Beyond SGD

Approximate large scale inference using variance reduced stochastic gradient langevin dynamics

Ayoub El Hanchi

October 8, 2019

Introduction

- Why is SGD so successful in training our models ?
- One hypothesis is that SGD samples from an approximate posterior, thereby avoiding overfitting and reaching solutions that generalize well.
- If this is the case, can we do better than SGD if we sample from a better approximation to the posterior ?
- MCMC is accurate but slow, VI is fast but offers no guarantees on the error. Both are rigid with respect to these properties.

Background

Denote by P the posterior distribution of the parameters given the data:

- MCMC builds a Markov chain $(\Theta_i)_{i\in\mathbb{N}}$ such that $\Theta_\infty \sim P$. Intuitively, it searches the space of probability measures for the posterior. Is there an optimal search direction ?

- VI considers a class of distributions D parameterized by $\phi \in \mathbb{R}^n$ and then solves the optimization problem:

$$\min_{Q\in D} D_{KL}(Q||P)$$

  using gradient descent. What we really want is to extend D to the whole space of probability measures. Can this be done in space ?

Langevin dynamics

Let $p(\Theta = \theta \mid X = x) = e^{-f(\theta)}$ be the density of the posterior distribution of the parameters given the data. Then it has been shown that the following update equation:

$$\Theta_{k+1} = \Theta_k - \alpha \nabla f(\Theta_k) dt + \mathcal{N}(0, 2\alpha)$$

is steepest descent in the space of probability measures with objective $D_{KL}(\cdot || P)$ and the 2-Wasserstein metric.

Variance reduction

- Given the striking similarity with gradient descent, it has been proposed to use only mini-batches to evaluate $\nabla f(\Theta_k)$, yielding the stochastic gradient Langevin dynamics algorithm (SGLD).
- Strictly in terms of performance, SGLD has been applied with good success in many problems, but the noise coming from the gradient estimation greatly affects the accuracy of the algorithm.
- Variance reduction techniques (such as SAG, SAGA, SVRG) allow significant reduction in this variance, yielding a more accurate algorithm.

Proposed project

Despite the strong theoretical foundations of variance reduced SGLD, no public implementation of these algorithms currently exists in the major frameworks. This is due mainly to the following:

- Some variance reduction techniques require the computation and storage of all individual gradients, which is not doable in large problems.
- Most frameworks do not support the computation of individual gradients. Doing it the naive way requires as many passes on the computational graph as there are examples.
- No one took the initiative yet ?

Proposed project

The goals of my proposed project are:

- Solving these problems and providing an efficient implementation of these algorithms in PyTorch.
- Comparing the accuracy of these algorithms with that of variational methods and more traditional MCMC.
- Evaluating the performance of these algorithms applied on standard large scale problems.