

# MAP estimation

Nishant Mehta

Lecture 12

# The perils of maximum likelihood estimation



# wins  $\uparrow = n_1$

$\uparrow$   
# wins =  $n_0$

Consider betting on two top chess players

Suppose the true probability that Kasparov wins is  $\theta = 0.5$  (unknown to us), so Kasparov and Karpov are evenly matched

Instead, we have observed two games. Kasparov won both games.

What is the MLE?  $\hat{\theta}_{MLE} = \frac{n_1}{n} = \frac{2}{2} = 1$

So, how much money should we bet on the next game? *all your money*

# Expected cross-entropy loss of $\hat{\theta}_{MLE}$

What is risk under cross-entropy loss when true parameter is  $\theta = 0.5$  and our estimate is  $\hat{\theta} = 1$ ?

$$\begin{aligned} E[\ell(Y, 1) \mid \theta = 0.5] &= E\left[-Y \log_{\hat{\theta}} 1 - (1-Y) \log_{1-\hat{\theta}} (1-1)\right] \\ &= E\left[-(1-Y) \log 0\right] \\ &= \frac{1}{2} \left( \underbrace{- (1-1) \log 0}_{0 \log 0} \right) + \frac{1}{2} \left( - (1-0) \log 0 \right) \\ &= \frac{1}{2} \cdot (-1) \cdot (-\infty) \\ &= +\infty \end{aligned}$$

# Intuition: Imaginary examples

Suppose we imagine that we have extra examples, one example for each class:

$$\tilde{n}_1 = n_1 + 1 \qquad \tilde{n}_0 = n_0 + 1$$

This is called *add-one smoothing*, a special case of a more general technique called *additive smoothing*.

Why might this be a good idea?

What happens to the MLE when we include these imaginary examples?

$$\hat{\theta} =$$

\* add-one smoothing is also called Laplace's rule of succession

# Intuition: Imaginary examples

Suppose we imagine that we have extra examples, one example for each class:

$$\tilde{n}_1 = n_1 + 1 \qquad \tilde{n}_0 = n_0 + 1$$

This is called *add-one smoothing*, a special case of a more general technique called *additive smoothing*.

Why might this be a good idea?

What happens to the MLE when we include these imaginary examples?

$$\hat{\theta} = \frac{\tilde{n}_1}{\tilde{n}_1 + \tilde{n}_0} = \frac{n_1 + 1}{n_1 + n_0 + 2}$$

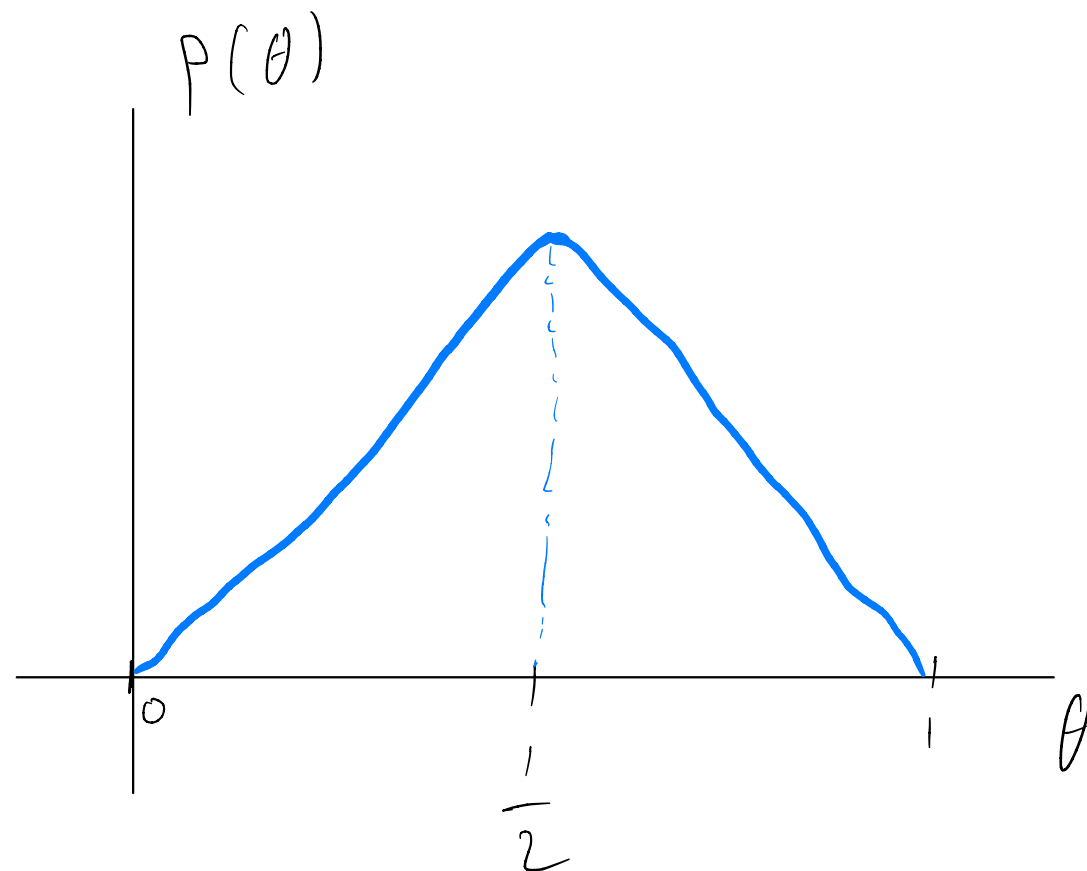
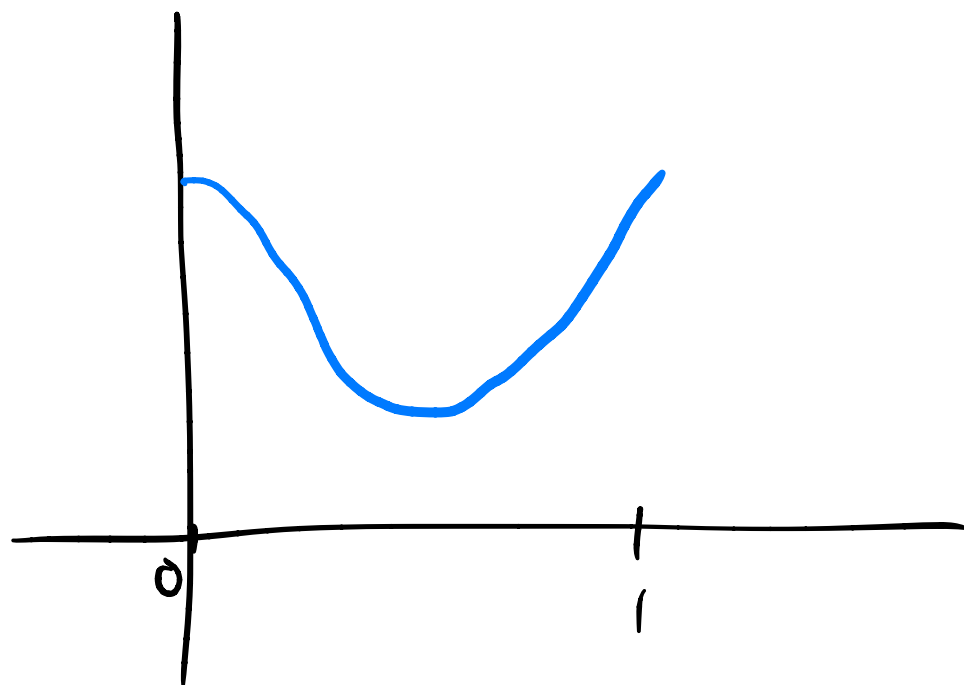
\* add-one smoothing is also called Laplace's rule of succession

# Prior distribution

*Prior distribution*  $P(\theta)$

Indicates our probability of belief that  $\theta$  is the true parameter, prior to seeing any evidence at all

Example: “probably” fair coin



# Posterior distribution

From Bayes rule, we have

$$P(\theta | D) =$$

This quantity is our probability of belief  $\theta$  is the true parameter, *a posteriori of the data*.

We call  $\theta \mapsto P(\theta | D)$  the *posterior distribution* over  $\Theta$

# Posterior distribution

Bernoulli  
 $\Theta = [0, 1]$

From Bayes rule, we have

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)} = \frac{P(D | \theta)P(\theta)}{\int_{\Theta} P(D | \theta)P(\theta)d\theta}$$

This quantity is our probability of belief that  $\theta$  is the true parameter, *a posteriori of the data*.

We call  $\theta \mapsto P(\theta | D)$  the *posterior distribution* over  $\Theta$



# Maximum A Posteriori estimate

From Bayes rule, we have

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)} = \frac{P(D | \theta)P(\theta)}{\int_{\Theta} P(D | \theta)P(\theta)d\theta}$$

This quantity is our probability of belief that  $\theta$  is the true parameter, *a posteriori of the data*.

We call  $\theta \mapsto P(\theta | D)$  the *posterior distribution* over  $\Theta$

The *Maximum a Posteriori estimate (MAP estimate)* of  $\theta$  is

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} P(\theta | D) = \arg \max_{\theta} \frac{P(D | \theta)P(\theta)}{P(D)} \leftarrow \text{ignore}$$

# Maximum A Posteriori estimate

From Bayes rule, we have

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)} = \frac{P(D | \theta)P(\theta)}{\int_{\Theta} P(D | \theta)P(\theta)d\theta}$$

This quantity is our probability of belief that  $\theta$  is the true parameter, *a posteriori of the data*.

We call  $\theta \mapsto P(\theta | D)$  the *posterior distribution* over  $\Theta$

The *Maximum a Posteriori estimate (MAP estimate)* of  $\theta$  is

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \arg \max_{\theta} P(\theta | D) = \arg \max_{\theta} \frac{P(D | \theta)P(\theta)}{P(D)} \leftarrow \text{ignore} \\ &= \arg \max_{\theta} P(D | \theta)P(\theta)\end{aligned}$$

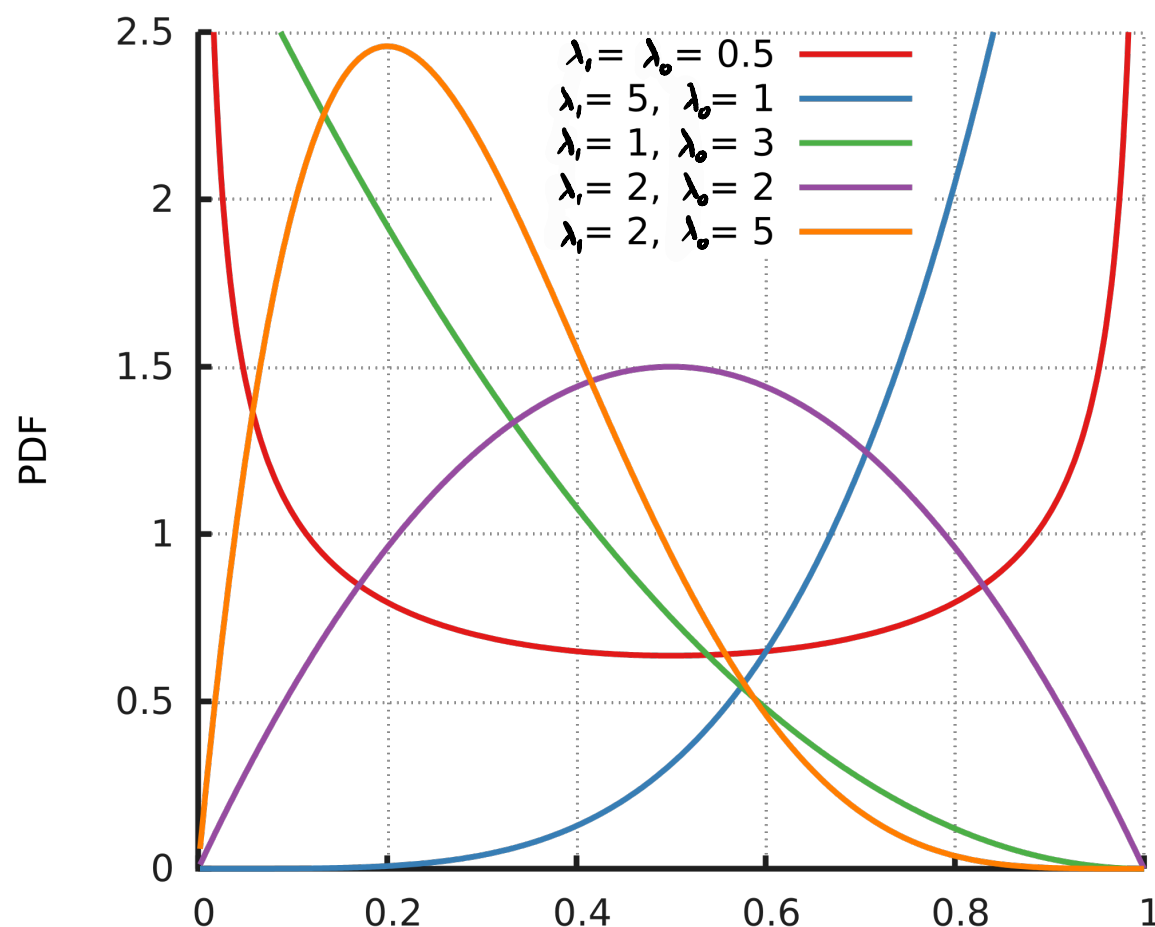
# Beta prior distribution

Suppose the examples are drawn i.i.d. from a Bernoulli distribution

A common choice of prior distribution is the *Beta distribution*

$$P(\theta) = \text{Beta}(\lambda_1, \lambda_0) = \frac{\theta^{\lambda_1-1}(1-\theta)^{\lambda_0-1}}{B(\lambda_1, \lambda_0)}$$

← normalization constant



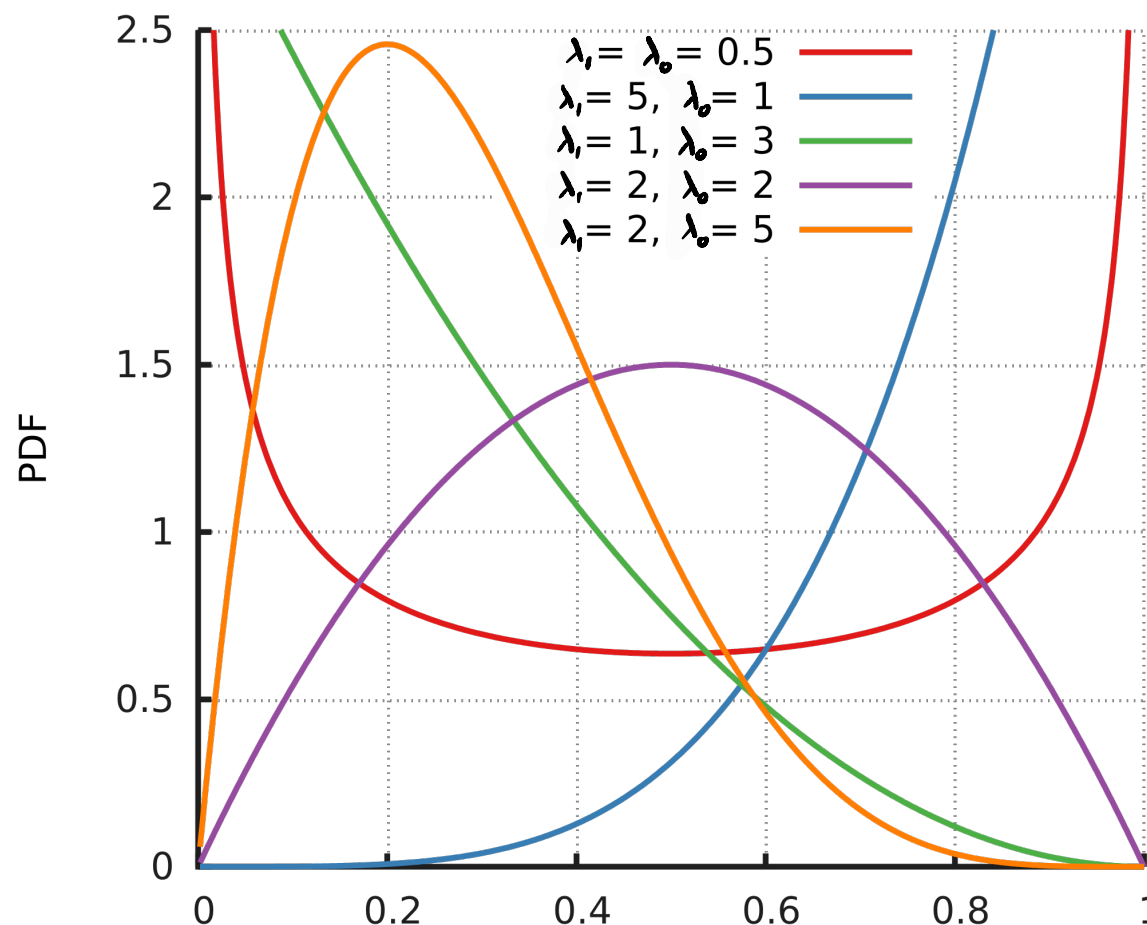
# Beta prior distribution

Suppose the examples are drawn i.i.d. from a Bernoulli distribution

A common choice of prior distribution is the *Beta distribution*

$$P(\theta) = \text{Beta}(\lambda_1, \lambda_0) = \frac{\theta^{\lambda_1-1}(1-\theta)^{\lambda_0-1}}{B(\lambda_1, \lambda_0)}$$

← normalization constant



If  $\lambda_1$  and  $\lambda_0$  are positive integers, then

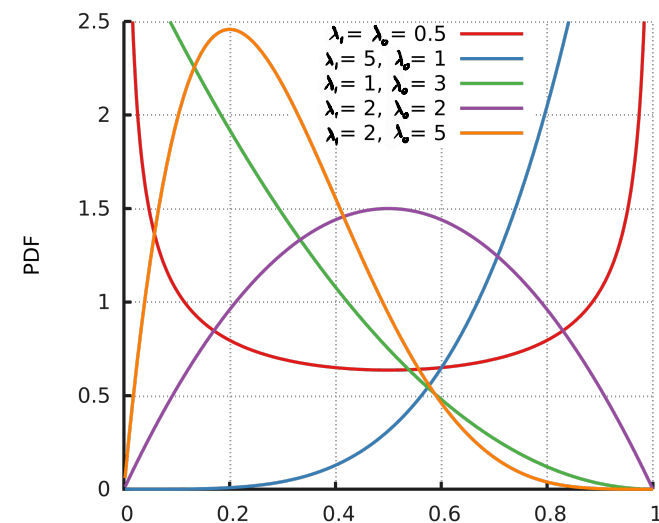
$$B(\lambda_1, \lambda_2) = \frac{(\lambda_1 - 1)!(\lambda_0 - 1)!}{(\lambda_1 + \lambda_0 - 1)!}$$

# Beta prior distribution

Suppose the examples are drawn i.i.d. from a Bernoulli distribution

A common choice of prior distribution is the *Beta distribution*

$$P(\theta) = \text{Beta}(\lambda_1, \lambda_0) = \frac{\theta^{\lambda_1-1}(1-\theta)^{\lambda_0-1}}{B(\lambda_1, \lambda_0)}$$



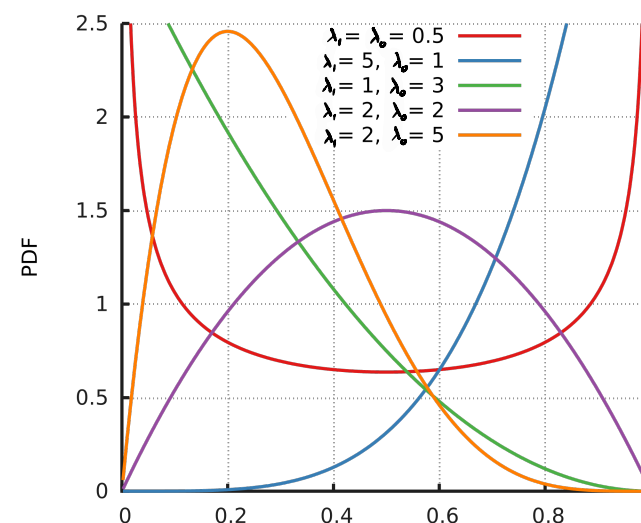
What is the MAP estimate when using a Beta prior?

# Beta prior distribution

Suppose the examples are drawn i.i.d. from a Bernoulli distribution

A common choice of prior distribution is the *Beta distribution*

$$P(\theta) = \text{Beta}(\lambda_1, \lambda_0) = \frac{\theta^{\lambda_1-1}(1-\theta)^{\lambda_0-1}}{B(\lambda_1, \lambda_0)}$$



What is the MAP estimate when using a Beta prior?

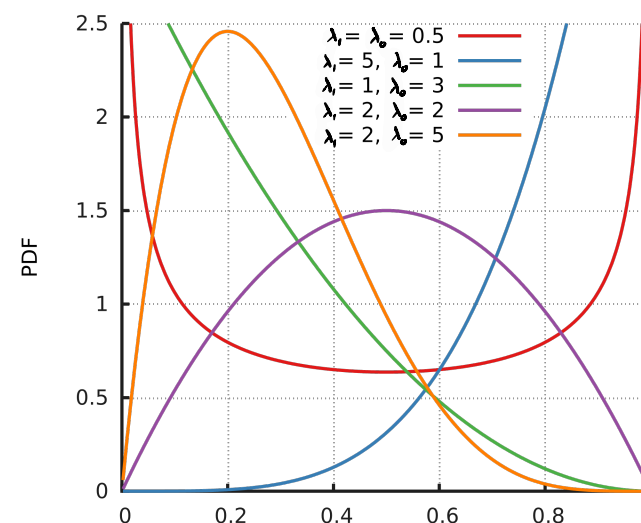
$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \arg \max_{\theta \in [0,1]} P(D | \theta)P(\theta) = \arg \max_{\theta \in [0,1]} \theta^{n_1}(1-\theta)^{n_0} \theta^{\lambda_1-1}(1-\theta)^{\lambda_0-1} \\ &= \arg \max_{\theta \in [0,1]} \theta^{n_1+\lambda_1-1}(1-\theta)^{n_0+\lambda_0-1}\end{aligned}$$

# Beta prior distribution

Suppose the examples are drawn i.i.d. from a Bernoulli distribution

A common choice of prior distribution is the *Beta distribution*

$$P(\theta) = \text{Beta}(\lambda_1, \lambda_0) = \frac{\theta^{\lambda_1-1}(1-\theta)^{\lambda_0-1}}{B(\lambda_1, \lambda_0)}$$



What is the MAP estimate when using a Beta prior?

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta \in [0,1]} P(D | \theta) P(\theta) = \arg \max_{\theta \in [0,1]} \theta^{n_1} (1-\theta)^{n_0} \theta^{\lambda_1-1} (1-\theta)^{\lambda_0-1}$$

$$= \arg \max_{\theta \in [0,1]} \theta^{n_1 + \lambda_1 - 1} (1-\theta)^{n_0 + \lambda_0 - 1}$$

$$\tilde{n}_1 = n_1 + \lambda_1 - 1$$

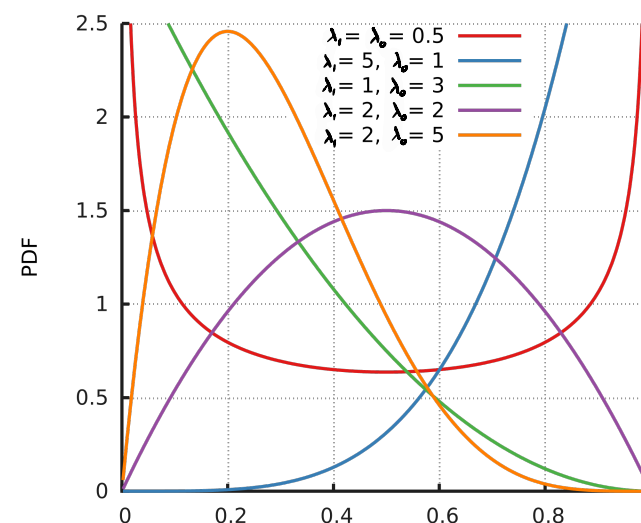
$$\tilde{n}_0 = n_0 + \lambda_0 - 1$$

# Beta prior distribution

Suppose the examples are drawn i.i.d. from a Bernoulli distribution

A common choice of prior distribution is the *Beta distribution*

$$P(\theta) = \text{Beta}(\lambda_1, \lambda_0) = \frac{\theta^{\lambda_1-1}(1-\theta)^{\lambda_0-1}}{B(\lambda_1, \lambda_0)}$$



What is the MAP estimate when using a Beta prior?

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta \in [0,1]} P(D | \theta) P(\theta) = \arg \max_{\theta \in [0,1]} \theta^{n_1} (1-\theta)^{n_0} \theta^{\lambda_1-1} (1-\theta)^{\lambda_0-1}$$

$$= \arg \max_{\theta \in [0,1]} \theta^{n_1 + \lambda_1 - 1} (1-\theta)^{n_0 + \lambda_0 - 1} = \arg \max_{\theta \in [0,1]} \theta^{\tilde{n}_1} (1-\theta)^{\tilde{n}_0}$$

$$\tilde{n}_1 = n_1 + \lambda_1 - 1$$

$$\tilde{n}_0 = n_0 + \lambda_0 - 1$$

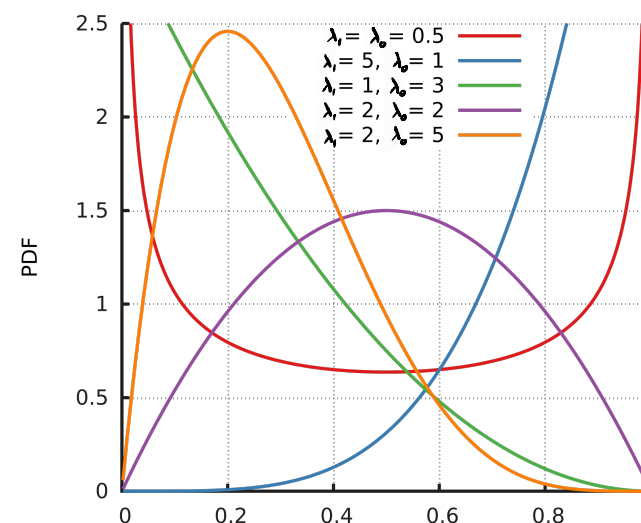


# Beta prior distribution

Suppose the examples are drawn i.i.d. from a Bernoulli distribution

A common choice of prior distribution is the *Beta distribution*

$$P(\theta) = \text{Beta}(\lambda_1, \lambda_0) = \frac{\theta^{\lambda_1-1}(1-\theta)^{\lambda_0-1}}{B(\lambda_1, \lambda_0)}$$



What is the MAP estimate when using a Beta prior?

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta \in [0,1]} P(D | \theta) P(\theta) = \arg \max_{\theta \in [0,1]} \theta^{n_1} (1-\theta)^{n_0} \theta^{\lambda_1-1} (1-\theta)^{\lambda_0-1}$$

$$= \arg \max_{\theta \in [0,1]} \theta^{n_1 + \lambda_1 - 1} (1-\theta)^{n_0 + \lambda_0 - 1} = \arg \max_{\theta \in [0,1]} \theta^{\tilde{n}_1} (1-\theta)^{\tilde{n}_0} = \frac{\tilde{n}_1}{\tilde{n}_1 + \tilde{n}_0}$$

$$\tilde{n}_1 = n_1 + \lambda_1 - 1$$

$$\tilde{n}_0 = n_0 + \lambda_0 - 1$$

# Conjugate prior

Note that the form of the posterior is again a Beta distribution

When the prior and posterior distributions have the same form, the prior is known as a *conjugate prior*

Benefits of a conjugate prior:

- Posterior is easy to interpret (if prior was easy to interpret)

- Computationally friendly (updating is easier)

# Additive smoothing

In *additive smoothing*, we add  $c$  imaginary positive examples and  $c$  imaginary negative examples, for parameter  $c > 0$

How should we set  $\lambda_1$  and  $\lambda_0$  to get additive smoothing?

$$\tilde{n}_1 = n_1 + c \iff \lambda_1 - 1 = c \iff \lambda_1 = c + 1$$

$$\tilde{n}_0 = n_0 + c \iff \lambda_0 = c + 1$$

$$\hat{\Theta}_{MAP} = \frac{\tilde{n}_1}{\tilde{n}_1 + \tilde{n}_0} = \frac{n_1 + c}{n + 2c} \quad n = n_1 + n_0$$

# MAP estimation $\longleftrightarrow$ regularized training

Just like with the MLE, we can write **MAP estimation** as the minimization of training error under cross-entropy loss...

... but now, we also have regularization!

$$\begin{aligned} \max_{\theta} \log (P(\mathcal{D}|\theta) P(\theta)) &\equiv \max_{\theta} \left[ \log P(\mathcal{D}|\theta) + \log P(\theta) \right] \\ &\equiv \min_{\theta} \left[ -\log P(\mathcal{D}|\theta) + \log \frac{1}{P(\theta)} \right] \\ &\equiv \min_{\theta} \left[ \underbrace{\sum_{j=1}^n \ell_{\log} (y_j, \theta)}_{\text{loss part}} + \underbrace{(\lambda_1 - 1) \log \frac{1}{\theta} + (\lambda_0 - 1) \log \frac{1}{1-\theta}}_{R(\theta)} \right] \end{aligned}$$

# Multiclass - One-hot encoding

In the multiclass case with  $K$  classes, there are two common choices of representation of the label

1) Standard representation:

$$Y \in \{1, 2, \dots, K\}$$

2) One-hot encoding (also called one-of- $K$  encoding):

$$Y \in \{0, 1\}^K \text{ with } Y_j = 1 \text{ if label is } j \text{ and } Y_j = 0 \text{ otherwise}$$

# MLE - Extension to multinoulli distribution

Suppose we have  $K$  classes. We use parameter vector  $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_K \end{pmatrix}$  satisfying  $\theta_j \in [0, 1]$  and  $\sum_{j=1}^K \theta_j = 1$

Log likelihood for *Multinoulli* (or *categorical*) distribution

$$\log P(Y = y) = \begin{cases} \log \theta_y & \text{(standard representation)} \\ \log \prod_{j=1}^K \theta_j^{y_j} = \sum_{j=1}^K y_j \log \theta_j & \text{(one-hot encoding)} \end{cases}$$

# MLE - Extension to multinoulli distribution

Suppose we have  $K$  classes. We use parameter vector  $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_K \end{pmatrix}$  satisfying  $\theta_j \in [0, 1]$  and  $\sum_{j=1}^K \theta_j = 1$

Log likelihood for *Multinoulli* (or *categorical*) distribution

$$\log P(Y = y) = \begin{cases} \log \theta_y & \text{(standard representation)} \\ \log \prod_{j=1}^K \theta_j^{y_j} = \sum_{j=1}^K y_j \log \theta_j & \text{(one-hot encoding)} \end{cases}$$

Multiclass cross-entropy loss

$$\log P(Y = y) = \begin{cases} -\log \theta_y & \text{(standard representation)} \\ \sum_{j=1}^K -y_j \log \theta_j & \text{(one-hot encoding)} \end{cases}$$

What is the MLE?

# MLE - Extension to multinoulli distribution

Suppose we have  $K$  classes. We use parameter vector  $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_K \end{pmatrix}$  satisfying  $\theta_j \in [0, 1]$  and  $\sum_{j=1}^K \theta_j = 1$

Log likelihood for *Multinoulli* (or *categorical*) distribution

$$\log P(Y = y) = \begin{cases} \log \theta_y & \text{(standard representation)} \\ \log \prod_{j=1}^K \theta_j^{y_j} = \sum_{j=1}^K y_j \log \theta_j & \text{(one-hot encoding)} \end{cases}$$

Multiclass cross-entropy loss

$$\log P(Y = y) = \begin{cases} -\log \theta_y & \text{(standard representation)} \\ \sum_{j=1}^K -y_j \log \theta_j & \text{(one-hot encoding)} \end{cases}$$

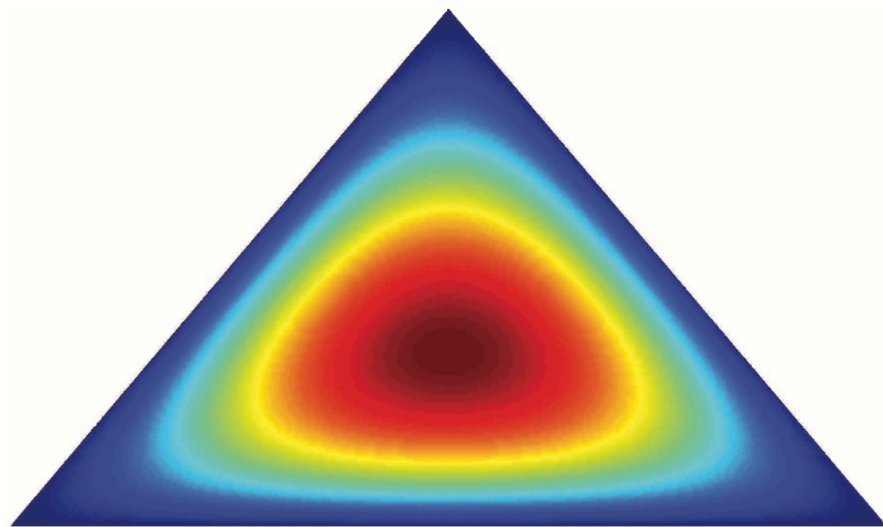
What is the MLE?  $\hat{\theta}_j = \frac{n_j}{n}$  ← number of examples with label  $j$



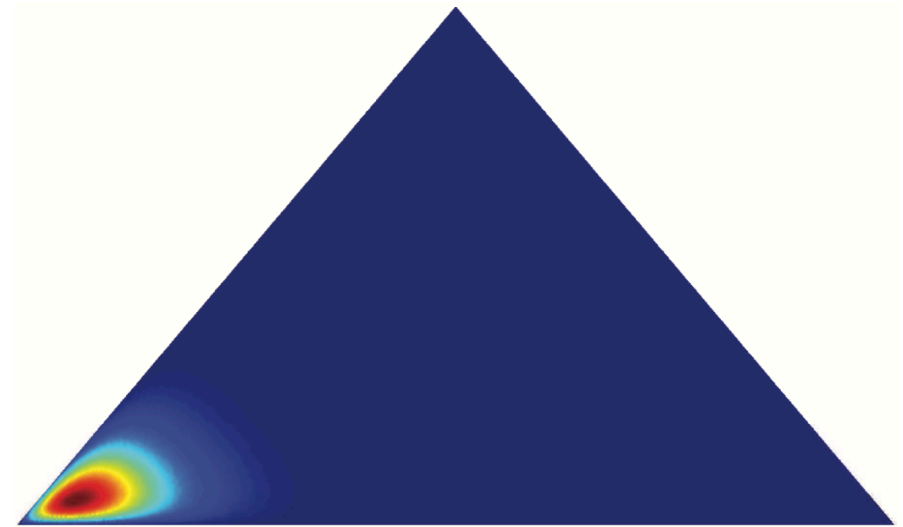
# MAP - Extension to multinoulli distribution

- Conjugate prior? *Dirichlet distribution*

- $P(\theta | \alpha) = \frac{1}{B(\alpha)} \prod_{j=1}^K \theta_j^{\alpha_j - 1}$  if  $\theta$  is probability vector (zero otherwise)



$$\alpha = (2, 2, 2)$$

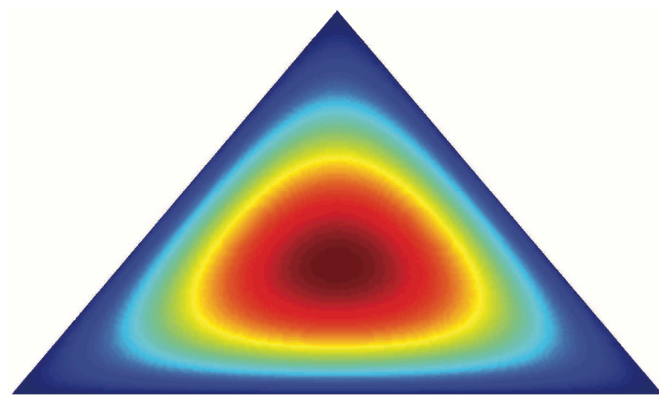


$$\alpha = (20, 2, 2)$$

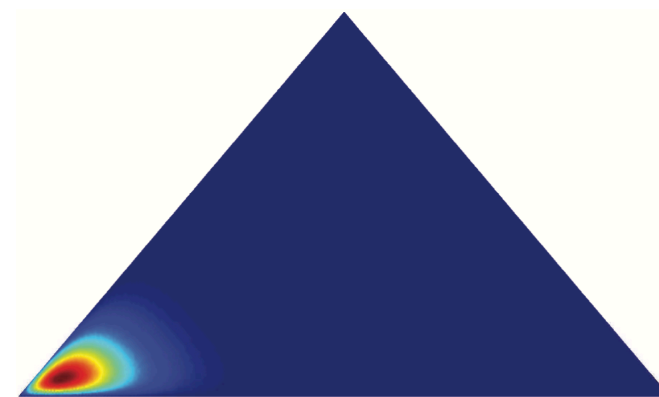
# MAP - Extension to multinoulli distribution

- Conjugate prior? *Dirichlet distribution*

- $$P(\theta | \alpha) = \frac{1}{B(\alpha)} \prod_{j=1}^K \theta_j^{\alpha_j - 1}$$
 if  $\theta$  is probability vector (zero otherwise)



$$\alpha = (2, 2, 2)$$



$$\alpha = (20, 2, 2)$$

- If we have  $N_j$  occurrences of class  $j$ , then posterior distribution is

$$P(\theta | D) = \frac{P(D | \theta)p(\theta)}{P(D)} \propto \prod_{j=1}^K \theta_j^{\alpha_j + n_j - 1}$$