Summary

$$\min_{x \in Q} f(x)$$

$Q \subseteq \mathbb{R}^n$ - bounded, $D = \operatorname{diam}(Q)$

$$\|f'(x)\| \le M \qquad \forall x \in Q$$

Stochastic Subgradient method.

$$x_{k+1} = \pi_Q\left(x_k - \frac{1}{\beta_k} g_k\right), \quad g_k \text{ - stochastic (sub) gradient}$$

$$\beta_k := \frac{1}{D}\sqrt{\|g_0\|^2 + \dots + \|g_k\|^2}$$

$$\mathbb{E}\left[f(\bar{x}_k) - f^*\right] \le \frac{(\sigma + M)D}{\sqrt{k}}$$

σ - variance: $\mathbb{E}\|g_k - f'(x_k)\|^2 \le \sigma^2$.

"Stochasticity" & "Non-smoothness" (convex)

$f: \mathbb{R}^n \to \mathbb{R}$ convex and smooth:

$$\|f'(y) - f'(x)\| \le L\|y - x\|$$

unconstrained minimization: $f'(x^*) = 0$.

Simple Analysis

$x := x^*$:

$$\|f'(y)\| \le L\|y - x^*\| \le LD = M$$

$D \ge \|x_0 - x^*\|$

Extra constraint: $Q = \{x \in \mathbb{R}^n : \|x - x_0\| \le D\}$

Smooth Problems

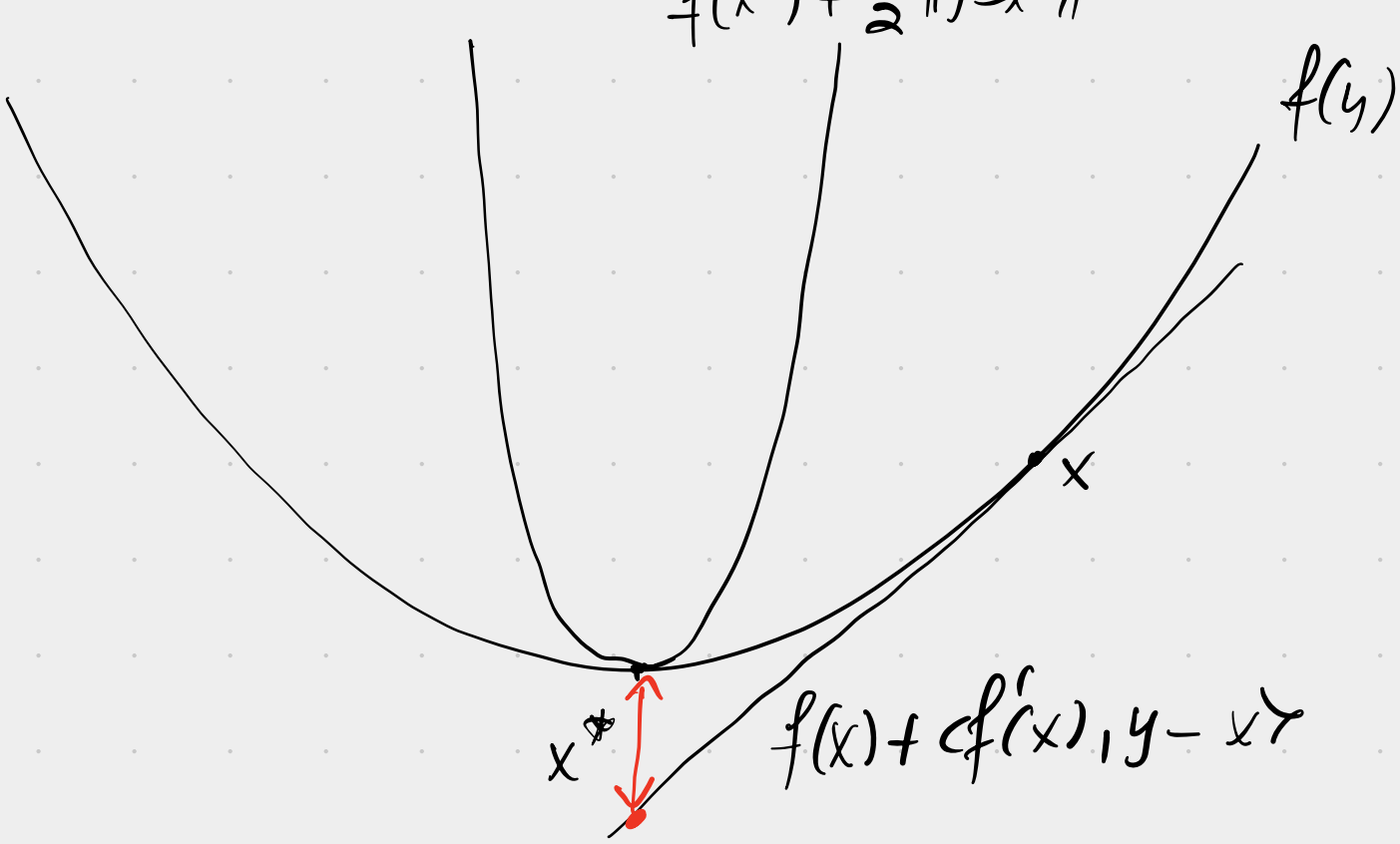
Useful Inequality: $\forall x, y \in \mathbb{R}^n$:

$$f(y) \geq f(x) + \langle f'(x), y - x \rangle + \frac{1}{2L}\|f'(y) - f'(x)\|^2$$

$y := x^*$:

$$\boxed{f(x^*) \geq f(x) + \langle f'(x), x^* - x \rangle + \frac{1}{2L}\|f'(x)\|^2} \quad \forall x$$

$f(x^*) + \frac{L}{2}\|y - x^*\|^2$



$$f(x) + \langle f'(x), y - x \rangle \leq f(y) \leq f(x^*) + \frac{L}{2}\|x^* - y\|^2 \quad \forall y$$

$$\star \quad f(x^*) \geq f(x) + \langle f'(x), x^* - x \rangle + \underbrace{\langle f'(x), y - x^* \rangle - \frac{L}{2}\|y - x^*\|^2}_{\to \max}$$

φ, $x^*_\varphi = x$, y, $\varphi'(y)$

$$= \frac{1}{2L}\|f'(x)\|^2$$

Rotate: Fix $x \in \mathbb{R}^n$; define

$$\varphi(y) = f(y) - \langle f'(x), y \rangle \Rightarrow \varphi'(y) = f'(y) - f'(x)$$

$$x^*_\varphi := x, \quad \varphi^* = f(x) - \langle f'(x), x \rangle$$

$$f(x) - \langle f'(x), x \rangle \geq f(y) - \langle f'(x), y \rangle + \langle f'(y) - f'(x), x - y \rangle$$

$$+ \frac{1}{2L}\|f'(y)-f'(x)\|^2 .$$

Corollary $\forall x, y \in \mathbb{R}^n$, $f:\mathbb{R}^n \to \mathbb{R}$, convex and smooth.

$$f(y) \geq f(x) + \langle f'(x), y-x\rangle + \frac{1}{2L}\|f'(y)-f'(x)\|^2 .$$

Stochastic Gradient Method

$$x_{n+1} = x_n - \gamma g_n , \qquad g_n = g(x_n, \xi_n)$$

$$\mathbb{E} g_n = f'(x_n)$$

$$\mathbb{E}\|g_n - f'(x_n)\|^2 \leq \sigma^2$$

$$\frac{1}{2}\|x_{n+1} - x^*\|^2 = \frac{1}{2}\|x_n - x^* - \gamma g_n\|^2 =$$

$$= \frac{1}{2}\|x_n - x^*\|^2 + \frac{\gamma^2}{2}\|g_n\|^2 - \gamma\langle g_n, x_n - x^*\rangle$$

Rearranging:

$$\frac{1}{2}\|x_n - x^*\|^2 + \frac{\gamma^2}{2}\|g_n\|^2 = \frac{1}{2}\|x_{n+1} - x^*\|^2 + \gamma\langle g_n, x_n - x^*\rangle$$

$$\mathbb{E}_{\xi_n}\|g_n - f'(x_n)\|^2 = \mathbb{E}_{\xi_n}\|g_n\|^2 - \|f'(x_n)\|^2$$

$$\Rightarrow \quad \mathbb{E}\|g_n\|^2 = \mathbb{E}\|g_n - f'(x_n)\|^2 + \mathbb{E}\|f'(x_n)\|^2$$

$$\mathbb{E}_{\xi_k}\langle g_n, x_n - x^*\rangle = \langle f'(x_n), x_n - x^*\rangle \geq f(x_n) - f^* + \frac{1}{2L}\|f'(x_n)\|^2$$

$$\mathbb{E}\langle g_n, x_n - x^*\rangle \geq \mathbb{E}[f(x_n) - f^*] + \frac{1}{2L}\mathbb{E}\|f'(x_n)\|^2$$

Denote: $R_k^2 := \mathbb{E}\|x_n - x^*\|^2$

Lemma:

$$\frac{1}{2}R_n^2 - \frac{1}{2}R_{n+1}^2 + \frac{\gamma^2}{2}\mathbb{E}\|g_n - \nabla f(x_n)\|^2 + \frac{\gamma^2}{2}\mathbb{E}\|\nabla f(x_n)\|^2 \geq$$

($\mathbb{E}\|g_n - \nabla f(x_n)\|^2 \leq \sigma^2$ — variance)

$$\geq \gamma\,\mathbb{E}[f(x_n) - f^*] + \frac{\gamma}{2L}\mathbb{E}\|\nabla f(x_n)\|^2$$

($\frac{\gamma}{2L}\mathbb{E}\|\nabla f(x_n)\|^2$ — smoothness)

Corollary $\gamma \leq \frac{1}{L}$:

$$\frac{1}{2}R_n^2 - \frac{1}{2}R_{n+1}^2 + \frac{\gamma^2}{2}\mathbb{E}\|g_n - \nabla f(x_n)\|^2 \geq \gamma\,\mathbb{E}[f(x_n) - f^*].$$

Apply variance assumption:

$$\frac{1}{2}R_n^2 - \frac{1}{2}R_{n+1}^2 + \frac{\gamma^2\sigma^2}{2} \geq \gamma\,\mathbb{E}[f(x_n) - f^*]$$

Telescope for the first K iterations:

$$\underbrace{\frac{1}{2}R_0^2 - \frac{1}{2}R_K^2} + \frac{\gamma^2\sigma^2 K}{2} \geq \gamma\sum_{i=0}^{K-1}\mathbb{E}[f(x_i) - f^*]$$

$\leq \frac{1}{2}R_0^2$

$$\geq \gamma K\,\mathbb{E}[f(\bar{x}_K) - f^*]$$

Jensen's inequality.

Finally:

$$\mathbb{E}[f(\bar{x}_K) - f^*] \leq \underbrace{\frac{R_0^2}{2\gamma K}}_{\leq \frac{\varepsilon}{2}} + \underbrace{\frac{\gamma\sigma^2}{2}}_{\leq \frac{\varepsilon}{2}} \longrightarrow \min_{\gamma}$$

$$\frac{R_0^2}{2\gamma_*^2 K} = \frac{\sigma^2}{2} \Rightarrow \gamma_* = \frac{R_0}{\sigma\sqrt{K}}, \quad \gamma \leq \frac{1}{L}$$

$$\gamma := \min\left\{\frac{1}{L}, \frac{R_0}{\sigma\sqrt{K}}\right\}$$

$$\frac{R_0^2}{2K}\max\left\{L, \frac{\sigma\sqrt{K}}{R_0}\right\} + \frac{R_0}{\sigma\sqrt{K}}\cdot\frac{\sigma^2}{2} = O\left(\frac{LR_0^2}{K} + \frac{\sigma R_0}{\sqrt{K}}\right)$$

Theorem:

$$\mathbb{E}\left[f(\bar{X}_K) - f^*\right] \le O\left(\frac{LR_0^2}{K} + \frac{\sigma R_0}{\sqrt{K}}\right) \le \varepsilon$$

$$K = O\left(\frac{LR_0^2}{\varepsilon} + \left[\frac{\sigma R_0}{\varepsilon}\right]^2\right)$$

stochastic oracle calls.

Variance Reduction

1. Mini-Batching.

m stochastic samples

$$g_k = \frac{1}{m}\sum_{i=1}^{m} g(x_k, \xi_{k,i}) \quad \Rightarrow \quad \mathbb{E}_{\xi_k} g_k = f'(x_k)$$

$$\mathbb{E}\left\|g_k - f'(x_k)\right\|^2 =$$

$$= \mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m}\underbrace{\left[g(x_k, \xi_{k,i}) - f'(x_k)\right]}_{\delta_i}\right\|^2 \quad (*)$$

$\mathbb{E}\delta_i = 0$

$$(*) \quad \mathbb{E}\left\langle \frac{1}{m}\sum_{i=1}^{m}\delta_i, \frac{1}{m}\sum_{j=1}^{m}\delta_j\right\rangle = \frac{1}{m^2}\sum_{i=1}^{m}\mathbb{E}\|\delta_i\|^2 + \frac{2}{m^2}\sum_{i<j}\overbrace{\mathbb{E}\langle\delta_i, \delta_j\rangle}^{=0}$$

$\leq \frac{1}{m}\sigma^2.$

$$\frac{LR_0^2}{\varepsilon} \stackrel{?}{=} \left[\frac{\sigma R_0}{\varepsilon\sqrt{m}}\right]^2 = \frac{\sigma^2 R_0^2}{m\varepsilon^2}$$

$$\Rightarrow \quad m = 1 + \frac{\sigma^2 R_0^2}{\varepsilon^2}\cdot\frac{\varepsilon}{LR_0^2} = \frac{\sigma^2}{\varepsilon L} + 1.$$

The iteration complexity: ← solving some subproblem

$$O\left(\frac{LR_0^2}{q}\right)$$

The total # samples:

$$O\left(\frac{LR_0^2}{\varepsilon}\right) * m = O\left(\boxed{\frac{LR_0^2}{\varepsilon}} + \boxed{\frac{\sigma^2 R_0^2}{\varepsilon^2}}\right)$$

Acceleration: $\sqrt{\frac{LR_0^2}{q}}$

$m \gtrsim 100$

---

2. SVRG

Finite-Sum Minimization:

$$f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x), \quad f_i(x) - \text{Lipschitz, convex}$$

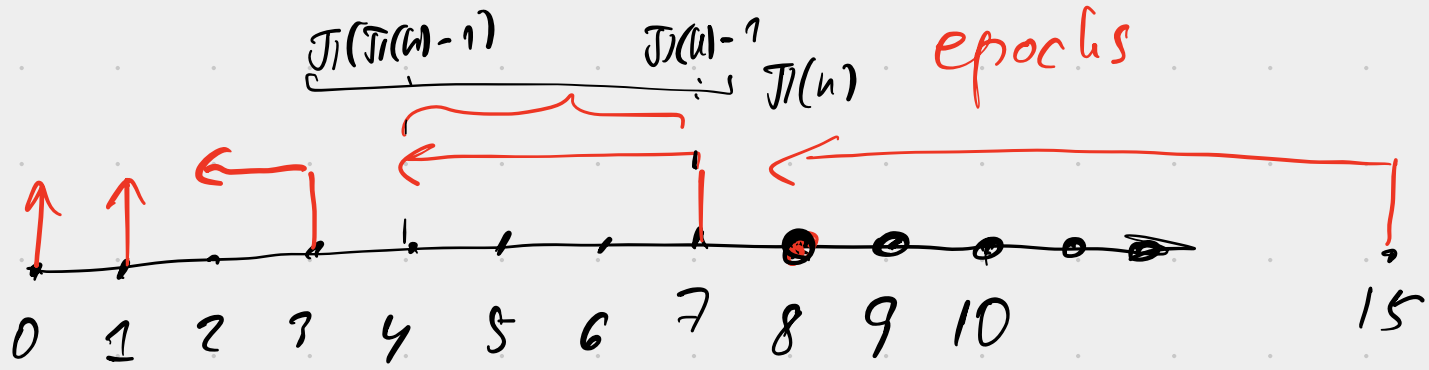
We use in the algorithm: sample i

$$g_k = f_i'(x_k) - f_i'(z_k) + f'(z_k)$$

z_k - fixed point from the past. ← compute it

Let $\pi(k)$ the largest power of 2 that is less than or equal to k.

$$\pi(k) = 2^{\lfloor \log_2 k \rfloor} \qquad \pi(0) = 0$$



$$z_u = x_{\pi(u)} \; ?$$

$$z_u := \frac{1}{\pi(u) - \pi(\pi(u)-1)} \sum_{i=\pi(\pi(u)-1)}^{\pi(u)-1} x_i$$

<u>lemma</u> $\mathbb{E}\, g_u = f'(x_u)$

$$\mathbb{E}\,\|g_u - f'(x_u)\|^2 \le 6L\,\mathbb{E}[f(x_u) - f^*] + 6L\,\mathbb{E}[f(z_u) - f^*]$$

$$\longrightarrow 0$$