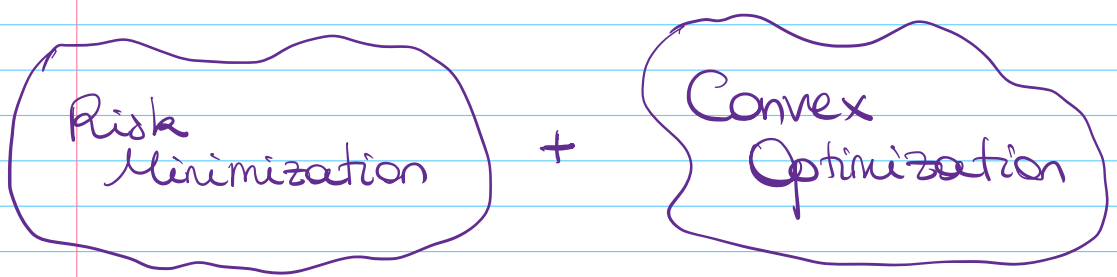
Stochastic Convex Optimization



Back to the basics :

$(x_1, y_1) \dots (x_n, y_n) \sim P$

- $\mathcal{X}$ set of inputs
- $\mathcal{Y}$ set of outputs

$\ell(y, y')$ loss function.

Goal: Find a predictor $f: \mathcal{X} \to \mathcal{Y}$ that minimizes the P-risk ← test risk

$$\boxed{R_P(f) = \mathbb{E}_P[\ell(f(X), Y)]}$$

To do so: introduce a class of predictors

$f_\theta(x) = \langle x, \theta \rangle$ $\quad \mathcal{F} = \{ f_\theta : \mathcal{X} \to \mathcal{Y} : \theta \in \mathbb{R}^d \}$

$\Rightarrow$ Consider $\theta^* = \operatorname{argmin}_\theta R_P(f_\theta)$

How we did so far:

$\frac{1}{n}\sum_i \ell(f_\theta(x_i), y_i)$

1. Approximate $R_P(f_\theta)$ by $R_n(f_\theta)$
2. Minimize $R_n(f_\theta)$ using GD.

$\hat{\theta} = (X^T X)^{-1} X^T y$ (or closed form in certain cases)

$$R_n(f_\theta) = \frac{1}{n}\sum_{i=1}^n \ell_\theta(f_\theta(x_i), y_i)$$

$$\nabla R_n(f_\theta) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta \ell_\theta(f_\theta(x_i), y_i)$$

n gradients to compute

GD: $\theta^{t+1} = \theta^t - s \nabla R_n(f_{\theta^t})$

$\Rightarrow$ T steps of GD $= O(nT)$

Can we do better ?

Idea: Apply GD directly on

$$\theta \mapsto R_P(f_\theta) = \mathbb{E}_P[\ell(f_\theta(X), Y)]$$

$$\nabla R_P(f_\theta) = \mathbb{E}_P[\nabla_\theta \ell(f_\theta(X), Y)]$$

... But we do not have access to $\nabla R_P(f_\theta)$

$\Rightarrow$ Estimate it by $\nabla_\theta \ell(f_\theta(x_i), y_i)$

$$\theta_{t+1} = \theta_t - s \nabla_\theta \ell(f_{\theta_t}(x_i), y_i)$$

STOCHASTIC GRADIENT DESCENT

$\rightsquigarrow$ 1 step of SGD $= O(1)$

① Stochastic Gradient Descent

A more general setting

$F: \mathbb{R}^d \longrightarrow \mathbb{R}$

SGD : • θ^1 initialization

For $t = 1 \dots T-1$:

• V^t random vector such that

$$\mathbb{E}[V^t | \theta^t] = \nabla F(\theta^t)$$

Unbiased estimate of the gradient

• $\theta^{t+1} = \theta^t - \eta_t V^t$

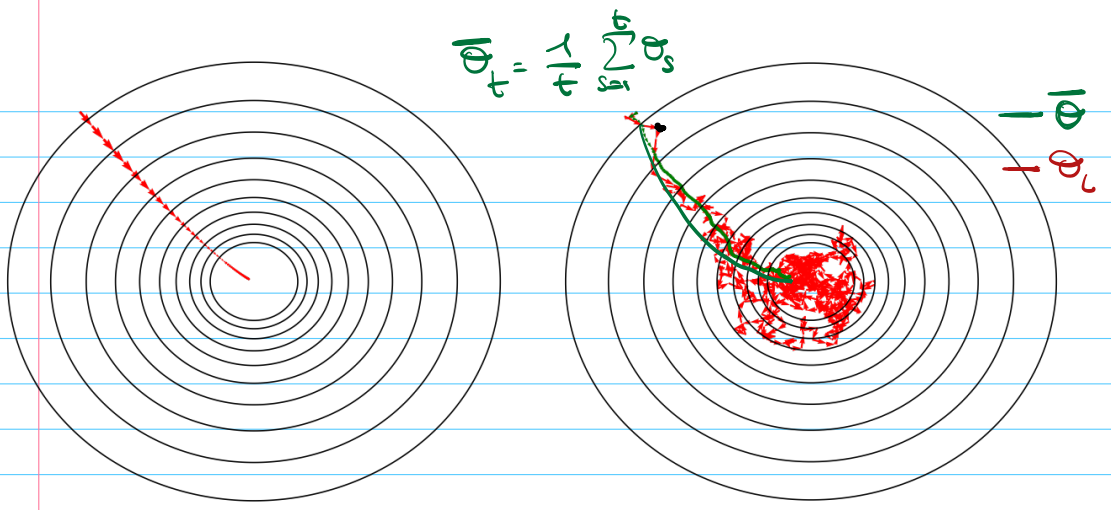
Output : average $\bar{\theta} = \frac{1}{T} \sum_{t=1}^{T} \theta_t$.

$a=b \rightsquigarrow \|\theta\|^2$

Toy Example : $F(\theta_1, \theta_2) = a\theta_1^2 + b\theta_2^2$ (r

$V^t = \nabla F(\theta_1^t, \theta_2^t) + \varepsilon$

↑ random noise



GRADIENT DESCENT

STOCHASTIC GRADIENT DESCENT

→ For risk minimization:

$$F(\theta) = R_P(f_\theta)$$

$T = n$ (number of samples)

$$V^i = \mathbb{E}\left[\nabla \ell(f_{\theta^i}(x_i), y_i)\right] = \nabla\, \mathbb{E}\left[\ell(f_{\theta^i}(x), y)\right]$$

$$\mathbb{E}[V^i \mid \theta^i] = \mathbb{E}[V^i] = \nabla F(\theta^i)$$

only depends on $(x_j, y_j)_{j<i}$

THEOREM :

$F : \mathbb{R}^d \to \mathbb{R}$ convex + differentiable

minimizer $\theta^* \in B(0; R)$ $\quad R = \|\theta^*\|$

$\mathbb{E}[\|v^t\|^2] \leq \rho^2$

Step size $\eta = \sqrt{\frac{4R^2}{\rho^2 T}}$.

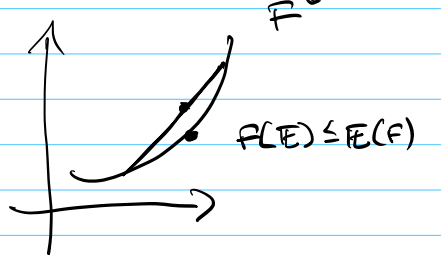
Initialization $\theta^0 \in B(0; R)$

$$\Rightarrow \mathbb{E}[F(\bar{\theta})] - F(\theta^*) \leq 2\frac{R\rho}{\sqrt{T}}$$

proof 1 : special case $v^t = \nabla F(\theta^t)$

proof 2 : General case

$$\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \theta_t$$



$$F(\bar{\theta}) - F(\theta^*) \leq \frac{1}{T}\sum_{t=1}^{T} F(\theta_t) - F(\theta^*)$$

$$F(\theta^*) \geq F(\theta^t) + \langle \nabla F(\theta^t), \theta^* - \theta^t \rangle$$

$$\langle \nabla F(\theta^t), \theta^t - \theta^* \rangle \geq F(\theta^t) - F(\theta^*)$$

(★) $$F(\bar{\theta}) - F(\theta^*) \leq \frac{1}{T}\sum_{t=1}^{T} \langle \nabla F(\theta^t), \theta^t - \theta^* \rangle$$

$$\mathbb{E}[v^t \mid \theta^t] = \nabla F(\theta^t)$$

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\langle v^t, \theta^* - \theta^t\rangle \mid \theta^t\right]$$

$$\mathbb{E}[F(\bar{\theta}) - F(\theta^*)] \le \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\langle v^t, \theta^* - \theta^t\rangle]$$

$$= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \langle v^t, \theta^* - \theta^t\rangle\right]$$

Claim: $\langle v^t, \theta^t - \theta^*\rangle = \frac{1}{2\delta}\left(\|\theta^t - \theta^*\|^2 - \|\theta^{t+1} - \theta^*\|^2\right) + \frac{\delta}{2}\|v^t\|^2$

$\theta^{t+1} = \theta^t - \delta v^t$

$\|\theta^t - \theta^* - \delta v^t\|^2$

$$\theta^{t+1} = \theta_t - \delta \nabla F(\theta^t) \qquad \leftarrow v^t$$

(*) **Claim:** $\langle \nabla F(\theta^t), \theta^* - \theta^t \rangle$

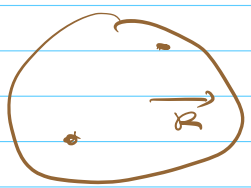
$$= \frac{1}{2\delta}\left(\|\theta^t - \theta^*\|^2 - \|\theta^{t+1} - \theta^*\|^2 \right) + \frac{\delta}{2} \|\nabla F(\theta^t)\|^2$$

$$\frac{1}{T}\sum_{t=1}^{T} \langle \nabla F(\theta^t), \theta^* - \theta^t \rangle = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{2\delta}\left(\|\theta^t - \theta^*\|^2 - \|\theta^{t+1} - \theta^*\|^2 \right)$$

$$+ \frac{1}{T}\frac{\delta}{2}\sum_{t=1}^{T} \|\nabla F(\theta^t)\|^2$$

$$\leq \frac{1}{T2\delta}\left(\underbrace{\|\theta^1 - \theta^*\|^2}_{\leq 4R^2} - \underbrace{\|\theta^T - \theta^*\|^2}_{\leq 0} \right) + \frac{\delta}{2}\rho^2$$

$\theta^*, \theta^1 \in B(0; R)$



$$\leq \frac{4R^2}{2T\delta} + \frac{\delta}{2}\rho^2.$$

$$\sum_{t=1}^{T} \left(\|\theta^t - \theta^*\|^2 - \|\theta^{t+1} - \theta^*\|^2 \right)$$

$$= \|\theta^1 - \theta^*\|^2 - \|\theta^2 - \theta^*\|^2$$
$$+ \|\theta^2 - \theta^*\|^2 - \|\theta^3 - \theta^*\|^2$$
$$+ \|\theta^3 - \dots - \|\theta^4$$
$$+ \dots$$
$$+ \|\theta^{T-1} - \dots\|^2 - \|\theta^T - \theta^*\|^2$$

Remark: In practice, a good idea is to discard the first iterates when computing $\bar{\theta}$: Forget about initialization

$$\bar{\theta} = \frac{1}{T - T_0} \sum_{t=T_0}^{T} \theta_t$$

Another Example: F_i → arbitrary functions

$$F(\theta) = \frac{1}{n} \sum_{i=1}^{n} F_i(\theta)$$

Let $I \sim \text{Unif}\{1 \dots n\}$ ($\frac{1}{n}$)

$$\Rightarrow F(\theta) = \mathbb{E}[F_I(\theta)] = \sum_{i=1}^{n} P(I=i) F_i(\theta)$$

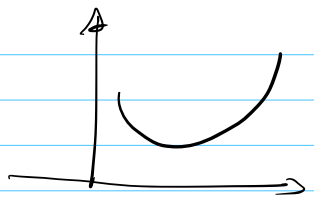
To get estimates of $\nabla F(\theta)$:

$$V^t = \nabla F_{I_t}(\theta) \qquad I_1 \dots I_T \text{ iid samples.}$$

$d=1$ β smooth: $F''(\theta) \le \beta$

α strongly cvx: $F''(\theta) \ge \alpha$

Comparison with GD:



Recall:

(1) F β-smooth + α-strongly convex

T steps of GD ; $s = 1/\beta$

$$F(\theta^T) - F(\theta^*) \le \exp\left(-\frac{\alpha}{\beta} T\right)$$

$T = O(\log(1/\varepsilon))$ steps to reach precision ε.

(2) F β-smooth + cvx

T steps of GD ; $s = 1/\beta$

$$F(\theta^T) - F(\theta^*) \le \frac{\beta}{T}$$

$T = O(1/\varepsilon)$ steps to reach precision ε.

(3) F cvx + diff

T steps of SGD

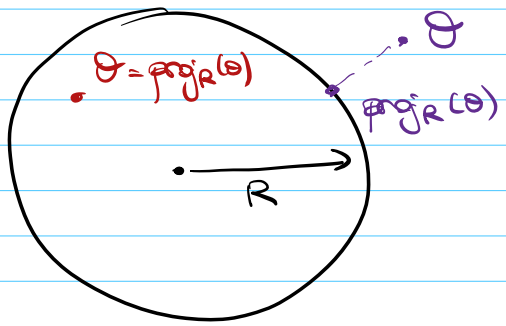
$$E[F(\bar{\theta})] - F(\theta^*) \le \frac{1}{\sqrt{T}}$$

Question : Can we improve the rate for SGD with more assumptions?

→ β - smoothness ? ⇝ no

→ α - strongly cvx ? ⇝ no yes ! 

$$\text{proj}_R(\theta) = \begin{cases} \theta & \text{if } \|\theta\| \le R \\ \frac{R}{\|\theta\|}\theta & \text{else.} \end{cases}$$



SGD with projection :

For $t = 1 \dots T-1$:

- V^t random vector such that

$$\mathbb{E}[V^t | \theta^t] = \nabla F(\theta^t)$$

- $\theta^{t+1} = \text{proj}_R(\theta^t - \delta_t V^t)$

Output : average $\bar{\theta} = \frac{1}{T} \sum_{t=1}^{T} \theta_t$.

THEOREM :

$F : \mathbb{R}^d \to \mathbb{R}$ α-strongly convex

minimizer $\theta^* \in B(0; R)$

$\mathbb{E}[\|V^t\|^2] \leq \rho^2$

Step size $\delta_t = 1/\alpha t$

Initialization $\theta^0 \in B(0; R)$

$$\Rightarrow \mathbb{E}[F(\bar{\theta})] - F(\theta^*) \leq \frac{\rho^2}{2\alpha T}(1 + \log T)$$

proof : $F(\bar{\theta}) - F(\theta^*) \le \frac{1}{T} \sum \left(F(\theta^t) - F(\theta^*)\right)$

$$\langle v^t, \theta^t - \theta^* \rangle = \frac{1}{2\delta_t}\left(\|\theta^t - \theta^*\|^2 - \|\theta^{t+1} - \theta^*\|^2\right) + \frac{\delta_t}{2}\|v^t\|^2$$

$$\mathbb{E}\left[F(\theta^t) - F(\theta^*)\right] \le \mathbb{E}\left[\langle \nabla F(\theta^t), \theta^t - \theta^* \rangle - \frac{\alpha}{2}\|\theta^t - \theta^*\|^2\right]$$

$\mathbb{E}[v^t \mid \theta^t] = \nabla F(\theta^t)$

$$\mathbb{E}\,\frac{1}{T}\sum\left(F(\theta^t) - F(\theta^*)\right) \le \mathbb{E}\left(\frac{1}{T}\sum_t \left(\langle \nabla F(\theta^t), \theta^t - \theta^* \rangle - \frac{\alpha}{2}\|\theta^t - \theta^*\|^2\right)\right)$$

$$= \mathbb{E}\left[\frac{1}{T}\sum_t \left(\langle v^t, \theta^t - \theta^* \rangle - \frac{\alpha}{2}\|\theta^t - \theta^*\|^2\right)\right]$$

$$\frac{\alpha_t}{2}\left(\|\theta^t - \theta^*\|^2 - \|\theta^{t+1} - \theta^*\|^2\right) + \frac{1}{2\alpha_t}\|v^t\|^2 \qquad \delta_t = \frac{1}{\alpha_t}$$

$$\mathbb{E}\left[\frac{1}{T}\sum_t \left(\frac{\alpha_t}{2}\|\theta^t - \theta^*\|^2 - \frac{\alpha_t}{2}\|\theta^{t+1} - \theta^*\|^2 + \frac{1}{2\alpha_t}\|v^t\|^2 - \frac{\alpha}{2}\|\theta^t - \theta^*\|^2\right)\right]$$

$$\frac{\alpha(t-1)}{2}\|\theta^t - \theta^*\|^2 = u_t$$

$$\frac{1}{T}\sum (u_t - u_{t+1}) = \frac{1}{T}(u_1 - u_T) = \frac{1}{T}(0 - u_T) \le 0$$

$$\leq \frac{1}{T}\mathbb{E}\left[\sum_t \frac{1}{2\alpha t}\|v_t\|^2\right]$$

$$\leq \frac{1}{T}\sum_t \frac{\rho^2}{2\alpha t} \leq \frac{\rho^2}{2\alpha T}\underbrace{\left(\sum_{t=1}^{T}\frac{1}{t}\right)}_{\approx \log T}$$

$$\int_1^T \frac{1}{t}\,dt = \log T$$

$\blacksquare$

② Application to Risk Minimization

Back to risk minimization

$$F(\theta) = R_P(f_\theta) \qquad \theta \in \mathbb{R}^k$$

minimizer θ^*

Sample $(X_1, Y_1) \dots (X_n, Y_n)$

Two methods to approximate θ^*:

(SGD) $\hat{\theta}_{SGD}$: n steps of SGD with gradients $\nabla_\theta \ell(f_\theta(X_i), Y_i)$

(GD-ER) $\hat{\theta}_T$: Gradient Descent for T steps applied to

$$R_n(f_\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i)$$

$$\mathbb{E}[R_P(f_{\hat{\theta}}) - R_P^*] = \underbrace{\mathbb{E}[R_P(f_{\hat{\theta}}) - R_P(f_{\theta^*})]}_{\text{optimization error}} + \underbrace{R_P(f_{\theta^*}) - R_P^*}_{\text{approximation error}}$$

Two Questions :

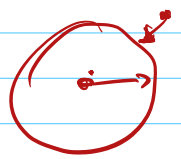
1. What is the minimal number n of samples required to get an optimization error smaller than ε ?
2. What is the associated time complexity ?

$\Rightarrow$ We answer (1) and (2) in the setting where : $\forall x \in \mathcal{X}, y \in \mathcal{Y}$

$\theta \mapsto \ell(f_\theta(x), y)$ is $\begin{cases} \alpha\text{- strongly convex} \\ \beta\text{- smooth} \end{cases}$

$|\langle x, \theta \rangle - y|^2$

(SGD) $V^i = \nabla \ell(f_{\theta^i}(x_i), y_i)$



$\Rightarrow \|V^i\| \leq \beta \|\theta^i - \theta^-\| \leq 2\beta R$

$\Rightarrow \mathbb{E}[R_\rho(f_{\theta_{SGD}}) - R_\rho(f_{\theta^*})] \lesssim \frac{1}{n}$

(1) $n = \tilde{O}(1/\varepsilon)$ → up to log factors

(2) Time complexity : $\tilde{O}(kn)$

$\theta_\infty = \operatorname{argmin}_\theta R_P(f_\theta)$ ←

(GD-ER) Let $\hat{\theta}_\infty = \operatorname{argmin}_\theta R_n(f_\theta)$.

$$R_P(f_{\theta_T}) - R_P(f_{\theta_\infty}) \le R_P(f_{\theta_T}) - R_n(f_{\hat{\theta}_T}) + R_n(f_{\theta_T}) - R_n(f_{\hat{\theta}_\infty}) + \boxed{R_n(f_{\hat{\theta}_\infty}) - R_n(f_{\theta_\infty})} \le 0 - R_P(f_{\theta_\infty}) + R_n(f_{\theta_\infty})$$

$R_n(f_{\hat{\theta}_\infty}) \le R_n(f_\theta) \; \forall \theta$

$$\le 2 \cdot \sup_\theta |R_n(f_\theta) - R_P(f_\theta)| + R_n(f_{\theta_T}) - R_n(f_{\theta_\infty})$$

estimation error $\lesssim \frac{1}{\sqrt{n}}$ $\qquad$ $\lesssim \exp(-\frac{\alpha}{\beta} T)$

To get

$$R_P(f_{\theta_T}) - R_P(f_{\theta_\infty}) \le \varepsilon$$

1. $n = O(1/\varepsilon^2)$ samples
2. $T = O(\log(\varepsilon^{-1}))$

Complexity $O(kTn) = \tilde{O}\left(\frac{k}{\varepsilon^2}\right)$

SUMMARY: in the { α-strongly convex, β-smooth } setting

Algo	Num. of samples	Complexity
SGD	$n = \tilde{O}(1/\varepsilon)$	$\tilde{O}(k/\varepsilon)$
GD	$n = O(1/\varepsilon^2)$	$\tilde{O}(k/\varepsilon^2)$

To go further...

Variance Reduction technique:

SVRG / SAGA

↳ see lecture notes for references.