$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$ — Empirical Risk Minimization Problem

<u>Problem 1</u>

$$\boxed{x_{k+1} = x_k - \eta \nabla f_i(x_k)}$$

SGD

i is sampled uniformly at random from $\{1, \dots, n\}$

The cost of computing full gradient is $\underline{O(n \cdot d)}$

stoch gradient is $\underline{\underline{O(d)}}$

$n \gg 1$

1) How many samples are left unseen in expectation after one epoch (n iterations)

$$X_i = \begin{cases} 1, & \text{if index } i \text{ is never sampled in one epoch} \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = ?$$

$$\parallel$$

$$\sum_{i=1}^{n} \mathbb{E}[X_i] = n \cdot \mathbb{E}[X_2]$$

$$\mathbb{E}[X_1] = \mathbb{P}(X_1 = 1)\cdot 1 + \mathbb{P}(X_1 = 0)\cdot 0$$
$$= \mathbb{P}(X_1 = 1)$$

$\mathbb{P}$(index i is sampled at current iteration) $= \frac{1}{n}$

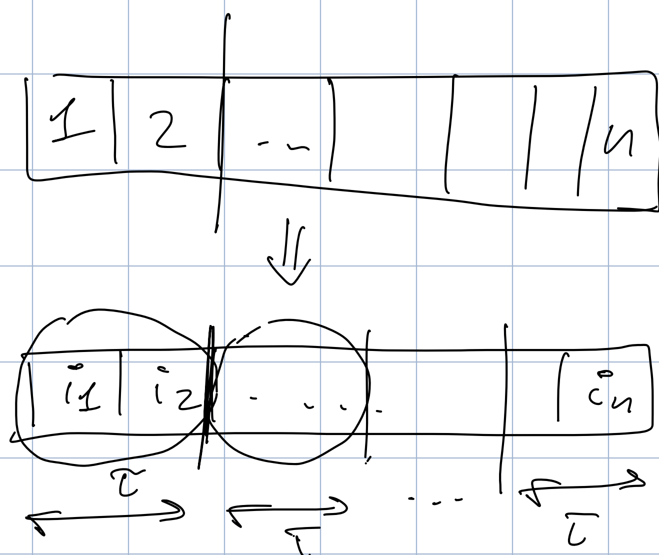
$\mathbb{P}$(index i is not sampled at current iteration) $= 1 - \frac{1}{n}$

$$\underline{\mathbb{P}(X_1 = 1) = \left(1 - \frac{1}{n}\right)^n} = \mathbb{E}[X_1]$$

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = n\left(1 - \frac{1}{n}\right)^n \xrightarrow{n\to\infty} \frac{n}{e} \approx 0.368\, n$$

$$\lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^n = e^{-1}$$

Random Reshuffling:



2) $\mathbb{E}[\|x_{k+1} - x^*\|^2] \geq \left(\frac{1}{2L}\right)^2 \mathbb{E}[\|\nabla f_i(x_k)\|^2]$ $\quad\left(\gamma = \frac{1}{2L}\right)$

$$\mathbb{E}\left[\|x_{k+1}-x^*\|^2\right] = \mathbb{E}\left[\|x_k - \eta \nabla f_i(x_k) - x^*\|^2\right] \quad \text{(step of SGD)}$$

$$= \mathbb{E}\left[\|x_k - x^*\|^2\right] - 2\eta\, \mathbb{E}\left[\langle x_k - x^*; \nabla f_i(x_k)\rangle\right]$$

$$+ \eta^2 \mathbb{E}\left[\|\nabla f_i(x_k)\|^2\right] \quad \text{(simple linear algebra)}$$

$$= \mathbb{E}\left[\|x_k - x^*\|^2\right] - 2\eta\, \mathbb{E}\left[\langle x_k - x^*; \nabla f(x_k)\rangle\right]$$

$$+ \eta^2 \mathbb{E}\left[\|\nabla f_i(x_k)\|^2\right] \quad \text{(unbiasedness of } \nabla f_i(x_k))$$

(Cauchy-Schwartz inequality says that $\forall a, b \in \mathbb{R}^d$

$$\underbrace{|\langle a, b\rangle| \leq \|a\| \cdot \|b\|}_{\Updownarrow} \longrightarrow -\langle a, b\rangle \geq -\|a\| \cdot \|b\|$$

$$-\|a\| \cdot \|b\| \leq \langle a, b\rangle \leq \|a\| \cdot \|b\|$$

$$\geq \mathbb{E}\left[\|x_k - x^*\|^2\right] - 2\eta\, \mathbb{E}\left[\|x_k - x^*\| \cdot \|\nabla f(x_k)\|\right] \quad \text{(Cauchy-Schwartz)}$$

$$+ \eta^2 \mathbb{E}\left[\|\nabla f_i(x_k)\|^2\right]$$

(We assume that f is L-smooth)

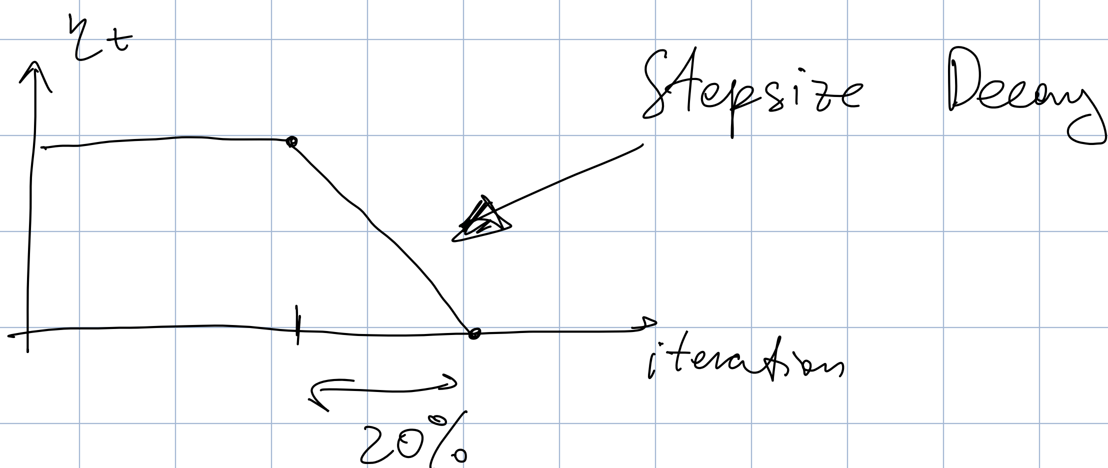
$$\|\nabla f(x_k)\| = \|\nabla f(x_k) - \nabla f(x^*)\| \leq L\|x_k - x^*\|$$

$$\geq \mathbb{E}\left[\|x_k - x^*\|^2\right] - 2\eta L\, \mathbb{E}\left[\|x_k - x^*\|^2\right]$$

$$+\eta^2 \mathbb{E}\left[\|\nabla f_i(x_k)\|^2\right]$$

$$\eta = \frac{1}{2L} \quad = \left(\frac{1}{2L}\right)^2 \mathbb{E}\left[\|\nabla f_i(x_k)\|^2\right]$$

$$\mathbb{E}\left[\|x_{k+1} - x^*\|^2\right] \geq \left(\frac{1}{2L}\right)^2 \mathbb{E}\left[\|\nabla f_i(x_k)\|^2\right] >$$



2) Variance Reduction

i is sampled

SVRG: $\underbrace{g_k} = \nabla f_i(x_k) - \nabla f_i(w_k) + \nabla f(w_k)$

SARAH: $\underbrace{g_k} = \nabla f_{i_k}(x_k) - \nabla f_{i_k}(w_k) + \nabla f_{i_{k-1}}(w_{k-1})$

L-SVRG

SAGA

Problem 2

Assumptions: f is μ-strongly convex; L-smooth
H-Lipschitz Hessian
B - Bounded stoch gradients.

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq H \cdot \|x-y\| \quad \text{(H-Lipschitz of Hes)}$$
$$\|\nabla f_i(x_k)\|^2 \leq B^2 \quad \forall i \in \{1, \dots, n\}$$

1) $g(x) = \|\nabla f(x)\|^2$

$$\nabla g(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} g(x) \\ \vdots \\ \frac{\partial}{\partial x_d} g(x) \end{pmatrix} \text{ - gradient of } g(x)$$

$$\frac{\partial}{\partial x_u} g(x) = \frac{\partial}{\partial x_u} \|\nabla f(x)\|^2 = \frac{\partial}{\partial x_u}\left(\sum_{e=1}^{d} \left(\frac{\partial f}{\partial x_e}\right)^2 \right)$$

$$= \sum_{e=1}^{d} \frac{\partial}{\partial x_u} \left(\frac{\partial f}{\partial x_e}\right)^2 = \sum_{e=1}^{d} 2 \cdot \frac{\partial f}{\partial x_e} \frac{\partial}{\partial x_u}\left(\frac{\partial f}{\partial x_e}\right)$$

$$= \sum_{e=1}^{d} 2 \frac{\partial^2 f}{\partial x_u \partial x_e} \cdot \frac{\partial f}{\partial x_e} =$$

$$\nabla^2 f(x) = \begin{array}{c} k \end{array} \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_k \partial x_1} & \cdots & \dfrac{\partial^2 f}{\partial x_k \partial x_d} \end{pmatrix} \quad \text{— Hessian}$$

$$= 2\left[\nabla^2 f(x) \cdot \nabla f(x)\right]_k$$

$$\boxed{\nabla g(x) = 2\nabla^2 f(x) \cdot \nabla f(x)}$$

Let's show that g is $\tilde{L}$-smooth

$$\|\nabla g(x) - \nabla g(y)\| = \|\underbrace{2\nabla^2 f(x)\nabla f(x)}_{=\nabla g(x)} - \underbrace{2\nabla^2 f(y)\nabla f(y)}_{=\nabla g(y)}\|$$

$$= 2\|\underbrace{\nabla^2 f(x)\nabla f(x) - \nabla^2 f(y)\nabla f(x)} + \underbrace{\nabla^2 f(y)\nabla f(x) - \nabla^2 f(y)\nabla f(y)}\|$$

$$= 2\|\underbrace{(\nabla^2 f(x) - \nabla^2 f(y))\nabla f(x)}_{a} + \underbrace{\nabla^2 f(y)(\nabla f(x) - \nabla f(y))}_{b}\|$$

(Triangle inequality: $\|a+b\| \le \|a\| + \|b\| \quad \forall a, b \in \mathbb{R}^d$)

$$\le 2\|\underbrace{(\nabla^2 f(x) - \nabla^2 f(y))}_{A}\underbrace{\nabla f(x)}_{x}\| + 2\|\underbrace{\nabla^2 f(y)}_{A}\underbrace{(\nabla f(x) - \nabla f(y))}_{x}\|$$

(for some matrix A we have $\|Ax\| \leq \|A\| \cdot \|x\|$)
some vector x

$$\leq 2 \underbrace{\|\nabla^2 f(x) - \nabla^2 f(y)\|} \cdot \|\nabla f(x)\| + 2\|\nabla^2 f(y)\| \cdot \underbrace{\|\nabla f(x) - \nabla f(y)\|}$$

$$\leq 2 \underbrace{H\|x-y\|} \cdot \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x) \right\| + 2\underbrace{\|\nabla^2 f(y)\|} \cdot \underbrace{L\|x-y\|}$$

Triangle inequality

$$\leq 2H\|x-y\| \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x)\| \right) + 2L \cdot L\|x-y\|$$

(∇f is L-Lipschitz $\Leftrightarrow$ $\nabla^2 f(x) \lesssim L \cdot Id$
($\|\nabla^2 f(x)\| \leq L$))

$$\leq 2H \cdot \|x-y\| \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} B \right)}_{B} + 2L^2 \|x-y\|$$

$$= \underbrace{2(BH + L^2)}_{\tilde{L}} \|x-y\|$$

$g(x)$ — $\tilde{L}$-smooth with $\tilde{L} = 2(BH + L^2)$

$$g(x_{k+1}) \leq g(x_k) + \langle \nabla g(x_k), x_{k+1} - x_k \rangle + \frac{\tilde{L}}{2} \| x_{k+1} - x_k \|^2$$
(smoothness inequality)

$$= g(x_k) + \langle 2 \nabla^2 f(x_k) \nabla f(x_k), -\eta \nabla f_i(x_k) \rangle + \frac{\tilde{L}}{2} \| -\eta \nabla f_i(x_k) \|^2$$

$$= g(x_k) - 2\eta \langle \nabla^2 f(x_k) \nabla f(x_k), \nabla f_i(x_k) \rangle + \frac{\tilde{L}\eta^2}{2} \| \nabla f_i(x_k) \|^2$$

Let's use unbiasedness of $\nabla f_i(x_k)$:

$$\mathbb{E}[g(x_{k+1})] \leq \mathbb{E}[g(x_k)] - 2\eta\, \mathbb{E}[\langle \nabla^2 f(x_k) \nabla f, \nabla f_i(x_k) \rangle] + \frac{\tilde{L}\eta^2}{2} \mathbb{E}[\| \nabla f_i(x_k) \|^2]$$

$$\mathbb{E}[g(x_{k+1})] \leq \mathbb{E}[g(x_k)] \underbrace{- 2\eta\, \mathbb{E}[\langle \nabla^2 f(x_k) \nabla f(x_k), \nabla f(x_k) \rangle]} + \frac{\tilde{L}\eta^2}{2} \mathbb{E}[\| \nabla f_i(x_k) \|^2]$$

$$\nabla^2 f(x_k) \succeq \mu \cdot I_d \implies$$

$$\Rightarrow \nabla f(x_k)^T \nabla^2 f(x_k) \nabla f(x_k) \geq \underbrace{\nabla f(x_k)^T \cdot \mu \cdot Id \cdot \nabla f(x_k)}_{= \mu \|\nabla f(x_k)\|^2}$$

$$\Rightarrow -\nabla f(x_k)^T \nabla^2 f(x_k) \nabla f(x_k) \leq -\mu \|\nabla f(x_k)\|^2$$

$$\mathbb{E}[g(x_{k+1})] \leq \mathbb{E}[g(x_k)] - 2\eta\mu \, \mathbb{E}[\|\nabla f(x_k)\|^2] + \frac{\tilde{L}\eta^2}{2} \mathbb{E}[\|\nabla f_i(x_k)\|^2]$$

$$= \mathbb{E}[g(x_k)] - 2\eta\mu \cdot \mathbb{E}[g(x_k)] + \frac{\tilde{L}\eta^2}{2} \underbrace{\mathbb{E}[\|\nabla f_i(x_k)\|^2]}_{\leq B^2}$$

$$\leq (1 - 2\eta\mu) \mathbb{E}[g(x_k)] + \frac{\tilde{L}\eta^2 B^2}{2}$$

Final inequality:

$$\mathbb{E}[g(x_{k+1})] \leq (1 - 2\eta\mu) \mathbb{E}[g(x_k)] + \frac{\tilde{L}\eta^2 B^2}{2}$$

$$\leq (1-2\eta\mu)\left((1-2\eta\mu)\mathbb{E}[g(x_{k-2})] + \frac{\tilde{L}\eta^2B^2}{2}\right) + \frac{\tilde{L}\eta^2B^2}{2}$$

$$= (1-2\eta\mu)^2\mathbb{E}[g(x_{k-2})] + \frac{\tilde{L}\eta^2B^2}{2}\left(1 + (1-2\eta\mu)\right)$$

$\dots$

$$\leq (1-2\eta\mu)^{k+1}\mathbb{E}[g(x_0)] + \frac{\tilde{L}\eta^2B^2}{2}\sum_{\ell=0}^{k}(\underbrace{1-2\eta\mu}_{q})^{\ell}$$

$$\sum_{\ell=0}^{\infty} q^{\ell} =: \frac{1}{1-q} \quad \text{for } q<1 \Rightarrow \sum_{\ell=0}^{k} q^{\ell} \leq \frac{1}{1-q}$$

$$\leq (1-2\eta\mu)^{k+1}\mathbb{E}[g(x_0)] + \frac{\tilde{L}\eta^2B^2}{2}\cdot\frac{1}{1-(1-2\eta\mu)}$$

$$= (1-2\eta\mu)^{k+1}\mathbb{E}[g(x_0)] + \frac{\tilde{L}\eta^2B^2}{2}\cdot\frac{1}{2\eta\mu}$$

$$\mathbb{E}[g(x_k)] \leq (1-2\eta\mu)^{k} g(x_0) + \frac{\tilde{L}\eta B^2}{4\mu}$$

Problem 3

1) L-smoothness of f:

$$f(x_{n+1}) \leq f(x_n) + \langle \nabla f(x_n), \underbrace{x_{n+1} - x_n} \rangle + \frac{L}{2} \|\underbrace{x_{n+1} - x_n}\|^2$$

$$= f(x_n) + \langle \nabla f(x_n), -\eta \nabla f_i(x_n) \rangle + \frac{L}{2} \|-\eta \nabla f_i(x_n)\|^2$$

$$= f(x_n) - \eta \langle \nabla f(x_n), \nabla f_i(x_n) \rangle + \frac{L\eta^2}{2} \|\nabla f_i(x_n)\|^2$$

$$\mathbb{E}[f(x_{n+1})] \leq \mathbb{E}[f(x_n)] - \eta \mathbb{E}[\langle \nabla f(x_n), \nabla f_i(x_n) \rangle] + \frac{L\eta^2}{2} \mathbb{E}[\|\nabla f_i(x_n)\|^2]$$

Unbiasedness

$$\underset{\|\nabla f_i(x_n)\| \leq B}{=} \mathbb{E}[f(x_n)] - \eta \mathbb{E}[\langle \nabla f(x_n), \nabla f(x_n) \rangle] + \frac{L\eta^2}{2} B^2$$

$$= \mathbb{E}[f(x_n)] - \eta \mathbb{E}[\|\nabla f(x_n)\|^2] + \frac{L\eta^2}{2} B^2$$

μ-PL inequality: $\|\nabla f(x_n)\|^2 \geq 2\mu (f(x_n) - f^*)$

$\Leftrightarrow -\|\nabla f(x_n)\|^2 \leq -2\mu (f(x_n) - f^*)$

$$\le \mathbb{E}[f(x_k)] - 2\eta\mu\, \mathbb{E}[f(x_k) - f^*] + \frac{L\eta^2 B^2}{2}$$

We get

$$\mathbb{E}[f(x_{k+1})] \le \mathbb{E}[f(x_k)] - 2\eta\mu\, \mathbb{E}[f(x_k) - f^*] + \frac{L\eta^2 B^2}{2}$$

Let's subst. $-f^*$

$$\mathbb{E}[f(x_{k+1}) - f^*] \le \mathbb{E}[f(x_k) - f^*] - 2\mu\eta\, \mathbb{E}[f(x_k) - f^*] + \frac{L\eta^2 B^2}{2}$$

2) $$\mathbb{E}[f(x_{k+1}) - f^*] \le \mathbb{E}[f(x_k) - f^*] - 2\mu \cdot \eta_k\, \mathbb{E}[f(x_k) - f^*] + \frac{L\eta_k^2 B^2}{2}$$

$$\eta_k = \frac{2k+1}{2\mu(k+1)^2}$$

$$\mathbb{E}[f(x_{k+1}) - f^*] \le \mathbb{E}[f(x_k) - f^*] - \cancel{2\mu} \cdot \frac{2k+1}{\cancel{2\mu}(k+1)^2}\, \mathbb{E}[f(x_k) - f^*]$$

$$+ \frac{L(2k+1)^2}{2\, 4\mu^2 (k+1)^4} B^2$$

$$\leq \left(1 - \frac{2k+1}{(k+1)^2}\right) \mathbb{E}\left[f(x_k) - f^*\right]$$

$$+ \frac{L(2k+2)^2}{2 \cdot 4\mu^2 (k+1)^4} B^2$$

$$= \left(\frac{(k+1)^2 - (2k+1)}{(k+1)^2}\right) \mathbb{E}\left[f(x_k) - f^*\right]$$

$$+ \frac{L \cdot 4 \cdot (k+1)^2}{2 \cdot 4\mu^2 (k+1)^4} B^2$$

$$= \frac{k^2}{(k+1)^2} \mathbb{E}\left[f(x_k) - f^*\right]$$

$$+ \frac{LB^2}{2\mu^2 (k+1)^2}$$

We obtain: $\mathbb{E}\left[f(x_{k+1}) - f^*\right] \leq \frac{k^2}{(k+1)^2} \mathbb{E}\left[f(x_k) - f^*\right]$

$$+\frac{LB^2}{2\mu^2(k+1)^2}$$

$$\mathbb{E}\Big[\underbrace{(k+1)^2(f(x_{k+1})-f^*)}_{\delta_f(k+1)}\Big] \le \mathbb{E}\Big[\underbrace{k^2(f(x_k)-f^*)}_{\delta_f(k)}\Big] + \frac{LB^2}{2\mu^2}$$

$$\begin{aligned}
\mathbb{E}[\delta_f(k+1)] &\le \mathbb{E}[\delta_f(k)] + \frac{LB^2}{2\mu^2} \\
&\le \mathbb{E}[\delta_f(k-1)] + 2\cdot\frac{LB^2}{2\mu^2} \\
&\le \mathbb{E}[\delta_f(k-2)] + 3\cdot\frac{LB^2}{2\mu^2} \\
&\quad \cdots\cdots \\
&\le \mathbb{E}[\delta_f(0)] + (k+1)\cdot\frac{LB^2}{2\mu^2}
\end{aligned}$$

$$\delta_f(0) = 0^2\cdot(f(x_0)-f^*) = 0$$

$$\mathbb{E}[\delta_f(k+1)] \le (k+1)\cdot\frac{LB^2}{2\mu^2}$$

This implies:

$$\mathbb{E}\left[k^2 (f(x_k) - f^*)\right] \le k \cdot \frac{LB^2}{2\mu^2}$$

$$\mathbb{E}\left[f(x_k) - f^*\right] \le \frac{1}{k} \cdot \frac{LB^2}{2\mu^2}$$ — convergence rate

<u>Or</u>: $\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - \dots$