Bayesian Estimation & Conditional Models

CMPUT 296: Basics of Machine Learning

Textbook §5.2-5.3

Logistics

- **Assignment 1 is due TODAY at 11:59pm**
- Assignment 2 will be available tomorrow
- Thought Questions #2: Ch.4-6 are due **October 8**

Recap: Parameter Estimation

- We are usually interested in predicting the value of unseen data X_{n+1} based on **training data** $\mathscr{D} = \{x_1, \ldots, x_n\}$
- Instead, we will want to choose a **model** $\hat{f}$ from a **hypothesis space** $\mathscr{F}$
  - Where the data are generated according to some "true" model f^*
  - $\mathscr{F}$ is often **parametric**: its members identified by **parameter** values
  - So choosing a **model** is equivalent to choosing a set of **parameter** values
- Two approaches to parameter estimation:

$$f_{\text{MAP}} = \arg\max_{f \in \mathscr{F}} p(f \mid \mathscr{D}) = \arg\max_{f \in \mathscr{F}} p(\mathscr{D} \mid f) p(f)$$

$$f_{\text{MLE}} = \arg\max_{f \in \mathscr{F}} p(\mathscr{D} \mid f)$$

Outline

1. Recap & Logistics
2. Bayesian Estimation
3. Conditional Distributions

Point Estimates

- Suppose we have a dataset $\mathscr{D}$ that was generated by a model $f(\cdot \mid \theta^*) \in \mathscr{F} = \{f(\cdot \mid \theta) \mid \theta \in \mathbb{R}\}$
- A **point estimate** asks: What is the **best single guess** for the parameter?
    - **MLE:** $\arg\max_{\theta} p(\mathscr{D} \mid \theta)$
    - **MAP:** $\arg\max_{\theta} p(\theta \mid \mathscr{D})$
    - Estimate of θ that has lowest **expected error**?

Bayes Estimator

The **Bayes estimator** is the point estimate that **minimizes the posterior risk** $c(\hat{\theta})$, where

$$c(\hat{\theta}) = \int_{\mathcal{F}} \ell(\theta, \hat{\theta}) p(\theta \mid \mathcal{D}) \, d\theta$$

The **loss** $\ell(\theta, \hat{\theta})$ expresses how "wrong" we are if we estimate $\hat{\theta}$ when the true answer is θ.

Bayes Estimator for Squared Loss

When $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$:

$$c(\hat{\theta}) = \int_{\mathscr{F}} (\theta - \hat{\theta})^2 p(\theta \mid \mathscr{D})\, d\theta$$

$$\iff \frac{\partial}{\partial \hat{\theta}} c(\hat{\theta}) = \frac{\partial}{\partial \hat{\theta}} \int_{\mathscr{F}} (\theta - \hat{\theta})^2 p(\theta \mid \mathscr{D})\, d\theta$$

$$= \int_{\mathscr{F}} \frac{\partial}{\partial \hat{\theta}} (\theta - \hat{\theta})^2 p(\theta \mid \mathscr{D})\, d\theta$$

$$= 2 \int_{\mathscr{F}} (\hat{\theta} - \theta) p(\theta \mid \mathscr{D})\, d\theta$$

$$= 2\hat{\theta} \int_{\mathscr{F}} p(\theta \mid \mathscr{D})\, d\theta - 2 \int_{\mathscr{F}} \theta p(\theta \mid \mathscr{D})\, d\theta$$

$$= 2\hat{\theta} - 2 \int_{\mathscr{F}} \theta p(\theta \mid \mathscr{D})\, d\theta \qquad \blacksquare$$

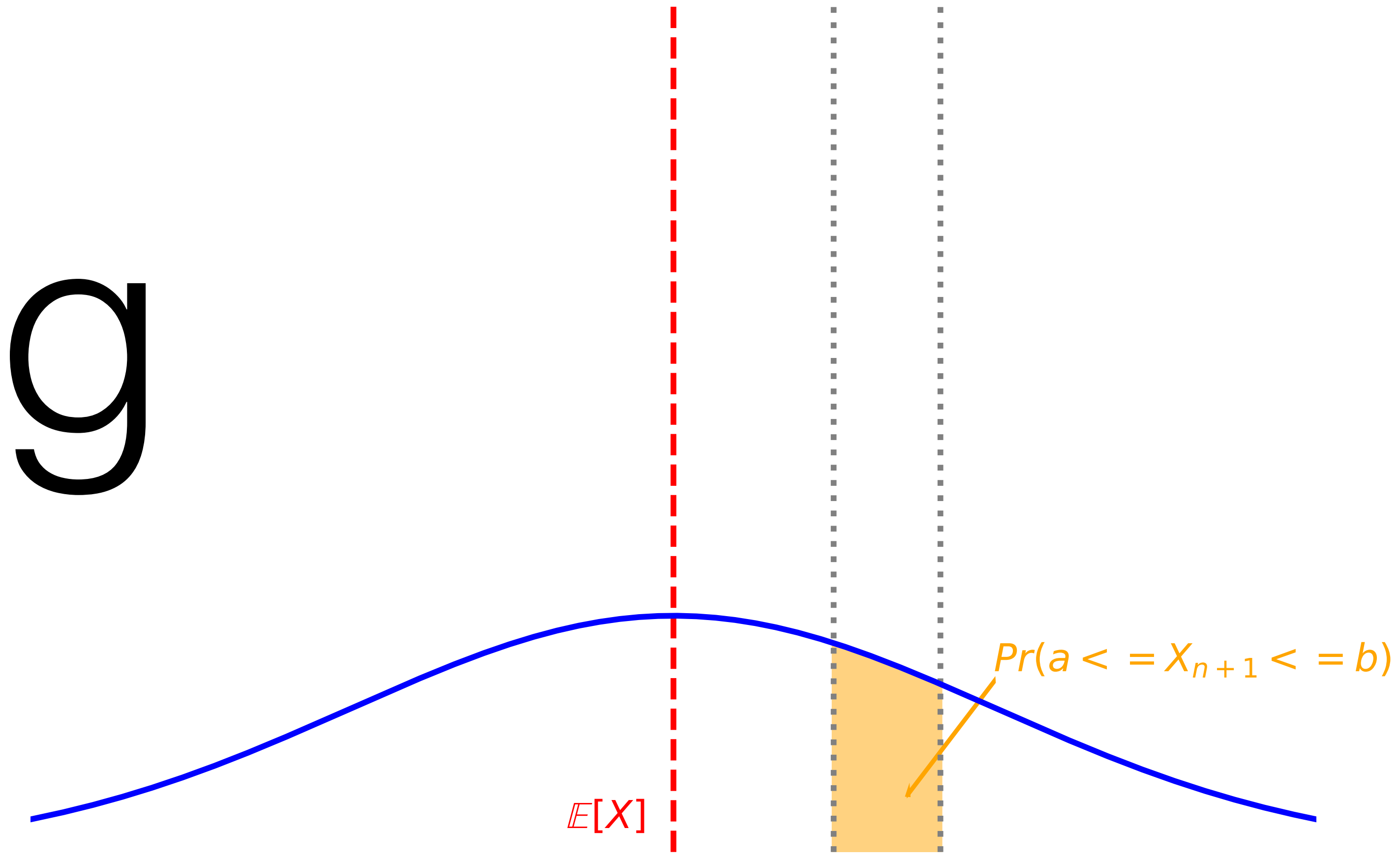
$$\frac{\partial}{\partial \hat{\theta}} \ell(\hat{\theta}) = 0$$

$$\iff 0 = 2\hat{\theta} - 2 \int_{\mathscr{F}} \theta p(\theta \mid \mathscr{D})\, d\theta$$

$$\iff \hat{\theta} = \int_{\mathscr{F}} \theta p(\theta \mid \mathscr{D})\, d\theta$$

$$= \mathbb{E}[\theta \mid \mathscr{D}] \qquad \blacksquare$$

Bayesian Reasoning



Question: How can we answer $\Pr(a \leq X_{n+1} \leq b)$?

1. **MLE:** $F\left(b \mid \theta_{\text{MLE}}\right) - F\left(a \mid \theta_{\text{MLE}}\right)$
2. **MAP:** $F\left(b \mid \theta_{\text{MAP}}\right) - F\left(a \mid \theta_{\text{MAP}}\right)$
3. **Bayes optimal estimator:** $F\left(b \mid \theta_{\text{B}}\right) - F\left(a \mid \theta_{\text{B}}\right)$
4. **Bayesian:** $\displaystyle\int_{\mathscr{F}} \left[F\left(b \mid \theta\right) - F\left(a \mid \theta\right)\right] p(\theta \mid \mathscr{D})\, d\theta$

$$= \mathbb{E}\left[F(b \mid \theta) - F(a \mid \theta) \,\middle|\, \mathscr{D}\right]$$

Question: Does this use of θ_B make sense? Why?

Example: Poisson Data with Gamma Prior

Example: Suppose dataset $\mathscr{D} = \{8,4,5,9,5,2\}$ is drawn i.i.d. from an unknown Poisson distribution, with parameter λ_0. We have a Gamma prior over λ; that is,

prior $p(\lambda) = \dfrac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)}$

and likelihood $p(\mathscr{D} \mid \lambda) = \dfrac{\lambda^{\left(\sum_{i=1}^n x_i\right)} e^{-n\lambda}}{\prod_{i=1}^n x_i!}$.

To compute the Bayes estimator, we will need the full **posterior** $p(\lambda \mid \mathscr{D})$, and not just the joint $p(\mathscr{D} \mid \lambda)p(\lambda) \propto p(\lambda \mid \mathscr{D})$. (**Why?**)

That means we need to compute the **model evidence** as well.

Poisson Data with Gamma Prior 2

$$p(\mathscr{D}) = \int_0^\infty p(\mathscr{D} \mid \lambda) p(\lambda)\, d\lambda$$

$$= \int_0^\infty \frac{\lambda^{\left(\sum_{i=1}^n x_i\right)} e^{-n\lambda}}{\prod_{i=1}^n x_i!} \cdot \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)}\, d\lambda$$

$$= \frac{\Gamma\left(k + \sum_{i=1}^n x_i\right)}{\theta^k \Gamma(k) \prod_{i=1}^n x_i! \left(n + \frac{1}{\theta}\right)^{\left(k + \sum_{i=1}^n x_i\right)}}$$

$$p(\lambda \mid \mathscr{D}) = \frac{\lambda^{\left(\sum_{i=1}^n x_i\right)} e^{-n\lambda}}{\cancel{\prod_{i=1}^n x_i!}} \cdot \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\cancel{\theta^k \Gamma(k)}} \cdot \frac{\cancel{\theta^k \Gamma(k) \prod_{i=1}^n x_i!} \left(n + \frac{1}{\theta}\right)^{\left(k + \sum_{i=1}^n x_i\right)}}{\Gamma\left(k + \sum_{i=1}^n x_i\right)}$$

$$= \frac{\lambda^{\left((k + \sum_{i=1}^n x_i) - 1\right)} \cdot e^{-\lambda(n + 1/\theta)} \cdot \left(n + \frac{1}{\theta}\right)^{\left(k + \sum_{i=1}^n x_i\right)}}{\Gamma\left(k + \sum_{i=1}^n x_i\right)}$$

$$= \frac{\lambda^{\left((k + \sum_{i=1}^n x_i) - 1\right)} \cdot e^{-\lambda(n + 1/\theta)}}{\left(\frac{1}{n + \frac{1}{\theta}}\right)^{\left(k + \sum_{i=1}^n x_i\right)} \cdot \Gamma\left(k + \sum_{i=1}^n x_i\right)}$$

i.e., a **Gamma distribution** with

$k' = k + \sum_{i=1}^{n} x_i$ and $\theta' = \frac{\theta}{n\theta + 1} = \frac{1}{n + 1/\theta}$

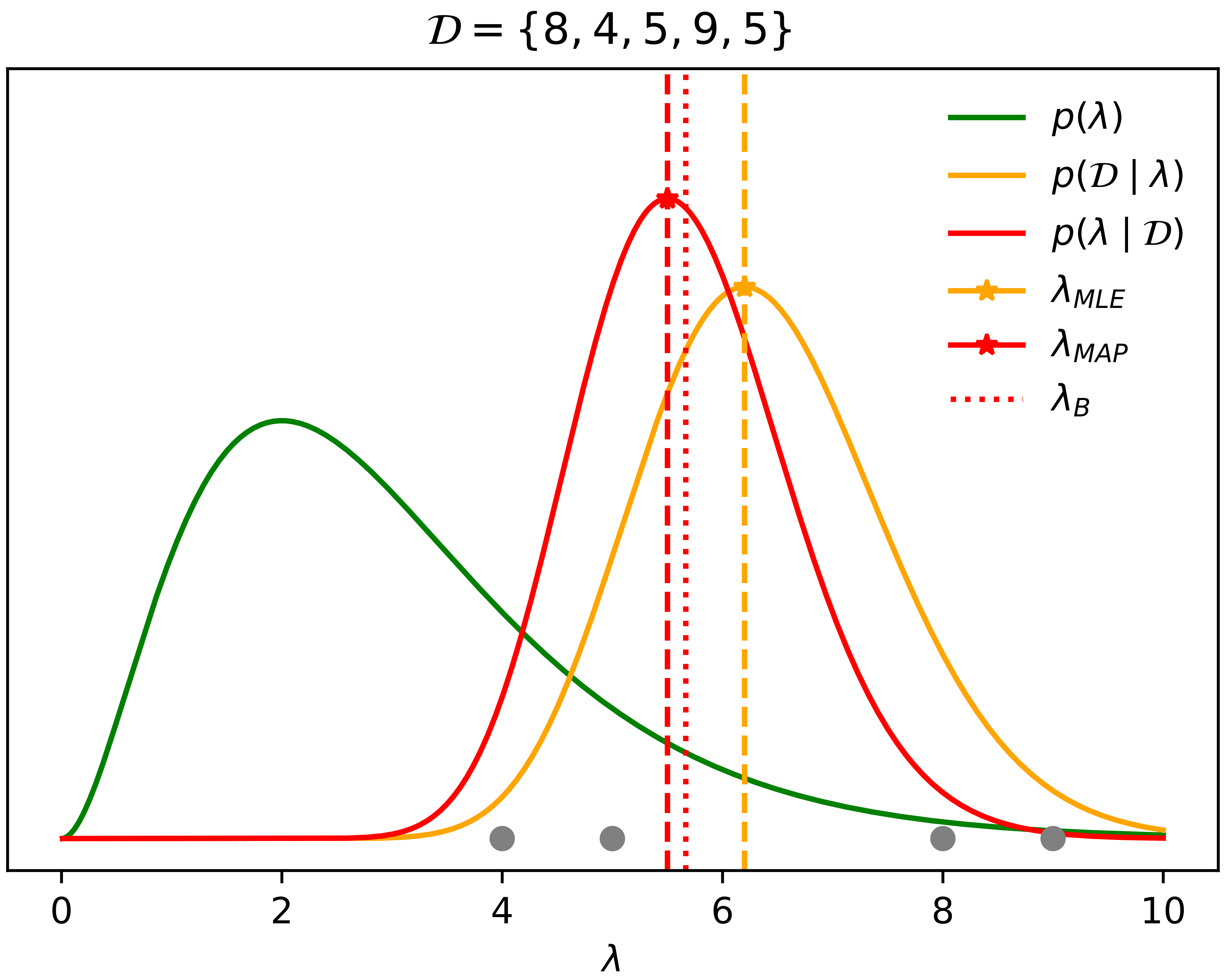
Fun fact: $\int_0^\infty x^{\alpha - 1} e^{-\beta x}\, dx = \frac{\Gamma(\alpha)}{\beta^\alpha}$

Conjugate Priors

- **Gamma** is a conjugate prior for the **Poisson** distribution
  - Starting from prior $\mathrm{Gamma}(k, \theta)$ and assuming a Poission likelihood, after seeing $x_1, \ldots, x_n$, posterior is $\mathrm{Gamma}\left(k + \sum_{i=1}^{n} s_i, \frac{1}{n + 1/\theta}\right)$
- Similarly, **Beta** is a conjugate prior for the **Binomial** distribution
  - Starting from prior $\mathrm{Beta}(a, b)$ and assuming a Binomial likelihood, after seeing n_T successes and n_F failures, posterior is $\mathrm{Beta}\left(a + n_T, b + n_F\right)$.

Poisson Data Example: Updating

Example: Suppose dataset $\mathscr{D} = \{8,4,5,9,5,2\}$ is drawn i.i.d. from an unknown Poisson distribution, with parameter λ_0 with prior $\mathrm{Gamma}(k = 3,\ \theta = 1)$:



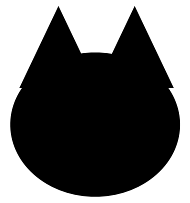
Advanced: Bayesian Methods with Nonconjugate Priors

- Conjugate priors are very convenient, and you should use them wherever possible
- **Question:** What can we do if the priors are **not conjugate**?
    - In general, the integral to compute $p(\mathscr{D})$ will be **intractable**
- The usual technique is variants of **Monte Carlo sampling**
    - *Basic idea:* Generate a bunch of random samples $\theta_1, \ldots, \theta_R \overset{i.i.d}{\sim} p(\theta \mid \mathscr{D})$

$$\mathbb{E}\left[\frac{1}{R}\sum_{r=1}^{R} f(\theta_r)\right] = \mathbb{E}[f(\theta)]$$

    - There are multiple techniques for generating random samples from an **unnormalized distribution** $q(\theta) \propto p(\theta)$
    - We can use one of these techniques to sample from $p(\mathscr{D} \mid \theta)p(\theta) \propto p(\theta \mid \mathscr{D})$

Conditional Models

Question: How can we ask "With what probability is  an image of a cat" using the models we have been learning?

- We want a **different distribution** depending on the **image**, and
- We want to be able to ask about **multiple images**
- Given an image described by pixels $\mathbf{x}$, we want something like

$$\Pr(Y = cat \mid X = \mathbf{x})$$

- Our models can be parameterized families of **conditional distributions**:

$$\mathcal{F} = \{f(y, x; \theta) \mid \theta \in \mathbb{R}^k\}$$

MLE, MAP, Bayesian Prediction for Conditional Models

Given a hypothesis space $\mathscr{F} = \{p(\cdot \mid \cdot, \theta) \mid \theta \in \mathbb{R}\}$ and a dataset $\mathscr{D} = \left\{(x_i, y_i)\right\}_{i=1}^{n}$ of observed **features** x_i and their corresponding **labels** y_i:

MLE: $p(y \mid x) = p(y \mid x, \theta_{\text{MLE}})$ where $\theta_{\text{MLE}} = \arg\max_{\theta} \sum_i \ln p(y_i \mid x_i, \theta)$

MAP: $p(y \mid x) = p(y \mid x, \theta_{\text{MLE}})$ where $\theta_{\text{MAP}} = \arg\max_{\theta} \ln p(\theta) + \sum_i \ln p(y_i \mid x_i, \theta)$

Bayesian: $p(y \mid x) = \int_{\mathscr{F}} p(y \mid x, \theta) p(\theta \mid \mathscr{D}) \, d\theta$

Question: What happened to θ_B?

Summary

- The MAP, MLE, and Bayes estimators for a model parameter are all **point estimates**
- MAP and MLE can be computed without computing $p(\mathscr{D})$
- **Conjugate priors** make it possible to perform Bayesian updates analytically
    - But many models don't have conjugate priors
- **Conditional models** allow us to change predictions based on **input features**
    - **MAP, MLE:** simply plug the features into the point estimate model
    - **Bayesian:** take posterior expected prediction over all models