# MTH707 Project Report: MCMC-Enabled Optimal Importance Sampling

Arqam Patel

We examine the following questions:

1. Are optimal importance sampling proposals feasible to sample from using MCMC?
2. Is it practically beneficial to use optimal IS estimates over vanilla Monte Carlo one?
3. Can we construct and use alternative IS proposals more viable, or amenable to MCMC?

## Optimal importance sampling proposals

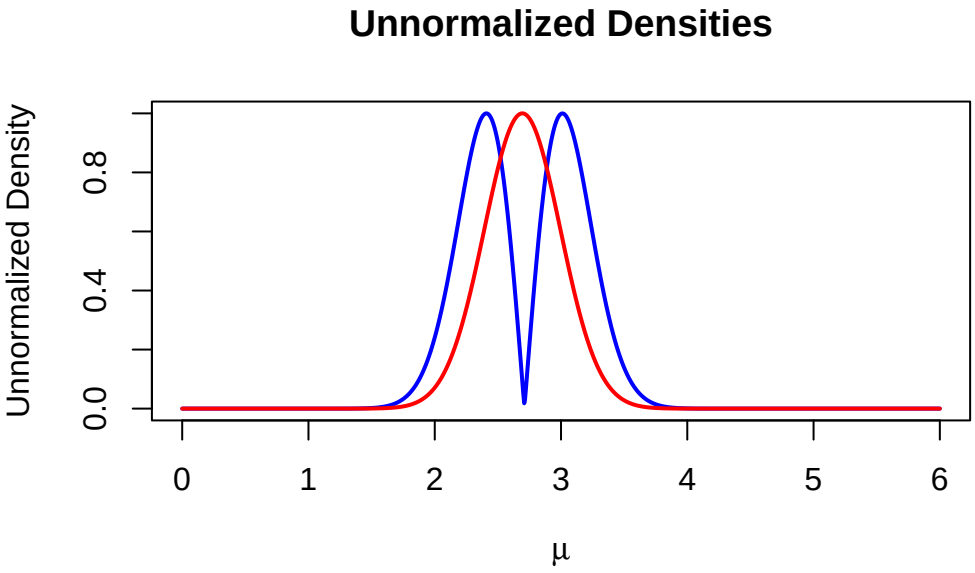
We find a detailed discussion of the various optimal IS proposals in Llorente et al. Note that these are derived under the assumption that our samples will be independent and identically distributed.

| | **Univariate $I$ and $f$** | **Multivariate I and f** |
|---|---|---|
| **Standard IS** | $q_{opt}(\theta) \propto \tilde{\pi}(\theta) \cdot \lvert f(\theta) \rvert$ | $q_{opt}(\theta) \propto \tilde{\pi}(\theta) \cdot \lVert \mathbf{f}(\theta) \rVert_2$ |
| **SNIS** | $q_{opt}(\theta) \propto \tilde{\pi}(\theta) \cdot \lvert f(\theta) - I \rvert$ | $q_{opt}(\theta) \propto \tilde{\pi}(\theta) \cdot \lVert \mathbf{f}(\theta) - \mathbf{I} \rVert_2$ |

Table 1: Optimal importance sampling proposals

### Interpretation and geometry

The optimal vanilla importance sampling proposal can be considered as upweighting regions with higher values of the function whose expectation we want to compute. The optimal self-normalised proposal can perhaps also be interpreted as upweighting regions with more 'outlying' function values (with respect to the expectation). The norm component takes a conical shape around the expectation (which is often near the mode in a unimodal setup), leading to a multimodal optimal SNIS density (plotted here for estimating the mean and variance of a simple 1D Gaussian posterior).




## Sampling from optimal proposals via MCMC

## Proximal computation

### Moreau Yosida envelope of L2 norm

Even assuming $\pi(\theta)$ to be differentiable, the norm term makes using derivative-based algorithms infeasible. Some prior work (Shukla et al, Pereyra et al) uses proximal methods to obtain a differentiable approximation to our MCMC target density and sample from it. However, these can not be used directly because of the condition of log-concavity of the MCMC target density.We can instead choose to selectively apply a Moreau envelope on the norm component of the density- which is convex. The envelope smooths the optimal density in the neighbourhood of the non differentiability.

We derive the general Moreau envelope of the L2 norm:

$$N_\lambda(\mathbf{x}) = \inf_{y\in\mathbb{R}^d}\left(\|y-\mu\|_2 + \frac{1}{2\lambda}\|x-y\|_2^2\right) = \inf_z\left\{\|z\| + \frac{1}{2\lambda}\|x-(z+\mu)\|^2\right\}.$$

Let $L(z) = \|z\| + \frac{1}{2\lambda}\|x-z-\mu\|^2$ denote our objective function. To find the infimum, we take the gradient of $L(z)$ with respect to $z$ and set it to zero:

$$\begin{aligned}\nabla L(z) = \frac{z}{\|z\|} - \frac{1}{\lambda}(x-z-\mu) &:= 0\\ \frac{z}{\lambda} + \frac{z}{\|z\|} &= \frac{1}{\lambda}(x-\mu)\\ z &= \frac{(x-\mu)}{\lambda(\frac{1}{\lambda}+\frac{1}{\|z\|})}\end{aligned}$$

Let $\|z\| = r$ and taking norm on both sides:

$$
\begin{aligned}
\|z\| &= \left\| \frac{x-\mu}{1+\frac{\lambda}{\|z\|}} \right\| \\
r &= \frac{r\|x-\mu\|}{r+\lambda} \\
r &= \|x-\mu\| - \lambda
\end{aligned}
$$

This solution is valid only if $\|x-\mu\| > \lambda$, and in that case we can directly substitute $z = \frac{x-\mu}{1+\frac{\lambda}{r}}$ and $r = \|x-\mu\| - \lambda$ into $L(z)$:

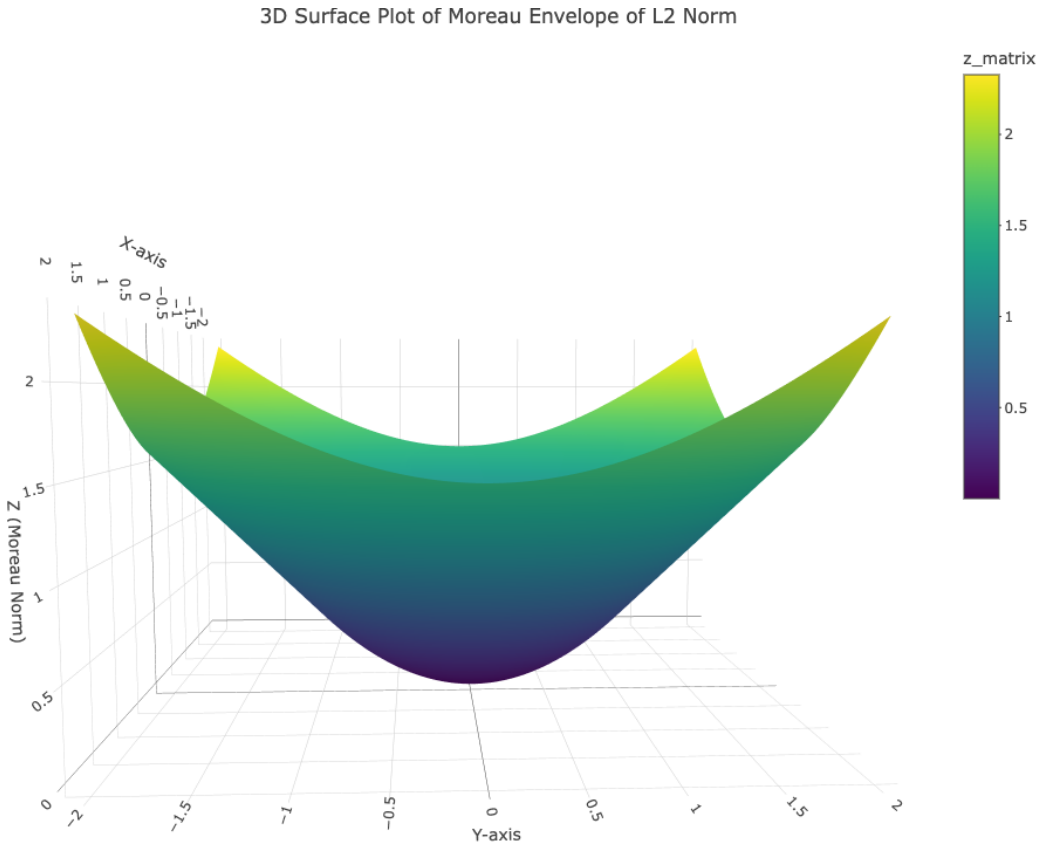
$$
\begin{aligned}
N_\lambda(x) &= (\|x-\mu\| - \lambda) + \frac{1}{2\lambda} \left\| x - \frac{x-\mu}{1+\frac{\lambda}{\|x-\mu\|-\lambda}} - \mu \right\|^2 \\
&= (\|x-\mu\| - \lambda) + \frac{1}{2\lambda} \left\| x - \mu - \frac{(\|x-\mu\| - \lambda)(x-\mu)}{\|x-\mu\|} \right\|^2 \\
&= (\|x-\mu\| - \lambda) + \frac{1}{2\lambda} \left\| \frac{\lambda(x-\mu)}{\|x-\mu\|} \right\|^2 \\
&= \|x-\mu\| - \frac{\lambda}{2}
\end{aligned}
$$

If instead $\|x-\mu\| \leq \lambda$, the infimum occurs at $z = 0$. We can prove this by showing $L(z) > L(0)$ for all $z \neq 0$. Consider the difference:

$$
\begin{aligned}
L(z) - L(0) &= \|z\| + \frac{1}{2\lambda}\|x-z-\mu\|^2 - \frac{1}{2\lambda}\|x-\mu\|^2 \\
&= \|z\| + \frac{1}{2\lambda}\left(\|x-\mu\|^2 - 2(x-\mu)^\top z + \|z\|^2\right) - \frac{1}{2\lambda}\|x-\mu\|^2 \\
&= \|z\| - \frac{1}{\lambda}(x-\mu)^\top z + \frac{1}{2\lambda}\|z\|^2 \\
&\geq \|z\| - \frac{1}{\lambda}\|x-\mu\| \cdot \|z\| + \frac{1}{2\lambda}\|z\|^2 \qquad \text{by Cauchy-Schwarz } |(x-\mu)^\top z| \leq \|x-\mu\| \cdot \|z\| \\
&\geq \|z\| \left(1 - \frac{\|x-\mu\|}{\lambda}\right) + \frac{1}{2\lambda}\|z\|^2 \\
&\geq 0
\end{aligned}
$$

Substuting $z = 0$ we get:

$$
N_\lambda(x) = \begin{cases} \frac{1}{2\lambda}\|x-\mu\|_2^2, & \|x-\mu\|_2 \leq \lambda \\ \|x-\mu\|_2 - \frac{\lambda}{2}, & \|x-\mu\|_2 > \lambda \end{cases}
$$




## Positivity condition

Self-normalised importance sampling requires $q(\theta) > 0$ wherever $\pi(\theta) > 0$ (Owen, pg 8). We can ensure this by adding a small constant epsilon to the norm. This also ensures our weights are bounded.

$$w(\theta) = \frac{\tilde{\pi}(\theta)}{\tilde{q}(\theta)} = \frac{1}{N_\lambda(f(\theta) - \mu) + \epsilon} \leq \frac{1}{\epsilon}$$

Thus, our quasi-optimal unnormalised density takes the form:

$$\hat{q}(\theta) \propto \tilde{\pi}(\theta) \cdot (N_\lambda\left(f(\theta) - \hat{\mu}\right) + \epsilon)$$

## Moreau-Yosida envelope of general log-density

For a general form target density $p(x)$ , we can use numerical optimisation methods to find the envelope (Pereyra 2015, Shukla et al 2025). Our objective is to find:

$$\pi_\lambda(\theta) = \sup_{u \in \mathbb{R}^d} \left\{ \frac{1}{k'} \pi(u) \exp\left( -\frac{\|\theta - u\|^2}{2\lambda} \right) \right\}$$

where $k'$ is a normalisation constant. Taking $\psi(\theta) = -\log \pi(\theta)$ as the negative log likelihood, we can reformulate this as finding:

$$\psi_\lambda(\theta) = \inf_{u \in \mathbb{R}^d} \left\{ -\log \pi(u) + \frac{\|\theta - u\|^2}{2\lambda} \right\}$$

In practice, to evaluate this at some value $\theta$, we need to find the proximal mapping:

$$\tilde{\theta} := \text{prox}_\psi^\lambda(\theta) = \arg\min_{u \in \mathbb{R}^d} \left\{ \psi(u) + \frac{\|\theta - u\|^2}{2\lambda} \right\} =: \arg\min_{u \in \mathbb{R}^d} f_\theta^\lambda(u)$$

Then, to compute $\psi_\lambda(x)$

$$\psi^\lambda(\theta) = \psi(\tilde{\theta}) + \frac{\|\theta - \tilde{\theta}\|^2}{2\lambda}$$

Using this we can evaluate $\pi_\lambda(\theta) = \exp(-\psi_\lambda(\theta))$. Further, we can also compute the log-gradient of

$$-\nabla \psi_\lambda(\theta) = \nabla \log \pi_\lambda(\theta) = \frac{\tilde{\theta} - \theta}{\lambda}$$

This can be used to implement gradient based algorithms on $\pi_\lambda(\theta)$ as a substitute for non-differentiable $\pi(\theta)$.

## Special case: Poisson regression

We now derive the algorithm for conducting PMALA on the IS optimal density in case of poisson regression, on which we have conducted simulation studies.

### Poisson Regression Model

Consider a Poisson regression model with $n$ observations and $d$ predictors. Given:

- Design matrix $X \in \mathbb{R}^{n \times d}$
- Response vector $y \in \mathbb{N}^n$
- Parameters $\theta \in \mathbb{R}^d$
- An estimate for our integral of interest $\hat{\mu}$

**Likelihood**: $y_i \sim \text{Poisson}(e^{X_i \theta})$

**Prior**: $\theta \sim \mathcal{N}(0, \sigma_0^2 I)$

The negative log-optimal density $\psi(\theta)$ is:

$$\psi(\theta) = \underbrace{-\sum_{i=1}^{n} [y_i (X_i \theta) - e^{X_i \theta}]}_{\text{Negative log-likelihood}} + \underbrace{\frac{1}{2\sigma_0^2} \|\theta\|^2}_{\text{Negative log-prior}} \underbrace{- \frac{1}{2} \log \|\theta - \hat{\mu}\|^2}_{\text{L2 norm term}} + C \tag{1}$$

Let the proximal loss function $f_\theta^\lambda(u) = \psi(u) + \frac{1}{2\lambda}\|\theta - u\|^2$. We consider the gradient and hessian of this expression in order to conduct Newton-Raphson optimization to find the arg min.

$$\begin{aligned}
\nabla\psi(u) &= \underbrace{-X^\top(y - e^{Xu})}_{\text{Log-likelihood grad}} + \underbrace{\frac{u}{\sigma_0^2}}_{\text{Prior grad}} - \underbrace{\frac{u-\hat{\mu}}{\|u-\hat{\mu}\|^2}}_{\text{L2 norm grad}} \\
\nabla_u f_\theta^\lambda(u) &= \nabla\psi(u) + \frac{u-\theta}{\lambda} \\
&= -X^\top(y - e^{Xu}) + \frac{u}{\sigma_0^2} - \frac{u-\hat{\mu}}{\|u-\hat{\mu}\|^2} + \frac{u-\theta}{\lambda} \\
\nabla^2\psi(u) &= \underbrace{X^\top W X}_{\text{Log-likelihood Hess}} + \underbrace{\frac{I_d}{\sigma_0^2}}_{\text{Prior Hess}} - \underbrace{\frac{I_d}{\|u-\hat{\mu}\|^2} + \frac{2(u-\hat{\mu})(u-\hat{\mu})^\top}{\|u-\hat{\mu}\|^4}}_{\text{L2 norm Hess}} \\
\nabla_u^2 f_\theta^\lambda(u) &= \nabla^2\psi(u) + \frac{I_d}{\lambda} \\
&= X^\top W X + \frac{I_d}{\sigma_0^2} - \frac{I_d}{\|u-\hat{\mu}\|^2} + \frac{2(u-\hat{\mu})(u-\hat{\mu})^\top}{\|u-\hat{\mu}\|^4} + \frac{I_d}{\lambda}
\end{aligned}$$

where $W = \text{diag}(e^{Xu})$

## PMALA Algorithm for optimal Poisson Regression IS target

**Initialize:** $\theta^{(0)} \neq \hat{\mu}$ (initial parameter vector) and $k = 0$ (iteration counter)

1. **While** $k < N$:

   a. Proximal Optimization:

   - Solve using Newton-Raphson:

$$\tilde{\theta}^{(k)} = \arg\min_u f_\theta^\lambda(u) = \arg\min_u \left[\psi(u) + \frac{1}{2\lambda}\|\theta^{(k)} - u\|^2\right]$$

   - The iterations will be:

$$u_{j+1}^{(k)} = u_j^{(k)} - \left(\nabla^2 f_\theta^\lambda\left(u_j^{(k)}\right)\right)^{-1} \cdot \nabla f_\theta^\lambda\left(u_j^{(k)}\right)$$

   b. Compute Proximal Gradient

$$g^{(k)} = \frac{\tilde{\theta}^{(k)} - \theta^{(k)}}{\lambda}$$

   c. Metropolis-Hastings Proposal: $\theta' \sim \mathcal{N}(\theta^{(k)} + \epsilon g^{(k)},\ 2\epsilon I_d)$

   d. Log Acceptance Probability:

$$\begin{aligned}
\log\alpha = &\underbrace{-\psi_\lambda(\theta') + \psi_\lambda(\theta^{(k)})}_{\text{Target ratio}} \\
&+ \underbrace{\frac{1}{4\epsilon}\|\theta^{(k)} - (\theta' + \epsilon g')\|^2 - \frac{1}{4\epsilon}\|\theta' - (\theta^{(k)} + \epsilon g^{(k)})\|^2}_{\text{Proposal ratio}}
\end{aligned}$$

e. Accept/Reject:

$$\theta^{(k+1)} = \begin{cases} \theta' & \text{w.p. } \min(1, \alpha) \\ \theta^{(k)} & \text{otherwise} \end{cases}$$

f. **Increment:** $k \leftarrow k + 1$

**Output:** Markov chain $\{\theta^{(0)}, \theta^{(1)}, ..., \theta^{(N)}\}$

## Proposed methods

We restrict our scope to the identity function, g(x) = x. We compare three samplers:

1. MALA: MALA on target $\pi(\theta)$

2. MNMALA: MALA on Moreau-norm adjusted target (SNIS) $\widehat{q}(\theta)$

3. PMALA: MALA on Moreau-envelope of optimal density (SNIS) $\pi_\lambda(\theta)$

For 2 & 3, we use the maximum likelihood estimate as $\widehat{\mu}$, the integral estimate, for approximating the optimal SNIS density.

### Numerical experiments

We consider the problem of Poisson regression on the epil dataset from package MASS, involving 6 covariates.

#### Variance

We observe very marginal variance reduction using PMALA, while MNMALA gives worse performance than simply MALA.

**Variance at 1e3 samples:**

| Parameter | PMALA | MNMALA | MALA |
|---|---|---|---|
| (Intercept) | 0.0002532014 | 0.0013638918 | 0.0009309689 |
| lbase | 0.0002370926 | 0.0010102440 | 0.0005255328 |
| trtprogabide | 0.0007315571 | 0.0023896189 | 0.0015707444 |
| lage | 0.0045597276 | 0.0087878501 | 0.0036886813 |
| V4 | 0.0002817426 | 0.0009914947 | 0.0005008303 |
| lbase:trtprogabide | 0.0009677908 | 0.0027321465 | 0.0009189163 |
| **Total Variance** | **0.007031112** | **0.017267236** | **0.008235674** |

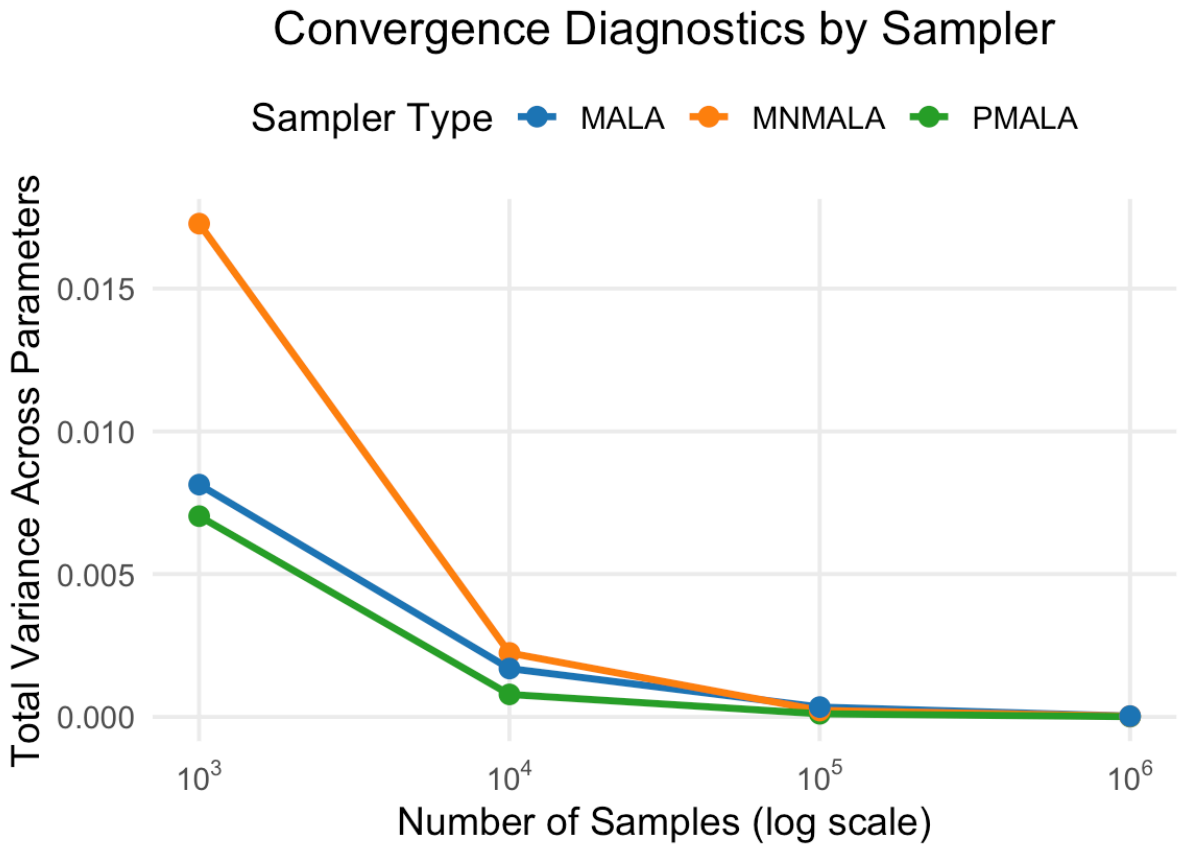
**Variance at 1e4 samples:**

| Parameter | PMALA | MNMALA | MALA |
|---|---|---|---|
| (Intercept) | 2.897564e-05 | 5.236143e-05 | 4.050534e-05 |
| lbase | 4.540592e-05 | 1.084788e-04 | 5.682031e-05 |
| trtprogabide | 7.251600e-05 | 2.675100e-04 | 1.029270e-04 |
| lage | 5.218886e-04 | 1.260951e-03 | 1.307645e-03 |
| V4 | 2.314834e-05 | 1.331611e-04 | 5.848811e-05 |

| Parameter | PMALA | MNMALA | MALA |
|---|---|---|---|
| lbase:trtprogabide | 9.571069e-05 | 4.182272e-04 | 1.279511e-04 |
| **Total Variance** | **7.89549e-04** | **2.23986e-03** | **1.69604e-03** |

**Variance at 1e5 samples:**

| Parameter | PMALA | MNMALA | MALA |
|---|---|---|---|
| (Intercept) | 4.748321e-06 | 8.700292e-06 | 1.238639e-05 |
| lbase | 4.517597e-06 | 1.249922e-05 | 1.246645e-05 |
| trtprogabide | 1.037464e-05 | 2.608106e-05 | 2.714461e-05 |
| lage | 7.631968e-05 | 1.270340e-04 | 2.541750e-04 |
| V4 | 1.566252e-06 | 1.064667e-05 | 6.565885e-06 |
| lbase:trtprogabide | 1.174674e-05 | 4.020578e-05 | 3.147683e-05 |
| **Total Variance** | **1.07768e-04** | **2.24264e-04** | **3.42386e-04** |




**Time**

| Algorithm | Mean time for 1e5 samples |
|---|---|
| PMALA | 76.91148 |
| MNMALA | 13.24839 |
| MALA | 12.81617 |

### Scalability

P-MALA uses Newton-Raphson and is thus susceptible to failure in case of high dimensional targets, due to singularity of the Hessian computations (since $f_\theta^\lambda$ is non-convex). Despite using generalised-inverses and gradient descent to attempt to circumvent this, we were unable to get PMALA to reliably work on the high dimensional football betting dataset (d'Angelo et al) due to issues of both computational cost and positive definiteness of Hessians, with gradient descent proving an insufficient substitute for Newton-Raphson and not converging.

## Literature review

Most prior research on the intersection of Markov chain samplers and importance sampling deals with using importance sampling to construct a more efficient estimator than standard Markov chain Monte Carlo. For example, Rudolf et al and Schuster et al propose self-normalised IS estimators that utilise all proposed samples, instead of just accepted ones in acceptance-rejection based Markov chain samplers.

Botev et al is more similar in spirit to what we seek to achieve here. It deals with sampling from an approximation of the optimal simple IS density using a Markov chain sampler, then estimating a one-dimensional integral using it.

## Conclusion

We explored two techniques for approximating the optimal self-normalised importance sampling target in order to get lower variance estimators than vanilla Monte Carlo. In the first, we apply a Moreau envelop on the L2 norm component of the optimal density and then use the resulting differentiable density as a target for Metropolis-adjusted Langevin sampling. In the second, we directly use the Moreau envelope of the optimal SNIS density.

Empirically, we observe that the first method does not incur as much computational overhead as the second. However, we find no significant evidence of variance reduction with the first method. With the second method, we observe very marginal variance reduction, that is offset by computation time and scalability constraints. The R scripts used for simulation experiments can be found at https://github.com/arqamrp/optimal_IS_MCMC.

## References


1. Botev, Z.I., L'Ecuyer, P. & Tuffin, B. Markov chain importance sampling with applications to rare event probability estimation. Stat Comput 23, 271–285 (2013). https://doi.org/10.1007/s11222-011-9308-2
2. Owen, A. B. Monte Carlo theory, methods and examples. (2013)
3. Llorente, F., & Martino, L. (2025). Optimality in importance sampling: a gentle survey. arXiv preprint arXiv:2502.07396.
4. Schuster, I., & Klebanov, I. (2020). Markov chain importance sampling—a highly efficient estimator for MCMC.*Journal of Computational and Graphical Statistics*,*30*(2), 260-268.

5. Rudolf, D., & Sprungk, B. (2020). On a Metropolis–Hastings importance sampling estimator.

6. Shukla A, Vats D, Chi E. (2024). MCMC Importance Sampling via Moreau-Yosida Envelopes. arXiv preprint arXiv:2501.02228v1

7. Pereyra M. (2015) Proximal Markov chain Monte Carlo algorithms arXiv preprint arXiv:1306.0187v4

8. D'Angelo L., Canale A (2022)., Efficient posterior sampling for Bayesian Poisson regression. arXiv preprint arXiv:2109.09520v3