Transport Elliptical Slice Sampling

arXiv: 2210.10644

Alberto Cabezas González

Department of Mathematics and Statistics
Lancaster University

October 29, 2022

Objectives

Sample from an unnormalized posterior distribution

$$\pi(x) \propto L(y|x)\pi_0(x).$$

- Use local pointwise information of the target distribution to generate a Markov chain of dependent samples.
- MCMC with transition kernel $q(x'|x)$ such that,

$$Q\pi := \int q(x'|x)\pi(x)dx = \pi(x'). \tag{1}$$

- Efficient: we estimate $\mu = \int f(x)\pi(x)dx$ with $\hat{\mu} = n^{-1}\sum_{i=1}^{n} f(x_i')$, then by CLT

$$\hat{\mu} \approx N(\mu, n^{-1}\sigma) \tag{2}$$

where $\sigma = Var(f(x_i')) + 2\sum_k Cov(f(x_i'), f(x_{i+k}'))$ if stationary.

Objectives

Sample from an unnormalized posterior distribution

$$\pi(x) \propto L(y|x)\pi_0(x).$$

- Efficient MCMC algorithms usually rely on using gradient information from the target distribution, i.e. discretized Hamiltonian or Langevin dynamics of a process stationary on our target distribution.
- Optimizing these algorithms to efficiently minimize both computations and correlations between sequential samples requires algorithmic parameters to be manually tuned.
- Accelerate sampling on modern computer architectures, e.g. utilizing GPUs or TPUs.
- Many, short chains (run in parallel) instead of few, long chains.
- Lockstep necessity when simulating parallel chains on modern vector oriented libraries (PyTorch, TensorFlow, JAX).

Elliptical Slice sampler

- Gradient-free MCMC with no tuning parameters [Murray et al., 2010].

Require: $x, L(\mathcal{D}|\cdot)$

1: $v \sim \mathcal{N}(0, \mathbb{I}_d)$
2: $w \sim \text{Uniform}(0, 1)$
3: $\log s \leftarrow \log L(\mathcal{D}|x) + \log w$
4: $\theta \sim \text{Uniform}(0, 2\pi)$
5: $[\theta_{min}, \theta_{max}] \leftarrow [\theta - 2\pi, \theta]$
6: $x' \leftarrow x \cos\theta + v \sin\theta$
7: **if** $\log L(\mathcal{D}|x') > \log s$ **then**
8: Return x'
9: **else**
10: **if** $\theta < 0$ **then**
11: $\theta_{min} \leftarrow \theta$
12: **else**
13: $\theta_{max} \leftarrow \theta$
14: **end if**
15: $\theta \sim \text{Uniform}(\theta_{min}, \theta_{max})$
16: Go to 6.
17: **end if**

- Assume our target $\pi(x) \propto L(y|x)\mathcal{N}(x|0, \mathcal{C})$.

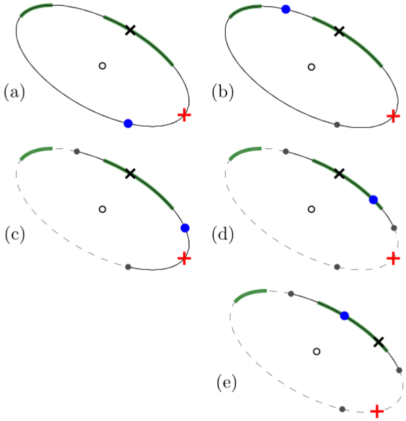

Figure: Elliptical Slice sampler

Normalizing flows

Theorem

Let $\mathcal{U} \subset \mathbb{R}^d$ be open and $T_\psi : \mathcal{U} \to \mathbb{R}^d$ be continuous, bijective and differentiable at every point in $\mathcal{U}$, then for every measurable $f : \mathbb{R}^d \to [0, \infty]$ and letting $\mathcal{X} = T_\psi(\mathcal{U})$

$$\int_{\mathcal{X}} f(x)dx = \int_{\mathcal{U}} f(T_\psi(u))|\det \nabla T_\psi(u)|du, \tag{3}$$

where ∇T is the Jacobian matrix of T.

- Choose a normalized and simple to sample from *reference density* $\phi(u)$.

$$\pi(x) = \phi(T_\psi^{-1}(x))|\det \nabla T_\psi^{-1}(x)| =: \hat{\phi}(x) \tag{4}$$

$$\phi(u) = \pi(T_\psi(u))|\det \nabla T_\psi(u)| =: \hat{\pi}(u), \tag{5}$$

- In practice we'll parametrize and optimize a T_ψ such that $\hat{\phi}(x) \approx \pi(x)$ and $\hat{\pi}(u) \approx \phi(u)$.

Normalizing flows

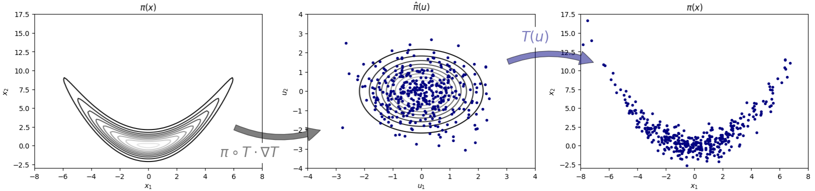

Figure: Sampling from the Banana density $\pi(x_1, x_2) \propto \exp\left(-[x_1^2/8 + \left(x_2 - x_1^2/4\right)^2]/2\right)$ using the transport map $T(u_1, u_2) = (\sqrt{8}u_1, u_2 + 2u_1^2)$ starts by transforming the target space to the reference space via a change of variables, drawing samples from an ellipsis on the extended reference space (not pictured) and pushing samples back to the target space.

Normalizing flows (parametrization)

- Wide class of linear and nonlinear functions which can be used [Kobyzev et al., 2020].
- Coupling architecture introduced by Dinh et al. [2014] with affine bijection.
- Consider the disjoint partition $x = (x^A, x^B) \in \mathbb{R}^p \times \mathbb{R}^{d-p}$. Then, one can define a transformation $G : \mathbb{R}^d \to \mathbb{R}^d$ by the formula

$$x^A = e^{\psi_1} \odot u^A + \psi_2 \tag{6}$$

$$x^B = u^B, \tag{7}$$

  given parameters $\Psi : \mathbb{R}^{d-p} \to \mathbb{R}^p \times \mathbb{R}^p$ learned only from the extended input, with Ψ a dense feedforward neural network.
- Easily inverted through a shift and scale with parameters $\Psi(x^B) = \Psi(u^B)$.
- The modulus determinant of its Jacobian matrix can be easily computed as $|\det \nabla G(x)| = \prod_{i=1}^{d} (e^{\psi_1})_i$ and $|\det \nabla G^{-1}(x)| = \prod_{i=1}^{d} (e^{-\psi_1})_i$.
- Arbitrary complexity by introducing a transformation $D : \mathbb{R}^d \to \mathbb{R}^d$ with the same structure as G but with the roles of the partitions reversed.

$$T_\psi = D_n \circ G_n \circ \cdots \circ D_1 \circ G_1, \qquad n \geq 1 \tag{8}$$

Normalizing flows (optimization)

- Minimize a divergence between our target density $\pi(x)$ and the *push-forward* reference density $\hat{\phi}(x)$.
- Kullback-Leibler divergence [KL; Kullback and Leibler, 1951] is arguably the most widely used and studied divergence.

$$\mathrm{KL}(\pi||\hat{\phi}) = \int \log \frac{\pi(x)}{\hat{\phi}(x)} \pi(x) dx. \tag{9}$$

- By LOTUS $\mathrm{KL}(\pi||\hat{\phi}) = KL(\hat{\pi}||\phi)$.
- KL divergence is asymmetric, i.e. $\mathrm{KL}(\pi||\hat{\phi}) \neq \mathrm{KL}(\hat{\phi}||\pi)$.
- Approximate inference minimizes (underestimating the real variance)

$$\mathrm{KL}(\phi(u)||\hat{\pi}(u)) \approx \frac{1}{M} \sum_{i=1}^{M} \log \frac{\phi(u_i)}{\hat{\pi}(u_i)}, \quad u_i \overset{iid}{\sim} \phi. \tag{10}$$

- We want to minimize (overestimating the real variance)

$$\mathrm{KL}(\pi(x)||\hat{\phi}(x)) \approx \frac{1}{k} \sum_{i=1}^{k} \log \frac{\pi(x_i)}{\hat{\phi}(x_i)}. \tag{11}$$

Transport Elliptical Slice Sampler

- Generalizes the elliptical slice sampler by targeting the extended state space $\pi(x)\phi(v)$.
- Given a map T_ψ such that $\hat{\pi}(u) \approx \phi(u)$, leave $\hat{\pi}(u)\phi(v)$ invariant.

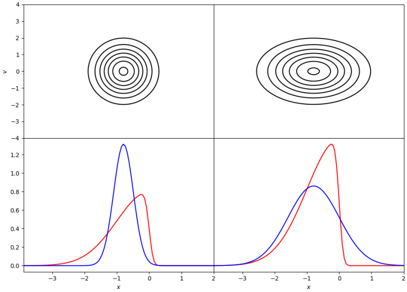

Figure: Ellipses corresponding to mean field approximations with underestimate (left) and overestimate (right) of the real variance.

Require: $u, T_\psi(\cdot), \hat{\pi}(\cdot)$

1: $v \sim \mathcal{N}(0, \mathbb{I}_d)$
2: $w \sim \text{Uniform}(0, 1)$
3: $\log s \leftarrow \log \hat{\pi}(u) + \log \phi(v) + \log w$
4: $\theta \sim \text{Uniform}(0, 2\pi)$
5: $[\theta_{min}, \theta_{max}] \leftarrow [\theta - 2\pi, \theta]$
6: $u' \leftarrow u \cos\theta + v \sin\theta$
7: $v' \leftarrow v \cos\theta - u \sin\theta$
8: **if** $\log \hat{\pi}(u') + \log \phi(v') > \log s$ **then**
9:     $x' \leftarrow T_\psi(u')$
10:     **Return** x', u'
11: **else**
12:     **if** $\theta < 0$ **then**
13:         $\theta_{min} \leftarrow \theta$
14:     **else**
15:         $\theta_{max} \leftarrow \theta$
16:     **end if**
17:     $\theta \sim \text{Uniform}(\theta_{min}, \theta_{max})$
18:     Go to 6.
19: **end if**

Adaptive Transport Elliptical Slice Sampler

Require: $u_{1:k}^{(0)}, h, m, N, \texttt{TESS}$
Set initial parameters of T_ψ and $\hat{\pi}$.
for $t \leftarrow 1, \dots, h$ **do** ▷ Warm-up
  for $i \leftarrow 1, \dots, k$ **do**
    $x_i^{(t)}, u_i^{(t)} \leftarrow \texttt{TESS}(u_i^{(t-1)}, T_\psi, \hat{\pi})$
  end for
  Update ψ in T_ψ by running m iterations of gradient descent on (12) using samples $x_{1:k}^{(t)}$.
end for
$u_{1:k}^{(0)} \leftarrow u_{1:k}^{(h)}$
for $t \leftarrow 1, \dots, N$ **do** ▷ Sampling
  for $i \leftarrow 1, \dots, k$ **do**
    $x_i^{(t)}, u_i^{(t)} \leftarrow \texttt{TESS}(u_i^{(t-1)}, T_\psi, \hat{\pi})$
  end for
end for
Return $x_{1:k}^{(1)}, \dots, x_{1:k}^{(N)}$

$$\mathrm{KL}(\pi(x)||\hat{\phi}(x)) \approx \frac{1}{k}\sum_{i=1}^{k} \log \frac{\pi(x_i)}{\hat{\phi}(x_i)} \qquad (12)$$

- Parameters must be learnt using samples from the target $\pi(x)$

$$\psi^* = \arg\min_{\psi \in \Psi} \mathrm{KL}(\pi||\hat{\phi}). \qquad (13)$$

- Alternates between optimizing ψ and sampling x, using k parallel chains, sequentially for h epochs.
- Minimizing $\mathrm{KL}(\pi||\hat{\phi})$ forces $\hat{\phi}(x)$ to cover the mass of $\pi(x)$.
- Overconfident approximation to the target variance, corrected using TESS.

Biochemical oxygen demand model

- $B(t) = \theta_0(1 - \exp(-\theta_1 t))$ for times $t < 5$.
- Set the parameters $\theta_0 = 1$ and $\theta_1 = 0.1$ and simulate $y(t_i)$ observations at times t_i evenly spaced in $[0, 5)$.

$$y(t_i) = \theta_0(1 - \exp(-\theta_1 t_i)) + \epsilon_i, \qquad i = 1, \ldots, 20, \tag{14}$$

  where $\epsilon_i \sim \mathcal{N}(0, \sigma_y^2)$ and fixed $\sigma_y^2 = 2 \times 10^{-4}$.
- Target posterior density is given by the likelihood $L(\mathbf{y}|\theta_0, \theta_1) = \prod_i \mathcal{N}(y(t_i); B(t_i; \theta_0, \theta_1), \sigma_y^2)$ and flat prior $\pi_0(\theta_0, \theta_1) \propto 1$.

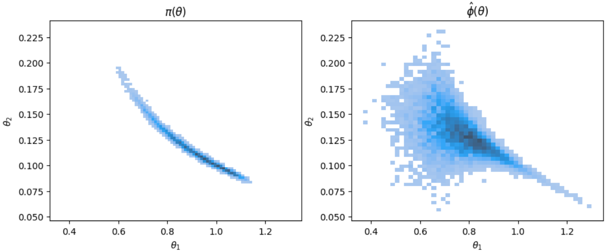

Figure: Samples from the target density $\pi(\theta)$ of the Biochemical oxygen demand model acquired by the TESS algorithm, mapped to $\hat{\phi}(\theta)$ (4), with diffeomorphism T_ψ learned using Adaptive TESS. With an approximation that overestimates the real variance (right) of our target (left) we are able to capture its global, non-Gaussian structure and explore it using a dimension independent and gradient-free method.

References

Iain Murray, Ryan Adams, and David MacKay. Elliptical slice sampling. In **Proceedings of the thirteenth international conference on artificial intelligence and statistics**, pages 541–548. JMLR Workshop and Conference Proceedings, 2010.

Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. **IEEE transactions on pattern analysis and machine intelligence**, 43(11):3964–3979, 2020.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. **arXiv preprint arXiv:1410.8516**, 2014.

Solomon Kullback and Richard A Leibler. On information and sufficiency. **The annals of mathematical statistics**, 22(1):79–86, 1951.