parameter estimation

1. maximum likelihood estimation (regularization)
2. empirical risk minimization (structural)

- MLE

Def: $X_1 \dots X_n \sim P_\theta(x)$

likelihood $L(x_i, \theta) \triangleq P_\theta(x_i)$

Def: likelihood for the entire dataset

$$L_n(\theta) \triangleq P_\theta(x_1, \dots, x_n)$$

Def: MLE

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L_n(\theta)$$

MLE maps $X_1 \dots X_n \to \hat{\theta}_n$

Def: log-likelihood $l_n(\theta) \triangleq \log L_n(\theta)$

Def: bias of $\hat{\theta}_n$ $\mathbb{E}\hat{\theta}_n - \underbrace{\theta}_{\text{truth}}$

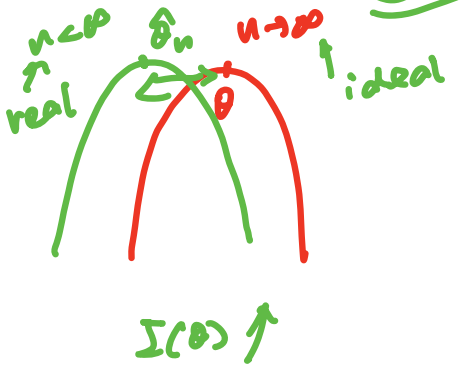
Def: if $\text{bias}(\hat{\theta}_n) = 0$ then $\hat{\theta}_n$ is the unbiased estimator

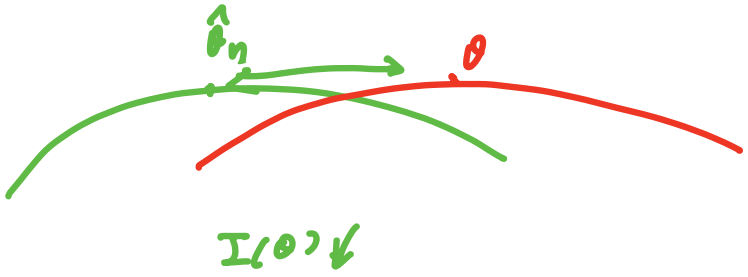
Theory of MLE

MLE is asymptotically normal and efficient

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \to \infty]{d} N(0, I^{-1}(\theta))$$

$I(\theta) \triangleq -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial \theta^2} \log P_\theta(X)\right]$ Fisher infor...





$f \in \mathcal{F}$ $\quad z_1 \cdots z_n \in \mathcal{Z}$

Empirical risk minimization.

$\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$ $\quad \ell$: loss

$L(f) \triangleq \mathbb{E}_P[\ell(f, z)]$ $\quad$ risk

($f \leftrightarrow \theta$, $z \leftrightarrow x$)

goal: $\min_f L(f)$ ← the population version

reality $\min_f \hat{L}_n(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(f, z_i)$ ($f \leftrightarrow \theta$, $z_i \leftrightarrow x_i$)

$-\log p_\theta(x_i)$

ERM

Example 1 (Binary classification problem)

$z = (x, y)$

x $\quad \{-1, +1\}$

$\mathbb{E}[\ell(f, z)] = \mathbb{E}[\mathbb{1}\{f(x)\, y \le 0\}]$

risk $\quad$ indicator

Example 2 (multiclass classification problem)

$y \in \{1, \dots, K\}$ $\quad f : \mathcal{X} \to \mathbb{R}^K$

$$\ell(f, x, y) = \mathbb{1}\Big\{\max_{i \neq y} f_i(x) \ge f_y(x)\Big\}$$

Example 3: (hinge or logistic) $\mathcal{F} = \{f_\theta : f_\theta(x) = \theta^T x\}$

$\ell_{hinge}(f_\theta, (x,y)) = [1 - y \underbrace{x^T\theta}_{f_\theta(x)}]_+$

$\ell_{logit}(f_\theta, (x,y)) = \log(1 + \exp(-y \underbrace{x^T\theta}_{f_\theta(x)}))$

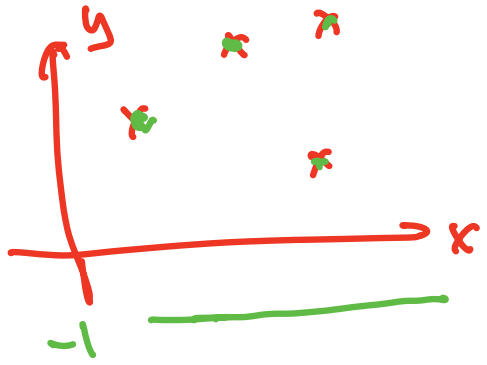
Given $z_1 \dots z_n$ pick $\hat{f}_n \in \mathcal{F}$ Generalization

$\hat{L}_n(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^n \ell(\hat{f}_n, z_i) \leq R(\hat{f}_n) + \underbrace{\varepsilon}_{\text{error}}$

$f_n(x) = y$ if (x,y) in the training dataset

$\hat{f}_n(x) = -1$ otherwise

$\ell(f, (x,y)) = (y - f(x))^2$



$\hat{L}_n(\hat{f}_n) = 0$

$R(\hat{f}_n) \nearrow$