Notes on statistics

Justin L. Ripley
ripley@illinois.edu

March 15, 2024

Contents

Abstract

I briefly review some of the basic notions of parametric Bayesian statistical inference that have come up in my research. These notes are not self contained and remain a work in progress! I assume some familiarity with probability theory and statistics, on the level of the first few chapters of [Was10]. I may sometimes implicitly assume some knowledge of differential geometry and partial differential equations. I try to cite sources whenever possible (whenever I can remember the source I learned something from), although the purpose of these notes are to serve more as a statistics "cheat sheet" than a formal review. If you think I am missing a reference please let me know. Please contact me if you find any errors!

The notation is: vectors/tensors are in **boldfont**. Indices are denoted with lower case latin letters, e.g. the i^{th} component of the vector $\boldsymbol{v}$ is $(\boldsymbol{v})_i = v_i$. We typically do not use boldfont when we explicitely write down indices. Repeated indices are summed over. Capital P always represents a probability distribution, $\boldsymbol{x}$ always represents an instantiation of measured data, $\boldsymbol{\theta}$ always represents model parameters. More generally, model parameters are represented by greek letters, while data is represented by latin letters. Random variables are always capatalized. Partial derivatives are denoted by ∂, and covariant derivatives by ∇_i (for our purposes, you can usually replace covariant derivatives with partial derivatives).

I thank Rohit Chandramouli for helpful conversations, and a lecture on model selection that inspired the creation of these notes, and Simone Mezzasoma for helpful comments that have led to a clearer presentation.

Chapter 1

Overview of parametric Bayesian statistics

1.1 Definitions

We use lower case latin latters to index vector/tensor components. We use a lower case latin letter in parenthesis to index a particular vector/tensor. We also bold font vectors. Repeated indices are summed (Einstein summation notation). We denote models with capital Latin letters, model parameters with lower case greek letters, and data with lower case latin letters. Notice that we use latin indices to index both model parameters and data with lower case latin indices, even though in general model parameters and data will live in different dimensional vector spaces. We will drop the instantiation index (the latin index in parenthesis) unless otherwise needed.

Here we focus on **parametric Bayesian statistics**. By parametric, we mean that we have explicit functional models for the probability distributions of parameters, and by Bayesian, we mean we mean that we are interested in the probability distribution of those parameters (and/or models), given the observed data.

Bayes theorem gives us

$$P\left(\boldsymbol{\theta}|\boldsymbol{x}, M\right) = \frac{P\left(\boldsymbol{x}|\boldsymbol{\theta}, M\right) P\left(\boldsymbol{\theta}, M\right)}{P\left(\boldsymbol{x}, M\right)}, \tag{1.1}$$

Here $P\left(\boldsymbol{x}|\boldsymbol{\theta}, M\right)$ is a statistical model M that reflects our beliefs about the data $\boldsymbol{x}$ given the values of the parameters $\boldsymbol{\theta}$ of a model M. The **posterior** $P\left(\boldsymbol{\theta}|\boldsymbol{x}, M\right)$ is a probability distribution for the model parameters $\boldsymbol{\theta}$ given $\boldsymbol{x}$. The **likelihood function** is $P\left(\boldsymbol{x}|\boldsymbol{\theta}, M\right)$, and is denoted by $\mathcal{L}\left(\boldsymbol{\theta}, M\right)$. The **prior distribution** $P\left(\boldsymbol{\theta}, M\right)$ quantifies our certainty of the model parameters $\boldsymbol{\theta}$ before we see the current data, and is often denote by $\pi\left(\boldsymbol{\theta}, M\right)$. The **evidence** [Ski06] (or marginal distribution of $\boldsymbol{x}$ [Was10]) $P\left(\boldsymbol{x}, M\right)$ essentially acts as a normalizing constant, as $P\left(\boldsymbol{\theta}|\boldsymbol{x}, M\right)$ must sum (integrate) to one. The evidence is often

denoted by $\mathcal{Z}(\boldsymbol{x}, M)$. If there are N independent observations of the data $\boldsymbol{x}$, the likelihood is

$$\mathcal{L}(\boldsymbol{\theta}, M) = \prod_{n=1}^{N} P\left(\boldsymbol{x}_{(n)}|\boldsymbol{\theta}, M\right). \tag{1.2}$$

We can write the evidence as the integral (or sum) over the model parameter values

$$\mathcal{Z}(\boldsymbol{x}, M) = \int d\theta \mathcal{L}(\boldsymbol{\theta}, M)\, \pi(\boldsymbol{\theta}). \tag{1.3}$$

Much of applied Bayesian statistics centers around finding efficients ways to evaluate the likelihood and evidence, given an assumed model $P(\mathbf{x}|\boldsymbol{\theta}, M)$ and prior $P(\boldsymbol{\theta}, M)$.

1.2 Parameter estimation

Assume you have one fixed model M. You can find the distribution of the parameters for the model, given a set of observed data, using Bayes theorem. Rewriting (6.1), we have

$$P(\boldsymbol{\theta}|\boldsymbol{x}, M) = \frac{\mathcal{L}(\boldsymbol{\theta}, M)\, \pi(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{x}, M)}. \tag{1.4}$$

The posterior probability distribution $P(\boldsymbol{\theta}|\boldsymbol{x}, M)$ for most problems is complicated and cannot be written in closed form. Determing the posterior can usually only be accomplished numerically. Additionally it can be computationally expensize to compute the posterior distribution, especially if there are many parameters in the model ($\boldsymbol{\theta}$ has many components).

This being said, it is straightforward to compute the relative probability of two different values of parameters $\boldsymbol{\theta}_{(n)}$ and $\boldsymbol{\theta}_{(m)}$. We have

$$\frac{P\left(\boldsymbol{\theta}_{(n)}|\boldsymbol{x}, M\right)}{P\left(\boldsymbol{\theta}_{(m)}|\boldsymbol{x}, M\right)} = \frac{\mathcal{L}\left(\boldsymbol{\theta}_{(n)}, M\right)}{\mathcal{L}\left(\boldsymbol{\theta}_{(m)}, M\right)} \frac{\pi\left(\boldsymbol{\theta}_{(n)}, M\right)}{\pi\left(\boldsymbol{\theta}_{(m)}, M\right)}. \tag{1.5}$$

We can write this in terms of the **likelihood ratio**

$$\lambda\left(\boldsymbol{\theta}_{(n)}, \boldsymbol{\theta}_{(m)}\right) \equiv \frac{\mathcal{L}\left(\boldsymbol{\theta}_{(n)}, M\right)}{\mathcal{L}\left(\boldsymbol{\theta}_{(m)}, M\right)}, \tag{1.6}$$

and the **prior odds**

$$R\left(\boldsymbol{\theta}_{(n)}, \boldsymbol{\theta}_{(m)}\right) \equiv \frac{\pi\left(\boldsymbol{\theta}_{(n)}, M\right)}{\pi\left(\boldsymbol{\theta}_{(m)}, M\right)}. \tag{1.7}$$

We discuss computational methods later, but we note that the value of $\boldsymbol{\theta}$ that maximizes $\mathcal{L}(\boldsymbol{\theta})$ is the **maximum likelihood estimator (MLE)**, and the value of $\boldsymbol{\theta}$ that maximizes $\mathcal{L}(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$ is the **maximum a posteriori probability estimator (MAP)**. Note that the MLE and MAP do not gives us any knowledge of the variance of those parameters–that requires knowledge of the full posterior probability distribution.

1.3 Model selection/Hypothesis testing

Second, you could have a collection of models $M_{(1)}, ..., M_{(N)}$. Given a set of observations, you may be interested in the relative ability of each model to explain the data. Using Bayes' theorem, we have

$$P\left(M_{(n)}|\boldsymbol{x}\right) = \frac{P\left(\boldsymbol{x}|M_{(n)}\right) P\left(M_{(n)}\right)}{P\left(\boldsymbol{x}\right)}. \tag{1.8}$$

This means that

$$\frac{P\left(M_{(n)}|\boldsymbol{x}\right)}{P\left(M_{(m)}|\boldsymbol{x}\right)} = \frac{P\left(\boldsymbol{x}|M_{(n)}\right)}{P\left(\boldsymbol{x}|M_{(m)}\right)} \frac{P\left(M_{(n)}\right)}{P\left(M_{(m)}\right)}. \tag{1.9}$$

Notice that we have essentially marginalized over the parameters of the models. That is, we have

$$P\left(\boldsymbol{x}|M_{(n)}\right) = \int d\theta P\left(\boldsymbol{x}|\boldsymbol{\theta}, M_{(n)}\right) P\left(\boldsymbol{\theta}\right) = \mathcal{Z}\left(\boldsymbol{x}, M_{(n)}\right). \tag{1.10}$$

We see that the odds ratio for two models is given by the ratio of the evidence for each model multiplied bythe prior odds for each model.

$$\frac{P\left(M_{(n)}|\boldsymbol{x}\right)}{P\left(M_{(m)}|\boldsymbol{x}\right)} = \frac{\mathcal{Z}\left(\boldsymbol{x}, M_{(n)}\right)}{\mathcal{Z}\left(\boldsymbol{x}, M_{(m)}\right)} \frac{P\left(M_{(n)}\right)}{P\left(M_{(m)}\right)}. \tag{1.11}$$

The odds ratio of the evidence is called the **Bayes factor**

$$B\left(M_{(n)}, M_{(m)}\right) \equiv \frac{\mathcal{Z}\left(\boldsymbol{x}, M_{(n)}\right)}{\mathcal{Z}\left(\boldsymbol{x}, M_{(m)}\right)}. \tag{1.12}$$

1.4 Choice of prior

Many people hold strong opinions about what a "good" choice of prior distribution for parameters should be, that often depends on the model in question and the field one is working in. Here we just review some of the terminology used in discussions on picking priors. Ultimately, there are at least as many (more serious) assumptions wrapped up in choosing a model to fit in parametric Bayesian statistics as there are in choosing a prior, so we just review the terminology used in picking priors here.

If the posterior probability distribution lies within the same probability distribution family (for a review of some different families, see Appendix B) as does the prior, then the posterior and prior are said to be **conjugate**, and the prior is a **conjugate prior**. Clearly the likelihood–that is the choice of model we are trying to fit for–plays a deciding role in determining if the prior is conjugate to the posterior. The notion of conjugate priors is mostly useful for analytic calculations, if we want to have a closed-form expression for the posterior.

So-called **uninformative priors** are meant to be used when you do not know much about the values the parameters you are trying to model. Ultimately though For example, the uniform distribution is often used as an uninformative prior, although even with a uniform you must choose bounds for it in order for the distribution to be normalizable (see Appendix B). Moreover, most distributions are not invariant under coordinate transformations. (see Appendix A). Consider an injective change of variables $\boldsymbol{\psi}(\boldsymbol{\theta})$. The PDF for $\boldsymbol{\psi}$ transforms as (see (A.14))

$$P(\boldsymbol{\psi}) = \frac{1}{|\det(J_{ij})|} P(\boldsymbol{\theta}), \tag{1.13}$$

where $J_{ij} = \partial\psi_i/\partial\theta_j$ is the Jacobian matrix. The **Jeffrey's prior** is a distribution that is proportional to the determinant Fisher information matrix

$$P(\boldsymbol{\theta}) \propto \sqrt{\det(F_{ij}(\boldsymbol{\theta}))}. \tag{1.14}$$

(see Appendix A), and does not change under the change of variables formula. This essentially follows from the fact that F_{ij} is a tensor, so that

$$\begin{aligned} \sqrt{\det(F_{ij}(\boldsymbol{\psi}))} &= \sqrt{\det(J_{ik}J_{kl}F_{kl}(\boldsymbol{\theta}))} \\ &= |\det(J_{ij})| \sqrt{\det(F_{ij}(\boldsymbol{\theta}))}. \end{aligned} \tag{1.15}$$

We see that the determinants of the Jacobian cancel each other out. While elegant, it is much more common (at least in the physics/astronomy literature) to nevertheless see the uniform prior being used when the authors profess ignorance about the expected value of the parameter in question.

In the limit of a large amount of data, so long as the prior does not *exclude* the best fit parameters to the model, different choices of prior should not dramatically affect the final estimated values for the paramters (the Bernstein-von Mises theorem is one concrete special case of this statement, see Appendix A).

Chapter 2

Model comparison

2.1 Bayes factor

To briefly review, in **parameter estimation**, one fintds the best fit parameters from the data given a model $h(\theta)$. What "best fit" means depends on the test statistic being used. Here we are concerend **Model selection**, which concerns finding which model better fits the data. In order to find the better fitting model, we compute the **Bayes factor**, which is the ratio of the evidence for each model

$$B_{2,1} \equiv \frac{P(d|H_2)}{P(d|H_1)}. \tag{2.1}$$

As a basic rule of thumb, if $B_{2,1} \sim 1$, then neither hypothesis is preferred compared to the other. If $B_{2,1} \ll 1$, then model 1 is preferred, while if $B_{2,1} \gg 1$, then model 2 is preferred. There are several subtleties to this interpretation, which we discuss more below.

If a model H has parameters $\boldsymbol{\theta}$, we can compute the likelihood by marginalizing over the models parameters for the likelihood (c.f. (1.3))

$$P(d|H) = \int d\theta P(d|\boldsymbol{\theta}, H) P(\boldsymbol{\theta}, H). \tag{2.2}$$

Doing this integral is typically challenging, since the dimension of the parameter space is very large, and the likelihood $P(d|\boldsymbol{\theta}, H)$ can be complicated (its functional form can only be guessed at in general). There are various approximations for how to compute this integral (analytically and numerically).

2.2 Nested models and the Savage-Dickey ratio

We consider a method to compute the Bayes factor for nested models. Consider a model M_1 which is nested in a model M_2. The model M_2 has one more parameter than M_1 (generalizing to more parameters is straightforward). We call the extra parameter λ. We call the rest of the parameters $\boldsymbol{\theta}$ nuisance parameters, as they do not distinguish the two models. In this setup we have that

$$P\left(d|\boldsymbol{\theta}, M_1\right) = P\left(d|\boldsymbol{\theta}, \lambda = \lambda_0, M_2\right), \tag{2.3}$$

where λ_0 is a constant.

The evidence of M_1 is

$$\begin{aligned} P\left(d|M_1\right) =& P\left(d|\lambda = \lambda_0, M_2\right) \\ =& \frac{P\left(\lambda = \lambda_0|d, M_2\right) P\left(d|M_2\right)}{P\left(\lambda = \lambda_0|M_2\right)}. \end{aligned} \tag{2.4}$$

We then see that the Bayes factor is

$$\begin{aligned} B_{2,1} =& \frac{P\left(d|H_2\right)}{P\left(d|H_1\right)} \\ =& \frac{P\left(\lambda = \lambda_0|M_2\right)}{P\left(\lambda = \lambda_0|d, M_2\right)}. \end{aligned} \tag{2.5}$$

This is the **Savage-Dickey ratio** [DL70]. We can also write this as

$$B_{2,1} = \left(\frac{\text{prior}}{\text{posterior}}\right)_{\lambda=\lambda_0}. \tag{2.6}$$

The advantage of this method is that you only need to compute the evidence of the model M_2, instead of computing the evidence of both M_2 and the nested model M_1. Also, you do not need to divide two noisy numbers (the evidence of model 1 and model 2), you only need to divide a known number (the prior) by one noisy number (the evidence of model 2).

2.3 Occam factor

For more discussion see for examle [Mac03]. We consider another measure of the power of a model to explain a given data set. The **Occam factor** is defined to be

$$O \equiv \frac{\text{posterior volume}}{\text{prior volume}} \sim \frac{\sigma_{\boldsymbol{\theta}|d}}{\sigma_{\boldsymbol{\theta}}}. \tag{2.7}$$

By volume, we mean the integral over parameter space of the probability distribution. Here $\sigma_{\boldsymbol{\theta}}$ is some measure of the variance of the prior probability distribution, and $\sigma_{\boldsymbol{\theta}|d}$ is variance

of the posterior probability distribution. The Occam factor measures how much the data shrinks the probability distribution as compared to its prior distribution. If the Occam factor is ~ 1, the data doesn't constrain the model well, since the variance parameters of the model do not shrink. We can interpret this as saying that the model does not explain the observed data well either. We can write

$$\text{evidence} \sim \text{max likelihood} \times \text{occam factor}. \tag{2.8}$$

From this, we see that the Occam factor accounts for the fact that models with more parameters can fit data better, and should be penalized for having more parameters.

For example, consider two hypothesis: H_1 and H_2. Say they are nested: $H_2(\boldsymbol{\theta}) \sim H_1(\boldsymbol{\theta}, \lambda)$. If λ is unconstrained, $O \sim 1$, and if λ is well-constrained, $O \ll 1$.

Chapter 3

Forecasting

When it is hard or expensive to collect data, it can be useful to predict (or forecast) how well parameters of a given model could be measured with simulated data. Forecasting can inform whether a a more in-depth analysis of a model on real data is worth doing–that is whether or not real data could place any meaningful measurement of the parameters of a model. Here we review several semi-analytic methods for forecasting.

3.1 Injection analysis

We consider a model $P(\boldsymbol{x}|\boldsymbol{\theta})$ with prior $P(\boldsymbol{\theta})$. For whatever reason, we do not have any data $\boldsymbol{x}$. For example, it may be expensive to collect data, so we do not want to collect it until we have some confidence that we could meaningfully measure parameters in the model $P(\boldsymbol{x}|\boldsymbol{\theta})$. An injection analysis involves determining the distribution of $P(\boldsymbol{\theta}_i|\boldsymbol{x}_0)$ where $\boldsymbol{x}_0$ is fake (generated) data set that we hope represents a characteristic realization of the data we may measure. In other words, we have "injected" the data $\boldsymbol{x}_0$ into our model. If we can meaningfully measure/determine the parameters $\boldsymbol{\theta}$ given $\boldsymbol{x}_0$, it may be worth collecting real data/observations.

A reasonable choice for $\boldsymbol{x}_0$ is to choose $\boldsymbol{x}_0$ to maximize $P(\boldsymbol{x}|\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ are the values of parameters that you expect to hope to measure. In equations we choose $\boldsymbol{x}_0$ to satisfy

$$\forall \boldsymbol{x}\ P(\boldsymbol{x}|\boldsymbol{\theta}_0) \leq P(\boldsymbol{x}_0|\boldsymbol{\theta}_0) \tag{3.1}$$

Sometimes it is worth adding a realization of noise, $\boldsymbol{n}$, to $\boldsymbol{x}_0$; we call this

$$\boldsymbol{x}_{0,n} = \boldsymbol{x}_0 + \boldsymbol{n} \tag{3.2}$$

For example, the components of $\boldsymbol{n}$ may be drawn from a Gaussian with zero mean and unit diagonal covariance matrix (although the choice of $\boldsymbol{n}$ will depend on your understanding nature of the experiment/observation). It is common to $\boldsymbol{n} = 0$, which can be considered the

"best" possible situation for recovering parameters. We then inject $\boldsymbol{x}_{0,n}$ into the likelihood, and sample on $\boldsymbol{\theta}$, that is we consider

$$P\left(\boldsymbol{\theta}|\boldsymbol{x}_{0,n}\right) = \frac{P\left(\boldsymbol{x}_{0,n}|\boldsymbol{\theta}\right) P\left(\boldsymbol{\theta}\right)}{P\left(\boldsymbol{x}_{0,n}\right)}. \tag{3.3}$$

The probability distribution $P\left(\boldsymbol{\theta}|\boldsymbol{x}_{0,n}\right)$ gives us an understanding of how well we could measure $\boldsymbol{\theta}_0$, given (near, if $\boldsymbol{n} \neq 0$) optimal data. Moreover, it can give us an idea of how the different components of $\boldsymbol{\theta}$ may be correlated with one another.

To determine $P\left(\boldsymbol{\theta}|\boldsymbol{x}\right)$, we either need to directly sample $\boldsymbol{\theta}$ from (3.3), or use an approximation method.

3.2 Fisher forecasting

Injection analysis with a multivariate normal approximation for the likelihood is called **Fisher forecasting**. Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimator for $\boldsymbol{\theta}$. Under appropriate regularity conditions, in the limit of a large number of observations, $P\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}\right)$ tends towards the normal distribution (in parameter space), with mean $\hat{\boldsymbol{\theta}}$ and covariance matrix

$$\Sigma_{ij}^{(F)} = \frac{1}{N} F_{ij}^{-1}\left(\hat{\boldsymbol{\theta}}\right). \tag{3.4}$$

Here N is the number of observations. This is known as the **Bernstein–von Mises theorem**. We provide a proof of this result in Appendix A.

Estimating the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is a challenging task of its own, that we do no explore further here. Using the Bernstein-von Mises theorem, the posterior probability distribution near $\hat{\boldsymbol{\theta}}$, in the limit of a large number of observations, is approximately

$$P\left(\boldsymbol{\theta}|\boldsymbol{x}\right) = \frac{\pi\left(\boldsymbol{\theta}\right)}{\mathcal{Z}\left(\boldsymbol{x}, M\right)} \exp\left[-\frac{1}{2}\left(\theta - \hat{\theta}\right)^i F_{ij}\left(\theta - \hat{\theta}\right)^j\right]. \tag{3.5}$$

If we assume a Gaussian prior on the parameters $\boldsymbol{\theta}$, then the posterior is a multivariate Gaussian with an inverse covariance matrix given by

$$\Sigma_{ij}^{-1} = N F_{ij} + \frac{1}{\sigma_i^2}\delta_{ij}. \tag{3.6}$$

That is, the posterior probability distribution within this approximation is

$$P\left(\boldsymbol{\theta}|\boldsymbol{x}\right) \approx \frac{1}{\sqrt{\left(2\pi\right)^k \det \Sigma}} \exp\left[-\frac{1}{2}\left(\theta - \hat{\theta}\right)^i \Sigma_{ij}^{-1}\left(\theta - \hat{\theta}\right)^j\right], \tag{3.7}$$

where k is the dimensionality of $\boldsymbol{\theta}$, that is the number of parameters.

To perform a **Fisher forecast** for a given model $P(\boldsymbol{x}|\boldsymbol{\theta})$, we pick a set of parameters $\boldsymbol{\theta}_0$, and then compute the Fisher information matrix (A.29). That is, we assume that $\boldsymbol{\theta}_0$ are the "true" model parameters, and also are the maximum likelihood estimators. We then "inject" those parameters into the likelihood, which we approximate as a multivariate Gaussian with inverse covariance matrix given by (3.6). This analysis can be useful to determine the strength of correlation between different the different components of θ^i (through the off-diagonal terms in Σ_{ij}). The diagonal of the covariance matrix additionally gives us the 1-σ error bars of the parameters. If we could make N measurements of the same data, each element in the covariance matrix would decrease $1/N$, as follows from (3.4). We see that the Fisher matrix can also give us a rough estimate of the number of observations N that are needed to make a $n-\sigma$ observation of a parameter θ^i.

Fisher forecasting is sometimes said to provide an optimal estimate of the variance of the parameters in a given measurement. This statement is justified by the **Cramér-Rao bound**, which states that the covariance matrix of an unbiased estimator for $\boldsymbol{\theta}$, $\boldsymbol{\Theta}$, (that is $\mathbb{E}[\boldsymbol{\Theta}(\boldsymbol{x})] = \boldsymbol{\theta}$) is bounded from below by the inverse of the Fisher information matrix

$$\Sigma_{ij}\Big|_{\boldsymbol{\theta}=\boldsymbol{\mu}_{\boldsymbol{\theta}}} \geq F_{ij}^{-1}(\boldsymbol{\theta})\,. \tag{3.8}$$

This bound should be interpreted with caution though, as (A.42) only holds for unbiased estimators to the parameters $\boldsymbol{\theta}$. Consider a general estimator $\boldsymbol{\Theta}(\boldsymbol{x})$, and denote its expectation by

$$\mathbb{E}[\boldsymbol{\Theta}(\boldsymbol{x})] = \boldsymbol{\psi}(\boldsymbol{\theta})\,. \tag{3.9}$$

The Cramér-Rao bound states that

$$\nabla_{\theta_m}\psi_i \nabla_{\theta_n}\psi_j F_{mn}^{-1}(\boldsymbol{\theta})\,. \tag{3.10}$$

If $\boldsymbol{\Theta}$ is an unbiased estimator ($\boldsymbol{\psi} = \boldsymbol{\theta}$), then (A.41) reduces to (A.42). We outline a proof of (A.41) in Appendix A.

Chapter 4

Times series analysis

We consider the problem of determining the signal from a data timestream. Calling the data $x(t)$, we then want to find a signal $s(t)$ given noise $n(t)$, where

$$x\left(t\right) = s\left(t\right) + n\left(t\right). \tag{4.1}$$

We model the noise $n(t)$ can be modeled as a **stochastic process** (which implies that $x(t)$ is a stochastic process) We assume x, s, n are all real functions. Our main goal is to derive the likelihood function for a time series of the form (4.1) when n takes the form of colored stationary noise, and to derive the matched filtering theorem.

4.1 Basic definitions

A function $n(t)$ is a **stochastic process**, if $n(t)$ is a random variable at each time t that is described by some probability distribution. This probability distribution may depend on t, and the previous history of values of x, for example. If t is a discrete variable, and the probability distribution for $n(t_i)$ depends on $n(t_{i-1})$, then $n(t_i)$ is a **Markov chain**. If the probability distribution for $n(t)$ is independent of t, then $n(t)$ is a **stationary process**.

4.2 Correlation and covariance

We denote the mean and variance of a time series $x(t)$ with μ_x and σ_x, respectively. We define the **covariance** between two stochastic processes x_1 and x_2 at times t_1 and t_2 to be

$$C_{x_1,x_2}\left(t_1,t_2\right) \equiv \mathbb{E}\left[\left(x_1\left(t_1\right) - \mu_{x_1(t_1)}\right)\left(x_2^*\left(t_2\right) - \mu^*_{x_2(t_2)}\right)\right]. \tag{4.2}$$

We define the **autocovariance** for a stocahstic process to be

$$K_x\left(t_1,t_2\right) \equiv C_{xx}\left(t_1,t_2\right). \tag{4.3}$$

We define the **autocorrelation** for a stochastic process to be

$$R_x(t_1, t_2) \equiv \mathbb{E}[x(t_1) x^*(t_2)] = K_x + \mu_x^2. \tag{4.4}$$

We define the **energy** of a time series s to be

$$E_x \equiv \int_{-\infty}^{\infty} dt\, |x(t)|^2 = \int_{-\infty}^{\infty} df\, |\tilde{x}(f)|^2 . \tag{4.5}$$

The last expression follows from Parseval's theorem. We define the **energy spectral density** to be

$$\hat{S}_x(f) \equiv |\tilde{x}(f)|^2 . \tag{4.6}$$

4.3 Stationary and weak-sense stationary stochastic processes

A stochastic process is said to be **stationary** if its joint probability distribution does not change under time shifts. A stochastic process is said to be **weak-sense stationary (WSS)** (or **wide-sense stationary**), then its first moment is independent of time, and if its autocorrelation function depends only on $\tau = t_1 - t_2$. For a WSS process we can write

$$R_x(\tau) = \mathbb{E}[x(t+\tau) x^*(t)], \tag{4.7}$$

where t is arbitrary. For a WSS/stationary process, we can write the expectation of x at a given instant as the average of x over all time

$$\mathbb{E}[x] = \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} dt x_T(t), \tag{4.8}$$

and similarly for functions of $x(t)$. Here $x_T(t)$ is defined to be

$$x_T(t) \equiv w_T(t)\, x(t), \tag{4.9}$$

$$w_T(t) \equiv \begin{cases} 1 & |t| < T/2 \\ 0 & \text{otherwise} \end{cases}. \tag{4.10}$$

In other words, we can write

$$R_x(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} dt x_T(t+\tau)\, x_T^*(t). \tag{4.11}$$

We emphasize that for WSS processes we can replace ensemble averages with time averages. This is extremely useful in practice, as we can then determine the statistical properties of WSS by taking repeated time measurements of observables of the time series. While it is

common to assume a given time stream is WSS, most real world data is at best approximately WSS, typically over a short time scale.

Stationary stochastic processes have support over the entire real line, so the energy integrals defined above typically diverge. For these processes, instead one looks at the **power**

$$\begin{aligned} P_x &\equiv \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} dt \left| x_T(t) \right|^2 \\ &= \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} df \left| \tilde{x}_T(f) \right|^2 \\ &= \int_{-\infty}^{\infty} df S_x(f) . \end{aligned} \tag{4.12}$$

One the last line we have used the **power spectral density**, which is defined to be

$$S_x(f) \equiv \lim_{T\to\infty} \frac{1}{T} \left| \tilde{x}_T(f) \right|^2 . \tag{4.13}$$

If $x(t)$ is real, then $\tilde{x}_T(-f) = \tilde{x}_T^*(f)$, and it is common to define the power spectral density to be

$$S_x(f) \equiv \lim_{T\to\infty} \frac{2}{T} \left| \tilde{x}_T(f) \right|^2 , \tag{4.14}$$

and to write

$$P_x = \int_0^{\infty} df S_x(f) . \tag{4.15}$$

Finally, we consider the expectation value of the Fourier transform of a stationary signal

$$\begin{aligned} \mathbb{E}\left[\tilde{x}(f') \tilde{x}^*(f) \right] &= \int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} dt' e^{-2\pi i (ft - f't')} \mathbb{E}\left[x(t) x(t') \right] \\ &= \int_{-\infty}^{\infty} dt e^{-2\pi i (f-f')t} \int_{-\infty}^{\infty} d\tau e^{-2\pi i f'\tau} \mathbb{E}\left[x(t+\tau) x(t) \right] \\ &= \int_{-\infty}^{\infty} dt e^{-2\pi i (f-f')t} \tilde{R}_x(f) \\ &= \delta(f - f') S_x(f) . \end{aligned} \tag{4.16}$$

On the third line we used that $x(t)$ was stationary.

4.4 Wiener-Khinchin theorem

The power spectral density and the Fourier transform of the autocorrelation are equal for WSS processes. This is known as the **Wiener–Khinchin theorem**. To prove this, we set $\tau \equiv t_1 - t_2$. We then write

$$\tilde{R}_x(f) = \int_{-\infty}^{\infty} d\tau e^{-2\pi i f \tau} R_x(\tau)$$

$$
\begin{aligned}
&= \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} d\tau e^{-2\pi i f\tau} \int_{-\infty}^{\infty} dt x_T\left(t+\tau\right) x_T^*\left(t\right) \\
&= \int_{-\infty}^{\infty} df' \int_{-\infty}^{\infty} d\tau e^{2\pi i (f'-f)\tau} \lim_{T\to\infty} \frac{1}{T} \left|\tilde{x}_T\left(f'\right)\right|^2 \\
&= S_x\left(f\right). \qquad (4.17)
\end{aligned}
$$

On the second line we used the formula for the autocorrelation function for WSS processes (4.11). On the third line we used the convolution theorem for Fourier transforms. On the last line we used that $\int d\tau e^{i\tau f} = \delta(f)$, and used the definition of $S_x(f)$ (4.13).

4.5 Gaussian white noise

As a special case of a stationary stochastic process, we first consider **Gaussian white noise**. By **Gaussian**, we mean that the probability distribution for $x\left(t\right)$ for each t_i is a Gaussian

$$x\left(t_i\right) \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right). \qquad (4.18)$$

By **white**, we mean that the $x\left(t_i\right)$ are uncorrelated, and that the means $\mu_i = 0$. The autocorrelation function then is

$$R_x\left(t_i, t_j\right) = \sigma^2 \delta_{ij}. \qquad (4.19)$$

Notice that σ^2 does not depend on t_i. We see that white noise is stationary. The autocorrelation function can be written in terms of $\tau \equiv t_i - t_j$ as

$$R_x\left(\tau\right) = \sigma^2 \delta\left(\tau\right). \qquad (4.20)$$

By the Wiener-Kinchin theorem, we can compute the power spectral density from the Fourier transform of the autocorrelation function

$$
\begin{aligned}
S_x\left(f\right) &= \int_{-\infty}^{\infty} dt e^{2\pi i f t} R_x\left(t\right) \\
&= \sigma^2. \qquad (4.21)
\end{aligned}
$$

We see for Gaussian white noise, the power spectral density is a constant–there is constant power across all frequencies. For real functions, the integral over frequencies goes from $[0, \infty)$, and we define

$$S_x\left(f\right) = 2\sigma^2. \qquad (4.22)$$

4.6 Likelihood function for a series of measurements with colored stationary noise

Our treatment roughly follows [CA11] (see also [Fin92]). We consider a series of a continuous time stream of observations $y\left(t\right)$. We assume that $y\left(t\right)$ can be related to a convolution of a

time stream drawn from Gaussian white noise

$$y(t) = \int_{-\infty}^{\infty} dt' \gamma(t - t') x(t'). \tag{4.23}$$

Here γ is the kernel and x is a time stream drawn from a Gaussian distribution with constant. We assume y, x, γ are all real functions. The Fourier transform gives us

$$\tilde{y}(f) = \tilde{\gamma}(f) \tilde{x}(f). \tag{4.24}$$

The power spectral density of y then is

$$S_y(f) = |\tilde{\gamma}(f)|^2 S_x(f) = |\tilde{\gamma}(f)|^2 \sigma. \tag{4.25}$$

The main point of adding the convolution is that we can consider processes with **colored noise**, that noise where the power spectral density can vary with frequency. We can do this by choosing some σ, and then choosing a γ such that $|\tilde{\gamma}(f)|^2$ gives us the spectral density we desire.

We note that $y(t)$ describes a WSS process, as

$$\begin{aligned}
\mathbb{E}[y(t_1) y(t_2)] =& \int_{-\infty}^{\infty} dt_1' \int_{-\infty}^{\infty} dt_2' \gamma(t_1 - t_1') \gamma(t_2 - t_2') \mathbb{E}[x(t_1') x(t_2')] \\
=& \sigma^2 \int_{-\infty}^{\infty} dt' \gamma(t_1 - t') \gamma(t_2 - t') \\
=& \sigma^2 \int_{-\infty}^{\infty} dt \gamma(\tau - t) \gamma(t).
\end{aligned} \tag{4.26}$$

On the second line we used (4.20), while on the third we set $t = t_2 - t'$.

We consider a discretized set of N points (evenly spaced) from $x(t)$ and $y(t)$. We write the vectors $\mathbf{x}$ and $\mathbf{y}$ where the componets are, e.g. $\mathbf{x}_i \equiv x(t_i)$. We define the discretized matrix $\Gamma_{ij} \equiv \gamma(t_i - t_j)$. We then have

$$y_i = \Gamma_{ij} x_j \tag{4.27}$$

We define the matrix

$$\Sigma_{ij} \equiv \frac{\sigma^2}{\Delta t} \delta_{ij}, \tag{4.28}$$

where $\Delta t \equiv T/(N-1)$, and N is the number of discretized points. We choose this scaling so that the autocorrelation of x_i approaches the correct behavior in the continuum limit, as we show below. The probability distribution for each $\mathbf{x}$ is

$$P_{\mathbf{x}}(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^N \frac{1}{\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2} x_i \Sigma_{ij}^{-1} x_j\right]. \tag{4.29}$$

The correlation for x is then

$$\begin{aligned}\mathbb{E}\left[x_i x_j\right] =& \int d^N x P_{\mathbf{X}}\left(\mathbf{x}\right) x_i x_j \\ =& \int d^N x \left(\frac{1}{\sqrt{2\pi}}\right)^N \frac{1}{\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2} x_i \Sigma^{-1}_{ij} x_j\right] x_i x_j \\ =& \sigma^2 \frac{\delta_{ij}}{\Delta t}.\end{aligned} \tag{4.30}$$

In the limit $\Delta t \to 0$, this approaches $\sigma^2 \delta\left(t_i - t_j\right)$,which is the autocorrelation function for Gaussian white noise; see (4.20).

We obtain the probability distribution for $\mathbf{y}$ under a linear transformation of variables (note the ordering of the indices)

$$P_{\mathbf{Y}}\left(\mathbf{y}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^N \frac{1}{\sqrt{\det \Sigma \det \Gamma}} \exp\left[-\frac{1}{2} y_i \Gamma^{-1}_{mi} \Sigma^{-1}_{mn} \Gamma^{-1}_{nj} y_j\right]. \tag{4.31}$$

This expression gives the probability the N draws. We now need to take the continuum limit. First we look at the argument of the exponential

$$\begin{aligned} y_i \Gamma^{-1} m_{mi} \Sigma^{-1}_{mn} \Gamma^{-1}_{nj} y_j =& \frac{1}{\sigma^2} x_i x_i \Delta t \\ \to& \frac{2}{S_x} \int_{t_s}^{t_f} dt \left|x\left(t\right)\right|^2 \\ \approx& \frac{2}{S_x} \int_{-\infty}^{\infty} dt \left|x\left(t\right)\right|^2 \\ =& \frac{4}{S_x} \int_0^{\infty} df \left|\tilde{x}\left(f\right)\right|^2 \\ =& 4 \int_{-\infty}^{\infty} df \frac{\left|\tilde{y}\left(f\right)\right|^2}{S_y}. \end{aligned} \tag{4.32}$$

Here t_s, t_f are the start and end times for the series $x(t)$. On the first line we used $x_i = \Gamma^{-1}_{ij} x_j$, and that $\Sigma^{-1}_{ij} = \sigma^{-1} \delta_{ij}$. On the second and third lines we converted the Riemann sum to an integral (we took the continuum limit). We approximated the start/end times with $\pm\infty$. On the last line we use $\tilde{x} = \tilde{y}/\tilde{\gamma}$, $S_y = \left|\tilde{\gamma}\left(f\right)\right|^2 S_x$, and that S_x is a constant, so we can pull it into the integral. Remember that we assume that x, y, γ are all real, so that for example $x\left(-f\right) = x^*\left(-f\right)$. Ignoring the constant normalization factor, we see that the probability density function (the likelihood function) for $y\left(t\right)$ is

$$P\left(y\left(t\right)\right) \propto \exp\left[-\frac{1}{2}\left(y, y\right)\right], \tag{4.33}$$

where we have defined the inner product

$$\left(a, b\right) \equiv 2 \int_0^{\infty} df \frac{a\left(f\right) b^*\left(f\right) + b\left(f\right) a^*\left(f\right)}{S_y\left(f\right)}. \tag{4.34}$$

Here $S_y(f)$ is the spectral noise density for the process, and a, b can represent the Fourier transform of particular draws. We interpret (4.33) as the likelihood function (up to normalization) for colored WSS noise. We call (4.34) a **matched filter**. We define the **signal to noise ratio (SNR)** for a signal s with noise n to be

$$\rho^2 \equiv (s,s) = 4\int_0^\infty df \frac{|s(f)|^2}{S_n(f)}. \tag{4.35}$$

4.7 Matched filter theorem

We next derive the optimal test statistic for extracting a singal from WSS colored noise. We consider a time series $x(t)$ that can be written as

$$x(t) = s(t) + n(t). \tag{4.36}$$

We assume that $n(t)$ can be written as a convolution with Gaussian white noise, as we described in Sec. (4.6). We assume we are searching for a signal $s(t)$ that we know how to compute. We these assumptions, we can compute the optimal test statistic to distinguish between the two following hypothesis:

Null hypothesis $\mathcal{H}_0$: $x(t) = n(t)$.

Alternative hypothesis $\mathcal{H}_1$: $x(t) = s_1(t) + n(t)$.

Here $s_1(t)$ is a signal we are guessing is in the data. We compare the two hypothesis by computing the likelihood ratio (likelihood for short for the rest of this section)

$$\Lambda(\mathcal{H}_1|x) \equiv \frac{P(x|\mathcal{H}_1)}{P(x|\mathcal{H}_0)}. \tag{4.37}$$

We use the likelihood function (4.33). We next show that $s_1 \propto s$ maximizes Λ, which is the **matched filtering theorem**.

If the null hypothesis is true, then the probability density function goes as

$$P(x|\mathcal{H}_0) \propto \exp\left[-\frac{1}{2}(x,x)\right]. \tag{4.38}$$

If the alternative hypothesis is true, then the probability density function goes as

$$P(x|\mathcal{H}_0) \propto \exp\left[-\frac{1}{2}(x-s_1, x-s_1)\right]. \tag{4.39}$$

We have used (4.34), with the noise power spectral density given by $S_n(f)$. The normalization factors cancel out in the likelihood ratio, and we are left with

$$\Lambda(\mathcal{H}_1|x) = \exp\left[-\frac{1}{2}(x-s_1, x-s_1) + \frac{1}{2}(x,x)\right]$$

$$= \exp\left[(x, s_1) - \frac{1}{2}(s_1, s_1)\right]. \tag{4.40}$$

The matched filtering theorem states that the likelihood ratio $\Lambda(\mathcal{H}_1|x)$ is maximized when $s_1 \propto s$. To show this, we first note that likelihood ratio is maximized when the log-likelihood ratio L is maximized. The log-likelihood is

$$L(s_1) \equiv (n, s_1) + (s, s_1) - \frac{1}{2}(s_1, s_1). \tag{4.41}$$

We only consider s_1 such that $(n, s_1) = 0$ (this also holds for the "true" signal s). Moreover, we fix $(s_1, s_1) = c_1$, where c_1 is a constant (otherwise the likelihood could be arbitrarily big or small by rescaling the amplitude of s_1) Maximizing the likelihood then reduces to maximizing

$$L(s_1) = (s, s_1) - \frac{1}{2}c_1. \tag{4.42}$$

By the Cauchy-Schwartz inequality, we have that

$$(s, s_1) \leq \sqrt{(s, s)}\sqrt{(s_1, s_1)}. \tag{4.43}$$

Equality only holds when $s_1 \propto s$. We conclude that choosing $s_1 \propto s$ maximizes the likelihood function (up to a proportionaly constant, which is fixed by the condition $(s_1, s_1) = c_1$).

The task of finding a signal in colored WSS noise then reduces to finding a filtering function s_1 that is orthogonal to the noise, and that maximizes the value of the matched filter (x, s_1). In practice, we can determine the noise profile of the detector by measuring the response of the detector in the (assumed) absence of any signal.

The matched filtering theorem is powerful, but it relies on several strong assumption that are only approximately met in practice. First, it assumes that we know what we are looking for–that is, that we have a **template bank** of templates $s_i(t)$ that we can convolve with the data. Even if we do have a template bank, it can be very computationally expensive to search for the s_i that fits the data best, especially if the parameter space for s_i is large. Efficiently evaluating the likelihood and searching through parameter space remains a topic of active research in, e.g. the gravitational wave astronomy community (for a review, see e.g. [CA11]). The matched filtering theorem also assumes the noise is stationary or WSS. Most kinds of detectors (say a phone line, or a gravitational wave detector) suffer from non-stationary noise, often called **glitches**. Provided those are well enough understood, they can be subtracted out of the signal, although in practice it can be difficult to completely remove glitches from a time stream.

Chapter 5

Markov Chains

First we define a few concepts about Markov chains. There is a large literature on Markov chains. Here we discuss only a few aspects of the theory. In particular we will mostly only discuss discrete Markov chains–that is Markov chains where that make discrete jumps in time.

We will borrow some of the concepts from the discussion on stochastic processes in Chptr. 4. A **Markov chain** is a sequence of random vectors (a stochastic process) $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ...\}$, where the probability distribution of $\boldsymbol{\theta}_{k+1}$ is dependent solely on $\boldsymbol{\theta}_n$. That is

$$P\left(\boldsymbol{\theta}_{k+1} = \boldsymbol{j} | \boldsymbol{\theta}_k = \boldsymbol{i}, \boldsymbol{\theta}_{k-1} = \boldsymbol{i}_{k-1}, \cdots \boldsymbol{\theta}_0 = \boldsymbol{i}_0\right) = P\left(\boldsymbol{\theta}_{k+1} = \boldsymbol{j} | \boldsymbol{\theta}_n = \boldsymbol{i}\right). \tag{5.1}$$

Here the $\boldsymbol{j}, \boldsymbol{i}_m$ label the states of the Markov chain–notice that the subscript of $\boldsymbol{\theta}$ denotes what step we're on in the Markov chain. We call $P\left(\boldsymbol{\theta}_{k+1} = \boldsymbol{j} | \boldsymbol{\theta}_n = \boldsymbol{i}\right)$ the **transition probability**. The probability distribution for $\boldsymbol{\theta}_0$, $P\left(\boldsymbol{\theta}_0\right)$, is called the **initial distribution**.

Notice by the conservation of probability, we have

$$\sum_{\boldsymbol{j} \in S} P\left(\boldsymbol{\theta}_{k+1} = \boldsymbol{j} | \boldsymbol{\theta}_n = \boldsymbol{i}\right) = 1, \tag{5.2}$$

where the sum over all states of the Markov chain. In other words, witha Markov chain, in the next step you must go *somewhere* with probability one–probability must be conserved.

If the transition probabilities are **stationary**, the transition probabilities do not depend on the step n in the chain. In that case we sometimes denote the transition probabilities with the simpler notation

$$P\left(\boldsymbol{\theta}_{k+1} = \boldsymbol{j} | \boldsymbol{\theta}_n = \boldsymbol{i}\right) \to P\left(\boldsymbol{j} | \boldsymbol{i}\right). \tag{5.3}$$

We will only consider stationary transition probabilities.

5.1 Representations of Markov chains

Some references set $P(\boldsymbol{j}|\boldsymbol{i}) = p_{\boldsymbol{ij}}$, others set $P(\boldsymbol{j}|\boldsymbol{i}) = p_{\boldsymbol{ji}}$. While I find the latter notation more natural, it seems like the most common notation is the first one (e.g. [Was10, PN14]), so we pick that one

$$P(\boldsymbol{j}|\boldsymbol{i}) \equiv p_{\boldsymbol{ij}}. \tag{5.4}$$

By (5.2), and (5.4), we see that the elements of each row must sum to one

$$\sum_{\boldsymbol{j}\in S} p_{\boldsymbol{ij}} = 1. \tag{5.5}$$

With Eq. (5.4), the PDF of a chain $\{\boldsymbol{\theta}_0 = \boldsymbol{i}_0, \boldsymbol{\theta}_1 = \boldsymbol{i}_i, ..., \boldsymbol{\theta}_k = \boldsymbol{i}_k\}$ is

$$P(\{\boldsymbol{\theta}_0 = \boldsymbol{i}_0, \boldsymbol{\theta}_1 = \boldsymbol{i}_1, ..., \boldsymbol{\theta}_n = \boldsymbol{i}_n\}) = p_{\boldsymbol{0}} p_{\boldsymbol{0,1}} p_{\boldsymbol{1,2}} \cdots p_{\boldsymbol{k-1,k}}, \tag{5.6}$$

When the states are scalars, we can set $p_{\boldsymbol{ij}} \to p_{ij}$ to represent the components of a matrix. Sometimes the transition probabilities are represented in terms of a **state transition matrix** P, where the rows/columns are given by

$$P_{ij} = p_{ij}. \tag{5.7}$$

Another common way to describe a Markov chain is through a **transition state diagram** (I find this to be especially useful for solving problems that involve Markov chains). One draws an arrow between points in the state space that have nonzero transition probabilities between them, and then labels each arrow with the transition probability.

5.2 Transition probabilities for scalar states

We next turn to a more vectorial representation of the evolution of probability distribution for scalar random variables. We consider an initial probability distribution of states, which we represent as a row vector

$$\boldsymbol{\pi}_0 = \left(P(\theta_0 = 1), \cdots, P(\theta_0 = n).\right). \tag{5.8}$$

The probability distribution of of states for the next step in the Markov chain is

$$P(\theta_1 = j) = \sum_{i\in S} P(j|i)\, P(\theta_0 = i). \tag{5.9}$$

In vectorial notation, we have

$$(\boldsymbol{\pi}_1)_i = (\boldsymbol{\pi}_0)_j\, P_{ji}. \tag{5.10}$$

Iterating, we see that the probability distribution at the k^{th} step is

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_0 P^k. \tag{5.11}$$

Finally, we note that from the law of total probability, we have

$$P\left(\theta_{m+n} = j | \theta_0 = i\right) = \sum_{k \in S} P\left(\theta_m = j | \theta_n = k\right) P\left(\theta_n = k | \theta_0 = i\right). \tag{5.12}$$

We can rewrite these as follows

$$p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^{(n)} p_{kj}^{(m)}. \tag{5.13}$$

These are called the **Chapmon-Kolmogorov equations**. Here we have used the notation

$$p_{ij}^{(n)} \equiv P\left(\theta_n = j | \theta_0 = i\right) = p_{ik_1} p_{k_2 k_3} \cdots p_{k_n j}. \tag{5.14}$$

5.3 Absorbing states

For a given state $\boldsymbol{i}$, if there is an integer n such that

$$P\left(\boldsymbol{X}_n = \boldsymbol{i} | \boldsymbol{X}_0 = \boldsymbol{i}\right) = 1, \tag{5.15}$$

we say the state $\boldsymbol{i}$ is **recurrent**. If instead for all n we have

$$P\left(\boldsymbol{X}_n = \boldsymbol{i} | \boldsymbol{X}_0 = \boldsymbol{i}\right) < 1, \tag{5.16}$$

we then say that the state is **transient**. We say a state $\boldsymbol{i}$ is **absorbing** if

$$P\left(\boldsymbol{X}_1 = \boldsymbol{i} | \boldsymbol{X}_0 = \boldsymbol{i}\right) = 1. \tag{5.17}$$

The condition (5.17) implies that once you are at state $\boldsymbol{i}$, we cannot move to any other state.

5.3.1 Calculating absorption probabilities

Let $\boldsymbol{j}_a$ is an absorbing state. The probability that a state $\boldsymbol{i}$ will eventually end up at $\boldsymbol{j}_a$ is called the **absorption probability**. We denote the absorption probability by

$$a_{\boldsymbol{i},\boldsymbol{j}_a} \equiv P\left(\text{absorption by } \boldsymbol{j}_a | \boldsymbol{X}_0 = \boldsymbol{i}\right). \tag{5.18}$$

To compute $a_{\boldsymbol{i},\boldsymbol{j}_a}$, we first set $a_{\boldsymbol{i},\boldsymbol{k}_a} = 0$ for any other absorbing states $\boldsymbol{k}_a \neq \boldsymbol{j}_a$. This is because for any other absorbing state, clearly we must have

$$a_{\boldsymbol{i},\boldsymbol{k}_a} = P\left(\text{absorption by } \boldsymbol{k}_a | \boldsymbol{X}_0 = \boldsymbol{j}_a\right) = \delta_{\boldsymbol{k}_a,\boldsymbol{j}_a}, \tag{5.19}$$

since $P\left(\text{absorption by } \boldsymbol{j}_a | \boldsymbol{X}_0 = \boldsymbol{j}_a\right) = 1$ for all absorbing states. To compute the $a_{\boldsymbol{i},b\boldsymbol{j}_a}$ for all $\boldsymbol{i}$ that are not absorbing, we use the law of total probability to derive a set of linear equations

$$\begin{aligned} a_{\boldsymbol{i},\boldsymbol{j}_a} =& P\left(\text{absorption by } \boldsymbol{i} | \boldsymbol{X}_0 = \boldsymbol{j}_a\right) \\ =& \sum_{\boldsymbol{k} \in S} P\left(\text{absorption by } \boldsymbol{j}_a | \boldsymbol{X}_1 = \boldsymbol{k}\right) P\left(\boldsymbol{X}_1 = \boldsymbol{k} | \boldsymbol{X}_0 = \boldsymbol{i}\right) \\ =& \sum_{\boldsymbol{k} \in S} a_{\boldsymbol{k},\boldsymbol{j}_a} P\left(\boldsymbol{k} | \boldsymbol{i}\right). \end{aligned} \tag{5.20}$$

That is we have (for all non-absorbing states)

$$a_{\boldsymbol{i},\boldsymbol{j}_a} = \sum_{\boldsymbol{k} \in S} p_{\boldsymbol{i}\boldsymbol{k}} a_{\boldsymbol{k},\boldsymbol{j}_a}. \tag{5.21}$$

On the last line we made use of the fact that since the transition probabilities are stationary, $a_{\boldsymbol{i},\boldsymbol{j}_a}$ does not depends only on $\boldsymbol{i}$ and $\boldsymbol{j}_a$, that is it does not depend on what step of the chain we are on. With $a_{\boldsymbol{k}_a,\boldsymbol{j}_a} = \delta_{\boldsymbol{k}_a,\boldsymbol{j}_a}$ for the absorbing states, (5.21) defines a set of linear equations that we can solve for the absorption probabilities to $\boldsymbol{j}_a$, $a_{\boldsymbol{k},\boldsymbol{j}_a}$, for the non-absorbing states.

5.3.2 Calculating mean absorption times

In addition to determine the probability of reaching a given absorbing state $\boldsymbol{j}_a$, another question of practical importance is determining the expectation of the number of steps it will take to reach *any* absorbing state. We denote the mean steps by

$$\mathbb{E}\left[\text{steps } T \text{ until absorption} | \boldsymbol{i}\right] = \mu_{\boldsymbol{i}}. \tag{5.22}$$

Clearly $\mu_{\boldsymbol{j}_a} = 0$ for all absorbing states $\boldsymbol{j}_a$. To compute $\mu_{\boldsymbol{i}}$ for non-absorbing states $\boldsymbol{i}$, we take one step, and use the law of total probability (iterated expectation)

$$\begin{aligned} \mu_{\boldsymbol{i}} =& \mathbb{E}\left[\text{steps } T \text{ until absorption} | \boldsymbol{\theta}_0 = \boldsymbol{i}\right] \\ =& 1 + \sum_{\boldsymbol{j}} \mathbb{E}\left[\text{steps } T \text{ until absorption} | \boldsymbol{\theta}_1 = \boldsymbol{j}\right] P\left(\boldsymbol{\theta}_1 = \boldsymbol{j} | \boldsymbol{\theta}_0 = \boldsymbol{i}\right) \\ =& 1 + \sum_{\boldsymbol{j}} \mu_{\boldsymbol{j}} P\left(\boldsymbol{j} | \boldsymbol{i}\right). \end{aligned} \tag{5.23}$$

That is we have (for all non-absorbing states)

$$\mu_{\boldsymbol{i}} = 1 + \sum_{\boldsymbol{j}} p_{\boldsymbol{i}\boldsymbol{j}} \mu_{\boldsymbol{j}}. \tag{5.24}$$

Equation (5.24) defines a set of linear equations for the expectation of the absorption time. Note that this is the expectation to be absorbed by *any* state, as opposed to (5.21), which defines the probability to be absorbed by a *specific* state $\boldsymbol{j}_a$.

Chapter 6

Numerical integration

We we discussed in Sec. 1, in parametric Bayesian statistics our goal is to determine the posterior probability distribution of the parameters of the model under consideration, given a set of measured data, or to determine the total evidence for the model.

A large portionof omputational, parameteric Bayesian statistics essentially consists of determining ways to compute high dimensional integrals. To understand why we need to compute integrals in parameteric Bayesian statistics we look again at Bayes theorem

$$P\left(\boldsymbol{\theta}|\boldsymbol{x}\right) = \frac{\mathcal{L}\left(\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)}{\mathcal{Z}\left(x\right)}, \tag{6.1}$$

where π is the prior and the the likelihood $\mathcal{L}$ and evidence $\mathcal{Z}$ are

$$\mathcal{L}\left(\boldsymbol{\theta}\right) = \prod_{i=1}^{N} P\left(\boldsymbol{x}_i|\boldsymbol{\theta}\right), \tag{6.2}$$

$$\mathcal{Z}\left(\boldsymbol{\theta}\right) = \int d^k\theta \mathcal{L}\left(\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right). \tag{6.3}$$

We have assumed the measurements of $\boldsymbol{x}$ have been taken idependently on one another in the equation for the likelihood. We assume that the parameters $\boldsymbol{\theta}$ are continuous. Clearly if we want to determine $P\left(\boldsymbol{\theta}|\boldsymbol{x}\right)$ directly, we need to compute the evidence $\mathcal{Z}$, which requires integrating over the likelihood. Beyond this though, many summary statistics of practical interest require computing an integral. For example, we may be interested in the expectation of $\boldsymbol{\theta}$ and its covariance matrix

$$\mu_i = \mathbb{E}\left[\theta_i\right] \equiv \int d^k\theta P\left(\boldsymbol{\theta}|\boldsymbol{x}\right)\theta_i, \tag{6.4a}$$

$$C_{ij} \equiv \mathbb{E}\left[\left(\theta_i - \mu_i\right)\left(\theta_j - \mu_j\right)\right] \equiv \int d^k\theta P\left(\boldsymbol{\theta}|\boldsymbol{x}\right)\left(\theta_i - \mu_i\right)\left(\theta_j - \mu_j\right), \tag{6.4b}$$

We cover methods to compute high dimensional integrals, as in many applications $k \gg 1$ (or at least, $k \gtrsim 10$). In this regime, it is usually computationally infeasible to compute (6.3)

using traditional deterministic methods such as the trapezoid rule or Gaussian quadrature. For example if $k = 10$, and if we have 10 quadrature points in each parameter direction, we will need to make $N \gtrsim 10^{10}$ evaluations for a trapsoid rule approximation ot the evidence. The likelihood function is often a highly complex function with sharp peaks, so many more than 10 grid points would be needed to resolve in each direction in order to properly resolve the posterior.

As far as I am aware, the most efficient way to compute high dimensional integrals is through stochastic/Monte Carlo methods. The fact that Monte Carlo methods are the best methods to compute many high dimensional integrals is somewhat surprising, as they have very slow rates of convergence. In general, the error of Monte Carlo integrals goes as $N^{-1/2}$, where N is the number of points used in the approximation. In one dimension, approximations as simple as the trapezoid rule converge to the correct answer as $1/N^2$ (e.g. [PTVF92]). This being said, the accuracy of methods such as the trapezoid rule rapidly deteriorate at higher dimension, while for Monte Carlo methods, the accuracy decreases as $N^{-1/2}$, regardless of the dimensionality of the problem, althouh the proportionality constant to this decrease strongly depends on the choice of algorithm one uses, and the dimensionality of the problem. We only consider stochastic/Monte Carlo integration methods in this chapter.

In effect, Bayesian parametric statistics reduces statistics to probability theory, and many problems in probability theory can be reduced to problems in the integration of complciated functions in high dimensional spaces. There are three main approaches to integration, **Riemann integration**, **Riemannian-Stieltjes integration**, and **Lebesgue integration**.

Monte Carlo integration can be thought of as providing an approximation to the Riemannian integral. We review Monte Carlo integration in Sec. 6.1. The most commonly used variant of Monte Carlo integration is **Markov chain Monte Carlo** (MCMC) integration, which can be thought of as approximating the Riemann-Stieltjes integral. We eview MCMC integration in Sec. 6.2. The Monte Carlo approximation of certain kinds of Lebesgue integrals goes under the name **Nested Sampling** (NS), which we review in Sec. 6.3. There are many excellent, long discussions of all these methods on the internet and elsewhere (e.g. [BGJM11, Ski06, HFM18]), so we only outline the main ideas.

Before continuing, we mention two applications where you do not need to compute an integral (and hence do not need to use the methods discussed here). If we only need to compute the ratio of the posterior for two parameters values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, we only need to determine $P(\boldsymbol{\theta}_1|\boldsymbol{x}) / P(\boldsymbol{\theta}_2|\boldsymbol{x}) = \mathcal{L}(\boldsymbol{\theta}_1) / \mathcal{L}(\boldsymbol{\theta}_2)$, which does not involve any integrals. We also do not need to compute any integrals if we only want the maximum of the posterior or the likelihood (the maximum likelihood estimator). We review some maxization methods in Chptr. 7.

6.1 Monte Carlo integration

Consider a function $f(\boldsymbol{\theta})$, and an integral over the domain Ω

$$I = \int_{\Omega} d^k \theta f(\boldsymbol{\theta}). \tag{6.5}$$

We can view (6.5) as the expectation of f over Ω, with respect to the uniform distribution $U(\Omega)$. In Monte-Carlo integration, we sample points uniformly on Ω, and then approximate I via

$$I_N = \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}_i). \tag{6.6}$$

Here N is the number of times we have sampled from $U(\Omega)$, and $\boldsymbol{x}_i$ are the sample points. Monte Carlo integration works if we can efficiently evaluate f. From the law of large numbers,

$$\lim_{N \to \infty} I_N = I. \tag{6.7}$$

The standard error of the mean goes as $N^{-1/2}$, which gives us our estimate for the error of this approximation; that is we can write (e.g. [PTVF92])

$$\int_{\Omega} d^k \theta f(\boldsymbol{\theta}) \approx V(\Omega) \left(\mathbb{E}[f] \pm \sqrt{\frac{\mathbb{V}[f]}{N}} \right), \tag{6.8}$$

where $V(\Omega)$ is the volume of Ω. We see that the convergence of Monte Carlo integration scales as $1/\sqrt{N}$, regardless of the dimensionality of the integral. This is the key property of stochastic integration methods, and what makes them widespread use in computing high dimensional integrals. In one dimension, almost any other quadrature method outperforms Monte Carlo integration (for example, the error to the trapezoid rule scales as $1/N^2$), but for higher dimensional integrals the convergence of most methods rapidly deteriorates.

We can think of Monte Carlo integration as an example of a stochastic approximation to the Riemann integral of f. Recall that the Riemann integral is the limit of the sum over $f(\boldsymbol{\theta}_i)$ multiplied by the volume of a small (possibly multidimensional) rectangle centered on $f(\boldsymbol{\theta}_i)$, which we call $V(\boldsymbol{\theta}_i)$.

$$\int d^k \theta f(\boldsymbol{\theta}) = \lim_{N \to \infty} \sum_{i=1}^{N} V(\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i). \tag{6.9}$$

In effect, in Monte Carlo integration we approximate $V(\boldsymbol{\theta}_i)$ with n/N, where n is the number of draws we made in that volume.

Monte Carlo integration works best if most of the integral I is not concentrated in a few small volume regions; that is if $f(\boldsymbol{\theta})$ is not too "peaked". Most likelihoods are strongly

peaked though–for example from the Bernstein–von Mises theorem (see Appendix A) we expect the likelihood to approximately behave as a multivariate normal function around the maximum likelihood estimator as the amount of data we collect goes to infinity. Moreover the covariance matrix elements scale as $1/N_d$, where N_d is the number of data points, so the distribution becomes increasingly localized near the maximum likelihood estimator, and more generally near other local maxima of the likelihood. That is, selecting $\boldsymbol{\theta}_i$ from the uniform distribution could mean that we are mostly sampling from places where $f(\boldsymbol{\theta}_i)$ is much smaller than near the peaks. In that case, we would be missing most what contributes to the integral (6.5), which slows down the rate of convergence (the prefactor in front of the asymptotic scaling of $1/\sqrt{N}$).

This motivates the introduction of integrations methods that preferentially sample from regions near the local maxima of the integrand in (6.5). We next discuss two such adaptive methods: Markov chain Monte Carlo (MCMC) methods, which can also be thought of as an adaptive approximation to the Riemann integral, and nested sampling methods, which can be though of as an adaptive approximation to the Lebesgue integral[1]

6.2 Markov chain Monte Carlo (MCMC)

The idea behind MCMC integration is to generate the random samples for the Monte Carlo integration of Ω dynamically, through a Markov Chain. To do this, we rewrite the integral (6.5) as follows

$$I = \int d^k\theta p(\boldsymbol{\theta})\, g(\boldsymbol{\theta}), \tag{6.10}$$

where

$$\int d^k\theta p(\boldsymbol{\theta}) = 1. \tag{6.11}$$

That is, we interpret $p(\boldsymbol{\theta})$ as a probability distribution. We can view (6.10) as a Riemann-Stieltjes integral,

$$I = \int dF(\boldsymbol{\theta})\, g(\boldsymbol{\theta}), \tag{6.12}$$

with the measure $dF(\boldsymbol{\theta}) \equiv d^k\theta p(\boldsymbol{\theta})$. We defined

$$F(\lambda) \equiv \int_0^{g(\boldsymbol{\theta})<\lambda} d^k\theta p(\boldsymbol{\theta}). \tag{6.13}$$

[1]For smooth functions–which is what the posterior distribution function $P(\boldsymbol{\theta}|\boldsymbol{x})$ is, provided the prior and our model $P(\boldsymbol{x}|\boldsymbol{\theta})$ are smooth–there is no substantive, practical difference between the Riemann and Lebesgue integral. Nevertheless we will see that there are different strengths and weaknesses to MCMC and nested sampling, unrelated to the kinds of integrals they are approximating.

Properly speaking, MCMC is a method for drawing samples from $p(\boldsymbol{\theta})$, for use in calculating integrals of the form (6.10). It turns out that a histogram of our sampling of $p(\boldsymbol{\theta})$ will begin to resemble $p(\boldsymbol{\theta})$ as the number of draws goes to infinity. For this reason, MCMC methods are often seen as ways to determine the "shape" or functional properties of $p(\boldsymbol{\theta})$. Here we take the perspective of numerical integration theory, so we think of $p(\boldsymbol{\theta}) = f(\boldsymbol{\theta})/g(\boldsymbol{\theta})$ as a weighting factor for our integration of (6.5). We refer to Chapter 5 for a more a discussion of Markov chains. We will only consider Markov chains with **stationary transition probabilities**, where the transition probabilities $P(\boldsymbol{\theta}_{n+1}|\boldsymbol{\theta}_n)$ do not depend on n.

Operationally, MCMC integration of (6.5) goes as follows.

1. We pick an initial point $\boldsymbol{\theta}_1$, and then generate new samples $\boldsymbol{\theta}_n$ based on a suitably chosen transition probability.

2. For the first few iterations of the Markov Chain, the points $\boldsymbol{\theta}_n$ will be highly correlated with our initial start point, but if one runs the Markov Chain for enough iterations, the points $\{\boldsymbol{\theta}_n\}$ will eventually converge to the target distribution $p(\boldsymbol{\theta})$.

3. Integration then proceeds as in Monte Carlo integration

$$I_N = \frac{1}{N}\sum_{i=1}^{N} g(\boldsymbol{\theta}_i), \tag{6.14}$$

   with similar convergence properties to Monte Carlo integration (the error will asymptotically go does as $1/\sqrt{N}$). The hope though is that the prefactor to the leading asymptotic decay will be much smaller than it would be for regular Monte Carlo integration.

We see that we can view (6.14) as an approximation to the Riemann–Stieltjes integral (6.12).

There are whole volumes on MCMC (e.g. [BGJM11]); for a nice shorter review see [HFM18]. Here we only outline what a "suitable" Markov Chain transition probability must satisfy, the Metropolis-Hastings algorithm, and some limitations of most MCMC methods.

An MCMC chain must eventually limit to a stationary distribution that is equal to $p(\boldsymbol{\theta})$. A sufficient (but not necessary) conditions for a Markov chain to have a stationary distribution $Q(\boldsymbol{\theta})$ is that the transition probabilities must satisfy the **detailed balance** condition

$$P(\boldsymbol{\theta}|\boldsymbol{\psi})\,Q(\boldsymbol{\psi}) = P(\boldsymbol{\psi}|\boldsymbol{\theta})\,Q(\boldsymbol{\theta}), \tag{6.15}$$

for any $\boldsymbol{\theta}, \boldsymbol{\psi}$. To see why detailed balance implies stationarity, we compute the probability of a transition to new step $\boldsymbol{\theta}_n$. The probability distribution for a new step $\boldsymbol{\theta}$ is equal to the integral (or sum, if there a discrete number of points) over all possible earlier points ψ. We assume those are distributed according to the probability distribution $Q(\boldsymbol{\psi})$. We then

show that the distribution for $\boldsymbol{\theta}$, $P(\boldsymbol{\theta})$, is equal to $Q(\boldsymbol{\theta})$, which implies that the chain is stationary. We have

$$\begin{aligned} P(\boldsymbol{\theta}) &= \int d^k\psi P(\boldsymbol{\theta}|\boldsymbol{\psi}) Q(\boldsymbol{\psi}) \\ &= \int d^k\psi P(\boldsymbol{\psi}|\boldsymbol{\theta}) Q(\boldsymbol{\theta}) \\ &= \frac{Q(\boldsymbol{\theta})}{P(\boldsymbol{\theta})} \int d^k\psi P(\boldsymbol{\psi}, \boldsymbol{\theta}) \\ &= Q(\boldsymbol{\theta}) . \end{aligned} \tag{6.16}$$

This proves existence of a stationary chain, but it does not prove uniqueness. Proving uniqueness of the stationary distribution is beyond the scope of these notes. Most practitioners simply ignore the question of uniqueness.

Finally, we discuss an example of a Markov Chain that satisfies the detailed balance condition for the target function $p(\boldsymbol{\theta})$ (lower case p; see (6.10)): the **Metropolis-Hastings algorithm**. Consider a point $\boldsymbol{\theta}$. We draw $\boldsymbol{\psi}$ from the **proposal probability** $Q(\boldsymbol{\psi}|\boldsymbol{\theta})$ (we are free to specify Q). We then draw a random variable x from the uniform distribution $U(0,1)$. We next compute the **acceptance probability**

$$r = \min\left(1, \frac{p(\boldsymbol{\theta})}{p(\boldsymbol{\psi})} \frac{Q(\boldsymbol{\theta}|\boldsymbol{\psi})}{Q(\boldsymbol{\psi}|\boldsymbol{\theta})}\right) . \tag{6.17}$$

If $x > r$, we jump to the point $\boldsymbol{\psi}$, otherwise, we stay at the point $\boldsymbol{\theta}$. To prove that transition probability in the Metropolis-Hastings algorithm satisfies the detailed balance condition, we rewrite the transition probability amplitude as being equal to the proposal probability times the acceptance probability

$$P(\boldsymbol{\psi}|\boldsymbol{\theta}) = r \times Q(\boldsymbol{\psi}|\boldsymbol{\theta}) . \tag{6.18}$$

We then have

$$\begin{aligned} P(\boldsymbol{\theta}|\boldsymbol{\psi}) p(\boldsymbol{\psi}) &= \min\left(p(\boldsymbol{\psi}) Q(\boldsymbol{\psi}|\boldsymbol{\theta}), p(\boldsymbol{\theta}) Q(\boldsymbol{\theta}|\boldsymbol{\psi})\right) \\ &= \min\left(p(\boldsymbol{\theta}) Q(\boldsymbol{\theta}|\boldsymbol{\psi}), p(\boldsymbol{\psi}) Q(\boldsymbol{\psi}|\boldsymbol{\theta})\right) \\ &= P(\boldsymbol{\psi}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) . \end{aligned} \tag{6.19}$$

6.3 Nested sampling

As with MCMC integration, we consider integrals of the form

$$I = \int d^k\theta p(\boldsymbol{\theta}) L(\boldsymbol{\theta}) . \tag{6.20}$$

We can only integrate positive definite functions with the nested sampling algorithm, which is why we use the slightly different notation of L instead of g here: we restrict to functions such that $L \geq 0$. This notation is motivated from the following: the main application of the nested sampling integration is to compute the evidence $\mathcal{Z}$, which is an integral of the prior probability distribution times the likelihood

$$\mathcal{Z} = \int d^k\theta \pi\left(\boldsymbol{\theta}\right) \mathcal{L}\left(\boldsymbol{\theta}\right). \tag{6.21}$$

Before we describe the algorithm, we first need to rewrite (6.20) as an integral over the level sets of $L\left(\boldsymbol{\theta}\right)$. To do this, we write (6.20) as Riemann-Stieltjes integral (6.12), and then integrate by parts. We define the function

$$X\left(\lambda\right) \equiv \int_{L(\boldsymbol{\theta})>\lambda} d^k\theta p\left(\boldsymbol{\theta}\right). \tag{6.22}$$

As λ increases, X decreases from 1 to 0. With this, we have

$$\begin{aligned} I = \int d^k\theta p\left(\boldsymbol{\theta}\right) L\left(\boldsymbol{\theta}\right) &= -\int dXL \\ &= -X\left(L\right) L\big|_{L=0}^{L=L_{max}} + \int_0^{L_{max}} dLX\left(L\right) \\ &= \int_0^{L_{max}} dLX\left(L\right). \end{aligned} \tag{6.23}$$

We assumed that $L_{min} = 0$ (which holds for the likelihood function), and used that $X = 0$ at $L = L_{max}$. We assume that we can invert $X\left(L\right)$ to the function $L\left(X\right)$. We can then rewrite (6.23) by integrating by parts, to obtain

$$I = -\int_1^0 dXL\left(X\right) = \int_0^1 dXL\left(X\right). \tag{6.24}$$

Unlike (6.20), (6.24) is a one-dimensional integral. In (6.24) we should think of L as the parameter in $X\left(L\right)$, not as $L\left(\boldsymbol{\theta}\right)$. Nested sampling provides a noisy approximation to (6.24), through a partitioning of the X interval, and hence provides a noisy approximation to the Lebesgue integral of (6.20).

To understand why (6.24) is the Lebesgue integral of (6.20), recall that the Lebesgue integral is the limit of the sum over $g_i \equiv g\left(\boldsymbol{\theta}_i\right)$ multiplied by the Lebesgue measure of the set E_i of points $\boldsymbol{\theta}_j$ for which $g\left(\boldsymbol{\theta}_j\right) \approx g\left(\boldsymbol{\theta}_i\right)$

$$I_N = \sum_{i=1}^{N} g_i \mu\left(E_i\right). \tag{6.25}$$

We can think of $g_i\mu\left(E_i\right)$ as the discretization of $dX\lambda\left(X\right)$.

The nested sampling algorithm goes as follow

1. We draw n points $\boldsymbol{\theta}_i$ from $p\left(\boldsymbol{\theta}\right)$, treating it as a probability distribution. Set $X_0 = 1$.

2. Repeat for N times, so you have the sequence $X_1, ..., X_n$ and $L_{min,1}, ..., L_{min,N}$. For the j^{th} iteration

   (a) Record the lowest value of $= L_{min,j} = L(\boldsymbol{\theta}_j)$. Set $X_j = e^{-j/n}$, or alternatively set $X_j = t_j X_{j-1}$, where t_j is drawn from the beta distribution Beta $(1, n)$.

   (b) Remove the value of $\boldsymbol{\theta}_j$ that minimizes $L(\boldsymbol{\theta})$, and then sample again from $p(\boldsymbol{\theta})$, until you get a point $\boldsymbol{\theta}_k$ such that $L(\boldsymbol{\theta}_k) > L_{min,j}$.

3. The integral can then be obtained by summing the L values via some quadrature rule

$$I_N = \sum_{i=1}^{N-1} w_i f(L_{min,i}). \tag{6.26}$$

   For example, for the trapezoid rule we would set $w_i = X_i - X_{i-1}$ and $f(L_{min,i}) = (L_{min,i} + L_{min,i-1})/2$.

In the context of Bayes rule, where I_N is our estimate of the evidence $\mathcal{Z}$, we can obtain an estimate for the posterior probability distribution via the rule

$$p(\boldsymbol{\theta}_i) \approx \frac{w_i f(L_{min,i})}{I_N}. \tag{6.27}$$

One of the tricky things to understand about the nested sampling method is the value of the measure of the likelihood $L_{min,i}$, X_i. Consider a sample from $p(\boldsymbol{\theta})$: $\{\theta_j\}$, subject to $L(\boldsymbol{\theta}_j) > L_{min,j-1}$. So long as we draw $\boldsymbol{\theta}_j$ from $p(\boldsymbol{\theta})$ subject to to the constraint $L(\boldsymbol{\theta}) > L_j$, the values of the volumes $X(\boldsymbol{\theta}_j)$ are drawn from $U(0, X_{j-1})$. This follows from the **probability integral transform**, which we review in Chptr. A. Then $X_j = t_j X_{j-1}$, where t_j is the largest of n uniformly distributed numbers in the interval $(0, 1)$. The number t_j is called the **shrinkage factor**. Notice that we have

$$X_j = \prod_{i=1}^{j} t_i. \tag{6.28}$$

The cumulative probability distribution function for the maximum of n randomly distributed numbers in that interval is

$$\begin{aligned} C.D.F.(t_{max}) &= P(\max\{t_1, ..., t_n\} < t_{max}) \\ &= (P(t < t_{max}))^n \\ &= t_{max}^n. \end{aligned} \tag{6.29}$$

The probability density function for the maximum is then the beta distribution Beta $(1, n)$, that is

$$P(t_{max}) = n t_{max}^{n-1}. \tag{6.30}$$

To estimate X_j then, we could take a draw from the Beta distribution, t, and multiply that by X_{j-1}. To get a (presumably) less noisy answer, we could set X_j to be its averaged expected value. We take the expectation of the log of X_j, to simplify the calculation of the expectation

$$\mathbb{E}\left[\log X_j\right] = \sum_{i=1}^{j} \mathbb{E}\left[\log t_i\right]. \tag{6.31}$$

As the t_i are independent, we can estimate the error of this approximation by computing the variance

$$\mathbb{V}\left[\log X_j\right] = \sum_{i=1}^{j} \mathbb{V}\left[\log t_i\right]. \tag{6.32}$$

The expectation value of the logarithm of t_{max} is

$$\begin{aligned}\mathbb{E}\left[\log t_{max}\right] =& \int_0^1 dt\ n t^{n-1}\ \log t \\ =& -\frac{1}{n}.\end{aligned} \tag{6.33}$$

The variance of the log of t_{max}

$$\begin{aligned}\mathbb{V}\left[\log t_{max}\right] =& \mathbb{E}\left[(\log t_{max})^2\right] - \left(\mathbb{E}\left[\log t_{max}\right]\right)^2 \\ =& \int_0^1 dt\ n t^{n-1}\ (\log t)^2 - \frac{1}{n^2} \\ =& \frac{1}{n^2}.\end{aligned} \tag{6.34}$$

Combining everything, we see that the shrinkage factor is approximately

$$\begin{aligned}\log X_j \approx& j\ \mathbb{E}\left[\log t_{max}\right] + \sqrt{j\ \mathbb{V}\left[\log t_{max}\right]} \\ =& -\frac{j}{n}\left(1 \pm \frac{1}{\sqrt{j}}\right).\end{aligned} \tag{6.35}$$

This gives us

$$X_j \approx e^{-j/n}. \tag{6.36}$$

Our approximation to the integral gets better as we add more points n, and as we take more steps N. We incur the biggest relative errors in the integration for the first few small steps j, but so long as L is highly peaked, and we take very small steps, those terms contribute very little to the total integral.

Chapter 7

Numerical optimization

While most often we want to compute integrals of the posterior probability distribution (for example, to compute the mean or covariance matrix of the posterior), sometimes it is informative to simply compute its maximum, or even just the maximum of the likelihood. Again we consider the posterior

$$P\left(\boldsymbol{\theta}|\boldsymbol{x}\right) = \frac{\mathcal{L}\left(\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)}{\mathcal{Z}\left(x\right)}, \tag{7.1}$$

The value of $\boldsymbol{\theta}$ that maximizes $\mathcal{L}\left(\boldsymbol{\theta}\right)$ is the **maximum likelihood estimator (MLE)**, and the value of $\boldsymbol{\theta}$ that maximizes $\mathcal{L}\left(\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)$ is the **maximum a posteriori probability estimator (MAP)**. Note that the MLE and MAP do not gives us any knowledge of the variance of those parameters–that requires knowledge of the full posterior probability distribution. This in turn requires integration of the likelihood.

From a numerical point of view, its convenient to consider numerical minimizers, and to find the MLE by finding the minimum of the negative log likelihood, which we call $\ell\left(\boldsymbol{\theta}\right)$

$$\ell\left(\boldsymbol{\theta}\right) \equiv -\log\mathcal{L}\left(\boldsymbol{\theta}\right). \tag{7.2}$$

It is convenient to consider the log likelihood, as the likelihood itself can vary drastically in value between its maxima and minima, which can be hard for a computer to resolve with finite precision arithematic. The likelihood, prior, evidence, and posterior are positive definite quantities as well, so there is no change of taking a logarithm of these quantities.

Here we review a few minimization methods. There is no best method that will work for all likelihoods, so we only review the basics of a few basic methods that underlie more complex optimization procedures. For concreteness we will focus on minimizing the negative log likelihood $\ell\left(\boldsymbol{\theta}\right)$.

7.1 Convex functions

We first consider the problem of optimizing convex functions. While the posterior is almost never convex, it is still useful to review this case first as the local max/min of a strictly convex function is the global maximum/minimum (this is almost never the case for non-convex functions). Because of this, some methods (namely, Newton's method and its extensions) used to find the minimum of functions try to convert the problem into one for finding the minimum of a convex function.

A **convex function** $f(\boldsymbol{\theta}) : X \to \mathbb{R}$ satisfies

$$f(t\boldsymbol{\theta}_1 + (1-t)\boldsymbol{\theta}_2) \leq tf(\boldsymbol{\theta}_1) + (1-t)f(\boldsymbol{\theta}_2), \tag{7.3}$$

for $t \in [0,1]$ and for all $\boldsymbol{\theta}_{1,2} \in X$ (for example, $X = \mathbb{R}^n$). A **strictly convex** function satisfies (7.4) except the $\leq$ is replaced by $<$, and $t \in (0,1)$ instead.

The local minimum of a convex function is the global minimum of the function. This is easy to show. Say $\boldsymbol{\theta}_*$ is a local minimum, and assume that we have found $\boldsymbol{\theta}$ such that $f(\boldsymbol{\theta}) < f(\boldsymbol{\theta}_*)$. Then we would have

$$\begin{aligned} f(t\boldsymbol{\theta}_* + (1-t)\boldsymbol{\theta}) \leq & tf(\boldsymbol{\theta}_*) + (1-t)f(\boldsymbol{\theta}) \\ < & tf(\boldsymbol{\theta}_*) + (1-t)f(\boldsymbol{\theta}_*) = f(\boldsymbol{\theta}_*). \end{aligned} \tag{7.4}$$

Setting $t = 1$, we encounter a contradiction, which concludes the argument. We see for a strictly convex function, a local minimum is a global minimum, and the global minimum is unique.

7.2 Gradient descent

First we consider a linear method for finding local minima–gradient descent. To understand this method, we Taylor series expand the negative log likelihood about a fiducial point $\boldsymbol{\theta}_0$

$$\ell = \ell_0 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{g}_0 + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{H}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \cdots, \tag{7.5}$$

where

$$g_{0,i} \equiv \nabla_i \ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \tag{7.6}$$

$$H_{0,ij} \equiv \nabla_i \nabla_j \ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}. \tag{7.7}$$

As a local minimum $\boldsymbol{\theta}_*$, the gradient of the function is zero, and the Hessian is positive definite. Near a local minimum then, we expect the gradient to be pointing "away" from the local minimum. Thus if move in the opposite direction to the gradient, we move in the

direction of the local minimum. In gradient descent then, we pick a fiducial value of $\boldsymbol{\theta}_0$, and then iterate the following

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \gamma_n \boldsymbol{g}_n. \tag{7.8}$$

Where $0 < \gamma_n$ is a scalar that one can introduce to make the change between steps be less large. We stop iterating when $|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n| < \epsilon$, where $|\cdots|$ is a norm of our choosing and $0 < \epsilon$ is a pre-set tolerance.

While gradient descent is an easy algorithm to implement, it suffers from a few problems. First, the method (if it converges) only find a local minimum, or possibly only a saddle point. Also, it can be tricky to find a good value of γ_n. If γ_n is too small, the method converges very slowly. If γ_n is too large, the method may never converge.

7.3 Stochastic gradient descent

Typically the gradient descent algorithm described in (7.2) will only find a local minimum of the target function $\ell(\boldsymbol{\theta})$. The idea behind stochastic gradient descent is to not take the full gradient of $\ell(\boldsymbol{\theta})$ when taking the step (7.8). That is, one instead computes

$$g_{n,i}^{\mathcal{I}_n} \equiv \nabla_i \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n} \delta_{\mathcal{I}_n i}, \tag{7.9}$$

where $\delta_{\mathcal{I}_n i} = 1$ if $i \in \mathcal{I}_n$ and is zero otherwise. Here $\mathcal{I}_n$ is a randomly chosen set if indices $\{i_{1,n}, ...i_{k,n}\}$. At each step we change the set $\mathcal{I}_n$. In stochastic gradient descent we then effectively take jumps in parameter space down randomly chosen gradients of the target function. Empirically, in many cases this allows the algorithm to jump out of local minima and saddle points of the target function. While there is no general method to determine if the algorithm has reached a global minimum, it almost always finds a minimum that is smaller than could be found with a simple gradient descent method.

In practice, a common way to use the algorithm is to pre-set several batches of indices $\mathcal{I}_1, ..., \mathcal{I}_p$, and then to shuffle through those sets in each iteration. The simplicity and effectiveness of stochastic gradient descent makes it a widely used algorithm.

7.4 Newton's method

We next consider a quadratic method for finding local minima–Newton's method (this is Newton's method for optimizing a function, not for finding the root to a function). To understand this method, we Taylor series expand the negative log likelihood about a fiducial point $\boldsymbol{\theta}_0$

$$\ell = \ell_0 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{g}_0 + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{H}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \cdots, \tag{7.10}$$

where

$$g_{0,i} \equiv \nabla_i \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \tag{7.11}$$

$$H_{0,ij} \equiv \nabla_i \nabla_j \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}. \tag{7.12}$$

We choose $\boldsymbol{\theta}$ to minimize the quadratic Taylor series expansion. *Assuming* that $\boldsymbol{H}_0$ is positive definite, minimizing the quadratic Taylor series expansion is an exercise in minimizing a convex function. The minimum is then located at the zero of the gradient of the second order Taylor series. We find that

$$\boldsymbol{g}_0 + \boldsymbol{H}_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = 0 \implies \boldsymbol{\theta} = \boldsymbol{\theta}_0 - \boldsymbol{H}_0^{-1}\boldsymbol{g}_0. \tag{7.13}$$

This motivates **Newton's method**. Starting with a fiducial point $\boldsymbol{\theta}_0$, we iterate in $\boldsymbol{\theta}_n$, where at each iteration we set

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \gamma_n \boldsymbol{H}_n^{-1}\boldsymbol{g}_n. \tag{7.14}$$

Where $0 < \gamma_n \leq 1$ is a scalar that one can introduce to make the change between steps be less large. We stop iterating when $|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n| < \epsilon$, where $|\cdots|$ is a norm of our choosing and $0 < \epsilon$ is a pre-set tolerance.

As with the gradient descent method, Newton's method may only find a local minimum of saddle point. There are also numerous technical problems with inversion of the Hessian. First, the Hessian matrix may be very large if there are many parameters, so it could be hard to invert (it may be ill conditioned). Additionally the Hessian could be singular, or nearly singular. We note that the Hessian may not be positive definite either at a given point (it often won't be), which in principle isn't fatal to the method, but depending on the size of the eigenvalues to the Hessian, large negative eigenvalues could dramatically change the value of $\boldsymbol{\theta}_{n+1}$ versus $\boldsymbol{\theta}_n$.

Chapter 8

Linear regression

Here we review some aspects of linear regression. We adopt a Bayesian viewpoint to motivate linear regression, although frequentist concepts may enter the discussion at several places.

Appendix A

Probability theory

A.1 Note on notation

We denote the probability of measuring $\boldsymbol{\theta}$ by $P(\boldsymbol{\theta})$, the joint probability by $P(\boldsymbol{\theta}, \boldsymbol{x})$, and the conditional probability by $P(\boldsymbol{\theta}|\boldsymbol{x})$. In probability theory it is common to denote a probability distribution function (PDF) as $f_{\boldsymbol{X}}(\boldsymbol{x})$ for a random variable $\boldsymbol{X}$. To get a probability from a continuous PDF, you need to perform an integral. In other words, technically $f_{\boldsymbol{X}}(\boldsymbol{X})$ is not a probability, but $\int_V d^n x f_{\boldsymbol{X}}(\boldsymbol{X})$ over some volume V is a probability. In situations where we may be refering to either a PDF or a probability, we will simply use P, which hopefully won't be too confusing.

When a random variables $\boldsymbol{X}$ has a PDF $f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})$, we say $\boldsymbol{X} \sim f(\boldsymbol{\theta})$, where f denotes that probability and $\boldsymbol{\theta}$ are the hyperparameters of the model.

A.2 Conditional probability and Bayes theorem

The conditional probability is

$$P(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{P(\boldsymbol{\theta}, \boldsymbol{x})}{P(\boldsymbol{x})}. \tag{A.1}$$

Using this, we have Bayes theorem

$$P(\boldsymbol{\theta}|\boldsymbol{x})\, P(\boldsymbol{x}) = P(\boldsymbol{x}|\boldsymbol{\theta})\, P(\boldsymbol{\theta}). \tag{A.2}$$

Bayes theorem is more often written as

$$P(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\boldsymbol{\theta})\, P(\boldsymbol{\theta})}{P(\boldsymbol{x})}. \tag{A.3}$$

As we discuss more in Chptr. 1, we can view $P(\boldsymbol{\theta}|\boldsymbol{x})$ as the distribution of model parameters given a model $P(\boldsymbol{x}|\boldsymbol{\theta})$, and some prior knowledge of the model parameters $P(\boldsymbol{\theta})$.

A.3 Cumulative distribution function

The **cumulative distribution function** (CDF) F for a probability distribution function $f_{\boldsymbol{\Theta}}$ is

$$\begin{aligned} F_{\boldsymbol{\Theta}}(\xi) &= P\left(\boldsymbol{\theta} < \boldsymbol{F}_{\boldsymbol{\Theta}}^{-1}(\xi)\right) \\ &= \int_{P(\boldsymbol{\theta})<\xi} d^k\theta f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}), \end{aligned} \tag{A.4}$$

where

$$\boldsymbol{F}_{\boldsymbol{\Theta}}^{-1}(\xi) \equiv \inf\{\boldsymbol{\theta} \ : \ F_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \geq \xi\}. \tag{A.5}$$

Note that $F_{\boldsymbol{\Theta}}\left(\boldsymbol{F}_{\boldsymbol{\Theta}}^{-1}(\boldsymbol{\theta})\right) = \boldsymbol{\theta}$. Sometimes in the literature you'll see $F_{\boldsymbol{\Theta}}^{-1}(\xi)$ – which we can efectively think of as the inverse of the CDF – is called the **percent point function**.

The **probability integral transform** states that the random variable Ξ defined to be

$$\Xi = F_{\boldsymbol{\Theta}}(\xi), \tag{A.6}$$

has the standard uniform distribution, that is $\Xi \sim U(0,1)$. To prove this, we look at the CDF of ξ

$$\begin{aligned} F_{\Xi}(\xi) &= P(\Xi < \xi) \\ &= P(F_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) < \xi) \\ &= P\left(\boldsymbol{\theta} < \boldsymbol{F}_{\boldsymbol{\Theta}}^{-1}(\xi)\right) \\ &= F_{\boldsymbol{\Theta}}\left(\boldsymbol{F}_{\boldsymbol{\Theta}}^{-1}(\xi)\right) \\ &= \xi. \end{aligned} \tag{A.7}$$

We see that the CDF of $F_{\Xi}(\xi)$ is the same as the CDF for the uniform distribution $U(0,1)$. We conclude that $\Xi \sim U(0,1)$. This fact is used in the nested sampling integration algorithm, which we discussed in Chpt. 6.

A.4 Functions of random variables

We consider two random (scalar) variables X and Y, with the join PDF $f_{X,Y}(x,y)$. Our goal is to find the PDF of the function $Z(X,Y)$ (generalizing to functions of a larger number of random variables is straighforward). To find this, we first determine the set

$$A_z = \{(x,y) \ : \ Z(x,y) \leq z\}. \tag{A.8}$$

Given this, we can then compute the CDF

$$F_Z(z) = \int_{A_z} dxdy f_{X,Y}(x,y). \tag{A.9}$$

From the CDF we can then compute the PDF by taking the derivative of the CDF

$$f_Z(z) = \frac{d}{dz} F_Z(z). \tag{A.10}$$

A.5 Change of variables for injective mappings

If we restrict ourselves to injective mappings between random variables, we can derive a simple closed-form expression for the PDF of a variable $\boldsymbol{\Psi}(\boldsymbol{\Theta})$. Notice that we do not need to restrict ourselves to scalar random variables here. Consider a PDF $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$. What is the probability distribution to $P(\boldsymbol{\psi}(\boldsymbol{\theta}))$, where $\boldsymbol{\psi}$ is some function of $\boldsymbol{\theta}$? The following remains unchanged under a change of variables

$$\begin{aligned}\int_V d^k\theta f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) &= \int_V d^k\psi f_{\boldsymbol{\Psi}}(\boldsymbol{\psi}) \\ &= \int_V d^k\theta \left|\det(J_{ij})\right| f_{\boldsymbol{\Psi}}(\boldsymbol{\psi}).\end{aligned} \tag{A.11}$$

We viewed V as a geometric volume (that is, it is independent of the coordinate choice we use). Here

$$J_{ij} \equiv \frac{\partial\psi^i}{\partial\theta_j}, \tag{A.12}$$

is the Jacobian matrix. Equating terms within the integral, we find that

$$f_{\boldsymbol{\Psi}}(\boldsymbol{\psi}) = \frac{1}{\left|\det(J_{ij})\right|} f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}). \tag{A.13}$$

As an example application of this formula, we consider the posterior probability distribution for $\boldsymbol{\psi}(\boldsymbol{\theta})$. If $\boldsymbol{\psi}(\boldsymbol{\theta})$ is not injective, we need to replace the RHS of (A.14) with a sum over the different values of $\boldsymbol{\theta}$ that map to the same $\boldsymbol{\psi}$. In other words, we have

$$f_{\boldsymbol{\Psi}}(\boldsymbol{\psi}) = \sum_{\boldsymbol{\theta}\,:\,\boldsymbol{\psi}(\boldsymbol{\theta})=\boldsymbol{\psi}} \frac{1}{\left|\det(J_{ij})\right|} f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}). \tag{A.14}$$

Note that in general this "sum" may in fact be an integral. From Bayes theorem (A.3), we then have

$$\begin{aligned}f_{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\boldsymbol{x}) &= \frac{f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\psi}(\boldsymbol{\theta}))\, f_{\boldsymbol{\Psi}}(\boldsymbol{\psi})}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\ &= \frac{1}{\left|\det(J_{ij})\right|} \frac{f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\psi}(\boldsymbol{\theta}))\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{f_{\boldsymbol{X}}(\boldsymbol{x})}.\end{aligned} \tag{A.15}$$

That is, to find the probability distribution for some function of the distribution parameters, we only need to find the probability distribution for the prior under that change in coordinates.

A.6 Expectation and covariance

We review a few basic definitions from probability theory, as they come up later in the notes. We define the **expectation** of a random variable $\boldsymbol{\Theta}(\boldsymbol{x})$ to be

$$\mathbb{E}[\boldsymbol{\Theta}] \equiv \int d^k x f_{\boldsymbol{X}}(\boldsymbol{x}) \boldsymbol{\Theta}(\boldsymbol{x}), \tag{A.16}$$

where $f_{\boldsymbol{X}}$ is the PDF of $\boldsymbol{x}$. Sometimes we denote the expectation with $\mu_{\boldsymbol{\Theta}}$. The **variance** is

$$\mathbb{V}_{ij}[\boldsymbol{\Theta}] \equiv \mathbb{E}\left[(\Theta_i - \mu_{\Theta_i})\left(\Theta_j - \mu_{\Theta_j}\right)\right]. \tag{A.17}$$

The **covariance** for two random variables $\boldsymbol{\Theta}, \boldsymbol{\Psi}$ is

$$\mathbb{C}_{ij}[\boldsymbol{\Theta}, \boldsymbol{\Psi}] \equiv \mathbb{E}\left[(\Theta_i - \mu_{\Theta_i})\left(\Psi_j - \mu_{\Psi_j}\right)\right]. \tag{A.18}$$

Note that we can think of the expectation of two scalar random variables Θ, Ψ as an inner product

$$\mathbb{E}[\Theta\Psi] = \langle \Theta, \Psi \rangle. \tag{A.19}$$

It is easy to see from (A.16) that (A.19) satisfies the properties of an inner product: $\langle \Theta, \Theta \rangle \geq 0$, $\langle \Theta, \Psi \rangle = \langle \Psi, \Theta \rangle$, and linearity.

Consider a random variable $\boldsymbol{\Theta}(\boldsymbol{\psi})$. Say we want this variable to represent another random variable, say the parameters of the posterior $\boldsymbol{\theta}$. We then call $\boldsymbol{\Theta}$ an **estimator** for $\boldsymbol{\theta}$. The **bias** of $\boldsymbol{\Theta}$ then is

$$\boldsymbol{b}(\boldsymbol{\Theta}) \equiv \mathbb{E}[\boldsymbol{\Theta}] - \boldsymbol{\theta}. \tag{A.20}$$

If $\boldsymbol{b} = 0$, then $\boldsymbol{\Theta}$ is an **unbiased estimator** for $\boldsymbol{\theta}$.

A.7 Characteristic/moment generating function

To prove the central limit theorem, first we introduce the Fourier transform (or **characteristic function**) of a probability distribution. Consider a random vector $\boldsymbol{X}$ with probability distribution $f_{\boldsymbol{X}}(\boldsymbol{x})$, that is $\boldsymbol{X} \sim f_{\boldsymbol{X}}$. We denote the moment generating function with $\psi_{\boldsymbol{X}}$, which is

$$\psi_{\boldsymbol{X}}(\boldsymbol{t}) \equiv \int_{-\infty}^{\infty} d^k x e^{i\boldsymbol{t}^T \boldsymbol{x}} f_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{E}\left[e^{i\boldsymbol{t}^T \boldsymbol{X}}\right]. \tag{A.21}$$

If we set $\boldsymbol{t} = -i\tilde{\boldsymbol{t}}$, then $\psi_{\boldsymbol{X}}$ is called the **moment generating function**. For most probability distributions, there is no meaningful difference between using $\boldsymbol{t}$ or $-i\tilde{\boldsymbol{t}}$ (there could

potentially only be a difference if the distribution had complex poles or branch cuts). Notice that

$$\mathbb{E}\left[X_{i_1}\cdots X_{i_l}\right] = \frac{1}{i^l}\nabla_{t_{i_1}}\cdots\nabla_{t_{i_l}}\psi_{\boldsymbol{X}}\left(\boldsymbol{t}\right). \tag{A.22}$$

That is, we can obtain the moments of the probability distribution from the characteristic function (although we need to divide by $1/i^l$).

Perhaps most importantly, notice that since the characteristic function for a probability distribution is the Fourier transform of the probability density, we can uniquely map a probability density to its characteristic function and back. That is, given a characteristic function, we can find the unique probability density that it corresponds to.

Consider a linear affine transformation of the random variable $\boldsymbol{X}$, which we call $\boldsymbol{Y} = a\boldsymbol{X}+\boldsymbol{b}$. We also call $\boldsymbol{y} = a\boldsymbol{x}+\boldsymbol{b}$. The probability distribution with the volume element remains unchanged $dyf_{\boldsymbol{Y}}\left(\boldsymbol{x}\right) = dxf_{\boldsymbol{X}}\left(\boldsymbol{x}\right)$. We conclude that

$$\begin{aligned}\psi_{\boldsymbol{Y}}\left(\boldsymbol{t}\right) =& \int_{-\infty}^{\infty} d^k y e^{i\boldsymbol{t}^T\boldsymbol{y}} f_{\boldsymbol{Y}}\left(\boldsymbol{y}\right)\\ =& e^{i\boldsymbol{t}^T\boldsymbol{b}}\int_{-\infty}^{\infty} d^k x e^{ia\boldsymbol{t}^T\boldsymbol{x}} f_{\boldsymbol{X}}\left(\boldsymbol{x}\right)\\ =& e^{i\boldsymbol{t}^T\boldsymbol{b}}\psi_{\boldsymbol{X}}\left(at\right).\end{aligned} \tag{A.23}$$

The characteristic function of a sum of independent variables is the product of the characteristic function for each variable. Define $\boldsymbol{Y} = \sum_{i=1}^n \boldsymbol{X}_i$, we then have

$$\begin{aligned}\psi_{\boldsymbol{Y}}\left(\boldsymbol{t}\right) =& \int_{-\infty}^{\infty} dx_1^k \cdots \int_{-\infty}^{\infty} dx_n^k e^{i\boldsymbol{t}^T\sum_i \boldsymbol{x}_i} f_{\boldsymbol{Y}}\left(\boldsymbol{x}_1,...,\boldsymbol{x}_n\right)\\ =& \prod_{i=1}^n \int_{-\infty}^{\infty} dx_i^k e^{i\boldsymbol{t}^T\boldsymbol{x}_i} f_{\boldsymbol{X}_i}\left(\boldsymbol{x}_i\right)\\ =& \prod_{i=1}^n \psi_{\boldsymbol{X}_i}\left(\boldsymbol{t}\right).\end{aligned} \tag{A.24}$$

The second line follows from $f_{\boldsymbol{Y}}\left(\boldsymbol{x}_1,...,\boldsymbol{x}_n\right) = \prod_i f_{\boldsymbol{X}_i}\left(\boldsymbol{x}_i\right)$, as all the variables are independent.

A.8 Central limit theorem

Let $\boldsymbol{X}_1,...,\boldsymbol{X}_n$ be n be independent and identically distributed random vectors (of dimension k each), and let each variable have mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The central limit theorem states that the probability distribution of the average of these variables,

$$\bar{\boldsymbol{X}}_n \equiv \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i, \tag{A.25}$$

limits to a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}/n$ as $n \to \infty$. Note that we made no assumption about the probability distribution for the $\boldsymbol{X}_i$, except that the probability distribution has a finite mean and variance. We can write the central limit theorem as

$$\lim_{n\to\infty} \sqrt{n}\bar{\boldsymbol{X}}_n \sim N_k\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \tag{A.26}$$

where N_k is the multivariate normal distribution. To prove this, we make use of the characteristic function for $\boldsymbol{X}_n$, which is

$$\begin{aligned}
\psi_{\boldsymbol{X}_n}\left(\boldsymbol{t}\right) =& \prod_{i=1}^{n} \psi_{\boldsymbol{X}_i}\left(\frac{\boldsymbol{t}}{n}\right) \\
=& \left(1 + i\frac{1}{n}\boldsymbol{t}^T\boldsymbol{\mu} + i^2\frac{1}{2}\frac{1}{n^2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t} + \mathcal{O}\left(\frac{1}{n^3}\right)\right)^n \\
=& \exp\left[i\boldsymbol{t}^T\boldsymbol{\mu} + i^2\frac{1}{2}\boldsymbol{t}^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{t}\right]\left(1 + \mathcal{O}\left(\frac{1}{n^3}\right)\right),
\end{aligned} \tag{A.27}$$

where $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}/n$. We used the identity

$$\lim_{n\to\infty}\left(1 + \frac{a}{n}\right)^n = e^a. \tag{A.28}$$

to leading order, the last line of (A.27) is the characteristic function for $N_k\left(\boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}\right)$ (see (B.8)). This concludes the proof.

A.9 Fisher information and the Bernstein–von Mises theorem

The **Fisher information** is the negative expectation of the Hessian of the log likelihood. In terms of components, we have

$$\begin{aligned}
F_{ij}\left(\boldsymbol{\theta}\right) \equiv& -\mathbb{E}_\theta\left[\nabla_{\theta^i}\nabla_{\theta^j}\ln P\left(\boldsymbol{x}|\boldsymbol{\theta}\right)\right] \\
=& -\int d^n x P\left(\boldsymbol{x}|\boldsymbol{\theta}\right)\nabla_{\theta^i}\nabla_{\theta^j}\ln P\left(\boldsymbol{x}|\boldsymbol{\theta}\right).
\end{aligned} \tag{A.29}$$

Here we view $P\left(\boldsymbol{x}|\boldsymbol{\theta}\right)$ as the PDF for $\boldsymbol{X}$, that has dependence on the **hyperparamters** $\boldsymbol{\theta}$. For example, we could imagine $P = \left(2\pi\sigma\right)^{-1/2}\exp\left[-\frac{1}{2}\left(x-\mu\right)\right]$, and that μ, σ as the hyperparameters of the model. We set the dimensionality of $\boldsymbol{x}$ to be n and the dimensionality of $\boldsymbol{\theta}$ to be k. The Fisher information can also be written as the variance of the **score function**. The score function is

$$s_i\left(\boldsymbol{x};\boldsymbol{\theta}\right) \equiv \nabla_{\theta^i}\ln P\left(\boldsymbol{x};\boldsymbol{\theta}\right). \tag{A.30}$$

The expectation of the score function is zero

$$\begin{aligned}\mathbb{E}\left[s_i\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right] =& \int d^n x P\left(\boldsymbol{x};\boldsymbol{\theta}\right)\nabla_{\theta^i}\ln P\left(\boldsymbol{x};\boldsymbol{\theta}\right)\\ =&\nabla_{\theta^i}\int d^n x P\left(\boldsymbol{x};\boldsymbol{\theta}\right) = 0.\end{aligned} \tag{A.31}$$

We then have

$$\begin{aligned}F_{ij}\left(\boldsymbol{\theta}\right) =& -\int d^n x P\left(\boldsymbol{x};\boldsymbol{\theta}\right)\nabla_{\theta^i}\nabla_{\theta^j}\ln P\left(\boldsymbol{x};\boldsymbol{\theta}\right)\\ =&\int d^n x\left[\frac{1}{P\left(\boldsymbol{x};\boldsymbol{\theta}\right)}\nabla_{\theta^i}P\left(\boldsymbol{x};\boldsymbol{\theta}\right)\nabla_{\theta^j}P\left(\boldsymbol{x};\boldsymbol{\theta}\right) - \nabla_{\theta^i}\nabla_{\theta^j}P\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right]\\ =&\int d^n x P\left(\boldsymbol{x};\boldsymbol{\theta}\right)\nabla_{\theta^i}\ln P\left(\boldsymbol{x};\boldsymbol{\theta}\right)\nabla_{\theta^j}\ln P\left(\boldsymbol{x};\boldsymbol{\theta}\right)\\ =&\mathbb{E}_\theta\left[s_i s_j\right]\\ =&\mathbb{V}_{\theta,ij}\left[\mathbf{s}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right].\end{aligned} \tag{A.32}$$

That is, the Fisher information is the variance of the score.

Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimator for $\boldsymbol{\theta}$. Under appropriate regularity conditions, the likelihood $\mathcal{L}\left(\boldsymbol{\theta}\right)$ tends towards a multivariate number with mean $\hat{\boldsymbol{\theta}}$ and covariance matrix given by the inverse Fisher information divided by the number of measurements of the data n, $\tilde{\boldsymbol{F}}^{-1} = \boldsymbol{F}^{-1}/n$. In equations, we have

$$\lim_{n\to\infty} P\left(\boldsymbol{\theta}|\boldsymbol{x}\right) = \lim_{n\to\infty}\frac{\pi\left(\boldsymbol{\theta}\right)\prod_{i=1}^n P\left(\boldsymbol{x}_i|\boldsymbol{\theta}\right)}{\mathcal{Z}\left(\boldsymbol{x}\right)} = N\left(\hat{\boldsymbol{\theta}},\tilde{\boldsymbol{F}}^{-1}\right). \tag{A.33}$$

This is known as the **Bernstein–von Mises theorem** (BvM theorem for short). We provide a rough sketch of how the proof goes. For more details see [Was10]. The log likelihood is

$$\ln\mathcal{L}\left(\boldsymbol{\theta}\right) = \sum_{i=1}^N \ln P\left(\boldsymbol{x}_i|\boldsymbol{\theta}\right) \tag{A.34}$$

We Taylor series expand the derivative of the log-likelihood to linear order about a point $\boldsymbol{\theta}_0$

$$\nabla_{\theta^i}\ln\mathcal{L}\left(\boldsymbol{\theta}\right) = \nabla_{\theta^i}\ln\mathcal{L}\left(\boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \nabla_{\theta^i}\nabla_{\theta^j}\ln\mathcal{L}\left(\boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\left(\theta-\theta_0\right)^j + \cdots. \tag{A.35}$$

Setting $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, relabeling $\boldsymbol{\theta}_0 \to \boldsymbol{\theta}$ (and dropping the $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, to reduce clutter), and rearranging gives us (we used that at the maximum likelihood estimator, $\hat{\boldsymbol{\theta}}$, the derivative of the likelihood is zero)

$$\sqrt{n}\left(\hat{\theta}^i - \theta^i\right) = -\left(\frac{1}{n}\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\ln\mathcal{L}\left(\boldsymbol{\theta}\right)\right)^{-1}_{ij}\left(\frac{1}{\sqrt{n}}\nabla_{\theta^j}\ln\mathcal{L}\left(\boldsymbol{\theta}\right)\right). \tag{A.36}$$

As $n\to\infty$, we see that

$$\lim_{n\to\infty}\nabla_{\theta^j}\ln\mathcal{L}\left(\boldsymbol{\theta}\right) = \frac{1}{\sqrt{n}}\lim_{n\to\infty}\sum_{m=0}^n \nabla_{\theta^j}\ln P\left(\boldsymbol{\theta}|\boldsymbol{x}_m\right)$$

$$\to \sim N_k\left(\mathbf{0}, \boldsymbol{F}\right). \tag{A.37}$$

This follows from the central limit theorem: the mean of the score is zero, and the variance of the score is the Fisher information. By the law of large numbers we can average

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\left(-\left(\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} \ln \mathcal{L}\left(\boldsymbol{\theta}\right)\right)_{ij}\right) &= \lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n}\left(-\nabla_{\theta^i}\nabla_{\theta^j} \ln P\left(\boldsymbol{\theta}|\boldsymbol{x}_m\right)\right) \\ &\to F_{ij}. \end{aligned} \tag{A.38}$$

Thus the variance of the limit is modified to be $\boldsymbol{F}^{-1}\boldsymbol{F}\boldsymbol{F}^{-1} = \boldsymbol{F}^{-1}$. We can then conclude that

$$\lim_{n\to\infty} \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \sim N_k\left(\mathbf{0}, \boldsymbol{F}^{-1}\right). \tag{A.39}$$

Or in other words

$$\lim_{n\to\infty} \sqrt{n}\hat{\boldsymbol{\theta}} \sim N_k\left(\hat{\boldsymbol{\theta}}, \boldsymbol{F}^{-1}\right). \tag{A.40}$$

We have not been careful by what we mean by "$\to$" and "$\sim$" here–in fact there are various notions of convergence that go into the full proof (see for example [Was10]).

We can understand the BvM theorem heuristically as follows. As we collect more data, the posterior probability becomes increasingly "peaked" near the maximum likelihood estimator. We can then Taylor series about the maximum of the log-likelihood to quadratic order. Exponentiating the log-likelihood gives us a multivariate normal with the inverse Fisher information as the covariance matrix.

A.10 Fisher information and the Cramér-Rao bound

Consider an estimator $\boldsymbol{\Theta}$ for model parameters $\boldsymbol{\theta}$. Let $\boldsymbol{\Sigma}$ is the covariance matrix for the estimator $\boldsymbol{\Theta}$, let $\mathbb{E}\left[\boldsymbol{\Theta}\right] = \boldsymbol{\psi}$, and let $\boldsymbol{F}$ be the Fisher information evaluated at $\boldsymbol{\psi}$. The Cramér-Rao bound states that

$$\Sigma_{ij} \geq \nabla_{\theta_m}\psi_i \nabla_{\theta_n}\psi_j F^{-1}_{mn}, \tag{A.41}$$

If $\boldsymbol{\Theta}$ is an unbiased estimator ($\boldsymbol{\psi} = \boldsymbol{\theta}$), then (A.41) reduces to

$$\Sigma_{ij} \geq F^{-1}_{ij}, \tag{A.42}$$

If $\boldsymbol{\Theta}$ is a biased estimator, then $\nabla_{\theta_i}\psi_j = \delta_{ij} + \nabla_{\theta_i} b_j$, where b_j is the bias. The Cramér-Rao bound can be used to interpret the Fisher matrix as an estimate for the lowest error one could achieve for an unbiased estimator. For biased estimators though, we see that the Fisher information does not give a lower bound on the elements of the covariance matrix, since it is possible that the bias could be negative, $\nabla_{\theta_i} b_j < 0$. That is, biased estimators can have

smaller covariance matrix elements than unbiased estimators. If the error from the bias is less than the error from the covariance (for example, if one only has a few measurements of noisy data), a biased estimator can sometimes be superior to an unbiased estimator in determining $\boldsymbol{\theta}$.

Here we provide the outline of a proof of (A.41). First we prove a generalization of the Cauchy-Schwartz inequality. Let $\boldsymbol{y}$ and $\boldsymbol{z}$ be random vectors (not necessarily of the same dimensionality). Then

$$\mathbb{V}_{ij}\left[\boldsymbol{z}\right] \geq \mathbb{C}_{ip}\left[\boldsymbol{z},\boldsymbol{y}\right]\mathbb{V}_{pq}\left[\boldsymbol{y}\right]^{-1}\mathbb{C}_{qj}\left[\boldsymbol{y},\boldsymbol{z}\right]. \tag{A.43}$$

To prove this, we define $\boldsymbol{u} \equiv \boldsymbol{y} - \boldsymbol{\mu_y}$ and $\boldsymbol{v} \equiv \boldsymbol{z} - \boldsymbol{\mu_z}$, so $\boldsymbol{\mu_u} = 0$ and $\boldsymbol{\mu_v} = 0$. For any matrix $\boldsymbol{A}$ we have the following matrix inequality (we insert the matrix in case $\boldsymbol{v}$ and $\boldsymbol{u}$ have different dimensionality)

$$\left(\boldsymbol{v} + \boldsymbol{A}\boldsymbol{u}\right)\left(\boldsymbol{v} + \boldsymbol{A}\boldsymbol{u}\right)^{T} \geq 0. \tag{A.44}$$

Taking the expectation of this and expanding, we have

$$\mathbb{E}\left[\boldsymbol{v}\boldsymbol{v}^{T}\right] + \boldsymbol{A}\mathbb{E}\left[\boldsymbol{u}\boldsymbol{v}^{T}\right] + \mathbb{E}\left[\boldsymbol{v}\boldsymbol{u}^{T}\right]\boldsymbol{A}^{T} + \boldsymbol{A}\mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^{T}\right]\boldsymbol{A}^{T} \geq 0. \tag{A.45}$$

Set $\boldsymbol{A} = -\mathbb{E}\left[\boldsymbol{u}\boldsymbol{v}^{T}\right]\mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^{T}\right]^{-1}$. The last two terms cancel, and we are left with

$$\mathbb{E}\left[\boldsymbol{v}\boldsymbol{v}^{T}\right] \geq \mathbb{E}\left[\boldsymbol{u}\boldsymbol{v}^{T}\right]\mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^{T}\right]^{-1}\mathbb{E}\left[\boldsymbol{u}\boldsymbol{v}^{T}\right]. \tag{A.46}$$

Re-introducing $\boldsymbol{y}$ and $\boldsymbol{z}$, and using the definition of the covariance (A.17) and variance (A.18), we have (A.43),

$$\mathbb{V}_{ij}\left[\boldsymbol{z}\right] \geq \mathbb{C}_{ip}\left[\boldsymbol{y},\boldsymbol{z}\right]\mathbb{V}_{pq}\left[\boldsymbol{y}\right]^{-1}\mathbb{C}_{qj}\left[\boldsymbol{y},\boldsymbol{z}\right]. \tag{A.47}$$

This completes the proof of the generalized Cauchy-Schwartz inequality.

We now prove (A.41). We use (A.43), and set

$$\boldsymbol{z} = \boldsymbol{\Theta}, \qquad \boldsymbol{y} = \boldsymbol{s}, \tag{A.48}$$

where $\boldsymbol{\Theta}$ is an estimator for $\boldsymbol{\theta}$, and $\boldsymbol{s}$ is the score (see (A.30)). The covariance between $\boldsymbol{\Theta}$ and $\boldsymbol{s}$ is

$$\begin{aligned}
\mathbb{C}_{ij}\left[\boldsymbol{\Theta},\boldsymbol{s}\right] =& \mathbb{E}\left[\left(\Theta_i - \mu_{\Theta_i}\right)\left(s_i - \mu_{s_j}\right)\right] \\
=& \mathbb{E}\left[\Theta_i s_j\right] \\
=& \int d^n x P\left(\boldsymbol{x}|\boldsymbol{\theta}\right)\Theta_i \nabla_{\theta_j} \ln P\left(\boldsymbol{x}|\boldsymbol{\theta}\right) \\
=& \nabla_{\theta_j}\mathbb{E}\left[\Theta_i\right] \\
=& \nabla_{\theta_j}\psi_i.
\end{aligned} \tag{A.49}$$

We also have

$$\mathbb{V}_{ij}\left[\boldsymbol{\Theta}\right] = \Sigma_{ij}, \qquad \mathbb{V}_{ij}\left[\boldsymbol{s}\right] = F_{ij}. \tag{A.50}$$

We have defined $\boldsymbol{\Sigma}$ to be the covariance matrix of $\boldsymbol{\Theta}$, and used that Fisher information is the variance of the score (see (A.32)), Plugging this all into (A.43), we obtain the Cramér-Rao bound

$$\Sigma_{ij}\Big|_{\boldsymbol{\theta}=\boldsymbol{\mu}_{\boldsymbol{\Theta}}} \geq \nabla_{\theta_p}\psi_i \nabla_{\theta_1}\psi_j F^{-1}_{pq}(\boldsymbol{\theta})\,. \tag{A.51}$$

Appendix B

Common probability distributions

We briefly review some common probability distributions and some of their properties.

B.1 Continuous distributions

B.1.1 Exponential family of distributions

A large number of PDFs fall under the exponential family, so we begin by reviewing this class of distributions. The PDF of the exponential family of distributions take the form

$$f_{\boldsymbol{X}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)=h\left(\boldsymbol{x}\right)g\left(\boldsymbol{\theta}\right)e^{\boldsymbol{\eta}^{T}\left(\boldsymbol{\theta}\right)\boldsymbol{T}\left(\boldsymbol{x}\right)}. \tag{B.1}$$

A large number of PDFs fall under the exponential family. The functional form of $h, g, \boldsymbol{\eta}, \boldsymbol{T}$ determine the particular instantiation of the distribution. The vector $\boldsymbol{\eta}$ is called the **natural parameter** of the exponential family. Consider a sequence of N independent and identical draws (IID) from an expoential family. The likelihood is

$$\mathcal{L}=g\left(\boldsymbol{\theta}\right)^{N}\left(\prod_{i=1}^{N}f\left(\boldsymbol{x}_{i}\right)\right)\exp\left[\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)\sum_{i=1}^{N}\boldsymbol{T}\left(\boldsymbol{x}_{i}\right)\right]. \tag{B.2}$$

We can view $\prod f\left(\boldsymbol{x}_{i}\right)$ as a normalization factor in the likelihood. We see then that in effect the likelihood depends only on $\boldsymbol{x}$ through $\boldsymbol{T}_{N}\equiv\sum_{i}\boldsymbol{T}\left(\boldsymbol{x}_{i}\right)$. This makes $\boldsymbol{T}_{N}$ a **sufficient statistic** for the exponential family of distributions. A sufficient (set of) statistic(s) completey describe the likelihood (or probability distribution).

B.1.2 Uniform distribution

The uniform distribution finds widespread use mostly because it is simple to manipulate and simple in understand. The uniform distribution is also commonly used as a "non-informative prior" (see Chpt. 1).

The uniform distribution over the interval (a, b) is denoted by $U(a, b)$. We write $X \sim U(a, b)$. The PDF is

$$f_X(x; a, b) = \begin{cases} a < x < b & \frac{1}{b-a} \\ \text{otherwise} & 0 \end{cases}. \tag{B.3}$$

The characteristic function is

$$\begin{aligned} \psi_X(t) &= \int_{-\infty}^{\infty} dx e^{ixt} f_X(x) \\ &= \frac{1}{b-a} \int_a^b dx e^{ixt} \\ &= \frac{1}{b-a} \frac{e^{ibt} - e^{iat}}{it}. \end{aligned} \tag{B.4}$$

The mean and variance are

$$\mu = \mathbb{E}[x] = \frac{b+a}{2}, \tag{B.5}$$

$$\sigma^2 = \mathbb{E}\left[(x-\mu)^2\right] = \frac{(b-a)^2}{12}. \tag{B.6}$$

From the mean and variance, we can determine a, b, and hence $U(a, b)$. We see that μ, σ^2 form a sufficient set of statistics.

B.1.3 Multivariate normal distribution

Many observed quantities in nature are approximately distrbuted according to the normal (or Gaussian) distribution. That the normal distribution appears so commonly in practice can be at least partially explained in part by the central limit theorem (see Appendix. A): the normal distribution is the limiting distribution of the mean of a large number of random variables drawn from any distribution with a finite mean and variance. This being said, there are plenty of cases where this does not happen, that is one may not be drawing from a random variable that is effectively the average of many of random variables of finite mean and variance. So while the normal distribution is commonly found in practice, it is certainly not the only probability distribution one encounters in practice.

The multivariate normal distribution in $\mathbb{R}^k$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The PDF is

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi \det \boldsymbol{\Sigma})^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]. \tag{B.7}$$

The characteristic function for the multivariate normal distribution plays an important role in the proof of the central limit theorem that we review in Appendix A. The characteristic function is

$$\begin{aligned}\psi_{\boldsymbol{X}}(\boldsymbol{t}) =& \int_{-\infty}^{\infty} d^k x e^{i\boldsymbol{t}^T\boldsymbol{x}} f_{\boldsymbol{X}}(\boldsymbol{x}) \\ =& \int_{-\infty}^{\infty} d^k x \frac{1}{\sqrt{2\pi}|\det \boldsymbol{\Sigma}|} \exp\left[i\boldsymbol{t}^T\boldsymbol{x} - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \\ =& \exp\left[i\boldsymbol{t}^T\boldsymbol{\mu} + i^2\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right] \\ & \times \int_{-\infty}^{\infty} d^k x \frac{1}{\sqrt{2\pi}|\det \boldsymbol{\Sigma}|} \exp\left[-\frac{1}{2}(\boldsymbol{x}-i\boldsymbol{\Sigma}\boldsymbol{t}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-i\boldsymbol{\Sigma}\boldsymbol{t}-\boldsymbol{\mu})\right] \\ =& \exp\left[i\boldsymbol{t}^T\boldsymbol{\mu} + i^2\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right]. \end{aligned} \tag{B.8}$$

The mean and covariance matrix are just $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$,

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}, \tag{B.9}$$

$$\boldsymbol{\Sigma} = \mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})\right] = \boldsymbol{\Sigma}. \tag{B.10}$$

Clearly μ, σ^2 form a sufficient set of statistics for the normal distribution.

B.1.4 Chi-square distribution

The chi-square distribution describes the distribution of the sum of the normalized, squared deviates from the mean of draws from the normal distribution. The chi-square distribution can be used to diagnose how well a given data set is approximated by the normal distribution with a given mean and variance.

Let $X_1, ..., X_N$ be independent samples drawn from $N(\mu, \sigma)$. We define the random variable

$$Q_N \equiv \sum_{i=1}^{N} \frac{(X_i - \mu)^2}{\sigma^2}. \tag{B.11}$$

The variable Q_N is then distributed according to the chi-square distribution with N degrees of freedom, $Q_N \sim \chi^2_N$. Understanding how the chi-squared PDF is derived is somewhat interesting, so we outline the details of the derivation here. To find the PDF for Q_N, f_{Q_N}, we first convert to the normalized variables

$$Z_i = \frac{X_i - \mu}{\sigma}. \tag{B.12}$$

The constraint (B.11) then is

$$Q_N = \sum_{i=1}^{N} Z_i^2. \tag{B.13}$$

We could apply the change of variables formula (A.14) to find f_{Q_N}. It is slightly easier to start from the observation that we want

$$f_{Q_N} dQ_N = \prod_{i=1}^{N} f_{Z_i} dZ_i, \tag{B.14}$$

subject to the constraint (B.13). We then convert to spherical polar coordinates. We set the radius $R^2 = Q_N$, and then a set of $N-1$ angular coordinates θ_j In spherical polar coordinates, the constraint (B.13) is particular simple: it just constrains the radius to a constant value $R = \sqrt{Q_N}$. The differential volume $dZ_1 \cdots dZ_N$ then reduces to the area of the $(N-1)$-sphere at radius R, $S_{N-1}(R)$, multiplied by dR Then (B.14) reduces to

$$f_{Q_N} dQ_N = S_{N-1}(R)\, dR \prod_{i}^{N} f_{Z_i}. \tag{B.15}$$

We have $dR = dQ_N/(2Q_N^{1/2})$ and

$$\begin{aligned} \prod_{i}^{N} f_{Z_i} &= (2\pi)^{-N/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{N} Z_i^2\right) \\ &= (2\pi)^{-N/2} \exp\left(-\frac{1}{2} Q_N\right). \end{aligned} \tag{B.16}$$

Finally we note that

$$S_{N-1}(R) = \frac{2R^{N-1}\pi^{N/2}}{\Gamma(N/2)}, \tag{B.17}$$

where Γ is the Gamma function. We conclude that the chi-square distribution with N degrees of freedom is

$$f_{Q_N}(q_N) = \frac{1}{2^{N/2}\Gamma(N/2)} q_N^{N/2-1} e^{-q_N/2}. \tag{B.18}$$

Remember this is the distribution for the squared normalized deviates from the mean, (B.11), and that we are restricting q_N to the positive real line. Notice that for $N > 2$, (B.18) is not peaked at $q_N = 0$.

More quantitatively, the mean and variance of the chi-squared distribution are

$$\mu = \mathbb{E}[q_N] = N, \tag{B.19}$$

$$\sigma^2 = \mathbb{V}\left[(q_N - \mu)^2\right] = 2N. \tag{B.20}$$

Notice that the mean grows as N increases.

B.1.5 Student's t-distribution

The chi-square distribution describes the distribution of (B.11), that is the sum of the normalized square deviates from the sample mean. We now consider the problem of inferring the posterior probability for the mean and variance of normal distribution, given data $D = \{x_1, ..., x_N\}$. Notice that in (B.11) we assumed that the random variables X_i were drawn from $N(\mu, \sigma^2)$, as we assumed that we knew μ and σ^2. We assume that we do not care about the variance.

$$\begin{aligned} P(\mu|D) &= \int d\sigma P(\mu, \sigma|D) \\ &= \int d\sigma \frac{P(D|\mu, \sigma) P(\mu, \sigma)}{P(D)}. \end{aligned} \tag{B.21}$$

B.1.6 Beta distribution

B.2 Discrete distributions

B.2.1 Binomial distribution

Consider a coin. The binomial distribution describes the number of heads you would obtain after n flips of the coin, assuming the probability of obtaining heads for each flip was $0 \leq p \leq 1$. Many other processes can be described by a binomial distribution.

The PDF of the binomial distribution is

$$f_X(x; p, n) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}. \tag{B.22}$$

The mean and variance are

$$\mu = \mathbb{E}[x] = np, \tag{B.23}$$

$$\sigma^2 = \mathbb{E}\left[(x-\mu)^2\right] = np(1-p). \tag{B.24}$$

Notice that the variance grows unbounded as we increase n.

B.2.2 Poisson distribution

The Poisson distribution often occurs in situations where one is counting up the number of occurances of a very rare event.

The PDF of the Poisson distribution is

$$f_X(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}. \tag{B.25}$$

We can see why they Poisson PDE describes the number of occurances of a very rare event by taking such a limit of the Binomial PDF. We take (B.22), and take the limit $p \to 0$ and $n \to \infty$ such that the combination $np = \lambda$ is finite. We set x to be a constant value, so that $n \gg x$. We then have

$$\begin{aligned}
\frac{n!}{x!\,(n-x)!}p^x\,(1-p)^{n-x} =& \frac{n!}{x!\,(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\
=& \frac{n!}{x!\,(n-x)!}\left(\frac{\lambda/n}{1-\lambda/n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n} \\
\to& \frac{1}{x!}\left(\frac{\lambda}{n}\right)^x e^{-\lambda}.
\end{aligned} \tag{B.26}$$

If we took the formal limit $n \to \infty$, this would reduce to zero. So instead, we do *not* take that exact limit.

The mean and variance of the Poisson distribution are

$$\mu = \mathbb{E}[x] = \lambda, \tag{B.27}$$

$$\sigma^2 = \mathbb{E}\left[(x-\mu)^2\right] = \lambda. \tag{B.28}$$

Appendix C

Common special functions

We briefly review the properties of some special functions that we use in the text.

C.1 Exponential

The exponential function is e^x. The Taylor series expansion of the exponential function about $x = 0$ is

$$e^x = \sum_{n=0}^{\infty} \frac{1}{n!} x^n, \tag{C.1}$$

this series has an infinite radius of convergence. Another useful formula is

$$e^x = \lim_{n\to\infty} \left(1 - \frac{x}{n}\right)^n. \tag{C.2}$$

C.2 Gamma

The gamma function is $\Gamma(x)$. It is a continuous generalization of the factorial function $x!$. The most common definitionof Γ is

$$\Gamma(x) = \int_0^{\infty} dt\; e^{-t}\, t^{x-1}. \tag{C.3}$$

The Gamma function satisfies

$$\Gamma(x) = (x-1)\,\Gamma(x-1). \tag{C.4}$$

We see that $\Gamma(1) = 1$, and that $\Gamma(-n) = \infty$ for all $n = 0, 1, 2, 3, ...$ Notice that heuristically this "had" to be so, as from (C.4)

$$\Gamma(1) = (1-1)\,\Gamma(0) = 1. \tag{C.5}$$

For this to hold, we "need" $\Gamma(0) = \infty$.

Appendix D

Fourier and other transforms

D.1 Brief review of complex analysis

For a function $A(t)$ that is singular at infinity, the **Cauchy principal value** is defined to be

$$\text{p.v.} \int_{-\infty}^{\infty} dt A(t) = \lim_{T\to\infty} \int_{-T}^{T} dt A(t) . \tag{D.1}$$

For complex-valued functions $A(z)$ that are singular at a point z_0, the Cachy principal value is defined to be the limit of the deformation of the integral C by a disk of radius ϵ centered around z_0

$$\text{p.v.} \int_{C} dz A(z) = \lim_{\epsilon\to 0^+} \int_{C(\epsilon)} dz A(z) . \tag{D.2}$$

This can also be written as

$$\text{p.v.} \int_{C} dz A(z) = \lim_{\epsilon\to 0^+} \left(\int_{-\infty}^{z_0-\epsilon} dz A(z) + \int_{z_0+\epsilon}^{\infty} dz A(z) \right) . \tag{D.3}$$

D.2 The Fourier transform

We briefly review Fourier transforms, along with a helpful transforms that are used in signal processing.

The one-dimensional **Fourier transform** and its inverse are

$$A(t) = \mathcal{F}^{-1}\left[\tilde{A}(f)\right](t) = \int_{-\infty}^{\infty} df e^{2\pi i f t} \tilde{A}(f) , \tag{D.4a}$$

$$\tilde{A}(f) = \mathcal{F}[A(t)](f) = \int_{-\infty}^{\infty} df e^{-2\pi i f t} A(t). \tag{D.4b}$$

The Fourier representation of the Dirac delta function $\delta(t)$ is

$$\tilde{\delta}(f) = \int_{-\infty}^{\infty} df e^{-2\pi i f t} \delta(t) = 1. \tag{D.5}$$

The **convolution** of two functions $A(t)$ and $B(t)$ are

$$(A * B)(t) \equiv \int_{-\infty}^{\infty} d\tau A(\tau) B(t-\tau) = \int_{-\infty}^{\infty} d\tau A(t-\tau) B(\tau). \tag{D.6}$$

The Fourier transform of the convlution is

$$\begin{aligned} \mathcal{F}[(A * B)(t)](f) &= \int_{-\infty}^{\infty} d\tau \int_{-\infty}^{\infty} df \int_{-\infty}^{\infty} df' e^{2\pi i f \tau} e^{2\pi i f'(t-\tau)} \tilde{A}(f) \tilde{B}(f') \\ &= \int_{-\infty}^{\infty} df e^{2\pi i f t} \tilde{A}(f) \tilde{B}(f). \end{aligned} \tag{D.7}$$

That is, convolution in real space is multiplication in frequency space.

D.3 The Laplace transform

The **Laplace transform** of a function $A(t)$ is

$$A(t) = \mathcal{L}^{-1}[A(\lambda)](t) = \frac{1}{2\pi i} \lim_{T\to\infty} \int_{\gamma - iT}^{\gamma + iT} d\lambda e^{\lambda t} \tilde{A}(\lambda), \tag{D.8a}$$

$$\tilde{A}(t) = \mathcal{L}[A(t)](\lambda) = \int_0^{\infty} dt e^{-\lambda t} A(t), \tag{D.8b}$$

Here γ is a real number so that the contour path of integration is in the region of convergence of $\tilde{A}(\lambda)$. In effect, the inverse Laplace transform is like the inverse Fourier transform.

D.4 The Hilbert transform

The **Hilbert transform** of a function $A(t)$ is

$$A(t) = \mathcal{H}^{-1}\left[\tilde{A}(\tau)\right](t) = -\frac{1}{\pi} \text{p.v.} \int_{-\infty}^{\infty} d\tau \frac{\tilde{A}(\tau)}{t-\tau}, \tag{D.9a}$$

$$\tilde{A}(\tau) = \mathcal{H}[A(t)](\tau) = \frac{1}{\pi} \text{p.v.} \int_{-\infty}^{\infty} dt \frac{A(t)}{\tau - t}. \tag{D.9b}$$

Appendix E

Stationary phase approximation

E.1 Stationary phase approximation

Here we review the **stationary phase approximation** for the Fourier transform. For more discussion see [BO99]. Consider a complex function, which we write as

$$B(t) = A(t) e^{i\phi(t)}. \tag{E.1}$$

The Fourier transform is

$$\tilde{B}(f) = \int_{-\infty}^{\infty} dt A(t) e^{i\phi(t) - 2\pi i f t}. \tag{E.2}$$

We imagine $A(t)$ is a slowly varying function, while $\phi(t)$ is rapidly varying. We then expect that the integral for $\tilde{B}(f)$ will be dominated by the stationary points of $\phi(t) - 2\pi f t$, that is the points where

$$\frac{d\phi}{dt} - 2\pi f = 0. \tag{E.3}$$

This can be more formally justified by the Riemann-Lebesgue lemma, which states that

$$\lim_{x\to\infty} \int_a^b dt e^{ixt} A(t) = 0, \tag{E.4}$$

provided $\int_a^b dt A(t)$ exists. We can extend $a, b \to \pm\infty$ so long as $A(t)$ is integrable. Going back to (E.2), we assume that $\phi(t) - 2\pi f t$ has one stationary point for each value of f, which we call $t_0(f)$. That is, $t_0(f)$ is defined to solve the stationary phase equation

$$\left.\frac{d\phi}{dt}\right|_{t=t_0} - 2\pi f = 0. \tag{E.5}$$

We Taylor series expand about the stationary point to quadratic order in ϕ,

$$\phi(t) - 2\pi f t = \phi(t_0) - 2\pi f t_0 + \frac{1}{2}\frac{d^2\phi}{dt^2}\Big|_{t=t_0}(t-t_0)^2 + \mathcal{O}\left[(t-t_0)^3\right], \tag{E.6}$$

insert this into (E.2), and obtain

$$\begin{aligned}\tilde{B}(f) \approx & A(t_0)\, e^{i\phi(t_0)-2\pi i f t_0}\int_{-\infty}^{\infty} dt \exp\left[i\frac{1}{2}\frac{d^2\phi}{dt^2}\Big|_{t=t_0}(t-t_0)^2\right] \\ = & \left[\frac{1}{2}\frac{d^2\phi}{dt^2}\Big|_{t=t_0}\right]^{-1/2} A(t_0)\, e^{i\phi(t_0)-2\pi i f t_0+i\pi/4}\int_{-\infty}^{\infty} dx e^{-x^2} \\ = & \left[\frac{1}{2\pi}\frac{d^2\phi}{dt^2}\Big|_{t=t_0}\right]^{-1/2} A(t_0)\, e^{i\phi(t_0)-2\pi i f t_0+i\pi/4}.\end{aligned} \tag{E.7}$$

Using the chain rule, we can write $\phi(t_0)$ and t_0 as integral equations in terms of the frequency. We have

$$t_0(f) = \int^f df' \frac{dt}{df}, \tag{E.8a}$$

$$\phi(t_0) = 2\pi \int^f df' \frac{dt}{df} f'. \tag{E.8b}$$

Defining $\dot{f} \equiv df/dt$, we see that we can write the phase of $\tilde{B}(f)$ as

$$\begin{aligned}\Psi \equiv & \phi(t_0) - 2\pi f t_0 + \frac{\pi}{4} \\ = & 2\pi \int^f df' \frac{1}{\dot{f}}(f' - f) + \frac{\pi}{4}.\end{aligned} \tag{E.9}$$

Bibliography

[BGJM11] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, Boca Raton, FL, USA, 2011.

[BO99] C.M. Bender and S.A. Orszag. *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*. Advanced Mathematical Methods for Scientists and Engineers. Springer, 1999.

[CA11] Jolien D. E. Creighton and Warren G. Anderson. *Gravitational-wave physics and astronomy: An introduction to theory, experiment and data analysis*. Wiley-VCH, 2011.

[DL70] James M. Dickey and B. P. Lientz. The Weighted Likelihood Ratio, Sharp Hypotheses about Chances, the Order of a Markov Chain. *The Annals of Mathematical Statistics*, 41(1):214 – 226, 1970.

[Fin92] Lee S. Finn. Detection, measurement and gravitational radiation. *Phys. Rev. D*, 46:5236–5249, 1992.

[HFM18] David W. Hogg and Daniel Foreman-Mackey. Data analysis recipes: Using Markov Chain Monte Carlo. *Astrophys. J. Suppl.*, 236(1):11, 2018.

[Mac03] David J. C. Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[PN14] H. Pishro-Nik. *Introduction to Probability, Statistics, and Random Processes*. Kappa Research, LLC, 2014.

[PTVF92] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, MA, USA, second edition, 1992.

[Ski06] John Skilling. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833 – 859, 2006.

[Was10] Larry Wasserman. *All of Statistics : A Concise Course in Statistical Inference*. Springer, New York, 2010.