Research proposal: Implementing variance reduced stochastic gradient Langevin dynamics

Ayoub El Hanchi

October 14, 2019

1 Background and Motivation

Optimization and sampling algorithms are the core algorithmic tools of machine learning and statistics. With the rise of deep learning, stochastic gradient descent (SGD) emerged as the clear winner among existing optimization algorithms, primarily due to its sub-linear per iteration time complexity in the number of data points, as well as its surprising ability to find solutions that generalize well even in very complex models. It was also observed to outperform more complex optimization algorithms that optimize better [15]. This suggests that SGD is not simply optimizing. One hypothesis is that it is sampling from a rough approximation to the posterior distribution [9].

On the sampling side, variational inference (VI) methods have taken over the more traditional Markov chain Monte Carlo (MCMC) approaches due to their ability to scale well with the number of data points, using SGD to optimize a variational objective. Here, on the other hand, it has been very difficult to compare the performance of VI approaches (which come with no guarantees on the error) with that of the more accurate MCMC methods on large scale problems. Traditional MCMC methods usually rely on an accept/reject step, which is very sensitive to stochasticity, making them unpractical for large scale problems.

Assuming that SGD is indeed sampling from an approximate posterior, it is reasonable to ask whether an algorithm that explicitly targets sampling from the posterior distribution can perform better than SGD. For it to be useful, any such algorithms must have a sub-linear per iteration time complexity in the number of data points. As a more accurate alternative to VI, [13] introduced the stochastic gradient Langevin dynamics (SGLD) algorithm as a way of accomplishing this, with update equation:

$$\theta_{k+1} = \theta_k - \alpha \nabla \widehat{f(\theta_k)} + \eta_k \quad \eta_k \sim \mathcal{N}(0, 2\alpha) \tag{1}$$

where $f(\theta)$ is the negative log posterior, $\alpha > 0$ is the step size, and $\widehat{\nabla f(\theta)}$ is an estimator of $\nabla f(\theta)$, which is usually taken to be the usual SGD estimator.

The difference between the SGLD update (1) and the SGD update is simply the addition of independent Gaussian noise. This strong resemblance is not just superficial. In fact, while SGD is an approximate version of gradient descent (GD) which in turn can be viewed as steepest descent under the euclidean norm, SGLD is an approximate version of Langevin dynamics which in turn can be interpreted as steepest descent under the 2-Wasserstein metric on the space of probability measures with objective the KL divergence between the distribution of the current iterate and the posterior distribution [8, 14].

SGLD has been applied successfully in practice, yielding similar performance compared to SGD, but it has failed to fulfill the promise of accurate sampling from the posterior distribution [10, 2]. This is due to the large noise introduced by the use of stochastic gradients, which severely impacts the accuracy of the algorithm.

Variance reduction techniques [12, 7, 5] were specifically designed in the optimization literature to tackle this problem and to accelerate convergence of SGD, but their popularity in training current models remains low given that they do not have the same generalization properties as SGD. In the case of SGLD, on the other hand, we do not have such concerns since we assume that more accurate sampling will yield better performance, and therefore variance reduction methods are very relevant. Luckily, given the resemblance between SGLD and SGD, adapting variance reduction technique for SGLD is straightforward, yielding algorithms with provably better convergence properties [6, 4, 3].

2 Objectives and Evaluation criteria

Despite the attractive theoretical properties of these variance reduced algorithms, no public implementation of them is currently available. Furthermore, they have not been extensively experimented with in the context of large problems. This is due in part to the following problems:

- Automatic differentiation engines do not offer per-example gradients, which are required in many variance reduced algorithms. The naive way of requesting each gradient sequentially is too wasteful and scales very poorly.
- Some variance reduction technique have prohibitive memory cost, requiring the storage of all the per-example gradients.

My goals in this project will be to:

- Propose solutions to the above problems.

- Implement these algorithms along with my solutions in PyTorch [11].
- Test these algorithms on standard large scale problems and evaluate their performance.
- If time permits, explore how my implementation can be used in Pyro[1], and use it to compare the accuracy of these algorithms with that of the NUTS sampler and variational methods in the context of graphical/hierarchical models.

I will know that I have succeeded if I am able to reach the first three goals stated above.

References

[1] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. oct 2018.

[2] Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of Stochastic Gradient Langevin Dynamics. nov 2018.

[3] Niladri S. Chatterji, Nicolas Flammarion, Yi-An Ma, Peter L. Bartlett, and Michael I. Jordan. On the Theory of Variance Reduction for Stochastic Gradient Monte Carlo. feb 2018.

[4] Arnak S. Dalalyan and Avetik G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. sep 2017.

[5] Aaron Defazio Ambiata, Francis Bach, and Simon Lacoste-Julien. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. Technical report.

[6] Avinava Dubey, Sashank J Reddi, Barnabás Póczos, Alexander J Smola, Eric P Xing, and Sinead A Williamson. Variance Reduction in Stochastic Gradient Langevin Dynamics. Technical report.

[7] Rie Johnson and Tong Zhang. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. Technical report.

[8] Richard Jordan, David Kinderlehrer, Felix Otto, and Siam J M A T H An A L. THE VARIATIONAL FORMULATION OF THE FOKKER-PLANCK EQUATION * In memory of Richard Duffin. Technical Report 1, 1998.

[9] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic Gradient Descent as Approximate Bayesian Inference. apr 2017.

[10] Tigran Nagapetyan, Andrew B. Duncan, Leonard Hasenclever, Sebastian J. Vollmer, Lukasz Szpruch, and Konstantinos Zygalakis. The True Cost of Stochastic Gradient Langevin Dynamics. jun 2017.

[11] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary Devito Facebook, A I Research, Zeming Lin, Alban Desmaison, Luca Antiga, Orobix Srl, and Adam Lerer. Automatic differentiation in PyTorch. Technical report.

[12] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing Finite Sums with the Stochastic Average Gradient. sep 2013.

[13] Max Welling, D Bren, and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. Technical report, 2010.

[14] Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. feb 2018.

[15] Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The Marginal Value of Adaptive Gradient Methods in Machine Learning. may 2017.