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

Problem: 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.

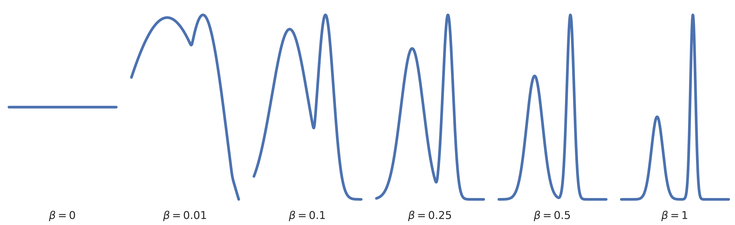
Tempered MCMC

Tempered MCMC is the most popular approach to sampling from multi-modal target distributions (see **?** 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.
- One of the most popular tempering algorithms is parallel tempering (PT) (**?**), where in parallel, K separate MCMC algorithms are run with each sampling from one of the tempered targets $\pi_k(\mathbf{x})$. Samples from neighboring Markov chains are exchanged (i.e. sample from chain k exchanged with chain $k-1$ or $k+1$) using a Metropolis-Hastings step. These exchanges improve the convergence of the Markov chain to the target of interest $\pi(\mathbf{x})$, however, information from low β_k targets is often slow to traverse up the temperature ladder.



Citations

Pseudo-Extended Method

We extend the state-space of the original target distribution eq. (1) by introducing N *pseudo-samples*, $\mathbf{x}_{1:N} = \{\mathbf{x}_i\}_{i=1}^N$, where the extended-target, defined on $\mathcal{X}^N$, is now of the same form as the *pseudo-marginal* target (**?**),

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

where we introduce an instrumental distribution $q(\mathbf{x}_i) \propto \exp\{-\delta(\mathbf{x}_i)\}$ with support covering that of $\pi(\mathbf{x})$.

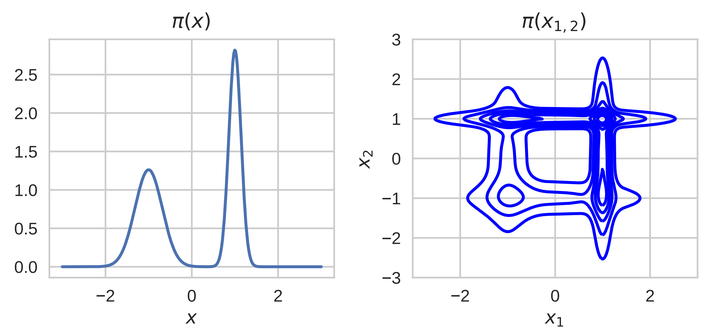


Figure 1:Original target density $\pi(\mathbf{x})$ (left) and extended target (right) with $N = 2$.

Unbiased Expectations

Let $\mathbf{x}_{1:N}$ be distributed according to the extended-target π^N. Weighting each sample with self-normalized weights proportional to $\gamma(\mathbf{x}_i)/q(\mathbf{x}_i)$, for $i = 1, \ldots, N$ gives samples from the target distribution, $\pi(\mathbf{x})$, in the sense that, 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

- A natural question for this method, "how do we choose $q(\mathbf{x})$?"
- Introduce pseudo-samples for temperatures $\{\beta_i\}_{i=1}^N$

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

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

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

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

Experiments

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

Boltzmann relaxation: The probability mass function,

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

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* (**?**), we introduce auxiliary variables $\mathbf{x} \in \mathbb{R}^d$ and transform the problem to sampling from $\pi(\mathbf{x})$

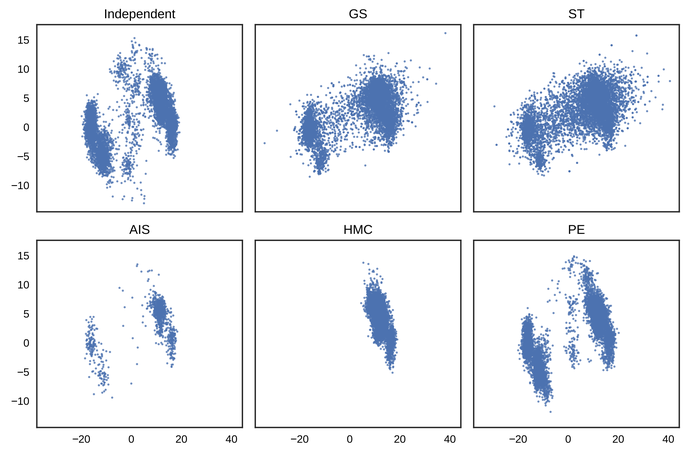


Figure 2:Two-dimensional projection of $10,000$ samples drawn from the target using each of the proposed methods, where the first plot gives the ground-truth sampled directly from the Boltzmann machine relaxation distribution. A temperature ladder of length 1,000 was used for both simulated tempering and annealed importance sampling.

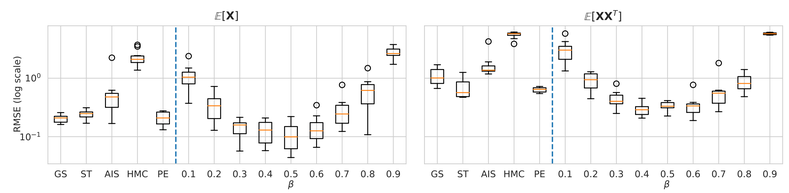


Figure 3:Root mean squared error (log scale) of the first and second moment of the target taken over 10 independent simulations and calculated for each of the proposed methods. Results labeled [0.1-0.9] correspond to pseudo-extended MCMC with fixed $\beta = [0.1 - 0.9]$.

Discussion