Properties of Estimators (and the MLE)

Is the MLE any good? More generally, how can we talk in measured terms about the efficacy of a particular estimator?

Let's start by carefully laying out our problem scenario, and what we mean by an "estimator". Suppose that we observe independent realizations $X_1 = \boldsymbol{x}_1, \ldots, X_N = \boldsymbol{x}_N$, $\boldsymbol{x}_n \in \mathbb{R}^D$ of a random vector with distribution

$$X_n \sim f_X(\boldsymbol{x}; \boldsymbol{\theta}_0),$$

for some fixed $\mathcal{T} \subset \boldsymbol{\theta}_0 \in \mathbb{R}^P$ (that is, there are P parameters that control the distribution of the X_n). An **estimator** based on these N observations is some fixed mapping $g(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) : \mathbb{R}^D \to \mathcal{T}$, defined for every N, that takes the observed data and produces[1] an estimate $\hat{\boldsymbol{\theta}}_N$.

With the randomness of the data taken into account, the estimate is a function of a sequence of random variables,

$$\hat{\boldsymbol{\Theta}}_N = g(X_1, \ldots, X_N),$$

and so is itself a random variable. We will discuss three properties of $\hat{\boldsymbol{\Theta}}_N$ that might be desirable:

Bias: An estimator is called *unbiased* if it is equal to $\boldsymbol{\theta}_0$ on average.

Consistency: An estimator is called *consistent* if it is guaranteed to return $\boldsymbol{\theta}_0$ as the amount of data we have goes to infinity.

Efficiency: An unbiased estimator estimator is called *efficient* if

[1] The MLE, for example, evaluates the log likelihood functions at the N different points for all values of $\boldsymbol{\theta}$, adds these together, then returns the $\boldsymbol{\theta}$ for which this sum is maximized.

it has the minimum mean-squared error of any possible estimator procedure for a particular value of N.

Bias

The bias of an estimator $\hat{\boldsymbol{\Theta}}$ is simply

$$\text{bias}(\hat{\boldsymbol{\Theta}}) = \text{E}[\hat{\boldsymbol{\Theta}}] - \theta_0.$$

An estimator is called **unbiased** if the bias is zero for all values of $\boldsymbol{\theta}_0 \in \mathcal{T}$.

The bias plays a role in describing the mean-squared error (MSE) of an estimator. For a scalar $\hat{\Theta}$, the MSE is

$$\begin{aligned}\text{MSE}(\hat{\Theta}) &= \text{E}[(\hat{\Theta} - \theta_0)^2] \\ &= \text{E}([\hat{\Theta} - \text{E}[\hat{\Theta}])^2] + (\text{E}[\hat{\Theta}] - \theta_0)^2 \\ &= \text{var}(\hat{\Theta}) + \text{bias}(\hat{\Theta})^2.\end{aligned}$$

For vector-valued estimators, this extends to

$$\begin{aligned}\text{MSE}(\hat{\boldsymbol{\Theta}}) &= \text{E}[\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}_0\|_2^2] \\ &= \text{trace}(\hat{\boldsymbol{R}}) + \|\,\text{E}[\hat{\boldsymbol{\Theta}}] - \boldsymbol{\theta}_0\|_2^2\end{aligned}$$

where $\hat{\boldsymbol{R}}$ is the covariance matrix for the estimator,

$$\hat{\boldsymbol{R}} = \text{E}[(\hat{\boldsymbol{\Theta}} - \text{E}[\hat{\boldsymbol{\Theta}}])(\hat{\boldsymbol{\Theta}} - \text{E}[\hat{\boldsymbol{\Theta}}])^{\text{T}}]$$

It is often the case that we can decrease the variance but at a cost of increasing the bias or vice versa — this is referred to as the **bias variance trade off**.

An estimator is called **asymptotically unbiased** if

$$\mathrm{E}[\hat{\boldsymbol{\Theta}}_N] \to \boldsymbol{\theta}_0 \quad \text{as} \quad N \to \infty,$$

for all $\boldsymbol{\theta}_0 \in \mathcal{T}$.

The MLE is sometimes unbiased and sometimes biased, but it is always asymptotically unbiased. We will not prove this here, but it will come out in our discussion of consistency in the next section.

Consistency

Consistent estimators return estimates that become arbitrarily close to $\boldsymbol{\theta}_0$ as N becomes large. To be a useful notion, this should be true not just for a particular data set, but should be "typical" behavior over large ensembles of data sets. This is made precise in the following way. We say that the estimator is **consistent** if for every $\epsilon > 0$ and $0 < \delta < 1$, there is an N (that of course depends on ϵ, δ) such that

$$\mathrm{P}\left(|\hat{\boldsymbol{\Theta}}_N - \boldsymbol{\theta}_0| > \epsilon\right) \leq \delta.$$

In words, this means that the probability that the estimate is off by more than ϵ can be made arbitrarily small with enough data.

Let's look at a particular example. Suppose that the X_n are binary-valued random variables,

$$X_n = \begin{cases} 1, & \text{with probability } \theta_0, \\ 0, & \text{with probability } 1 - \theta_0, \end{cases}$$

and let $\hat{\boldsymbol{\Theta}}_N$ be the sample mean (which we have already seen is also

the MLE):

$$\hat{\mathbf{\Theta}}_N = \frac{1}{N}\sum_{n=1}^{N} X_n.$$

It is clear that $\hat{\mathbf{\Theta}}_N$ is unbiased, as

$$\mathrm{E}[\hat{\mathbf{\Theta}}_N] = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}[X_n] = \theta_0,$$

and the variance is

$$\mathrm{var}(\hat{\mathbf{\Theta}}_N) = \frac{1}{N^2}\sum_{n=1}^{N} \mathrm{var}(X_n) = \frac{1}{N^2}\sum_{n=1}^{N} \theta_0(1-\theta_0) = \frac{\theta_0(1-\theta_0)}{N}.$$

The Chebyshev inequality (see the Technical Details section) says that

$$\mathrm{P}\left(|\hat{\mathbf{\Theta}}_N - \theta| > \epsilon\right) \;\leq\; \frac{\theta_0(1-\theta_0)}{N\epsilon^2},$$

that is for any $0 < \delta < 1$,

$$\mathrm{P}\left(|\hat{\mathbf{\Theta}}_N - \theta| > \epsilon\right) \;\leq\; \delta \quad \text{when} \quad N > \frac{\theta_0(1-\theta_0)}{\epsilon^2\delta}.$$

As $\theta(1-\theta) \leq 1/4$ for all $\theta \in [0,1]$, we can ensure the probability above when

$$N > \frac{1}{4\epsilon^2\delta}.$$

So, the sample mean is a consistent estimator for the probability that a binary random variable is equal to 1.

Consistency of the MLE

We can establish the consistency of the MLE from a series of independent samples under fairly general conditions. As above, suppose that

$$X_n \sim f_X(\boldsymbol{x}; \boldsymbol{\theta}_0), \quad \text{for some } \boldsymbol{\theta}_0 \in \mathcal{T}. \tag{1}$$

The MLE takes observations $X_1 = \boldsymbol{x}_1, \ldots, X_N = \boldsymbol{x}_N$ and then finds the maximum of the log likelihood:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta} \in \mathcal{T}} \ell(\boldsymbol{\theta}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \;=\; \arg\max_{\boldsymbol{\theta} \in \mathcal{T}} \sum_{n=1}^{N} \ell(\boldsymbol{\theta}; \boldsymbol{x}_n),$$

where

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}_n) = \log f_X(\boldsymbol{x}_n; \boldsymbol{\theta}).$$

To make this problem well-posed, we will assume that the collection of probability models $\{f_X(\boldsymbol{x}; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \mathcal{T}\}$ is **identifiable**. This means that for each value of $\boldsymbol{\theta} \in \mathcal{T}$, $f_X(\boldsymbol{x}; \boldsymbol{\theta})$ is a different density function. More precisely, it means that if I draw random a X using the probability law (1), then evaluate[2] two different densities at this point, then there is at least some chance I get different answers:

$$\mathrm{P}\left(f_X(X; \boldsymbol{\theta}_1) \neq f_X(X; \boldsymbol{\theta}_2)\right) > 0 \quad \text{for all } \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{T}, \quad \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2.$$

Of course, for different sets of observations $\{\boldsymbol{x}_n\}$, the log likelihood function $\ell(\boldsymbol{\theta}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ will be different, and hence so will it maximizer $\hat{\boldsymbol{\theta}}_{\text{MLE}}$. Taking the random nature of the observations into

[2] It is worth re-emphasizing here that the quantities $f_X(X; \boldsymbol{\theta}_1)$ and $f_X(X; \boldsymbol{\theta}_2)$ are random variables — they are created by evaluating densities at a common random point.

account, we can think of

$$\ell(\boldsymbol{\theta}; X_1, \ldots, X_N) = \sum_{n=1}^{N} \ell(\boldsymbol{\theta}; X_n),$$

as a random function of $\boldsymbol{\theta}$ — the draw of the X_n will determine what this function is. This random function has some well-defined expected value at every point $\boldsymbol{\theta} \in \mathcal{T}$; we call this function the

$$\text{expected log likelihood} \quad = \quad \mathrm{E}[\ell(\boldsymbol{\theta}; X_1, \ldots, X_N)].$$

While $\ell(\boldsymbol{\theta}; X_1, \ldots, X_N)$ is a random function, the expected log likelihood $\mathrm{E}[\ell(\boldsymbol{\theta}; X_1, \ldots, X_n)]$ is a deterministic function of $\boldsymbol{\theta}$. Note that since the X_n are iid,

$$\mathrm{E}[\ell(\boldsymbol{\theta}; X_1, \ldots, X_N)] = N \, \mathrm{E}[\ell(\boldsymbol{\theta}; X)],$$

where $X \sim f_X(\boldsymbol{x}; \boldsymbol{\theta}_0)$.

The key fact is that if the X_n are drawn as in (1), the the expecetd log likelihood peaks exactly where we need it to:

$$\arg\max_{\boldsymbol{\theta} \in \mathcal{T}} \mathrm{E}[\ell(\boldsymbol{\theta}; X_1, \ldots, X_N)] = \boldsymbol{\theta}_0. \tag{2}$$

To prove this, we start by noting that for any $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0$,

$$\mathrm{E}\left[\frac{f_X(X; \boldsymbol{\theta}_1)}{f_X(X; \boldsymbol{\theta}_0)}\right] \quad \leq \quad 1.$$

Indeed, if $\mathcal{X}_0$ is the support of the density $f(\boldsymbol{x}; \boldsymbol{\theta}_0)$,

$$\mathcal{X}_0 = \{\boldsymbol{x} \in \mathbb{R}^D \quad : \quad f_X(\boldsymbol{x}; \boldsymbol{\theta}_0) > 0\},$$

then

$$\begin{aligned}
\mathrm{E}\left[\frac{f_X(X;\boldsymbol{\theta}_1)}{f_X(X;\boldsymbol{\theta}_0)}\right] &= \int_{\boldsymbol{x}\in\mathcal{X}_0} \frac{f_X(\boldsymbol{x};\boldsymbol{\theta}_1)}{f_X(\boldsymbol{x};\boldsymbol{\theta}_0)}\, f_X(\boldsymbol{x};\boldsymbol{\theta}_0)\ \mathrm{d}\boldsymbol{x} \\
&= \int_{\boldsymbol{x}\in\mathcal{X}_0} f_X(\boldsymbol{x};\boldsymbol{\theta}_1)\ \mathrm{d}\boldsymbol{x} \\
&\leq 1.
\end{aligned}$$

Now consider any $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0$. We have

$$\begin{aligned}
\mathrm{E}[\ell(\boldsymbol{\theta}_1;X)] - \mathrm{E}[\ell(\boldsymbol{\theta}_0;X)] &= \mathrm{E}\left[\log\left(\frac{f_X(X;\boldsymbol{\theta}_1)}{f_X(X;\boldsymbol{\theta}_0)}\right)\right] \\
&\leq \mathrm{E}\left[\frac{f_X(X;\boldsymbol{\theta}_1)}{f_X(X;\boldsymbol{\theta}_0)} - 1\right], \quad \text{since } \log t \leq t-1, \\
&\leq 0.
\end{aligned}$$

Since $\log(t) = t-1$ only at $t=1$, the only way the first $\leq$ above holds with equality is if $\mathrm{P}\left(f_X(X;\boldsymbol{\theta}_1) = f_X(X;\boldsymbol{\theta}_0)\right) = 1$, which our identifiability condition forbids. Thus

$$\mathrm{E}[\ell(\boldsymbol{\theta}_1;X)] \ < \ \mathrm{E}[\ell(\boldsymbol{\theta}_0;X)] \quad \text{for all} \quad \boldsymbol{\theta}_1 \in \mathcal{T},\ \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0,$$

which establishes (2).

So if we could maximize the expected log likelihood, we would get exactly the right answer every time. Of course, we don't have access to this function; the MLE instead maximizes the sample mean:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}\in\mathcal{T}} \frac{1}{N}\sum_{n=1}^{N} \ell(\boldsymbol{\theta};\boldsymbol{x}_n).$$

By the weak law of large numbers,

$$\frac{1}{N}\sum_{n=1}^{N}\ell(\boldsymbol{\theta};\boldsymbol{x}_n) \;\to\; \mathrm{E}[\ell(\boldsymbol{\theta};X)], \quad \text{for all } \boldsymbol{\theta}\in\mathcal{T},$$

and so the maxima of the left- and right-hand sides will also eventually coincide.

We easily can make this perfectly precise when $\mathcal{T}$ is finite. In this case, we know that there is some $\gamma > 0$ such that

$$\mathrm{E}[\ell(\boldsymbol{\theta};X)] \le \mathrm{E}[\ell(\boldsymbol{\theta}_0;X)] - \gamma, \quad \text{for all } \boldsymbol{\theta}\neq\boldsymbol{\theta}_0.$$

Under the very mild assumption that $\mathrm{var}(\ell(\boldsymbol{\theta};X)) < \infty$ for all $\boldsymbol{\theta}$, the Chebyshev inequality tells us that for every $\epsilon > 0$ there exists an N such that

$$\mathrm{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N}\ell(\boldsymbol{\theta};\boldsymbol{x}_n) - \mathrm{E}[\ell(\boldsymbol{\theta};X)]\right| > \gamma/2\right) \;\le\; \frac{4\,\mathrm{var}(\ell(\boldsymbol{\theta};X))}{N\gamma^2},$$

and so

$$\begin{aligned}
&\mathrm{P}\left(\max_{\boldsymbol{\theta}\in\mathcal{T}}\left|\frac{1}{N}\sum_{n=1}^{N}\ell(\boldsymbol{\theta};\boldsymbol{x}_n) - \mathrm{E}[\ell(\boldsymbol{\theta};X)]\right| > \gamma/2\right)\\
&\quad\le \sum_{\boldsymbol{\theta}\in\mathcal{T}}\mathrm{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N}\ell(\boldsymbol{\theta};\boldsymbol{x}_n) - \mathrm{E}[\ell(\boldsymbol{\theta};X)]\right| > \gamma/2\right)\\
&\quad\le \sum_{\boldsymbol{\theta}\in\mathcal{T}}\frac{4\,\mathrm{var}(\ell(\boldsymbol{\theta};X))}{N\gamma^2}\\
&\quad\le \frac{4m|\mathcal{T}|}{N\gamma^2},
\end{aligned}$$

where $m = \max_{\boldsymbol{\theta} \in \mathcal{T}} \mathrm{var}(\ell(\boldsymbol{\theta}; X))$ and $|\mathcal{T}|$ is the number of elements in the set $\mathcal{T}$. The point is that this probability goes to 0 as N gets large, so that probability that

$$\frac{1}{N} \sum_{n=1}^{N} \ell(\boldsymbol{\theta}_0; \boldsymbol{x}_n) > \mathrm{E}[\ell(\boldsymbol{\theta}_0; X)] - \frac{\gamma}{2}$$

and

$$\frac{1}{N} \sum_{n=1}^{N} \ell(\boldsymbol{\theta}; \boldsymbol{x}_n) < \mathrm{E}[\ell(\boldsymbol{\theta}; X)] + \frac{\gamma}{2}, \quad \text{for all } \boldsymbol{\theta} \neq \boldsymbol{\theta}_0,$$

goes to 1. Thus for N large enough,

$$\frac{1}{N} \sum_{n=1}^{N} \ell(\boldsymbol{\theta}; \boldsymbol{x}_n) \; < \; \frac{1}{N} \sum_{n=1}^{N} \ell(\boldsymbol{\theta}_0; \boldsymbol{x}_n), \quad \text{for all } \boldsymbol{\theta} \neq \boldsymbol{\theta}_0,$$

and so

$$\mathrm{P}\left(\arg\max_{\boldsymbol{\theta} \in \mathcal{T}} \frac{1}{N} \sum_{n=1}^{N} \ell(\boldsymbol{\theta}; \boldsymbol{x}_m) = \arg\max_{\boldsymbol{\theta} \in \mathcal{T}} \mathrm{E}[\ell(\boldsymbol{\theta}; X)] = \boldsymbol{\theta}_0 \right) \to 1,$$

as $N \to \infty$. This establishes the consistency of the MLE.

When $\mathcal{T}$ is an infinite set, we need additional assumptions on the density functions $f_X(\boldsymbol{x}; \boldsymbol{\theta})$. Using arguments not too dissimilar from the above, you can show that if

1. the $f_X(\boldsymbol{x}; \boldsymbol{\theta})$ have the same support for all $\boldsymbol{\theta} \in \mathcal{T}$,
2. the $\mathrm{E}[\ell(\boldsymbol{\theta}; X)]$ are all differentiable, and
3. $\boldsymbol{\theta}_0$ is away from the boundary of $\mathcal{T}$,

then the MLE is consistent.

Efficiency

For unbiased estimators, where $\mathrm{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}_0$ and so

$$\mathrm{MSE}(\hat{\boldsymbol{\Theta}}) = \mathrm{var}(\hat{\boldsymbol{\Theta}}),$$

there is a **fundamental limit** on the MSE that no (unbiased) estimator can surpass. This is called the **Cramer-Rao lower bound**, and it is a massive result from classical statistics. Estimators that achieve this limit as the data size gets large are called **efficient**.

The key quantity in deriving this fundamental limit is the **score function**:

$$\boldsymbol{s}(\boldsymbol{\theta}; \boldsymbol{x}) = \nabla_\theta \ell(\boldsymbol{\theta}; \boldsymbol{x}).$$

As you can see, this is a vector-valued function of $\boldsymbol{\theta}$ that returns the gradient of the log likelihood at a fixed data point $\boldsymbol{x}$. By definition, the MLE looks for a value of $\boldsymbol{\theta}$ such that $\boldsymbol{s}(\boldsymbol{\theta}; \boldsymbol{x}) = \mathbf{0}$.

What is the behavior of this score function around the true value of the parameters $\boldsymbol{\theta} = \boldsymbol{\theta}_0$? As we would hope, the expectation of the score at $\boldsymbol{\theta}_0$ with respect to a random draw of $X \sim f_X(\boldsymbol{x}; \boldsymbol{\theta}_0)$ is indeed $\mathbf{0}$. To see this, we compute

$$\begin{aligned}
\mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}; X)] &= \int \nabla_\theta \ell(\boldsymbol{\theta}; X)\, f_X(\boldsymbol{x}; \boldsymbol{\theta}_0)\ \mathrm{d}\boldsymbol{x} \\
&= \int \nabla_\theta f_X(\boldsymbol{x}; \boldsymbol{\theta}) \frac{f_X(\boldsymbol{x}; \boldsymbol{\theta}_0)}{f_X(\boldsymbol{x}; \boldsymbol{\theta})}\ \mathrm{d}\boldsymbol{x} \quad \text{(chain rule)},
\end{aligned}$$

so[3]

$$\begin{aligned}
\mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0; X)] &= \int \nabla_\theta \, f_X(\boldsymbol{x}; \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \, \mathrm{d}\boldsymbol{x} \\
&= \nabla_\theta \left(\int f_X(\boldsymbol{x}; \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{x} \right) \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \nabla_\theta \, 1|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \mathbf{0}.
\end{aligned}$$

So no matter what $\boldsymbol{\theta}_0$ is, in expectation the score function crosses $\mathbf{0}$ at the right place.

Intuitively, how effectively we are able to estimate $\boldsymbol{\theta}_0$ will depend on the *variation* of the score function around $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. If $\boldsymbol{s}(\boldsymbol{\theta}; X)$ behaves very predictably around $\boldsymbol{\theta}_0$, the score function will pass though $\mathbf{0}$ in almost the same place for typical draws of the data, meaning that the data carry a lot of information about what $\boldsymbol{\theta}_0$ actually is. If $\boldsymbol{s}(\boldsymbol{\theta}; X)$ has wide variation around $\boldsymbol{\theta}_0$, then we will have a harder time estimating $\boldsymbol{\theta}_0$ from the data.

Mathematically, this variation is quantified with the covariance matrix of the score function:

$$\begin{aligned}
\boldsymbol{J}(\boldsymbol{\theta}_0) &= \mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0; X)\boldsymbol{s}(\boldsymbol{\theta}_0; X)^{\mathrm{T}}] \\
&= \mathrm{E}[\nabla_\theta \log f_X(X; \boldsymbol{\theta}_0) \nabla_\theta \log f_X(X; \boldsymbol{\theta}_0)^{\mathrm{T}}]
\end{aligned}$$

This is called the **Fisher information matrix** at $\boldsymbol{\theta}_0$, and as we see below it is indeed what determines our fundamental limit.

[3] Switching the integral and the gradient here actually requires some very mild smoothness conditions on how the densities change as a function of $\boldsymbol{\theta}$.

Let $\hat{\boldsymbol{\Theta}}$ be any unbiased estimator, $\mathrm{E}[\hat{\boldsymbol{\Theta}}] = \boldsymbol{\theta}_0$, that is formed from a *single sample*:

$$\hat{\boldsymbol{\Theta}} = g(X_1),$$

where $g : \mathbb{R}^D \to \mathbb{R}^P$ is the mapping that actually implements the estimation algorithm. After deriving a bound on the performance of such single-sample estimators, we will quickly be able to generalize to multiple samples.

It is a fact that the cross-correlation between the error $\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}_0$ and the score vector at $\boldsymbol{\theta}_0$ is the identity:

$$\mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0; X)(\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}}] = \mathbf{I}.$$

To see this, we write down what is means for $\hat{\boldsymbol{\Theta}}$ to be unbiased[4],

$$\mathrm{E}[(\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}}] = \int f_X(\boldsymbol{x}; \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}} \ \mathrm{d}\boldsymbol{x} = \mathbf{0}^{\mathrm{T}},$$

meaning

$$\int f_X(\boldsymbol{x}; \boldsymbol{\theta}_0)\hat{\boldsymbol{\theta}}^{\mathrm{T}} \ \mathrm{d}\boldsymbol{x} = \int f_X(\boldsymbol{x}; \boldsymbol{\theta}_0)\boldsymbol{\theta}_0^{\mathrm{T}} \ \mathrm{d}\boldsymbol{x}.$$

Taking the gradient of both sides yields

$$\nabla_\theta \left(\int f_X(\boldsymbol{x}; \boldsymbol{\theta})\hat{\boldsymbol{\theta}}^{\mathrm{T}} \ \mathrm{d}\boldsymbol{x} \right) \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \nabla_\theta \left(\int f_X(\boldsymbol{x}; \boldsymbol{\theta})\boldsymbol{\theta}^{\mathrm{T}} \ \mathrm{d}\boldsymbol{x} \right) \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}. \tag{3}$$

[4]Note that $\hat{\boldsymbol{\Theta}}$ is a function of the data — inside the integral here, this is a deterministic function of $\boldsymbol{x}$, so we write it lowercase $\hat{\boldsymbol{\theta}}$. We are not explicitly denoting the dependence of $\hat{\boldsymbol{\theta}}$ on $\boldsymbol{x}$, but it is there.

Switching[5] the gradient and the integral and applying the chain rule to the right-hand side above yields

$$\begin{aligned}
&\nabla_\theta \left(\int f_X(\boldsymbol{x};\boldsymbol{\theta})\boldsymbol{\theta}^{\mathrm{T}} \ \mathrm{d}\boldsymbol{x} \right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&\quad = \int \nabla_\theta \left(f_X(\boldsymbol{x};\boldsymbol{\theta})\boldsymbol{\theta}^{\mathrm{T}} \right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \ \mathrm{d}\boldsymbol{x} \\
&\quad = \int \nabla_\theta (f_X(\boldsymbol{x};\boldsymbol{\theta}))\boldsymbol{\theta}^{\mathrm{T}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \ \mathrm{d}\boldsymbol{x} + \int f_X(\boldsymbol{x};\boldsymbol{\theta})\nabla_\theta(\boldsymbol{\theta}^{\mathrm{T}})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \ \mathrm{d}\boldsymbol{x} \\
&\quad = \int \nabla_\theta (f_X(\boldsymbol{x};\boldsymbol{\theta}_0))\boldsymbol{\theta}_0^{\mathrm{T}} \ \mathrm{d}\boldsymbol{x} + \int f_X(\boldsymbol{x};\boldsymbol{\theta})\mathbf{I} \ \mathrm{d}\boldsymbol{x} \\
&\quad = \int \nabla_\theta (f_X(\boldsymbol{x};\boldsymbol{\theta}_0))\boldsymbol{\theta}_0^{\mathrm{T}} \ \mathrm{d}\boldsymbol{x} + \mathbf{I}.
\end{aligned}$$

Returning to (3), pulling the gradient inside the integral on the left hand side and combing with the above yields

$$\int \nabla_{\boldsymbol{\theta}} f_X(\boldsymbol{x};\boldsymbol{\theta}_0) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right)^{\mathrm{T}} \mathrm{d}\boldsymbol{x} = \mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0; X)(\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}}] = \mathbf{I}.$$

The pieces are now in place for our lower bound. Again, let $\hat{\boldsymbol{\Theta}}$ be any unbiased estimator, and denote its covariance matrix as

$$\hat{\boldsymbol{R}} = \mathrm{E}[(\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}}].$$

Recall that the MSE of $\hat{\boldsymbol{\Theta}}$ is

$$\mathrm{MSE}(\hat{\boldsymbol{\Theta}}) = \mathrm{E}[\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}_0\|_2^2] = \mathrm{trace}(\hat{\boldsymbol{R}}).$$

Now let Q be the random vector

$$Q = \begin{bmatrix} \hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}_0 \\ \boldsymbol{s}(\boldsymbol{\theta}_0; X) \end{bmatrix}.$$

[5] Again, we need regularity conditions on the densities to make this completely valid.

The covariance for Q is then

$$\mathrm{E}[QQ^{\mathrm{T}}] = \begin{bmatrix} \hat{\boldsymbol{R}} & \mathbf{I} \\ \mathbf{I} & \boldsymbol{J}(\boldsymbol{\theta}_0). \end{bmatrix}$$

Since $\mathrm{E}[QQ^{\mathrm{T}}]$ is symmetric positive semi-definite, so is

$$\begin{bmatrix} \mathbf{I} & -\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{R}} & \mathbf{I} \\ \mathbf{I} & \boldsymbol{J}(\boldsymbol{\theta}_0). \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{R}} - \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{J}(\boldsymbol{\theta}_0). \end{bmatrix}$$

Block diagonal matrices are symmetric positive semi-definite if and only if their blocks are sym+semi-def. Thus

$$\hat{\boldsymbol{R}} - \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1} \quad \text{is symmetirc positive semi-definite,}$$

which we write as

$$\hat{\boldsymbol{R}} \succeq \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}.$$

In particular,

$$\mathrm{MSE}(\hat{\boldsymbol{\Theta}}) = \mathrm{trace}(\hat{\boldsymbol{R}}) \geq \mathrm{trace}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1})$$

This result is called the **Cramer-Rao lower bound**.

The analysis for estimates formed from multiple independent samples is not too different. Indeed, we could treat this problem by just drawing a single random vector in $\mathbb{R}^{DN}$, where the vector can be partitioned into N independent vectors of length D. In the end, the analysis goes precisely as the above, only with the score function

$$\begin{aligned} \boldsymbol{s}(\boldsymbol{\theta}; X_1, \ldots, X_N) &= \nabla_\theta \ell(\boldsymbol{\theta}; X_1, \ldots, X_N) \\ &= \sum_{n=1}^{N} \nabla_\theta \ell(\boldsymbol{\theta}; X_n). \end{aligned}$$

The the Fisher information matrix becomes

$$\begin{aligned}
\boldsymbol{J}_N(\boldsymbol{\theta}_0) &= \mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0; X_1, \ldots, X_N)\boldsymbol{s}(\boldsymbol{\theta}_0; X_1, \ldots, X_N)^{\mathrm{T}}] \\
&= \sum_{n=1}^{N}\sum_{m=1}^{N} \mathrm{E}[\nabla_\theta \ell(\boldsymbol{\theta}_0; X_n)\nabla_\theta \ell(\boldsymbol{\theta}_0; X_m)^{\mathrm{T}}] \\
&= \sum_{n=1}^{N} \mathrm{E}[\nabla_\theta \ell(\boldsymbol{\theta}_0; X_n)\nabla_\theta \ell(\boldsymbol{\theta}_0; X_n)^{\mathrm{T}}] \\
&= N\boldsymbol{J}(\boldsymbol{\theta}_0).
\end{aligned}$$

Thus any unbiased estimator $\hat{\boldsymbol{\Theta}}_N$ that works from N samples must have a covariance matrix $\hat{\boldsymbol{R}}_N$ that obeys

$$\mathrm{E}[(\hat{\boldsymbol{\Theta}}_N - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\Theta}}_N - \boldsymbol{\theta}_0)^{\mathrm{T}}] = \hat{\boldsymbol{R}}_N \succeq \frac{1}{N}\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}.$$

Cramer-Rao Lower Bound. Let $X_1, \ldots, X_N$ be iid random vectors in $\mathbb{R}^D$ with distribution

$$X_n \sim f_X(\boldsymbol{x}; \boldsymbol{\theta}_0).$$

Let $\hat{\boldsymbol{\Theta}}_N = g(X_1, \ldots, X_N) : \mathbb{R}^{D \times N} \to \mathbb{R}^D$ be any estimator with $\mathrm{E}[\hat{\boldsymbol{\Theta}}_N] = \boldsymbol{\theta}_0$. Then

$$\mathrm{E}[(\hat{\boldsymbol{\Theta}}_N - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\Theta}}_N - \boldsymbol{\theta}_0)^{\mathrm{T}}] = \hat{\boldsymbol{R}}_N \succeq \frac{1}{N} \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1},$$

and in particular

$$\mathrm{MSE}(\hat{\boldsymbol{\Theta}}) = \mathrm{E}[\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}_0\|_2^2] \geq \frac{1}{N} \operatorname{trace}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}),$$

where

$$\boldsymbol{J}(\boldsymbol{\theta}_0) = \mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0; X) \boldsymbol{s}(\boldsymbol{\theta}_0; X)^{\mathrm{T}}]$$

Example: Let's return to our binary random variable example, where

$$X = \begin{cases} 1, & \text{with probability } \theta_0, \\ 0, & \text{with probability } 1 - \theta_0. \end{cases}$$

Recall that with

$$\hat{\Theta}_N = \frac{1}{N} \sum_{n=1}^{N} X_n$$

we have

$$\mathrm{var}(\Theta_N) = \frac{\theta_0(1 - \theta_0)}{N}.$$

We compute the Fisher information matrix (actually scalar in this case) as follows. We have

$$\ell(x;\theta) = \theta^x(1-\theta)^{1-x}, \quad \nabla_\theta \ell(x;\theta) = \frac{x}{\theta} - \frac{(1-x)}{(1-\theta)},$$

(since x is a scalar in this case, the gradient above is really just a regular derivative). Then

$$\begin{aligned} J(\theta) &= \mathrm{E}\left[\left(\frac{X}{\theta} - \frac{(1-X)}{(1-\theta)}\right)^2\right] \\ &= \frac{\mathrm{E}[X^2(1-\theta)^2 - 2X(1-X)\theta(1-\theta) + \theta^2(1-X)^2]}{\theta^2(1-\theta)^2} \\ &= \frac{\theta(1-\theta)^2 + \theta^2(1-\theta)}{\theta^2(1-\theta)^2} \\ &= \frac{1}{\theta(1-\theta)}. \end{aligned}$$

Thus[6]

$$\frac{1}{N}\,\mathrm{trace}(J(\theta_0)^{-1}) = \frac{\theta_0(1-\theta_0)}{N}.$$

This matches the variance, so the sample mean (which is also the MLE) is efficient in this case.

[6] Again, $J(\theta_0)$ is a scalar here, so $\mathrm{trace}(J(\theta_0)^{-1}) = 1/J(\theta_0)$.

Efficiency of the MLE. We will not prove the efficiency of the MLE. Rather we will simply state a more general and very powerful result. Under smoothness conditions on the expected log likelihood $\mathrm{E}[\ell(\boldsymbol{\theta}; X)]$ in the neighborhood of $\boldsymbol{\theta}_0$ (which mostly say that $\mathrm{E}[\ell(\boldsymbol{\theta}; X)]$ has three well-behaved derivatives), the distribution of the error in the MLE becomes Gaussian:

$$(\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}_0) \to Z \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}/N).$$

This means that the MLE is asymptotically unbiased, and that its MSE approaches the Cramer-Rao lower bound.

If you want more detailed proofs for the asymptotics of the MLE, a good reference is:
L. Wasserman, *All of Statistics*, 2nd ed, Springer, 2010.
http://www.stat.cmu.edu/~larry/all-of-statistics/

Technical Details: Weak Law of Large Numbers

In this section, we show that under very mild conditions, the sample mean converges to the true mean. The only condition is that the underlying distribution has finite variance.

We start by stating the main result precisely. Let X be a random variable with pdf $f_X(x)$, mean $\mathrm{E}[X] = \mu$, and variance $\mathrm{var}(X) = \sigma^2 < \infty$. We observe *samples* of X labeled $X_1, X_2, \ldots, X_N$. The X_i are independent of one another, and they all have the same distribution as X. We will show that the sample mean formed from a sample of size N:

$$M_N = \frac{1}{N}(X_1 + X_2 + \cdots + X_N),$$

obeys

$$\mathrm{P}\left(|M_N - \mu| > \epsilon\right) \;\leq\; \frac{\sigma^2}{N\epsilon^2},$$

where $\epsilon > 0$ is an arbitrarily small number. In the expression above, M_N is the only thing which is random; μ and σ^2 are fixed underlying properties of the distribution, N is the amount of data we see, and ϵ is something we can choose arbitrarily.

Notice that no matter how small ϵ is, the probability on the right hand side above goes to zero as $N \to \infty$. That is, for any fixed $\epsilon > 0$,

$$\lim_{N\to\infty} \mathrm{P}\left(|M_N - \mu| > \epsilon\right) = 0.$$

This result is follows from two simple but important tools known as the *Markov* and *Chebyshev* inequalities.

Markov inequality

Let X be a random variable that only takes positive values:

$$f_X(x) = 0, \quad \text{for} \;\; x < 0, \quad \text{or} \quad F_X(0) = 0.$$

Then

$$\boxed{\mathrm{P}\left(X \geq a\right) \;\leq\; \frac{\mathrm{E}[X]}{a} \quad \text{for all } a > 0.}$$

For example, the probability that X is more than 5 times its mean is 1/5, 10 times the mean is 1/10, etc. And this holds for **any distribution**.

The Markov inequality is easy to prove:

$$\begin{aligned}
\mathrm{E}[X] &= \int_0^\infty x f_X(x) \; \mathrm{d}x \\
&\geq \int_a^\infty x f_X(x) \; \mathrm{d}x \\
&\geq \int_a^\infty a f_X(x) \; \mathrm{d}x \\
&= a \cdot \mathrm{P}\left(X \geq a\right)
\end{aligned}$$

and so $P(X \geq a) \leq \frac{\mathrm{E}[X]}{a}$.

Again, this is a very general statement in that we have assumed nothing about X other than it is positive. The price for the generality is that the bound is typically very loose, and does not usually capture the behavior of $\mathrm{P}\left(X \geq a\right)$. We can, however, cleverly apply the Markov inequality to get something slightly more useful.

Chebyshev inequality

The main use of the Markov inequality turns out to be its use in deriving other, more accurate deviation inequalities. Here we will use it to derive the **Chebyshev inequality**, from which the weak law of large numbers will follow immediately.

Chebyshev inequality: If X is a random variable with mean μ and variance σ^2, then

$$\boxed{\mathrm{P}\left(|X-\mu|>c\right) \;\leq\; \frac{\sigma^2}{c^2} \quad \text{for all } c>0.}$$

The Chebyshev inequality follows immediately from the Markov inequality in the following way. No matter what range of values X takes, the quantity $|X-\mu|^2$ is always positive. Thus

$$\mathrm{P}\left(|X-\mu|^2>c^2\right) \;\leq\; \frac{\mathrm{E}[|X-\mu|^2]}{c^2}=\frac{\sigma^2}{c^2}.$$

Since squaring $(\cdot)^2$ is monotonic (invertible) over positive numbers,

$$\mathrm{P}\left(|X-\mu|^2>c^2\right)=\mathrm{P}\left(|X-\mu|>c\right) \;\leq\; \frac{\sigma^2}{c^2}.$$

We now have a bound which depends on the mean and the variance of X; this leads to a more accurate approximation of the probability.

The weak law of large numbers (WLLN)

We now turn to the behavior of the the sample mean

$$M_N = \frac{X_1 + X_2 + \cdots + X_N}{N},$$

where again the X_i are iid random variables with $\mathrm{E}[X_i] = \mu$ and $\operatorname{var} X_i = \sigma^2$. We know that

$$\mathrm{E}[M_N] = \frac{\mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_N]}{N} = \frac{N\mu}{N} = \mu,$$

and since the X_i are independent,

$$\operatorname{var}(M_N) = \frac{\operatorname{var}(X_1) + \operatorname{var}(X_2) + \cdots + \operatorname{var}(X_N)}{N^2} = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}.$$

For any $\epsilon > 0$, a direct application of the Chebyshev inequality tells us that

$$\mathrm{P}\left(|M_N - \mu| > \epsilon\right) \ \leq \ \frac{\sigma^2}{N\epsilon^2}.$$

The point is that this gets arbitrarily small as $N \to \infty$ no matter what ϵ was chosen to be. We have established, in some sense, that even though $\{M_N\}$ is a sequence of random numbers, it converges to something deterministic, namely μ.

WLLN: Let $X_1, X_2, \ldots$ be iid random variables as above. For **every** $\epsilon > 0$, we have

$$\mathrm{P}\left(|M_N - \mu| > \epsilon\right) = \mathrm{P}\left(\left|\frac{X_1 + \cdots + X_N}{N} - \mu\right| > \epsilon\right) \longrightarrow 0,$$

as $N \to \infty$.

One of the philosophical consequences of the WLLN is that it tells us that probabilities can be estimated through **empirical frequencies**. Suppose I want to estimate the probability of and event A occurring related to some probabilistic experiment. We run a series of (independent) experiments, and set $X_i = 1$ if A occurred in experiment i, and $X_i = 0$ otherwise. Then given $X_1, \ldots, X_N$, we estimate the probability of A in a completely reasonable way, by computing the percentage of times it occurred:

$$p_{\text{empirical}} = \frac{X_1 + \cdots + X_N}{N}.$$

The WLLN tells us that

$$p_{\text{empirical}} \to \mathrm{P}\left(A\right), \quad \text{as } N \to \infty.$$

This lends some mathematical weight to our interpretation of probabilities as *relative frequencies*.

All of the above of course applies to functions of random variables. That is, if X is a random variable, and $g(X)$ is a function of that random variable with

$$\mathrm{var}(g(X)) = \mathrm{E}[(g(X) - \mathrm{E}[g(X)])^2] < \infty,$$

then given independent realizations $X_1, \ldots, X_N$, we have

$$\frac{1}{N} \sum_{n=1}^{N} g(X_n) \to \mathrm{E}[g(X)]$$

as $N \to \infty$.