Lecture 18: Stochastic Optimization

Yudong Chen

1 Setup

The algorithms we’ve seen so far have access to a first order oracle, which returns the exact (sub)gradient at a given point, plus potentially the function value.

\[ x \in X \rightarrow 1\text{st order oracle} \rightarrow g_x \in \partial f(x) \text{ (\(\nabla f(x)\) if \(f\) is differentiable)} \]

Stochastic optimization: We are given a noisy version of the (sub)gradient:

\[ x \in X \rightarrow 1\text{st order stochastic oracle} \rightarrow \tilde{g}(x, \xi) \]

Here \(\tilde{g}(x, \xi)\) is a stochastic estimate of some \(g_x \in \partial f(x)\), where \(\xi\) is a random variable (representing the randomness in the stochastic estimate).

Remark 1. Some models also assume access to stochastic estimates of the function value \(f(x)\). We do not need that here.

1.1 Examples

Example 1. \(\tilde{g}(x, \xi) = g_x + \xi\), where \(\xi\) is additive noise from, e.g., inaccurate measurements in physical systems. Sometimes, the noise is added intentionally (for privacy).

Example 2. Finite sum minimization:

\[ f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \]

and \(n\) is large. We can take \(\tilde{g}(x, \xi) = \nabla f_i(x)\), where \(i\) is an integer sampled uniformly at random from \(\{1, 2, \ldots, n\}\).

Example 3. Empirical risk minimization (ERM): We want to minimize

\[ f(x) = \mathbb{E}_{(x,y)\sim \Pi_{\text{data}}} [l(x;a,b)], \]

but we do not know how to compute the expectation exactly. Suppose we have collected \(n\) data points \((a_i, b_i)\) that come from the distribution \(\Pi_{\text{data}}\). We can consider minimizing the empirical loss

\[ f_{\text{emp}}(x) = \frac{1}{n} \sum_{i=1}^{n} l(x; a_i; b_i). \]

When \(n \to \infty\), \(f_{\text{emp}} \to f\). Here we may view \(\tilde{g}(x, \xi) = \nabla f_{\text{emp}}(x)\) as a noisy estimate of \(\nabla f(x)\).
1.2 Assumptions

Consider the problem

\[ \min_{x \in X} f(x). \]  

(P)

We assume that

- \( f \) is convex and \( M \)-Lipschitz (w.r.t. \( \|\cdot\|_2 \)).
- \( X \) is closed, convex and nonempty. The projection \( P_X \) can be efficiently computed.
- For all \( x \in X \), it holds that
  
  (unbiased estimate) \( \mathbb{E}_\xi [\bar{g}(x, \xi)] = g_x \in \partial f(x) \),

  (bounded variance) \( \mathbb{E}_\xi \|\bar{g}(x, \xi) - g_x\|^2_2 \leq \sigma^2 < \infty \).

2 Stochastic (projected sub)gradient descent

Consider the following S-PSubGD algorithm:

\[ x_{k+1} = \arg\min_{u \in X} \left\{ a_k \langle \bar{g}(x_k, \xi_k), u - x_k \rangle + \frac{1}{2} \|u - x_k\|^2 \right\} \]

\[ = P_X (x_k - a_k \bar{g}(x_k, \xi_k)), \]

where \( a_k > 0 \) is the stepsize to be chosen later.

2.1 Convergence analysis

In the sequel, we assume that \( \xi_0, \xi_1, \ldots, \xi_k, \ldots \) are independent and identically distributed (i.i.d.). To avoid cluttered notation, we introduce the shorthands \( g_k \equiv g_{x_k} \) (true subgradient) and \( \bar{g}_k \equiv \bar{g}(x_k, \xi_k) \) (noisy subgradient).

As in the previous lecture, we analyze the averaged iterate \( x_k^{\text{out}} := \frac{1}{A_k} \sum_{i=0}^{k} a_i x_i \), where \( A_k := \sum_{i=0}^{k} a_i \), and we use the same \( U_k, L_k \) and \( G_k \):

upper bound: \( U_k := \frac{1}{A_k} \sum_{i=0}^{k} a_i f(x_i) \geq f(x_k^{\text{out}}) \),

lower bound: \( L_k := \frac{1}{A_k} \sum_{i=0}^{k} a_i f(x_i) + \frac{1}{A_k} \sum_{i=0}^{k} a_i \langle g_i, x^* - x_i \rangle \leq f(x^*) \),

optimality gap bound: \( G_k := U_k - L_k = -\frac{1}{A_k} \sum_{i=0}^{k} a_i \langle g_i, x^* - x_i \rangle \geq f(x_k^{\text{out}}) - f(x^*) \).

The analysis is similar to last lecture, except that we need to keep track of the stochastic error \( g_k - \bar{g}_k \). We have

\[ A_0 G_0 = -a_0 \langle g_0, x^* - x_0 \rangle , \]
and
\[ A_k G_k - A_{k-1} G_{k-1} = -a_k \langle g_k, x^* - x_k \rangle = a_k \langle g_k, x_k - x_{k+1} \rangle + a_k \langle g_k, x_{k+1} - x^* \rangle \]
\[ = a_k \langle g_k, x_k - x_{k+1} \rangle + a_k \langle \tilde{g}_k, x_{k+1} - x^* \rangle + a_k \langle g_k - \tilde{g}_k, x_{k+1} - x^* \rangle. \]

Note that \( x_{k+1} = P_{\mathcal{X}} (x_k - a_k \tilde{g}_k) \) satisfies the minimum principle:
\[ \langle a_k \tilde{g}_k + x_{k+1} - x_k, x^* - x_{k+1} \rangle \geq 0, \]
hence
\[ a_k \langle \tilde{g}_k, x_{k+1} - x^* \rangle \leq \langle x_{k+1} - x_k, x^* - x_{k+1} \rangle \]
\[ = \frac{1}{2} \| x_k - x^* \|^2 - \frac{1}{2} \| x_{k+1} - x^* \|^2 - \frac{1}{2} \| x_k - x_{k+1} \|^2. \]

It follows that
\[ A_k G_k - A_{k-1} G_{k-1} \]
\[ \leq a_k \langle \tilde{g}_k, x_k - x_{k+1} \rangle + \left( \frac{1}{2} \| x_k - x^* \|^2 - \frac{1}{2} \| x_{k+1} - x^* \|^2 - \frac{1}{2} \| x_k - x_{k+1} \|^2 \right) + a_k \langle g_k - \tilde{g}_k, x_{k+1} - x^* \rangle \]
\[ \leq a_k^2 M^2 + \frac{1}{2} \| x_k - x^* \|^2 - \frac{1}{2} \| x_{k+1} - x^* \|^2 + a_k \langle g_k - \tilde{g}_k, x_{k+1} - x^* \rangle. \]

We take expectation of both sides. By the Law of Iterated Expectation,\(^1\) we can write
\[ \mathbb{E} [\text{RHS}] = \mathbb{E} \left[ \mathbb{E} [\text{RHS} | \xi^{k-1}_0] \right], \]
where \( \xi^{k-1}_0 := (\xi_0, \ldots, \xi_{k-1}) \) denotes all the previous randomness in iterations 0 through \( k-1 \) (not including \( \xi_k \)). Observe that
\[ \mathbb{E} \left[ a_k \langle g_k - \tilde{g}_k, x_{k+1} - x^* \rangle | \xi^{k-1}_0 \right] \]
\[ = a_k \mathbb{E} \left[ \langle g_k - \tilde{g}_k, x_{k+1} \rangle | \xi^{k-1}_0 \right] \]
\[ = a_k \mathbb{E} \left[ \langle g_k - \tilde{g}_k, P_{\mathcal{X}} (x_k - a_k \tilde{g}_k) \rangle | \xi^{k-1}_0 \right] \]
\[ = a_k \mathbb{E} \left[ \langle g_k - \tilde{g}_k, P_{\mathcal{X}} (x_k - a_k \tilde{g}_k) - P_{\mathcal{X}} (x_k - a_k g_k) \rangle | \xi^{k-1}_0 \right] \]
\[ \leq a_k \mathbb{E} \left[ \| g_k - \tilde{g}_k \|^2 \cdot \| P_{\mathcal{X}} (x_k - a_k \tilde{g}_k) - P_{\mathcal{X}} (x_k - a_k g_k) \|^2 | \xi^{k-1}_0 \right] \]
\[ \leq a_k \mathbb{E} \left[ a_k \| g_k - \tilde{g}_k \|^2 | \xi^{k-1}_0 \right] \]
\[ \leq a_k^2 \sigma^2. \]

\(^1\)Also known as Law of Total Expectation, or Tower Rule.
It follows that
\[
\mathbb{E} [A_k G_k - A_{k-1} G_{k-1}] \leq \mathbb{E} \left[ \frac{1}{2} \|x_k - x^*\|_2^2 - \frac{1}{2} \|x_{k+1} - x^*\|_2^2 \right] + \frac{a_k^2 (M^2 + 2\sigma^2)}{2}.
\]

Summing both sides over \(k\) and telescoping, we get
\[
\mathbb{E} \left[ f(x_{\text{out}}^K) - f(x^*) \right] \leq \mathbb{E} [G_K] \leq \frac{\|x_0 - x^*\|_2^2 + (M^2 + 2\sigma^2) \sum_{k=0}^K a_k^2}{2A_K}.
\]

The expression on the right-hand side is the same as what we got the last time for projected sub-gradient descent (PSubGD), except for having \(M^2 + 2\sigma^2\) in place of \(M^2\). The rest of the analysis is similar to that for PSubGD:

- Using constant stepsize \(a_k = \frac{\|x_0 - x^*\|_2}{\sqrt{M^2 + 2\sigma^2} \sqrt{k+1}}, \forall k\), we get an \(O\left(\frac{1}{\sqrt{K}}\right)\) convergence rate.
- Same discussion about anytime algorithm, unknown/unbounded diameter of \(\mathcal{X}\), unknown \(M\), etc.