Pseudo-Extended Markov Chain Monte Carlo

Christopher Nemeth[1], Fredrik Lindsten[2], Maurizio Filippone[3] & James Hensman[4]

[1]Lancaster University, [2]Linköping University, [3]EURECOM and [4]PROWLER.io

Background

Setup: Let π be a target probability density on $\mathbb{R}^d$ defined for all $\mathbf{x} \in \mathcal{X} := \mathbb{R}^d$ by

$$\pi(\mathbf{x}) := \frac{\gamma(\mathbf{x})}{Z} = \frac{\exp\{-\phi(\mathbf{x})\}}{Z}, \tag{1}$$

where $\phi : \mathcal{X} \to \mathbb{R}$ is a continuously differentiable function and Z is the normalizing constant. In the Bayesian setting, this would be the posterior, $\pi(\mathbf{x}) = p(\mathbf{y}|\mathbf{x})\pi_0(\mathbf{x})/p(\mathbf{y})$, where $p(\mathbf{y}|\mathbf{x})$, $\pi_0(\mathbf{x})$ and $p(\mathbf{y})$ are the likelihood, prior and normalizing constant, respectively.

Evaluation: Often $\pi(\mathbf{x})$ does not have a closed-form solution so we can use Markov chain Monte Carlo (MCMC) algorithms to sample from $\pi(\mathbf{x})$.

Problem: MCMC algorithms struggle to sample from $\pi(\mathbf{x})$ when it is *multimodal*.

Tempered MCMC

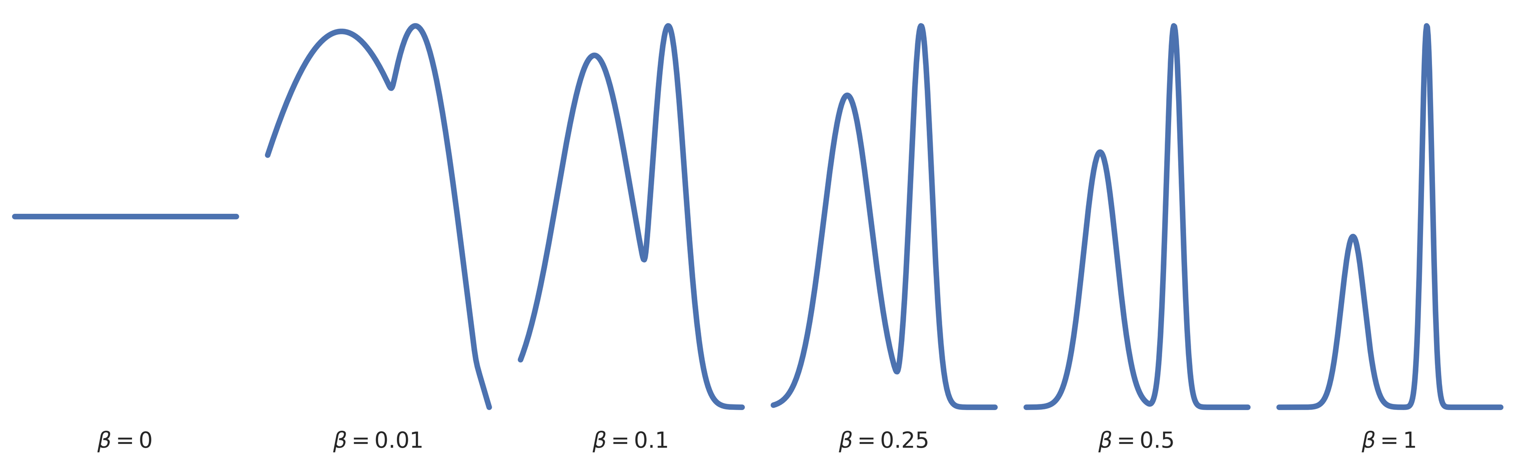
Tempered MCMC is the most popular approach to sampling from multi-modal target distributions (see Jasra et al. (2007) for a full review). The main idea behind tempered MCMC is to sample from a sequence of tempered targets,

$$\pi_k(\mathbf{x}) \propto \exp\{-\beta_k \phi(\mathbf{x})\}, \qquad k = 1, \ldots, K,$$

where β_k is a tuning parameter referred to as the **temperature** that is associated with $\pi_k(\mathbf{x})$.

A sequence of temperatures, commonly known as the **ladder**, is chosen **a priori**, where $0 = \beta_1 < \beta_2 < \ldots < \beta_K = 1$.

The intuition behind tempered MCMC is that **when β_k is small, the modes of the target are flattened out** making it easier for the MCMC sampler to traverse through the regions of low density separating the modes.



References

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37:697–725.

Graham, M. M. and Storkey, A. J. (2017). Continuously tempered Hamiltonian Monte Carlo. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, pages 1–12.

Hertz, J. A., Krogh, A. S., and Palmer, R. G. (1991). *Introduction to the theory of neural computation*, volume 1. Basic Books.

Jasra, A., Stephens, D. A., and Holmes, C. C. (2007). On population-based simulation for static inference. *Statistics and Computing*, 17(3):263–279.

Pseudo-Extended Method

We propose to **extend the state-space of the original target** eq. (1) by introducing N **pseudo-samples**, $\mathbf{x}_{1:N} = \{\mathbf{x}_i\}_{i=1}^N$, where the extended-target is now of the same form as the **pseudo-marginal** target (Andrieu and Roberts, 2009),

$$\pi^N(\mathbf{x}_{1:N}) := \frac{1}{N}\sum_{i=1}^N \pi(\mathbf{x}_i)\prod_{j\neq i} q(\mathbf{x}_j) = \frac{1}{Z}\left\{\frac{1}{N}\sum_{i=1}^N \frac{\gamma(\mathbf{x}_i)}{q(\mathbf{x}_i)}\right\} \times \prod_i q(\mathbf{x}_i), \tag{2}$$

where we introduce an instrumental distribution $q(\mathbf{x}_i) \propto \exp\{-\delta(\mathbf{x}_i)\}$ with support covering that of $\pi(\mathbf{x})$. See Fig. 1 for an illustration with $N = 2$.

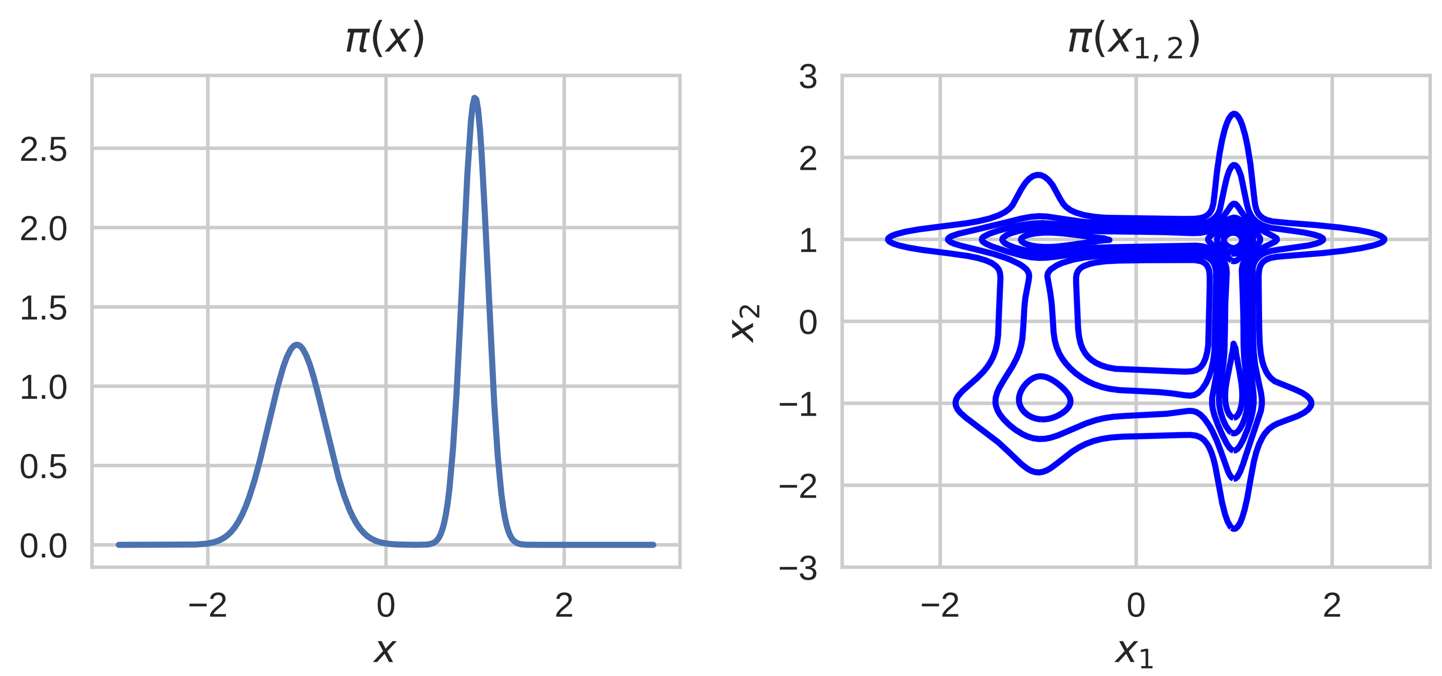


Figure 1: Original target density $\pi(\mathbf{x})$ (left) and extended target (right) with $N = 2$.

Unbiased Expectations

If $\mathbf{x}_{1:N}$ are distributed according to the extended-target π^N, then weighting each sample with self-normalized weights proportional to $\gamma(\mathbf{x}_i)/q(\mathbf{x}_i)$, gives samples from the original target, $\pi(\mathbf{x})$, where for an arbitrary integrable f,

$$\mathbb{E}_{\pi^N}\left[\frac{\sum_{i=1}^N f(\mathbf{x}_i)\gamma(\mathbf{x}_i)/q(\mathbf{x}_i)}{\sum_{i=1}^N \gamma(\mathbf{x}_i)/q(\mathbf{x}_i)}\right] = \mathbb{E}_\pi[f(\mathbf{x})].$$

Tempering targets with instrumental distributions

- Choosing $q(\mathbf{x})$ can be as challenging as sampling from $\pi(\mathbf{x})$.
- In pseudo-extended MCMC, unlike importance sampling, we don't need to sample from $q(\mathbf{x})$, just evaluate it. So we let $q(\mathbf{x}) = \pi(\mathbf{x})^\beta$ be a tempered version of the target and learn the temperature parameters $\pi(\beta)$. Therefore,

$$q(\mathbf{x}, \beta) = \frac{\gamma_\beta(\mathbf{x})g(\beta)}{C},$$

where $g(\beta)$ can be evaluated point-wise and C is a normalizing constant. Plugging $q(\mathbf{x}, \beta)$ into (2) gives

$$\pi^N(\mathbf{x}_{1:N}, \beta_{1:N}) := \frac{1}{ZC^{N-1}}\left\{\frac{1}{N}\sum_{i=1}^N \frac{\gamma(\mathbf{x}_i)\pi(\beta_i)}{\gamma_{\beta_i}(\mathbf{x}_i)g(\beta_i)}\right\}\prod_{j=1}^N \gamma_{\beta_j}(\mathbf{x}_j)g(\beta_j), \tag{3}$$

We apply MCMC directly on (3) and do not need to sample from $q(\mathbf{x}, \beta)$, just evaluate it.

Experiments

In the paper we consider the following targets: **Mixture of Gaussians**, **Boltzmann machine relaxation model**, and **Sparse logistic regression**. Plots for the Boltzmann example are given below and other results can be found in the paper.

Boltzmann relaxation: The probability mass function,

$$P(\mathbf{s}) = \frac{1}{Z_b}\exp\left\{\frac{1}{2}\mathbf{s}^\top\mathbf{W}\mathbf{s} + \mathbf{s}^\top\mathbf{b}\right\}, \quad \text{with} \quad Z_b = \sum_{\mathbf{s}\in\mathcal{S}}\exp\left\{\frac{1}{2}\mathbf{s}^\top\mathbf{W}\mathbf{s} + \mathbf{s}^\top\mathbf{b}\right\}, \tag{4}$$

is defined on the binary space $\mathbf{s} \in \{-1, 1\}^{d_b} := \mathcal{S}$, where $\mathbf{W}$ is a $d_b \times d_b$ real symmetric matrix and $\mathbf{b} \in \mathbb{R}^{d_b}$ are the model parameters. Using the **Gaussian integral trick** (Hertz et al., 1991), we introduce auxiliary variables $\mathbf{x} \in \mathbb{R}^d$ and transform the problem to sampling from $\pi(\mathbf{x})$

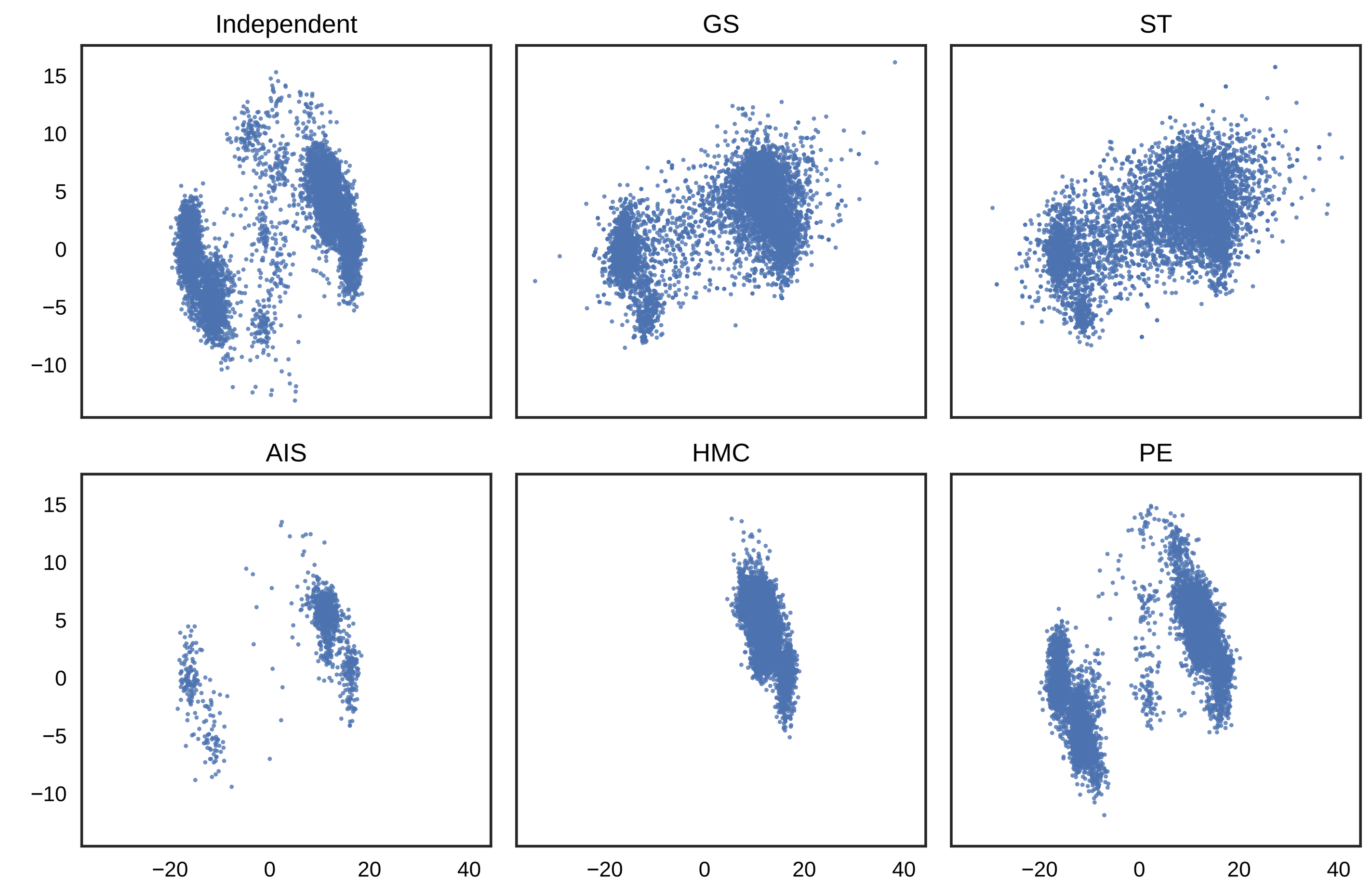


Figure 2: Two-dimensional projection of $10,000$ samples drawn from the target using methods (Independent - ground truth; GS - Graham and Storkey (2017) ; ST- Simulated Tempering; AIS - Annealed Importance Sampling; HMC - Hamiltonian Monte Carlo; PE - Pseudo-extended).

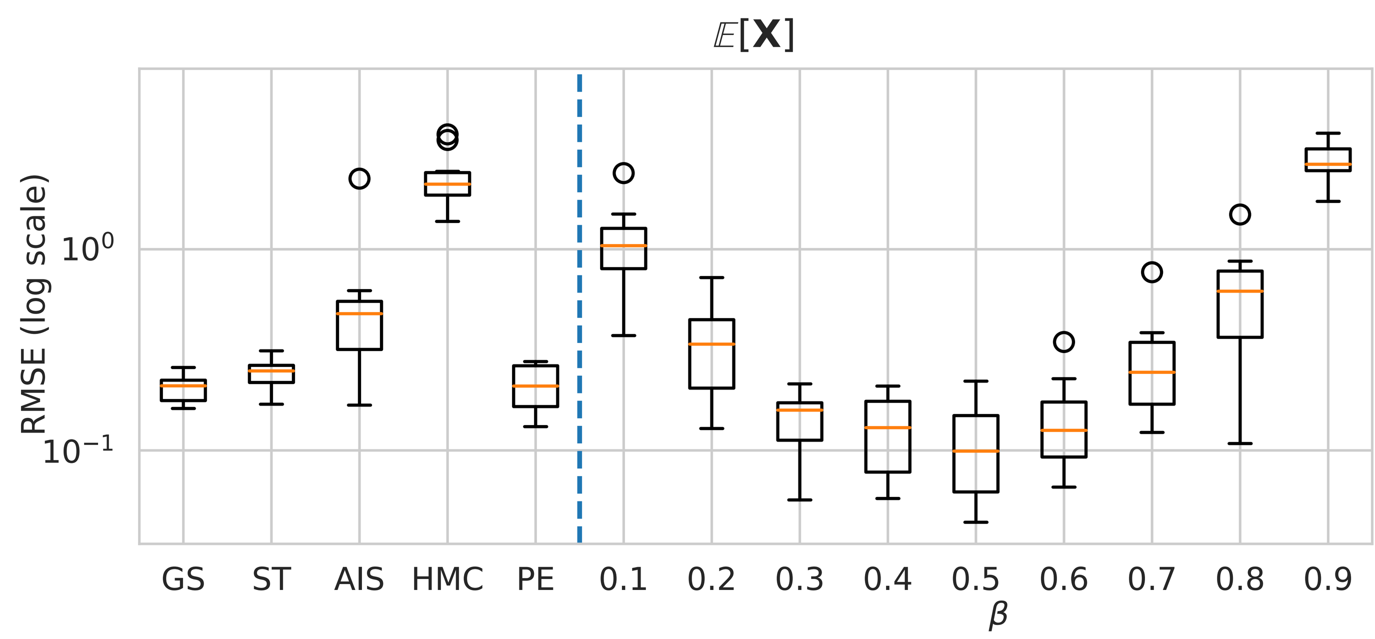


Figure 3: Root mean squared error (log scale) of the first and second moment of the target taken over 10 independent simulations and calculated for each of the proposed methods. Results labeled [0.1-0.9] correspond to pseudo-extended MCMC with fixed $\beta = [0.1 - 0.9]$.