**IEMS 402: Statistical Learning** **Winter 2024-2025**

# Lecture 6: Fisher Information

*Lecturer: Yiping Lu* *Scribes: Heyuan Yao*



## 6.1 Fisher Information

**Definition 6.1** *If $\{P_\theta\}_{\theta\in\Theta}$ is a family of models, the fisher information matrix of $P_\theta$ is*

$$I_\theta = \mathop{\mathbf{E}}_{P_\theta}[\nabla_\theta \ell_\theta (\nabla_\theta \ell_\theta)^T]. \tag{6.1}$$

**Theorem 6.2** *For models which are "nice", meaning that we can switch the order of differentiation with respect to $\theta$, and expectation with respect to $x$, then it holds that*

$$I_\theta = \mathrm{Cov}(\nabla \ell_\theta) = -\mathbf{E}[\nabla^2 \ell_\theta]. \tag{6.2}$$

**Proof:** We begin by showing that the expectation of the score function is constant.

$$\mathop{\mathbf{E}}_{P_\theta}[\nabla \ell_\theta] = \int \frac{\nabla p_\theta}{p_\theta} p_\theta = \nabla \int p_\theta = \nabla 1 = 0, \tag{6.3}$$

using that $\nabla \ell_\theta = \frac{\nabla p_\theta}{p_\theta}$. Again moving the derivatives out of the integral, we also have that

$$\mathop{\mathbf{E}}_{P_\theta}[\frac{\nabla^2 p_\theta}{p_\theta}] = \nabla^2 \int p_\theta = \nabla^2 1 = 0. \tag{6.4}$$

These equations imply that

$$\mathbf{E}[\nabla^2 \ell_\theta] = \mathbf{E}[\frac{\nabla^2 p_\theta}{p_\theta} - \frac{\nabla p_\theta \nabla p_\theta^T}{p_\theta^2}] = -\mathbf{E}[\nabla \ell_\theta \nabla \ell_\theta^T] = -I_\theta \tag{6.5}$$

Using that the expected value of the score function is zero, we get that

$$\mathbf{Cov}(\nabla \ell_\theta) = \mathbf{E}[(\nabla \ell_\theta - \mathbf{E}[\nabla \ell_\theta])(\nabla \ell_\theta - \mathbf{E}[\nabla \ell_\theta])^T] = \mathbf{E}[\nabla \ell_\theta (\nabla \ell_\theta)^T] = I_\theta. \tag{6.6}$$

■

## 6.2 Variance of MLE

If $\{P_\theta\}_{\theta\in\Theta}$ is "nice", then we showed last lecture that

$$\sqrt{n}(\hat{\theta}_n^{MLE} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma_*), \tag{6.7}$$

where $\Sigma_* = (P\nabla^2\ell_{\theta^*})^{-1}(\mathrm{Cov}(\nabla\ell_{\theta^*}))(P\nabla^2\ell_{\theta^*})^{-1}$. With the same regularity conditions as in the previous theorem we thus get that $\Sigma_* = (P\nabla^2\ell_{\theta^*})^{-1}(\mathrm{Cov}(\nabla\ell_{\theta^*}))(P\nabla^2\ell_{\theta^*})^{-1} = (-I_{\theta^*})^{-1}I_{\theta^*}(-I_{\theta^*})^{-1} = (I_{\theta^*})^{-1}$.

Thus, under suitable regularity conditions, the MLE converges to a normal distribution with variance equal to the inverse Fisher information. Hence a "large" fisher information implies a small asymptotic variance, and a "small" fisher information implies a large asymptotic variance. One way to understand this for matrices is through the PSD ordering of matrices. We say that $A \succeq B$ if $A - B$ is PSD, this is the PSD ordering of matrices. Another way is to pick a vector w and look at the limiting distribution along w:

$$\sqrt{n}(\langle w, \hat{\theta}_n^{MLE}\rangle - \langle w, \theta^*\rangle) \xrightarrow{d} \mathcal{N}(0, w^T I_{\theta^*}^{-1} w) \tag{6.8}$$

## 6.3 Intuition and Identities for FI

To see the intuition behind FI, we start with a simple example.

**Example 1** *Consider normal distribution $\mathcal{N}(\theta, \sigma^2)$ with $\sigma^2$ known. Then one can show that $I_\theta = \frac{1}{\sigma^2}$.*

Intuitively, FI gives us a way to measure the amount of information that a random variable contains about some unknown parameter $\theta$ of the random variable's assumed probability distribution. In Example 1, A smaller $\sigma$ (less variance) means the data points are more concentrated around the mean $\theta$. This concentration makes it easier to estimate $\theta$ precisely since the observations are less spread out. Therefore, less variance (smaller $\sigma$) implies more information about $\theta$, hence a larger Fisher Information value. Therefore, FI intuitively conveys that as our certainty (or precision) in the data increases (i.e., lower variance), our ability to accurately estimate $\theta$ also increases.

**Identity 6.3** *If $X, Y$ are independent random variables with joint density $\mathbb{P}_\theta(x, y)$, then $I_{xy}(\theta) = I_x(\theta) + I_y(\theta)$.*

**Proof:** Since $X, Y$ are independent, $\mathbb{P}_\theta(x, y) = f_\theta(x)g_\theta(y)$, where $f_\theta(x), g_\theta(y)$ denote the probability density functions of $X, Y$ respectively. It follows that

$$\nabla \log \mathbb{P}_\theta = \nabla \log f_\theta + \nabla \log g_\theta$$

Taking covariance of the LHS, we have the following identity:

$$\mathrm{Cov}(\nabla \log \mathbb{P}_\theta) = \mathrm{Cov}(\nabla \log f_\theta) + \mathrm{Cov}(\nabla \log g_\theta)$$

Using definition of Fisher Information completes the proof. ■

**Corollary 6.4** *If $X_1, ..., X_n$ are i.i.d. samples from distribution $\mathbb{P}_\theta$, then $I_{X_1,...,X_n}(\theta) = nI(\theta)$.*

**Identity 6.5** *$I_{f(\theta)} = \frac{I_\theta}{(f'(\theta))^2}$ (be smart if $f$ is not invertible!)*

This reflects the intuition that if $f$ is really flat around $\theta$, then getting an accurate approximation to $f(\theta)$ should be easier (require less data) than approximating $\theta$.

**Proof:** Observe that

$$\frac{\partial \ell_\theta}{\partial f(\theta)} = \frac{\partial \ell_\theta}{\partial \theta}\frac{\partial \theta}{\partial f(\theta)} = \frac{\partial \ell_\theta}{\partial \theta}\frac{1}{f'(\theta)}.$$

It follows that

$$I_{f(\theta)} = E[(\frac{\partial \ell_\theta}{\partial f(\theta)})(\frac{\partial \ell_\theta}{\partial f(\theta)})^T] = \frac{1}{(f'(\theta))^2} E[(\frac{\partial \ell_\theta}{\partial \theta})(\frac{\partial \ell_\theta}{\partial \theta})^T] = \frac{I_\theta}{(f'(\theta))^2}.$$

Tselil: the statement of the lemma is only valid for scalar-valued $\theta$ and $f(\theta)$; a more general statement applies but then one has to use the Jacobian matrix $f'$. ■

The above leads us to observe that the Fisher Information is sensitive to the parameterization of the model, which can lead to counterintuitive behavior. For example:

**Example 2** *Given normal distribution $\mathcal{N}(\mu, \sigma^2)$, where $\sigma$ is known and $\mu$ is unknown, we consider function $f(\mu) = \sqrt{\mu}$. Then from Identity 6.5 we have that $I_{\sqrt{\mu}} = \frac{4\mu}{\sigma^2}$.*

Example 2 is "weird" because in Example 1, the amount of information that $\mathcal{N}(\mu, \sigma^2)$ carries did not depend on $\mu$ (though this also makes sense, in light of the fact that $f(x) = \sqrt{x}$ changes quickly near $x = 0$ and therefore should be harder to approximate within e.g. multiplicative factors for small $\mu$). As we shall soon see, this affects the downstream constructed confidence intervals, etc.

## 6.4 Optimality of MLE

**Theorem 6.6 (Cramér-Rao Bound)** *If $\delta_n$ is any unbiased estimator of $\theta$, $E[(\theta - \delta_n)(\theta - \delta_n)^T] \succeq \frac{1}{n} I_\theta^{-1}$.*

**Proof:** Suppose scalar function $t_n := X_1, ..., X_n \to \mathbb{R}$ is such that $g(\theta) = E[t_n]$ for some function $g(\cdot)$ and parameter $\theta$. We will show that the variance of $t_n$ is lower bounded by $\frac{1}{n}(\nabla g)^T I_\theta^{-1}(\nabla g)$, or equivalently:

$$\mathrm{Var}(t_n) \geqslant \frac{1}{n}(\nabla g)^T I_\theta^{-1}(\nabla g). \tag{6.9}$$

First, to see the sufficiency of inequality 6.9 to the conclusion, observe that for any $u \in \mathbb{R}^d$, we choose $t_n = \langle u, \delta_n \rangle, \nabla g = u$. Together, inequality 6.9 can be re-written as

$$\mathrm{Var}(t_n) = u^T (E[(\delta_n - \theta)(\delta_n - \theta)^T])u \geqslant u^T (\frac{1}{n} I_\theta^{-1})u \tag{6.10}$$

Recall that $A \preceq B$ if and only if for all $u \in \mathbb{R}^d$, $u^T A u \leqslant u^T B u$. This concludes our proof of sufficiency of inequality 6.9.

It remains to show that inequality 6.9 holds. To this end, fix any $w \in \mathbb{R}^d$. By Cauchy-Schwarz inequality, the following holds:

$$\mathrm{Var}(t_n) \geqslant \frac{\mathrm{Cov}^2(t_n, \langle w, \nabla \ell_{\theta^n} \rangle)}{\mathrm{Var}(\langle w, \nabla \ell_{\theta^n} \rangle)} \tag{6.11}$$

Here, the denominator

$$\mathrm{Var}(\langle w, \nabla \ell_{\theta^n} \rangle) = w^T \mathrm{Cov}(\nabla \ell_{\theta^n}) w = w^T (nI(\theta)) w.$$

Now notice that

$$\mathrm{Cov}(t_n, \langle w, \nabla \ell_{\theta^n} \rangle) = E[(t_n - g) \sum_i (\nabla (\ell_{\theta^n})_i w_i)] \tag{6.12}$$

$$= \langle w, E[(t_n - g) \nabla \ell_{\theta^n}] \rangle \tag{6.13}$$

Furthermore,

$$E[(t_n - g)\nabla \ell_{\theta^n}] = E[t_n \nabla \ell_{\theta^n}] - gE[\nabla \ell_{\theta^n}] \tag{6.14}$$

$$= \int t_n \frac{\nabla \mathbb{P}_{\theta^n}}{\mathbb{P}_{\theta^n}} \mathbb{P}_{\theta^n} d\mu - 0 \tag{6.15}$$

$$= \nabla \int t_n \mathbb{P}_{\theta^n} d\mu \tag{6.16}$$

$$= \nabla g(\theta). \tag{6.17}$$

Substituting equation 6.17 into equation 6.13, then for all $w \in \mathbb{R}^d$, we can rewrite inequality 6.11 as the following:

$$\mathrm{Var}(t_n) \geqslant \frac{\langle w, \nabla g \rangle^2}{n w^T I_\theta w}. \tag{6.18}$$

In particular, choose $w = I_\theta^{-1} \nabla g$. Then inequality 6.18 becomes

$$\mathrm{Var}(t_n) \geqslant \frac{(\nabla g^T I_\theta^{-1} \nabla g)^2}{n(\nabla g)^T I_\theta^{-1} I_\theta I_\theta^{-1} \nabla g} = \frac{(\nabla g)^T I_\theta^{-1} (\nabla g)}{n}. \tag{6.19}$$

■

A useful corollary derived from Theorem 6.6 is the following.

**Corollary 6.7** *If $\delta_n$ is any unbiased estimator of $\theta$, then*

$$E[\|\theta - \delta_n\|^2] \geqslant \frac{1}{n} \mathrm{Tr}(I_\theta^{-1}).$$

The Cramér-Rao Bound (CRB) provides a fundamental limit on the variance of unbiased estimators. Generally, an estimator is said to be "efficient" if it reaches the lower bound set by the Cramér-Rao Bound. If an estimator achieves this minimum variance, it's considered to be the best unbiased estimator in terms of variance. MLE is particularly important in this context because under certain regularity conditions, it is asymptotically efficient. This means that as the sample size increases, the variance of the MLE approaches the Cramér-Rao Lower Bound. In other words, for large samples, the MLE achieves the lowest possible variance among all unbiased estimators, making it the most efficient estimator in this limit. More on this topic will be discussed in the next lectures.

## 6.5 Influnce Function

This section refers to and follows the notion setting of the work by Koh & Liang (2017). You can find more detailed information in `https://proceedings.mlr.press/v70/koh17a/koh17a.pdf`.

Consider a prediction problem from some input space $\mathcal{X}$ (e.g., images) to an output space $\mathcal{Y}$ (e.g., labels). We are given training points $z_1, \ldots, z_n$, where $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. For a point $z$ and parameters $\theta \in \Theta$, let $L(z, \theta)$ be the loss, and let $\frac{1}{n} \sum_{i=1}^n L(z_i, \theta)$ be the empirical risk. The empirical risk minimizer is given by $\hat{\theta} := \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n L(z_i, \theta)$.We assume that the empirical risk is twice-differentiable and strictly convex in $\theta$. And remember that in our class, **we simply let $L$ be the log-likelihood $l$.**

Our goal is to understand the effect of training points on a model's predictions. We formalize this goal by asking the counterfactual: **how would the model's predictions change if we did not have this training point, or if we reweight this point?**

Let us begin by studying the change in model parameters due to removing a point $z$ from the training set. Formally, this change is $\hat{\theta}_{-z} - \hat{\theta}$, where

$$\hat{\theta}_{-z} := \arg\min_{\theta \in \Theta} \sum_{z_i \neq z} L(z_i, \theta).$$

However, retraining the model for each removed $z$ is prohibitively slow.

Fortunately, influence functions give us an efficient approximation. The idea is to compute the parameter change if $z$ were **upweighted** by some small $\epsilon$. From now w.o.l.g. we let $z = z_n = (x_n, y_n)$, and it gives us new parameters

$$\hat{\theta}_\epsilon := \arg\min_{\theta \in \Theta} \left( \frac{1}{n} \sum_{i=1}^{n} L(z_i, \theta) + \epsilon L(z_n, \theta). \right)$$

To approximate $\hat{\theta}_\epsilon$, an idea is to calculate the **influence function** $\frac{d\hat{\theta}_\epsilon}{d\epsilon}$. From our optimization course we note $\hat{\theta}_\epsilon$ satisfies the first-order condition

$$\left( \sum_{i=1}^{n} \nabla_\theta L(z_i, \hat{\theta}_\epsilon) + \epsilon \nabla_\theta L(z_n, \hat{\theta}_\epsilon) \right) = 0,$$

for all fixed $\theta$. By taking gradient with respect to $\epsilon$, the above equation then implies, with $H_\theta L(z_i, \theta) := \nabla^2_\theta L(z_i, \theta)$ denoting the Hessian of $L$,

$$\sum_{i=1}^{n} H_\theta L(z_i, \hat{\theta}_\epsilon) \frac{d\hat{\theta}_\epsilon}{d\epsilon} + \epsilon H_\theta L(z_n, \hat{\theta}_\epsilon) \frac{d\hat{\theta}_\epsilon}{d\epsilon} + \nabla_\theta L(z_n, \hat{\theta}_\epsilon) = 0.$$

Although this expression is not easy to compute, we note that at least when $\epsilon = 0$, the influence of upweighting $z = z_n$ on the parameters $\hat{\theta}$ is given by the following closed form

$$\mathcal{I}_{\text{up,params}}(z) := \frac{d\hat{\theta}_\epsilon}{d\epsilon}|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \, \nabla_\theta L(z, \hat{\theta}), \tag{6.20}$$

where $H_{\hat{\theta}} := \frac{1}{n} \sum_{i=1}^{n} H_\theta L(z_i, \hat{\theta})$ is the Hessian and is positive definite (PD) by assumption. And we remember that from our optimization class (450-II), we can compute $H_{\hat{\theta}}^{-1}$ using the **conjugate gradient (CG)** method: Since $H_{\hat{\theta}} \succ 0$ by assumption, $H_{\hat{\theta}}^{-1} v \equiv \arg\min_t \{t^\top H_{\hat{\theta}} t - v^\top t\}$. We can solve this with CG approaches that only require the evaluation of $H_{\hat{\theta}} t$, which takes $O(np)$ time, without explicitly forming $H_{\hat{\theta}}$. While an exact solution takes $p$ CG iterations, in practice we can get a good approximation with fewer iterations.

Since removing a point $z$ is the same as upweighting it by $\epsilon = -\frac{1}{n}$, we can **linearly approximate the parameter change due to removing $z$ without retraining the model** by computing

$$\hat{\theta}_{-z} - \hat{\theta} \approx -\frac{1}{n} \mathcal{I}_{\text{up,params}}(z).$$

The influence function has several interesting applications, where interested readers may find Section 5 in Koh & Liang pretty useful. One example is that if we have the influence function by perturbing a training

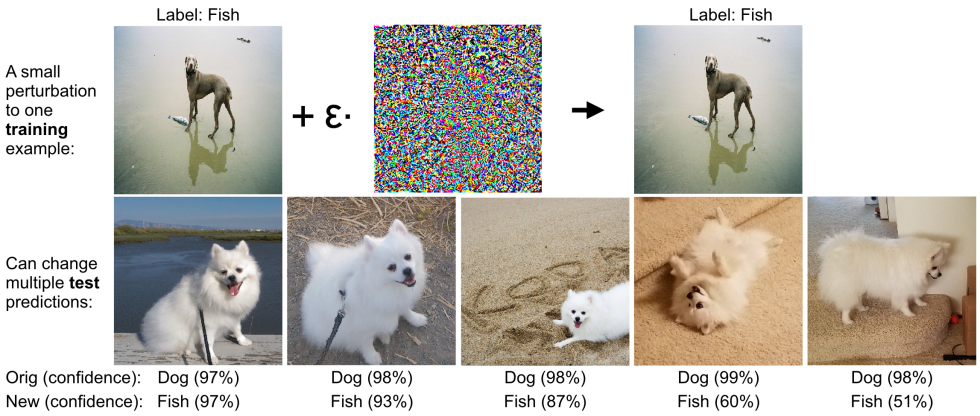


Figure 6.1: Koh and Liang targeted a set of 30 test images featuring Koh's dog in a variety of poses and backgrounds. By maximizing the average loss over these 30 images, they found a visually imperceptible change to the particular training image (shown on top) that flipped predictions on 16 test images.

image input $z_i$[1], we can construct$\tilde{z}_i$, an adversarial version of a $z_i$, by introducing some noise based on the influence function and result a visually indistinguishable from real test images but completely fool a classifier. Figure 6.1 shows that such an indistinguishable training-set attack may flip multiple test predictions.

We also refer interested readers to the work `https://arxiv.org/pdf/2006.14651` by Basu et al (2020), where they suggested that the influence functions are, however, not well understood in the context of deep learning with non-convex loss functions. In particular, they shared the following three findings:

1. influence estimates are fairly accurate for shallow networks, while for deeper networks the estimates are often erroneous;
2. for certain network architectures and datasets, training with weight-decay regularization is important to get high-quality influence estimates;
3. The accuracy of influence estimates can vary significantly depending on the examined test points.

[1]This influence approximates the effect when we perturb one test data $z = (x, y)$ to $z' = (x + \delta, y)$, which is not the influence function $\mathcal{I}_{\text{up,params}}$ we formulated before.