$P(x|\theta)$ $\quad D = \{x_1, \dots, x_n\}$ $\quad$ iid

$P(x|\theta)$ → ?

frequentist → $\hat{\theta}_{ML} = \arg\max_{\theta} P(D|\theta)$

Bayesian → θ is random var.

$\theta \sim P(\theta) \xrightarrow{D} \theta \sim P(\theta|D)$

$$\hat{\theta}_{ML} = \arg\max_{\theta} P(D|\theta) = \arg\max_{\theta} P(x_1, \dots, x_n|\theta) \overset{i.i.d}{=} \arg\max_{\theta} \prod_{i=1}^{n} P(x_i|\theta)$$

✓ $\hat{\theta} \longrightarrow \theta$

$n \to \infty$

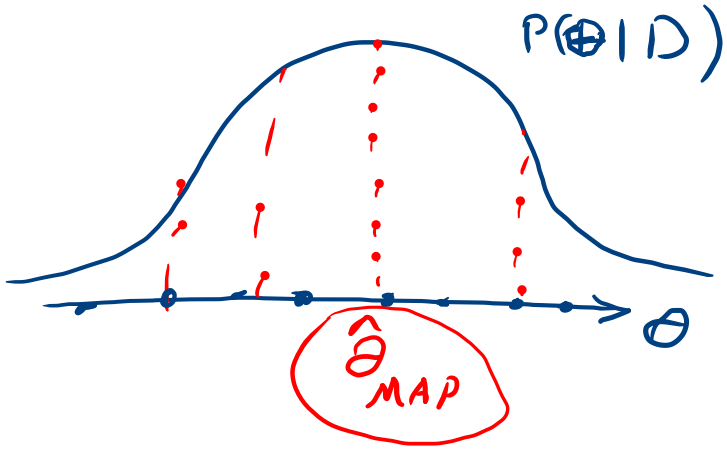
$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$E\left[(\hat{\theta} - \theta)^2\right]$

UMVE

$P(x|\theta) = ?$

$P(\theta) \xrightarrow{D} P(\theta|D)$



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

θ_1 θ_2

$$P(x|D) = \int P(x|\theta)\, P(\theta|D)\, d\theta = E_{\theta \sim P(\theta|D)}\left[P(x|\theta) \right]$$

predictive distribution

$P(x|\overline{x_1, \dots, x_n})$

$$p(x|D) = \int p(x,\theta|D)\, d\theta = \int p(x|\theta,D)\, p(\theta|D)\, d\theta$$

$p(x|\theta)$

$$D \longrightarrow \theta \longrightarrow x$$

$$= \int p(x|\theta)\, p(\theta|D)\, d\theta$$

$$= E\left[p(x|\theta)\right]$$

$\theta \sim p(\theta|D)$

$$p(x,y) = p(x)\, p(y|x)$$

1

$\theta \sim p(\theta) \longrightarrow p(\theta|D)$

$\int f_1(\theta)\, p(\theta|D)\, d\theta$

$= E[f_1(\theta)]$

Conjugate Prior

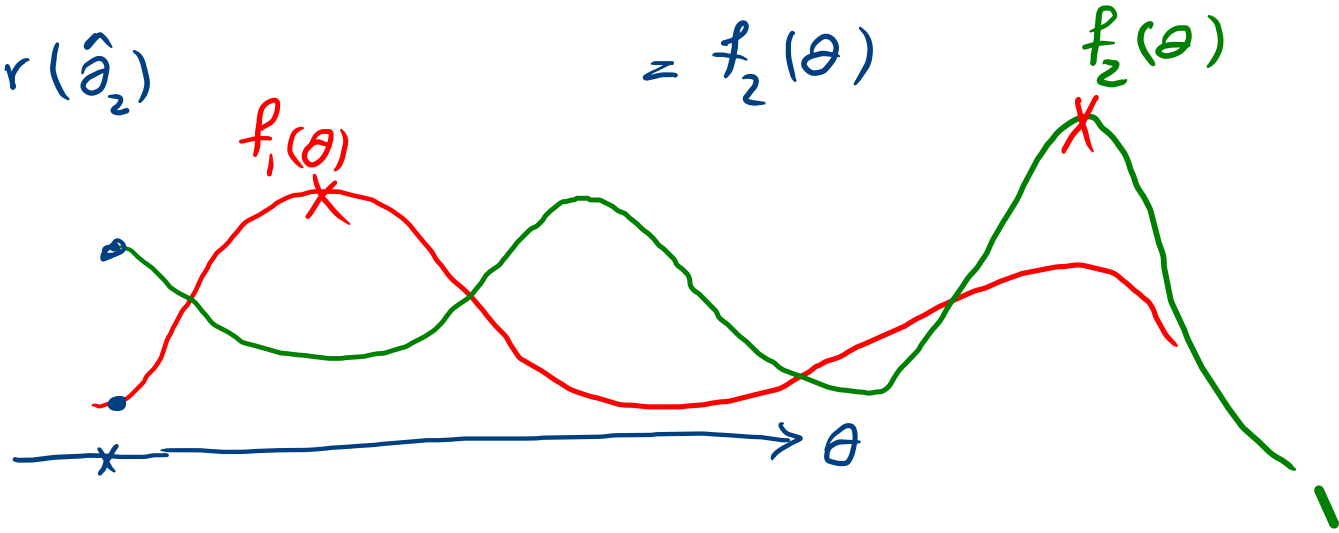
$\longrightarrow \hat{\theta}_1$

$\longrightarrow \hat{\theta}_2$

minimax

$var(\hat{\theta}_1) = E[(\hat{\theta} - \theta)^2] = f_1(\theta)$

$var(\hat{\theta}_2) = f_2(\theta)$



$$P(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x \in \{0, 1, \dots\}$$

$$O\left(\frac{\alpha}{\beta^2}\right)$$

$$\text{Gamma}(\lambda|\alpha,\beta) = c\,\lambda^{\alpha-1} e^{-\beta\lambda}$$

hyper-parameter

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \propto P(D|\theta)P(\theta) = \underbrace{\prod_{i=1}^{n} P(x_i|\theta)}_{P(x_1,\dots,x_n|\theta)} P(\theta)$$

$$= \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \; c\,\lambda^{\alpha-1} e^{-\beta\lambda} = c'\,\lambda^{\sum x_i + \alpha - 1} e^{-(n+\beta)\lambda}$$

$$= \text{Gamma}\left(\lambda \,\middle|\, \underbrace{\sum_{i=1}^{n} x_i + \alpha}_{\alpha'}, \underbrace{n+\beta}_{\beta'}\right)$$

$n \to \infty$

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)\, d\mu}$$
$$= \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu),$$

$N(\underline{\mu}, \Sigma)$?

$\mu \sim N(\mu_0, \sigma_0^2)$

where α is a **normalization factor** that depends on D but is independent of μ.

This equation shows how the observation of a set of training samples **affects our ideas about the true value of μ;**

$p(x_k|\mu) \sim N(\mu, \sigma^2)$ $\qquad$ $p(\mu) \sim N(\mu_0, \sigma_0^2)$

$$p(\mu|\mathcal{D}) = \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}^{p(\mu)}$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu-x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right]$$

$$= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right],$$

factors that **do not depend on μ** have been absorbed into the constants α, α', and α''.

Given $\mathcal{D} = \{x_1, \ldots, x_n\}$, we obtain

$$p(\mu|\mathcal{D}) \propto \prod_{i=1}^{n} p(x_i|\mu)p(\mu)$$

$$\propto \exp\left[-\frac{1}{2}\left(\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\mu_0}{\sigma_0^2}\right)\mu\right)\right]$$

$$= N(\mu_n, \sigma_n^2)$$

$p(\mu)$ is said to be a ***conjugate prior***

Where

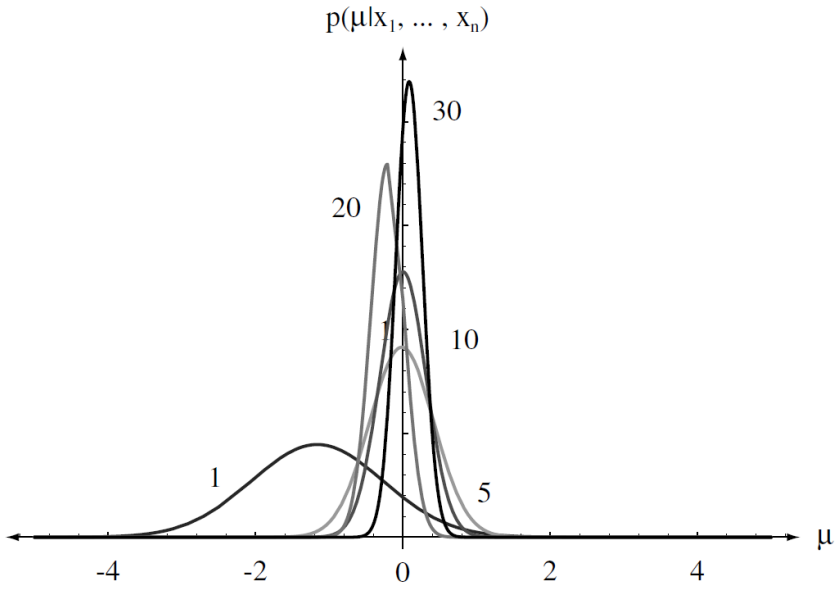
$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\bar{x}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$$

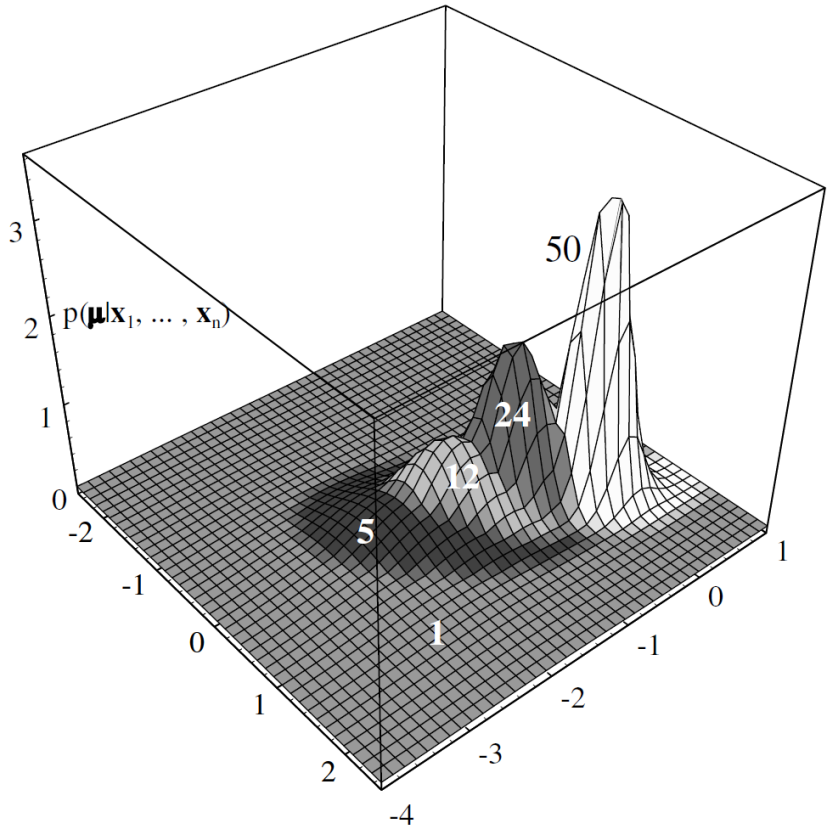
$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}.$$

σ_n^2 uncertainty about nth guess

$$\sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}.$$

- Since σ_n^2 **decreases monotonically** with *n*- **approaching σ^2/n** as *n* approaches infinity
- Each additional observation **decreases our uncertainty** about the true value of μ.
- As *n* increases, **p(μ|D)** becomes more and more **sharply peaked**, approaching a **Dirac delta** function





The Gaussian Case

- μ_0 is our **best prior guess** and σ_0^2 is the **uncertainty about** this guess.
- μ_n is our best guess **after observing n sample** in **D** and σ_n^2 is the uncertainty about this guess.
- μ_n always lies between $\bar{x}_n$ and μ_0 with coefficients that are non-negative **and sum to one**.

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\bar{x}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

  - If $\sigma_0 = 0$, then $\mu_n = \mu_0$ (**no observation can change** our prior opinion).
  - If $\sigma_0 >> \sigma$, then $\mu_n = \bar{x}_n$ (we are **very uncertain** about our prior guess).
  - Otherwise, μ_n approaches $\bar{x}_n$ as n **approaches infinity**

Class-conditional density

- Given the posterior density p(μ|***D***), the conditional density p(x|***D***) can be computed as

$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu|\mathcal{D})\ d\mu \\
&= \int \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]d\mu \\
&= \frac{1}{2\pi\sigma\sigma_n}\exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] f(\sigma,\sigma_n),
\end{aligned}
$$

$$
f(\sigma,\sigma_n) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu-\frac{\sigma_n^2 x+\sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right] d\mu.
$$

predictive dist →

$$
\boxed{p(x|\mathcal{D}) = N(\mu_n, \sigma^2+\sigma_n^2)}
$$

- Where
  - the **conditional mean μ_n** is treated as if it were the true mean,
  - the known **variance is increased** to account for our **lack of exact knowledge** of the mean μ.

The multivariate case

$$p(\mathbf{x}|\boldsymbol{\mu}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where μ is the only unknown parameter with a prior distribution

$$p(\boldsymbol{\mu}) = N(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}) \quad (\boldsymbol{\Sigma}, \boldsymbol{\mu_0} \text{ and } \boldsymbol{\Sigma_0} \text{ are all known}).$$

Given D = {x_1, . . . , x_n}, we obtain

$$p(\boldsymbol{\mu}|\mathcal{D}) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}^T\left(n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\mathbf{x}_i + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)\right)\right].$$

The Multivariate Gaussian Case

It follows that

$$p(\boldsymbol{\mu}|\mathcal{D}) = N(\boldsymbol{\mu_n}, \boldsymbol{\Sigma_n})$$

where

$$\boldsymbol{\mu_n} = \boldsymbol{\Sigma_0}\left(\boldsymbol{\Sigma_0} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\hat{\mu}_n} + \frac{1}{n}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma_0} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu_0},$$

$$\boldsymbol{\Sigma_n} = \frac{1}{n}\boldsymbol{\Sigma_0}\left(\boldsymbol{\Sigma_0} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}.$$

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\bar{x}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}.$$

Class-conditional density

- Given the posterior density p(μ|***D***), the conditional density p(x|***D***) can be computed as

$$p(\mathbf{x}|\mathcal{D}) = N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n)$$

- Which can be viewed as the **sum of** a random vector μ with

$$p(\boldsymbol{\mu}|\mathcal{D}) = N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$$

- and an **independent random vector y** with

$$p(\mathbf{y}) = N(0, \boldsymbol{\Sigma}).$$

The Bernoulli Case

- Consider P(x|**q**) = Bernoulli(**q**) where **q** is the unknown parameter with a prior distribution

  p(**q**)=Beta(a,b) (a and b are both known).

$$f(x;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

- Given D = {x_1, . . . , x_n}, we obtain

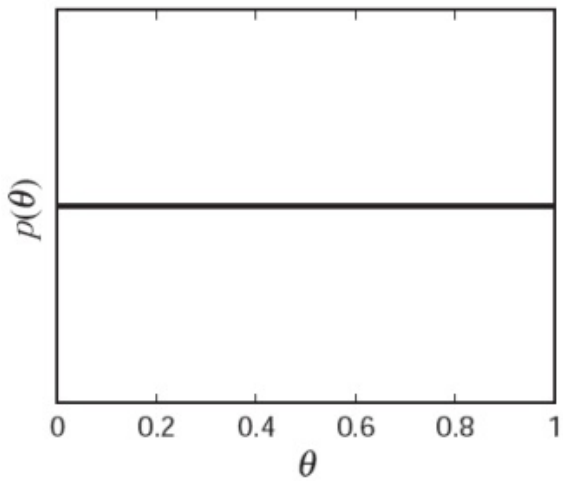
$$p(\theta|\mathcal{D}) = \mathrm{Beta}\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i\right).$$

$x \sim \text{Bernoulli}(\theta) \Rightarrow P(x) = \theta^x (1-\theta)^{1-x} \quad x \in \{0, 1\}$

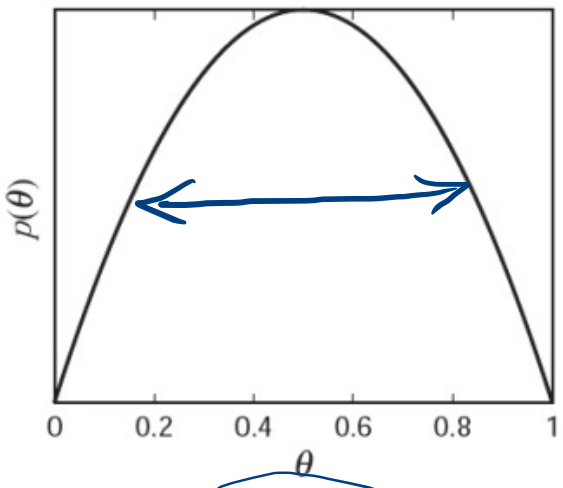
$\theta \sim P(\theta)$ $\qquad 0 \leq \theta \leq 1$

$\text{Beta}(\alpha, \beta)$

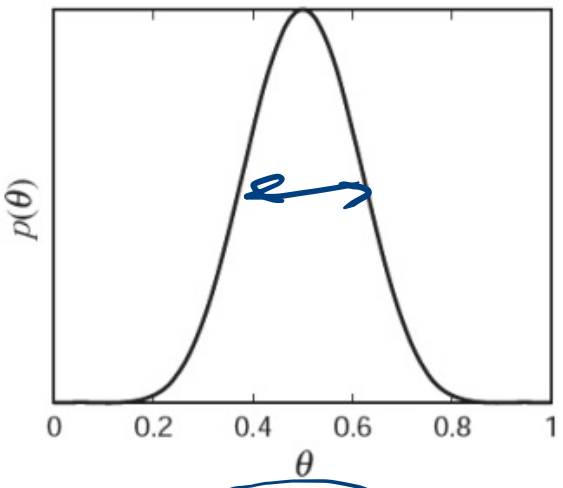
Beta(α, β)



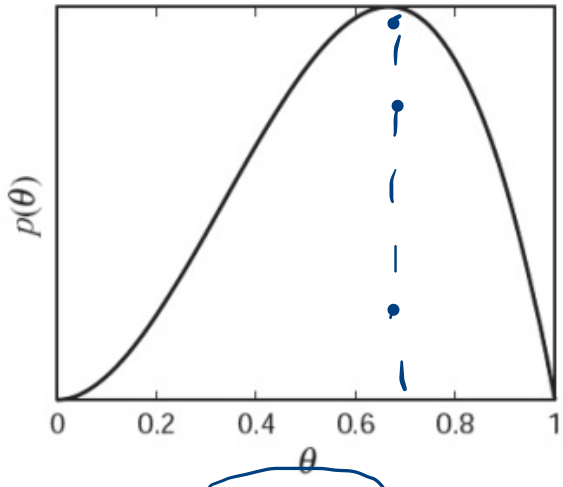
Beta(1,1)



Beta(2,2)

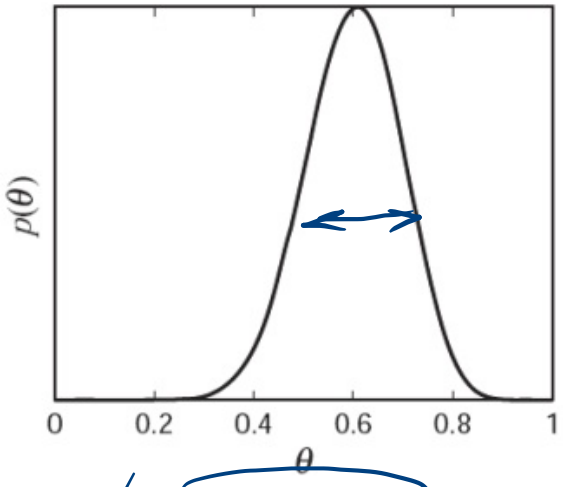


Beta(10,10)

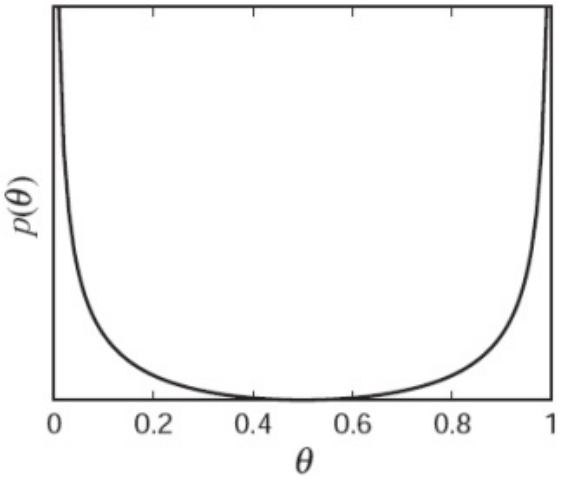


Beta(3,2)

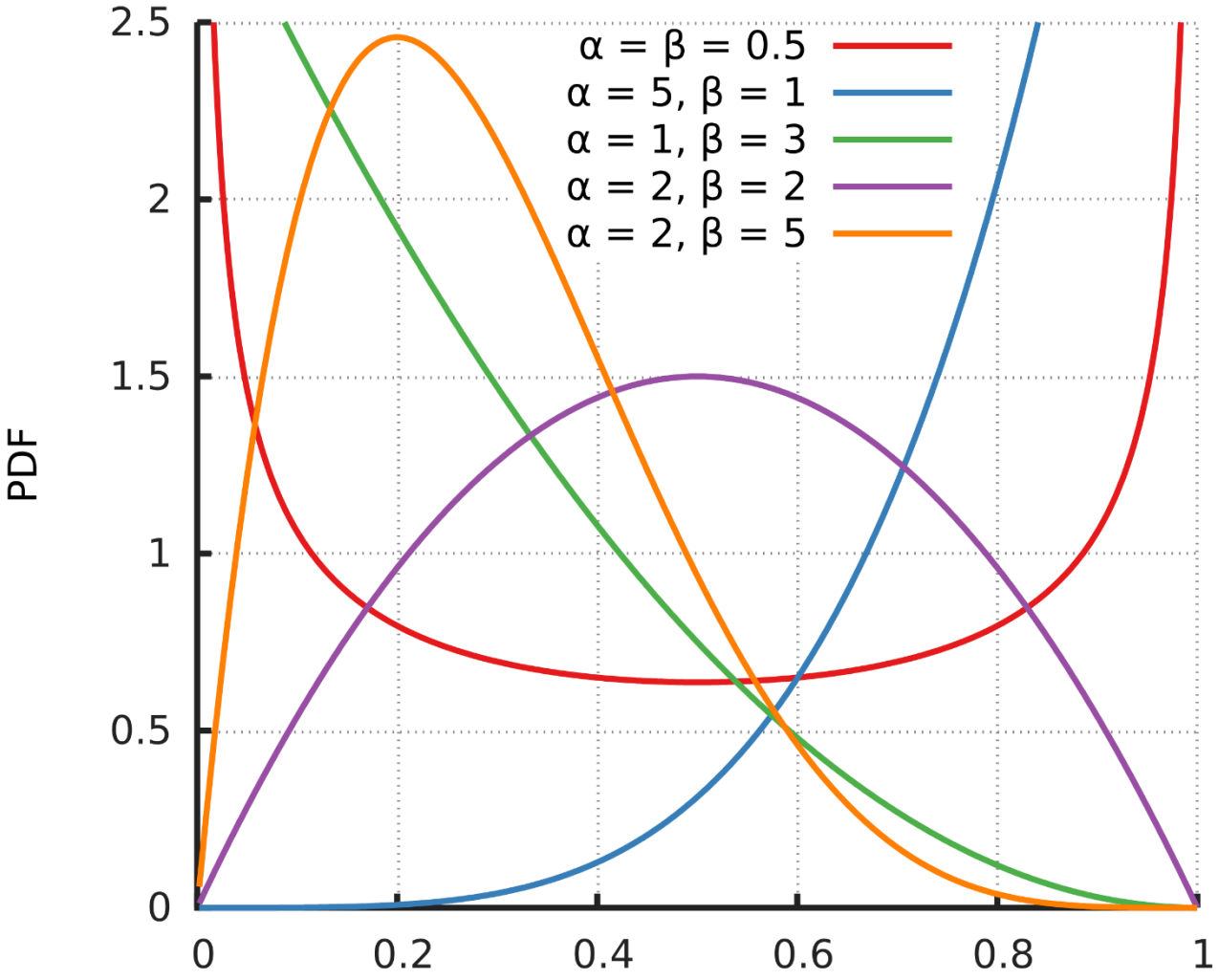
α > β



Beta(15,10)



Beta(0.5,0.5)



$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

$\alpha > 0$ shape (real)

$\beta > 0$ shape (real)

$x \in [0,1]$ or $x \in (0,1)$

$$\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx,$$

$$\mathrm{E}[X] = \frac{\alpha}{\alpha+\beta}$$

Mode $\frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$

$$\mathrm{var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The Bernoulli Case

- The Bayes estimate of **q** can be computed as the expected value of p(**q**|D), i.e.,

$$\hat{\theta} = \frac{\alpha + \sum_{i=1}^{n} x_i}{\alpha + \beta + n}$$

$$= \left(\frac{n}{\alpha + \beta + n}\right) \frac{1}{n} \sum_{i=1}^{n} x_i + \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right) \frac{\alpha}{\alpha + \beta}.$$

Conjugate Priors

- A conjugate prior is one which, when multiplied with the probability of the observation, gives a posterior probability **having the same functional form** as the prior.
- This relationship **allows** the **posterior** to **be used as a prior** in further computations.

pdf generating the sample	*corresponding conjugate prior*
Gaussian	Gaussian
Exponential	Gamma
Poisson	Gamma
Binomial	Beta
Multinomial	Dirichlet

Recursive Bayes Learning

- What about the **convergence** of p(x|D) to p(x)?
- Given $\mathcal{D}^n = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$, for $n > 1$

$$p(\mathcal{D}^n|\boldsymbol{\theta}) = p(\mathbf{x_n}|\boldsymbol{\theta})p(\mathcal{D}^{n-1}|\boldsymbol{\theta})$$

and

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}, \qquad p(\boldsymbol{\theta}|\mathcal{D}^n) = \frac{p(\mathbf{x_n}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D}^{n-1})}{\int p(\mathbf{x_n}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D}^{n-1})\, d\boldsymbol{\theta}}$$

where

$$p(\boldsymbol{\theta}|\mathcal{D}^0) = p(\boldsymbol{\theta})$$

- Quite useful if the distributions can be represented using **only a few parameters** (**sufficient statistics**).

Recursive Bayes Learning

- Consider the Bernoulli case P(x|**q**) = Bernoulli(**q**)

  where p(**q**)=Beta(a,b), the Bayes estimate of **q** is

$$\hat{\theta} = \frac{\alpha}{\alpha + \beta}.$$

- Given the training set D = {x_1, . . . , x_n}, we obtain

$$p(\theta|\mathcal{D}) = \mathbf{Beta}(\alpha + m, \beta + n - m)$$

where

$$m = \sum_{i=1}^{n} x_i = \#\{x_i | x_i = 1, x_i \in \mathcal{D}\}.$$

Recursive Bayes Learning

- The Bayes estimate of **q** becomes

$$\hat{\theta} = \frac{\alpha + m}{\alpha + \beta + n}.$$

- Then, given a **new training** set

$$\mathcal{D}' = \{x_1, \ldots, x_{n'}\}$$

- We obtain

$$p(\theta|\mathcal{D}, \mathcal{D}') = \text{Beta}(\alpha + m + m', \beta + n - m + n' - m')$$

- Where

$$m' = \sum_{i=1}^{n'} x_i = \#\{x_i | x_i = 1, x_i \in \mathcal{D}'\}.$$

Recursive Bayes Learning

- The Bayes estimate of **q** becomes

$$\hat{\theta} = \frac{\alpha + m + m'}{\alpha + \beta + n + n'}.$$

- Thus, recursive Bayes learning involves **only keeping the counts m** (related to **sufficient statistics** of Beta) and the number of training samples n.

Comparison of MLEs and Bayes estimates

	MLE	*Bayes*
computational complexity	differential calculus, gradient search	multidimensional integration
interpretability	point estimate	weighted average of models
prior information	assume the parametric model $p(\mathbf{x}\|\boldsymbol{\theta})$	assume the models $p(\boldsymbol{\theta})$ and $p(\mathbf{x}\|\boldsymbol{\theta})$ but the resulting distribution $p(\mathbf{x}\|\mathcal{D})$ may not have the same form as $p(\mathbf{x}\|\boldsymbol{\theta})$

If there is **much data** (strongly peaked p(**q**|D)) and the prior **p(q) is uniform, then the Bayes estimate and MLE** are **equivalent**.

Classification Error

- To apply these results to **multiple classes**, separate the training samples to c subsets $D_1, \ldots, D_c$, with the samples in D_i belonging to class w_i, and then **estimate each density** $p(x|w_i,D_i)$ **separately**.

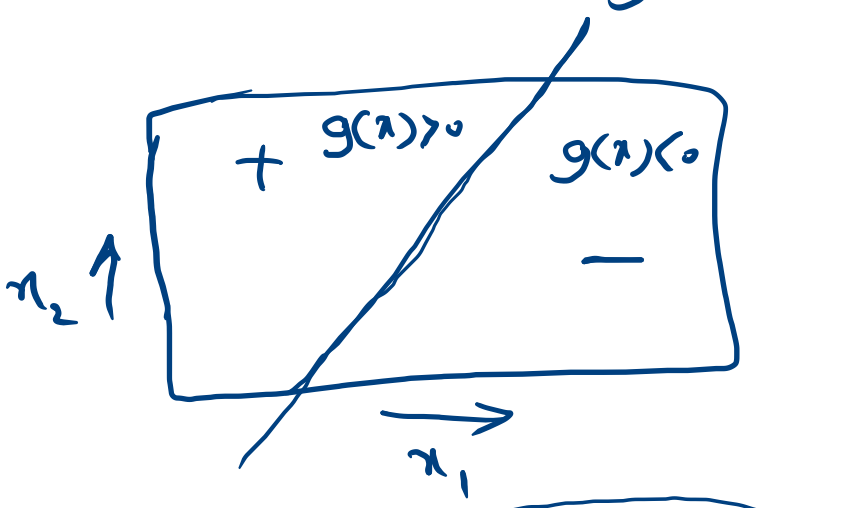
- Different **sources** of error:
  - **Bayes error**: due to overlapping class-conditional densities (related to the features used; inherent property of the problem and **can never be eliminated**).
  - **Model error**: due to i**ncorrect model**.
  - **Estimation error**: due to estimation from a finite sample (can be reduced by increasing the amount of training data).

Logistic Regression:

- Discriminative classifier → $P(y|x)$ ✓
- Bayes Classifier
- Linear Classifier

Binary //

$$\frac{P(y=1|x)}{P(y=0|x)} \underset{0}{\overset{1}{\gtrless}} 1$$



$P(y|x)$

$g(x)$

$$g(x) = \ln \frac{P(y=1|x)}{1 - P(y=1|x)} \underset{0}{\overset{1}{\gtrless}} 0$$

$$g(x) = w^T x + w_0$$

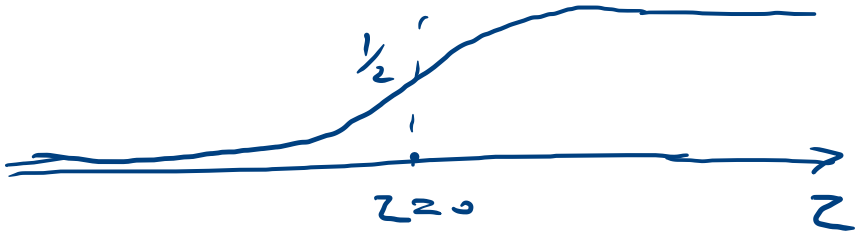
$$g(x) = \ln \frac{P(y=1|x)}{1 - P(y=1|x)} = \omega^T x$$

$$g(x) = \omega^T x + \omega_0$$

$$\omega = \begin{bmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_d \end{bmatrix} \qquad x = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$$P(y=1|x) = \frac{1}{1 + e^{-\omega^T x}}$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



$$P(y=1|x) = \sigma(\omega^T x)$$

$$P(y|x) = \theta^y (1-\theta)^{1-y} = \sigma(\omega^T x)^y \left(1 - \sigma(\omega^T x)\right)^{1-y}$$

$\overset{?}{=}$ $\overset{?}{=}$

$$P(y|x) = \sigma(w^T x)^y \left(1 - \sigma(w^T x)\right)^{1-y}$$

$\rightarrow D = \{(x_1, y_1), \dots, (x_n, y_n)\}$

$$\hat{w}_{ML} = \arg\max_w \ln P(D|w)$$

$$\log P(D|w) = \log P(y_1, \dots, y_n | x_1, \dots, x_n, w) \overset{\text{i.i.d}}{=} \sum_{i=1}^{n} \log P(y_i|x_i, w)$$

$$= \sum_{i=1}^{n} y_i \log \sigma(w^T x_i) + (1 - y_i) \log\left(1 - \sigma(w^T x_i)\right) = L(w)$$

$$\nabla_w L(w) = \sum_{i=1}^{n} x_i \left(y_i - \sigma(w^T x_i)\right)$$

$\hat{w}_{ML}$

$\sigma(w^T x_i) = \frac{1}{1 + e^{-w^T x_i}}$

$p(x|\theta)$

$p(y|x,\theta)$

$\arg\max_{\theta} \; p(x_1, \cdots, x_n|\theta)$

$p(y_1 \cdots y_n | x_1 \cdots x_n, \theta)$