CSE599i: Online and Adaptive Machine Learning **Winter 2018**

Lecture 8: Regression

Lecturer: Kevin Jamieson *Scribes: Behnoosh Parsa, Kuikui Liu*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

To answer many questions in science we need to understand the relationship between variables defined in an experimental setup and the outputs (or observations) of the experiment. Regression analysis has been one of the main methods used for this purpose, which includes many techniques for modeling and analyzing several variables.

1 Introduction

Let's first introduce the problem generally. Typically in science, we encounter some phenomenon that can be described via the following relation: $y = f(x; \theta)$, where the symbols are typically interpreted as follows:

- y is the **observable response** of an experiment; typically represented via some vector in $\mathbb{R}^m$ for some $m \geq 1$
- x are the "conditions under which the experiment is conducted", which are **freely chosen at the experimenter's discretion prior to conducting the experiment**; typically range over some **experimental domain** Ω such as (a subset of) $\mathbb{R}^d$
- θ is the vector of parameters that influence the experiment's outcome but are unknown to the experimenter, and are typically attributed to **nature**; typically range over some unknown **parameter domain** Θ such as (a convex subset of) $\mathbb{R}^{d'}$ but are **not** random
- f is the function that maps the inputs of nature θ and the experimenter's setup x to the results, and describes patterns observed in the phenomenon being studied; typically ranges over some known set of functions $\mathcal{F}$

To capture the noise or stochasticity of the process it is common to introduce a random variable $\epsilon \sim \mathcal{N}(0, \sigma^2)$ into the model: $y = f(x; \theta) + \epsilon$. In this model y has mean $f(x; \theta)$, and the variance σ^2.

There are two common scenarios under which these problems are introduced in the literature.

(1) **Parametric Estimation:** f is parameterized by θ in a way such that knowing θ **completely determines** f. One can think of the linear regression setting, where $f(x) = \theta^\top x$. Here, it suffices to devise methods to estimate θ.

(2) **Nonparametric Estimation:** It is known that f falls in some large class of functions (such as the class of Lipschitz functions) but we may not know the form of f explicitly. Here, learning θ no longer suffices to estimate f everywhere so we will need methods to estimate f directly without knowing θ. In this case, the role of θ will be suppressed.

1.1 The Process of Experimental Design

Experimental Design is the process by which the experimenter chooses points $x_1, \ldots, x_n$ (corresponding to various conditions under which to run experiments) which give responses $y_1, \ldots, y_n$. The goal is to choose the $x_1, \ldots, x_n$ so as to maximize the information gained, which then allows for the most accurate estimate of f. Another way of thinking about experimental design is as a learning process under noise.

Definition 1 (Estimation Strategy)**.** *An **estimation strategy scheme** $\{\hat{f}_i, S_i\}_{i=1}^n$ **for** f is a sequence of pairs $\hat{f}_i, S_i$, referred to as **estimation strategies for** f, where*

- *each S_i specifies a distribution over the domain Ω of the x independently from f, from which we draw the next experiment point x_i, and is known as a **sampling strategy***
- *each $\hat{f}_i$ is some estimator of f that may depend on $x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}$*

***Remark* 1.** We will want to harness randomness when choosing experiment points x_i, which is why we've defined a sampling strategy as a distribution S_i over Ω. Furthermore, this definition encompasses deterministic sampling strategies, which is when S_i places probability 1 at a single point $x_i \in \Omega$.

There are two types of learning one typically considers:

- **Passive Learning:** Each x_i is chosen independently of $x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}$. One can think of the experimenter a priori fixing all $x_1, \ldots, x_n$ testing environments and then afterwards, running all of the experiments.
- **Active Learning:** Each x_i may be drawn from a distribution that depends on $x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}$ (it is adaptive). In particular, S_i may depend on $x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}$. One can think of the experimenter trying one experiment, and then based on the entire history of experiments run so far and their results, decide on what the next most useful experiment to run is.

Typical questions include the following:

- What are the tradeoffs between the following quantities: the number of samples n, the variance of the noise σ^2, the dimensionality d of the input space $\Omega \subset \mathbb{R}^d$, the amount of tolerable error $\eta > 0$, and the probability of exceeding the tolerable error $\delta > 0$?
- How much more power does adaptivity provide you? More precisely, when can active learning surpass the proven limits of passive learning?

1.2 Preliminaries

Here, we gather some necessary preliminary definitions and theorems. For convenience, whenever we write "pdf", we mean a "probabilistic density function".

Definition 2 (Jacobian Matrix)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be a continuously differentiable function. The **Jacobian** $\partial f(x)/\partial x$ **of** f is the $m \times d$ matrix*

$$\left[\frac{\partial}{\partial x_1} f, \ldots, \frac{\partial}{\partial x_d} f\right]$$

where each $\frac{\partial}{\partial x_i} f$ is an m-dimensional vector with each entry j corresponding $\frac{\partial}{\partial x_i} f_j(x)$.

Definition 3 (Kullback-Liebler Divergence)**.** *For two pdfs p, q, we define the **Kullback-Liebler Divergence** (or **relative entropy**) of p w.r.t. q as*

$$KL(p, q) = \mathbb{E}_{y \sim p}\left[\log \frac{p(y)}{q(y)}\right]$$

***Remark* 2.** We will show later in the notes that this quantity is nonnegative and zero only if p, q are equal almost everywhere. However, we note that $KL(p, q)$ is not symmetric in general and does not satisfy the Triangle Inequality. Hence, while it can be thought of as a metric between probability distributions, one should take great care not to use the symmetry and Triangle Inequality axioms of metrics when working with this quantity.

Theorem 1 (Jensen's Inequality)**.** *Let $f : \Omega \to \mathbb{R}$ be a convex function on a convex subset $\Omega \subset \mathbb{R}^n$, and let μ be a probability measure on Ω. Then $f(\mathbb{E}_{x \sim \mu}[x]) \leq \mathbb{E}_{x \sim \mu}[f(x)]$. Furthermore, if f is strictly convex, we have strict inequality, i.e. $f(\mathbb{E}_{x \sim \mu}[x]) < \mathbb{E}_{x \sim \mu}[f(x)]$.*

2 Fisher Information

We begin by studying parametric estimation. Recall that now, estimating θ allows one to estimate $f(x;\theta)$ for all $x \in \Omega$.

Definition 4 (Fisher Information Matrix)**.** *Let X be a random variable with pdf $p(x;\theta)$ that depends on θ and has continuous first-order partial derivatives in x, θ. Then the* ***Fisher Information Matrix*** *$\mathcal{I}_X(\theta)$* ***of*** *θ* ***w.r.t.*** *X is the $d \times d$ matrix with entries given by*

$$\mathcal{I}_X(\theta)_{ij} = \mathbb{E}_{x\sim p(x;\theta)}\left[\left(\frac{\partial}{\partial\theta_i}\log p(x;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log p(x;\theta)\right)\right]$$

As the subscript X can be inferred from the context, we will drop the subscript and write $\mathcal{I}(\theta)$.

***Remark* 3.** The intuition here is that $\mathcal{I}(\theta)$ measures the amount of information X contains about θ (see below). Let's make a few observations about the Fisher Information.

- $\mathcal{I}_X(\theta)$ is symmetric.
- $\mathcal{I}_X(\theta)$ is positive semidefinite, which may be proven via the following observation: if one defines

  $$u(x;\theta) \overset{\text{def}}{=} \nabla_\theta \log p(x;\theta) = \left(\frac{\partial}{\partial\theta_1}\log p(x;\theta), \ldots, \frac{\partial}{\partial\theta_d}\log p(x;\theta)\right)^{\mathrm{T}} \in \mathbb{R}^d$$

  then $\mathcal{I}_X(\theta)$ is precisely $\mathbb{E}_{x\sim p(x;\theta)}[u(x;\theta)u(x;\theta)^\top]$. Thus, if $y \in \mathbb{R}^d$ is any vector, we have by linearity of expectation that

  $$y^\top \mathcal{I}_X(\theta) y = y^\top \mathbb{E}_{x\sim p(x;\theta)}[u(x;\theta)u(x;\theta)^\top]y = \mathbb{E}_{x\sim p(x;\theta)}[\langle u(x;\theta), y\rangle^2] \geq 0$$

- If p has continuous second order partial derivatives, then may in fact rewrite the Fisher Information as

  $$\mathcal{I}_X(\theta)_{ij} = -\mathbb{E}_{x\sim p(x;\theta)}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(x;\theta)\right] = -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathbb{E}_{x\sim p(x;\theta)}\left[\log p(x;\theta)\right]$$

  which is simply the Hessian of what is essentially the **relative entropy/Kullback-Leibler divergence of X w.r.t. θ**. In fact, this connection can be made formal, as we shall see in the following lemma.

We now formalize the relation between the Hessian of the likelihood function $\log p(x;\theta)$ with the Fisher Information. For convenience and cleanliness of notation, we write it out in the 1-dimensional case, where all the main ideas are already captured; the generalization to higher dimensions is obvious (just change the second-order one-dimensional derivative to a partial w.r.t. θ_i and the other to a partial w.r.t. θ_j).

Lemma 1.

$$\mathbb{E}\left[\frac{\partial^2 \log p(x;\theta)}{\partial\theta^2}\right] = -\mathcal{I}(\theta)$$

Proof. First, observe that

$$\frac{\partial \log\left(p(x;\theta)\right)}{\partial\theta} = \frac{1}{p(x;\theta)}\frac{\partial p(x;\theta)}{\partial\theta} \tag{1}$$

Expanding the second order derivative and then applying linearity of expectation, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2 \log\left(p(x;\theta)\right)}{\partial\theta^2}\right] &= \mathbb{E}\left[\frac{\partial}{\partial\theta}\frac{\partial \log\left(p(x;\theta)\right)}{\partial\theta}\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta}\left(\frac{1}{p(x;\theta)}\frac{\partial p(x;\theta)}{\partial\theta}\right)\right] \\
&= \mathbb{E}\left[\frac{-1}{p(x;\theta)^2}\left(\frac{\partial}{\partial\theta}p(x;\theta)\right)^2 + \frac{1}{p(x;\theta)}\frac{\partial^2}{\partial\theta^2}p(x;\theta)\right] \\
&= -\mathbb{E}\left[\left(\frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta}p(x;\theta)\right)^2\right] + \mathbb{E}\left[\frac{1}{p(x;\theta)}\frac{\partial^2}{\partial\theta^2}p(x;\theta)\right]
\end{aligned}$$

Observe the first term on the right-hand side is precisely

$$-\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log p(x;\theta)\right)^2\right] = -\mathcal{I}(\theta)$$

and hence, it suffices to show the second term on the right-hand side is zero. This can be done via the following trick.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{p(x;\theta)}\frac{\partial^2}{\partial\theta^2}p(x;\theta)\right] &= \int\left(\frac{\partial^2}{\partial\theta^2}p(x;\theta)\right)\frac{1}{p(x;\theta)}p(x;\theta)\,dx = \int\left(\frac{\partial^2}{\partial\theta^2}p(x;\theta)\right)\,dx \\
&= \frac{\partial^2}{\partial\theta^2}\int p(x;\theta)\,dx = \frac{\partial^2}{\partial\theta^2}1 = 0
\end{aligned}$$

□

Let us look at some examples for computing the Fisher Information.

Example 1 (Linear Regression with Gaussian Noise)**.** Consider $y_i \sim \mathcal{N}(\theta^\top x_i, \sigma^2)$ and recall that the pdf of y_i is given by $p(y_i \mid x_i, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(y_i - x_i^\top\theta)^2}{2\sigma^2}}$, where the x_i are fixed deterministic vectors in $\mathbb{R}^d$.

$$\frac{\partial \log\left(p(y_i \mid x_i,\theta)\right)}{\partial\theta} = \frac{1}{\sigma^2}(y_i - x_i^\top\theta)x_i \tag{2}$$

$$\frac{\partial^2 \log\left(p(y_i \mid x_i,\theta)\right)}{\partial\theta^2} = -\frac{1}{\sigma^2}x_i x_i^\top = \mathbb{E}_{y_i}\left[\frac{\partial^2 \log\left(p(y_i \mid x_i,\theta)\right)}{\partial\theta^2}\right] \tag{3}$$

$$\mathbb{E}_{y_i}\left[\left(\frac{\partial \log(p(y_i|x_i,\theta))}{\partial\theta}\right)\left(\frac{\partial \log(p(y_i|x_i,\theta))}{\partial\theta}\right)^\top\right] = \frac{1}{\sigma^4}\mathbb{E}\left[x_i x_i^\top(y_i - x_i^\top\theta)^2\right] = \frac{1}{\sigma^4}x_i x_i^\top\mathbb{E}_{y_i}[(y_i - x_i^\top\theta)^2] \tag{4}$$

$$= \frac{1}{\sigma^4}x_i x_i^\top\sigma^2 = \frac{1}{\sigma^2}x_i x_i^\top \tag{5}$$

Note we have written down the expressions for both

$$\mathbb{E}_{y_i}\left[\frac{\partial^2 \log\left(p(y_i \mid x_i,\theta)\right)}{\partial\theta^2}\right] \qquad \text{and} \qquad \mathbb{E}_{y_i}\left[\left(\frac{\partial \log(p(y_i|x_i,\theta))}{\partial\theta}\right)\left(\frac{\partial \log(p(y_i|x_i,\theta))}{\partial\theta}\right)^\top\right]$$

just to work explicitly through both definitions of the Fisher Information for twice continuously differentiable pdfs. With these facts, we see that the Fisher Information in this case is $\mathcal{I}(\theta) = \frac{1}{\sigma^n}\sum_{i=1}^n x_i x_i^\top$. Note that it is **independent** of θ.

Furthermore, it is inversely dependent on σ^2. This intuitively makes sense since the higher the variance, the more noisy each y_i is and hence, the less the information each y_i has about the parameter θ.

Example 2 (Logistic Regression)**.** Let's consider a logistic regression problem with observations $y_i \in \{-1, 1\}$. Since these observations are independent Bernoulli random variables, the likelihood for the logistic regression model is given by

$$l(\theta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)} = \prod_{i=1}^{n}(1-p_i)\left(\frac{p_i}{1-p_i}\right)^{y_i} \tag{6}$$

where $p_i = p(y_i = 1 \mid x_i, \theta) = \frac{\exp(\sum_{j=1}^{p}\theta_j x_{ij})}{1+\exp(\sum_{j=1}^{p}\theta_j x_{ij})}$.

Then, introducing for convenience a covariate $x_{i0} \equiv 1$ for all i that captures the intercept term, the log-likelihood is

$$l(\theta_0, \ldots, \theta_p) = \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i) \tag{7}$$

$$= \sum_{i=1}^{n}\left(y_i \sum_{j=1}^{p}\theta_j x_{ij} - \log\left(1 + \exp(\sum_{j=1}^{p}\theta_j x_{ij})\right)\right) \tag{8}$$

Then the first and second derivative of the likelihood function is

$$\frac{\partial l(\theta)}{\partial \theta_m} = \sum_{i=1}^{n} x_{im}\left(y_i - \frac{\exp(\sum_{j=1}^{p}\theta_j x_{ij})}{1+\exp(\sum_{j=1}^{p}\theta_j x_{ij})}\right) \tag{9}$$

$$\frac{\partial^2 l(\theta)}{\partial \theta_m \partial \theta_l} = -\sum_{i=1}^{n} x_{im}x_{il}\frac{\exp(\sum_{j=1}^{p}\theta_j x_{ij})}{(1+\exp(\sum_{j=1}^{p}\theta_j x_{ij}))^2} = -X_m^\top W X_l \tag{10}$$

where we have set $X_j = (x_{1,j}, \ldots, x_{n,j})$ as the j^{th} column of the matrix of covariates, and defined the $n \times n$ diagonal matrix W as

$$W := W(\theta) = \text{diag}\left(\frac{\exp(\sum_{j=1}^{p}\theta_j x_{1j})}{(1+\exp(\sum_{j=1}^{p}\theta_j x_{1j}))^2}, \ldots, \frac{\exp(\sum_{j=1}^{p}\theta_j x_{nj})}{(1+\exp(\sum_{j=1}^{p}\theta_j x_{nj}))^2}\right) \tag{11}$$

Therefore, the Fisher information matrix for the logistic regression is $I_Y(\theta) = X^\top W X$ .Note that the Fisher information is a function of θ.

2.1 The Cramér-Rao Lower Bound

A common strategy in statistics is to develop a random quantity X that is an unbiased estimator for the parameter θ of interest. For example, when trying to estimate the probability p of a coin landing on heads, one can flip the coin n times and look at the proportion of tosses that landed on heads. This is an unbiased estimator for p.

When pursuing this strategy, one needs to bound the variance of the estimator, as the estimator may be very poor on any given instantiation. For example, in the coin example we just discussed, if we only used a single flip, then our estimator would be 0 or 1 depending on the outcome of the single flip. If $p = 1/2$, this estimator would **always** be 1/2 far away from the true value of p.

In this section, we consider a strong lower bound on the variance of any unbiased estimator. Later, we will see an example of this lower bound applied to Linear Regression and compare the variance achieved via maximum likelihood estimation to this lower bound. This example will show you how you can use the CRLB to certify the (asymptotic) optimality of an estimator that you've devised for your task.

Because we are working in high dimensions, we will work with the covariance matrix of an estimator.

Theorem 2 (Cramér-Rao Lower Bound (CRLB))**.** *Let X be a random variable with pdf $p(x;\theta)$ which depends on θ and is continuously differentiable in x, θ. Let $T, \psi : \mathbb{R}^d \to \mathbb{R}^m$ be continuously differentiable functions such that $T(X)$ be an unbiased estimator of $\psi(\theta)$, which only depends on $x \sim p(x;\theta)$, for every θ. Then*

$$\mathrm{cov}_{X\sim p(x;\theta)}[T(X)] \succeq \frac{\partial\psi(\theta)}{\partial\theta} \cdot \mathcal{I}_X(\theta)^{-1} \cdot \left(\frac{\partial\psi(\theta)}{\partial\theta}\right)^\top$$

In particular, if $\psi(\theta) = \theta$, i.e. that $T(X)$ is an unbiased estimator for θ, then $\mathrm{cov}_{X\sim p(x;\theta)}(T(X)) \succeq \mathcal{I}_X(\theta)^{-1}$.

We give an easy proof in one dimension. The full proof is given in Appendix A.

Proof. (Proof for single parameter)

Let $T(X)$ be a an unbiased estimator of $\psi(\theta)$, that is, $\mathbb{E}[T(X)] = \psi(\theta)$. The goal is to prove that, for all θ,

$$\mathrm{Var}(T(x)) \geq \frac{[\psi\prime(\theta)]^2}{\mathcal{I}_X(\theta)}$$

where $\psi\prime(\theta) = \frac{\partial}{\partial\theta}\psi(\theta)$.

Let x be a random variable with probability density function $p(x;\theta)$. Define the score function $V = \frac{\partial}{\partial\theta}\log p(x;\theta)$ which equals $\frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta}p(x;\theta)$ by the Chain Rule. Then $\mathbb{E}[V] = 0$ because

$$\mathbb{E}[V] = \int p(x;\theta)\left[\frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta}p(x;\theta)\right]dx = \frac{\partial}{\partial\theta}\int p(x;\theta)dx = 0$$

where the integral and partial derivative have been interchanged.

If we consider covariance $\mathrm{cov}(V,T)$ of V and T, we have $\mathrm{cov}(V,T) = \mathbb{E}[(V - \mathbb{E}[V])(T - \mathbb{E}[T])] = \mathbb{E}[V(T - \mathbb{T})] = \mathbb{E}[VT] - \mathbb{E}[V]\mathbb{E}[T] = \mathbb{E}[VT]$ (since $\mathbb{E}[V] = 0$), then by expanding this expression, we have

$$\begin{aligned}\mathrm{cov}(V,T) &= \mathbb{E}\left(T\left[\frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta}p(x;\theta)\right]\right) = \int T(x)\left[\frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta}p(x;\theta)\right]p(x;\theta)dx \\ &= \frac{\partial}{\partial\theta}\left[\int T(x)p(x;\theta)dx\right] = \psi\prime(\theta)\end{aligned}$$

where again, the continuous differentiability allows us to exchange the order of integration and differentiation.

The Cauchy-Schwarz inequality shows that:

$$\sqrt{\mathrm{Var}(T)\,\mathrm{Var}(V)} \geq |\,\mathrm{cov}(V,T)| = |\psi\prime(\theta)|$$

therefore,

$$\mathrm{Var}(T) \geq |\mathrm{cov}(V,T)| = \frac{[\psi\prime(\theta)]^2}{\mathrm{Var}(V)} = \frac{[\psi\prime(\theta)]^2}{\mathcal{I}_X(\theta)}$$

which proves the proposition. □

***Remark* 4.** We will write $T(X)$ instead of $T_\theta(X)$ for convenience. However, note that the distribution of $T(X)$ implicitly depends on θ, as X is distributed according to $p(x;\theta)$.

3 Maximum Likelihood Estimators, Fisher Information, and Experimental Design

In this section, we will consider maximum likelihood estimation (MLE) as a method of devising estimators for experimental design.

Assumption 1. *The observations $\{y_i\}_{i=1}^n$ are given by:*

$$y_i = x_i^T \theta_* + \epsilon_i, \qquad i \in \{1, \ldots, n\} \tag{12}$$

where the $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d., and θ_* is unknown.

Definition 5. *Maximum likelihood estimate (MLE):*

$$\hat{\theta}_{MLE} = \arg\max_\theta \prod_{i=1}^n p(y_i|\theta) = \arg\max_\theta \sum_{i=1}^n \log(p(y_i|\theta)) \tag{13}$$

Example 3. Normally distributed observations: Let $\{y_i\}_{i=1}^n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ is an arbitrary number (one dimension). In other words:

$$p(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

then the maximum likelihood estimate for the mean μ is:

$$\hat{\mu}_{MLE} = \arg\max_{\hat{\mu}} \sum_{i=1}^n -\frac{(y_i - \hat{\mu})^2}{2}$$

Therefore, by computing the gradient and setting to zero, the MLE estimate would be:

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^n y_i$$

In the context of linear regression, we would have that $y_i = x_i^\top \theta_* + \epsilon_i \sim \mathcal{N}(x_i^\top \theta_*, \sigma^2)$ since $x_i^\top \theta_*$ is a deterministic (but unknown) quantity. For example, we were allowed to choose any vector in $\mathbb{R}^d$ for x_i (which may not be possible in general due to physical constraints of the experiments you can design), one could just choose $x_i = e_i$, the ith standard basis vector, for $i = 1, \ldots, n$, in which case we would have $x_i^\top \theta_* = \theta_*(i)$ and $y_i \sim \mathcal{N}(\theta_*(i), \sigma^2)$. In this case, the y_i themselves would be estimators for the components of θ_*, and by repeating each experiment sufficiently many times, the variance of the y_i would drop.

3.1 Consistency of MLE in General

In this subsection, we will show that the MLE $\hat{\theta}_{MLE}$ converges asymptotically to the true parameter value θ_*. In particular, this guarantees that after sufficiently many samples are drawn, the $\hat{\theta}_{MLE}$ will be a good estimator for θ_*, and hence justifies why the MLE can be useful at all. We say an estimator $\hat{\theta}$ is **consistent** if it converges in probability to θ_* as the number of samples increases.

Lemma 2. *Fix* ***any*** *estimator θ. As n goes to infinity, the log-likelihood of Y w.r.t. θ converges almost surely to the true parameter of the distribution:*

$$\frac{1}{n} \sum_{i=1}^n \log\left(p(y_i|\theta)\right) \overset{a.s.}{\to} \mathbb{E}_{y \sim p(Y|\theta_*)}\left[\log p(Y|\theta)\right]$$

Similarly,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log\left(p(y_i|\theta)\right) &\overset{a.s.}{\to} \frac{\partial}{\partial\theta}\mathbb{E}_{y\sim p(Y|\theta_*)}\left[\log p(Y|\theta)\right]\\ \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log\left(p(y_i|\theta)\right) &\overset{a.s.}{\to} \frac{\partial^2}{\partial\theta^2}\mathbb{E}_{y\sim p(Y|\theta_*)}\left[\log p(Y|\theta)\right]\end{aligned}$$

In particular, this holds for the maximum likelihood estimator $\hat{\theta}_{MLE}$.

Proof. The first convergence is an immediate consequence of the **Law of Large Numbers (LLN)**. The remaining two convergence claims follow from the first convergence claim, and linearity of expectation and partial differentiation. We suppress technical regularity conditions such as differentiation under the integral. □

In the following we prove that the KL divergence between two probability distributions p, q is a nonnegative quantity and that it is zero only if the two distributions are equal almost everywhere. Crucially, we will use it to show that θ_* is a (the) parameter θ that maximizes $\mathbb{E}_{y\sim p(Y|\theta_*)}[\log p(y \mid \theta)]$, which intuitively is the case; the best "estimator" for θ_* asymptotically is θ_* is itself. We want this because we already have the corresponding fact that the MLE maximizes the empirical log-likelihood $\frac{1}{n}\sum_{i=1}^{n}\log p(y_i \mid \theta)$ (which by the above, converges to $\mathbb{E}_{y\sim p(Y|\theta_*)}[\log p(y \mid \theta)]$).

Lemma 3. *$KL(p,q) \geq 0$ for every pair of distributions p, q. Furthermore, $KL(p,q) = 0$ only if $\{x : p(x) \neq q(x)\}$ has 0 Lebesgue measure.*

Proof. We will show $-KL(p,q) \leq 0$.

$$\begin{aligned}-KL\left(p,q\right) &= \mathbb{E}_{y\sim p}\left[\log\left(\frac{q(y)}{p(y)}\right)\right]\\ &\leq \log\left(\mathbb{E}_{y\sim p}\left[\frac{q(y)}{p(y)}\right]\right) && \text{by Jensen's inequality and concavity of logarithm}\\ &= \log\left(\int\left(\frac{q(y)}{p(y)}\right)p(y)\,dy\right)\\ &= \log\left(\int q(y)\,dy\right) = \log(1) = 0\end{aligned}$$

Now, if $\{x : p(x) \neq q(x)\} = \{x : \log(q(x)/p(x)) \neq 0\}$ has nonzero Lebesgue measure[1], then since $\log(x)$ is **strictly** concave, Jensen's Inequality is a strict inequality, whence $-KL(p,q) < 0$. □

Corollary 1. *θ_* maximizes $\mathbb{E}_{y\sim p(Y|\theta_*)}[\log p(y \mid \theta)]$ over all choices of θ, for any θ_*. Furthermore, if $\{y : p(y \mid \theta_*) \neq p(y \mid \theta)\}$ has nonzero Lebesgue measure for every choice of $\theta \neq \theta_*$, then θ_* is the unique maximizer of $\mathbb{E}_{y\sim p(Y|\theta_*)}[\log p(y \mid \theta)]$ over all choices of θ.*

[1] We will not give a precise definition of the Lebesgue measure, as throughout this note, we will only need to distinguish between sets with zero or nonzero Lebesgue measure. One should think of a set having zero Lebesgue measure as a set on which the integral of any function is always zero; from the perspective of taking integrals (in particular, expectations), sets of zero Lebesgue measure might as well be empty sets. More generally, given a function $f : \mathbb{R}^n \to \mathbb{R}$, one can modify the function arbitrarily on a set of zero Lebesgue measure without changing $\int_{\mathbb{R}^n} f(x)\,dx$.

All countable sets (including $\mathbb{Q}^n$) in $\mathbb{R}^n$ have zero Lebesgue measure. Any set which has nonempty interior (contains an open ball of positive radius) has nonzero Lebesgue measure. One can also think of the Lebesgue measure of a set as its volume.

Functions f, g which differ on a set of nonzero Lebesgue measure are "nontrivially different". For example, if f, g are continuous functions which differ at a single point x, then f, g differ on a set of nonzero Lebesgue measure (not because $\{x\}$ has nonzero Lebesgue measure but because continuity of f, g forces them to differ on an ϵ-radius ball around x for sufficiently small ϵ).

Proof. Let θ be any parameter. For convenience, set $p = p(.|\theta_*)$ and $q = p(.|\theta)$. Hence, we may rewrite the expectations of interest as $\mathbb{E}_{y\sim p(Y|\theta_*)}[\log p(y \mid \theta)] = \mathbb{E}_{y\sim p}[\log q(y)]$ and $\mathbb{E}_{y\sim p(Y|\theta_*)}[\log p(y \mid \theta_*)] = \mathbb{E}_{y\sim p}[\log p(y)]$. Now, observe that by Lemma 2

$$\mathbb{E}_{y\sim p}[\log p(y)] - \mathbb{E}_{y\sim p}[\log q(y)] = \mathbb{E}_{y\sim p}\left[\log \frac{p(y)}{q(y)}\right] = KL(p,q) \geq 0$$

The uniqueness claim follows from the fact that by Lemma 2, we have strict inequality $KL(p,q) > 0$ if $\{y : p(y) \neq q(y)\}$ has nonzero Lebesgue measure. □

With all of these tools, we can now prove the following theorem, which justifies why the MLE is a commonly used estimator.

Theorem 3. *Let $\hat{\theta}_{MLE}(n)$ denote the maximum likelihood estimator obtained from samples $y_1,\dots,y_n$. Then $\hat{\theta}_{MLE}(n) \to \theta_*$ as $n \to \infty$ almost surely. In particular, the MLE is (asymptotically)* ***consistent****.*

Proof. For convenience, we will assume $p(y \mid \theta)$ varies jointly smoothly in y, θ and that $\{y : p(y \mid \theta_*) \neq p(y \mid \theta)\}$ has nonzero Lebesgue measure for every choice of $\theta \neq \theta_*$. By Lemma 1, we have $\frac{1}{n}\sum_{i=1}^n \log p(y_i \mid \theta)$ converges to $\mathbb{E}_{y\sim p(Y|\theta_*)}[\log p(y \mid \theta)]$ as $n \to \infty$ almost surely (recall this was due to the Law of Large Numbers). Since $\hat{\theta}_{MLE}(n)$ maximizes $\frac{1}{n}\sum_{i=1}^n \log p(y_i \mid \theta)$ (by definition), θ_* maximizes $\mathbb{E}_{y\sim p(Y|\theta_*)}[\log p(y \mid \theta)]$ by Corollary 1, and $p(y \mid \theta)$ varies jointly smoothly in y, θ, we must have $\hat{\theta}_{MLE}(n) \to \theta_*$ as $n \to \infty$ almost surely. □

3.2 Adding in the CLT

We've now shown that the MLE is consistent. However, we can use even stronger theorems than the LLN. In particular, we can now use the Central Limit Theorem (CLT) and obtain more quantitative results.

Before we begin, we remind the reader that here, the Fisher Information is with respect to the random variable $Y \sim p(y;\theta_*)$. The reason is that our estimator $\hat{\theta}_{MLE}$ is a function of samples y_i of the random variable Y. For convenience, we will however write $\mathcal{I}(\theta_*)$ to mean $\mathcal{I}_Y(\theta_*)$.

Now, let's recall the CLT.

Theorem 4 (Central Limit Theorem)**.** *Suppose $\{X_1, X_2, \dots\}$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. For $n \in \mathbb{N}$, define the random variables $S_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge* ***in distribution*** *to a normal $\mathcal{N}(0,\sigma^2)$, which may be written as*

$$\sqrt{n}\,(S_n - \mu) \xrightarrow{D} \mathcal{N}(0,\sigma^2)$$

We will use this theorem later to study the MLE asymptotically in a more quantitative fashion. But more now, we leave it aside and begin by looking at the Taylor expansion of the log-likelihood function around θ^* to get a better intuition:

$$\log\left(p(y|\theta)\right) = \log\left(p(y|\theta')\right)\Big|_{\theta'=\theta_*} + \frac{\partial \log\left(p(y|\theta')\right)}{\partial \theta'}\Big|_{\theta'=\theta_*}(\theta - \theta_*) + \frac{1}{2}\frac{\partial^2 \log\left(p(y|\theta')\right)}{\partial \theta'^2}\Big|_{\theta'=\theta_*}(\theta-\theta_*)^2 + \mathcal{O}((\theta-\theta_*)^3)$$

where we use $\mathcal{O}$ to suppress terms of order 3 (which contain derivatives of order 3 and above) given by Taylor's Remainder Theorem (we will justify this very soon in the footnote; for now just take it as a notational convenience). Taking the derivative with respect to θ will result in:

$$\frac{\partial \log\left(p(y|\theta)\right)}{\partial \theta} = \frac{\partial \log\left(p(y|\theta')\right)}{\partial \theta'}\Big|_{\theta'=\theta_*} + \frac{\partial^2 \log\left(p(y|\theta')\right)}{\partial \theta'^2}\Big|_{\theta'=\theta_*}(\theta-\theta_*) + \mathcal{O}((\theta-\theta_*)^2)$$

where, similarly, $\mathcal{O}$ is used to suppress terms of order 2 (which contain derivatives of order 3, not 2, due to the differentiation). From (13) we know that the derivative of the objective function is zero at the optima ($\hat{\theta}_{MLE}$) so,

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log\left(p(y_i|\theta)\right)\Big|_{\theta=\hat{\theta}_{MLE}} = 0$$

Thus, combining this with the Taylor expansion, multiplying on both sides by $1/\sqrt{n}$, and ignoring higher order terms[2], we obtain

$$\begin{aligned} 0 = \frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log\left(p(y_i|\theta)\right)\Big|_{\theta=\hat{\theta}_{MLE}} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\log\left(p(y_i|\theta')\right)}{\partial\theta'}\Big|_{\theta'=\theta_*} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial^2\log\left(p(y_i|\theta')\right)}{\partial\theta'^2}\Big|_{\theta'=\theta_*}(\theta-\theta_*) \\ &= \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\frac{\partial\log\left(p(y_i|\theta')\right)}{\partial\theta'}\Big|_{\theta'=\theta_*} + \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log\left(p(y_i|\theta')\right)}{\partial\theta'^2}\Big|_{\theta'=\theta_*}\right)\cdot\sqrt{n}(\hat{\theta}_{MLE}-\theta_*) \end{aligned}$$

Therefore, we have, using the almost sure convergence results stated above,

$$\begin{aligned} \sqrt{n}(\hat{\theta}_{MLE}-\theta_*) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\log\left(p(y_i|\theta')\right)}{\partial\theta'}\Big|_{\theta'=\theta_*}\right)\left(-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log\left(p(y_i|\theta')\right)}{\partial\theta'^2}\Big|_{\theta'=\theta_*}\right)^{-1} \\ &\overset{\text{a.s.}}{\to} \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\log\left(p(y_i|\theta')\right)}{\partial\theta'}\Big|_{\theta'=\theta_*}\right)\mathbb{E}_{y\sim p(Y|\theta_*)}\left[-\frac{\partial^2\log p(Y\mid\theta_*)}{\partial\theta^2}\right]^{-1} \end{aligned}$$

With this, we have by combining Lemma 1 (equivalent definition of Fisher Information via second order partials) and Lemma 2

$$\mathbb{E}_{y\sim p(Y|\theta_*)}\left[-\frac{\partial^2\log p(Y\mid\theta)}{\partial\theta^2}\right]\overset{\text{a.s.}}{\to}\mathcal{I}(\theta_*)$$

and, combining Lemma 2 **with** the Central Limit Theorem,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\log\left(p(y_i|\theta')\right)}{\partial\theta'}\Big|_{\theta'=\theta_*}\right)\overset{D}{\to}\mathcal{N}(0,\mathcal{I}(\theta_*))$$

Thus, we have $\sqrt{n}(\hat{\theta}_{MLE}-\theta_*)\overset{D}{\to}\mathcal{I}(\theta_*))^{-1}\mathcal{N}(0,\mathcal{I}(\theta_*)) = \mathcal{N}(0,\mathcal{I}(\theta_*)^{-1})$; therefore, $\hat{\theta}_{MLE}\overset{D}{\to}\mathcal{N}(\theta_*,\frac{1}{n}\mathcal{I}(\theta_*)^{-1})$.

3.2.1 Cramér Rao Lower Bound Comparison

Note in this setting, $\hat{\theta}_{MLE}$ is an unbiased estimator for θ_*. Hence, one can ask what the CRLB gives. It turns out that the CRLB gives

$$\mathbb{E}\left[(\hat{\theta}_{MLE}-\theta_*)(\hat{\theta}_{MLE}-\theta_*)^{\mathrm{T}}\right]\succeq\frac{1}{n}\mathcal{I}(\theta_*)^{-1} \tag{14}$$

Since $\hat{\theta}_{MLE}\overset{\text{a.s.}}{\to}\mathcal{N}(\theta_*,\frac{1}{n}\mathcal{I}(\theta_*)^{-1})$, this shows that $\hat{\theta}_{MLE}$ is **asymptotically efficient**.

[2] We will assume for convenience that $\frac{1}{\sqrt{n}}\mathcal{O}(2)\to 0$ a.s. as $n\to\infty$. Note that here, $\mathcal{O}(2)$ looks like $2\sum_{i=1}^{n}\frac{\partial^3\log(p(y_i|\theta'))}{\partial\theta'^3}(\theta'-\theta_*)^2\Big|_{\theta'=\hat{\theta}_{MLE}}$ since we're summing the entire series over the n samples $y_1,\ldots,y_n$ so the $\mathcal{O}(2)$ implicitly depends on n. To reemphasize, we are assuming this dependence on n has mild growth in that $\frac{1}{\sqrt{n}}\mathcal{O}(2)\to 0$ a.s. as $n\to\infty$

3.3 MLE and Two-Stage Active Learning for Nonlinear Regression

Now, we turn to how the MLE can be applied to the nonlinear setting. Specifically, we now consider the problem is experimental design when f is nonlinear. This is work of Chaudhuri-Mykland [1]. We begin by defining our "optimality" condition. This is how we will measure the quality of our experimental design.

Definition 6 (Locally D-Optimal Design)**.** *A* ***locally D-optimal design at*** θ *is a distribution* ξ^* *on* Ω *that maximizes*

$$\det\left(\int_\Omega \mathcal{I}_X(\theta)\, d\xi^*(X)\right)$$

or, equivalently, minimizes

$$\det\left(\int_\Omega \mathcal{I}_X(\theta)\, d\xi^*(X)\right)^{-1}$$

We will develop an estimation scheme whose Fisher Information estimates converge to the total Fisher Information under a locally D-optimal design.

Before we proceed, let's briefly mention why this might be a reasonable criterion for experimental design. Based on the preceding lecture, we know that D-optimality is proportional to the volume of the **confidence ellipsoid**. Furthermore, we know that the level sets of the Gaussian $\mathcal{N}(0, \mathcal{I}(\theta_*))$, which by the above analysis is the limiting distribution of $\sqrt{n}(\hat{\theta}_{MLE} - \theta_*)$, are precisely scalings of the ellipsoid corresponding to $\mathcal{I}(\theta_*)$.

Here, we will consider the following strategy:

Algorithm 1 MLE Active Learn

```
Input: number of experiments $n$, number of initial stage experiments $n_1$, prior information
via prior information, select $X_1, \ldots, X_{n_1}$ design points (say, uniformly distributed over $\Omega \subset \mathbb{R}^d$ if there is no prior information)
observe response $Y_1, \ldots, Y_{n_1}$
estimate $\theta$ via $\theta^*_{n_1}$ based on $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_1}$
for $t = n_1 + 1, \ldots, n$ do // active learning phase
    select design point $X_t$ maximizing det $\left(\sum_{\tau=1}^{t} \mathcal{I}_{X_\tau}(\theta_{t-1})\right)$ (determinant of total Fisher Information)
    observe response $Y_t$
    estimate $\theta$ via $\theta^*_t$ dependent on $X_1, \ldots, X_t, Y_1, \ldots, Y_t$
end for
return final estimate $\theta^*_n$
```

At this point, we still have not specified "how to use prior information to choose $X_1, \ldots, X_{n_1}$" and how the estimators θ^*_t are defined. Basically, the former can be answered by selecting an evenly distribution of $X_1, \ldots, X_{n_1}$ (i.e. a diverse set of points), which works for most situations encountered. For the latter, we will essentially use maximum likelihood estimates.

In fact, one can prove that, under some technical conditions, $\frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_{X_i}(\theta)$ converges to $\int_\Omega \mathcal{I}_X(\theta)\, d\xi^*(X)$ in probability as $n \to \infty$, where ξ^* is a locally D-optimal design at θ. We refer the reader to Theorem 3.5 of [1] for the precise statement and proof.

4 Passive vs. Active Learning for Nonparametric Estimation

Here, we will study estimating f at all points x without estimating θ. For simplicity, we will assume f maps Ω to $\mathbb{R}$. Furthermore, we will restrict attention to when $\Omega = [0, 1]^d$. However, we will allow f to range over much larger classes of functions than, for example, linear functions. We will see that there are strong lower bounds for convergence. This is work of Castro-Willett-Nowak [2].

Here, we will measure the strength of an estimator $\hat{f}$ of f by how far $\hat{f}$ is from f, which is measured by

$$\left|\left|\hat{f} - f\right|\right|_2 \overset{\text{def}}{=} \left(\int_{\mathbb{R}^d} |\hat{f}(x) - f(x)|^2 \, dx\right)^{1/2}$$

Let's first describe things at a high-level but with maximum generality. For this, we will fix a class of functions $\mathcal{F}$, the elements of which we will try to produce estimates for. We want to devise a scheme that will give us a good estimation strategy for every $f \in \mathcal{F}$.

Definition 7 (Risk)**.** *Let $f_n \in \mathcal{F}$. The L_2^2* ***risk of the estimation strategy*** *$(\hat{f}_n, S_n)$ is the expected distance between the produced estimator and f, that is*

$$R(\hat{f}_n, S_n, f) \overset{\text{def}}{=} \mathbb{E}_{\hat{f}_n, S_n}[\left|\left|\hat{f} - f\right|\right|_2^2]$$

where the expectation is taken over possible estimators $\hat{f}_n$ that can result from first randomly sampling $X_n \sim S_n$, and the history of samples $X_1, \ldots, X_{n-1}, Y_1, \ldots, Y_{n-1}$. The ***maxima risk of*** *$(\hat{f}_n, S_n)$ is simply $\sup_{f \in \mathcal{F}} R(\hat{f}_n, S_n, f)$.*

We are interested in asymptotically bounding

$$\underbrace{\inf_{(\hat{f}_n, S_n) \in \mathcal{S}_n}}_{\text{best strategy}} \underbrace{\sup_{f \in \mathcal{F}}}_{\text{worst case}} R(\hat{f}_n, S_n, f)$$

from below (to show that it is "difficult to estimate $\mathcal{F}$" in the worst case) and from above (to show that we "can estimate $\mathcal{F}$" decently well); here $\mathcal{S}_n$ will denote the space of valid estimation strategies conditioned on preceding output $X_1, \ldots, X_{n-1}, Y_1, \ldots, Y_{n-1}$. Note $\mathcal{S}_n$ is small under passive learning because strategies that depend on $X_1, \ldots, X_{n-1}, Y_1, \ldots, Y_{n-1}$ are no longer valid. It will be clear from context whether this is the space of passive or active estimation strategies.

4.1 Well-Studied Classes of Functions

Definition 8 (Locally Hölder Smooth)**.** *A function $f : \Omega \to \mathbb{R}$ is* ***locally Hölder smooth at x with constants*** *$L, \alpha > 0$ if*

- *it has continuous partial derivatives up to order $k = \lfloor \alpha \rfloor$ at x ($k = \lfloor \alpha \rfloor$ is the maximal integer such that $k < \alpha$)*
- *there exists $\epsilon > 0$ such that for all $y \in \mathbb{R}^d$ satisfying $||x - y|| < \epsilon$, $|f(y) - P_x(y)| \leq L \, ||x - y||^\alpha$*

where $P_x(\cdot)$ is the degree-k Taylor polynomial of f expanded around x. f is ***Hölder smooth with constants*** *$L, \alpha > 0$ if it is locally Hölder smooth with constants $L, \alpha > 0$ for all $x \in \mathbb{R}^d$. We denote the class of such functions $\Sigma(L, \alpha)$.*

***Remark* 5.** The second condition says that the degree-k Taylor polynomial of f expanded around x gives an additive ϵ error, i.e. $|f(y) - P_x(y)| \leq \epsilon$, when $||x - y|| \leq (\epsilon/L)^{1/\alpha}$. At a high level, this says that in order for $P_x(y)$ to be a good approximation of $f(y)$, then x (which can be thought of as a design point) only needs to be polynomially close to y (which can be thought of as a point where we use our estimator to approximate f because we have not sampled it).

For more intuition, it may be helpful to consider the 1-dimensional case, and observe that

$$|f(y) - P_x(y)| = \left| f(y) - \sum_{j=0}^{k} \frac{f^{(j)}(x) \cdot (y-x)^j}{j!} \right| \leq |f(y) - f(x)| + \sum_{j=1}^{k} \frac{|f^{(j)}(x)|}{j!} \cdot ||x-y||^j$$

One can then think that if f and its derivatives up to kth order satisfy a Lipschitz continuity condition, then the above is bounded by $L \cdot ||x-y||$. Replacing $||x-y||$ with $||x-y||^{\alpha}$ in some sense relaxes "how continuous" we need f and its derivatives to be.

Theorem 5 (Theorem 1 from [2]). *Under the requirements of the passive learning model, we have*

$$\inf_{(\hat{f}_n, S_n) \in \mathcal{S}_n} \sup_{f \in \Sigma(L,\alpha)} R(\hat{f}_n, S_n, f) \geq \Omega\left(n^{-\frac{2\alpha}{2\alpha+d}}\right)$$

where the constant hidden in Ω depends on $L, \alpha, \sigma^2 > 0$.

Theorem 6 (Theorem 3 from [2]). *Under the requirements of the active learning model, we have*

$$\inf_{(\hat{f}_n, S_n) \in \mathcal{S}_n} \sup_{f \in \Sigma(L,\alpha)} R(\hat{f}_n, S_n, f) \geq \Omega\left(n^{-\frac{2\alpha}{2\alpha+d}}\right)$$

where the constant hidden in Ω depends on $L, \alpha, \sigma^2 > 0$.

We note that both theorems are tight in the sense that we have estimation schemes that attain these lower bounds. **This shows that active learning has no asymptotic advantage over passive learning for functions ranging in the class** $\Sigma(L, \alpha)$.

5 Appendix A: Proof of the Cramér-Rao Lower Bound

To prove this theorem, we will need a few lemmas. Before we begin, we remark that the continuously differentiability assumption made is sufficient to permit the interchanging of order between differentiation and integration; we will not belabor this point and assume all such manipulations as valid. Furthermore, we will drop the $x \sim f(x;\theta)$ notation when writing expectation and covariances; all integrals (and hence expectations and covariances) are taken to mean integrating over $x \sim f(x;\theta)$ for a fixed θ. This is for convenience and cleanliness of notation.

First, we write down two standard lemmas from statistics.

Lemma 4. *Let X, Y be two scalar random variables. Then their correlation coefficient $\rho(X,Y) \overset{\text{def}}{=} \frac{\text{cov}(X,Y)}{\sigma(X)\sigma(Y)}$ satisfies $|\rho(X,Y)| \leq 1$.*

Lemma 5. *Let X, Y be two vector-valued random variables and a, b be any fixed vectors. Then $\text{cov}(a^\top X, b^\top Y) = a^\top \text{cov}(X,Y) b$.*

Proof.

$$\begin{aligned} \text{cov}(a^\top X, b^\top Y) &= \mathbb{E}[a^\top X Y^\top b] - \mathbb{E}[a^\top X]\mathbb{E}[Y^\top b] = a^\top \mathbb{E}[XY^\top] b - a^\top \mathbb{E}[X]\mathbb{E}[Y]^\top b \\ &= a^\top (\mathbb{E}[XY^\top] - \mathbb{E}[X]\mathbb{E}[Y]^\top) b = a^\top \text{cov}(X,Y) b \end{aligned}$$

□

Lemma 6. *Assume the same conditions on X, θ, T, ψ, f as above.*

(1) For every $i = 1, \ldots, d$, $\frac{\partial}{\partial \theta_i} \log f(X;\theta)$ has mean 0 as a random variable (function of X).

(2) For every $i = 1, \ldots, d$,

$$\mathbb{E}_{x\sim f(x;\theta)}\left[T(x)\frac{\partial}{\partial \theta_i}\log f(x;\theta)\right] = \frac{\partial}{\partial \theta_i}\psi(\theta)$$

where the differentiation and equality here are taken entrywise (recall that $T, \psi : \mathbb{R}^d \to \mathbb{R}^m$).

Proof. Observe that via the Chain Rule,

$$\frac{\partial}{\partial \theta_i}\log f(x;\theta) = \frac{1}{f(x;\theta)}\frac{\partial}{\partial \theta_i} f(x;\theta)$$

whence

$$\frac{\partial}{\partial \theta_i} f(x;\theta) = f(x;\theta)\cdot\frac{\partial}{\partial \theta_i}\log f(x;\theta)$$

With this in hand, the remaining is done via computation.

1. For (1), we use a cute little trick where we know that probability densities integrate to 1 and the derivative of a constant function is 0

$$\begin{aligned}\mathbb{E}_{x\sim f(x;\theta)}\left[\frac{\partial}{\partial \theta_i}\log f(x;\theta)\right] &= \int_{\mathbb{R}^d}\left[\frac{\partial}{\partial \theta_i}\log f(x;\theta)\right]\cdot f(x;\theta)\,dx = \int_{\mathbb{R}^d}\frac{\partial}{\partial \theta_i} f(x;\theta)\,dx \\ &= \frac{\partial}{\partial \theta_i}\underbrace{\int_{\mathbb{R}^d} f(x;\theta)\,dx}_{=1\text{ constant}} = 0\end{aligned}$$

   (This is actually the exact same trick that allows you to prove that the Fisher information is the Hessian of the relative entropy; out of laziness, we refer the interested reader here: https://en.wikipedia.org/wiki/Fisher_information#Relation_to_relative_entropy)

2. For (2), additionally using the fact that $T(x)$ is an unbiased estimator for $\psi(\theta)$,

$$\begin{aligned}\mathbb{E}_{x\sim f(x;\theta)}\left[T(x)\frac{\partial}{\partial \theta_i}\log f(x;\theta)\right] &= \int_{\mathbb{R}^d} T(x)\left[\frac{\partial}{\partial \theta_i}\log f(x;\theta)\right] f(x;\theta)\,dx = \int_{\mathbb{R}^d} T(x)\frac{\partial}{\partial \theta_i} f(x;\theta)\,dx \\ &= \frac{\partial}{\partial \theta_i}\int_{\mathbb{R}^d} T(x) f(x;\theta)\,dx = \frac{\partial}{\partial \theta_i}\psi(\theta)\end{aligned}$$

□

Corollary 2. *Assume the same conditions on X, θ, T, ψ, f as above.*

(1)

$$\mathrm{cov}_{x\sim f(x;\theta)}\left[T(x), \nabla_\theta \log f(x;\theta)\right] = \frac{\partial \psi(\theta)}{\partial \theta}$$

(2) $\mathrm{cov}_{x\sim f(x;\theta)}[(\nabla_\theta \log f(x;\theta))(\nabla_\theta \log f(x;\theta))^\top] = \mathcal{I}_X(\theta)$

Proof. We consider the matrix entrywise. Recall that the covariance satisfies

$$\begin{aligned}\mathrm{cov}_{x\sim f(x;\theta)}\left[T(x), \nabla_\theta \log f(x;\theta)\right] = \mathbb{E}_{x\sim f(x;\theta)}&\left[T(x)\left(\nabla_\theta \log f(x;\theta)\right)^\top\right] \\ &- \mathbb{E}_{x\sim f(x;\theta)}[T(x)]\cdot\mathbb{E}_{x\sim f(x;\theta)}\left[\nabla_\theta \log f(x;\theta)\right]^\top\end{aligned}$$

Applying (1) from the previous lemma, we have that the second term immediately above is zero because $\mathbb{E}_{x\sim f(x;\theta)}\left[\nabla_\theta \log f(x;\theta)\right] = 0$. Thus, it suffices to show that the first term equals $\partial\psi(\theta)/\partial\theta$. But

$$\mathbb{E}_{x\sim f(x;\theta)}\left[T(x)\left(\nabla_\theta \log f(x;\theta)\right)^\top\right] = \mathbb{E}_{x\sim f(x;\theta)}\left[\left(T(x)\frac{\partial}{\partial\theta_1}\log f(x;\theta),\ldots,T(x)\frac{\partial}{\partial\theta_d}\log f(x;\theta)\right)\right]$$

where each entry in the parentheses is a column vector, which by (2) of the preceding lemma, equals

$$\left(\frac{\partial}{\partial_1}\psi(\theta),\ldots,\frac{\partial}{\partial\theta_d}\psi(\theta)\right) = \frac{\partial\psi(\theta)}{\partial\theta}$$

This proves (1). (2) follows immediately from the definition of $\mathcal{I}_X(\theta)$. □

With all of these tools in hand, we ready to prove Cramér-Rao.

Proof of CRLB. Let a, b be any vectors and consider the correlation coefficient ρ of $a^\top T(x)$ with $b^\top \nabla_\theta \log f(x;\theta)$. Using Lemma 2,

$$\rho = \frac{a^\top \operatorname{cov}[T(x), \nabla_\theta \log f(x;\theta)]b}{\sqrt{(a^\top \operatorname{cov}[T(x),T(x)]a)\cdot(b^\top \operatorname{cov}[\nabla_\theta \log f(x;\theta), \nabla_\theta \log f(x;\theta)]b)}}$$

Using Corollary 1, $\operatorname{cov}[T(x),\nabla_\theta \log f(x;\theta)] = \partial\psi(\theta)/\partial\theta$ and $\operatorname{cov}[\nabla_\theta \log f(x;\theta),\nabla_\theta \log f(x;\theta)] = \mathcal{I}_X(\theta)$. With this, we have that

$$\rho = \frac{a^\top(\partial\psi(\theta)/\partial\theta)b}{\sqrt{(a^\top \operatorname{cov}[T(x)]a)(b^\top \mathcal{I}_X(\theta)b)}} \le 1$$

Now, we this bound holds for any choice of a, b. Specifically, choose x arbitrary and $a = x, b = -\mathcal{I}_X(\theta)^{-1}\cdot(\partial\psi(\theta)/\partial\theta)\cdot x$. Then we have

$$-\frac{x^\top(\partial\psi(\theta)/\partial\theta)\mathcal{I}_X(\theta)^{-1}(\partial\psi(\theta)/\partial\theta)x}{\sqrt{(x^\top \operatorname{cov}[T(x)]x)(x^\top(\partial\psi(\theta)/\partial\theta)\mathcal{I}_X(\theta)^{-1}(\partial\psi(\theta)/\partial\theta)x)}} = -\sqrt{\frac{x^\top(\partial\psi(\theta)/\partial\theta)\mathcal{I}_X(\theta)^{-1}(\partial\psi(\theta)/\partial\theta)x}{x^\top \operatorname{cov}[T(x)]x}} \le 1$$

Squaring both sides and moving things around, we have

$$x^\top\left(\operatorname{cov}[T(x)] - \left(\frac{\partial\psi(\theta)}{\partial\theta}\right)\mathcal{I}_X(\theta)^{-1}\left(\frac{\partial\psi(\theta)}{\partial\theta}\right)^\top\right)x \ge 0$$

Since x was arbitrary, we conclude that

$$\operatorname{cov}[T(x)] \succeq \left(\frac{\partial\psi(\theta)}{\partial\theta}\right)\mathcal{I}_X(\theta)^{-1}\left(\frac{\partial\psi(\theta)}{\partial\theta}\right)^\top$$

as claimed. □

5.1 Appendix B: A Theorem from Learning Theory for Lower Bounds

Theorem 7 (Main theorem of Risk Minimization; Theorem 13 from [2])**.** *Let Θ be a class of models (functions from Ω to $\mathbb{R}$). Associated with each model $\theta \in \Theta$, we have a probability measure P_θ on Ω. Let $M \ge 2$ be an integer, let $0 < \gamma < 1/8$ be a number, and let $d(\cdot,\cdot) : \Theta \times \Theta \to \mathbb{R}$ be a semimetric. Suppose we have $\{\theta_0,\ldots,\theta_M\} \in \Theta$ such that*

- *$d(\theta_j,\theta_k) \ge 2s > 0$ for all $0 \le j,k \le M$*
- *$P_{\theta_j} \ll P_{\theta_0}$ for all $1 \le j \le M$*

- $\frac{1}{M}\sum_{j=1}^{M} KL(P_{\theta_j} \,||\, P_{\theta_0}) \leq \gamma \log M$

Then

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} P_\theta \left[d(\hat{\theta},\theta) \geq s\right] \geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\gamma - 2\sqrt{\frac{\gamma}{\log M}}\right) > 0$$

***Remark* 6.** Let's do some more hand-waving to understand this theorem.

- A semimetric is a notion of distance that satisfies all properties of regular metrics except perhaps the Triangle Inequality; metrics are semimetrics. In place of $d(\cdot,\cdot)$, one should think of L_2-distance (just as what we are considering in this section).
- For two measures μ,ν, we write $\nu \ll \mu$ to mean that $\mu(A) = 0 \implies \nu(A) = 0$ for every event $A \subset \Omega$; in English, we say that μ "dominates" ν. So when we write $P_{\theta_j} \ll P_{\theta_0}$, one can think of this informally as saying that $\text{supp}(p_{\theta_j}) \subset \text{supp}(p_{\theta_0})$, where p_{θ_i} is the density function of the measure P_{θ_i} for all i.
- Essentially, this is what the theorem is saying: "if your model class Θ contains some models $\theta_0,\ldots,\theta_M$ that are very different but the distributions associated with them are very similar, then this model class has high risk". The reason this is intuitively true is the following: Suppose you sampling points from P_{θ_i} for $i \neq 0$ from which you construct your estimator $\hat{\theta}$. Since P_{θ_i} is "close" to P_{θ_0}, the points you sample will "likely work" for estimating θ_0 as well as for θ_i. Since your estimator $\hat{\theta}$ is based on these sampled points, it will necessarily have difficulty distinguishing between θ_0 and θ_i. But since θ_0,θ_i have large distance w.r.t. d, your estimate $\hat{\theta}$ will be poor for both θ_0,θ_i.
- The power of this theorem is that we have **not made any assumption on what probability measures P_θ are assigned to each model θ.** In particular, there can be arbitrarily complex dependencies between the distributions P_θ, and so this theorem can be used for lower bounds for both the passive setting and the active setting.
- The bigger M is, the greater the lower bound is. Typically, we will want M to be superpolynomially large in the dimension (and this can be achieved for $\Theta = \Sigma(L,\alpha)$).

Corollary 3 (Corollary 2 from [2])**.** *Under the assumptions of the preceding theorem,*

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}[d(\hat{\theta},\theta)^2] \geq s^2\frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\gamma - 2\sqrt{\frac{\gamma}{\log M}}\right) > cs^2$$

where $c = c(\gamma, M) > 0$ is a constant.

Proof. This follows via Markov's Inequality. Specifically,

$$P_\theta\left[d(\hat{\theta},\theta) \geq s\right] = P_\theta\left[d(\hat{\theta},\theta)^2 \geq s^2\right] \leq \frac{\mathbb{E}[d(\hat{\theta},\theta)^2]}{s^2}$$

Thus,

$$\mathbb{E}[d(\hat{\theta},\theta)^2] \geq s^2 P_\theta\left[d(\hat{\theta},\theta) \geq s\right]$$

and the result by taking $\inf_{\hat{\theta}} \sup_{\theta\in\Theta}$ on both sides and applying the preceding theorem. □

The way to prove lower bounds for passive and active learning is then to take $\Theta = \Sigma(L,\alpha)$ and construct many functions $f_0,\ldots,f_M \in \Sigma(L,\alpha)$ that are separated by a large distance under the L_2 metric and apply the theorems above. We refer the reader to the paper for the construction.

References

[1] Probal Chaudhuri and Per A. Mykland. Nonlinear experiments: Optimal design and inference based on likelihood. *Journal of the American Statistical Association*, 88(422):538–546, 1993.

[2] Rebecca Willett Rui Castro and Robert Nowak. Faster rates in regression via active learning, 2005.

[3] Friedrich Pukelsheim. *Optimal Design of Experiments (Classics in Applied Mathematics) (Classics in Applied Mathematics, 50).* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2006.