CS281 Section 1: Math Review

Scott Linderman

1. Probability

   The "probability of an event" is more subtle than one might think. Is it the number of times that event would occur in repeated experiments; or is it a more abstract notion of our uncertainty about that event? These questions are at the heart of the *Bayesian* v.s. *frequestist* debate. We will not dwell on this here, but Murphy gives a nice overview of the philosophical differences between the two approaches and points to more detailed references on the subject. Instead, we'll begin with a review of basic probability definitions.

   (a) PMFs and PDFs

   Consider a random variable X which takes on states from a finite or countably infinite *state space* $\mathcal{X}$. We denote the probability of the event $X = x$ by $p(X = x)$ or $p(x)$ for short. The function p is a *probability mass function* and must satisfy the requirements

   $$0 \leq p(x) \leq 1 \forall x \in \mathcal{X}, \tag{1}$$

   $$\sum_{x \in \mathcal{X}} p(x) = 1. \tag{2}$$

   If we instead have continuous state space, for example $\mathcal{X} = \mathbb{R}$, then it does not make sense to talk about the probability of an individual state $x \in \mathcal{X}$ because there are uncountable infinitely many states, each with proability zero. Instead we talk about the probability that a random variable takes on a value in an interval. Define the *cumulative distribution function* $F(x) = p(X \leq x)$. This must be monotonically non-decreasing. Then we have,

   $$p(a < x \leq b) = F(b) - F(a). \tag{3}$$

   Finally, we define the *probability density function* $f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x)$ such that

   $$p(a < x \leq b) = \int_a^b f(x) \mathrm{d}x. \tag{4}$$

   (b) Conditional Probabilities

   We will often talk about the probability of an event $X = x$ given that event $Y = y$ occurred. This is written as

   $$p(x \,|\, y) = \frac{p(x, y)}{p(y)}, \tag{5}$$

   and is only defined if $p(y) > 0$.

There are two fundamental rule sof probability that we will use over and over again, namely the *sum rule* and the *product rule*.

The sum rule tells us how to arrive at a *marginal distribution* from a *joint distribution*. Namely,

$$p(x) = \sum_y p(x, y). \tag{6}$$

The product rule tells us how to decompose a joint distribution into the product of a marginal distribution and a joint distribution:

$$p(x, y) = p(x \mid y)p(y). \tag{7}$$

This can be combined with the definition of conditional probability to yield *Bayes' Rule*

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} \tag{8}$$

$$= \frac{p(y \mid x)p(x)}{\sum_x p(y, x)}. \tag{9}$$

It is helpful to think of these in more concrete terms. Suppose D is a random variable representing our observed data, and θ is a set of latent parameters that "caused" to that data, in the generative modeling context. We often wish to reason about the probability distribution of the parameters, θ after observing some data. Using Bayes' rule, we write this as:

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}. \tag{10}$$

We call $p(\theta \mid D)$ the *posterior distribution* over parameters. On the right hand side, $p(D \mid \theta)$ is the *likelihood* of the parameters after observing data when we are referring to it as a function of θ. Note that this is *not* a distribution over θ in that it does not integrate to 1! Finally, $p(\theta)$ is the prior distribution over the parameters.

(c) Expectation

Distributions are often characterized by their *moments*, for example their mean and variance. The mean, or expected value, is defined as

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} xp(x), \tag{11}$$

where X is a discrete random variable taking on values from state space $\mathcal{X}$. The variance is a measure of the "spread" of a distribution, and is defined as

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] \tag{12}$$

$$= \sum_{x \in \mathcal{X}} (x - \mathbb{E}[X])^2 p(x). \tag{13}$$

(d) Important Distributions

Murphy gives a nice overview of common distributions, but I would also recommend the Appendix B of the Bishop book, as well as Chapter 2 of Prof. Blitzein and Prof. Morris's forthcoming book, "Probability for Statistical Science" if you can get your hands on it.

2. Estimating synaptic strength

Suppose you are working in an experimental neuroscience lab where you are measuring the strength of a synaptic connection between two neurons. When the presynaptic neuron spikes, the voltage in the postsynaptic neuron changes in what is known as a "post-synaptic potential (PSP)." One measure of strength is the amplitude of the PSP. Your means of recording from the cells is via patch clamp, a method which gives you access to noisy measurements of the potential. You believe the noise is Gaussian distributed with zero mean and unknown variance σ^2 about the true mean, μ. You would like to infer the most likely values of μ and σ^2).

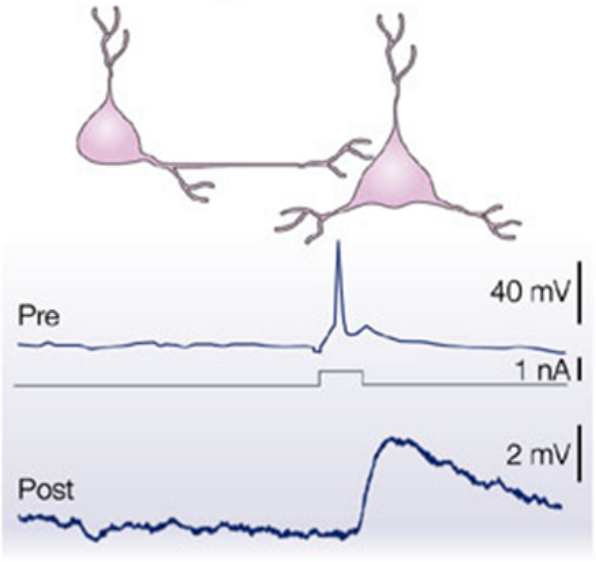

Figure 1: Example of a synaptic connection. The presynaptic neuron on the left makes an excitatory connection, i.e. a "synapse," onto the postsynaptic neuron on the right. When the presynaptic neuron spikes (top trace), the postsynaptic neuron exhibits a post-synaptic potential, or a brief increase in its voltage. Adapted from Debanne, Dominique. "Information processing in the axon." *Nature Reviews Neuroscience* 5.4 (2004): 304-316.

Suppose that your measurements of the PSP's are independent. The conditional probability of the measurements $\{w_n\}_{n=1}^N$ given the mean and variance is then

$$p(\{w_n\}_{n=1}^N \mid \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(w_n \mid \mu, \sigma^2). \tag{14}$$

As a function of α and σ^2, this is called the *likelihood* of the parameters.

Our goal is to find the most likely set of parameters (μ, σ^2) that maximize this likelihood function, that is, the *maximum likelihood estimate (MLE)* of the parameters. A common trick that we will make use of, both for ease of computing derivatives and gradients, as well as for numerical stability, is to *work with log-probabilities*. Since log is a monotonically increasing function, maximizing the log-likelihood is the same as maximizing the likelihood itself.

In this case,

$$\log p(\{w_n\}_{n=1}^N \mid \mu, \sigma^2) = \sum_{n=1}^N \mathcal{N}(w_n \mid \mu, \sigma^2) \tag{15}$$

$$= \sum_{n=1}^N -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(w_n - \mu)^2 \tag{16}$$

$$= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \sum_{n=1}^N \frac{1}{2\sigma^2}(w_n - \mu)^2. \tag{17}$$

Taking partial derivatives with respect to μ and setting to zero yields

$$\frac{\partial \log p(\{w_n\}_{n=1}^N \mid \mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{n=1}^N w_n - \mu) = 0 \tag{18}$$

$$N\mu = \sum_{n=1}^N w_n \tag{19}$$

$$\mu_{MLE} = \frac{1}{N}\sum_{n=1}^N w_n. \tag{20}$$

Doing the same for σ^2 yields

$$\frac{\partial \log p(\{w_n\}_{n=1}^N \mid \mu, \sigma^2)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \sum_{n=1}^N \frac{1}{2(\sigma^2)^2}(w_n - \mu)^2 = 0 \tag{21}$$

$$0 = -\frac{1}{2\sigma^2}\left[-N + \frac{1}{\sigma^2}\sum_{n=1}^N (w_n - \mu)^2\right] \tag{22}$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{n=1}^N (w_n - \mu)^2. \tag{23}$$

The maximum likelihood parameter estimates $(\mu_{MLE}, \sigma^2_{MLE})$ must simultaneously satisfy equations 20 and 23, so we can plug the value of μ_{MLE} into equation 23.

3. Maximum likelihood estimate for the multivariate Gaussian

   Suppose that now we have access to considerably more high-tech recording methods that allow us to optically record the potential in various parts of the cell simultaneously. Due to the cell geometry, these measurements could have interesting covariance structure. Now we

would like to model the distribution of the potential in the various parts of the postsynaptic neuron. We'll model this with a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance Σ.

Our former approach can be generalized to the multivariate setting, but our partial derivatives will become gradients and we'll have to recal some facts from linear algebra.

Now we have,

$$p(\{\boldsymbol{w}_n\} \mid \boldsymbol{\mu}, \Sigma) = \prod_n (2\pi)^{-D/2} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{w}_n - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{w}_n - \boldsymbol{\mu}) \right\}. \tag{24}$$

Taking logs,

$$\log p(\{\boldsymbol{w}_n\} \mid \boldsymbol{\mu}, \Sigma) = \sum_n -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\boldsymbol{w}_n - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{w}_n - \boldsymbol{\mu}). \tag{25}$$

Instead of partial derivatives, in the multivariate case we take gradients, the multivariate equivalent of a derivative, with respect to the vector $\boldsymbol{\mu}$.

Aside: The *gradient* of a function $f(\boldsymbol{x}) = y$ that maps the vector input $\boldsymbol{x} \in \mathbb{R}^M$ to a scalar output in $y \in \mathbb{R}$ is defined as

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial y}{x_1} \\ \vdots \\ \frac{\partial y}{x_M} \end{bmatrix}. \tag{26}$$

If we have a function $F(\boldsymbol{x}) = \boldsymbol{y}$ where $\boldsymbol{x} \in \mathbb{R}^M$ and $\boldsymbol{y} \in \mathbb{R}^N$, we concatenate the transposed gradient vectors with respect to each output dimension into an $N \times M$ matrix known as the *Jacobian*:

$$\nabla_{\boldsymbol{x}} F(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial y_1}{x_1} & \cdots & \frac{\partial y_1}{x_M} \\ \vdots & & \vdots \\ \frac{\partial y_N}{x_1} & \cdots & \frac{\partial y_N}{x_M} \end{bmatrix}. \tag{27}$$

Now, returning to our multivariate normal problem, we see that we must compute the gradient of the likelihood with respect to μ and with respect to Σ. The mean is the easier one. First, we drop the dependence on terms which do not contain $\boldsymbol{\mu}$ to get

$$\log p(\{\boldsymbol{w}_n\} \mid \boldsymbol{\mu}, \Sigma) = \text{const.} + \sum_n -\frac{1}{2}(b\boldsymbol{w}_n^T \Sigma^{-1} \boldsymbol{w}_n - \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{w}_n - \boldsymbol{w}_n^T \Sigma^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu}) \tag{28}$$

$$= \text{const.} - \frac{1}{2}\sum_n -2\boldsymbol{w}_n^T \Sigma^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} \tag{29}$$

$$= \text{const.} - \frac{N}{2}\boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} + \sum_n \boldsymbol{w}_n^T \Sigma^{-1} \boldsymbol{\mu}, \tag{30}$$

where we have used the trick that the transpose of a scalar is still a scalar.

Then we take the gradient

$$\nabla_{\boldsymbol{\mu}} \log p(\{\boldsymbol{w}_n\} \mid \boldsymbol{\mu}, \Sigma) = -\frac{N}{2}(\Sigma^{-1} + \Sigma^{-T})\boldsymbol{\mu} + \sum_n \Sigma^{-T}\boldsymbol{w}_n = 0 \tag{31}$$

$$0 = -N\Sigma^{-1}\boldsymbol{\mu} + \Sigma^{-1}\sum_n \boldsymbol{w}_n \tag{32}$$

$$\boldsymbol{\mu}_{MLE} = \frac{1}{N}\sum_n \boldsymbol{w}_n. \tag{33}$$

Note that $\Sigma^{-T} = (\Sigma^{-1})^T$. We have repeatedly used the fact that since Σ is symmetric, Σ^{-1} is as well.

Computing the MLE covariance matrix Σ_{MLE} will require a few more useful tricks. First, we will look at the gradient with respect to the inverse covariance matrix $\Lambda = \Sigma^{-1}$. Second, we'll make use of the *trace trick*:

$$\boldsymbol{x}^T A \boldsymbol{x} = \mathrm{tr}(\boldsymbol{x}^T A \boldsymbol{x}) = \mathrm{tr}(A x x^T). \tag{34}$$

Thus,

$$\log p(\{\boldsymbol{w}_n\} \mid \boldsymbol{\mu}, \Lambda) = \text{const.} + \frac{N}{2}\log|\Lambda| - \frac{1}{2}\sum_n \mathrm{tr}\left[(\boldsymbol{w}_n - \boldsymbol{\mu})(\boldsymbol{w}_n - \boldsymbol{\mu})^T\Lambda\right] \tag{35}$$

$$= \text{const.} + \frac{N}{2}\log|\Lambda| - \frac{1}{2}\mathrm{tr}\left[\sum_n \left((\boldsymbol{w}_n - \boldsymbol{\mu})(\boldsymbol{w}_n - \boldsymbol{\mu})^T\right)\Lambda\right]. \tag{36}$$

Rather tha using the "nabla" notation, the common form is

$$\frac{\partial \log p(\{\boldsymbol{x}_n\} \mid \boldsymbol{\mu}, \Lambda)}{\partial \Lambda} = \frac{N}{2}\Lambda^{-T} - \frac{1}{2}\sum_n (\boldsymbol{w}_n - \boldsymbol{\mu})(\boldsymbol{w}_n - \boldsymbol{\mu})^T = 0 \tag{37}$$

$$\Lambda^{-T} = \Lambda^{-1} = \Sigma = \frac{1}{N}\sum_n (\boldsymbol{w}_n - \boldsymbol{\mu})(\boldsymbol{w}_n - \boldsymbol{\mu})^T \tag{38}$$

4. Bayesian inference given knowledge of voltage distribution*

   *This example was not covered in section. Bayesian inference will be covered in greater detail next week.

   Suppose we have prior knowledge that the postsynaptic potential varies smoothly as we travel along the neuron. For simplicity, suppose the cell is one dimensional, as if we are looking only one branch of the neuron. Furthermore, suppose we have centered the data such that the mean should be about zero. We discretize the branch into D equally spaced compartments, and model the mean potential in those compartments as a vector $\boldsymbol{\mu} \in \mathbb{R}^D$. Our knowledge of smoothness between adjacent comparments can be expressed as a prior distribution over the vector $\boldsymbol{\mu}$.

   In particular, suppose that we have the following model:

$$\mu_1 \sim \mathcal{N}(0, \eta^2) \tag{39}$$

$$\mu_d \mid \mu_{d-1} \sim N(\mu_{d-1}, \eta^2) \quad \text{for } d = 2 \ldots D. \tag{40}$$

Suppose ν^2 is known. We will also simplify the observation model for PSPs given the mean. It will still be a multivariate Gaussian, but we will assume a spherical Gaussian noise distribution with known variance σ^2:

$$p(\{\boldsymbol{w}_n\}_{n=1}^N \mid \boldsymbol{\mu}, \sigma^2) = \prod_{n=1}^N \mathcal{N}(\boldsymbol{w}_n \mid \boldsymbol{\mu}, \sigma^2 \mathbb{I}). \tag{41}$$

For simplicity, now assume σ^2 is known. Our goal is to infer the posterior distribution $p(\boldsymbol{\mu} \mid \{w_n\}, \sigma^2, \eta^2)$, and find the value of $\boldsymbol{\mu}$ that maximizes this posterior distribution, i.e. the *maximum a posteriori estimate*.

By Bayes' rule, this is

$$p(\boldsymbol{\mu} \mid \{w_n\}, \sigma^2, \nu^2) \propto p(\{w_n\} \mid \boldsymbol{\mu}, \sigma^2, \eta^2) p(\boldsymbol{\mu}). \tag{42}$$

Note that we skipped an application of the product rule for $p(\{w_n\}, \sigma^2, \nu^2 \mid \boldsymbol{\mu})$.

First we need to translate our prior knowledge about pairwise smoothness between adjacent entries in $\boldsymbol{\mu}$ into a joint distribution. By the product rule and conditional independencies of the model,

$$p(\boldsymbol{\mu}) = p\left(\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_D \end{bmatrix}\right) = p(\mu_1) \prod_{d=2}^D p(\mu_d \mid \mu_1, \ldots, \mu_{d-1}) \tag{43}$$

$$= p(\mu_1) \prod_{d=2}^D p(\mu_d \mid \mu_{d-1}). \tag{44}$$

Plugging in the normal distributions we get

$$\log p(\boldsymbol{\mu}) = \text{const.} \; - \frac{\mu_1^2}{2\eta^2} - \frac{(\mu_2 - \mu_1)^2}{2\eta^2} - \ldots - \frac{(\mu_D - \mu_{D-1})^2}{2\eta^2} \tag{45}$$

$$= \text{const.} \; - \frac{1}{2\eta^2} \left[\sum_{d=1}^{D-1} \left(2\mu_d^2 - 2\mu_{d+1}\mu_d\right) + \mu_D^2 \right] \tag{46}$$

$$= \text{const.} \; - \frac{1}{2\eta^2} \begin{bmatrix} \mu_1 & \ldots & \mu_D \end{bmatrix} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & & \ddots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_D \end{bmatrix} \tag{47}$$

We recognize this as a quadratic form, which implies that the prior over $\boldsymbol{\mu}$ is a multivariate normal.

$$p(\boldsymbol{\mu}) \sim \mathcal{N}(\boldsymbol{\mu} \mid \mathbf{0}, \eta^2 L^{-1}), \tag{48}$$

where L is the Laplacian matrix in equation 47.

Now our posterior distribution is

$$p(\boldsymbol{\mu} \mid \{\boldsymbol{w}_n\}_{n=1}^N, \sigma^2, \eta^2) \propto \prod_{n=1}^N \mathcal{N}(\boldsymbol{w}_n \mid \boldsymbol{\mu}, \sigma^2 \mathbb{I}) \mathcal{N}(\boldsymbol{\mu} \mid \mathbf{0}, \eta^2 L^{-1}). \tag{49}$$

In log space (neglecting constants with respect to $\boldsymbol{\mu}$) this is

$$\log p(\boldsymbol{\mu} \mid \{\boldsymbol{w}_n\}_{n=1}^N, \sigma^2, \eta^2) = \text{const.} + \sum_n -\frac{1}{2\sigma^2}(\boldsymbol{w}_n - \boldsymbol{\mu})^T(\boldsymbol{w}_n - \boldsymbol{\mu}) - \frac{1}{2\eta^2}\boldsymbol{\mu}^T L \boldsymbol{\mu} \tag{50}$$

$$= \text{const.} - \frac{1}{2\sigma^2}\sum_n \left(\boldsymbol{w}_n^T \boldsymbol{w}_n - 2\boldsymbol{w}_n^T \boldsymbol{\mu} + \boldsymbol{\mu}^T \boldsymbol{\mu}\right) - \frac{1}{2\eta^2}\boldsymbol{\mu}^T L \boldsymbol{\mu} \tag{51}$$

$$= -\frac{2}{2\sigma^2}\left(\sum_n \boldsymbol{w}_n\right)^T \boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^T \left(\frac{N}{\sigma^2}\mathbb{I} + \frac{1}{\eta^2}L\right)\boldsymbol{\mu}. \tag{52}$$

To put this in a multivariate Gaussian form we must complete the square. That is, we must equate

$$ax^2 + bx + c = -\frac{1}{2\sigma^2}(x - \mu)^2 \tag{53}$$

$$\implies \quad \sigma^2 = -\frac{1}{2a} \tag{54}$$

$$\mu = -\frac{b}{2a}. \tag{55}$$

In matrix form this is

$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_\mu, \Sigma_\mu) \tag{56}$$

$$\text{where} \qquad \Sigma_\mu = \left(\frac{N}{\sigma^2}\mathbb{I} + \frac{1}{\eta^2}L\right)^{-1}, \tag{57}$$

$$\boldsymbol{\mu}_\mu = \Sigma_\mu \left(-\frac{1}{\sigma^2}\sum_n \boldsymbol{w}_n\right). \tag{58}$$