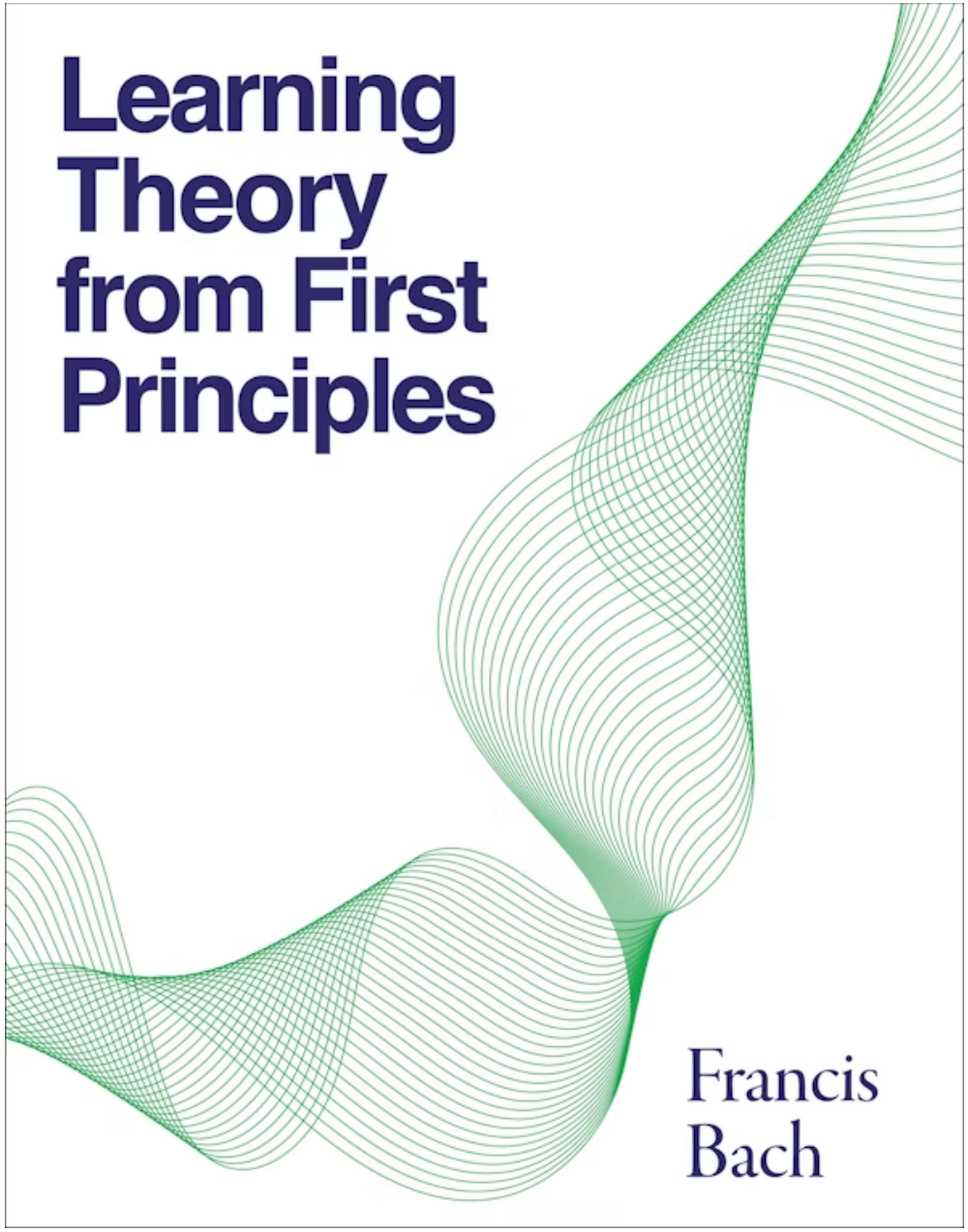
Adaptivity in Machine Learning



. Available online
. Feel free to say!

---

3 lectures
1. SGD
2. kernels
3. Neural networks

Simple notes online

Supervised learning

data $(x_i, y_i) \in X \times Y \quad i = 1, \dots, n$ IID

goal: to estimate $f: X \to Y$

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

$\ell(y, z) \longrightarrow$ regression $Y = \mathbb{R}$

$\ell(y, z) = (y - z)^2$

classification $Y = \{-1, 1\}$

0-1 loss $\ell(y, z) = 1_{y \neq z}$

Expected risk: $R(f) = \mathbb{E}[\ell(Y, f(X))]$ ≈ test error

function $X \to Y$ wrt test distribution

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

optimal function f_* $\simeq$ Bayes predictor

target function

$\simeq \operatorname{argmin} R(f)$

$\Rightarrow R^* = R(f_*)$

Bayes risk

Formula $R(f) = \mathbb{E}(L(y, f(x)))$

$= \int \left(\int \mathbb{E} L(y, f(x)) \, q(y|x) \, dy \right) p(x) \, dx$

$\mathbb{E}(L(y, f(x)) | x)$

$\Rightarrow f^*(x) = \operatorname{argmin}_{z \in Y} \mathbb{E}(L(y, z) | x)$

square loss

$\longrightarrow \mathbb{E}[y|x]$

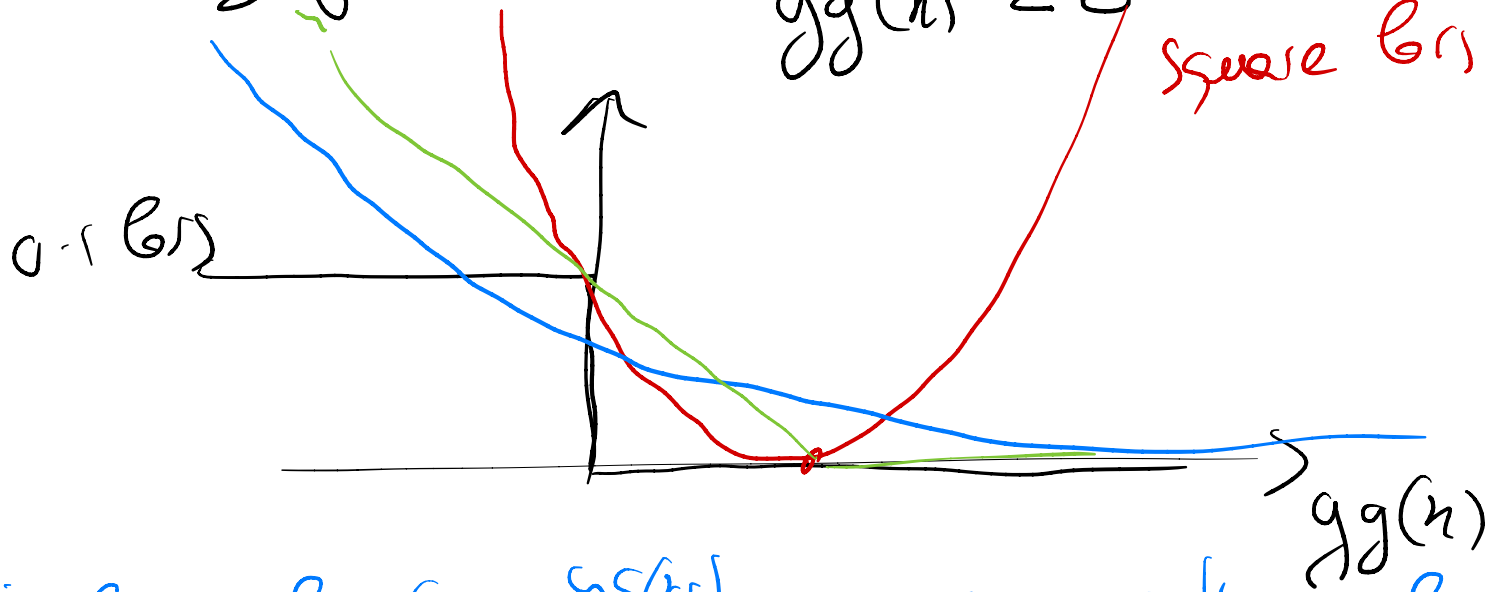
0-1 loss $f^*(x) = \operatorname{sign}\left(\mathbb{E}(y|x)\right) \; | x$

Convex surrogates for classification

Instead of estimating $f : \mathcal{X} \to \{-1, 1\}$

we estimate $g : \mathcal{X} \to \mathbb{R}$ and $f = \mathrm{sign}(g)$

0-1 loss of $f(x) = \mathbb{1}_{yg(x) < 0}$



logistic loss : $\log(1+e^{-yg(x)}) \simeq$ cross entropy loss

hinge loss : $(1 - yg(x))_+$

$\Rightarrow$ ASSUME : $\ell(y, z)$ is G-lipschitz continuous in z

Analysis : $\mathcal{F} = \{ f_\theta : x \to \mathbb{R} \} = \{ f_\theta : x \to \mathbb{R}, \ \theta \in \Theta \}$

Goal : find $\hat{f} \in \mathcal{F}$ s.t. $R(\hat{f}) - R_*$ small

decomposition : $R(\hat{f}) - R_* = \left(R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f) \right) + \left(\inf_{f \in \mathcal{F}} R(f) - R_* \right)$

estimation error ≥ 0 — approximation error — $R(f_{\mathcal{F}}^*)$

Empirical risk minimizer

$$\hat{f} \approx \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}(f)$$

Estimation error $= \left[R(\hat{f}) - \hat{R}(\hat{f}) \right] + \left(\hat{R}(\hat{f}) - \hat{R}(f_{\mathcal{F}}^*) \right) + \left(\hat{R}(f_{\mathcal{F}}^*) - R(f_{\mathcal{F}}^*) \right)$

$\leq \hat{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \hat{R}(f)$ — optimization error

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

$\mathbb{E} \hat{R}(f) = R(f)$ if f fixed

$\leq \sup_{f \in \mathcal{F}} R(f) - \hat{R}(f)$ — uniform deviation

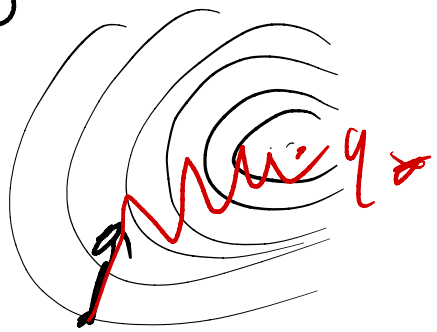
gradient descent : $F: \mathbb{R}^d \to \mathbb{R}$ convex, B-Lipschitz continuous

ex: ① $F(\theta) = \hat{R}(\beta_\theta)$ with $\beta_\theta(x) = \theta^T \varphi(x)$ where $\varphi: \mathcal{X} \to \mathbb{R}^d$
(feature map)

convex loss G-Lipschitz continuous
and $\|\varphi(x)\| \le R$ for all x $\Rightarrow$ $B = GR$
Convex

② $F(\theta) = R(\beta_\theta)$

gradient descent = iterative alg. $\theta_k = \theta_{k-1} - \gamma F'(\theta_{k-1})$

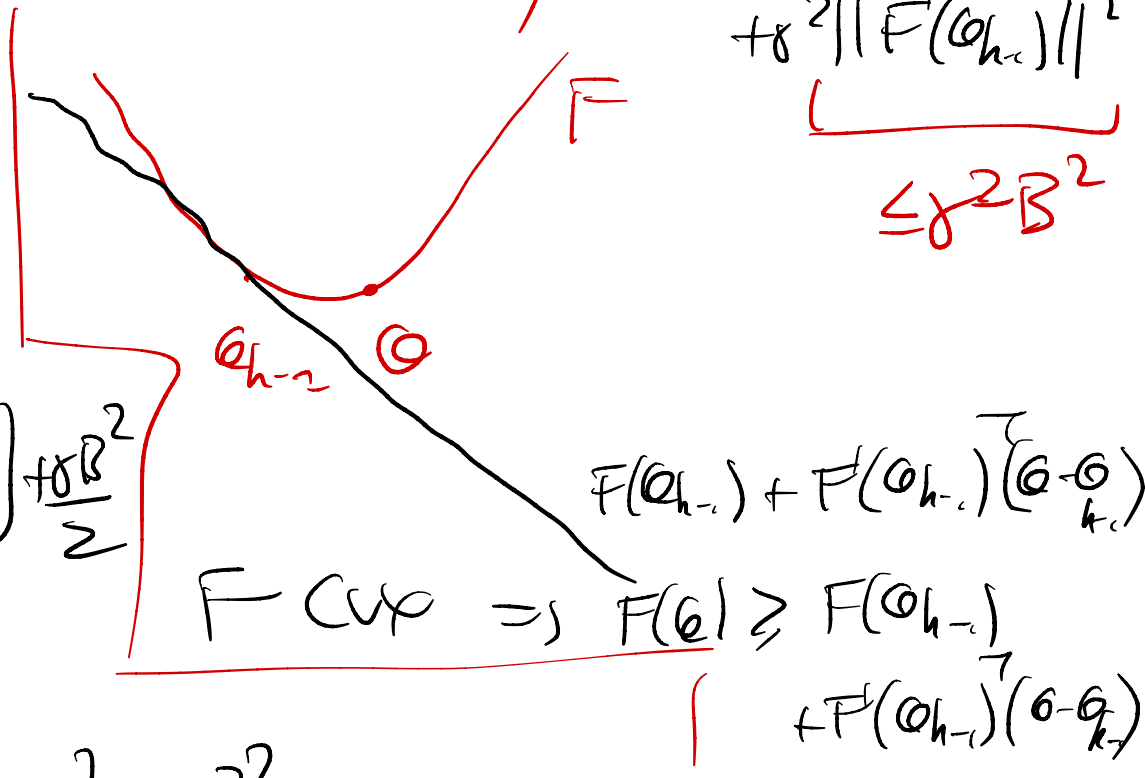


Analysis: $\theta_{k} = \Pi_D\left(\theta_{k-1} - \gamma F'(\theta_{k-1})\right)$ projects over ball of center and radius D

For any $\theta \in \mathbb{R}^d$ s.t. $\|\theta\| \le D$

$$\|\theta_k - \theta\|^2 \le \|\theta_{k-1} - \gamma F'(\theta_{k-1}) - \theta\|^2 = \|\theta_{k-1} - \theta\|^2 - 2\gamma F'(\theta_{k-1})^\top(\theta - \theta_{k-1}) + \gamma^2 \|F'(\theta_{k-1})\|^2$$

$$\le \|\theta_{k-1} - \theta\|^2 + \gamma^2 B^2 - 2\gamma\left[F(\theta_{k-1}) - F(\theta)\right]$$



$\le \gamma^2 B^2$

$$\Leftrightarrow F(\theta_{k-1}) - F(\theta) \le \frac{1}{2\gamma}\left[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2\right] + \frac{\gamma B^2}{2}$$

F CVX $\Rightarrow F(\theta) \ge F(\theta_{k-1}) + F'(\theta_{k-1})^\top(\theta - \theta_{k-1})$

summing from $k=1$ to t

$$\frac{1}{t}\sum_{k=1}^{t} F(\theta_{k-1}) - F(\theta) \le \frac{1}{2\gamma t}\|\theta_0 - \theta\|^2 + \frac{\gamma B^2}{2}$$

(Jensen's)

$$F\left(\frac{1}{t}\sum_{k=1}^{t}\theta_{k-1}\right) \le F(\theta) + \frac{1}{2\gamma t}\|\theta_0 - \theta\|^2 + \frac{\gamma B^2}{2} \le \frac{BD}{\sqrt{t}}$$

$\theta_0 = 0$ $\quad D^2/2\gamma t + \gamma B^2/2$ $\quad$ if $\gamma \propto \frac{1}{\sqrt{t}}$

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

$$F'(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell'(y_i, \theta^T\varphi(x_i))\,\varphi(x_i)$$

$$F(\text{averaged iterate}) - F(\theta_*) \leq \frac{GRD}{\sqrt{t}}$$

\# iterations

NB: uniform deviation bounds $\frac{GRD}{\sqrt{n}}$

⇒ optimization is "free" if $t = n$

\# directions

⇒ $O(n^2)$ access to gradients

SGD = minimize $F(\theta) = \mathbb{E}\, \ell(y, \theta^\top \varphi(x))$

access to unbiased gradients

$$\theta_k = \theta_{k-1} - \gamma g_k$$

$$\mathbb{E}(g_k \mid \theta_{k-1}) = F'(\theta_{k-1})$$

stochastic gradient

$$\Rightarrow g_k = \frac{\partial}{\partial \theta} \ell(y_k, \varphi_\theta(x_k))$$

at $\theta = \theta_{k-1}$

F = test error

Result:

$$\mathbb{E}\, F\left(\frac{1}{n} \sum_{k=1}^{n} \theta_{k-1}\right) \leq F(\theta) + \frac{GRD}{\sqrt{n}}$$

↳ training data

if $k \leq n$

complexity : $O(n)$

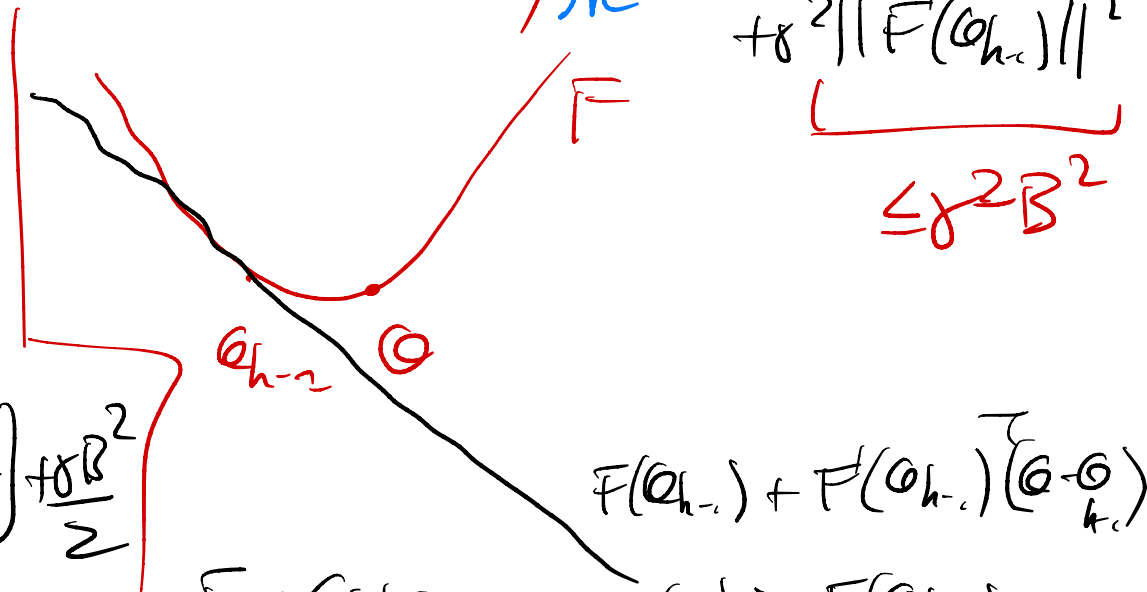
access to grad.

Analysis: $\theta_{k} = \Pi_D\left(\theta_{k-1} - \gamma \underbrace{F'(\theta_{k-1})}_{g_k}\right)$ projects over ball of center and radius D

For any $\theta \in \mathbb{R}^d$, $\|\theta\| \le D$

$$\|\theta_k - \theta\|^2 \le \|\theta_{k-1} - \gamma F'(\theta_{k-1}) - \theta\|^2 = \|\theta_{k-1} - \theta\|^2 - 2\gamma \underbrace{F'(\theta_{k-1})}_{g_k}{}^\top (\theta - \theta_{k-1}) + \gamma^2 \underbrace{\|F'(\theta_{k-1})\|^2}_{\le B^2}$$

$$\le \|\theta_{k-1} - \theta\|^2 + \gamma^2 B^2 - 2\gamma \left[F(\theta_{k-1}) - F(\theta)\right]$$



F CVX $\Rightarrow F(\theta) \ge F(\theta_{k-1}) + F'(\theta_{k-1})^\top(\theta - \theta_{k-1})$

$$F(\theta_{k-1}) - F(\theta) \le \frac{1}{2\gamma}\left[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2\right] + \frac{\gamma B^2}{2}$$

averaging from $k=1$ to t

$$\frac{1}{t}\sum_{k=1}^{t} F(\theta_{k-1}) - F(\theta) \le \frac{1}{2\gamma t}\|\theta_0 - \theta\|^2 + \frac{\gamma B^2}{2}$$

(Jensen's)

$$F\left(\frac{1}{t}\sum_{k=1}^{t} \theta_{k-1}\right) \le F(\theta) + \underbrace{\frac{1}{2\gamma t}\|\theta_0 - \theta\|^2 + \frac{\gamma B^2}{2}}_{\le D^2/2\gamma t + \gamma B^2/2} \le \frac{BD}{\sqrt{t}} \quad \text{if } \gamma = \frac{1}{\sqrt{t}}$$