Summary of previous lecture

Optimization: ① least-squares } condition number
② smooth optimization }
③ non-smooth optimization + SGD

Data: $(x_i, y_i) \in \mathcal{X} \times \mathbb{R} / \{-1,1\}$ learn a function $f : \mathcal{X} \to \mathbb{R}$

Expected risk: $R(f) = \mathbb{E}\, \ell(y, f(x))$ testing data

Empirical risk $\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$

ERM: class of models $\mathcal{F}$, $\hat{f} \in \arg\min_{f \in \mathcal{F}} \hat{R}(f)$

goal: find $\hat{f}$ s.t. $\hat{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \hat{R} \leq \varepsilon$

Estimation error:

$$\underbrace{R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f)}_{R(f_{\mathcal{F}}^*)} = R(\hat{f}) - \hat{R}(\hat{f}) + \underbrace{\hat{R}(\hat{f}) - \hat{R}(f_{\mathcal{F}}^*)}_{\leq 0 \;\Rightarrow\; \varepsilon} + \hat{R}(f_{\mathcal{F}}^*) - R(f_{\mathcal{F}}^*)$$

$$\leq \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| + \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| + \varepsilon$$

Tool of choice: Rademacher complexity $= O\left(\frac{1}{\sqrt{n}}\right)$

Optimization = forget about ML

Goal = find a minimizer of $F:\mathbb{R}^d \to \mathbb{R}$

Motivation: ML $F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$

linear model

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

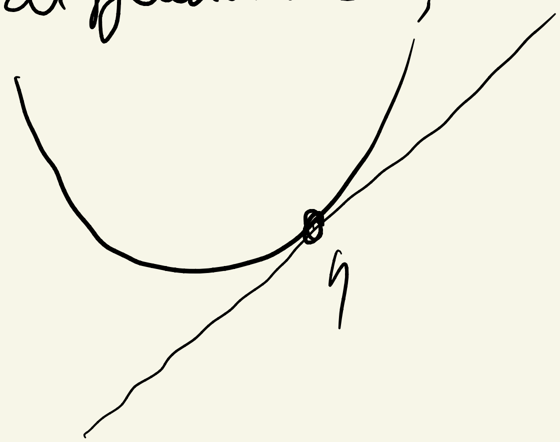
(1) Convex vs non convex

Positive semi definite matrix

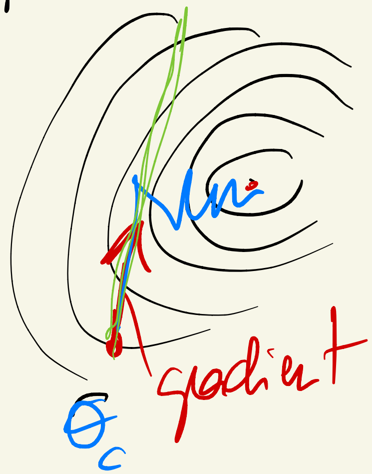
F: convex

- if F twice differentiable, $F''(\theta) \succcurlyeq 0 \quad \forall \theta$
- if F differentiable, F above all of its tangents

$$F(\theta) \geq \underbrace{F(\eta) + F'(\eta)^T(\theta - \eta)}_{\text{tangent at } \eta}$$



gradient descent (Cauchy, 1847)



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

step-size > 0 < constant / decaying with fixed schedule / line search

- quadratic
- smooth
- Non-smooth

quadratic function: $F(\theta) = \frac{1}{2}\theta^T H \theta - c^T\theta$, convex $H \succcurlyeq 0$

H invertible

η_* minimizer unique

$$F'(\eta_*) = \underbrace{0 = H\eta_* - c}_{\text{Normal equations}} \Rightarrow \eta_* = H^{-1}c$$

gradient descent: $\theta_t = \theta_{t-1} - \gamma[H\theta_{t-1} - \overbrace{c}^{H\eta_*}] = \theta_{t-1} - \gamma[H\theta_{t-1} - H\eta_*]$

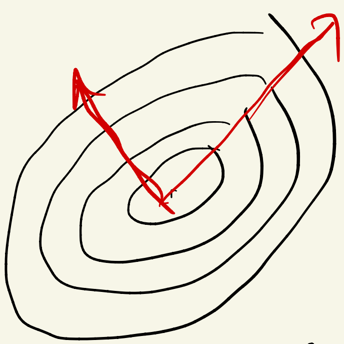
Linear iteration

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

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

Eigenvalue decomposition: $H = \sum_{i=1}^{d} \underbrace{\lambda_i}_{\text{eigenvalue}} \underbrace{u_i u_i^T}_{\text{eigenvector}}$

$(I - \gamma H)^t$ has same eigenvectors, and $\boxed{(1 - \lambda_i \gamma)^t}$ as eigenvalues



① when is it convergent? if $|1 - \lambda_i \gamma| < 1$, $\forall i$

② how fast?

$\Leftrightarrow 1 - \lambda_i \gamma > -1$

$\gamma < \frac{2}{\lambda_i}$ $\forall i$

$\boxed{\gamma < 2/L}$

Def: μ = smallest eigenvalue

L = largest eigenvalue

$0 < \mu < L$

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

$\theta_t - \eta_* = (1-\gamma H)^t(\theta_0 - \eta_*)$

$\|\theta_t - \eta_*\|^2 \le \Big(\underline{\text{largest eigenvalue of } (1-\gamma H)^t}\Big)^2 \|\theta_0 - \eta_*\|^2$

$\|Mz\|^2 \le \|z\|^2 \big(\text{largest eig of } M\big)^2$

with $\gamma = \frac{1}{L}$ $\quad (1-\gamma H)^t = \big(\underbrace{1 - \tfrac{1}{L}H}_{\text{positive}}\big)^t$

with largest eig. $\left(1-\frac{\mu}{L}\right)^t$ because μ is the smallest eig. of H

$\|\theta_t - \eta_*\|^2 \le \left(1-\frac{\mu}{L}\right)^{2t}\|\theta_0 - \eta_*\|^2$ linear convergence, exponential convergence

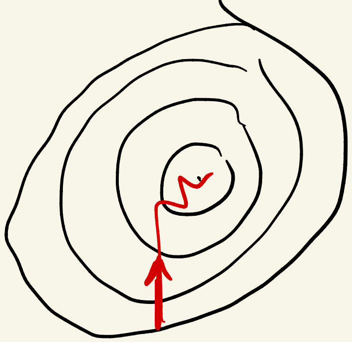
$\le \exp\left(-\frac{2\mu}{L}t\right)\|\theta_0 - \eta_*\|^2$

$1-\alpha \le e^{-\alpha}$

$\le \boxed{\exp(-2t/\kappa)}\,\|\theta_0 - \eta_*\|^2$ with $\kappa = \frac{L}{\mu} \geq 1$ condition number

$= \varepsilon$

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



small $\kappa = \frac{L}{\mu}$

large $\kappa = \frac{L}{\mu}$

How small is μ?
How big is κ?
$H = \frac{1}{n}\Phi^\top\Phi$ for least squares

$$\theta_t - \eta_* = (1-\gamma H)^t(\theta_0 - \eta_*) \qquad \gamma = 1/L$$

$$F(\theta_t) - F(\eta_*) = \frac{1}{2}(\theta_t-\eta_*)^T H (\theta_t - \eta_*)$$ by Taylor expansion around η_*

$F'(\eta_*) = 0 \quad F''(\eta_*) = H$

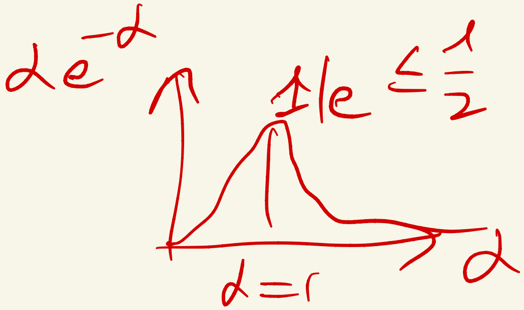
$$= \frac{1}{2}(\theta_0-\eta_*)^T \underline{H}(1-\gamma H)^{2t}(\theta_0-\eta_*)$$

$$\leq \frac{1}{2}\left(\text{largest eig. of } \underline{H}\left(1-\frac{1}{L}H\right)^{2t}\right) \times \|\theta_0-\eta_*\|^2$$

$$\leq \frac{1}{2} \sup_{\lambda\in[\mu,L]} \lambda\left(1-\frac{1}{L}\lambda\right)^{2t} \|\theta_0-\eta_*\|^2$$

$$\lambda\left(1-\frac{\lambda}{L}\right)^{2t} \leq \lambda e^{-2t\lambda/L} = \underbrace{\frac{2t\lambda}{L} e^{-2t\lambda/L}}_{\leq 1/2} \frac{L}{2t} \leq \frac{L}{4t}$$

$1-\alpha \leq e^{-\alpha}$



- Consequence = $F(\theta_t) - F(\eta_*) \leq \dfrac{L}{4t}\|\theta_0 - \eta_*\|^2$
- Adaptivity
- Optimality with acceleration

linear: $\kappa \to \sqrt{\kappa}$ / $\frac{1}{t} \to 1/t^2$

True as well for convex functions: Spectrum of $H \subset [\mu, L]$

Assumption: F is smooth "$\Leftrightarrow$" $\forall \theta$, $F''(\theta)$ has eigenvalues less than L

10h18

10h40

F is strongly convex "$\Leftrightarrow$" $\forall \theta$, $F''(\theta)$ ——— larger than μ

Only if F is twice differentiable

Lemmas: ① if F is smooth

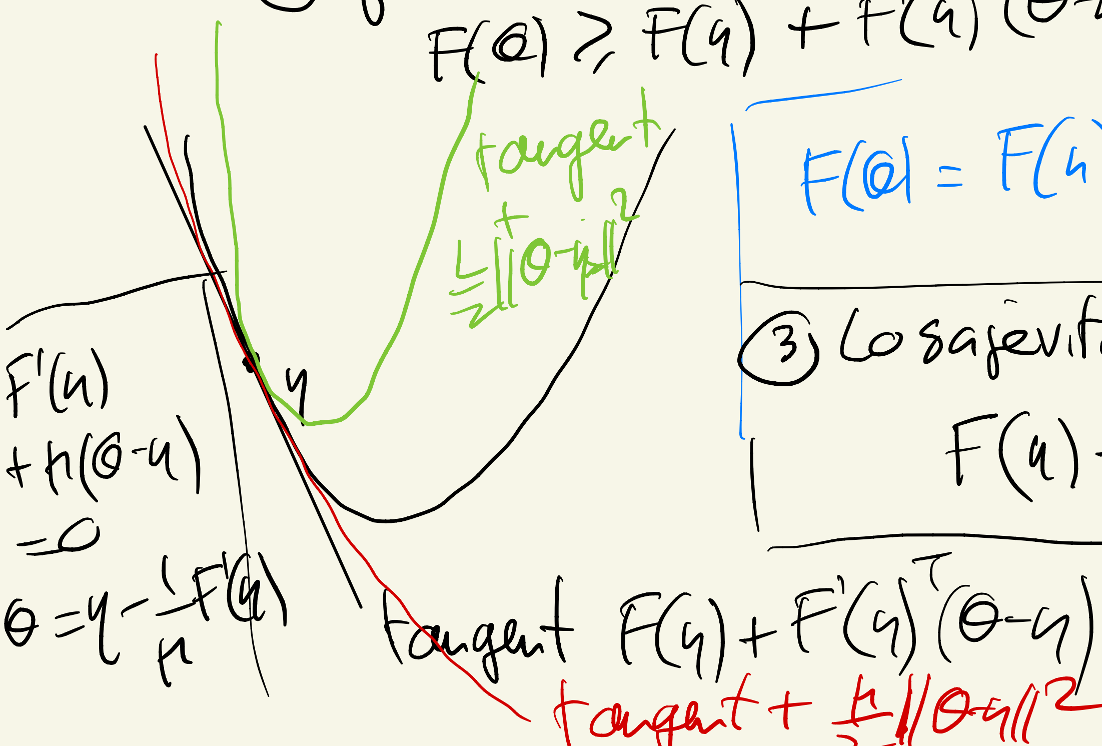
$$F(\theta) \leq F(\eta) + F'(\eta)^T(\theta-\eta) + \boxed{\frac{L}{2}\|\theta-\eta\|^2}$$

② if F is strongly convex

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

proof: Taylor expansion

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



③ Lojasevitch inequality

$$F(\eta) - F(\eta_*) \leq \frac{1}{2\mu}\|F'(\eta)\|^2$$

Proof: $F(\eta_*) \geq F(\eta) - \frac{1}{2\mu}\|F'(\eta)\|^2$

Proof of convergence of gradient descent

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

smoothness

$$F(\theta_t) \leq F(\theta_{t-1}) + F'(\theta_{t-1})^T(\theta_t - \theta_{t-1}) + \frac{L}{2}\|\theta_t - \theta_{t-1}\|^2$$

$$\leq F(\theta_{t-1}) + F'(\theta_{t-1})^T(-\gamma F'(\theta_{t-1})) + \frac{L}{2}\|\gamma F'(\theta_{t-1})\|^2$$

using GD iteration

$$= F(\theta_{t-1}) - (\gamma - \underbrace{\frac{L}{2}\gamma^2}_{\gamma/2})\|F'(\theta_{t-1})\|^2 \qquad \gamma = 1/L$$

$$F(\theta_t) \leq F(\theta_{t-1}) - \frac{\gamma}{2}\underbrace{\|F'(\theta_{t-1})\|^2}_{\geq 2\mu\left(F(\theta_{t-1}) - F(\eta_*)\right)}$$

Losajewitz ineq.

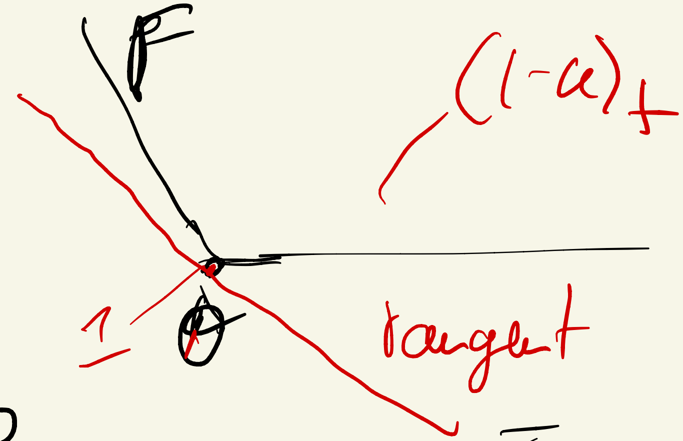
$$F(\theta_t) - F(\eta_*) \leq \underbrace{(1 - \gamma\mu)}_{1-\mu/L}\left(F(\theta_{t-1}) - F(\eta_*)\right)$$

other result

$$\frac{L}{t}\|\theta_0 - \eta_*\|^2$$

$$\leq \left(1 - \frac{\mu}{L}\right)^t \left[F(\theta_0) - F(\eta_*)\right]$$

Non-smooth optimization



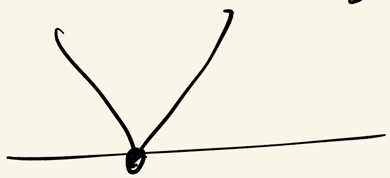
F convex, not necessarily differentiable

gradient may not exist

use "subgradient" at θ : any vector z

such that $F(\eta) \geq F(\theta) + z^\top(\eta - \theta)$

Example : $|\theta|$



at $\theta > 0$, gradient $= 1$

$\theta < 0$, gradient $= -1$

$\theta = 0$, subgradient : any element of $[-1, 1]$

Example from ML

$$F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(1 - y_i \theta^\top \varphi(x_i)\right)_+$$

SVM

$B = \max_i \|\varphi(x_i)\|$

Assumption : F is convex and Lipschitz-continuous

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

Proposition : $\forall s$ subgradient of F, $\|s\|_2 \leq B$

Subgradient method: $\theta_t = \theta_{t-1} - \underbrace{\gamma_t}_{\text{decaying stepsize}} \underbrace{F'(\theta_{t-1})}_{\text{any subgradient of } F \text{ at } \theta_{t-1}}$

Convergence proof:

Assumption: η_* is a global minimizer

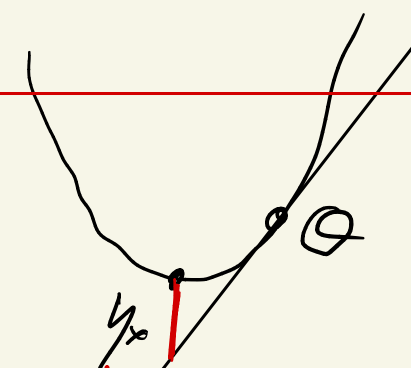
$$\|\theta_t - \eta_*\|^2 = \|\theta_{t-1} - \eta_* - \gamma_t F'(\theta_{t-1})\|^2$$
$$= \|\theta_{t-1} - \eta_*\|^2 - 2\gamma_t \underbrace{F'(\theta_{t-1})^\top(\theta_{t-1} - \eta_*)}_{\geq F(\theta_{t-1}) - F(\eta_*)} + \underbrace{\gamma_t^2 \|F'(\theta_{t-1})\|^2}_{\leq \gamma_t^2 B^2}$$

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

Lemma: $\forall \theta$: $F(\theta) - F(\eta_*) \leq F'(\theta)^\top(\theta - \eta_*)$

Proof: $F(\eta_*) \geq F(\theta) + F'(\theta)^\top(\eta_* - \theta)$

function above tangent



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

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

$$2\sum_{t=1}^{s}\gamma_t\Big[\underbrace{F(\theta_{t-1})}_{\ge \min\limits_{t\in\{1,\dots,s\}} F(\theta_{t-1})} - F(\theta_*)\Big] \le \underbrace{\|\theta_0-\eta_*\|^2 - \cancel{\|\theta_s-\eta_*\|^2}}_{\text{telescoping sum}} + \sum_{t=1}^{s}\gamma_t^2 B^2$$

$$\left(2\sum_{t=1}^{s}\gamma_t\right)\left[\min_t F(\theta_{t-1}) - F(\eta_*)\right] \le$$

$$\min_{t\in\{1,\dots,s\}} F(\theta_{t-1}) - F(\eta_*) \le \frac{\|\theta_0-\eta_*\|^2}{2\sum_{t=1}^{s}\gamma_t} + \frac{\sum_{t=1}^{s}\gamma_t^2}{\sum_{t=1}^{s}\gamma_t}B^2$$

$\gamma_t=\gamma,\ \forall t$ $\qquad \dfrac{}{s\gamma} + \gamma B^2$

$\gamma_t = 1/t^\alpha \to \alpha = 1$ $\quad \sum_{t=1}^{s}\frac{1}{t} \sim \log(s)$ $\qquad \alpha = 1/2$

if $\gamma_t = \frac{\gamma}{\sqrt{t}}$; $\sum_{t=1}^{S} \gamma_t = \gamma \sum_{t=1}^{S} \underbrace{\frac{1}{\sqrt{t}}}_{\geq \frac{1}{\sqrt{S}}} \geq \gamma S / \sqrt{S} = \gamma \sqrt{S}$

$$\sum_{t=1}^{S} \gamma_t^2 = \gamma^2 \sum_{t=1}^{S} \frac{1}{t} \leq \gamma^2 (1 + \log S)$$

compare with integral

$$\sum_{t=1}^{S} \frac{1}{t} \leq 1 + \sum_{t=2}^{S} \frac{1}{t} \leq 1 + \int_1^{S-1} \frac{du}{u}$$

$$\leq 1 + \log(S-1)$$

$$\min_{t \in \{1, \dots, S\}} F(\theta_{t-1}) - F(\eta_*) \leq \frac{\frac{1}{2} \|\theta_0 - \eta_*\|^2}{\gamma \sqrt{S}} + \frac{1}{2} B^2 \gamma \frac{1 + \log S}{\sqrt{S}}$$

$\longrightarrow 0$ as $S \longrightarrow +\infty$

as $\frac{\log S}{\sqrt{S}}$

$\gamma_t = 1/t^\alpha$

$\sum \gamma_t = S^{1-\alpha}$

$\sum \gamma_t^2 \sim$

Stochastic gradient descent

Version 1: Minimize $\frac{1}{n}\sum_{i=1}^{n} F_i(\theta)$ "finite sum" ex: empirical risk

$= \mathbb{E}\, G(\theta, z)$ using empirical dist

Algo: $\theta_t = \theta_{t-1} - \gamma_t \underbrace{F'_{i(t)}}_{\text{index chosen at random in } \{1,\dots,n\} \text{ independently}}(\theta_{t-1})$

will converge to $\theta_* = \arg\min \frac{1}{n}\sum F_i$

Version 2: $F(\theta) = \mathbb{E}_z\, G(\theta, z)$ "expectation" ex: expected risk

Algo: $\theta_t = \theta_{t-1} - \gamma_t\, G'(\theta_{t-1}, \underline{z_t})$ independent observation

$\boxed{t = n}$

will converge to $\theta_* = \arg\min F$

$$\theta_t = \theta_{t-1} - \gamma_t G'(\theta_{t-1}, z_t) \qquad \boxed{E_z G(\theta, z) = F(\theta)}$$

$$\|\theta_t - \theta_*\|^2 = \|\theta_{t-1} - \theta_*\|^2 - 2\gamma_t(\theta_{t-1} - \theta_*)^T G'(\theta_{t-1}, z_t)$$

$$\underbrace{+\gamma_t^2 \|G'(\theta_{t-1}, z_t)\|^2}_{\leq \gamma_t^2 B^2}$$

$E(\quad | \theta_{t-1})$

$$E(\|\theta_t - \theta_*\|^2 | \theta_{t-1}) = \|\theta_{t-1} - \theta_*\|^2 + \gamma_t^2 B^2$$

$$-2\gamma_t \underbrace{(\theta_{t-1} - \theta_*)^T F(\theta_{t-1})}_{\geq F(\theta_{t-1}) - F(\theta_*)}$$

$$\mathbb{E} \|\theta_t - \theta_*\|^2 \leq \mathbb{E} \|\theta_{t-1} - \theta_*\|^2 + \gamma_t^2 B^2$$

$$- 2\gamma_t \left(\mathbb{E}(F(\theta_{t-1})) - F(\theta_*) \right)$$

$$\sum_{t=1}^{s} \gamma_t \mathbb{E}(F(\theta_t)) - F(\theta_*) \leq \_\_\_\_\_\_$$

Jensen's inequality

$$\sum_{t=1}^{s} \gamma_t F(\theta_t) \geq \left(\sum_{t=1}^{s} \gamma_t\right) F\left(\frac{\sum_{t=1}^{s} \gamma_t \theta_t}{\sum_{t=1}^{s} \gamma_t}\right)$$

$\bar{\theta}_s$

$$\mathbb{E} F(\theta) \geq F(\mathbb{E}\theta)$$

$\min F(\theta_{t-1})$

$t \in \{1, \dots, s\}$

$$\gamma_t = \frac{\gamma}{\sqrt{t}}$$

$$\boxed{\mathbb{E}\left(F(\bar{\theta}_s)\right) - F(\theta_*) \leq \frac{1}{2\gamma\sqrt{s}} \|\theta_0 - \theta_*\|^2 + \frac{\gamma B^2 (1 + \log s)}{2\sqrt{s}}}$$

$s = n$

$1/\sqrt{n}$

complexity $O(n)$