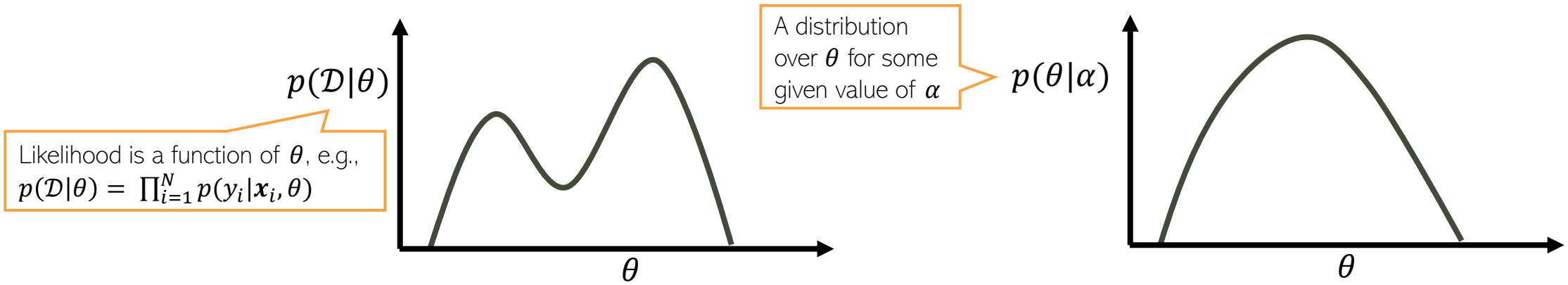
Probabilistic ML: Some Basic Ideas

CS772A: Probabilistic Machine Learning

Piyush Rai

Probabilistic ML Modeling: The Basic Ingredients

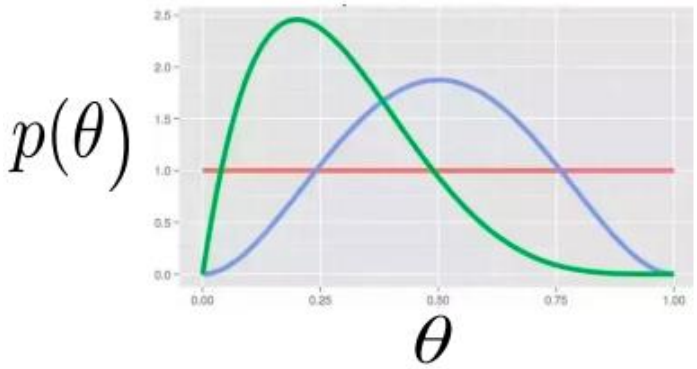
- Likelihood model $p(\mathcal{D}|\theta)$ for data $\mathcal{D}$; prior distribution $p(\theta|\alpha)$ over parameters θ



- Likelihood defined in terms of distribution(s) we assume data is generated from
  - It's like a measure of "fit" between observed data and each possible value of parameters
  - Its negative is like the "loss function" (high likelihood value = low loss; and vice-versa)
- Prior specifies our prior knowledge about θ before we have seen the data
  - It also acts as a regularizer for θ (will see the reason formally later)
- Note: The prior itself depends on other parameters α (also unknown)
  - These are sometimes called "hyperparameters" (can set by hand or estimate from data)

The Prior: Where does it come from?

- The prior $p(\theta|\alpha)$ plays an important role in probabilistic/Bayesian modeling
  - Reflects our prior beliefs about possible parameter values before seeing the data



  - Can be "subjective" or "objective" (also a topic of debate, which we won't get into)
  - Subjective: Prior (our beliefs) derived from past experiments
  - Objective: Prior represents "neutral knowledge" (e.g.. uniform, vague prior)
  - Can also be seen as a regularizer (connection with non-probabilistic view)

Parameter Estimation

- The parameters θ are unknown and need to be estimated from training data $\mathcal{D}$
- When estimating θ, we may take one of the following approaches

Approach 1

- θ has an unknown with fixed value
- Estimate the single best estimate of θ by optimizing a loss function

$$\hat{\theta} = \text{argmin}_{\theta}\, \mathcal{L}(\mathcal{D}; \theta)$$

Approach 2

- Treat θ as a random variable
- Estimate θ by computing its distribution conditioned on $\mathcal{D}$

Posterior distribution $p(\theta|\mathcal{D})$

- Approach 2 also gives uncertainty about our estimate of θ; Approach 1 doesn't
  - But possible to estimate uncertainty in θ even with Approach 1 (e.g., using ensembles)
- Approach 1 is also a simplified/special case/approximation of Approach 2
- Can also take a hybrid (Approach 2 for some parameters; Approach 1 for others)

The Posterior Distribution

- The posterior distribution is computed using Bayes rule (Bayesian inference)

$$p(\theta|\mathcal{D},\alpha) = \frac{p(\mathcal{D},\theta|\alpha)}{p(\mathcal{D}|\alpha)} = \frac{p(\mathcal{D}|\theta,\alpha)p(\theta|\alpha)}{\int p(\mathcal{D}|\theta,\alpha)p(\theta|\alpha)d\theta}$$

$$= \frac{p(\mathcal{D}|\theta)p(\theta|\alpha)}{\int p(\mathcal{D}|\theta)p(\theta|\alpha)d\theta} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

Assuming α is known so the posterior is conditioned on α as well

Given θ, the data is conditionally independent of the prior's hyperparameters α so $p(\mathcal{D}|\theta,\alpha) = p(\mathcal{D}|\theta)$

- Marginal likelihood is an important quantity

$$p(\mathcal{D}|\alpha) = \int p(\mathcal{D}|\theta)p(\theta|\alpha)d\theta = \mathbb{E}_{p(\theta|\alpha)}p(\mathcal{D}|\theta)]$$

The average" likelihood (average taken w.r.t. all values of θ from the prior distribution)

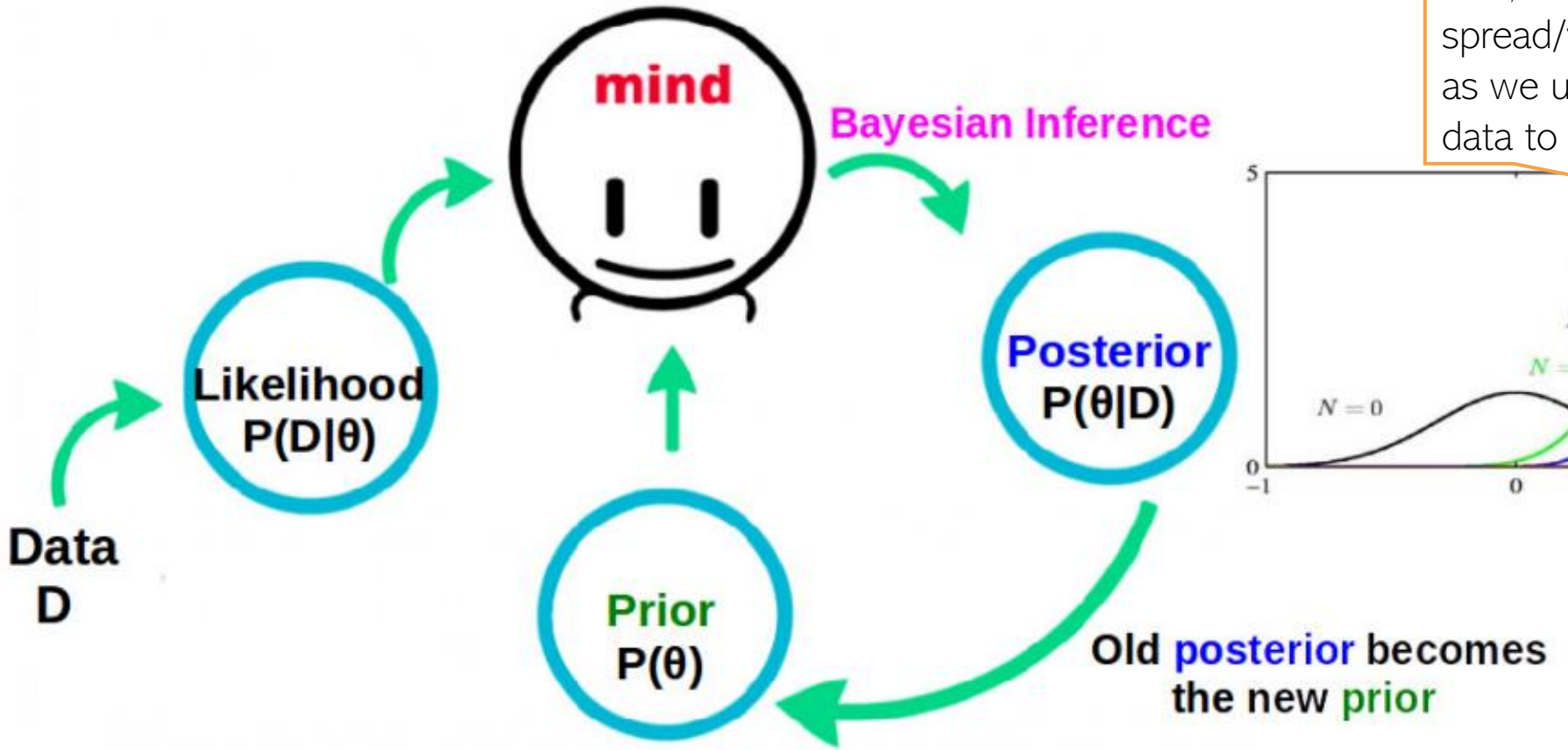
Hard to compute in general (that's why posterior is difficult to compute in general) but be computed exactly in some cases

We can use it also to find the best value of hyperparameters α

For example, $\hat{\alpha} = \text{argmax}_\alpha \log p(\mathcal{D}|\alpha)$

"Online" Nature of Bayesian Inference Updates

- Bayesian inference can naturally be done in an online fashion



Point Estimation

- Recall that the posterior is

$$p(\theta|\mathcal{D},\alpha) = \frac{p(\mathcal{D}|\theta)p(\theta|\alpha)}{p(\mathcal{D}|\alpha)}$$

Intractable mainly because the marginal likelihood (the denominator on the RHS is intractable in general)

Intractable to compute except for some very simple models or if the likelihood and prior are conjugate (discussed later) to each other

However, point estimation throws away all the uncertainty information about θ

- If posterior is intractable, can use MLE/MAP to get point estimates
  - Maximum likelihood (ML) estimation: Find θ for which likelihood is highest

Meaning the observed data has the largest probability for this value of θ

$$\hat{\theta}_{ML} = \operatorname{argmax}_\theta \log p(\mathcal{D}|\theta) = \operatorname{argmin}_\theta -\log p(\mathcal{D}|\theta) = \operatorname{argmin}_\theta NLL(\theta)$$

Negative Log likelihood (equivalent to a loss function)

  - Maximum a posteriori (MAP) estimation: Find θ with largest posterior prob.

$$\hat{\theta}_{MAP} = \operatorname{argmax}_\theta \log p(\theta|\mathcal{D},\alpha) = \operatorname{argmax}_\theta [\log p(\mathcal{D}|\theta) + \log p(\theta|\alpha)]$$

$$= \operatorname{argmin}_\theta [NLL(\theta) - \log p(\theta|\alpha)]$$

Like MLE with info from prior added

Akin to a regularizer added to the loss

The regularizer hyperparameter is part of prior

The Predictive Distribution

- Predictive distribution is the distribution of test data $\mathcal{D}_*$ given training data $\mathcal{D}$
- In the general form, we can write it as

$p(\mathcal{D}_*|\mathcal{D})$ is known as posterior predictive distribution (PPD)

$$p(\mathcal{D}_*|\mathcal{D}) = \int p(\mathcal{D}_*, \theta|\mathcal{D})\, d\theta$$

In the posterior, not showing prior's hyperparameters α for brevity of notation

$$= \int p(\mathcal{D}_*|\theta, \mathcal{D}) p(\theta|\mathcal{D})\, d\theta$$

PPD is more robust (less chance of overfitting) than plug-in prediction because we aren't relying on a single best estimate

Assuming observations are i.i.d. given θ

$$= \int p(\mathcal{D}_*|\theta) p(\theta|\mathcal{D})\, d\theta$$

The "averaged" prediction using all possible θ values with each prediction weighted by how important θ is as per the posterior distribution

This expectation may not be computable exactly and may itself require approximations (will see later)

An expectation over the posterior distribution (averaging over the posterior)

$$= \mathbb{E}_{p(\theta|\mathcal{D})}[p(\mathcal{D}_*|\theta)]$$

- If we only have point estimate of θ (say $\hat{\theta}$ obtained from MLE/MAP) then

This approximation of PPD is called "plug-in" predictive distribution

$$p(\mathcal{D}_*|\mathcal{D}) \approx p(\mathcal{D}_*|\hat{\theta})$$

Because now the posterior is just a point mass at $\hat{\theta}$

A "Shortcut": PPD using Marginal Likelihood

- PPD, by definition, is obtained by the following marginalization

$$p(\mathcal{D}_*|\mathcal{D}) = \int p(\mathcal{D}_*|\theta)p(\theta|\mathcal{D})\, d\theta$$

- Can also compute PPD without computing the posterior! Some ways:

1. Using a ratio of marginal likelihoods as follows

Follows simply from Bayes rule

$$p(a|b) = \frac{p(a,b)}{p(b)}$$

$$p(\mathcal{D}_*|\mathcal{D}) = \frac{p(\mathcal{D}_*, \mathcal{D})}{p(\mathcal{D})}$$

Joint marginal likelihood for training and test data

Marginal likelihood for training data

2. If $p(\mathcal{D}_*|\mathcal{D})$ can be obtained easily from the joint $p(\mathcal{D}_*, \mathcal{D})$
   - Note that the PPD $p(\mathcal{D}_*|\mathcal{D})$ is also a conditional distribution
   - For some distributions (e.g., Gaussian), conditionals can be easily derived from joint

Will see this being used we we study Gaussian Process (GP)

Bernoulli Observation Model

Estimating a Coin's Bias

- Consider a sequence of N coin toss outcomes (observations)
- Each observation y_n is a binary random variable. Head: $y_n = 1$, Tail: $y_n = 0$
- Each y_n is assumed generated by a **Bernoulli distribution** with param $\theta \in (0,1)$

Probability of a head

Likelihood or observation model

$$p(y_n|\theta) = \text{Bernoulli}(y_n|\theta) = \theta^{y_n}(1-\theta)^{1-y_n}$$

- Here θ the unknown param (probability of head). Let's do MLE
- Log-likelihood: $\sum_{n=1}^{N} \log p(y_n|\theta) = \sum_{n=1}^{N} [y_n \log \theta + (1-y_n)\log(1-\theta)]$

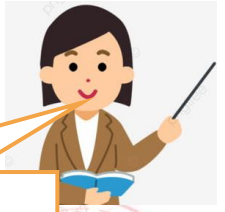
assuming i.i.d. data

- Maximizing log-lik, or minimizing neg. log-lik (NLL) w.r.t. θ gives

$$\theta_{MLE} = \frac{\sum_{n=1}^{N} y_n}{N}$$

I tossed a coin 5 times – gave 1 head and 4 tails. Does it means $\theta = 0.2$?? The MLE approach says so. What is I see 0 head and 5 tails. Does it mean $\theta = 0$?

Thus MLE solution is simply the fraction of heads! ☺ Makes intuitive sense!



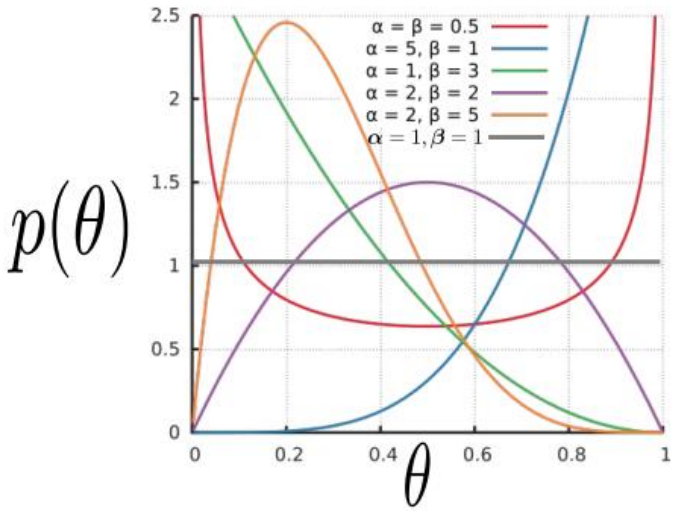
Indeed, with a small number of training observations, MLE may overfit and may not be reliable. An alternative is MAP estimation which can incorporate a prior distribution over θ

Estimating a Coin's Bias

- Let's do MAP estimation for the bias of the coin
- Each likelihood term is Bernoulli

$$p(y_n|\theta) = \text{Bernoulli}(y_n|\theta) = \theta^{y_n}(1-\theta)^{1-y_n}$$

- Also need a prior since we want to do MAP estimation
- Since $\theta \in (0,1)$, a reasonable choice of prior for θ would be Beta distribution



$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

The gamma function

Using $\alpha = 1$ and $\beta = 1$ will make the Beta prior a uniform prior

α and β (both non-negative reals) are the two hyperparameters of this Beta prior

Can set these based on intuition, cross-validation, or even learn them

Estimating a Coin's Bias

- The log posterior for the coin-toss model is log-lik + log-prior

$$LP(\theta) = \sum_{n=1}^{N} \log p(y_n|\theta) + \log p(\theta|\alpha,\beta)$$

- Plugging in the expressions for Bernoulli and Beta and ignoring any terms that don't depend on θ, the log posterior simplifies to

$$LP(\theta) = \sum_{n=1}^{N} [y_n \log \theta + (1-y_n)\log(1-\theta)] + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)$$

- Maximizing the above log post. (or min. of its negative) w.r.t. θ gives

$$\theta_{MAP} = \frac{\sum_{n=1}^{N} y_n + \alpha - 1}{N + \alpha + \beta - 2}$$

Using $\alpha = 1$ and $\beta = 1$ gives us the same solution as MLE

Recall that $\alpha = 1$ and $\beta = 1$ for Beta distribution is in fact equivalent to a uniform prior (hence making MAP equivalent to MLE)

Prior's hyperparameters have an interesting interpretation. Can think of $\alpha - 1$ and $\beta - 1$ as the number of heads and tails, respectively, before starting the coin-toss experiment (akin to "pseudo-observations")

Such interpretations of prior's hyperparameters as being "pseudo-observations" exist for various other prior distributions as well (in particular, distributions belonging to "exponential family" of distributions

The Posterior Distribution

- Let's do fully Bayesian inference and compute the posterior distribution
- Bernoulli likelihood: $p(y_n|\theta) = \text{Bernoulli}(y_n|\theta) = \theta^{y_n}(1-\theta)^{1-y_n}$
- Beta prior: $p(\theta) = \text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The posterior can be computed as

Number of heads (N_1)

Number of tails (N_0)

$$\theta^{\sum_{n=1}^N y_n}(1-\theta)^{N-\sum_{n=1}^N y_n}$$

Hyperparams α, β not shown for brevity

$$p(\theta|\boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{p(\boldsymbol{y})} = \frac{p(\theta)\prod_{n=1}^N p(y_n|\theta)}{p(\boldsymbol{y})} = \frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{n=1}^N \theta^{y_n}(1-\theta)^{1-y_n}}{\int \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{n=1}^N \theta^{y_n}(1-\theta)^{1-y_n} d\theta}$$

- Here, even without computing the denominator (marg lik), we can identify the posterior
  - It is Beta distribution since $p(\theta|\boldsymbol{y}) \propto \theta^{\alpha+N_1-1}(1-\theta)^{\beta+N_0-1}$
  - Thus $p(\theta|\boldsymbol{y}) = \text{Beta}(\theta|\alpha+N_1, \beta+N_0)$

Hint: Use the fact that the posterior must integrate to 1
$\int p(\theta|\boldsymbol{y})d\theta = 1$

Exercise: Show that the normalization constant equals
$$\frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+\sum_{n=1}^N y_n)\Gamma(\beta+N-\sum_{n=1}^N y_n)}$$

- Here, finding the posterior boiled down to simply "multiply, add stuff, and identify"
- Here, posterior has the same form as prior (both Beta): property of conjugate priors

Conjugacy and Conjugate Priors

- Many pairs of distributions are conjugate to each other
  - Bernoulli (likelihood) + Beta (prior) ⇒ Beta posterior
  - Binomial (likelihood) + Beta (prior) ⇒ Beta posterior
  - Multinomial (likelihood) + Dirichlet (prior) ⇒ Dirichlet posterior
  - Poisson (likelihood) + Gamma (prior) ⇒ Gamma posterior
  - Gaussian (likelihood) + Gaussian (prior) ⇒ Gaussian posterior
  - and many other such pairs ..

Not true in general, but in some cases (e.g., the variance of the Gaussian likelihood is fixed)

- Tip: If two distr are conjugate to each other, their functional forms are similar
  - Example: Bernoulli and Beta have the forms

$$\text{Bernoulli}(y|\theta) = \theta^{y}\,(1-\theta)^{1-y}$$

$$\text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

This is why, when we multiply them while computing the posterior, the exponents get added and we get the same form for the posterior as the prior but with just updated hyperparameter. Also, we can identify the posterior and its hyperparameters simply by inspection

- More on conjugate priors when we look at exponential family distributions

Predictive Distribution

- Suppose we want to compute the prob that the next outcome $\boldsymbol{y_{N+1}}$ will be head (=1)
- The posterior predictive distribution (averaging over all $\boldsymbol{\theta}$'s weighted by their respective posterior probabilities)

$$p(y_{N+1}=1|\boldsymbol{y}) = \int_0^1 p(y_{N+1}=1,\theta|\boldsymbol{y})\,d\theta = \int_0^1 p(y_{N+1}=1|\theta)p(\theta|\boldsymbol{y})\,d\theta$$

$$= \int_0^1 \theta \times p(\theta|\boldsymbol{y})\,d\theta$$

$$= \mathbb{E}_{p(\theta|\boldsymbol{y})}[\theta]$$

$$= \frac{\alpha + N_1}{\alpha + \beta + N}$$

Expectation of θ w.r.t. the Beta posterior distribution $p(\theta|\boldsymbol{y}) = \text{Beta}(\theta|\alpha + N_1, \beta + N_0)$

- Therefore the PPD will be

$$p(y_{N+1}|\boldsymbol{y}) = \text{Bernoulli}(y_{N+1}|\mathbb{E}_{p(\theta|\boldsymbol{y})}[\theta])$$

For models where likelihood and prior are conjugate to each other, the PPD can be computed easily in closed form (more on this when we talk about exponential family distributions)

- The plug-in predictive distribution using a point estimate $\hat{\theta}$ (e.g., using MLE/MAP)

$$p(y_{N+1}=1|\boldsymbol{y}) \approx p(y_{N+1}=1|\hat{\theta}) = \hat{\theta} \quad \Rightarrow \quad p(y_{N+1}|\boldsymbol{y}) = \text{Bernoulli}(y_{N+1}|\hat{\theta})$$

Multinoulli Observation Model

The Posterior Distribution

MLE/MAP left as an exercise

- Assume N discrete obs $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$ with each $y_n \in \{1,2,\ldots,K\}$, e.g.,
  - y_n represents the outcome of a dice roll with K faces
  - y_n represents the class label of the n^{th} example in a classification problem (total K classes)
  - y_n represents the identity of the n^{th} word in a sequence of words
- Assume likelihood to be multinoulli with unknown params $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_K]$

These sum to 1

$$p(y_n|\pi) = \textbf{multinoulli}(y_n|\pi) = \prod_{k=1}^{K} \pi_k^{\mathbb{I}[y_n=k]}$$

Generalization of Bernoulli to $K > 2$ discrete outcomes

- $\boldsymbol{\pi}$ is a vector of probabilities ("probability vector"), e.g.,
  - Biases of the K sides of the dice
  - Prior class probabilities in multi-class classification ($p(y_n = k) = \pi_k$)
  - Probabilities of observing each word of the K words in a vocabulary
- Assume a conjugate prior (Dirichlet) on $\boldsymbol{\pi}$ with hyperparams $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_K]$

Called the concentration parameter of the Dirichlet (assumed known for now)

Large values of $\boldsymbol{\alpha}$ will give a Dirichlet peaked around its mean (next slides illustrates this)

Each $\alpha_k \geq 0$

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

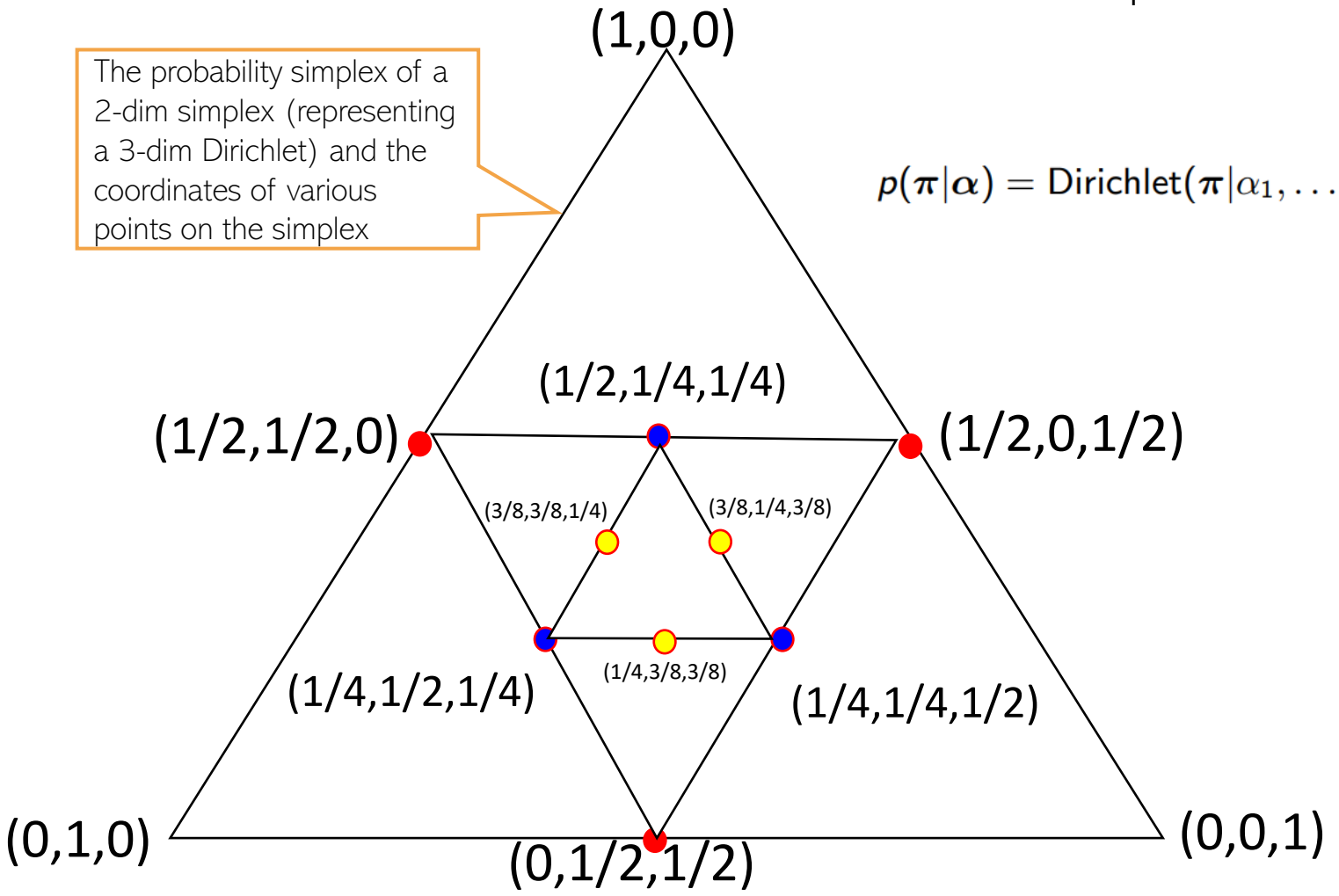
Generalization of Beta to K-dimensional probability vectors

Brief Detour: Dirichlet Distribution

Basically, probability vectors

- An important distribution. Models non-neg. vectors $\boldsymbol{\pi}$ that also sum to one
- A random draw from K-dim Dirich. will be a point under $(K\text{-}1)$-dim probability simplex

The probability simplex of a 2-dim simplex (representing a 3-dim Dirichlet) and the coordinates of various points on the simplex



$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1,\ldots,\alpha_K) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \pi_k^{\alpha_k - 1} = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^K \pi_k^{\alpha_k - 1}$$

$$\text{Mean} = \left[\frac{\alpha_1}{\sum_{k=1}^K \alpha_k}, \ldots, \frac{\alpha_K}{\sum_{k=1}^K \alpha_k}\right]$$

$$\text{Mode} = \left[\frac{\alpha_1 - 1}{\sum_{k=1}^K \alpha_k - K}, \ldots, \frac{\alpha_K - 1}{\sum_{k=1}^K \alpha_k - K}\right] (\alpha_k > 1)$$

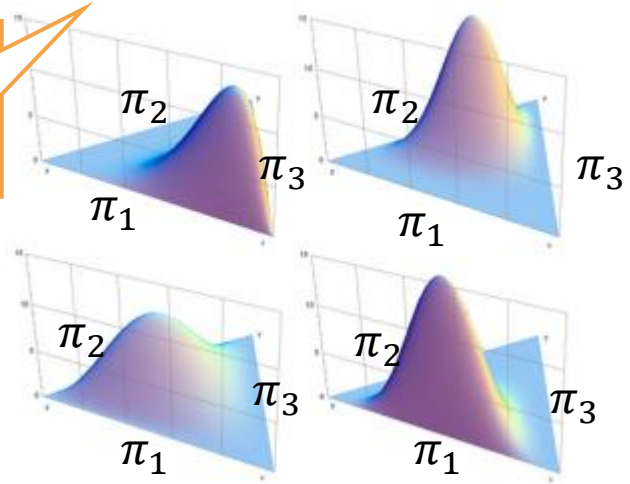
$$\text{var}(\pi_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \qquad \alpha_0 = \sum_{k=1}^K \alpha_k$$

Brief Detour: Dirichlet Distribution

- A visualization of Dirichlet distribution for different values of concentration param

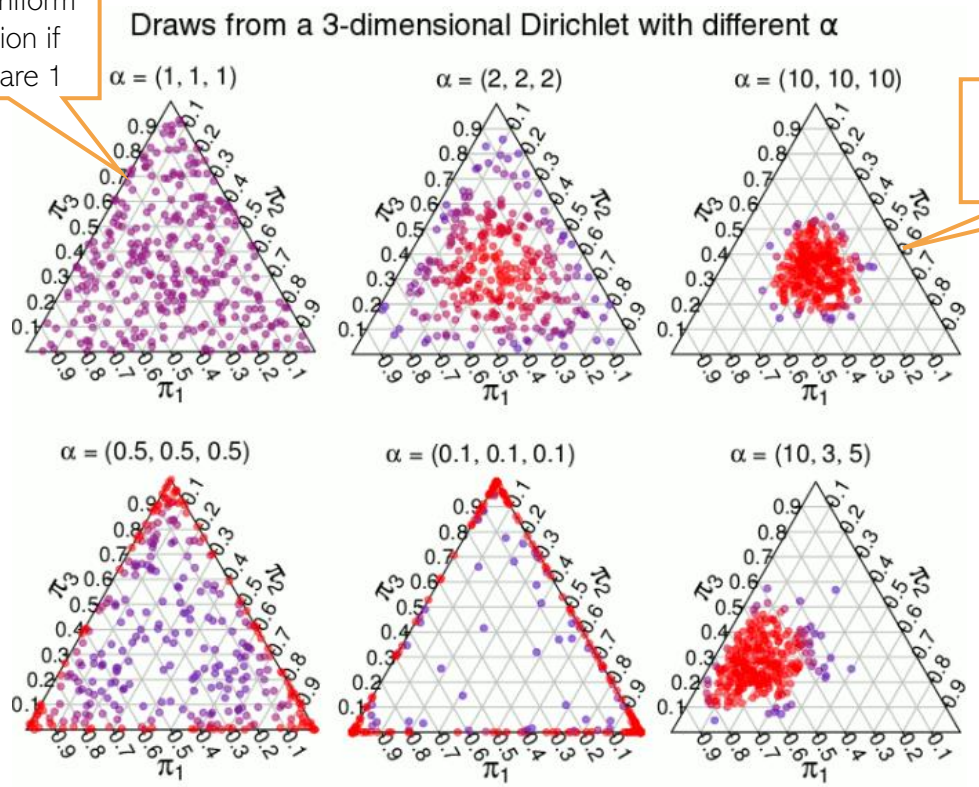
Visualizations of PDFs of some 3-dim Dirichlet distributions (each generated using a different conc. Param vector $\boldsymbol{\alpha}$)

$\boldsymbol{\alpha}$ controls the shape of the Dirichlet (just like Beta distribution's hyperparameters)



Like a uniform distribution if all α_k's are 1

All α_k's large results in peak around the center of the simplex



- Interesting fact: Can generate a K-dim Dirichlet random variable by independently generating K gamma random variables and normalizing them to sum to 1

The Posterior Distribution

- Posterior $p(\boldsymbol{\pi}|\boldsymbol{y})$ is easy to compute due to conjugacy b/w multinoulli and Dir.

Likelihood

Prior

$$p(\boldsymbol{\pi}|\boldsymbol{y},\boldsymbol{\alpha}) = \frac{p(\boldsymbol{\pi},\boldsymbol{y}|\boldsymbol{\alpha})}{p(\boldsymbol{y}|\boldsymbol{\alpha})} = \frac{p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{y}|\boldsymbol{\pi},\boldsymbol{\alpha})}{p(\boldsymbol{y}|\boldsymbol{\alpha})} = \frac{p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{y}|\boldsymbol{\pi})}{p(\boldsymbol{y}|\boldsymbol{\alpha})}$$

Don't need to compute for this case because of conjugacy

Marg-lik = $\int p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{y}|\boldsymbol{\pi})\mathrm{d}\boldsymbol{\pi}$

- Assuming y_n's are i.i.d. given $\boldsymbol{\pi}$, $p(\boldsymbol{y}|\boldsymbol{\pi}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{\pi})$, and therefore

$$p(\boldsymbol{\pi}|\boldsymbol{y},\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \times \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{\mathbb{I}[y_n=k]} = \prod_{k=1}^{K} \pi_k^{\alpha_k + \sum_{n=1}^{N} \mathbb{I}[y_n=k] - 1}$$

- Even without computing marg-lik, $p(\boldsymbol{y}|\boldsymbol{\alpha})$, we can see that the posterior is Dirichlet
- Denoting $N_k = \sum_{n=1}^{N} \mathbb{I}[y_n = k]$, number of observations with with value k

$$p(\boldsymbol{\pi}|\boldsymbol{y},\boldsymbol{\alpha}) = \mathrm{Dirichlet}(\boldsymbol{\pi}|\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K)$$

Similar to number of heads and tails for the coin bias estimation problem

- Note: $N_1,, N_2 \ldots, N_K$ are the sufficient statistics for this estimation problem
  - We only need the suff-stats to estimate the parameters and values of individual observations aren't needed (another property from exponential family of distributions – more on this later)

The Predictive Distribution

- Finally, let's also look at the posterior predictive distribution for this model
- PPD is the prob distr of a new $y_* \in \{1,2,\ldots,K\}$, given training data $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$

Will be a multinoulli. Just need to estimate the probabilities of each of the K outcomes

$$p(y_*|\boldsymbol{y},\boldsymbol{\alpha}) = \int p(y_*|\boldsymbol{\pi})\boldsymbol{p}(\boldsymbol{\pi}|\boldsymbol{y},\boldsymbol{\alpha})\boldsymbol{d\pi}$$

- $p(y_*|\boldsymbol{\pi}) = \text{multinoulli}(y_*|\boldsymbol{\pi})$, $p(\boldsymbol{\pi}|\boldsymbol{y},\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K)$
- Can compute the posterior predictive probability for each of the K possible outcomes

$$p(y_* = k|\boldsymbol{y},\boldsymbol{\alpha}) = \int p(y_* = k|\boldsymbol{\pi})\boldsymbol{p}(\boldsymbol{\pi}|\boldsymbol{y},\boldsymbol{\alpha})\boldsymbol{d\pi}$$

$$= \int \pi_k \times \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K)d\pi$$

$$= \frac{\alpha_k + N_k}{\sum_{k=1}^{K} \alpha_k + N}$$ (Expectation of π_k w.r.t the Dirichlet posterior)

- Thus PPD is multinoulli with probability vector $\left\{\frac{\alpha_k + N_k}{\sum_{k=1}^{K} \alpha_k + N}\right\}_{k=1}^{K}$

Note how these probabilities have been "smoothened" due to the use of the prior + the averaging over the posterior

A similar effect was achieved in the Beta-Bernoulli model, too

- Plug-in predictive will also be multinoulli but with prob vector given by the point estimate of $\boldsymbol{\pi}$