Course Notes for
Advanced Probabilistic Machine Learning

John Paisley
Department of Electrical Engineering
Columbia University

Fall 2014

Abstract

These are lecture notes for the seminar *ELEN E9801 Topics in Signal Processing: "Advanced Probabilistic Machine Learning"* taught at Columbia University in Fall 2014. They are transcribed almost verbatim from the handwritten lecture notes, and so they preserve the original bulleted structure and are light on the exposition. Some lectures therefore also have a certain amount of reiterating in them, so some statements may be repeated a few times throughout the notes.

The purpose of these notes is to (1) have a cleaner version for the next time it's taught, and (2) make them public so they may be helpful to others. Since the exposition comes during class, often via student questions, these lecture notes may come across as too fragmented depending on the reader's preference. Still, I hope they will be useful to those not in the class who want a streamlined way to learn the material at a fairly rigorous level, but not yet at the hyper-rigorous level of many textbooks, which also mix the fundamental results with the fine details. I hope these notes can be a good primer towards that end. As with the handwritten lectures, this document does not contain any references.

The twelve lectures are split into two parts. The first eight deal with several stochastic processes fundamental to Bayesian nonparametrics: Poisson, gamma, Dirichlet and beta processes. The last four lectures deal with some advanced techniques for posterior inference in Bayesian models. Each lecture was between 2 and 2-1/2 hours long.

Contents

Part I

Topics in Bayesian Nonparametrics

Chapter 1

Poisson distribution and process, superposition and marking theorems

- The Poisson distribution is perhaps the fundamental discrete distribution and, along with the Gaussian distribution, one of the two fundamental distributions of probability.

Importance:	Poisson	→	discrete r.v.'s
	Gaussian	→	continuous r.v.'s

Definition: A random variable $X \in \{0, 1, 2, \dots\}$ is Poisson distributed with parameter $\lambda > 0$ if

$$P(X = n|\lambda) = \frac{\lambda^n}{n!}e^{-\lambda}, \tag{1.1}$$

denoted $X \sim \text{Pois}(\lambda)$.

Moments of Poisson

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} nP(X = n|\lambda) = \sum_{n=1}^{\infty} \frac{\lambda^n}{(n-1)!}e^{-\lambda} = \lambda \underbrace{\sum_{n=1}^{\infty} \frac{\lambda^{n-1}}{(n-1)!}e^{-\lambda}}_{=1} \tag{1.2}$$

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{n=1}^{\infty} n^2 P(X = n|\lambda) = \lambda \sum_{n=1}^{\infty} \frac{n\lambda^{n-1}}{(n-1)!}e^{-\lambda} = \lambda \sum_{n=0}^{\infty} \frac{(n+1)\lambda^n}{n!}e^{-\lambda} \\ &= \lambda(\mathbb{E}[X] + 1) = \lambda^2 + \lambda \end{aligned} \tag{1.3}$$

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda \tag{1.4}$$

Sums of Poisson r.v.'s (take 1)

- Sums of Poisson r.v.'s are also Poisson. Let $X_1 \sim \text{Pois}(\lambda_1)$ and $X_2 \sim \text{Pois}(\lambda_2)$. Then $X_1 + X_2 \sim \text{Pois}(\lambda_1 + \lambda_2)$.

Important interlude: Laplace transforms and sums of r.v.'s

- Laplace transforms give a very easy way to calculate the distribution of sums of r.v.'s (among other things).

Laplace transform

- Let $X \sim p(X)$ be a positive random variable and let $t > 0$. The Laplace transform of X is

$$\mathbb{E}[e^{-tX}] = \int_X e^{-tx} p(x)\, dx \qquad \text{(sums when appropriate)} \tag{1.5}$$

Important property

- There is a one-to-one mapping between $p(x)$ and $\mathbb{E}[e^{-tX}]$. That is, if $p(x)$ and $q(x)$ are two distributions and $\mathbb{E}_p[e^{-tX}] = \mathbb{E}_q[e^{-tX}]$, then $p(x) = q(x)$ for all x. (p and q are the same distribution)

Sums of r.v.'s

- Let $X_1 \overset{ind}{\sim} p(x)$, $X_2 \overset{ind}{\sim} q(x)$ and $Y = X_1 + X_2$. What is the distribution of Y?

- Approach: Take the Laplace transform of Y and see what happens.

$$\mathbb{E}e^{-tY} = \mathbb{E}e^{-t(X_1+X_2)} = \underbrace{\mathbb{E}[e^{-tX_1}e^{-tX_2}] = \mathbb{E}[e^{-tX_1}]\mathbb{E}[e^{-tX_2}]}_{\text{by independence of } X_1 \text{ and } X_2} \tag{1.6}$$

- So we can multiply the Laplace transforms of X_1 and X_2 and see if we recognize it.

Sums of Poisson r.v.'s (take 2)

- The Laplace transform of a Poisson random variable has a very important form that should be memorized.

$$\mathbb{E}e^{-tX} = \sum_{n=0}^{\infty} e^{-tn}\frac{\lambda^n}{n!}e^{-\lambda} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda e^{-t})^n}{n!} = e^{-\lambda}e^{\lambda e^{-t}} = e^{\lambda(e^{-t}-1)} \tag{1.7}$$

- Back to the problem: $X_1 \overset{ind}{\sim} \text{Pois}(\lambda_1)$, $X_2 \overset{ind}{\sim} \text{Pois}(\lambda_2)$, $Y = X_1 + X_2$.

$$\mathbb{E}e^{-tY} = \mathbb{E}[e^{-tX_1}]\mathbb{E}[e^{-tX_2}] = e^{\lambda_1(e^{-t}-1)}e^{\lambda_2(e^{-t}-1)} = e^{(\lambda_1+\lambda_2)(e^{-t}-1)} \tag{1.8}$$

  We recognize that the last term is the Laplace transform of a Pois$(\lambda_1 + \lambda_2)$ random variable. We can therefore conclude that $Y \sim \text{Pois}(\lambda_1 + \lambda_2)$.

- Another way of saying this is that, if we draw $Y_1 \sim \text{Pois}(\lambda_1 + \lambda_2)$ and $X_1 \sim \text{Pois}(\lambda_1)$ and $X_2 \sim \text{Pois}(\lambda_2)$ and define $Y_2 = X_1 + X_2$. Then Y_1 is equal to Y_2 *in distribution.* (i.e., they may not be equal, but they have the same distribution. We write this as $Y_1 \overset{d}{=} Y_2$.)

- The idea extends to sums of more than two. Let $X_i \sim \text{Pois}(\lambda_i)$. Then $\sum_i X_i \sim \text{Pois}(\sum_i \lambda_i)$ since

$$\mathbb{E}e^{-t\sum_i X_i} = \prod_i \mathbb{E}e^{-tX_i} = e^{(\sum_i \lambda_i)(e^{-t}-1)}. \tag{1.9}$$

A conjugate prior for λ

- What if we have $X_1, \dots, X_N$ that we believe to be generated by a Pois(λ) distribution, but we don't know λ?

- One answer: Put a prior distribution on λ, $p(\lambda)$, and calculate the posterior of λ using Bayes' rule.

Bayes' rule (review)

$$\begin{aligned} P(A,B) &= P(A|B)P(B) = P(B|A)P(A) && (1.10) \\ &\Downarrow && \\ P(A|B) &= \frac{P(B|A)P(A)}{P(B)} && (1.11) \\ \text{posterior} &= \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} && (1.12) \end{aligned}$$

Gamma prior

- Let $\lambda \sim$ Gam(a,b), where $p(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}$ is a gamma distribution. Then the posterior of λ is

$$\begin{aligned} p(\lambda|X_1,\dots,X_N) &\propto p(X_1,\dots,X_N|\lambda)p(\lambda) = \left[\prod_{i=1}^{N} \frac{\lambda^{X_i}}{X_i!}e^{-\lambda}\right] \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda} \\ &\propto \lambda^{a+\sum_{i=1}^N X_i - 1} e^{-(b+N)\lambda} && (1.13) \\ &\Downarrow \\ p(\lambda|X_1,\dots,X_N) &= \text{Gam}(a + \textstyle\sum_{i=1}^N X_i, b+N) && (1.14) \end{aligned}$$

Note that $\mathbb{E}[\lambda|X_1,\dots,X_N] = \frac{a+\sum_{i=1}^N X_i}{b+N} \approx$ Empirical average of X_i
(Makes sense because $\mathbb{E}[X|\lambda] = \lambda$)
$\mathbb{V}[\lambda|X_1,\dots,X_N] = \frac{a+\sum_{i=1}^N X_i}{(b+N)^2} \approx$ Empirical average$/N$
(Get more confident as we see more X_i)

- The gamma distribution is said to be the conjugate prior for the parameter of the Poisson distribution because the posterior is also gamma.

Poisson–Multinomial

- A sequence of Poisson r.v.'s is closely related to the multinomial distribution as follows:

  Let $X_i \overset{ind}{\sim}$ Pois(λ_i) and let $Y = \sum_{i=1}^N X_i$.

  Then what is the distribution of $\vec{X} = (X_1, \dots, X_N)$ given Y?

We can use basic rules of probability. . .

$$\begin{aligned} P(X_1,\ldots,X_N) &= P(X_1,\ldots,X_N, Y = \textstyle\sum_{i=1}^N X_i) \leftarrow (Y \text{ is a deterministic function of } X_{1:N}) \\ &= P(X_1,\ldots,X_N | Y = \textstyle\sum_{i=1}^N X_i) P(Y = \textstyle\sum_{i=1}^N X_i) \qquad (1.15) \end{aligned}$$

And so

$$P(X_1,\ldots,X_N | Y = \textstyle\sum_{i=1}^N X_i) = \frac{P(X_1,\ldots,X_N)}{P(Y=\sum_{i=1}^N X_i)} = \frac{\prod_i P(X_i)}{P(Y=\sum_{i=1}^N X_i)} \qquad (1.16)$$

- We know that $P(Y) = \text{Pois}(Y; \sum_{i=1}^N \lambda_i)$, so

$$\begin{aligned} P(X_1,\ldots,X_N | Y = \textstyle\sum_i X_i) &= \left[\prod_{i=1}^N \frac{\lambda_i^{X_i}}{X_i!} e^{-\lambda_i}\right] \Big/ \left[\frac{(\sum_{i=1}^N \lambda_i)^{\sum_{i=1}^N X_i}}{(\sum_{i=1}^N X_i)!} e^{-\sum_{i=1}^N \lambda_i}\right] \\ &= \frac{Y!}{X_1! \cdots X_N!} \prod_{i=1}^N \left(\frac{\lambda_i}{\sum_{j=1}^N \lambda_j}\right)^{X_i} \qquad (1.17) \\ &\quad \Downarrow \\ &\quad \text{Mult}(Y; p_1,\ldots,p_N), \quad p_i = \lambda_i / \textstyle\sum_j \lambda_j \end{aligned}$$

- What is this saying?

  1. Given the sum of N independent Poisson r.v.'s, the individual values are distributed as a multinomial using the normalized parameters.
  2. We can sample $X_1,\ldots,X_N$ in two *equivalent* ways
     a) Sample $X_i \sim \text{Pois}(\lambda_i)$ independently
     b) First sample $Y \sim \text{Pois}(\sum_j \lambda_j)$, then $(X_1,\ldots,X_N) \sim \text{Mult}\Big(Y; \frac{\lambda_1}{\sum_j \lambda_j}, \ldots, \frac{\lambda_N}{\sum_j \lambda_j}\Big)$

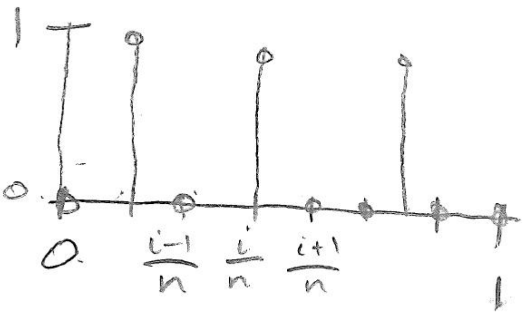
Poisson as a limiting case distribution

- The Poisson distribution arises as a limiting case of the sum over many binary events each having small probability.

Binomial distribution and Bernoulli process

- Imaging we have an array of random variables X_{nm}, where $X_{nm} \sim \text{Bern}\left(\frac{\lambda}{n}\right)$ for $m = 1,\ldots,n$ and fixed $0 \leq \lambda \leq n$. Let $Y_n = \sum_{m=1}^n X_{nm}$ and $Y = \lim_{n\to\infty} Y_n$. Then $Y_n \sim \text{Bin}\left(n, \frac{\lambda}{n}\right)$ and $Y \sim \text{Pois}(\lambda)$.

Picture: Have n coins with bias λ/n evenly spaced between $[0,1]$. Go down the line and flip each independently. In the limit $n\to\infty$, the total # of 1's is a Pois(λ) r.v.



Proof:

$$\lim_{n\to\infty} P(Y_n = k|\lambda) = \lim_{n\to\infty} \frac{n!}{(n-k)!k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \tag{1.18}$$

$$= \lim_{n\to\infty} \underbrace{\left[\frac{n(n-1)\cdots(n-k+1)}{n^k}\right]}_{\to 1} \underbrace{\left[\left(1-\frac{\lambda}{n}\right)^{-k}\right]}_{\to 1} \underbrace{\left[\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{n}\right]}_{\to \underbrace{\frac{\lambda^k}{k!}e^{-\lambda}}_{=\text{Pois}(k;\lambda)}}$$

So $\lim_{n\to\infty}$ Bin $\left(n, \frac{\lambda}{n}\right) =$ Pois(λ).

<u>A more general statement</u>

- Let λ_{nm} be an array of positive numbers such that
  1. $\sum_{m=1}^{n} \lambda_{nm} = \lambda < \infty$
  2. $\lambda_{nm} < 1$ and $\lim_{n\to\infty} \lambda_{nm} = 0$ for all m

Let $X_{nm} \sim$ Bern(λ_{nm}) for $m = 1, \ldots, n$. Let $Y_n = \sum_{m=1}^{n} X_{nm}$ and $Y = \lim_{n\to\infty} Y_n$. Then $Y \sim$ Pois(λ).

Proof: (use Laplace transform)

1. $\mathbb{E}\, e^{-tY} = \lim\limits_{n\to\infty} \mathbb{E}\, e^{-tY_n} = \lim\limits_{n\to\infty} \mathbb{E}\, e^{-t\sum_{m=1}^{n} X_{nm}} = \lim\limits_{n\to\infty} \prod\limits_{m=1}^{n} \mathbb{E}\, e^{-tX_{nm}}$
2. $\mathbb{E}\, e^{-tX_{nm}} = \lambda_{nm}e^{-t\cdot 1} + (1-\lambda_{nm})e^{-t\cdot 0} = 1 - \lambda_{nm}(1-e^{-t})$
3. So $\mathbb{E}\, e^{-tY} = \lim\limits_{n\to\infty} \prod\limits_{m=1}^{n} \big(1-\lambda_{nm}(1-e^{-t})\big) = \lim\limits_{n\to\infty} e^{\sum_{m=1}^{n} \ln(1-\lambda_{nm}(1-e^{-t}))}$
4. $\ln(1-\lambda_{nm}(1-e^{-t})) = -\sum\limits_{s=1}^{\infty} \frac{1}{s}\lambda_{nm}^{s}(1-e^{-t})^s \quad$ because $0 \le 1-\lambda_{nm}(1-e^{-t}) < 1$
5. $\sum\limits_{m=1}^{n} \ln(1-\lambda_{nm}(1-e^{-t})) = -\underbrace{\left(\sum\limits_{m=1}^{n} \lambda_{nm}\right)}_{=\lambda}(1-e^{-t}) - \underbrace{\sum\limits_{s=2}^{\infty} \frac{1}{s}\left(\sum\limits_{m=1}^{n} \lambda_{nm}^{s}\right)(1-e^{-t})^s}_{\to 0 \text{ as } n\to\infty}$

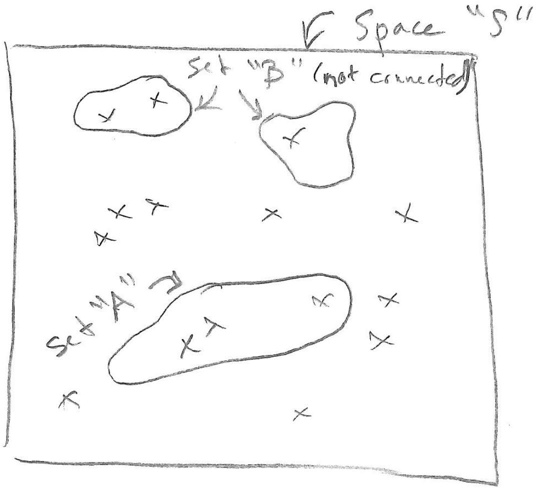
So $\mathbb{E}\, e^{-tY} = e^{-\lambda(1-e^{-t})}$. Therefore $Y \sim$ Pois(λ).

Poisson process

- In many ways, the Poisson process is no more complicated than the previous discussion on the Poisson distribution. In fact, the Poisson process can be thought of as a "structured" Poisson distribution, which should hopefully be more clear below.

Intuitions and notations

- S : a space (think $\mathbb{R}^d$ or part of $\mathbb{R}^d$)
- Π : a random countable subset of S (i.e., a random # of points and their locations)
- $A \subset S$: a subset of S
- $N(A)$: a counting measure. Counts how many points in Π fall in A (i.e., $N(A) = |\Pi \cap A|$)
- $\mu(\cdot)$: a measure on S
  $\hookrightarrow \mu(A) \geq 0$ for $|A| > 0$
  $\mu(\cdot)$ is non-atomic
  $\hookrightarrow \mu(A) \to 0$ as $|A| \to \emptyset$



Think of μ as a scaled probability distribution that is continuous so that $\mu(\{x\}) = 0$ for all points $x \in S$.

Poisson processes

- A Poisson process Π is a countable subset of S such that
  1. For $A \subset S$, $N(A) \sim \text{Pois}(\mu(A))$
  2. For disjoint sets $A_1, \ldots, A_k$, $N(A_1), \ldots, N(A_k)$ are independent Poisson random variables.

$N(\cdot)$ is called a "Poisson random measure". (See above for mapping from Π to $N(\cdot)$)

Some basic properties

- The most basic properties follow from the properties of a Poisson distribution.
  a) $\mathbb{E}N(A) = \mu(A) \to$ therefore μ is sometimes referred to as a "mean measure"
  b) If $A_1, \ldots, A_k$ are disjoint, then

$$N(\textstyle\bigcup_{i=1}^k A_i) = \sum_{i=1}^k N(A_i) \sim \text{Pois}\Big(\sum_{i=1}^k \mu(A_i)\Big) = \text{Pois}\Big(\mu(\bigcup_{i=1}^k A_i)\Big) \tag{1.19}$$

Since this holds for $k \to \infty$ and $\mu(A_i) \searrow 0$, $N(\cdot)$ is "infinitely divisible".

c) Let $A_1, \dots, A_k$ be disjoint subsets of S. Then

$$P(N(A_1) = n_1, \dots, N(A_k) = n_k | N(\bigcup_{i=1}^k A_i) = n) = \frac{P(N(A_1)=n_1,\dots,N(A_k)=n_k)}{P(N(\bigcup_{i=1}^k A_i)=n)} \tag{1.20}$$

Notice from earlier that $N(A_i) \Leftrightarrow X_i$. Following the same exact calculations,

$$P(N(A_1) = n_1, \dots, N(A_k) = n_k | N(\bigcup_{i=1}^k A_i) = n) = \frac{n!}{n_1! \cdots n_k!} \prod_{i=1}^k \left(\frac{\mu(A_i)}{\mu(\bigcup_{j=1}^k A_j)} \right)^{n_i} \tag{1.21}$$

Drawing from a Poisson process (break in the basic properties)

- Property (c) above gives a very simple way for drawing $\Pi \sim$ PP(μ), though some thought is required to see why.

  1. Draw the total number of points $N(S) \sim$ Pois($\mu(S)$).
  2. For $i = 1, \dots, N(S)$ draw $X_i \overset{iid}{\sim} \mu/\mu(S)$. In other words, normalize μ to get a probability distribution on S.
  3. Define $\Pi = \{X_1, \dots, X_{N(S)}\}$.

d) Final basic property (of these notes)

$$\mathbb{E}\, e^{-tN(A)} = e^{\mu(A)(e^{-t}-1)} \longrightarrow \text{ an obvious result since } N(A) \sim \text{Pois}(\mu(A)) \tag{1.22}$$

Some more advanced properties

Superposition Theorem: Let $\Pi_1, \Pi_2, \dots$ be a countable collection of independent Poisson processes with $\Pi_i \sim$ PP(μ_i). Let $\Pi = \bigcup_{i=1}^\infty \Pi_i$. Then $\Pi \sim$ PP(μ) with $\mu = \sum_{i=1}^\infty \mu_i$.

Proof: Remember from the definition of a Poisson process we have to show two things.

1. Let $N(A)$ be the PRM (Poisson random measure) associated with PP (Poisson process) Π and $N_i(A)$ with Π_i. Clearly $N(A) = \sum_{i=1}^\infty N_i(A)$, and since $N_i(A) \sim$ Pois($\mu_i(A)$) by definition, it follows that $N(A) \sim$ Pois($\sum_{i=1}^\infty \mu_i(A)$).
2. Let $A_1, \dots, A_k$ be disjoint. Then $N(A_1), \dots, N(A_k)$ are independent because $N_i(A_j)$ are independent for all i and j.

Restriction Theorem: If we restrict Π to a subset of S, we still have a Poisson process. Let $S_1 \subset S$ and $\Pi_i = \Pi \cap S_1$. Then $\Pi_1 \sim$ PP(μ_1), where $\mu_1(A) = \mu(S_1 \cap A)$. This can be thought of as setting $\mu = 0$ outside of S_1, or just looking at the subspace S_1 and ignoring the rest of S.

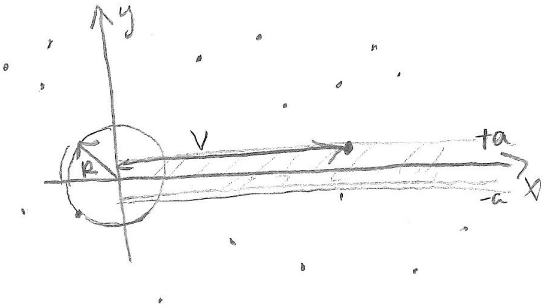
Mapping Theorem: This says that one-to-one function $y = f(x)$ preserve the Poisson process. That is, if $\Pi_x \sim$ PP(μ) and $\Pi_y = f(\Pi_x)$, then Π_y is also a Poisson process with the proper transformation made to μ. (See Kingman for details. We won't use this in this class.)

Example

Let N be a Poisson random measure on $\mathbb{R}^2$ with mean measure $\mu(A) = c|A|$.

$|A|$ is the area of A ("Lebesgue measure")

Intuition : Think trees in a forest.



Question 1: What is the distance R from the origin to the nearest point?

Answer: We know that $R > r$ if $N(B_r) = 0$, where B_r = ball of radius r. Since these are equivalent events, we know that

$$P(R > r) = P(N(B_r) = 0) = e^{-\mu(B_r)} = e^{-c\pi r^2}. \tag{1.23}$$

Question 2: Let each atom be the center of a disk of radius a. Take our line of sight as the x-axis. How far can we see?

Answer: The distance V is equivalent to the farthest we can extend a rectangle D_x with y-axis boundaries of $[-a, a]$. We know that $V > x$ if $N(D_x) = 0$. Therefore

$$P(V > x) = P(N(D_x) = 0) = e^{-\mu(D_x)} = e^{-2acx}. \tag{1.24}$$

Marked Poisson processes

- The other major theorem of these notes relates to "marking" the points of a Poisson process with a random variable.
- Let $\Pi \sim$ PP(μ). For each $x \in \Pi$, associate a r.v. $y \sim p(y|x)$. We say that y has "marked" x. The results is also a Poisson process.

Theorem: Let μ be a measure on space S and $p(y|x)$ a probability distribution on space M. For each $x \in \Pi \sim$ PP(μ) draw $y \sim p(y|x)$ and define $\Pi^* = \{(x_i, y_i)\}$. Then Π^* is a Poisson process on $S \times M$ with mean measure $\mu^* = \mu(dx)p(y|x)dy$.

Comment: If $N^*(C) = |\Pi^* \cap C|$ for $C \subset S \times M$, this says that $N^*(C) \sim$ Pois$(\mu^*(C))$, where $\mu^*(C) = \int_C \mu(dx)p(y|x)dy$.

Proof: Need to show that $\mathbb{E}e^{-tN^*(C)} = \exp\{\int_C (e^{-t}-1)\mu(dx)p(y|x)dy\}$

1. Note that $N^*(C) = \sum_{i=1}^{N(S)} \mathbb{1}\{(x_i, y_i) \in C\}$. $N(S)$ is PRM associated with $\Pi \sim$ PP(μ).
2. Recall tower property: $\mathbb{E}f(A,B) = \mathbb{E}[\mathbb{E}[f(A,B)|B]]$.
3. Therefore, $\mathbb{E}e^{-tN^*(C)} = \mathbb{E}\left[\mathbb{E}\left[\exp\left\{-t\sum_{i=1}^{N(S)} \mathbb{1}\{(x_i,y_i) \in C\}\right\}|\Pi\right]\right]$
4. Manipulating :

$$\begin{aligned}\mathbb{E}e^{-tN^*(C)} &= \mathbb{E}\left[\prod_{i=1}^{N(S)} \mathbb{E}[e^{-t\mathbb{1}\{(x_i,y_i)\in C\}}|\Pi]\right] \qquad (1.25)\\ &= \mathbb{E}\left[\prod_{i=1}^{N(S)} \int_M \left\{e^{-t\cdot 1}\mathbb{1}\{(x_i,y_i)\in C\} + e^{-t\cdot 0}\mathbb{1}\{(x_i,y_i)\notin C\}\right\} p(y_i|x_i)dy_i\right]\end{aligned}$$

5. Continuing :

$$\begin{aligned}\mathbb{E}e^{-tN^*(C)} &= \underbrace{\mathbb{E}\left[\prod_{i=1}^{N(S)} [1 + \int_M (e^{-t}-1)\mathbb{1}\{(x_i,y_i)\in C\}p(y_i|x_i)dy_i]\right]}_{\text{use } \mathbb{1}\{(x_i,y_i)\notin C\} = 1 - \mathbb{1}\{(x_i,y_i)\in C\}}\\ &= \underbrace{\mathbb{E}\left[\prod_{i=1}^{n}\left[1 + \int_S\int_M (e^{-t}-1)\mathbb{1}\{(x,y)\in C\}p(y|x)dy\frac{\mu(dx)}{\mu(S)}|N(S)=n\right]\right]}_{\text{Tower again using Poisson-multinomial representation}}\\ &= \underbrace{\mathbb{E}\left[\left((1 + \int_C (e^{-t}-1)p(y|x)dy\frac{\mu(dx)}{\mu(S)}\right)^{N(S)}\right]}_{N(S)\sim \text{Pois}(\mu(S))} \qquad (1.26)\end{aligned}$$

6. Recall that if $n \sim$ Pois(λ), then

$$\mathbb{E}z^n = \sum_{n=0}^{\infty} z^n \frac{\lambda^n}{n!}e^{-\lambda} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{(z\lambda)^n}{n!} = e^{\lambda(z-1)}.$$

Therefore, this last expectation shows that

$$\begin{aligned}\mathbb{E}e^{-tN^*(C)} &= \exp\left\{\mu(S)\int_C (e^{-t}-1)p(y|x)dy\frac{\mu(dx)}{\mu(S)}\right\}\\ &= \exp\left\{\int_C (e^{-t}-1)\mu(dx)p(y|x)dy\right\}, \qquad (1.27)\end{aligned}$$

thus $N^*(C) \sim$ Pois($\mu^*(C)$).

Example 1: Coloring

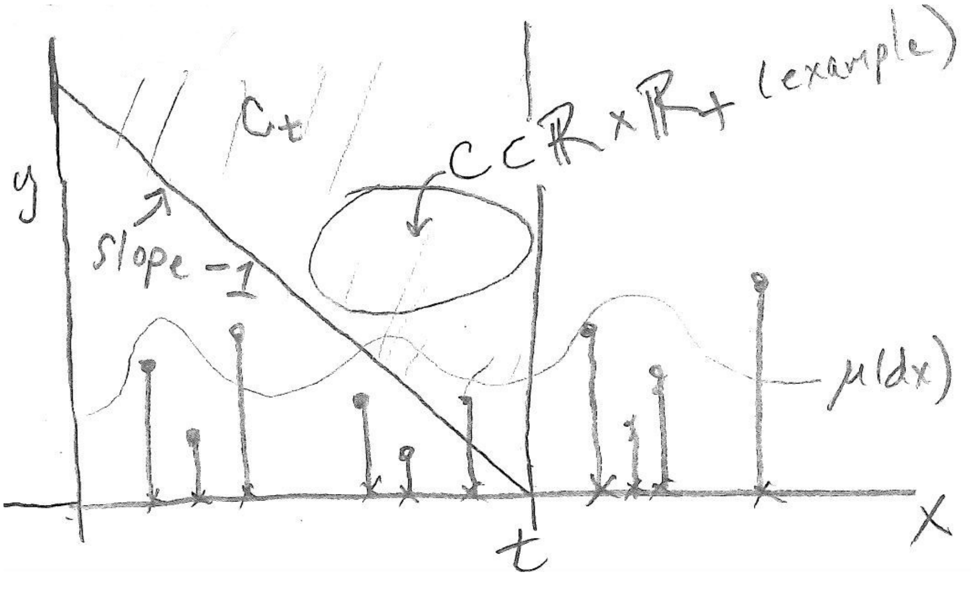
- Let $\Pi \sim \text{PP}(\mu)$ and let an $x \in \Pi$ be randomly colored from among K colors. Denote the color by y with $P(y=i)=p_i$. Then $\Pi^* = \{(x_i, y_i)\}$ is a PP on $S \times \{1, \dots, K\}$ with mean measure $\mu^*(dx \cdot \{y\}) = \mu(dx)\prod_{i=1}^K p_i^{\mathbb{1}(y=i)}$. If we want to restrict Π^* to the ith color, called Π_i^*, then we know that $\Pi_i^* \sim \text{PP}(p_i\mu)$. We can also restrict to two colors, etc.

Example 2: Using the extended space

- Image we have a 24-hour store and customers arrive according to a PP with mean measure μ on $\mathbb{R}$ (time). In this case, let $\mu(\mathbb{R}) = \infty$, but $\mu([a,b]) < \infty$ for finite a, b (which gives the expected number of arrivals between times a and b). Imagine a customer arriving at time x stays for duration $y \sim p(y|x)$. At time t, what can we say about the customers in the store?
  - $\Pi \sim \text{PP}(\mu)$ and $\Pi^* = \{(x_i, y_i)\}$ for $x_i \in \Pi$ is $\text{PP}(\mu(dx)p(y|x)dy)$ because it's a marked Poisson process.
  - We can construct the marked Poisson process like below. The counting measure $N^*(C) \sim \text{Pois}(\mu^*(C)),\ \mu^* = \mu(dx)p(y|x)dy$.
  - The points in C_t (below) are those that arrive before time t and are still there at time t. It follows that

$$N^*(C_t) \sim \text{Pois}\left(\int_0^t \mu(dx)\int_{t-x}^{\infty} p(y|x)dy\right).$$

    $N^*(C_t)$ is the number of customers in the store at time t.



Chapter 2

Completely random measures, Campbell's theorem, gamma process

Poisson process review

- Recall the definition of a Poisson random measure.

PRM definition: Let S be a space and μ a non-atomic (i.e., diffuse, continuous) measure on it (think a positive function). A random measure N on S is a PRM with mean measure μ if

a) For every subset $A \subset S$, $N(A) \sim \text{Pois}(\mu(A))$.

b) For disjoint sets $A_1, \ldots, A_k$, $N(A_1), \ldots, N(A_k)$ are independent r.v.'s

Poisson process definition: Let $X_1, \ldots, X_{N(S)}$ be the $N(S)$ points in N (a random number) of measure equal to one. Then the *point process* $\Pi = \{X_1, \ldots, X_{N(S)}\}$ is a Poisson process, denoted $\Pi \sim \text{PP}(\mu)$.

Recall that to draw this (when $\mu(S) < \infty$) we can

a) Draw $N(S) \sim \text{Pois}(\mu(S))$

b) For $i = 1, \ldots, N(S)$, draw $X_i \overset{iid}{\sim} \mu/\mu(S)$ ← normalize μ to get a probability measure.

Functions of Poisson processes

- Often models will take the form of a function of an underlying Poisson process: $\sum_{x\in\Pi} f(x)$.

Example (see Sec. 5.3 of Kingman): Imagine that star locations are distributed as $\Pi \sim \text{PP}(\mu)$. They're marked independently with a mass $m \sim p(m)$, giving a marked PP $\Pi^* \sim \text{PP}(\mu \times p\, dm)$. The gravitational field at, e.g., the origin is

$$\sum_{(x,m)\in\Pi^*} f((x,m)) = \sum_{(x,m)\in\Pi^*} \frac{Gm_x}{\|x\|_2^3} x \qquad (G \text{ is a constant from physics})$$

- We can analyze these sorts of problems using PP techniques
- We will be more interested in the context of "completely random measures" in this class.

Functions: The finite case

- Let $\Pi \sim$ PP(μ) and $|\Pi| < \infty$ (with probability 1). Let $f(x)$ be a positive function. Let $\mathcal{M} = \sum_{x\in\Pi} f(x)$. We calculate its Laplace transform (for $t < 0$):

$$\mathbb{E}e^{t\mathcal{M}} = \mathbb{E}e^{t\sum_{x\in\Pi} f(x)} = \mathbb{E}\left[\prod_{i=1}^{|\Pi|} e^{tf(x_i)}\right] \leftarrow \text{recall two things (below)} \tag{2.1}$$

Recall:

1. $|\Pi| = N(S)$ ← Poisson random measure for Π
2. Tower property $\mathbb{E}g(x,y) = \mathbb{E}[\mathbb{E}[g(x,y)|y]]$.

So:

$$\mathbb{E}\left[\prod_{i=1}^{N(S)} e^{tf(x_i)}\right] = \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^{N(S)} e^{tf(x_i)}|N(S)\right]\right] = \mathbb{E}\left[\mathbb{E}\left[e^{tf(x)}\right]^{N(S)}\right]. \tag{2.2}$$

Since $N(S) \sim$ Pois$(\mu(S))$, we use the last term to conclude

$$\begin{aligned}\mathbb{E}\left[\prod_{i=1}^{N(S)} e^{tf(x_i)}\right] &= \exp\{\mu(S)(\mathbb{E}[e^{tf(x)}] - 1)\} \\ &= \exp\int_S \mu(dx)(e^{tf(x)} - 1)\end{aligned}$$

$\left(\text{since } \mathbb{E}e^{tf(x)} = \int_S \frac{\mu(dx)}{\mu(S)} e^{tf(x)}\right)$ ↗ ↑

recall that $\mathbb{E}[(e^t)^{N(A)}] = \exp\int_A \mu(dx)(e^t - 1)$.

$(f(x) = 1$ in this case$)$

And so for functions $\mathcal{M} = \sum_{x\in\Pi} f(x)$ of Poisson processes Π with an almost sure finite number of points

$$\mathbb{E}e^{t\mathcal{M}} = \exp\int_S \mu(dx)(e^{tf(x)} - 1). \tag{2.3}$$

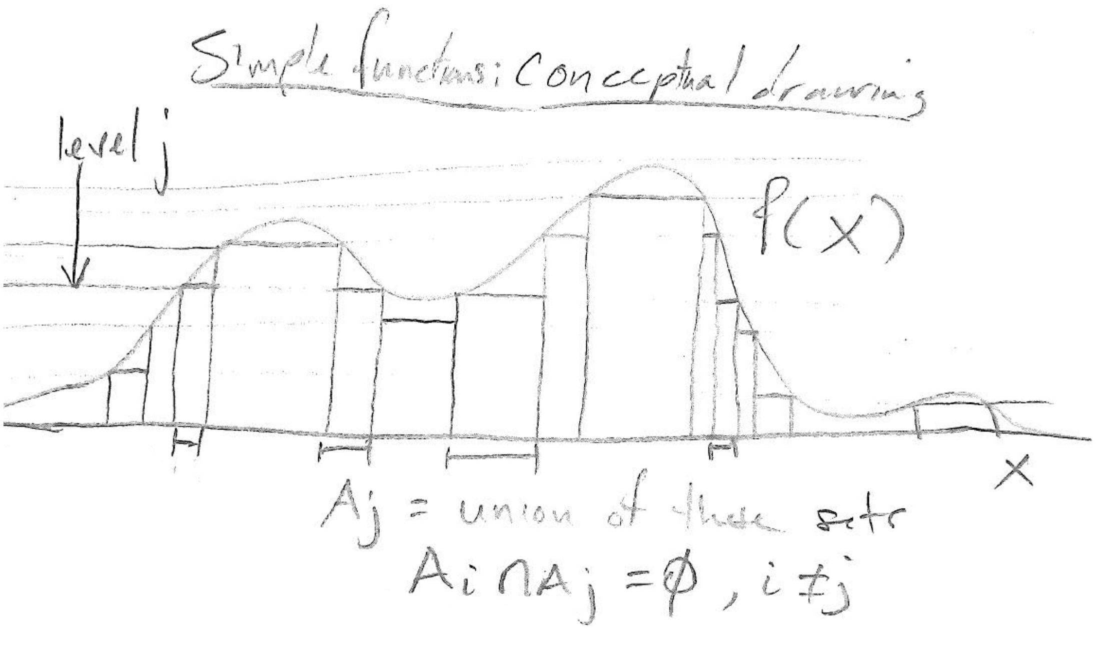
The infinite case

- In the case where $\mu(S) = \infty$, $N(S) = \infty$ with probability one. We therefore use a different proof technique that gives the same result.
- Approximate $f(x)$ with simple functions: $f_k = \sum_{i=1}^{2^k} a_i \mathbb{1}_{A_i}$ using $k2^k$ equally-spaced levels in the interval $(0, k)$; $A_i = [\frac{i}{2^k}, \frac{i+1}{2^k})$ and $a_i = \frac{i}{2^k}$. In the limit $k \to \infty$, the width of those levels $\frac{1}{2^k} \searrow 0$, and so $f_k \to f$.
- That's not a proof, but consider that $f_k(x) = \max_{i \leq k2^k} \frac{i}{2^k} \mathbb{1}(\frac{i}{2^k} < x)$, so $f_k(x) \nearrow x$ as $k \to \infty$.
- Important notation change: $\mathcal{M} = \sum_{x \in \Pi} f(x) \Leftrightarrow \int_S N(dx) f(x)$.
- Approximate f with f_k. Then with the notation change, $\mathcal{M}_k = \sum_{x \in \Pi} f_k(x) \Leftrightarrow \sum_{i=1}^{2^k} a_i N(A_i)$.
- The Laplace functional is

$$\begin{aligned}
\mathbb{E}e^{t\mathcal{M}} &= \lim_{k\to\infty} \mathbb{E}e^{t\mathcal{M}_k} = \lim_{k\to\infty} \mathbb{E} \prod_{i=1}^{2^k} e^{ta_i N(A_i)} && \leftarrow N(A_i) \text{ and } N(A_j) \text{ are independent} \\
&= \lim_{k\to\infty} \exp\left\{ \sum_{i=1}^{2^k} \mu(A_i)(e^{ta_i} - 1) \right\} && \leftarrow N(A_i) \sim \text{Pois}(\mu(A_i)) \\
&= \exp \int_S \mu(dx)(e^{tf(x)} - 1) && \leftarrow \text{integral as limit of infinitesimal sums} \qquad (2.4)
\end{aligned}$$

Mean and variance of $\mathcal{M}$: Using ideas from moment generating functions, it follows that

$$\underbrace{\mathbb{E}(\mathcal{M}) = \int_S \mu(dx) f(x),}_{= \frac{d}{dt}\mathbb{E}e^{t\mathcal{M}}|_{t=0}} \qquad \underbrace{\mathcal{V}(\mathcal{M}) = \int_S \mu(dx) f(x)^2}_{= \frac{d^2}{dt^2}\mathbb{E}e^{t\mathcal{M}}|_{t=0} - \mathbb{E}(\mathcal{M})^2} \tag{2.5}$$



Finiteness of $\int_S N(dx)f(x)$

- The next obvious question when $\mu(S) = \infty$ (and thus $N(S) = \infty$) is if $\int_S N(dx)f(x) < \infty$. Campbell's theorem gives the necessary and sufficient conditions for this to be true.

- Campbell's Theorem: Let $\Pi \sim \mathrm{PP}(\mu)$ and N the PRM. Let $f(x)$ be a non-negative function on S. Then with probability one,

$$\mathcal{M} = \int_S N(dx)f(x) \begin{cases} < \infty & \text{if } \int_S \min(f(x),1)\mu(dx) < \infty \\ = \infty & \text{otherwise} \end{cases} \tag{2.6}$$

Proof: For $u > 0$, $e^{-u\mathcal{M}} = e^{-u\mathcal{M}}\mathbb{1}(\mathcal{M} < \infty) \nearrow \mathbb{1}(\mathcal{M} < \infty)$ as $u \searrow 0$. By dominated convergence, as $u \searrow 0$

$$\mathbb{E}e^{-u\mathcal{M}} = \mathbb{E}[e^{-u\mathcal{M}}\mathbb{1}(\mathcal{M} < \infty)] \nearrow \mathbb{E}[\mathbb{1}(\mathcal{M} < \infty)] = P(\mathcal{M} < \infty) \tag{2.7}$$

In our case: $P(\mathcal{M} < \infty) = \lim_{u\searrow 0} \exp \int_S \mu(dx)(e^{-uf(x)} - 1)$.

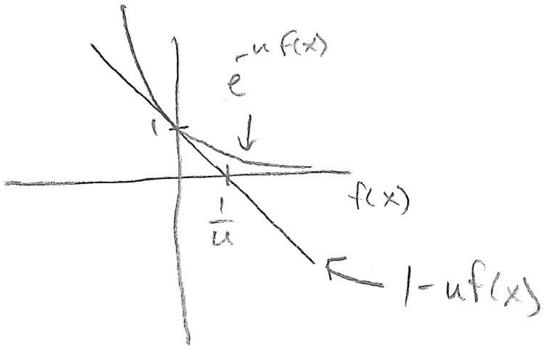
Sufficiency

1. For $P(\mathcal{M} < \infty) = 1$ as in the theorem, we need $\lim_{u\searrow 0} \int_S \mu(dx)(e^{-uf(x)} - 1) = 0$.

2. For $0 < u < 1$ we have

   $1-f(x) < 1-uf(x) < e^{-uf(x)} \quad \longrightarrow$

   (Figure: $e^{-uf(x)}$ is convex in $f(x)$ and $1 - uf(x)$ is a 1st order Taylor expansion of $e^{-uf(x)}$ at $f(x) = 0$)



3. Therefore $1 - e^{-uf(x)} < f(x)$. Also, $1 - e^{-uf(x)} < 1$ trivially.

4. So:
$$0 \le \int_S \mu(dx)(1 - e^{-uf(x)}) \le \int_S \mu(dx)\min(f(x),1) \tag{2.8}$$

5. If $\int_S \mu(dx)\min(f(x),1) < \infty$, then by dominated convergence

$$\lim_{u\searrow 0} \int_S \mu(dx)(1 - e^{-uf(x)}) = \int_S \mu(dx)\Big(1 - \exp\Big\{\lim_{u\searrow 0} -uf(x)\Big\}\Big) = 0 \tag{2.9}$$

- This proves sufficiency. For necessity we can show that (see, e.g., Cinlar II.2.13)

$$\int_S \min(f(x),1)\mu(dx) = \infty \implies \int_S \mu(dx)(1 - e^{-uf(x)}) = \infty \implies \mathbb{E}e^{-u\mathcal{M}} = 0, \forall u \tag{2.10}$$

Completely random measures (CRM)

- *Definition of measure*: The set function μ is a measure on the space S if

  1. $\mu(\emptyset) = 0$
  2. $\mu(A) \geq 0$ for all $A \subset S$
  3. $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ when $A_i \cap A_j = \emptyset, i \neq j$

- *Definition of completely random measure*: The set function M is a completely random measure on the space S if it satisfies $\#1$ to $\#3$ above and

  1. $M(A)$ is a random variable
  2. $M(A_1), \ldots, M(A_k)$ are independent for disjoint sets A_i

- Example: Let N be the counting measure associated with $\Pi \sim$ PP(μ). It's a CRM.

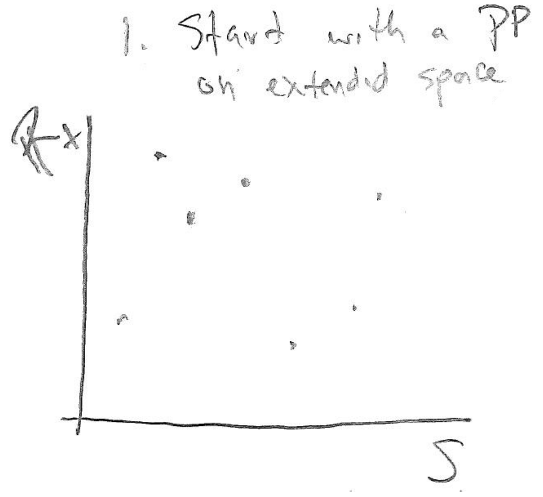
- We will be interested in the following situation: Let $\Pi \sim$ PP(μ) and mark each $\theta \in \Pi$ with a r.v. $\pi \sim \lambda(\pi)$, $\pi \in \mathbb{R}_+$. Then $\Pi^* = \{(\theta, \pi)\}$ is a PP on $S \times \mathbb{R}_+$ with mean measure $\mu(d\theta)\lambda(\pi)d\pi$

- If $N(d\theta, d\pi)$ is the counting measure for Π^*, then $N(C) \sim$ Pois$(\int_C \mu(d\theta)\lambda(\pi)d\pi)$.

- For $A \subset S$, let $M(A) = \int_A \int_0^{\infty} N(d\theta, d\pi)\pi$. Then M is a CRM on S.

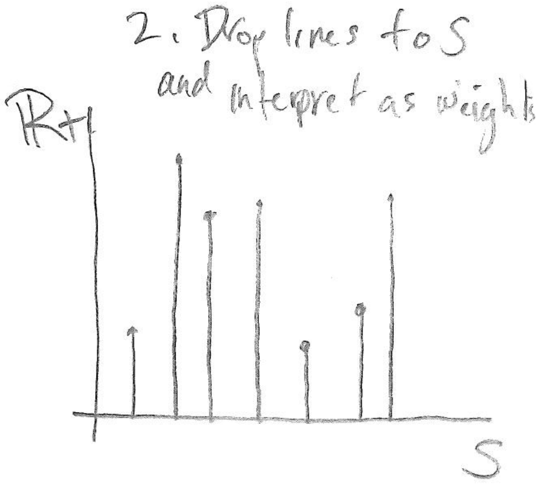
- M is a special case of sums of functions of Poisson processes with $f(\theta, \pi) = \pi$. Therefore we know that
$$\mathbb{E}e^{tM(A)} = \exp \int_A \int_0^{\infty} (e^{t\pi} - 1)\mu(d\theta)\lambda(\pi)d\pi. \tag{2.11}$$

- This works both ways: If we define M and show it has this Laplace transform, then we *know* there is a marked Poisson process "underneath" it with mean measure equal to $\mu(d\theta)\lambda(\pi)d\pi$.



marked PP on $S \times \mathbb{R}_+$ with mean measure $\mu(dx) \times \lambda(\pi)d\pi$

$\longrightarrow$



completely random measure on S.
$M(A) = \int_A \int_0^{\infty} N(d\theta, d\pi)\pi$

Gamma processes

- Definition: Let μ be a non-atomic measure on S. Then G is a gamma process if for all $A \subset S$, $G(A) \sim \text{Gam}(\mu(A), c)$ and $G(A_1), \dots, G(A_k)$ are independent for disjoint $A_1, \dots, A_k$. We write $G \sim \text{GaP}(\mu, c)$. ($c > 0$)

- Before trying to intuitively understand G, let's calculate its Laplace transform. For $t < 0$,

$$\mathbb{E}e^{tG(A)} = \int_0^\infty \frac{c^{\mu(A)}}{\Gamma(\mu(A))} G(A)^{\mu(A)-1} e^{-G(A)(c-t)} dG(A) = \left(\frac{c}{c-t}\right)^{\mu(A)} \tag{2.12}$$

- Manipulate this term as follows (and watch the magic!)

$$\begin{aligned}
\left(\frac{c}{c-t}\right)^{\mu(A)} &= \exp\left\{-\mu(A)\ln\frac{c-t}{c}\right\} \\
&= \exp\left\{-\mu(A)\int_c^{c-t}\frac{1}{s}ds\right\} \\
&= \exp\left\{-\mu(A)\int_c^{c-t}ds\int_0^\infty e^{-\pi s}d\pi\right\} \\
&= \exp\left\{-\mu(A)\int_0^\infty d\pi\int_c^{c-t}e^{-\pi s}ds\right\} \quad \text{(switched integrals)} \\
&= \exp\left\{\mu(A)\int_0^\infty (e^{t\pi}-1)\pi^{-1}e^{-c\pi}d\pi\right\} \qquad (2.13)
\end{aligned}$$

Therefore, G has an underlying Poisson random measure on $S \times \mathbb{R}_+$

$$G(A) = \int_A\int_0^\infty N(d\theta, d\pi)\pi, \quad N(d\theta, d\pi) \sim \text{Pois}(\mu(d\theta)\pi^{-1}e^{-c\pi}d\pi) \tag{2.14}$$

- The mean measure of N is $\mu(d\theta)\pi^{-1}e^{-c\pi}d\pi$ on $S \times \mathbb{R}_+$. We can use this to answer questions about $G \sim \text{GaP}(\mu, c)$ using the Poisson process perspective.

  1. How many total atoms? $\int_S\int_0^\infty \mu(d\theta)\pi^{-1}e^{-c\pi}d\pi = \infty \quad \Rightarrow \quad$ infinite # w.p. 1
  (Tells us that there are an infinite number of points in any subset $A \subset S$ that have nonzero mass according to G)
  2. How many atoms $\geq \epsilon > 0$? $\int_S\int_\epsilon^\infty \mu(d\theta)\pi^{-1}e^{-c\pi}d\pi < \infty \quad \Rightarrow \quad$ finite # w.p. 1
  (w.r.t. #1, further tells us only a finite number have mass greater than ϵ)
  3. Campbell's theorem: $f(\theta, \pi) = \pi \rightarrow \int_S\int_0^\infty \min(\pi, 1)\mu(d\theta)\pi^{-1}e^{-c\pi}d\pi < \infty$, therefore

  $$G(A) = \int_A\int_0^\infty N(d\theta, d\pi)\pi < \infty \text{ w.p. } 1$$

  (Tells us if we summed up the infinite number of nonzero masses in any set A, we would get a finite number even though we have an infinite number of nonzero things to add)

- Aside: We already knew #3 by definition of G, but this isn't always the order in which CRMs are defined. Imagine starting the definition with a mean measure on $S \times \mathbb{R}_+$.

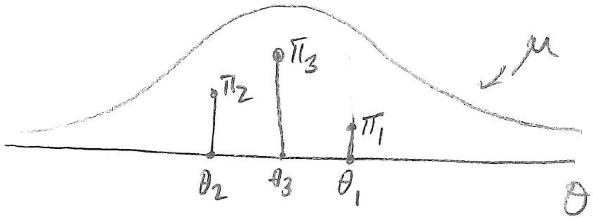
- #3 shouldn't feel mysterious at all. Consider $\sum_{n=1}^\infty \frac{1}{n^2}$. It's finite, but for each n, $\frac{1}{n^2} > 0$ and there are an infinite number of settings for n. "Infinite jump processes" such as the gamma process replace deterministic sequences like $(1, 1/2^2, 1/3^2, \dots)$ with something random.

Gamma process as a limiting case

- Is there a more intuitive way to understand the gamma process?
- Definition: Let μ be a non-atomic measure on S and $\mu(S) < \infty$. Let $\pi_i \overset{iid}{\sim} \text{Gam}(\frac{\mu(S)}{K}, c)$ and $\theta_i \overset{iid}{\sim} \mu/\mu(S)$ for $i = 1, \dots, K$. If $G_k = \sum_{i=1}^K \pi_i \delta_{\theta_i}$, then $\lim_{K\to\infty} G_K = G \sim \text{GaP}(\mu, c)$.

Picture: In the limit $K \to \infty$, we have more and more atoms with smaller and smaller weights

$$G_K(A) = \sum_{i=1}^K \pi_i \delta_{\theta_i}(A) = \sum_{i=1}^K \pi_i \mathbb{1}(\theta_i \in A)$$



Proof: Use the Laplace transform

$$\begin{aligned}
\mathbb{E}e^{tG_K(A)} &= \mathbb{E}e^{t\sum_{i=1}^K \pi_i \mathbb{1}(\theta_i \in A)} = \mathbb{E}\prod_{i=1}^K e^{t\pi_i \mathbb{1}(\theta_i\in A)} = \mathbb{E}[e^{t\pi\mathbb{1}(\theta\in A)}]^K \\
&= \mathbb{E}\left[e^{t\pi}\mathbb{1}(\theta\in A) + \mathbb{1}(\theta\notin A)\right]^K \\
&= \left[\mathbb{E}[e^{t\pi}]P(\theta\in A) + P(\theta\notin A)\right]^K \\
&= \left[\left(\frac{c}{c-t}\right)^{\frac{\mu(S)}{K}} \frac{\mu(A)}{\mu(S)} + 1 - \frac{\mu(A)}{\mu(S)}\right]^K \\
&= \left[1 + \frac{\mu(A)}{\mu(S)}\left(\left(\frac{c}{c-t}\right)^{\frac{\mu(S)}{K}} - 1\right)\right]^K \\
&= \left[1 + \frac{\mu(A)}{\mu(S)}\left(\sum_{n=1}^\infty \frac{(\ln\frac{c}{c-t})^n}{n!}\left(\frac{\mu(S)}{K}\right)^n\right)\right]^K \quad \leftarrow \text{(exponential power series)} \\
&= \left[1 + \frac{\mu(A)}{\mu(S)}\left(\frac{\mu(S)}{K}\ln\frac{c}{c-t} + O(1/K^2)\right)\right]^K \qquad (2.15)
\end{aligned}$$

- In the limit $K \to \infty$, this last equation converges to

$$\exp\left\{\frac{\mu(A)}{\mu(S)}\mu(S)\ln\frac{c}{c-t}\right\} = \left(\frac{c}{c-t}\right)^{\mu(A)}.$$

(recall that $\lim_{K\to\infty}(1 + \frac{a}{K} + O(K^{-2}))^K = e^a$)

- This is the Laplace transform of a $\text{Gam}(\mu(A), c)$ random variable.
- Therefore, $G_K(A) \to G(A) \sim \text{Gam}(\mu(A), c)$, which we've already defined and analyzed as a gamma process.

Chapter 3

Beta processes and the Poisson process

A sparse coding latent factor model

- We have a $d \times n$ matrix Y. We want to factorize it as follows:

$$ d\left[\ \overset{n}{Y}\ \right] \approx d\left[\ \overset{K}{\Theta}\ \right] \times \left(K\left[\ \overset{n}{W}\ \right] \circ \left[\ \overset{n}{Z}\ \right] K \right) $$

where

$$\left.\begin{array}{ll} \theta_i \sim p(\theta) & i = 1, \ldots, K \\ w_j \sim p(w) & j = 1, \ldots, n \\ z_j \in \{0,1\}^K & j = 1, \ldots, n \end{array}\right\}$$

"sparse coding" because each vector Y_j only possesses the columns of Θ indicated by z_j (want $\sum_i z_{ji} \ll K$)

- Example: Y could be

  a) gene data of n people,

  b) patches extracted from an image for denoising (called "dictionary learning")

- We want to define a "Bayesian nonparametric" prior for this problem. By this we mean that

  1. The prior can allow $K \to \infty$ and remain well defined
  2. As $K \to \infty$, the *effective rank* is finite (and relatively small)
  3. The model somehow learns this rank from the data (not discussed)

A "beta sieves" prior

- Let $\theta_i \sim \mu/\mu(S)$ and w_j be drawn as above. Continue this generative model by letting

$$\begin{aligned} z_{ji} &\overset{iid}{\sim} \text{Bern}(\pi_i), j = 1, \dots, n & (3.1)\\ \pi_i &\sim \text{Beta}\left(\alpha\frac{\gamma}{K}, \alpha\Big(1 - \frac{\gamma}{K}\Big)\right), i = 1, \dots, K & (3.2)\end{aligned}$$

- The set (θ_i, π_i) are paired. π_i gives the probability an observation picks θ_i. Notice that we expect $\pi_i \to 0$ as $K \to \infty$.
- Construct a completely random measure $H_K = \sum_{i=1}^K \pi_i \delta_{\theta_i}$.
- We want to analyze what happens when $K \to \infty$. We'll see that it converges to a *beta process*.

Asymptotic analysis of beta sieves

- We have that $H_K = \sum_{i=1}^K \pi_i \delta_{\theta_i}$, $\pi_i \overset{iid}{\sim} \text{Beta}\left(\alpha\gamma/K, \alpha(1-\gamma/K)\right)$, $\theta_i \overset{iid}{\sim} \mu/\mu(S)$, where $\gamma = \mu(S) < \infty$. We want to understand $\lim_{K\to\infty} H_K$.
- Look at the Laplace transform of $H_K(A)$. Let $H(A) = \lim_{K\to\infty} H_K(A)$. Then

$$\mathbb{E}e^{tH(A)} = \lim_{K\to\infty} \mathbb{E}e^{tH_K(A)} = \underbrace{\lim_{K\to\infty} \mathbb{E}e^{t\sum_{i=1}^K \pi_i \mathbb{1}(\theta_i \in A)} = \lim_{K\to\infty} \mathbb{E}[e^{t\pi\mathbb{1}(\theta\in A)}]^K}_{\text{sum} \to \text{product and use i.i.d. fact}} \qquad (3.3)$$

- Focus on $\mathbb{E}e^{t\pi\mathbb{1}(\theta\in A)}$ for a particular K-level approximation. We have the following (long) sequence of equalities:

$$\begin{aligned}
\mathbb{E}e^{t\pi\mathbb{1}(\theta\in A)} &= \mathbb{E}[e^{t\pi}\mathbb{1}(\theta \in A) + \mathbb{1}(\theta \notin A)] = P(\theta \in A)\mathbb{E}e^{t\pi} + P(\theta \notin A)\\
&= 1 + \frac{\mu(A)}{\mu(S)}\Big(\mathbb{E}e^{t\pi} - 1\Big) \quad \leftarrow \quad \mathbb{E}e^{t\pi} = 1 + \sum_{s=1}^{\infty} \frac{t^s}{s!} \prod_{r=0}^{s-1} \frac{\frac{\alpha\gamma}{K} + r}{\alpha + r}\\
&= 1 + \frac{\mu(A)}{\mu(S)} \sum_{s=1}^{\infty} \frac{t^s}{s!} \prod_{r=0}^{s-1} \frac{\frac{\alpha\gamma}{K} + r}{\alpha + r} \quad \leftarrow \quad \text{plugging in} \uparrow\\
&= 1 + \frac{\mu(A)}{K} \sum_{s=1}^{\infty} \frac{t^s}{s!} \prod_{r=1}^{s-1} \frac{r}{\alpha + r} + O(\tfrac{1}{K^2}) \quad \leftarrow \quad \text{separate out } r = 0\\
&= 1 + \frac{\mu(A)}{K} \sum_{s=1}^{\infty} \frac{t^s}{s!} \frac{\alpha\Gamma(\alpha)\Gamma(s)}{\Gamma(\alpha + s)} + O(\tfrac{1}{K^2}) \quad \leftarrow \quad \text{since } \Gamma(\alpha + 1) = \alpha\Gamma(\alpha)\\
&= 1 + \frac{\mu(A)}{K} \sum_{s=1}^{\infty} \frac{t^s}{s!} \int_0^1 \alpha\pi^{s-1}(1-\pi)^{\alpha-1} d\pi + O(\tfrac{1}{K^2}) \quad \leftarrow \quad \text{normalizer of Beta}(s, \alpha)\\
&= 1 + \frac{\mu(A)}{K} \int_0^1 \sum_{s=1}^{\infty} \Big(\frac{(t\pi)^s}{s!}\Big)\alpha\pi^{-1}(1-\pi)^{\alpha-1} d\pi + O(\tfrac{1}{K^2})\\
&= 1 + \frac{\mu(A)}{K} \int_0^1 (e^{t\pi} - 1)\alpha\pi^{-1}(1-\pi)^{\alpha-1} d\pi + O(\tfrac{1}{K^2})
\end{aligned}$$

Again using $\lim_{K\to\infty}(1 + a/K + O(K^{-2})^K = e^a$, it follows that

$$\lim_{K\to\infty} \mathbb{E}[e^{t\pi\mathbb{1}(\theta\in A)}]^K = \exp\left\{\mu(A)\int_0^1 (e^{t\pi}-1)\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi\right\}. \tag{3.4}$$

- We therefore know that

  1. H is a completely random measure.
  2. It has an associated underlying Poisson random measure $N(d\theta, d\pi)$ on $S \times [0,1]$ with mean measure $\mu(d\theta)\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi$.
  3. We can write $H(A)$ as $\int_A\int_0^1 N(d\theta, d\pi)\pi$.

Beta process (as a CRM)

- Definition: Let $N(d\theta, d\pi)$ be a Poisson random measure on $S \times [0,1]$ with mean measure $\mu(d\theta)\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi$, where μ is a non-atomic measure. Define the CRM $H(A)$ as $\int_A\int_0^1 N(d\theta, d\pi)\pi$. Then H is called a beta process, $H \sim \text{BP}(\alpha, \mu)$ and

$$\mathbb{E}e^{tH(A)} = \exp\left\{\mu(A)\int_0^1 (e^{t\pi}-1)\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi\right\}.$$

- (We just saw how we can think of H as the limit of a finite collection of random variables. This time we're just starting from the definition, which we could proceed to analyze regardless of the beta sieves discussion above.)

- Properties of H: Since H has a Poisson process representation, we can use the mean measure to calculate its properties (and therefore the asymptotic properties of the beta sieves approximation).

  - Finiteness: Using Campbell's theorem, $H(A)$ is finite with probability one, since

$$\int_A\int_0^1 \underbrace{\min(\pi, 1)}_{=\pi}\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi\mu(d\theta) = \mu(A) < \infty \ \text{ (by assumption about } \mu) \tag{3.5}$$

  - Infinite jump process: $H(A)$ is constructed from an infinite number of jumps, almost all infinitesimally small, since

$$N(A\times(0,1]) \sim \text{Pois}\Big(\mu(A)\int_0^1 \alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi\Big) = \text{Pois}(\infty) \tag{3.6}$$

  - Finite number of "big" jumps: There are only a finite number of jumps greater than any $\epsilon > 0$ since

$$N(A\times[\epsilon,1]) \sim \text{Pois}\Big(\underbrace{\mu(A)\int_\epsilon^1 \alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi}_{<\,\infty \text{ for } \epsilon\,>\,0}\Big) \tag{3.7}$$

  As $\epsilon \to 0$, the value in the Poisson goes to infinity, so the infinite jump process arises in this limit. Since the integral over the magnitudes is finite, this infinite number of atoms is being introduced in a "controlled" way as a function of ϵ (i.e., not "too quickly")

- Reminder and intuitions: All of these properties are over instantiations of a beta process, and so all statements are made with probability one.
  - It's not absurd to talk about beta processes that don't have an infinite number of jumps, or integrate to something infinite ("not absurd" in the way that it is absurd to talk about a negative value drawn from a beta distribution).
  - The support of the beta process includes these events, but they have probability zero, so any $H \sim \text{BP}(\alpha, \mu)$ is guaranteed to have the properties discussed above.
  - It's easy to think of H as one random variable, but as the beta sieves approximation shows, H is really a collection of an infinite number of random variables.
  - The statements we are making about H above aren't like asking whether a beta random variable is greater than 0.5. They are larger scale statements about properties of this infinite collection of random variables as a whole.

- Another definition of the beta process links it to the beta distribution and our finite approximation:

- <u>Definition II</u>: Let μ be a non-atomic measure on S. For all <u>infinitesimal</u> sets $d\theta \in S$, let

  $$H(d\theta) \sim \text{Beta}\{\alpha\mu(d\theta), \alpha(1-\mu(d\theta))\},$$

  then $H \sim \text{BP}(\alpha, \mu)$.

- We aren't going to prove this, but the proof is actually very similar to the beta sieves proof.

- Note the difference from the gamma process, where $G(A) \sim \text{Gam}(\mu(A), c)$ for any $A \subset S$. The beta distribution only comes in the infinitesimal limit. That is

  $$H(A) \not\sim \text{Beta}\{\alpha\mu(A), \alpha(1-\mu(A))\},$$

  when $\mu(A) > 0$. Therefore, we can only write beta distributions on things that equal zero with probability one... Compare this with the limit of the beta sieves prior.

- Observation: While $\mu(\{\theta\}) = \mu(d\theta) = 0$, $\int_A \mu(\{\theta\})d\theta = 0$ but $\int_A \mu(d\theta) = \mu(A) > 0$.

- This is a major difference between a measure and a function: μ is a measure, not a function. It also seems to me a good example of why these additional concepts and notations are necessary, e.g., why we can't just combine things like $\mu(A) = \int_A p(\theta)d\theta$ into one single notation, but instead talk about the "measure" μ and it's associated density $p(\theta)$ such that $\mu(d\theta) = p(\theta)d\theta$.

- This leads to discussions involving the Radon-Nikodym theorem, etc. etc.

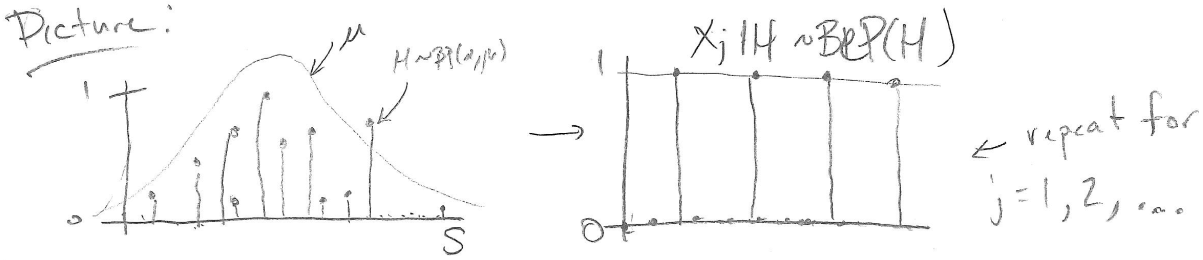
- The level of our discussion stops at an appreciation for why these types of theorems exist and are necessary (as overly-obsessive as they may feel the first time they're encountered), but we won't re-derive them.

Bernoulli process

- The Bernoulli process is constructed from the infinite limit of the "z" sequence in the process $z_{ji} \sim \text{Bern}(\pi_i), \pi_i \overset{iid}{\sim} \text{Beta}\left(\alpha\gamma/K, \alpha(1-\gamma/K)\right), i = 1, \dots, K$. The random measure $X_j^{(K)} = \sum_{i=1}^{K} z_{ji}\delta_{\theta_i}$ converges to a "Bernoulli process" as $K \to \infty$.

- Definition: Let $H \sim \text{BP}(\alpha, \mu)$. For each atom of H (the θ for which $H(\{\theta\}) > 0$), let $X_j(\{\theta\})|H \overset{ind}{\sim} \text{Bern}(H(\{\theta\}))$. Then X_j is a Bernoulli process, denoted $X_j|H \sim \text{BeP}(H)$.

- Observation: We know that H has an infinite number of locations θ where $H(\{\theta\}) > 0$ because it's a discrete measure (the Poisson process proves that). Therefore, X is infinite as well.



Some questions about X

1. How many 1's in X_j?
2. For $X_1, \dots, X_n | H \sim \text{BeP}(H)$, how many total locations are there with at least one X_j equaling one there (marginally speaking, with H integrated out)?

$$\text{i.e., what is} \quad \left|\{\theta : \textstyle\sum_{j=1}^{n} X_j(\{\theta\}) > 0\}\right|$$

- The Poisson process representation of the BP makes calculating this relatively easy. We start by observing that the X_j are marking the atoms, and so we have a marked Poisson process (or "doubly marked" since we can view (θ, π) as a marked PP as well).

Beta process marked by a Bernoulli process

- Definition: Let $\Pi \sim \text{PP}(\mu(d\theta)\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi)$ on $S \times [0,1]$ be a Poisson process underlying a beta process. For each $(\theta, \pi) \in \Pi$ draw a binary vector $z \in \{0,1\}^n$ where $z_i|\pi \overset{iid}{\sim} \text{Bern}(\pi)$ for $i = 1, \dots, n$. Denote the distribution on z as $Q(z|\pi)$. Then $\Pi^* = \{(\theta, \pi, z)\}$ is a marked Poisson process with mean measure $\mu(d\theta)\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi Q(z|\pi)$.

- There is therefore a Poisson process underlying the joint distribution of the hierarchical process

$$H \sim \text{BP}(\alpha, \mu), \quad X_i|H \overset{iid}{\sim} \text{BeP}(H), \; i = 1, \dots, n.$$

- We next answer the two questions about X asked above, starting with the second one.

Question: What is $K_n^+ = \Big|\{\theta : \sum_{j=1}^n X_j(\{\theta\}) > 0\}\Big|$?

Answer: The transition distribution $Q(z|\pi)$ gives the probability of a vector z at a *particular* location $(\theta, \pi) \in \Pi$ (notice Q doesn't depend on θ).

All we care about is whether $z \in C = \{0,1\}^n \backslash \vec{\mathbf{0}}$ (i.e., has a 1 in it)

We make the following observations:

- The probability $Q(C|\pi) = P(z \in C|\pi) = 1 - P(z \notin C|\pi) = 1 - (1-\pi)^n$.

- If we restrict the marked PP to C, we get the distribution on the value of K_n^+:

$$K_n^+ = N(S, [0,1], C) \sim \text{Pois}\Big(\int_S \int_0^1 \mu(d\theta)\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi \underbrace{Q(C|\pi)}_{=\,1-(1-\pi)^n} \Big). \tag{3.8}$$

- It's worth stopping to remember that N is a *counting* measure, and think about what exactly $N(S,[0,1],C)$ is counting.
  - ∗ $N(S,[0,1],C)$ is asking for the number of times event C happens (an event related to z), not caring about what the corresponding θ or π are (hence the S and $[0,1]$).
  - ∗ i.e., it's counting the thing we're asking for, K_n^+.

- We can show that $1-(1-\pi)^n = \sum_{i=0}^{n-1} \pi(1-\pi)^i \leftarrow$ geometric series

- It follows that

$$\begin{aligned}\int_S \int_0^1 \mu(d\theta)\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi(1-(1-\pi)^n) &= \mu(S)\sum_{i=0}^{n-1}\alpha\int_0^1 (1-\pi)^{\alpha+i-1}d\pi \\ &= \mu(S)\sum_{i=0}^{n-1}\frac{\alpha\Gamma(1)\Gamma(\alpha+i)}{\Gamma(\alpha+i+1)} \\ &= \sum_{i=0}^{n-1}\frac{\alpha\mu(S)}{\alpha+i} \qquad (3.9)\end{aligned}$$

- Therefore $K_n^+ \sim \text{Pois}\Big(\sum_{i=0}^{n-1}\frac{\alpha\mu(S)}{\alpha+i}\Big)$.

- Notice that as $n \to \infty$, $K_n^+ \to \infty$ with probability one, and that $\mathbb{E}K_n^+ \approx \alpha\mu(S)\ln n$.

- Also notice that we get the answer to the first question for free. Since X_j are i.i.d., we can treat each one marginally as if it were the first one.

- If $n = 1$, $X(S) \sim \text{Pois}(\mu(S))$. That is, the number of ones in each Bernoulli process is Pois$(\mu(S))$-distributed.

Chapter 4

Beta processes and size-biased constructions

The beta process

- Definition (review): Let $\alpha > 0$ and μ be a finite non-atomic measure on S. Let $C \in S \times [0, 1]$ and N be a Poisson random measure with $N(C) \sim \text{Pois}(\int_C \mu(d\theta)\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi)$. For $A \subset S$ define $H(A) = \int_A \int_0^1 N(d\theta, d\pi)\pi$. Then H is a beta process, denoted $H \sim \text{BP}(\alpha, \mu)$.

Intuitive picture (review)

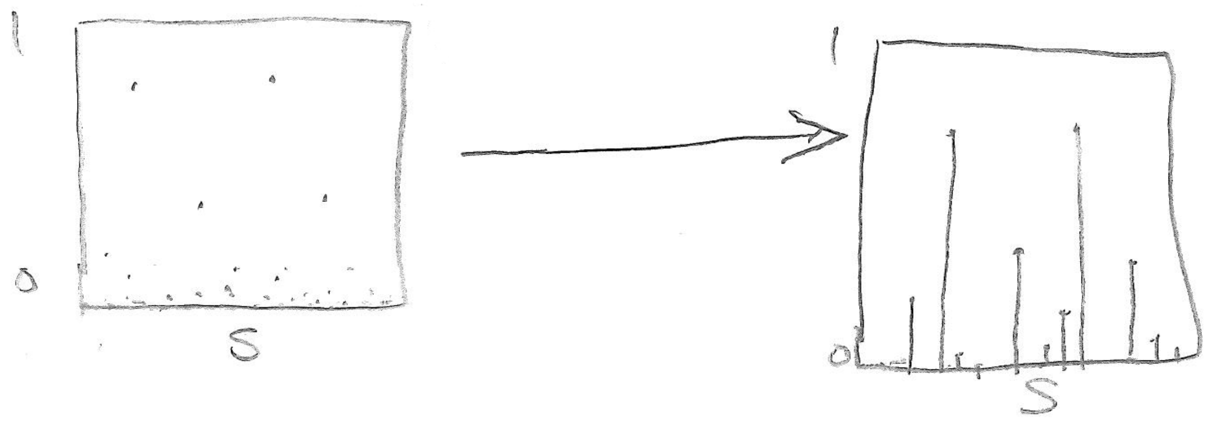


Figure 4.1 (left) Poisson process (right) CRM constructed form Poisson process. If $(d\theta, d\pi)$ is a point in the PP, $N(d\theta, d\pi) = 1$ and $N(d\theta, d\pi)\pi = \pi$. $H(A)$ is adding up π's in the set $A \times [0, 1]$.

Drawing from this prior

- In general, we know that if $\Pi \sim \text{PP}(\mu)$, we can draw $N(S) \sim \text{Pois}(\mu(S))$ and $X_1, \ldots, X_{N(S)} \overset{iid}{\sim} \mu/\mu(S)$ and construct Π from the $X_i's$.

- Similarly, we have the reverse property that if $N \sim \text{Pois}(\gamma)$ and $X_1, \ldots, X_N \overset{iid}{\sim} p(X)$, then the set $\Pi = \{X_1, \ldots, X_N\} \sim \text{PP}(\gamma p(X)dX)$. (This inverse property will be useful later.)

- Since $\int_S \int_0^1 \mu(d\theta)\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi = \infty$, this approach obviously won't work for drawing $H \sim$ BP(α, μ).

- The method of partitioning $[0, 1]$ and drawing $N(S\times(a, b])$ followed by

$$\theta_i^{(a)} \sim \mu/\mu(S), \quad \pi_i^{(a)} \sim \frac{\alpha\pi^{-1}(1-\pi)^{\alpha-1}}{\int_a^b \alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi}\mathbb{1}(a < \pi_i^{(a)} \leq b) \tag{4.1}$$

  is possible (using the Restriction theorem from Lecture 1, independence of PP's on disjoint sets, and the first bullet of this section), but not as useful for Bayesian models.

- The goal is to find *size-biased* representations for H that are more straightforward. (i.e., that involve sampling from standard distributions, which will hopefully make inference easier)

Size-biased representation I (a "restricted beta process")

- Definition: Let $\alpha = 1$ and μ be a non-atomic measure on S with $\mu(S) = \gamma < \infty$. Generate the following independent set of random variables

$$V_i \overset{iid}{\sim} \text{Beta}(\gamma, 1), \quad \theta_i \overset{iid}{\sim} \mu/\mu(S), \quad i = 1, 2, \ldots \tag{4.2}$$

  Let $H = \sum_{i=1}^{\infty}\left(\prod_{j=1}^{i} V_j\right)\delta_{\theta_i}$. Then $H \sim$ BP$(1, \mu)$.

  Proof: The proof uses the limiting case of the following finite approximation

- Let $\pi_i \sim$ Beta$(\frac{\gamma}{K}, 1), \theta_i \sim \mu/\mu(S)$ for $i = 1, \ldots, K$. Let $H_K = \sum_{i=1}^{K}\pi_i\delta_{\theta_i}$. Then $\lim_{K\to\infty} H_K \sim$ BP$(1, \mu)$. The proof is similar to the one last lecture.

  Question 1: As $K \to \infty$, what is $\pi_{(1)} = \max\{\pi_1, \ldots, \pi_K\}$?

  Answer: Look at the CDF's. We want the function $P(\pi_{(1)} < V_1)$ for a $V_1 \in [0, 1]$. Because the π_i are independent,

$$P(\pi_{(1)} < V_1) = \lim_{K\to\infty} P(\pi_1 < V_1, \ldots, \pi_K < V_1) = \lim_{K\to\infty}\prod_{i=1}^{K} P(\pi_i < V_1) \tag{4.3}$$

  - $P(\pi_i < V_1) = \int_0^{V_1} \frac{\gamma}{K}\pi_i^{\frac{\gamma}{K}-1}d\pi_i = V_1^{\frac{\gamma}{K}}$
  - $\lim_{K\to\infty}\prod_{i=1}^{K} P(\pi_i < V_1) = V_1^{\gamma}$
  - Therefore, $\pi_{(1)} = V_1, \quad V_1 \sim$ Beta$(\gamma, 1)$

Question 2: What is the second largest, denoted $\pi_{(2)} = \lim_{K\to\infty} \max\{\pi_1, \dots, \pi_K\}\backslash\{\pi_{(1)}\}$?

Answer: This is a little more complicated, but answering how to get $\pi_{(2)}$ shows how to get the remaining $\pi_{(i)}$.

$$\begin{aligned}
P(\pi_{(2)} < t|\pi_{(1)} = V_1) &= \prod_{\pi_i \neq \pi_{(1)}} P(\pi_i < t|\pi_i < V_1) \ \leftarrow \text{condition is each } \pi_i < \pi_{(1)} = V_1 \\
&= \prod_{\pi_i \neq \pi_{(1)}} \frac{P(\pi_i < t, \pi_i < V_1)}{P(\pi_i < V_1)} \\
&= \prod_{\pi_i \neq \pi_{(1)}} \frac{P(\pi_i < t)}{P(\pi_{(1)} = V_1)} \ \leftarrow \text{since } t < V_1\text{, first event contains second} \\
&= \lim_{K\to\infty} \prod_{\pi_i \neq \pi_{(1)}} \frac{\int_0^t \frac{\gamma}{K}\pi_i^{\frac{\gamma}{K}-1} d\pi_i}{\int_0^{V_1} \frac{\gamma}{K}\pi_i^{\frac{\gamma}{K}-1} d\pi_i} \\
&= \lim_{K\to\infty} \left[\left(\frac{t}{V_1}\right)^{\frac{\gamma}{K}}\right]^{K-1} = \left(\frac{t}{V_1}\right)^{\gamma} \qquad (4.4)
\end{aligned}$$

- So the density $p(\pi_{(2)}|\pi_{(1)} = V_1) = V_1^{-1}\gamma \left(\frac{\pi_{(2)}}{V_1}\right)^{\gamma-1}$. $\pi_{(2)}$ has support $[0, V_1]$.
- Change of variables: $V_2 := \pi_{(2)}/V_1 \quad \to \quad \pi_{(2)} = V_1 V_2,\ d\pi_{(2)} = V_1 dV_2$.
- Plugging in, $P(V_2|\pi_{(1)} = V_1) = V_1^{-1}\gamma V_2^{\gamma-1} \cdot \underbrace{V_1}_{\text{Jacobian}} = \gamma V_2^{\gamma-1} = \text{Beta}(\gamma, 1)$
- The above calculation has shown two things:

  1. V_2 is independent of V_1 (this is an instance of a "neutral-to-the-right process")
  2. $V_2 \sim \text{Beta}(\gamma, 1)$

- Since $\pi_{(2)}|\{\pi_{(1)} = V_1\} = V_1 V_2$ and V_1, V_2 are independent, we can get the value of $\pi_{(2)}$ using previously drawn V_1 and then drawing V_2 from $\text{Beta}(\gamma, 1)$ distributions.
- The same exact reasoning follows for $\pi_{(3)}, \pi_{(4)}, \dots$
- For example, for $\pi_{(3)}$, we have $P(\pi_{(3)} < t|\pi_{(2)} = V_1V_2, \pi_{(1)} = V_1) = P(\pi_{(3)} < t|\pi_{(2)} = V_1V_2)$ because conditioning on $\pi_{(2)} = V_1V_2$ restricts $\pi_{(3)}$ to also satisfy condition of $\pi_{(1)}$.
- In other words, if we force $\pi_{(3)} < \pi_{(2)}$ by conditioning, we get the additional requirement $\pi_{(3)} < \pi_{(1)}$ for free, so we can condition on the $\pi_{(i)}$ immediately before.
- Think of V_1V_2 as a single non-random (i.e., already known) value by the time we get to $\pi_{(3)}$. We can exactly follow the above sequence after making the correct substitutions and re-indexing.

Size-biased representation II

Definition: Let $\alpha > 0$ and μ be a non-atomic measure on S with $\mu(S) < \infty$. Generate the following random variables:

$$\begin{aligned} C_i &\sim \text{Pois}\Big(\frac{\alpha\mu(S)}{\alpha+i-1}\Big), \quad i = 1, 2, \dots \\ \pi_{ij} &\sim \text{Beta}(1, \alpha+i-1), \quad j = 1, \dots, C_i \\ \theta_{ij} &\sim \mu/\mu(S), \quad j = 1, \dots, C_i \end{aligned} \tag{4.5}$$

Define $H = \sum_{i=1}^{\infty}\sum_{j=1}^{C_i}\pi_{ij}\delta_{\theta_{ij}}$. Then $H \sim \text{BP}(\alpha, \mu)$.

Proof:

- We can use Poisson processes to prove this. This is a good example of how easy a proof can become when we recognize a hidden Poisson process and calculate it's mean measure.
  - Let $H_i = \sum_{j=1}^{C_i}\pi_{ij}\delta_{\theta_{ij}}$. Then the set $\Pi_i = \{(\theta_{ij}, \pi_{ij})\}$ is a Poisson process because it contains a Poisson-distributed number of i.i.d. random variables.
  - As a result, the mean measure of Π_i is

    $$\underbrace{\frac{\alpha\mu(S)}{\alpha+i-1}}_{\text{Poisson \# part}} \times \underbrace{(\alpha+i-1)(1-\pi)^{\alpha+i-2}d\pi}_{\text{distribution on } \pi} \times \underbrace{\mu(d\theta)/\mu(S)}_{\text{distribution on } \theta} \tag{4.6}$$

    We can simplify this to $\alpha\mu(d\theta)(1-\pi)^{\alpha+i-2}d\pi$. We can justify this with the marking theorem (π marks θ), or just thinking about the joint distribution of (θ, π).
  - $H = \sum_{i=1}^{\infty} H_i$ by definition. Equivalently $\Pi = \bigcup_{i=1}^{\infty}\Pi_i$.
  - By the superposition theorem, we know that Π is a Poisson process with mean measure equal to the sum of the mean measures of each Π_i.
  - We can calculate this directly:

    $$\sum_{i=1}^{\infty}\alpha\mu(d\theta)(1-\pi)^{\alpha+i-2}d\pi = \alpha\mu(d\theta)(1-\pi)^{\alpha-2}\underbrace{\sum_{i=1}^{\infty}(1-\pi)^{i}}_{=\frac{1-\pi}{\pi}}d\pi \tag{4.7}$$

  - Therefore, we've shown that Π is a Poisson process with mean measure

    $$\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi\mu(d\theta).$$

  - In other words, this second size-biased construction is the CRM constructed from integrating a PRM with this mean measure against the function $f(\theta, \pi) = \pi$ along the π dimension. This is the definition of a beta process.

Size-biased representation III

- Definition: Let $\alpha > 0$ and μ be a non-atomic measure on S with $\mu(S) < \infty$. The following is a constructive definition of $H \sim \text{BP}(\alpha, \mu)$.

$$C_i \sim \text{Pois}(\mu(S)), \quad V_{ij}^{(\ell)} \sim \text{Beta}(1, \alpha), \quad \phi_{ij} \sim \mu/\mu(S) \tag{4.8}$$

$$H = \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} V_{ij}^{(i)} \prod_{\ell=1}^{i-1} (1 - V_{ij}^{(\ell)}) \delta_{\theta_{ij}} \tag{4.9}$$

- Like the last construction, the weights are decreasing in expectation as a function of i.

$$H = \sum_{j=1}^{C_1} V_{1j} \delta_{\theta_{1j}} + \sum_{j=1}^{C_2} V_{2j}^{(2)} (1 - V_{2j}^{(1)}) \delta_{\theta_{2j}} + \sum_{j=1}^{C_3} V_{3j}^{(3)} (1 - V_{3j}^{(2)})(1 - V_{3j}^{(1)}) \delta_{\theta_{3j}} + \cdots \tag{4.10}$$

- The structure is also very similar. We have a Poisson-distributed number of atoms in each group and they're marked with an independent random variable.

- Therefore, we can write $H = \sum_{i=1}^{\infty} H_i$, where each H_i has a corresponding Poisson process Π_i with mean measure $\mu(S) \times (\mu(d\theta)/\mu(S)) \times \lambda_i(\pi) d\pi$, where $\lambda_i(\pi)$ is the distribution of

$$\pi = V_i \prod_{j=1}^{i-1} (1 - V_j), \quad V_i \sim \text{Beta}(1, \alpha).$$

- By the superposition theorem, H has an underlying Poisson process Π with mean measure equal to the sum of each H_i's mean measures: $\mu(d\theta) \sum_{i=1}^{\infty} \lambda_i(\pi) d\pi$.

- Therefore, all that remains is to calculate this sum (which is a little complicated).

Proof:

- We focus on $\lambda_i(\pi)$, which is the distribution on $\pi = f(V_1, \ldots, V_i)$, where $f(V_1, \ldots, V_i) = V_i \prod_{j=1}^{i-1} (1 - V_j), V_j \sim \text{Beta}(1, \alpha)$.

- *Lemma*: Let $T \sim \text{Gam}(i-1, \alpha)$. Then $e^{-T} \overset{d}{=} \prod_{j=1}^{i-1} (1 - V_j)$.

- *Proof*: Define $\xi_j = -\ln(1 - V_j)$. We can show by a change of variables that $\xi_j \sim \text{Exp}(\alpha)$. The function $-\ln \prod_{j=1}^{i-1} (1 - V_j) = \sum_{j=1}^{i-1} \xi_j$. Since the V_j are independent, the ξ_j are independent. We know that sums of i.i.d. exponential r.v.'s are gamma distributed, so $T = \sum_{j=1}^{i-1} \xi_j$ is distributed as $\text{Gam}(i-1, \alpha)$. That is, $-\ln \prod_{j=1}^{i-1} (1 - V_j) \overset{d}{=} T \sim \text{Gam}(i-1, \alpha)$ and the result follows because the same function of two equally distributed r.v.'s is also equally distributed.

- We split the proof into two cases, $i = 1$ and $i > 1$.

- Case $i = 1$: $V_{1j} \sim \text{Beta}(1, \alpha)$, therefore $\pi_{1j} = V_{1j} \sim \lambda_1(\pi) d\pi = \alpha(1 - \pi)^{\alpha - 1} d\pi$.

- Case $i > 1$: $V_{ij} \sim \text{Beta}(1, \alpha), T_{ij} \sim \text{Gam}(i-1, \alpha), \pi_{ij} = V_{ij}e^{-T_{ij}}$. We need to find the density of π_{ij}. Let $W_{ij} = e^{-T_{ij}}$. Then changing variables,

$$p_{W_i}(w|\alpha) = \frac{\alpha^{i-1}}{(i-2)!} w^{\alpha-1}(-\ln w)^{i-2}. \tag{4.11}$$

↑

plug $T_{ij} = -\ln W_{ij}$ into gamma distribution and multiply by Jacobian

- Therefore $\pi_{ij} = V_{ij}W_{ij}$ and using the product distribution formula

$$\begin{aligned} \lambda_i(\pi|\alpha) &= \int_\pi^1 w^{-1} p_V(\pi/w|\alpha) p_{W_i}(w|\alpha) dw \\ &= \frac{\alpha^i}{(i-2)!} \int_\pi^1 w^{\alpha-1}(-\ln w)^{i-2}(1-\pi/w)^{\alpha-1} dw \\ &= \frac{\alpha^i}{(i-2)!} \int_\pi^1 w^{-1}(-\ln w)^{i-2}(w-\pi)^{\alpha-1} dw \end{aligned} \tag{4.12}$$

- This integral doesn't have a closed form solution. However, recall that we only need to calculate $\mu(d\theta)\sum_{i=1}^\infty \lambda_i(\pi)d\pi$ to find the mean measure of the underlying Poisson process.

$$\mu(d\theta)\sum_{i=1}^\infty \lambda_i(\pi)d\pi = \mu(d\theta)\lambda_1(\pi)d\pi + \mu(d\theta)\sum_{i=2}^\infty \lambda_i(\pi)d\pi \tag{4.13}$$

$$\begin{aligned} \mu(d\theta)\sum_{i=2}^\infty \lambda_i(\pi)d\pi &= \mu(d\theta)\sum_{i=2}^\infty d\pi \frac{\alpha^i}{(i-2)!}\int_\pi^1 w^{-1}(-\ln w)^{i-2}(w-\pi)^{\alpha-1}dw \\ &= \mu(d\theta)d\pi\alpha^2 \int_\pi^1 w^{-1}(w-\pi)^{\alpha-1}dw \underbrace{\sum_{i=2}^\infty \frac{(-\alpha\ln w)^{i-2}}{(i-2)!}}_{=\, e^{-\alpha \ln w} \,=\, w^{-\alpha}} \\ &= \mu(d\theta)d\pi\alpha^2 \underbrace{\int_\pi^1 w^{-(\alpha+1)}(w-\pi)^{\alpha-1}dw}_{=\, \frac{(w-\pi)^\alpha}{\alpha\pi w^\alpha}\Big|_\pi^1} \\ &= \mu(d\theta)\frac{\alpha(1-\pi)^\alpha}{\pi}d\pi \end{aligned} \tag{4.14}$$

- Adding $\mu(d\theta)\lambda_1(\pi)d\pi$ from Case 1 with this last value,

$$\mu(d\theta)\sum_{i=1}^\infty \lambda_i(\pi)d\pi = \mu(d\theta)\alpha\pi^{-1}(1-\pi)^{\alpha-1}d\pi. \tag{4.15}$$

- Therefore, the construction corresponds to a Poisson process with mean measure equal to that of a beta process. It's therefore a beta process.

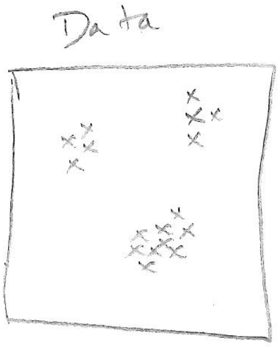
Chapter 5

Dirichlet processes and a size-biased construction

- We saw how beta processes can be useful as a Bayesian nonparametric prior for latent factor (matrix factorization) models.

- We'll next discuss BNP priors for mixture models.

Quick review



- 2-dimensional data generated from Gaussian with unknown mean and known variance.
- There are a small set of possible means and an observations picks one of them using a probability distribution.
- Let $G = \sum_{i=1}^{K} \pi_k \delta_{\theta_i}$ be the mixture distribution on mean parameters – θ_i: ith mean, π_i: probability of it

- For the nth observation,
  1. $c_n \sim \text{Disc}(\pi)$ picks mean index
  2. $x_n \sim N(\theta_{c_n}, \Sigma)$ generates observation

Priors on G

- Let μ be a non-atomic *probability* measure on the parameter space.

- Since π is a K-dimensional probability vector, a natural prior is Dirichlet.

Dirichlet distribution: A distribution on probability vectors

- Definition: Let $alpha_1, \ldots, \alpha_K$ be K positive numbers. The Dirichlet distribution density function is defined as

$$\text{Dir}(\pi|\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{K} \pi_i^{\alpha_i - 1} \tag{5.1}$$

- Goals: The goals are very similar to the beta process.
  1. We want $K \to \infty$
  2. We want the parameters $\alpha_1, \dots, \alpha_K$ to be such that, as $K \to \infty$, things are well-defined.
  3. It would be nice to link this to the Poisson process somehow.

  Dirichlet random vectors and gamma random variables

- Theorem: Let $Z_i \sim \text{Gam}(\alpha_i, b)$ for $i = 1, \dots, K$. Define $\pi_i = Z_i / \sum_{j=1}^K Z_j$ Then
$$(\pi_1, \dots, \pi_K) \sim \text{Dir}(\alpha_1, \dots, \alpha_K). \tag{5.2}$$
  Furthermore, π and $Y = \sum_{j=1}^K Z_j$ are independent random variables.

- Proof: This is just a change of variables.
  - $p(Z_1, \dots, Z_K) = \prod_{i=1}^K p(Z_i) = \prod_{i=1}^K \frac{b^{\alpha_i}}{\Gamma(\alpha_i)} Z_i^{\alpha_i - 1} e^{-bZ_i}$
  - $(Z_1, \dots, Z_K) := f(Y, \pi) = (Y\pi_1, \dots, Y\pi_{K-1}, Y(1 - \sum_{i=1}^{K-1} \pi_i))$
  - $P_{Y,\pi}(Y, \pi) = P_Z(f(Y, \pi)) \cdot |J(f)| \quad \leftarrow J(\cdot) = \text{Jacobian}$
  - $J(f) = \begin{bmatrix} \frac{\partial f_1}{\partial \pi_1} & \cdots & \frac{\partial f_1}{\partial \pi_{K-1}} & \frac{\partial f_1}{\partial Y} \\ & \ddots & & \\ \frac{\partial f_K}{\partial \pi_1} & \cdots & \frac{\partial f_K}{\partial \pi_{K-1}} & \frac{\partial f_K}{\partial Y} \end{bmatrix} = \begin{bmatrix} Y & 0 & \cdots & \pi_1 \\ 0 & Y & 0 & \pi_2 \\ 0 & 0 & \ddots & \vdots \\ -Y & -Y & \cdots & 1 - \sum_{i=1}^{K-1} \pi_i \end{bmatrix}$
  - And so $|J(f)| = Y^{K-1}$
  - Therefore
$$\begin{aligned} P_Z(f(Y,\pi))|J(f)| &= \prod_{i=1}^K \frac{b^{\alpha_i}}{\Gamma(\alpha_i)} (Y\pi_i)^{\alpha_i - 1} e^{-bY\pi_i} Y^{K-1}, \qquad (\pi_K := 1 - \sum_{i=1}^{K-1} \pi_i) \\ &= \underbrace{\left[\frac{b^{\sum_i \alpha_i}}{\Gamma(\sum_i \alpha_i)} Y^{\sum_i \alpha_i - 1} e^{-bY} \right]}_{\text{Gam}(\sum_i \alpha_i, b)} \underbrace{\left[\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^K \pi_i^{\alpha_i - 1} \right]}_{\text{Dir}(\alpha_1, \dots, \alpha_K)} \end{aligned} \tag{5.3}$$

- We've shown that:
  1. A Dirichlet distributed probability vector is a normalized sequence of independent gamma random variables with a constant scale parameter.
  2. The sum of these gamma random variables is independent of the normalization because their joint distribution can be written as a product of two distributions.
  3. This works in reverse: If we want to draw an independent sequence of gamma r.v.'s, we can draw a Dirichlet vector and scale it by an independent gamma random variable with first parameter equal to the sum of the Dirichlet parameters (and second parameter set to whatever we want).

Dirichlet process

- Definition: Let $\alpha > 0$ and μ a non-atomic *probability* measure on S. For all partitions of S, $A_1, \dots, A_k$, where $A_i \cap A_j = \emptyset$ for $i \neq j$ and $\cup_{i=1}^K A_i = S$, define the random measure G on S such that
$$(G(A_1), \dots, G(A_k)) \sim \text{Dir}(\alpha\mu(A_1), \dots, \alpha\mu(A_k)). \tag{5.4}$$
Then G is a Dirichlet process, denoted $G \sim \text{DP}(\alpha\mu)$.

Dirichlet processes via the gamma process

- Pick a partition of S, $A_1, \dots, A_k$. We can represent $G \sim \text{DP}(\alpha\mu)$ as the normalization of gamma distributed random variables,
$$(G(A_1), \dots, G(A_k)) = \Big(\frac{G'(A_1)}{G'(S)}, \dots, \frac{G'(A_k)}{G'(S)}\Big), \tag{5.5}$$
$$G'(A_i) \sim \text{Gam}(\alpha\mu(A_i), b), \qquad G'(S) = G'(\cup_{i=1}^k A_i) = \sum_{i=1}^k G'(A_i) \tag{5.6}$$

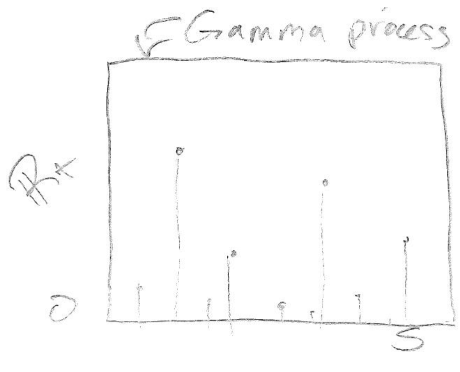
- Looking at the definition and how G' is defined, we realize that $G' \sim \text{GaP}(\alpha\mu, b)$. Therefore, a Dirichlet process is simply a normalized gamma process.

- Note that $G'(S) \sim \text{Gam}(\alpha, b)$. So $G'(S) < \infty$ with probability one and so the normalization G is well-defined.

Gamma processes and the Poisson process

- Recall that the gamma process is constructed from a Poisson process.

- Gamma process: Let N be a Poisson random measure on $S \times \mathbb{R}_+$ with mean measure $\alpha\mu(d\theta)z^{-1}e^{-bz}dz$. Define $G'(A_i) = \int_{A_i}\int_0^\infty N(d\theta, dz)z$. Then $G' \sim \text{GaP}(\alpha\mu, b)$.



- Since the DP is a rescaled GaP, this shares the same properties (from Campbells theorem)
  - For example, it's an infinite jump process.
  - However, the DP is not a CRM like the GaP since $G(A_i)$ and $G(A_j)$ are not independent for disjoint sets A_i and A_j. This should be clear since G has to integrate to 1.

Dirichlet process as limit of finite approximation

- This is very similar to the previous discussion on limits of finite approximations to the gamma and beta process.

- Definition: Let $\alpha > 0$ and μ a non-atomic probability measure on S. Let

$$G_K = \sum_{i=1}^{K} \pi_i \delta_{\theta_i}, \qquad \pi \sim \text{Dir}(\alpha/K, \dots, \alpha/K), \qquad \theta_i \overset{iid}{\sim} \mu \tag{5.7}$$

Then $\lim_{K\to\infty} G_K = G \sim \text{DP}(\alpha\mu)$.

- Rough proof: We can equivalently write

$$G_K = \sum_{i=1}^{K} \left(\frac{Z_i}{\sum_{j=1}^{K} Z_j} \right) \delta_{\theta_i}, \qquad Z_i \sim \text{Gam}(\alpha/K, b), \qquad \theta_i \sim \mu \tag{5.8}$$

- If $G'_K = \sum_{i=1}^{K} Z_i \delta_{\theta_i}$, we've already proven that $G'_K \to G' \sim \text{GaP}(\alpha\mu, b)$. G_K is thus the limit of the normalization of G'_K. Since $\lim_{K\to\infty} G'_K(S)$ is finite almost surely, we can take the limit of the numerator and denominator of the gamma representation of G_K separately. The numerator converges to a gamma process and the denominator its normalization. Therefore, G_K converges to a Dirichlet process.

Some comments

- This infinite limit of the finite approximation results in an infinite vector, but the original definition was of a K dimensional vector, so is a Dirichlet process infinite or finite dimensional? Actually, the finite vector of the definition is constructed from an infinite process:

$$G(A_j) = \lim_{K\to\infty} G_K(A_j) = \lim_{K\to\infty} \sum_{i=1}^{K} \pi_i \delta_{\theta_i}(A_j). \tag{5.9}$$

- Since the partition $A_1, \dots, A_k$ of S is of a continuous space we have to be able to let $K \to \infty$, so there has to be an infinite-dimensional process underneath G.

- The Dirichlet process gives us a way of defining priors on infinite discrete probability distributions on this continuous space S.

- As an intuitive example, if S is a space corresponding to the mean of a Gaussian, the Dirichlet process gives us a way to assign a probability to every possible value of this mean.

- Of course, by thinking of the DP in terms of the gamma and Poisson processes, we know that an infinite number of means will have probability zero, and infinite number will also have non-zero probability, but only a small handful of points in the space will have substantial probability. The number and locations of these atoms are random and learned during inference.

- Therefore, as with the beta process, size-biased representations of $G \sim \text{DP}(\alpha\mu)$ are needed.

A “stick-breaking” construction of $G \sim$ DP$(\alpha\mu)$

- Definition: Let $\alpha > 0$ and μ be a non-atomic probability measure on S. Let

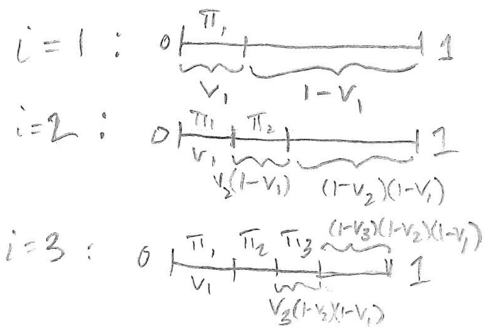
$$V_i \sim \text{Beta}(1, \alpha), \qquad \theta_i \sim \mu \tag{5.10}$$

independently for $i = 1, 2, \ldots$ Define

$$G = \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1} (1 - V_j) \delta_{\theta_i}. \tag{5.11}$$

Then $G \sim$ DP$(\alpha\mu)$.

- Intuitive picture: We start with a unit length stick and break off proportions.



$$\begin{aligned} G \;=\; & V_1 \delta_{\theta_1} + \\ & V_2(1 - V_1)\delta_{\theta_2} + \\ & V_3(1 - V_2)(1 - V_1)\delta_{\theta_3} + \cdots \end{aligned}$$

$(1 - V_2)(1 - V_1)$ is what’s left after the first two breaks. We take proportion V_3 of that for θ_3 and leave $(1 - V_3)(1 - V_2)(1 - V_1)$ ↙

Getting back to finite Dirichlets

- Recall from the definition that $(G(A_1), \ldots, G(A_K)) \sim$ Dir$(\alpha\mu(A_1), \ldots, \alpha\mu(A_K))$ for all partitions $A_1, \ldots, A_K$ of S.

- Using the stick-breaking construction, we need to show that the vector formed by

$$G(A_k) = \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1} (1 - V_j) \delta_{\theta_i}(A_k) \tag{5.12}$$

for $k = 1, \ldots, K$ is distributed as Dir$(\alpha\mu_1, \ldots, \alpha\mu_K)$, where $\mu_k = \mu(A_k)$.

- Since $P(\theta_i \in A_k) = \mu(A_k)$, $\delta_{\theta_i}(A_k)$ can be equivalently represented by a K-dimensional vector $e_{Y_i} = (0, \ldots, 1, \ldots, 0)$, with the 1 in the position Y_i and $Y_i \sim$ Disc$(\mu_1, \ldots, \mu_K)$ and the rest 0.

- Letting $\pi_i = G(A_i)$, we therefore need to show that if

$$\pi = \sum_{i=1}^{\infty} V_i \prod_{j=1}^{\infty} (1 - V_j) e_{Y_i}, \quad V_i \overset{iid}{\sim} \text{Beta}(1, \alpha), \quad Y_i \overset{iid}{\sim} \text{Disc}(\mu_1, \ldots, \mu_K) \tag{5.13}$$

Then $\pi \sim$ Dir$(\alpha\mu_1, \ldots, \alpha\mu_K)$.

- Lemma: Let $\pi \sim \text{Dir}(a_1 + b_1, \dots, a_K + b_K)$. We can equivalently represent this as

$$\begin{aligned} \pi = VY + (1-V)W, \quad V \sim \text{Beta}(\textstyle\sum_k a_k, \sum_k b_k), \\ Y \sim \text{Dir}(a_1, \dots, a_K), \quad W \sim \text{Dir}(b_1, \dots, b_K) \end{aligned} \tag{5.14}$$

Proof: Use the normalized gamma representation: $\pi_i = Z_i / \sum_j Z_j, \quad Z_i \sim \text{Gam}(a_i + b_i, c)$.

  - We can use the equivalence

$$Z_i^Y \sim \text{Gam}(a_i, c), \quad Z_i^W \sim \text{Gam}(b_i, c) \quad \Leftrightarrow \quad Z_i^Y + Z_i^W \sim \text{Gam}(a_i + b_i, c) \tag{5.15}$$

  - Splitting into two random variables this way we have the following normalized gamma representation for π

$$\begin{aligned} \pi &= \left(\frac{Z_1^Y + Z_1^W}{\sum_i Z_i^Y + Z_i^W}, \dots, \frac{Z_K^Y + Z_K^W}{\sum_i Z_i^Y + Z_i^W} \right) \qquad (5.16) \\ &= \underbrace{\left(\frac{\sum_i Z_i^Y}{\sum_i Z_i^Y + Z_i^W} \right)}_{V \sim \text{Beta}(\sum_i a_i, \sum_i b_i)} \underbrace{\left(\frac{Z_1^Y}{\sum_i Z_i^Y}, \dots, \frac{Z_K^Y}{\sum_i Z_i^Y} \right)}_{Y \sim \text{Dir}(a_1, \dots, a_K)} \\ &\quad + \underbrace{\left(\frac{\sum_i Z_i^W}{\sum_i Z_i^Y + Z_i^W} \right)}_{1-V} \underbrace{\left(\frac{Z_1^W}{\sum_i Z_i^W}, \dots, \frac{Z_K^W}{\sum_i Z_i^W} \right)}_{W \sim \text{Dir}(b_1, \dots, b_K)} \end{aligned}$$

  - From the previous proof about normalized gamma r.v.'s, we know that the sums are independent from the normalized values. So V, Y, and W are all independent.

- *Proof of stick-breaking construction*:

  - Start with $\pi \sim \text{Dir}(\alpha\mu_1, \dots, \alpha\mu_K)$.

  - Step 1:

$$\begin{aligned} \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^K \pi_i^{\alpha_i - 1} &= \left(\sum_{j=1}^K \pi_j \right) \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^K \pi_i^{\alpha_i - 1} \qquad (5.17) \\ &= \sum_{j=1}^K \frac{\alpha\mu_j}{\alpha\mu_j} \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^K \pi_i^{\alpha_i + e_j(i) - 1} \\ &= \sum_{j=1}^K \mu_j \underbrace{\frac{\Gamma(1 + \sum_i \alpha_i)}{\Gamma(1 + \alpha\mu_j) \prod_{i \neq j} \Gamma(\alpha_i)} \prod_{i=1}^K \pi_i^{\alpha_i + e_j(i) - 1}}_{= \text{Dir}(\alpha\mu + e_j)} \end{aligned}$$

  - Therefore, a hierarchical representation of $\text{Dir}(\alpha\mu_1, \dots, \alpha\mu_K)$ is

$$Y \sim \text{Discrete}(\mu_1, \dots, \mu_K), \quad \pi \sim \text{Dir}(\alpha\mu + e_Y). \tag{5.18}$$

- Step 2:
  From the lemma we have that $\pi \sim \text{Dir}(\alpha\mu + e_Y)$ can be expanded into the equivalent hierarchical representation $\pi = VY' + (1-V)\pi'$, where

$$V \sim \text{Beta}(\underbrace{\textstyle\sum_i e_Y(i)}_{=1}, \underbrace{\textstyle\sum_i \alpha\mu_i}_{=\alpha}), \quad \underbrace{Y' \sim \text{Dir}(e_Y)}_{=e_Y \text{ with probability } 1}, \quad \pi' \sim \text{Dir}(\alpha\mu_1, \dots, \alpha\mu_K) \quad (5.19)$$

- Combining Steps 1& 2:
  We will use these steps to recursively break down a Dirichlet distributed random vector an infinite number of times. If

$$\pi = Ve_Y + (1-V)\pi', \quad (5.20)$$

$$V \sim \text{Beta}(1,\alpha), \quad Y \sim \text{Disc}(\mu_1, \dots, \mu_K), \quad \pi' \sim \text{Dir}(\alpha\mu_1, \dots, \alpha\mu_K),$$

  Then from steps 1 & 2, $\pi \sim \text{Dir}(\alpha\mu_1, \dots, \alpha\mu_K)$.

- Notice that there are independent $\text{Dir}(\alpha\mu_1, \dots, \alpha\mu_K)$ r.v.'s on both sides. We "broke down" the one on the left, we can continue by "breaking down" the one on the right:

$$\pi = V_1 e_{Y_1} + (1-V_1)(V_2 e_{Y_2} + (1-V_2)\pi'') \quad (5.21)$$

$$V_i \overset{iid}{\sim} \text{Beta}(1,\alpha), \quad Y_i \overset{iid}{\sim} \text{Disc}(\mu_1, \dots, \mu_K), \quad \pi'' \sim \text{Dir}(\alpha\mu_1, \dots, \alpha\mu_K),$$

  π is still distributed as $\text{Dir}(\alpha\mu_1, \dots, \alpha\mu_K)$.

- Continue this an infinite number of times:

$$\pi = \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1} e_{Y_i}, \quad V_i \overset{iid}{\sim} \text{Beta}(1,\alpha), \quad Y_i \overset{iid}{\sim} \text{Disc}(\mu_1, \dots, \mu_K). \quad (5.22)$$

  Still, following each time the right-hand Dirichlet is expanded we get a $\text{Dir}(\alpha\mu_1, \dots, \alpha\mu_K)$ random variable. Since $\lim_{T\to\infty} \prod_{j=1}^{T}(1-V_j) = 0$, the term pre-multiplying this RHS Dirichlet vector equals zero and the limit above results, which completes the proof.

- Corollary:

  If G is drawn from $\text{DP}(\alpha\mu)$ using the stick-breaking construction and $\beta \sim \text{Gam}(\alpha, b)$ independently, then $\beta G \sim \text{GaP}(\alpha\mu, b)$. Writing this out,

$$\beta G = \sum_{i=1}^{\infty} \beta \Big(V_i \prod_{j=1}^{i-1}(1-V_j) \Big) \delta_{\theta_i} \quad (5.23)$$

- We therefore get a method for drawing a gamma process almost for free. Notice that α appears in both the DP and gamma distribution on β. These parameters must be the same value for βG to be a gamma process.

Chapter 6

Dirichlet process extensions, count processes

Gamma process to Dirichlet process

- Gamma process: Let $\alpha > 0$ and μ a non-atomic probability measure on S. Let $N(d\theta, dw)$ be a Poisson random measure on $S \times \mathbb{R}_+$ with mean measure $\alpha\mu(d\theta)we^{-cw}dw$, $c > 0$. For $A \subset S$, let $G'(A) = \int_A \int_0^\infty N(d\theta, dw)w$. Then G' is a gamma process, $G' \sim$ GaP$(\alpha\mu, c)$, and $G'(A) \sim$ Gam$(\alpha\mu(A), c)$.

- Normalizing a gamma process: Let's take G' and normalize it. That is, define $G(d\theta) = G'(d\theta)/G'(S)$. ($G'(S) \sim$ Gam(α, c), so it's finite w.p. 1). Then G is called a Dirichlet process, written $G \sim$ DP$(\alpha\mu)$.

  Why? Take S and partition it into K disjoint regions, i.e., $(A_1, \dots, A_K)$, $A_i \cap A_j = \emptyset$, $i \neq j$, $\cup_i A_i = S$. Construct the vector

$$(G(A_1), \dots, G(A_K)) = \left(\frac{G'(A_1)}{G'(S)}, \dots, \frac{G'(A_K)}{G'(S)} \right). \tag{6.1}$$

  Since each $G'(A_i) \sim$ Gam$(\alpha\mu(A_i), c)$, and $G'(S) = \sum_{i=1}^K G'(A_i)$, it follows that

$$(G(A_1), \dots, G(A_K)) \sim \text{Dir}(\alpha\mu(A_1), \dots, \alpha\mu(A_K)). \tag{6.2}$$

  This is the *definition* of a Dirichlet process.

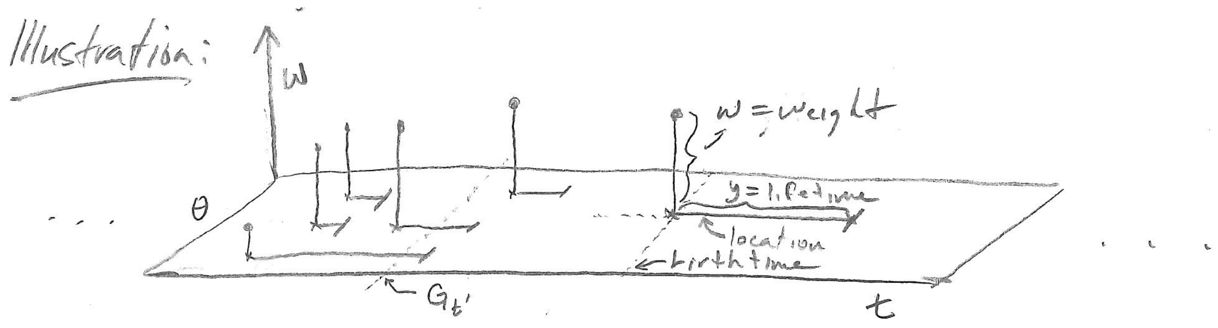
- The Dirichlet process has many extensions to suit the structure of different problems.
- We'll look at four, two that are related to the underlying normalized gamma process, and two from the perspective of the stick-breaking construction.
- The purpose is to illustrate how the basic framework of Dirichlet process mixture modeling can be easily built into more complicated models that address problems not perfectly suited to the basic construction.
- Goal is to make it clear how to continue these lines of thinking to form new models.

Example 1: Spatially and temporally normalized gamma processes

- Imagine we wanted a temporally evolving Dirichlet process. Clusters (i.e., atoms, θ) may arise and die out at different times (or exist in geographical regions)

Time-evolving model: Let $N(d\theta, dw, dt)$ be a Poisson random measure on $S \times \mathbb{R}_+ \times \mathbb{R}$ with mean measure $\alpha\mu(d\theta)w^{-1}e^{-cw}dwdt$. Let $G'(d\theta, dt) = \int_0^\infty N(d\theta, dw, dt)w$. Then G' is a gamma process with added "time" dimension t.

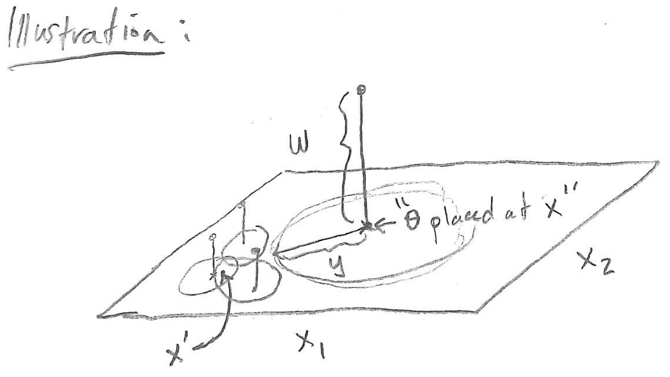
- (There's nothing new from what we've studied: Let $\theta^* = (\theta, t)$ and $\alpha\mu(d\theta^*) = \alpha\mu(d\theta)dt$.)
- For each atom (θ, t) with $G'(d\theta, dt) > 0$, add a marking $y_t(\theta) \overset{ind}{\sim}$ Exp(λ).
- We can think of $y_t(\theta)$ as the lifetime of parameter θ born at time t.
- By the marking theorem, $N^*(d\theta, dw, dt, dy) \sim$ Pois $\left(\alpha\mu(d\theta)w^{-1}e^{-cw}dwdt\lambda e^{-\lambda y}dy\right)$



- At time t', construct the Dirichlet process G'_t by normalizing over all atoms "alive" at time t. (Therefore, ignore atoms already dead or yet to be born.)

Spatial model: Instead of giving each atom θ a time-stamp and "lifetime," we might want to give it a location and "region of influence".

- Replace $t \in \mathbb{R}$ with $x \in \mathbb{R}^2$ (e.g., latitude-longitude). Replace dt with dx.
- Instead of $y_t(\theta) \sim$ Exp(λ) = lifetime, $y_x(\theta) \sim$ Exp(λ) = radius of ball at x.



- G'_x is the DP at location x'.
- It is formed by normalizing over all atoms θ for which

$$x' \in \text{ ball of radius } y_x(\theta) \text{ at } x$$

Example 2: Another time-evolving formulation

- We can think of other formulations. Here's one where time is discrete. (We will build up to this with the following two properties).

- Even though the DP doesn't have an underlying PRM, the fact that it's constructed from a PRM means we can still benefit from its properties.

Superposition and the Dirichlet process: Let $G_1' \sim \text{GaP}(\alpha_1\mu_1, c)$ and $G_2' \sim \text{GaP}(\alpha_2\mu_2, c)$. Then $G_{1+2}' = G_1' + G_2' \sim \text{GaP}(\alpha_1\mu_1 + \alpha_2\mu_2, c)$. Therefore,

$$G_{1+2} = \frac{G_{1+2}'}{G_{1+2}'(S)} \sim \text{DP}(\alpha_1\mu_1 + \alpha_2\mu_2). \tag{6.3}$$

We can equivalently write

$$G_{1+2} = \underbrace{\frac{G_1'(S)}{G_{1+2}'(S)}}_{\text{Beta}(\alpha_1,\alpha_2)} \times \underbrace{\frac{G_1'}{G_1'(S)}}_{\text{DP}(\alpha_1\mu_1)} + \frac{G_2'(S)}{G_{1+2}'(S)} \times \underbrace{\frac{G_2'}{G_2'(S)}}_{\text{DP}(\alpha_2\mu_2)} \sim \text{DP}(\alpha_1\mu_1 + \alpha_2\mu_2) \tag{6.4}$$

From the lemma last week, these two DP's and the beta r.v. are all independent.

- Therefore,

$$G = \pi G_1 + (1-\pi)G_2, \quad \pi \sim \text{Beta}(\alpha_1, \alpha_2), \quad G_1 \sim \text{DP}(\alpha_1\mu_1), \quad G_2 \sim \text{DP}(\alpha_2\mu_2) \tag{6.5}$$

is equal in distribution to $G \sim \text{DP}(\alpha_1\mu_1 + \alpha_2\mu_2)$.

Thinning of gamma processes (a special case of the marking theorem)

- We know that we can construct $G' \sim \text{GaP}(\alpha\mu, c)$ from the Poisson random measure $N(d\theta, dw) \sim \text{Pois}\,(\alpha\mu(d\theta)w^{-1}e^{-cw}dw)$. Mark each point (θ, w) in N with a binary variable $z \sim \text{Bern}(p)$. Then
$$N(d\theta, dw, z) \sim \text{Pois}\left(p^z(1-p)^z\alpha\mu(d\theta)w^{-1}e^{-cw}dw\right). \tag{6.6}$$

- If we view $z = 1$ as "survival" and $z = 0$ as "death," then if we only care about the atoms that survive, we have

$$N_1(d\theta, dw) = N(d\theta, dw, z = 1) \sim \text{Pois}(p\alpha\mu(d\theta)w^{-1}e^{-cw}dw). \tag{6.7}$$

- This is called "thinning." We see that $p \in (0, 1)$ down-weights the mean measure, so we only expect to see a fraction p of what we saw before.

- Still, a normalized thinned gamma process is a Dirichlet process

$$\dot{G}' \sim \text{GaP}(p\alpha\mu, c), \quad \dot{G} = \frac{\dot{G}'}{\dot{G}'(S)} \sim \text{DP}(p\alpha\mu). \tag{6.8}$$

- What happens if we thin twice? We're marking with $z \in \{0,1\}^2$ and restricting to $z = [1,1]$,

$$\ddot{G}' \sim \text{GaP}(p^2\alpha\mu, c) \quad \rightarrow \quad \ddot{G} \sim \text{DP}(p^2\alpha\mu). \tag{6.9}$$

- Back to the example, we again want a time-evolving Dirichlet process where new atoms are born and old atoms die out.

- We can easily achieve this by introducing new gamma processes and thinning old ones.

A dynamic Dirichlet process:

At time t: 1. Draw $G_t^* \sim \text{GaP}(\alpha_t\mu_t, c)$.

2. Construct $G_t' = G_t^* + \dot{G}_{t-1}'$, where $\dot{G}_{t-1}'$ is the gamma process at time $t-1$ thinned with parameter p.

3. Normalize $G_t = G_t'/G_t'(S)$.

- Why is G_t still a Dirichlet process? Just look at G_t':
  - Let $G_{t-1}' \sim \text{GaP}(\hat{\alpha}_{t-1}\hat{\mu}_{t-1}, c)$.
  - Then $\dot{G}_{t-1}' \sim \text{GaP}(p\hat{\alpha}_{t-1}\hat{\mu}_{t-1}, c)$ and $G_t' \sim \text{GaP}(\alpha_t\mu_t + p\hat{\alpha}_{t-1}\hat{\mu}_{t-1}, c)$.
  - So $G_t \sim \text{DP}(\alpha_t\mu_t + p\hat{\alpha}_{t-1}\hat{\mu}_{t-1})$.

- By induction,

$$G_t \sim \text{DP}(\alpha_t\mu_t + p\alpha_{t-1}\mu_{t-1} + p^2\alpha_{t-2}\mu_{t-2} + \cdots + p^{t-1}\alpha_1\mu_1). \tag{6.10}$$

- If we consider the special case where $\alpha_t\mu_t = \alpha\mu$ for all t, we can simplify this Dirichlet process

$$G_t \sim \text{DP}\left(\frac{1-p^t}{1-p}\alpha\mu\right). \tag{6.11}$$

In the limit $t \rightarrow \infty$, this has the steady state

$$G_\infty \sim \text{DP}\left(\frac{1}{1-p}\alpha\mu\right). \tag{6.12}$$

- Stick-breaking construction (review for the next process)

We saw that if $\alpha > 0$ and μ is any probability measure, atomic or non-atomic or mixed, then we can draw $G \sim \text{DP}(\alpha\mu)$ as follows:

$$V_i \overset{iid}{\sim} \text{Beta}(1,\alpha), \qquad \theta_i \overset{iid}{\sim} \mu, \qquad G = \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1}(1-V_j)\delta_{\theta_i} \tag{6.13}$$

- It's often the case where we have grouped data. For example, groups of documents where each document is a set of words.

- We might want to model each group (indexed by d) as a mixture G_d. Then, for observation n in group d, $\theta_n^{(d)} \sim G_d, x_n^{(d)} \sim p(x|\theta_n^{(d)})$.

- We might think that each group shares the same set of highly probable atoms, but has different distributions on them.

- The result is called a mixed-membership model.

Mixed-membership models and the hierarchical Dirichlet process (HDP)

- As the stick-breaking construction makes clear, when μ is non-atomic simply drawing each $G_d \overset{iid}{\sim} \text{DP}(\alpha\mu)$ won't work because it places all probability mass on a disjoint set of atoms.

- The HDP fixes this by "discretizing the base distribution."

$$G_d \,|\, G_0 \overset{iid}{\sim} \text{DP}(\beta G_0), \qquad G_0 \sim \text{DP}(\alpha\mu). \tag{6.14}$$

- Since G_0 is discrete, G_d has probability on the same subset of atoms. This is very obvious by writing the process with the stick-breaking construction:

$$G_d = \sum_{i=1}^{\infty} \pi_i^{(d)} \delta_{\theta_i}, \qquad (\pi_1^{(d)}, \pi_2^{(d)}, \dots) \sim \text{Dir}(\alpha p_1, \alpha p_2, \dots) \tag{6.15}$$

$$p_i = V_i \prod_{j=1}^{i-1}(1 - V_j), \qquad V_i \overset{iid}{\sim} \text{Beta}(1, \alpha), \qquad \theta_i \overset{iid}{\sim} \mu.$$

- Nested Dirichlet processes

  The stick-breaking construction is totally general: μ can be any distribution.

  What if $\mu \to \text{DP}(\alpha\mu)$? That is, we define the base distribution to be a Dirichlet process.

$$G \sim \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1}(1 - V_j)\delta_{G_i}, \quad V_i \overset{iid}{\sim} \text{Beta}(1, \alpha), \quad G_i \overset{iid}{\sim} \text{DP}(\alpha\mu). \tag{6.16}$$

  (We write G_i to link to the DP, but we could have written $\theta_i \overset{iid}{\sim} \text{DP}(\alpha\mu)$ since that's what we've been using.)

- We now have a mixture model of mixture models. For example:

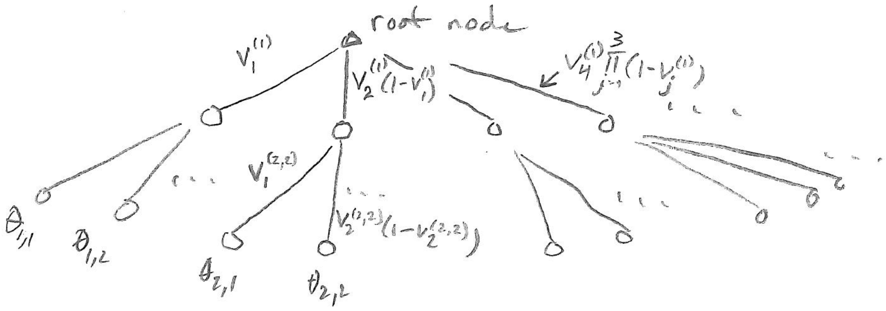
  1. A group selects $G^{(d)} \sim G$ (picks mixture G_i according to probability $V_i \prod_{j<i}(1 - V_j)$).

2. Generates all of its data using this mixture. For the nth observation in group d, $\theta_n^{(d)} \sim G^{(d)}$, $X_n^{(d)} \sim p(X|\theta_n^{(d)})$.

- In this case we have all-or-nothing sharing. Two groups either share the atoms <u>and</u> the distribution on them, or they share nothing.

<u>Nested Dirichlet process trees</u>

- We can nest this further. Why not let μ in the nDP be a Dirichlet process also? Then we would have a three level tree.



- We can then pick paths down this tree to a leaf node where we get an atom.

<u>Count Processes</u>

- We briefly introduce count processes. With the Dirichlet process, we often have the generative structure

$$G \sim \text{DP}(\alpha\mu), \quad \theta_j^*|G \overset{iid}{\sim} G, \quad G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}, \quad j = 1, \ldots, N \tag{6.17}$$

- What can we say about the count process $n(\theta) = \sum_{j=1}^{N} \mathbb{1}(\theta_j^* = \theta)$?

- Recall the following equivalent processes:

$$\begin{aligned} G' &\sim \text{GaP}(\alpha\mu, c) \; (6.18) \\ n(\theta)|G' &\sim \text{Pois}(G'(\theta)) \, (6.19) \end{aligned} \qquad \underline{\text{and}} \qquad \begin{aligned} G' &\sim \text{GaP}(\alpha\mu, c) \;\; (6.20) \\ n(S) &\sim \text{Pois}(G'(S)) \; (6.21) \\ \theta_{1:n(S)}^* &\sim G'/G'(S) \quad (6.22) \end{aligned}$$

- We can therefore analyze this using the underlying marked Poisson process. However, notice that we have to let the data size be random and Poisson distributed.

<u>Marking theorem:</u> Let $G' \sim \text{GaP}(\alpha\mu, c)$ and mark each (θ, w) for which $G'(\theta) = w > 0$ with the random variable $n|w \sim \text{Pois}(w)$. Then (θ, w, n) is a marked Poisson process with mean measure $\alpha\mu(d\theta)w^{-1}e^{-cw}dw\frac{w^n}{n!}e^{-w}$.

- We can restrict this to n by integrating over θ and w.

Theorem: The number of atoms having k counts is

$$\#_k \sim \text{Pois}\Big(\int_S \int_0^\infty \alpha\mu(d\theta)w^{-1}e^{-cw}dw\frac{w^k}{k!}e^{-w}\Big) = \text{Pois}\Big(\frac{\alpha}{k}\Big(\frac{1}{1+c}\Big)^k\Big) \tag{6.23}$$

Theorem: The total number of uniquely observed atoms is also Poisson

$$\#_{\text{unique}} = \sum_{k=1}^\infty \#_k \sim \text{Pois}\Big(\sum_{k=1}^\infty \frac{\alpha}{k}\Big(\frac{1}{1+c}\Big)^k\Big) = \text{Pois}(\alpha \ln(1+\frac{1}{c})) \tag{6.24}$$

Theorem: The total number of counts is $n(S)|G' \sim \text{Pois}(G'(S))$, $G' \sim \text{GaP}(\alpha\mu, c)$. So $\mathbb{E}[n(S)] = \frac{\alpha}{c}$ ($= \sum_{k=1}^\infty k\mathbb{E}\#_k$)

Final statement: Let $c = \frac{\alpha}{N}$. If we expect a dataset of size N to be drawn from $G \sim \text{DP}(\alpha\mu)$, we expect that dataset to use $\alpha\ln(\alpha+N) - \alpha\ln\alpha$ unique atoms from G.

A quick final count process

- Instead of gamma process $\longrightarrow$ Poisson counts, we could have beta process $\longrightarrow$ negative binomial counts.

- Let $H = \sum_{i=1}^\infty \pi_i\delta_{\theta_i} \sim \text{BP}(\alpha, \mu)$.

- Let $n(\theta) = \text{negBin}(r, H(\theta))$, where the negative binomial random variable counts how many "successes" there are, with $P(\text{success}) = H(\theta)$ until there are r "failures" with $P(\text{failure}) = 1 - H(\theta)$.

- This is another count process that can be analyzed using the underlying Poisson process.

Chapter 7

Exchangeability, Dirichlet processes and the Chinese restaurant process

DP's, finite approximations and mixture models

- DP: We saw how, if $\alpha > 0$ and μ is a probability measure on S, for every finite partition $(A_1, \ldots, A_k)$ of S, the random measure

  $$(G(A_1)), \ldots, G(A_k)) \sim \text{Dir}(\alpha\mu(A_1), \ldots, \alpha\mu(A_k))$$

  defines a Dirichlet process.

- Finite approximation: We also saw how we can approximate $G \sim \text{DP}(\alpha\mu)$ with a finite Dirichlet distribution,

  $$G_K = \sum_{i=1}^{K} \pi_i \delta_{\theta_i}, \quad \pi_i \sim \text{Dir}\Big(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\Big), \quad \theta \overset{iid}{\sim} \mu.$$

- Mixture models: Finally, the most common setting for these priors is in mixture models, where we have the added layers

  $$\theta_j^* | G \sim G, \quad X_j | \theta_j^* \sim p(X|\theta_j^*), \;\; j = 1, \ldots, n. \qquad (Pr(\theta_j^* = \theta_i | G) = \pi_i)$$

- The values of $\theta_1^*, \ldots, \theta_n^*$ induce a clustering of the data.

- If $\theta_j^* = \theta_{j'}^*$ for $j \neq j'$ then X_j and $X_{j'}$ are "clustered" together since they come from the same distribution.

- We've thus far focused on G. We now focus on the clustering of $X_1, \ldots, X_n$ induced by G.

Polya's Urn model (finite Dirichlet)

- To simplify things, we work in the finite setting and replace the parameter θ_i with its index i. We let the indicator variables $c_1, \ldots, c_n$ represent $\theta_1^*, \ldots, \theta_n^*$ such that $\theta_j^* = \theta_{c_j}$.

- Polya's Urn is the following process for generating $c_1, \dots, c_n$:

  1. For the first indicator, $c_1 \sim \sum_{i=1}^{K} \frac{1}{K}\delta_i$
  2. For the nth indicator, $c_n | c_1, \dots, c_{n-1} \sim \sum_{j=1}^{n-1} \frac{1}{\alpha+n-1}\delta_{c_j} + \frac{\alpha}{\alpha+n-1}\sum_{i=1}^{K}\frac{1}{K}\delta_i$

- In words, we start with an urn having $\frac{\alpha}{K}$ balls of color i for each of K colors. We randomly pick a ball, put it back in the urn and put another ball of the same color in the urn.

- Another way to write #2 above is to define $n_i^{(n-1)} = \sum_{j=1}^{n-1} \mathbb{1}(c_j = i)$. Then

$$c_n | c_1, \dots, c_{n-1} \sim \sum_{i=1}^{K} \frac{\frac{\alpha}{K} + n_i^{(n-1)}}{\alpha + n - 1}\delta_i.$$

  To put it most simply, we're just sampling the next color from the empirical distribution of the urn at step n.

- What can we say about $p(c_1 = i_1, \dots, c_n = i_n)$ (write as $p(c_1, \dots, c_n)$) under this prior?

  1. By the chain rule of probability, $p(c_1, \dots, c_n) = \prod_{j=1}^{n} p(c_j | c_1, \dots, c_{j-1})$.
  2. $p(c_j = i | c_1, \dots, c_{j-1}) = \dfrac{\frac{\alpha}{K} + n_i^{(j-1)}}{\alpha + j - 1}$
  3. Therefore,

$$p(c_{1:n}) \;=\; p(c_1)p(c_2|c_1)p(c_3|c_1,c_2)\cdots \;=\; \prod_{j=1}^{n} \frac{\frac{\alpha}{K} + n_{c_j}^{(j-1)}}{\alpha + j - 1} \tag{7.1}$$

- A few things to notice about $p(c_1, \dots, c_n)$

  1. The denominator is simply $\prod_{j=1}^{n}(\alpha + j - 1)$
  2. $n_{c_j}^{(j-1)}$ is incrementing by one. That is, after $c_{1:n}$ we have the counts $(n_1^{(n)}, \dots, n_K^{(n)})$. For each $n_i^{(n)}$ the numerator will contain $\prod_{s=1}^{n_i^{(n)}}(\frac{\alpha}{K} + s - 1)$.
  3. Therefore,

$$p(c_1, \dots, c_n) \;=\; \frac{\prod_{i=1}^{K}\prod_{s=1}^{n_i^{(n)}}(\frac{\alpha}{K} + s - 1)}{\prod_{j=1}^{n}(\alpha + j - 1)} \tag{7.2}$$

- Key: The key thing to notice is that this does not depend on the order of $c_1, \dots, c_n$. That is, if we permuted $c_1, \dots, c_n$ such that $c_j = i_{\rho(j)}$, where $\rho(\cdot)$ is a permutation of $(1, \dots, n)$, then

$$p(c_1 = i_1, \dots, c_n = i_n) \;=\; p(c_1 = i_{\rho(1)}, \dots, c_n = i_{\rho(n)}).$$

- The sequence $c_1, \dots, c_n$ is said to be "exchangeable" in this case.

Exchangeability and independent and identically distributed (iid) sequences

- Independent sequences are exchangeable

$$p(c_1,\dots,c_n) = \prod_{i=1}^{n} p(c_i) = \prod_{i=1}^{n} p(c_{\rho(i)}) = p(c_{\rho(1)},\dots,c_{\rho(n)}). \tag{7.3}$$

- Exchangeable sequences aren't necessarily independent (exchangeability is "weaker"). Think of the urn. c_j is clearly not independent of $c_1,\dots,c_{j-1}$.

Exchangeability and de Finetti's

- de Finetti's theorem: A sequence is exchangeable if and only if there is a parameter π with distribution $p(\pi)$ for which the sequence is independent and identically distributed given π.

- In other words, for our problem there is a probability vector π such that $p(c_{1:n}|\pi) = \prod_{j=1}^{n} p(c_j|\pi)$.

- The problem is to find $p(\pi)$

$$\begin{aligned}
p(c_1,\dots,c_n) &= \int p(c_1,\dots,c_n|\pi)p(\pi)d\pi \\
&= \int \prod_{j=1}^{n} p(c_j|\pi)p(\pi)d\pi \\
&= \int \prod_{j=1}^{n} \pi_{c_j} p(\pi)d\pi \\
&= \int \prod_{i=1}^{k} \pi_i^{n_i^{(n)}} p(\pi)d\pi \\
\downarrow \qquad & \qquad\quad \downarrow \\
\frac{\prod_{i=1}^{K}\prod_{s=1}^{n_i^{(n)}}(\frac{\alpha}{K}+s-1)}{\prod_{j=1}^{n}(\alpha+j-1)} &= \mathbb{E}_{p(\pi)}\left[\prod_{i=1}^{K} \pi_i^{n_i^{(n)}}\right] \qquad (7.4)
\end{aligned}$$

- Above, the first equality is always true. The second one is by de Finetti's theorem since $c_1,\dots,c_n$ is exchangeable. (We won't proven this theorem, we'll just use it.) The following results. In the last equality, the left hand side was previously shown and the right hand side is what the second to last line is equivalently written as.

- By de Finetti and exchangeability of $c_1,\dots,c_n$, we therefore arrive at an expression for the moments of π according to the still unknown distribution $p(\pi)$.

- Because the moments of a distribution are unique to that distribution (like the Laplace transform), $p(\pi)$ has to be Dir$(\frac{\alpha}{K}, \dots, \frac{\alpha}{K})$, since plugging this in for $p(\pi)$ we get

$$\begin{aligned}
\mathbb{E}_{p(\pi)}\left[\prod_{i=1}^K \pi_i^{n_i^{(n)}}\right] &= \int \prod_{i=1}^K \pi_i^{n_i} \frac{\Gamma(\alpha)}{\Gamma(\frac{\alpha}{K})^K} \prod_{i=1}^K \pi_i^{\frac{\alpha}{K}-1} d\pi \\
&= \frac{\Gamma(\alpha)\prod_{i=1}^K \Gamma(\frac{\alpha}{K}+n_i)}{\Gamma(\frac{\alpha}{K})^K \Gamma(\alpha+n)} \underbrace{\int \frac{\Gamma(\alpha+n)}{\prod_{i=1}^K \Gamma(\frac{\alpha}{K}+n_i)} \prod_{i=1}^K \pi^{n_i+\frac{\alpha}{K}-1} d\pi}_{=\text{Dir}(\frac{\alpha}{K}+n_1,\dots,\frac{\alpha}{K}+n_k)} \\
&= \frac{\Gamma(\alpha)\prod_{i=1}^K \Gamma(\frac{\alpha}{K})\prod_{s=1}^{n_i}(\frac{\alpha}{K}+s-1)}{\Gamma(\frac{\alpha}{K})^K \Gamma(\alpha)\prod_{j=1}^n(\alpha+j-1)} \\
&= \frac{\prod_{i=1}^K \prod_{s=1}^{n_i^{(n)}}(\frac{\alpha}{K}+s-1)}{\prod_{j=1}^n(\alpha+j-1)} \qquad (7.5)
\end{aligned}$$

- This holds for all n and $(n_1, \dots, n_k)$. Since a distribution is defined by its moments, the result follows.

- Notice that we didn't *need* de Finetti since we could just hypothesize the existence of a π for which $p(c_{1:n}|\pi) = \prod_i p(c_i|\pi)$. It's more useful when the distribution is more "non-standard," or to prove that a π doesn't exist.

- Final statement: As $n \to \infty$, the distribution $\sum_{i=1}^K \frac{n_i^{(n)}+\frac{\alpha}{K}}{\alpha+n}\delta_i \to \sum_{i=1}^K \pi_i^* \delta_i$.
  - This is because there exists a π for which $c_1, \dots, c_n$ are iid, and so by the law of large numbers the point π^* exists and $\pi^* = \pi$.
  - Since $\pi \sim$ Dir$(\frac{\alpha}{K}, \dots, \frac{\alpha}{K})$, it follows that the empirical distribution converges to a *random* vector that is distributed as Dir$(\frac{\alpha}{K}, \dots, \frac{\alpha}{K})$.

The infinite limit (Chinese restaurant process)

- Let's go back to the original notation:

$$\theta_j^*|G_K \sim G_K, \quad G_K = \sum_{i=1}^K \pi_i \delta_{\theta_i}, \quad \pi \sim \text{Dir}(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}), \quad \theta_i \overset{iid}{\sim} \mu.$$

- Following the exact same ideas (only changing notation). The urn process is

$$\theta_n^*|\theta_1^*, \dots, \theta_{n-1}^* \sim \sum_{i=1}^K \frac{\frac{\alpha}{K} = n_i^{(n-1)}}{\alpha+n-1}\delta_{\theta_i}, \quad \theta_i \overset{iid}{\sim} \mu.$$

- We've proven that $\lim_{K\to\infty} G_K = G \sim$ DP$(\alpha\mu)$. We now take the limit of the corresponding urn process.

- Re-indexing: At observation n, re-index the atoms so that $n_j^{(n-1)} > 0$ for $j = 1, \ldots, K_{n-1}^+$ and $n_j^{(n-1)} = 0$ for $j > K_{n-1}^+$. ($K_{n-1}^+ =$ # unique values in $\theta_{1:n-1}^*$) Then

$$\theta_n^* | \theta_1^*, \ldots, \theta_{n-1}^* \sim \sum_{i=1}^{K_{n-1}^+} \frac{\frac{\alpha}{K} + n_i^{(n-1)}}{\alpha + n - 1} \delta_{\theta_i} + \frac{\alpha}{\alpha + n - 1} \sum_{i=1+K_{n-1}^+}^{K} \frac{1}{K} \delta_{\theta_i}. \tag{7.6}$$

- Obviously for $n \gg K$, $K_{n-1}^+ = K$ very probably, and just the left term remains. However, we're interested in $K \to \infty$ before we let n grow. In this case

  1. $\dfrac{\frac{\alpha}{K} + n_i^{(n-1)}}{\alpha + n - 1} \longrightarrow \dfrac{n_i^{(n-1)}}{\alpha + n - 1}$
  2. $\displaystyle\sum_{i=1+K_{n-1}^+}^{K} \frac{1}{K} \delta_{\theta_i} \longrightarrow \mu.$

- For #2, if you sample K times from a distribution and create a uniform measure on those samples, then in the infinite limit you get the original distribution back. Removing $K_{n-1}^+ < \infty$ of those atoms doesn't change this (we won't prove this).

The Chinese restaurant process

- Let $\alpha > 0$ and μ a probability measure on S. Sample the sequence $\theta_1^*, \ldots, \theta_n^*$, $\theta^* \in S$ as follows:

  1. Set $\theta_1^* \sim \mu$
  2. Sample $\theta_n^* | \theta_1^*, \ldots, \theta_{n-1}^* \sim \sum_{j=1}^{n-1} \frac{1}{\alpha+n-1} \delta_{\theta_j^*} + \frac{\alpha}{\alpha+n-1} \mu$

  Then the sequence $\theta_1^*, \ldots, \theta_n^*$ is a Chinese restaurant process.

- Equivalently define $n_i^{(n-1)} = \sum_{j=1}^{n-1} \mathbb{1}(\theta_j^* = \theta_j)$. Then

$$\theta_n^* | \theta_1^*, \ldots, \theta_{n-1}^* \sim \sum_{i=1}^{K_{n-1}^+} \frac{n_i^{(n-1)}}{\alpha + n - 1} \delta_{\theta_i} + \frac{\alpha}{\alpha + n - 1} \mu.$$

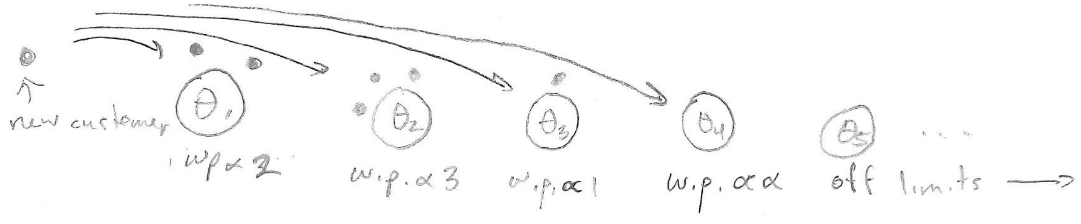
- As $n \to \infty$, $\frac{\alpha}{\alpha+n-1} \to 0$ and

$$\sum_{i=1}^{K_{n-1}^+} \frac{n_i^{(n-1)}}{\alpha + n - 1} \delta_{\theta_i} \longrightarrow G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i} \sim \text{DP}(\alpha\mu). \tag{7.7}$$

- Notice with the limits that K first went to infinity and then n went to infinity. The resulting empirical distribution is a Dirichlet process because for finite K the de Finetti mixing measure is Dirichlet and the infinite limit of this finite Dirichlet is a Dirichlet process (as we've shown).

Chinese restaurant analogy

- An infinite sequence of tables each have a dish (parameter) placed on it that is drawn iid from μ. The nth customer sits at an occupied table with probability proportional to the number of customers seated there, or selects the first unoccupied table with probability proportional to α.



- "Dishes" are parameters for distributions, a "customer" is a data point that uses its dish to create its value.

Some properties of the CRP

- Cluster growth: What does the number of unique clusters K_n^+ look like as a function of n?

$K_n^+ = \sum_{j=1}^n \mathbb{1}(\theta_j^* \neq \theta_{\ell<j}^*)$ ← this event occurs when we select a new table

$$\mathbb{E}[K_n^+] = \sum_{j=1}^n \mathbb{E}[\mathbb{1}(\theta_j^* \neq \theta_{\ell<j}^*)] = \sum_{j=1}^n P(\theta_j^* \neq \theta_{\ell<j}^*) = \sum_{j=1}^n \frac{\alpha}{\alpha+j-1} \approx \alpha \ln n$$

Where does this come from? Each θ_j^* can pick a new table. θ_j^* does so with probability $\frac{\alpha}{\alpha+j-1}$.

- Cluster sizes: We saw last lecture that if we let the number of observations n be random, where

$$n|y \sim \text{Pois}(y), \qquad y \sim \text{gam}(\alpha, \alpha/n),$$

then we can analyze cluster size and number using the Poisson process.

  - $\mathbb{E}[n] = \mathbb{E}[\mathbb{E}[n|y]] = \mathbb{E}[y] = n$ ← expected number of observations
  - $K_n^+ \sim \text{Pois}(\alpha \ln(\alpha + n) - \alpha \ln \alpha)$ ← total # clusters
  - Therefore $\mathbb{E}[K_n^+] = \alpha \ln((\alpha + n)/\alpha)$ (compare with above where n is not random)
  - $\sum_{i=1}^{K_n^+} \mathbb{1}(n_i^{(n)} = k) \sim \text{Pois}\left(\frac{\alpha}{K}(\frac{n}{\alpha+n})^K\right)$ ← number of clusters with k observations

- It's important to remember that n is random here. So in #2 and #3 above, we *first* generate n and then sample this many times from the CRP. For example, $\mathbb{E}[K_n^+]$ is slightly different depending on whether n is random or not.

Inference for the CRP

- Generative process: $X_n|\theta_n^* \sim p(X|\theta_n^*),\ \theta_n^*|\theta_{1:n-1}^* \sim \sum_{i=1}^n \frac{1}{\alpha+n-1}\delta_{\theta_i^*} + \frac{\alpha}{\alpha+n-1}\mu$. $\alpha > 0$ is "concentration" parameter and we assume μ is non-atomic probability measure.

- Posterior inference: Given the data $X_1, \ldots, X_N$ and parameters α and μ, the goal of inference is to perform the inverse problem of finding $\theta_1^*, \ldots, \theta_N^*$. This gives the unique parameters $\theta_{1:K_N} = \text{unique}(\theta_{1:N}^*)$ and the partition of the data into clusters.

- Using Bayes rule doesn't get us far (recall that $p(B|A) = \frac{p(A|B)p(B)}{p(A)}$).

$$p(\theta_1, \theta_2, \ldots, \theta_{1:N}^*|X_{1:N}) = \left[\prod_{j=1}^N p(X_j|\theta_j^*)\right] p(\theta_{1:N}^*) \prod_{i=1}^\infty p(\theta_i) \Big/ \text{intractable normalizer} \quad (7.8)$$

- Gibbs sampling: We can't calculate the posterior analytically, but we can sample from it: Iterate between sampling the atoms given the assignments and then sampling the assignments given the atoms.

Sampling the atoms $\theta_1, \theta_2, \ldots$

- This is the easier of the two. For the K_N unique clusters in $\theta_1^*, \ldots, \theta_N^*$ at iteration t, we need to sample $\theta_1, \ldots, \theta_{K_N}$.

Sample θ_i: Use Bayes rule,

$$p(\theta_i|\theta_{-1}, \theta_{1:N}^*, X_{1:N}) \propto \underbrace{\left[\prod_{j:\theta_j^*=\theta_i} p(X_j|\theta_i)\right]}_{\text{likelihood}} \times \underbrace{p(\theta_i)}_{\text{prior }(\mu)} \quad (7.9)$$

- In words, the posterior of θ_i depends only on the data assigned to the ith cluster according to $\theta_1^*, \ldots, \theta_N^*$.

- We simply select this subset of data and calculate the posterior of θ_i on this subset. When μ and $p(X|\theta)$ are conjugate, this is easy.

Sampling θ_j^* (seating assignment for X_j)

- Use exchangeability of $\theta_1^*, \ldots, \theta_N^*$ to treat X_j as if it were the last observation,

$$p(\theta_j^*|X_{1:N}, \Theta, \theta_{-j}^*) \propto p(X_j|\theta_j^*, \Theta)p(\theta_j^*|\theta_{-j}^*) \quad \leftarrow \text{ also conditions on "future" } \theta_n^* \quad (7.10)$$

- Below is the sampling algorithm followed by the mathematical derivation

$$\text{set} \quad \theta_j^* = \begin{cases} \theta_i & \text{w.p.} \propto p(X_j|\theta_i)\sum_{n\neq j} \mathbb{1}(\theta_n^* = \theta_i), \quad \theta_i \in \text{unique}\{\theta_{-j}^*\} \\ \theta_{new} \sim p(\theta|X_j) & \text{w.p.} \propto \alpha \int p(X_j|\theta)p(\theta)d\theta \end{cases} \quad (7.11)$$

- The first line should be straightforward from Bayes rule. The second line is trickier because we have to account for the infinitely remaining parameters. We'll discuss the second line next.

- First, return to the finite model and take then take the limit (and assume the appropriate re-indexing).

- Define: $n_i^{-j} = \#\{\theta_n^* : \theta_n^*, n \neq j\}$, $K_{-j} = \#\text{unique}\{\theta_{-j}^*\}$.

- Then the prior on θ_j^* is

$$\theta_j^* | \theta_{-j}^* \sim \sum_{i=1}^{K_{-j}} \frac{n_i^{-j}}{\alpha + n - 1} \delta_{\theta_i} + \frac{\alpha}{\alpha + n - 1} \sum_{i=1}^{K} \frac{1}{K} \delta_{\theta_i} \tag{7.12}$$

- The term $\sum_{i=1}^{K} \frac{1}{K} \delta_{\theta_i}$ overlaps with the K_{-j} atoms in the first term, but we observe that $K_{-j}/K \to 0$ as $K \to \infty$.

- First: What's the probability a new atom is used in the infinite limit ($K \to \infty$)?

$$p(\theta_j^* = \theta_{new} | X_j, \theta_{-j}^*) \propto \lim_{K \to \infty} \alpha \sum_{i=1}^{K} \frac{1}{K} p(X_j | \theta_i) \tag{7.13}$$

Since $\theta_i \overset{iid}{\sim} \mu$,

$$\lim_{K \to \infty} \sum_{i=1}^{K} \frac{1}{K} p(X_j | \theta_i) = \mathbb{E}_\mu [p(X_j | \theta)] = \int p(X_j | \theta) \mu(d\theta). \tag{7.14}$$

Technically, this is the probability that an atom is selected from the second part of (7.12) above. We'll see why this atom is therefore "new" next.

- Second: Why is $\theta_{new} \sim p(\theta | X_j)$? (And why is it new to begin with?)

- Given that $\theta_j^* = \theta_{new}$, we need to find the index i so that $\theta_{new} = \theta_i$ from the second half of (7.12).

$$\begin{aligned} p(\theta_{new} = \theta_i | X_j, \theta_j^* = \theta_{new}) &\propto p(X_j | \theta_{new} = \theta_i) p(\theta_{new} = \theta_i | \theta_j^* = \theta_{new}) \\ &\propto \lim_{K \to \infty} p(X_j | \theta_i) \frac{1}{K} \Rightarrow p(X_j | \theta) \mu(d\theta) \end{aligned} \tag{7.15}$$

So $p(\theta_{new} | X_j) \propto p(X_j | \theta) \mu(d\theta)$.

Therefore, given that the atom associated with X_j is selected from the second half of (7.12), the probability it coincides with an atom in the first half equals zero (and so it's "new" with probability one). Also, the atom itself is distributed according to the posterior given X_j.

Chapter 8

Exchangeability, beta processes and the Indian buffet process

Marginalizing (integrating out) stochastic processes

- We saw how the Dirichlet process gives a discrete distribution on model parameters in a clustering setting. When the Dirichlet process is integrated out, the cluster assignments form a Chinese restaurant process:

$$\underbrace{p(\theta_1^*, \ldots, \theta_N^*)}_{\text{Chinese restaurant process}} = \int \underbrace{\prod_{n=1}^{N} p(\theta_n^*|G)}_{\text{i.i.d. from discrete dist.}} \underbrace{p(G)}_{\text{DP}} dG \tag{8.1}$$

- There is a direct parallel between the beta-Bernoulli process and the "Indian buffet process":

$$\underbrace{p(Z_1*, \ldots, Z_N)}_{\text{Indian buffet process}} = \int \underbrace{\prod_{n=1}^{N} p(Z_n|H)}_{\text{Bernoulli process}} \underbrace{p(H)}_{\text{BP}} dH \tag{8.2}$$

- As with the DP→CRP transition, the BP→IBP transition can be understood from the limiting case of the finite BP model.

Beta process (finite approximation)

- Let $\alpha > 0$, $\gamma > 0$ and μ a non-atomic probability measure. Define

$$H_K = \sum_{i=1}^{K} \pi_i \delta_{\theta_i}, \quad \pi_i \sim \text{Beta}(\alpha \tfrac{\gamma}{K}, \alpha(1 - \tfrac{\gamma}{K})), \quad \theta_i \sim \mu. \tag{8.3}$$

Then $\lim_{K \to \infty} H_K = H \sim \text{BP}(\alpha, \gamma\mu)$. (See Lecture 3 for proof.)

Bernoulli process using H_K

- Given H_K, we can draw the Bernoulli process $Z_n^K | H_K \sim BeP(H_K)$ as follows:

$$Z_n^K = \sum_{i=1}^{K} b_{in}\delta_{\theta_i}, \quad b_{in} \sim \text{Bernoulli}(\pi_i). \tag{8.4}$$

Notice that b_{in} should also be marked with K, which we ignore. Again we are particularly interested in $\lim_{K\to\infty}(Z_1^K, \dots, Z_N^K)$.

- To derive the IBP, we first consider

$$\lim_{K\to\infty} p(Z_{1:N}^K) = \lim_{K\to\infty} \int \prod_{n=1}^{N} p(Z_n^K | H_K) p(H_K) dH_K. \tag{8.5}$$

- We can think of $Z_1^K, \dots, Z_N^K$ in terms of a binary matrix, $B_K = [b_{in}]$, where

1. each row corresponds to atom θ_i and $b_{in} \overset{iid}{\sim} \text{Bern}(\pi_i)$
2. each column corresponds to a Bernoulli process, Z_n^K

Important: The rows of B are *independent* processes.

- Consider the process $b_{in}|pi_i \overset{iid}{\sim} \text{Bern}(\pi_i)$, $\pi_i \sim \text{Beta}(\alpha\frac{\gamma}{K}, \alpha(1-\frac{\gamma}{K}))$. The marginal process $b_{i1}, \dots, b_{iN}$ follows an urn model with two colors.

Polya's urn (two-color special case)

1. Start with an urn having $\alpha\gamma/K$ balls of color 1 and $\alpha(1-\gamma/K)$ balls of color 2.
2. Pick a ball at random, pit it back and put a second one of the same color

Mathematically:

1. $b_{i1} \sim \dfrac{\gamma}{K}\delta_1 + (1-\dfrac{\gamma}{K})\delta_0$

2. $b_{i,N+1}|b_{i1}, \dots, b_{iN} \sim \dfrac{\frac{\alpha\gamma}{K} + n_i^{(N)}}{\alpha + N}\delta_1 + \dfrac{\alpha(1-\frac{\gamma}{K}) + N - n_i^{(N)}}{\alpha + N}\delta_0$

where $n_i^{(N)} = \sum_{j=1}^{N} b_{ij}$. Recall from exchangeability and deFinetti that

$$\lim_{K\to\infty} \frac{n_i^{(N)}}{N} \longrightarrow \pi_i \sim \text{Beta}(\alpha\tfrac{\gamma}{K}, \alpha(1-\tfrac{\gamma}{K})) \tag{8.6}$$

- Last week we proved this in the context of the finite symmetric Dirichlet, $\pi \sim \text{Dir}(\frac{\alpha}{K}, \dots, \frac{\alpha}{K})$. The beta distribution is the two-dimensional special case of the Dirichlet and the proof can be applied to any parameterization besides symmetric.

- In the Dirichlet→CRP limiting case, K corresponds to the number of colors in the urn and $K \to \infty$ with the starting number of each color $\frac{\alpha}{K} \to 0$.

- In this case, there are always only two colors. However, the *number of urns* equals K, so the number of urn processes is going to infinity as the first parameter of each beta goes to zero.

An intuitive derivation of the IBP: Work with the urn representation of B_K. Again, b_{in} is entry (i, n) of B_K and the generative process for B_K is

$$b_{i,N+1}|b_{i1},\dots,b_{iN} \sim \frac{\frac{\alpha\gamma}{K} + n_i^{(N)}}{\alpha+N}\delta_1 + \frac{\alpha(1-\frac{\gamma}{K}) + N - n_i^{(N)}}{\alpha+N}\delta_0 \tag{8.7}$$

where $n_i^{(N)} = \sum_{j=1}^N b_{ij}$. Each row of B_K is associated with a $\theta_i \sim_{iid} \mu$ so we can reconstruct $Z_n^K = \sum_{i=1}^K b_{in}\delta_{\theta_i}$ using what we have.

- Let's break down $\lim_{K\to\infty} B_K$ into two cases.

Case $n = 1$: We ask how many ones are in the first column of B?

$$\lim_{K\to\infty} \sum_{i=1}^K b_{i1} \sim \lim_{K\to\infty} \text{Bin}(K, \gamma/K) = \text{Pois}(\gamma) \tag{8.8}$$

So Z_1 has Pois(γ) ones. Since the θ associated with these ones are i.i.d., we can "ex post facto" draw them i.i.d. from μ and re-index.

Case $n > 1$: For the remaining Z_n, we break this into two subcases.

- Subcase $n_i^{(n-1)} > 0$: $\quad b_{in}|b_{i1},\dots,b_{i,n-1} \sim \frac{n_i^{(n-1)}}{\alpha+n-1}\delta_1 + \frac{\alpha+n-1-n_i^{(n-1)}}{\alpha+n-1}\delta_0$
- Subcase $n_i^{(n-1)} = 0$: $\quad b_{in}|b_{i1},\dots,b_{i,n-1} \sim \lim_{K\to\infty} \frac{\alpha\frac{\gamma}{K}}{\alpha+n-1}\delta_1 + \frac{\alpha(1-\frac{\gamma}{K})+n-1}{\alpha+n-1}\delta_0$

- For each i, $\left(\frac{\alpha\gamma}{\alpha+n-1}\right)\frac{1}{K}\delta_1 \to 0\delta_1$, but there are also an infinite number of these indexes i for which $n_i^{(n-1)} = 0$. Is there a limiting argument we can again make to just ask how many ones there are total for these indexes with $n_i^{(n-1)} = 0$?

- Let $K_n = \#\{i : n_i^{(n-1)} > 0\}$, which is finite almost surely. Then

$$lim_{K\to 0}\sum_{i=1}^K b_{in}\mathbb{1}(n_i^{(n-1)} = 0) \sim \lim_{K\to\infty} \text{Bin}\left(K - K_n, \Big(\frac{\alpha\gamma}{\alpha+n-1}\Big)\frac{1}{K}\right) = \text{Pois}\Big(\frac{\alpha\gamma}{\alpha+n-1}\Big) \tag{8.9}$$

- So there are $\text{Pois}\Big(\frac{\alpha\gamma}{\alpha+n-1}\Big)$ new locations for which $b_{in} = 1$. Again, since the atoms are i.i.d. regardless of the index, we can simply draw them i.i.d. from μ and re-index.

Putting it all together: The Indian buffet process

- For $n = 1$: Draw $C_1 \sim \text{Pois}(\gamma)$, $\theta_1, \dots, \theta_{C_1} \overset{iid}{\sim} \mu$ and set $Z_1 = \sum_{i=1}^{C_1} \delta_{\theta_1}$.

- For $n > 1$: Let $K_{n-1} = \sum_{j=1}^{n-1} C_j$. For $i = 1, \dots, K_{n-1}$, draw

$$b_{in}|b_{i,1:n-1} \sim \frac{n_i^{(n-1)}}{\alpha+n-1}\delta_1 + \frac{\alpha+n-1-n_i^{(n-1)}}{\alpha+n-1}\delta_0. \tag{8.10}$$

Then draw $C_n \sim \text{Pois}\Big(\frac{\alpha\gamma}{\alpha+n-1}\Big)$ and $\theta_{K_{n-1}+1}, \dots, \theta_{K_{n-1}+C_n} \overset{iid}{\sim} \mu$ and set

$$Z_n = \sum_{i=1}^{K_{n-1}} b_{in}\delta_{\theta_i} + \sum_{i'=K_{n-1}+1}^{K_{n-1}+C_n} \delta_{\theta_i}. \tag{8.11}$$

- By exchangeability and deFinetti, $\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} Z_i \to H \sim \text{BP}(\alpha, \gamma\mu)$.

The IBP story: Start with α customers not eating anything.

1. A customer walks into an Indian buffet with an infinite number of dishes and samples $\text{Pois}(\gamma)$ of them.
2. The nth customer arrives and samples from the previously sampled dishes with probability proportional to the number of previous customers who sampled it, and then samples $\text{Pois}\Big(\frac{\alpha\gamma}{\alpha+n-1}\Big)$ new dishes.

- In modeling scenarios, each dish corresponds to a factor (e.g., a one-dimensional subspace) and a customer samples a subset of factors.

- Clearly, after n customers there are $\sum_{i=1}^{n} C_i \sim \text{Pois}\Big(\sum i = 1^n \frac{\alpha\gamma}{\alpha+i-1}\Big)$ dishes that have been sampled. (See Lecture 3 for another derivation of this quantity.)

BP constructions and the IBP

- From Lecture 4: Let $\alpha, \gamma > 0$ and μ a non-atomic probability measure. Let

$$C_i \sim \text{Pois}\Big(\frac{\alpha\gamma}{\alpha+i-1}\Big), \quad \pi_{ij} \sim \text{Beta}(1, \alpha+i-1), \quad \theta_{ij} \sim \mu \tag{8.12}$$

and define $H = \sum_{i=1}^{\infty}\sum_{j=1}^{C_i} \pi_{ij}\delta_{\theta_{ij}}$. Then $H \sim \text{BP}(\alpha, \gamma\mu)$.

- We proved this using the Poisson process theory. Now we'll show how this construction can be arrived at in the first place.

- Imagine an urn with β_1 balls of one color and β_2 of another. The distribution on the first draw from this urn is
$$\frac{\beta_1}{\beta_1+\beta_2}\delta_1 + \frac{\beta_2}{\beta_1+\beta_2}\delta_0.$$
Let $(b_1, b_2, \dots)$ be the urn process with this initial configuration. Then as we have seen, in the limit $N \to \infty$, $\frac{1}{N}\sum_{i=1}^N b_i \to \pi \sim \text{Beta}(\beta_1, \beta_2)$.

- Key: We can skip the whole urn procedure if we're only interested in π. That is, if we draw from the urn once, look at it and see it's color one, then the urn distribution is
$$\frac{\beta_1+1}{1+\beta_1+\beta_2}\delta_1 + \frac{\beta_2}{1+\beta_1+\beta_2}\delta_0$$
The questions to ask are what will happen if we continue? What is the different between this "posterior" configuration and another urn where this is defined to be the initial setting? It shouldn't be hard to convince yourself that, in this case, as $N \to \infty$,
$$\frac{1}{N}\sum_{i=1}^N b_i | b_1 = 1 \;\to\; \pi \sim \text{Beta}(\beta_1 + 1, \beta_2).$$

- This is because the sequence $(b_1 = 1, b_2, b_3, \dots)$ is equivalent to an urn process $(b_2, b_3, \dots)$ where the initial configuration are $\beta_1 + 1$ of color one and β_2 of color 2.

- Furthermore, we know from deFinetti and exchangeability that if $Z_1, \dots, Z_N$ are from an IBP, then $\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N Z_i \;\to\; H \sim \text{BP}(\alpha, \gamma\mu)$. Right now we're only interested in this limit.

Derivation of the construction via the IBP

- For each Z_n, there are $C_n \sim \text{Pois}\left(\frac{\alpha\gamma}{\alpha+n-1}\right)$ new "urns" introduced with the initial configuration $\frac{1}{\alpha+n}\delta_1 + \frac{\alpha+n-1}{\alpha+n}\delta_0$. From this point on, each of these new urns can be treated as independent processes. Notice that this is the case for every instantiation of the IBP.

- With the IBP, we continue this urn process. Instead, we can ask the limiting distribution of the urn immediately after instantiating it. We know from above that the new urns created at step n will converge to random variables drawn independently from a Beta$(1, \alpha + n - 1)$ distribution. We can draw this probability directly.

- The results is the construction written above.

Part II

Topics in Bayesian Inference

Chapter 9

EM and variational inference

- We've been discussing Bayesian nonparametric priors and their relationship to the Poisson process up to this point (first 2/3 of the course). We now turn to the last 1/3 which will be about model inference (with tie-ins to Bayesian nonparametrics).

- In this lecture, we'll review two deterministic optimization-based approaches to inference: the EM algorithm for MAP inference and an extension of EM to approximate Bayesian posterior inference called variational inference.

  General setup

  1. Prior: We define a model with a set of parameters θ and a prior distribution on them $p(\theta)$.
  2. Likelihood: Conditioned on the model parameters, we have a way of generating the set of data X according to the distribution $p(X|\theta)$.
  3. Posterior: By Baye's rule, the posterior distribution of the parameters θ is

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int p(X|\theta)p(\theta)d\theta} \tag{9.1}$$

- Of course, the problem is that for most models the denominator is intractable, and so we can't actually calculate $p(\theta|X)$ analytically.

- The first, most obvious solution is to ask for the most probable θ according to $p(\theta|X)$. Since

$$p(\theta|X) = p(X|\theta)p(\theta)\Big/\underbrace{\int p(X|\theta)p(\theta)d\theta}_{\text{constant wrt } \theta}$$

  we know that

$$\arg\max_\theta \; p(\theta|X) = \arg\max_\theta \; p(X|\theta)p(\theta) = \underbrace{\arg\max_\theta \; \ln p(X|\theta) + \ln p(\theta)}_{\text{for computational convenience}} \tag{9.2}$$

- Therefore, we can come up with an algorithm for finding a local optimal solution to the (usually) non-convex problem $\arg\max_\theta \; p(X,\theta)$.

- This often leads to a problem (that will be made more clear in an example later). The function $\ln p(X, \theta)$ is sometimes hard to optimize with respect to θ (e.g., it might require messy gradients).

- The EM algorithm is a very general method for optimizing $\ln p(X, \theta)$ w.r.t. θ in a "nice" way (e.g., closed form updates).

- With EM, we expand the joint likelihood $p(X, \theta)$ using some auxiliary variables c so that we have $p(X, \theta, c)$ and $p(X, \theta) = \int p(X, \theta, c)dc$. c is something picked entirely out of convenience. For now, we keep it abstract and highlight this with an example later.

Build-up to EM

- Observe that by basic probability, $p(c|X, \theta)p(X, \theta) = p(X, \theta, c)$. Therefore,

$$\underbrace{\ln p(X, \theta)}_{\text{thing we want}} = \underbrace{\ln p(X, \theta, c) - \ln p(c|X, \theta)}_{\text{expanded version equal to thing we want}} \tag{9.3}$$

- We then introduce any probability distribution on c, $q(c)$, and follow the the sequence of steps:

  1. $\ln p(X, \theta) = \ln p(X, \theta, c) - \ln p(c|X, \theta)$
  2. $q(c) \ln p(X, \theta) = q(c) \ln p(X, \theta, c) - q(c) \ln p(c|X, \theta)$
  3. $\int q(c) \ln p(X, \theta)dc = \int q(c) \ln p(X, \theta, c)dc - \int q(c) \ln p(c|X, \theta)dc$
  4. $\ln p(X, \theta) = \int q(c) \ln p(X, \theta, c)dc - \int q(c) \ln q(c)dc$
     $- \int q(c) \ln p(c|X, \theta)dc + \int q(c) \ln q(c)dc$ (add and subtract same thing)
  5. $$\ln p(X, \theta) = \underbrace{\int q(c) \ln \frac{p(X, \theta, c)}{q(c)}dc}_{\mathcal{L}(q,\theta)} + \underbrace{\int q(c) \ln \frac{q(c)}{p(c|X, \theta)}dc}_{KL(q\|p)}$$

- We have decomposed the objective function $\ln p(X, \theta)$ into a sum of two terms

  1. $\mathcal{L}(q, \theta)$: A function we can (hopefully) calculate
  2. $KL(q\|p)$: The KL-divergence between q and p. Recall that $KL \geq 0$ and $KL = 0$ only when $q = p$.

- The practical usefulness isn't immediately obvious. For now, we analyze it as an algorithm for optimizing $\ln p(X, \theta)$ w.r.t. θ. That is, how can we use the right side to find a local optimal solution of the left side?

- We will see that the following algorithm gives a sequence of values for θ that are monotonically increasing $p(X,\theta)$ and can be shown to converge to a local optimum of $p(X,\theta)$.

The Expectation-Maximization Algorithm

E-step: Change $q(c)$ by setting $q(c) = p(c|X,\theta)$. Obviously, we're assuming $p(c|X,\theta)$ can be calculated in closed form. What does this do?

1. Sets $KL(q\|p) = 0$.
2. Since θ is fixed, $\ln p(X,\theta)$ isn't changed here, therefore since $KL \geq 0$ in general, by changing q so that $KL(q\|p) = 0$, we increase $\mathcal{L}(q,\theta)$ so that $\mathcal{L}(q,\theta) = \ln p(X,\theta)$.

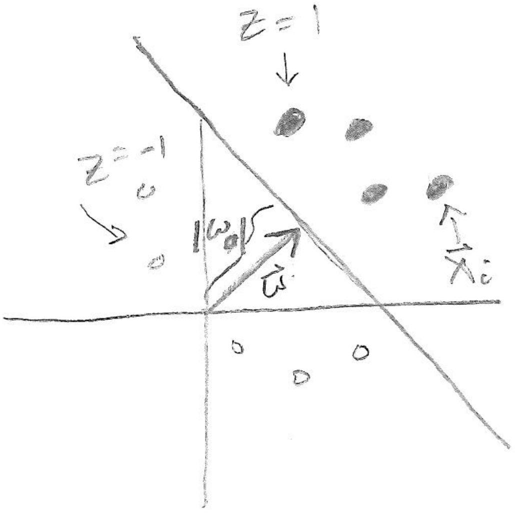
M-step: $\mathcal{L}(q,\theta) = \mathbb{E}_q[\ln p(X,\theta,c)] - \mathbb{E}_q[\ln q(c)]$ has now been modified as a function of θ. Therefore, we can maximize $\mathcal{L}(q,\theta)$ w.r.t. θ. What does this do?

1. Increases $\mathcal{L}(q,\theta)$ (obvious).
2. Since θ has changed, $q(c) \neq p(c|X,\theta)$ anymore, so $KL(q\|p) > 0$.
3. Therefore,

$$\ln p(X,\theta_{\text{old}}) = \mathbb{E}_q\left[\ln \frac{p(X,\theta_{\text{old}},c)}{q(c)}\right] < \mathbb{E}_q\left[\ln \frac{p(X,\theta_{\text{new}},c)}{q(c)}\right] + KL(q\|p) = \ln p(X,\theta_{\text{new}}) \tag{9.4}$$

Example: Probit classification

- We're given data pairs $D = \{(x_i, z_i)\}$ for $i = 1,\ldots,N$, $x_i \in \mathbb{R}^d$ and $z_i \in \{-1,+1\}$. We want to build a linear classifier $f(x^T w - w_0) : x \to \{-1,+1\}$.



$$x_i^T w - w_0 \begin{cases} > 0 & \text{for all points to upper right} \\ < 0 & \text{for all points to lower left} \\ = 0 & \text{all points on line} \end{cases}$$

Classify using the probit function:

$$f(x^T w - w_0) = \Phi\left(\frac{x^T w - w0}{\sigma}\right) = p(z = 1|w,x)$$

Φ is the CDF of a standard normal distribution

Simplification of notation: Let $w \leftarrow [w_0 w^T]^T$ and $x_i \leftarrow [1 x_i^T]^T$.

- The model variable $\theta = w$ in this case. Let $w \sim N(0, cI)$ be the prior.
- The joint likelihood is $p(\boldsymbol{z}, w|\boldsymbol{x}) = p(w)\prod_{i=1}^N p(z_i|w,x_i)$.

- So we would like to optimize

$$\ln p(w)+\sum_{i=1}^{N}\ln p(z_i|w,x_i) = -\frac{1}{2c}w^Tw+\sum_{i=1}^{N} z_i \ln \Phi\Big(\frac{x_i^Tw}{\sigma}\Big)+\sum_{i=1}^{N}(1-z_i)\ln\left[1-\Phi\Big(\frac{x_i^Tw}{\sigma}\Big)\right] \tag{9.5}$$

with respect to w, which is hard because of $\Phi(\cdot)$.

Solution via EM

- Introduce the latent variables ("hidden data") $y1,\dots,y_N$ such that $y_i|x_i,w \sim N(x_i^Tw,\sigma^2)$ and let $z_i = \mathbb{1}(y_i > 0)$. Marginalizing this expanded joint likelihood,

$$\begin{aligned} p(z_i=1|x_i,w) &= \int_{-\infty}^{\infty} p(y_i,z_i=1|x_i,w)dy_i \qquad (9.6)\\ &= \int_{-\infty}^{\infty} \mathbb{1}(y_i>0)N(y_i|x_i^Tw,\sigma^2)dy_i = \Phi\Big(\frac{x_i^Tw}{\sigma}\Big)\end{aligned}$$

so the expanded joint likelihood $p(z,y,w|x)$ has the correct marginal distribution.

- From the EM algorithm we know that

$$\begin{aligned}\ln p(z,w|x) &= \mathbb{E}_q\left[\ln\frac{p(z,y,w|x)}{q(y)}\right]+KL(q(y)\|p(y|z,w,x)) \qquad (9.7)\\ &= \sum_{i=1}^{N}\int q(y)\ln\frac{p(z_i,y_i,w|x_i)}{q(y)}dy+\int q(y)\ln\frac{q(y)}{p(y|z,w,x)}dy\end{aligned}$$

Q: Can we simplify $q(y)=q(y_1,\dots,y_N)$?

A: We need to set $q(y)=p(y|z,w,x)$. Since $p(y|z,w,x)=\prod_{i=1}^{N}p(y_i|z_i,w,x_i)$, we know that we can write $q(y)=\prod_{i=1}^{N}q(y_i)$ (but we still don't know what it is).

- E-step

1. For each i, set $q(y_i)=p(y_i|z_i,w,x_i)\propto p(z_i|y_i)p(y_i|x_i,w)$. Since this second term equals $\mathbb{1}(y_i\in\mathbb{R}_{z_i})N(y_i|x_i^Tw,\sigma^2)$, it follows that $q(y_i)=TN_{z_i}(x_i^Tw,\sigma^2)$, which is a truncated normal on the z_i half of $\mathbb{R}$.
2. Construct the objective $\sum_i \mathbb{E}_{q(y_i)}[\ln p(z_i,y_i,w|x_i)]$. We can ignore $\sum_i \mathbb{E}_{q(y_i)}[\ln q(y_i)]$ because it doesn't depend on w, which is what we want to optimize over next.

$$\begin{aligned}\sum_{i=1}^{N}\mathbb{E}_q\ln p(z_i,y_i,w|x_i) &= \sum_{i=1}^{N}\mathbb{E}_q\underbrace{[-\tfrac{1}{2\sigma^2}(y_i-x_i^Tw)^2]}_{\ln p(y_i|x_i,w)\ +\ \text{const.}}+\sum_{i=1}^{N}\mathbb{E}_q\ln p(z_i|y_i)\\ &\quad -\tfrac{1}{2c}w^Tw+\text{const.}\leftarrow\ \ln p(w) \qquad (9.8)\\ &= \sum_{i=1}^{N}-\tfrac{1}{2\sigma^2}(\mathbb{E}_q[y_i]-x_i^Tw)^2-\tfrac{1}{2c}w^Tw+\text{const. w.r.t. } w\end{aligned}$$

- M-step

  1. Optimize the objective from the E-step with respect to w. We can take the gradient w.r.t. w and see this is equal to

$$w = \Big(\frac{1}{c}I + \frac{1}{\sigma^2}\sum_{i=1}^{N} x_i x_i^T\Big)^{-1}\Big(\sum_{i=1}^{N} x_i \mathbb{E}_q[y_i]/\sigma^2\Big) \tag{9.9}$$

  - The expectation $\mathbb{E}_q[y_i]$ is of the truncated Normal r.v. with distribution $q(y_i)$, which can be looked up in a book or on Wikipedia. The expectation looks complicated, but it's simply an equation that can be evaluated using what we have. The final algorithm can be seen in this equation: After updating w, update q. This updates $\mathbb{E}_q[y_i]$ and so we can again update w. The code would look like iterations between updating w and updating $\mathbb{E}_q[y_i]$ for each i. Notice that this is a nice, closed form iterative algorithm. The output w maximizes $p(z, w|x)$.

Extension of EM to a BNP modification of the probit model

Model extension: Imagine that x is very high dimensional. We might think that only a few dimensions of x are relevant for predicting z. We can use a finite approximation to the gamma process to address this.

Gamma process prior: Let $c_j \overset{iid}{\sim} \text{Gam}(\frac{\alpha}{d}, b)$, $w_j \sim N(0, c_j)$. From the GaP theory we know that when $d \ll \alpha$ most values of $c_j \approx 0$, but there are some that will be large. In this context, c_j are variances, so the model says that most dimensions of w should be zero, in which case the corresponding dimension of x is not used in predicting $z = f(x^T w)$.

Log joint likelihood: The EM equation is

$$\ln p(z, w|x) = \int q(y, c) \ln \frac{p(z, y, w, c|x)}{q(y, c)} dydc \; + \int q(y, c) \ln \frac{q(y, c)}{p(y, c|z, w, x)} dydc \tag{9.10}$$

We can think of y again as the "hidden data" that doesn't have an interpretation as a model variable. c can be thought of as a model variable and we want to integrate out the uncertainty of this variable to infer a point estimate of w.

E-step:

1. Set $q(y, c) = p(y, c|z, w, x)$. In this case y and c are conditionally independent of each other,

$$p(y, c|z, w, x) = p(y|z, w, x)p(c|w).$$

   We can calculate these as before:

$$\begin{aligned} q(y_i) &= p(y_i|z_i, w, x_i) = TN_{z_i}(x_i^T w, \sigma^2) & (9.11) \\ q(c_j) &= p(c_j|w_j) = \text{GiG}(w_j^2, \alpha, b) \;\leftarrow\; \text{generalized inverse Gaussian} & (9.12) \end{aligned}$$

2. Construct the objective function

   $\sum_i \mathbb{E}_q \ln p(z_i, y_i, w, c|x_i) = \sum_i \mathbb{E}_q \ln p(z_i|y_i) + \sum_i \mathbb{E}_q \ln p(y_i|x_i, w) + \mathbb{E}_q \ln p(w|c) + \text{const.}$

M-step:

1. Optimize the objective from the E-step with respect to w. We can take the gradient with respect to w and set to zero to find that

$$w = \Big(\text{diag}(\mathbb{E}_q[c_j^{-1}]) + \frac{1}{\sigma^2}\sum_{i=1}^N x_i x_i^T\Big)^{-1}\Big(\sum_{i=1}^N x_i \mathbb{E}_q[y_i]/\sigma^2\Big) \tag{9.13}$$

Variational inference

- We motivate variational inference in the context of the model we've been discussing and then give a general review of it.

- Generative process:

$$z_i = \text{sign}(y_i), \quad y_i \sim N(x_i^T w, \sigma^2), \quad w_j \sim N(0, c_j), \quad c_j \sim \text{Gam}(\tfrac{a}{d}, b) \tag{9.14}$$

- We saw how we could maximize $\ln p(z, w|x)$ easily by including distributions on y and c. Could we do the same for w?

$$\ln p(z|x) = \int q(y, c, w) \ln \frac{p(z, y, w, c|x)}{q(y, c, w)} dydcdw + \int q(y, c, w) \ln \frac{q(y, c, w)}{p(y, c, w|z, x)} dydcdw \tag{9.15}$$

- Short answer: Probably not. What has changed?

  1. Before, we were maximizing $\ln p(z, w|x)$ w.r.t. w. This meant we needed to calculate $p(y, c|z, x, w)$. Fortunately, since y and c are independent conditioned on w, we can calculate $p(y, c|z, x, w) = p(y|z, x, w)p(c|w)$ analytically using the point estimate of w.

  2. Therefore, we could set $q(y) = p(y|z, x, w)$ and $q(c) = p(c|w)$ to make $KL = 0$ and optimize $\mathcal{L}$.

  3. This time, we can't factorize $p(y, c, w|z, x)$, so we can't find $q(y, c, w) = p(y, c, w|z, x)$ in closed form to make $KL = 0$.

- Some other observations

  1. $\ln p(z|x)$ is a constant since it depends only on the data z, x and the underlying model.

  2. Therefore, it doesn't make sense to talk about maximizing $\ln p(z|x)$ like before when we were using EM to maximize $\ln p(z, w|x)$ with respect to w.

  3. If we could find $q(y, c, w)$ to make $KL(q\|p) = 0$, we would have the full posterior of the model! (That's the whole purpose of inference.)

  4. What made EM for $\ln p(z, w|x)$ easy was that we could write $q(y, c) = q(y)q(c)$ and still find optimal distributions.

- Variational inference idea

  1. We can't write $q(y, c, w) = q(y)q(c)q(w)$ and still have $KL(q\|p) = 0$, i.e., can't have $q(y)q(c)q(w) = p(y, c, w|z, x)$.
  2. What if we make an *approximation* that $p(y, c, w|z, x) \approx q(y)q(c)q(w)$?
  3. We *still* have that

$$\ln p(z|x) = \underbrace{\int q(y)q(c)q(w) \ln \frac{p(z, y, w, c|x)}{q(y)q(c)q(w)} dydcdw}_{\mathcal{L}(q)} + \underbrace{\int q(y)q(c)q(w) \ln \frac{q(y)q(c)q(w)}{p(y, c, w|z, x)} dydcdw}_{KL(q\|p)} \tag{9.16}$$

  4. However, $KL \neq 0$ (ever)
  5. Notice though that, since $\ln p(z|x) =$ const. and $KL > 0$, if we can maximize $\mathcal{L}$ wrt q, we can minimize KL between q and p.

Variational inference (in general)

- Have data x and model with variables $\theta_1, \dots, \theta_m$.

- Choose probability distributions $q(\theta_i)$ for each θ_i.

- Construct the equality

$$\begin{aligned}\underbrace{\ln p(x)}_{\text{log evidence (constant)}} &= \underbrace{\int \Big(\prod_{i=1}^{m} q(\theta_i)\Big) \ln \frac{p(x, \theta_1, \dots, \theta_m)}{\prod_{i=1}^{m} q(\theta_i)} d\theta_1 \dots d\theta_m}_{\mathcal{L}\text{:variational lower bound on log evidence}} \\ &\quad + \underbrace{\int \Big(\prod_{i=1}^{m} q(\theta_i)\Big) \ln \frac{\prod_{i=1}^{m} q(\theta_i)}{p(\theta_1, \dots, \theta_m|x)} d\theta_1 \dots d\theta_m}_{\text{KL divergence between } \prod_{i=1}^{m} q(\theta_i) \text{ and } p(\theta_1,\dots,\theta_m|x)}\end{aligned} \tag{9.17}$$

Observations

1. As before, $\ln p(x)$ is a constant (that we don't know)
2. $KL(q\|p) \geq 0$ (as always), therefore $\mathcal{L} \leq \ln p(x)$
3. As with EM, $p(x, \theta_1, \dots, \theta_m)$ is something we can write (joint likelihood)
4. $\mathcal{L} = \underbrace{\mathbb{E}_q \ln p(x, \theta_1, \dots, \theta_m)}_{\text{expected log joint likelihood}} \underbrace{- \sum_{i=1}^{m} \int q(\theta_i) \ln q(\theta_i) d\theta_i}_{\text{individual entropies of each } q}$ is a closed form objective function (currently by assumption, not always true!)
5. We've defined the distribution $q(\theta_i)$ but not its parameters. $\mathcal{L}$ is a function of these parameters
6. By finding parameters of each $q(\theta_i)$ to maximize $\mathcal{L}$, we are equivalently finding parameters to minimize $KL(q\|p)$.

Chapter 10

Stochastic variational inference

Bayesian modeling review

- Recall that we have a set of data X and of model parameters $\Theta = \{\theta_i\}$ for $i = 1, \dots, M$.

- We have defined a model for the data, $p(X|\Theta)$, and a prior for the model, $p(\Theta)$.

- The goal is to learn Θ and capture a level of uncertainty, so by Bayes rule, we want

$$p(\Theta|X) = \frac{p(X|\Theta)p(\Theta)}{p(X)}. \tag{10.1}$$

- We can't calculate $p(X)$ so we need approximate methods.

Variational inference

- Using the EM algorithm as motivation, we set up the "master equation"

$$\underbrace{\ln p(X)}_{\text{constant marginal likelihood}} = \underbrace{\int q(\Theta) \ln \frac{p(X,\Theta)}{q(\Theta)} d\Theta}_{\text{variational lower bound, } \mathcal{L}} + \underbrace{\int q(\Theta) \ln \frac{q(\Theta)}{p(\Theta|X)} d\Theta}_{\text{KL-divergence between } q \text{ and desired } p(\Theta|X)} \tag{10.2}$$

- We have freedom to pick $q(\Theta)$, but that means we have to define it.

- There is a trade-off:
  1. Pick $q(\Theta)$ to minimize KL (ideally set $q = p(\Theta|X)$!)
  2. We don't know $p(\Theta|X)$, so pick q that's easy to optimize.

- What do we mean by "optimize"?
Answer: We can pick a $q(\Theta)$ so that we have a closed-form function $\mathcal{L}$. Since $\ln p(X)$ is a constant, by maximizing $\mathcal{L}$ wrt the parameters of q, we are minimizing the non-negative KL-divergence. Each increase in $\mathcal{L}$ finds a q closer to $p(\Theta|X)$ according to KL.

Picking $q(\theta_i)$

- The variational objective function is

$$\mathcal{L} = \int \Big(\prod_{i=1}^{m} q(\theta_i) \Big) \ln p(X, \Theta) d\Theta - \sum_{i=1}^{m} \int q(\theta_i) \ln q(\theta_i) d\theta_i \tag{10.3}$$

- Imagine we have $q(\theta_j)$ for $j \neq i$. Then we can isolate $q(\theta_i)$ as follows:

$$\begin{aligned} \mathcal{L} &= \int q(\theta_i) \mathbb{E}_{-q_i} \left[\ln p(X, \Theta) \right] d\theta_i - \int q(\theta_i) \ln q(\theta_i) d\theta_i - \sum_{j \neq i} \int q(\theta_j) \ln q(\theta_j) d\theta_j \\ &= \int q(\theta_i) \ln \frac{\frac{1}{Z} \exp\{\mathbb{E}_{-q_i} \left[\ln p(X, \Theta) \right]\}}{q(\theta_i)} d\theta_i \ + \ \text{constant wrt } \theta_i \end{aligned} \tag{10.4}$$

- Z is the normalizing constant: $Z = \int \exp\{\mathbb{E}_{-q_i} \left[\ln p(X, \Theta) \right]\} d\theta_i$. It's a constant wrt θ_i, so we can cancel it out by including the appropriate term in the constant.

- Therefore, $\mathcal{L} = -KL(q(\theta_i) \| \exp\{\mathbb{E}_{-q_i} \left[\ln p(X, \Theta) \right]\}/Z) + \text{const wrt } \theta_i$.

- To maximize this over all $q(\theta_i)$, we need to set $KL = 0$ (since $KL \geq 0$). Therefore,

$$q(\theta_i) \propto \exp\{\mathbb{E}_{-q_i} \left[\ln p(X, \Theta) \right]\}. \tag{10.5}$$

- In words: We can find $q(\theta_i)$ by calculating the expected log joint likelihood using the other q's and exponentiating the results (and normalizing it)

- Notice that we get both the form and parameters for $q(\theta_i)$ this way.

Example: Sparse regression

- Data: Labeled data (x_n, z_n), $x_n \in \mathbb{R}^d$, $z_n \in \{-1, 1\}$, d large
- Model: $z_n = \text{sign}(y_n), y_n \sim N(x_n^T w, \sigma^2), w_j \sim N(0, c_j), c_j \sim \text{Gam}(\frac{a}{d}, b)$
- Recall: Last week, we did EM MAP for w, meaning we got a point estimate of w and integrated out y and c. In that case, we could set $q(c_j)$ and $q(y_n)$ to their exact posteriors conditioned on w because they were independent.

Variational inference

Step 1: Pick a factorization, say $q(y, c, w) = \left[\prod_n q(y_n) \right] \left[\prod_j q(c_j) \right] q(w)$

Step 2: Find optimal q distributions

- $q(y_n) \propto \exp\{\underbrace{\cancel{\mathbb{E}_{-q}}[\ln \mathbb{1}(z_n = \text{sign}(y_n))]}_{\text{determines support of } y_n \ (\mathbb{R}_+ \text{ or } \mathbb{R}_-)} + \underbrace{\mathbb{E}_{-q}[\ln p(y_n|x_n, w)]}_{\text{Determines function and params}} \} = TN_{z_n}(x_n^T \mathbb{E}_q w, \sigma^2)$

  This would have been *much* harder by taking derivatives of $\mathcal{L}$

- $$\begin{aligned} q(c_j) &\propto \exp\{\mathbb{E}_{-q}[\ln p(w_j|c_j)] + \cancel{\mathbb{E}_q} \ln p(c_j)\} \\ &\propto c_j^{-\frac{1}{2}} \exp\{-\tfrac{1}{2c_j}\mathbb{E}_q[w_j^2]\} c_j^{\frac{a}{d}-1} \exp\{-bc_j\} = GiG(2b, \mathbb{E}_q[w_j^2], \tfrac{a}{d} - \tfrac{1}{2}) \end{aligned}$$

  Again, it would have been harder to define $GiG(\tau_1, \tau_2, \tau_3)$ and take derivatives of $\mathcal{L}$

- $q(w) \propto \exp\{\mathbb{E}_{-q}[\ln p(\vec{y}|X, w)] + \mathbb{E}_{-q}[\ln p(w|\vec{c})]\} = N(\mu, \Sigma)$ (a standard calculation)

  Problem: $w \in \mathbb{R}^d$ and we're assuming d is large (e.g., 10,000). Therefore, calculating the $d \times d$ matrix Σ is impractical.

  A solution: Use further factorization. Let $q(w) = \prod_j q(w_j)$. We can again find the optimal form of $q(w_j)$ to be $N(\mu_j, \sigma_j^2)$.

  More tricks: In this case, updating $q(w_j)$ depends on $q(w_{i \neq j})$. This can make things slow to converge. Instead, it works better to directly work with $\mathcal{L}$ and update the vectors $\vec{\mu} = (\mu_1, \dots, \mu_d)$ and $\vec{\sigma}^2 = (\sigma_1^2, \dots, \sigma_d^2)$ by setting, e.g., $\nabla_{\vec{\mu}} \mathcal{L} = 0$.

Conjugate exponential family models (CEF)

- Variational inference in CEF models has a nice generic form. Here, conjugate means the prior/likelihood of all model variables are conjugate. Exponential means all distributions are in the exponential family.

- Review and notation: Pick a model variable, θ_i. We have a likelihood and prior for it:
  - Likelihood: Assume a conditionally independent likelihood

$$p(w_1, \dots, w_n|\eta(\theta_i)) = \prod_{n=1}^N p(w_n|\eta_i) = \left[\prod_{n=1}^N h(w_n)\right] \exp\left\{\eta_i^T \sum_{n=1}^N t(w_n) - NA(\eta_i)\right\} \quad (10.6)$$

    $\eta(\theta_i) \equiv \eta_i$ is natural parameter, $t(w_n)$ is sufficient statistic vector, $A(\eta_i)$ is log-normalizer

  - Prior: A conjugate prior is $p(\eta_i) = f(\chi, \nu) \exp\{\eta_i^T \chi - \nu A(\eta)\}$

  - Posterior: $p(\eta_i|\vec{w}) = f(\chi', \nu') \exp\{\eta_i^T \chi' - \nu' A(\eta_i)\}, \ \chi' = \chi + \sum_n t(w_n), \ \nu' = \nu + N$

Variational Inference

- We have variational distributions on each w_n, $q(w_n)$ and on $q(\theta_i)$. Using the log → expectation → exponential "trick"

$$q(\theta_i) \propto \exp\left\{\eta(\theta_i)^T \Big(\chi + \sum_{n=1}^N \mathbb{E}_q[t(w_n)]\Big) - (\nu + N)A(\eta(\theta_i))\right\}. \quad (10.7)$$

- Therefore, we immediately see that the optimal q distribution for a model variable with a prior conjugate to the likelihood is in the same family as the prior.

- In fact, we see that with variational inference, we take the expected value of the sufficient statistics that would be used to calculate the true conditional posterior.

- Contrast this with Gibbs sampling, where we have samples of $t(w_1), \dots, t(w_N)$ and we sample a new value of θ_i directly using these sufficient statistics (that are themselves changing with each iteration because they are also sampled).

The more direct calculation for CEF models

- Imagine instead that we wanted to work directly with the variational objective function, $\mathcal{L}$.

  - Generically, $\mathcal{L} = \mathbb{E}_q[\ln p(X, \theta)] - \mathbb{E}_q[\ln q(\theta)]$. In terms of the CEF form we've been talking about, this is

$$\mathcal{L} = \mathbb{E}_q[\ln p(w_{1:N}, \theta_i)] - \mathbb{E}_q[\ln q(\theta_i)] + \text{things that don't depend on } q(\theta_i)$$

- We've defined $q(\theta_i) = f(\chi', \nu') \exp\{\eta(\theta_i)^T \chi' - \nu' A(\eta(\theta_i))\}$. Therefore focusing just on θ_i,

$$\begin{aligned}\mathcal{L}_{\theta_i} &= \mathbb{E}_q[\eta_i^T(\chi + \textstyle\sum_n t(w_n)) - (\nu + N)A(\eta_i)] - \mathbb{E}_q[\eta_i^T \chi' - \nu' A(\eta_i)] - \ln f(\chi', \nu') \\ &= \mathbb{E}_q[\eta_i]^T \Big(\chi + \sum_{n=1}^N \mathbb{E}_q[t(w_n)] - \chi'\Big) - (\nu + N - \nu')\mathbb{E}_q[A(\eta_i)] - \ln f(\chi', \nu') \quad (10.8)\end{aligned}$$

- We need some equalities from the conjugate exponential family:

$$\begin{aligned}\int q(\theta_i) d\theta_i = 1 \;\Rightarrow\; \int \nabla_{\chi'} q(\theta_i) d\theta_i &= 0 \quad (10.9) \\ &= \int (\nabla_{\chi'} \ln f(\chi', \nu') + \eta_i) q(\theta_i) d\theta_i \\ \text{which means} \;\Rightarrow\; \mathbb{E}_q \eta_i &= -\nabla_{\chi'} \ln f(\chi', \nu')\end{aligned}$$

$$\begin{aligned}\int q(\theta_i) d\theta_i = 1 \;\Rightarrow\; \int \frac{d}{d\nu'} q(\theta_i) d\theta_i &= 0 \quad (10.10) \\ &= \int \Big(\frac{d}{d\nu'} \ln f(\chi', \nu') + A(\eta_i)\Big) q(\theta_i) d\theta_i \\ \text{which means} \;\Rightarrow\; \mathbb{E}_q A(\eta_i) &= -\frac{d}{d\nu'} \ln f(\chi', \nu')\end{aligned}$$

- Therefore, the general form of the objective for θ_i is

$$\mathcal{L}_{\theta_i} = -\nabla_{\chi'} \ln f(\chi', \eta')^T(\chi + \textstyle\sum_n \mathbb{E}_q t(w_n) - \chi') - \frac{d}{d\nu'} \ln f(\chi', \nu')(\nu + N - \nu') - \ln f(\chi', \nu') \quad (10.11)$$

- Our goal is to maximize this with respect to (χ', ν'). If we take the gradient of $\mathcal{L}$ with respect to these parameters and set it to zero,

$$\nabla_{(\chi',\nu')}\mathcal{L} = -\left[\begin{array}{cc} \frac{d^2 \ln f}{d\chi' d\chi'^T} & \frac{d^2 \ln f}{d\chi' d\nu'} \\ \frac{d^2 \ln f}{d\nu' d\chi'^T} & \frac{d^2 \ln f}{d\nu'^2} \end{array}\right] \left[\begin{array}{c} \chi + \sum_n \mathbb{E}_q[t(w_n)] - \chi' \\ \nu + N - \nu' \end{array}\right] = 0 \tag{10.12}$$

we see that we get exactly the same updates that we derived before using the "log-expectation-exponentiate" approach. We can simply ignore the matrix and set the right vector equal to zero.

Variational inference from the gradient perspective

- The most general way to optimize an objective function is via gradient ascent.

- In the variational inference setting, this means updating $q(\theta)i|\psi_i)$ according to $\psi_i \leftarrow \psi_i + \rho M \nabla_{\psi_i}\mathcal{L}$, where ρ is step size and M is a positive definite matrix.

- Ideally, we want to maximize with respect to ψ_i, meaning set ψ_i such that $\nabla_{\psi_i}\mathcal{L} = 0$. We saw how we could do that in closed form with CEF models.

- Equivalently, if $\left[\begin{array}{c} \chi' \\ \nu' \end{array}\right] \leftarrow \left[\begin{array}{c} \chi' \\ \nu' \end{array}\right] + \rho M \nabla_{(\chi',\nu')}\mathcal{L}$, we can set $\rho = 1$ and $M = -\left[\frac{d^2 \ln f}{d \cdot d \cdot^T}\right]^{-1}$. Then this step would move directly to the closed form solution for ψ_i (i.e., where $\nabla_{\psi_i}\mathcal{L} = 0$).

- We observe that $M = -\left[\frac{d^2 \ln f}{d \cdot d \cdot^T}\right]^{-1} = -\left[\frac{d^2 \ln q(\theta_i)}{d \cdot d \cdot^T}\right]^{-1}$. This second matrix is the inverse of the Fisher information matrix. When M is set to this value, we are said to be moving in the direction of the natural gradient.

- Therefore, in sum, when we update χ' and ν' in closed form as previously discussed, we are implicitly taking a full step (since $\rho = 1$) in the direction of the natural gradient (because of setting of M) and landing directly on the optimum. The important point being stressed here is that there is an implied gradient algorithm being executed when we calculate our closed-form parameter updates in variational inference for CEF models.

- This last point motivates the following discussion on scalable extensions of variational inference for CEF models.

- In many data modeling problems we have two issues:
  - N is huge (# documents in corpus, # users, etc)
  - Calculating $q(w_n)$ requires non-trivial amount of work

  Both of these issues lead to slow algorithms. (How slow? e.g., one day = one iteration)

- We'll next show how stochastic optimization can be applied to this problem in a straightforward way for CEF models.

Simple notation

- Let θ be a "global variable," meaning it's something that interacts with each unit of data.

- Let w_n be a set of "local variables" and associated data. Given θ, w_n and $w_{n'}$ are independent of each other.

Examples:

- The Gaussian mixture model, $\theta = \{\underbrace{\pi}_{\text{mixing weights}}, \underbrace{\mu_{1:K}, \Sigma_{1:K}}_{\text{Gaussian params}}\}, \quad w_n = \{\underbrace{x_n}_{\text{data}}, \underbrace{c_n}_{\text{cluster}}\}$.
- LDA, $\theta = \{\underbrace{\beta_1, \dots, \beta_k}_{\text{topics}}\}, \quad w_n = \{\underbrace{\theta_n}_{\text{topic dist}}, \underbrace{\vec{c}_n}_{\text{word alloc}}, \underbrace{\vec{x}_n}_{\text{word obs}}\}$

- The variational objective function is,

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_q\Big[\ln \frac{p(w_n|\theta)}{q(w_n)}\Big] + \mathbb{E}_q \ln p(\theta) - \mathbb{E}_q \ln q(\theta). \tag{10.13}$$

- Imagine if we subsampled $S_t \subset \{1, \dots, N\}$ and created

$$\mathcal{L}_t = \frac{N}{|S_t|} \sum_{n \in S_t} \mathbb{E}_q\Big[\ln \frac{p(w_n|\theta)}{q(w_n)}\Big] + \mathbb{E}_q \ln p(\theta) - \mathbb{E}_q \ln q(\theta). \tag{10.14}$$

- There are $\binom{N}{|S_t|}$ possible subsets of $\{1, \dots, N\}$ of size $|S_t|$, each having probability $1/\binom{N}{|S_t|}$.

- Each w_n appears in $\binom{N-1}{|S_t|-1}$ of the $\binom{N}{|S_t|}$ subsets.

- Therefore, we can calculate $\mathbb{E}\mathcal{L}_t$ summing over the random subsets as follows

$$\begin{aligned} \mathbb{E}\mathcal{L}_t &= \binom{N}{|S_t|}^{-1} \sum_{S_t} \mathcal{L}_t \qquad (10.15) \\ &= \frac{N}{|S_t|} \sum_{n=1}^{N} \binom{N}{|S_t|}^{-1} \binom{N-1}{|S_t|-1} \mathbb{E}_q\Big[\ln \frac{p(w_n|\theta)}{q(w_n)}\Big] + \mathbb{E}_q \ln p(\theta) - \mathbb{E}_q \ln q(\theta). \end{aligned}$$

- Observe that $\binom{N}{|S_t|}^{-1}\binom{N-1}{|S_t|-1} = \frac{|S_t|}{N}$. Therefore, $\mathbb{E}\mathcal{L}_t = \mathcal{L}$.

Stochastic variational inference (SVI)

- Idea: What if for each iteration we subsampled a random subset of w_n indexed by S_t, optimized their $q(w_n)$ only and then took a gradient step on parameters of $q(\theta)$ using $\mathcal{L}_t$ instead of $\mathcal{L}$? We would have SVI.

- <u>Stochastic variational inference</u>: Though the technique can be made more general, we restrict our discussion to CEF models.

- Method:

  1. At iteration t, subsample $w_1, \dots, w_N$ according to randomly generated index set S_t.
  2. Construct $\mathcal{L}_t = \dfrac{N}{|S_t|} \displaystyle\sum_{n \in S_t} \mathbb{E}_q \Big[\ln \frac{p(w_n|\theta)}{q(w_n)} \Big] + \mathbb{E}_q \ln p(\theta) - \mathbb{E}_q \ln q(\theta)$.
  3. Optimize $q(w_n)$ for $n \in S_t$.
  4. Update $q(\theta|\psi)$ by setting $\psi \leftarrow \psi + \rho_t M \nabla_\psi \mathcal{L}_t$.

<u>A closer look at Step 4 (the key step)</u>

- By following the same calculations as before (and letting $\psi = [\chi', \nu']$,

$$\nabla_{(\chi',\nu')} \mathcal{L} = - \left[\begin{array}{cc} \frac{d^2 \ln f}{d\chi' d\chi'^T} & \frac{d^2 \ln f}{d\chi' d\nu'} \\ \frac{d^2 \ln f}{d\nu' d\chi'^T} & \frac{d^2 \ln f}{d\nu'^2} \end{array} \right] \left[\begin{array}{c} \chi + \frac{N}{|S_t|} \sum_{n \in S_t} \mathbb{E}_q[t(w_n)] - \chi' \\ \nu + |S_t| - \nu' \end{array} \right] \tag{10.16}$$

- If we set $M = - \left[\begin{array}{cc} \frac{d^2 \ln f}{d\chi' d\chi'^T} & \frac{d^2 \ln f}{d\chi' d\nu'} \\ \frac{d^2 \ln f}{d\nu' d\chi'^T} & \frac{d^2 \ln f}{d\nu'^2} \end{array} \right]^{-1}$, then this simplifies nicely.

$$\left[\begin{array}{c} \chi' \\ \nu' \end{array} \right] \leftarrow (1 - \rho_t) \left[\begin{array}{c} \chi' \\ \nu' \end{array} \right] + \rho_t \left[\begin{array}{c} \chi + \frac{N}{|S_t|} \sum_{n \in S_t} \mathbb{E}_q[t(w_n)] \\ \nu + |S_t|' \end{array} \right] \tag{10.17}$$

- What has changed?
  - We are now looking at subsets of data.
  - Therefore, we don't set $\rho_t = 1$ like before. In general, stochastic optimization requires $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 < \infty$ for convergence.

- What is this doing?
  - Before, we were making a full update using all the data $(\rho_t = 1)$. Now, since we only look at a subset, we average the "full update" (restricted to the subset) with the most recent value of the parameters χ' and ν'. Since $\rho_t \to 0$, we weight this new information less and less.

Chapter 11

Variational inference for non-conjugate models

- Variational inference

  Have: Data X, model variables $\Theta = \{\theta_i\}$, factorized distribution $q(\Theta) = \prod_i q(\theta_i)$.

  Setup:
  $$\underbrace{\ln p(X)}_{\text{constant}} = \underbrace{\int \left(\prod_i q(\theta_i)\right) \ln \frac{p(X,\Theta)}{\prod_i q(\theta_i)} d\Theta}_{\text{variational objective } \mathcal{L}} + \underbrace{\int \left(\prod_i q(\theta_i)\right) \ln \frac{\prod_i q(\theta_i)}{p(\Theta|X)} d\Theta}_{\text{KL divergence} \geq 0}$$

  Variational inference: Find parameters of each $q(\theta_i)$ to maximize $\mathcal{L}$, which simultaneously minimizes $KL(q\|p(\Theta|X))$.

  Potential problems: This assumes that we can actually compute $\mathcal{L} = \mathbb{E}_q \ln p(X,\Theta) - \mathbb{E}_q \ln q(\Theta)$. Even if we use the "trick" from before, where we set $q(\theta_i) \propto \exp\{\mathbb{E}_{-q} \ln p(X,\Theta)\}$, we still need to be able to take these expectations.

- Examples: We'll give three examples where this problem arises, followed by possible solutions.

  1. Poisson matrix factorization has a generative portion containing (loosely) something like $Y \sim \text{Pois}(a_1 b_1 + \cdots + a_k b_k)$. Y is the data nad a_i, b_i are r.v.'s we want to learn.
     - Log-joint: $\ln p(Y,a,b) = Y \ln \sum_i a_i b_i - \sum_i a_i b_i + \sum_i \ln p(a_i)p(b_i) + \text{const.}$
     - Problem: $\mathbb{E} \ln \sum_i a_i b_i$ is intractable for continuous distribution $q(a,b) = \prod_i q(a_i)q(b_i)$.
  2. Logistic normal models look like $c_n \sim \text{Disc}(p)$, with $p_i \propto e^{x_i}$ and $x \sim N(\mu, \Sigma)$.
     - Log-joint: $\ln p(c,x) = \sum_i x_i \sum_n \mathbb{1}(c_n = i) - N \ln \sum_i e^{x_i} + \ln p(x)$.
     - Problem: $-\mathbb{E} \ln \sum_i e^{x_i}$ is intractable, where $q(x) = N(m, S)$.
  3. Logistic regression has the process $y_n \sim \sigma(x_n^T w)\delta_1 + (1 - \sigma(x_n^T w))\delta_{-1}$, $\sigma(a) = \frac{1}{1+e^{-a}}$.
     - Log-joint: $\ln p(y, w|X) = -\sum_n \ln(1 + e^{-y_n x_n^T w}) + \ln p(w)$.
     - Problem: $-\mathbb{E} \ln(1 + e^{-y_n x_n^T w})$ is intractable using continuous $q(w)$.

- Instant solution: Setting $q(\theta) = \delta_\theta$ corresponds to a point estimate of θ. We don't need to do the intractable integral in that case.

- One solution to intractable objectives is to find a more tractable lower bound of the function that's intractable (since we're in the maximization regime).
  - The tighter the lower bound the better the function is being approximated. Therefore, not all lower bounds are equally good.
  - Lower bounds often use additional parameters that control how tight they are. These are auxiliary variables that are optimized along with q.

- Solutions to the three problems

1. $\ln()$ is concave. Therefore, $\ln \mathbb{E}x \geq \mathbb{E}\ln x$. If we introduce $p \in \Delta_K$, then

$$\mathbb{E}\ln\sum_i a_i b_i = \mathbb{E}\ln\sum_i p_i \frac{a_i b_i}{p_i} \geq \mathbb{E}\left[\sum_i p_i \ln\frac{a_i b_i}{p_i}\right] = \sum_i p_i \mathbb{E}[\ln a_i + \ln b_i] - \sum_i p_i \ln p_i \tag{11.1}$$

   Often $\mathbb{E}\ln x$ is tractable. After updating $q(a_i)$ and $q(b_i)$, we also need to update the vector p to tighten the bound.

2. $-\ln()$ is convex, therefore we can use a first-order Taylor expansion

$$-\mathbb{E}\ln\sum_i e^{x_i} \geq -\ln\xi - \mathbb{E}\left[(\sum_i e^{x_i} - \xi)\frac{d\ln z}{dz}\Big|_\xi\right] = -\ln\xi - \frac{\sum_i \mathbb{E}e^{x_i} - \xi}{\xi} \tag{11.2}$$

   Usually we can calculate $\mathbb{E}e^{x_i}$ (MGF). We then update ξ after $q(x)$.

3. Finding a lower bound for $-\ln(1 + e^{-y_n x_n^T w})$ is more challenging. One bound due to Jaakola and Jordan in 2000 is

$$-\ln(1 + e^{-y_n x_n^T w}) \geq \ln\sigma(\xi_n) + \frac{1}{2}(y_n x_n^T w - \xi_n) - \lambda(\xi_n)(w x_n x_n^T w - \xi_n^2) \tag{11.3}$$

   where $\lambda(\xi_n) = (2\sigma(\xi_n) - 1)/(4\xi_n)$.
   - In this case there is a xi_n for each x_n since the functions we want to bound are different for each observation.
   - Though it's more complicated, we can take all expectations now since the bound is quadratic in w.
   - In fact, when $q(w)$ and $p(w)$ are Gaussian, the solution for q is in closed form.
   - After solving for $q(w)$, we can get closed form updates of ξ_n.

- The way we update all auxiliary variables, p, ξ and ξ_n, is by differentiation and setting to zero.

- We next give a more concrete example found in Bayesian nonparametric models.

- A size-biased representation of the beta process (review)

- Definition: Let $\gamma > 0$ and μ a non-atomic probability measure. For $k = 1, 2, \dots$, draw $\theta_k \overset{iid}{\sim} \mu$ and $V_j \overset{iid}{\sim}$ Beta$(\gamma, 1)$. If we define $H = \sum_{k=1}^{\infty} \big(\prod_{j=1}^{k} V_j \big) \delta_{\theta_k}$ then $H \sim$ BP$(1, \gamma\mu)$.

  Proof: See notes from Lecture 4 for the proof.

- Bernoulli process (review)

- Recall that if we have H, then we can use it to draw a sequence of Bernoulli processes $\vec{Z}_n | H \overset{iid}{\sim}$ BeP(H), where $\vec{Z}_n = \sum_{k=1}^{\infty} z_{nk} \delta_{\theta_k}$, $z_{nk} \sim$ Bern$\big(\prod_{i=1}^{k} V_j \big)$.

  ($\vec{Z}_n$ is also often treated as latent and part of data generation)

Variational inference for the beta process (using this representation)

- We simply focus on inference for $V_1, V_2, \dots$. Inference for θ_k and $\vec{Z}_n$ are problem-specific (and possibly don't have non-conjugacy issues)

- Let everything else in the model be represented by X. X contains all the data and any variables besides $\vec{Z}$ and Θ (I'll switch to Z instead of $\vec{Z}$ now)

- Joint likelihood: The joint likelihood is the first thing we need to write out. This can factorize as
$$p(X, \Theta, Z, V) = p(X|\Theta, Z) p(Z|V) p(V) p(\Theta). \tag{11.4}$$

- Therefore, to update the variational distribution $q(V)$ given all other q distributions, we only need to focus on $p(Z|V)p(V)$.

- For now, we assume $q(Z)$ is a delta function. That is, we have a 0-or-1 point estimate for the values in Z.

- So we focus on $p(Z|V)p(V)$ with Z a point estimate. According to the prior/likelihood definition, this further factorizes as

$$\begin{aligned} p(Z, V) &= \underbrace{\left[\prod_{n=1}^{N} \prod_{k=1}^{\infty} p(z_{nk} | V_1, \dots, V_k) \right]}_{\text{likelihood}} \underbrace{\left[\prod_{k=1}^{\infty} p(V_k) \right]}_{\text{prior}} \quad \leftarrow \text{ truncate at } T \text{ atoms} \\ &\propto \left[\prod_{n=1}^{N} \prod_{k=1}^{T} \Big(\prod_{j=1}^{k} V_j \Big)^{z_{nk}} \Big(1 - \prod_{j=1}^{k} V_j \Big)^{1 - z_{nk}} \right] \left[\prod_{k=1}^{T} V_j^{\gamma - 1} \right] \end{aligned} \tag{11.5}$$

- We pick $q(V) = \prod_{k=1}^{T} q(V_k)$ and leave the form undefined for now.

- The variational objective related to V is

$$\mathcal{L}_V = \sum_{n=1}^{N}\sum_{k=1}^{T}\left[z_{nk}\sum_{j=1}^{k}\mathbb{E}_q \ln V_j + (1-z_{nk})\mathbb{E}_q \ln(1-\textstyle\prod_{j=1}^{k} V_j)\right] + \sum_{k=1}^{T}(\gamma-1)\mathbb{E}_q \ln V_j - \sum_{k=1}^{T}\mathbb{E}_q \ln q \tag{11.6}$$

<u>Problem</u>: The expectation $\mathbb{E}_q \ln(1-\prod_{j=1}^{k} V_j)$ is intractable.

<u>Question</u>: How can we update $q(V_k)$?

<u>A solution</u>: Try lower bounding $\ln(1-\prod_{j=1}^{k} V_j)$ in such a way that it works.

<u>Recall</u>: An aside from Lecture 3 stated that $1-\prod_{j=1}^{k} V_j = \sum_{i=1}^{k}(1-V_i)\prod_{j=1}^{i-1} V_j$. Therefore, we can try looking for a lower bound using this instead.

- Since $\ln$ is a concave function, $\ln \mathbb{E}X \geq \mathbb{E}\ln X$. Therefore, if we introduce a k-dimensional probability vector p, we have

$$\begin{aligned} \ln\Big(\sum_{i=1}^{k} p_i(1-V_i)\prod_{j=1}^{i-1} V_j/p_i\Big) &\geq \sum_{i=1}^{k} p_i \ln\left[\frac{(1-V_i)\prod_{j=1}^{i-1} V_j}{p_i}\right] \\ &= \sum_{i=1}^{k} p_i\Big(\ln(1-V_i) + \sum_{j=1}^{i-1}\ln V_j\Big) - \sum_{i=1}^{k} p_i \ln p_i \qquad (11.7) \end{aligned}$$

Therefore,

$$\mathbb{E}_q \ln\Big(1-\sum_{j=1}^{k} V_j\Big) \geq \sum_{i=1}^{k} p_i\Big(\mathbb{E}_q \ln(1-V_i) + \sum_{j=1}^{i-1}\mathbb{E}_q \ln V_j\Big) - \sum_{i=1}^{k} p_i \ln p_i \tag{11.8}$$

(Aside) This use of p is the variational lower bound idea and is often how variational inference is presented in papers:

$$\ln p(X) = \ln\int p(X,\Theta)d\Theta = \ln\int q(\Theta)\frac{p(X,\Theta)}{q(\Theta)}d\Theta \geq \int q(\Theta)\ln\frac{p(X,\Theta)}{q(\Theta}d\Theta \equiv \mathcal{L}. \tag{11.9}$$

- We simply replace $\mathbb{E}_q \ln\Big(1-\prod_{i=1}^{k} V_i\Big)$ with this lower bound and hope it fixes the problem. Since we have this problematic term for all $k = 1,\dots,T$, we have T auxiliary probability vectors $p^{(1)},\dots,p^{(T)}$.

- We can still use the method of setting $q(V_k) = \exp\{\mathbb{E}_{-q}\ln p(Z,V)\}$ to find the best $q(V_k)$ but replacing $p(Z,V)$ with the lower bound of it. Therefore, $q(V_k)$ is the best q for the lower bound version of $\mathcal{L}$ (the function we're now optimizing), not the true $\mathcal{L}$.

- In this case, we find that we get a (very complicated) analytical solution for $q(V_k)$:

$$\begin{aligned} q(V_k) &\propto V_k^{\gamma+\sum_{m=k}^{T}\sum_{n=1}^{N} z_{nk}+\sum_{m=k+1}^{T}\sum_{n=1}^{N}(1-z_{nk}(\sum_{i=k+1}^{m} p_i^{(m)})-1} \times \\ & \quad (1-V_k)^{1+\sum_{m=k}^{T}\sum_{n=1}^{N}(1-z_{nk})p_k^{(m)}-1} \end{aligned} \tag{11.10}$$

But this simply has the form of a beta distribution with parameters that can be calculated "instantaneously" by a computer.

Un-addressed question: What do we set the probability vector $p^{(m)}$ equal to?

Answer: We can treat it like a new model parameter and optimize over it along with the variational parameters. In general, if we want to optimize p in $\sum_{i=1}^{k} p_i \mathbb{E} y_i - \sum_{i=1}^{k} p_i \ln p_i$, we can set

$$p_i \propto \exp\{\mathbb{E} y_i\}. \tag{11.11}$$

In this case, $\mathbb{E} y_i = \mathbb{E}\ln(1-V_i) + \sum_{j=1}^{i-1}\mathbb{E}\ln V_j$, which is fast to compute.

- We can iterate between updating $p^{(1)}, \ldots, p^{(T)}$ and $q(V_1), \ldots, q(V_T)$ before moving on to the other q distribution, or update each once and move to the other q's before returning.

Directly optimizing $\mathcal{L}$

- Rather than introduce approximations, we would like to find other ways to optimize $\mathcal{L}$. Specifically, is there a way to optimize it without taking expectations at all?

Setup: Coordinate ascent

Given $q(\theta)\prod_i q(\theta_i)$, the following is the general way to optimize $\mathcal{L}$.

1. Pick a $q(\theta_i|\phi_i) \;\leftarrow\; \phi_i$ are variational parameters
2. Set $\nabla_{\phi_i}\mathcal{L} = 0$ w.r.t. ϕ_i (if possible) or $\phi_i \leftarrow \phi_i + \rho\nabla_{\phi_i}\mathcal{L}$ (if not)
3. After updating each $q(\theta_i)$ this way, repeat.

Problem: For $q(\theta_i|\phi_i)$, $\mathbb{E}_q \ln p(X,\Theta)$ isn't (totally) in closed form.

Detail: Let

$$\ln p(X,\Theta) = \underbrace{f_1(X,\theta_i)}_{\text{problem}} + \underbrace{f_2(X,\Theta)}_{\text{no problem}} \quad \rightarrow \quad \mathcal{L} = \mathbb{E}f_1 + \underbrace{\mathbb{E}f_2 - \mathbb{E}\ln q}_{= h(X,\Psi)} \tag{11.12}$$

f_1 isn't necessarily the log of an entire distribution, so f_1 and f_2 could contain θ_i. Ψ contains all the variational parameters of q.

- What we need is $\nabla_{\psi_i}\mathcal{L} = \underbrace{\nabla_{\psi_i}\mathbb{E}f_1(X,\theta_i)}_{\text{hard}} + \underbrace{\nabla_{\psi_i}h(X,\Theta)}_{\text{easy}}.$

Solution (version 1): Create an unbiased estimate of $\nabla_{\psi_i}\mathbb{E}f_1$

$$\begin{aligned}\nabla_{\psi_i}\mathbb{E}f_1 &= \nabla_{\psi_i}\int f_1(X,\theta_i)q(\theta_i|\psi_i)d\theta_i \\ &= \int f_1(X,\theta_i)q(\theta_i|\psi_i)\nabla_{\psi_i}\ln q(\theta_i|\psi_i)d\theta_i \end{aligned} \tag{11.13}$$

We use the log trick for this $q\nabla_\psi \ln q = \frac{q}{q}\nabla_\psi q = \nabla_\psi q$.

- Use Monte Carlo integration:

$$\mathbb{E}_q X \approx \frac{1}{S}\sum_{s=1}^{S} X_s, \quad X_s \overset{iid}{\sim} q \quad \rightarrow \quad \lim_{S\to\infty}\frac{1}{S}\sum_{s=1}^{S} X_s = \mathbb{E}_q X \tag{11.14}$$

Therefore, we can use the following approximation,

$$\begin{aligned}\int f_1(X,\theta_i)q(\theta_i|\psi_i)\nabla_{\psi_i}\ln q(\theta_i|\psi_i)d\theta_i &= \mathbb{E}_q[f_1(X,\theta_i)\nabla_{\psi_i}\ln q(\theta_i|\psi_i)] \\ &\approx \underbrace{\frac{1}{S}\sum_{s=1}^{S} f_1(X,\theta_i^{(s)})\nabla_{\psi_i}\ln q(\theta_i^{(s)}),}_{\equiv \overline{\nabla_{\psi_i}\mathbb{E}f_1}} \quad \theta_i^{(s)} \overset{iid}{\sim} q(\theta_i|\psi_i)\end{aligned} \tag{11.15}$$

Therefore, $\nabla_{\psi_i}\mathcal{L} \approx \overline{\nabla_{\psi_i}\mathbb{E}f_1} + \nabla_{\psi_i}h(X,\Psi)$ and RHS is an unbiased approximation, meaning the expectation of the RHS equals the LHS.

Problem with this

- Even though the expectation of this approximation equals the truth, the variance might be huge, meaning S should be large.

Review: If $\hat{X}_S = \frac{1}{S}\sum_{s=1}^{S} x_s, x_s \overset{iid}{\sim} q$. Then $\mathbb{E}\hat{X}_S = \mathbb{E}_q X$ and $Var(\hat{X}_S) = Var(X)/S$. If we want $Var(\hat{X}_s) < \epsilon$, then $S > Var(X)/\epsilon$ (possible huge!)

Need: Variance reduction, which is a way for approximating $\mathbb{E}f(\theta)$ that is unbiased, but has smaller variance. One such method is to use a control variate.

Control variates: A way of approximating $\mathbb{E}_q f(\theta)$ with less variance than using Monte Carlo integration as described above.

Introduce a function $g(\theta)$ of the same variables such that

- $g(\theta)$ and $f(\theta)$ are highly correlated
- $\mathbb{E}g(\theta)$ can be calculated (unlike $\mathbb{E}f(\theta)$)

Procedure:

- Replace $f(\theta)$ with $\hat{f}(\theta) = f(\theta) - a(g(\theta) - \mathbb{E}g(\theta))$ in $\mathcal{L}$, with $a \in \mathbb{R}$
- Replace $\frac{1}{S}\sum_{s=1}^{S} f(\theta_s)\nabla \ln q(\theta_s)$ with $\frac{1}{S}\sum_{s=1}^{S} \hat{f}(\theta_s)\nabla \ln q(\theta_s)$

Why this works better:

- $\mathbb{E}_q\hat{f}(\theta) = \mathbb{E}f(\theta) - a(\mathbb{E}g(\theta) - \mathbb{E}[\mathbb{E}g(\theta)]) = \mathbb{E}f(\theta) \leftarrow$ unbiased
- $Var(\hat{f}(\theta)) = \mathbb{E}[(\hat{f}(\theta) - \mathbb{E}\hat{f}(\theta))^2] = Var(f(\theta)) - 2aCov(f(\theta), g(\theta)) + a^2Var(g(\theta))$

- Setting $a = \frac{Cov(f,g)}{Var(f)}$ minimizes $V(\hat{f}(\theta))$ above. In this case, we can see that $\frac{Var(\hat{f}(\theta))}{Var(f(\theta))} = 1 - Corr(f,g)^2$.

- Therefore, when f and g are highly correlated, the variance of $\hat{f}$ is much less than f. This translates to fewer samples to approximate $\mathbb{E}\hat{f}(\theta)$ within a accuracy threshold than needed for $\mathbb{E}f(\theta)$ using the same threshold.

- In practice, the optimal value $a = \frac{Cov(f,g)}{Var(f)}$ can be approximated each iteration using samples.

Modified coordinate ascent

1. Pick a $q(\theta_i|\psi_i)$
2. If $\mathcal{L}$ is closed form w.r.t. $q(\theta_i|\psi_i)$, optimize the normal way. Else, approximate $\nabla_{\psi_i}\mathcal{L}$ by $\nabla^*_{\psi_i}\mathcal{L} = \overline{\nabla_{\psi_i}\mathbb{E}f_1} + \nabla_{\psi_i}h(X,\Psi)$ and set $\psi_i \leftarrow \psi_i + \rho\nabla^*_{\psi_i}\mathcal{L}$. Possibly taking multiple steps.
3. After each $q(\theta_i|\psi_i)$ has been updated, repeat.

Picking control variates

- Lower bounds: By definition, a tight lower bound is highly correlated with the function it approximates. It's also picked to give closed form expectations.
- Taylor expansions (2nd order): A second order Taylor expansion is not a bound, but often it is a better approximation than a bound. in this case,

$$g(\theta) = f(\mu) + (\theta - \mu)^T\nabla f(\mu) + \frac{1}{2}(\theta - \mu)^T\nabla^2 f(\mu)(\theta - \mu). \tag{11.16}$$

  If the second moments of $q(\theta)$ can be calculated, then this is a possible control variate. μ could be the mean of $q(\theta|\psi)$ given the current value of ψ.

Chapter 12

Scalable MCMC inference

Abstract

In the last class, I present two recent papers that address scalability for Monte Carlo Metropolis Hastings (MCMC) inference:

1. M. Welling and Y.W. Teh, "Bayesian Learning via Stochastic Gradient Langevin Dynamics," *International Conference on Machine Learning*, 2011.
2. D. Maclaurin and R.P. Adams, "Firefly Monte Carlo: Exact MCMC with Subsets of Data," *Uncertainty in Artificial Intelligence*, 2014.

These presentations were given using slides, which I transcribe verbatim.

- Tons of digitized data that can potentially be modeled:
  - Online "documents" (blogs, reviews, tweets, comments, etc.)
  - Audio, video & images (Pandora, YouTube, Flickr, etc.)
  - Social media, medicine, economic, astronomy, etc. etc.
- It used to be that the problem was turning a mathematical model into reality in the most basic sense:
  - How do we learn it? (computer innovations)
  - Yes, but how do we *learn* it? (MCMC, optimization methods)
- Given the massive amount of data now readily accessible by a computer, all of which we want to use, the new question is:
  - Yes, but how do we learn it *in a reasonable amount of time*?
- Up to (and often including) now, the default solution has been to sub-sample a manageable amount and throw the rest away.

- Justification:

  – The large data we have is itself a small subsample of the set of all *possible* data.

  – So if we think that's a reasonable statistical representation, what's wrong with a little less (relatively speaking)?

- Example 1: Netflix has 17K movies, 480K users, 100M ratings
  – There are 8,160M *potential* ratings (1.2% measured), so if we remove users to learn about the movies, what can it hurt?

- Example 2: We have 2M newspaper articles of $\times 1000$ more existing articles. Why not subsample to learn the topics?

- We want to (and should) use all data since the more data we have, the more information we can extract.

- This translates to more "expressive" (i.e., complicated) probabilistic models in the context of this class.

- This hinges on having the computational resources to do it:

  - More powerful computers,
  - Recognizing and exploiting parallelization,
  - Using inference tricks to process all the data more efficiently

- Recent successes in large scale machine learning have mostly been optimization-based approaches:

  - Stochastic optimization has a huge literature
  - Stochastic variational inference ports this over to the problem of approximate posterior learning

- The common denominator of these methods is that they process small subsets of the data at a time.

- Example: With SVI we saw how each piece of data had its own local variables and all data shared global variables.

  - Pick a subset of the data,
  - update their local variables,
  - then perform a gradient step on the global variables.

- This idea of processing different subsets of data for each iteration is a general one.

- Can it be extended to MCMC inference?

- We'll review two recent ML papers on this:

  - M. Welling and Y.W. Teh, "Bayesian Learning via Stochastic Gradient Langevin Dynamics," ICML 2011.

  - D. Maclaurin and R.P. Adams, "Firefly Monte Carlo: Exact MCMC with Subsets of Data," UAI 2014.

  Bayesian Learning via Stochastic Gradient Langevin Dynamics

- This SGLD algorithm combines two things:

  - Stochastic gradient optimization

  - Langevin dynamics, which incorporates noise to the updates

- The stochastic gradients and Langevin dynamics are combined in such a way that MCMC sampling results.

- Considers a simple modeling framework with approximations
  (Alert: possible future research!)

- Let θ denote model vector and $p(\theta)$ its prior distribution.

- Let $p(x|\theta)$ be the probability of data x given model variable θ.

- The posterior distribution of N observations $X = \{x_1, \ldots, x_N\}$ is assumed to have the form

$$p(\theta|X) \propto p(\theta) \prod_{i=1}^{N} p(x_i|\theta).$$

- How do we work with this? Rough outline of following slides:
  - Maximum a posteriori with gradient methods
  - Stochastic gradient methods for MAP (big data extension)
  - Metropolis-Hastings MCMC sampling
  - Langevin diffusions (gradient + MH-type algorithm)
  - Stochastic gradient for Langevin diffusions

- A *maximum a posteriori* estimate of θ is

$$\theta_{\text{MAP}} = \arg\max_{\theta} \ \ln p(\theta) + \sum_{i=1}^{N} \ln p(x_i|\theta).$$

- This can be (locally) optimized using gradient ascent

  $$\theta_{t+1} = \theta_t + \Delta\theta_t$$

  where $\Delta\theta_t$ is the weighted gradient of the log joint likelihood

  $$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \ln p(\theta_t) + \sum_{i=1}^{N} \nabla \ln p(x_i|\theta_t)\right).$$

- Notice that this requires evaluating $p(x_i|\theta_t)$ for every x_i.

- *Stochastic gradient ascent* optimizes exactly the same objective function (or finds a local optimal solution).

- The only difference is that for each step t, a subset of $n \ll N$ observations is used in the step $\theta_{t+1} = \theta_t + \Delta\theta_t$:

  $$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \ln p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \ln p(x_{ti}|\theta_t)\right).$$

- x_{ti} is the ith observation in subset t:
  - This set is usually talked of as being randomly generated
  - In practice, contiguous chunks of n are often used

- For the *stochastic* gradient

  $$\Delta\theta_t = \tfrac{\epsilon_t}{2}\left(\nabla \ln p(\theta_t) + \tfrac{N}{n}\textstyle\sum_{i=1}^{n} \nabla \ln p(x_{ti}|\theta_t)\right)$$

  there are two requirements on ϵ_t to provably ensure convergence (see Leon Bottou's papers, e.g., one in 1998),

  $$\sum_{t=1}^{\infty} \epsilon_t = \infty, \qquad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty.$$

- Intuitively:
  - The first one ensures that the parameters can move as far as necessary to find the local optimum.
  - The second one ensures that the "noise" from using $n \ll N$ does not have "infinite impact" and the algorithm converges.

- Often used: $\epsilon_t = \left(\frac{a}{b+t}\right)^{\gamma}$ with $\gamma \in (0.5, 1]$.

- MAP (and ML) approaches are good tools to have, but they don't capture uncertainty and can lead to overfitting.

- A typical Bayesian approach is to use MCMC sampling, which is a set of techniques that gives samples from the posterior.

- One MCMC approach is "random walk Metropolis-Hasting":

  - Sample $\eta_t \sim N(0, \epsilon I)$ and set $\theta^*_{t+1} = \theta_t + \eta_t$
    – i.e., $\theta^*_{t+1} \sim q(\theta^*_{t+1}|\theta_t) = N(\theta_t, \epsilon I)$.

  - Set $\theta_{t+1} = \theta^*_{t+1}$ with probability $\frac{p(X,\theta^*_{t+1})q(\theta_t|\theta^*_{t+1})}{p(X,\theta_t)q(\theta^*_{t+1}|\theta_t)} = \frac{p(X,\theta^*_{t+1})}{p(X,\theta_t)}$, otherwise $\theta_{t+1} = \theta_t$.

- We evaluate the complete joint likelihood to calculate the probability of acceptance.

- After $t > T$, we can treat θ_t as samples from the posterior distribution $p(\theta|X)$ and use them for empirical averages.

- Langevin dynamics combines gradient methods with the idea of random walk MH sampling.

- Instead of treating t as an index, think of it as a continuous time and let

$$\theta_{t+\Delta t} = \theta_t + \Delta\theta_t$$

$$\Delta\theta_t = \frac{\epsilon\Delta t}{2}\nabla \ln p(X, \theta_t) + \sqrt{\epsilon}\eta_t, \quad \eta_t \sim N(0, \Delta t I).$$

- We have:

  - A random Gaussian component like random walk MH
  - A derivative component like gradient ascent.

- We could then have an accept/reject step similar to the one on the previous slide (but q doesn't cancel out).

- What happens to $\theta_{t+\Delta t} - \theta_t = \Delta\theta_t$ when $\Delta t \to 0$?

$$d\theta_t = \frac{\epsilon}{2}\nabla \ln p(X, \theta_t)dt + \sqrt{\epsilon}dW_t.$$

- We can show that we will always accept the new location.

- On RHS: The right term randomly explores the space, while the left term pulls in the direction of higher density.

- The path that θ maps out mimics the posterior density (beyond the scope of class to prove this).

- What does this mean?
  –If we wait and collect samples of θ every Δt amount of time (sufficiently large), we can treat them as samples from $p(\theta|X)$.

- We would naturally want to think of discretizing this.

- i.e., set

$$\theta_{t+\Delta t} = \theta_t + \Delta\theta_t$$
$$\Delta\theta_t = \frac{\epsilon\Delta t}{2}\nabla \ln p(X,\theta_t) + \eta_t, \quad \eta_t \sim N(0,\epsilon\Delta t I).$$

  for $0 < \Delta t \ll 1$, and accept with probability 1.

- (Notice an equivalent representation is used above.)
- The authors consider replacing $\epsilon\Delta t$ with ϵ_t such that

$$\textstyle\sum_{t=1}^{\infty}\epsilon_t = \infty, \qquad \sum_{t=1}^{\infty}\epsilon_t^2 < \infty.$$

- This leads to the insight and contribution of the paper:

  Replace $\nabla \ln p(X,\theta_t)$ with its stochastic gradient.

- That is, let $\theta_{t+1} = \theta_t + \Delta\theta_t$ with

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \ln p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n}\nabla \ln p(x_{ti},\theta_t)\right) + \eta_t$$
$$\eta_t \sim N(0,\epsilon_t I).$$

  again with $\sum_{t=1}^{\infty}\epsilon_t = \infty$, $\sum_{t=1}^{\infty}\epsilon_t^2 < \infty$.

- Notice that the *only* difference between this and stochastic gradient for MAP is the addition of η_t.
- The authors' give intuition why we can skip the accept/reject step by setting $\sum_{t=1}^{\infty}\epsilon_t = \infty$, $\sum_{t=1}^{\infty}\epsilon_t^2 < \infty$, and also why the following converges to the Langevin dynamics

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \ln p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n}\nabla \ln p(x_{ti},\theta_t)\right) + \sqrt{\epsilon_t}\eta_t$$
$$\eta_t \sim N(0,I) \leftarrow \text{notice the equivalent representation}$$

- It's not intended to be a proof, so we won't go over it.
- However notice that:
  - $\sqrt{\epsilon_t}$ is *huge* (relatively) compared to ϵ_t for $0 < \epsilon_t \ll 1$.
  - The sums of both terms individually are infinite, so both provide an "infinite" amount of information.

  - The sum of the variances of both terms is *finite* for the first and *infinite* for the second, so sub-sampling has finite impact.

- The paper considers the toy problem

$$\theta_1 \sim N(0, \sigma_1^2), \quad \theta_2 \sim N(0, \sigma_2^2),$$

$$x_i \sim \frac{1}{2} N(\theta_1, \sigma_x^2) + \frac{1}{2} N(\theta_1 + \theta_2, \sigma_x^2).$$

- Set $\theta_1 = 0$, $\theta_2 = 1$, sample $N = 100$ points.

- Set $n = 1$ and process the data set 10000 times.

- (Showed figures and experimental details from the paper.)

Firefly Monte Carlo: Exact MCMC with Subsets of Data

- We're again in the setting where we have a lot of data that's conditionally independent given a model variable θ.

- MCMC algorithms such as random walk Metropolis-Hastings are time consuming:

  Set $\theta_{t+1} \leftarrow \theta_{t+1}^* \sim N(\theta_t, \epsilon I)$ w.p. $\frac{p(\theta_{t+1}^*) \prod_i p(x_i|\theta_{t+1}^*)}{p(\theta_t) \prod_i p(x_i|\theta_t)}$.

- This requires many evaluations in the joint likelihood.

- Evaluating subsets for each t is promising, but previous methods are only approximate or asymptotically correct.

- This paper presents a method for performing *exact* MCMC sampling by working with subsets.

- High level intuition: Introduce a binary switch for each observation and only process those for which the switch is on.

- The switches change randomly each iteration, so they are like fireflies (1 = light, 0 = dark). Hence "Firefly Monte Carlo".

- This does require nice properties, tricks and simplifications which make this a non-trivial method for each new model.

- Again, we have a parameter of interest θ with prior $p(\theta)$.

- We have N observations $X = \{x_1, \ldots, x_N\}$ with N huge.

- The likelihood factorizes, therefore the posterior

$$p(\theta|X) \propto p(\theta) \prod_{n=1}^{N} p(x_n|\theta).$$

- For notational convenience let $L_n(\theta) = p(x_n|\theta)$, so

$$p(\theta|X) \propto p(\theta) \prod_{n=1}^{N} L_n(\theta).$$

- The initial setup for FlyMC is remarkably straightforward:

  - Introduce a new function B, with $B_n(\theta)$ the evaluation on x_n with parameter θ, such that $0 < B_n(\theta) \leq L_n(\theta)$ for all θ.

  - With each x_n associate an auxiliary variable $z_n \in \{0, 1\}$ distributed as

$$p(z_n|x_n, \theta) = \left[\frac{L_n(\theta) - B_n(\theta)}{L_n(\theta)}\right]^{z_n} \left[\frac{B_n(\theta)}{L_n(\theta)}\right]^{1-z_n}$$

- The *augmented* posterior is

$$p(\theta, Z|X) \propto p(\theta) \prod_{n=1}^{N} p(x_n|\theta) p(z_n|x_n, \theta).$$

- Summing out z_n from $p(\theta) \prod_{n=1}^{N} p(x_n|\theta) p(z_n|x_n, \theta)$ leaves the original joint likelihood in tact as required.

- Therefore, this is an auxiliary variable sampling method:

  - Before: Sample θ with Metropolis-Hastings
  - Now: Iterate sampling each z_n and then θ with MH.

- As with all MCMC methods, to get samples from the desired $p(\theta|X)$, keep the θ samples and throw away the Z samples.

- What has been done so far is always possible as a general rule. The question is if it makes things easier (here, faster).

- To this end, we need to look at the augmented joint likelihood. For a particular term,

$$\begin{aligned} p(x_n|\theta)p(z_n|x_n,\theta) &= L_n(\theta)\left[\frac{L_n(\theta)-B_n(\theta)}{L_n(\theta)}\right]^{z_n}\left[\frac{B_n(\theta)}{L_n(\theta)}\right]^{1-z_n} \\ &= \begin{cases} L_n(\theta)-B_n(\theta) & \text{if } z_n=1 \\ B_n(\theta) & \text{if } z_n=0 \end{cases} \end{aligned}$$

- For each iteration, we only need to evaluate the likelihood $L_n(\theta)$ at points where $z_n = 1$.
  Alarm bells: What about $B_n(\theta)$?

- What about all the $B_n(\theta)$, which appear regardless of z_n?

- Look at the augmented joint likelihood:

$$p(X, Z, \theta) = \hat{p}(\theta) \prod_{n:z_n=1} \hat{L}_n(\theta),$$

$$\hat{p}(\theta) = p(\theta)\prod_{n=1}^{N} B_n(\theta), \quad \hat{L}_n(\theta) = \frac{L_n(\theta)-B_n(\theta)}{B_n(\theta)}.$$

- When we sample a new θ using MH, we compute $\frac{p(X,Z,\theta^*_{t+1})}{p(X,Z,\theta_t)}$.

- We have to evaluate
  - $p(\theta) \rightarrow$ fast
  - $\hat{L}_n(\theta)$ for each $n : z_n = 1 \rightarrow$ fast *if* $\sum_n z_n \ll N$
  - $\prod_{n=1}^{N} B_n(\theta) \rightarrow$ fast?

- This therefore indicates the two requirements we have about $B_n(\theta)$ for this to work well:
  - $B_n(\theta)$ is close to $L_n(\theta)$ for many n and θ so that few $z_n = 1$.
  - $\prod_{n=1}^{N} B_n(\theta)$ has to be quickly evaluated.

- Condition 1 states that $B_n(\theta)$ is a tight lower bound of $L_n(\theta)$ – Recall that $p(z_n = 1|x_n, \theta) = \frac{L_n(\theta)-B_n(\theta)}{L_n(\theta)}$.

- Condition 2 essentially means that B allows us to only have to evaluate a function of θ and *statistics* of $X = \{x_1, \ldots, x_N\}$.

- (Showed intuitive figure from paper.)

- This leaves two issues:

  - Actually choosing the function $B_n(\theta)$.
  - Sampling z_n for each n.

- Notice that to sample $z_n \sim Bern\left(\frac{L_n(\theta)-B_n(\theta)}{L_n(\theta)}\right)$, we have to evaluate $L_n(\theta)$ and $B_n(\theta)$ for each n.

- So our computations have actually increased.

- The paper presents methods for dealing with this as well.
- Problem 1: Choosing $B_n(\theta)$. Solution: Problem specific.

- Example: Logistic regression. For $x_n \in \mathbb{R}^d$ and $y_n \in \{-1, 1\}$,

$$L_n(\theta) = (1 + \mathrm{e}^{-y_n x_n^T \theta})^{-1} \geq \mathrm{e}^{a(\xi)(y_n x_n^T \theta)^2 + \frac{1}{2}(y_n x_n^T \theta) + c(\xi)} = B_n(\theta).$$

- $a(\xi)$ and $c(\xi)$ are functions of auxiliary parameter ξ.

- The product is
$$\prod_{n=1}^{N} B_n(\theta) = \mathrm{e}^{a(\xi)\theta^T (\sum_n x_n x_n^T)\theta + \frac{1}{2}\theta^T (\sum_n y_n x_n) + c(\xi)}$$

- We can calculate the necessary statistics from the data in advance. $\prod_{n=1}^{N} B_n(\theta)$ is then extremely fast to compute.
- Problem 2: Sampling z_n. Solution: Two general methods.

- Simplest: Sample new z_n for random subset each iteration.

- Alternative: Introduce MH step to propose switching light→dark and dark→light.

- Focus on z_n:

  - Propose $z_n \to z_n'$ by sampling $z_n' \sim q(z_n'|z_n)$.
  - Accept with probability $\frac{p(z_n'|x_n,\theta)q(z_n|z_n')}{p(z_n|x_n,\theta)q(z_n'|z_n)}$.

- Immediate observation: Always accept if $z_n' = z_n$.
- Focus on z_n:

  - Propose $z_n \to z_n'$ by sampling $z_n' \sim q(z_n'|z_n)$.

- Accept with probability $\frac{p(z'_n|x_n,\theta)q(z_n|z'_n)}{p(z_n|x_n,\theta)q(z'_n|z_n)}$.

- Case $z'_n = 0, z_n = 1$: Sample total from $Bern(\sum_n z_n, q_{1\to 0})$, then pick which x_n uniformly w.o. replacement.
  Observation: Can re-use $L_n(\theta)$ and $B_n(\theta)$ if rejected.

- Case $z'_n = 1, z_n = 0$: Sample from $Bern(N - \sum_n z_n, q_{0\to 1})$, then pick which x_n uniformly w.o. replacement.
  Observation: Want $q_{0\to 1}$ small to reduce evaluations of L, B.

- Final consideration: We want $B_n(\theta)$ tight, which is done by setting its auxiliary parameters.

- The authors discuss this in their paper for the logistic regression problem.

- It's an important consideration since it will control how many $z_n = 1$ in an iteration
  Tighter $B_n(\theta)$ means fewer $z_n = 1$ means faster algorithm.

- This procedure may require running a fast algorithm in advance (e.g., stochastic MAP).

- (Showed figures and experiments from the paper.)