Adaptive Importance Sampling for Finite-Sum Optimization and Sampling with Decreasing Step-Sizes

Ayoub El Hanchi David A. Stephens

McGill University

December 2020

Setup

- Loss/Log-density : $f(x) = \sum_{i=1}^{N} f_i(x)$
- SGD: $x_{t+1} = x_t - \alpha_t \hat{g}^t$
- SGLD: $x_{t+1} = x_t - \alpha_t \hat{g}^t + \mathcal{N}(0, 2\alpha_t)$
- where $\hat{g}^t$ is an estimator for $\nabla f(x_t)$
- We require: $\mathbb{E}[\hat{g}^t] = \nabla f(x_t)$
- We run the algorithm for T iterations, and assume that the step sizes $\{\alpha_t\}_{t=1}^{T}$ are decreasing.

Setup

- The usual estimator is formed by sampling I_t uniformly and defining:
$$\hat{g}^t = N \nabla f_{I_t}(x_t)$$
- More generally, we can sample $I_t \sim p^t$, and estimate the gradient with:
$$\hat{g}^t = \frac{1}{p^t_{I_t}} \nabla f_{I_t}(x_t)$$
.
- Question: How can we choose $\{p^t\}_{t=1}^T$ so as to minimize the variance of $\hat{g}^t \rightarrow$ accelerate convergence ?

Setup

- It is not hard to show that picking

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

  minimizes the variance of $\hat{g}^t$.
- Unfortunately, this means that computing the optimal distribution is as expensive as computing the full gradient.
- Alternatively, we can formulate the problem as an online learning problem and look for a no-regret algorithm.

Online learning formulation

- At each time step t we will choose a distribution p^t from which we sample I_t, and we only get back $\|\nabla f_{I_t}(x_t)\|_2$ as feedback.
- Cost function given by: $c_t(p) = \sum_{i=1}^{N} \frac{1}{p_i} \|\nabla f_i(x_t)\|_2^2$
- Goal: design an algorithm with sub-linear expected 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 $\mathbb{R}^N$.

Main Idea

- While we don't have access to $\|\nabla f_i(x_t)\|_2$, we can store $\|h_i^t\|_2$, the norm of the last seen gradient of the i^{th} function at time t. (à la SAGA)
- Naively, we could then choose $p_i^t \propto \|h_i^t\|_2$, but this runs the risk of assigning near zero probability to some index, from which our algorithm cannot recover.

Main Idea

- Instead, at time t, we lower bound the probability of picking any index by ε_t and use the following sequence:

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

where:

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

- Our analysis suggests a specific decay rate for $\{\varepsilon_t\}_{t=1}^T$ to achieve optimal dynamic regret.
- We also give an explicit algorithm for the computation of $\{p^t\}_{t=1}^T$.

Theory

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.

Experiments