# Stochastic Gradient Langevin Dynamics

Duruo Li

CONTENTS

# Contents

# Intuitions: Why?

Stochastic gradient langevin dynamics (SGLD):

**Bayesian Methods in BIG DATA Era**

Difficulty: MCMC (whole dataset)

Advantage: avoid/mitigate overfitting + measure uncertainty

↓

**Goal**: simulate the posterior distribution efficiently

# Intuitions: how come?

- 1. Stochastic gradient descent

a. stochastic: quick convergence

b. optimization: gradient descent => best “point”

- 2. Langevin dynamics

MCMC(sampling) => best “distribution”

1) Meet the goal

2) “Sharing area”

# Intuitions: methods

- Stochastic Gradient Descent (SGD)

0. Gradient descent

**Goal**: “best” parameter w => minimize Loss(w;X)

**Intuition**: slide to the bottom ; direction <= gradient

1. Stochastic

Q: what is stochastic? A: selection of data

# Intuitions: methods

Langevin Dynamics (LD)

0. Stochastic differential equation (guassian process)

1. Sampling (MCMC)

**Goal**: a chain (posterior distribution)

**Intuition**: exist a potential field/force s.t. ...

slide, not stop at bottom(MAP), oscillate around

# Technical Details: how to seam?

SGLD:

$$\Delta\theta_t = \frac{\boxed{\epsilon_t}}{2}\left(\nabla \log p(\theta_t) + \boxed{\frac{N}{n}}\sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta_t)\right) + \boxed{\eta_t}$$

$$\eta_t \sim N(0, \epsilon_t) \qquad (4)$$

1. Stochastic Gradient Decent

$$\Delta\theta_t = \frac{\boxed{\epsilon_t}}{2}\left(\nabla \log p(\theta_t) + \boxed{\frac{N}{n}}\sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta_t)\right) \qquad (1)$$

$$\sum_{t=1}^{\infty} \epsilon_t = \infty \qquad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty \qquad (2)$$

2. Langevin Dynamics

$$\Delta\theta_t = \frac{\epsilon}{2}\left(\nabla \log p(\theta_t) + \sum_{i=1}^{N} \nabla \log p(x_i|\theta_t)\right) + \boxed{\eta_t}$$

$$\eta_t \sim N(0, \epsilon) \qquad (3)$$

**Difficulty**:

1) discrete “LD” 2) converge to target distribution?

# Technical Details: difficulty 1

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n}\nabla \log p(x_{ti}|\theta_t)\right) + \eta_t$$

$$\eta_t \sim N(0, \epsilon_t) \qquad (4)$$

- Discrete “LD”

**Solution**: add accept/reject procedure (MH)?

* when $\varepsilon_t$ is very small, rejection rate ≈ 0

$$\sum_{t=1}^{\infty}\epsilon_t = \infty \qquad \boxed{\sum_{t=1}^{\infty}\epsilon_t^2 < \infty} \qquad (2)$$

# Technical Details: difficulty 2

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n}\nabla \log p(x_{ti}|\theta_t)\right) + \eta_t$$
$$\eta_t \sim N(0, \epsilon_t) \qquad (4)$$

- Converge to target distribution?

**Solution**: $\sum_{t=1}^{\infty}\epsilon_t = \infty \qquad \sum_{t=1}^{\infty}\epsilon_t^2 < \infty \qquad (2)$

1st phase: Stochastic gradient (“speed up”) →

2nd phase: Langevin dynamics (sampling)

# Experiment: mixture of Guassians

$$\theta_1 \sim N(0, \sigma_1^2)\ ; \qquad \theta_2 \sim N(0, \sigma_2^2)$$

$$x_i \sim \tfrac{1}{2} N(\theta_1, \sigma_x^2) + \tfrac{1}{2} N(\theta_1 + \theta_2, \sigma_x^2)$$

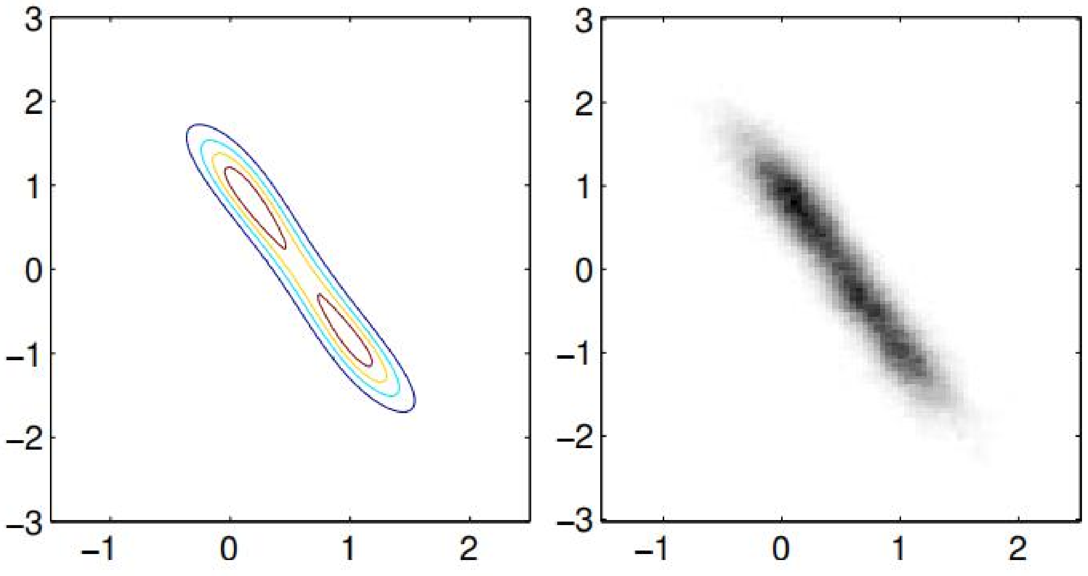

*Figure 1.* True and estimated posterior distribution.

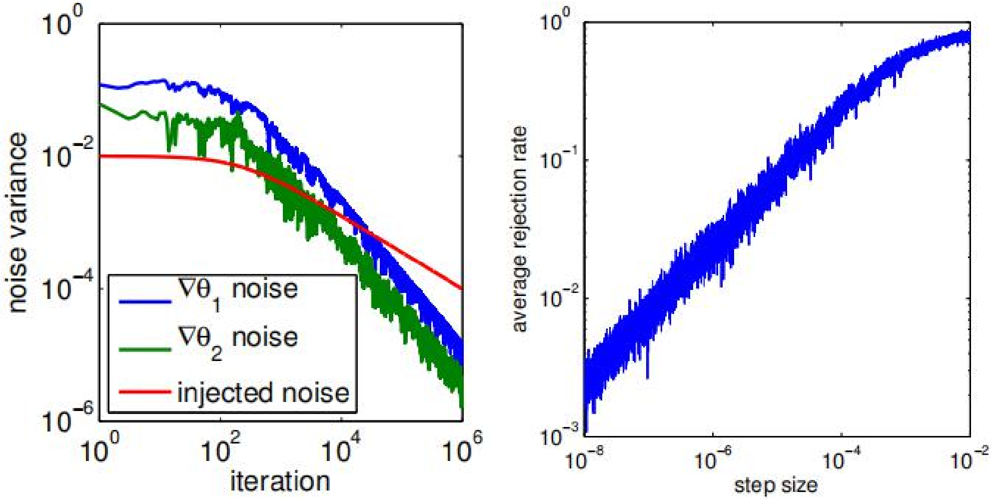


*Figure 2.* Left: variances of stochastic gradient noise and injected noise. Right: rejection probability versus step size. We report the average rejection probability per iteration in each sweep through the dataset.

# Future

**Existing problem**: step size $\varepsilon_t \rightarrow 0$, change slowly

**Possible solutions**:

1) Threshold: rejection rate ≈ 0 => ε stop decreases

2) Other MCMC methods: use SGD burn-in, then...

Thank You :)