Optimization of proposal distribution for the Metropolis-Hastings algorithm

Agenda

- Goals of this talk
- Recall of the MH algorithm
- The lower bound on the acceptance rate
- Problem statement for two settings
- Density-based setting
- Sample-based setting

Agenda

- Goals of this talk
- Recall of the MH algorithm
- The lower bound on the acceptance rate
- Problem statement for two settings
- Density-based setting
- Sample-based setting

Disclaimer for the rigorous police

This talk is about some work in progress

More experiments are coming!

Agenda

- Goals of this talk
- Recall of the MH algorithm
- The lower bound on the acceptance rate
- Problem statement for two settings
- Density-based setting
- Sample-based setting

The Metropolis-Hastings algorithm

Target distribution $p(x)$
Proposal distribution $q(x'|x)$

1. sample proposal point $x' \sim q(x'|x)$,
   given previously accepted point x

2. accept $\begin{cases} x', & \text{if} \quad \frac{p(x')q(x|x')}{p(x)q(x'|x)} > u, \quad u \sim \text{Uniform}[0,1] \\ x, & \text{otherwise} \end{cases}$

If $q_\phi(x'|x) = q_\phi(x')$, we obtain the *independent* MH algorithm

Acceptance rate of the MH algorithm

$$\mathrm{AR} = \mathbb{E}_\xi \min\{1, \xi\} = \int dx dx' p(x) q(x'|x) \min\left\{1, \frac{p(x')q(x|x')}{p(x)q(x'|x)}\right\}$$

$$\xi = \frac{p(x')q(x|x')}{p(x)q(x'|x)}, \quad x \sim p(x), \;\; x' \sim q(x'|x)$$

Agenda

- Goals of this talk
- Recall of the MH algorithm
- The lower bound on the acceptance rate
- Problem statement for two settings
- Density-based setting
- Sample-based setting

Lower bounding the acceptance rate

$$\mathbb{E}_\xi \min\{1, \xi\} = 1 - \frac{1}{2}\mathbb{E}_\xi |\xi - 1| = 1 - \mathrm{TV}\left(p(x')q(x|x')\middle\|p(x)q(x'|x)\right)$$

$$\mathrm{AR} \geq 1 - \sqrt{\frac{1}{2} \cdot \mathrm{KL}\left(p(x')q(x|x')\middle\|p(x)q(x'|x)\right)}$$

Optimization problems

Optimization of the acceptance rate

$$\mathrm{TV}\left(p(x')q_\phi(x|x')\middle\|p(x)q_\phi(x'|x)\right) \to \min_\phi$$

Optimization of the lower bound on the acceptance rate

$$\mathrm{KL}\left(p(x')q_\phi(x|x')\middle\|p(x)q_\phi(x'|x)\right) \to \min_\phi$$

In terms of expectation

Optimization of the acceptance rate

$$\mathop{\mathbb{E}}_{\substack{x \sim p(x) \\ x' \sim q_\phi(x'|x)}} \left| \frac{p(x')q_\phi(x|x')}{p(x)q_\phi(x'|x)} - 1 \right| \to \min_\phi$$

Optimization of the lower bound on the acceptance rate

$$\mathop{\mathbb{E}}_{\substack{x \sim p(x) \\ x' \sim q_\phi(x'|x)}} \log \frac{p(x)q_\phi(x'|x)}{p(x')q_\phi(x|x')} \to \min_\phi$$

Agenda

- Goals of this talk
- Recall of the MH algorithm
- The lower bound on the acceptance rate
- Problem statement for two settings
- Density-based setting
- Sample-based setting

Two settings

Setting	**Target distribution**	**Proposal distribution**	**Density Ratio**
Density-based	given $\hat{p}(x) \propto p(x)$	explicit model $q(x')$	explicit
Sample-based	set of samples $X \sim p(x)$	implicit model $q(x')$ implicit model $q(x'\|x)$	learned discriminator

Agenda

- Goals of this talk
- Recall of the MH algorithm
- The lower bound on the acceptance rate
- Problem statement for two settings
- Density-based setting
- Sample-based setting

Collapsing to the delta-function

For the symmetric kernel

$$q_\phi(x'|x) = q_\phi(x|x')$$

The objective is

$$\mathbb{E}_{\substack{x \sim p(x) \\ x' \sim q_\phi(x'|x)}} \left| \frac{p(x')}{p(x)} - 1 \right| \to \min_\phi$$

The independent proposal

Optimization of the acceptance rate

$$\mathcal{L}(p, q_\phi) = \mathbb{E}_{\substack{x \sim p(x) \\ x' \sim q_\phi(x')}} \left| \frac{p(x')q_\phi(x)}{p(x)q_\phi(x')} - 1 \right| \to \min_\phi$$

Optimization of the lower bound on the acceptance rate

$$\mathcal{L}(p, q_\phi) = \mathbb{E}_{\substack{x \sim p(x) \\ x' \sim q_\phi(x')}} \log \frac{p(x)q_\phi(x')}{p(x')q_\phi(x)} \to \min_\phi$$

Algorithm in the density-based setting

Algorithm 1 Optimization of proposal distribution in density-based case

---

Require: explicit probabilistic model $q_\phi(x')$
Require: density of target distribution $\hat{p}(x) \propto p(x)$
  while ϕ not converged **do**
    sample $\{x'_k\}_{k=1}^K \sim q_\phi(x')$
    sample $\{x_k\}_{k=1}^K \sim p(x)$ using independent MH with current proposal q_ϕ
    $\mathcal{L}(p, q_\phi) \simeq \frac{1}{K}\sum_{k=1}^K \left| \frac{p(x'_k)q_\phi(x_k)}{p(x_k)q_\phi(x'_k)} - 1 \right|$ ▷ approximate loss with finite number of samples
    $\phi \leftarrow \phi - \alpha \nabla_\phi \mathcal{L}(p, q_\phi)$ ▷ perform gradient descent step
  end while
  return optimal parameters ϕ

For the lower bound on the acceptance rate

Algorithm 1 Optimization of proposal distribution in density-based case

---

Require: explicit probabilistic model $q_\phi(x')$
Require: density of target distribution $\hat{p}(x) \propto p(x)$
  while ϕ not converged **do**
    sample $\{x'_k\}_{k=1}^K \sim q_\phi(x')$
    sample $\{x_k\}_{k=1}^K \sim p(x)$ using independent MH with current proposal q_ϕ
    $\mathcal{L}(p, q_\phi) \simeq \frac{1}{K}\sum_{k=1}^K \log \frac{p(x_k)q_\phi(x'_k)}{p(x'_k)q_\phi(x_k)}$ ▷ approximate loss with finite number of samples
    $\phi \leftarrow \phi - \alpha \nabla_\phi \mathcal{L}(p, q_\phi)$ ▷ perform gradient descent step
  end while
  return optimal parameters ϕ

In the case of the Bayesian inference

Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$

Likelihood $p(\mathcal{D}|\theta) = \prod_i p(y_i|x_i, \theta)$

Prior $p(\theta)$

We want to obtain the predictive distribution

$$p(y|x) = \mathbb{E}_{p(\theta|\mathcal{D})} p(y|x, \theta)$$

samples from the posterior

Let’s sample using the MH algorithm!

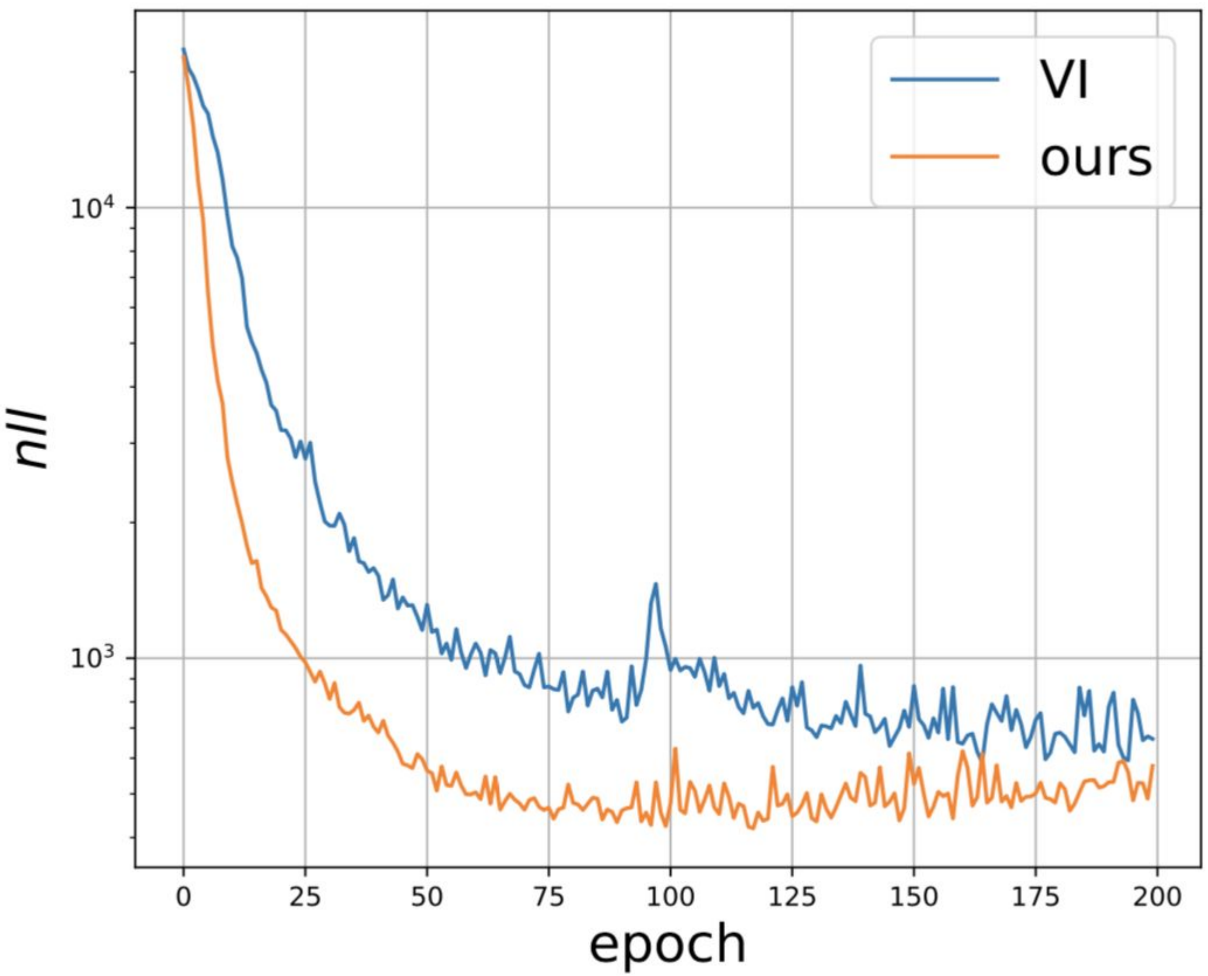
Optimization of the lower bound on the acceptance rate

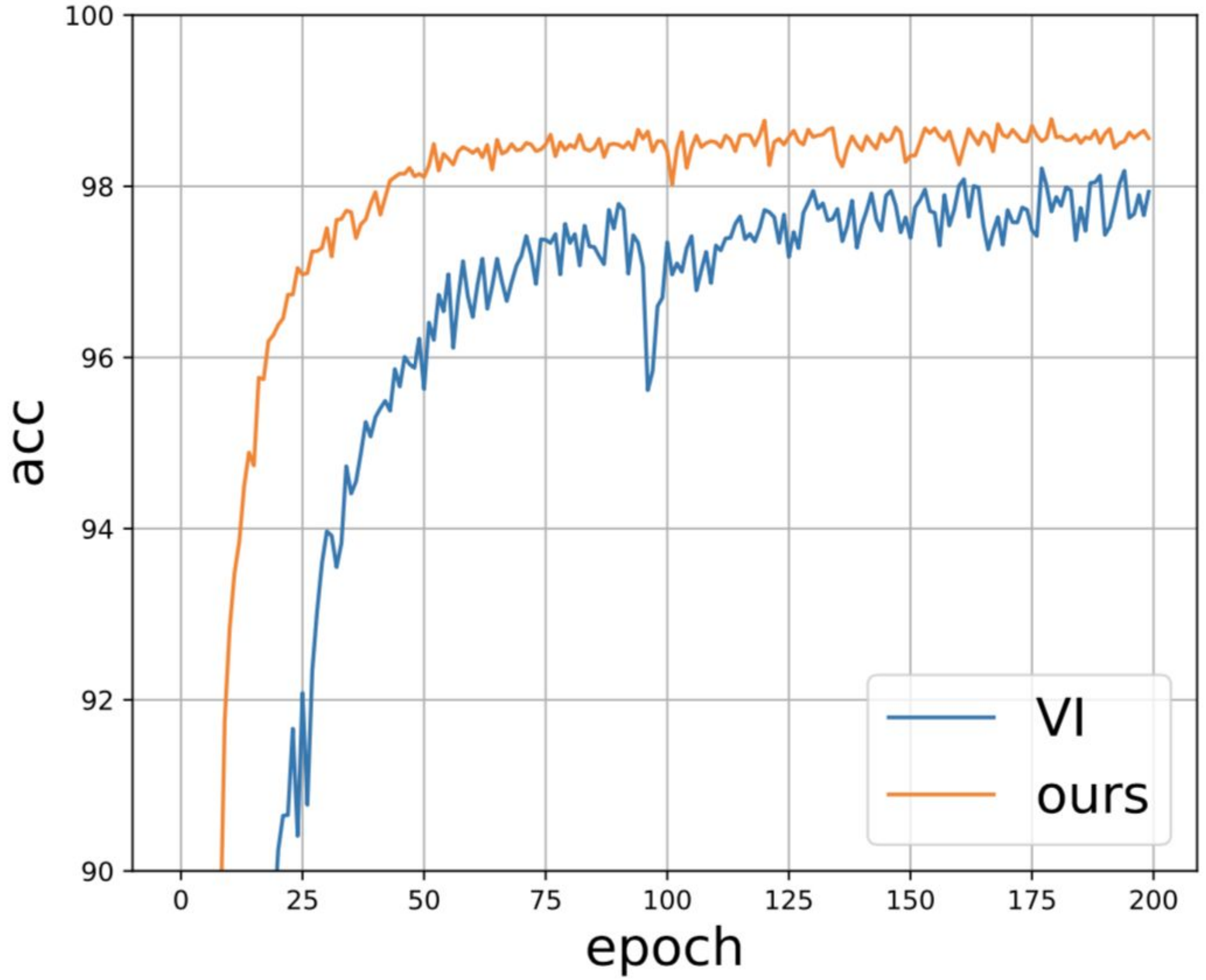
$$\mathcal{L}\Big(p(\theta|\mathcal{D}), q_\phi(\theta)\Big) = \mathrm{KL}\Big(p(\theta'|\mathcal{D})q_\phi(\theta)\Big\|p(\theta|\mathcal{D})q_\phi(\theta')\Big) \to \min_\phi$$

$$\mathrm{KL}\Big(p(\theta'|\mathcal{D})q_\phi(\theta)\Big\|p(\theta|\mathcal{D})q_\phi(\theta')\Big) = \mathrm{KL}\Big(q_\phi(\theta)\Big\|p(\theta|\mathcal{D})\Big) + \mathrm{KL}\Big(p(\theta'|\mathcal{D})\Big\|q_\phi(\theta')\Big) \to \min_\phi$$

$$\underbrace{-\mathbb{E}_{\theta\sim q_\phi(\theta)}\sum_{i=1}^{N}\log p(y_i|x_i,\theta) + \mathrm{KL}(q_\phi(\theta)\|p(\theta))}_{-\mathrm{ELBO}} - \mathbb{E}_{\theta\sim p(\theta|\mathcal{D})}\log q_\phi(\theta) \to \min_\phi$$

Reduced LeNet-5 on MNIST





Agenda

- Goals of this talk
- Recall of the MH algorithm
- The lower bound on the acceptance rate
- Problem statement for two settings
- Density-based setting
- Sample-based setting

Objectives

$$\mathcal{L}(p, q_\phi) = \mathbb{E}_{\substack{x \sim p(x) \\ x' \sim q_\phi(x'|x)}} \left| \frac{p(x')q_\phi(x|x')}{p(x)q_\phi(x'|x)} - 1 \right| \to \min_\phi$$

easy to sample

hard to estimate ratios

$$\mathcal{L}(p, q_\phi) = \mathbb{E}_{\substack{x \sim p(x) \\ x' \sim q_\phi(x'|x)}} \log \frac{p(x)q_\phi(x'|x)}{p(x')q_\phi(x|x')} \to \min_\phi$$

Density-ratio estimation

Objective for the discriminator

$$-\mathbb{E}_{\substack{x \sim p(x) \\ x' \sim q_\phi(x'|x)}} \log D(x, x') - \mathbb{E}_{\substack{x \sim p(x) \\ x' \sim q_\phi(x'|x)}} \log(1 - D(x', x)) \to \min_D$$

Optimal discriminator

$$D(x, x') = \frac{p(x)q_\phi(x'|x)}{p(x)q_\phi(x'|x) + p(x')q_\phi(x|x')}$$

Some ambiguity

For the optimal discriminator

$$D(x, x') = \frac{p(x)q_\phi(x'|x)}{p(x)q_\phi(x'|x) + p(x')q_\phi(x|x')}$$

We have

$$D(x, x') = 1 - D(x', x)$$

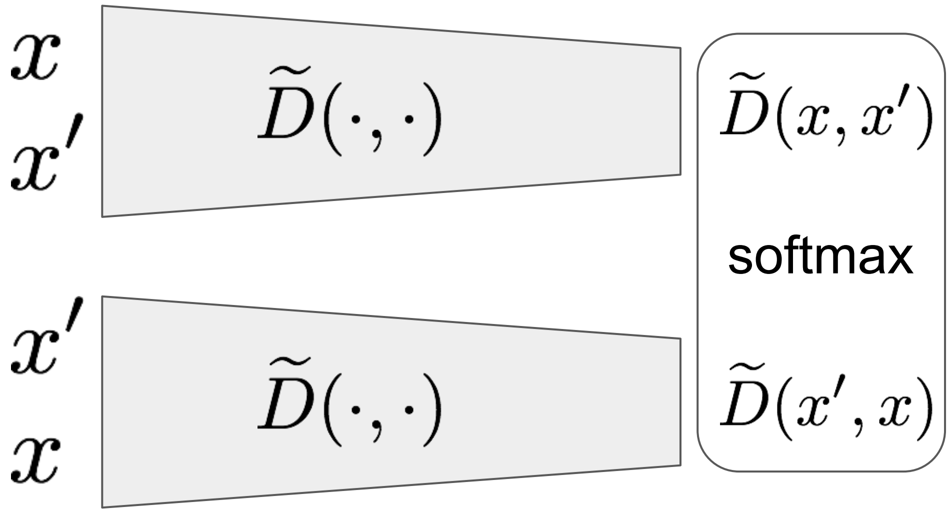
But what we should do on practice?

$$\frac{p(x)q_\phi(x'|x)}{p(x')q_\phi(x|x')} \approx \frac{D(x, x')}{1 - D(x, x')} \approx \frac{1 - D(x', x)}{D(x', x)} \approx \frac{1 - D(x', x)}{1 - D(x, x')} \approx \frac{D(x, x')}{D(x', x)}$$

Learn the discriminator of special structure

$$D(x,x') = \frac{\exp(\widetilde{D}(x,x'))}{\exp(\widetilde{D}(x,x')) + \exp(\widetilde{D}(x',x))}$$

Where $\widetilde{D}(\cdot,\cdot)$ is the convolutional neural network



Algorithm in the sample-based setting

Algorithm 2 Optimization of proposal distribution in sample-based case

Require: implicit probabilistic model $q_\phi(x' \mid x)$
Require: large set of samples $X \sim p(x)$
for n iterations **do**
  sample $\{x_k\}_{k=1}^K \sim X$
  sample $\{x'_k\}_{k=1}^K \sim q_\phi(x'|x)$
  train discriminator D by optimizing 13
  $\mathcal{L}(p, q_\phi) \approx \frac{1}{K}\sum_{k=1}^K \left|\frac{1-D(x_k,x'_k)}{D(x_k,x'_k)} - 1\right|$ ▷ approximate loss with finite number of samples
  $\phi \leftarrow \phi - \alpha\nabla_\phi\mathcal{L}(p, q_\phi)$ ▷ perform gradient descent step
end for
return parameters ϕ

For the lower bound on the acceptance rate

Algorithm 2 Optimization of proposal distribution in sample-based case

Require: implicit probabilistic model $q_\phi(x' \mid x)$
Require: large set of samples $X \sim p(x)$
  for n iterations **do**
    sample $\{x_k\}_{k=1}^K \sim X$
    sample $\{x'_k\}_{k=1}^K \sim q_\phi(x'|x)$
    train discriminator D by optimizing 13
    $\mathcal{L}(p, q_\phi) \approx \frac{1}{K}\sum_{k=1}^K \log\frac{D(x_k, x'_k)}{1-D(x_k, x'_k)}$ ▷ approximate loss with finite number of samples
    $\phi \leftarrow \phi - \alpha\nabla_\phi\mathcal{L}(p, q_\phi)$ ▷ perform gradient descent step
  end for
  return parameters ϕ

The little difference between LB and AR

$$D(x,x') = \frac{1}{1+\exp\left(-(\widetilde{D}(x,x') - \widetilde{D}(x',x))\right)} = \frac{1}{1+\exp(-d(x,x'))}$$

$$\frac{\partial L_{\mathrm{AR}}}{\partial x'} = \frac{1}{D^2(x,x')}\frac{\partial D(x,x')}{\partial x'} = \exp(-d(x,x'))\frac{\partial d(x,x')}{\partial x'}$$

$$\frac{\partial L_{\mathrm{LB}}}{\partial x'} = \frac{1}{(1-D(x,x'))D(x,x')}\frac{\partial D(x,x')}{\partial x'} = \frac{\partial d(x,x')}{\partial x'}$$

Independent proposal

with MH correction

without MH correction

Markov chain proposal

AR optimization

Lower bound optimization

Markov chain proposal

The end!