1 Assumptions for convergence rate of stochastic sub-gradient for convex functions

- f is convex
- $\mathbb{E}\left[\|g_i(x_k)\|_2^2 \mid x_k\right] \leq \sigma^2$ for any sub-gradient g_i.

2 Proof for convergence rate of stochastic sub-gradient for convex functions

Since we do not assume Lipschitz continuity, we cannot use the upper bound that is based on FToC. We have to work directly with the distance to x^*, the minimizer of f. Using the stochastic sub-gradient algorithm we have that

$$x_{k+1} := x_k - \alpha_k g_i(x_k),$$

where $g_i(x_k)$ is a sub-gradient of the loss function f_i at point x_k, i.e., $g_i(x_k) \in \partial f_i(x_k)$. We have that

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &= \|x_k - \alpha_k g_i(x_k) - x^*\|_2^2 \\ &= \|x_k - x^*\|_2^2 - 2\alpha_k g_i(x_k)^T(x_k - x^*) + \alpha_k^2\|g_i(x_k)\|_2^2.\end{aligned}$$

Taking conditional expectation and assuming that α_k are not random variables we get

$$\mathbb{E}\left[\|x_{k+1} - x^*\|_2^2 \mid x_k\right] = \underbrace{\|x_k - x^*\|_2^2}_{\text{old distance}} \underbrace{- 2\alpha_k \mathbb{E}\left[g_i(x_k)^T(x_k - x^*) \mid x_k\right]}_{\text{progress}} + \alpha_k^2 \underbrace{\mathbb{E}\left[\|g_i(x_k)\|_2^2 \mid x_k\right]}_{\text{“variance”}}.$$

Note that the third term is not exactly the variance, i.e., $\mathbb{E}\left[\|g_i(x_k) - \mathbb{E}\left[g_i(x_k) \mid x_k\right]\|_2^2 \mid x_k\right]$, but we will abuse the terminology for this proof. By the third assumption we have that

$$\mathbb{E}\left[\|x_{k+1} - x^*\|_2^2 \mid x_k\right] \leq \|x_k - x^*\|_2^2 - 2\alpha_k \mathbb{E}\left[g_i(x_k)^T(x_k - x^*) \mid x_k\right] + \alpha_k^2\sigma^2. \tag{1}$$

Let's focus on the middle term in the right hand side of the above equation. By convexity of f and the fact that $g_i(x_k)$ is an unbiased estimator, i.e., $\mathbb{E}\left[g_i(x_k) \mid x_k\right] = g(x_k)$ we have that

$$-\mathbb{E}\left[g_i(x_k)^T(x_k - x^*) \mid x_k\right] = g(x_k)^T(x^* - x_k) \leq f^* - f(x_k) = -(f(x_k) - f^*).$$

Using this inequality in (1)

$$\mathbb{E}\left[\|x_{k+1} - x^*\|_2^2 \mid x_k\right] \leq \|x_k - x^*\|_2^2 - 2\alpha_k(f(x_k) - f^*) + \alpha_k^2\sigma^2$$

Let's re-arrange

$$2\alpha_k(f(x_k) - f^*) \leq \|x_k - x^*\|_2^2 - \mathbb{E}\left[\|x_{k+1} - x^*\|_2^2 \mid x_k\right] + \alpha_k^2\sigma^2$$

Let's take expectation over the first $t+1$ iterations and sum over the first $t+1$ iterations to get

$$\sum_{k=0}^{t} 2\alpha_k \mathbb{E}\left[f(x_k) - f^*\right] \leq \sigma^2 \sum_{k=0}^{t} \alpha_k^2 + \sum_{k=0}^{t} \mathbb{E}\left[\|x_k - x^*\|_2^2\right] - \mathbb{E}\left[\|x_{k+1} - x^*\|_2^2\right]. \tag{2}$$

Let's define

$$\bar{x}_{t+1} := \mathbb{E}\left[x_k\right] = \frac{1}{t+1}\sum_{k=0}^{t} x_t.$$

Using Jensen's inequality $f(\mathbb{E}\left[x_k\right]) \leq \mathbb{E}\left[f(x_k)\right]$ (holds only for convex functions) we get that

$$\sum_{k=0}^{t} 2\alpha_k \mathbb{E}\left[f(x_k) - f^*\right] \geq \sum_{k=0}^{t} 2\alpha_k f(\mathbb{E}\left[x_k\right]) - f^* = (f(\bar{x}_{t+1}) - f^*)\sum_{k=0}^{t} 2\alpha_k$$

Using this in (2) we get

$$(f(\bar{x}_{t+1}) - f^*)\sum_{k=0}^{t} 2\alpha_k \leq \sigma^2 \sum_{k=0}^{t} \alpha_k^2 + \sum_{k=0}^{t} \mathbb{E}\left[\|x_k - x^*\|_2^2\right] - \mathbb{E}\left[\|x_{k+1} - x^*\|_2^2\right].$$

The right hand side telescopes

$$\begin{aligned}(f(\bar{x}_{t+1}) - f^*)\sum_{k=0}^{t} 2\alpha_k &\leq \sigma^2 \sum_{k=0}^{t} \alpha_k^2 + \|x_0 - x^*\|_2^2 - \mathbb{E}\left[\|x_{t+1} - x^*\|_2^2\right] \\ &\leq \sigma^2 \sum_{k=0}^{t} \alpha_k^2 + \|x_0 - x^*\|_2^2.\end{aligned}$$

This gives us

$$f(\bar{x}_{t+1}) - f^* \leq \frac{\sigma^2 \sum_{k=0}^{t} \alpha_k^2}{2\sum_{k=0}^{t} \alpha_k} + \frac{\|x_0 - x^*\|_2^2}{2\sum_{k=0}^{t} \alpha_k}$$

We now have to make a decision about selecting α_j's.

- Decreasing step-sizes: if $\alpha_j = 1/j$ then $\sum_{j=0}^{t} \alpha_j = \mathcal{O}(\log t)$ and $\sum_{j=0}^{t} \alpha_j^2 = \mathcal{O}(1)$, which gives rate $\mathcal{O}(1/\log t)$.
- Large decreasing step-size: if $\alpha_j = 1/\sqrt{j}$ then $\sum_{j=0}^{t} \alpha_j = \mathcal{O}(\sqrt{t})$ and $\sum_{j=0}^{t} \alpha_j^2 = \mathcal{O}(\log t)$, which gives rate $\mathcal{O}(\log t/\sqrt{t})$.
- Constant step-size: if $\alpha_j = \alpha$ for some constant α, then $\sum_{j=0}^{t} \alpha_j = (t+1)\alpha$ and $\sum_{j=0}^{t} \alpha_j^2 = (t+1)\alpha^2$, which gives rate $\mathcal{O}(1/t + \alpha)$. This means implies that the minimum expected distance $f(x_k) - f^*$ will never go to zero and then the algorithm only converges to a neighborhood of a stationary point.