24 Decision and Estimation Asymptotics

In this section, we further develop and apply our asymptotic analysis to some basic hypothesis testing and estimation settings of interest.

24.1 Asymptotics of Hypothesis Testing

In this section, we employ the large deviation analysis tools we developed so far to characterize the asymptotic behavior of hypothesis testing. We start with the analysis of the likelihood ratio test, study the behavior of both types of errors and explore the geometry of the corresponding probability space. We then discuss the special cases of the classical and Bayesian approaches.

Let y be the random variable whose values are used to disambiguate between the two hypotheses in question. Under the null hypothesis H_0, y is distributed according to the probability distribution p_0; under the alternative hypothesis H_1, y is distributed according to the probability distribution p_1. Let $\mathbf{y} = [y_1, \dots y_N]^{\mathrm{T}}$ be a sequence of N i.i.d. samples of the random variable $\mathbf{y}$. For any constant γ, the log-likelihood ratio test

$$\frac{1}{N} \log \frac{p_1^N(\mathbf{y})}{p_0^N(\mathbf{y})} = \frac{1}{N} \sum_{n=1}^{N} \log \frac{p_1(y_n)}{p_0(y_n)} \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \gamma \tag{1}$$

induces decision regions

$$\mathcal{R}_0 = \left\{ \mathbf{y} \in \mathcal{Y}^N \colon \frac{1}{N} \sum_{n=1}^{N} \log \frac{p_1(y_n)}{p_0(y_n)} \leq \gamma \right\} \text{ and } \mathcal{R}_1 = \left\{ \mathbf{y} \in \mathcal{Y}^N \colon \frac{1}{N} \sum_{n=1}^{N} \log \frac{p_1(y_n)}{p_0(y_n)} \geq \gamma \right\}. \tag{2}$$

Here we allow a bit of imprecision by including sequences $\mathbf{y}$ for which the test is met with equality in both sets. This makes our proofs easier, and as we already saw before, does not make much of a difference for the actual test.

The method of types allows us to map sets of observations to sets of probability distributions. Using this mapping, we construct the two corresponding sets on the probability simplex:

$$\mathcal{S}_0 = \left\{ p \in \mathcal{P}^{\mathcal{Y}} \colon \mathbb{E}_p \left[\log \frac{p_1(\mathsf{y})}{p_0(\mathsf{y})} \right] \leq \gamma \right\} \quad \text{and} \quad \mathcal{S}_1 = \left\{ p \in \mathcal{P}^{\mathcal{Y}} \colon \mathbb{E}_p \left[\log \frac{p_1(\mathsf{y})}{p_0(\mathsf{y})} \right] \geq \gamma \right\}. \tag{3}$$

It is easy to see the connection between the two kinds of sets:

$$\mathcal{R}_i = \left\{ \mathbf{y} \in \mathcal{Y}^N \colon \hat{p}(\cdot; \mathbf{y}) \in \mathcal{S}_i \cap \mathcal{P}_N \right\} \quad \text{for } i = 0, 1. \tag{4}$$

Here we consider the case when $p_0 \in \mathcal{S}_0$ and $p_1 \in \mathcal{S}_1$. Since

$$\mathbb{E}_{p_0}\left[\log \frac{p_1(\mathsf{y})}{p_0(\mathsf{y})}\right] = -\,\mathbb{E}_{p_0}\left[\log \frac{p_0(\mathsf{y})}{p_1(\mathsf{y})}\right] = -D(p_0\|p_1), \tag{5}$$

$$\mathbb{E}_{p_1}\left[\log \frac{p_1(\mathsf{y})}{p_0(\mathsf{y})}\right] = D(p_1\|p_0), \tag{6}$$

the corresponding range of values of the threshold γ is between these two values:

$$-D(p_0\|p_1) \le \gamma \le D(p_1\|p_0). \tag{7}$$

We discuss the behavior of the test when the threshold is outside this range later in this section.

Using the relationship between the sets of sequences $\mathcal{R}_i$ and the sets of probability distributions $\mathcal{S}_i$, we can characterize the probability of errors in hypothesis testing. Specifically, Sanov's Theorem implies that both the probability of false alarm and the probability of miss decay exponentially with the number of samples N:

$$P_{\mathrm{F}} = \mathbb{P}\left(\hat{H}(\mathbf{y}) = H_1 \;\middle|\; \mathsf{H} = H_0\right) \tag{8}$$

$$= P_0\{\mathcal{R}_1\} = P_0\{\mathcal{S}_1 \cap \mathcal{P}_N\} \doteq 2^{-ND(p_0^*\|p_0)}, \tag{9}$$

where p_0^* is the I-projection of p_0 onto $\mathcal{S}_1$, and

$$P_{\mathrm{M}} = 1 - P_{\mathrm{D}} = \mathbb{P}\left(\hat{H}(\mathbf{y}) = H_0 \;\middle|\; \mathsf{H} = H_1\right) \tag{10}$$

$$= P_1\{\mathcal{R}_0\} = P_1\{\mathcal{S}_0 \cap \mathcal{P}_N\} \doteq 2^{-ND(p_1^*\|p_1)}, \tag{11}$$

where p_1^* is the I-projection of p_1 onto $\mathcal{S}_0$.

Let us consider the geometry of this problem. The linear family of probability distributions

$$\mathcal{L} = \left\{p \in \mathcal{P}^{\mathcal{Y}} \colon\ \mathbb{E}_p\left[\log \frac{p_1(\mathsf{y})}{p_0(\mathsf{y})}\right] = \gamma\right\} \tag{12}$$

forms the boundary between $\mathcal{S}_0$ and $\mathcal{S}_1$. It is easy to see that both p_0^* and p_1^* belong to $\mathcal{L}$, otherwise we could decrease the corresponding distance between the distribution and its I-projection.

Now we construct the exponential family of weighted geometric means of the distributions p_0 and p_1,

$$\mathcal{E}_{p_0,p_1}(x) = \left\{p \in \mathcal{P}^{\mathcal{Y}} \colon p(y;x) = \frac{p_0(y)^{1-x} p_1(y)^x}{Z(x)}\right\} \tag{13}$$

$$= \left\{p \in \mathcal{P}^{\mathcal{Y}} \colon p(y;x) = \exp\left\{x \log \frac{p_1(y)}{p_0(y)} - \alpha(x) + \log p_0(y)\right\}\right\} \tag{14}$$

for $0 \leq x \leq 1$, and note that it uses p_0 as the base distribution and $t(y) = \log(p_1(y)/p_0(y))$ as the natural statistic. This exponential family therefore intersects the linear family $\mathcal{L}$ at p_0^*, which is the I-projection of the distribution p_0 onto $\mathcal{L}$.

The same exponential family has an alternative parameterization that uses $\tilde{x} = 1 - x$ to index its members:

$$\mathcal{E}_{p_1,p_0}(\tilde{x}) = \left\{ p \in \mathcal{P}^{\mathcal{Y}} \colon p(y;\tilde{x}) = \frac{p_0(y)^{\tilde{x}} p_1(y)^{1-\tilde{x}}}{Z(\tilde{x})} \right\} \tag{15}$$

$$= \left\{ p \in \mathcal{P}^{\mathcal{Y}} \colon p(y;\tilde{x}) = \exp\left\{ \tilde{x} \log \frac{p_0(y)}{p_1(y)} - \alpha(\tilde{x}) + \log p_1(y) \right\} \right\} \tag{16}$$

for $0 \leq \tilde{x} \leq 1$. Similarly, the linear family $\mathcal{L}$ has an alternative representation

$$\mathcal{L} = \left\{ p \in \mathcal{P}^{\mathcal{Y}} \colon \mathbb{E}_p \left[\log \frac{p_0(\mathsf{y})}{p_1(\mathsf{y})} \right] = -\gamma \right\}. \tag{17}$$

Based on these alternative representations, we conclude that p_1^*, which is the I-projection of p_1 onto $\mathcal{L}$, is also an intersection of $\mathcal{E}_{p_0,p_1}$ and $\mathcal{L}$. This proves that $p_1^* = p_0^*$, i.e., the I-projections of p_0 and p_1 onto $\mathcal{L}$ coincide.

The resulting hypothesis testing information geometry is depicted in Fig. 1.

Applying large deviation analysis we conclude that $\mathbb{E}_{p(\cdot;x)}\left[\log(p(\mathsf{y};x)/p_0(\mathsf{y}))\right]$ and $D(p(\cdot;x)\|p_0)$ are monotonically increasing functions of the parameter x. Using the alternative representation, we can also show that $\mathbb{E}_{p(\cdot;x)}\left[\log(p(\mathsf{y};x)/p_1(\mathsf{y}))\right]$ and $D(p(\cdot;x)\|p_1)$ are monotonically decreasing functions of the parameter x. For any member of the exponential family,

$$\mathbb{E}_{p(\cdot;x)} \left[\log \frac{p_1(\mathsf{y})}{p_0(\mathsf{y})} \right] = \mathbb{E}_{p(\cdot;x)} \left[\log \frac{p(\mathsf{y};x)}{p_0(\mathsf{y})} \right] - \mathbb{E}_{p(\cdot;x)} \left[\log \frac{p(\mathsf{y};x)}{p_1(\mathsf{y})} \right] \tag{18}$$

$$= D(p(\cdot;x)\|p_0) - D(p(\cdot;x)\|p_1) \tag{19}$$

is a monotonically increasing function of x. The distribution $p_* = p_0^* = p_1^*$ is the member of the exponential family that satisfies

$$\mathbb{E}_{p_*} \left[\log \frac{p_1(\mathsf{y})}{p_0(\mathsf{y})} \right] = D(p_*\|p_0) - D(p_*\|p_1) = \gamma. \tag{20}$$

We now go back to the expressions for P_{F} and P_{M} in (9) and (11) and arrive at the following asymptotic characterization:

$$P_{\mathrm{F}} \doteq 2^{-N D(p_*\|p_0)}, \quad P_{\mathrm{M}} \doteq 2^{-N D(p_*\|p_1)}, \tag{21}$$

where

$$p_*(y) = \frac{1}{Z(x^*)} p_0(y)^{1-x^*} p_1(y)^{x^*} \tag{22}$$

with x^* chosen such that $D(p_*\|p_0) - D(p_*\|p_1) = \gamma$.

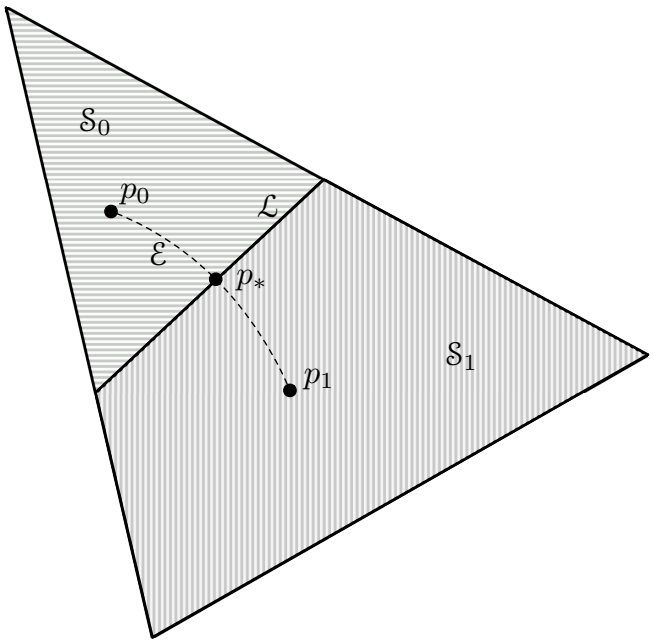

Figure 1: The geometry of hypothesis testing with data distributed i.i.d. according to p_0 or p_1. The false alarm probability is exponentially small with Chernoff exponent $D(p_*\|p_0)$, where p_* is the I-projection of p_0 onto $\mathcal{S}_1$, where the boundary of $\mathcal{S}_1$ is determined by γ.

24.1.1 Classical Hypothesis Testing

In the classical setting, we place an upper limit on one of the error terms and minimize the other one. Here we constrain P_{M} and minimize P_{F}. The analysis for bounding P_{F} and minimizing P_{M} is virtually identical.

Eq. (21) implies that to keep P_{M} from decaying exponentially with N, we need to set $\gamma = D(p_1\|p_0)$, implying $p_* = p_1$. In this case,

$$P_{\mathrm{F}} \doteq 2^{-ND(p_1\|p_0)}, \quad P_{\mathrm{M}} \doteq 1, \tag{23}$$

i.e., P_{M} decays subexponentially while the Chernoff exponent for P_{F} is equal to $D(p_1\|p_0)$. A variant of this fact is often referred to as Stein's Lemma: when $P_{\mathrm{M}} \leq \alpha$, the fastest rate of decay for P_{F} is equal to $D(p_1\|p_0)$.

This setup is a bit unnatural as we fight the exponential behavior of the errors by growing the threshold on the likelihood ratio exponentially in the number of samples N. Since γ is equal to the logarithm of the threshold in the likelihood ratio test, we effectively commit to growing the threshold $\eta = 2^{N\gamma} = 2^{ND(p_1\|p_0)}$ exponentially with N to force one of the errors to decay subexponentially.

24.1.2 Bayesian Hypothesis Testing

In the Bayesian setting, our goal is to minimize the expected cost

$$\mathbb{E}[C] = C_{01}P_0P_{\mathrm{F}} + C_{10}P_1P_{\mathrm{M}} \doteq C_{01}P_0 2^{-ND(p_*\|p_0)} + C_{10}P_1 2^{-ND(p_*\|p_1)} \tag{24}$$

$$\doteq 2^{-N\min\{D(p_*\|p_0), D(p_*\|p_1)\}}. \tag{25}$$

To maximize the rate of the cost decay, we must maximize the smallest of $D(p_*\|p_0)$ and $D(p_*\|p_1)$. Since $D(p(\cdot;x)\|p_0)$ monotonically increases with x and $D(p(\cdot;x)\|p_1)$ monotonically decreases with x, the highest rate of cost decay in (25) is achieved for p_* that satisfies $D(p_*\|p_0) = D(p_*\|p_1)$. This correspond to $\gamma = 0$, and $\eta = 2^{N\gamma} = 1$.

We also note that as $N \to \infty$, the effect of the prior and of the costs is eliminated. In this regime, the Bayesian likelihood ratio test reduces to the ML likelihood ratio test. The probability of error decays exponentially, with the Chernoff exponent equal to $D(p_*\|p_0) = D(p_*\|p_1)$.

24.1.3 Special Case: Extreme Values of Threshold

Before concluding this section, we analyze the special case when both probability distributions p_0 and p_1 belong to the same set $\mathcal{S}_1$. This corresponds to placing the threshold in the LRT so low that the typical sequences with respect to p_0 are still assigned to the decision region $\mathcal{R}_1$, i.e.,

$$\gamma < \mathbb{E}_{p_0}\left[\log \frac{p_1(\mathsf{y})}{p_0(\mathsf{y})}\right] = -D(p_0\|p_1). \tag{26}$$

Similar to the analysis earlier in this section, we can show that the I-projections of both distributions onto $\mathcal{S}_0$ coincide:

$$p_* = \arg\min_{p \in \mathcal{S}_0} D(p||p_1) = \arg\min_{p \in \mathcal{S}_0} D(p||p_0). \tag{27}$$

Then Sanov's theorem implies that the probability that a sequence $\mathbf{y}$ of N i.i.d. samples of y has a type that belongs to $\mathcal{S}_0$ decays exponentially with N, whether the data is generated by p_0 or by p_1:

$$P_{\mathrm{M}} = P_1\{\mathcal{S}_0 \cap \mathcal{P}_N\} \doteq 2^{-ND(p_*||p_1)}, \tag{28}$$

$$1 - P_{\mathrm{F}} = P_0\{\mathcal{S}_0 \cap \mathcal{P}_N\} \doteq 2^{-ND(p_*||p_0)}. \tag{29}$$

In other words, the probability of miss P_{M} decays exponentially, but the probability of false alarm P_{F} approaches 1 as $N \to \infty$.

The other special case of an exceedingly high threshold leads to a virtually identical analysis. If $\gamma > D(p_1\|p_0)$, $N \to \infty$, both the probability of false alarm and the probability of true detection decay exponentially.

24.2 Asymptotics of Parameter Estimation

As we saw earlier, there are many ways to evaluate the quality of estimators. The analysis of their asymptotic behavior is richer still. To develop our results, we begin by developing the concepts of convergence we will need for our development.

24.2.1 Convergence of Random Sequences

In this section, we define some notions of convergence of sequences of random variables that we need for our development.

Definition 1. *A sequence of random variables $\mathsf{z}_1, \mathsf{z}_2, \ldots$ is said to converge* almost surely *(or* with probability one*) to a random variable z if*

$$\mathbb{P}\left(\lim_{N\to\infty} \mathsf{z}_N = \mathsf{z}\right) = 1. \tag{30}$$

We express this convergence using the notation $\mathsf{z}_N \xrightarrow{\text{a.s.}} \mathsf{z}$.

By comparison, recall that the weak law of large numbers (WLLN) focused on convergence in probability, defined as follows.

Definition 2. *A sequence of random variables $\mathsf{z}_1, \mathsf{z}_2, \ldots$ is said to converge* in probability *(or* with high probability*) to a random variable z if for every $\epsilon > 0$ we have*

$$\lim_{N\to\infty} \mathbb{P}\left(|\mathsf{z}_N - \mathsf{z}| < \epsilon\right) = 1. \tag{31}$$

We express this convergence using the notation $\mathsf{z}_N \xrightarrow{\text{p}} \mathsf{z}$.

Almost-sure convergence is stronger than convergence in probability, as the following alternative characterization emphasizes.[1]

Fact 1 (Alternative Characterization of Almost Sure Convergence)**.** *For a collection of random variables* $\mathsf{z}, \mathsf{z}_1, \mathsf{z}_2, \ldots$*, we have* $\mathsf{z}_N \xrightarrow{\text{a.s.}} \mathsf{z}$ *if and only if for every* $\epsilon > 0$*, there exists an* N_0 *such that*

$$\mathbb{P}\left(|\mathsf{z}_N - \mathsf{z}| < \epsilon \text{ for all } N > N_0\right) = 1.$$

In other words, for an almost-sure convergent sequence, when a sample path gets within ϵ of z, it remain within that tolerance forever thereafter (with probability one).

It is also useful to note that almost-sure convergence follows when the convergence in probability is sufficiently fast, i.e., when $\mathbb{P}\left(|\mathsf{z}_N - \mathsf{z}| > \epsilon\right)$ decays sufficiently strongly with N. For example, we have the following fact

Fact 2. *For a collection of random variables* $\mathsf{z}, \mathsf{z}_1, \mathsf{z}_2, \ldots$*, we have* $\mathsf{z}_N \xrightarrow{\text{a.s.}} \mathsf{z}$ *if for every* $\epsilon > 0$ *we have*

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|\mathsf{z}_n - \mathsf{z}| > \epsilon\right) < \infty.$$

In contrast, a useful notion of convergence that is *weaker* than convergence in probability is convergence in *distribution*.

Definition 3. *A sequence of random variables* $\mathsf{z}_1, \mathsf{z}_2, \ldots$ *is said to converge* in distribution *(or* in law*) to a random variable* z *if*

$$\lim_{N\to\infty} P_{\mathsf{z}_N}(z) = P_{\mathsf{z}}(z) \tag{32}$$

for all z *at which* $P_{\mathsf{z}}(\cdot)$ *is continuous. We express this convergence using the notation* $\mathsf{z}_N \xrightarrow{\text{d}} \mathsf{z}$*.*

Although we omit a proof, an equivalent characterization of convergence in distribution is that

$$\lim_{N\to\infty} \mathbb{E}\left[g(\mathsf{z}_N)\right] = \mathbb{E}\left[g(\mathsf{z})\right], \quad \text{for all bounded, continuous functions } g(\cdot).$$

Moreover, convergence in distribution implies that for some events $\mathcal{A} \subset \mathbb{R}$ for which probabilities can be evaluated, we have

$$\mathbb{P}_{p_{\mathsf{z}_N}}[\mathcal{A}] = \int_{\mathcal{A}} p_{\mathsf{z}_N}(c)\,\mathrm{d}c \quad \to \quad \mathbb{P}_{p_{\mathsf{z}}}[\mathcal{A}] = \int_{\mathcal{A}} p_{\mathsf{z}}(c)\,\mathrm{d}c, \qquad \text{as } N \to \infty. \tag{33}$$

[1] As an illustration of this distinction, there are examples of sequences $\mathsf{z}_1, \mathsf{z}_2, \ldots$ that converge in probability to z but with probability one *no* individual sequence converges, i.e., $\mathbb{P}\left(\lim_{N\to\infty} \mathsf{z}_N = \mathsf{z}\right) = 0$.

However, in general, convergence in distribution certainly does *not* imply (33) for *all* such events $\mathcal{A}$. Thus Definition 3 can be viewed as a relatively weak form of convergence in distribution. To obtain such a result requires a stronger notion of convergence in distribution, such as convergence in *divergence*.[2]

Definition 4. *A sequence of random variables* $\mathsf{z}_1, \mathsf{z}_2, \ldots$ *is said to converge* in divergence *(or* strongly in law*) to a random variable* z *if*

$$\lim_{N\to\infty} D(p_{\mathsf{z}_N} \| p_{\mathsf{z}}) = 0. \tag{34}$$

We express this convergence using the notation $\mathsf{z}_N \xrightarrow{\mathrm{D}} \mathsf{z}$.

It is this stronger notion of convergence in distribution that is often needed in inference applications.[3] To confirm that it is indeed stronger, it suffices to note that for arbitrary densities p and q we have, for an arbitrary y,

$$\begin{aligned}
|P(y) - Q(y)| &= \left| \int_{-\infty}^{y} p(b)\,\mathrm{d}b - \int_{-\infty}^{y} q(b)\,\mathrm{d}b \right| \\
&= \left| \int_{-\infty}^{y} [p(b) - q(b)]\,\mathrm{d}b \right| \\
&\le \int_{-\infty}^{y} |p(b) - q(b)|\,\mathrm{d}b \\
&\le \int_{-\infty}^{+\infty} |p(b) - q(b)|\,\mathrm{d}b \\
&\le \sqrt{D(p\|q)/(2\log e)},
\end{aligned}$$

where to obtain the last inequality we have used (a tight version of) Pinsker's inequality for continuous alphabets. Hence if $p, p_1, p_2, \ldots$ are such that $D(p_N\|p) \to 0$, then $|P_N(y) - P(y)| \to 0$ as well.

The following theorem asserts that continuous transformations of random variables preserve their mode of convergence.

Theorem 1 (Continuous Mapping Theorem)**.** *If* $\mathsf{z}, \mathsf{z}_1, \mathsf{z}_2, \ldots$ *are defined over* $\mathbb{R}^k$ *and if* $g\colon \mathbb{R}^k \mapsto \mathbb{R}$ *is a continuous function, then:*

$$\mathsf{z}_N \xrightarrow{\mathrm{d}} \mathsf{z} \quad \Rightarrow \quad g(\mathsf{z}_N) \xrightarrow{\mathrm{d}} \mathsf{z} \tag{35}$$

$$\mathsf{z}_N \xrightarrow{\mathrm{p}} \mathsf{z} \quad \Rightarrow \quad g(\mathsf{z}_N) \xrightarrow{\mathrm{p}} \mathsf{z} \tag{36}$$

$$\mathsf{z}_N \xrightarrow{\mathrm{a.s.}} \mathsf{z} \quad \Rightarrow \quad g(\mathsf{z}_N) \xrightarrow{\mathrm{a.s.}} \mathsf{z} \tag{37}$$

More generally the theorem holds whenever the set of points at which $g(\cdot)$ is discontinuous has probability zero under p_{z}.

[2] Another example of such a sufficiently stronger notion is convergence in *density*, i.e., $p_{\mathsf{z}_N}(z) \to p_{\mathsf{z}}(z)$ as $N \to \infty$, for each z.

[3] Recall that it is what we needed in our approximation results on exponential families.

24.2.2 Convergence of Sample-Averages

Equipped with the preceding notions of convergence, we now discuss how sample averages of i.i.d. random variables behave with respect to these notions, which our analysis of estimators will rely on.

Theorem 2 (Strong Law of Large Numbers (SLLN))**.** *Let $\mathsf{w}_1, \ldots, \mathsf{w}_N$ be a set of i.i.d. random variables with mean μ and $\mathbb{E}\left[|\mathsf{w}_n|\right] < \infty$. Then*

$$\frac{1}{N}\sum_{n=1}^{N} \mathsf{w}_n \xrightarrow{\text{a.s.}} \mu, \quad \text{as } N \to \infty, \tag{38}$$

i.e., the sample-average converges almost surely to μ.

By comparison, the weak law states that

$$\frac{1}{N}\sum_{n=1}^{N} \mathsf{w}_n \xrightarrow{\text{p}} \mu.$$

Note that the strong law does not require stronger conditions than what we required in our statement of the weak law, in an earlier installment of the notes.

For analyzing estimators, we need a more general version of the law of large numbers, which is called the *uniform* law.

Theorem 3 (Uniform Law of Large Numbers (ULLN))**.** *Let $\mathsf{w}_1, \ldots, \mathsf{w}_N$ be a set of i.i.d. random variables, let Θ be a compact parameter set, and let g be a function such that: i) $g(w;\theta)$ is continuous in $\theta \in \Theta$, and ii) there exists $g_+(w) > 0$ such that $\mathbb{E}\left[g_+(\mathsf{w})\right] < \infty$ and $|g(w;\theta)| \leq g_+(w)$ for all w and $\theta \in \Theta$. Then*

$$\max_{\theta\in\Theta} \left| \frac{1}{N}\sum_{n=1}^{N} g(\mathsf{w}_n;\theta) - \mu(\theta) \right| \xrightarrow{\text{a.s.}} 0, \quad \text{as } N \to \infty, \tag{39}$$

where $\mu(\theta) = \mathbb{E}\left[g(\mathsf{w};\theta)\right]$, i.e., the sample-average almost surely converges uniformly[4] to its mean.

This is the *strong* uniform law of large numbers, corresponding to almost sure convergence (i.e., with probability one). There is also a weak version, where the convergence is in probability (i.e., with high probability).

The following celebrated theorem establishes that the distribution of a sample-average converges, in an appropriate sense, to a Gaussian.

[4] As a (nonrandom) illustration of the kind of convergence phenomena this result is trying to avoid, consider a function $f_N \colon [0,1] \mapsto [0,1]$ for $N = 1, 2, \ldots$, where $f_N(x) = \mathbb{1}_{1/N^2 \leq x \leq 1/N}$. Then $f_n(x) \to 0$ as $n \to \infty$ for each $x \in [0,1]$, i.e., $f_N(x)$ convergences *pointwise* to $f(x) \equiv 0$. However $\max_{x\in[0,1]} |f_N(x) - f(x)| = \max_{x\in[0,1]} |f_N(x)| = 1$ for all $N \geq 1$, so $f_N(x)$ does *not* converge uniformly to $f(x) \equiv 0$.

Theorem 4 (Central Limit Theorem (CLT))**.** *Let $\mathsf{w}_1, \ldots, \mathsf{w}_N$ be a set of i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Then*[5]

$$\sqrt{N}\left[\frac{1}{N}\sum_{n=1}^{N}\mathsf{w}_n - \mu\right] \xrightarrow{\mathrm{d}} \mathtt{N}(0,\sigma^2), \quad \text{as } N \to \infty. \tag{40}$$

Note that (40) expresses that for large N the sample average behaves like

$$\frac{1}{N}\sum_{n=1}^{N}\mathsf{w}_n \sim \mathtt{N}\left(\mu, \frac{\sigma^2}{N}\right).$$

As an aside, a theorem due to Pólya establishes that if $\mathsf{z}_N \xrightarrow{\mathrm{d}} \mathsf{z}$ and $P_{\mathsf{z}}(\cdot)$ is continuous, then the convergence is uniform, i.e.,

$$\lim_{N\to\infty}\max_{z}|P_{\mathsf{z}_N}(z) - P_{\mathsf{z}}(z)| = 0.$$

Hence, the convergence in the CLT is uniform. A useful bound on the (specifically, $1/\sqrt{N}$) rate of this uniform convergence is provided by what is referred to as the Berry-Esseen Theorem.

There is also a stronger CLT corresponding to convergence in divergence, which is particularly useful for inference applications.

Theorem 5 (Strong Central Limit Theorem (SCLT))**.** *Let $\mathsf{w}_1, \ldots, \mathsf{w}_N$ be a set of i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Then*

$$\mathsf{e}_N \triangleq \sqrt{N}\left[\frac{1}{N}\sum_{n=1}^{N}\mathsf{w}_n - \mu\right] \xrightarrow{\mathrm{D}} \mathtt{N}(0,\sigma^2), \quad \text{as } N \to \infty \tag{41}$$

if and only if $D(p_{\mathsf{e}_N}\|N(0,\sigma^2)) < \infty$ for some N.

24.2.3 Asymptotics of ML Estimates

Let $\{p_{\mathsf{y}}(\cdot;a)\colon a \in \mathcal{X}\}$ be a parameterized family of distributions, y be a random variable distributed according to a family member $p_{\mathsf{y}}(\cdot;x)$. We start by constructing the normalized log-likelihood of N i.i.d. samples drawn from an arbitrary member of the family $p_{\mathsf{y}}(\cdot;a)$, viewed as a function of a:

$$\ell_{\mathbf{y}}^{N}(a) = \frac{1}{N}\sum_{n=1}^{N}\log p_{\mathsf{y}}(\mathsf{y}_n;a). \tag{42}$$

[5]Here we adopt a slight abuse of notation, since we really should write $\xrightarrow{\mathrm{d}} \mathsf{z} \sim \mathtt{N}(0,\sigma^2)$. However, we will use this convenient alternate notation interchangeably in the sequel, since the risk of confusion is minimal.

Since $\ell_{\mathbf{y}}^N(a)$ depends on the random variables $\mathsf{y}_1, \ldots, \mathsf{y}_N$, it is also a random variable. At the same time, we can view it as a (probabilistic) function of a. It is easy to see that

$$\mathbb{E}_{p_{\mathsf{y}}(\cdot;x)} \left[\log p_{\mathsf{y}}(\mathsf{y}; a)\right] = -D(p_{\mathsf{y}}(\cdot; x) \| p_{\mathsf{y}}(\cdot; a)) - H(p_{\mathsf{y}}(\cdot; x)). \tag{43}$$

The LLN therefore implies that

$$\ell_{\mathbf{y}}^N(a) \to g(a) \triangleq -D(p_{\mathsf{y}}(\cdot; x) \| p_{\mathsf{y}}(\cdot; a)) - H(p_{\mathsf{y}}(\cdot; x)). \tag{44}$$

In other words, with probability one the likelihood function $L_N(a; \mathbf{y})$ we observe for any particular sequence $\mathbf{y}$ will approach $g(a)$ as $N \to \infty$. Since the maximum of the function $g(a)$ is achieved at the true value of the parameter x, we expect the ML estimate of x to be close to x.

The following theorem makes this framework of reasoning precise establishes conditions under which this indeed happens.

Theorem 6. *Consider a model family $\{p_{\mathsf{y}}(\cdot; a) \colon a \in \mathcal{X}\}$ for a parameter set $\mathcal{X} \subset \mathbb{R}$, and let $\mathsf{y}_1, \ldots, \mathsf{y}_N$ be i.i.d. according to the model $p_{\mathsf{y}}(\cdot; x)$ for some $x \in \mathcal{X}$. Then if:*

1. *$\mathcal{X}$ is compact;*
2. *$p_{\mathsf{y}}(y; a)$ is continuous in $a \in \mathcal{X}$ for all y and continuous in y in a neighborhood of $a = x$;*
3. *there exists an ℓ_y^+ such that $\mathbb{E}\left[|\ell_{\mathsf{y}}^+|\right] < \infty$ and*

$$\tilde{\ell}_y(a) \triangleq \log \frac{p_{\mathsf{y}}(y; a)}{p_{\mathsf{y}}(y; x)} \leq \tilde{\ell}_y^+, \quad \textit{for all } a \in \mathcal{X} \textit{ and } y; \textit{ and}$$

4. *the model is identifiable in the sense that $x' = x$ if $p_{\mathsf{y}}(y; x') = p_{\mathsf{y}}(y; x)$ (almost everywhere);*

then the ML estimator $\hat{x}_N(\mathsf{y}_1, \ldots, \mathsf{y}_N)$ satisfies

$$\hat{x}_N(\mathsf{y}_1, \ldots, \mathsf{y}_N) \xrightarrow{\text{a.s.}} x, \tag{45}$$

i.e., $\hat{x}_N(\mathsf{y}_1, \ldots, \mathsf{y}_N)$ converges to x as $N \to \infty$ with probability one.

When (45) holds for an estimator, we say that the estimator is *consistent*. More precisely, we say it is *strongly* consistent, to distinguish this from the scenario where $\hat{\mathsf{x}}_N \xrightarrow{\text{p}} x$, in which case the estimator is weakly consistent. Thus, Theorem 6 establishes that the ML estimator is consistent.

Proof. First, for $a \in \mathcal{X}$ let

$$\tilde{\ell}_{\mathbf{y}}^N(a) \triangleq \frac{1}{N} \sum_{n=1}^{N} \tilde{\ell}_{y_n}(a).$$

Then

$$\hat{x}_N(y_1, \ldots, y_N) = \arg\max_{a \in \mathcal{X}} \tilde{\ell}_{\mathbf{y}}^N(a), \tag{46}$$

and

$$\mathbb{E}\left[\tilde{\ell}_{\mathbf{y}}^N(a)\right] = -D(p_y(\cdot; x) \| p_y(\cdot; a)) \leq 0. \tag{47}$$

Next, note that

$$\tilde{\ell}_{\mathbf{y}}^N(\hat{x}_N) = \max_{a \in \mathcal{X}} \tilde{\ell}_{\mathbf{y}}^N(a) \geq \tilde{\ell}(x) = 0. \tag{48}$$

Now for $\rho > 0$ define the compact set

$$\mathcal{X}_\rho \triangleq \{a \in \mathcal{X} \colon |a - x| \geq \rho\},$$

and let

$$\bar{D}_\rho = \max_{a \in \mathcal{X}_\rho} -D(p_y(\cdot; x) \| p_y(\cdot; a)) < 0, \tag{49}$$

where the strict inequality follows from the fact that $x \notin \mathcal{X}_\rho$. Finally, let $x_\rho \in \mathcal{X}_\rho$ denote the value a at which the maximum in (49) is achieved.

Then when the model is such that the ULLN can be applied for the parameter set $\mathcal{X}_\rho$, we know that for any choice of $\epsilon > 0$, there exists an N_0 such that for all $N > N_0$,

$$\left|\tilde{\ell}_{\mathbf{y}}^N(a) + D(p_y(\cdot; x) \| p_y(\cdot; a))\right| < \epsilon, \quad \text{for all } a \in \mathcal{X}_\rho, \tag{50}$$

with probability one. So let us choose, say, $\epsilon = -\bar{D}_\rho/2$, so (50) implies

$$\tilde{\ell}_{\mathbf{y}}^N(a) < -\bar{D}/2 - D(p_y(\cdot; x) \| p_y(\cdot; a)), \quad \text{for all } a \in \mathcal{X}_\rho. \tag{51}$$

Hence

$$\begin{aligned} \max_{a \in \mathcal{X}_\rho} \tilde{\ell}_{\mathbf{y}}^N(a) &\leq -\bar{D}_\rho/2 + \max_{a \in \mathcal{X}_\rho} -D(p_y(\cdot; x) \| p_y(\cdot; a)) \\ &= -\bar{D}_\rho/2 + \bar{D}_\rho = \bar{D}_\rho/2 \\ &< 0. \end{aligned} \tag{52}$$

Comparing (52) with (48) we conclude that $\hat{x}_N(\mathsf{y}_1, \ldots, \mathsf{y}_N) \notin \mathcal{X}_\rho$ as $N \to \infty$ with probability one, i.e., $|\hat{x}_N(\mathsf{y}_1, \ldots, \mathsf{y}_N) - x| < \rho$. Since $\rho > 0$ is arbitrary, we conclude $\hat{x}_N(\mathsf{y}_1, \ldots, \mathsf{y}_N) \xrightarrow{\text{a.s.}} x$. □

Not only does the ML estimator $\hat{\mathsf{x}}_N$ tend toward the correct value x for large N, but its fluctuations about x are Gaussian in the large N regime. In particular, a consistent ML estimator behaves for large N like

$$\hat{\mathsf{x}}_N \sim \mathtt{N}\left(x, \frac{1}{N J_y(x)}\right). \tag{53}$$

The following theorem more precisely characterizes this behavior under suitable conditions (that, for example, allow the needed ULLN to be applied).

Theorem 7. *Consider a model family* $\{p_{\mathsf{y}}(\cdot;a)\colon a\in\mathcal{X}\}$ *for a parameter set* $\mathcal{X}\subset\mathbb{R}$*, and let* $\mathsf{y}_1,\dots,\mathsf{y}_N$ *be i.i.d. according to the model* $p_{\mathsf{y}}(\cdot;x)$ *for some* $x\in\mathcal{X}$*. If*

1. $\mathcal{X}$ *is open;*
2. $\partial^2 p_{\mathsf{y}}(y;a)/\partial a^2$ *is continuous in* $a\in\mathcal{X}$ *for all* y*, and we have the usual regularity condition*
$$\int \frac{\partial}{\partial a} p_{\mathsf{y}}(y;a)\,\mathrm{d}y = 0;$$
3. *there exists* $\ddot{\ell}_y^+$ *such that* $\mathbb{E}\left[\ddot{\ell}_{\mathsf{y}}^+;x\right]<\infty$ *and*
$$|\ddot{\ell}_y(a)|\le \ddot{\ell}_y^+, \quad \textit{for all } a \textit{ in a neighborhood of } x \textit{ and all } y,$$
where
$$\ell_y(a)\triangleq \log p_{\mathsf{y}}(y;a); \tag{54}$$
4. $J_{\mathsf{y}}(x)>0$*; and*
5. *the ML estimator* $\hat{\mathsf{x}}_N=\hat{x}_N(\mathsf{y}_1,\dots,\mathsf{y}_N)$ *is (strongly) consistent;*

then
$$\sqrt{N}(\hat{\mathsf{x}}_N-x)\xrightarrow{\mathrm{d}}\mathtt{N}(0,1/J_{\mathsf{y}}(x)). \tag{55}$$

Proof. Let
$$\ell_{\mathbf{y}}^N(a)\triangleq\frac{1}{N}\sum_{n=1}^N \ell_{\mathsf{y}_n}(a), \tag{56}$$
which we note is a sample-average of i.i.d. random variables. Then by Taylor's Theorem[6] we can write the score function as
$$\dot{\ell}_{\mathbf{y}}^N(a)=\dot{\ell}_{\mathbf{y}}^N(x)+(a-x)\,\ddot{\ell}_{\mathbf{y}}^N(a') \tag{58}$$
for some a' between a and x. Let us choose $a=\hat{\mathsf{x}}_N$, in which case, since $\hat{\mathsf{x}}_N$ is achieved at a stationary point of the likelihood function (because $\mathcal{X}$ is open), we have
$$0=\dot{\ell}_{\mathbf{y}}^N(\hat{\mathsf{x}}_N)=\dot{\ell}_{\mathbf{y}}^N(x)+(\hat{\mathsf{x}}_N-x)\,\ddot{\ell}_{\mathbf{y}}^N(x'), \quad \text{for some } x' \text{ between } \hat{\mathsf{x}}_N \text{ and } x,$$

[6]The special case of Taylor's Theorem we need states that if a function f is differentiable, then for any t and t_0 we can write
$$f(t)=f(t_0)+(t-t_0)\dot{f}(t') \tag{57}$$
for some t' between t and t_0.

whence

$$\sqrt{N}(\hat{x}_N - x) = \frac{\sqrt{N}\dot{\ell}_{\mathbf{y}}^N(x)}{-\ddot{\ell}_{\mathbf{y}}^N(x')}. \tag{59}$$

Now since, since

$$\dot{\ell}_{\mathbf{y}}^N(a) = \frac{1}{N}\sum_{n=1}^{N} \dot{\ell}_{y_n}(a)$$

is also the sample-average of i.i.d. random variables for any a, and since, as we saw in our development of the Cramér-Rao bound,

$$\mathbb{E}\left[\dot{\ell}_{\mathbf{y}}^N(x)\right] = 0 \quad \text{and} \quad \mathbb{E}\left[\left(\dot{\ell}_{\mathbf{y}}^N(x)\right)^2\right] = J_y(x)$$

under our regularity conditions, via the CLT we conclude that the numerator of the right-hand side of (59) satisfies

$$\sqrt{N}\dot{\ell}_{\mathbf{y}}^N(x) \xrightarrow{\text{d}} \mathtt{N}(0, J_y(x)).$$

Turning next to the denominator of the right-hand side of (59), we note that if

$$-\ddot{\ell}_{\mathbf{y}}^N(x') \xrightarrow{\text{a.s.}} J_y(x) \tag{60}$$

then by the Continuous Mapping Theorem we obtain (55), since $\operatorname{var}[\mathsf{z}/J_y(x)] = (\operatorname{var}\mathsf{z})/J_y(x)^2$ for any random variable z.

Thus, it remains only to show (60). Proceeding, since[7] $\mathbb{E}\big[\ddot{\ell}_{\mathbf{y}}(a); x\big]$ is continuous in a and since, as we also showed in our development of the Cramér-Rao bound, $\mathbb{E}\big[\ddot{\ell}_{\mathbf{y}}(x); x\big] = -J_y(x)$, for any $\epsilon > 0$ there exists a $\rho > 0$ such that

$$\left|\mathbb{E}\left[\ddot{\ell}_{\mathbf{y}}(a); x\right] + J_y(x)\right| < \epsilon/2$$

whenever $a \in \bar{\mathbb{S}}_\rho \triangleq \{a' \colon |a' - x| \leq \rho\}$.

Now since

$$\ddot{\ell}_{\mathbf{y}}^N(a) = \frac{1}{N}\sum_{n=1}^{N} \ddot{\ell}_{y_n}(a)$$

is also a sample-average of i.i.d. random variables, by the ULLN there exists n_1 such that for all $n > n_1$

$$\max_{a \in \bar{\mathbb{S}}_\rho} \left|\ddot{\ell}_{\mathbf{y}}(a) - \mathbb{E}\left[\ddot{\ell}_{\mathbf{y}}(a); x\right]\right| < \epsilon/2 \tag{61}$$

Moreover, since $\hat{x}_N$ is (strongly) consistent, there exists n_2 such that for all $n > n_2$ we have $\hat{x}_N \in \bar{\mathbb{S}}_\rho$ and, in turn, $x' \in \bar{\mathbb{S}}_\rho$, since x' lies between $\hat{x}_N$ and x.

[7] Recall our notational convention is such that $\mathbb{E}[f(\mathsf{y}); x] = \mathbb{E}_{p_y(\cdot;x)}[f(\mathsf{y})]$ for a given f.

Hence, for all $n > \max(n_1, n_2)$ we have

$$\begin{aligned}
\left|-\ddot{\ell}_{\mathbf{y}}(x') - J_y(x)\right| &= \left|\ddot{\ell}_{\mathbf{y}}(x') + J_y(x)\right| \\
&= \left|\ddot{\ell}_{\mathbf{y}}(x') - \mathbb{E}\left[\ddot{\ell}_{\mathbf{y}}(x'); x\right] + \mathbb{E}\left[\ddot{\ell}_{\mathbf{y}}(x'); x\right] + J_y(x)\right| \\
&\leq \left|\ddot{\ell}_{\mathbf{y}}(x') - \mathbb{E}\left[\ddot{\ell}_{\mathbf{y}}(x'); x\right]\right| + \left|\mathbb{E}\left[\ddot{\ell}_{\mathbf{y}}(x'); x\right] + J_y(x)\right| \\
&\leq \underbrace{\max_{a \in \bar{\mathbb{S}}_\rho} \left|\ddot{\ell}_{\mathbf{y}}(a) - \mathbb{E}\left[\ddot{\ell}_{\mathbf{y}}(a); x\right]\right|}_{\leq \epsilon/2} + \underbrace{\left|\mathbb{E}\left[\ddot{\ell}_{\mathbf{y}}(x'); x\right] + J_y(x)\right|}_{\leq \epsilon/2} \\
&\leq \epsilon,
\end{aligned}$$

which establishes (60) as desired. □

We observe that Theorem 7 further establishes that the variance of the ML estimator asymptotically satisfies the Cramér-Rao bound in a particular sense. More generally, we have the following definition.

Definition 5. *Let $\mathsf{y}_1, \ldots, \mathsf{y}_N$ be i.i.d. according to the model $p_y(\cdot; x)$ for some $x \in \mathcal{X}$. An estimator $\hat{\mathsf{x}}_N = \hat{x}_N(\mathsf{y}_1, \ldots, \mathsf{y}_N)$ is* asymptotically efficient *if*

$$\operatorname{var}\left[\sqrt{N}(\hat{\mathsf{x}}_N - x)\right] \to 1/J_y(x), \quad \text{as } N \to \infty. \tag{62}$$

Hence, Theorem 7 establishes that the ML estimator is *asymptotically efficient*.

It is straightforward to extend the results of this section to the case of vector parameters. In such analysis, the inverse of the Fisher information matrix, for example, determines the asymptotic covariance structure of the estimator.

24.2.4 Mismatched ML Estimation

In the preceding analysis, the true model lies in the parameterized class, i.e. $x \in \mathcal{X}$. However, in practice, we cannot always know we have chosen $\mathcal{X}$ large enough such that this is the case. Let us examine what happens when $x \notin \mathcal{X}$.

If the data is generated by a probability distribution q, the SLLN implies

$$\frac{1}{N}\sum_{n=1}^{N} \log \frac{p_y(\mathsf{y}_n; x)}{q(\mathsf{y}_n)} \xrightarrow{\text{a.s.}} -D(q(\cdot) \| p_y(\cdot; x)), \tag{63}$$

therefore

$$p_{\mathbf{y}}(\mathbf{y}; x) = \prod_{n=1}^{N} p_y(\mathsf{y}_n; x) \doteq e^{-ND(q(\cdot)\|p_y(\cdot;x))} \prod_{n=1}^{N} q(\mathsf{y}_n). \tag{64}$$

If q is not in the family $p_{\mathsf{y}}(\cdot; x)$, the ML estimate is equal to the value of the parameter x^* that minimizes $D(q(\cdot)\|p_{\mathsf{y}}(\cdot; x))$. Since $D(q(\cdot)\|p_{\mathsf{y}}(\cdot; x))$ is equal to the log-loss penalty for making inference based on $p_{\mathsf{y}}(\cdot; x)$ when the observations are generated by q, ML estimation achieves asymptotically optimal solution with respect to the log-loss function.

24.2.5 Asymptotics of MAP Estimation

In the Bayesian case, we model the hidden parameter as a random variable x distributed according to the probability distribution $p_{\mathsf{x}}(\cdot)$. To understand the asymptotic behavior of the MAP estimates, we consider the log ratio of the posterior probability distribution of x and the posterior probability of the true parameter value x_0, given a sequence $\mathbf{y}$ of N i.i.d. samples of the random variable y:

$$\frac{1}{N}\log\frac{p_{\mathsf{x}|\mathbf{y}}(x|\mathbf{y})}{p_{\mathsf{x}|\mathbf{y}}(x_0|\mathbf{y})} = \frac{1}{N}\log\frac{p_{\mathsf{x}}(x)}{p_{\mathsf{x}}(x_0)} + \frac{1}{N}\sum_{n=1}^{N}\log\frac{p_{\mathsf{y}}(\mathsf{y}_n|x)}{p_{\mathsf{y}}(\mathsf{y}_n|x_0)}. \tag{65}$$

As $N \to \infty$, the first term vanishes while the second term does not. Instead, it converges to $g(x) = -D(p_{\mathsf{y}}(\cdot|x_0)\|p_{\mathsf{y}}(\cdot|x))$. The estimates we construct by maximizing the posterior probability distribution become increasingly close to those constructed based on maximizing the likelihood. This shows again that asymptotically, the importance of priors is diminished. However, choosing a prior might affect the rate of convergence, as we will see in the next section when we analyze model capacity.

24.3 Further Reading

The following texts provide useful supplementary reading:

Cover and Thomas, *Elements of Information Theory,* provide an in-depth discussion and analysis of typical sets, the method of types and hypothesis testing.

Serfling, *Approximation Theorems of Mathematical Statistics,* offers a detailed analysis of asymptotic behavior of parameter estimation and hypothesis testing.

Ferguson, *A Course in Large Sample Theory,* has a careful, self-contained treatment of several aspects of estimation asymptotics.