Topic 4: Parameter Estimation

Instructor: Daniel L. Pimentel-Alarcón

4.1 Introduction

In many applications we observe data x drawn according to a distribution p with an unknown parameter $\theta^\star$, and we want to *learn* or *estimate* such parameter. We write this as

$$x \sim \mathsf{p}(\mathrm{x}|\theta^\star), \quad \theta^\star \in \Theta.$$

In other words, we want to find the parameter $\theta^\star$ that generated our data x.

Example 4.1 (Psichokinesis)**.** Imagine controlling things with your mind. For example, skipping a song by just thinking about it. One way to do this is by putting D sensors in your head (in your headphones, for example). These sensors would record small voltages generated by your brain, store them in a vector $\boldsymbol{x} \in \mathbb{R}^{\mathrm{D}}$, and send them to a machine (phone, computer, server, etc.). The machine should interpret $\boldsymbol{x}$ and skip the song if that is what you thought about (see Figure 4.1).

We can expect that whenever you think of skipping a song, $\boldsymbol{x}$ will be composed of an *skip* signal $\boldsymbol{\mu}_{\blacktriangleright\!\!\blacktriangleright}^\star \in \mathbb{R}^{\mathrm{D}}$ plus other stuff, i.e.,

$$\boldsymbol{x} = \boldsymbol{\mu}_{\blacktriangleright\!\!\blacktriangleright}^\star + \boldsymbol{\eta},$$

where $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ represents noise. If we knew $\boldsymbol{\mu}_{\blacktriangleright\!\!\blacktriangleright}^\star$, we could setup a hypothesis test, similar to Example 3.6:

$$\begin{aligned} H_0: \ \boldsymbol{x} &\sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) && \Rightarrow \text{do nothing}, \\ H_1: \ \boldsymbol{x} &\sim \mathcal{N}(\boldsymbol{\mu}_{\blacktriangleright\!\!\blacktriangleright}^\star, \sigma^2\mathbf{I}) && \Rightarrow \text{skip song}. \end{aligned}$$

However, we do not know $\boldsymbol{\mu}_{\blacktriangleright\!\!\blacktriangleright}^\star$, and hence we want to *learn* it, or more precisely, *estimate* it. To this end, we can put the sensors in your head, ask you to *think* of skipping a song, and record the response vector $\boldsymbol{x}$. We can repeat this experiment N times to obtain i.i.d. samples $\{\mathbf{x}_{\mathrm{i}}\}_{\mathrm{i}=1}^{\mathrm{N}}$ according to $\mathcal{N}(\boldsymbol{\mu}_{\blacktriangleright\!\!\blacktriangleright}^\star, \sigma^2\mathbf{I})$. We don't know $\boldsymbol{\mu}_{\blacktriangleright\!\!\blacktriangleright}^\star$, but we know that $\{\mathbf{x}_{\mathrm{i}}\}_{\mathrm{i}=1}^{\mathrm{N}}$ were generated according to $\mathcal{N}(\boldsymbol{\mu}_{\blacktriangleright\!\!\blacktriangleright}^\star, \sigma^2\mathbf{I})$, so the idea is to find the $\boldsymbol{\mu}_{\blacktriangleright\!\!\blacktriangleright}^\star$ that most likely generated $\{\mathbf{x}_{\mathrm{i}}\}_{\mathrm{i}=1}^{\mathrm{N}}$.

In this example, $\mathsf{p}(\mathbf{x}|\theta^\star)$ is the gaussian probability density function with parameters $\theta^\star = \{\boldsymbol{\mu}_{\blacktriangleright\!\!\blacktriangleright}^\star, \sigma^2\}$, and we want to estimate $\boldsymbol{\mu}_{\blacktriangleright\!\!\blacktriangleright}^\star$.

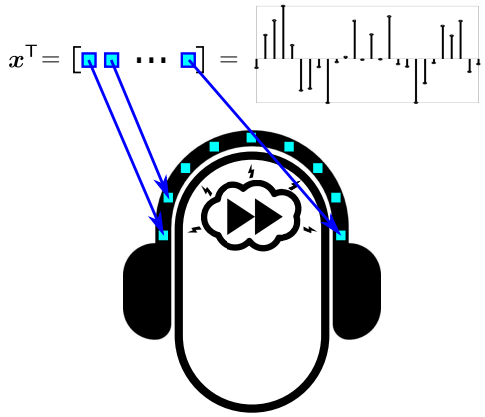

Figure 4.1: Sensors record small voltages generated by your brain and store them in a *signal* vector $\boldsymbol{x} \in \mathbb{R}^{\mathrm{D}}$. A machine (phone, computer, server, etc.) should interpret $\boldsymbol{x}$ and *skip* the song if that is what you thought about. See Example 4.1.

Remark 4.1 (Notation)**.** In general, we will use * to denote the *true* parameter that we want to estimate (e.g., $\theta^\star$), and $\hat{}$ to denote an estimator (e.g., $\hat{\theta}$).

4.2 Maximum Likelihood

Again, suppose we are given data $x \sim \mathsf{p}(\mathrm{x}|\theta^\star)$, and we want to estimate $\theta^\star$. One way to do this is by finding the parameter $\theta \in \Theta$ that *most likely* generated our data x.

Definition 4.1 (Maximum likelihood estimator (MLE))**.**

$$\hat{\theta}_{ML} \;:=\; \arg\max_{\theta \in \Theta} \; \mathsf{p}(\mathrm{x}|\theta).$$

Remark 4.2 (Likelihood)**.** The term *likelihood* is often a source of confusion. Recall that in an estimation problem we are given an *instance* of a random variable, i.e., we observe data $x = \mathrm{x}$ drawn according to some probability distribution $\mathsf{p}(\mathrm{x}|\theta^\star)$, and we want to find a parameter $\theta \in \Theta$ that *likely* generated x. The *likelihood* is nothing more than the *probability* that some θ is the parameter that generated x. In other words, the likelihood is $\mathsf{p}(\mathrm{x}|\theta)$ [evaluated at $x = \mathrm{x}$], and thinking of θ as the variable.

Example 4.2. Let $x_1, \ldots, x_{\mathrm{N}} \overset{iid}{\sim} \mathcal{N}(\mu^\star, \sigma^2)$ with σ^2 known and $\mu^\star$ unknown. To find $\hat{\mu}_{ML}$ we use elemental techniques from optimization: take the derivative, set to zero, and solve for the desired

parameter. First observe that since $\mathsf{p} > 0$,

$$\begin{aligned}\hat{\mu}_{ML} &= \arg\max_{\mu\in\mathbb{R}} \mathsf{p}(\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{N}|\mu) = \arg\max_{\mu\in\mathbb{R}} \log \mathsf{p}(\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{N}|\mu) = \arg\max_{\mu\in\mathbb{R}} \log\left(\prod_{\mathrm{i}=1}^{\mathrm{N}} \mathsf{p}(\mathrm{x}_\mathrm{i}|\mu)\right)\\ &= \arg\max_{\mu\in\mathbb{R}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathsf{p}(\mathrm{x}_\mathrm{i}|\mu) = \arg\max_{\mu\in\mathbb{R}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\mathrm{x}_\mathrm{i}-\mu}{\sigma}\right)^2}\right)\\ &= \arg\max_{\mu\in\mathbb{R}} \underbrace{\sum_{\mathrm{i}=1}^{\mathrm{N}} \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)}_{\text{constant}} + \sum_{\mathrm{i}=1}^{\mathrm{N}} \log\left(e^{-\frac{1}{2}\left(\frac{\mathrm{x}_\mathrm{i}-\mu}{\sigma}\right)^2}\right) = \arg\max_{\mu\in\mathbb{R}} -\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{1}{2}\left(\frac{\mathrm{x}_\mathrm{i}-\mu}{\sigma}\right)^2.\end{aligned}$$

Taking derivative with respect to (w.r.t.) μ we have

$$-\frac{\partial}{\partial\mu}\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{1}{2}\left(\frac{\mathrm{x}_\mathrm{i}-\mu}{\sigma}\right)^2 = \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\mathrm{x}_\mathrm{i}-\mu}{\sigma}.$$

Setting to zero and solving for μ we obtain:

$$\hat{\mu}_{ML} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x}_\mathrm{i}.$$

4.3 Bias and Variance

Notice that an estimator $\hat{\theta}$ is a function of the observed data $x = \mathrm{x}$. But since x is a random variable, our estimator will also have some variability. Depending on the particular x that we observe, sometimes our estimator may be better or worse. Hence it is often useful to use the mean (bias) and variance as a measures of the stability of our estimator.

Definition 4.2 (Bias, variance)**.** The *bias* and *variance* of an estimator $\hat{\boldsymbol{\theta}} \in \mathbb{R}^\mathrm{K}$ is defined as

$$\begin{aligned}\mathsf{bias}(\hat{\boldsymbol{\theta}}) &:= \boldsymbol{\theta}^\star - \mathsf{E}[\hat{\boldsymbol{\theta}}],\\ \mathsf{var}(\hat{\boldsymbol{\theta}}) &:= \mathsf{E}\left[\left\|\hat{\boldsymbol{\theta}} - \mathsf{E}[\hat{\boldsymbol{\theta}}]\right\|_2^2\right],\end{aligned}$$

where the expectations are w.r.t. the *true* probability distribution of x, i.e., w.r.t. $\boldsymbol{\theta}^\star$. If $\mathsf{bias}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ we say that $\hat{\boldsymbol{\theta}}$ is *unbiased*.

Intuitively, the bias tells us how close we will get to the true parameter on average. The variance tells us how far we expect to be from that average. If an estimator is unbiased but has large variance, it is likely to be inaccurate for our particular sample $x = \mathrm{x}$. Similarly, if an estimator has low variance but is biased, we expect it to be consistently inaccurate (see Figure 4.2). Ideally, we want an unbiased estimator with low variance, meaning that we expect it to be close to the true parameter.

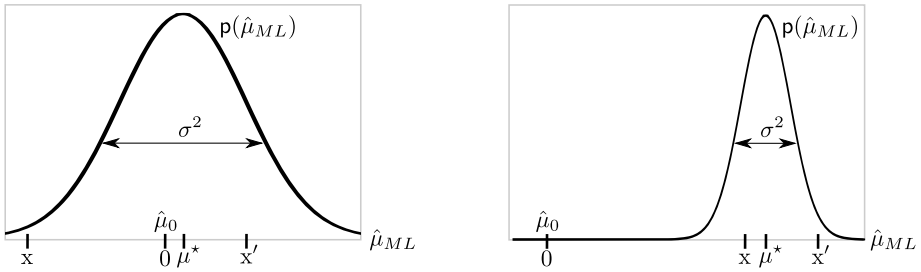

Figure 4.2: An estimator is generally a function of the data. For instance, consider $\hat{\mu}_{ML} = x$ from Example 4.3. Since $x \sim \mathcal{N}(\mu^\star, \sigma^2)$, so does $\hat{\mu}_{ML}$. For example, $\hat{\mu}_{ML}$ could take the value x or x′. **Left**: If σ is large, we can expect $\hat{\mu}_{ML}$ to be far from $\mu^\star$. In this case the *constant* estimator $\hat{\mu}_0$ (which has zero variance) might be better than $\hat{\mu}_{ML}$. **Right**: If σ is small, we can expect $\hat{\mu}_{ML}$ to be close to $\mu^\star$. In this case, $\hat{\mu}_{ML}$ would be better than $\hat{\mu}_0$.

Example 4.3. Let $x \sim \mathcal{N}(\mu^\star, \sigma^2)$, and consider two estimators: the constant estimator $\hat{\mu}_0 = 0$ and the MLE $\hat{\mu}_{ML} = x$. Then

$$\begin{aligned}
\mathsf{bias}(\hat{\mu}_0) &= \mu^\star - \mathsf{E}[\hat{\mu}_0] = \mu^\star - \mathsf{E}[0] = \mu^\star - 0 = \mu^\star, \\
\mathsf{var}(\hat{\mu}_0) &= \mathsf{E}\left[(\hat{\mu}_0 - \mathsf{E}[\hat{\mu}_0])^2\right], = \mathsf{E}\left[(0-0)^2\right] = 0,
\end{aligned}$$

$$\begin{aligned}
\mathsf{bias}(\hat{\mu}_{ML}) &= \mu^\star - \mathsf{E}[\hat{\mu}_{ML}] = \mu^\star - \mathsf{E}[x] = \mu^\star - \mu^\star = 0, \\
\mathsf{var}(\hat{\mu}_{ML}) &= \mathsf{E}\left[(\hat{\mu}_{ML} - \mathsf{E}[\hat{\mu}_{ML}])^2\right] = \mathsf{E}\left[(x-\mu^\star)^2\right] = \mathsf{var}(x) = \sigma^2.
\end{aligned}$$

If $\mu^\star$ is small, $\hat{\mu}_0$ will always be close to $\mu^\star$, regardless of the particular x that we observe. In addition, if σ is big, depending on the particular x that we observe, $\hat{\mu}_{ML}$ could be far from $\mu^\star$. In this case $\hat{\mu}_0$ would be better. For similar arguments, if $\mu^\star$ is big and σ is small, $\hat{\mu}_{ML}$ would be better. See Figure 4.2 to build some intuition.

4.4 Error/Loss and Risk

Example 4.3 shows that some estimators are more accurate than others. As we saw, the bias and variance are good indicators of an estimator's performance. As we will see, these are particular cases of a more general way to measure how good an estimator is: the *risk*. To define it, we first need to introduce the concept of *error*, often also known as *loss*.

Definition 4.3 (Error/loss)**.** An *error* or *loss* function is a mapping $\ell : \Theta \to \mathbb{R}_+$ that measures the distance between $\theta^\star$ and $\hat{\theta}$.

Example 4.4. Here are some common loss functions, with $\Theta = \mathbb{R}^{\mathrm{K}}$:

- 0/1 loss:

$$\ell_0(\hat{\boldsymbol{\theta}}) \;:=\; \mathbb{1}_{\{\hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^\star\}} \;=\; \begin{cases} 0 & \text{if} \quad \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star, \\ 1 & \text{if} \quad \hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^\star. \end{cases}$$

- ℓ_1 loss (absolute error):

$$\ell_1(\hat{\boldsymbol{\theta}}) \;:=\; \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_1 \;=\; \sum_{\mathrm{k}=1}^{\mathrm{K}} |\hat{\theta}_\mathrm{k} - \theta^\star_\mathrm{k}|.$$

- ℓ_2 loss (squared error):

$$\ell_2(\hat{\boldsymbol{\theta}}) \;:=\; \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2^2 \;=\; (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\mathsf{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) \;=\; \sum_{\mathrm{k}=1}^{\mathrm{K}} (\hat{\theta}_\mathrm{k} - \theta^\star_\mathrm{k})^2.$$

- ℓ_l *log*-probability:

$$\ell_l(\hat{\boldsymbol{\theta}}) \;:=\; -\log \mathsf{p}(\mathbf{x}|\hat{\boldsymbol{\theta}}),$$

  which measures the distance from $\boldsymbol{\theta}^\star$ because $\boldsymbol{x}$ is truly distributed $\mathsf{p}(\mathbf{x}|\boldsymbol{\theta}^\star)$. We can interpret $\hat{\theta}$ as $\arg\min_\theta \ell_l(\theta)$

Recall that an estimator $\hat{\theta}$ is a function of the observed data $x = \mathrm{x}$. But since x is a random variable, our estimator will also have some variability. Depending on the particular x that we observe, sometimes our estimate may be better or worse. Hence a better indicator of our estimator's performance is the *expected* error/loss, also known as *risk*.

Definition 4.4 (Risk)**.**

$$\mathsf{Risk}(\hat{\theta}) \;:=\; \mathsf{E}\left[\ell(\hat{\theta})\right].$$

Example 4.5. Continuing with Example 4.4,

$$\begin{aligned}
\mathsf{Risk}_0(\hat{\boldsymbol{\theta}}) \;&:=\; \mathsf{E}\left[\ell_0(\hat{\boldsymbol{\theta}})\right] \;=\; \mathsf{E}\left[\mathbb{1}_{\{\hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^\star\}}\right] \;=\; \mathsf{P}(\hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^\star), \\
\mathsf{Risk}_1(\hat{\boldsymbol{\theta}}) \;&:=\; \mathsf{E}\left[\ell_1(\hat{\boldsymbol{\theta}})\right] \;=\; \mathsf{E}\left[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_1\right], \\
\mathsf{Risk}_2(\hat{\boldsymbol{\theta}}) \;&:=\; \mathsf{E}\left[\ell_2(\hat{\boldsymbol{\theta}})\right] \;=\; \mathsf{E}\left[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2^2\right] \;=:\; \mathsf{MSE}(\hat{\boldsymbol{\theta}}), \\
\mathsf{Risk}_l(\hat{\boldsymbol{\theta}}) \;&:=\; \mathsf{E}\left[\ell_l(\hat{\boldsymbol{\theta}})\right] \;=\; \mathsf{E}\left[-\log \mathsf{p}(\mathbf{x}|\hat{\boldsymbol{\theta}})\right],
\end{aligned}$$

where MSE stands for *mean squared error*.

The next proposition shows that the bias and variance have a tight relation with a particular kind of risk, namely with $\mathsf{Risk}_2 = \mathsf{MSE}$.

Proposition 4.1 ($\mathsf{MSE} = \mathsf{bias}^2 + \mathsf{var}$)**.**

$$\mathsf{MSE}(\hat{\boldsymbol{\theta}}) \;=\; \mathsf{bias}(\hat{\boldsymbol{\theta}})^{\mathsf{T}}\mathsf{bias}(\hat{\boldsymbol{\theta}}) \;+\; \mathsf{var}(\hat{\boldsymbol{\theta}}).$$

Proof. Write

$$\begin{aligned}
\mathsf{MSE}(\hat{\boldsymbol{\theta}}) \;&=\; \mathsf{E}\left[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2^2\right] \;=\; \mathsf{E}\Big[\|(\hat{\boldsymbol{\theta}} \underbrace{-\mathsf{E}[\hat{\boldsymbol{\theta}}]}_{0}) + (\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star)\|_2^2\Big] \\
&= \mathsf{E}\left[\Big((\hat{\boldsymbol{\theta}} - \mathsf{E}[\hat{\boldsymbol{\theta}}]) + (\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star)\Big)^{\mathsf{T}}\Big((\hat{\boldsymbol{\theta}} - \mathsf{E}[\hat{\boldsymbol{\theta}}]) + (\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star)\Big)\right] \\
&= \mathsf{E}\left[(\hat{\boldsymbol{\theta}} - \mathsf{E}[\hat{\boldsymbol{\theta}}])^{\mathsf{T}}(\hat{\boldsymbol{\theta}} - \mathsf{E}[\hat{\boldsymbol{\theta}}]) + 2(\hat{\boldsymbol{\theta}} - \mathsf{E}[\hat{\boldsymbol{\theta}}])^{\mathsf{T}}(\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star) + (\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star)^{\mathsf{T}}(\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star)\right] \\
&= \mathsf{E}\left[(\hat{\boldsymbol{\theta}} - \mathsf{E}[\hat{\boldsymbol{\theta}}])^{\mathsf{T}}(\hat{\boldsymbol{\theta}} - \mathsf{E}[\hat{\boldsymbol{\theta}}])\right] + 2\mathsf{E}\Big[(\hat{\boldsymbol{\theta}} - \mathsf{E}[\hat{\boldsymbol{\theta}}])^{\mathsf{T}} \underbrace{(\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star)}_{\text{constant}}\Big] + \mathsf{E}\Big[(\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star)^{\mathsf{T}} \underbrace{(\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star)}_{\text{constant}}\Big] \\
&= \mathsf{E}\left[\left\|\hat{\boldsymbol{\theta}} - \mathsf{E}[\hat{\boldsymbol{\theta}}]\right\|_2^2\right] + 2\mathsf{E}\left[\hat{\boldsymbol{\theta}} - \mathsf{E}[\hat{\boldsymbol{\theta}}]\right]^{\mathsf{T}}(\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star) + \mathsf{E}\left[\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star\right]^{\mathsf{T}}(\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star) \\
&= \underbrace{\mathsf{E}\left[\left\|\hat{\boldsymbol{\theta}} - \mathsf{E}[\hat{\boldsymbol{\theta}}]\right\|_2^2\right]}_{\mathsf{var}(\hat{\boldsymbol{\theta}})} + 2\underbrace{(\mathsf{E}[\hat{\boldsymbol{\theta}}] - \mathsf{E}[\hat{\boldsymbol{\theta}}])^{\mathsf{T}}}_{\mathbf{0}}(\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star) + \underbrace{(\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star)^{\mathsf{T}}}_{\mathsf{bias}(\hat{\boldsymbol{\theta}})}\underbrace{(\mathsf{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^\star)}_{\mathsf{bias}(\hat{\boldsymbol{\theta}})}.
\end{aligned}$$

□

Example 4.6. Consider Example 4.1. Then

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_{ML} \;&=\; \operatorname*{arg\,max}_{\boldsymbol{\mu}\in\mathbb{R}^k} \mathsf{p}(\mathbf{x}_1, \ldots, \mathbf{x}_N|\boldsymbol{\mu}) \;=\; \operatorname*{arg\,max}_{\boldsymbol{\mu}\in\mathbb{R}^D} \prod_{i=1}^N \mathsf{p}(\mathbf{x}_i|\boldsymbol{\mu}) \;=\; \operatorname*{arg\,max}_{\boldsymbol{\mu}\in\mathbb{R}^D} \log \prod_{i=1}^N \mathsf{p}(\mathbf{x}_i|\boldsymbol{\mu}) \\
&= \operatorname*{arg\,max}_{\boldsymbol{\mu}\in\mathbb{R}^D} \sum_{i=1}^N \log \mathsf{p}(\mathbf{x}_i|\boldsymbol{\mu}) \;=\; \operatorname*{arg\,max}_{\boldsymbol{\mu}\in\mathbb{R}^D} \sum_{i=1}^N \log\left(\frac{1}{(\sqrt{2\pi}\sigma)^D} e^{-\frac{1}{2\sigma^2}(\mathbf{x}_i-\boldsymbol{\mu})^{\mathsf{T}}(\mathbf{x}_i-\boldsymbol{\mu})}\right) \\
&= \operatorname*{arg\,max}_{\boldsymbol{\mu}\in\mathbb{R}^D} \underbrace{\sum_{i=1}^N \log \frac{1}{(\sqrt{2\pi}\sigma)^D}}_{\text{constant}} - \sum_{i=1}^N \underbrace{\frac{1}{2\sigma^2}}_{\text{constant}} (\mathbf{x}_i - \boldsymbol{\mu})^{\mathsf{T}}(\mathbf{x}_i - \boldsymbol{\mu}) \\
&= \operatorname*{arg\,max}_{\boldsymbol{\mu}\in\mathbb{R}^D} -\sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})^{\mathsf{T}}(\mathbf{x}_i - \boldsymbol{\mu}) \;=\; \operatorname*{arg\,max}_{\boldsymbol{\mu}\in\mathbb{R}^D} -\sum_{i=1}^N \left(\mathbf{x}_i^{\mathsf{T}}\mathbf{x}_i - 2\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\mu} + \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu}\right) \\
&= \operatorname*{arg\,max}_{\boldsymbol{\mu}\in\mathbb{R}^D} \sum_{i=1}^N \left(2\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\mu} - \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu}\right).
\end{aligned}$$

Now we use our usual tricks: take derivative w.r.t. $\boldsymbol{\mu}$, set to zero and solve for $\boldsymbol{\mu}$. To learn more about how to take derivatives w.r.t. vectors and matrices see *Old and new matrix algebra useful for statistics*

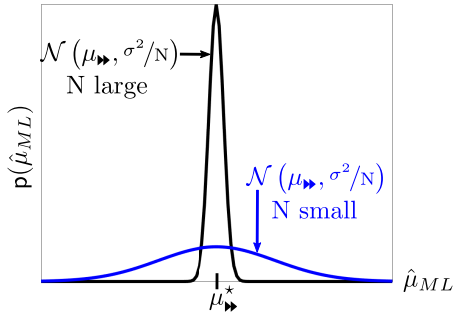

Figure 4.3: Example 4.6 shows that $\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbf{x}_{\mathrm{i}}$. This implies that $\hat{\boldsymbol{\mu}}_{ML} \sim \mathcal{N}(\boldsymbol{\mu}^\star_{\blacklozenge\blacklozenge}, \frac{\sigma^2}{\mathrm{N}}\mathbf{I})$. As N grows, $\mathsf{var}(\hat{\boldsymbol{\mu}}_{ML})$ shrinks, and we expect $\hat{\boldsymbol{\mu}}_{ML}$ to be closer to $\boldsymbol{\mu}^\star$.

by Thomas P. Minka.

$$\frac{\partial}{\partial \boldsymbol{\mu}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \left(2\mathbf{x}_{\mathrm{i}}^{\mathsf{T}}\boldsymbol{\mu} - \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu}\right) \;=\; \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\partial}{\partial \boldsymbol{\mu}} \left(2\mathbf{x}_{\mathrm{i}}^{\mathsf{T}}\boldsymbol{\mu} - \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu}\right) \;=\; \sum_{\mathrm{i}=1}^{\mathrm{N}} \left(2\mathbf{x}_{\mathrm{i}} - 2\boldsymbol{\mu}\right) \;=\; 2\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbf{x}_{\mathrm{i}} - 2\mathrm{N}\boldsymbol{\mu} \;=\; \mathbf{0}.$$

It follows that

$$\hat{\boldsymbol{\mu}}_{ML} \;=\; \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbf{x}_{\mathrm{i}}.$$

Next notice that

$$\mathsf{E}[\hat{\boldsymbol{\mu}}_{ML}] \;=\; \mathsf{E}\left[\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbf{x}_{\mathrm{i}}\right] \;=\; \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathsf{E}\left[\mathbf{x}_{\mathrm{i}}\right] \;=\; \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \boldsymbol{\mu}^\star_{\blacklozenge\blacklozenge} \;=\; \boldsymbol{\mu}^\star_{\blacklozenge\blacklozenge}.$$

Hence $\mathsf{bias}(\hat{\boldsymbol{\mu}}_{ML}) = \mathbf{0}$. On the other hand,

$$\mathsf{var}(\hat{\boldsymbol{\mu}}_{ML}) \;=\; \mathsf{var}\Big(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbf{x}_{\mathrm{i}}\Big) \;=\; \frac{1}{\mathrm{N}^2} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathsf{var}\left(\mathbf{x}_{\mathrm{i}}\right) \;=\; \frac{1}{\mathrm{N}^2} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{D}\sigma^2 \;=\; \frac{\mathrm{D}\sigma^2}{\mathrm{N}}.$$

Hence

$$\mathsf{MSE}(\hat{\boldsymbol{\mu}}_{ML}) \;=\; \frac{\mathrm{D}\sigma^2}{\mathrm{N}}.$$

4.5 Excess Risk/Regret and KL Divergence

Each sample x has a certain probability of being observed, namely $\mathsf{p}(\mathrm{x}|\theta^\star)$. An other way to analyze how good is an estimator is by comparing $\mathsf{p}(\mathrm{x}|\theta^\star)$ against the probability of the same sample under the estimator, i.e., $\mathsf{p}(\mathrm{x}|\hat{\theta})$. This is the idea behind the concept of *excess risk* or *regret*: to compare $\mathsf{p}(\mathrm{x}|\theta^\star)$ against $\mathsf{p}(\mathrm{x}|\hat{\theta})$.

Equivalently, we can compare $\log \mathsf{p}(\mathrm{x}|\theta^\star)$ against $\log \mathsf{p}(\mathrm{x}|\hat{\theta})$. This is done thinking about cases where we have i.i.d. samples $x_1, \ldots, x_{\mathrm{N}}$, so that $\log \mathsf{p}(\mathbf{x}|\cdot)$ factors nicely into $\sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathsf{p}(\mathrm{x}_{\mathrm{i}}|\cdot)$. Finally, since samples are random variables, it is better to compare the expected values of these quantities, which is precisely what

the *regret* compares.

Definition 4.5 (Excess risk/regret)**.**

$$\mathsf{Regret}(\hat{\theta}) \;:=\; \mathsf{E}\big[\log \mathsf{p}(\mathbf{x}|\theta^\star)\big] \;-\; \mathsf{E}\big[\log \mathsf{p}(\mathbf{x}|\hat{\theta})\big].$$

The next proposition shows that the regret is the *Kullback-Leibler divergence* (KL divergence).

Proposition 4.2 (Regret = KL divergence)**.**

$$\mathsf{Regret}(\hat{\theta}) \;=\; \mathsf{D}\left(\mathsf{p}(\mathbf{x}|\theta^\star)\middle\|\mathsf{p}(\mathbf{x}|\hat{\theta})\right) \;:=\; \int \mathsf{p}(\mathbf{x}|\theta^\star)\log\frac{\mathsf{p}(\mathbf{x}|\theta^\star)}{\mathsf{p}(\mathbf{x}|\hat{\theta})}d\mathbf{x}.$$

Proof.

$$\mathsf{E}\big[\log \mathsf{p}(\mathbf{x}|\theta^\star)\big] - \mathsf{E}\big[\log \mathsf{p}(\mathbf{x}|\hat{\theta})\big] \;=\; \mathsf{E}\big[\log \mathsf{p}(\mathbf{x}|\theta^\star) - \log \mathsf{p}(\mathbf{x}|\hat{\theta})\big] \;=\; \mathsf{E}\big[\log\frac{\mathsf{p}(\mathbf{x}|\theta^\star)}{\log \mathsf{p}(\mathbf{x}|\hat{\theta})}\big].$$

□

D is a well-studied function. Proposition 4.2 allows us to translate from Regret to D and use all the machinery and results known for D.

Example 4.7. Suppose $x_1, \ldots, x_\mathrm{N} \overset{iid}{\sim} \mathcal{N}(\mu^\star, {\sigma^\star}^2)$. Then the regret of any estimator $\hat{\theta} = \{\hat{\mu}, \hat{\sigma}\}$ is the well-known KL divergence of two gaussian distributions:

$$\mathsf{Regret}(\hat{\theta}) \;=\; \log\Big(\frac{\hat{\sigma}}{\sigma^\star}\Big) + \frac{\sigma^{\star 2} + (\mu^\star - \hat{\mu})^2}{2\hat{\sigma}^2} - \frac{1}{2}.$$

Here is an other example of the usefulness of Proposition 4.2:

Proposition 4.3 (Information inequality)**.**

$$\mathsf{D}\left(\mathsf{p}(\mathbf{x}|\theta^\star)\middle\|\mathsf{p}(\mathbf{x}|\hat{\theta})\right) \;\geq\; 0$$

with equality if and only if $\mathsf{p}(\mathbf{x}|\theta^\star) = \mathsf{p}(\mathbf{x}|\hat{\theta})$.

Proof. See Theorem 2.6.3 in *Elements of Information Theory* by Cover and Thomas, second edition. □

Proposition 4.3 implies that (as intuition suggests) in expectation, the true distribution $\mathsf{p}(\mathbf{x}|\theta)$ is more likely to have produced $\mathbf{x}$ than *any* other distribution $\mathsf{p}(\mathbf{x}|\hat{\theta})$. See *Elements of Information Theory* by Cover and Thomas to learn more about KL-divergence, entropy and other useful information theory functions and results.

4.6 Estimators of Functions

Often we want to estimate a function g of $\theta^\star$.

Example 4.8. Suppose we observe $x_1, \ldots, x_\mathrm{N} \overset{iid}{\sim} Poisson(\lambda^\star)$, with $\lambda^\star \in \mathbb{R}$ unknown. We want to know the probability γ that a new sample x is larger than $\lambda^\star$. Here $\gamma = g(\lambda^\star) = \mathsf{P}(x \geq \lambda^\star)$.

The next theorem shows that the MLE of a function is the function of the MLE.

Theorem 4.1 (Invariance of the MLE)**.** Let $x_1, \ldots, x_\mathrm{N} \overset{iid}{\sim} \mathsf{p}(\mathrm{x}|\theta^\star)$. Let $\gamma^\star = g(\theta^\star)$ for some function $g : \Theta \twoheadrightarrow \Gamma$. Then the MLE of $\gamma^\star$, defined as

$$\hat{\gamma}_{ML} \;:=\; \arg\max_{\gamma\in\Gamma} \Big(\max_{\theta\in g^{-1}(\gamma)} \mathsf{p}(\mathbf{x}|\theta) \Big),$$

is given by $\hat{\gamma}_{ML} = g(\hat{\theta}_{ML})$. Here g^{-1} denotes the inverse image of g, i.e., $g^{-1}(\gamma) = \{\theta \in \Theta : g(\theta) = \gamma\}$.

Think of $\gamma^\star$ as a parameter (that is a function of $\theta^\star$) that we also want to estimate.

Proof. Γ is defined as the range of g, so even if g is not one-to-one, the sets $\{g^{-1}(\gamma)\}_{\gamma\in\Gamma}$ form a partition of Θ (see Figure 4.4 to build some intuition). Therefore,

$$\bigcup_{\gamma\in\Gamma} g^{-1}(\gamma) \;=\; \Theta,$$

which in turns implies

$$\max_{\gamma\in\Gamma} \Big(\max_{\theta\in g^{-1}(\gamma)} \mathsf{p}(\mathbf{x}|\theta) \Big) \;=\; \max_{\theta\in\Theta} \mathsf{p}(\mathbf{x}|\theta).$$

□

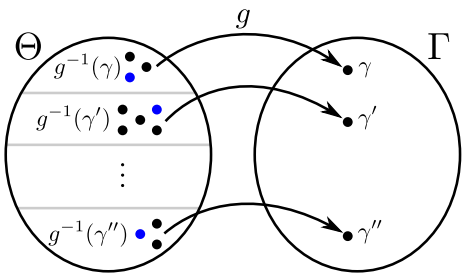

Figure 4.4: In Theorem 4.1, since $\{g^{-1}(\gamma)\}_{\gamma\in\Gamma}$ forms a partition of Θ, taking the max over Θ is the same as first taking the max of over each individual $g^{-1}(\gamma)$, highlighted in blue, and then taking the max over these maximums. More precisely, $\max_{\theta\in\Theta}\cdot = \max_{\gamma\in\Gamma}(\max_{\theta\in g^{-1}(\gamma)}\cdot)$.

Example 4.9. Continuing with Example 4.8, we can use Theorem 4.1 to first compute $\hat{\lambda}_{ML}$ and then obtain $\hat{\gamma}_{ML} = g(\hat{\lambda}_{ML}) = \mathsf{P}(X > \hat{\lambda}_{ML})$.

$$\begin{aligned}
\hat{\lambda}_{ML} &= \arg\max_{\lambda\in\mathbb{R}} \prod_{\mathrm{i}=1}^{\mathrm{N}} \mathsf{P}(x = \mathrm{x_i}|\lambda) = \arg\max_{\lambda\in\mathbb{R}} \log\prod_{\mathrm{i}=1}^{\mathrm{N}} \mathsf{P}(x = \mathrm{x_i}|\lambda) = \arg\max_{\lambda\in\mathbb{R}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathsf{P}(x = \mathrm{x_i}|\lambda) \\
&= \arg\max_{\lambda\in\mathbb{R}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \log \frac{\lambda^{\mathrm{x_i}} e^{-\lambda}}{\mathrm{x_i}!} = \arg\max_{\lambda\in\mathbb{R}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \log\lambda^{\mathrm{x_i}} - \sum_{\mathrm{i}=1}^{\mathrm{N}} \lambda - \underbrace{\sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathrm{x_i}!}_{\text{constant}} \\
&= \arg\max_{\lambda\in\mathbb{R}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i}\log\lambda - \mathrm{N}\lambda = \arg\max_{\lambda\in\mathbb{R}} \log\lambda \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i} - \mathrm{N}\lambda.
\end{aligned}$$

It follows that

$$\frac{\partial}{\partial\lambda}\Big(\log\lambda \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i} - \mathrm{N}\lambda\Big) = \frac{1}{\lambda}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i} - \mathrm{N}.$$

Setting to zero and solving for λ we have:

$$\hat{\lambda}_{ML} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i}.$$

It follows by Theorem 4.1 that

$$\hat{\gamma}_{ML} = g(\hat{\lambda}_{ML}) = \mathsf{P}(x > \hat{\lambda}_{ML}) = \mathsf{P}\Big(x > \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i}\Big).$$

4.7 Asymptotics

Example 4.6 shows that the larger N, i.e., the more *training* samples we have, the closer $\hat{\boldsymbol{\mu}}_{ML}$ will be to the true *skip* signal $\boldsymbol{\mu}^{\star}_{\blacktriangleright\!\blacktriangleright}$ that we want to estimate (see Figure 4.3). This is a desirable property in an estimator called *consistency*.

Definition 4.6 (Consistency)**.** We say $\hat{\theta}$ is *consistent* if $\hat{\theta} \to \theta^\star$ as $\mathrm{N} \to \infty$. More precisely, $\hat{\theta}$ is *consistent w.r.t.* Risk if $\mathsf{Risk}(\hat{\theta}) \to 0$ as $\mathrm{N} \to \infty$.

Example 4.10. $\hat{\boldsymbol{\mu}}_{ML}$ from Example 4.6 is consistent w.r.t. $\mathsf{Risk} = \mathsf{MSE}$, because

$$\mathsf{MSE}(\hat{\boldsymbol{\mu}}_{ML}) \;=\; \mathsf{E}\left[\|\hat{\boldsymbol{\mu}}_{ML} - \boldsymbol{\mu}^\star_{\blacktriangleright\blacktriangleright}\|_2^2\right] \;=\; \frac{\mathrm{D}\sigma^2}{\mathrm{N}} \;\xrightarrow{\mathrm{N}\to\infty}\; 0,$$

i.e., $\hat{\boldsymbol{\mu}}_{ML} \to \boldsymbol{\mu}^\star_{\blacktriangleright\blacktriangleright}$ as $\mathrm{N} \to \infty$.

As mentioned before, an estimator is a function of the data we observe, and hence it is a random variable. The next theorem states that if we observe a large number N of i.i.d. samples, then regardless of the distribution of these data, $\hat{\theta}_{ML}$ will be distributed normal, centered at $\theta^\star$, and with variance decreasing with N, implying consistency.

Theorem 4.2 (Asymptotic distribution of the MLE)**.** Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\mathrm{N} \overset{iid}{\sim} \mathsf{p}(\mathbf{x}|\boldsymbol{\theta}^\star)$, with $\boldsymbol{\theta}^\star \in \mathbb{R}^\mathrm{K}$. Let

$$L(\boldsymbol{\theta}) \;:=\; \sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathsf{p}(\mathbf{x}_\mathrm{i}|\boldsymbol{\theta}).$$

Suppose $\dfrac{\partial L(\boldsymbol{\theta})}{\partial \theta_\mathrm{k}}$ and $\dfrac{\partial^2 L(\boldsymbol{\theta})}{\partial \theta_\mathrm{k} \theta_\ell}$ exist for every $\mathrm{k}, \ell \in \{1, \ldots, \mathrm{K}\}$. Then

$$\hat{\boldsymbol{\theta}}_{ML} \;\overset{\mathrm{N}\to\infty}{\sim}\; \mathcal{N}\Big(\boldsymbol{\theta}^\star, \frac{1}{\mathrm{N}} \boldsymbol{I}^{-1}_{\boldsymbol{\theta}^\star}\Big),$$

where $\boldsymbol{I}_{\boldsymbol{\theta}^\star}$ is the *Fisher-information matrix*, whose elements are defined as:

$$[\boldsymbol{I}_{\boldsymbol{\theta}^\star}]_{\mathrm{k}\ell} \;:=\; -\mathsf{E}\left[\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \theta_\mathrm{k} \partial \theta_\ell}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}\right].$$

Proof. The proof follows as a consequence of the central limit theorem, which says that if $x_1, \ldots, x_\mathrm{N}$ are i.i.d. with mean μ and variance σ^2, then $\frac{1}{\sqrt{\mathrm{N}}} \sum_{\mathrm{i}=1}^{\mathrm{N}} x_\mathrm{i}$ is asymptotically distributed $\mathcal{N}(\mu, \sigma^2)$. For a detailed proof see Theorem 10.1.12 in *Statistical Inference* by George Casella and Roger L. Berger, second edition. □

Example 4.11 (Exam-type question)**.** We want to find out which of two treatments is more effective. So we give treatment k to n_k people, and record the number x_k of people that react favorably to treatment k ($\mathrm{k} = 1, 2$). We can model this as $x_\mathrm{k} \sim \mathit{Binomial}(\mathrm{n}_\mathrm{k}, \mathrm{p}^\star_\mathrm{k})$, where x_1 is independent of x_2. Then the difference $\delta^\star = \mathrm{p}^\star_1 - \mathrm{p}^\star_2$ would be a good indicator of which treatment is more effective, and by how much. To estimate $\delta^\star$ (which is a function of $\mathrm{p}^\star_1$ and $\mathrm{p}^\star_2$), we can first estimate $\mathrm{p}^\star_1$ and $\mathrm{p}^\star_2$, and

then use Theorem 4.1. First notice that,

$$\hat{p}_k \;=\; \underset{p_k\in[0,1]}{\arg\max}\; \mathsf{P}(x_k = x_k|p_k) \;=\; \underset{p_k\in[0,1]}{\arg\max}\; \log \mathsf{P}(x_k = x_k|p_k) \;=\; \underset{p_k\in[0,1]}{\arg\max}\; L(p_k).$$

Next we will use our usual tricks: take derivative, set to zero and solve for the desired parameters.

$$\begin{aligned}\frac{\partial L(p_k)}{\partial p_k} \;&=\; \frac{\partial}{\partial p_k}\log\left(\binom{n_k}{x_k} p_k^{x_k}(1-p_k)^{n_k-x_k}\right) \;=\; \frac{\partial}{\partial p_k}\Big(\underbrace{\log\binom{n_k}{x_k}}_{\text{constant}} + \log p_k^{x_k} + \log(1-p_k)^{n_k-x_k}\Big)\\ &=\; \frac{\partial}{\partial p_k}\big(x_k\log p_k + (n_k-x_k)\log(1-p_k)\big) \;=\; \frac{x_k}{p_k} - \frac{n_k-x_k}{1-p_k}.\end{aligned}$$

Setting to zero,

$$0 \;=\; \frac{x_k(1-p_k)-(n_k-x_k)p_k}{p_k(1-p_k)} \;=\; x_k - \cancel{x_k p_k} - n_k p_k + \cancel{x_k p_k} \;=\; x_k - n_k p_k,$$

and solving for p_k we obtain $\hat{p}_k \;=\; \frac{x_k}{n_k}$. It follows by Theorem 4.1 that

$$\hat{\delta}_{ML} \;=\; \hat{p}_1 - \hat{p}_2 \;=\; \frac{x_1}{n_1} - \frac{x_2}{n_2}.$$

Theorems 4.1 and 4.2 imply that $\hat{\delta}_{ML} \to \delta^\star$ as $n_1, n_2 \to \infty$ (can you show this?). But we can also use them to quantify how accurate our estimator $\hat{\delta}_{ML}$ will be. For example, if we only test $n_k = 1$ people per treatment, our estimator might not be very reliable. We want to know how many people we need to test to guarantee that $\hat{\delta}_{ML}$ is pretty close to the true $\delta^\star$. More precisely, we want

$$\mathsf{P}\left(\left|\hat{\delta}_{ML} - \delta^\star\right| \geq \beta\right) \leq \alpha.$$

Recall that the sum of binomials is only binomial if they share the same p. Hence we do not know the distribution of $\hat{\delta}_{ML}$. Fortunately, we can use Theorem 4.2 to know its asymptotic distribution. To this end, we first need to compute $\boldsymbol{I}_{p_k^\star}$, so write:

$$\begin{aligned}\frac{\partial^2 L(p_k)}{\partial^2 p_k} \;&=\; \frac{\partial}{\partial p_k}\Big(\frac{x_k}{p_k} - \frac{n_k-x_k}{1-p_k}\Big) \;=\; x_k\frac{\partial p_k^{-1}}{\partial p_k} - (n_k-x_k)\frac{\partial(1-p_k)^{-1}}{\partial p_k}\\ &=\; -x_k p_k^{-2} - (n_k-x_k)(1-p_k)^{-2} \;=\; -\frac{x_k}{p_k^2} - \frac{n_k-x_k}{(1-p_k)^2}.\end{aligned}$$

Hence

$$\begin{aligned}\boldsymbol{I}_{p_k^\star} \;&=\; -\mathsf{E}\left[\frac{\partial^2 L(p_k)}{\partial^2 p_k}\Big|_{p_k=p_k^\star}\right] \;=\; -\mathsf{E}\left[\left(-\frac{x_k}{p_k^2} - \frac{n_k-x_k}{(1-p_k)^2}\right)\Big|_{p_k=p_k^\star}\right]\\ &=\; \mathsf{E}\left[\frac{x_k}{p_k^{\star 2}} + \frac{n_k-x_k}{(1-p_k^\star)^2}\right] \;=\; \frac{\mathsf{E}[x_k]}{p_k^{\star 2}} + \frac{n_k-\mathsf{E}[x_k]}{(1-p_k^\star)^2}\\ &=\; \frac{n_k p_k}{p_k^{\star 2}} + \frac{n_k - n_k p_k}{(1-p_k^\star)^2} \;=\; \frac{n_k}{p_k^\star} + \frac{n_k(1-p_k)}{(1-p_k^\star)^2}\\ &=\; \frac{n_k}{p_k^\star} + \frac{n_k}{1-p_k^\star} \;=\; n_k\frac{1-p_k^\star+p_k^\star}{p_k^\star(1-p_k^\star)} \;=\; \frac{n_k}{p_k^\star(1-p_k^\star)},\end{aligned}$$

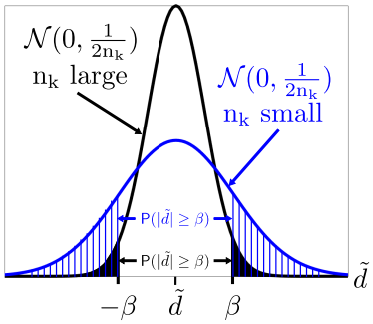

Figure 4.5: In Example 4.11, $\mathsf{P}(|\hat{\delta}_{ML} - \delta^\star| \geq \beta) \leq \mathsf{P}(|\tilde{\delta}| \geq \beta)$, where $\tilde{\delta} \sim \mathcal{N}(0, \frac{1}{2n_k})$. As n_k grows, $\mathsf{var}(\tilde{\delta})$ shrinks, and hence so does $\mathsf{P}(|\tilde{\delta}| \geq \beta)$. We want to find the n_k to guarantee that $\mathsf{P}(|\tilde{\delta}| \geq \beta) \leq \alpha$. This will also imply that $\mathsf{P}(|\hat{\delta}_{ML} - \delta^\star| \geq \beta) \leq \alpha$.

which implies

$$\boldsymbol{I}^{-1}_{p_k^\star} = \frac{p_k^\star(1-p_k^\star)}{n_k}.$$

By Theorem 4.2, $\hat{p}_k \overset{N\to\infty}{\sim} \mathcal{N}\left(p_k^\star, \frac{1}{N}\frac{p_k^\star(1-p_k^\star)}{n_k}\right)$, with $N = 1$ (because we only observe one x_1 and one x_2). Since sums of gaussians are gaussians, it follows that

$$\hat{\delta}_{ML} \overset{N\to\infty}{\sim} \mathcal{N}\left(p_1^\star - p_2^\star, \frac{p_1^\star(1-p_1^\star)}{n_1} + \frac{p_2^\star(1-p_2^\star)}{n_2}\right).$$

Now suppose for simplicity that we will test the same patients for each treatment, i.e., $n_1 = n_2$. Then this simplifies to

$$\hat{\delta}_{ML} \overset{N\to\infty}{\sim} \mathcal{N}\left(\delta^\star, \frac{p_1^\star(1-p_1^\star) + p_2^\star(1-p_2^\star)}{n_k}\right).$$

Next observe that since $p(1-p) \leq 1/4$ for every $p \in [0,1]$, the asymptotic variance of $\hat{\delta}_{ML}$ is bounded by $\frac{1}{2n_k}$. Letting $\tilde{\delta}$ be a $\mathcal{N}(0, \frac{1}{2n_k})$ distributed random variable,

$$\mathsf{P}\left(\left|\hat{\delta}_{ML} - \delta^\star\right| \geq \beta\right) \leq \mathsf{P}\left(\left|\tilde{\delta}\right| \geq \beta\right) = 2Q_{0,\frac{1}{2n_k}}(\beta) = 2Q(\sqrt{2n_k}\beta).$$

See Figure 4.5 to build some intuition. If we want $2Q(\sqrt{2n_k}\beta) \leq \alpha$, then we need to test

$$n_k \geq \left(\frac{Q^{-1}(\alpha/2)}{\sqrt{2}\beta}\right)^2 \tag{4.1}$$

patients per treatment.

4.8 Experiments

It is often good to run some experiments to verify that our findings are correct. For instance, we can run a simulation of Example 4.11.

Recall that the ultimate goal in Example 4.11 is to determine which of two treatments is better, and by how

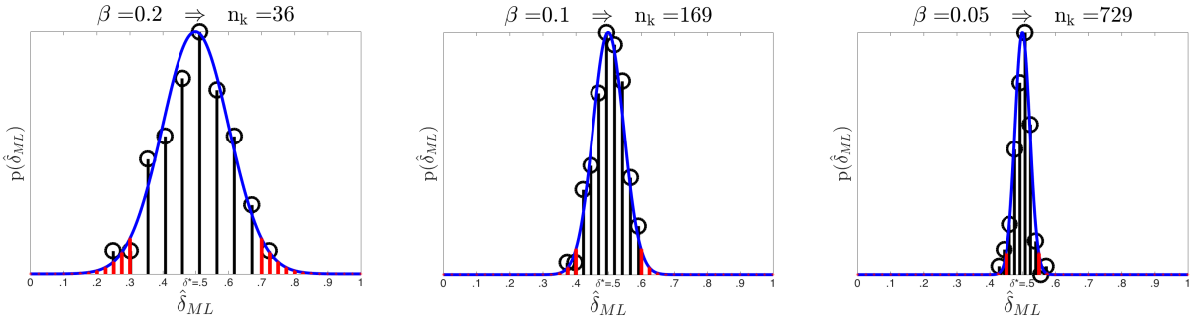

Figure 4.6: Results of the simulation of Example 4.11. In **black** is the histogram of $\hat{\delta}_{ML}$ over T = 100 trials. In **blue** is the asymptotic distribution of $\hat{\delta}_{ML}$. In **red** is the probability α that $|\hat{\delta}_{ML} - \delta^\star| \geq \beta$. Notice that as $\mathrm{n_k}$ grows, the variance of $\hat{\delta}_{ML}$ decreases, and so $\hat{\delta}_{ML}$ gets closer and closer to $\delta^\star$. The code for this simulation is in the appendix.

much. This is determined by $\delta^\star$, so our goal is to estimate $\delta^\star$. To this end, we need to know how many patients $\mathrm{n_k}$ we need to test on each treatment to be $(1-\alpha)\%$ sure that our estimate $\hat{\delta}_{ML}$ will be within β error of $\delta^\star$, i.e., we want $\mathsf{P}(|\hat{\delta}_{ML} - \delta^\star| \geq \beta) \leq \alpha$. Using Theorems 4.1 and 4.2 we found that $\mathrm{n_k}$ needs to be as in (4.1). We can run a simulation to verify whether this is accurate as follows.

First generate random vectors $\boldsymbol{y}_\mathrm{k} \in \mathbb{R}^{\mathrm{n_k}}$ with i.i.d. Bernoulli($\mathrm{p_k^\star}$) entries. The $\mathrm{i^{th}}$ entry in $\boldsymbol{y}_\mathrm{k}$ simulates whether patient i reacted favorably to treatment k. Then x_k is simply the sum of the entries in $\boldsymbol{y}_\mathrm{k}$. From our results in Example 4.11 we know that $\hat{\delta}_{ML} = \frac{x_1}{\mathrm{n_1}} - \frac{x_2}{\mathrm{n_2}}$.

We can repeat this experiment T trials, plot a histogram of $\hat{\delta}_{ML}$, and verify whether $\hat{\delta}_{ML}$ truly is within β of $\delta^\star$ on $(1-\alpha)\%$ of the trials. Furthermore, we can compare it with the asymptotic distribution obtained by Theorem 4.2. Figure 4.6 shows some results. The code for this simulation is in the appendix.

4.9 Cramer-Rao Lower Bound

Theorem 4.2 tells us that $\hat{\theta}_{ML} \to \theta^\star$ at a rate of $1/\sqrt{\mathrm{N}}$ (because $\mathsf{var}(\hat{\theta}_{ML}) = \mathsf{E}[(\hat{\theta}_{ML} - \theta^\star)^2] = \frac{1}{\mathrm{N}}\boldsymbol{I}_{\boldsymbol{\theta}^\star}^{-1}$). Intuitively, this means that we need about N samples to be within $1/\sqrt{\mathrm{N}}$ error. This raises the question: is there a better estimator? For example, one that achieves $1/\mathrm{N}$ error with the same number of samples? The next theorem shows that this is *not* the case. More precisely, it shows that there exists *no* estimator with a faster convergence rate than $1/\sqrt{\mathrm{N}}$.

Theorem 4.3 (Cramer-Rao Lower Bound)**.** Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\mathrm{N} \overset{iid}{\sim} \mathsf{p}(\mathbf{x}|\boldsymbol{\theta}^\star)$, with $\boldsymbol{\theta}^\star \in \mathbb{R}^\mathrm{K}$. Suppose $\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_\mathrm{k}}$ and $\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \theta_\mathrm{k} \theta_\ell}$ exist for every $\mathrm{k}, \ell \in \{1, \ldots, \mathrm{K}\}$. Let $\hat{\boldsymbol{\theta}}$ be an unbiased estimator of $\boldsymbol{\theta}^\star$. Define the *error covariance matrix* as

$$\widehat{\boldsymbol{C}} \;:=\; \mathsf{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\mathsf{T}\right].$$

Then

$$\widehat{\boldsymbol{C}} \;\succeq\; \frac{1}{\mathrm{N}}\boldsymbol{I}_{\boldsymbol{\theta}^\star}^{-1},$$

where $\succeq$ means that the eigenvalues of $\widehat{\boldsymbol{C}} - \frac{1}{\mathrm{N}}\boldsymbol{I}_{\boldsymbol{\theta}^\star}^{-1}$ are ≥ 0

Proof. We will prove this for D = K = 1, i.e., $x_1, \ldots, x_N, \theta \in \mathbb{R}$ are scalars. The general result follows by similar arguments. Since $\hat{\theta}$ is unbiased, we know

$$0 \;=\; \mathsf{E}[\hat{\theta} - \theta^\star] = \int (\hat{\theta} - \theta^\star)\mathsf{p}(\mathbf{x}|\theta^\star)d\mathbf{x}.$$

Taking derivatives w.r.t. θ we have:

$$\begin{aligned} 0 \;&=\; \frac{\partial}{\partial\theta}\int (\hat{\theta}-\theta)\mathsf{p}(\mathbf{x}|\theta)d\mathbf{x}\Big|_{\theta=\theta^\star} \;=\; \int \frac{\partial}{\partial\theta}\underbrace{(\hat{\theta}-\theta)}_{\mathrm{u}}\,\underbrace{\mathsf{p}(\mathbf{x}|\theta)}_{\mathrm{v}}\,d\mathbf{x}\Big|_{\theta=\theta^\star} \\ &=\; \underbrace{\int -\mathsf{p}(\mathbf{x}|\theta)d\mathbf{x}\Big|_{\theta=\theta^\star}}_{-1} \;+\; \int (\hat{\theta}-\theta)\frac{\partial \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta}d\mathbf{x}\Big|_{\theta=\theta^\star}. \end{aligned}$$

Recall that

$$\frac{\partial \log \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta} \;=\; \frac{1}{\mathsf{p}(\mathbf{x}|\theta)}\frac{\partial \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta}.$$

Plugging this into the previous equation, we get

$$\begin{aligned} 1 \;&=\; \int (\hat{\theta}-\theta)\frac{\partial \log \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta}\mathsf{p}(\mathbf{x}|\theta)d\mathbf{x}\Big|_{\theta=\theta^\star} \;=\; \mathsf{E}\Big[(\hat{\theta}-\theta)\cdot\frac{\partial \log \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta}\Big|_{\theta=\theta^\star}\Big] \\ &\leq\; \underbrace{\sqrt{\mathsf{E}\big[(\hat{\theta}-\theta^\star)^2\big]}}_{\sqrt{\mathsf{var}(\hat{\theta})}}\;\sqrt{\mathsf{E}\left[\Big(\frac{\partial \log \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta}\Big)^2\Big|_{\theta=\theta^\star}\right]}, \end{aligned} \tag{4.2}$$

where the last step follows by the Cauchy-Schwartz inequality: for two random vectors $\boldsymbol{x}, \boldsymbol{y}$, we define their inner product as $\langle \boldsymbol{x}, \boldsymbol{y}\rangle := \mathsf{E}[\boldsymbol{x}^\mathsf{T}\boldsymbol{y}]$ and their norm as $\|\boldsymbol{x}\|_2^2 := \langle \boldsymbol{x}, \boldsymbol{x}\rangle = \mathsf{E}[\boldsymbol{x}^\mathsf{T}\boldsymbol{x}]$. Then the Cauchy-Schwartz inequality ($|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \|\mathbf{x}\|\|\mathbf{y}\|$) becomes:

$$\big|\mathsf{E}[\boldsymbol{x}^\mathsf{T}\boldsymbol{y}]\big| \;\leq\; \sqrt{E[\boldsymbol{x}^\mathsf{T}\boldsymbol{x}]\;\mathsf{E}[\boldsymbol{y}^\mathsf{T}\boldsymbol{y}]}.$$

We will now show that the second term in (4.2) is $-\sqrt{\boldsymbol{I}_{\theta^\star}}$. Write:

$$\begin{aligned} \frac{\partial^2 \log \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta^2} \;&=\; \frac{\partial}{\partial\theta}\left(\frac{\partial \log \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta}\right) \;=\; \frac{\partial}{\partial\theta}\left(\frac{1}{\mathsf{p}(\mathbf{x}|\theta)}\frac{\partial \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta}\right) \\ &=\; -\frac{1}{\mathsf{p}(\mathbf{x}|\theta)^2}\frac{\partial \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta}\frac{\partial \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta} \;+\; \frac{1}{\mathsf{p}(\mathbf{x}|\theta)}\frac{\partial^2 \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta^2} \\ &=\; -\Big(\underbrace{\frac{1}{\mathsf{p}(\mathbf{x}|\theta)}\frac{\partial \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta}}_{\frac{\partial \log \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta}}\Big)^2 \;+\; \frac{1}{\mathsf{p}(\mathbf{x}|\theta)}\frac{\partial^2 \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta^2}. \end{aligned}$$

For the second term write:

$$\mathsf{E}\left[\frac{1}{\mathsf{p}(\mathbf{x}|\theta)}\frac{\partial^2 \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta^2}\right] \;=\; \int \frac{1}{\mathsf{p}(\mathbf{x}|\theta)}\frac{\partial^2 \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta^2}\mathsf{p}(\mathbf{x}|\theta)d\mathbf{x} \;=\; \int \frac{\partial^2 \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta^2}d\mathbf{x} \;=\; \frac{\partial^2}{\partial\theta^2}\underbrace{\int \mathsf{p}(\mathbf{x}|\theta)d\mathbf{x}}_{1} \;=\; 0.$$

It follows that

$$\begin{aligned} \mathsf{E}\left[\Big(\frac{\partial \log \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta}\Big)^2\Big|_{\theta=\theta^\star}\right] \;&=\; \mathsf{E}\left[\frac{\partial^2 \log \mathsf{p}(\mathbf{x}|\theta)}{\partial\theta^2}\Big|_{\theta=\theta^\star}\right] \;=\; \mathsf{E}\left[\frac{\partial^2 \log \prod_{\mathrm{i}=1}^{\mathrm{N}} \mathsf{p}(\mathrm{x_i}|\theta)}{\partial\theta^2}\Big|_{\theta=\theta^\star}\right] \\ &=\; \mathsf{E}\left[\frac{\partial^2 \sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathsf{p}(\mathrm{x_i}|\theta)}{\partial\theta^2}\Big|_{\theta=\theta^\star}\right] \;=\; \mathrm{N}\mathsf{E}\left[\frac{\partial^2 \log \mathsf{p}(\mathrm{x_i}|\theta)}{\partial\theta^2}\Big|_{\theta=\theta^\star}\right] \;=\; -\mathrm{N}\boldsymbol{I}_{\theta^\star} \end{aligned}$$

Plugging this in (4.2) and taking squares, we have:

$$\mathsf{var}(\hat{\theta}) \;\geq\; \frac{1}{\mathrm{N}} \boldsymbol{I}_{\theta^\star}^{-1},$$

as desired. □

4.10 Mixtures

Theorem 4.3 shows that the MLE is optimal in the sense that it achieves *the best* convergence rate to the true parameter. However, it is not always easy, or even possible, to derive or compute the MLE.

Example 4.12. Suppose we observe $x_1, \ldots, x_{\mathrm{N}}$ from a *mixture* of K gaussians (see Figure 4.7):

$$x_{\mathrm{i}} \overset{iid}{\sim} \mathsf{p}(\mathrm{x}|\boldsymbol{\theta}^\star) \;=\; \sum_{\mathrm{k}=1}^{\mathrm{K}} \rho_{\mathrm{k}}^\star \, \mathcal{N}(\mu_{\mathrm{k}}^\star, \sigma_{\mathrm{k}}^{\star 2}) \;=\; \sum_{\mathrm{k}=1}^{\mathrm{K}} \rho_{\mathrm{k}}^\star \, \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{k}}^\star} e^{-\frac{1}{2}\left(\frac{\mathrm{x}-\mu_{\mathrm{k}}^\star}{\sigma_{\mathrm{k}}^\star}\right)^2}.$$

where $\rho_1, \ldots, \rho_{\mathrm{K}} \geq 0$ and $\sum_{\mathrm{k}=1}^{\mathrm{K}} \rho_{\mathrm{k}} = 1$. ρ_{k} is the probability that a sample corresponds to the k^{th} gaussian. In this case $\boldsymbol{\theta}^\star = [\theta_1^\star \;\cdots\; \theta_{\mathrm{k}}^\star]^{\mathsf{T}}$, where $\theta_{\mathrm{k}}^\star = \{\rho_{\mathrm{k}}^\star, \mu_{\mathrm{k}}^\star, \sigma_{\mathrm{k}}^\star\}$, and computing the MLE

$$\arg\max_{\boldsymbol{\theta}\in\Theta} \; \mathsf{p}(\mathbf{x}|\boldsymbol{\theta}) \;=\; \arg\max_{\boldsymbol{\theta}\in\Theta} \; \log \mathsf{p}(\mathbf{x}|\boldsymbol{\theta}).$$

is not so easy because the log of a sum does not factor nicely, i.e., $\mathsf{p}(\mathbf{x}|\boldsymbol{\theta})$ is no longer convex.

In cases like these where it is not easy to compute the MLE. Estimating mixtures is an active field of research, as they are good models for classification. So, how would you do it? Give it some thought. Maybe you come up with a great idea and you'd become famous! ;)

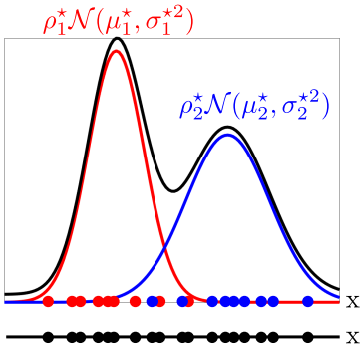

Figure 4.7: See Example 4.12. In a *mixture* each sample $\mathrm{x_i}$ is drawn according to one out of K distributions that we want to estimate. The challenge is that we don't know which distribution generated each $\mathrm{x_i}$, and so we don't know which $\mathrm{x_i}$'s should be used to estimate the k^{th} parameter. In this illustration we only observe the *black* points. If we knew which one is red and which one is blue, we could use only the red to estimate $\{\mu_1^\star, \sigma_1^\star\}$ and only the blue to estimate $\{\mu_2^\star, \sigma_2^\star\}$. The challenge is that we don't know the colors, and so we also have to estimate which distribution each point belongs to. Here $\rho_{\mathrm{k}}^\star \in [0, 1]$ models the fraction of points that correspond to the k^{th} distribution, which also has to be estimated. So, how would you do it? Give it some thought. Maybe you come up with a great idea and you'd become famous! ;)

A Code for Simulation of Example 4.11

```
clear all; close all; clc;

% ===== Code to simulate the asymptotic distribution of delta_hat =====
% ===== See Example 4.11.

T = 100;         % Number of trials.
beta = 0.1;      % Error margin that we want.
alpha = 0.05;    % Allowed probability of having bigger error that we want.
p1_star = 0.75; % True probability of reacting favorably to treatment 1.
p2_star = 0.25; % True probability of reacting favorably to treatment 1.
delta_star = p1_star-p2_star;   % Parameter we want to estimate.
delta_hat = zeros(T,1);         % Estimates of delta_star over trials.

% Number of patients to test each treatment, to guarantee desired accuracy.
nk = ceil(norminv(alpha/2)/(sqrt(2)*beta))^2;

for t=1:T,
    % Yk contains the outcomes of the patients given treatment k,
    % 1=reacted favorably, 0=not favorably.
    Y1 = rand([nk,1]) < p1_star;
    Y2 = rand([nk,1]) < p2_star;

    % Xk is the number of patients that reacted favorably to treatment k.
    X1 = sum(Y1);
    X2 = sum(Y2);

    % hat_pk is the estimate of pk_star.
    p1_hat = X1 / nk;
    p2_hat = X2 / nk;

    delta_hat(t) = p1_hat - p2_hat;
end

% Plot results.
figure(1);
axes('Box','on');
hold on;
[h,rangeh] = hist(delta_hat);
stem(rangeh,h,'k','LineWidth',4,'MarkerSize',20);

% Asymptotic distribution of delta_hat.
range = 0:.001:1;
mu_delta = delta_star;
sigma_delta = sqrt((p1_star*(1-p1_star) + p2_star*(1-p2_star))/nk);
asymptotic_dist = normpdf(range,mu_delta,sigma_delta);
plot(range,asymptotic_dist/max(asymptotic_dist)*max(h),'b','LineWidth',4);

% Highlight beta and the probability alpha.
bounds_alpha = [0:.025:delta_star-beta,delta_star+beta:0.025:1];
region_alpha = normpdf(bounds_alpha,mu_delta,sigma_delta);
stem(bounds_alpha,region_alpha/max(asymptotic_dist)*max(h),'r',...
    'LineWidth',5,'MarkerSize',1);

% Make figure look sexy.
axis tight;
set(gca,'XTick',[0:.1:1],'xticklabel',...
    {'0','.1','.2','.3','.4','\delta*=.5','.6','.7','.8','.9','1'},...
    'fontsize',15);
set(gca,'YTick',[],'yticklabel',[]);
ylabel('p$(\hat{\delta}_{ML})$','Interpreter','latex','fontsize',25);
```

```
61  xlabel('$\hat{\delta}_{ML}$','Interpreter','latex','fontsize',30);
62  title(['$\beta=$',num2str(beta),'$\ \ \ \Rightarrow \ \ {\rm n}_{\rm ...
        k}=$',num2str(nk)],'Interpreter','latex','fontsize',30);
63
64  % Save figure.
65  set(gcf, 'renderer','default');
66  figurename = ['asymptotic_beta',num2str(beta),',nk',num2str(nk),'.pdf'];
67  saveas(gcf,figurename);
```