Oops I took a gradient: Gibbs with gradients

Scalable sampling for discrete distributions (Grathwohl et al., 2021)

Geelon So, agso@eng.ucsd.edu
Time series reading group — November 11, 2021

Sampling problem

Goal: sample from distribution π over discrete space $\mathcal{X}$ of the form:

$$\log \pi(x) = f(x) - \log Z,$$

where $f(x)$ is the unnormalized log-probability of x and Z is a normalization constant.

- We consider $\mathcal{X} = \{0, 1\}^D$, or more generally, $\mathcal{X} = \{0, 1, \ldots, K\}^D$.

Metropolis-Hastings algorithm

- Initialize X_0 arbitrarily from $\mathcal{X}$
- For $t = 0, 1, \ldots, T-1$
  - Sample from proposal distribution $X' \sim q(x' \mid X_t)$
  - Accept proposal with probability $A(X', X)$
    - If accept, set $X_{t+1} \leftarrow X'$
    - Otherwise, set $X_{t+1} \leftarrow X_t$
- Return X_T approximately drawn from π

Derivation of MH

Metropolis-Hastings designs a Markov process $P(x' \mid x)$ with stationary distribution π.

- **Key idea:** π is a stationary distribution of a Markov process if:

$$\pi(x)P(x' \mid x) = \pi(x')P(x \mid x'). \tag{$*$}$$

- Decompose $P(x' \mid x)$ as $q(x' \mid x)A(x', x)$
- If we want the decomposition to satisfy condition $(*)$, we need:

$$\frac{A(x', x)}{A(x, x')} = \frac{P(x')}{P(x)} \frac{q(x \mid x')}{q(x' \mid x)}.$$

- An example of an acceptance probability satisfying this is:

$$A(x', x) = \min\left\{1, \frac{P(x')}{P(x)} \frac{q(x \mid x')}{q(x' \mid x)}\right\}.$$

MNIST example

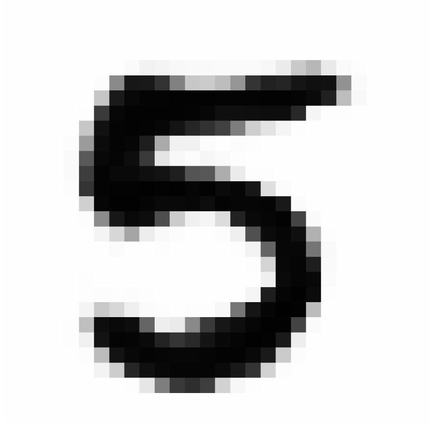

Figure 1: Most pixels are in the background, so will likely not change. This amounts to a wasted computation almost every time MH proposes changing a background pixel.

Hamming update

Consider the MH algorithm with proposal distribution:

$$q(x' \mid x) = \sum_{i \in [D]} q(x' \mid x, i) q(i \mid x),$$

where x' differ from x only on one coordinate:

- Choose a random coordinate according to $q(i \mid x)$ over $[D]$.
- Then, propose a change x' on only the ith coordinate.

MNIST example: can choose $q(i \mid x)$ to focus on pixels at the edge of a digit.

Locally-informed proposals

Consider the **locally-informed proposal**,

$$q_\tau(x' \mid x) \propto \exp\left(\frac{1}{\tau}\big(f(x') - f(x)\big)\right) \mathbf{1}\{\|x - x'\|_1 \leq r\}.$$

- This tends to propose a change x' around x that increases the local likelihood.
- The temperature τ controls how aggressively the local likelihood is optimized in,

$$A(x', x) = \min\left\{1,\ \frac{P(x')}{P(x)}\frac{q(x \mid x')}{q(x' \mid x)}\right\}.$$

  - If τ is too low, reverse transition probability collapsed.
  - If τ is too high, does not take local likelihood into accout.
- Previous work: $\tau = 2$ is the optimal locally-informed proposal (Zanella, 2020).

Efficient computation of locally-informed proposals

Problem: how to efficiently compute $f(x') - f(x)$ over the Hamming r-ball?

- Many useful discrete distributions have a related continuous distribution.

Distribution	$\log \pi(x) + \log Z$
Categorical	$x^T \theta$
Poisson	$x \log \lambda - \log \Gamma(x+1)$
RBM	$\sum_i \mathrm{softplus}(Wx + b)_i + c^T x$
Ising	$x^T W x + b^T x$
Deep EBM	$f_\theta(x)$

Table 1: Examples of discrete distributions with a differentiable extension to continuous space.

Approximation via gradients

For simplicity, consider the binary setting $\mathcal{X} = \{-\frac{1}{2}, \frac{1}{2}\}^D$.

- The likelihood ratios of flipping each bit is:

$$\tilde{d}(x) = -\mathrm{sign}(x) \odot \nabla f(x),$$

so that $\tilde{d}_i(x) \approx f(x_{-i}) - f(x)$ where x_{-i} corresponds to flipping the ith bit.

Algorithm: Gibbs with gradients

When the proposal x' satisfies $\|x' - x\|_1 \leq 1$, it corresponds to flipping at most one bit. Set proposal distribution $q(x_{-i} \mid x)$ to:

$$q(i \mid x) \propto \text{Categorical}\left(\text{softmax}\left(\frac{1}{2}\tilde{d}(x)\right)\right).$$

Algorithm: Gibbs with gradients

Input: unnormalized log-probability f and current sample x

- Compute $\tilde{d}(x) = -\text{sign}(x) \odot \nabla f(x)$
- Compute $q(i \mid x) \propto \text{Categorical}\big(\text{softmax}(\frac{1}{2}\tilde{d}(x))\big)$
- Sample index $i \sim q(i \mid x)$
- Propose flipping the ith coordinate
- Accept proposal with probability:

$$\min\left\{1, \exp\left(f(x') - f(x)\right) \cdot \frac{q(i \mid x')}{q(i \mid x)}\right\}.$$

Relationship to continuous relaxations

Prior works have made use of gradient information by:

- Transport the discrete sampling problem into a continuous relaxation
- Perform updates in continuous space (e.g. SVGD, MALA, HMC)
- Transform back to discrete space

Issue: poor scalability in high-dimensions, introduces additional hyperparameters

Experiments: restricted Boltzmann machines and MNIST

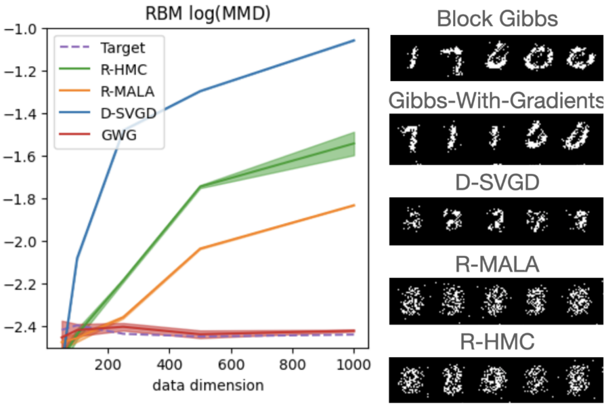

Figure 2: Ground truth distribution is given by Block Gibbs, where data distributions run up to 1000 dimensions. Plot compares Gibbs-with-Gradients with prior algorithms (lower is better).

Experiments: deep energy-based models and image generation

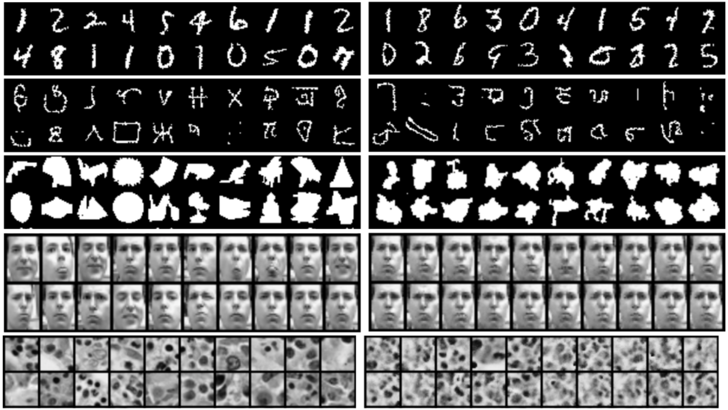

Figure 3: (*Left*) data. (*Right*) samples from ResNet EBM where samples are generated with annealed Markov chain using 300,000 GwG steps (MNIST, omniglot, Caltech Silhouettes, Frey Faces, Histopathology).

Approximation analysis

- Recall that q_τ where $\tau = 2$ is the optimal locally-informed proposal (Zanella, 2020).
- We've constructed an approximation to q_2 using gradients (Grathwohl et al., 2021).
- This paper shows that the approximated proposal distribution is at most a constant factor less efficient than q_2.

Quantities of analysis

- **Asymptotic variance** of a Markov kernel Q with stationary distribution π

$$\mathrm{Var}_\pi(h, Q) = \lim_{T\to\infty} \frac{1}{T}\mathrm{Var}\left(\sum_{t=1}^{T} h(x_t)\right),$$

  where $h : \mathcal{X} \to \mathbb{R}$ and $X_{t+1} \sim Q(x' \| X_t)$ and $X_1 \sim \pi$.
  - The smaller $\mathrm{Var}_\pi(h, Q)$, the more efficient the MCMC estimation of $\mathbb{E}_\pi[h]$.
- **Spectral gap** is defined:

$$\mathrm{Gap}(Q) = 1 - \lambda_2,$$

  where λ_2 is the second largest eigenvalue of the transition probability matrix of Q.
  - The larger the gap, the faster the mixing time.

Approximate proposal is efficient

Suppose that we have:

- f is the unnormalized log-probability, $\pi(x) = \frac{1}{Z} \exp\big(f(x)\big)$
- $q_2(x' \mid x)$ is the optimal locally balanced proposal
- $q^{\nabla}(x' \mid x)$ is the gradient-based approximation
- $Q(x', x)$ and $Q^{\nabla}(x', x)$ are the Markov transition kernel defined by MH

Theorem

If f is L-smooth, then:

(a) $\mathrm{Var}_\pi(h, Q^{\nabla}) \leq \frac{1}{c}\mathrm{Var}_\pi(h, Q) + \frac{1-c}{c}\mathrm{Var}_\pi(h)$

(b) $\mathrm{Gap}(Q^{\nabla}) \geq c \cdot \mathrm{Gap}(Q)$,

where $c = e^{-\frac{1}{2}L}$.

Practical implications of theorem

Since the decrease in efficiency has to do with the smoothness of f, if there is a choice for the functional representation, choose one that minimizes the Lipschitz constant.

Proof ingredients

The proof uses a result from Zanella (2020), reducing the problem to showing:

$$Q^{\nabla}(x', x) \geq c \cdot Q(x', x).$$

- Make use of L-smoothness of f, so that:

$$\|f(x') - f(x) - \nabla f(x)^T(x' - x)\| \leq \frac{L}{2}\|x' - x\|^2.$$

References

Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, and Chris J Maddison. Oops i took a gradient: Scalable sampling for discrete distributions. *arXiv preprint arXiv:2102.04509*, 2021.

Giacomo Zanella. Informed proposals for local mcmc in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865, 2020.