BAYESIAN LEARNING VIA STOCHASTIC GRADIENT LANGEVIN DYNAMICS

Zakaria Boulkhir & Omar Iken

Bayesian Machine Learning
Master 2 Data Science

February 2, 2022

Université de Lille

Context

MCMC for Bayesian posterior inference

$$f(\theta) = p(\theta|x_1, \ldots, x_N) = \frac{p(\theta) \prod_{i=1}^{N} p(x_i \theta)}{p(x_1, \ldots, x_N)} \propto p(\theta) \prod_{i=1}^{N} p(x_i|\theta) \tag{1}$$

Sampling Process

- Initialize θ_0
- At each step t, draw a candidate θ'.
- Compute acceptance rate: $\rho = \min\left\{1, \frac{f(\theta')q(\theta_t|\theta')}{f(\theta_t)q(\theta'|\theta_t)}\right\}$
- Draw $u \sim \mathrm{N}(0, 1)$, and accept if $u < \rho$, otherwise, reject.

If we samples long enough, we will arrive at a (stationary) distribution matching the posterior!

Motivation

- Emergence of large-scale datasets.
- Bayesian MCMC methods require access to the entire dataset, which is time consuming (computations of ρ).
- Stochastic optimization methods: very successful for large scale machine learning by using batches.

Outline

1 Stochastic Gradient Langevin Dynamics (SGLD)

2 Experiments

3 Thoughts & Takeaways

Stochastic Gradient Langevin Dynamics (SGLD)

Preliminaries

Stochastic Gradient Optimization [Robbins & Monro, 1951]

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

Robbins & Monro Conditions: $\sum_{t=1}^{\infty} \epsilon_t = \infty, \qquad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty$

It does not really do Bayesian posterior inference. Since θ converges to a single value while we want instead a distribution.

Langevin Dynamics [Neal, 2010] Langevin dynamics injects noise into the parameter updates in such a way that the trajectory of the parameters will converge to the full posterior distribution rather than just the maximum a posteriori mode.

$$\Delta\theta_t = \frac{\epsilon}{2} \left(\nabla \log p(\theta_t) + \sum_{i=1}^{N} \nabla \log p(x_i|\theta_t) \right) + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \epsilon) \tag{3}$$

It uses full-batch! Let's instead use a mini-batch

Stochastic Gradient Langevin Dynamics [Welling & Teh., 2011]

SGLD

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

1. Change the gradient of the Langevin dynamics to mini-batch estimation.
2. Inject noise into the parameter updates such that the parameter trajectory converges to the full posterior distribution.
3. Make the step size go to zero.

$$\epsilon_t \xrightarrow[t\to\infty]{} 0 \implies \rho \to 1 \implies \text{No more Metropolis-Hastings acceptance test.}$$

Posterior sampling phase

- Initial phase: the stochastic gradient noise will dominate and the algorithm will mimic a stochastic gradient ascent algorithm.
- Later phase: the injected noise will dominate and the algorithm will imitate a Langevin dynamics MH algorithm.

$$\alpha = \frac{\epsilon_t N^2}{4n}\lambda_{max}(M^{1/2}V_s M^{1/2}) \ll 1 \tag{5}$$

Where λ_{max} is the largest eigenvalue, M is the preconditioning matrix and V is empirical variance of the parameters.

Experiments

Experiments [1]

Mixture of Gaussians

$$x_i \sim N(\theta_1, \sigma_x^2) + \frac{1}{2} N(\theta_1 + \theta_2, \sigma_x^2) \text{ where } \sigma_x^2 = 2$$

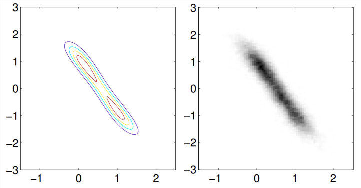

Figure 1: True and estimated posterior dist.

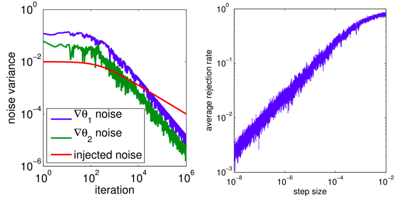

Figure 2: Convergence to Langevin Dynamics

Logistic Regression

- $p(y_i|x_i) = \sigma(y_i \beta^T x_i)$
- $\frac{\partial}{\partial\beta} \log p\left(y_i \mid x_i\right) = \sigma\left(-y_i \beta^T x_i\right) y_i x_i$
- $p(beta) = -\operatorname{sign}(\beta)$

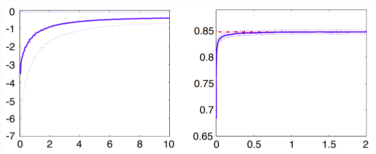

Figure 3: Average log joint probability per data item (left) and accuracy on test set (right)

Thoughts & Takeaways

Thoughts & Takeaways

About the paper

- Max Welling & Yee Whye Teh.
- ICML: Proceedings of the 28th International Conference on International Conference on Machine Learning.
- June 2011.

Merits of the paper

- The paper is will written and most of the parts are well explained.
- The experiments are done using a variety of basic and well known models.
- Details of the experiments are provided.

Weaknesses

- We felt a big lack of theoretical groundings, as most proofs are just explained intuitively.
- The part about Independent Components Analysis is much complicated to process.
- Details of the experiments are given, but no code is provided for reproducibility.

References

[1] Max Welling and Yee W Teh. "Bayesian learning via stochastic gradient Langevin dynamics". In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer. 2011, pp. 681–688.