The Problem

\[ x^* = \arg \min_{x \in \mathbb{R}^n} \left[ f(x) \right] \]

We assume \( f \) are differentiable and \( f \) is quasi strongly convex.

Stochastic Reformulation

Stochastic reformulation of (1) is the problem:

\[ \min_{x \in \mathbb{R}^n} \mathbb{E}_{\theta}[f(x)] = \min_{x \in \mathbb{R}^n} \mathbb{E}_{\theta}[\sum_{i=1}^n v_i(x)] \]

where \( \{v_1, \ldots, v_n\} \in \mathbb{R}^n \) is a random vector for which

\[ \mathbb{E}[v_i] = 0, \quad \forall i \in \{1, \ldots, n\} \]

• Equivalence: (2) is equivalent to (1) since \( \mathbb{E}_{\theta}[f(x)] = f(x) \) also note that \( \mathbb{E}_{\theta}[\nabla f(x)] = \nabla f(x) \), which can be seen via

\[ \mathbb{E}_{\theta}[\nabla f(x)] = \nabla f(x) \]

• We propose to solve (1) by applying SGD to (2):

\[ \nabla f(x) - \gamma \sum_{i=1}^n v_i(x) \]

where \( \gamma \) is a step size and \( \gamma > 0 \) is a step size.

Example: Arbitrary Sampling

A sampling is a random set-valued mapping \( S \) with values being subsets of \( \{1, \ldots, n\} \). A sampling is defined by assigning probabilities to all \( 2^n \) subsets of \( \{1, \ldots, n\} \).

• A sampling is proper if \( \mathbb{P}|E| > 0 \) for all \( E \in \{1, \ldots, n\} \).

• Each proper sampling \( S \) gives rise to a sampling vector \( v \):

\[ v = \text{Diag}(v_1, \ldots, v_n) \]

where \( v_i \) is the \( i \)-th standard unit basis vector in \( \mathbb{R}^n \). It is easy to see that \( \mathbb{E}[v_i] = 1 \) Indeed, just notice that \( v_i = 1 \) if \( i \in S \) and \( v_i = 0 \) if \( i \notin S \).

Main Contributions

• We introduce and study a flexible stochastic reformulation (see (2)) of the finite-sum problem (1), and study SGD applied to this reformulation (5).

• We establish linear global convergence of SGD applied to the stochastic reformulation. As a by-product, we establish linear convergence of SGD under the arbitrary sampling paradigm [2].

• Our results require very weak assumptions. In particular, we do not assume bounded second moment of the gradients for every \( x \) (only at \( x^* \), see (8)). We rely on the expected smoothness assumption (7) [3, 4].

• Optimal minibatch size: We establish formulas for the optimal dependence of the stepsize on the minibatch size.

• Learning schedule: We provide a formula for when SGD should switch from a constant stepsize to a decreasing stepsize (see (9)).

• Interpolated models: We extend the findings in [5], and show that optimal minibatch size is 1 for independent sampling and sampling with replacement.

Theorem 1

Choose \( \gamma \in (0, \frac{1}{2}) \), then SGD (5) satisfies:

\[ \mathbb{E}[\|x - x^*\|^2] \leq (1 - \gamma \|x - x^*\|^2) \|x^* - x^*\|^2 + 2\gamma^2 \]

In particular, with stepsize \( \gamma \), we have

\[ k \geq \max \left\{ \frac{2\gamma^2}{\mu}, \frac{\gamma^2}{\|
abla f(x^*)\|^2} \right\} \Rightarrow \mathbb{E}[\|x - x^*\|^2] \leq \epsilon \]

Proof. Let \( x^0 = x_0 \) and \( \gamma \mathbb{E}[\|\nabla f(x^k)\|^2] \). Taking expectation conditioned on \( x^k \) we obtain

\[ \mathbb{E}[\|x^{k+1} - x^*\|^2] \leq (1 - \gamma \|x^k - x^*\|^2) \|x^k - x^*\|^2 + 2\gamma^2 (1 - \gamma \|x^k - x^*\|^2) \|x^k - x^*\|^2 + \gamma^2 \mathbb{E}[\|\nabla f(x^k)\|^2] \]

Assumption 7. Then SGD (5) satisfies:

\[ \mathbb{E}[\|x - x^*\|^2] \leq (1 - \gamma \|x - x^*\|^2) \|x^* - x^*\|^2 + 2\gamma^2 (1 - \gamma \|x^* - x^*\|^2) \|x^* - x^*\|^2 + \gamma^2 \mathbb{E}[\|\nabla f(x^*)\|^2] \]

Sublinear Convergence with Constant and Later Decreasing Step Size

In the next theorem we propose a stepping stepsize strategy: first use a constant stepsize, and at some point step to \( O(1/k) \) stepsize. This leads to \( O(1/k) \) rate.

References


