Problem Class: $f : \mathbb{R}^n \to \mathbb{R}$

$$f^* = \inf_{x \in \mathbb{R}^n} f(x) > -\infty$$

Goal: Find $\bar{x}$: $\|f'(\bar{x})\|_* \leq \varepsilon$.

Smoothness: $\|f'(y) - f'(x)\|_* \leq L \cdot \|y - x\| \quad \forall x, y \in \mathbb{R}^n$.

Init. $x_0 \in \mathbb{R}^n$. Iterate, $k \geq 0$:

$$x_{k+1} = x_k^+(M_k) = \arg\min_{y \in \mathbb{R}^n} \left[f(x_k) + \langle f'(x_k), y - x_k \rangle + \frac{M_k}{2} \|y - x_k\|^2 \right]$$

Norm is Euclidean: $x_{k+1} = x_k - \frac{1}{M} f'(x_k)$.

For $M \geq L$: $f(x_k) - f(x_k^+(M)) \geq \frac{1}{2M} \|f'(x_k)\|_*^2$.

In theory: $\boxed{M := L}$.

Adaptive Search.

$$M_k \approx L$$

Gradient Method with Adaptive Search.

Init.: $x_0 \in \mathbb{R}^n$, $\varepsilon > 0$, $M_0 > 0$

Iterations, $k \geq 0$:

1. If $\|f'(x_k)\|_* \leq \varepsilon$ then
   return x_k.

2. For $t \geq 0$ iterate:
   - Set $M_k^+ = M_k \cdot 2^t$.
   - Try gradient step: $x_k^+ := x_k^+(M_k^+)$
   - If $f(x_k) - f(x_k^+) \geq \frac{1}{2M_k^+}\|f'(x_k)\|_*^2$ then

     break and go to step 3.

3. $x_{k+1} = x_k^+$, $M_{k+1} = M_k^+ \cdot \frac{1}{2}$.

As soon as $M_k \cdot 2^t \geq L$ then we exit from inner loop.

<u>Proposition 1</u> For $k > 0$:

$$M_k \leq M_* := \max\{M_0, L\}.$$

<u>Proof</u> For $k = 0$. $M_0 \leq M_*$. true.

Consider $k \geq 0$.

Denote $t_k \geq 0$ the value of parameter t on step 2 that triggered the break.

$$M_{k+1} = M_k \cdot 2^{t_k - 1}.$$

Case 1 $t_k = 0$. $\Rightarrow M_{k+1} = \frac{1}{2} M_k \leq \frac{1}{2} \cdot M_* \leq M_*$.

$t_k > 0$. $\Rightarrow M_{k+1} \leq L \leq M_*$. $\square$

Corollary: To find $\bar{x}$: $\|f'(\bar{x})\|_* \leq \varepsilon$ it's enough to do

$$K = \left\lceil \frac{4 \cdot \max\{M_0, L\}(f(x_0) - f^*)}{\varepsilon^2} \right\rceil.$$

Proposition 2 Let N_K the total number of oracle calls during the first K iterations.

$$N_K \leq 2K + \max\{0, 1 + \log_2 \frac{L}{M_0}\}.$$

Proof.

$$2^{t_k - 1} = \frac{M_{k+1}}{M_k} \Rightarrow t_k = 1 + \log_2 \frac{M_{k+1}}{M_k}.$$

$$N_K = \sum_{i=0}^{K-1} (1 + t_i) = 2 \cdot K + \sum_{i=0}^{K-1} \log_2 \frac{M_{i+1}}{M_i}$$

$$= 2K + \log_2 \frac{M_K}{M_0} \leq 2K + \log_2 \frac{M_*}{M_0}.$$

$\square$.

Stochastic Gradient Method.

Example (Expensive)

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \quad , \quad f_i : \mathbb{R}^n \to \mathbb{R}$$

N – size of dataset.

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

– expensive to spend $O(N)$

but, it's affordable $f_i'(x)$

Example (Impossible)

$$f(x) = \mathbb{E}_\xi F(x, \xi)$$

$F(\cdot, \xi)$ is smooth, for every ξ.

Example (We can, but we don't want)

e.g. when $f_i(x)$ reveals user data.

Stochastic Oracle

$\|\cdot\| = \|\cdot\|_2$ euclidean.

Stochastic first-order oracle by

$$g(x, \xi) \in \mathbb{R}^n$$

Ex. Finite sum: $\xi = i \sim \{1, \dots, N\}$

$g(x, \xi) = f_i'(x)$

1. Unbiased estimator: $\forall x \in \mathbb{R}^n$

$$\mathbb{E}_\xi [g(x, \xi)] = f'(x)$$

2. Bounded variance: $\sigma > 0$ s.t. $\forall x \in \mathbb{R}^n$

$$\mathbb{E}_\xi \|g(x, \xi) - f'(x)\|^2 \leq \sigma^2. \quad (2)$$

$$\|g(x, \xi) - f'(x)\|^2 = \|g(x, \xi)\|^2 + \|f'(x)\|^2 - 2\langle g(x, \xi), f'(x)\rangle$$

$$\sigma^2 \geq \mathbb{E}\|g(x,\xi) - f'(x)\|^2 = \mathbb{E}\|g(x,\xi)\|^2 - \|f'(x)\|^2$$

$$(2) \iff \mathbb{E}\|g(x,\xi)\|^2 \leq \|f'(x)\|^2 + \sigma^2.$$

Stochastic Gradient Method.

Initialization: $x_0 \in \mathbb{R}^n$, $M > 0$, number of iterations $K \geq 1$.

For $k = 0 \dots K-1$:

1. Sample ξ_k
2. Compute $g_k = g(x_k, \xi_k)$
3. Gradient Step:

$$x_{k+1} = x_k - \frac{1}{M} g_k.$$

Sample $j \in \{0, \dots, K-1\}$ uniformly and return $\bar{x}_K = x_j$.

Problem:

$$\min_x f(x)$$

Find a point $\bar{x}$: $\mathbb{E}\|f'(\bar{x})\| \leq \varepsilon$.

Proposition Let $M > 0$.

$$x_{k+1} = x_k - \frac{1}{M} g_k \text{ with } g_k = g(x_k, \xi_k)$$

Then

$$\mathbb{E}_{\xi_k}\left[f(x_k) - f(x_{k+1})\right] \geq \frac{1}{M}\|f'(x_k)\|^2 \cdot \left(\underbrace{1 - \frac{L}{2M}}\right) - \frac{L}{2M^2}\sigma^2 .$$

Remark $M \geq L \iff 1 - \frac{L}{2M} \geq \frac{1}{2}$.

We obtain:

$$\mathbb{E}_{\xi_k}\left[f(x_k) - f(x_{k+1})\right] \geq \frac{1}{2M}\|f'(x_k)\|^2 - \frac{L}{2M^2}\sigma^2$$

If we $\sigma = 0 \Rightarrow$ the same as for GD.

Stochastic Gradient Descent (SGD)

Proof Smoothness:

$$f(x_{k+1}) \leq f(x_k) + \langle f'(x_k), \underbrace{x_{k+1} - x_k}_{= -\frac{1}{M} g_k} \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

$$= f(x_k) - \frac{1}{M}\langle f'(x_k), g_k\rangle + \frac{L}{2M^2}\|g_k\|^2$$

Rearrange:

$$\mathbb{E}_{\xi_k}\left[f(x_k) - f(x_{k+1})\right] \geq \mathbb{E}_{\xi}\left[\frac{1}{M}\langle f'(x_k), g_k\rangle - \frac{L}{2M^2}\|g_k\|^2\right]$$

$$= \frac{1}{M}\|f'(x_k)\|^2 - \frac{L}{2M^2}\|f'(x_k)\|^2 - \frac{L}{2M^2}\sigma^2 \quad \square.$$

Theorem $M \geq L$. Consider $K \geq 1$ iterations.

$$\mathbb{E}\|f'(\overline{x}_K)\|^2 \leq \frac{2M(f(x_0) - f^*)}{K} + \frac{L}{M}\sigma^2.$$

Proof.

$\mathbb{E}[\cdot]$ the expectation w.r.t. all randomness $\xi_1, \ldots, \xi_K, j$.

$\mathbb{E}_{\xi}[\cdot]$ the expectation w.r.t all $\xi_1, \ldots, \xi_K$.

$$\mathbb{E}_{\xi}\left[f(x_k) - f(x_{k+1})\right] \geq \frac{1}{2M}\mathbb{E}_{\xi}\left[\|f'(x_k)\|^2\right] - \frac{L}{2M^2}\sigma^2.$$

Telescope it for $K \geq 1$ iterations:

$$f(x_0) - f^* \geq \mathbb{E}_{\xi}\left[f(x_0) - f(x_k)\right] \geq \frac{1}{2M}\sum_{i=0}^{K-1}\mathbb{E}_{\xi}\left[\|f'(x_i)\|^2\right]$$

$$- K \cdot \frac{L}{2M^2}\sigma^2.$$

$$\frac{2M}{K}(f(x_0) - f^*) \geq \frac{1}{K}\sum_{i=0}^{K-1}\mathbb{E}_{\xi}\left[\|f'(x_i)\|^2\right]$$

$$- \frac{L}{M}\sigma^2 =$$

$$= \mathbb{E}\left[\|f'(\bar{x}_K)\|^2\right] - \frac{L}{M}\sigma^2 \quad \square$$

Stepsize Tuning

$M \geq L$.

$$\mathbb{E}\left[\|f'(\bar{x}_k)\|^2\right] \leq \frac{2MF_0}{K} + \frac{L}{M}\sigma^2$$

where $F_0 = f(x_0) - f^*$.

Naive: $M = L$:

$$\mathbb{E}\left[\|f'(\bar{x}_k)\|^2\right] \leq \frac{2LF_0}{K} + \sigma^2.$$

Now:

1. $\dfrac{2MF_0}{K} \leq \dfrac{\varepsilon^2}{2}$

2. $\dfrac{L}{M}\sigma^2 \leq \dfrac{\varepsilon^2}{2} \Rightarrow \boxed{M = L \cdot \underbrace{\max\left\{1, \frac{2\sigma^2}{\varepsilon^2}\right\}}_{\leq 1 + \frac{2\sigma^2}{\varepsilon^2}}}$.

(1) $\boxed{K = 1 + \left\lfloor \dfrac{4MF_0}{\varepsilon^2} \right\rfloor}$

Corollary To find a random point $\bar{x} \in \mathbb{R}^n$ s.t. $\mathbb{E}\|f'(\bar{x})\| \leq \varepsilon$ it's enough to do

$$K = O\left(L \cdot (f(x_0) - f^*) \cdot \left[\frac{1}{\varepsilon^2} + \frac{\sigma^2}{\varepsilon^4}\right]\right)$$

stochastic oracle calls.

$\Rightarrow$ optimal complexity.