1 Stochastic Optimization

We are interested in constrained minimization of $f : \mathbb{R}^n \to \mathbb{R}$

$$\text{minimize}_{x\in\mathcal{X}} \ \left[f(x) = \mathbb{E}\left[F(x,\xi)\right]\right]$$

where $\mathcal{X} \subset \mathbb{R}^n$ is closed, bounded convex set. ξ is a random variable, and $F(\cdot,\xi)$ is convex for all $\xi \in \Xi$, and therefore $f(\cdot)$ is convex. For uniform p_ξ over finite alphabets of size n, the problem reduces to finite sum problem

$$\text{minimize}_{x\in\mathcal{X}} \ \left[f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right]$$

Assume we can

1. Sample $\xi_1, \xi_2, \cdots \overset{i.i.d.}{\sim} p_\xi$
2. Given $(x,\xi) \in \mathcal{X} \times \Xi$, a first order oracle that returns a subgradient vector $G(x,\xi) \in \partial_x F(x,\xi)$. We also assume that G is unbiased, i.e. $g(x) := \mathbb{E}\left[G(x,\xi)\right] \in \partial f(x)$

1.1 Stochastic Gradient Method

We can show that if $f \in \mathscr{S}^1_{L,\mu}$, the choice of $\alpha_k = \mathcal{O}(1/k)$ yields sublinear convergence of $\mathcal{O}(\frac{1}{\epsilon})$ for last iterates. If $f \in \mathscr{F}^1_L$, the choice of $\alpha_k = \mathcal{O}(\frac{1}{\sqrt{k}})$ yields a sublinear convergence of $\mathcal{O}(\frac{1}{\epsilon^2})$ for average iterates. Stochastic gradient method (or Stochastic Approximation (SA) algorithms) solves the problem by

$$x^{k+1} = \mathcal{P}_{\mathcal{X}}\left(x^k - \alpha_k G(x^k, \xi_k)\right)$$

where $\alpha_k > 0$ are stepsizes, $\mathcal{P}_{\mathcal{X}}(y) = \arg\min_{x\in\mathcal{X}} \frac{1}{2}\|x-y\|_2^2$ is the euclidean projection onto a convex set. It is important to note that the current iterate x^k are functions of random variables $x^k := x^k(\xi_{[k-1]})$ where $\xi_{[k-1]} = (\xi_1, \cdots, \xi_{k-1})$, and therefore are random variables themselves. In addition, $x^k \perp\!\!\!\perp \xi_k$.

1.2 Convergence

Derivations copied from [7], [8] and slides. We assume

1. bounded variance for stochastic subgradient, which translates to $\mathbb{E}_\xi\left[G(x,\xi)\right] \le M^2$ given $x \in \mathcal{X}$.
2. bounded $\mathcal{X}$ where radius given by $D_{\mathcal{X}} = \max_{x\in\mathcal{X}} \|x - x^*\|_2$.

We outline implications of some assumptions

1. If f is convex, then

$$f(x') \ge f(x) + \langle g(x), x' - x\rangle \qquad \forall\, x, x' \in \mathcal{X} \tag{1}$$

2. If f has L lipschitz continuous gradients, then

$$\left\|\nabla f(x') - \nabla f(x)\right\| \le L\left\|x' - x\right\| \qquad \forall\, x, x' \in \mathcal{X} \tag{2}$$

$$f(x) - f(x^*) \le \frac{1}{2}L\left\|x - x^*\right\| \qquad \forall\, x \in \mathcal{X} \tag{descent lemma}$$

3. If f is μ-strongly convex, then

$$\left\langle \nabla f(x') - \nabla f(x), x' - x \right\rangle \geq \mu \left\| x' - x \right\|_2^2 \qquad \forall\, x \in \mathcal{X} \tag{3}$$

$$\mu \left\| x - x^* \right\|^2 \leq \mu \left\langle g(x) - g(x^*), x - x^* \right\rangle = \left\langle g(x), x - x^* \right\rangle \qquad \forall\, x \in \mathcal{X},\ g(x) \in \partial f(x) \tag{4}$$

We first derive some preliminary results. Using iterated expecatation, we have

$$\begin{aligned}
\mathbb{E}\left[\left\langle G(x^k, \xi_k), x^k - x^* \right\rangle\right] &= \mathbb{E}_{\xi_{[k-1]}}\left[\mathbb{E}_{\xi_k}\left[\left\langle G(x^k(\xi_{[k-1]}), \xi_k), x^k(\xi_{[k-1]}) - x^* \right\rangle\right] \mid \xi_{[k-1]}\right] \\
&= \mathbb{E}_{\xi_{[k-1]}}\left[\left\langle \mathbb{E}_{\xi_k}\left[G(x^k(\xi_{[k-1]}), \xi_k) \mid \xi_{[k-1]}\right], x^k(\xi_{[k-1]}) - x^* \right\rangle\right] \\
&= \mathbb{E}\left[\left\langle g(x^k), x^k - x^* \right\rangle\right]
\end{aligned} \tag{5}$$

where the expectation is taken w.r.t $\xi_{[k-1]}$. We first derive a bound on $R_k = \left\| x^k - x^* \right\|_2^2$ and $r_k = \mathbb{E}\left[R_k\right]$,

$$\begin{aligned}
R_{k+1} &= \left\| x^k - x^* \right\|^2 \\
&= \left\| \mathcal{P}_{\mathcal{X}}\left(x^k - \alpha_k G(x^k, \xi_k)\right) - \mathcal{P}_{\mathcal{X}}(x^*) \right\| && (\mathcal{P}_{\mathcal{X}}(x^*) = x^*) \\
&\leq \left\| x^k - \alpha_k G(x^k, \xi_k) - x^* \right\|^2 && (\text{nonexpansive of } \mathcal{P}(\cdot)\ \left\| \mathcal{P}_{\mathcal{X}}(x') - \mathcal{P}_{\mathcal{X}}(x) \right\| \leq \left\| x' - x \right\|) \\
&\leq R^k - 2\alpha_k \left\langle G(x^k, \xi_k), x^k - x^* \right\rangle + \alpha_k^2 \left\| G(x^k, \xi_k) \right\|^2 \\
r_{k+1} &\leq r_k - 2\alpha_k \mathbb{E}\left[\left\langle G(x^k, \xi_k), x^k - x^* \right\rangle\right] + \alpha_k^2 \mathbb{E}\left[\left\| G(x^k, \xi_k) \right\|^2\right] && (\text{Take expectation w.r.t. } \xi_{[k]}) \\
&= r_k - 2\alpha_k \mathbb{E}\left[\left\langle g(x^k), x^k - x^* \right\rangle\right] + \alpha_k^2 M^2 && (\text{By (5) and bounded variance})
\end{aligned}$$

1.2.1 Strongly Convex Case

If $f \in \mathscr{S}^1_{L,\mu}$, using (4), we have

$$r_{k+1} \leq r_k - 2\alpha_k \mathbb{E}\left[\left\| x^k - x^* \right\|^2\right] + \alpha_k^2 M^2 = (1 - 2\mu\alpha_k) r_k + \alpha_k^2 M^2$$

If we choose $\alpha_k = \theta/(k+1)$, where $\theta > 1/(2\mu)$. It could be shown by induction that [7]

$$r_k \leq \frac{c_\theta}{k+1} \qquad \text{where} \qquad c_\theta = \max\left\{ \frac{2\theta^2 M^2}{2\mu\theta - 1}, r_0 \right\}$$

By (descent lemma), we derive bound on the objective value

$$\mathbb{E}\left[f(x^k) - f(x^*)\right] \leq \frac{1}{2} L \mathbb{E}\left[\left\| x^k - x^* \right\|^2\right] \leq \frac{L c_\theta}{2(k+1)}$$

Therefore, the choice of $\alpha_k = \mathcal{O}(\frac{1}{\epsilon})$ yields last iterate convergence rate of $\mathcal{O}(\frac{1}{\epsilon})$

1.2.2 Convex Case

[7] indicates that we need to increase the stepsize ($\mathcal{O}(\frac{1}{k})$ to $\mathcal{O}(\frac{1}{\sqrt{k}})$)) to ensure faster convergence rate for general convex problems, at the cost of *more noisy* trajectory. To suppress the noise, we use average iterates $\left\{x^k\right\}$ rather than last iterates as solution to the problem.

$$
\begin{aligned}
r_{k+1} &\le r_k - 2\alpha_k \mathbb{E}\left[\left\langle g(x^k), x^k - x^*\right\rangle\right] + \alpha_k^2 M^2 \\
2\alpha_k \mathbb{E}\left[f(x^k) - f(x^*)\right] &\le 2\alpha_k \mathbb{E}\left[\left\langle g(x^k), x^k - x^*\right\rangle\right] \le r_k - r_{k+1} + \alpha_k^2 M^2 && \text{(Rearrange, and by 1)} \\
\sum_{i=1}^{k} \left(2\alpha_i \mathbb{E}\left[f(x^i) - f(x^*)\right]\right) &\le \sum_{i=1}^{k} \left(r_i - r_{i+1} + \alpha_i M^2\right) = r_1 + \sum_{i=1}^{k} \alpha_i^2 && \text{(Telescope)} \\
\sum_{i=1}^{k} \gamma_i \mathbb{E}\left[(f(x^i) - f(x^*))\right] &= \mathbb{E}\left[\sum_{i=1}^{k} \gamma_i (f(x^i) - f(x^*))\right] \le \frac{r_1 + M^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i} \\
& \text{(Let } \gamma_i = \alpha_i / \textstyle\sum_i \alpha_i\text{. Divide by } 2\textstyle\sum_i \alpha_i\text{. Use linearity of expectation)} \\
\mathbb{E}\left[f(\tilde{x}^k) - f(x^*)\right] &\le \frac{r_1 + M^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i} \\
& \text{(Let } \tilde{x}^k = \textstyle\sum_{i=1}^{k} \gamma_i x^i\text{. } f(\tilde{x}^k) \le \textstyle\sum_i \gamma_i f(x^i) \text{ by convexity of } f\text{. } \textstyle\sum_i \gamma_i = 1\text{)}
\end{aligned}
$$

We derive tighest bound by finding minimal value of $\alpha_k = \alpha$ of the bound.

$$
\mathbb{E}\left[f(\tilde{x}^k) - f(x^*)\right] \le \frac{D_{\mathcal{X}} M}{\sqrt{k}} \qquad \alpha_k = D_{\mathcal{X}} / (M\sqrt{k})
$$

Therefore, the choice of $\alpha_k = \mathcal{O}(\frac{1}{\sqrt{k}})$ yields average iterate convergence rate of $\mathcal{O}(\frac{1}{\epsilon^2})$