\mathrm{KL}(\rho\|\pi)/n\Big\} + 192\log(1/\delta)/n$$

ρ ranges over $\mathscr{P}(\Theta)$: no entropy and no covering condition, any prior, and absolute constants

UNDER A PRIOR MASS CONDITION

$$\pi\Big(\big\{\theta : \mathscr{H}^2(P_*,P_\theta) \le \mathscr{H}^2(P_*,P_{\theta_0}) + r\big\}\Big) \;\ge\; (r/c)^{d_\pi} \qquad \text{for every } r \le r_0$$

$$\mathbb{E}_{\theta\sim\hat{p}_\lambda}\big[\mathscr{H}^2(P_*,P_\theta)\big] \;\le\; 104\,\mathscr{H}^2(P_*,P_{\theta_0}) + \mathcal{O}\big(d_\pi\log(n)/n\big) \qquad \text{Cor. 4.14}$$

$\theta_0 \in \operatorname{argmin}_\theta \mathscr{H}^2(P_*,P_\theta)$, and d_π the Catoni dimension at π: minimax up to the $\log n$

WHAT CONTAMINATION COSTS

$$P_* = (1-\varepsilon)P_{\theta_0} + \varepsilon Q \;\Longrightarrow\; \mathbb{E}_{\theta\sim\hat{p}_\lambda}\big[\mathscr{H}^2(P_*,P_\theta)\big] \;\le\; 104\varepsilon + \mathcal{O}\big(d_\pi\log(n)/n\big)$$

The bound uses $\mathscr{H}^2$ alone, and contamination enters additively, at order ε.

Certifying a variational approximation

VARIATIONAL INFERENCE, WHERE THERE WAS NONE

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

DEF. 4.22

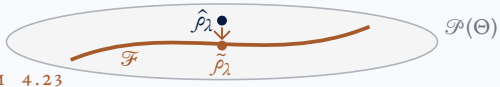


Certifying a variational approximation

VARIATIONAL INFERENCE, WHERE THERE WAS NONE

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

DEF. 4.22



CHAPTER 4, THEOREM 4.23

$$\mathbb{E}_{\theta\sim\tilde{\rho}_\lambda}\left[\mathscr{H}^2(P_*,P_\theta)\right] \;\le\; \inf_{\rho\in\mathcal{F}}\left\{104\,\mathbb{E}_{\theta\sim\rho}\left[\mathscr{H}^2(P_*,P_\theta)\right] + \frac{384\,\mathrm{KL}(\rho\|\pi)}{n}\right\} + \frac{192\log(1/\delta)}{n}$$

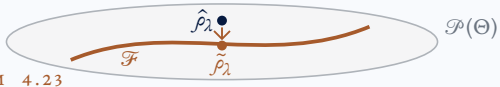
the constants of the exact $\hat{\rho}_\lambda$, unchanged: $\mathcal{F}$ pays only for what it cannot reach

Certifying a variational approximation

VARIATIONAL INFERENCE, WHERE THERE WAS NONE

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

DEF. 4.22



CHAPTER 4, THEOREM 4.23

$$\mathbb{E}_{\theta\sim\widetilde{\rho}_\lambda}\left[\mathscr{H}^2(P_*,P_\theta)\right] \;\le\; \inf_{\rho\in\mathcal{F}}\left\{104\,\mathbb{E}_{\theta\sim\rho}\left[\mathscr{H}^2(P_*,P_\theta)\right] + \frac{384\,\mathrm{KL}(\rho\|\pi)}{n}\right\} + \frac{192\log(1/\delta)}{n}$$

the constants of the exact $\widehat{\rho}_\lambda$, unchanged: $\mathcal{F}$ pays only for what it cannot reach

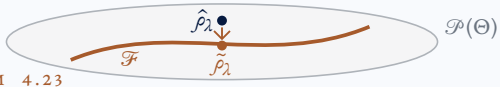
FOR A RATE, AND ONLY FOR A RATE

$$\exists\,\rho_n \in \mathcal{F} \text{ with } \mathbb{E}_{\rho_n}\left[\mathscr{H}^2(P_*,P_\theta)\right] \le 1/n \text{ and } \mathrm{KL}(\rho_n\|\pi) \le d_\pi \log n$$

Certifying a variational approximation

VARIATIONAL INFERENCE, WHERE THERE WAS NONE

$$\Lambda_\lambda^{\mathcal{F}}(\theta;\pi) = \sup_{\rho' \in \mathcal{F}} \left\{ \mathbb{E}_{\theta' \sim \rho'} \left[\widehat{R}_\psi(\theta, \theta') \right] - \mathrm{KL}(\rho' \| \pi)/\lambda \right\} \;\le\; \Lambda_\lambda(\theta;\pi) \qquad \text{Def. 4.22}$$



Chapter 4, Theorem 4.23

$$\mathbb{E}_{\theta \sim \tilde{\rho}_\lambda} \left[\mathscr{H}^2(P_*, P_\theta) \right] \;\le\; \inf_{\rho \in \mathcal{F}} \left\{ 104 \mathbb{E}_{\theta \sim \rho} \left[\mathscr{H}^2(P_*, P_\theta) \right] + \frac{384\, \mathrm{KL}(\rho \| \pi)}{n} \right\} + \frac{192 \log(1/\delta)}{n}$$

the constants of the exact $\hat{\rho}_\lambda$, unchanged: $\mathcal{F}$ pays only for what it cannot reach

FOR A RATE, AND ONLY FOR A RATE

$$\exists\, \rho_n \in \mathcal{F} \text{ with } \mathbb{E}_{\rho_n} \left[\mathscr{H}^2(P_*, P_\theta) \right] \le 1/n \text{ and } \mathrm{KL}(\rho_n \| \pi) \le d_\pi \log n$$

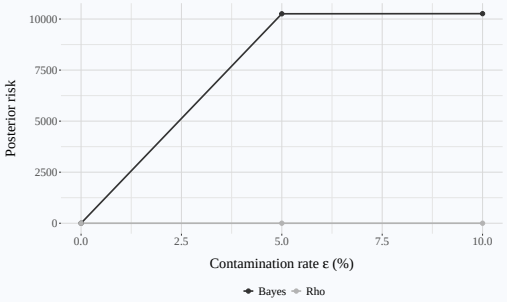
The variational minimiser solves a saddle point, which we solve by stochastic extragradient.

Contamination, in simulation

POSTERIOR RISK OVER 1000 REPLICATIONS, AGAINST ε

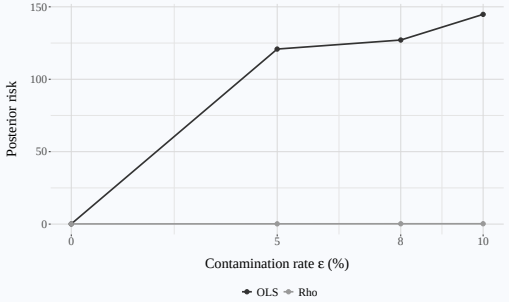
$P_\theta = \mathcal{U}([0, \theta])$, $\theta > 0$, and $\theta_0 = 1$

$P_* = (1 - \varepsilon)\mathcal{U}([0, 1]) + \varepsilon\mathcal{U}([101, 102])$, $n = 200$



$Y_i = \langle W_i, \beta^\star \rangle + \xi_i$, $\beta^\star \in \mathbb{R}^{10}$ five-sparse

$\Sigma_{jk} = 0.7^{|j-k|}$, ξ_i Pareto-contaminated, $n = 100$

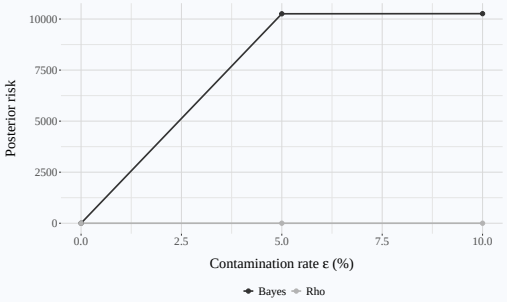


Contamination, in simulation

POSTERIOR RISK OVER 1000 REPLICATIONS, AGAINST ε

$P_\theta = \mathcal{U}([0, \theta])$, $\theta > 0$, and $\theta_0 = 1$

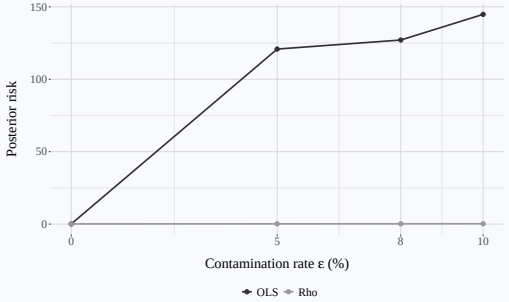
$P_* = (1 - \varepsilon)\mathcal{U}([0, 1]) + \varepsilon\mathcal{U}([101, 102])$, $n = 200$



$\hat{\theta}_{\text{MLE}} = \max_i X_i$, at ≈ 102

$Y_i = \langle W_i, \beta^\star \rangle + \xi_i$, $\beta^\star \in \mathbb{R}^{10}$ five-sparse

$\Sigma_{jk} = 0.7^{|j-k|}$, ξ_i Pareto-contaminated, $n = 100$



least squares passes 150, while the posterior stays under 4

Part 3 in one page

GUARANTEES THE ρ-POSTERIOR DID NOT HAVE

- the same oracle inequality for the approximation one actually computes Thm. 4.23
- fixed-design regression under a misspecified noise density, the errors adding Thm. 4.17
- a saddle point of condition number $\mathcal{O}(1)$, solved at a rate free of n Thm. 4.33

None of the three is available for the ρ-posterior as originally defined, whose criterion cannot be evaluated.

Part 3 in one page

GUARANTEES THE ρ-POSTERIOR DID NOT HAVE

- the same oracle inequality for the approximation one actually computes Thm. 4.23
- fixed-design regression under a misspecified noise density, the errors adding Thm. 4.17
- a saddle point of condition number $\mathcal{O}(1)$, solved at a rate free of n Thm. 4.33

None of the three is available for the ρ-posterior as originally defined, whose criterion cannot be evaluated.

IN PROGRESS

tighter relaxations of the supremum, to remove the $\log n$ and to allow localised priors

شكرًا | Thank you

Related publications

- On importance sampling and independent Metropolis–Hastings with an unbounded weight function
  Deligiannidis, Jacob, Khribch & Wang, *Annals of Statistics*, accepted
- Convergence of Statistical Estimators via Mutual Information Bounds
  Khribch & Alquier, submitted
- Robust Bayesian Inference via Variational Approximations of Generalized ρ-Posteriors
  Khribch & Alquier, submitted

