IFT 6269: Probabilistic Graphical Models **Fall 2020**

Lecture 26 — December 4

Lecturer: Jose Gallego Posada *Scribe: Naga Karthik*

Disclaimer: These notes have only been lightly proofread.

Bayesian Non-Parametrics

The aim of this lecture is to briefly introduce Stochastic processes and dive a little deeper into Gaussian processes and Dirichlet processes. However, note that this lecture will not cover the implementational aspects and hyperparameter selection. For a more in-depth/complete understanding of these topics, the readers are referred to the works of David MacKay, Yee Whye Teh, Kilian Weinberger, Tamara Broderick, and Michael Jordan.

26.1 Stochastic Process

A stochastic process is defined as a collection of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a sample space, $\mathcal{F}$ is a sigma-algebra, and $\mathbb{P}$ is a probability measure; and the random variables, indexed by some set T, all take values in the same space S, measurable with respect to some σ-algebra.

Mathematically, it can be understood as: For every element $t \in T$, there is a random variable X (from a common probability space) indexed by t. Therefore, we have a simple definition as:

$$\{X(t) \mid t \in T\}$$

26.1.1 Examples of Stochastic Process

The following are some examples of stochastic processes:

- Single random variable: Consider any singleton set T, say $\{1\}$ for instance. Then, the collection of random variables indexed by T is given by $\{X_1\}$.
- IID random variables: Consider the finite set $\{1, 2, \ldots, n\}$ to be the indexing set T. The corresponding collection of random variables is given by $\{X_1, X_2, \ldots, X_n\}$, where X_i and X_j are independent and follow a common distribution X^*.
- Deterministic function: Let f be a function $f : T \to S$, given the index set T, the collection is defined as $\{X(t) := \delta_{f(t)} \mid t \in T\}$, where δ is the Dirac distribution.
- Wiener process, Poisson process, etc.
- Gaussian process
- Dirichlet process

26.2 Gaussian Processes

The goal is to do inference based on the inputs X and targets t and use that for prediction when presented with an unseen test input x^*. Inference is essentially calculating the posterior distribution over the models $y(\cdot)$ using Bayes' theorem:

$$\mathbb{P}(y(\cdot) \mid \boldsymbol{t}, \boldsymbol{X}) = \frac{\mathbb{P}(\boldsymbol{t} \mid y(\cdot), \boldsymbol{X})\mathbb{P}(y(\cdot))}{\mathbb{P}(\boldsymbol{t} \mid \boldsymbol{X})} \tag{26.1}$$

where, $\mathbb{P}(\boldsymbol{t} \mid y(\cdot), \boldsymbol{X})$ is the likelihood, $\mathbb{P}(y(\cdot))$ is the prior over functions, and $\mathbb{P}(\boldsymbol{t} \mid \boldsymbol{X})$ is the evidence. It is important to note that the prior $\mathbb{P}(y(\cdot))$ cannot be easily defined because it has to cover an infinite-dimensional space.

The posterior distribution is further used for prediction as it allows us to obtain the predictive distribution $\mathbb{P}(t^* \mid \boldsymbol{x}^*, \boldsymbol{t}, \boldsymbol{X})$ defined as:

$$\mathbb{P}(t^* \mid \boldsymbol{x}^*, \boldsymbol{t}, \boldsymbol{X}) = \int \mathbb{P}(t^* \mid y(\cdot), x^*)\mathbb{P}(y(\cdot) \mid \boldsymbol{t}, \boldsymbol{X})dy \tag{26.2}$$

where, $\mathbb{P}(y(\cdot) \mid \boldsymbol{t}, \boldsymbol{X})$ is computed from equation (26.1).

26.2.1 Gaussian Properties

We do a quick review of Gaussian properties as they are useful in the treatment of Gaussian processes. Consider the following setting where there are two random variables X, Y and they are Gaussian distributed with mean $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{XY}^\top & \boldsymbol{\Sigma}_Y \end{bmatrix} \right)$$

The following are some of the important properties of Gaussian distributions:

Property 26.2.1 *(Tractable) Normalization* - $\int \tilde{p}(x)dx = (2\pi)^{d/2}|\Sigma|^{1/2}$

- Allows us to have a closed-form solution and is also easier for sampling.

Property 26.2.2 *Marginalization* - $X \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$

- X is distributed according to a Gaussian with mean $\boldsymbol{\mu}_X$ and covariance $\boldsymbol{\Sigma}_X$. Can be proved using the *characteristic function* of a Gaussian.

Property 26.2.3 *Conditioning* - $X \mid Y = \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_Y), \boldsymbol{\Lambda}_X^{-1})$

- Where $\boldsymbol{\Lambda} := \boldsymbol{\Sigma}^{-1}$. Note that the conditional distribution is also a Gaussian.

Property 26.2.4 *Addition* - $X + Y = \mathcal{N}(\boldsymbol{\mu}_X + \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_Y + 2\boldsymbol{\Sigma}_{XY})$

- Addition of two Gaussian random variables is also a Gaussian.

Property 26.2.5 *Product of densities* - $\mathcal{N}(x \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \cdot \mathcal{N}(x \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \propto \mathcal{N}(x \mid \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$

- Note: This is **not** a product of two Gaussian random variables, rather, it's a product of two Gaussian densities that results in an un-normalized Gaussian density.

26.2.2 Linear Regression

The goal is to perform inference on a linear model. We start with the following assumptions:

- Locations $\{\boldsymbol{x}_i \mid i \in [1, N]\}$
- Basis functions $\{\phi_h(\boldsymbol{x})\}_{h=1}^{H}$ - some transformations of the input $\boldsymbol{x}$ using H basis functions.
- Linear model $y(\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$
- Prior $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \sigma_w^2 \boldsymbol{I})$ - a Gaussian defined over the weights $\boldsymbol{w}$ with isotropic covariance.
- Observation noise $t \mid y \sim \mathcal{N}(y, \sigma_\nu^2)$ - typically defined as $t = y + \sigma_\nu \epsilon$, where y is the "true value" that is not observed directly.

Given these assumptions, we have that the joint distribution of the true values of the model $\boldsymbol{y}$ is a Gaussian given by: $\boldsymbol{y}_N \sim \mathcal{N}(\boldsymbol{0}, \sigma_w^2 \Phi_N \Phi_N^\top)$ where Φ_N is the $n \times d$ design matrix where each $\boldsymbol{\phi}(\boldsymbol{x_i})$ is put as a row. Since we don't observe the true values directly but observe a noisy version of them, we arrive at a similar conclusion for the targets $\boldsymbol{t}$ which is: $\boldsymbol{t}_N \sim \mathcal{N}(\boldsymbol{0}, \sigma_w^2 \boldsymbol{\Phi}_N \boldsymbol{\Phi}_N^\top + \sigma_\nu^2 \boldsymbol{I})$.

Remark 26.2.1 *The term $\boldsymbol{\Phi}_N \boldsymbol{\Phi}_N^\top$ is interesting because considering a value $(\boldsymbol{\Phi}_N \boldsymbol{\Phi}_N^\top)_{ij}$, even before we observe the values of the noisy random variables, our prior encodes our belief that if features of datapoint i and features of datapoint j are similar, then the corresponding values of $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ will be close to each other.*

Remark 26.2.2 *The covariance matrix $(\sigma_w^2 \boldsymbol{\Phi}_N \boldsymbol{\Phi}_N^\top + \sigma_\nu^2 \boldsymbol{I})$ can be generalized to any positive semi-definite (PSD) similarity matrix. This has deep connections to the theory of kernel regression.*

Note: From here on, the similarity matrix is represented using C wherein $C(x_i, x_j)$ refers to the similarity between X_i and X_j.

26.2.3 Definition of a Gaussian Process

The probability distribution of a function $y(\boldsymbol{x})$ is a Gaussian process if for any finite selection of points $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)$, the vector $[y(\boldsymbol{x}_1), y(\boldsymbol{x}_2), \ldots, y(\boldsymbol{x}_N)]$ follows a Gaussian distribution. Recalling what we defined earlier for stochastic processes, we have that:

- Index set T - $\mathcal{X}$ - the set of possible inputs $\boldsymbol{x}$
- Value space S - $\mathbb{R}^l$ (can also be multi-dimensional)
- Random variables - Gaussians (it's essentially the function values acting as random variables for checking the distribution they follow)

Now, given all this information, **we finally observe the targets $\boldsymbol{t_N} = \{t_i \mid i \in [1, N]\}$** and would like to infer a new target $\boldsymbol{t_{N+1}}$ for a new test point $\boldsymbol{x_{N+1}}$. The similarity matrix for a new datapoint is given by:

$$\boldsymbol{C}_{N+1} = \begin{bmatrix} \boldsymbol{C}_N & \boldsymbol{k} \\ \boldsymbol{k}^\top & \kappa \end{bmatrix} \tag{26.3}$$

where, $\boldsymbol{C}_N$ is the collection of all the similarity values for the N datapoints in the space, an element in $\boldsymbol{k}$, say k_j, is the similarity value between j-th and the new datapoint X_{N+1}, i.e. $k_j = C(X_{N+1}, X_j)$, and κ is the similarity value between the new datapoint with itself, i.e. $C(X_{N+1}, X_{N+1})$.

The predictive distribution is now given by:

$$\mathbb{P}(t_{N+1} \mid \boldsymbol{t}_N) = \frac{\mathbb{P}(t_{N+1}, \boldsymbol{t}_N)}{\mathbb{P}(\boldsymbol{t}_N)} \propto \exp\left[\frac{-1}{2}\begin{bmatrix} \boldsymbol{t}_N & t_{N+1} \end{bmatrix} \boldsymbol{C}_{\boldsymbol{N+1}}^{-1} \begin{bmatrix} \boldsymbol{t}_N \\ t_{N+1} \end{bmatrix}\right] \tag{26.4}$$

For notational simplicity, the dependence on $\boldsymbol{\Phi}_N$ is omitted here. Note that this is only a distribution over t_{N+1}. Upon further simplification we have the following result:

$$t^* \mid t_1, \ldots, t_N, x_1, \ldots, x_n, x^* \sim \mathcal{N}(\boldsymbol{k}^\top \boldsymbol{C}_{\boldsymbol{N}}^{-1} t_N, \kappa - \boldsymbol{k}^\top \boldsymbol{C}_{\boldsymbol{N}}^{-1} \boldsymbol{k}) \tag{26.5}$$

Remark 26.2.3 *The dependence is on $\boldsymbol{C}_N$ (and not $\boldsymbol{C}_{N+1}$). So, for prediction over multiple test points, we only need to calculate the inverse of the similarity matrix over the data space $\boldsymbol{C}_N$ once, augment the corresponding $\boldsymbol{k}$ for all the new test points and then compute the Gaussian with mean and variance (using simple linear algebra).*

Remark 26.2.4 *The posterior mean is linear w.r.t $\boldsymbol{t}_N$ weighted by the term $\boldsymbol{k}^\top \boldsymbol{C}_{\boldsymbol{N}}^{-1}$. For instance, if we have positive correlation between elements, then a higher target value would correspond to a new test point distribution with higher mean.*

Note: A demo showing the working of a Gaussian process can be found at the following link: http://chifeng.scripts.mit.edu/stuff/gp-demo/

26.2.4 GP Summary

Problem: y is a function i.e. "an infinite-dimensional vector". But the multivariate Gaussian distribution is only defined for **finite** dimensional vectors.
Definition: A GP is a (potentially infinite) collection of random variables such that the joint distribution of **every finite subset** of them is a multivariate Gaussian.

If we start with the assumption that the function values are distributed according to a Gaussian prior i.e. $y(X_1), \ldots, y(X_N) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_N)$, then the posterior distribution is also a Gaussian of the form:

$$t^* \mid t_1, \ldots, t_N, x_1, \ldots, x_n, x^* \sim \mathcal{N}(\boldsymbol{k}^\top \boldsymbol{C}_{\boldsymbol{N}}^{-1} t_N, \kappa - \boldsymbol{k}^\top \boldsymbol{C}_{\boldsymbol{N}}^{-1} \boldsymbol{k})$$

26.3 Dirichlet Processes

Considering the generative modelling perspective, we define our model (say, a latent variable model) as follows:

- Assignments $z_n \mid \boldsymbol{\rho} \sim \text{Categorical}(\rho_1, \ldots, \rho_K)$ - the class assignments are obtained from a Categorical distribution with $\rho_i \geq 0$ and $\sum_i \rho_i = 1$.
- Class conditionals $\boldsymbol{x}_n \mid z_n \sim \mathcal{N}(\boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}) = F_{z_n}$ - the class conditionals depend on a specific value F_{z_n} of the latent variable.

We also make some prior assumptions as follows:

- Number of classes K is known.
- Class proportions $\boldsymbol{\rho} \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$ - following a Dirichlet distribution.
- Class parameters $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = H$ - the parameters defining the behaviour of F_{z_n} are themselves sampled from a Gaussian with fixed mean and covariance, called H.

26.3.1 Dirichlet Distribution

The Dirichlet distribution is formally defined as:

$$\mathbb{P}(\rho_1, \ldots, \rho_K \mid \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\Sigma_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \rho_k^{\alpha_k - 1} \tag{26.6}$$

Remark 26.3.1 *The Dirichlet distribution can be thought of as a distribution over distributions. This is because whenever we sample a random variable from this distribution, we get a collection of $\rho_1, \ldots \rho_K$ which in turn specifies a Categorical distribution. Note that $\rho_1, \ldots \rho_K$ belong to a simplex of $K - 1$ dimensions.*

Remark 26.3.2 *The Dirichlet distribution is a direct extension of the Beta distribution defined only over ρ_1, ρ_2 and α_1, α_2.*

An important property of the Gamma distribution is given below. It behaves similar to the factorial function.

Property 26.3.1

$$\Gamma(z + 1) = z\Gamma(z) \qquad \textit{where,} \quad z \in \mathbb{C} \tag{26.7}$$

Property 26.3.2 *Collapsibility property aka Aggregation*

If $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K) \sim Dir(\alpha_1, \ldots, \alpha_K)$, then if the random variables at the i-th and j-th index are dropped and replaced with their sum, the new distribution is also Dirichlet with:

$$\boldsymbol{\rho}' = (\rho_1, \ldots, \rho_i + \rho_j, \ldots, \rho_K) \sim Dir(\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_K) \tag{26.8}$$

It can be useful in deriving the marginal distribution of ρ_i from the joint above.

26.3.2 Understanding the influence of Dirichlet parameters

Let us consider the following extension of the collapsibility property where we fix a ρ_i and sum over all others:

$$\left(\rho_i, \sum_{j\neq i} \rho_j\right) \sim \text{Dir}\left(\alpha_i, \sum_{j\neq i} \alpha_j\right)$$

Recall that if we have a collection of random variables following a Dirichlet distribution, then each random variable follows a Beta distribution i.e. $\rho_i \sim \text{Beta}(\alpha_i, \alpha_0 - \alpha_i)$, where, $\alpha_0 = \sum_i \alpha_i$. Likewise, we can also compute the mean, variance and the covariance as:

$$\mathbb{E}[\rho_i] = \frac{\alpha_i}{\alpha_i + \alpha_0 - \alpha_i} = \frac{\alpha_i}{\alpha_0} \tag{26.9}$$

$$\text{Var}(\rho_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \tag{26.10}$$

$$\text{Cov}(\rho_i, \rho_j) = \frac{-\alpha_i\alpha_j}{\alpha_0^2(\alpha_0 + 1)} \tag{26.11}$$

Remark 26.3.3 *Note that in equation 26.10 if α_i is too small or if $\alpha_i \approx \alpha_0$ the variance will be small (i.e. ≈ 0). Similarly, considering the scale of the parameters, if all the parameters are large, the denominator becomes much larger and hence the variance will be small.*

Remark 26.3.4 *The covariance in equation 26.11 is negative to account for the fact that the random variables lie in a probability simplex (summing up to 1), so, intuitively, if one of the variables gets large then the others have to be small for this to be true.*

Figure 26.1 shows an example of how different values of Dirichlet parameters determine the spread of the probability density function (pdf) on the 2-simplex. In the top-left image, $\mathbb{E}[\rho_i] = 0.33$ and $\alpha_0 = 4.5$ showing concentric shapes for each random variable. In the top-right image, $\alpha_0 = 15$ which shows that the variance is small hence the concentric shapes are concentrated in the middle of the simplex. For the bottom-right image, $\mathbb{E}[\rho_3] = \frac{8}{14} = 0.57$ which explains the skewness towards the random variable ρ_3.

26.3.3 Dirichlet Simulation

The goal here is to outline a mechanism for sampling from a Dirichlet distribution. We have:

$$\rho_1 \sim \text{Beta}(\alpha_1, \sum_{k=1}^{K} \alpha_k - \alpha_1) \quad \perp \quad \frac{(\rho_2, \ldots, \rho_K)}{1 - \rho_1} \sim \text{Dir}(\alpha_2, \ldots, \alpha_K) \tag{26.12}$$

$$\nu_2 \sim \text{Beta}(\alpha_2, \sum_{k=2}^{K} \alpha_k - \alpha_2) \qquad \rho_2 = (1 - \nu_1)\nu_2 \tag{26.13}$$

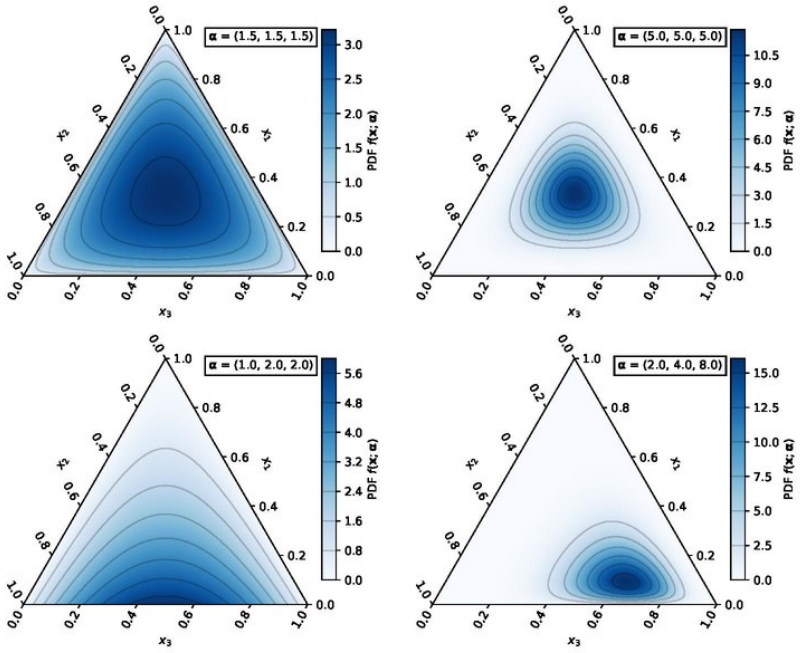

Figure 26.1: Probability density functions for the Dirichlet distribution on the 2-simplex. Source: Dirichlet distribution wikipedia.

$$\vdots \qquad\qquad \vdots$$

$$\nu_l \sim \text{Beta}(\alpha_l, \sum_{k=l}^{K} \alpha_k - \alpha_l) \qquad\qquad \rho_l = \left(\prod_{k=1}^{l-1}(1-\nu_k)\right)\nu_l \tag{26.14}$$

$$\rho_K = 1 - \sum_{k=1}^{K-1} \rho_k \tag{26.15}$$

Remark 26.3.5 *Consider equation 26.12 where ρ_1 is sampled from a Beta distribution. If the distribution is re-normalized excluding the most recent sample (ρ_1 in this case), we again obtain a Dirichlet distribution without the corresponding parameter α_1 for the random variable ρ_1. In other words, the re-normalized distribution, once ρ_1 is observed, only depends on the values of the remaining random variables.*

Remark 26.3.6 *Considering equation 26.13, note that ν_2 denotes the sample from the re-normalized distribution (without ρ_1). Therefore, ρ_2 is obtained by multiplying ν_2 with the remaining portion denoted by $(1-\nu_1)$.*

Therefore, we have the following outline where we sample from a Beta distribution, re-normalize it, sample another variable from the updated distribution, re-normalize it, and so

on until we reach $K-1$ random variables. The K-th can then be sampled by subtracting all the previously sampled variables from 1 (equation 26.15).

Note: This is also analogous to the stick-breaking mechanism where, with each sampled variable we chop off a certain portion of the stick and continue sampling from the remaining portion.

26.3.4 How do we choose K?

We don't! Instead, we directly consider an infinite-dimensional mixture model. The goal here is to extend our knowledge of sampling from a finite Dirichlet to an infinite setting when we don't know K to begin with.

The solution is to keep continuing the "stick-breaking" process till infinity i.e.

$$\begin{aligned}
\nu_1 &\sim \text{Beta}(\alpha_1, \beta_1) & \rho_1 &= \nu_1 \\
\nu_2 &\sim \text{Beta}(\alpha_2, \beta_2) & \rho_2 &= (1-\nu_1)\nu_2 \\
&\vdots & &\vdots \\
\nu_l &\sim \text{Beta}(\alpha_l, \beta_l) & \rho_l &= \left(\prod_{k=1}^{l-1}(1-\nu_k)\right)\nu_l \\
&\vdots & &\vdots
\end{aligned}$$

However, the values for α_l and β_l must be chosen carefully for proper normalization as they dictate whether the portion of the stick chopped off is too small or too large.

Towards that end, a method proposed by Griffiths, Engen and McCloskey (GEM) is to use the following strategy:

$$\nu_l = \text{Beta}(1, \alpha) \qquad \rho_1, \rho_2, \ldots \sim \text{GEM}(\alpha)$$

Remark 26.3.7 *Using the GEM strategy, we can observe that* $\mathbb{E}[\nu_l] = \frac{1}{1+\alpha}$, *where* $\mathbb{E}[\nu_l] \approx 1$ *if* $\alpha << 1$ *and* $\mathbb{E}[\nu_l] \approx 0$ *if* $\alpha >> 1$.

26.3.5 Random Measures

We have an infinite collection of proportions $\rho_1, \rho_2, \ldots \sim \text{GEM}(\alpha)$ and a distribution over the parameters of the distribution $\phi_1, \phi_2, \ldots \overset{\text{iid}}{\sim} H$ (e.g. $\mathcal{N}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$), which comprise two sources of randomness. Recall the generative modelling setting in the beginning of section 26.3 where we had the parameters of the distribution specified by a fixed mean depending on the class and a shared covariance matrix. So, we have $S \in \mathbb{R}^d \times \{\boldsymbol{\Sigma}\}$, with each point in

this space being a vector that is parametrized by the mean (for each class) and the shared covariance matrix. We then define a *random measure* over the space S:

$$G = \sum_{k=1}^{\infty} \rho_k \delta_{\phi_k} \tag{26.16}$$

Its interpretation is as follows: we sample infinitely many proportions ρ's and the parameters of the distribution ϕ's according to the distribution H (which is also a distribution on the space S) and define a random measure such that for each location ϕ_k in the space S, a Dirac distribution is assigned with a height (given by the corresponding proportion ρ_k) that is associated with the location ϕ_k.

The sampling is then done as follows:

$$z_n \mid G \sim G \tag{26.17}$$

$$\boldsymbol{x}_n \mid z_n \sim F_{z_n} \tag{26.18}$$

Equation 26.17 tells that once we have defined the random measure G, we sample the class assignments from G and note that realizations of z_n are nothing but ϕ_k's parameterized by $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$. Given the class assignments z_n, we can then obtain the observations by sampling from F, which depends on z_n by $\mathcal{N}(\boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma})$.

Why is G a random measure?

First, note that G is also a distribution on S. Consider a subset of the space S, say A, and apply the random measure G on the subset A denoting it as $G(A) \in [0, 1]$. Now, we see that for any subset in S, the result of applying this measure on that subset will be random because G itself is a random measure with ρ's and ϕ's being its sources of randomness. Since $G(A)$ is random, we can analyze it further by calculating its expectation $\mathbb{E}_G[G(A)]$.

$$\begin{aligned}
\mathbb{E}_G[G(A)] &= \mathbb{E}_G\left[\sum_k \rho_k \delta_{\phi_k}(A)\right] \\
&= \mathbb{E}_\rho\left[\sum_k \rho_k \mathbb{E}_{\phi_k}[\delta_{\phi_k}(A)]\right] \\
&= \mathbb{E}_\rho\left[\sum_k \rho_k H(A)\right] \quad (\because \delta_{\phi_k}(A) = 1 \text{ if } \phi_k \text{ lies in } A\ )
\end{aligned}$$

This is because the $\mathbb{E}_{\phi_k}[\delta_{\phi_k}(A)]$ is the probability that ϕ_k belongs to A. But, ϕ's are defined according to H, therefore, the probability *measure* w.r.t which it is obtained is H.

$$= H(A) \quad (\because H(A) \text{ is independent of } \rho)$$

Remark 26.3.8 *This result tells us that whatever random measures G we obtain by creating subsets on the space S, their expectation always results in the probability measure H being applied to the subsets on the space S. In other words, the probability that a random measure assigns to a subset in expectation is same as the probability that H assigns to the subset.*

26.3.6 Definition of a Dirichlet Process

Given a measurable space $(S, \mathcal{M})$, a base probability distribution H and $\alpha > 0$, the probability distribution of a (random) measure G is a Dirichlet process if for any finite partition $\{A_1, \ldots, A_r\}$ of S, the vector $[G(A_1), \ldots, G(A_r)] \sim \text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r))$. As for every stochastic process, we have:

- Random variables - Dirichlet (every component of random vector arising from a finite partition of the space S will follow a Dirichlet distribution)
- Index set T - $\mathcal{M}$ (the "measureable" subsets of the space)

Exercise to the Reader: Explore the behavior of a Dirichlet process $\text{DP}(\alpha, H)$ when: (i) $\alpha \to 0$? and (ii) $\alpha \to \infty$? This should also lead you to understand why α is known as the concentration parameter.

26.3.7 DP Summary

Problem: $\rho_1, \rho_2, \ldots$ is an infinite-dimensional probability vector. But, the Dirichlet distribution is defined for **finite** dimensional spaces.

Definiton: A DP is a distribution over probability measures such that for **every finite partition**, the probabilities of the partition elements follow a joint Dirichlet distribution.

$$[G(A_1), \ldots, G(A_r)] \sim \text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r))$$