Basics of Parameter Estimation in Probabilistic Models

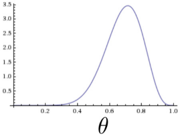
Piyush Rai

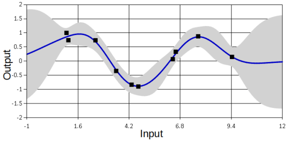
Probabilistic Machine Learning (CS772A)

Aug 3, 2017

Recap: Why a Probabilistic Approach to ML?

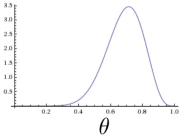
- Gives a principle way to model the uncertainty in data, parameters, and in future predictions

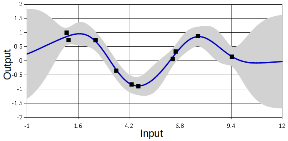




Recap: Why a Probabilistic Approach to ML?

- Gives a principle way to model the uncertainty in data, parameters, and in future predictions

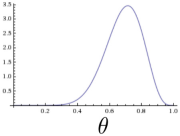


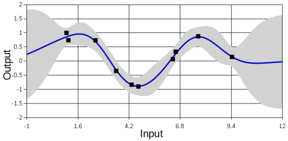


- Can construct latent variable models for data

Recap: Why a Probabilistic Approach to ML?

- Gives a principle way to model the uncertainty in data, parameters, and in future predictions

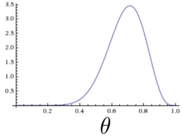


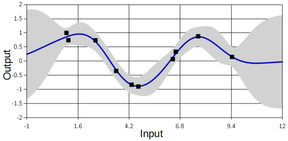


- Can construct latent variable models for data
- Can construct complex models in “modular” fashion by combining simpler probabilistic models

Recap: Why a Probabilistic Approach to ML?

- Gives a principle way to model the uncertainty in data, parameters, and in future predictions



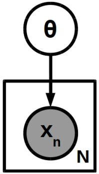


- Can construct latent variable models for data
- Can construct complex models in "modular" fashion by combining simpler probabilistic models
- Probabilistic modeling gives us a "language" to do such things naturally

Probabilistic Modeling

- Assume data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from a probabilistic model with unknown parameters θ

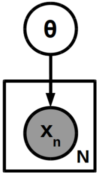
$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \sim p(\boldsymbol{x}|\theta)$$



Probabilistic Modeling

- Assume data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from a probabilistic model with unknown parameters θ

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \sim p(\boldsymbol{x}|\theta)$$

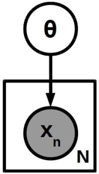


- Likelihood function $p(\boldsymbol{x}|\theta)$ or the "observation model" specifies how data is generated
  - It is also the probability of the observed data, given θ

Probabilistic Modeling

- Assume data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from a probabilistic model with unknown parameters θ

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \sim p(\boldsymbol{x}|\theta)$$

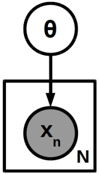


- Likelihood function $p(\boldsymbol{x}|\theta)$ or the "observation model" specifies how data is generated
  - It is also the probability of the observed data, given θ
- Prior distribution $p(\theta)$ specifies how likely different parameter values are *a priori*
  - As we'll see later, is also corresponds to imposing a "regularizer" over θ

Probabilistic Modeling

- Assume data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from a probabilistic model with unknown parameters θ

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \sim p(\boldsymbol{x}|\theta)$$

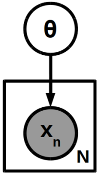


- Likelihood function $p(\boldsymbol{x}|\theta)$ or the "observation model" specifies how data is generated
  - It is also the probability of the observed data, given θ
- Prior distribution $p(\theta)$ specifies how likely different parameter values are *a priori*
  - As we'll see later, is also corresponds to imposing a "regularizer" over θ
- Goal: To estimate the unknowns of the model (θ in this case), given the observed data $\mathbf{X}$

Probabilistic Modeling

- Assume data $\mathbf{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ generated from a probabilistic model with unknown parameters θ

$$\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \sim p(\boldsymbol{x}|\theta)$$

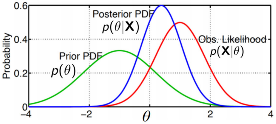


- Likelihood function $p(\boldsymbol{x}|\theta)$ or the "observation model" specifies how data is generated
  - It is also the probability of the observed data, given θ
- Prior distribution $p(\theta)$ specifies how likely different parameter values are *a priori*
  - As we'll see later, is also corresponds to imposing a "regularizer" over θ
- Goal: To estimate the unknowns of the model (θ in this case), given the observed data $\mathbf{X}$ (and use these estimates to make predictions about future data)

Estimating Parameters via (Bayesian) Inference

- Can use the Bayes rule to compute the posterior **distribution** over parameters

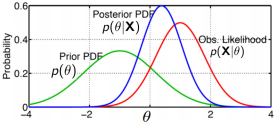
$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$



Estimating Parameters via (Bayesian) Inference

- Can use the Bayes rule to compute the posterior **distribution** over parameters

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

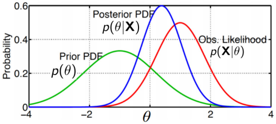


- This is known as Bayesian inference. The posterior also captures the uncertainty in θ

Estimating Parameters via (Bayesian) Inference

- Can use the Bayes rule to compute the posterior **distribution** over parameters

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

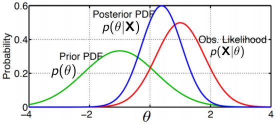


- This is known as Bayesian inference. The posterior also captures the uncertainty in θ
- The **marginal likelihood**: $p(\mathbf{X}) = \int p(\mathbf{X}|\theta)p(\theta)d\theta$

Estimating Parameters via (Bayesian) Inference

- Can use the Bayes rule to compute the posterior **distribution** over parameters

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

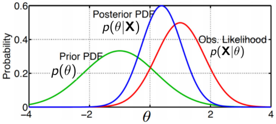


- This is known as Bayesian inference. The posterior also captures the uncertainty in θ
- The **marginal likelihood**: $p(\mathbf{X}) = \int p(\mathbf{X}|\theta)p(\theta)d\theta$
  - Hard-to-compute in general but very useful (e.g., for model-selection, hyperparam. estimation, etc.)

Estimating Parameters via (Bayesian) Inference

- Can use the Bayes rule to compute the posterior **distribution** over parameters

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

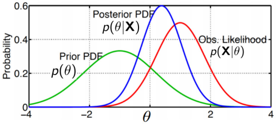


- This is known as Bayesian inference. The posterior also captures the uncertainty in θ
- The **marginal likelihood**: $p(\mathbf{X}) = \int p(\mathbf{X}|\theta)p(\theta)d\theta$
  - Hard-to-compute in general but very useful (e.g., for model-selection, hyperparam. estimation, etc.)
- In general, inferring the posterior $p(\theta|\mathbf{X})$ is a hard problem (because $p(\mathbf{X})$ can be intractable)

Estimating Parameters via (Bayesian) Inference

- Can use the Bayes rule to compute the posterior **distribution** over parameters

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$



- This is known as Bayesian inference. The posterior also captures the uncertainty in θ
- The **marginal likelihood**: $p(\mathbf{X}) = \int p(\mathbf{X}|\theta)p(\theta)d\theta$
  - Hard-to-compute in general but very useful (e.g., for model-selection, hyperparam. estimation, etc.)
- In general, inferring the posterior $p(\theta|\mathbf{X})$ is a hard problem (because $p(\mathbf{X})$ can be intractable)
- Note: It's easy if the likelihood and prior are "conjugate" to each other (we will see it later)

Estimating Parameters via Point Estimation

- Point estimation is a computationally cheaper alternative to full Bayesian inference

Estimating Parameters via Point Estimation

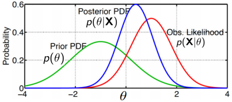
- Point estimation is a computationally cheaper alternative to full Bayesian inference
- Recall the definition of the posterior distribution over parameters

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

Estimating Parameters via Point Estimation

- Point estimation is a computationally cheaper alternative to full Bayesian inference
- Recall the definition of the posterior distribution over parameters

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

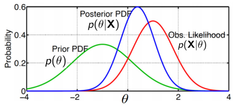


- Point Estimation finds the **single "best" estimate** of the parameters via optimization.

Estimating Parameters via Point Estimation

- Point estimation is a computationally cheaper alternative to full Bayesian inference
- Recall the definition of the posterior distribution over parameters

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$



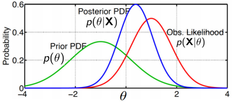
- Point Estimation finds the **single "best" estimate** of the parameters via optimization. E.g.,
  - Maximum likelihood estimation (MLE)

$$\hat{\theta} = \arg\max_{\theta} \log p(\mathbf{X}|\theta)$$

Estimating Parameters via Point Estimation

- Point estimation is a computationally cheaper alternative to full Bayesian inference
- Recall the definition of the posterior distribution over parameters

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$



- Point Estimation finds the **single "best" estimate** of the parameters via optimization. E.g.,
  - Maximum likelihood estimation (MLE)

$$\hat{\theta} = \arg\max_{\theta} \log p(\mathbf{X}|\theta)$$

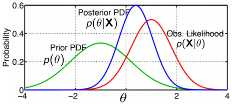
  - Maximum-a-Posteriori (MAP) estimation

$$\hat{\theta} = \arg\max_{\theta} \log p(\theta|\mathbf{X})$$

Estimating Parameters via Point Estimation

- Point estimation is a computationally cheaper alternative to full Bayesian inference
- Recall the definition of the posterior distribution over parameters

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$



- Point Estimation finds the **single "best" estimate** of the parameters via optimization. E.g.,
  - Maximum likelihood estimation (MLE)

$$\hat{\theta} = \arg\max_{\theta} \log p(\mathbf{X}|\theta)$$

  - Maximum-a-Posteriori (MAP) estimation

$$\hat{\theta} = \arg\max_{\theta} \log p(\theta|\mathbf{X})$$

  - .. and some other methods such as "method of moments"

Point Estimation via MLE

- MLE finds the parameter θ that maximizes the (log-) likelihood $p(\mathbf{X}|\theta)$

$$\mathcal{L}(\theta) = \log p(\mathbf{X}|\theta) = \log p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \theta)$$

Point Estimation via MLE

- MLE finds the parameter θ that maximizes the (log-) likelihood $p(\mathbf{X}|\theta)$

$$\mathcal{L}(\theta) = \log p(\mathbf{X}|\theta) = \log p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \theta)$$

- If the observations are i.i.d., $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)$

Point Estimation via MLE

- MLE finds the parameter θ that maximizes the (log-) likelihood $p(\mathbf{X}|\theta)$

$$\mathcal{L}(\theta) = \log p(\mathbf{X}|\theta) = \log p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \theta)$$

- If the observations are i.i.d., $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)$
- Maximum Likelihood parameter estimation

$$\boxed{\hat{\theta}_{MLE} = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta)}$$

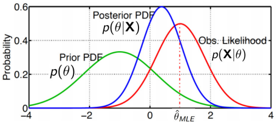
Point Estimation via MLE

- MLE finds the parameter θ that maximizes the (log-) likelihood $p(\mathbf{X}|\theta)$

$$\mathcal{L}(\theta) = \log p(\mathbf{X}|\theta) = \log p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \theta)$$

- If the observations are i.i.d., $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)$
- Maximum Likelihood parameter estimation

$$\boxed{\hat{\theta}_{MLE} = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta)}$$



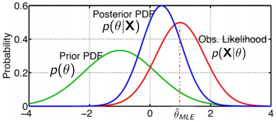
Point Estimation via MLE

- MLE finds the parameter θ that maximizes the (log-) likelihood $p(\mathbf{X}|\theta)$

$$\mathcal{L}(\theta) = \log p(\mathbf{X}|\theta) = \log p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \theta)$$

- If the observations are i.i.d., $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)$
- Maximum Likelihood parameter estimation

$$\boxed{\hat{\theta}_{MLE} = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta)}$$



- Note: MLE is "consistent", i.e., as $N \longrightarrow \infty$, $\hat{\theta}$ converges to the true θ (will see it later)

Point Estimation via MAP

- MAP estimation finds the parameter θ that maximizes the (log-) posterior probability $p(\theta|\mathbf{X})$

$$\mathcal{L}(\theta) = \log p(\theta|\mathbf{X}) = \log \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})}$$

Point Estimation via MAP

- MAP estimation finds the parameter θ that maximizes the (log-) posterior probability $p(\theta|\mathbf{X})$

$$\mathcal{L}(\theta) = \log p(\theta|\mathbf{X}) = \log \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})}$$

- Again assuming i.i.d. observations, and noting that $p(\mathbf{X})$ is independent of θ

$$\boxed{\hat{\theta}_{MAP} = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) + \log p(\theta)}$$

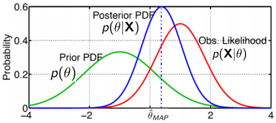
Point Estimation via MAP

- MAP estimation finds the parameter θ that maximizes the (log-) posterior probability $p(\theta|\mathbf{X})$

$$\mathcal{L}(\theta) = \log p(\theta|\mathbf{X}) = \log \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})}$$

- Again assuming i.i.d. observations, and noting that $p(\mathbf{X})$ is independent of θ

$$\boxed{\hat{\theta}_{MAP} = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) + \log p(\theta)}$$



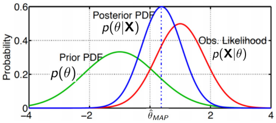
Point Estimation via MAP

- MAP estimation finds the parameter θ that maximizes the (log-) posterior probability $p(\theta|\mathbf{X})$

$$\mathcal{L}(\theta) = \log p(\theta|\mathbf{X}) = \log \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})}$$

- Again assuming i.i.d. observations, and noting that $p(\mathbf{X})$ is independent of θ

$$\boxed{\hat{\theta}_{MAP} = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) + \log p(\theta)}$$



- Note: When the prior is uniform, MAP and MLE solutions are identical

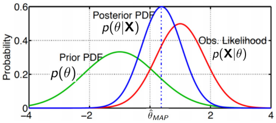
Point Estimation via MAP

- MAP estimation finds the parameter θ that maximizes the (log-) posterior probability $p(\theta|\mathbf{X})$

$$\mathcal{L}(\theta) = \log p(\theta|\mathbf{X}) = \log \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})}$$

- Again assuming i.i.d. observations, and noting that $p(\mathbf{X})$ is independent of θ

$$\boxed{\hat{\theta}_{MAP} = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) + \log p(\theta)}$$



- Note: When the prior is uniform, MAP and MLE solutions are identical
- Despite using the prior, MAP is NOT considered a Bayesian approach (still gives a point estimate)

Point Estimation (MLE/MAP) vs Loss Function Minimization

- Recall the maximum Likelihood parameter estimation procedure

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta)$$

Point Estimation (MLE/MAP) vs Loss Function Minimization

- Recall the maximum Likelihood parameter estimation procedure

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta)$$

- We can also think of it as minimizing the **negative** log-likelihood (NLL)

$$\boxed{\hat{\theta}_{MLE} = \arg\min_{\theta} NLL(\theta)}$$

where $NLL(\theta) = -\sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta)$ is called the negative log-likelihood

Point Estimation (MLE/MAP) vs Loss Function Minimization

- Recall the maximum Likelihood parameter estimation procedure

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta)$$

- We can also think of it as minimizing the **negative** log-likelihood (NLL)

$$\boxed{\hat{\theta}_{MLE} = \arg\min_{\theta} NLL(\theta)}$$

where $NLL(\theta) = -\sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta)$ is called the negative log-likelihood

- Likewise, MAP parameter estimation can be shown to have the following form

$$\boxed{\hat{\theta}_{MAP} = \arg\min_{\theta} NLL(\theta) - \log p(\theta)}$$

Point Estimation (MLE/MAP) vs Loss Function Minimization

- Recall the maximum Likelihood parameter estimation procedure

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta)$$

- We can also think of it as minimizing the **negative** log-likelihood (NLL)

$$\boxed{\hat{\theta}_{MLE} = \arg\min_{\theta} NLL(\theta)}$$

where $NLL(\theta) = -\sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta)$ is called the negative log-likelihood

- Likewise, MAP parameter estimation can be shown to have the following form

$$\boxed{\hat{\theta}_{MAP} = \arg\min_{\theta} NLL(\theta) - \log p(\theta)}$$

- Note that NLL is like a loss function and $-\log p(\theta)$ is like a regularizer on θ

Point Estimation (MLE/MAP) vs Loss Function Minimization

- Recall the maximum Likelihood parameter estimation procedure

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta)$$

- We can also think of it as minimizing the **negative** log-likelihood (NLL)

$$\boxed{\hat{\theta}_{MLE} = \arg\min_{\theta} NLL(\theta)}$$

where $NLL(\theta) = -\sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta)$ is called the negative log-likelihood

- Likewise, MAP parameter estimation can be shown to have the following form

$$\boxed{\hat{\theta}_{MAP} = \arg\min_{\theta} NLL(\theta) - \log p(\theta)}$$

- Note that NLL is like a loss function and $-\log p(\theta)$ is like a regularizer on θ
- Thus MLE is like empirical loss/risk minization (ERM) and MAP is like regularized ERM

MLE via a simple example

- Consider a sequence of N coin tosses (call head $= 0$, tail $= 1$)
- The n^{th} outcome $\boldsymbol{x}_n$ is a binary random variable $\in \{0, 1\}$

MLE via a simple example

- Consider a sequence of N coin tosses (call head $= 0$, tail $= 1$)
- The n^{th} outcome $\boldsymbol{x}_n$ is a binary random variable $\in \{0, 1\}$
- Assume θ to be probability of a head (parameter we wish to estimate)

MLE via a simple example

- Consider a sequence of N coin tosses (call head $= 0$, tail $= 1$)
- The n^{th} outcome $\boldsymbol{x}_n$ is a binary random variable $\in \{0, 1\}$
- Assume θ to be probability of a head (parameter we wish to estimate)
- Each likelihood term $p(\boldsymbol{x}_n \mid \theta)$ is Bernoulli: $p(\boldsymbol{x}_n \mid \theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$

MLE via a simple example

- Consider a sequence of N coin tosses (call head $= 0$, tail $= 1$)
- The n^{th} outcome $\boldsymbol{x}_n$ is a binary random variable $\in \{0, 1\}$
- Assume θ to be probability of a head (parameter we wish to estimate)
- Each likelihood term $p(\boldsymbol{x}_n \mid \theta)$ is Bernoulli: $p(\boldsymbol{x}_n \mid \theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- Log-likelihood: $\sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta) = \sum_{n=1}^{N} \boldsymbol{x}_n \log\theta + (1-\boldsymbol{x}_n)\log(1-\theta)$

MLE via a simple example

- Consider a sequence of N coin tosses (call head $= 0$, tail $= 1$)
- The n^{th} outcome $\boldsymbol{x}_n$ is a binary random variable $\in \{0, 1\}$
- Assume θ to be probability of a head (parameter we wish to estimate)
- Each likelihood term $p(\boldsymbol{x}_n \mid \theta)$ is Bernoulli: $p(\boldsymbol{x}_n \mid \theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- Log-likelihood: $\sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta) = \sum_{n=1}^{N} \boldsymbol{x}_n \log \theta + (1 - \boldsymbol{x}_n) \log(1-\theta)$
- Taking derivative of the log-likelihood w.r.t. θ, and setting it to zero gives

$$\hat{\theta}_{MLE} = \frac{\sum_{n=1}^{N} \boldsymbol{x}_n}{N}$$

MLE via a simple example

- Consider a sequence of N coin tosses (call head $= 0$, tail $= 1$)
- The n^{th} outcome $\boldsymbol{x}_n$ is a binary random variable $\in \{0, 1\}$
- Assume θ to be probability of a head (parameter we wish to estimate)
- Each likelihood term $p(\boldsymbol{x}_n \mid \theta)$ is Bernoulli: $p(\boldsymbol{x}_n \mid \theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- Log-likelihood: $\sum_{n=1}^N \log p(\boldsymbol{x}_n \mid \theta) = \sum_{n=1}^N \boldsymbol{x}_n \log \theta + (1-\boldsymbol{x}_n)\log(1-\theta)$
- Taking derivative of the log-likelihood w.r.t. θ, and setting it to zero gives

$$\hat{\theta}_{MLE} = \frac{\sum_{n=1}^N \boldsymbol{x}_n}{N}$$

- $\hat{\theta}_{MLE}$ in this example is simply the fraction of heads!

MLE via a simple example

- Consider a sequence of N coin tosses (call head $= 0$, tail $= 1$)
- The n^{th} outcome $\boldsymbol{x}_n$ is a binary random variable $\in \{0, 1\}$
- Assume θ to be probability of a head (parameter we wish to estimate)
- Each likelihood term $p(\boldsymbol{x}_n \mid \theta)$ is Bernoulli: $p(\boldsymbol{x}_n \mid \theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- Log-likelihood: $\sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta) = \sum_{n=1}^{N} \boldsymbol{x}_n \log \theta + (1 - \boldsymbol{x}_n) \log(1-\theta)$
- Taking derivative of the log-likelihood w.r.t. θ, and setting it to zero gives

$$\hat{\theta}_{MLE} = \frac{\sum_{n=1}^{N} \boldsymbol{x}_n}{N}$$

- $\hat{\theta}_{MLE}$ in this example is simply the fraction of heads!
- MLE doesn't have a way to express our prior belief about θ. Can be problematic especially when the number of observations is very small (e.g., suppose very few or zero heads when N is small).

MAP via a simple example

- MAP estimation can incorporate a prior $p(\theta)$ on θ

MAP via a simple example

- MAP estimation can incorporate a prior $p(\theta)$ on θ
- Since $\theta \in (0,1)$, one possibility can be to assume a Beta prior

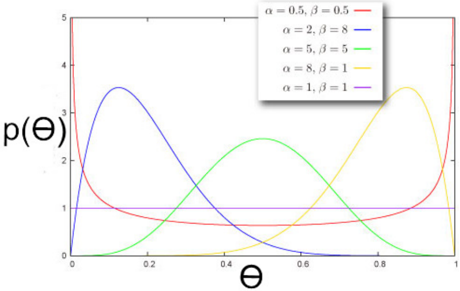
$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

MAP via a simple example

- MAP estimation can incorporate a prior $p(\theta)$ on θ
- Since $\theta \in (0,1)$, one possibility can be to assume a Beta prior

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

- α, β are called hyperparameters of the prior (these can have intuitive meaning; we'll see shortly)

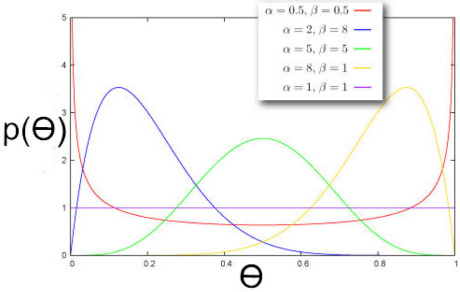


MAP via a simple example

- MAP estimation can incorporate a prior $p(\theta)$ on θ
- Since $\theta \in (0,1)$, one possibility can be to assume a Beta prior

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

- α, β are called hyperparameters of the prior (these can have intuitive meaning; we'll see shortly)



- Note that each likelihood term is still a Bernoulli: $p(\boldsymbol{x}_n|\theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$

MAP via a simple example (contd.)

- The log posterior probability $= \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) + \log p(\theta)$

MAP via a simple example (contd.)

- The log posterior probability $= \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) + \log p(\theta)$
- Ignoring the constants w.r.t. θ, the log posterior probability:

$$\sum_{n=1}^{N}\{\boldsymbol{x}_n \log\theta + (1-\boldsymbol{x}_n)\log(1-\theta)\} + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)$$

MAP via a simple example (contd.)

- The log posterior probability $= \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) + \log p(\theta)$
- Ignoring the constants w.r.t. θ, the log posterior probability:

$$\sum_{n=1}^{N}\{\boldsymbol{x}_n \log\theta + (1-\boldsymbol{x}_n)\log(1-\theta)\} + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)$$

- Taking derivative w.r.t. θ and setting to zero gives

$$\hat{\theta}_{MAP} = \frac{\sum_{n=1}^{N}\boldsymbol{x}_n + \alpha - 1}{N+\alpha+\beta-2}$$

MAP via a simple example (contd.)

- The log posterior probability $= \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) + \log p(\theta)$
- Ignoring the constants w.r.t. θ, the log posterior probability:

$$\sum_{n=1}^{N}\{\boldsymbol{x}_n \log\theta + (1-\boldsymbol{x}_n)\log(1-\theta)\} + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)$$

- Taking derivative w.r.t. θ and setting to zero gives

$$\hat{\theta}_{MAP} = \frac{\sum_{n=1}^{N}\boldsymbol{x}_n + \alpha - 1}{N+\alpha+\beta-2}$$

- Note: For $\alpha=1,\beta=1$, i.e., $p(\theta) = \text{Beta}(1,1)$ (equivalent to a uniform prior), $\hat{\theta}_{MAP} = \hat{\theta}_{MLE}$

MAP via a simple example (contd.)

- The log posterior probability $= \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) + \log p(\theta)$
- Ignoring the constants w.r.t. θ, the log posterior probability:

$$\sum_{n=1}^{N}\{\boldsymbol{x}_n \log\theta + (1-\boldsymbol{x}_n)\log(1-\theta)\} + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)$$

- Taking derivative w.r.t. θ and setting to zero gives

$$\hat{\theta}_{MAP} = \frac{\sum_{n=1}^{N}\boldsymbol{x}_n + \alpha - 1}{N+\alpha+\beta-2}$$

- Note: For $\alpha = 1, \beta = 1$, i.e., $p(\theta) = \text{Beta}(1,1)$ (equivalent to a uniform prior), $\hat{\theta}_{MAP} = \hat{\theta}_{MLE}$
- What hyperparameters represent intuitively? Hyperparameters of the prior (in this case α, β) can often be thought of as "pseudo-observations".

MAP via a simple example (contd.)

- The log posterior probability $= \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) + \log p(\theta)$
- Ignoring the constants w.r.t. θ, the log posterior probability:

$$\sum_{n=1}^{N}\{\boldsymbol{x}_n \log \theta + (1 - \boldsymbol{x}_n)\log(1-\theta)\} + (\alpha - 1)\log\theta + (\beta - 1)\log(1-\theta)$$

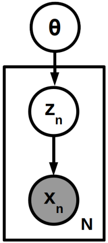
- Taking derivative w.r.t. θ and setting to zero gives

$$\hat{\theta}_{MAP} = \frac{\sum_{n=1}^{N}\boldsymbol{x}_n + \alpha - 1}{N + \alpha + \beta - 2}$$

- Note: For $\alpha = 1, \beta = 1$, i.e., $p(\theta) = \text{Beta}(1,1)$ (equivalent to a uniform prior), $\hat{\theta}_{MAP} = \hat{\theta}_{MLE}$
- What hyperparameters represent intuitively? Hyperparameters of the prior (in this case α, β) can often be thought of as "pseudo-observations".
  - $\alpha - 1$, $\beta - 1$ are the expected numbers of heads and tails, respectively, before seeing any data

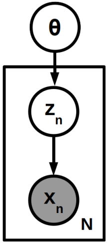
Parameter Estimation in Models with Latent Variables

- Becomes harder if there are latent variables in the model



Parameter Estimation in Models with Latent Variables

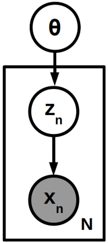
- Becomes harder if there are latent variables in the model



- Reason: Computing $\log p(\boldsymbol{x}|\theta) = \log \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}|\theta)$ is hard

Parameter Estimation in Models with Latent Variables

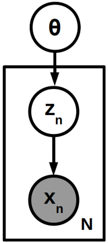
- Becomes harder if there are latent variables in the model



- Reason: Computing $\log p(\boldsymbol{x}|\theta) = \log \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}|\theta)$ is hard
- Doing parameter estimation in such cases requires using methods such as expectation-maximization (EM), MCMC, or variational methods

Parameter Estimation in Models with Latent Variables

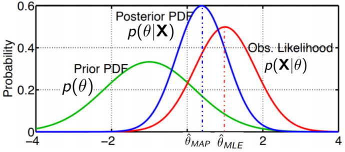
- Becomes harder if there are latent variables in the model



- Reason: Computing $\log p(\boldsymbol{x}|\theta) = \log \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}|\theta)$ is hard
- Doing parameter estimation in such cases requires using methods such as expectation-maximization (EM), MCMC, or variational methods
- More on this when we will look at latent variable models

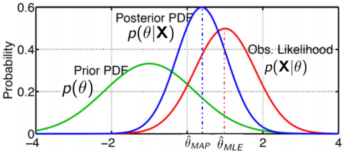
Point Estimation vs Full Posterior

- MLE and MAP only provide us with a best "point estimate" of θ
- MLE does not incorporate any prior knowledge about parameters
- MAP does incorporate prior knowledge but still only gives a point estimate



Point Estimation vs Full Posterior

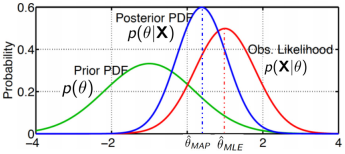
- MLE and MAP only provide us with a best "point estimate" of θ
- MLE does not incorporate any prior knowledge about parameters
- MAP does incorporate prior knowledge but still only gives a point estimate



- Point estimate doesn't capture the uncertainty about the parameter θ

Point Estimation vs Full Posterior

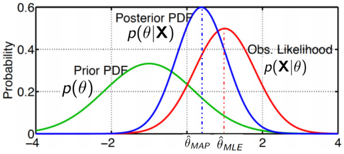
- MLE and MAP only provide us with a best "point estimate" of θ
- MLE does not incorporate any prior knowledge about parameters
- MAP does incorporate prior knowledge but still only gives a point estimate



- Point estimate doesn't capture the uncertainty about the parameter θ
- The full posterior distribution $p(\theta|\mathbf{X})$, although harder to estimate, gives a more complete picture

Point Estimation vs Full Posterior

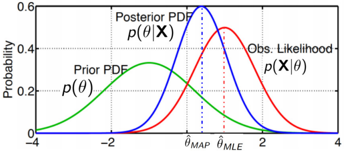
- MLE and MAP only provide us with a best "point estimate" of θ
- MLE does not incorporate any prior knowledge about parameters
- MAP does incorporate prior knowledge but still only gives a point estimate



- Point estimate doesn't capture the uncertainty about the parameter θ
- The full posterior distribution $p(\theta|\mathbf{X})$, although harder to estimate, gives a more complete picture
- In some cases, however, we can analytically compute the full posterior (e.g., when the prior distribution is **"conjugate"** to the likelihood distribution)

Point Estimation vs Full Posterior

- MLE and MAP only provide us with a best "point estimate" of θ
- MLE does not incorporate any prior knowledge about parameters
- MAP does incorporate prior knowledge but still only gives a point estimate



- Point estimate doesn't capture the uncertainty about the parameter θ
- The full posterior distribution $p(\theta|\mathbf{X})$, although harder to estimate, gives a more complete picture
- In some cases, however, we can analytically compute the full posterior (e.g., when the prior distribution is **"conjugate"** to the likelihood distribution)
- In other cases, it can be approximated via **approximate Bayesian inference** methods such as MCMC and variational inference (more on this later during the semester)

Inferring the Full Posterior: Simple Case (Coin-Toss Example)

- Recall that each likelihood term was Bernoulli: $p(\boldsymbol{x}_n|\theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- The prior $p(\theta)$ was Beta: $p(\theta) = \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

Inferring the Full Posterior: Simple Case (Coin-Toss Example)

- Recall that each likelihood term was Bernoulli: $p(\boldsymbol{x}_n|\theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- The prior $p(\theta)$ was Beta: $p(\theta) = \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The posterior distribution will be proportional to the product of likelihood and prior

$$\begin{aligned} p(\theta|\mathbf{X}) &\propto \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)p(\theta) \\ &\propto \theta^{\alpha+\sum_{n=1}^{N}\boldsymbol{x}_n-1}(1-\theta)^{\beta+N-\sum_{n=1}^{N}\boldsymbol{x}_n-1} \end{aligned}$$

Inferring the Full Posterior: Simple Case (Coin-Toss Example)

- Recall that each likelihood term was Bernoulli: $p(\boldsymbol{x}_n|\theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- The prior $p(\theta)$ was Beta: $p(\theta) = \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The posterior distribution will be <u>proportional</u> to the product of likelihood and prior

$$
\begin{aligned}
p(\theta|\mathbf{X}) &\propto \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)p(\theta) \\
&\propto \theta^{\alpha+\sum_{n=1}^{N}\boldsymbol{x}_n-1}(1-\theta)^{\beta+N-\sum_{n=1}^{N}\boldsymbol{x}_n-1}
\end{aligned}
$$

- From simple inspection, note that the posterior $p(\theta|\mathbf{X}) = \text{Beta}(\alpha + \sum_{n=1}^{N}\boldsymbol{x}_n, \beta + N - \sum_{n=1}^{N}\boldsymbol{x}_n)$

Inferring the Full Posterior: Simple Case (Coin-Toss Example)

- Recall that each likelihood term was Bernoulli: $p(\boldsymbol{x}_n|\theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- The prior $p(\theta)$ was Beta: $p(\theta) = \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The posterior distribution will be <u>proportional</u> to the product of likelihood and prior

$$
\begin{aligned}
p(\theta|\mathbf{X}) &\propto \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)p(\theta) \\
&\propto \theta^{\alpha+\sum_{n=1}^{N}\boldsymbol{x}_n-1}(1-\theta)^{\beta+N-\sum_{n=1}^{N}\boldsymbol{x}_n-1}
\end{aligned}
$$

- From simple inspection, note that the posterior $p(\theta|\mathbf{X}) = \text{Beta}(\alpha + \sum_{n=1}^{N}\boldsymbol{x}_n, \beta + N - \sum_{n=1}^{N}\boldsymbol{x}_n)$
- Here, finding the posterior boiled down to simply "multipy, add stuff, and identify the distribution"

Inferring the Full Posterior: Simple Case (Coin-Toss Example)

- Recall that each likelihood term was Bernoulli: $p(\boldsymbol{x}_n|\theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- The prior $p(\theta)$ was Beta: $p(\theta) = \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The posterior distribution will be proportional to the product of likelihood and prior

$$
\begin{aligned}
p(\theta|\mathbf{X}) &\propto \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)p(\theta) \\
&\propto \theta^{\alpha+\sum_{n=1}^{N}\boldsymbol{x}_n-1}(1-\theta)^{\beta+N-\sum_{n=1}^{N}\boldsymbol{x}_n-1}
\end{aligned}
$$

- From simple inspection, note that the posterior $p(\theta|\mathbf{X}) = \text{Beta}(\alpha + \sum_{n=1}^{N}\boldsymbol{x}_n, \beta + N - \sum_{n=1}^{N}\boldsymbol{x}_n)$
- Here, finding the posterior boiled down to simply "multipy, add stuff, and identify the distribution"
- Note: Can verify (exercise) that the normalization constant $= \frac{\Gamma(\alpha+\sum_{n=1}^{N}\boldsymbol{x}_n)\Gamma(\beta+N-\sum_{n=1}^{N}\boldsymbol{x}_n)}{\Gamma(\alpha+\beta+N)}$
  - To verify, make use of the fact that $\int p(\theta|\mathbf{X})d\theta = 1$

Inferring the Full Posterior: Simple Case (Coin-Toss Example)

- Recall that each likelihood term was Bernoulli: $p(\boldsymbol{x}_n|\theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- The prior $p(\theta)$ was Beta: $p(\theta) = \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The posterior distribution will be proportional to the product of likelihood and prior

$$
\begin{aligned}
p(\theta|\mathbf{X}) &\propto \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)p(\theta) \\
&\propto \theta^{\alpha+\sum_{n=1}^{N}\boldsymbol{x}_n-1}(1-\theta)^{\beta+N-\sum_{n=1}^{N}\boldsymbol{x}_n-1}
\end{aligned}
$$

- From simple inspection, note that the posterior $p(\theta|\mathbf{X}) = \text{Beta}(\alpha + \sum_{n=1}^{N}\boldsymbol{x}_n, \beta + N - \sum_{n=1}^{N}\boldsymbol{x}_n)$
- Here, finding the posterior boiled down to simply "multipy, add stuff, and identify the distribution"
- Note: Can verify (exercise) that the normalization constant $= \frac{\Gamma(\alpha+\sum_{n=1}^{N}\boldsymbol{x}_n)\Gamma(\beta+N-\sum_{n=1}^{N}\boldsymbol{x}_n)}{\Gamma(\alpha+\beta+N)}$
  - To verify, make use of the fact that $\int p(\theta|\mathbf{X})d\theta = 1$
- Here, the posterior has the same form as the prior (both Beta): property of **conjugate priors**.

Conjugate Priors

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

Conjugate Priors

- Many pairs of distributions are conjugate to each other. E.g.,
  - Bernoulli (likelihood) + Beta (prior) $\Rightarrow$ Beta posterior
  - Binomial (likelihood) + Beta (prior) $\Rightarrow$ Beta posterior
  - Multinomial (likelihood) + Dirichlet (prior) $\Rightarrow$ Dirichlet posterior
  - Poisson (likelihood) + Gamma (prior) $\Rightarrow$ Gamma posterior
  - Gaussian (likelihood) + Gaussian (prior) $\Rightarrow$ Gaussian posterior
  - and many other such pairs ..
- Easy to identify if two distributions are conjugate to each other: their functional forms are similar
  - E.g., recall the forms of Bernoulli and Beta

$$\text{Bernoulli} \propto \theta^x(1-\theta)^{1-x}, \quad \text{Beta} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Conjugate Priors

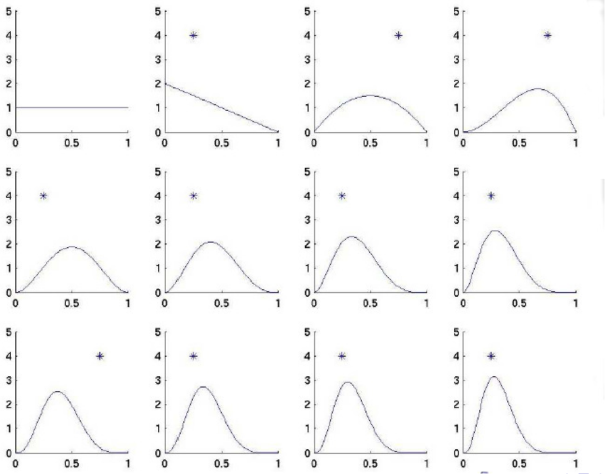
- Many pairs of distributions are conjugate to each other. E.g.,
  - Bernoulli (likelihood) + Beta (prior) $\Rightarrow$ Beta posterior
  - Binomial (likelihood) + Beta (prior) $\Rightarrow$ Beta posterior
  - Multinomial (likelihood) + Dirichlet (prior) $\Rightarrow$ Dirichlet posterior
  - Poisson (likelihood) + Gamma (prior) $\Rightarrow$ Gamma posterior
  - Gaussian (likelihood) + Gaussian (prior) $\Rightarrow$ Gaussian posterior
  - and many other such pairs ..
- Easy to identify if two distributions are conjugate to each other: their functional forms are similar
  - E.g., recall the forms of Bernoulli and Beta

$$\text{Bernoulli} \propto \theta^{x}(1-\theta)^{1-x}, \quad \text{Beta} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

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

Posterior Evolution with Observed Data

- Assume starting with a uniform prior (equivalent to Beta(1,1)) in the coin-toss example and observing a sequence of heads and tails



Information within the Posterior

- The posterior distribution for the coin-toss example $p(\theta|\mathbf{X}) = \text{Beta}(\theta|\alpha + N_1, \beta + N_0)$ where
  - N_1: number of observed heads
  - $N_0 = N - N_1$: number of observed tails

Information within the Posterior

- The posterior distribution for the coin-toss example $p(\theta|\mathbf{X}) = \text{Beta}(\theta|\alpha + N_1, \beta + N_0)$ where
  - N_1: number of observed heads
  - $N_0 = N - N_1$: number of observed tails
- The posterior summarizes everything about θ. In particular,
  - Mean of this Beta posterior: $\mathbb{E}[\theta|\mathbf{X}] = \frac{\alpha+N_1}{\alpha+N_1+\beta+N_0} = \frac{\alpha+N_1}{\alpha+\beta+N}$
  - Variance of this posterior: $\text{var}[\theta|\mathbf{X}] = \frac{(\alpha+N_1)(\beta+N_0)}{(\alpha+N_1+\beta+N_0)^2(\alpha+N_1+\beta+N_0+1)}$

Information within the Posterior

- The posterior distribution for the coin-toss example $p(\theta|\mathbf{X}) = \text{Beta}(\theta|\alpha + N_1, \beta + N_0)$ where
  - N_1: number of observed heads
  - $N_0 = N - N_1$: number of observed tails
- The posterior summarizes everything about θ. In particular,
  - Mean of this Beta posterior: $\mathbb{E}[\theta|\mathbf{X}] = \frac{\alpha+N_1}{\alpha+N_1+\beta+N_0} = \frac{\alpha+N_1}{\alpha+\beta+N}$
  - Variance of this posterior: $\text{var}[\theta|\mathbf{X}] = \frac{(\alpha+N_1)(\beta+N_0)}{(\alpha+N_1+\beta+N_0)^2(\alpha+N_1+\beta+N_0+1)}$
- Mode of this Beta posterior (MAP estimate): $\theta_{MAP} = \frac{\alpha+N_1-1}{\alpha+N_1-1+\beta+N_0-1} = \frac{\alpha+N_1-1}{\alpha+\beta+N-2}$

Information within the Posterior

- The posterior distribution for the coin-toss example $p(\theta|\mathbf{X}) = \text{Beta}(\theta|\alpha + N_1, \beta + N_0)$ where
  - N_1: number of observed heads
  - $N_0 = N - N_1$: number of observed tails
- The posterior summarizes everything about θ. In particular,
  - Mean of this Beta posterior: $\mathbb{E}[\theta|\mathbf{X}] = \frac{\alpha+N_1}{\alpha+N_1+\beta+N_0} = \frac{\alpha+N_1}{\alpha+\beta+N}$
  - Variance of this posterior: $\text{var}[\theta|\mathbf{X}] = \frac{(\alpha+N_1)(\beta+N_0)}{(\alpha+N_1+\beta+N_0)^2(\alpha+N_1+\beta+N_0+1)}$
- Mode of this Beta posterior (MAP estimate): $\theta_{MAP} = \frac{\alpha+N_1-1}{\alpha+N_1-1+\beta+N_0-1} = \frac{\alpha+N_1-1}{\alpha+\beta+N-2}$
- MLE is the same as MAP with uniform or Beta(α, β) prior with $\alpha = \beta = 1$: $\theta_{MLE} = \frac{N_1}{N}$

Information within the Posterior

- The posterior distribution for the coin-toss example $p(\theta|\mathbf{X}) = \text{Beta}(\theta|\alpha + N_1, \beta + N_0)$ where
  - N_1: number of observed heads
  - $N_0 = N - N_1$: number of observed tails
- The posterior summarizes everything about θ. In particular,
  - Mean of this Beta posterior: $\mathbb{E}[\theta|\mathbf{X}] = \frac{\alpha+N_1}{\alpha+N_1+\beta+N_0} = \frac{\alpha+N_1}{\alpha+\beta+N}$
  - Variance of this posterior: $\text{var}[\theta|\mathbf{X}] = \frac{(\alpha+N_1)(\beta+N_0)}{(\alpha+N_1+\beta+N_0)^2(\alpha+N_1+\beta+N_0+1)}$
- Mode of this Beta posterior (MAP estimate): $\theta_{MAP} = \frac{\alpha+N_1-1}{\alpha+N_1-1+\beta+N_0-1} = \frac{\alpha+N_1-1}{\alpha+\beta+N-2}$
- MLE is the same as MAP with uniform or $\text{Beta}(\alpha, \beta)$ prior with $\alpha = \beta = 1$: $\theta_{MLE} = \frac{N_1}{N}$
- Note: Doing MAP ($\theta_{MAP} = \arg\max_\theta P(\theta|\mathbf{X})$) and MLE ($\theta_{MLE} = \arg\max_\theta P(\mathbf{X}|\theta)$) directly will also give us the same answers as above. But we won't get the full posterior in these cases.

Using Posterior for Making Predictions

- How do we use the inferred posterior $P(\theta|\mathbf{X})$ over θ to <u>predict</u> something?

Using Posterior for Making Predictions

- How do we use the inferred posterior $P(\theta|\mathbf{X})$ over θ to <u>predict</u> something?
- Let's say we want to compute the probability that the next outcome will be a head

Using Posterior for Making Predictions

- How do we use the inferred posterior $P(\theta|\mathbf{X})$ over θ to predict something?
- Let's say we want to compute the probability that the next outcome will be a head
- We can do posterior averaging (averaging over all θ weighted by their posterior probabilities):

$$P(\boldsymbol{x}_{N+1} = 1|\mathbf{X}) \quad = \quad \int_0^1 P(\boldsymbol{x}_{N+1} = 1|\theta)P(\theta|\mathbf{X})d\theta$$

Using Posterior for Making Predictions

- How do we use the inferred posterior $P(\theta|\mathbf{X})$ over θ to predict something?
- Let's say we want to compute the probability that the next outcome will be a head
- We can do posterior averaging (averaging over all θ weighted by their posterior probabilities):

$$\begin{aligned}
P(\boldsymbol{x}_{N+1} = 1|\mathbf{X}) &= \int_0^1 P(\boldsymbol{x}_{N+1} = 1|\theta)P(\theta|\mathbf{X})d\theta \\
&= \int_0^1 \theta \times \text{Beta}(\theta|\alpha + N_1, \beta + N_0)
\end{aligned}$$

Using Posterior for Making Predictions

- How do we use the inferred posterior $P(\theta|\mathbf{X})$ over θ to predict something?
- Let's say we want to compute the probability that the next outcome will be a head
- We can do posterior averaging (averaging over all θ weighted by their posterior probabilities):

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

Using Posterior for Making Predictions

- How do we use the inferred posterior $P(\theta|\mathbf{X})$ over θ to predict something?
- Let's say we want to compute the probability that the next outcome will be a head
- We can do posterior averaging (averaging over all θ weighted by their posterior probabilities):

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

Using Posterior for Making Predictions

- How do we use the inferred posterior $P(\theta|\mathbf{X})$ over θ to predict something?
- Let's say we want to compute the probability that the next outcome will be a head
- We can do posterior averaging (averaging over all θ weighted by their posterior probabilities):

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

- Therefore the posterior predictive distribution: $p(\boldsymbol{x}_{N+1}|\mathbf{X}) = \text{Bernoulli}(\boldsymbol{x}_{N+1} \mid \mathbb{E}[\theta|\mathbf{X}])$

Using Posterior for Making Predictions

- How do we use the inferred posterior $P(\theta|\mathbf{X})$ over θ to predict something?
- Let's say we want to compute the probability that the next outcome will be a head
- We can do posterior averaging (averaging over all θ weighted by their posterior probabilities):

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

- Therefore the posterior predictive distribution: $p(\boldsymbol{x}_{N+1}|\mathbf{X}) = \text{Bernoulli}(\boldsymbol{x}_{N+1} \mid \mathbb{E}[\theta|\mathbf{X}])$
- Note that the predictive distribution doesn't depend on a single value of θ (θ_{MLE} or θ_{MAP})