Optimization:

$\boxed{\begin{array}{l} F:\mathbb{R}^d \to \mathbb{R} \\ \theta^* \text{ a minimizer} \end{array}}$

$\neq \theta^*$

① Classical GD on a function $F:\mathbb{R}^d \to \mathbb{R}$.

example: $F(\theta) = \hat{R}(f_\theta)$

- smooth problems $\to$ quadratics
- non smooth problems $\Rightarrow \frac{1}{\sqrt{t}}$

$t = \#$ iterations

② SGD $\to$ finite sums (train)

$\hookrightarrow$ expectation (test)

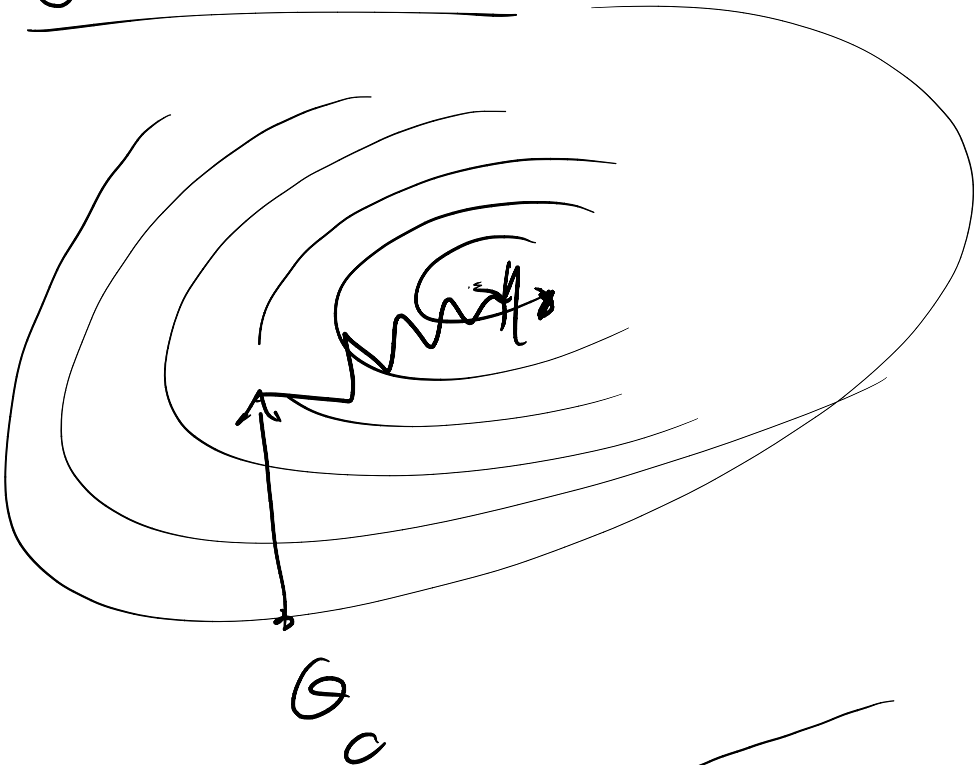
Gradient descent:

$$\theta_t = \theta_{t-1} - \gamma_t F'(\theta_{t-1})$$

step-size > 0

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

- line search
- constant $\gamma_t = \gamma$



Notion of condition number

least - squares : $F(\theta) = \frac{1}{2n} \|y - \Phi\theta\|_2^2$

Hessian : $H = \frac{1}{n}\Phi^\top \Phi \in \mathbb{R}^{d \times d}$

$$F'(\theta) = 0 \iff \frac{1}{n}\Phi^\top(\Phi\theta - y) = 0$$

$$H\eta_* = \frac{1}{n}\Phi^\top y \qquad \eta_* \text{ a minimizer of } F$$

$$\theta_t = \theta_{t-1} - \gamma F'(\theta_{t-1}) = \theta_{t-1} - \gamma \frac{1}{n}\Phi^\top(\Phi\theta_{t-1} - y)$$

$$= \theta_{t-1} - \gamma\Big[H\theta_{t-1} - \underbrace{\frac{1}{n}\Phi^\top y}_{H\eta_*}\Big]$$

$$\theta_t - \eta_* = \theta_{t-1} - \eta_* - \gamma H(\theta_{t-1} - \eta_*)$$

$$= (I - \gamma H)(\theta_{t-1} - \eta_*)$$

$$\boxed{\theta_t - \eta_* = (I - \gamma H)^t(\theta_0 - \eta_*)}$$

Criteria: $F(\theta_t) - F(q_*) = \frac{1}{2}(\theta_t - q_*)^T \underline{H} (\theta_t - q_*)$

a_t //

$$\theta_t - q_* = (I - \gamma H)^t (\theta_0 - q_*)$$

$$a_t = \frac{1}{2}(\theta_0 - q_*)^T H (I - \gamma H)^{2t} (\theta_0 - q_*)$$

H symmetric eigenvalues $[\mu, L]$ $\quad \mu > 0$, $L > \mu$

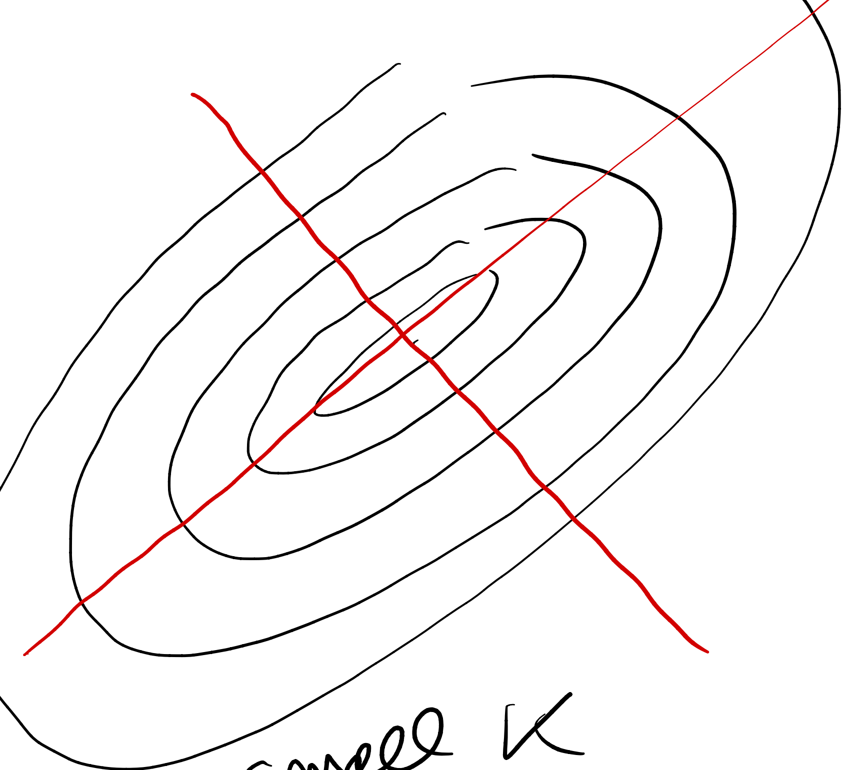
eigenvalues of $(I - \gamma H)^{2t}$ are $(1 - \gamma\lambda)^{2t}$ for $\lambda \in$ spectrum(H)

when is $|1 - \gamma\lambda| \leq 1$ $\quad \forall \lambda \in (\mu, L)$?

if $\gamma = \frac{1}{L}$ then $|1 - \gamma\lambda|^{2t} \in \left[0, \left(1 - \frac{\mu}{L}\right)^{2t}\right]$

$$a_t \leq \underbrace{\frac{1}{2}(\theta_0 - q_*)^T H (\theta_0 - q_*)}_{F(\theta_0) - F(q_*)} \left(1 - \frac{\mu}{L}\right)^{2t}$$

- exponential
- linear cvce



$$F(\theta_t) - F(\theta_*) \le \left(1-\frac{\mu}{L}\right)^{2t}\left(F(\theta_0) - F(\theta_*)\right)$$

$$\left(1-\frac{1}{K}\right)^{2t}$$

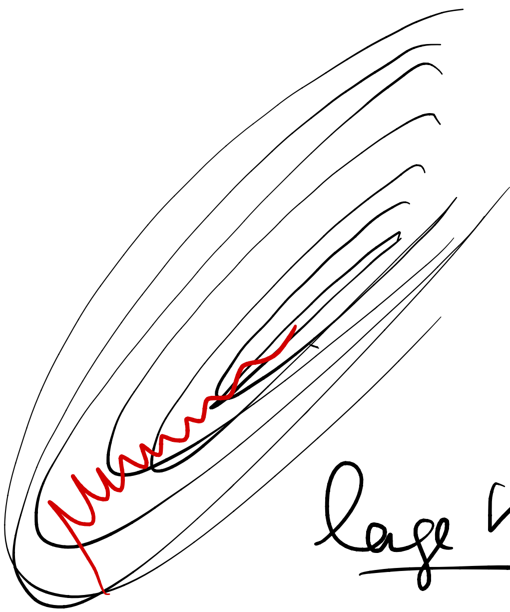
$\frac{L}{\mu} = K \geq 1$ condition #

$$\le e^{-\frac{\mu}{L}2t}$$

$$\varepsilon = e^{-2t/K}$$

(using $1-\alpha \le e^{-\alpha}$)

$$t = \frac{K}{2}\log\frac{1}{\varepsilon}$$



⚠ in practice, μ very small unless regularization is added

$$\boxed{\lambda_n \sim \frac{1}{n} \quad \frac{1}{\sqrt{n}}}$$

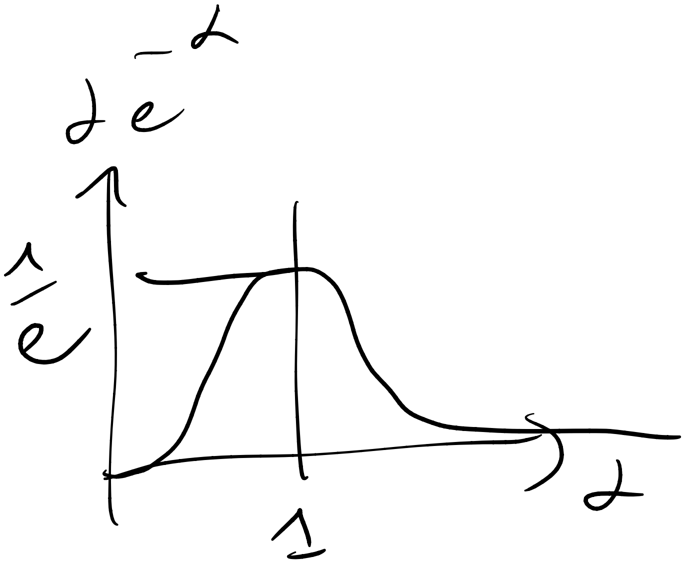
What if $\mu = 0$? $\frac{1}{2}(\theta_0 - \eta_*)^T \underbrace{H(1-\gamma H)^{2t}}(\theta_0 - \eta_*)$

$$\lambda(1-\gamma\lambda)^{2t} \le \lambda e^{-\gamma\lambda 2t}$$ because $1-\gamma\lambda \in (0,1)$

$$= \frac{\gamma\lambda 2t\, e^{-\gamma\lambda 2t}}{2\gamma t} \le \frac{\sup_{\alpha \ge 0} \alpha e^{-\alpha}}{2\gamma t}$$

$$\le \frac{1}{2e}\frac{1}{\gamma t}$$

$$\le \frac{1}{4\gamma t} = \frac{L}{4t}$$



$$\boxed{a_t \le \frac{L}{8t}\|\theta_0 - \eta_*\|_2^2}$$

Improvements: Conjugate gradient, Nesterov, $\kappa \to \sqrt{\kappa}$ $\frac{1}{t} \to \frac{1}{t^2}$

Summary $F(\theta_t) - F(\eta_\theta) \le \left(F(\theta_0) - F(\eta_\theta)\right)\left(1 - \frac{1}{\kappa}\right)^{2t}$

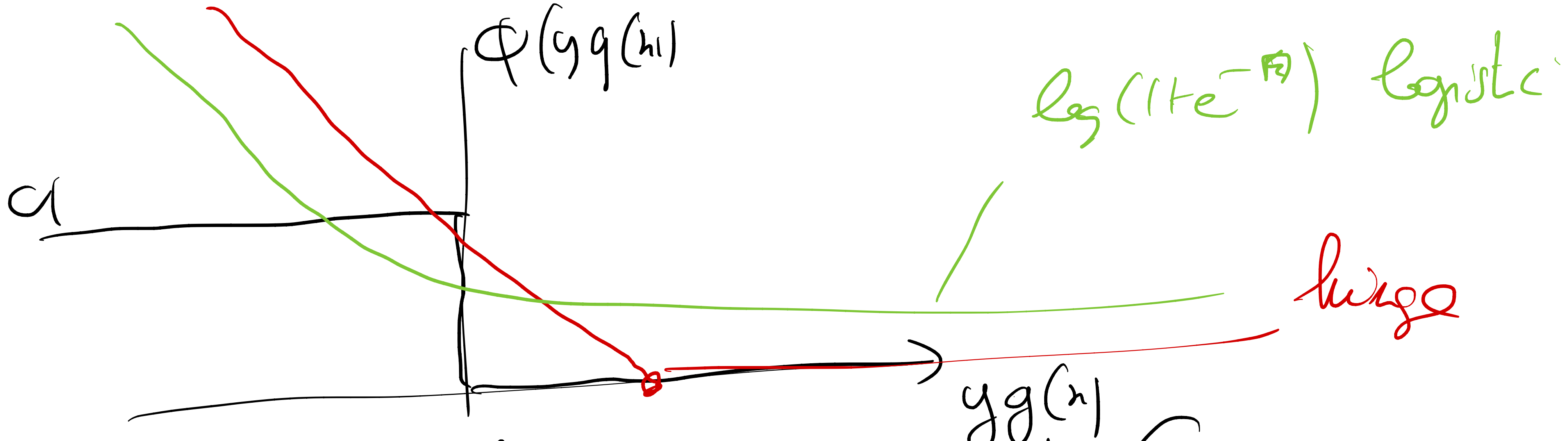
$$\le \frac{L}{8t}\|\theta_0 - \eta_\theta\|_2^2$$

For ridge regression : $F(\theta) = \frac{1}{2n}\|y - \Phi\theta\|_2^2 + \frac{\lambda}{2}\|\theta\|_2^2$

$$\kappa = \frac{\lambda_{max}(\Sigma) + \lambda}{\underbrace{\lambda_{min}(\Sigma)}_{=0} + \lambda} \simeq \frac{\lambda_{max}(\Sigma)}{\lambda} + 1$$

Same rates (up to constant) for smooth strongly convex.

if $\forall \theta \in \mathbb{R}^d$ $F''(\theta)$ has eigenvalues in $[\mu, L]$.



ϕ is Lipschitz-continuous with constant G

$$F(\theta) = \frac{1}{n}\sum_i \phi(y_i\theta^T\varphi(x_i))$$

$$F'(\theta) = \frac{1}{n}\sum_i \phi'(y_i\theta^T\varphi(x_i))\, y_i\, \underline{\varphi(x_i)}$$

$\|\cdot\| \le R$ $\quad$ B

$$\|F'(\theta)\|_2 \le \frac{1}{n}\sum_{i=1}^{n} GR = GR$$

if $\|\varphi(x)\| \le R$ a.s

Goal: $\min_{\theta \in \mathbb{R}^d} F(\theta)$ with F convex

$\|F'(\theta)\|_2 \leq B \quad \forall \theta \in \mathbb{R}^d$

Alg: $\theta_t = \theta_{t-1} - \gamma F'(\theta_{t-1})$

Proof technique: find a Lyapunov function $V : \mathbb{R}^d \to \mathbb{R}_+$

- $V(\theta_t) \leq (1-\alpha) V(\theta_{t-1})$
  $\leq (1-\alpha)^t V(\theta_0)$

- $V(\theta_t) \leq V(\theta_{t-1}) - W(\theta_{t-1})$
  $\Downarrow$
  $W(\theta_{t-1}) \leq V(\theta_{t-1}) - V(\theta_t)$

telescoping sum

average of $W(\theta_t) \leq \dfrac{V(\theta_0)}{t}$

$F(\theta)$ with F convex

d $\|F'(\theta)\|_2 \leq B \quad \forall \theta \in \mathbb{R}^d$

$$\theta_t = \theta_{t-1} - \gamma F'(\theta_{t-1})$$

$\|\theta_{t-1} - \eta_* - \gamma F'(\theta_{t-1})\|_2^2$

$\|\theta_{t-1} - \eta_*\|_2^2 - 2\gamma F'(\theta_{t-1})^T(\theta_{t-1} - \eta_*) + \gamma^2 \underbrace{\|F'(\theta_{t-1})\|_2^2}_{\leq B^2}$

$\leq \|\theta_{t-1} - \eta_*\|_2^2 - 2\gamma\left[F(\theta_{t-1}) - F(\eta_*)\right] + \gamma^2 B^2$

$F(\theta_{t-1}) - F(\eta_*) \leq \frac{1}{2\gamma}\left[\|\theta_{t-1} - \eta_*\|_2^2 - \|\theta_t - \eta_*\|_2^2\right] + \frac{\gamma B^2}{2}$

$(\eta_* - \theta)$

$+$

$(\theta - \eta_*)$

$\frac{1}{T}\sum_{t=1}^{T} F(\theta_{t-1}) - F(\eta_*) \leq \frac{1}{2\gamma T}\|\theta_0 - \eta_*\|_2^2 + \frac{\gamma B^2}{2}$

Jensen

$F\left(\frac{1}{T}\sum_{t=1}^{T}\theta_{t-1}\right) - F(\eta_*) \leq \frac{1}{2\gamma T}\|\theta_0 - \eta_*\|_2^2 + \frac{\gamma B^2}{2}$

$$F\left(\frac{1}{T}\sum_{t=1}^{T}\theta_{t-1}\right)-F(\eta_*) \le \underbrace{\frac{1}{2\gamma T}\|\theta_0-\eta_*\|_2^2+\frac{\gamma B^2}{2}}_{\text{inf} \Rightarrow \gamma = \frac{\|\theta_0-\eta_*\|_2}{B\sqrt{T}}} \le \frac{B\|\theta_0-\eta_*\|_2}{\sqrt{T}}$$

inf γ $\Rightarrow \gamma = \frac{\|\theta_0-\eta_*\|_2}{B\sqrt{T}}$ if horizon T is known in advance

⚠ Not practical

See back for $\gamma_t \propto \frac{1}{\sqrt{t}} \Rightarrow$ bound $\frac{\log t}{\sqrt{t}}$

"anytime" algorithm

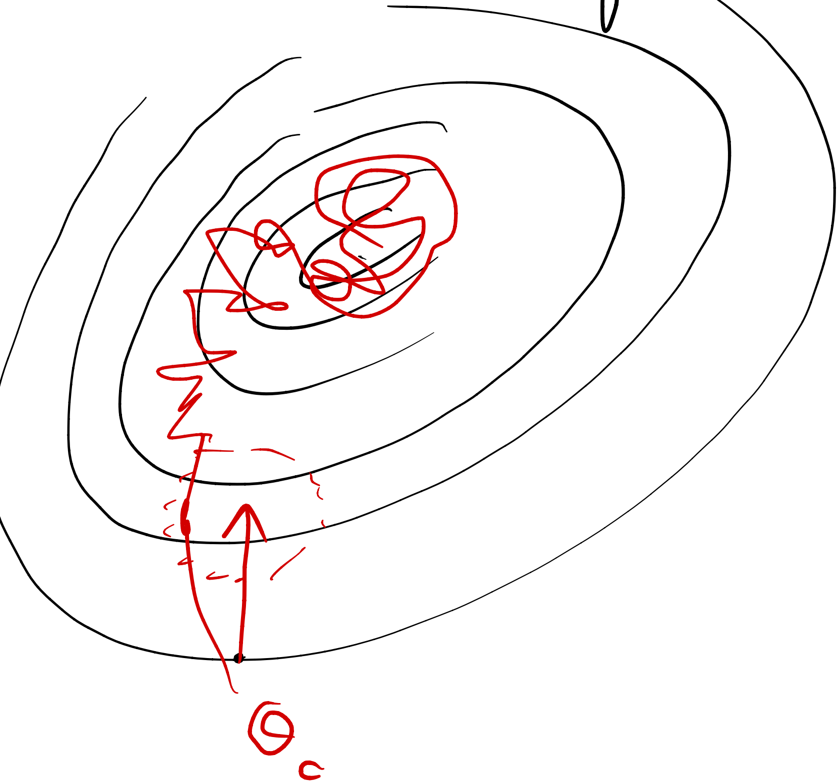
Summary for ML with linear models ① GD on $\frac{GRD}{\sqrt{t}}$ ← # of iterations

② uniform deviations: $\frac{GRD}{\sqrt{n}}$ ← # of observations

$\Rightarrow$ $t = n$ is enough

$O(n^2)$ accesses to gradients

Stochastic gradient "descent" (1951 Robbins Monro).



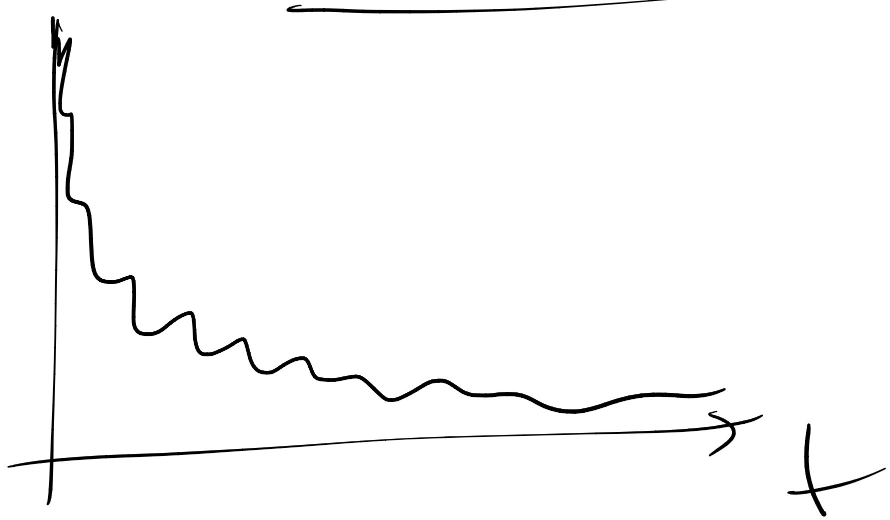
$$\theta_t = \theta_{t-1} - \gamma g_t$$

$$E[g_t \mid \theta_{t-1}] = F'(\theta_{t-1})$$

"unbiased"

variance does not go to zero

Related work: online learning. Ch 11



Machine learning set up

$$\theta_t = \theta_{t-1} - \gamma g_t$$

(1) Empirical risk minimization.

$$F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h_\theta(x_i))$$

sto. grad: $g_t = \left.\frac{\partial \ell(y_{i(t)}, h_\theta(x_{i(t)}))}{\partial \theta}\right|_{\theta = \theta_{t-1}}$

with $i(t) \in \{1, \dots, n\}$ selected un. family at random

then $\mathbb{E}(g_t \mid \theta_{t-1}) = F(\theta_{t-1})$

$$\mathbb{E}\left[F(\theta_T) - F(\eta_*)\right] = \frac{D}{\sqrt{T}}$$

over choice of $i(t)$, $t = 1, \dots, T$)

$F(\theta) = \hat{R}(h_\theta)$

$\eta_* = \text{ERM}$

② Minimization of the testing error

$$F(\theta) = \mathbb{E}\left[\ell(y, f_\theta(x))\right], \quad \theta_t = \theta_{t-1} - \gamma g_t$$

$$g_t = \left.\frac{\partial \ell(y_t, f_\theta(x_t))}{\partial \theta}\right|_{\theta=\theta_{t-1}}$$

$t = 1, \ldots, n. \quad \Rightarrow$ single pass SGD

$$\mathbb{E}[g_t \mid \theta_{t-1}] = \mathbb{E}\left[\left.\frac{\partial}{\partial \theta} \ell(y_t, f_\theta(x_t))\right|_{\theta=\theta_{t-1}} \,\middle|\, \theta_{t-1}\right]$$

$$= \frac{\partial}{\partial \theta} \underbrace{\mathbb{E}\left[\left.\ell(y_t, f_\theta(x_t))\right|_{\theta=\theta_{t-1}} \,\middle|\, \theta_{t-1}\right]}_{= F'(\theta_{t-1})}$$

wrt data

because (x_t, y_t) are iiD

Goal: $\min_{\theta \in \mathbb{R}^d} F(\theta)$ with F convex $\|g_t\|_2 \leq B$

Alg: $\theta_t = \theta_{t-1} - \gamma g_t$ $\qquad \mathbb{E}(g_t | \theta_{t-1}) =$

$$\|\theta_t - \eta_*\|_2^2 = \|\theta_{t-1} - \eta_* - \gamma g_t\|_2^2$$

$$= \|\theta_{t-1} - \eta_*\|_2^2 - 2\gamma g_t^T(\theta_{t-1} - \eta_*) + \gamma^2 \overbrace{\|g_t\|_2^2}^{\leq B^2}$$

$$\mathbb{E}\left[\|\theta_t - \eta_*\|^2 | \theta_{t-1}\right] = \|\theta_{t-1} - \eta_*\|_2^2 - 2\gamma F'(\theta_{t-1})^T(\theta_{t-1} - \eta_*) + \gamma^2 B^2$$

$$\leq \|\theta_{t-1} - \eta_*\|_2^2 - 2\gamma\left[F(\theta_{t-1}) - F(\eta_*)\right] + \gamma^2 B^2$$

$$\mathbb{E}\left[F(\theta_{t-1})\right] - F(\eta_*) \leq \frac{1}{2\gamma}\left[\mathbb{E}\left(\|\theta_{t-1} - \eta_*\|_2^2 \cancel{|\theta_{t-1}}\right) - \mathbb{E}\|\theta_t - \eta_*\|_2^2\right] + \frac{\gamma B^2}{2}$$

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E} F(\theta_{t-1}) - F(\eta_*) \leq \frac{1}{2\gamma T}\|\theta_0 - \eta_*\|_2^2 + \frac{\gamma B^2}{2}$$

Jensen

$$\mathbb{E} F\left(\frac{1}{T}\sum_{t=1}^{T} \theta_{t-1}\right) - F(\eta_*) \leq \frac{1}{2\gamma T}\|\theta_0 - \eta_*\|_2^2 + \frac{\gamma B^2}{2}$$

Result: $\mathbb{E}\, F\left(\frac{1}{T}\sum_{t=1}^{T}\theta_{t-1}\right) - F(\theta_*) \leq \frac{BD}{\sqrt{T}}$

$D \geq \|\theta_0 - \theta_*\|_2^2$

① $F(\theta) = \hat{R}(f_\theta)$ TRAIN

$\mathbb{E}$ wrt to the choice of points

$(i(t))$

② $F(\theta) = R(f_\theta)$ TEST

$\mathbb{E}$ wrt to the data. $T = n$

$$\boxed{\mathbb{E}\, R(f_{\text{averaged iterate}}) \leq \frac{D}{\sqrt{n}}}$$

Generalization result

Q: what about strong convexity

$F(\theta) + \frac{\mu}{2}\|\theta\|_2^2 \longrightarrow \frac{B^2}{\mu n}$ instead of $\frac{1}{\sqrt{n}}$

Q: can we have exponential rates?

Finite sums: $F(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$

$f_i(\theta) = \ell(y_i, \theta^T \varphi(x_i))$

μ-strongly convex

variance reduction $\Rightarrow$ exp rates

$$\theta_t = \theta_{t-1} - \gamma \left[f'_{i(t)}(\theta_{t-1}) + z_{i(t)}^{t-1} \right]$$

SAGA $\quad \exp\left(-\min\left(\frac{1}{n}; \frac{\mu}{R^2}\right) t\right)$

Tomorrow :

1. kernel methods

$f_\theta(x) = \theta^T \varphi(x)$

↳ inf. dim.

2. neural nets.