Lecture 6 Fisher Information

IEMS 402 Statistical Learning

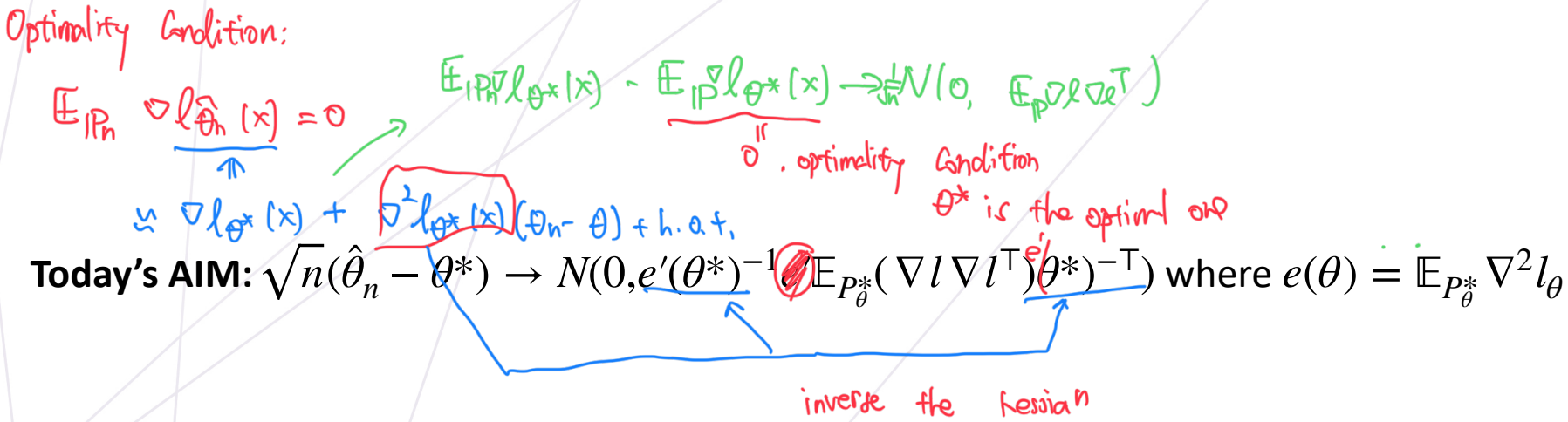
Northwestern

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: $\sqrt{n}(\hat{\theta}_n - \theta^*) \to N(0, e'(\theta^*)^{-1} \mathbb{E}_{P^*_\theta}(\nabla l \nabla l^T) e'(\theta^*)^{-T})$ where $e(\theta) = \mathbb{E}_{P^*_\theta} \nabla^2 l_\theta$

Asymptotic theory

Theorem

Let $X_i \overset{\text{iid}}{\sim} P_{\theta_0}$ *and assume* $\widehat{\theta}_n = \text{argmax}_\theta \, P_n \ell_\theta(X)$ *is consistent. Define the covariance*

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

Under the previous assumptions,

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

- "typically" $\Sigma_\theta = -(P_\theta \nabla^2 \ell_\theta(X))^{-1} = \text{Cov}_\theta(\dot{\ell}_\theta)$

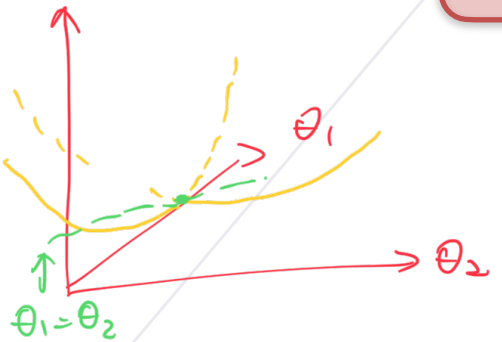
Proof

Bias-variance trade-off in Asymptotic?

Not Required

$\theta_1 = \theta_2$

- $Y = \theta_1 X + \theta_2 X$ learn θ_1, θ_2
- $Y = \theta_1 X + \theta_1 X$ learn θ_1



Var = Var of Estimator
project to $\theta_1 = \theta_2$.

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 $T : \mathcal{X} \to \mathbb{R}^d$ such that P_θ has density*

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

with respect to μ, 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

Fisher Information

Asymptotic Theory for Max like-lihood

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

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

Thm. $\sqrt{n}(\theta_n - \theta_*) \to N(0, I_{\theta*}^{-1})$ $I_{\theta*}$ is the Fisher Information

Claim. $\mathbb{E}_{P_\theta}[\nabla^2 l_\theta] = -\mathbb{E}[\nabla l_\theta \nabla l_\theta^T]$ $\nabla l_\theta = \frac{\nabla P}{P}$

$\frac{\nabla^2 P_\theta}{P_\theta} - \frac{\nabla P \nabla P^T}{P_\theta^2}$

Today's AIM: $\sqrt{n}(\hat{\theta}_n - \theta^*) \to N(0, e'(\theta^*)^{-1} e' \mathbb{E}_{P_{\hat{\theta}}^*}(\nabla l \nabla l^T)\theta^*)^{-T})$ where $e(\theta) = \mathbb{E}_{P_{\hat{\theta}}^*} \nabla^2 l_\theta$

$\mathbb{E}\frac{\nabla^2 P_\theta}{P_\theta} = \int \nabla^2 P_\theta = \nabla^2 \int P_\theta = 0$

Fisher Information

Definition (Fisher information)

For a model family $\{P_\theta\}$ on $\mathcal{X}$, the *Fisher information* is

$$I(\theta) := \mathbb{E}_\theta[\nabla \ell_\theta(X) \nabla \ell_\theta(X)^\top]$$

- when $\mathbb{E}$ and ∇ are interchangable, then $I(\theta) = -\mathbb{E}[\nabla^2 \ell_\theta(X)]$

Claim. $\mathbb{E}\, \nabla \ell(\theta) = 0$ ① optimality condition of max-likelihood

$$\mathbb{E}_P[\nabla \log p] = \mathbb{E}_P\left[\frac{\nabla p}{p}\right] = \int \nabla p = \nabla \int p = \nabla 1 = 0$$

$\nabla \ln(x) = \frac{1}{x}$ is a key structure!!

Score function!

$$\nabla \ell_\theta(x) = \left[\frac{\partial}{\partial \theta_j} \log p_\theta(x)\right]_{j=1}^{d} \in \mathbb{R}^d$$

$$\nabla^2 \ell_\theta(x) = \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_\theta(x)\right]_{i,j=1}^{d} \in \mathbb{R}^{d \times d},$$

Cramér–Rao lower bound

Thm. If $\left[\mathbb{E}\,\hat{\theta}_n\right] = \theta^*$ ⇐ unbias.

$$\mathrm{Var}(\hat{\theta}_n) \geq I_{\theta^*}^{-1}$$

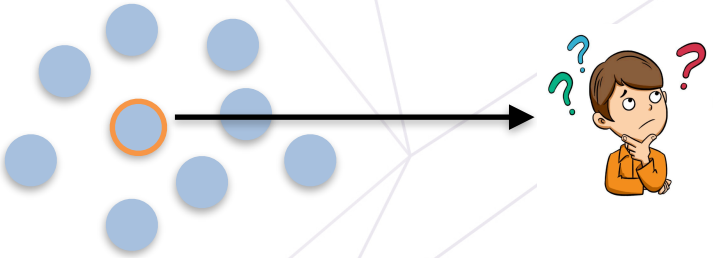
Pf. $\mathbb{E}\left[\hat{\theta}\nabla \log p(\theta)\right] = 1$

$$\mathbb{E}\int \hat{\theta}\frac{\nabla p}{p} = \int \hat{\theta}\nabla p = \nabla\theta^* = I.$$

then use $\mathrm{Cov}(X,Y) \geq \mathrm{Var}(X)\mathrm{Var}(Y)$

Influence Function

influence function

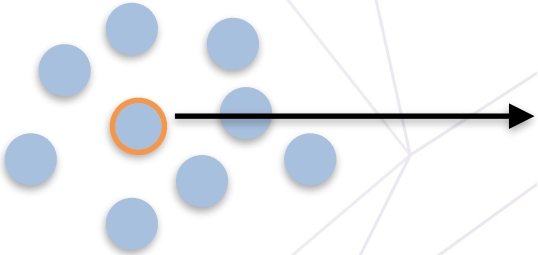


What is the influence that we delete the data?

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \left(\sum_{i=1}^{n} l(x_i, y_i; \theta) + l(x_n, y_n; \theta) \right)$$

$$\hat{\theta}_{-} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} l(x_i, y_i; \theta)$$

influence function



What is the influence that we delete the data?

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \left(\sum_{i=1}^{n} l(x_i, y_i; \theta) + l(x_n, y_n; \theta) \right)$$

$$\hat{\theta}_{-} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} l(x_i, y_i; \theta)$$

$$\hat{\theta}_{\epsilon} = \arg\min_{\theta} \frac{1}{n} \left(\sum_{i=1}^{n} l(x_i, y_i; \theta) + \epsilon l(x_n, y_n; \theta) \right)$$

Influence function: $\frac{d\hat{\theta}_{\epsilon}}{d\epsilon}$

How to compute that?

Influence function

$$\hat{\theta}_\epsilon = \arg\min_\theta \frac{1}{n}\left(\sum_{i=1}^{n} l(x_i, y_i; \theta) + \epsilon l(x_n, y_n; \theta)\right)$$

AIM: $\frac{d\hat{\theta}_\epsilon}{d\epsilon}$

$$\left(\sum_{i=1}^{n} \nabla_\theta l(x_i, y_i; \hat{\theta}_\epsilon) + \epsilon \nabla_\theta l(x_n, y_n; \hat{\theta}_\epsilon)\right) = 0$$

Influence function

$$\hat{\theta}_\epsilon = \arg\min_\theta \frac{1}{n}\left(\sum_{i=1}^{n} l(x_i, y_i; \theta) + \epsilon l(x_n, y_n; \theta)\right)$$

AIM: $\frac{d\hat{\theta}_\epsilon}{d\epsilon}$

$$\left(\sum_{i=1}^{n} \nabla_\theta l(x_i, y_i; \hat{\theta}_\epsilon) + \epsilon \nabla_\theta l(x_n, y_n; \hat{\theta}_\epsilon)\right) = 0$$

$$\mathcal{I}_{\text{up,params}}(z) \stackrel{\text{def}}{=} \left.\frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta}),$$

Hessian of all data

Gradient of the data of interest

Take gradient respect to ϵ

How to compute this?

$$\sum_{i=1}^{n} H_\theta l(x_i, y_i; \hat{\theta}_\epsilon)\frac{d\hat{\theta}_\epsilon}{d\epsilon} + \epsilon H_\theta l(x_n, y_n; \hat{\theta}_\epsilon)\frac{d\hat{\theta}_\epsilon}{d\epsilon} + \nabla_\theta l(x_n, y_n; \hat{\theta}_\epsilon) = 0$$

Applications

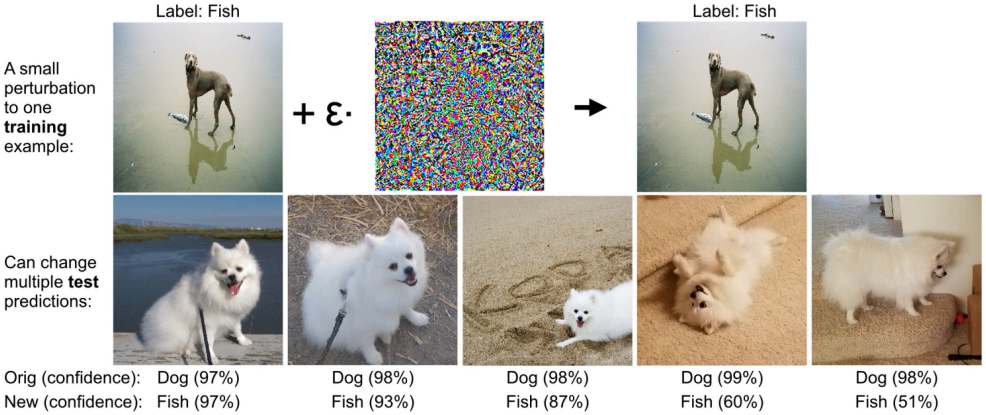
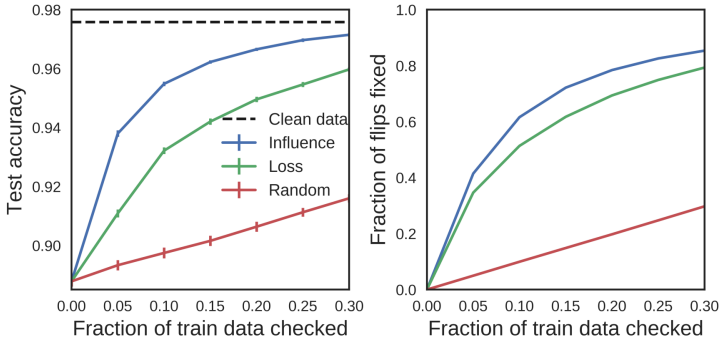


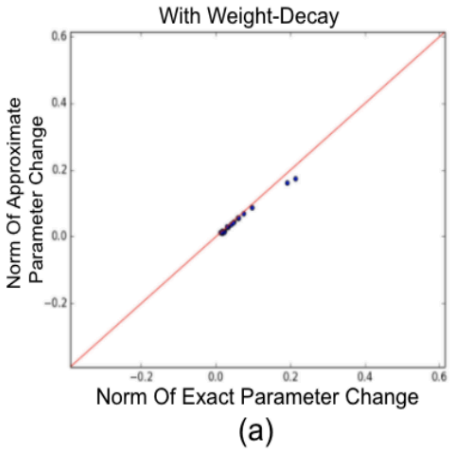
Figure 5. **Training-set attacks.** We targeted a set of 30 test images featuring the first author's dog in a variety of poses and backgrounds. By maximizing the average loss over these 30 images, we found a visually-imperceptible change to the particular training image (shown on top) that flipped [illegible] test imag[illegible]

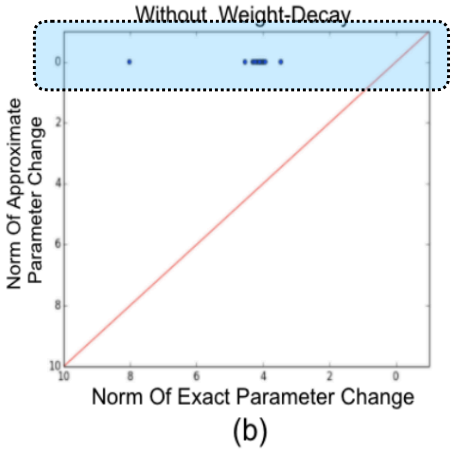


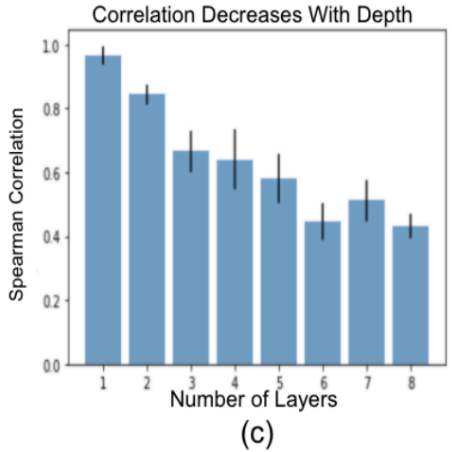
https://arxiv.org/pdf/1703.04730

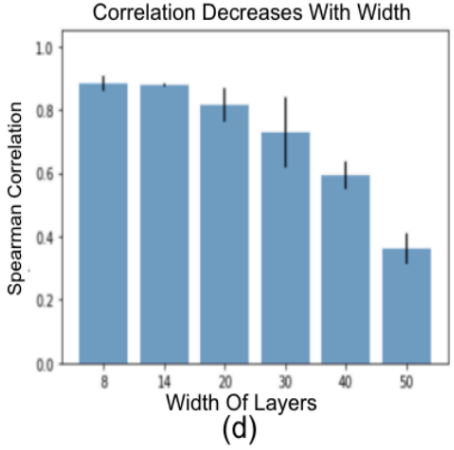
Checking mislabeled data

However







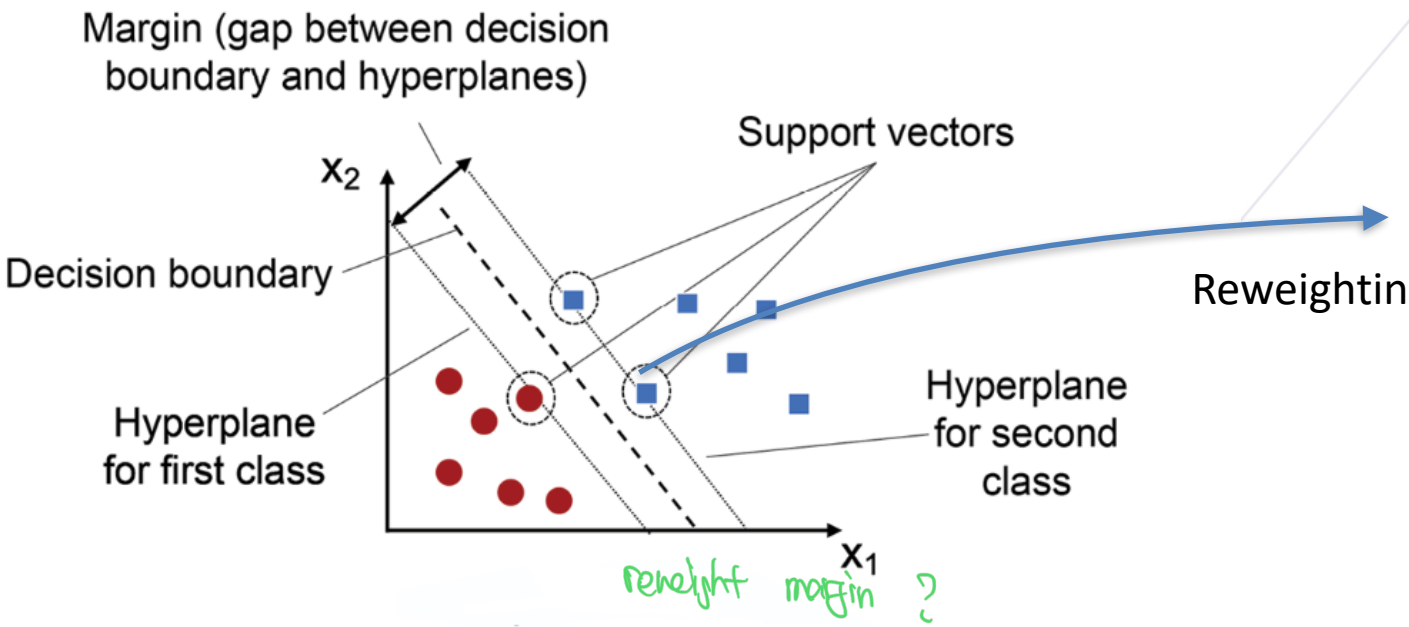


https://arxiv.org/pdf/2006.14651

Overparameterize: SVM example

Conjecture: IF = 0 for overparameterized models

Open!



Reweighting no longer works!