# Bayesian Methods

David Rosenberg

New York University

April 22, 2015

# Frequentist or "Classical" Statistics

- Probability model with parameter $\theta \in \Theta$

$$\{p(y;\theta) \mid \theta \in \Theta\},$$

  where $p(y;\theta)$ is either a PDF or a PMF.
- Assume that $p(y;\theta)$ governs the world we are observing.
- In **frequentist statistics**, the **parameter** $\theta$ is a
  - **fixed constant** (i.e. not random) and is
  - **unknown** to us.
- If we knew $\theta$, there would be no need for statistics.
- Instead of $\theta$, we have a **sample** $\mathcal{D} = \{y_1, \ldots, y_n\}$ i.i.d. $p(y;\theta)$.
- Statistics is about how to use $\mathcal{D}$ in place of $\theta$.

# Point Estimation

- One type of statistical problem is **point estimation**.
- A **statistic** $s = s(\mathcal{D})$ is any function of the data.
- A statistic $\hat{\theta} = \hat{\theta}(\mathcal{D})$ is a **point estimator** if $\hat{\theta} \approx \theta$.
- Desirable statistical properties of point estimators:
  - **Consistency:** As data size $n \to \infty$, we get $\hat{\theta} \to \theta$.
  - **Efficiency:** (Roughly speaking) For large $n$, $\hat{\theta}$ achieves accuracy at least as good as any other estimator.
  - e.g. **maximum likelihood estimation** is consistent and efficient under reasonable conditions.
- In frequentist statistics, you can make up any estimator you want.
  - Justify its use by showing it has desirable properties.

# Bayesian Statistics

- Major viewpoint change In **Bayesian statistics**:
    - parameter $\theta \in \Theta$ is a **random variable**.
- New ingredient: the **prior distribution:**
    - a distribution on parameter space $\Theta$.
    - Reflects our belief about $\theta$.
    - Must be chosen before seeing any data.

# The Bayesian Method

1. Define the model:
   - Choose a distribution $p(\theta)$, called the **prior distribution**.
   - Choose a probability model or "**likelihood model**", now written as:

$$\{p(y \mid \theta) \mid \theta \in \Theta\}.$$

2. After observing $\mathcal{D}$, compute the **posterior distribution** $p(\theta \mid \mathcal{D})$.
3. Decide the **action** based on $p(\theta \mid \mathcal{D})$.

# The Posterior Distribution

- By Bayes rule, can write the posterior distribution as

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}.$$

- **likelihood:** $p(\mathcal{D} \mid \theta)$
- **prior:** $p(\theta)$
- **marginal likelihood**: $p(\mathcal{D})$.
- Note: $p(\mathcal{D})$ is just a normalizing constant for $p(\theta \mid \mathcal{D})$. Can write

$$\underbrace{p(\theta \mid \mathcal{D})}_{\text{posterior}} \sim \underbrace{p(\mathcal{D} \mid \theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}.$$

# Recap and Interpretation

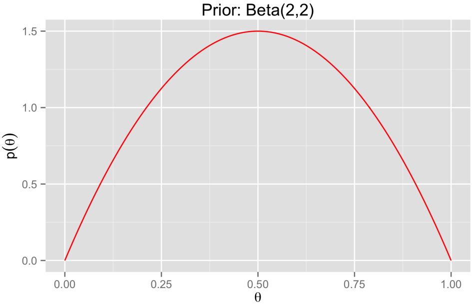
- Prior represents belief about $\theta$ before observing data $\mathcal{D}$.
- Posterior represents the rationally "updated" beliefs after seeing $\mathcal{D}$.
- All inferences and action-taking are based on the posterior distribution.
- In the Bayesian approach,
  - No issue of "choosing a procedure" or justifying an estimator.
  - Only choices are the **prior** and the **likelihood mode**l.
  - For decision making, need a **loss function**.
  - Everything after that is **computation**.

# Example: Coin Flipping

- Suppose we have a coin, possibly biased

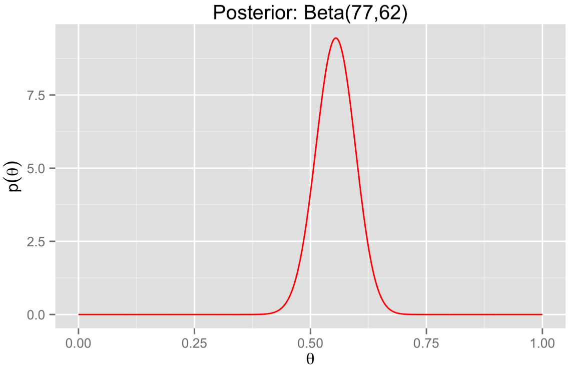
$$\mathbb{P}(\text{Heads} \mid \theta) = \theta.$$

- **Parameter space** $\theta \in \Theta = [0, 1]$.
- **Prior distribution:** $\theta \sim \text{Beta}(2, 2)$.



# Example: Coin Flipping

- Next, we gather some data $\mathcal{D} = \{H, H, T, T, T, T, T, H, \ldots, T\}$:
- Heads: 75      Tails: 60
  - $\hat{\theta}_{\text{MLE}} = \frac{75}{75+60} \approx 0.556$
- **Posterior distribution:** $\theta \mid \mathcal{D} \sim \textbf{Beta}(77, 62)$:



# What to do with the Posterior Distribution?

- Look at it.
- Extract a point estimate of $\theta$ (e.g. mean or mode of posterior).
- Extract "**credible set**" for $\theta$ (a Bayesian confidence interval).
    - e.g. Interval $[a, b]$ is a 95% **credible set** if

$$\mathbb{P}\left(\theta \in [a, b] \mid \mathcal{D}\right) \geqslant 0.95$$

- The most "Bayesian" approach is **Bayesian decision theory**:
    - Choose a loss function.
    - Find action minimizing "posterior risk".

# Bayesian Decision Theory

- Ingredients:
  - **Action space** $\mathcal{A}$.
  - **Parameter space** $\Theta$.
  - **Loss function**: $\ell : \mathcal{A} \times \Theta \to \mathbf{R}$.
  - **Prior**: Distribution $p(\theta)$ on $\Theta$.
- The **posterior risk** of an action $a \in \mathcal{A}$ is

$$
\begin{aligned}
r(a) &:= \mathbb{E}\left[\ell(\theta, a) \mid \mathcal{D}\right] \\
&= \int \ell(\theta, a) p(\theta \mid \mathcal{D})\, d\theta.
\end{aligned}
$$

  - It's the **expected loss under the posterior.**
- A **Bayes action** $a^*$ is an action that minimizes posterior risk:

$$
r(a^*) = \min_{a \in \mathcal{A}} r(a)
$$

# Bayesian Point Estimation

- General Setup:
  - Data $\mathcal{D}$ generated by $p(y \mid \theta)$, for unknown $\theta \in \Theta$.
  - Want to produce a **point estimate** for $\theta$.
- Choose the following:
  - **Loss** $\ell(\hat{\theta}, \theta) = \left(\theta - \hat{\theta}\right)^2$
  - **Prior** $p(\theta)$ on $\Theta$.
- Find **action** $\hat{\theta} \in \Theta$ that minimizes **posterior risk:**

$$\begin{aligned} r(\hat{\theta}) &= \mathbb{E}\left[\left(\theta - \hat{\theta}\right)^2 \mid \mathcal{D}\right] \\ &= \int \left(\theta - \hat{\theta}\right)^2 p(\theta \mid \mathcal{D})\, d\theta \end{aligned}$$

# Bayesian Point Estimation: Square Loss

- Find **action** $\hat{\theta} \in \Theta$ that minimizes **posterior risk**

$$r(\hat{\theta}) = \int \left(\theta - \hat{\theta}\right)^2 p(\theta \mid \mathcal{D})\, d\theta.$$

- Differentiate:

$$\begin{aligned}
\frac{dr(\hat{\theta})}{d\hat{\theta}} &= -\int 2\left(\theta - \hat{\theta}\right) p(\theta \mid \mathcal{D})\, d\theta \\
&= -2\int \theta p(\theta \mid \mathcal{D})\, d\theta + 2\hat{\theta} \underbrace{\int p(\theta \mid \mathcal{D})\, d\theta}_{=1} \\
&= -2\int \theta p(\theta \mid \mathcal{D})\, d\theta + 2\hat{\theta}
\end{aligned}$$

# Bayesian Point Estimation: Square Loss

- Derivative of posterior risk is

$$\frac{dr(\hat{\theta})}{d\hat{\theta}} = -2\int \theta p(\theta \mid \mathcal{D})\, d\theta + 2\hat{\theta}.$$

- First order condition $\frac{dr(\hat{\theta})}{d\hat{\theta}} = 0$ gives

$$\begin{aligned} \hat{\theta} &= \int \theta p(\theta \mid \mathcal{D})\, d\theta \\ &= \mathbb{E}[\theta \mid \mathcal{D}] \end{aligned}$$

- **Bayes action** for **square loss** is the posterior mean.

# Bayesian Point Estimation: Absolute Loss

- **Loss:** $\ell(\theta, \hat{\theta}) = \left|\theta - \hat{\theta}\right|$
- **Bayes action** for **absolute loss** is the **posterior median.**
  - That is, the median of the distribution $p(\theta \mid \mathcal{D})$.
  - Show with approach similar to what was used in Homework #1.

# Bayesian Point Estimation: Zero-One Loss

- Suppose $\Theta$ is discrete (e.g. $\Theta = \{\text{english}, \text{french}\}$)
- **Zero-one loss:** $\ell(\theta, \hat{\theta}) = 1(\theta \neq \hat{\theta})$
- **Posterior risk**:

$$
\begin{aligned}
r(\hat{\theta}) &= \mathbb{E}\left[1(\theta \neq \hat{\theta}) \mid \mathcal{D}\right] \\
&= \mathbb{P}\left(\theta \neq \hat{\theta} \mid \mathcal{D}\right) \\
&= 1 - \mathbb{P}\left(\theta = \hat{\theta} \mid \mathcal{D}\right) \\
&= 1 - p(\hat{\theta} \mid \mathcal{D})
\end{aligned}
$$

- **Bayes action** is

$$
\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, p(\theta \mid \mathcal{D})
$$

- This $\hat{\theta}$ is called the **maximum a posteriori (MAP)** estimate.
- The MAP estimate is the **mode** of the posterior distribution.

## Bayesian Point Estimation: Custom Loss Function

- Suppose $\Theta$ is discrete (e.g. $\Theta = \{\text{english}, \text{french}\}$)
- **Loss function** $\ell(\hat{\theta}, \theta)$:

$$
\begin{aligned}
\ell(\text{french}, \text{english}) &= 10 \\
\ell(\text{english}, \text{french}) &= 1 \\
\ell(\text{english}, \text{english}) &= 0 \\
\ell(\text{french}, \text{french}) &= 0
\end{aligned}
$$

- **Posterior risk**:

$$
\begin{aligned}
r(\text{french}) &= 10p(\text{english} \mid \mathcal{D}) + 0p(\text{french} \mid \mathcal{D}) \\
r(\text{english}) &= 1p(\text{french} \mid \mathcal{D}) + 0p(\text{english} \mid \mathcal{D})
\end{aligned}
$$

- **Bayes action** is english iff $r(\text{english}) > r(\text{english})$, i.e. when

$$
\frac{p(\text{french} \mid \mathcal{D})}{p(\text{english} \mid \mathcal{D})} = 10.
$$

# Bayesian Conditional Models

- Input space $\mathcal{X} = \mathbf{R}^d$ Output space $\mathcal{Y} = \mathbf{R}$
- **Conditional probability model**, or **likelihood model**:

$$\{p(y \mid x, \theta) \mid \theta \in \Theta\}$$

  - Conditional here refers to the conditioning on the input $x$.
  - Means that $x$'s are known and not governed by our probability model.

# Gaussian Regression Model

- Input space $\mathcal{X} = \mathbf{R}^d$ Output space $\mathcal{Y} = \mathbf{R}$
- **Conditional probability model**, or **likelihood model**:

$$y \mid x, \theta \sim \mathcal{N}\left(\theta^T x, \sigma^2\right),$$

  for some known $\sigma^2 > 0$.
- **Parameter space** $\Theta = \mathbf{R}^d$.
- **Data:** $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
  - Write $y = (y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$.
  - Assume $y_i$'s are **conditionally independent**, given $x$ and $\theta$.

# Gaussian Likelihood

- The **likelihood** of $\theta \in \Theta$ for the data $\mathcal{D}$ is

$$
\begin{aligned}
p(y \mid x, \theta) &= \prod_{i=1}^{n} p(y_i \mid x_i, \theta) \qquad \text{by conditional independence.} \\
&= \prod_{i=1}^{n} \left[ \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{(y_i - \theta^T x_i)^2}{2\sigma^2} \right) \right]
\end{aligned}
$$

- Recall from the GLM lecture[1] that the **MLE** is

$$
\begin{aligned}
\theta^*_{\text{MLE}} &= \underset{\theta \in \mathbf{R}^d}{\arg\min}\, p(y \mid x, \theta) \\
&= \underset{\theta \in \mathbf{R}^d}{\arg\min} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2
\end{aligned}
$$

[1] `https://davidrosenberg.github.io/ml2015/docs/8.Lab.glm.pdf`, slide 5.

# Priors and Posteriors

- Choose a Gaussian **prior distribution** $p(\theta)$ on $\Theta$:

$$\theta \sim \mathcal{N}(0, \Sigma_0)$$

  for some **covariance matrix** $\Sigma_0 \succ 0$ (i.e. $\Sigma_0$ is spd).

- **Posterior distribution**

$$\begin{aligned}
p(\theta \mid \mathcal{D}) &= p(\theta \mid x, y) \\
&= p(y \mid x, \theta)\, p(\theta) / p(y) \\
&\propto p(y \mid x, \theta) p(\theta) \\
&= \prod_{i=1}^{n} \left[ \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{(y_i - \theta^T x_i)^2}{2\sigma^2} \right) \right] \quad \textbf{(likelihood)} \\
&\quad \times |2\pi\Sigma_0|^{-1/2} \exp\left( -\frac{1}{2} \theta^T \Sigma_0^{-1} \theta \right) \quad \textbf{(prior)}
\end{aligned}$$

# Example in 1-Dimension

- Input space $\mathcal{X} = [-1, 1]$ Output space $\mathcal{Y} = \mathbf{R}$
- Basic Gaussian regression model:

$$y = w_0 + w_1 x + \varepsilon,$$

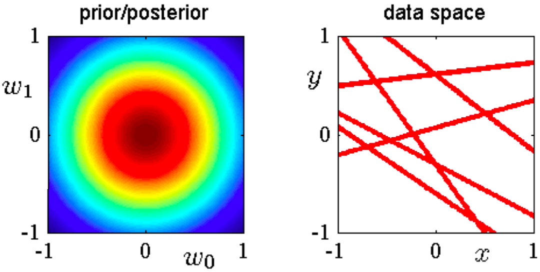
where $\varepsilon \sim \mathcal{N}(0, 0.2^2)$.

- Written another way, the **likelihood model** is

$$y \mid x, \theta = (w_0, w_1) \quad \sim \quad \mathcal{N}\left(w_0 + w_1 x \,,\, 0.2^2\right).$$

# Example in 1-Dimension

- **Prior distribution:** $\theta = (w_0, w_1) \sim \mathcal{N}\left(0, \frac{1}{2}I\right)$




- On right, plots of $y = w_0 + w_1 x$ for random $(w_0, w_1) \sim p(\theta) = \mathcal{N}(0, \frac{1}{2}I)$.

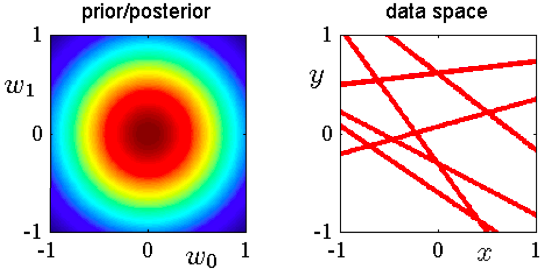
Bishop's PRML Fig 3.7

# Example in 1-Dimension

- Consider $y$ and $x$ related as $y = w_0 + w_1 x + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 0.2^2)$.
- **Conditional probability model**, or **likelihood model**:

$$y \mid x, \theta = (w_0, w_1) \quad \sim \quad \mathcal{N}\left(w_0 + w_1 x \,,\, 0.2^2\right).$$

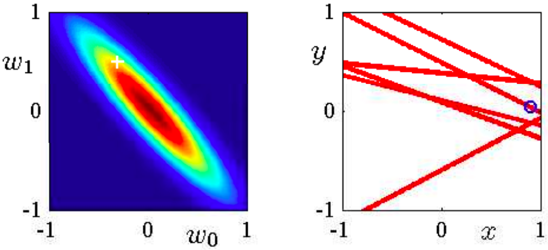
- **Prior distribution:** $\theta = (w_0, w_1) \sim \mathcal{N}\left(0, \frac{1}{2} I\right)$




- On right, plots of $y = w_0 + w_1 x$ for random $(w_0, w_1) \sim p(\theta) = \mathcal{N}(0, \frac{1}{2} I)$.

Bishop's PRML Fig 3.7

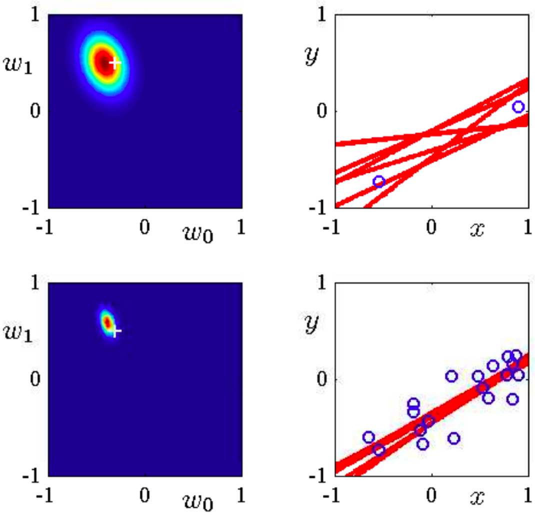
# Example in 1-Dimension: 1 Observation




- On left, the white cross indicates the true parameter values.
- On right, the blue circle indicates the training observation.

Bishop's PRML Fig 3.7

# Example in 1-Dimension: 2 and 20 Observations




Bishop's PRML Fig 3.7

# Predictive Distribution

- Given a new input point $x_{\text{new}}$, how to predict $y_{\text{new}}$ ?
- **Predictive distribution**

$$
\begin{aligned}
& p(y_{\textbf{new}} \mid x_{\textbf{new}}, \mathcal{D}) \\
= & \int p(y_{\textbf{new}} \mid x_{\textbf{new}}, \theta, \mathcal{D}) p(\theta \mid \mathcal{D})\, d\theta \\
= & \int p(y_{\textbf{new}} \mid x_{\textbf{new}}, \theta) p(\theta \mid \mathcal{D})\, d\theta
\end{aligned}
$$

- For Gaussian regression, posterior and predictive distributions have closed forms.

# Closed Form for Posterior

- Model:

$$\begin{aligned} \theta &\sim \mathcal{N}(0, \Sigma_0) \\ y_i \mid x, \theta &\text{ i.i.d. } \mathcal{N}(\theta^T x_i, \sigma^2) \end{aligned}$$

- Design matrix $X$ Response column vector $y$
- **Posterior distribution is a Gaussian distribution:**

$$\begin{aligned} \theta \mid \mathcal{D} &\sim \mathcal{N}(\mu_P, \Sigma_P) \\ \Sigma_\mathbf{P} &= \left(\sigma^{-2} X^T X + \Sigma_0^{-1}\right)^{-1} \\ \mu_\mathbf{P} &= \sigma^{-2} \Sigma_\mathbf{P} X^T y \end{aligned}$$

- **Posterior Variance** $\Sigma_P$ gives us a natural **uncertainty measure.**

See Rasmussen and Williams' *Gaussian Processes for Machine Learning*, Ch 2.1. http://www.gaussianprocess.org/gpml/chapters/RW2.pdf

# Closed Form for Posterior

- **Posterior distribution is a Gaussian distribution:**

$$\begin{aligned} \theta \mid \mathcal{D} &\sim \mathcal{N}(\mu_P, \Sigma_P) \\ \Sigma_P &= \left(\sigma^{-2} X^T X + \Sigma_0^{-1}\right)^{-1} \\ \mu_P &= \sigma^{-2} \Sigma_P X^T y \end{aligned}$$

- The **MAP estimator** and the **posterior mean** are given by

$$\mu_P = \left(X^T X + \sigma^2 \Sigma_0^{-1}\right)^{-1} X^T y$$

- Look familiar?
- For the prior variance $\Sigma_0 = \frac{\sigma^2}{\lambda} I$, we get

$$\mu_P = \left(X^T X + \lambda I\right)^{-1} X^T y,$$

which is of course the ridge regression solution.

# Posterior Mean and Posterior Mode (MAP)

- Posterior density for $\Sigma_0 = \frac{\sigma^2}{\lambda} I$:

$$p(\theta \mid \mathcal{D}) \quad \propto \quad \underbrace{\exp\left(-\frac{\lambda}{2\sigma^2}\|\theta\|^2\right)}_{\textbf{prior}} \underbrace{\prod_{i=1}^{n} \exp\left(-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right)}_{\textbf{likelihood}}$$

- To find MAP, sufficient to minimize the log posterior:

$$\begin{aligned} \hat{\theta}_{\mathbf{MAP}} &= \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \left[-\log p(\theta \mid \mathcal{D})\right] \\ &= \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \underbrace{\sum_{i=1}^{n} (y_i - \theta^T x_i)^2}_{\textbf{log-likelihood}} + \underbrace{\lambda \|\theta\|^2}_{\textbf{log-prior}} \end{aligned}$$

- Which is the ridge regression objective.

# Closed Form for Predictive Distribution

- **Model**:

$$\begin{array}{rcl} \theta & \sim & \mathcal{N}(0, \Sigma_0) \\ y_i \mid x, \theta & \text{i.i.d.} & \mathcal{N}(\theta^T x_i, \sigma^2) \end{array}$$

- **Predictive Distribution**

$$p(y_{\text{new}} \mid x_{\text{new}}, \mathcal{D}) = \int p(y_{\text{new}} \mid x_{\text{new}}, \theta) p(\theta \mid \mathcal{D})\, d\theta.$$

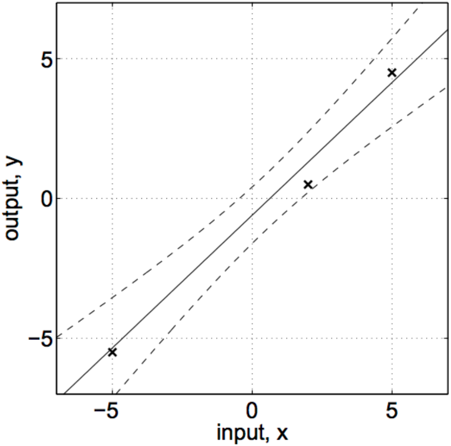
  - Averages over prediction for each $\theta$, weighted by posterior distribution.

- **Closed form:**

$$\begin{array}{rcl} y_{\text{new}} \mid x_{\text{new}}, \mathcal{D} & \sim & \mathcal{N}(\eta_{\text{new}}, \sigma_{\text{new}}) \\ \mu_{\text{new}} & = & \mu_P^T x_{\text{new}} \\ \sigma_{\text{new}} & = & \underbrace{x_{\text{new}}^T \Sigma_P x_{\text{new}}}_{\text{from variance in } \theta} + \underbrace{\sigma^2}_{\text{inherent variance in } y} \end{array}$$

# Predictive Distributions

- With predictive distributions, can draw error bands:




Rasmussen and Williams' *Gaussian Processes for Machine Learning*, Fig.2.1(b)

# Bayesian Predictive Distributions vs GLMs

- Gaussian regression with MLE, from our GLM lecture:
  - produces a Gaussian for each input $x$.

$$x \mapsto \mathcal{N}\left(x^T \theta_{\text{MLE}}, \sigma^2\right)$$

- Bayesian predictive distributions:
  - produce a Gaussian for each input $x$

$$x \mapsto \mathcal{N}\left(\theta_{\text{ridge}}^T x, \underbrace{x_{\text{new}}^T \Sigma_{\text{P}} x_{\text{new}}}_{\text{from variance in } \theta} + \underbrace{\sigma^2}_{\text{inherent variance in } y}\right)$$

- In Bayesian version
  - equivalent to using a **regularized** least squares fit
  - variance has additional piece from uncertainty in $\theta$

# Coin Flipping

- **Parameter space** $\theta \in \Theta = [0,1]$:

$$\mathbb{P}(\text{Heads} \mid \theta) = \theta.$$

- Data $\mathcal{D} = \{H, H, T, T, T, T, T, H, \ldots, T\}$
  - $n_h$: number of heads
  - $n_t$: number of tails
- **Conditional Independence Assumption**:
  - Conditioned on $\theta$, repeated flips are independent
- **Likelihood model** (Bernoulli Distribution):

$$p(\mathcal{D} \mid \theta) = \theta^{n_h} (1-\theta)^{n_t}$$

  - (probability of getting the flips in the order they were received)

# Coin Flipping: Beta Prior

- **Prior:**

$$\begin{aligned} \theta &\sim \text{Beta}(h,t) \\ p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1} \end{aligned}$$

- **Mean of Beta distribution:**

$$\mathbb{E}\theta = \frac{h}{h+t}$$

- Interpret $h$ and $t$ as the number of heads/tails received in a prior experiment.
  - Then $\mathbb{E}\theta$ is the obvious MLE and plug-in estimate for $\theta$.
- For fixed $\mathbb{E}\theta$, $\mathrm{Var}(\theta)$ decreases as number of flips $n = h + t$ grows.

# Coin Flipping: Posterior

- **Prior:**

$$\begin{aligned} \theta &\sim \text{Beta}(h,t) \\ p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1} \end{aligned}$$

- **Likelihood model**:

$$p(\mathcal{D} \mid \theta) = \theta^{n_h}(1-\theta)^{n_t}$$

- **Posterior density:**

$$\begin{aligned} p(\theta \mid \mathcal{D}) &\propto p(\theta)p(\mathcal{D} \mid \theta) \\ &\propto \theta^{h-1}(1-\theta)^{t-1} \times \theta^{n_h}(1-\theta)^{n_t} \\ &= \theta^{h-1+n_h}(1-\theta)^{t-1+n_t} \end{aligned}$$

# Posterior is Beta

- **Prior:**

$$\begin{aligned} \theta &\sim \textbf{Beta}(h,t) \\ p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1} \end{aligned}$$

- **Posterior density:**

$$p(\theta \mid \mathcal{D}) \propto \theta^{h-1+n_h}(1-\theta)^{t-1+n_t}$$

- So

$$\theta \mid \mathcal{D} \sim \text{Beta}(h+n_h, t+n_t)$$

- It's as though we continued our experiment by adding more flips.

# Conjugate Prior Examples

- A prior is conjugate for a likelihood model if the posterior is in the same "family" as the prior.

1. If prior is a beta distribution, and likelihood model is a Bernoulli distribution, then posterior is a beta distribution.
   - Prior and posterior in the same family $\Longrightarrow$ **Beta is a conjugate prior for Bernoulli**
2. If prior is a Gaussian distribution, and likelihood model is a Gaussian distribution, then posterior is a Gaussian distribution.
   - Prior and posterior in the same family $\Longrightarrow$ **Gaussian is a conjugate prior for Gaussian**

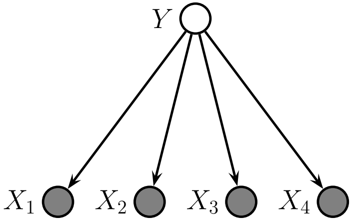
Conjugacy of the prior is really a statement about the prior **family.**

# Conjugate Prior Family

- Let $\pi$ be a family of prior distributions on $\Theta$.
- Let $P$ be likelihood model with parameter space $\Theta$.
- We say that $\pi$ **is conjugate to** $P$ if for any prior in $\pi$, the posterior is always in $\pi$.
- Trivial Example:
  - The family of all probability distributions is conjugate to any likelihood model.
- Every exponential family has a nontrivial conjugate prior family. (KPM Section 9.2)

# Naive Bayes: A Generative Model for Classification

- $\mathcal{X} = \left\{ \left( X_1, X_2, X_3, X_4 \right) \in \{0,1\}^4 \right\}$ $\quad\mathcal{Y} = \{0,1\}$ be a class label.
- Consider the Bayesian network depicted below:




- BN structure implies joint distribution factors as:

$$p(x_1, x_2, x_3, x_4, y) = p(y)p(x_1 \mid y)p(x_2 \mid y)p(x_3 \mid y)p(x_4 \mid y)$$

- Features $X_1, \ldots, X_4$ are independent given the class label $Y$.

KPM Figure 10.2(a).

# Parameterized Expression for Joint Distribution

- **Parameters**:

$$\mathbb{P}(Y=1)=\theta_y \qquad \mathbb{P}(X_i=1 \mid Y=1)=\theta_{i1} \qquad \mathbb{P}(X_i=1 \mid Y=0)=\theta_{i0}$$

- **Joint distribution** is

$$\begin{aligned}
& p(x_1,\ldots x_d,y) \\
= \quad & p(y)\prod_{i=1}^{n} p(x_i \mid y) \\
= \quad & (\theta_y)^y\,(1-\theta_y)^{1-y} \\
& \times \prod_{i=1}^{n} (\theta_{i1})^{yx_i}\,(1-\theta_{i1})^{y(1-x_i)}\,(\theta_{i0})^{(1-y)x_i}\,(1-\theta_{i0})^{(1-y)(1-x_i)}
\end{aligned}$$

# Maximum Likelihood Estimators for Naive Bayes

- Training set $\mathcal{D} = \left\{ \left(x^1, y^1\right), \ldots \left(x^n, y^n\right) \right\}$.
- Obvious "plug-in" estimators for the Naive Bayes model are also MLEs:

$$
\begin{aligned}
\mathbb{P}(Y=1) \quad &\approx \quad \hat{\theta}_y = \frac{1}{n}\sum_{i=1}^{n} 1(y^i = 1) \\
\mathbb{P}(X_i = 1 \mid Y = 1) \quad &\approx \quad \hat{\theta}_{i1} = \frac{\sum_{j=1}^{n} 1(y^j = 1 \text{ and } x_i^j = 1)}{\sum_{j=1}^{n} 1(y^j = 1)} \\
\mathbb{P}(X_i = 1 \mid Y = 0) \quad &= \quad \hat{\theta}_{i0} = \frac{\sum_{j=1}^{n} 1(y^j = 0 \text{ and } x_i^j = 1)}{\sum_{j=1}^{n} 1(y^j = 0)}
\end{aligned}
$$

# Example: SPAM Classification

- Label $Y \in \mathcal{Y} = \{\text{SPAM}, \text{HAM}\}$.
- Features $X_i \in \{0, 1\}$.
- Bag of words representation:

$$X_i = 1(\text{Email contains word "Private\_Jet"})$$

- After parameter estimation, prediction done with

$$p(\text{SPAM}|x) \propto p(\text{SPAM}) \prod_{i=1}^{d} \hat{p}(x_i \mid \text{SPAM}).$$

- Each $\hat{p}(x_i \mid y)$ is the estimated probability that $x_i$ would be observed (or not) in a SPAM message.
- Issue: What if we never see $X_1 = 1$ when $Y = \text{SPAM}$ in $\mathcal{D}$?
  - Then whenever we see $X_1 = 1$, we will predict $p(\text{SPAM} \mid x) = 0$.

# The Zero Count Issue

- If any conditional probabilities $\mathbb{P}(X_i = x_i \mid y)$ get estimated as 0,
    - we'll predict 0 probability for some $y$ whenever $x_i$ is observed.
- This is bad:
    - Never want to predict probability 0 if something is possible.
- Worse: This occurrence is not unusual at all for small sample sizes or rare features.

# Laplace Smoothing

- One traditional fix to the 0 count issue is called **Laplace Smoothing.**
- Idea is to add 1 to every empirical count.
- To estimate $\mathbb{P}(X_i = 1 \mid Y = 1)$, use

$$\hat{\theta}_{i1} = \frac{1 + \sum_{j=1}^{n} 1(y^j = 1 \text{ and } x_i^j = 1)}{1 + \sum_{j=1}^{n} 1(y^j = 1)}.$$

- The added 1 is called a **pseudocount**.
- Like assuming every outcome that can occur was observed at least once.
- Seems to solve the problem – but is there a more principled approach?

# Bayesian Naive Bayes

- **Parameters**:

$$\mathbb{P}(Y=1)=\theta_y \qquad \mathbb{P}(X_i=1 \mid Y=1)=\theta_{i1} \qquad \mathbb{P}(X_i=1 \mid Y=0)=\theta_{i0}$$

- Put a Beta prior distribution on each parameter.
- **Option 1:** Use posterior mean as point estimate for each parameter, then continue as before.
  - Laplace smoothing is a special case, in which priors are all $\text{Beta}(1,1)$.
- **Option 2: Go full Bayesian**.
  - No parameter estimates. Base everything on posterior $\theta \mid \mathcal{D}$.
- Predict with the predictive distribution:

$$y \mid x, \mathcal{D}$$

  - Recall, this is integrating out the parameter $\theta$ w.r.t. the posterior distribution.