

# Lecture 5 Asymptotic Normality

IEMS 402 Statistical Learning

Northwestern

# Bias

**Lemma 3** The bias of  $\hat{p}_h$  satisfies:

$$\sup_{p \in \Sigma(\beta, L)} |p_h(x) - p(x)| \leq ch^\beta \quad (14)$$

for some  $c$ .

**Proof.** We have

$$\begin{aligned} |p_h(x) - p(x)| &= \left| \int \frac{1}{h^d} K(\|u - x\|/h) p(u) du - p(x) \right| = \int \frac{1}{h^d} K(\|u - x\|/h) (p(u) - p(x)) du \\ &= \left| \int K(\|v\|) (p(x + hv) - p(x)) dv \right| \xrightarrow{\text{Taylor Expn}} p_{x,\beta}(x + hv) + p(x + hv) - p_{x,\beta}(x + hv) \quad \text{poly of } v \\ &\stackrel{\text{"H\"older"} \text{ and smooth}}{\leq} \left| \int K(\|v\|) (p(x + hv) - p_{x,\beta}(x + hv)) dv \right| + \left| \int K(\|v\|) (p_{x,\beta}(x + hv) - p(x)) dv \right|. \end{aligned}$$

The first term is bounded by  $Lh^\beta \int K(s)|s|^\beta$  since  $p \in \Sigma(\beta, L)$ . The second term is 0 from the properties on  $K$  since  $p_{x,\beta}(x + hv) - p(x)$  is a polynomial of degree  $\beta$  (with no constant term).  $\square$

① because higher order term cancel out the poly nominal.

# Variance

**Lemma 4** *The variance of  $\hat{p}_h$  satisfies:*

$$\sup_{p \in \Sigma(\beta, L)} \text{Var}(\hat{p}_h(x)) \leq \frac{c}{nh^d} \quad \begin{array}{l} \text{dependency over dimension} \\ \text{using } n \text{ data} \end{array} \quad (15)$$

for some  $c > 0$ .

**Proof.** We can write  $\hat{p}(x) = n^{-1} \sum_{i=1}^n Z_i$  where  $Z_i = \frac{1}{h^d} K\left(\frac{\|x-X_i\|}{h}\right)$ . Then,

$$\begin{aligned} \text{Var}(Z_i) &\leq \mathbb{E}(Z_i^2) = \frac{1}{h^{2d}} \int K^2\left(\frac{\|x-u\|}{h}\right) p(u) du = \frac{h^d}{h^{2d}} \int K^2(\|v\|) p(x+hv) dv \\ &\leq \frac{\sup_x p(x)}{h^d} \int K^2(\|v\|) dv \leq \frac{c}{h^d} \end{aligned}$$

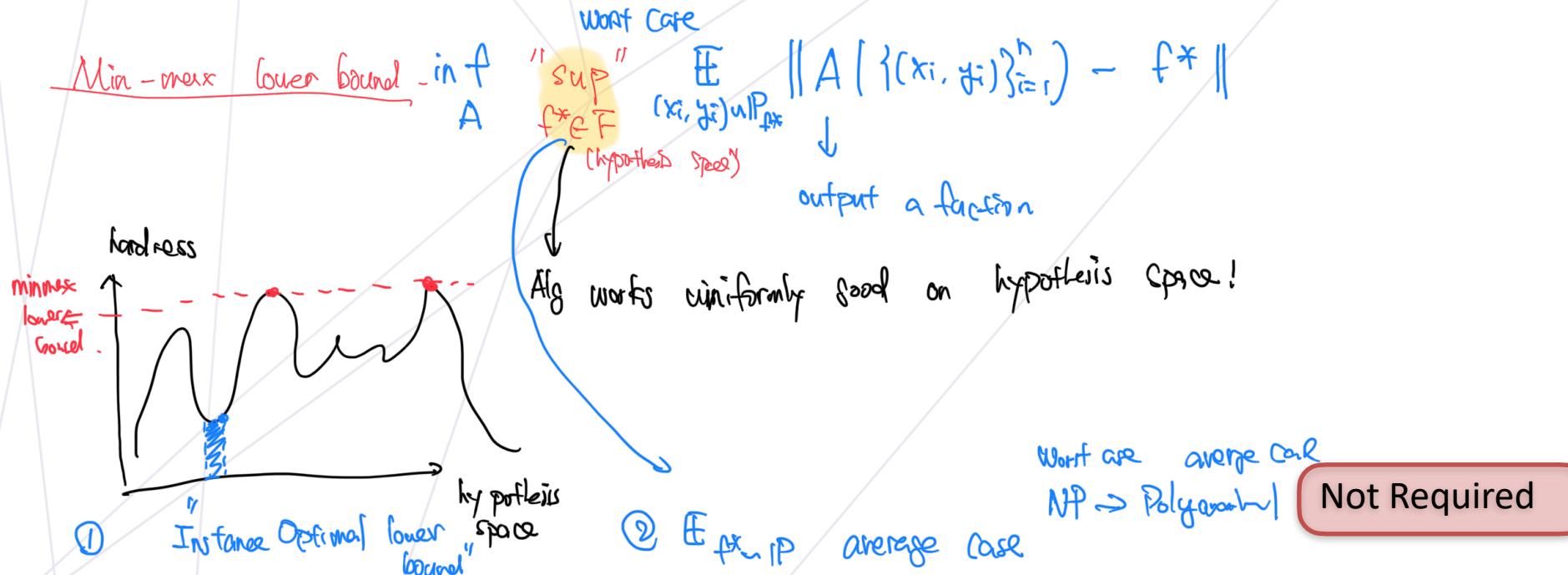
gives out a  $h^d$   
normalizing constant  
change of variable

for some  $c$  since the densities in  $\Sigma(\beta, L)$  are uniformly bounded. The result follows.  $\square$

# Why our result is optimal in 1d

<http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf>

Lecture 2.2



# Why our result is optimal in 1d

[https://web.stanford.edu/class/ee378c/lecture7\\_annotated.pdf](https://web.stanford.edu/class/ee378c/lecture7_annotated.pdf)

① reduce "Estimation lower bound" to "Classification lower bound"

$$f_1, f_2 \in \mathcal{F}^*$$

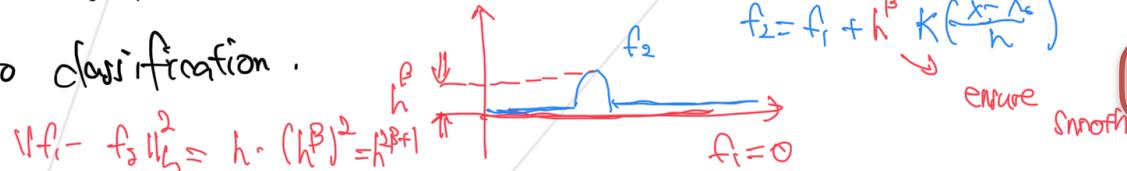
$(\{(x_i, y_i)\}_{i=1}^n)$  dataset of  $n$  data. → ask you:  
which function generate  
the dataset?

If you can't do classification task and  $\|f_1 - f_2\| = \epsilon$

then you can't do estimation better than  $\epsilon/2$

② Consider distribution of  $(\{(x_i, y_i)\}_{i=1}^n)$  too close. smaller than  $o(1)$

⇒ can't do classification.

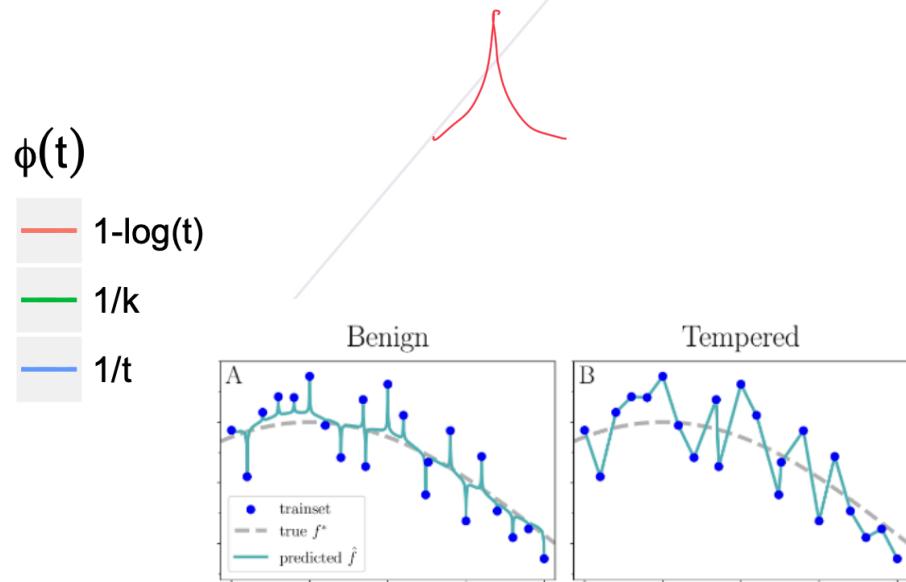
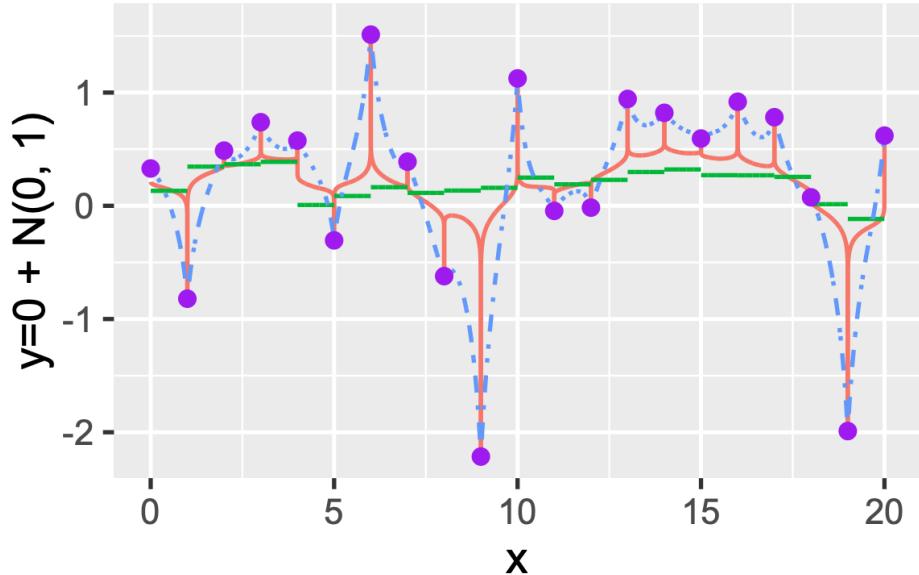


$$f_2 = f_1 + h^B K\left(\frac{x-x_c}{h}\right)$$

ensure smooth

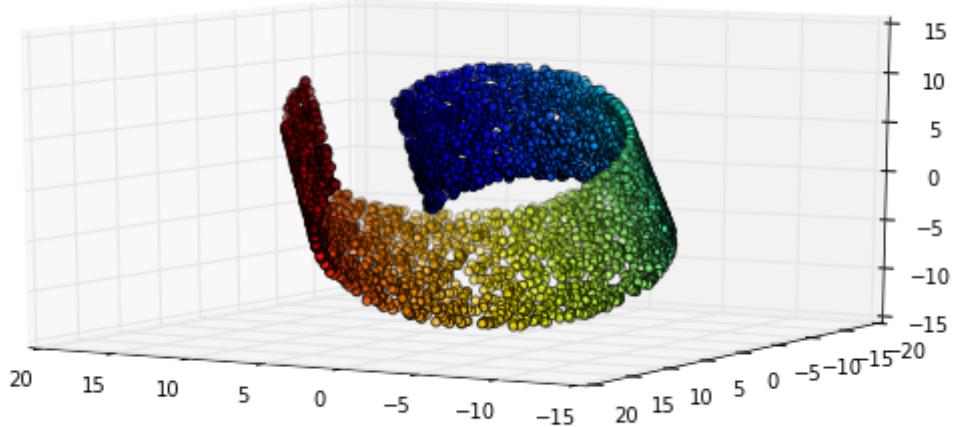
Not Required

# Ok... Interpolation...(1-NN)



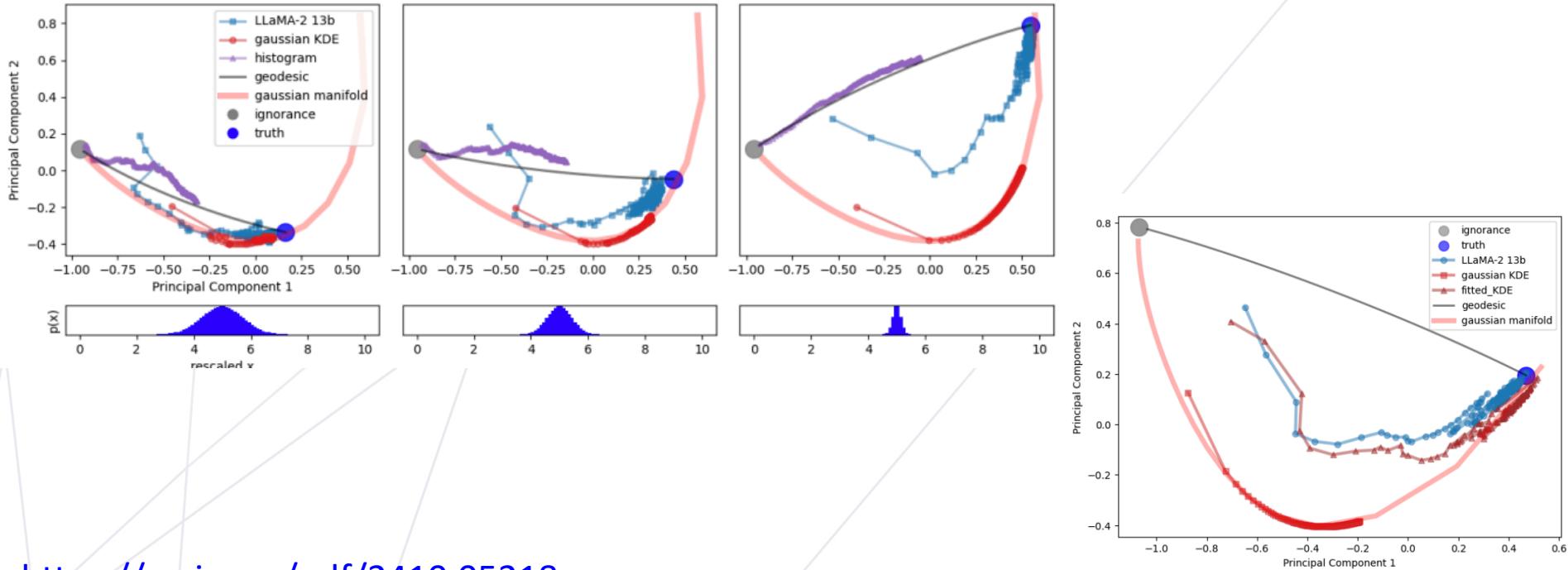
Xing Y, Song Q, Cheng G. Benefit of interpolation in nearest neighbor algorithms. SIAM Journal on Mathematics of Data Science, 2022, 4(2): 935-956.

# Open Questions



Bias computation on manifold: Section 8.1 in <https://arxiv.org/abs/2407.09286>

# LLM learns “Optimized” Kernel



<https://arxiv.org/pdf/2410.05218>

# Delta Methods

<https://web.stanford.edu/class/stats300b/ScribeNotes/2021/lecture-03.pdf>

<https://web.stanford.edu/class/stats300b/ScribeNotes/2021/lecture-04.pdf>

# Aim of asymptotic theory

$$r_n(T_n - \theta) \rightarrow T$$

Estimator using  $n$  data

$r_n \rightarrow \infty$  is deterministic

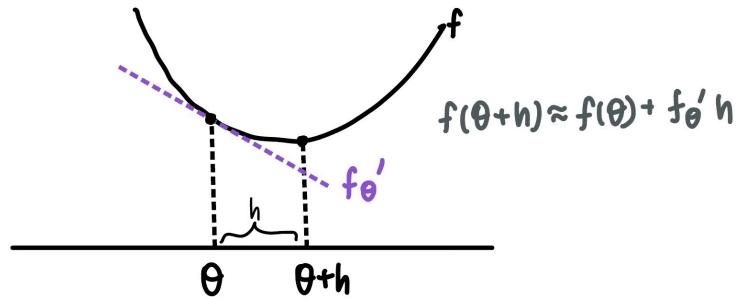
Asymptotic distribution

# Delta Methods

from central limit theorem we know  $r_n(T_n - \theta) \rightarrow T$

Question: What is the asymptotic distribution of  $\Phi(T_n)$

Idea: Taylor Expansion



# Delta method

**Thm** If  $r_n(T_n - \theta) \rightarrow T$ , then  $r_n(\Phi(T_n) - \Phi(\theta)) \rightarrow \underline{\phi'(\theta)T}$

Jacobian Matrix  $[\Phi'(\theta)]_{ij} = \frac{\partial \phi_i(\theta)}{\partial \theta_j}$

$$T \stackrel{\text{def}}{=} N(0, \Sigma)$$

$$\Sigma \stackrel{\text{def}}{=} E[X^T X]$$

Homework 4!

# Example

Example (The delta method for quadratics)

Assume  $X_i \stackrel{\text{iid}}{\sim} P$  with  $\mathbb{E}[X] = \theta \neq 0$ ,  $\text{Cov}(X) = \Sigma$ , and set  $\phi(h) = \frac{1}{2} \|h\|_2^2$ . Then

$$\sqrt{n} \left( \frac{1}{2} \left\| \frac{1}{n} \sum_{i=1}^n X_i \right\|_2^2 - \frac{1}{2} \|\theta\|_2^2 \right) \xrightarrow{d} \mathcal{N} \left( 0, \theta^T \Sigma \theta \right)$$

# Example

Example (Delta method for sample variance)

For  $X_i$  i.i.d. with  $\text{Var}(X_i) = \sigma^2$  and  $\mathbb{E}[X_i^4] < \infty$ , let

$$S_n^2 := \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2.$$

Then for  $\phi(x, y) = y - x^2$  we have  $S_n^2 = \phi(\bar{X}_n, \bar{X}_n^2)$ , and

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[X^4] - \mathbb{E}[X^2]^2) \stackrel{\text{dist}}{=} \mathcal{N}(0, \text{Var}(X^2)).$$

# Higher-Order Delta Method

What happens if  $\phi'(\theta) = 0$ ?

$$\begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$
$$= a_{11}x^2 + (a_{12} + a_{21})xy + a_{22}y^2$$
$$r_n^2(\Phi(T_n) - \Phi(\theta)) \rightarrow \frac{1}{2} \underbrace{T^T \nabla^2 \Phi(\theta) T}_{\text{second term in Taylor Expansion}}$$

$$\sqrt{n}(X_n - 0) \rightarrow N(0, 1)$$

$$\phi(x) = x^2 \Rightarrow \phi'(0) = 0$$

$$\sqrt{n}(X_n^2 - 0) \rightarrow 0$$

$$n(X_n^2 - 0) \rightarrow \mathcal{X}(0, 2)$$

This tells me that

If  $X_n \rightarrow 0$  at  $O(\frac{1}{\sqrt{n}})$

$X_n^2 \rightarrow 0$  at  $O(\frac{1}{n})$

# Example

recall KL-divergence between distributions

$$D_{\text{KL}}(P\|Q) := \int dP \log \frac{dP}{dQ} = \int p \log \frac{p}{q} d\mu$$

## Example

Let  $X_i \in \{0, 1\}$ ,  $X_i \sim P_\theta := \text{Bernoulli}(\theta)$  (i.e.  $\mathbb{E}[X_i] = \theta$ ). For  $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$ ,

$$nD_{\text{KL}}(P_{\hat{\theta}_n}\|P_\theta) \xrightarrow{d} \frac{1}{2} W^2 \quad \text{and} \quad nD_{\text{KL}}(P_\theta\|P_{\hat{\theta}_n}) \xrightarrow{d} \frac{1}{2} W^2$$

for  $W \sim \mathcal{N}(0, 1)$

# Asymptotic Normality

# Asymptotic Theory for ERM?

what is the asymptotic distribution of  $\hat{\theta}_n := \arg \min \mathbb{E}_{P_n} l_\theta(x)$

**For example:** maximum likelihood  $l_\theta(x) := \log P_\theta(x)$

**Today's AIM:**  $\sqrt{n}(\hat{\theta}_n - \theta^*) \rightarrow N(0, e'(\theta^*)^{-1} e' \mathbb{E}_{P_{\theta^*}} (\nabla l \nabla l^\top) \theta^*)^{-1}$  where  $e(\theta) = \mathbb{E}_{P_{\theta^*}} \nabla^2 l_\theta$

# Asymptotic theory

## Theorem

Let  $X_i \stackrel{\text{iid}}{\sim} P_{\theta_0}$  and assume  $\hat{\theta}_n = \operatorname{argmax}_{\theta} P_n \ell_{\theta}(X)$  is consistent.

Define the covariance

$$\Sigma_{\theta} := (P_{\theta} \nabla^2 \ell_{\theta}(X))^{-1} \operatorname{Cov}_{\theta}(\nabla \ell_{\theta}(X)) (P_{\theta} \nabla^2 \ell_{\theta}(X))^{-1}$$

Under the previous assumptions,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\theta_0})$$

- ▶ “typically”  $\Sigma_{\theta} = -(P_{\theta} \nabla^2 \ell_{\theta}(X))^{-1} = \operatorname{Cov}_{\theta}(\dot{\ell}_{\theta})$

# Proof

# Bias-variance trade-off in Asymptotic?

Not Required

Duchi J, Ruan F. Asymptotic optimality in stochastic optimization. arXiv preprint  
arXiv:1612.05612, 2016.

# Moment Estimator

if we know  $e(\theta) = \mathbb{E}_{X \sim P_\theta}[F(X)]$ , we define  $e(\hat{\theta}_n) = \mathbb{E}_{\mathbb{P}_n} f(X)$

# Inverse Function Theorem

$$(F^{-1})'(t) = \frac{\partial}{\partial t} F^{-1}(t) = (F'(F^{-1}(t)))^{-1}.$$

# Hints for future research

$f(\theta) = \arg \min_f F_\theta(f)$ , What is  $f'(\theta)$ ?

Not Required

# Exponential Family

**Definition 3.1.**  $\{P_\theta\}_{\theta \in \Theta}$  is a regular exponential family if there is a sufficient statistic  $T : \mathcal{X} \rightarrow \mathbb{R}^d$  such that  $P_\theta$  has density

$$P_\theta = \exp(\theta^T T(x) - A(\theta))$$

with respect to  $\mu$ , where  $A(\theta) = \log \int e^{\theta^T T(x)} d\mu(x)$ .

**Fact:** Moment estimator for exp family using moment  $T$  equals to ERM estimator