

# Bayesian Inference

Data Analysis for Networks - DataNets'19  
Anastasios Giovanidis

Sorbonne-LIP6



May 22, 2019

# Bibliography

- B.1 Christopher M. Bishop, "Pattern Recognition and Machine Learning", Springer 2006.
  - B.2 H. Pishro-Nik, "Introduction to probability, statistics, and random processes", available at <https://www.probabilitycourse.com>, Kappa Research LLC, 2014.
- 👉 Chapter 8.3, 8.4

## Heads or Tails?

A. Giovanidis 2019

Suppose we toss a coin three times: (H, H, H)



What can we say about the probability to get heads (H) in the next toss?

## Probability of Heads

We remind the frequentist estimation (Sample Mean):

$$\hat{\Theta} = \bar{X} = \frac{1 + 1 + 1}{3} = 1$$

☞ The estimated probability for heads (H) is 1, thus we expect surely to get heads next time we throw the coin.

Is this a good estimate?

# Probability of Heads

We remind the frequentist estimation (Sample Mean):

$$\hat{\Theta} = \bar{X} = \frac{1 + 1 + 1}{3} = 1$$

☞ The estimated probability for heads (H) is 1, thus we expect surely to get heads next time we throw the coin.

Is this a good estimate?

This is the best we can do, given the information we have.

## Limited experience

A. Giovanidis 2019

In the "Heads or Tails" game, we can repeat the experiment several times, until we get a good "frequentist" estimate of the chance to fall Heads (H).

If the coin is fair, the unknown parameter will obviously be  $1/2$ . The sample mean will "eventually" converge to this value because of zero bias.

But, there are also other events that cannot be repeated many times:

Will the Arctic ice cap have disappeared by the end of the century?

# Revise Uncertainty

A. Giovanidis 2019



☞ By obtaining fresh data, we can revise every year our opinion on the rate of ice loss, given some previous idea that we had.

# Thomas Bayes (1701-1761)

A. Giovanidis 2019



- ▶ Theologist, scientist, mathematician.
- ▶ **Inverse Probability** "Essay towards solving a problem in the doctrine of chances" (1764)
- ▶ The name "Bayes Theorem" was given by Poincaré.



## Pierre-Simon Laplace (1749-1827)

A. Giovanidis 2019



- ▶ "Best mathematician in France" at that time.
- ▶ "Théorie Analytique des Probabilités" (1812)

## Bayes rule

Back to our estimation problem. Suppose that we observe data  $\mathcal{D} = \{x_1, \dots, x_n\}$ , and we want to estimate  $\theta$ .

In the Heads-Tails example, the estimate was the probability of Heads.

Bayes rule, assumes a **prior distribution**  $f_{\Theta}(\theta)$  over the value of  $\theta$ .

$$f_{\Theta|\mathcal{D}}(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta) \cdot f_{\Theta}(\theta)}{P(\mathcal{D})}$$

The **posterior density**  $f_{\Theta|\mathcal{D}}(\theta|\mathcal{D})$  can be used to infer  $\Theta$ .

Bayes rule assumes that **the unknown is a random variable  $\Theta$  rather than fixed and deterministic.**

## Prior and Posterior distributions

- ▶  $P(\mathcal{D}|\theta)$  is just the **likelihood function** ! How probable is the observed data given the parameter  $\theta$  and the distribution.
- ▶  $P(\mathcal{D})$  is the overall probability to observe the data

$$P(\mathcal{D}) = \int P(\mathcal{D}|\theta) f_{\Theta}(\theta) d\theta.$$

Note: It is a normalisation constant.

Bayes theorem in simple words

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

## Prior and Posterior distributions

- ▶  $P(\mathcal{D}|\theta)$  is just the **likelihood function** ! How probable is the observed data given the parameter  $\theta$  and the distribution.
- ▶  $P(\mathcal{D})$  is the overall probability to observe the data

$$P(\mathcal{D}) = \int P(\mathcal{D}|\theta) f_{\Theta}(\theta) d\theta.$$

Note: It is a normalisation constant.

Bayes theorem in simple words

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

☞ The prior distribution summarises our initial **uncertainty** over the parameter value  $\theta$ , and the posterior, how this uncertainty is updated after the data is taken into account.

# Application: Wireless Communications

A. Giovanidis 2019

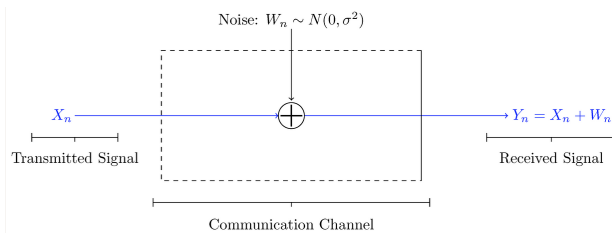


Figure: Source H. Pishro-Nik (B.2)

# Application: Wireless Communications

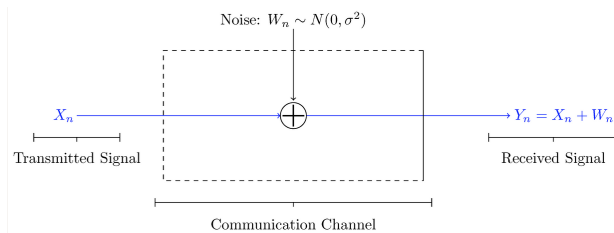


Figure: Source H. Pishro-Nik (B.2)

👉 We want to estimate  $X_n$  based on the received  $Y_n$ , and assuming we know the prior distribution. Then, the posterior (pdf) is

$$f_{X_n|Y_n}(x|y) = \frac{f_{Y_n|X_n}(y|x) \cdot f_X(x)}{f_Y(y)}.$$

## Application: Spam filter

Given that a certain word  $W$  appears in an email, is it **Spam** or **Ham**?

The Software applies Bayes' theorem (**PMF**):

$$P(S|W) = \frac{P(W|S) \cdot P(S)}{P(W|S) \cdot P(S) + P(W|H) \cdot P(H)}$$

$Pr(S W)$	probability that a message is a spam, given it contains "W"
$Pr(S)$	overall probability that any message is spam
$Pr(W S)$	probability that the word "W" appears in spam messages
$Pr(H)$	overall probability that any given message is not spam
$Pr(W H)$	probability that the word "W" appears in "ham" messages.

## Example: Inference

☞ 3 coins in my pocket

1. Biased 3:1 in favour of Tails
2. Fair coin
3. Biased 3:1 in favour of Heads

I randomly pick one coin, flip it and get Heads (H). What is the probability that I have chosen coin No.3?

### INPUT

$X = 1$  means Heads,  $X = 0$  means Tails,  $\theta$  is the mean.

Prior:  $P(\theta = 0.25) = P(\theta = 0.5) = P(\theta = 0.75) = \frac{1}{3}$ .



## Example: Inference cont'd

		Prior	Likelihood	Posterior	Posterior Norm.
Coin	$\theta$	$P(\theta)$	$P(X = 1 \theta)$	$P(X = 1 \theta)P(\theta)$	$\frac{P(X=1 \theta)P(\theta)}{P(X=1)}$
No.1	0.250	0.333	0.250	0.083	0.167
No.2	0.500	0.333	0.500	0.167	0.333
No.3	0.750	0.333	0.750	0.250	0.500

where, the normalising constant is

$$P(X = 1) = 0.083 + 0.167 + 0.250 = 0.500.$$

👉 **Answer:** I have chosen No.3 with probability 50%, No.2 with probability 25% and No.1 with probability 33.3%.

Coin No.3 is both the ML estimate as well as the "MAP" estimate.

# MAP Estimator

**Maximum Likelihood (ML) estimator**

$$\theta_{ML} = \arg \max_{\theta} P(\mathcal{D} \mid \theta)$$

**Maximum A Posteriori (MAP) estimator**

$$\begin{aligned}\theta_{MAP} &= \arg \max_{\theta} P(\theta \mid \mathcal{D}) \\ &= \arg \max_{\theta} P(\mathcal{D} \mid \theta) \cdot P_{\Theta}(\theta)\end{aligned}$$

Note: When the prior is a uniform distribution, then  $P_{\Theta}(\theta)$  is a constant and  $\theta_{ML} = \theta_{MAP}$ .

☞ The MAP is a summary statistic of the posterior which corresponds to the **mode** (arg max).

## Normal distribution

Consider a single real-valued variable  $x$  that follows a Gaussian distribution

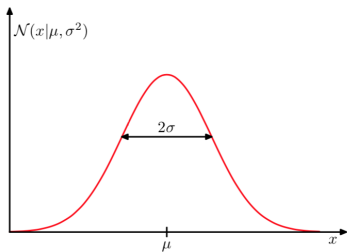
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\},$$

- ▶ with **mean**  $\mu$ ,
- ▶ **variance**  $\sigma^2$ ,
- ▶ **standard deviation**  $\sigma$  (derived as  $\sqrt{\text{Var}(X)}$ ),
- ▶ and **precision**  $\beta = 1/\sigma^2$ .

## Gaussian PDF: properties

A. Giovanidis 2019

- ▶ Positive:  $\mathcal{N}(x|\mu, \sigma^2) > 0$ ,
- ▶ Valid probability density:  $\int_{-\infty}^{+\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1$
- ▶ Mean:  $\mathbb{E}[X] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x dx = \mu$ ,
- ▶ Second moment:  $\mathbb{E}[X^2] = \mu^2 + \sigma^2$ ,
- ▶ Variance:  $\mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mu^2 + \sigma^2 - \mu^2 = \sigma^2$ .



Source: Bishop (B.2)

## Gaussian inference

A. Giovanidis 2019

- Data

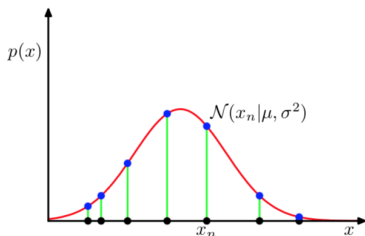
$$\mathcal{D} = \{x_1, \dots, x_N\}$$

- Data i.i.d. from Gaussian PDF

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Unknown parameters

$$\theta = \{\mu, \sigma^2\}$$



Source: Bishop (B.2)

## Gaussian ML

A. Giovanidis 2019

► Likelihood

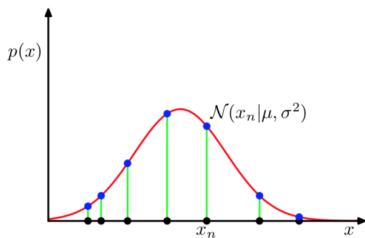
$$P(\mathcal{D}|\theta) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2)$$

► Maximum likelihood

$$\theta_{ML} = \arg \max_{\theta} P(\mathcal{D}|\theta)$$

► equivalent problem

$$\theta_{ML} = \arg \max_{\theta} \log P(\mathcal{D}|\theta)$$



Source: Bishop (B.2)

## Gaussian ML solution

$$(\mu_{ML}, \sigma_{ML}) = \arg \max_{\mu, \sigma} \left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log(2\pi) \right\}$$

- Maximise first over  $\mu$

$$\frac{\partial \log P(\mathcal{D}|\theta)}{\partial \mu} = 0 \Rightarrow \mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n =: \bar{X}_n$$

- Then, maximise over  $\sigma^2$

$$\frac{\partial \log P(\mathcal{D}|\theta)}{\partial \sigma^2} = 0 \Rightarrow \sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2.$$

## Problems related to ML solution

The available dataset, could be a result of i.i.d Gaussian realisations, but the available values contain uncertainty. Let us observe the extreme case for  $N = 1$

- ▶ ML estimate of  $\mu$

$$\mu_{ML} = x_1$$

- ▶ and ML estimate of  $\sigma^2$

$$\sigma_{ML}^2 = (x_1 - \mu_{ML})^2 = (x_1 - x_1)^2 = 0.$$

How about the Bayesian approach?



## Gaussian MAP

- Assume for simplicity known  $\{\sigma^2\}$  variance.

Unknown parameter  $\theta = \{\mu\}$ .

- Likelihood

$$P(\mathcal{D}|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2)$$

- Combine likelihood with a **Gaussian prior** over  $\mu$

$$P(\mu) = \mathcal{N}(\mu | m_0, s_0^2)$$

- The **posterior** is proportional to

$$P(\mu | \mathcal{D}, \sigma^2) \propto P(\mathcal{D}|\mu, \sigma^2) \cdot P(\mu)$$

## Bayesian update

$$\begin{aligned}
 P(\mu \mid \mathcal{D}, \sigma^2) &\propto P(\mathcal{D} \mid \mu, \sigma^2) \cdot P(\mu) \\
 &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \cdot \frac{1}{\sqrt{2\pi s_0^2}} \exp\left(-\frac{(\mu - m_0)^2}{2s_0^2}\right) \\
 &= \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{\sqrt{2\pi s_0^2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \mu)^2 - \frac{1}{2s_0^2} (\mu - m_0)^2\right) \\
 &= C_1 \cdot e^{-\frac{1}{2\sigma^2} \sum_{i=1}^N (x_i^2 + \mu^2 - 2\mu x_i) - \frac{1}{2s_0^2} (\mu^2 + m_0^2 - 2\mu m_0)} \\
 &= C_2 \cdot \exp\left(-\frac{1}{2\hat{\sigma}_N^2} \left[\mu^2 - 2\mu \hat{\sigma}_N^2 \left(\frac{N\mu_{ML}}{\sigma^2} + \frac{m_0}{s_0^2}\right) + C_3\right]\right).
 \end{aligned}$$

$C_2$  and  $C_3$  are such that  $\frac{P(\mathcal{D} \mid \mu, \sigma^2) \cdot P(\mu)}{Z}$  is a probability density function.  
For the Gaussian pdf the max coincides with the mean due to symmetry!

## Iteration

- ▶  $P(\mu \mid \mathcal{D}, \sigma^2) \sim \mathcal{N}(\mu \mid \hat{\mu}_N, \hat{\sigma}_N^2)$  .
- ▶  $\frac{1}{\hat{\sigma}_N^2} = \frac{1}{s_0^2} + \frac{N}{\sigma^2}$
- ▶  $\hat{\mu}_N = \frac{\sigma^2}{Ns_0^2 + \sigma^2} m_0 + \frac{Ns_0^2}{Ns_0^2 + \sigma^2} \mu_{ML}, \quad \mu_{ML} = \frac{1}{N} \sum_{i=1}^N x_i.$

where  $\hat{\mu}_N, \hat{\sigma}_N^2$  are the Bayesian MAP estimates.

### Limiting cases

	$N = 0$	$N \rightarrow \infty$
$\hat{\sigma}_N^2$	$s_0^2$	0
$\hat{\mu}_N$	$m_0$	$\mu_{ML}$

## Posterior for the mean

- Posterior of the Gaussian mean for increasing data size  $N$   
(The variance reduces with  $N$ !)

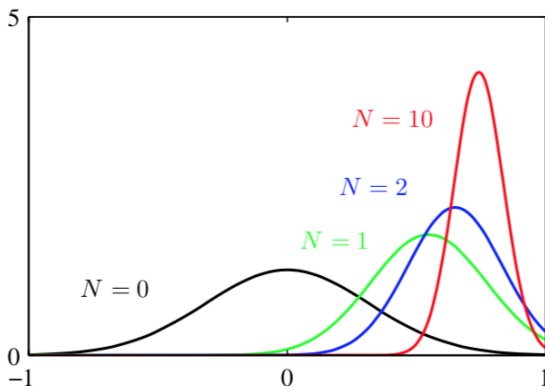


Figure: Bishop (B.2), p.99

## Conjugate Priors

☞ In the above case:

Observe already that the posterior distribution has the same shape (Gaussian) as the prior!

$P(\theta)$  is a **conjugate prior** for a particular likelihood  $P(\mathcal{D} \mid \theta)$  if the posterior is of the same functional form as the prior.

For all members of the **exponential family** it is possible to construct a conjugate prior.

$$P(\mathbf{x} \mid \theta) = h(\mathbf{x})g(\theta) \exp(\theta^T u(\mathbf{x})).$$

## Conjugate Priors for Gaussian variance

► Likelihood

$$P(\mathcal{D}|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2)$$
$$\stackrel{\lambda:=1/\sigma^2}{=} \left(\frac{\lambda}{2\pi}\right)^{N/2} \exp\left\{-\frac{\lambda}{2} \sum_{i=1}^N (x_i - \mu)^2\right\}$$

► For **known mean**, the suitable prior is:

$$P(\lambda) = \text{Gam}(\lambda \mid a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda).$$

Note:  $\Gamma(x) := \int_0^\infty u^{x-1} e^{-u} du$ . Also  $\Gamma(x+1) = x\Gamma(x)$ . It holds:  $\Gamma(1) = 1$  and hence  $\Gamma(x+1) = x!$  when  $x$  is integer.

Properties:  $\mathbb{E}[\lambda] = \frac{a}{b}$ , and  $\text{Var}[\lambda] = \frac{a}{b^2}$ .

## Conjugate Priors for Gaussian (general)

Likelihood reformulated

$$\begin{aligned} P(\mathcal{D}|\mu, \sigma^2) &\stackrel{\lambda:=1/\sigma^2}{=} \left(\frac{\lambda}{2\pi}\right)^{N/2} \exp\left\{-\frac{\lambda}{2} \sum_{i=1}^N (x_i - \mu)^2\right\} \\ &\propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu \sum_{i=1}^N x_i - \frac{\lambda}{2} \sum_{i=1}^N x_i^2\right\} \\ &\propto \exp\left\{-\frac{\beta\lambda}{2} \left(\mu - \frac{c}{\beta}\right)^2\right\} \lambda^{\beta/2} \exp\left\{-\left(d - \frac{c^2}{2\beta}\right)\lambda\right\}. \end{aligned}$$

☞ For **unknown mean and variance** the conjugate prior is

$$P(\mu, \lambda) = \mathcal{N}(\mu \mid \mu_0, (\beta\lambda)^{-1}) \cdot \text{Gam}(\lambda \mid a, b).$$

**Normal-Gamma distribution** (coupling between  $\mu$  and  $\lambda$ )

## Bernoulli inference

Consider a number  $N$  of Bernoulli realisations with parameter  $\mu$

$$P(x = 1 \mid \mu) = \mu.$$

The probability distribution for Bernoulli is given by

$$\text{Bernoulli}(x \mid \mu) = \mu^x(1 - \mu)^{1-x},$$

which has variance  $\text{Var}[x] = \mu(1 - \mu)$ .

The **Likelihood function**, given i.i.d. observations from  $P(x = 1 \mid \mu)$

$$P(\mathcal{D} \mid \mu) = \prod_{i=1}^N \mu^{x_i}(1 - \mu)^{1-x_i}.$$



## Bernoulli ML

From the **Maximum Likelihood** estimate, we get:

$$\begin{aligned}\mu_{ML} &= \arg \max_{\mu} \log P(\mathcal{D} \mid \mu) \\ &= \arg \max_{\mu} \sum_{i=1}^N \log P(x_i \mid \mu) \\ &= \arg \max_{\mu} \sum_{i=1}^N \{x_i \log(\mu) + (1 - x_i) \log(1 - \mu)\}\end{aligned}$$

$$\frac{d}{d\mu} \log P(\mathcal{D} \mid \mu) = 0 \quad \Rightarrow \quad \sum_{i=1}^N \frac{x_i}{\mu} = \sum_{i=1}^N \frac{1 - x_i}{1 - \mu}$$

$$\mu_{ML} = \frac{1}{N} \sum_{i=1}^N x_i.$$

## Binomial Heads

Equivalently, we see that

$$\mu_{ML} = \frac{m(N)}{N},$$

where  $m(N)$  is the number of Heads (H) in a Heads-Tails experiment of size  $N$ .

The number  $m(N)$  follows the [Binomial distribution](#)

$$m(N) \sim \text{Binomial}(m \mid N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m},$$

where  $\binom{N}{m} = \frac{N!}{(N-m)!m!}$  is the number of choosing  $m$  objects out of  $N$  identical ones.

## Binomial Distribution

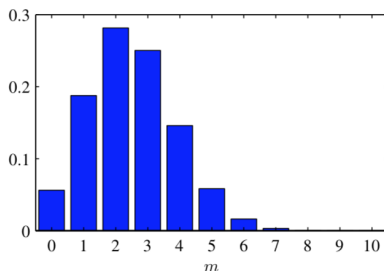


Figure: Bishop (B.2), p.70

- ▶  $\mathbb{E}[m] := \sum_{m=0}^N m \text{Binomial}(m \mid N, \mu) = N\mu.$
- ▶  $\text{Var}[m] := \sum_{m=0}^N (m - \mathbb{E}[m])^2 \text{Binomial}(m \mid N, \mu) = N\mu(1 - \mu).$

## Bernoulli ML issues

☞ The ML estimator for the Bernoulli is based strongly on the available data, and tends to **severely overfit** the estimated value for small data-sets.

Remember the Heads-Tails example  $\{H, H, H\}$ .

$$\mu_{ML} = \frac{1}{3} \sum_{i=1}^3 x_i = \frac{1+1+1}{3} = 1.$$

**Prediction:** From the above the coin should always (a.s.) give Heads !

## Bayesian approach

- We will use the Bayesian approach and will propose a **conjugate prior** that keeps the same shape when multiplied by the Likelihood function.
- ▶ We saw that the Likelihood function is

$$P(\mathcal{D} \mid \mu) = \prod_{i=1}^N \mu^{x_i} (1 - \mu)^{1-x_i} = \mu^{\textcolor{red}{m}} (1 - \mu)^{\textcolor{red}{\ell}},$$

where  $m$  is the count of ( $H$ ),  $\ell$  is the count of ( $T$ ) and  $\ell = N - m$ .

## Bayesian approach

- We will use the Bayesian approach and will propose a **conjugate prior** that keeps the same shape when multiplied by the Likelihood function.
- ▶ We saw that the Likelihood function is

$$P(\mathcal{D} \mid \mu) = \prod_{i=1}^N \mu^{x_i} (1 - \mu)^{1-x_i} = \mu^m (1 - \mu)^\ell,$$

where  $m$  is the count of ( $H$ ),  $\ell$  is the count of ( $T$ ) and  $\ell = N - m$ .

- ▶ The **Beta function** has the conjugate property

$$\text{Beta}(\mu \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1 - \mu)^{b-1}.$$

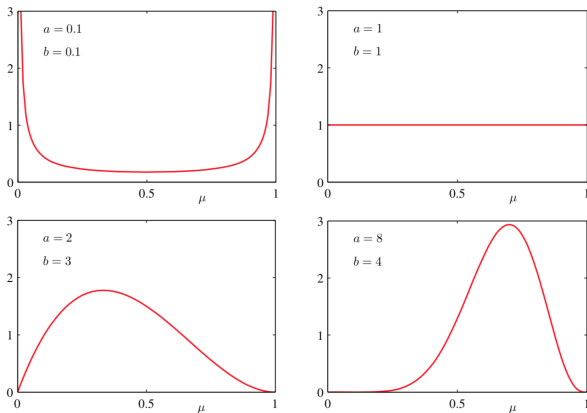
## Beta Moments

The mean and variance of the Beta distribution are given by

$$\mathbb{E}[\mu] = \frac{a}{a+b},$$

$$\text{Var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}.$$

## Beta plots



**Figure 2.2** Plots of the beta distribution  $\text{Beta}(\mu|a, b)$  given by (2.13) as a function of  $\mu$  for various values of the hyperparameters  $a$  and  $b$ .

Figure: Bishop (B.2), p.72



## Bernoulli posterior

- Suppose  $\text{Beta}(\mu|a, b)$  is the **prior distribution** for  $\mu$  and multiply with the binomial likelihood function.

👉 Posterior distribution  $\text{Beta}(\mu|a + m, b + \ell)$ :

$$P(\mu \mid m, \ell, a, b) \propto \mu^{m+a-1}(1 - \mu)^{\ell+b-1}.$$

## Bernoulli posterior cont'd

Taking into account the normalisation, we have:

$$P(\mu \mid m, \ell, a, b) = \frac{\Gamma(m+a+\ell+b)}{\Gamma(m+a)\Gamma(\ell+b)} \mu^{m+a-1} (1-\mu)^{\ell+b-1}.$$

□ Observing  $m$  Heads in data, adds  $m$  to  $a$ . Similarly, observing  $\ell$  Tails in data adds  $\ell$  to  $b$ . → These **hyperparameters** can be seen as the **effective number of observations of  $x = 1$  and  $x = 0$** .

□ The posterior probability distribution has an **updated mean**

$$P(x = 1 \mid \mathcal{D}) = \frac{m+a}{m+a+\ell+b}$$

When  $m, \ell \rightarrow \infty$ :  $P(x = 1 \mid \mathcal{D}) \approx \frac{m}{m+\ell} = \frac{m}{N} = \mu_{ML}$ .

# Sequential Learning

- ▶ Very often we do not have the whole dataset  $\mathcal{D}$  available.
- ▶ Data arrives sequentially, and we need to **update** our estimates using the new info.
- ▶ e.g.  $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_t, \dots$
- ▶ In the simplest case, each data-set consists of 1 single new data (measurement)

**Question:** How do the ML and MAP (Bayesian) estimators update sequentially?

# Sequential ML

A. Giovanidis 2019

Consider we have observed data from  $\mathcal{D}' = \{x_1, \dots, x_{N-1}\}$ , estimate  $\mu_{ML}^{(N-1)}$ , and then observe  $x_N$  and update the estimate.

► Gaussian, Bernoulli:

$$\begin{aligned}\mu_{ML}^{(N)} &= \frac{1}{N} \sum_{i=1}^N x_i \\ &= \dots \\ &= \mu_{ML}^{(N-1)} + \frac{1}{N} \left( x_N - \mu_{ML}^{(N-1)} \right)\end{aligned}$$

## Robbins-Monro algorithm

In the more general case, we can use following sequential algorithm:

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1} \frac{\partial}{\partial \theta^{(N-1)}} \left[ -\log P(x_N | \theta^{(N-1)}) \right]$$

where

- ▶  $\lim_{N \rightarrow \infty} a_N = 0$ ,
- ▶  $\sum_{N=1}^{\infty} a_N = \infty$ ,
- ▶  $\sum_{N=1}^{\infty} a_N^2 < \infty$

Note that in the case of Gaussian  $-\log P(x|\mu_{ML}, \sigma^2) = \frac{1}{2\sigma^2} (x - \mu_{ML})^2$ ,

$$\mu_{ML}^{(N)} = \mu_{ML}^{(N-1)} + a_{N-1} \frac{1}{\sigma^2} (x_N - \mu_{ML}^{(N-1)}) .$$

What is  $a_{N-1}$  for the Gaussian ML?

# Sequential MAP

- ▶ We consider again a data set  $\mathcal{D}'$  of  $N - 1$  data points, and observation  $x_N$ .
- ▶ Posterior distribution:

$$\begin{aligned} P(\theta \mid \{\mathcal{D}', x_N\}) &\propto \prod_{i=1}^N P(x_i \mid \theta) \cdot P(\theta) \\ &= \left[ \prod_{i=1}^{N-1} P(x_i \mid \theta) \cdot P(\theta) \right] P(x_N \mid \theta) \\ &= P(\theta \mid \mathcal{D}') \cdot P(x_N \mid \theta). \end{aligned}$$

The **posterior distribution** after  $N - 1$  observations, becomes the **new prior**!

## Exercise 1: RADAR

A radar scans a surface for dangerous targets every time unit [hour].

- ▶ The detection mechanism of the radar can detect a real target in 99% of all cases (True Positive).
- ▶ It happens that in 2% of scans there is a False Alarm (False Positive).
- ▶ From statistics that a real target appears every 1000 time units.

**Question:** What is the probability that an alarm by the radar corresponds to a true target?

# Solution 1: RADAR

A. Giovanidis 2019

- ▶  $P(\text{Alarm} \mid \text{Target}) = 0.99$
- ▶  $P(\text{Alarm} \mid \text{Nothing}) = 0.02$
- ▶  $P(\text{Target}) = 0.001$

$$\begin{aligned} P(\text{Target} \mid \text{Alarm}) &= \frac{P(\text{Alarm} \mid \text{Target}) \cdot P(\text{Target})}{P(\text{Alarm})} \\ &= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.02 \cdot 0.999} \\ &\approx 0.05. \end{aligned}$$



## Exercise 2: WIFI

A user wants to use a public WiFi shared by others. At different times per day, the user has different access probability:

1. When there are not many users connected (GOOD):  
 $P(\text{Access}) = 99/100$ .
2. When there are many users online (BAD):  
 $P(\text{Access}) = 50/100$ .

☞ Suppose a user connects, and the first frame sent goes through.

**Question:** What is the probability that the second frame goes through as well? (i.i.d.)

## Solution 2: WIFI

The user does not know if the channel is GOOD or BAD, so let us choose a prior distribution

$$P(GOOD) = P(BAD) = 0.5.$$

We want to compute:

$$P(X_2 = 1 \mid X_1 = 1) = \frac{P(X_2 = 1, X_1 = 1)}{P(X_1 = 1)}.$$

We have the following information:

$$P(X_i = 1 \mid GOOD) = 0.99 \quad P(X_i = 1 \mid BAD) = 0.5.$$

## Solution 2: WIFI cont'd

$$\begin{aligned}P(X_2 = 1, X_1 = 1) &= P(X_2 = 1|GOOD)P(X_1 = 1|GOOD)P(GOOD) \\&+ P(X_2 = 1|BAD)P(X_1 = 1|BAD)P(BAD) \\&= (0.99)^2 \frac{1}{2} + (0.50)^2 \frac{1}{2}\end{aligned}$$

Also,

$$\begin{aligned}P(X_1 = 1) &= P(X_1 = 1|GOOD)P(GOOD) + P(X_1 = 1|BAD)P(BAD) \\&= 0.99 \frac{1}{2} + 0.50 \frac{1}{2}\end{aligned}$$

Altogether,

$$\begin{aligned}P(X_2 = 1 | X_1 = 1) &= \frac{(0.99)^2 \frac{1}{2} + (0.50)^2 \frac{1}{2}}{0.99 \frac{1}{2} + 0.50 \frac{1}{2}} \\&= \frac{(0.99)^2 + (0.50)^2}{0.99 + 0.50} \approx 82,6\%\end{aligned}$$

## Solution 2: WIFI cont'd

A. Giovanidis 2019

The states of the two efforts two access are **not independent!**

$$82.6\% \approx P(X_2 = 1 \mid X_1 = 1) \neq P(X_2 = 1) = \frac{1}{2}(0.99 + 0.50) \approx 75\%$$

A. Giovanidis 2019

**END**