CS 726: Stochastic Convex Optimization

Jelena Diakonikolas

Fall 2022

All of the methods we have seen so far were working under the assumption that the algorithm has access to the exact value of the gradient or subgradient. This assumption is often violated in practice, especially when it comes to problems arising in data science, where the objective function itself is often expressed as an expectation over an unknown distribution and thus exact (sub)gradient and/or function evaluations are impossible.

Much of the ideas we have seen in previous lectures will extend to the stochastic case. In particular, we can pretty much carry out most of the analysis using the arguments from previous lectures. The only thing that will change is that we will need to deal with bounding stochastic errors. In this lecture, we will see an example of how to do this for projected subgradient descent for nonsmooth convex optimization problems; however, as already mentioned, the same ideas can be transferred to convergence analyses of other algorithms and setups we saw in previous lectures.

1 Setup

We begin by describing the setup we will be working with. Recall that our basic optimization problem is

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}). \tag{P}$$

For this lecture, we will additionally be assuming that:

- The norm associated with the space $\|\cdot\|$ is arbitrary but fixed and its dual norm is $\|\cdot\|_*$;
- f is M-Lipschitz continuous w.r.t. $\|\cdot\|$, convex, and minimized by some $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$;
- Given any point $\mathbf{x} \in \mathcal{X}$, we have access to a subgradient estimate $\widetilde{\mathbf{g}}(\mathbf{x},\xi)$, where ξ is a random variable drawn from an unknown distribution $\mathcal{D}$, independent of history, with the following properties:

$$\begin{aligned} \mathbb{E}_{\xi\sim\mathcal{D}}[\widetilde{\mathbf{g}}(\mathbf{x},\xi)] = \mathbf{g}_{\mathbf{x}}, \ \text{ for some } \ \mathbf{g}_{\mathbf{x}} \in \partial f(\mathbf{x}) \quad &\text{(unbiased estimate)}, \\ \mathbb{E}_{\xi\sim\mathcal{D}}[\|\widetilde{\mathbf{g}}(\mathbf{x},\xi) - \mathbf{g}_{\mathbf{x}}\|_*^2] \le \sigma^2 < \infty \quad &\text{(bounded variance)}. \end{aligned}$$

- $\mathcal{X}$ is closed and convex. We have access to efficiently computable projections for $\mathcal{X}$.

We now provide some examples that give rise to the described stochastic setup (stochastic subgradient estimation).

Example 1.1. Perhaps the most direct example of stochastic (sub)gradient estimate is the case $\widetilde{\mathbf{g}}(\mathbf{x},\xi) = \mathbf{g}_{\mathbf{x}} + \xi$, where ξ is zero-mean finite-variance noise vector. Such examples are encountered in e.g., industrial applications, where ξ represents measurement noise. Sometimes, noise is added intentionally to (sub)gradients in applications involving notions of (differential) privacy to prevent disclosing sensitive information but at the same time allow learning from datasets containing data such as e.g., medical or financial records.

Example 1.2. Statistical learning often involves solving stochastic optimization problems of the form

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\xi\sim\mathcal{D}}[\ell(\mathbf{x},\xi)], \tag{1}$$

where ℓ is a loss function. These problems are usually referred to as the (population) risk (or loss) minimization problems. Since the objective function $f(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}}[\ell(\mathbf{x},\xi)]$ itself is an expectation in this case, its subgradients can be expressed as $\mathbf{g}^f_{\mathbf{x}} = \mathbb{E}_{\xi\sim\mathcal{D}}[\mathbf{g}^\ell(\mathbf{x},\xi)]$, where $\mathbf{g}^f_{\mathbf{x}} \in \partial f(\mathbf{x})$, $\mathbf{g}^\ell(\mathbf{x},\xi) \in \partial\ell(\mathbf{x},\xi)$. Thus, it is immediate that

$\widetilde{\mathbf{g}}(\mathbf{x}, \xi) = \mathbf{g}^{\ell}(\mathbf{x}, \xi)$ is an unbiased estimate. The finite variance assumption is easily satisfied in this case, by assuming that the loss function ℓ is either Lipschitz continuous or that it is smooth and the feasible set $\mathcal{X}$ is bounded (you can argue that in both cases the subgradient will be bounded, which is sufficient for guaranteeing finite variance, using Young's inequality; one of these two assumptions will almost always hold).

Example 1.3. Finite-sum minimization problems are of the form:

$$\min_{\mathbf{x} \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}). \tag{2}$$

Such problems can be encountered more broadly but also arise as empirical versions of the population risk problems (1), using an i.i.d. sample $\{\xi_1, \xi_2, \ldots, \xi_n\}$ drawn from $\mathcal{D}$ and $f_i(\mathbf{x}) = \ell(\mathbf{x}, \xi_i)$. In this case, the subgradient of the objective function $f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x})$ can be estimated by sampling i from n uniformly at random and setting $\widetilde{\mathbf{g}}(\mathbf{x}, i) = \mathbf{g}_{\mathbf{x}}^{f_i}$, where $\mathbf{g}_{\mathbf{x}}^{f_i}$. Clearly, this is an unbiased estimate. The finite variance assumption can be satisfied when e.g., each f_i is Lipschitz continuous or each f_i is smooth and the set $\mathcal{X}$ is bounded, along the same lines as in previous example.

2 Stochastic (Projected Sub)Gradient Descent

The algorithm that we will analyze in this lecture is the stochastic variant of projected subgradient descent we saw in the last lecture. In particular, the algorithm can be stated as follows: starting with $\mathbf{x}_0 \in \mathcal{X}$, for $k \geq 0$ the algorithm updates its iterates as

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{y} \in \mathcal{X}} \left\{ a_k \left\langle \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k), \mathbf{y} \right\rangle + \frac{1}{2} \|\mathbf{y} - \mathbf{x}_k\|_2^2 \right\}, \tag{S-PSubGD}$$

where, as before a_k is the step size and ξ_k is drawn from an unknown fixed distribution $\mathcal{D}$, independent of history.

We now discuss how to extend the analysis from the last lecture to analyze (S-PSubGD). This time, we will only be able to bound the optimality gap *in expectation*, as we are working with the stochastic setup.

Our approximate gap will be exactly the same as last time, with the output point defined by $\mathbf{x}_k^{\text{out}} = \frac{1}{A_k} \sum_{i=0}^{k} a_i \mathbf{x}_i$, upper bound defined by $U_k = \frac{1}{A_k} \sum_{i=0}^{k} a_i f(\mathbf{x}_i) \geq f(\mathbf{x}_k^{\text{out}})$, and lower bound defined by $L_k = \frac{1}{A_k} \sum_{i=0}^{k} a_i \big(f(\mathbf{x}_i) + \langle \mathbf{g}_{\mathbf{x}_i}, \mathbf{x}^* - \mathbf{x}_i \rangle \big)$, where, as before, $\mathbf{g}_{\mathbf{x}_i} \in \partial f(\mathbf{x}_i)$. Thus:

$$f(\mathbf{x}_k^{\text{out}}) - f(\mathbf{x}^*) \leq G_k = -\frac{1}{A_k} \sum_{i=0}^{k} a_i \left\langle \mathbf{g}_{\mathbf{x}_i}, \mathbf{x}^* - \mathbf{x}_i \right\rangle. \tag{3}$$

Same as before, we will carry out the analysis by bounding individual summation terms $-a_i \langle \mathbf{g}_{\mathbf{x}_i}, \mathbf{x}^* - \mathbf{x}_i \rangle$ from the definition of G_k. However, using the same approach as last time fails in this case, as the iterates of the algorithm are defined w.r.t. $\widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k)$ in place of $\mathbf{g}_{\mathbf{x}_k}$. A simple "fix" is to correct the analysis by replacing $\mathbf{g}_{\mathbf{x}_k}$ by $\widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k)$, as follows:

$$-a_k \left\langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}^* - \mathbf{x}_k \right\rangle = a_k \left\langle \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k), \mathbf{x}_k - \mathbf{x}^* \right\rangle + a_k \left\langle \mathbf{g}_{\mathbf{x}_k} - \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k), \mathbf{x}_k - \mathbf{x}^* \right\rangle. \tag{4}$$

This works out nicely, because we are only aiming at proving a convergence bound that holds in expectation, and $\mathbb{E}_{\xi_k \sim \mathcal{D}}[a_k \langle \mathbf{g}_{\mathbf{x}_k} - \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k), \mathbf{x}_k - \mathbf{x}^* \rangle] = 0$, as $\mathbf{x}_k, \mathbf{x}^*$ are independent of ξ_k (by our starting assumptions) and $\mathbb{E}_{\xi_k \sim \mathcal{D}}[\mathbf{g}_{\mathbf{x}_k} - \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k)] = \mathbf{0}$, by the assumed unbiasedness of $\widetilde{\mathbf{g}}$. Thus,

$$\mathbb{E}_{\xi_k \sim \mathcal{D}}[-a_k \left\langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}^* - \mathbf{x}_k \right\rangle] = \mathbb{E}_{\xi_k \sim \mathcal{D}}[a_k \left\langle \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k), \mathbf{x}_k - \mathbf{x}^* \right\rangle], \tag{5}$$

which allows us to focus on bounding $a_k \left\langle \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k), \mathbf{x}_k - \mathbf{x}^* \right\rangle$, using the same argument as last time. To avoid repetition (or you can simply look up the analysis from last time), we have that

$$a_k \left\langle \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k), \mathbf{x}_k - \mathbf{x}^* \right\rangle \leq a_k \left\langle \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k), \mathbf{x}_{k+1} - \mathbf{x}^* \right\rangle + \frac{1}{2} \|\mathbf{x}^* - \mathbf{x}_k\|_2^2 - \frac{1}{2} \|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2 - \frac{1}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2. \tag{6}$$

To complete the analysis, last time we bounded $a_k \langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}_{k+1} - \mathbf{x}^* \rangle - \frac{1}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2$ by $\frac{{a_k}^2 M^2}{2}$, using the bound on the subgradients $\|\mathbf{g}_{\mathbf{x}_k}\|_*$, which follows from f being M-Lipschitz continuous. We cannot do the same for the

gradient estimate $\widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k)$ however, as nothing in our assumptions guarantees that it is bounded. On the other hand, our bounded variance assumption ensures that $\mathbb{E}_{\xi_k \sim \mathcal{D}}[\|\widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k) - \mathbf{g}_{\mathbf{x}_g}\|_*^2]$ is bounded by $\sigma^2 < \infty$, thus we can hope to bounded everything in expectation after adding and subtracting $\langle \mathbf{g}_{\mathbf{x}}, \mathbf{x}_{k+1} - \mathbf{x}^* \rangle$ in (6). In particular, we have that

$$a_k \langle \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k), \mathbf{x}_k - \mathbf{x}^* \rangle \leq a_k \langle \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k) - \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}_{k+1} - \mathbf{x}^* \rangle + \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 - \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2 + \frac{{a_k}^2 M^2}{2}, \quad (7)$$

and in what remains we focus on bounding $\mathbb{E}_{\xi_k \sim \mathcal{D}}[a_k \langle \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k) - \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}_{k+1} - \mathbf{x}^* \rangle]$. As discussed before, by the unbiasedness of $\widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k)$, we have that $\mathbb{E}_{\xi_k \sim \mathcal{D}}[a_k \langle \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k) - \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}_{k+1} \rangle]$. The remaining expectation need not be zero however, as $\mathbf{x}_{k+1} = P_{\mathcal{X}}(\mathbf{x}_k - a_k \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k))$ depends on ξ_k. The key insight here is that (exact) projected subgradient descent step $P_{\mathcal{X}}(\mathbf{x}_k - a_k \mathbf{g}_{\mathbf{x}_k})$ is independent of ξ_k, which allows us to write

$$\begin{aligned}
&\mathbb{E}_{\xi_k \sim \mathcal{D}}[a_k \langle \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k) - \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}_{k+1} - \mathbf{x}^* \rangle] \\
&= \mathbb{E}_{\xi_k \sim \mathcal{D}}[a_k \langle \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k) - \mathbf{g}_{\mathbf{x}_k}, P_{\mathcal{X}}(\mathbf{x}_k - a_k \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k)) \rangle] \\
&= a_k \mathbb{E}_{\xi_k \sim \mathcal{D}}[\langle \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k) - \mathbf{g}_{\mathbf{x}_k}, P_{\mathcal{X}}(\mathbf{x}_k - a_k \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k)) - P_{\mathcal{X}}(\mathbf{x}_k - a_k \mathbf{g}_{\mathbf{x}_k}) \rangle] \\
&\leq a_k \mathbb{E}_{\xi_k \sim \mathcal{D}}[\|\widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k) - \mathbf{g}_{\mathbf{x}_k}\|_2 \|P_{\mathcal{X}}(\mathbf{x}_k - a_k \widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k)) - P_{\mathcal{X}}(\mathbf{x}_k - a_k \mathbf{g}_{\mathbf{x}_k})\|_2] \\
&\leq {a_k}^2 \mathbb{E}_{\xi_k \sim \mathcal{D}}[\|\widetilde{\mathbf{g}}(\mathbf{x}_k, \xi_k) - \mathbf{g}_{\mathbf{x}_k}\|_2^2] \\
&\leq {a_k}^2 \sigma^2,
\end{aligned}$$

where we have used the non-expansive property of the projection operator (proved in the lecture on Constrained, Projection-Based Optimization).

Combining the last inequality with (7) and (5), we now get that

$$\mathbb{E}_{\xi_k \sim \mathcal{D}}[-a_k \langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}^* - \mathbf{x}_k \rangle] \leq \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 - \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2 + \frac{{a_k}^2 (M^2 + 2\sigma^2)}{2}. \quad (8)$$

Let us now consider the algorithm output after K iterations. Based on (8), for any $k \in \{0, 1, \ldots, K\}$ (by taking the expectation w.r.t. all randomness in the algorithm on both sides),

$$\mathbb{E}_{\{\xi_0, \xi_1, \ldots, \mathbf{x}_K\} \sim \mathcal{D}^{K+1}}[-a_k \langle \mathbf{g}_{\mathbf{x}_k}, \mathbf{x}^* - \mathbf{x}_k \rangle] \leq \mathbb{E}_{\{\xi_0, \xi_1, \ldots, \mathbf{x}_K\} \sim \mathcal{D}^{K+1}} \left[\frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 - \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2 + \frac{{a_k}^2 (M^2 + 2\sigma^2)}{2}\right].$$

Hence, we can now conclude from (3) that

$$\mathbb{E}_{\{\xi_0, \xi_1, \ldots, \mathbf{x}_K\} \sim \mathcal{D}^{K+1}}[f(\mathbf{x}_k^{\text{out}}) - f(\mathbf{x}^*)] \leq \mathbb{E}_{\{\xi_0, \xi_1, \ldots, \mathbf{x}_K\} \sim \mathcal{D}^{K+1}}[G_k] \leq \frac{\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 + {a_k}^2 (M^2 + 2\sigma^2)}{2A_k}. \quad (9)$$

The expression on the right-hand side is the same as what we got the last time for projected subgradient descent, except for having $M^2 + 2\sigma^2$ in place of M^2. Hence, the same conclusions about convergence under different choices of step sizes a_k apply here is well, by replacing M^2 by $M^2 + 2\sigma^2$ in the discussion.