AMATH 562
Advanced Stochastic Processes
Homework 1

Lucas Cassin Cruz Burke

Due: January 16, 2023

1. Write about the relationship between the mathematical theory of probability and its applications to real-world data.

   Solution: In probability theory, as in all areas of applied mathematics, we seek a logically consistent mathematical framework which we can use to model and understand the real world. The mathematical framework must of course come from pure mathematics, which has developed extremely powerful tools such as measure theory and calculus which are ideally suited to the study of probability theory and stochastic processes. However, in the course of applying these to real world problems we often run into paradoxes arising from real-world limitations which do not exist in the realm of pure mathematics. Many central mathematical results rely upon the existence of limits of infinite sequences or on smooth real-valued functions, however in the physical world of observed quantities we have no reason to believe that these objects exist, or even make sense to talk about. Even if we restrict our theory to "well-defined and well-behaved limits of finite sets", as Jaynes implores us, we must still contend with having basic concepts, such as probability measures or random variables, whose real-world meaning is frustratingly ambiguous, despite being perfectly well-defined mathematically.

   Equations (1) and (2) are two examples of how this tension arises when trying to interpret the mathematical theory of probability in the context of real-world data. Each of these equations convey what appears at first glance to be intuitive statements. For instance, the law of large numbers for empirical mean value (1) captures our intuitive understanding that, for a fair coin, averaging its value over larger and larger samples should eventually get us closer to its expected value $\mathbb{E}X$. However upon closer inspection there are several issues with this which must be explored. For one, both mathematical expressions require the existence of a sequence of independent and identically distributed variables X_i. Mathematically there is no issue with this statement. But in a physical system such as, for instance, a coin-flipping experiment, one must either flip n separate coins at once, or flip the same coin n different times. By modeling this process as a sequence of i.i.d. random variables we have already introduced a major assumption about our experiment which seems incompatible with the real experimental setup.

Moreover, both (1) and (2) define statistical properties of X in terms of infinite sums, which clearly do not exist in the context of real-world data. Even more troubling is that these are sums of random variables, which are functions on Ω used to model empirical statistical quantities, and yet the rhs of both (1) and (2) are abstract mathematical concepts. The rhs of (2) is a probability, which is defined on $\mathcal{F}$, a completely different space from the random variables it is supposed to be derived from. To make sense of these expressions, we must first take great care to define what we mean by convergence of random variables. In this case, the limiting process is defined in terms of convergence in distribution, or weak convergence. While this again makes the expressions mathematically sensible, the trade-off is that the quantities on the rhs are not defined in terms of X_i at all, but in terms of their distribution.

(1) and (2) demonstrate that even for extremely simple probabilistic concepts there is constant tension between the mathematical theory of probability and its application in the real world, where pure mathematical concepts of infinite limits and smooth real valued functions do not exist. They also show that while the mathematical objects we use to model probabilistic systems are well-defined mathematically, it is not always clear what they correspond to in a real-world setting. Disagreements over how one should interpret the rhs of (1) and (2) have been waged for more than a hundred years, and as the nature of big data and statistics continues to shift it is likely that the dominant interpretation will shift as well.

2. Give two examples $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ with $X_1(\omega)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ with $X_2(\omega')$ and show that in both cases the cumulative probability function is given by

$$\mathbb{P}_1\left(X_1(\omega) > x\right) = \mathbb{P}_2\left(X_2(\omega') > x\right) = e^{-rx}$$

Solution: Consider the measurable space $(\mathbb{R}^+, \mathcal{F})$ where $\mathcal{F} = \mathcal{B}(\mathbb{R}^+)$. We define two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ on $\mathcal{F}$ by $\mathbb{P}_1(dx) = dx$ and $\mathbb{P}_2(dx) = re^{-rx}dx$.

Now consider the random variables $X_1(x) = re^{-rx}$ and $X_1(x) = x$.

3. Consider a reference measure $\mathbb{P}_1$ and a collection of continuously parameterized measures $\mathbb{P}_2(\theta)$. Assume the RND $Z(\omega;\theta) = \frac{d\mathbb{P}_2}{d\mathbb{P}_1}(\omega;\theta)$ is smooth with respect to θ.

Now let

$$\mathcal{I}_k(\theta) = -\mathbb{E}^{\mathbb{P}_2}\left[\frac{\partial^k}{\partial\theta^k}\log\left(\frac{d\mathbb{P}_2}{d\mathbb{P}_1}(\omega;\theta)\right)\right]$$

Assuming expectations and differentiations with respect to θ are interchangeable, show the following:

(a)

$$\mathcal{I}_0(\theta) = -\int_\Omega \left(\frac{d\mathbb{P}_2}{d\mathbb{P}_1}(\omega;\theta)\right)\log\left(\frac{d\mathbb{P}_2}{d\mathbb{P}_1}(\omega;\theta)\right)\mathbb{P}_1(dw)$$

This is called the Shannon entropy of $\mathbb{P}_2$ w.r.t. the measure $\mathbb{P}_1$.

Solution: We begin with the definition for $\mathcal{I}_0(\theta)$.

$$\begin{aligned}\mathcal{I}_0(\theta) &= -\mathbb{E}^{\mathbb{P}_2}\left[\log\left(\frac{d\mathbb{P}_2}{d\mathbb{P}_1}(\omega;\theta)\right)\right] \\ &= -\int_\Omega \log Z\mathbb{P}_2(d\omega)\end{aligned}$$

We can now use the RND to change the measure used in the expectation from $\mathbb{P}_2$ to $\mathbb{P}_1$.

$$\begin{aligned}\mathcal{I}_0(\theta) &= -\int_\Omega \log Z\frac{d\mathbb{P}_2}{d\mathbb{P}_1}\mathbb{P}_1(d\omega) \\ &= -\int_\Omega \left(\frac{d\mathbb{P}_2}{d\mathbb{P}_1}(\omega;\theta)\right)\log\left(\frac{d\mathbb{P}_2}{d\mathbb{P}_1}(\omega;\theta)\right)\mathbb{P}_1(dw)\end{aligned}$$

Which is the Shannon entropy of $\mathbb{P}_2$ w.r.t. the measure $\mathbb{P}_1$.

(b)

$$\mathcal{I}_1(\theta) = 0$$

Solution: We begin by writing $\mathcal{I}_1(\theta)$ from the above definition. We have

$$\begin{aligned}I_1(\theta) &= -\mathbb{E}^{\mathbb{P}_2}\left[\frac{\partial}{\partial\theta}\log\left(\frac{d\mathbb{P}_2}{d\mathbb{P}_1}\right)\right] \\ &= -\mathbb{E}^{\mathbb{P}_2}\left[\frac{\partial}{\partial\theta}\log Z\right] \\ &= -\int_\Omega \frac{Z_\theta}{Z}\mathbb{P}_2(d\omega)\end{aligned}$$

We now use the RND to change the measure used in the expectation from $\mathbb{P}_2$ to $\mathbb{P}_1$.

$$I_1(\theta) = -\int_\Omega \frac{Z_\theta}{Z}\frac{d\mathbb{P}_2}{d\mathbb{P}_1}\mathbb{P}_1(d\omega)$$

But $Z = d\mathbb{P}_2/d\mathbb{P}_1$, so after canceling terms we are left with

$$I_1(\theta) = -\int_\Omega Z_\theta\mathbb{P}_1(d\omega) = -\int_\Omega \frac{\partial}{\partial\theta}Z\mathbb{P}_1(d\omega) = -\frac{\partial}{\partial\theta}\int_\Omega Z\mathbb{P}_1(d\omega)$$

where in the last step we have utilized our assumption that expectation and differentiation with respect to θ are interchangeable.
Now, by the definition of the RND, $Z\mathbb{P}_1(d\omega) = \mathbb{P}_2(d\omega)$, and so we are able to write this expression as

$$I_1(\theta) = -\frac{\partial}{\partial\theta}\int_\Omega \mathbb{P}_2(d\omega)$$

But since $\mathbb{P}_2$ is a probability measure, it must be that $\int_\Omega \mathbb{P}_2(d\omega) = 1$. Hence we find

$$I_1(\theta) = -\frac{\partial}{\partial\theta}(1) = 0$$

Which is what we wanted to show.

(c)

$$\mathcal{I}_2(\theta) = \mathbb{E}^{\mathbb{P}_2}\left[\left(\frac{\partial}{\partial\theta}\log\left(\frac{d\mathbb{P}_2}{d\mathbb{P}_1}(\omega;\theta)\right)\right)^2\right] \geq 0$$

This is known as the *Fisher information.*

Solution: By the definition of $\mathcal{I}_2(\theta)$, we have

$$\mathcal{I}_2(\theta) = -\mathbb{E}^{\mathbb{P}_2}\left[\frac{\partial^2}{\partial\theta^2}\log\left(\frac{d\mathbb{P}_2}{d\mathbb{P}_1}\right)\right] = -\mathbb{E}^{\mathbb{P}_2}\left[\log Z\right]''$$

where primes indicate differentiation w.r.t. θ. We begin by evaluating the double derivative of $\log Z$.

$$\mathcal{I}_2(\theta) = -\mathbb{E}^{\mathbb{P}_2}\left[\frac{Z'}{Z}\right]' = -\mathbb{E}^{\mathbb{P}_2}\left[\frac{Z''}{Z} - \frac{Z'^2}{Z^2}\right] = -\mathbb{E}^{\mathbb{P}_2}\left[\frac{Z''}{Z}\right] + \mathbb{E}^{\mathbb{P}_2}\left[\frac{Z'^2}{Z^2}\right]$$

Let us focus on the first expectation on the rhs. Similar to (b), we can use the RND to change our expectation measure from $\mathbb{P}_2$ to $\mathbb{P}_1$, and then swap the order of expectation and differentiation.

$$\mathbb{E}^{\mathbb{P}_2}\left[\frac{Z''}{Z}\right] = \int_\Omega \frac{Z''}{Z}\mathbb{P}_2(d\omega) = \int_\Omega Z''\mathbb{P}_1(d\omega) = \frac{\partial^2}{\partial\theta^2}\int_\Omega Z\mathbb{P}_1(d\omega) = \frac{\partial^2}{\partial\theta^2}\int_\Omega \mathbb{P}_2(d\omega)$$

and since $\mathbb{P}_2$ is a probability measure, $\int_\Omega \mathbb{P}_2(d\omega) = 1$, and so

$$\mathbb{E}^{\mathbb{P}_2}\left[\frac{Z''}{Z}\right] = \frac{\partial^2}{\partial\theta^2}(1) = 0$$

And so, returning to our above expression for $\mathcal{I}_2(\theta)$, we are left with only one term, which is

$$\mathcal{I}_2(\theta) = \mathbb{E}^{\mathbb{P}_2}\left[\frac{Z'^2}{Z^2}\right] = \mathbb{E}^{\mathbb{P}_2}\left[\left(\frac{Z'}{Z}\right)^2\right] = \mathbb{E}^{\mathbb{P}_2}\left[\left(\frac{\partial}{\partial\theta}\log Z\right)^2\right]$$

And so we find our result

$$\mathcal{I}_2(\theta) = \mathbb{E}^{\mathbb{P}_2}\left[\left(\frac{\partial}{\partial\theta}\log\left(\frac{d\mathbb{P}_2}{d\mathbb{P}_1}(\omega;\theta)\right)\right)^2\right] \geq 0$$

4. The Legendre-Fenchel transform (MLN §3.5) is given by

$$\Lambda^*(x) = \sup_{t\in\mathbb{R}}\{xt - \Lambda(t)\}$$

for $x \in \mathbb{R}$.

Assuming that the $\Lambda(t)$ is strictly convex and twice differentiable, then the supremum in the equation is given by

$$\Lambda^*(x) = \left\{ \begin{array}{c} \Lambda^*(t) = x(t)t - \Lambda(t) \\ x(t) = \Lambda'(t) \end{array} \right.$$

This gives the function $\Lambda^*(x)$ in a parametric form in terms of t as a continuous parameter. Show that this equation implies the following:

(a) $\Lambda^*(x)$ is also convex.

Solution: We note that since $\Lambda(t)$ is strictly convex, the function $x(t) = \Lambda'(t)$ must be strictly monotonically increasing, and hence one-to-one and invertible. We may therefore consider the function $t(x)$, and write

$$\Lambda^*(x) = xt(x) - \Lambda(t(x))$$

Differentiating this once gives results in

$$\Lambda^{*\prime}(x) = t(x) + xt'(x) - \Lambda'(t)t'(x)$$

and since $\Lambda'(t) = x$, this is simply

$$\Lambda^{*\prime}(x) = t(x)$$

It is a well-known result that the inverse of a strictly monotonically increasing function is also strictly monotonically increasing, which means that

$$\Lambda^{*\prime\prime}(x) = t'(x) > 0$$

Since the second derivative of $\Lambda^*(x)$ is strictly positive definite, we conclude that $\Lambda^*(x)$ is strictly convex.

(b) An inverse, dual relation

$$\Lambda(t) = \sup_{x\in\mathbb{R}}\{tx - \Lambda^*(x)\}$$

Solution: Since we have shown that $\Lambda^*(x)$ is also strictly convex, we may again compute the supremum in this equation by differentiation, just as above, to give a parameterized solution for $\Lambda(t)$.

$$\Lambda(t) = tx - \Lambda^*(x)$$

where $t(x) = \Lambda^{*\prime}(x)$. But we already know that

$$\Lambda^*(x) = xt - \Lambda(t)$$

for $x(t) = \Lambda'(t)$. Plugging this directly into the above expression for $\Lambda(t)$ gives us

$$\Lambda(t) = tx - xt + \Lambda(t) = \Lambda(t)$$

Hence we conclude that for convex, twice differentiable functions, the Legendre-Fenchel transform is involutive, and that $\Lambda(t)$ and $\Lambda^*(x)$ are dual.

(c) With the pair of convex functions $\Lambda(x)$ and $\Lambda^*(t)$ defined above, show that for any real x and t,

$$\Lambda(t) + \Lambda^*(x) - tx \geq 0$$

What is the condition for equality to hold?

Solution: We can rearrange this expression to the form

$$\Lambda(t) + \Lambda^*(x) - tx = \Lambda(t) - (tx - \Lambda^*(x))$$

We now use the dual relation introduced in (b) to write $\Lambda(t)$ in terms of $\Lambda^*(x)$. This results in the following expression:

$$\sup_{x\in\mathbb{R}}\{tx - \Lambda^*(x)\} - (tx - \Lambda^*(x)) \geq 0$$

This expression is clearly positive definite by definition of supremum. Equality occurs when

$$tx - \Lambda^*(x) = \sup_{x\in\mathbb{R}}\{tx - \Lambda^*(x)\}$$

which we have already seen occurs when $t = \Lambda^{*\prime}(x)$. We could have instead chosen to write $\Lambda^*(x)$ in terms of $\Lambda(t)$, which would have resulted in the equality condition being $x = \Lambda'(t)$. Due to the convexity of $\Lambda(t)$ and $\Lambda^*(x)$, and their resulting duality under the Legendre-Fenchel transform, these two conditions are equivalent. That is

$$t = \Lambda^{*\prime}(x) \iff x = \Lambda'(t) \implies \Lambda(t) + \Lambda^*(x) - tx = 0$$

5. For $\Omega = \{1, 2, \ldots, n\}$ and two probability measures $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_n)$ and $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, where $\nu_i, p_i > 0$, the Shannon relative entropy is given by

$$H[\boldsymbol{\nu} \parallel \mathbf{p}] = \sum_{i=1}^{n} \nu_i \ln\left(\frac{\nu_i}{p_i}\right)$$

(a) Show that for any two probability measures $\boldsymbol{\nu} >> 0$ and $\mathbf{p} >> 0$

$$H[\boldsymbol{\nu} \parallel \mathbf{p}] \geq 0$$

This is known as Jensen's inequality.

Solution: We have

$$H[\boldsymbol{\nu} \parallel \mathbf{p}] = \sum_{i=1}^{n} \nu_i \ln\left(\frac{\nu_i}{p_i}\right) = \sum_{i=1}^{n} \nu_i \ln \nu_i - \sum_{i=1}^{n} \nu_i \ln p_i$$

And so

$$H[\boldsymbol{\nu} \parallel \mathbf{p}] \geq 0 \iff \sum_{i=1}^{n} \nu_i \ln \nu_i \geq \sum_{i=1}^{n} \nu_i \ln p_i$$

To show that this is true for any two probability measures $\boldsymbol{\nu}, \mathbf{p} >> 0$ we can use the method of Lagrange multipliers to find the maximal value of $\sum_{i=1}^{n} \nu_i \ln p_i$ with respect to $\mathbf{p}$ subject to the constraint $\sum_{i=1}^{n} p_i = 1$.

We begin by considering the Lagrangian function

$$\mathcal{L}(p_1, \ldots, p_n, \lambda) = \sum_{i=1}^{n} \nu_i \ln p_i + \lambda(1 - \sum_{i=1}^{n} p_i)$$

We would like to find $\mathbf{p}, \lambda$ such that $\nabla_{p_1,\ldots,p_n,\lambda}\mathcal{L} = 0$. For the components p_i we have

$$\mathcal{L}_{p_i} = \frac{\nu_i}{p_i} - \lambda = 0 \Rightarrow p_i = \frac{\nu_i}{\lambda}$$

While for λ we have

$$\mathcal{L}_{\lambda} = 1 - \sum_{i=1}^{n} p_i = 0 \Rightarrow \sum_{i=1}^{n} p_i = 1$$

Solving for λ, and using the fact that $\sum \nu_i = 1$, we find that critical points of $\mathcal{L}$ occur when

$$\sum_{i=1}^{n} \frac{\nu_i}{\lambda} = 1 \Rightarrow \frac{1}{\lambda} \sum_{i=1}^{n} \nu_i = 1 \Rightarrow \frac{1}{\lambda} = 1 \Rightarrow \lambda = 1$$

This means that the probability measure $\mathbf{p}$ which maximizes the quantity $\sum_{i=1}^{n} \nu_i \log p_i$ is in fact $\mathbf{p} = \boldsymbol{\nu}$. That is,

$$\max_{\sum p_i = 1} \sum_{i=1}^{n} \nu_i \ln p_i = \sum_{i=1}^{n} \nu_i \ln \nu_i$$

And from this it follows that, for any two probability measures $\boldsymbol{\nu}, \mathbf{p} >> 0$,

$$\sum_{i=1}^{n} \nu_i \ln \nu_i \geq \sum_{i=1}^{n} \nu_i \ln p_i \iff H[\boldsymbol{\nu} \parallel \mathbf{p}] \geq 0$$

We conclude that the Shannon relative entropy of any two probability measures $\boldsymbol{\nu}, \mathbf{p} >> 0$ is positive semi-definite, with equality when $\mathbf{p} = \boldsymbol{\nu}$.

(b) Show that the Legendre-Fenchel transform of $H[\boldsymbol{\nu}||\mathbf{p}]$ is given by

$$\sup_{\boldsymbol{\nu}>>0} \left\{ \sum_{i=1}^{n} \epsilon_i \nu_i - H[\boldsymbol{\nu}||\mathbf{p}] \right\} = \ln \sum_{i=1}^{n} p_i e^{\epsilon_i}$$

in which $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ is the conjugate variable to $\boldsymbol{\nu}$.

Solution: We will compute the Legendre-Fenchel transform of $H[\boldsymbol{\nu}||\mathbf{p}]$ using the method of Lagrange multipliers once again. We seek a $\boldsymbol{\nu}$ which maximizes the quantity

$$\sum_{i=1}^{n} \epsilon_i \nu_i - H[\boldsymbol{\nu}||\mathbf{p}] = \sum_{i=1}^{n} \epsilon_i \nu_i - \sum_{i=1}^{n} \nu_i \ln\left(\frac{\nu_i}{p_i}\right)$$

subject to the constraint $\sum \nu_i = 1$.

We begin by constructing the Lagrangian function

$$\mathcal{L}(\nu_1, \ldots, \nu_n, \lambda) = \sum_{i=1}^{n} \epsilon_i \nu_i - \sum_{i=1}^{n} \nu_i \ln\left(\frac{\nu_i}{p_i}\right) + \lambda \left(1 - \sum_{i=1}^{n} \nu_i\right)$$

We would like to find $\boldsymbol{\nu}, \lambda$ such that $\nabla_{\nu_1,\ldots,\nu_n,\lambda}\mathcal{L} = 0$. For the components ν_i we have

$$\mathcal{L}_{\nu_i} = \epsilon_i - \ln\left(\frac{\nu_i}{p_i}\right) - 1 - \lambda = 0 \Rightarrow \nu_i = p_i e^{\epsilon_i - 1 - \lambda}$$

While for λ we have

$$\mathcal{L}_\lambda = 1 - \sum_{i=1}^{n} \nu_i = 1 - \sum_{i=1}^{n} p_i e^{\epsilon_i - 1 - \lambda} = 0$$

$$\Rightarrow e^{1+\lambda} = \sum_{i=1}^{n} p_i e^{\epsilon_i}$$

$$\Rightarrow \lambda = \ln\left(\sum_{i=1}^{n} p_i e^{\epsilon_i}\right) - 1$$

Plugging this value into our prior expression for ν_i, we find that the ν_i which maximizes our expression is

$$\nu_i = p_i e^{\epsilon_i - 1 - \lambda} = \frac{p_i e^{\epsilon_i}}{\sum_i p_i e^{\epsilon_i}} = A p_i e^{\epsilon_i}$$

where $A^{-1} = \sum_i p_i e^{\epsilon_i}$ is a normalization factor.

Plugging this value into our above expression for the Legendre-Fenchel transform of $H[\boldsymbol{\nu}||\mathbf{p}]$ gives us

$$\begin{aligned} H^*[\boldsymbol{\epsilon}||\mathbf{p}] &= \sup_{\boldsymbol{\nu}>>0}\left\{\sum_{i=1}^{n}\epsilon_i\nu_i - \sum_{i=1}^{n}\nu_i\ln\left(\frac{\nu_i}{p_i}\right)\right\} \\ &= \sum_{i=1}^{n}\epsilon_i\nu_i - \sum_{i=1}^{n}\nu_i\ln\left(Ae^{\epsilon_i}\right) \\ &= -\ln A\sum_{i=1}^{n}\nu_i \\ &= \ln A^{-1} = \ln\sum_{i=1}^{n}p_ie^{\epsilon_i} \end{aligned}$$

Which is what we wanted to show.

6. Let Ω be a simply connected compact domain in $\mathbb{R}^m$. Consider the statistical mechanical energy function $E(\mathbf{x})$ for $x \in \Omega$, and the sequence of probability measures whose density functions w.r.t. the Lebesgue measure is

$$f^{(n)}(\mathbf{x}) = A_ne^{-nE(\mathbf{x})}$$

where A_n is the normalization factor

$$A_n^{-1} = \int_{\Omega}e^{-nE(\mathbf{x})}d\mathbf{x}$$

which is assumed to satisfy

$$\lim_{n\to\infty}\frac{\log A_n}{n} = 0$$

Note that this is the Bolzmann-Gibbs distribution where n takes the role of inverse temperature β.

We now add some additional structure by letting $\mathbf{x} = (x_1, \mathbf{y})$, where $x_1 \in \mathbb{R}$ and $\mathbf{y} = (x_2, \ldots, x_m) \in \mathbb{R}^{m-1}$. Now let $f_1^{(n)}(x)$ be the marginal distribution

$$f_1^{(n)}(x) = A_n\int_{\Omega\cap\mathbb{R}^{m-1}}e^{nE(x,\mathbf{y})}d\mathbf{y}.$$

Where we assume

$$\lim_{n\to\infty}\frac{1}{n}\log f_1^{(n)}(x) = -\Lambda^*(x)$$

(a) Show that the n-scaled cumulant generating function

$$\log\int f_1^{(n)}(x)e^{ntx}dx$$

has the limit

$$\lim_{n\to\infty}\frac{1}{n}\log\int f_1^{(n)}(x)e^{ntx}dx = \max_{x\in\mathbb{R}}\{tx-\Lambda^*(x)\}$$

Assume all functions are sufficiently smooth to allow one to freely exchange limits and integration w.r.t x.

Solution: We have

$$\lim_{n\to\infty}\frac{1}{n}\log\int f_1^{(n)}(x)e^{ntx}dx = \lim_{n\to\infty}\frac{1}{n}\log\int f_1^{(n)}(x)e^{ntx}dx$$

We have assumed smoothness and

$$\lim_{n\to\infty}\frac{1}{n}\log f_1^{(n)}(x) = -\Lambda^*(x) \Rightarrow \lim_{n\to\infty} f_1^{(n)}(x) = e^{-n\Lambda^*(x)}$$

and so we may write the rhs as

$$\lim_{n\to\infty}\frac{1}{n}\log\int f_1^{(n)}(x)e^{ntx}dx = \lim_{n\to\infty}\frac{1}{n}\log\int e^{-n\Lambda^*(x)}e^{ntx}dx$$

Now, from Laplace's method for integrals we have the result (Qian §1.28)

$$\lim_{n\to\infty}\frac{1}{n}\log\int_a^b g(x)e^{nh(x)}dx = \max_{x\in[a,b]} h(x)$$

and so by letting $h(x) = tx - \Lambda^*(x)$ and $a, b \to \mp\infty$, we can use this identity to rewrite the rhs of the above equation as

$$\lim_{n\to\infty}\frac{1}{n}\log\int f_1^{(n)}(x)e^{ntx}dx = \max_{x\in\mathbb{R}}\{tx-\Lambda^*(x)\}$$

which is what we wanted to show.

(b) Denoting

$$\Lambda(t) = \max_{x\in\mathbb{R}}\{tx-\Lambda^*(x)\}$$

show that one can obtain $\Lambda^*(x)$ parametrically as

$$\Lambda^*(x) = \begin{cases} \Lambda^*(t) = -\frac{d}{d(1/t)}\left(\frac{\Lambda(t)}{t}\right) \\ x(t) = \Lambda'(t) \end{cases}$$

Solution: Let $u = 1/t$. Then by the chain rule

$$-\frac{d}{du}\left(\frac{\Lambda(t)}{t}\right) = t^2\frac{d}{dt}\left(\frac{\Lambda(t)}{t}\right) = t^2\left(\frac{\Lambda'(t)}{t}-\frac{\Lambda(t)}{t^2}\right) = t\Lambda'(t)-\Lambda(t)$$

And letting $x(t) = \Lambda'(t)$, we have

$$-\frac{d}{du}\left(\frac{\Lambda(t)}{t}\right) = tx(t)-\Lambda(t)$$

Hence we recover our original definition for $\Lambda^*(x)$:

$$\Lambda^*(x) = \left\{ \begin{array}{c} \Lambda^*(t) = tx(t) - \Lambda(t) \\ x(t) = \Lambda'(t) \end{array} \right.$$

7. Let $\mathbf{G} = (q_{kl})_{K\times K}$ be the infinitesimal generator, or transition probability rate, for a continuous-time Markov chain $\mathbf{X}_t \in \mathcal{S} = \{1, 2, \ldots, K\}$:

$$\Pr\{\mathbf{X}_{t+dt} = l|\mathbf{X}_t = k\} = \left\{ \begin{array}{cc} q_{kl}dt & l \neq k \\ 1 - \sum_{m\neq k} q_{km}dt & l = k \end{array} \right.$$

Let us assume that $\mathbf{G}$ has rank $K-1$ and is diagonalizable with eigenvalues $\lambda_1 = 0, \lambda_2, \ldots, \lambda_K$. It can be shown using the Perron-Frobenius theorem that the real part of λ_k is negative for all $k \geq 2$.

(a) Let τ be a fixed time interval. Show that

$$\mathbf{P} = e^{\tau\mathbf{G}} := \sum_{j=0}^{\infty} \frac{1}{j!}(\tau\mathbf{G})^j$$

is a transition probability matrix for a Markov chain, which has the same invariant probability as $\mathbf{G}$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$.

Solution: Let

$$\mathbf{G} = \mathbf{U}^{-1}\boldsymbol{\Lambda}\mathbf{U}$$

be an eigenvector decomposition of $\mathbf{G}$, and let $\boldsymbol{\nu}_i = \{\nu_1, \ldots, \nu_K)$ denote the ith left-eigenvector of $\mathbf{G}$. Then

$$\mathbf{P} = \sum_{j=0}^{\infty} \frac{1}{j!}(\tau\mathbf{U}^{-1}\boldsymbol{\Lambda}\mathbf{U})^j = \mathbf{U}^{-1}\left(\sum_{j=0}^{\infty} \frac{1}{j!}(\tau\boldsymbol{\Lambda})^j\right)\mathbf{U} = \mathbf{U}^{-1}e^{\tau\boldsymbol{\Lambda}}\mathbf{U}$$

It follows that $\mathbf{P}$ shares the same set of eigenvectors $\{\boldsymbol{\nu}_i\}$ with $\mathbf{G}$, but with eigenvalues $e^{\tau\lambda_i} = \{1, e^{\tau\lambda_2}, \ldots, e^{\tau\lambda_K}\}$. The first of these eigenvectors is the invariant probability $\boldsymbol{\pi}$, since $\lambda_1 = 0$ and $e^{\tau\lambda_1} = 1$, hence we have

$$\boldsymbol{\pi}\mathbf{G} = \mathbf{0} \iff \boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$$

(b) Let $\boldsymbol{\xi}(i)$ be a real-valued function for $i \in \mathcal{S}$. Let $\mathbf{X}_0, \mathbf{X}_\tau, \ldots, \mathbf{X}_{m\tau}, \ldots$ be the sample path of the discrete-time Markov chain. Now let

$$\overline{\boldsymbol{\xi}}^{(m)} = \frac{1}{m}\sum_{j=0}^{m-1} \boldsymbol{\xi}(X_{j\tau})$$

and show that

$$\lim_{m\to\infty} \mathbb{E}\left[\bar{\boldsymbol{\xi}}^{(m)}\right] = \sum_{k=1}^{K} \pi_k \boldsymbol{\xi}(k)$$

Solution: We have

$$\mathbb{E}\left[\bar{\boldsymbol{\xi}}^{(m)}\right] = \mathbb{E}\left[\frac{1}{m}\sum_{j=0}^{m-1} \boldsymbol{\xi}(X_{j\tau})\right] = \frac{1}{m}\sum_{j=0}^{m-1} \mathbb{E}[\boldsymbol{\xi}(X_{j\tau})]$$