Bayesian learning via stochastic gradient Langevin dynamics

Zhengyang LAN

Université de Lille

January 11, 2021

Introduction and Overview

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

Mini batch Stochastic Gradient:

- Can converge to the MAP but no uncertainty.

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

Langevin dynamics:

- Have uncertainty but need to compute full gradient at each iteration.

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

Stochastic Gradient Langevin dynamics:

- Can sample from true posterior, just need to compute gradient of mini batch at each iteration.

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

Sampling threshold

From Stochastic Gradient phase to Langevin dynamics phase.

When?

Condition proposed:

$$\frac{\epsilon_t N^2}{4n}\lambda_{\max}(M^{\frac{1}{2}} V_s M^{\frac{1}{2}}) = \alpha \ll 1$$

The $V(\theta_t)$ can be estimated by:

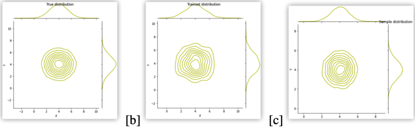
$$V(\theta_t) \approx \frac{N^2}{n^2}\sum_{i=1}^{n}(s_{ti} - \bar{s}_t)(s_{ti} - \bar{s}_t)^T$$

For threshold:

$$I_F \approx NV_s \Rightarrow \Sigma_\theta \approx I_F^{-1} \Rightarrow \epsilon_t \approx \frac{4\alpha n}{N}\lambda_{min}(\Sigma_\theta)$$

Apply on real data

- 2D normal distribution



[a] [b] [c]

(a) True distribution (b) Train distribution (c) Sample distribution

Figure 1: Distribution

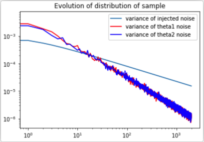


Figure 2: Variance of noise

- Linear regression

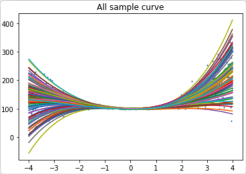


Figure 3: All sample curve

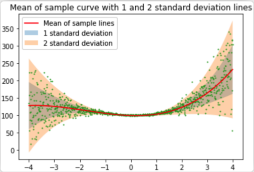


Figure 4: Mean curve and 1,2 standard deviation lines

Thinking

- $Vs \propto \frac{N^2}{n}$
- More computation for threshold - More *batchsized* $\times\ \theta$ times gradients in each iteration.

Creativity

- Different MCMC
- Different optimization algorith

Conclusion

- Main contribution - idea of sample follow true posterior.
- Inspired - Keep up with the cutting edge

Thank you for your attention!