Optimization: $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$

loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$

Risk $\begin{cases} \text{expected} : \mathbb{E}\,\ell(y, f_\theta(x)) = R(f_\theta) \\ \text{empirical} : \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) \end{cases}$

model $f_\theta, \theta \in \Theta$

$\hat{\theta}$ is an estimator.

$$R(f_{\hat{\theta}}) - \inf_{\theta \in \Theta} R(f_\theta) = R(f_{\hat{\theta}}) - \hat{R}(f_{\hat{\theta}}) + \hat{R}(f_{\hat{\theta}}) - \hat{R}(f_{\theta_*}) + \hat{R}(f_{\theta_*}) - R(f_{\theta_*})$$

$R(f_{\theta_*})$: estimation error

$\leq \sup_{\theta \in \Theta} |R(f_\theta) - \hat{R}(f_\theta)|$: uniform deviation

$\hat{\theta} \in \operatorname{argmin}_{\theta \in \Theta} \hat{R}(f_\theta)$

todeg: $\leq \hat{R}(f_{\hat{\theta}}) - \inf_{\theta \in \Theta} \hat{R}(f_\theta)$

$\leq \varepsilon$

optimization error

optimisation : $F : \mathbb{R}^d \to \mathbb{R}$ → $F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$

→ $F(\theta) = \mathbb{E}[\ell(y, f_\theta(x))]$

$\theta_* \in \operatorname{argmin} F$

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

$\Phi \in \mathbb{R}^{n \times d}$: design matrix / data

$y \in \mathbb{R}^n$ response vector

argmin θ_*

$F(\theta) = \frac{1}{n}\Phi^T(\Phi\theta - y) \in \mathbb{R}^d$

"normal equations"

θ_* characterized by $F(\theta_*) = 0 \iff \frac{1}{n}\Phi^T\Phi\theta_* = \frac{\Phi^T y}{n}$

$H = \hat{\Sigma}$ covariance matrix

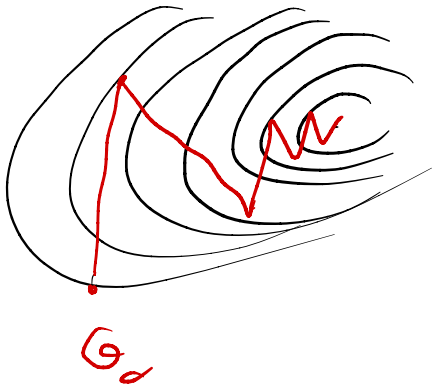
Gradient descent

$\theta_0 \in \mathbb{R}^d$

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

γ_t: step size / learning rate

- line search (with prob)
- constant step size $\boxed{\gamma_t = \gamma}$



Least-squares:

$$\theta_t = \theta_{t-1} - \gamma F'(\theta_{t-1}) = \theta_{t-1} - \gamma \left[\underbrace{\Phi^T \Phi \theta_{t-1}}_{n} - \underbrace{\Phi^T y}_{n} \right]$$

$H \theta_{t-1} - H \eta_*$

$$\theta_t - \eta_* = \theta_{t-1} - \eta_* - \gamma H \left[\theta_{t-1} - \eta_* \right]$$

$$= \left[I - \gamma H \right] \left(\theta_{t-1} - \eta_* \right) = (I - \gamma H)^t (\theta_0 - \eta_*)$$

Eigenvalue decomposition $H = \sum_{i=1}^{d} \lambda_i e_i e_i^T$ $\quad e_i^T e_j = 1_{i=j}$

Eigenvalues of $(I - \gamma H)^t \to (1 - \gamma \lambda_i)^t$ $\quad \in [\mu, L]$

where $\mu = \lambda_{min}(H) \geq 0$, $L = \lambda_{max}(H)$

$\boxed{\gamma = 1/L}$

$$\theta_t - \eta_\infty = (I - \gamma H)^t (\theta_0 - \eta_\infty)$$

$\gamma = \frac{1}{L}$, eigenvalues of $(I - \gamma H) \in \left(0, 1 - \frac{\mu}{L}\right)$

$\lambda_i \in (\mu, L) \Rightarrow 1 - \gamma\lambda_i \in (1 - \gamma L, 1 - \gamma\mu) = \left[0, 1 - \frac{\mu}{L}\right]$.

Convex: $F(\theta) - F(\eta_\infty)$ always possible

~~$\|\theta - \eta_\infty\|_2^2$~~ reserved for strongly cvx pb $\mu > 0$

$$F(\theta) - F(\eta_\infty) = F'(\eta_\infty)^T(\theta - \eta_\infty) + \frac{1}{2}(\theta - \eta_\infty)^T H (\theta - \eta_\infty)$$

$F'(\eta_\infty) = 0$ exact 2nd order Taylor expansion

$$F(\theta_t) - F(\eta_\infty) = \frac{1}{2}(\theta_0 - \eta_\infty)^T H (I - \gamma H)^{2t} (\theta_0 - \eta_\infty)$$

$$= \frac{1}{2}\left[H^{1/2}(\theta_0 - \eta_\infty)\right]^T (I - \gamma H)^{2t} \left[H^{1/2}(\theta_0 - \eta_\infty)\right)$$

$$\leq \frac{1}{2}\|H^{1/2}(\theta_0 - \eta_\infty)\|^2 \cdot \left(1 - \frac{\mu}{L}\right)^{2t}$$

$F(\theta_0) - F(\eta_\infty)$ $\in (0, 1)$

Exponential convergence of $\mu > 0$ $\kappa = \lambda_{\min}(\Phi^T\Phi/n)$ $d \leq n$

Lemma:
if $\mathrm{eig}(M) \in (0, \alpha)$
$\beta^T M \beta \leq \alpha \|\beta\|_2^2$

$$F(Q_t) - F(\eta_*) \le \left(1 - \frac{1}{L}\right)^{2t} \left(F(Q_0) - F(\eta_*)\right)$$

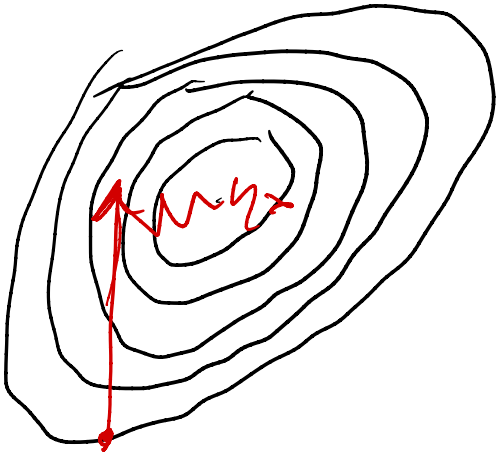
$$\le e^{-2t/L} \left(\quad \right)$$

$$e^{-2t/\kappa} = \varepsilon \Leftrightarrow \boxed{t = \frac{\kappa}{2} \log 1/\varepsilon}$$

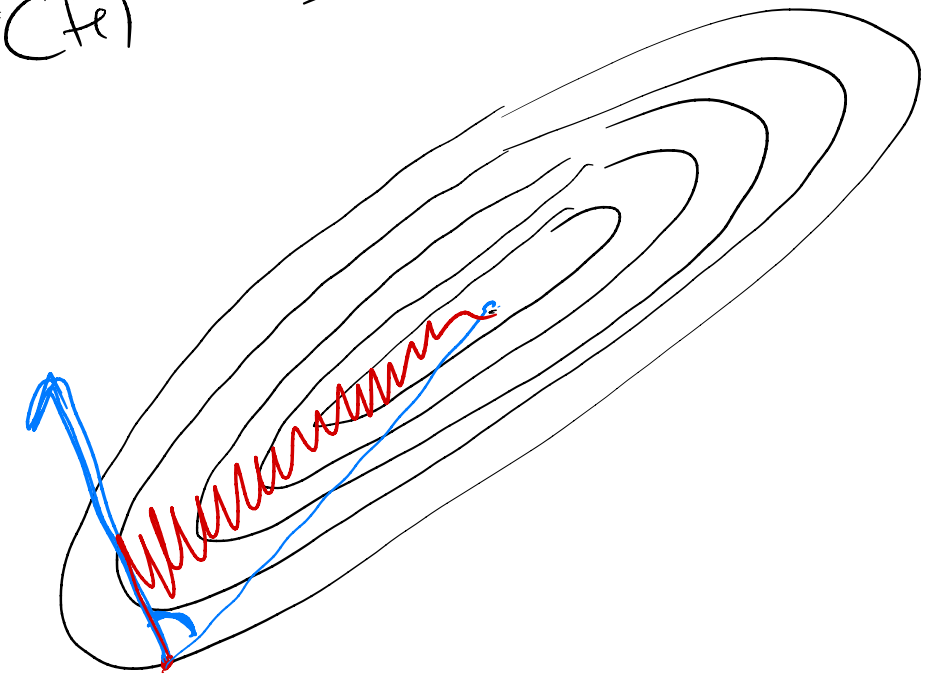
with $\kappa = \frac{L}{\mu}$ condition number

$$= \frac{\lambda_{max}(H)}{\lambda_{min}(H)} \geq 1$$

Lemma
$1 - \alpha \le e^{-\alpha}$
$\alpha > 0$ of eq



κ small



κ large

$\gamma = 0$: $F(\theta_t) - F(\eta_t) = \frac{1}{2}(\theta_t - \eta_t) H (1 - \delta t)^{2t} (\theta_t - \eta_t)$

eigenvalues $\lambda_i (1 - \frac{\lambda_i}{L})^{2t} \leq \lambda_i e^{-\frac{2t\lambda_i}{L}}$



$= \frac{2t\lambda_i}{L} e^{-2t\lambda_i/L} \times \frac{L}{2t}$

$\leq \left(\sup_{\alpha \geq 0} \alpha e^{-\alpha} \right) \frac{L}{2t}$

$\leq \frac{L}{4t}$

$\gamma = \frac{1}{L}$

$$F(\theta_t) - F(\eta_t) \leq \frac{L}{8t} \|\theta_0 - \eta_0\|^2$$

$$\leq \left(1 - \frac{1}{L}\right)^{2t} \left[F(\theta_0) - F(\eta_0) \right]$$

⚠ check homogeneity

⚠ Adaptivity

⚠ "unimportable"



⚠ band

Blog post

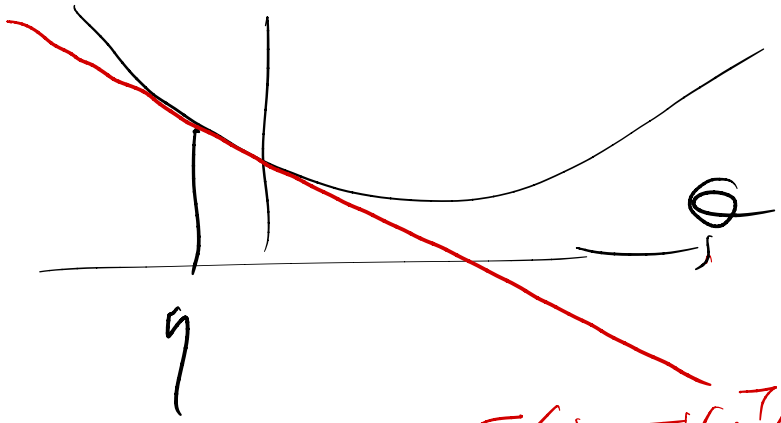
Convex functions : $F : \mathbb{R}^d \to \mathbb{R}$ convex — Boyd & vandenberghe

Assume F is differentiable

F is cvx $\Leftrightarrow$ $\forall y \in \mathbb{R}^d$, $\forall x$ $F(x) \geq F(y) + F'(y)^T(x-y)$

if F is twice differentiable

$\Leftrightarrow$ $F''(x)$ is PSD $\forall x \in \mathbb{R}^d$



Corollary 1 : y_* minimizer $\Leftrightarrow F'(y_*) = 0$

proof = using tangent at y_*.

Corollary 2 : $\forall y \in \mathbb{R}^d$ $F(y) - F(y_*) \leq F'(y)^T(y - y_*)$

y_* minimizer

$C^0 C \to C^0 3C$

in ML, which fct are cvx?

$$F(\theta) = \mathbb{E}\left[\ell(y, f_\theta(x))\right]$$

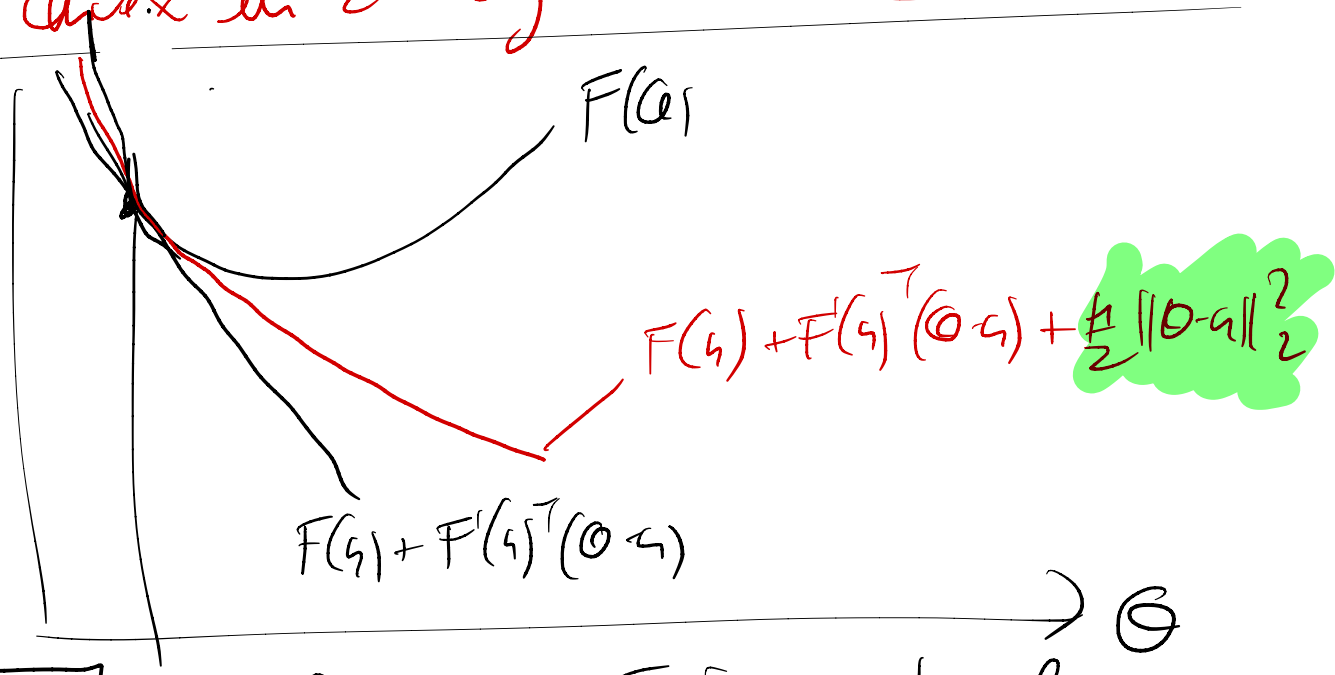
test

empirical

f_θ need to be linear in θ

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

cvx in 2nd argument : "weak"



μ-strong cvxity

strict

Lemma : F is μ-strongly cvx

$\Leftrightarrow$ $F(\theta) - \frac{\mu}{2}\|\theta\|_2^2$ is cvx.

Corollary : F cvx $\Rightarrow$ $F + \frac{\mu}{2}\|\cdot\|^2$ μ-strongly cvx

Prop : if F twice diff.

F μ-strongly cvx

$\Leftrightarrow$ $\forall \theta$, eig of $F''(\theta)$ are $\geq \mu$

Prop: if F is μ-strongly cvx, then it has a unique minimizer x^* and $\forall \theta$, $F(\theta) - F(x^*) \leq \frac{1}{2\mu} \|F'(\theta)\|^2$

proof: by definition of μ-strongly

$$F(\eta) \geq F(\theta) + F'(\theta)^T(\eta - \theta) + \frac{\mu}{2}\|\theta - \eta\|_2^2$$

$\downarrow \min_\eta$ $\qquad \downarrow \inf_\eta$

$$F(x^*) \geq F(\theta) - \frac{1}{2\mu}\|F'(\theta)\|_2^2$$

Polyak, Lojasiewicz inequality

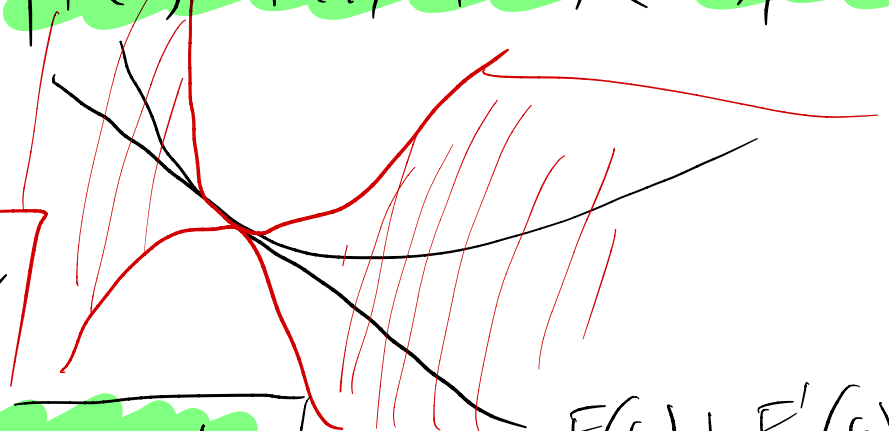
Lemma

$\inf_z \; z^T a - \frac{1}{2}\|z\|^2 = \frac{1}{2}\|a\|^2$

pf: gradient is $a - z$, so $z^* = a$ ...

Smoothness: $F: \mathbb{R}^d \to \mathbb{R}$ is L-smooth

iff $\forall \theta, \eta$, $|F(\theta) - F(\eta) - F'(\eta)^T(\theta - \eta)| \leq \frac{L}{2}\|\theta - \eta\|_2^2$



$F(\eta) + F'(\eta)^T(\theta - \eta) + \frac{L}{2}\|\theta - \eta\|_2^2$

Prop: if F twice diff

F is L-smooth

$\iff \forall \theta$, eig of $F''(\theta) \leq L$

$F(\eta) + F'(\eta)^T(\theta - \eta)$

lim inf, what is μ and L?

For ML. $F(\theta) = \mathbb{E}\left(\ell(y, \theta^\top \varphi(x))\right)$

$F'(\theta) = \mathbb{E}\left(\ell'(y, \theta^\top \varphi(x))\, \varphi(x)\right)$

lemma: gradient of $\theta \mapsto g(\theta^\top a)$ is $g'(\theta^\top a)\, a$

hessian is $g''(\theta^\top a)\, aa^\top$

proof: partial derivatives

Smooth (ex)

↳ derivative wrt to 2nd variable

$F''(\theta) = \mathbb{E}\left[\ell''(y, \theta^\top \varphi(x))\, \varphi(x)\varphi(x)^\top\right]$

ex: least-squares $\ell''(\,) = 1 \Rightarrow F''(\theta) = \mathbb{E}\left[\varphi(x)\varphi(x)^\top\right]$ covariance matrix

$\ell(y, u) = \frac{1}{2}(y-u)^2$

$\ell'(y, u) = y - u$

$\ell''(\,) = 1$

logistic = exercise $\quad \ell'' \leq 1/4$

strong-convexity = $\mathbb{E}\left(\ell(y, \theta^\top \varphi(x))\right) + \frac{\mu}{2}\|\theta\|^2$

if $|\ell'| \leq G$, $\|\varphi(x)\| \leq R$

$\Rightarrow \|F'(\theta)\|_2 \leq GR$

Convergence of GD for L-smooth, μ-strongly cvx function F

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

Need for Lyapunov fund V

$$0 \le V(\theta_t) \le V(\theta_{t-1})(1-\alpha) \le V(\theta_0)(1-\alpha)^t$$

Blog post for "automatic" proof

smoothness ↓

$$F(\theta_t) - F(\eta_*) \le F(\theta_{t-1}) - F(\eta_*) + F'(\theta_{t-1})^\top(\theta_t - \theta_{t-1}) + \frac{L}{2}\|\theta_t - \theta_{t-1}\|^2$$

$$= F(\theta_{t-1}) - F(\eta_*) + F'(\theta_{t-1})^\top\left(-\frac{1}{L}F'(\theta_{t-1})\right) + \frac{L}{2}\left\|-\frac{1}{L}F'(\theta_{t-1})\right\|_2^2$$

$$= F(\theta_{t-1}) - F(\eta_*) - \frac{1}{2L}\|F'(\theta_{t-1})\|^2$$

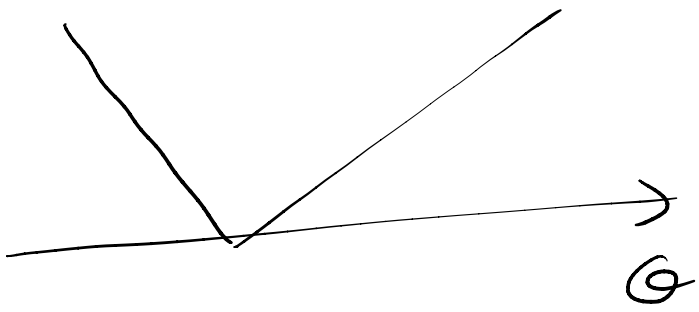
Łojasiewicz inequality ↓

$$\le F(\theta_{t-1}) - F(\eta_*) - \frac{1}{2L}\left[2\mu\left(F(\theta_{t-1}) - F(\eta_*)\right)\right]$$

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

Exponential (linear) convergence

Nonsmooth optimization: $F: \mathbb{R}^d \to \mathbb{R}$ convex



Lipschitz continuous

$\forall \theta, \eta, \quad |F(\theta) - F(\eta)| \le B \|\theta - \eta\|_2$

$\Leftrightarrow \quad \|F'(\theta)\|_2 \le B$

if F differentiable

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

proof (Assuming η_* is the minimizer of F).

$$\|\theta_t - \eta_*\|_2^2 = \|\theta_{t-1} - \eta_* - \gamma F'(\theta_{t-1})\|^2 \quad \text{by def of GD}$$

$$= \|\theta_{t-1} - \eta_*\|^2 + \gamma^2 \underbrace{\|F'(\theta_{t-1})\|^2}_{\le B^2} - 2\gamma \underbrace{F'(\theta_{t-1})^\top (\theta_{t-1} - \eta_*)}_{\ge F(\theta_{t-1}) - F(\eta_*) \text{ by convexity}}$$

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

$$F(\theta_{t-1}) - F(\eta_*) \le \frac{1}{2\gamma}\left[\|\theta_{t-1} - \eta_*\|^2 - \|\theta_t - \eta_*\|^2\right] + \frac{\gamma B^2}{2} \quad \to \quad \frac{1}{T}\sum_{t=1}^{T}$$

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

≤ 0

$\hookrightarrow \ge \min_{t=1,\dots,T} F(\theta_{t-1})$ or $\ge F\left(\frac{1}{T}\sum_{t=1}^{T} \theta_{t-1}\right)$ (Jensen's inequality)

$$\left.\begin{array}{l}\min_{t=1,\dots,T} F(\theta_{t-1}) \\ F\left(\frac{1}{T}\sum_{t=1}^{T}\theta_{t-1}\right)\end{array}\right\} \le \frac{1}{2\gamma T}\underbrace{\|\theta_0-\theta_*\|^2}_{\le D^2} + \frac{\gamma B^2}{2} \le \frac{D^2}{2\gamma T} + \frac{\gamma B^2}{2}$$

$$\frac{d}{d\gamma}(\cdot)=0 \iff \gamma_* = \frac{D}{B\sqrt{T}}.$$

T = "horizon"

$$\le \frac{BD}{\sqrt{T}}$$

See back for $\gamma_t = \frac{D}{B\sqrt{t}}$ "anytime" step size $\Rightarrow \frac{BD}{\sqrt{T}}\log T$

If $B = GR \Rightarrow$ optimum error is $\frac{GRD}{\sqrt{T}}$ after T iterations

Recall: uniform deviations $\frac{GRD}{\sqrt{n}}$ with n observations

GD can be stopped after $T = \Omega(n)$ iterations

But how many accesses to gradients? $\Rightarrow nT$ evaluations $\sim \boxed{n^2}$

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

SGD ① $F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell(y_i, f_\theta(x_i))}_{F_i(\theta)}$

empirical risk

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

↳ $i(t)$ random in $\{1,\dots,n\}$ uniform and independently

sampling with replacement

$\mathbb{E}[g_t \mid \theta_{t-1}] = \frac{1}{n}\sum_{i=1}^{n} F_i'(\theta_{t-1}) = F'(\theta_{t-1})$

② $F(\theta) = \mathbb{E}[\ell(y, f_\theta(x))]$

expected risk.

$\theta_t = \theta_{t-1} - \gamma \underbrace{\nabla_\theta \ell(y_t, f_\theta(x_t))\big|_{\theta=\theta_{t-1}}}_{g_t}$

unbiased gradient

independent draw from the test dist.

$t \le n$

↑ # observations

$\mathbb{E}(g_t \mid \theta_{t-1}) = \mathbb{E}\left(\nabla_\theta \ell(y, f_\theta(x))\big|_{\theta=\theta_{t-1}}\right)$

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

Summary = $\theta_t = \theta_{t-1} - \gamma g_t$, $\mathbb{E}(g_t \mid \theta_{t-1}) = F'(\theta_{t-1})$