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.

$$x \in \mathcal{X} \longrightarrow \boxed{\begin{array}{c}\text{1st order}\\ \text{oracle}\end{array}} \longrightarrow \begin{array}{l} g_x \in \partial f(x) \ \ (\nabla f(x) \text{ if } f \text{ is differentiable}) \\ \qquad\qquad\ \ \text{maybe also } f(x) \end{array}$$

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

$$x \in \mathcal{X} \longrightarrow \boxed{\begin{array}{c}\text{1st order}\\ \text{stochastic oracle}\end{array}} \longrightarrow \tilde{g}(x, \xi)$$

Here $\tilde{g}(x, \xi)$ is a stochastic estimate of some $g_x \in \partial f(x)$, where ξ 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 it here.

1.1 Examples

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

Example 2. Finite sum minimization: Want to minimize

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

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

More generally, we can take $\tilde{g}(x, \xi) = \frac{1}{n} \sum_{i \in S} \nabla f_i(x)$, where S is a random subset of $\{1, \ldots n\}$; here $\xi = S$ is sometimes called a mini-batch.

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

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

but we do not know how to exactly compute the expectation above. Suppose we have collected n data points (a_i, b_i) that come from the distribution Π_{data}. As an approximation we minimize 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 view $\tilde{g}(x, \xi) = \nabla f_{\text{emp}}(x)$ as a noisy estimate of $\nabla f(x)$.

Assumptions for this note: $\min_{x \in \mathcal{X}} f(x)$.

1. f is cvx, M-Lipschitz w.r.t. $\|\cdot\|_2$. (f may not be differentiable).
2. $\mathcal{X}$ is closed, cvx, nonempty. Projection $P_{\mathcal{X}}(\cdot)$ can be computed.

3. $\forall x \in \mathcal{X}$, it holds:

$$\begin{cases} \text{Unbiased estimate:} & \mathbb{E}_{\xi}[\tilde{g}(x,\xi)] = g_x \in \partial f(x). \\ \text{Bounded variance.} & \mathbb{E}_{\xi}\left[\|\tilde{g}(x,\xi) - g_x\|_2^2\right] \le \sigma^2 < \infty. \end{cases}$$

2 Stochastic (projected sub)gradient descent

Consider the following S-PSubGD algorithm:

$$x_{k+1} = \underset{u \in \mathcal{X}}{\operatorname{argmin}} \left\{ a_k \langle \tilde{g}(x_k, \xi_k), u - x_k \rangle + \frac{1}{2}\|u - x_k\|_2^2 \right\}$$
$$= P_{\mathcal{X}}\left(x_k - a_k \tilde{g}(x_k, \xi_k)\right),$$

where $a_k > 0$ is the stepsize to be chosen later.

Convergence Analysis:

$\xi_0, \xi_1, \dots, \xi_k$: i.i.d.

True (sub)grad: $g_k \doteq g_{x_k}$. noisy (sub)grad: $\tilde{g}_k = \tilde{g}(x_k, \xi_k)$

Similar framework: $x_k^{out} = \frac{1}{A_k}\sum_{i=0}^{k} a_i x_i$. $A_k := \sum_{i=0}^{k} a_i$.

Upbd: $U_k := \frac{1}{A_k}\sum_{i=0}^{k} a_i f(x_i) \ge f(x_k^{out})$, by cvx of f.

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

$\Rightarrow G_k = U_k - L_k = -\frac{1}{A_k}\sum_{i=0}^{k} a_i \langle g_i, x^* - x_i\rangle \ge f(x_k^{out}) - f(x^*)$.

$A_0 G_0 = -\langle g_0, x^* - x_0\rangle$.

$A_k G_k - A_{k-1}G_{k-1} = -a_k\langle g_k, x^* - x_k\rangle = a_k\langle g_k, x_k - x_{k+1}\rangle + a_k\langle g_k, x_{k+1} - x^*\rangle$

$= \underbrace{a_k\langle g_k, x_k - x_{k+1}\rangle}_{\le a_k\|g_k\|_*\|x_k - x_{k+1}\| \le M a_k\|x_k - x_{k+1}\|} + a_k\langle \tilde{g}_k, x_{k+1} - x^*\rangle + \underbrace{a_k\langle g_k - \tilde{g}_k, x_{k+1} - x^*\rangle}_{\text{stochastic term.}}$

$x_{k+1} = P_{\mathcal{X}}(x_k - a_k\tilde{g}_k)$

By minimum principle, $\forall y \in \mathcal{X}$, $\langle x_{k+1} - x_k + a_k \tilde{g}_k, y - x_{k+1} \rangle \geq 0$.

Set $y = x^*$ $\quad a_k \langle \tilde{g}_k, x_{k+1} - x^* \rangle \leq \langle x_{k+1} - x_k, x^* - x_{k+1} \rangle$

$$= \frac{1}{2}\|x_k - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 - \frac{1}{2}\|x_k - x_{k+1}\|^2$$

$$A_k G_k - A_{k-1} G_{k-1} \leq \frac{1}{2}\|x_k - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 \underbrace{- \frac{1}{2}\|x_k - x_{k+1}\|^2 + M a_k \|x_k - x_{k+1}\|}_{\leq \frac{1}{2} M^2 a_k^2} + a_k \langle g_k - \tilde{g}_k, x_{k+1} - x^* \rangle$$

$$\Rightarrow \mathbb{E}[A_k G_k - A_{k-1} G_{k-1}] \leq \frac{1}{2} \mathbb{E}\left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right] + \frac{1}{2} M^2 a_k^2 + \mathbb{E}[a_k \langle g_k - \tilde{g}_k, x_{k+1} - x^* \rangle]$$

$$\underset{\uparrow \atop \xi_{0:k}}{\mathbb{E}}[\langle g_k - \tilde{g}_k, x_{k+1} - x^* \rangle] = \mathbb{E}_{\xi_{0:k-1}} \mathbb{E}_{\xi_k}[\langle g_k - \tilde{g}_k, x_{k+1} - x^* \rangle \mid \xi_{0:k-1}]$$

$\mathbb{E}_{\xi_k}[\langle g_k - \tilde{g}_k, x_{k+1} - x^* \rangle \mid \xi_{0:k-1}]$. Let's handle this term,

$$\mathbb{E}[\langle g_k - \tilde{g}_k, x^* \rangle \mid \xi_{0:k-1}] \underset{\text{linearity}}{=} \langle g_k - \mathbb{E}_{\xi_k}[\tilde{g}_k \mid \xi_{0:k-1}], x^* \rangle$$

$\tilde{g}_k$ indep of $\xi_{0:k-1}$ $\quad = \langle g_k - \mathbb{E}_{\xi_k}[\tilde{g}_k], x^* \rangle \underset{\text{unbiased estimate}}{=} 0$.

$(\tilde{g}_k \sim \xi_k)$

$$\therefore \mathbb{E}_{\xi_k}[\langle g_k - \tilde{g}_k, x_{k+1} - x^* \rangle \mid \xi_{0:k-1}] = \mathbb{E}_{\xi_k}[\langle g_k - \tilde{g}_k, x_{k+1} \rangle \mid \xi_{0:k-1}]$$

$$= \mathbb{E}_{\xi_k}[\langle g_k - \tilde{g}_k, P_{\mathcal{X}}(x_k - a_k \tilde{g}_k) \rangle \mid \xi_{0:k-1}]$$

$$\mathbb{E}[\langle X, Y \rangle] \underset{X, Y \text{ independent.}}{=} \langle \mathbb{E}[X], \mathbb{E}[Y] \rangle$$

$$\mathbb{E}_{\xi_k}[\langle g_k - \tilde{g}_k, P_{\mathcal{X}}(x_k - a_k g_k) \rangle \mid \xi_{0:k-1}]$$

$$= \langle \mathbb{E}_{\xi_k}[g_k - \tilde{g}_k \mid \xi_{0:k-1}], \mathbb{E}_{\xi_k}[P_{\mathcal{X}}(x_k - a_k g_k) \mid \xi_{0:k-1}] \rangle$$

$$= \langle \underbrace{g_k - \mathbb{E}_{\xi_k}[\tilde{g}_k]}_{=0}, \cdots \rangle = 0.$$

多加1个为0项.

$$\mathbb{E}_{\xi_k}[\langle g_k - \tilde{g}_k, P_X(x_k - a_k \tilde{g}_k)\rangle \mid \xi_{0:k-1}] = \mathbb{E}_{\xi_k}[\langle g_k - \tilde{g}_k, P_X(x_k - a_k \tilde{g}_k) - P_X(x_k - a_k g_k) \mid \xi_{0:k-1}]$$

$$\leq \mathbb{E}_{\xi_k}\left[\|g_k - \tilde{g}_k\|_2 \cdot \|P_X(x_k - a_k \tilde{g}_k) - P_X(x_k - a_k g_k)\|_2 \mid \xi_{0:k-1}\right]$$ Cauchy

$$\leq \mathbb{E}_{\xi_k}\left[a_k \|g_k - \tilde{g}_k\|_2^2 \mid \xi_{0:k-1}\right]$$ $P_X(\cdot)$, non-expansive

$$= a_k \mathbb{E}_{\xi_k}\left[\|g_k - \tilde{g}_k\|_2^2\right]$$ independence.

$$\leq a_k \sigma^2$$ Bounded var.

Hence

$$\mathbb{E}[A_k G_k - A_{k-1} G_{k-1}] \leq \frac{1}{2}\mathbb{E}\left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right] + \frac{1}{2} M^2 a_k^2 + a_k \cdot a_k \sigma^2$$

$$= \frac{1}{2}\mathbb{E}\left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right] + \frac{a_k^2}{2}(M^2 + 2\sigma^2)$$

we set x_0, no randomness.

Sum,

$$\mathbb{E}[f(x_K^{out}) - f(x^*)] \leq \mathbb{E}[G_K] \leq \frac{\mathbb{E}[\|x_0 - x^*\|^2] + \sum_{k=1}^{K} a_k^2 (M^2 + 2\sigma^2)}{2A_K}$$

$$= \frac{\|x_0 - x^*\|^2 + (M^2 + 2\sigma^2)\sum_{k=0}^{K} a_k^2}{2A_K}$$

$2\sigma^2$: sto term.

Then analysis is similar to PsubGD.

Apply $a_k = \dfrac{\|x_0 - x^*\|_2}{\sqrt{M^2 + 2\sigma^2} \cdot \sqrt{K+1}}$

$$\mathbb{E}[f(x_K^{out}) - f(x^*)] \le \frac{\|x_0 - x^*\|^2}{A_K} = \frac{\sqrt{M^2+2\sigma^2}\,\|x_0-x^*\|_2}{\sqrt{K+1}}$$

Observation:

① $\sigma = 0$. Generalize to Psub GD.

② $O(\frac{1}{\sqrt{K}})$, convergence rate

③ We can discuss the case. $\{$ $\mathcal{X}$'s diameter. M? σ? unknown.

3 Analysis of SGD in other settings (Optional)

In this section, we state without proof several additional convergence results for (projected) stochastic (sub)gradient descent.[2] As before, we assume that f is convex and the stochastic gradient $g(x, \xi)$ is unbiased, but we will consider other additional properties of f and $g(x, \xi)$.

3.1 Role of smoothness

Still assume that stochastic gradient has variance bounded by σ^2; see equation (1). We make the additional assumption that f is L-smooth (w.r.t. $\|\cdot\|$). Let $D := \max_{x,y\in\mathcal{X}} \|x - y\|_2$ be the diameter of $\mathcal{X}$. With a constant stepsize $a_k = \frac{1}{L+(\sigma/D)\sqrt{(K+1)/2}}, \forall k$, one can show that

$$\mathbb{E}f\left(x_K^{\text{out}}\right) - f(x^*) \le D\sigma\sqrt{\frac{2}{K+1}} + \frac{LD^2}{K+1}. \tag{3}$$

When K is large, the first term on the RHS dominates and thus we have an $O(1/\sqrt{K})$ rate. This rate is essentially the same as the bound (2) for nonsmooth f. Therefore, smoothness does not offer much benefit in the stochastic setting. In contrast, in the deterministic setting, smoothness leads to the faster rates of $O(1/K)$ (for GD) and $O(1/K^2)$ (for AGD).

3.2 Role of strong convexity

Going back to the setting with M-Lipschitz f. Still assume that stochastic gradient has variance bounded by σ^2; see equation (1). We make the additional assumption that f is m-strongly convex (w.r.t. $\|\cdot\|_2$). Note that this is possible only when $\mathcal{X}$ is bounded. [3]

For the diminishing stepsize $a_k = \frac{2}{m(k+2)}$, we have

$$\mathbb{E}f\left(\sum_{k=0}^{K} \frac{2(k+1)}{(K+1)(K+2)} x_k\right) - f(x^*) \le \frac{2(M^2+\sigma^2)}{m(K+2)}. \tag{4}$$

This $O(1/K)$ rate is better than the $O(1/\sqrt{K})$ rate for non-strongly convex f.

3.3 More general noise

We now consider a more general form of noise assumption: there exist some $L_g \ge 0$ and $B \ge 0$ such that for all $x \in \mathcal{X}$:

$$\mathbb{E}\left[\|g(x,\xi)\|_2^2\right] \le L_g^2 \|x - x^*\|_2^2 + B^2. \tag{5}$$

We consider three cases.

3.3.1 $L_g = 0$, $B > 0$, convex f

This setting is a slight generalization of the previous assumption (1) of M-Lipschitz f and σ^2-bounded variance. In particular, the assumption (1) implies that

$$\begin{aligned}\mathbb{E}\left[\|g(x,\xi)\|_2^2\right] &= \|\mathbb{E}[g(x,\xi)]\|_2^2 + \mathbb{E}_\xi\left[\|\tilde{g}(x,\xi) - g_x\|_2^2\right]\\ &= \|g_x\|_2^2 + \mathbb{E}_\xi\left[\|\tilde{g}(x,\xi) - g_x\|_2^2\right] \le M^2 + \sigma^2.\end{aligned}$$

Therefore, the more general assumption (5) is satisfied with $L_g = 0$ and $B^2 = M^2 + \sigma^2$. In this case, using the constant stepsize $a_k = \frac{\|x_0 - x^*\|_w}{B\sqrt{K+1}}, \forall k$, we have

$$\mathbb{E}\left[f(x_K^{\text{out}}) - f(x^*)\right] \le \frac{\|x_0 - x^*\|_2 B}{\sqrt{K+1}}.$$

This bound is essentially the same as the bound (2) proved earlier.

3.3.2 $L_g > 0$, $B = 0$, m-strongly convex f

In this setting, we have $\mathbb{E}\left[\|g(x,\xi)\|_2^2\right] \to 0 = \nabla f(x^*)$ as $x \to x^*$. That is, the stochastic gradient becomes more and more accurate near x^*. Moreover, we have

$$\begin{aligned}L_g^2 \|x - x^*\|_2^2 &\ge \mathbb{E}\left[\|g(x,\xi)\|_2^2\right] & \\ &\ge \|\mathbb{E}[g(x,\xi)]\|_2^2 & \text{Jensen's}\\ &= \|\nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(x^*)\|_2^2, & \text{unbiased, } \nabla f(x^*) = 0\end{aligned}$$

[3]For a strongly convex function, its subgradient grows linearly away from x^*: $\|\nabla f(x)\|_2 \ge \frac{m}{2}\|x - x^*\|_2$, hence $\|\nabla f(x)\| \le M$ cannot be over the entire $\mathbb{R}^d$.

so the gradient of f satisfies a "Lipschitz-like" assumption.

With a constant stepsize $a_k = \frac{m}{L_g^2}, \forall k$, we have

$$\mathbb{E}\, \|x_K - x^*\|_2^2 \leq \left(1 - \frac{m^2}{L_g^2}\right)^K \|x_0 - x^*\|^2.$$

We have geometric convergence thanks to strong convexity and the Lipschitz-like property. The contraction factor is $1 - \frac{m^2}{L_g^2}$, which is worse than the $1 - \frac{m}{L}$ (for GD) and $1 - \sqrt{\frac{m}{L}}$ (for AGD) factors we saw in the deterministic setting with m-strong convexity and L-Lipschitz gradient.

3.3.3 $L_g > 0, B > 0$, m-strongly convex f

With a diminishing stepsize $a_k = \frac{1}{2m(L_g^2/2m^2+k)}$, we have

$$\mathbb{E}\, \|x_K - x^*\|_2^2 \leq \frac{c_0 B^2}{2m(L_g^2/2m^2 + K)}.$$

For large K, this is an $O(1/K)$ rate.