Stochastic MCMC techniques

Dmitry Vetrov

Research professor at HSE

Lab leader at Samsung AI Center

Head of Bayesian methods research group

http://bayesgroup.ru

Deep|Bayes

What stochasticity we are talking about?

We want to sample from posterior distribution

$$p(\theta|X) \propto p(\theta) \prod_{i=1}^{N} p(x_i|\theta)$$

Full dataset!

How to use minibatches instead of full dataset?

Minibatch MCMC techniques

Dmitry Vetrov

Research professor at HSE

Lab leader at Samsung AI Center

Head of Bayesian methods research group

http://bayesgroup.ru

DeepBayes

Variational Bayes vs. MCMC

	MCMC	IPM	Variational Bayes
Bias	No	??	Strong
Sampling/Ensembling	Inefficient	??	Efficient
Density	No	??	Yes
Likelihood	Needed	??	Needed

Metropolis-Hastings

$$\alpha(\theta, \theta') = \frac{p(\theta'|X)q(\theta|\theta')}{p(\theta|X)q(\theta'|\theta)}$$

$$\alpha(\theta, \theta') = \frac{p_0(\theta') \prod_{i=1}^{N} p(x_i|\theta')q(\theta|\theta')}{p_0(\theta) \prod_{i=1}^{N} p(x_i|\theta)q(\theta'|\theta)}$$

Full dataset!

$p_0(\theta)$ – prior distribution

Accept θ' if

$$\alpha(\theta, \theta') > u, \qquad u \sim \text{Uniform}[0,1]$$

An Efficient Minibatch Acceptance Test for Metropolis-Hastings

Deniel Seita, Xinlei Pan, Haoyu Chen, Jhon Canny

Barker lemma

$$\Delta(\theta, \theta') = \log \frac{p_0(\theta') \prod_{i=1}^{N} p(x_i|\theta') q(\theta|\theta')}{p_0(\theta) \prod_{i=1}^{N} p(x_i|\theta) q(\theta'|\theta)}$$

For any function $g(s)$ such that $g(s) = \exp(s)\, g(-s)$,
$\alpha(\theta, \theta') \triangleq g(\Delta(\theta, \theta'))$ satisfies detailed balance.

What does it mean?

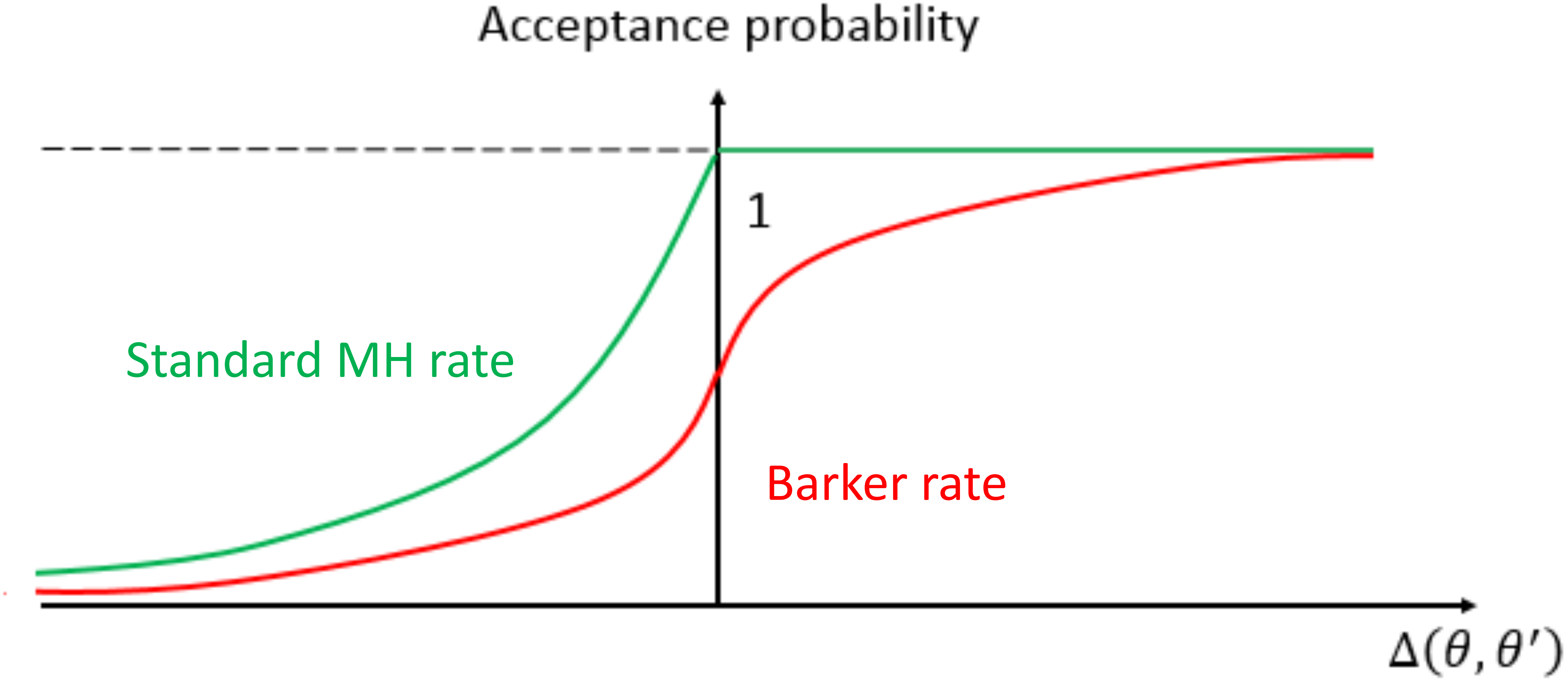
$$\Delta(\theta, \theta') = \log \frac{p_0(\theta') \prod_{i=1}^{N} p(x_i|\theta') q(\theta|\theta')}{p_0(\theta) \prod_{i=1}^{N} p(x_i|\theta) q(\theta'|\theta)}$$

If g satisfies Barker lemma, then performing the test

$$g(\Delta(\theta, \theta')) > u, \qquad u \sim \text{Uniform}[0,1]$$

we sample from true posterior distribution!

Acceptance rate



Barker acceptance function

Let $g(s) = \frac{1}{1+\exp(-s)}$, then test

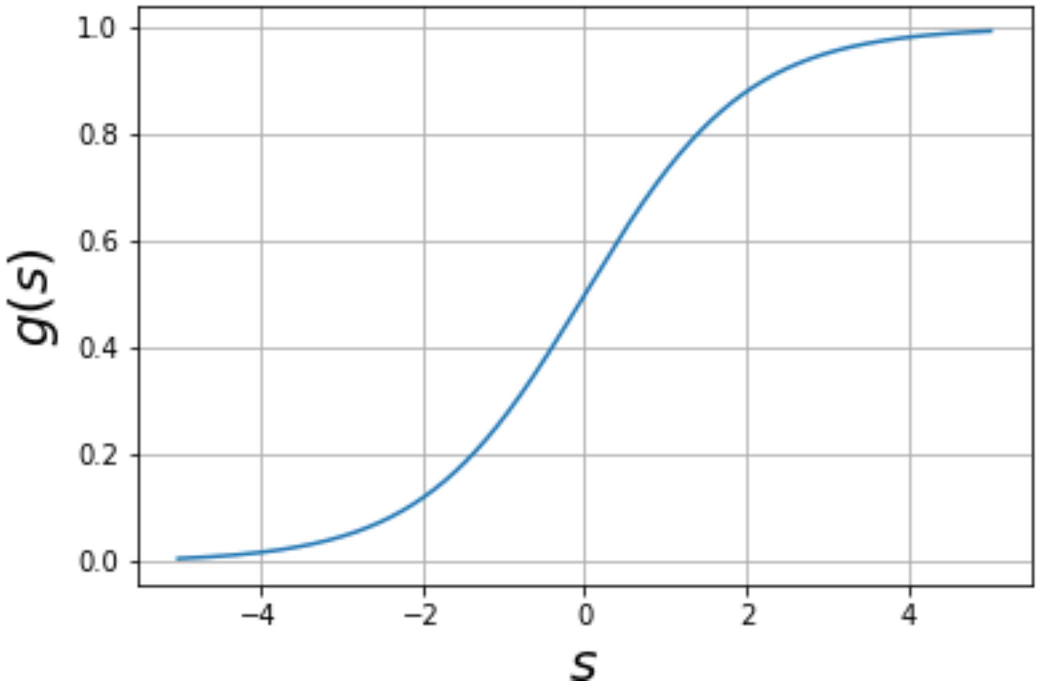
$$g\big(\Delta(\theta, \theta')\big) > u, \qquad u \sim \mathrm{Uniform}[0,1]$$

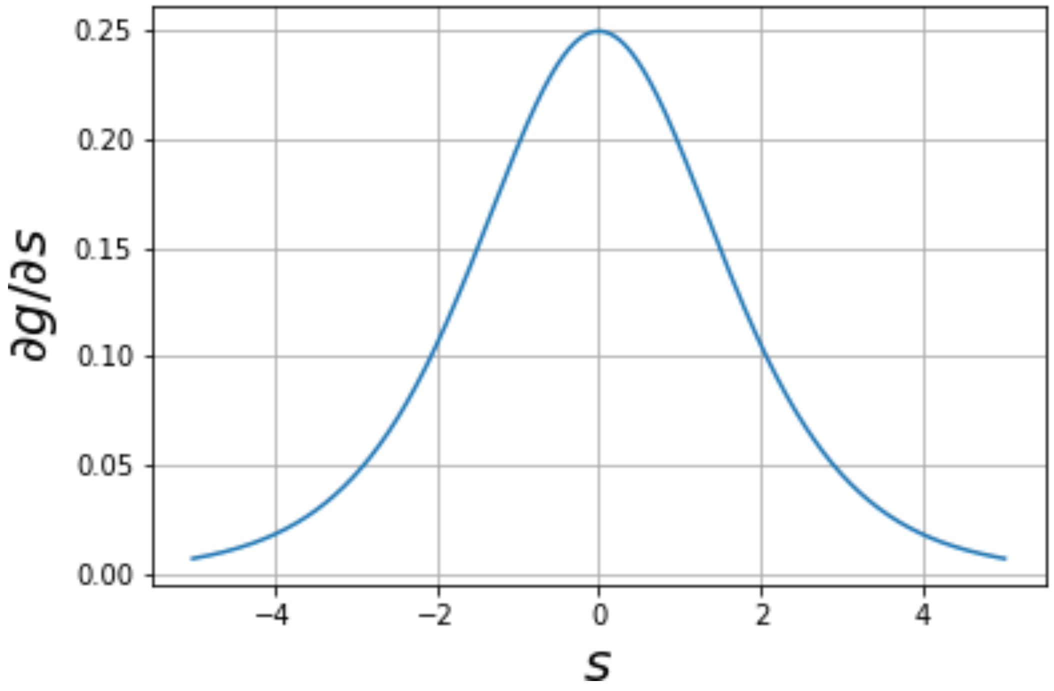
satisfies detailed balance and

$$\Delta(\theta, \theta') > X = g^{-1}(u), \qquad u \sim \mathrm{Uniform}[0,1]$$

also satisfies detailed balance

$g^{-1}(u)$ – sample from logistic distribution





$$X = X_{log} \sim \text{Logistic}(0,1)$$

$$-X_{log} \sim \text{Logistic}(0,1)$$

New acceptance test

$$\Delta(\theta,\theta') = \sum_i^N \log\frac{p(x_i|\theta')}{p(x_i|\theta)} + \log\frac{p_0(\theta')q(\theta|\theta')}{p_0(\theta)q(\theta'|\theta)}$$

Accept θ' if

$$\Delta(\theta,\theta') > X = g^{-1}(u), \qquad u \sim \text{Uniform}[0,1]$$

Or equivalently

$$\Delta(\theta,\theta') + X_{log} > 0, \qquad X_{log} \sim \text{Logistic}(0,1)$$

Exact, but we still use full dataset to sample one point

Minibatch acceptance test

$$\Delta^*(\theta, \theta') = \frac{N}{b} \sum_{i=1}^{b} \log \frac{p(x_i|\theta')}{p(x_i|\theta)} + \log \frac{p_0(\theta')q(\theta|\theta')}{p_0(\theta)q(\theta'|\theta)}$$

$$\Delta^* = \Delta + \mathrm{X}_{\mathrm{norm}}, \qquad \mathrm{X}_{\mathrm{norm}} \sim \overline{\mathcal{N}}\left(0, \sigma^2(\Delta^*)\right)$$

$\mathrm{X}_{\mathrm{norm}}$ – approximately normal distribution (Central Limit Theorem)

$$\Delta_i = N \log \frac{p(x_i|\theta')}{p(x_i|\theta)} + \log \frac{p_0(\theta')q(\theta|\theta')}{p_0(\theta)q(\theta'|\theta)}$$

$$\sigma^2(\Delta^*) = \sum_{i=1}^{b} \left(\Delta_i - \overline{\Delta}\right)^2$$

How to use Δ^* instead of Δ?

$$\Delta(\theta,\theta') = \sum_{i}^{N} \log \frac{p(x_i|\theta')}{p(x_i|\theta)} + \log \frac{p_0(\theta')q(\theta|\theta')}{p_0(\theta)q(\theta'|\theta)}$$

Accept θ' if

$$\Delta(\theta,\theta') + X_{log} > 0, \qquad X_{log} \sim \mathrm{Logistic}(0,1)$$

Our current test

But for minibatches we have value

$$\Delta^* = \Delta + \mathrm{X}_{\mathrm{norm}}, \qquad \mathrm{X}_{\mathrm{norm}} \sim \overline{\mathcal{N}}\left(0, \sigma^2(\Delta^*)\right)$$

Logistic noise decomposition

Let's decompose

$$X_{log} = X_{norm} + X_{corr}, \qquad X_{norm} \sim \mathcal{N}(0, \sigma^2),$$

where X_{corr} – correction distribution with PDF $C_\sigma(x)$

Not true!

If

$$\Delta^* = \Delta + X_{norm}, \qquad X_{norm} \sim \mathcal{N}(0, \sigma^2),$$

Then

$$\Delta + X_{log} = \underbrace{\Delta + X_{norm}}_{\Delta^*} + X_{corr} = \Delta^* + X_{corr}$$

Big picture

1. Evaluate

$$\Delta^*(\theta, \theta') = \frac{N}{b} \sum_{i=1}^{b} \log \frac{p(x_i|\theta')}{p(x_i|\theta)} + \log \frac{p_0(\theta')q(\theta|\theta')}{p_0(\theta)q(\theta'|\theta)}$$

2. Sample

$$X_{corr} \sim \text{Correction Distribution}\big(\sigma^2(\Delta^*)\big)$$

3. Accept θ' if

$$\Delta^* + X_{corr} > 0$$

4. Otherwise repeat θ

We still have some questions

- How to sample from correction distribution?
- What error we have if we assume that

$$\Delta^* = \Delta + X_{norm}, \qquad X_{norm} \sim \mathcal{N}\left(0, \sigma^2(\Delta^*)\right)$$

Instead of

$$\Delta^* = \Delta + X_{norm}, \qquad X_{norm} \sim \overline{\mathcal{N}}\left(0, \sigma^2(\Delta^*)\right)$$

We still have some questions

- How to sample from correction distribution?
- What error we have if we assume that

$$\Delta^* = \Delta + X_{norm}, \qquad X_{norm} \sim \mathcal{N}\left(0, \sigma^2(\Delta^*)\right)$$

Instead of

$$\Delta^* = \Delta + X_{norm}, \qquad X_{norm} \sim \overline{\mathcal{N}}\left(0, \sigma^2(\Delta^*)\right)$$

PDF of the correction distribution

$$X_{log} = X_{norm} + X_{corr}, \qquad X_{norm} \sim \mathcal{N}(0, \sigma^2)$$

Φ_σ – CDF of $\mathcal{N}(0, \sigma^2)$

C_σ – PDF of corresponding correction distribution

$. * .$ – convolution operation

$$CDF(X_{norm} + X_{corr}) = \Phi_\sigma * C_\sigma$$

$$\mathbb{P}\{X + Y < t\} = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{t-x} p_x(x) p_y(y) dy = \int_{-\infty}^{+\infty} dx p_x(x) F_y(t - x)$$

PDF of the correction distribution

$$C_\sigma^* = \operatorname*{argmin}_{C_\sigma} \sup |\Phi_\sigma * C_\sigma - S|$$

S −CDF of Logistic Distribution

After discretization on the uniform grid [-20,20]

$$C_\sigma^* = \operatorname*{argmin}_{C_\sigma} \max_{\mathrm{i} \in \mathrm{I}} \left| \sum_{j \in J} \Phi_\sigma(X_i - Y_j) C_\sigma(Y_j) - S(X_i) \right|$$

Define

$$M_{ij} = \Phi_\sigma(X_i - Y_j), \qquad u_j = C_\sigma(Y_j), \qquad v_i = S(X_i)$$

Then

$$u^* = (M^T M + \lambda I)^{-1} M^T v$$

PDF of the correction distribution

S —CDF of Logistic Distribution

After discretization on the uniform grid [-20, 20]

$$C_\sigma^* = \underset{C_\sigma}{\operatorname{argmin}} \left\| \sum_{j \in J} \Phi_\sigma(X_i - Y_j) C_\sigma(Y_j) - S(X_i) \right\|_2^2 + \lambda \sum_j C_\sigma(Y_j)^2$$

Define

$$M_{ij} = \Phi_\sigma(X_i - Y_j), \qquad u_j = C_\sigma(Y_j), \qquad v_i = S(X_i)$$

Then

$$u^* = (M^T M + \lambda E)^{-1} M^T v$$

PDF of the correction distribution

Bu instead of solving

$$u^* = \underset{u}{\operatorname{argmin}} \max_{\mathrm{i} \in \mathrm{I}} |Mu - v|, \qquad u > 0$$

Let’s solve

$$u^* = \underset{u}{\operatorname{argmin}} \|Mu - v\|_2^2 + \lambda \|u\|_2^2$$

$$u^* = (M^T M + \lambda I)^{-1} M^T v$$

And show empirically that error is negligible

Precomputing correction distribution

Note that PDF $C_\sigma(x)$ depends on variance σ^2 of normal distribution $\mathcal{N}(0, \sigma^2)$

$$\underbrace{\Delta + \mathcal{N}(0, \sigma^2(\Delta^*))}_{\approx \Delta^*} + X_{corr}\left(\sigma^2(\Delta^*)\right) =$$

$$\Delta + \underbrace{\mathcal{N}\left(0, \sigma^2(\Delta^*)\right) + \mathcal{N}\left(0{,}1 - \sigma^2(\Delta^*)\right)}_{\mathcal{N}(0{,}1)} + X_{corr}(\sigma^2 = 1)$$

We can imitate standard normal noise by adding $\mathrm{X_{nc}} \sim \mathcal{N}\left(0{,}1 - \sigma^2(\Delta^*)\right)$

Good We can use precomputed $X_{corr}(\sigma^2 = 1)$

Bad We need to sample minibatches until $\sigma^2(\Delta^*) < 1$

We still have some questions

- How to sample from correction distribution?
- What error we have if we assume that

$$\Delta^* = \Delta + X_{norm}, \qquad X_{norm} \sim \mathcal{N}\left(0, \sigma^2(\Delta^*)\right)$$

Instead of

$$\Delta^* = \Delta + X_{norm}, \qquad X_{norm} \sim \overline{\mathcal{N}}\left(0, \sigma^2(\Delta^*)\right)$$

Bounding acceptance probability error

$$X_i = N \log \frac{p(x_i|\theta')}{p(x_i|\theta)} - \sum_i^N \log \frac{p(x_i|\theta')}{p(x_i|\theta)}$$

$$\approx N \log \frac{p(x_i|\theta')}{p(x_i|\theta)} - \frac{N}{b} \sum_i^b \log \frac{p(x_i|\theta')}{p(x_i|\theta)}$$

Authors bound error of acceptance probability

$$\sup_y |\mathbb{P}\{\Delta^* + X_{nc} + X_{corr} < y\} - S(y - \Delta)| \leq \frac{6.4\mathbb{E}|X|^3 + 2\mathbb{E}|X|}{\sqrt{b}} = \varepsilon$$

S −CDF of Logistic Distribution

Bounds stationary distribution

$\hat{p}$, p – stationary distributions of true and approximate transition operators $\hat{\tau}$ and τ

If

$$|\widehat{\mathbb{P}}\{\text{acceptance}\} - \mathbb{P}\{\text{acceptance}\}| < \varepsilon$$

And true operator has contraction property

$$d(\tau q, p) < \eta d(q, p),$$

Where $d(q, p)$ – total variation distance

Then

$$d(\hat{p}\ , p) < \frac{\varepsilon}{1 - \eta}$$

Korattikara, Anoop, Yutian Chen, and Max Welling. "Austerity in MCMC land: Cutting the Metropolis-Hastings budget." *International Conference on Machine Learning*. 2014.

Algorithm

1. Sample candidate $\theta' \sim \mathrm{q}(\theta'|\theta)$
2. Increase minibatch until

$$\sigma^2(\Delta^*) < 1 \qquad \text{and} \qquad \varepsilon < \delta$$

3. Accept θ' if

$$\Delta^* + X_{nc} + X_{corr} > 0$$

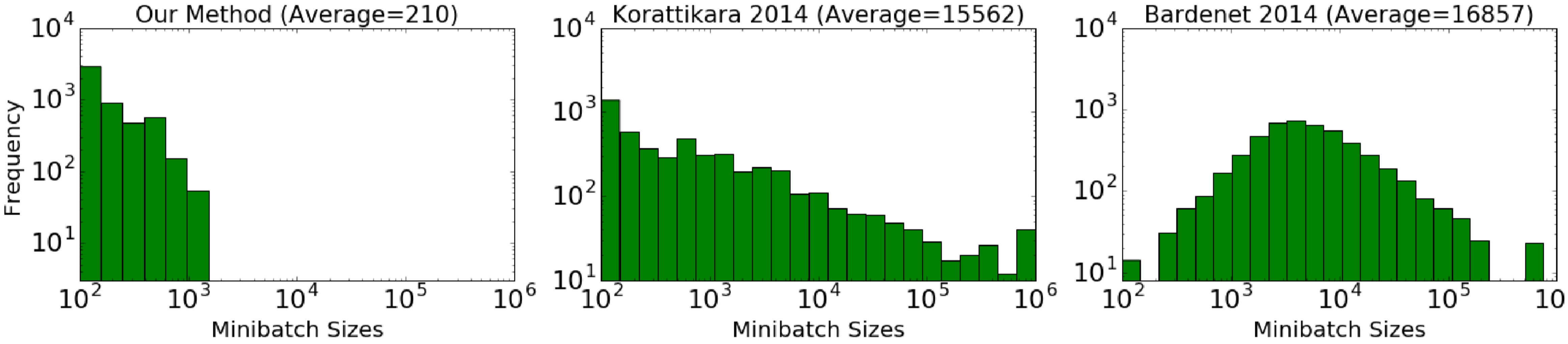
$$X_{nc} \sim \mathcal{N}(0,1 - \sigma^2(\Delta^*))$$

$$X_{corr} \sim \text{Correction Distribution}(\sigma^2 = 1)$$

4. Otherwise keep old θ

Efficiency

Dataset of 10^6 points sampled from mixture of Gaussians



5 min break

Langevin Dynamics

Makes use of the gradient of log-density

$$\Delta\theta_t = \frac{\varepsilon}{2}\nabla \log p(\theta) + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \varepsilon)$$

In Bayesian Inference

$$\Delta\theta_t = \frac{\varepsilon}{2}\left(\nabla \log p(\theta_t) + \boxed{\sum_{i=1}^{N} \nabla \log p(x_i|\theta_t)}\right) + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \varepsilon)$$

Full dataset!

Bayesian Learning via Stochastic Gradient Langevin Dynamics

Max Welling, Yee Whye Teh

Stochastic Gradient Langevin Dynamics

Estimate gradient in Langevin Dynamics on minibatch $= \{x_{t_1}, \dots, x_{t_n}\}$

$$\Delta\theta_t = \frac{\varepsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p\left(x_{t_i} \middle| \theta_t\right)\right) + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \varepsilon_t)$$

~~Proof~~ Intuitive Analysis of SGLD

True gradient

$$\mathrm{g}(\theta) = \nabla \log p(\theta) + \sum_{i=1}^{N} \nabla \log p(x_i|\theta)$$

Deviations

$$\mathrm{h_t}(\theta) = \nabla \log p(\theta) + \frac{N}{n} \sum_{i=1}^{n} \nabla \log p\left(x_{t_i} \middle| \theta\right) - g(\theta),$$

$$h_t(\theta) \sim \overline{\mathcal{N}}(0, V_t(\theta))$$

Intuitive Analysis of SGLD

Given

$$\sum_{t=1}^{\infty} \varepsilon_t = \infty \qquad \sum_{t=1}^{\infty} \varepsilon_t^2 < \infty$$

We can find subsequence $t_1 < t_2 < \cdots$ such that

$$\lim_{s \to \infty} \sum_{t=t_s+1}^{t_{s+1}} \varepsilon_t = \varepsilon_0,$$

Where $0 < \varepsilon_0 < 1$ is initial step

After one step

$$\Delta\theta_t = \frac{\varepsilon_t}{2}\big(g(\theta_t) + h_t(\theta_t)\big) + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \varepsilon_t)$$

After several steps

$$\Delta\theta = \sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2}\big(g(\theta_t) + h_t(\theta_t)\big) + \mathcal{N}\left(0, \sum_{t=t_s+1}^{t_{s+1}} \varepsilon_t\right) =$$

$$\sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2}\Big(g(\theta_t) - g\big(\theta_{t_s}\big) + g\big(\theta_{t_s}\big)\Big) + \sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2} h_t(\theta_t) + \mathcal{N}\left(0, \sum_{t=t_s+1}^{t_{s+1}} \varepsilon_t\right)$$

For s big enough

One step of Langevin dynamics

Langevin noise

Has order of $O\left(\sqrt{\varepsilon_0}\right)$

$$\Delta\theta = \overbrace{\frac{\varepsilon_0}{2} g\left(\theta_{t_s}\right) + \mathcal{N}(0, \varepsilon_0)} + \sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2}\left(g(\theta_t) - g\left(\theta_{t_s}\right)\right) + \sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2} h_t(\theta_t)$$

Systematic error

Non-zero mean

Has order of $O(\varepsilon_0^2)$

"Random" error

Zero-mean

Has order of $O(\varepsilon_0)$

Bounding systematic error

Firstly we bound $\left\|\theta_t - \theta_{t_s}\right\|_2 \ \forall\, t \in [t_s + 1, t_{s+1}]$

$$\left\|\theta_t - \theta_{t_s}\right\|_2 \le \left\| \sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2}\big(g(\theta_t) + h_t(\theta_t)\big) + \mathcal{N}\left(0, \sum_{t=t_s+1}^{t_{s+1}} \varepsilon_t\right) \right\|_2$$

$$\le \sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2} \|g(\theta_t)\|_2 + \left\| \sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2} h_t(\theta_t) \right\|_2 + \|\mathcal{N}(0, \varepsilon_0)\|_2 = O(\varepsilon_0)$$

Assuming that $\|\mathrm{g}(\theta)\|_2$ and $\|h_t(\theta_t)\|$ have some upper bounds

Bounding systematic error

Assuming gradient Lipschitz continuity $\left\|g(\theta_t) - g\left(\theta_{t_s}\right)\right\|_2 \leq L\left\|\theta_t - \theta_{t_s}\right\|_2$

$$\left\|\sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2}\left(g(\theta_t) - g\left(\theta_{t_s}\right)\right)\right\| \leq O(\varepsilon_0) \sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2} = O(\varepsilon_0^2)$$

$O(\varepsilon_0^2)$ is negligible compared to $\frac{\varepsilon_0}{2} g\left(\theta_{t_s}\right)$

$$\sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2}\left(g(\theta_t) - g\left(\theta_{t_s}\right)\right) \ll \frac{\varepsilon_0}{2} g\left(\theta_{t_s}\right)$$

Systematic error True gradient

Analysis of "random" error

$$\sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2} h_t(\theta_t) \sim \overline{\mathcal{N}}\left(0, \sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t^2}{4} V_t(\theta_t)\right)$$

$$\sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t^2}{4} V_t(\theta_t) \leq V\left(\sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2}\right)^2 = V\varepsilon_0^2$$

Variance $V\varepsilon_0^2$ is negligible compared to ε_0

$$\underbrace{\sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2} h_t(\theta_t)}_{\text{“Random” error}} \ll \underbrace{\mathcal{N}(0, \varepsilon_0)}_{\text{Langevin dynamics noise}}$$

Several steps of SGLD ≈ one step of LD

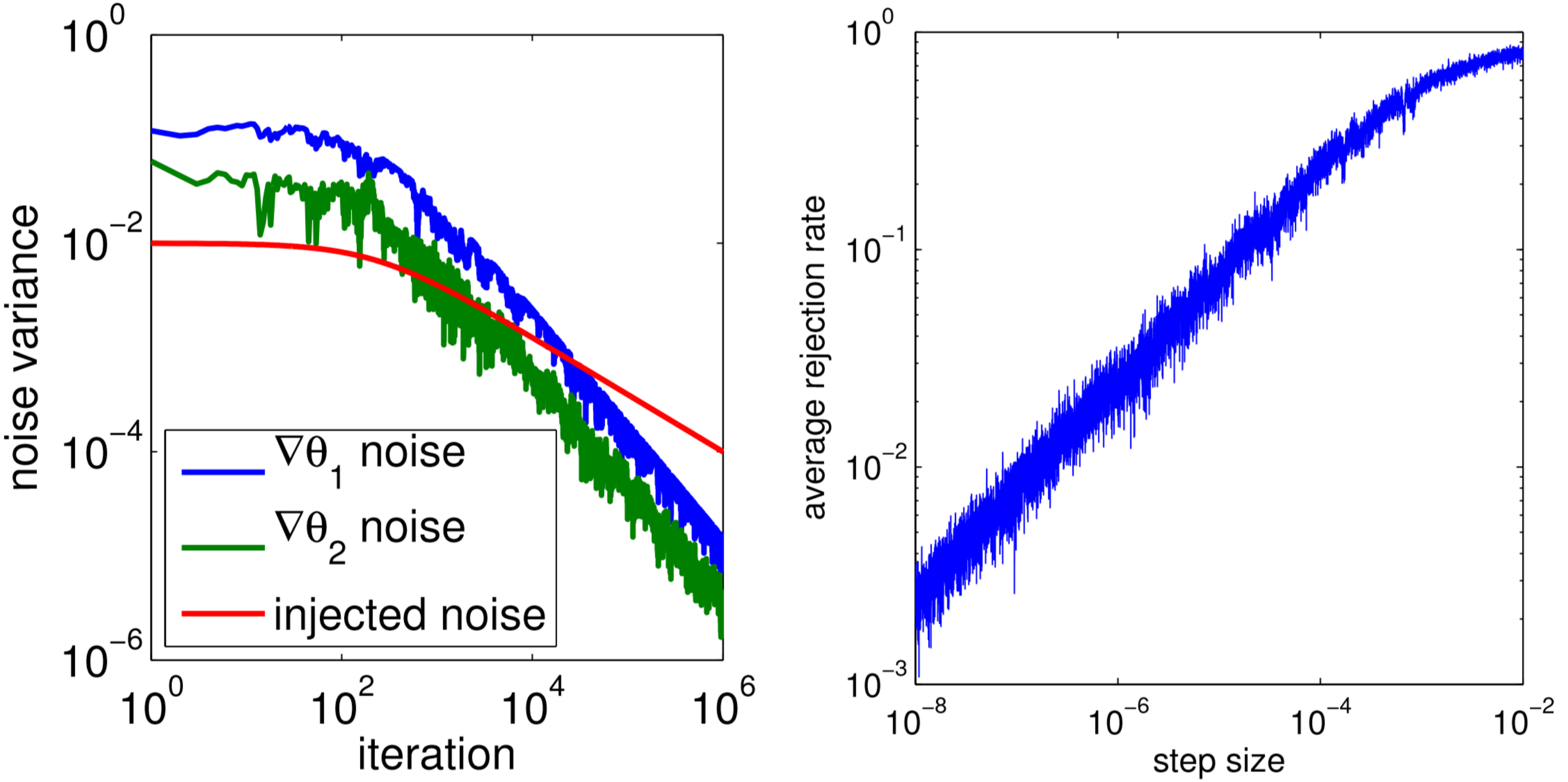
$$\Delta\theta = \sum_{t=t_s+1}^{t_{s+1}} \frac{\varepsilon_t}{2}\big(g(\theta_t) + h_t(\theta_t)\big) + \mathcal{N}\left(0, \sum_{t=t_s+1}^{t_{s+1}} \varepsilon_t\right) \approx$$
$$\approx \frac{\varepsilon_0}{2} g\left(\theta_{t_s}\right) + \mathcal{N}(0, \varepsilon_0)$$

For ε_0 small enough we can ignore M-H test

OR

Perform minibatch M-H test!

Empirical analysis on toy problem



Conclusion

- Stochastic MCMC is a new-generation methods of sampling from posterior conditioned on large dataset
- Makes use of mini-batching and stochastic optimization
- Higher rejection rates but MUCH cheaper iterations

Acceptance rate