Chapter 9
Classical Limit Theorems

In the last section, we have shown that if the paramter-to-observable map is linear, i.e. $h(m) = \mathscr{A}m$, and both the noise and the prior models are Gaussian, then the MAP point and the posterior covariance matrix are exactly the solution and the inverse of the Hessian of the Tikhonov functional, respectively. Moreover, since the posterior is Gaussian, the MAP point is identically the mean, and hence the posterior distribution is completely characterized. In practice, $h(m)$ is typically nonlinear. Consequently, the posterior distribution is no longer Gaussian. Nevertheless, the MAP point is still the solution of the Tikhonov functional, though the mean and the convariance matrix are to be determined. The question is how to estimate the mean and the covariance matrix of a non-Gaussian density.

We begin by recalling the definition the mean

$$\overline{m} = \mathbb{E}[m],$$

and a natural idea is to approximate the integral by some numerical integration. For example, suppose $S = [0,1]$ and then we can divide S into N intervals, each of which has length of $1/N$. Using a rectangle rule gives

$$\overline{m} \approx \frac{(m_1 + \ldots + m_N)}{N}. \tag{9.1}$$

But this kind of method cannot be extended to $S = \mathbb{R}^n$. This is where the central limit theorem and law of large numbers come to rescue. They say that the simple formula (9.1) is still valid with a simple error estimation expression.

9.1 Some classical limit theorems

Theorem 9.1 (The central limit theorem (CLT)). *Assume that real valued random variables $m_1, \ldots$ are independent and identically distributed (iid), each with expectation $\overline{m}$ and variance σ^2. Then*

$$Z_N = \frac{1}{\sigma\sqrt{N}}(m_1 + m_2 + \cdots + m_N) - \frac{\overline{m}}{\sigma}\sqrt{N}$$

converges, in distribution[1]*, to a standard normal random variable. In particular,*

$$\lim_{N\to\infty} \mathbb{P}[Z_N \le m] = \frac{1}{2\pi}\int_{-\infty}^{m} \exp\left(-\frac{t^2}{2}\right) dt \tag{9.2}$$

Proof. The proof is elementary, though technical, using the concept of characteristic function (Fourier transform of a random variable). A complete proof can be consulted from [16].

Theorem 9.2 (Strong law of large numbers (LLN)). *Assume random variables* $m_1, \ldots$ *are independent and identically distributed (iid), each with finite expectation* $\overline{m}$ *and finite variance* σ^2*. Then*

$$\lim_{N\to\infty} S_N \overset{def}{=} \frac{1}{N}(m_1 + m_2 + \cdots + m_N) = \overline{m} \tag{9.3}$$

almost surely[2].

Proof. A beautiful, though not classical, proof of this theorem is based on backward martingale, tail σ-algebra, and uniform integrability. Let's accept it in this note and see [16] for a complete proof.

Remark 9.1. The central limit theorem says that no matter what the underlying common distribution looks like, the sum of iid random variables, when properly scaled and centralized, converges in distribution to a standard normal distribution. The strong law of large numbers, on the other hand, states that the average of the sum is, as expected in the limit, precisely the mean of the common distribution with probability one.

Exercise 9.1 (Numerical verification of CLT and LLN). In Matlab, draw N iid samples from the standard uniform distribution, for which we know that the mean is 0.5.

1. Plot the histogram of Z_N as a function of N and observe the convergence of the histogram to the standard normal distribution. Estimate the mean and variance of Z_N as a function of N and see whether they converege to 0 and 1.
2. Plot S_N as a function of N and see whether it converges to 0.5.

•

[1] Convergence in distribution is also known as weak convergence and it is beyond the scope of this introductory note. You can think of the distribution of Z_n is more and more like the standard normal distribution as $n \to \infty$, and it is precisely (9.2).

[2] Almost sure convergence is the same as convergence with probability one, that is, the event on which the convergence (9.3) does not happen has zero probability.

Both the central limit theorem (CLT) and the strong law of large numbers (LLN) are useful, particularly LLN, and we use them routinely. For example, if we are given an iid sample $\{m_1, m_2, \cdots, m_N\}$ from a common distribution $\pi(m)$, the first thing we should do is perhaps to compute the the sample mean S_N to estimate the actual mean $\overline{m}$. From LLN we know that the sample mean can be as close as desired if N is sufficiently large. A question immediately arises is whether we can estimate the error between the sample mean and the truth mean, given a finite N. Let us first give an answer based on a simple application of the CLT. Since the sample $\{m_1, m_2, \cdots, m_N\}$ satisfies the condition of the CLT, we know that Z_N converges to $\mathcal{N}(0,1)$. It follows that, at least for sufficiently large N, the mean squared error between z_N and 0 can be estimated as

$$1 \approx \mathbb{V}ar[Z_N] \overset{\text{def}}{=} \|Z_N - \overline{Z_N}\|^2_{L^2(S,\mathbb{P})} \approx \|Z_N - 0\|^2_{L^2(S,\mathbb{P})} \overset{\text{def}}{=} \mathbb{E}\left[(Z_N - 0)^2\right],$$

which, after some simple algebra manipulations, can be rewritten as

$$\|S_N - \overline{m}\|^2_{L^2(S,\mathbb{P})} \approx \frac{\sigma^2}{N} \tag{9.4}$$

Exercise 9.2. Show that (9.4) holds. •

The result (9.4) shows that the error of the sample mean S_N in the $L^2(S,\mathbb{P})$-norm goes to zero like $1/\sqrt{N}$. One should be aware of the popular statement that the error goes to zero like $1/\sqrt{N}$ independent of dimension is not entirely correct because the variance σ^2, and hence the standard deviation σ, of the underlying distribution $\pi(m)$ may depend on the dimension n.

If we are a little bit delicate, we may not feel completely comfortable with the error estimate (9.4) since we can rewrite it as

$$\|S_N - \overline{m}\|_{L^2(S,\mathbb{P})} = C\frac{\sigma}{\sqrt{N}},$$

and we are not sure how big C is and the dependence of C on N. Let us attempt to determine C. We have

$$\begin{aligned}\|S_N - \overline{m}\|^2_{L^2(S,\mathbb{P})} &= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{i=1}^N (m_i - \overline{m})\right)\left(\sum_{j=1}^N (m_j - \overline{m})\right)\right] \\ &= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{i=1}^N (m_i - \overline{m})^2\right)\right] = \frac{1}{N^2}\sum_{i=1}^N \sigma^2 = \frac{\sigma^2}{N},\end{aligned}$$

where we have used $\overline{m} = \frac{1}{N}\sum_{i=1}^N \overline{m}$ in the first equality, $\mathbb{E}[(m_i - \overline{m})(m_j - \overline{m})] = 0$ if $i \neq j$ in the second equality since m_i, $i = 1, \ldots, N$ are iid random variables, and the definition of variance in the third equality. So in fact $C = 1$.

In practice, we rarely work with m directly but indirectly via some mapping $g : S \to T$. We have that $g(m_i)$, $i = 1,\ldots,N$ are iid[3] if m_i, $i = 1,\ldots,N$ are iid.

Exercise 9.3. Suppose the density of m is $\pi(m)$ and $z = g(m)$ is differentially invertible, i.e. $m = g^{-1}(z)$ exists and differentiable, what is the density of $g(m)$? •

Perhaps, one of the most popular and practical problems is to evaluate the mean of g, i.e.,

$$I \overset{\text{def}}{=} \mathbb{E}[G(m)] = \int_S g(m)\,\pi(m)\,dm, \tag{9.5}$$

which is an integral in $\mathbb{R}^n$.

Exercise 9.4. Define $z = g(m) \in T$, the definition of the mean in (4.4) gives

$$\mathbb{E}[G(m)] \equiv \mathbb{E}[Z] \overset{\text{def}}{=} \int_T z\pi_Z(z)\,dz.$$

Derive formula (9.5). •

Again, we emphasize that using any numerical integration methods that you know of for integral (9.5) is either infeasible or prohibitedly expensive when the dimension n is large, and hence not scalable. The LLN provides a reasonable answer if we can draw iid samples $\{g(m_1),\ldots,g(m_N)\}$ since we know that

$$\lim_{N\to\infty} \underbrace{\frac{1}{N}(g(m_1)+\ldots+g(m_N))}_{I_N} = I$$

Do you trivially see this?

with probability 1. Moreover, as shown above, the mean squared error is given by

$$\|I_N - I\|^2_{L^2(T,\mathbb{P})} = \mathbb{E}\left[(I_N - I)^2\right] = \frac{\mathbb{V}ar[G(m)]}{N}.$$

Again, the error decreases to zero like $1/\sqrt{N}$ "independent" of the dimension of T, but we need to be careful with such a statement unless $\mathbb{V}ar[G(m)]$ DOES NOT depend on the dimension.

A particular function g of interest is the following

$$g(m) = (m - \overline{m})(m - \overline{m})^T,$$

whose expectation is precisely the covariance matrix

$$\Gamma = cov(m) = \mathbb{E}\left[(m - \overline{m})(m - \overline{m})^T\right] = \mathbb{E}[G].$$

The average I_N in this case is known as the sample (aka empirical) covariance matrix. Denote

[3] We avoid technicalities here, but g needs to be a Borel function for the statement to be true.

$$\hat{\Gamma} = \frac{1}{N}\sum_{i=1}^{N}(m_i - \overline{m})(m_i - \overline{m})^T$$

as the sample covariance matrix. Clearly, $\hat{\Gamma}$ converges almost surely to Γ by LLN. Note that $\overline{m}$ is typically not available in practice, and we have to resort to a computable approximation

$$\hat{\Gamma} = \frac{1}{N}\sum_{i=1}^{N}(m_i - \hat{m})(m_i - \hat{m})^T, \tag{9.6}$$

with

$$\hat{m} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} m_i$$

denoting the sample mean. One can show that $\hat{\Gamma}$ converges to Γ but it needs more techinical argument on convergence in probability[4], and hence is omitted. The easier question to answer is whether $\hat{\Gamma}$ is an *unbiased estimator* of Γ, and the detail is in Exercise 9.5.

Exercise 9.5. It turns out that (9.6) is a biased estimator for Γ. We define $\tilde{m}$ an unbiased estimator for m if $\mathbb{E}[\tilde{m}] = m$. For example, $\hat{m}$ is an unbiased estimator for $\overline{m}$. Indeed

$$\mathbb{E}[\hat{m}] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[m_i] = \overline{m},$$

since, again, m_i are iid with mean $\overline{m}$.

1. Show that
$$\mathbb{E}[(\hat{m} - \overline{m})(\hat{m} - \overline{m})] = \frac{\Gamma}{N}.$$
2. Start from
$$J = \sum_{i=1}^{N}(m_i - \hat{m})(m_i - \hat{m})^T = \sum_{i=1}^{N}(m_i - \overline{m} + \overline{m} - \hat{m})(m_i - \overline{m} + \overline{m} - \hat{m})^T$$
to show that
$$\mathbb{E}[J] = (N-1)\Gamma,$$
and then conclude that
$$\frac{J}{N-1} = \frac{1}{N-1}\sum_{i=1}^{N}(m_i - \hat{m})(m_i - \hat{m})^T$$
is an unbiased estimator for Γ. So the “correct scaling” in $\hat{\Gamma}$ should be $N-1$ instead of N. For sufficient large N, the difference is however negligible.

•

[4] The key is the Slutsky’s theorem on convergence in probability of sequence of random variables.

9.2 Appendix

Definition 9.1 (Convergence in distribution). A sequence of random variables m_N, $N \in \mathbb{N}$, converges in distribution to m if μ_{m_N} converges weakly to μ_m, i.e.,

$$\int f \, d\mu_{m_N} \overset{N\to\infty}{\to} \int f \, d\mu_m$$

for any bounded and continuous function f.

Definition 9.2 (Almost surely convergence). A sequence of random variables m_N, $N \in \mathbb{N}$, converges almost surely to m if

$$\mathbb{P}\left[\{m_N - m \neq 0\}\right] = 0,$$

or equivalently

$$\mathbb{P}\left[\{m_N - m = 0\}\right] = 1.$$