Parameter Estimation in Probabilistic Models (Contd.)

CS772A: Probabilistic Machine Learning

Piyush Rai

Plan Today

- Two simple examples of parameter estimation in probabilistic models
  - Beta (prior) – Bernoulli (likelihood) observation model
  - Dirichlet (prior) – Multinomial (likelihood) observation model
- Conjugate priors
- "Reading" a posterior distribution

An Important Point: PPD using Marginal Likelihood

- PPD for a model $\boldsymbol{m}$, by definition, is obtained by the following marginalization

$$p(\boldsymbol{x}_*|\mathbf{X}, m) = \int p(\boldsymbol{x}_*|\theta, m)p(\theta|\mathbf{X}, m)\, d\theta$$

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

1. Using a ratio of marginal likelihoods as follows

Follows simply from Bayes rule $p(a|b) = \frac{p(a,b)}{p(b)}$

$$p(\boldsymbol{x}_*|\mathbf{X}, m) = \frac{p(\boldsymbol{x}_*, \mathbf{X}|m)}{p(\mathbf{X}|m)}$$

Joint marginal likelihood for training and test data

Marginal likelihood for training data

2. If $p(\boldsymbol{x}_*|\mathbf{X}, m)$ can be obtained easily from the joint distribution $p(\boldsymbol{x}_*, \mathbf{X}|m)$
   - Note that the PPD $p(\boldsymbol{x}_*|\mathbf{X}, m)$ 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)

Estimating a Beta-Bernoulli Model

Estimating a Coin's Bias: MLE

- 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). Want to estimate it using MLE

assuming i.i.d. data

- Log-likelihood: $\sum_{n=1}^{N} \log p(y_n|\theta) = \sum_{n=1}^{N} [y_n \log\theta + (1-y_n)\log(1-\theta)]$
- Maximizing log-lik, or minimizing neg. log-lik (NLL) w.r.t. θ gives

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

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

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

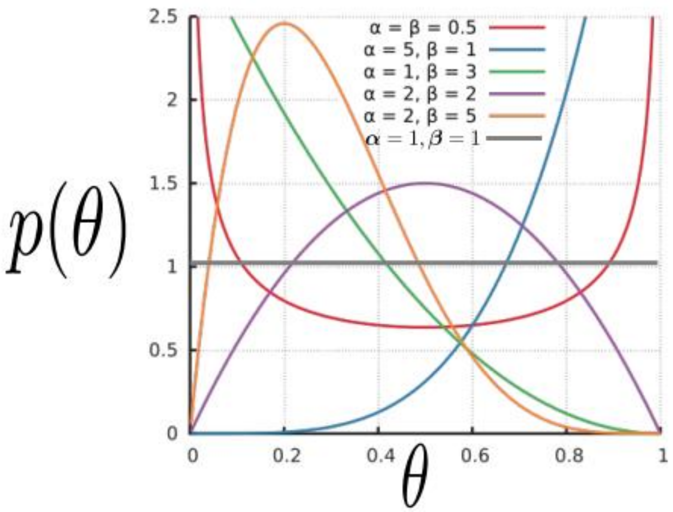
Indeed, with a small number of training observations, MLE may overfit and may not be reliable. An alternative is MAP estimation which can incorporate a prior distribution over θ

Estimating a Coin's Bias: MAP

- Let's again consider the coin-toss problem (estimating 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

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

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

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

Estimating a Coin's Bias: MAP

- 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 $\boldsymbol{\theta}$, 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. $\boldsymbol{\theta}$ gives

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

Using $\boldsymbol{\alpha} = \mathbf{1}$ and $\boldsymbol{\beta} = \mathbf{1}$ gives us the same solution as MLE

Recall that $\boldsymbol{\alpha} = \mathbf{1}$ and $\boldsymbol{\beta} = \mathbf{1}$ 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

Estimating a Coin's Bias: Fully Bayesian Inference

- In fully Bayesian inference, we 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

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

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

Number of heads (N_1)

Number of tails (N_0)

- 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 $\frac{\Gamma(\alpha+\sum_{n=1}^N x_n)\Gamma(\beta+N-\sum_{n=1}^N x_n)}{\Gamma(\alpha+\beta+N)}$

- 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

Making Predictions

- Suppose we want to compute the prob that the next outcome $\boldsymbol{x}_{N+1}$ will be head (=1)
- The plug-in predictive distribution using a point estimate $\hat{\boldsymbol{\theta}}$ (e.g., using MLE/MAP)

$$p(x_{N+1}=1|\mathbf{X}) \approx p(x_{N+1}=1|\hat{\theta}) = \hat{\theta} \quad \text{or equivalently} \quad p(x_{N+1}|\mathbf{X}) \approx \text{Bernoulli}(x_{N+1} \mid \hat{\theta})$$

- The posterior predictive distribution (averaging over all $\boldsymbol{\theta}$'s weighted by their respective posterior probabilities)

$$\begin{aligned} p(x_{N+1}=1|\mathbf{X}) &= \int_0^1 P(x_{N+1}=1|\theta)p(\theta|\mathbf{X})d\theta \\ &= \int_0^1 \theta \times \text{Beta}(\theta|\alpha+N_1, \beta+N_0)d\theta \\ &= \mathbb{E}[\theta|\mathbf{X}] \\ &= \frac{\alpha+N_1}{\alpha+\beta+N} \end{aligned}$$

Expectation of $\boldsymbol{\theta}$ w.r.t. the Beta posterior distribution

- Therefore the PPD is $p(x_{N+1}|\mathbf{X}) = \text{Bernoulli}(x_{N+1} \mid \mathbb{E}[\theta|\mathbf{X}])$

Estimating a Dirichlet-Multinoulli Model

Bayesian Inference for Multinoulli/Multinomial

- Assume N discrete obs $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$ with each $x_n \in \{1,2,\dots,K\}$, e.g.,
  - x_n represents the outcome of a dice roll with K faces
  - x_n represents the class label of the n^{th} example in a classification problem (total K classes)
  - x_n represents the identity of the n^{th} word in a sequence of words
- Assume likelihood to be multinoulli with unknown params $\boldsymbol{\pi} = [\pi_1, \pi_2, \dots, \pi_K]$

These sum to 1

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

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 $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_K]$

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

Large values of α will give a Dirichlet peaked around its mean (next slide illustrates this)

Each $\alpha_k \geq 0$

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

Generalization of Beta to K-dimensional probability vectors

Brief Detour: Dirichlet Distribution

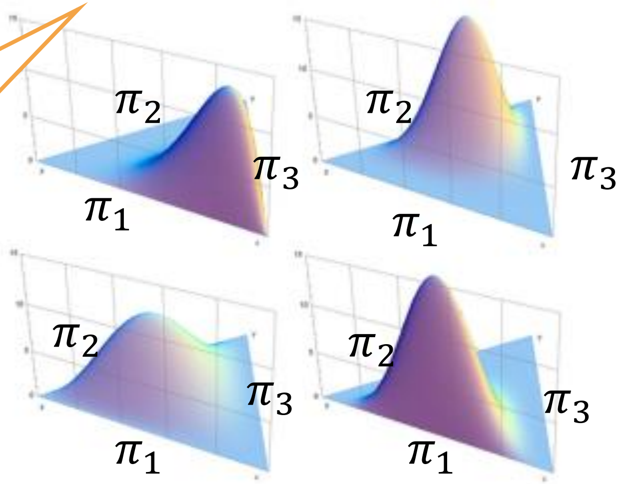
Basically, probability vectors

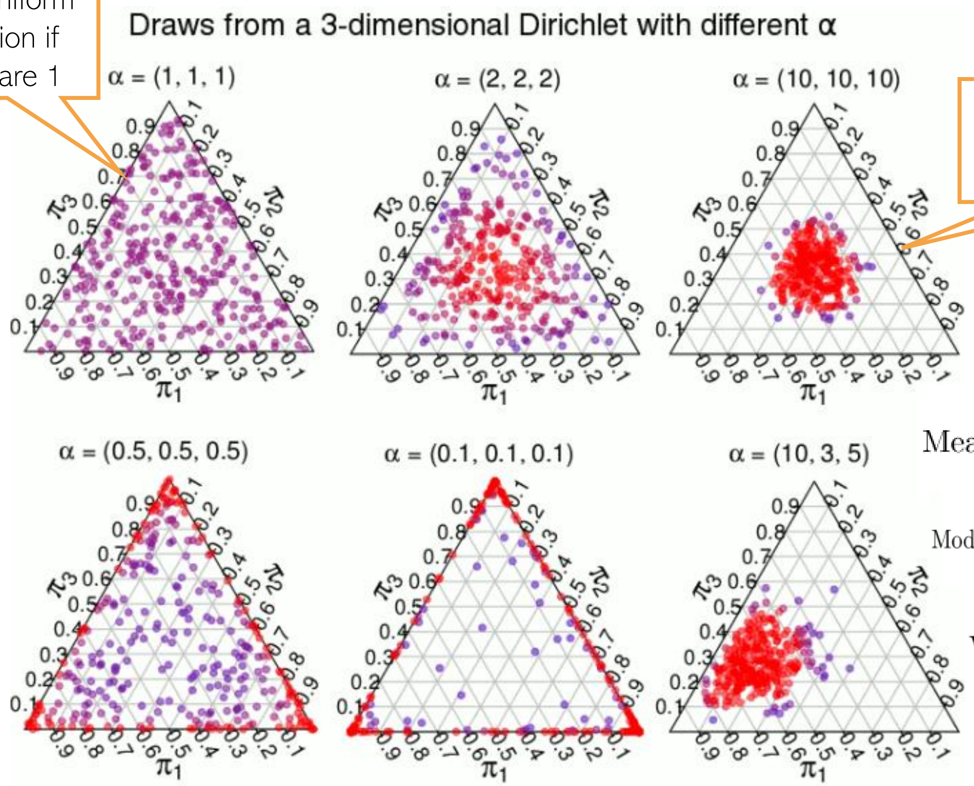
- An important distribution. Models non-neg. vectors $\boldsymbol{\pi}$ that also sum to one
- A random draw from K-dim Dirich. will be a point under (K-1)-dim probability simplex

Like a uniform distribution if all α_k's are 1

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)





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

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

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

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

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

Bayesian Inference for Multinoulli

Likelihood

Prior

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

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

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

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

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

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

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

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

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)

Bayesian Inference for Multinoulli

- Finally, let's also look at the posterior predictive distribution for this model
- PPD is the prob distr of a new $\boldsymbol{x}_* \in \{1,2,\ldots,K\}$, given training data $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$

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

$$p(\boldsymbol{x}_*|\mathbf{X},\boldsymbol{\alpha}) = \int p(\boldsymbol{x}_*|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{X},\boldsymbol{\alpha})d\boldsymbol{\pi}$$

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

$$p(\boldsymbol{x}_* = k|\mathbf{X},\boldsymbol{\alpha}) = \int p(\boldsymbol{x}_* = k|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{X},\boldsymbol{\alpha})d\boldsymbol{\pi}$$

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

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

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

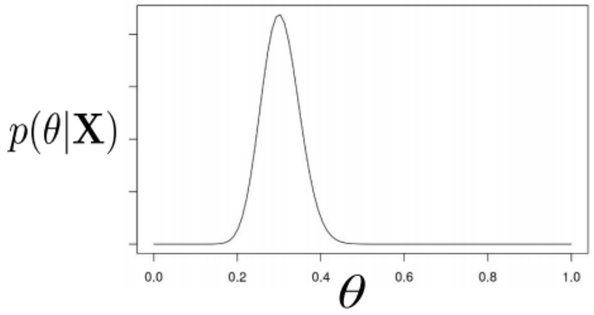
Note how these probabilities have been "smoothened" due to the use of the prior + the averaging over the posterior

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 $\boldsymbol{\pi}$

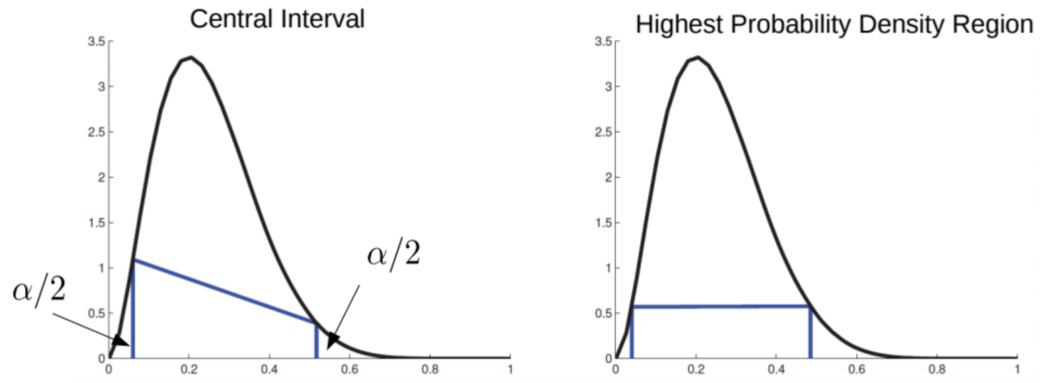
"Reading" the Posterior Distribution

- Posterior provides us a holistic view about $\boldsymbol{\theta}$ given observed data
- A simple unimodal posterior for a scalar parameter $\boldsymbol{\theta}$ might look something like

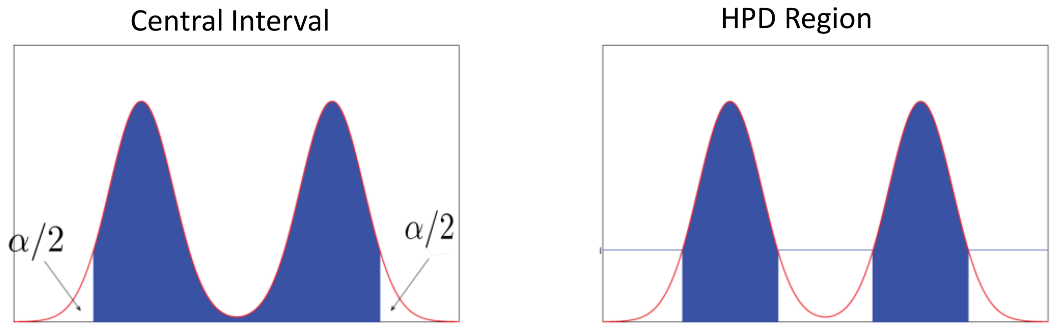


- Various types of estimates regarding $\boldsymbol{\theta}$ can be obtained from the posterior, e.g.,
  - Mode of the posterior (same as the MAP estimate)
  - Mean and median of the posterior
  - Variance/spread of the posterior (uncertainty in our estimate of the parameters)
  - Any quantile (say $0 < \alpha < 1$ quantile) of the posterior, e.g., $\boldsymbol{\theta}_*$ s.t. $p(\boldsymbol{\theta} \leq \boldsymbol{\theta}_*) = \alpha$
  - Various types of intervals/regions

"Reading" the Posterior Distribution



Also defined for multi-modal posteriors



- $100(1-\alpha)\%$ Credible Interval: Region in which $1-\alpha$ fraction of posterior's mass resides

$$\mathcal{C}_\alpha(\mathbf{X}) = (\ell, u) : p(\ell \leq \theta \leq u|\mathbf{X}) = 1-\alpha$$

Computing central interval or HPD usually requires inverting CDFs

- Credible Interval is not unique (there can be many $100(1-\alpha)\%$ intervals)
- Central Interval is a symmetrized version of Credible Interval ($\alpha/2$ mass on each tail)
- Another useful interval: The $(1-\alpha)$ Highest Probability Density (HPD) region

$$\mathcal{C}_\alpha(\mathbf{X}) = \{\theta : p(\theta|\mathbf{X}) \geq p^*\} \quad \text{s.t.} \quad 1-\alpha = \int_{\theta:p(\theta|\mathbf{X})\geq p^*} p(\theta|\mathbf{X})d\theta$$