NSF@75: Advancing Statistical Science for a Data-Driven World

July 16th, 2025

Importance Sampling & MCMC

Presenter: Quan Zhou

Department of Statistics, Texas A&M University

Acknowledgment

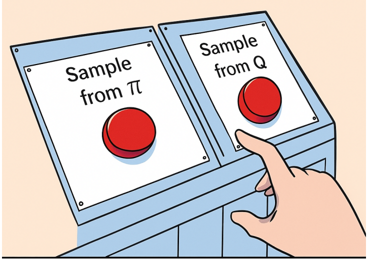
This talk is based on my recent works co-authored with

- Guanxun Li, Beijing Normal University at Zhuhai, China
- Hyunwoong (Woody) Chang, University of Texas at Dallas
- Aaron Smith, University of Ottawa, Canada

The research presented in this talk is supported by NSF DMS-2245591, DMS-2311307.

Questions to be Addressed

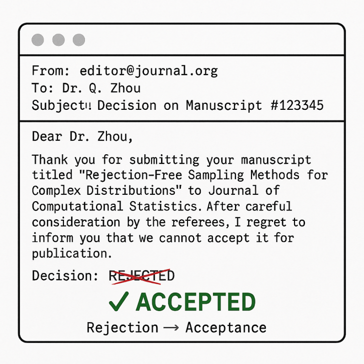
Suppose we want to approximate a distribution Π, and we can sample from either Π or another distribution Q. Which to choose?



By ChatGPT.

Questions to be Addressed

For various Metropolis–Hastings schemes, can we skip the rejection step and always accept the proposal?



By ChatGPT.

Importance Sampling

Π : target probability distribution; Q : trial probability distribution.

$$\int f \mathrm{d}\Pi = \int \left(f \frac{\mathrm{d}\Pi}{\mathrm{d}Q} \right) \mathrm{d}Q.$$

Define $w = \mathrm{d}\Pi / \mathrm{d}Q$.

Estimating the expectation of f with samples from Π
$\Longrightarrow$ estimating the expectation of fw with samples from Q

Importance Sampling Estimators

Let $X_i \sim Q$. Importance sampling estimator:

$$\widehat{\Pi}_{Q,n}(f) := \frac{1}{n} \sum_{i=1}^{n} f(X_i) w(X_i).$$

Self-normalized importance sampling estimator:

$$\widetilde{\Pi}_{Q,n}(f) := \frac{\sum_{i=1}^{n} f(X_i) w(X_i)}{\sum_{i=1}^{n} w(X_i)}.$$

w only needs to be evaluated up to a normalizing constant.

Variances of Importance Sampling Estimators

Let f be centered, i.e., $\int f \mathrm{d}\Pi = 0$. Then,

$$\begin{aligned}\sigma^2(Q, f) &:= \lim_{n\to\infty} n\mathrm{Var}\left(\widetilde{\Pi}_{Q,n}(f)\right) \\ &= n\mathrm{Var}\left(\widehat{\Pi}_{Q,n}(f)\right) \\ &= \int f^2 w \,\mathrm{d}\Pi.\end{aligned}$$

What is the optimal choice of Q?

Variances of Importance Sampling Estimators

For a fixed, centered f, the optimal Q minimizing $\sigma^2(Q, f)$ satisfies

$$\frac{\mathrm{d}Q}{\mathrm{d}\Pi}(x) \propto |f(x)|.$$

Unless f is constant, there exists some Q such that importance sampling is more efficient than direct sampling from Π.

What if f is not fixed? *Then maybe it is optimal to sample from Π?*

Minimax Optimal Trial Distribution

Define the "maximum risk" of Q by

$$R(Q) = \sup_{f:\ \int f \mathrm{d}\Pi = 0, \int f^2 \mathrm{d}\Pi = 1} \sigma^2(Q, f).$$

So $R(\Pi) = 1$.

We say Q^* is minimax optimal if

$$R(Q^*) = \inf_Q R(Q).$$

Minimax Optimal Trial Distribution

Theorem

Π is minimax optimal if and only if Π does not have an atom with probability mass > 0.5.

Theorem

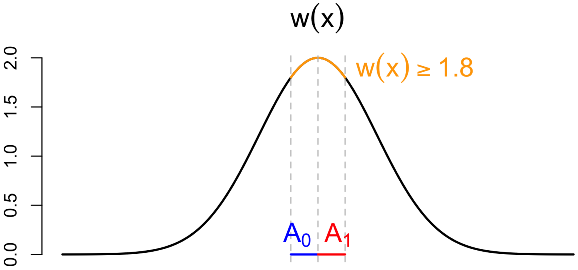
If $\Pi(\{x^*\}) = p > 0.5$, then the minimax optimal Q^* is given by

$$Q^*(\{x^*\}) = \frac{1}{2}, \text{ and } \frac{\mathrm{d}\Pi}{\mathrm{d}Q^*}(x) = 2(1-p) \text{ for } x \neq x^*.$$

(So x^* receives largest importance weight equal to $2p$.) Further,

$$R(Q^*) = 4p(1-p).$$

How to construct the worst test function?



$f(x) = c\mathbb{1}_{A_0}(x) - c\mathbb{1}_{A_1}(x)$ where c is s.t. $\int f^2 \mathrm{d}\Pi = 1$.

Then $\sigma^2(Q, f) = \int f^2 w \, \mathrm{d}\Pi \geq 1.8$.

Key Takeaways

Suppose Π is concentrated on a small set A. As long as f does not vary wildly over A, it is probably better to assign larger importance weights to states in A and smaller weights to those outside.

Of course, in most applications, we don't know where A is. Further, i.i.d. sampling is often not feasible.

A practical solution: let Q have density $q(x) \propto \pi(x)^{\beta}$ for some $\beta \in (0, 1)$ and use MCMC to draw samples from Q.

Markov Chain Importance Sampling

Let $(X_i)_{i\geq 1}$ be a Markov chain with stationary density $q(x) \propto \pi(x)^\beta$. We can still use the self-normalized importance sampling estimator:

$$\widetilde{\Pi}_{Q,n}(f) \coloneqq \frac{\sum_{i=1}^n f(X_i)w(X_i)}{\sum_{i=1}^n w(X_i)},$$

where $w(x) \propto \pi(x)^{1-\beta}$.

We call this scheme *importance-tempered MCMC* [3, 10].

Setup for Theoretical Analysis

$$\widetilde{\Pi}_{Q,n}(f) := \frac{\sum_{i=1}^{n} f(X_i) w(X_i)}{\sum_{i=1}^{n} w(X_i)},$$

If we view $w(X_i)$ as the *time* the chain stays at X_i, then $\widetilde{\Pi}_{Q,n}(f)$ becomes a simple time average of a continuous-time process.

If we further replace each $w(X_i)$ with an exponential random variable with mean $w(X_i)$, this continuous-time process becomes a *continuous-time Markov chain* with generator

$$(\mathscr{A} g)(x) = \frac{1}{w(x)} \int_{\mathcal{X}} [g(y) - g(x)] \, \mathcal{T}(x, \mathrm{d}y),$$

where $\mathcal{T}$ is the transition kernel of the discrete-time Markov chain $(X_i)_{i \geq 1}$.

Uniform and Geometric Ergodicity

Definition

We say a Markov process $(Y_t)_{t\geq 0}$ with state space $\mathcal{X}$ and invariant distribution Π is geometrically ergodic, if for each $x \in \mathcal{X}$, there exist constants $C(x) < \infty$ and $\theta \in (0,1)$ such that

$$d_{\mathrm{TV}}(\mathrm{Law}(Y_t \mid Y_0 = x),\, \Pi) \leq C(x)\theta^t, \quad \forall\, t > 0,$$

where d_{TV} denotes the total variation distance.
If $\sup_{x\in\mathcal{X}} C(x) < \infty$, we say $(Y_t)_{t\geq 0}$ is uniformly ergodic.

Ergodicity of Metropolis–Hastings Algorithms

Let Π be a positive continuous distribution on $\mathbb{R}$. For any random walk Metropolis–Hastings algorithm with a "local" proposal scheme, it is well known that [6]

1. it cannot be uniformly ergodic;
2. it is geometrically ergdoic if and only if Π has sub-exponential tails.

Ergodicity of Importance-tempered Metropolis–Hastings

Consider our importance-tempered MCMC scheme with $(X_i)_{i\geq 1}$ generated from a random walk Metropolis–Hastings algorithm targeting π^β. Let $(Y_t)_{t\geq 0}$ denote the corresponding continuous-time Markov chain.

Theorem

$(Y_t)_{t\geq 0}$ is uniformly ergodic if Π has sub-exponential tails.

Ergodicity of Importance-tempered Metropolis–Hastings

Theorem

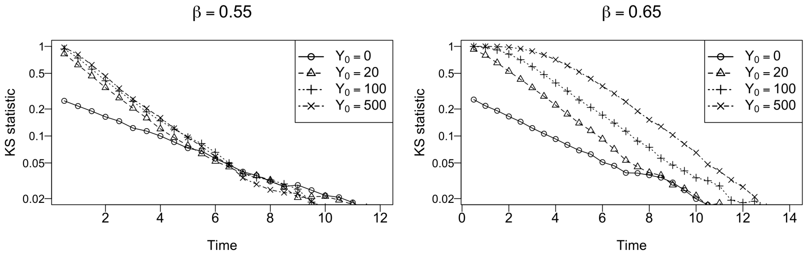
Let $\gamma > 1$ *and* Π *have density*

$$\pi(x) = \frac{\gamma - 1}{2}(1 + |x|)^{-\gamma}, \quad \forall\, x \in \mathbb{R},$$

Then $(Y_t)_{t \geq 0}$ *is uniformly ergodic if and only if*

$$\frac{1}{\gamma} < \beta < \frac{\gamma - 2}{\gamma}.$$

Numerical Illustration



Simulation of the continuous-time Markov chain $(Y_t)_{t\geq 0}$ with Π being t_4. The Kolmogorov–Smirnov test statistic compares t_4 with the distribution of Y_t over 10^4 replicates. According to our theory, $(Y_t)_{t\geq 0}$ is uniformly ergodic if and only if $0.2 < \beta < 0.6$.

No Warm-up Iterations Needed

HEALTH & FITNESS

Don't Warm Up? You're Going to Get Injured

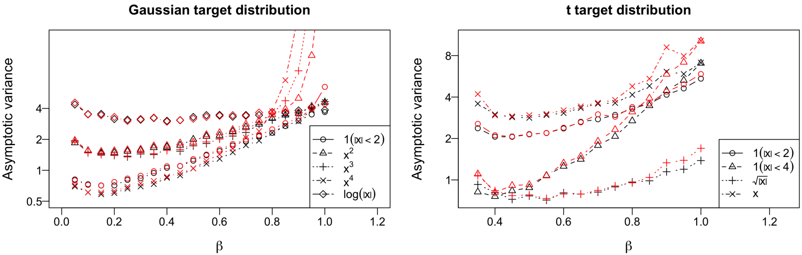
A cold muscle is a muscle at risk.

Laura Williams • Dec 4, 2017 6:47 PM EST

Odilon Dimier/Getty Images

From `mensjournal.com`

Numerical Illustration

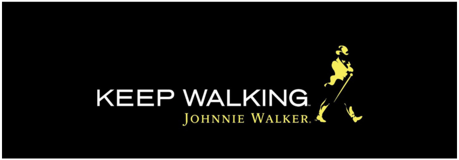


Simulation of the importance-tempered Metropolis–Hastings algorithm with initial value $X_0 \approx 0$ (black) or $X_0 = 10$ (red). Asymptotic variance is estimated over $2,000$ replicates and scaled by $\sigma^2(\Pi, f)$.

Making Metropolis–Hastings Rejection-free

Let $\mathcal{K}$ denote the transition kernel of the proposal scheme of a Metropolis–Hastings Algorithms. If $\mathcal{K}$ has a stationary distribution Q, then we can simply run $\mathcal{K}$ (i.e., accept every proposal) and correct for the bias by importance weighting.

It probably won't work (well) if $\mathcal{K}$ is a naive random walk proposal scheme. But if $\mathcal{K}$ is an *informed* scheme, this idea is almost always effective.



Example: Importance Tempering of MTM

Locally balanced MTM on general state spaces

Let $\mathcal{K}(x,\cdot)$ denote a symmetric proposal with density κ. Let h be a function s.t. $h(u) = u\,h(u^{-1})$ for $u \geq 0$.

An iteration of MTM at state x with m tries:

1. Draw $y_1,\ldots,y_m$ from $\mathcal{K}(x,\cdot)$.
2. Select y from $y_1,\ldots,y_m$ with probability $\propto h(\pi(y)/\pi(x))$.
3. Draw $x_1,\ldots,x_{m-1}$ from $\mathcal{K}(y,\cdot)$. *Set $x_m = x$.*
4. Accept y with probability

$$\min\left\{1, \frac{Z_h(x,y_1,\ldots,y_m)}{Z_h(y,x_1,\ldots,x_m)}\right\},$$

where $Z_h(x,y_1,\ldots,y_m) = \sum_{k=1}^{m} h(\pi(y_k)/\pi(x))$.

Example: Importance Tempering of MTM

Multiple-try importance tempering

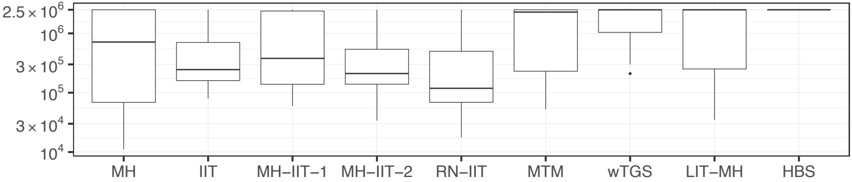
In Step 4, we can actually just accept y and assign to the previous state x importance weight $1/Z_h(x, y_1, \ldots, y_m)$. In the next iteration, the m candidate neighboring states of y are NOT resampled.

No extra computational cost for obtaining the importance weight.

Why is it correct? One can show that this algorithm is just a Markov chain importance sampling algorithm on an *augmented* space with auxiliary variables being the m candidate neighboring states.

Numerical Examples

A variable selection problem with $n = 1,000$ and $p = 5,000$



Box plot for the number of posterior calls (truncated at 2.5M) needed to find the best model. We consider a setting described in [9], where the design matrix has high collinearity, and the signal-to-noise ratio is intermediate. RN-IIT is a variant of the multiple-try importance tempering on discrete spaces. MTM: [1]; wTGS: [10]; LIT-MH: [12]; HBS: [8].

Concluding Remarks

- Importance tempering seems always better than MH for utilizing informed proposals. See [5] for more examples.
- Mixing time and asymptotic variance analysis is more challenging. For results on discrete spaces, see [11].
- The balancing function h needs to be chosen with caution.
- Importance tempering perspective opens doors to devising new MCMC schemes that are more efficient than existing ones.

Thank you!

Slides available at `https://zhouquan34.github.io`

- QZ. "From minimax optimal importance sampling to uniformly ergodic importance-tempered MCMC." `arXiv:2506.19186`.
- G. Li, A. Smith and QZ. "Importance is important: Generalized Markov chain importance sampling methods." `arXiv:2304.06251`.

[1] Hyunwoong Chang, Changwoo Lee, Zhao Tang Luo, Huiyan Sang, and Quan Zhou. Rapidly mixing multiple-try Metropolis algorithms for model selection problems. *Advances in Neural Information Processing Systems*, 35: 25842–25855, 2022.

[2] Philippe Gagnon, Florian Maire, and Giacomo Zanella. Improving multiple-try Metropolis with local balancing. *arXiv preprint arXiv:2211.11613*, 2022.

[3] Robert Gramacy, Richard Samworth, and Ruth King. Importance tempering. *Statistics and Computing*, 20:1–7, 2010.

[4] Anthony Lee and Krzysztof Łatuszyński. Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation. *Biometrika*, 101(3):655–671, 2014.

[5] Guanxun Li, Aaron Smith, and Quan Zhou. Importance is important: A guide to informed importance tempering methods. *arXiv preprint arXiv:2304.06251*, 2023.

[6] Kerrie L Mengersen and Richard L Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1): 101–121, 1996.

References II

[7] Jeffrey S Rosenthal, Aki Dote, Keivan Dabiri, Hirotaka Tamura, Sigeng Chen, and Ali Sheikholeslami. Jump Markov chains and rejection-free Metropolis algorithms. *Computational Statistics*, pages 1–23, 2021.

[8] Michalis K Titsias and Christopher Yau. The Hamming ball sampler. *Journal of the American Statistical Association*, 112(520):1598–1611, 2017.

[9] Yun Yang, Martin J Wainwright, and Michael I Jordan. On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, 44(6):2497–2532, 2016.

[10] Giacomo Zanella and Gareth Roberts. Scalable importance tempering and Bayesian variable selection. *Journal of the Royal Statistical Society Series B*, 81(3):489–517, 2019.

[11] Quan Zhou and Aaron Smith. Rapid convergence of informed importance tempering. pages 10939–10965, 2022.

[12] Quan Zhou, Jun Yang, Dootika Vats, Gareth O Roberts, and Jeffrey S Rosenthal. Dimension-free mixing for high-dimensional bayesian variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1751–1784, 2022.