Adaptive Importance Sampling for Finite-Sum Optimization and Sampling.

Ayoub El Hanchi and David Stephens

McGill University

Setup

Given $f : \mathbb{R}^d \to \mathbb{R}$ of finite-sum form:

$$f(x) \coloneqq \sum_{i=1}^{N} f_i(x)$$

we want to efficiently:

- minimize it if it is an objective function using SGD:

$$x_{t+1} = x_t - \alpha_t \hat{g}^t$$

- sample from the distribution with density $e^{-f(x)}$ using SGLD:

$$x_{t+1} = x_t - \alpha_t \hat{g}^t + \xi_t \qquad \xi_t \sim \mathcal{N}(0, 2\alpha_t)$$

where in both cases:

- we run the algorithm for a total number of iterations T.
- $\hat{g}^t$ is an unbiased estimator of $\nabla f(x_t)$.
- $(\alpha_t)_{t=1}^T$ is a sequence of decreasing step sizes.

Main Idea

Let $\left\|h_i^t\right\|_2$ be the last observed gradient norm of the function f_i at time t. Then we consider the following surrogate cost function for each time step t:

$$\tilde{c}_t(p) \coloneqq \sum_{i=1}^{N} \frac{1}{p_i} \left\|h_i^t\right\|_2^2$$

Naively, we could just minimize this surrogate cost at each time step. But we run the risk of assigning zero probability to some index, from which we cannot recover. Instead, we perform the optimization over the restricted simplex:

$$\Delta(\varepsilon_t) \coloneqq \left\{ p \in \mathbb{R}^N \mid p_i \geq \varepsilon_t, \ \sum_{i=1}^{N} p_i = 1 \right\}$$

for some specified decreasing sequence $\{\varepsilon_t\}_{t=1}^T$. Our proposed algorithm can therefore be written as:

$$p^t \in \operatorname*{argmin}_{p \in \Delta(\varepsilon_t)} \sum_{i=1}^{N} \frac{1}{p_i} \left\|h_i^t\right\|_2^2$$

We give an efficient algorithm for the computation of this sequence in the paper.

Estimators

The usual estimator is formed by:

- sampling I_t uniformly from $[N] \coloneqq \{1, \dots, N\}$.
- defining:

$$\hat{g}^t = N \nabla f_{I_t}(x_t)$$

But we can consider more general unbiased estimators by:

- sampling I_t according to a specified distribution p^t on $[N]$.
- defining:

$$\hat{g}^t = \frac{1}{p_{I_t}^t} \nabla f_{I_t}(x_t)$$

It is easy to show that if we had access to the all the gradients $(\nabla f_i(x_t))_{i=1}^N$, then the variance minimizing sequence of distributions is given by:

$$p_i^t \propto \left\|\nabla f_i(x_t)\right\|_2$$

Of course we don't have access to all the individual gradients. How can we design a sequence of distributions $(p^t)_{t=1}^T$ that well approximates the variance minimizing ones ?

Theory

Under appropriate assumptions, we can show that our proposed sequence achieves sub-linear dynamic regret for both SGD and SGLD.

Theorem. *Assuming the functions $\{f_i(x)\}_{i=1}^N$ are smooth and have bounded gradients, if SGD is run with the given probabilities $\{p^t\}_{t=1}^T$ and step-sizes $\mathcal{O}(1/t)$, then:*

$$\mathbb{E}\left[Regret_D(T)\right] \leq \mathcal{O}(T^{2/3})$$

and for SGLD:

$$\mathbb{E}\left[Regret_D(T)\right] \leq \mathcal{O}(T^{5/6})$$

under the same conditions.

The validity of the theorem depends crucially on the choice of the sequence $(\varepsilon_t)_{t=1}^T$. As an example, for SGD with step sizes $O(1/t)$, one can pick the sequence:

$$\varepsilon_t \coloneqq \frac{1}{N^{2/3}(N+t)^{1/3}}$$

to satisfy the theorem.

Online Learning Formulation

We embed the problem in the online learning framework. At each step t, we:

- choose a distribution p^t.
- sample $I_t \sim p^t$ and compute $\hat{g}^t$.
- receive $\left\|\nabla f_{I_t}(x_t)\right\|_2$ as feedback.

The cost function is given by the trace of the covariance matrix of the gradient estimator:

$$c_t(p) \coloneqq \sum_{i=1}^{N} \frac{1}{p_i} \left\|\nabla f_i(x_t)\right\|_2^2$$

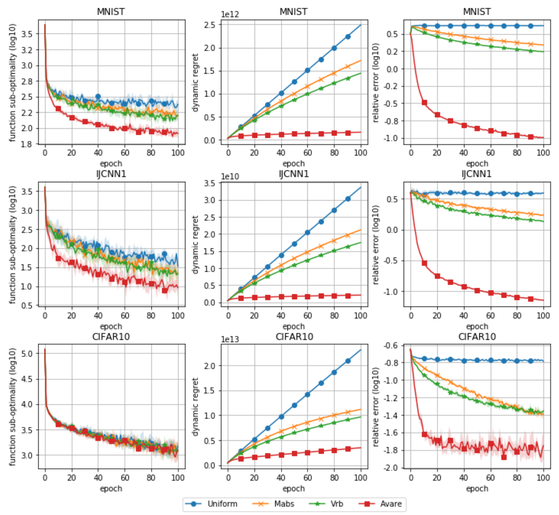
And the goal is to design an algorithm with small regret. Previous work has focused on static regret, we consider the stronger dynamic regret:

$$\text{Regret}_D(T) = \sum_{t=1}^{T} \left[c_t(p^t) - \min_{p \in \Delta} c_t(p) \right]$$

where Δ is the probability simplex in $\mathbb{R}^N$.

Experiments

We named our algorithm *Avare* for adaptive variance minimization. Here we compare its performance with previously developed methods for adaptive importance sampling.
The relative error is defined as: $[c_t(p^t) - \min_{p \in \Delta} c_t(p)] \,/\, \min_{p \in \Delta} c_t(p)$.



References

Stochastic Optimization with Bandit Sampling, Farnood Salehi, L. Elisa Celis, and Patrick Thiran, 2017.
Online Variance Reduction for Stochastic Optimization, Zalan Borsos, Andreas Krause, and Kfir Y. Levy, 2018.