Parameter Estimation: Maximum Likelihood (ML), Maximum a Posteriori (MAP), and Bayesian Inference

Lecturer:
Jiří Matas

Authors:
Ondřej Drbohlav, Jiří Matas

Centre for Machine Perception
Czech Technical University, Prague
http://cmp.felk.cvut.cz

Last update: 11.10.2019

Probability Estimation

Both in the Bayesian Decision Theory and the Non-Bayesian Methods lectures, it has been assumed that all the necessary probabilities (priors, conditionals) are known.

In practice, the probabilities almost always need to be estimated from the training data.

Probability Distribution Estimation Methods (1/2)

According to the form of the model for the distribution:

- **Parametric**. The distribution has a known form of a function which has parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_D)$. The number of parameters is low.
  Example: the normal distribution $\mathcal{N}(x \,|\, \mu, \sigma^2)$: the parameters to be estimated are $\boldsymbol{\theta} = \{\mu, \sigma^2\}$. The parameter space is two-dimensional.
- **Non-parametric**.: The same as with the Parametric models, but the number of parameters to be estimated is very high. Note the apparent contradiction in the terminology (high number of parameters to estimate $\rightarrow$ "non-parametric" method?). This is because the term 'parameter' often disappears from the estimating methods procedure.
  Example: K-nearest neighbors; Parzen window; histogram.

  To be discussed: complexity of estimating e.g. mixtures:

$$p(x) = \sum_{i=1}^{D} \pi_i \mathcal{N}(\mu_i, \sigma_i^2)\,, \tag{1}$$

  or parameters of feed-forward neural nets.

Probability Distribution Estimation Methods (2/2)

Learning principles:

- Maximum Likelihood
- Maximum A Posteriori
- Bayesian Inference

Maximum Likelihood (ML) Principle

- The training set $\mathcal{T}$ is available, $\mathcal{T} = \{(x_1, k_1), (x_2, k_2), ..., (x_N, k_N)\}$.
- The parametric form of the likelihood $L(\boldsymbol{\theta}) = p(\mathcal{T}|\boldsymbol{\theta})$ is known.
- **Note** that the likelihood function $L(\boldsymbol{\theta})$ is a function of the parameters θ, for fixed observations $\mathcal{T}$. In particular, $L(\boldsymbol{\theta})$ does not sum up to 1.

ML principle
The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ for the observed data $\mathcal{T}$ is defined as:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{T}|\boldsymbol{\theta}). \tag{2}$$

The argument for this formulation is, informally, "if the parameters are correct then they will give larger probabilities for the observations, compared to wrong parameters".

Usually, the parameters for different classes are independent (no shared parameters between classes). In that case, the likelihood function $p(\mathcal{T}|\boldsymbol{\theta})$ can be factorized to

$$p(\mathcal{T} \,|\, \boldsymbol{\theta}) = p(\mathcal{T}_1 \,|\, \boldsymbol{\theta}_1) p(\mathcal{T}_2 \,|\, \boldsymbol{\theta}_2) ... p(\mathcal{T}_K \,|\, \boldsymbol{\theta}_K) \tag{3}$$

where $\mathcal{T}_k = \{x \colon (x, l) \in \mathcal{T} \wedge l = k\}$ is the training set for class k. The parameters $\boldsymbol{\theta}_k$ for individual classes can be estimated independently. $\Rightarrow$ **In the subsequent text, we will drop the class index k. All analysis will be done "per class".**

Maximum Likelihood (ML) Estimation

Consider the observations $\mathcal{T} = \{x_1, x_2, ..., x_N\}$ and the known parametric form of the likelihood function $L(\boldsymbol{\theta}) = p(\mathcal{T}|\boldsymbol{\theta})$. The ML estimate of $\hat{\boldsymbol{\theta}}$ is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, L(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\text{argmax}}\, p(\mathcal{T}|\boldsymbol{\theta})\,. \tag{19a}$$

- If samples in $\mathcal{T}$ are independent and identically distributed (i.i.d) then $p(\{x_1, x_2, ..., x_N\}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(x_i|\boldsymbol{\theta})$, and the ML estimate for the class is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(x_i|\boldsymbol{\theta})\,. \tag{4}$$

- The argument $\hat{\boldsymbol{\theta}}$ maximizing likelihood in Eq. (4) equals the argument maximizing the log-likelihood (as logarithm is an increasing function). This fact will be often be taken advantage of in calculations.

Example 1: Binomial Distribution (ML) (1)

There are red and green socks in the drawer. N socks have been drawn randomly from the drawer, with replacement. The result is:

$$R \text{ red socks} \tag{5}$$

$$G \text{ green socks } (G = N - R) \tag{6}$$

Compute the Maximum Likelihood estimate for the actual percentage π of the red socks in the drawer.

Analysis. For an individual draw, $P(\text{red}|\pi) = \pi$ and $P(\text{green}|\pi) = 1 - \pi$. For N independent measurements with an outcome as in Eqs. (5, 6), the likelihood is

$$p(R, N|\pi) = \binom{N}{R} \pi^R (1 - \pi)^{N-R} . \tag{7}$$

Note: Consider a training set in a slightly different form: it is an ordered sequence of observations $\mathcal{T} = (\text{red}, \text{green}, \text{green}, ..., \text{green})$, with R observations "red" and G observations "green", as before. What is the likelihood function for these observations? How does it differ from Eq. (7)? Will it matter for the ML estimation result?

Example 1: Binomial Distribution (ML) (2)

(copied from the previous slide:)

$$p(R, N|\pi) = \binom{N}{R} \pi^R (1-\pi)^{N-R} . \tag{7}$$

Taking the derivative of $p(R, N|\pi)$ with respect to π and setting it to zero gives

$$\binom{N}{R} R\, \pi^{R-1} (1-\pi)^{N-R} - \binom{N}{R} \pi^R (N-R)(1-\pi)^{N-R-1} = 0 , \tag{8}$$

and thus

$$R(1-\pi) - (N-R)\pi = 0 \tag{9}$$

which implies

$$\hat{\pi}_{ML} = \frac{R}{N} . \tag{10}$$

The ML solution is the fraction of the red socks within the socks drawn.

Example 2: Normal Distribution (ML) (1)

Let the conditional probability of a class be normal. Assume that the observations $\mathcal{T} = \{x_1, x_2, ..., x_N\}$ are i.i.d. and find the ML estimate for the mean and variance. The likelihood to be maximized is:

$$P(\mathcal{T}|\mu,\sigma) = \frac{1}{\sigma^N\sqrt{(2\pi)^N}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right]. \tag{11}$$

We require that the partial derivatives w.r.t. both μ and σ vanish:

$$\frac{\partial P(\mathcal{T}|\mu,\sigma)}{\partial\mu} = \quad P(\mathcal{T}|\mu,\sigma)\frac{1}{\sigma^2}\left(\sum_{i=1}^{N}(x_i-\mu)\right) = 0 \tag{12}$$

$$\frac{\partial P(\mathcal{T}|\mu,\sigma)}{\partial\sigma} = -P(\mathcal{T}|\mu,\sigma)\frac{N}{\sigma} + P(\mathcal{T}|\mu,\sigma)\frac{1}{\sigma^3}\left(\sum_{i=1}^{N}(x_i-\mu)^2\right) = 0 \tag{13}$$

The first and second equations imply, respectively, the terms on the right. The ML estimator for mean is the sample mean (as before in Example 1) and the ML estimator for variance is the sample variance, with sample mean Eq. (14) plugged into Eq. (15).

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i\,, \tag{14}$$

$$\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu}_{ML})^2. \tag{15}$$

Example 2: Normal Distribution (ML) (2)

Let us try now with maximizing the log-likelihood:

$$L(\mathcal{T}|\mu,\sigma) = \ln P(\mathcal{T}|\mu,\sigma) = -N\ln\sigma - \frac{N}{2}\ln 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\,. \tag{16}$$

Again, setting the partial derivatives w.r.t. μ and σ to zero yeilds

$$\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) = 0\,, \tag{17}$$

$$-\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2 = 0\,, \tag{18}$$

which leads to the same solution as before.

Maximum Likelihood—Features, Problems (1)

Why ML estimators?

Under very general conditions, ML is

- Asymptotically unbiased: as the number of observations N grows to infinity, ML estimate approaches the actual parameters $\boldsymbol{\theta}_0$ ($\lim_{N\to\infty} E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}_0$)
- Asymptotically consistent: sequence of estimates converges in probability to $\boldsymbol{\theta}_0$ as N grows to infinity ($\lim_{N\to\infty} \text{prob}\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \leq \epsilon\} = 1$)
- Asymptotically efficient
- Asymptotically normal (pdf of ML estimates as $N \to \infty$ approached Gaussian.)

Maximum Likelihood—Features, Problems (2)

With low number of observations, the ML estimates can be counter-intuitive. Consider the following examples:

- Binomial distribution, coin tossing, $\mathcal{T} = \{H, H, H\}$. The ML estimate is $\pi_{\text{head}} = 1$ (completely unfair coin). Would you believe that estimate?
- Normal distribution, estimating x-coordinate of a particle. A range of feasible μ's can be known a priori, but the sample mean taken from a few observations can be outside this range.

These examples demonstrate that employing a prior knowledge (or belief) about the parameters to be estimated would be beneficial, if available.

Maximum A Posteriori (MAP) Estimation

- The set of observations $\mathcal{T}$ is $\mathcal{T} = \{x_1, x_2, ..., x_N\}$.
- The parametric form of the likelihood function $L(\boldsymbol{\theta}) = p(\mathcal{T}|\boldsymbol{\theta})$ is known.
- The prior distribution $p(\boldsymbol{\theta})$ of the model parameters $\boldsymbol{\theta}$ is known.

MAP principle

The maximum a posteriori estimate $\hat{\boldsymbol{\theta}}$ of the distribution parameters for the observed data $\mathcal{T}$ is defined as:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta} \,|\, \mathcal{T}). \tag{19}$$

The posterior $p(\boldsymbol{\theta}|\mathcal{T})$ can be computed from $p(\mathcal{T}|\boldsymbol{\theta})$ and the prior $p(\boldsymbol{\theta})$ using the Bayes formula:

$$p(\boldsymbol{\theta}|\mathcal{T}) = \frac{p(\mathcal{T}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{T})}\,. \tag{20}$$

The denominator of Eq. (20) is *independent* of the parameters $\boldsymbol{\theta}$, and the solution $\hat{\boldsymbol{\theta}}$ can be found by maximizing the nominator only:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta} \,|\, \mathcal{T}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \frac{p(\mathcal{T}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{T})} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{T}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{21}$$

which has practical implications and shows the difference w.r.t. the ML approach: The term to be maximized is the **product of the likelihood** (as in the ML) **and the prior** on $\boldsymbol{\theta}$ which "shifts" the optimum $\hat{\boldsymbol{\theta}}$ when the number of observations is low.

Example 1, Binomial Distribution (MAP) (1)

Recall that

$$p(R, N|\pi) = \binom{N}{R} \pi^R (1-\pi)^{N-R}, \tag{7}$$

where N is the total number of socks drawn, of which R are the red ones, $G = N - R$ are the green ones and π is the percentage of red socks in the sock population to be estimated.

We need a suitable prior on π. A lucky coincidence would be if the prior $p(\pi)$ would take the same functional form in π as the above equation, that is,

$$p(\pi) \sim \pi^A (1-\pi)^B. \tag{22}$$

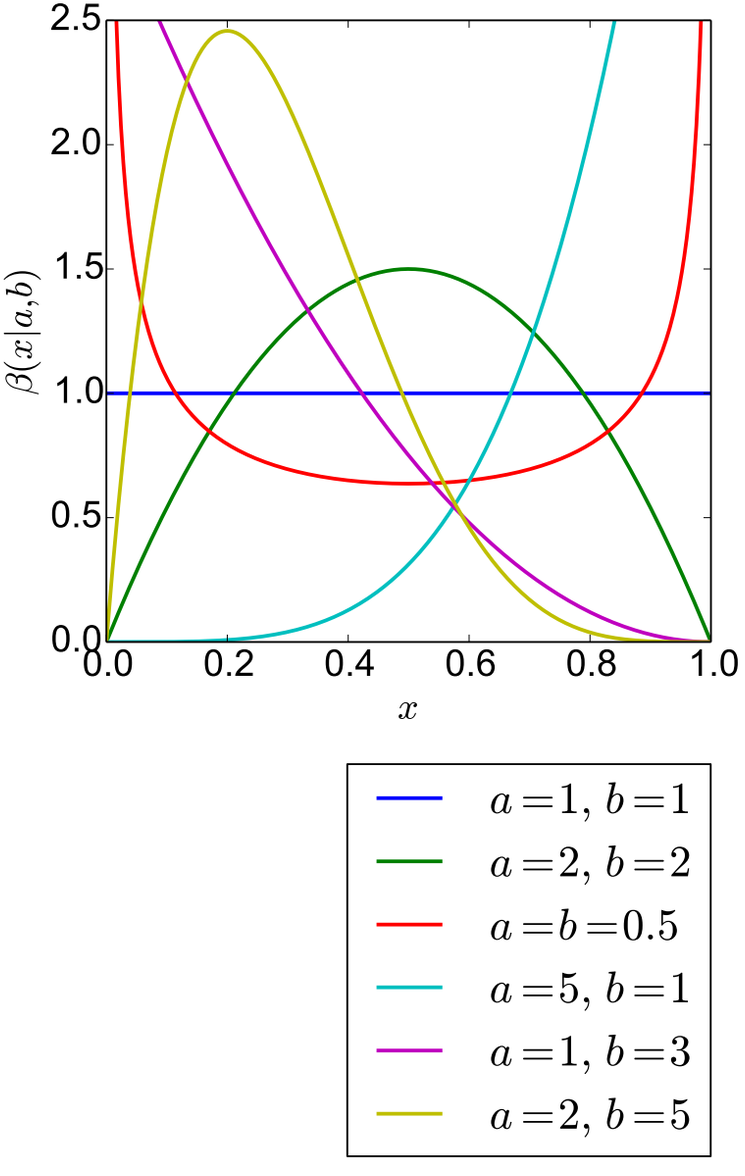
This would imply that the the product of likelihood and the prior would be

$$p(R, N|\pi)p(\pi) \sim \pi^R (1-\pi)^{N-R} \pi^A (1-\pi)^B = \pi^{R+A} (1-\pi)^{N-R+B}. \tag{23}$$

Such prior $p(\pi)$ is known under the name Beta distribution. The maximization of this term is already done, as due to this functional form it is the same as the ML solution for $(R + A)$ red socks out of the total number of $(N + A + B)$. Thus, for such a prior,

$$\hat{\pi}_{MAP} = \frac{R + A}{N + A + B}. \tag{24}$$

Example 1, Binomial Distribution (MAP) (2)



The prior distribution $p(\pi) \sim \pi^A(1-\pi)^B$ is known as the **Beta distribution**, and is defined as:
(note the subtle change $A \to a-1, B \to b-1$)

$$\beta(\pi|a,b) = \frac{\pi^{a-1}(1-\pi)^{b-1}}{\int_0^1 \pi^{a-1}(1-\pi)^{b-1}\mathrm{d}\pi} = \frac{1}{B(a,b)}\pi^{a-1}(1-\pi)^{b-1} \tag{25}$$

where $B(a,b)$, the normalizing constant, is the Beta function. Using the β distribution, the term $p(R,N|\pi)p(\pi)$ can be rewritten as

$$p(R,N|\pi)p(\pi) \sim \pi^{R+A}(1-\pi)^{N-R+B} \tag{26}$$

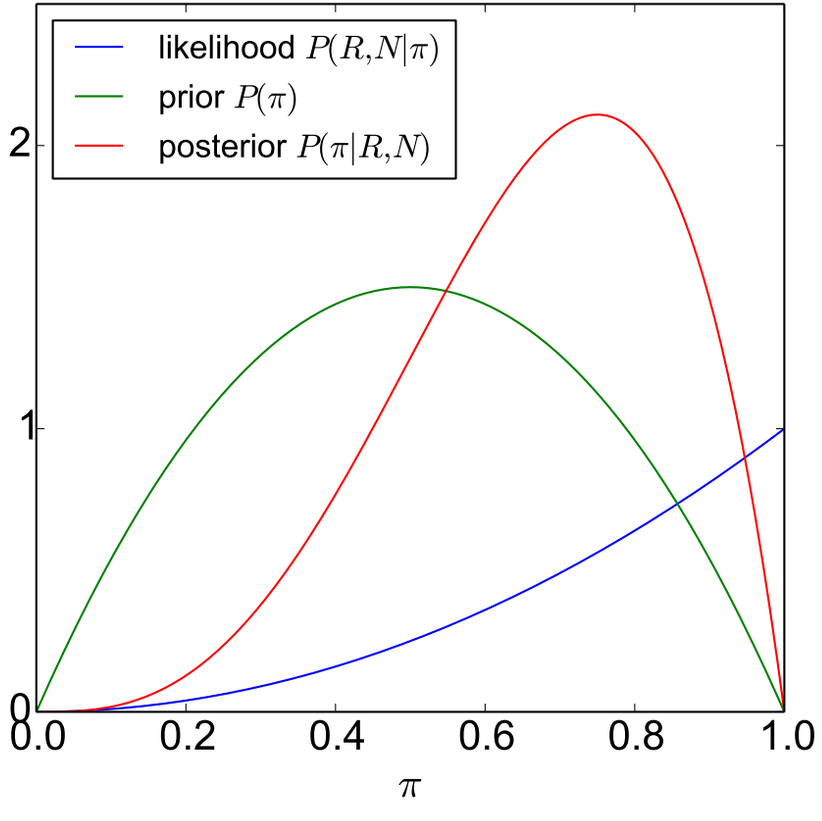
$$\sim \beta(R+A+1, N-R+B+1)\,. \tag{27}$$

Note that, indeed, the **posterior** $p(\pi|R,N) = \beta(R+A+1, N-R+B+1)$.

Example 1, Binomial Distribution (MAP) (3)

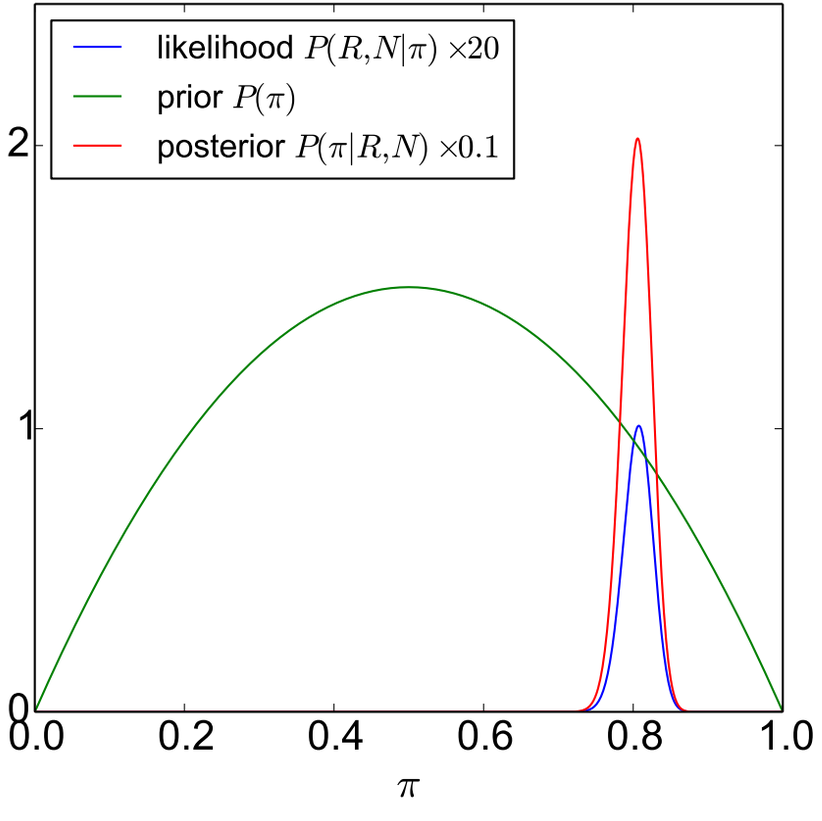
Examples:

- $N = 2$ socks drawn
- both of them are red $(R = 2)$
- the prior is set to $\beta(r|2, 2) \sim r(1 - r)$



- $\hat{\pi}_{ML} = 2/2 = 1$
- $\hat{\pi}_{MAP} = (2 + 1)/(2 + 2) = 3/4$

- $N = 400$ socks drawn
- of which $R = 323$ are red ones
- the prior is set to $\beta(r|2, 2) \sim r(1 - r)$



- $\hat{\pi}_{ML} = 323/400$
- $\hat{\pi}_{MAP} = 324/401$

Example 1, Binomial Distribution (MAP) (4)

$$\hat{\pi}_{MAP} = \frac{R + A}{N + A + B} \tag{24}$$

Note that the parameters A, B of the prior $p(\pi) \sim \pi^A (1 - \pi)^B$ behave as "*virtual*" observations; it is as if A red socks and B green socks have been already observed before any real observation has been done.

The parameters of the prior (here A, B) are generally called **hyperparameters**. This name distinguishes them from the parameters of the probabilistic model which are to be estimated.

Example 2: Normal Distr. with unknown μ (1)

Consider the normal distribution $p(\mathcal{T} \,|\, \mu, \sigma^2)$, and for simplicity let the variance σ^2 be known and equal to $\sigma^2 = 1$. Estimate the mean μ.
The likelihood is

$$L(\mu) = p(\mathcal{T}|\mu) = \prod_{i=1}^{N} p(x_i|\mu) = \frac{1}{\sqrt{(2\pi)^N}} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2\right]. \tag{28}$$

Consider the prior $p(\mu)$ on μ to be distributed normally:

$$p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{1}{2}\frac{(\mu - \mu_0)^2}{\sigma_0^2}\right]. \tag{29}$$

The MAP estimate of μ will be found as

$$\mu_{MAP} = \underset{\mu}{\text{argmax}}\, p(\mathcal{T}|\mu)p(\mu)\,. \tag{30}$$

Example 2: Normal Distr. with Unknown μ (MAP) (2)

The exponent of the likelihood $p(\mathcal{T}|\mu)$ (recall $\sigma^2 = 1$ is fixed and known) can be rewritten as

$$-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2 = -\frac{1}{2}\left(\sum_{i=1}^{N}x_i^2 - 2\mu\sum_{i=1}^{N}x_i + N\mu^2\right) = \tag{31}$$

$$-\frac{1}{2}N\left(\mu^2 - 2\mu\underbrace{\frac{\sum_{i=1}^{N}x_i}{N}}_{\mu'} + C_1\right) = -\frac{N}{2}\left(\mu-\mu'\right)^2 + C_2\,, \tag{32}$$

where μ' is the sample mean of $\mathcal{T}$ and C_1 and C_2 are constants independent of μ. The posterior $p(\mu|\mathcal{T})$ can then be written again as a normal distribution (mean μ_c, var. σ_c^2):

$$p(\mu|\mathcal{T}) \sim \frac{\exp\left[-\frac{1}{2}N(\mu-\mu')^2\right]\mathcal{N}\left(\mu|\mu_0,\sigma_0^2\right)}{Z'(\mathcal{T})} = \frac{\mathcal{N}\left(\mu|\mu',\frac{1}{N}\right)\mathcal{N}\left(\mu|\mu_0,\sigma_0^2\right)}{Z''(\mathcal{T})} = \mathcal{N}\left(\mu|\mu_c,\sigma_c^2\right) \tag{33}$$

where $Z'(\mathcal{T})$ and $Z''(\mathcal{T})$ are suitable normalizing constants.

Example 2: Normal Distr. with Unknown μ (MAP) (3)

It is easy to see that making a product of two normal distributions $\mathcal{N}\left(\mu|\mu_1,\sigma_1^2\right)$ and $\mathcal{N}\left(\mu|\mu_2,\sigma_2^2\right)$ and normalizing it results in another normal distribution $\mathcal{N}\left(\mu|\mu_c,\sigma_c^2\right)$, as looking at the exponent of the product, there are terms:

$$\frac{(\mu-\mu_1)^2}{\sigma_1^2}+\frac{(\mu-\mu_2)^2}{\sigma_2^2}=\mu^2\left[\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right]-2\mu\left[\frac{\mu_1}{\sigma_1^2}+\frac{\mu_2}{\sigma_2^2}\right]+D_1 \tag{34}$$

$$=\left[\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right]\left(\mu-\left[\frac{\mu_1}{\sigma_1^2}+\frac{\mu_2}{\sigma_2^2}\right]/\left[\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right]\right)^2+D_2 \tag{35}$$

where D_1 and D_2 are constants which do not even need to be evaluated, as they will be factored into the normalization provided by the term for normal distribution itself. Pairing the parameters and the terms, we obtain

$$\mu_c=\left[\frac{\mu_1}{\sigma_1^2}+\frac{\mu_2}{\sigma_2^2}\right]/\left[\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right] \qquad \sigma_c^2=\left[\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right]^{-1}. \tag{36}$$

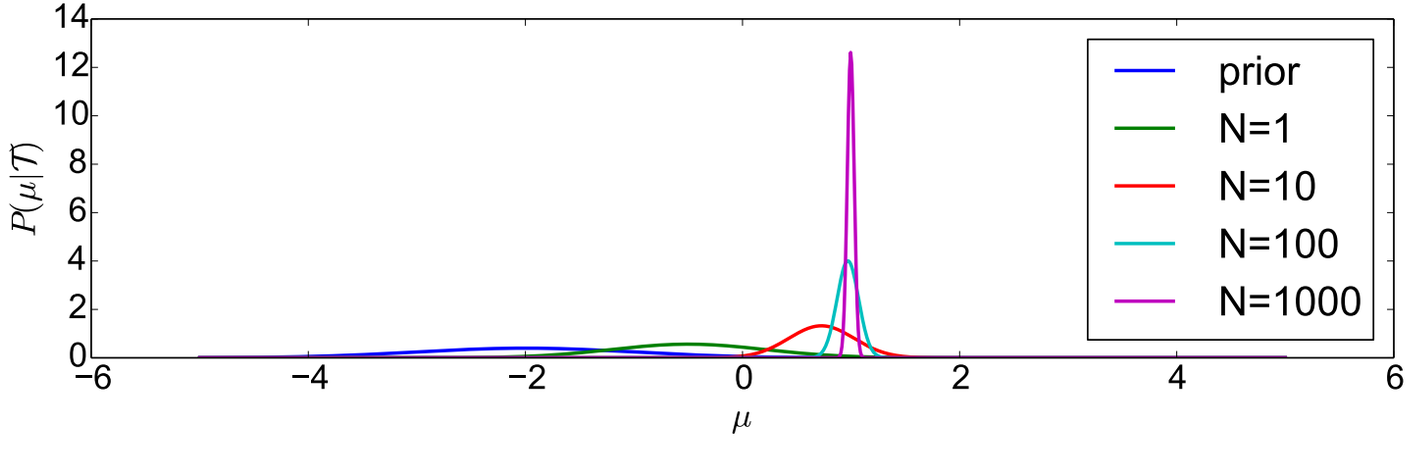
Thus for the case studied here,

$$P(\mu|\mathcal{T})=\frac{\mathcal{N}\left(\mu|\mu',\frac{1}{N}\right)\mathcal{N}\left(\mu|\mu_0,\sigma_0^2\right)}{Z''(\mathcal{T})}=\mathcal{N}\left(\mu|\frac{\left[N\mu'+\frac{\mu_0}{\sigma_0^2}\right]}{[N+1/\sigma_0^2]},\frac{1}{N+1/\sigma_0^2}\right). \tag{37}$$

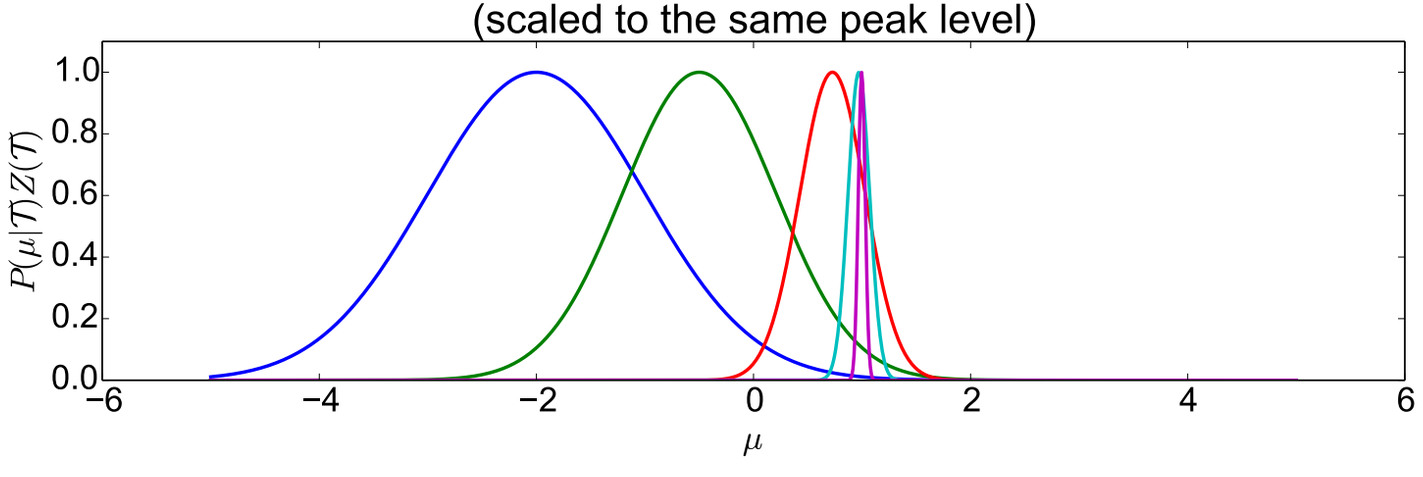
Example 1: Normal Distr. with unknown μ, Conj. Prior (4)

Thus when the likelihood is normal in μ and the prior is also normal in μ, the posterior is normal in μ as well.

Example: The posterior of μ for for sample mean $\mu' = 1$ and the prior $p(\mu) = \mathcal{N}\left(\mu|\mu_0 = -2, \sigma_0^2 = 1\right)$, for different sizes of the observation set:



(scaled to the same peak level)



Conjugate Prior

In both Example 1 and Example 2, the priors had a functional form which was similar to the form of the likelihood.

This enabled us to make an easy derivation of the MAP formulas.

In general, such "suitable" priors are called **conjugate**.

Bayesian Inference

- The requirements are the same as in MAP (it is necessary to know $p(\mathcal{T}|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$)
- ML and MAP search for the maximum of likelihood and likelihood combined with prior, respectively.
- The Bayesian Inference does not pick such "best" solution, but instead minimizes the risk $R(\theta)$ of the estimate θ (quadratic loss function and one-dimensional estimation problem considered here):

$$R(\theta) = \int_{-\infty}^{\infty} p(t|\mathcal{T})(t-\theta)^2 \mathrm{d}t \tag{38}$$

$$\theta_{BI} = \operatorname*{argmin}_{\theta} R(\theta) \tag{39}$$

- This leads to

$$\theta_{BI} = \int_{-\infty}^{\infty} t\, p(t|\mathcal{T}) \mathrm{d}t\,. \tag{40}$$

- It is very convenient when the prior has a suitable form (*conjugate prior*.)

Example 1, Binomial Distribution (Bayes Inference)

We know (from the MAP analysis on slide 15) that for the prior $p(\pi) = \beta(A+1, B+1)$, the posterior $p(\pi|R,N)$ is

$$p(\pi|R,N) = \beta(R+A+1, [N-R]+B+1)\,. \tag{41}$$

The estimate π_{BI} obtained by Bayesian Inference is

$$\pi_{BI} = \int_0^1 \pi\, p(\pi|R,N)\mathrm{d}\pi = \tfrac{R+A+1}{N+A+B+2}\,, \tag{42}$$

where we have used the known fact about the β distribution that the expected value for $\beta(a,b)$ is $a/(a+b)$.

Example: Consider $R=2$, $N=2$ and the **uniform** prior, $p(\pi) = r^0(1-r)^0$ (thus $A=0, B=0$.) Then $\pi_{BI} = (R+1)/(N+2) = 3/4$ $\quad(\pi_{ML} = \pi_{MAP} = R/N = 1.)$

Useful known properties of the Beta distribution $\beta(a,b)$:

- mode (= maximum value): $\frac{a-1}{a+b-2}$ (agrees with our computations)
- mean (expected value): $\frac{a}{a+b}$
- variance: $\frac{ab}{(a+b)^2(a+b+1)}$

Estimator Properties: Bias (1/5)

"Is an estimator biased?"

This question has the following meaning. Let us say that we observe $N = 5$ data points $x_1, x_2, ..., x_5$ and estimate $\hat{\mu}_{ML}$ and $\hat{\sigma}^2_{ML}$ from them. These estimates will most likely be different from the true parameters μ and σ of the distribution $\mathcal{N}(x|\mu, \sigma)$. But how about if we do this repeatedly, that is:

for i=1 to K **do**
  get a 5-tuple
  compute $\mu_{ML}(i)$, $\sigma^2_{ML}(i)$
end for
Average the obtained values: $\mu'_{ML} = \frac{1}{K}\sum_{i=1}^{K}\mu_{ML}(i)$, $\sigma'^2_{ML} = \frac{1}{K}\sum_{i=1}^{K}\sigma^2_{ML}(i)$

If such a procedure produces true parameter values in the limit as $K \to \infty$ then the estimator is unbiased. Otherwise, it is biased.

Mathematically, (for an example of μ) this is written as

$$\mu - \mathrm{E}\left[\frac{1}{5}\sum_{i=1}^{5} x_i\right] = 0 \text{ iff the estimator is unbiased.} \tag{43}$$

where E is the expected value operator which integrates over the entire distribution of 5-tuples.

Estimator Properties: Bias (2/5), Sample Mean

This expected value is:

$$\mathrm{E}\left[\frac{1}{5}\sum_{i=1}^{5} x_i\right] = \iiiiint_{-\infty}^{\infty} \underbrace{\frac{1}{5}(x_1 + ... + x_5)}_{\text{estimator}} \underbrace{\mathcal{N}(x_1|\mu,\sigma)\,\mathcal{N}(x_2|\mu,\sigma)\cdot ... \cdot \mathcal{N}(x_5|\mu,\sigma)}_{\text{5-tuples distribution}} \mathrm{d}x_1 ... \mathrm{d}x_5 \tag{44}$$

$$= \frac{1}{5}\underbrace{\int_{-\infty}^{\infty} x_1 \mathcal{N}(x_1|\mu,\sigma)\,\mathrm{d}x_1}_{\mu} \underbrace{\int_{-\infty}^{\infty} \mathcal{N}(x_2|\mu,\sigma)\,\mathrm{d}x_2 \cdot ... \cdot \int_{-\infty}^{\infty} \mathcal{N}(x_5|\mu,\sigma)\,\mathrm{d}x_5}_{1}$$

$$+ 4 \text{ analogous terms} \tag{45}$$

$$= \frac{1}{5} 5\mu = \mu\,. \tag{46}$$

As the expected value of the sample mean estimator is μ, the estimator is unbiased. The same is obviously true for arbitrary N.

Estimator Properties: Bias (3/5), Sample Variance

Is the variance estimator also unbiased? The expected value of the estimator is

$$\mathrm{E}\left[\sigma^2_{ML}\right] = \int \ldots \int_{-\infty}^{\infty} \underbrace{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2}_{\text{estimator}} \mathcal{N}\left(x_1|\mu,\sigma\right) \ldots \mathcal{N}\left(x_N|\mu,\sigma\right) \mathrm{d}x_1 \ldots \mathrm{d}x_N \,. \tag{47}$$

The estimator is

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left((x_i - \mu) - \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)\right)^2 \tag{48}$$

and thus by substituting $x_i \leftarrow (x_i - \mu)$ Eq. (47) is rewritten as

$$\mathrm{E}\left[\sigma^2_{ML}\right] = \int \ldots \int_{-\infty}^{\infty} \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j\right)^2 \mathcal{N}\left(x_1|0,\sigma\right) \ldots \mathcal{N}\left(x_N|0,\sigma\right) \mathrm{d}x_1 \ldots \mathrm{d}x_N \,. \tag{49}$$

Estimator Properties: Bias (4/5), Sample Variance

(copied from the previous slide:)

$$\mathrm{E}\left[\sigma_{ML}^2\right] = \int \ldots \int_{-\infty}^{\infty} \frac{1}{N} \sum_{i=1}^{N} \left(x_i - \frac{1}{N} \sum_{j=1}^{N} x_j \right)^2 \mathcal{N}\left(x_1|0,\sigma\right) \ldots \mathcal{N}\left(x_N|0,\sigma\right) \mathrm{d}x_1 \ldots \mathrm{d}x_N \,. \tag{49}$$

Next, we will use the identity

$$\frac{1}{N} \sum_{i=1}^{N} \left(x_i - \frac{1}{N} \sum_{j=1}^{N} x_j \right)^2 = \underbrace{\frac{1}{N} \sum_{i=1}^{N} x_i^2}_{T_1} - \underbrace{\left(\frac{\sum_{i=1}^{N} x_i}{N} \right)^2}_{T_2} \tag{50}$$

$\mathrm{E}[x_k^2]$ is σ^2 by an analogous construction as used in the derivation of expected value of mean. Thus,

$$\mathrm{E}[T_1] = \mathrm{E}\left[\frac{1}{N} \sum_{i=1}^{N} x_i^2 \right] = \frac{1}{N} N \sigma^2 = \sigma^2. \tag{51}$$

Estimator Properties: Bias (5/5), Sample Variance

As for T_2, it is sufficient to note that

$$\mathrm{E}\left[\left(\sum_{i=1}^{N} x_i\right)^2\right] = \sum_{i=1}^{N} \underbrace{\mathrm{E}[x_i^2]}_{\sigma^2} - \sum_{i \neq j} \underbrace{\mathrm{E}[x_i x_j]}_{0} \tag{52}$$

Taking the two results together,

$$\mathrm{E}[\sigma_{ML}^2] = \sigma^2 - \frac{1}{N}\sigma^2 = \frac{N-1}{N}\sigma^2. \tag{53}$$

The ML estimator for the variance σ^2 is thus *biased.*

The unbiased version of the variance estimator is

$$\frac{N}{N-1}\sigma_{ML}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_{ML})^2\,, \tag{54}$$

where μ_{ML} is the sample mean $\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

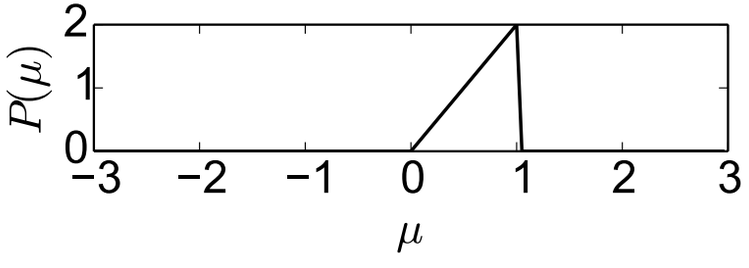
Example 3: Normal Distr. with Unknown μ (MAP), Non-conjugate prior (1)

Consider again the normal distribution $p(\mathcal{T} \,|\, \mu, \sigma^2)$, and let the variance σ^2 be known ($\sigma^2 = 1$.) Estimate the mean μ. As before, we have

$$L(\mu) = p(\mathcal{T}|\mu) = \prod_{i=1}^{N} p(x_i|\mu) = \frac{1}{\sqrt{(2\pi)^N}} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2\right]. \tag{55}$$

Let us consider the prior on μ to have the form (note: this is clearly not a conjugate prior)

$$p(\mu) = \begin{cases} 2\mu, & 0 < \mu \leq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{56}$$

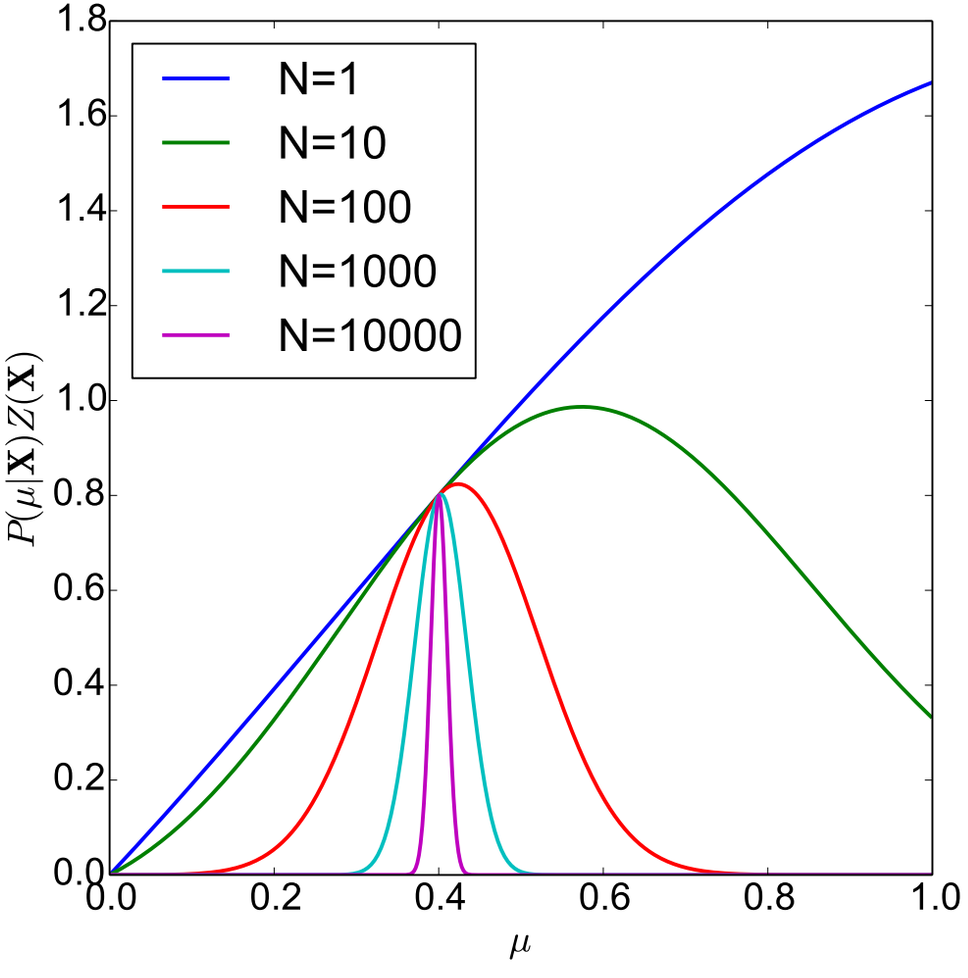


The MAP estimate of μ will be found as

$$\mu_{MAP} = \underset{\mu}{\text{argmax}}\, p(\mathcal{T}|\mu)p(\mu)\,. \tag{57}$$

Note that $p(\mathcal{T}|\mu)\,p(\mu)$ can attain maximum either inside the inverval $0 < \mu < 1$, or at its border $\mu = 1$ (not at the other border $\mu = 0$, as $p(0) = 0$.)

Example 3: Normal Distr. with Unknown μ (MAP), Non-conjugate prior (2)



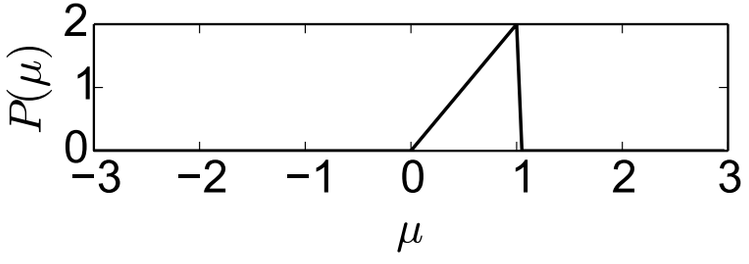
Left: $p(\mathcal{T} \mid \mu)p(\mu)$ evaluated for sample mean $\mu' = 0.4$ and increasing cardinality of the observation set $\mathcal{T}$. Note:

- variance of the distribution decreases as N grows;
- the distribution is quite close to the prior for $N = 1$;
- influence of the prior decreases with increasing N.

Example 3: Normal Distr. with Unknown μ (MAP), Non-conjugate prior (3)

$$L(\mu) = p(\mathcal{T}|\mu) = \frac{1}{\sqrt{(2\pi)^N}} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2\right]. \tag{55}$$

$$p(\mu) = \begin{cases} 2\mu, & 0 < \mu \leq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{56}$$



Taking the log of $p(\mathcal{T}|\mu)p(\mu)$ gives (for the interval $0 < \mu < 1$):

$$\ln p(\mathcal{T}|\mu)p(\mu) = \ln p(\mathcal{T}|\mu) + \ln p(\mu) = -N/2 \ln 2\pi - \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2 + \ln 2\mu \tag{58}$$

Taking the derivative w.r.t. μ and setting it to zero gives

$$\frac{\partial \ln p(\mathcal{T}|\mu)p(\mu)}{\partial \mu} = \sum_{i=1}^{N}(x_i - \mu) + \frac{1}{\mu} = 0\,, \qquad (0 < \mu < 1). \tag{59}$$

Note that this is a decreasing function of μ and thus there can be **at most one** solution for μ in the considered interval.

Example 3: Normal Distr. with Unknown μ (MAP), Non-conjugate prior (4)

Denoting $S = \sum_{i=1}^{N} x_i$, this is rewritten as

$$S - N\mu + \frac{1}{\mu} = 0 \Rightarrow N\mu - \frac{1}{\mu} = S, \qquad (0 < \mu < 1). \tag{60}$$

It is easily checked that for any S, $\mu > 0$ can be found such that this equation holds. Taken with the previous observation that there is at most 1 solution, there is exactly 1 solution for $\mu > 0$.

Multiplying by μ, we get

$$N\mu^2 - \mu S - 1 = 0. \tag{61}$$

The roots of this quadratic equation are

$$\mu = \frac{S \pm \sqrt{S^2 + 4N}}{2N} = \frac{1}{2}\frac{S}{N} \pm \frac{1}{2}\sqrt{\frac{S^2}{N^2} + \frac{4}{N}} = \begin{cases} \mu^+ > 0 \\ \mu^- < 0 \end{cases} . \tag{62}$$

Only μ^+ (always > 0) can be the solution of Eq. (60) if $\mu^+ < 1$. The root μ^- can never be the solution to it, as it is always < 0.

If $\mu^+ < 1$ then $\mu_{MAP} = \mu^+$. Otherwise the maximum is attained at the right border of the interval, and $\mu_{MAP} = 1$.

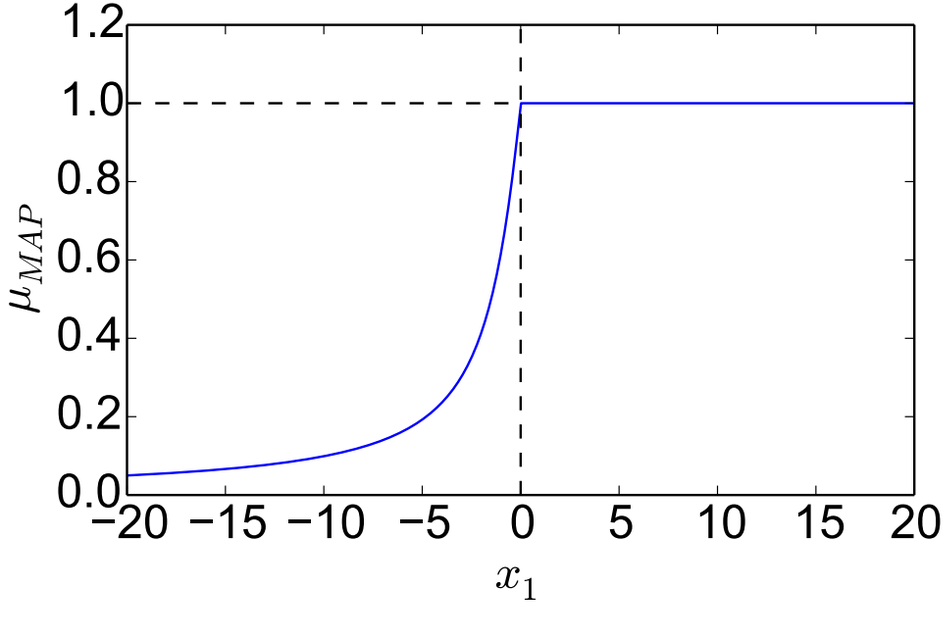
Example 3: Normal Distr. with Unknown μ (MAP), Non-conjugate prior (5)

In conclusion, the MAP solution is the following:

- Compute $S = \sum_{i=1}^{N} x_i$
- Compute $\mu^+ = \frac{1}{2}\frac{S}{N} + \frac{1}{2}\sqrt{\frac{S^2}{N^2} + \frac{4}{N}}$.
- If $\mu^+ < 1$ then $\mu_{MAP} = \mu^+$ else $\mu_{MAP} = 1$.

Example: Consider the training set consisting of a single observation $\mathcal{T} = \{x_1\}$.

The estimate μ_{MAP}:



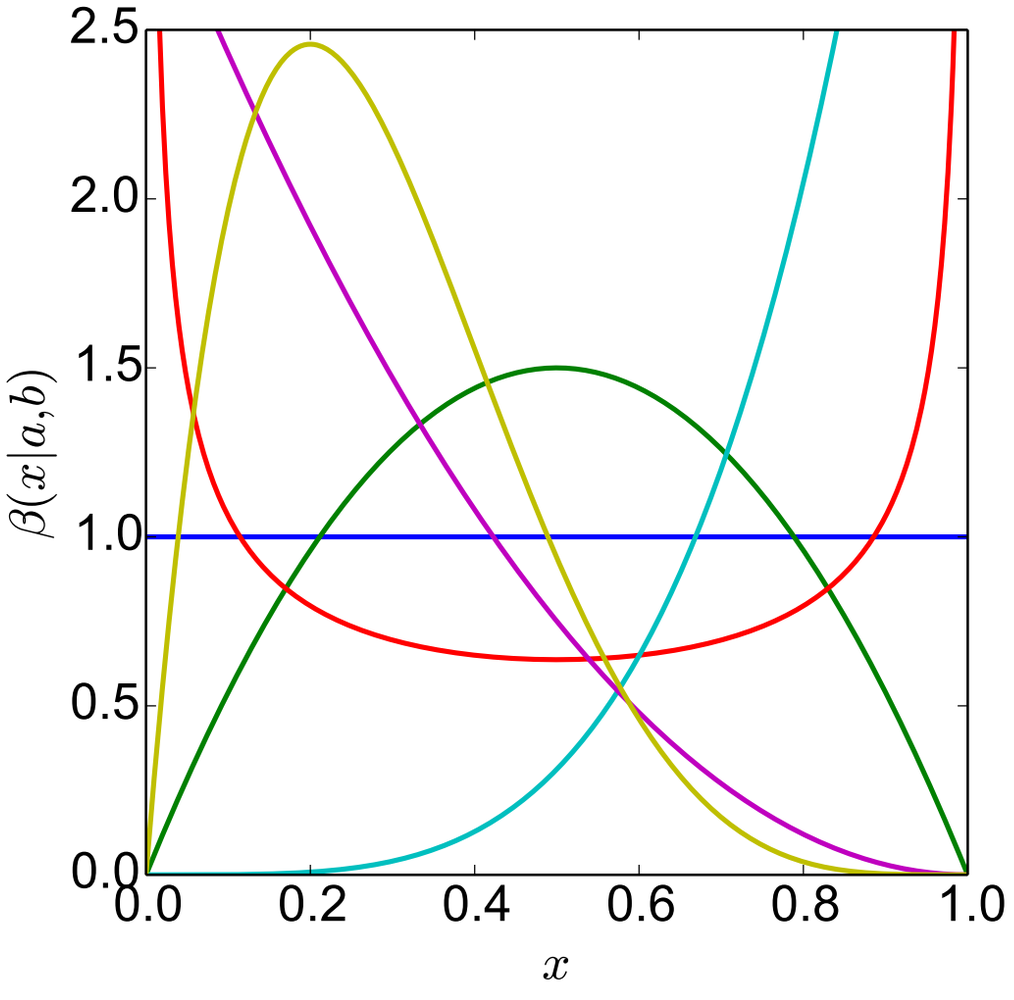
Note that for all $x_1 > 0$, $\mu_{MAP} = 1$.

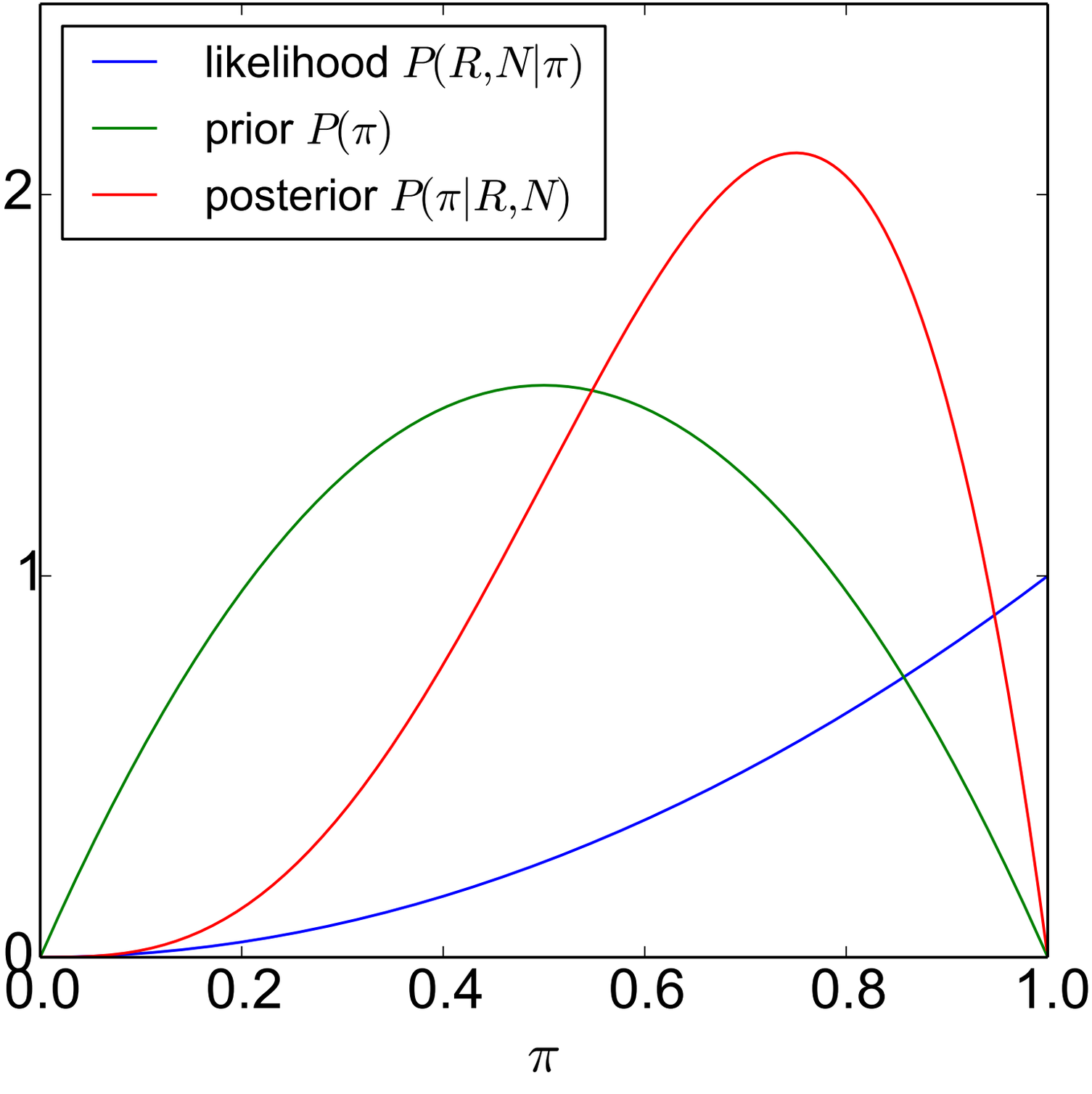
Also note that, as $N \to \infty$,

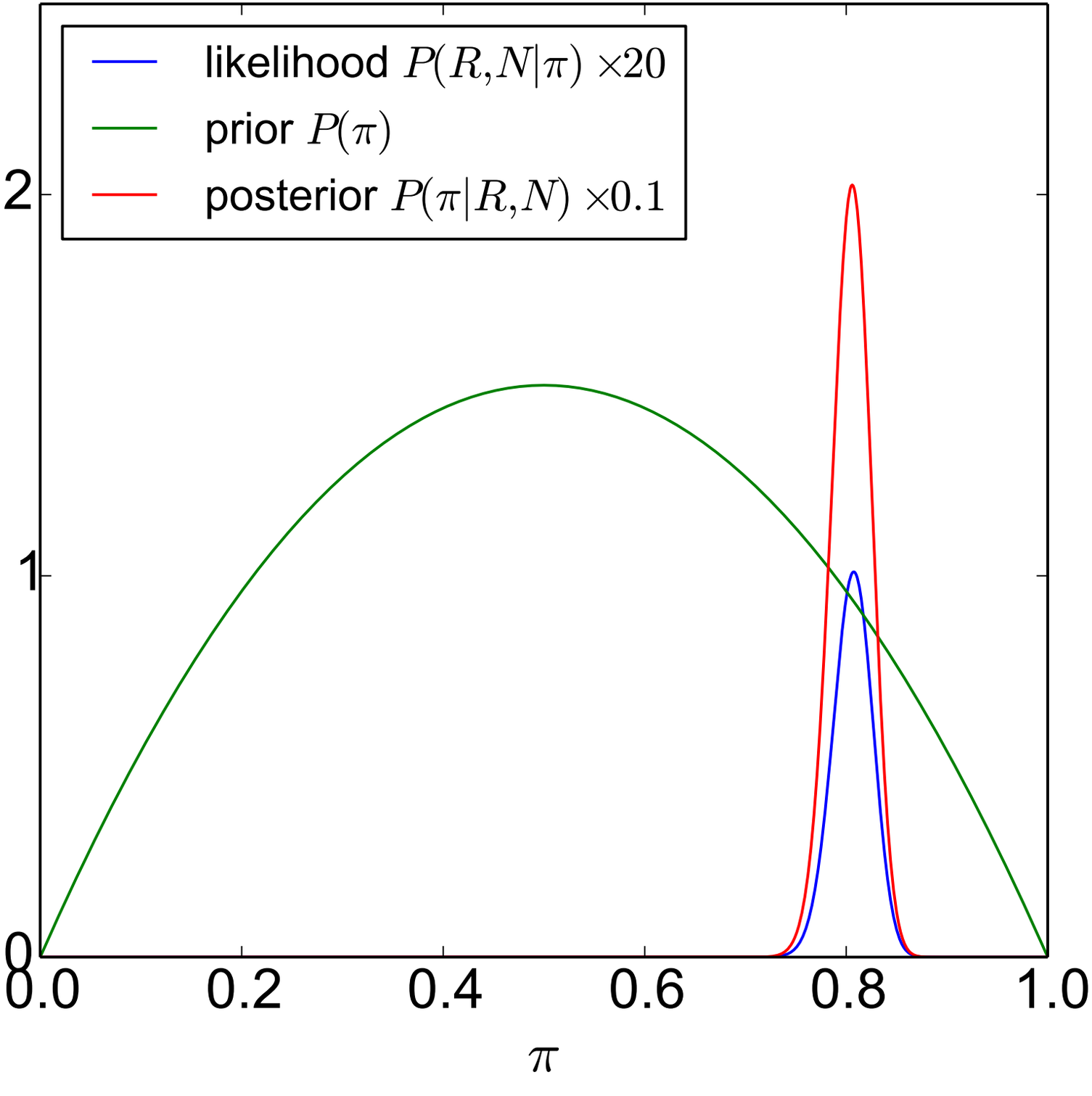
$$\mu^+ \to \frac{1}{2}\left(\frac{S}{N} + \frac{|S|}{N}\right). \tag{63}$$

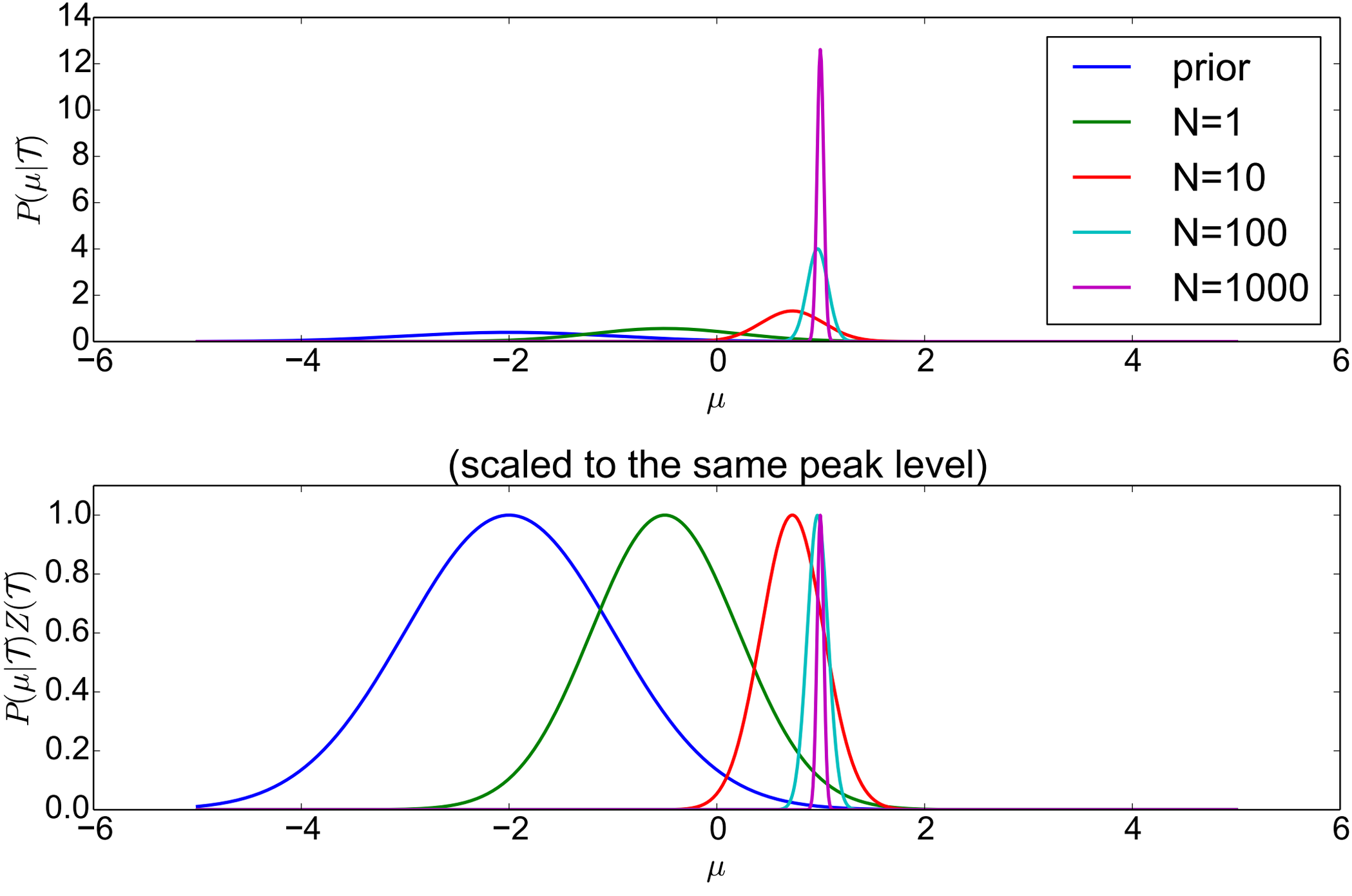
For a high number of observations and $S > 0$, the solution will converge to the ML solution $\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N}$. This is the usual behavior of MAP vs. ML: The prior distribution of parameters (here μ) shifts the solution towards the values which are a priori more probable, but as evidence grows the influence of the prior shades away.

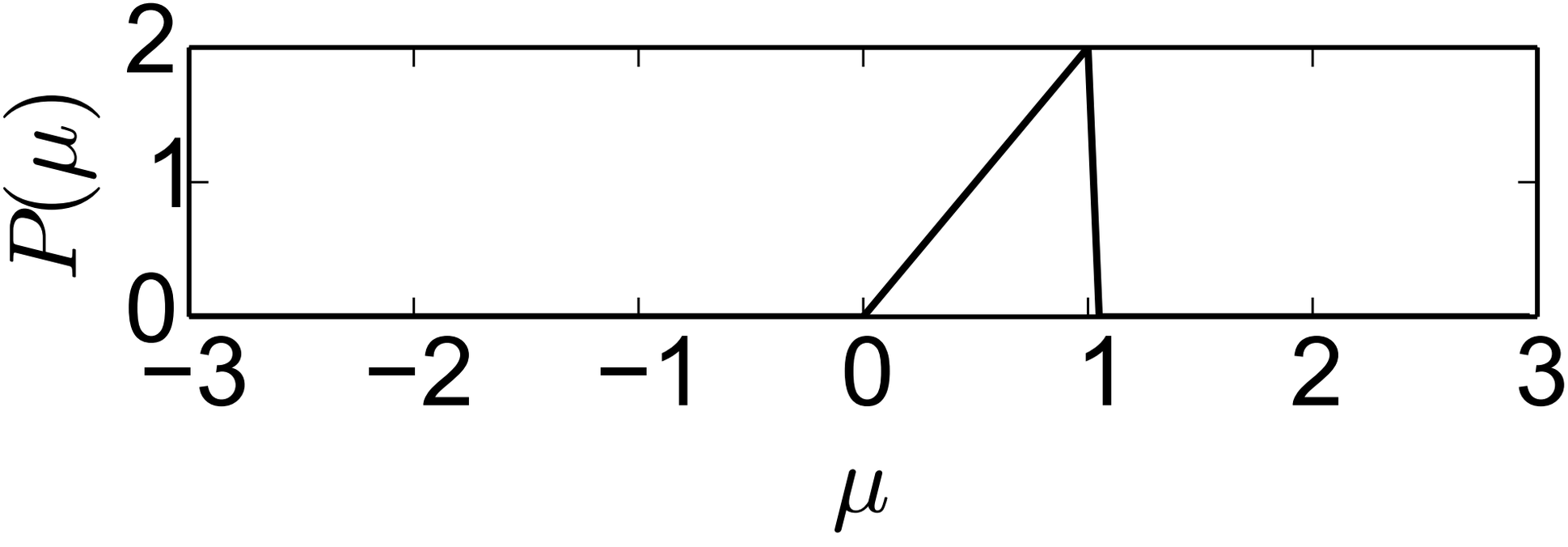

m p

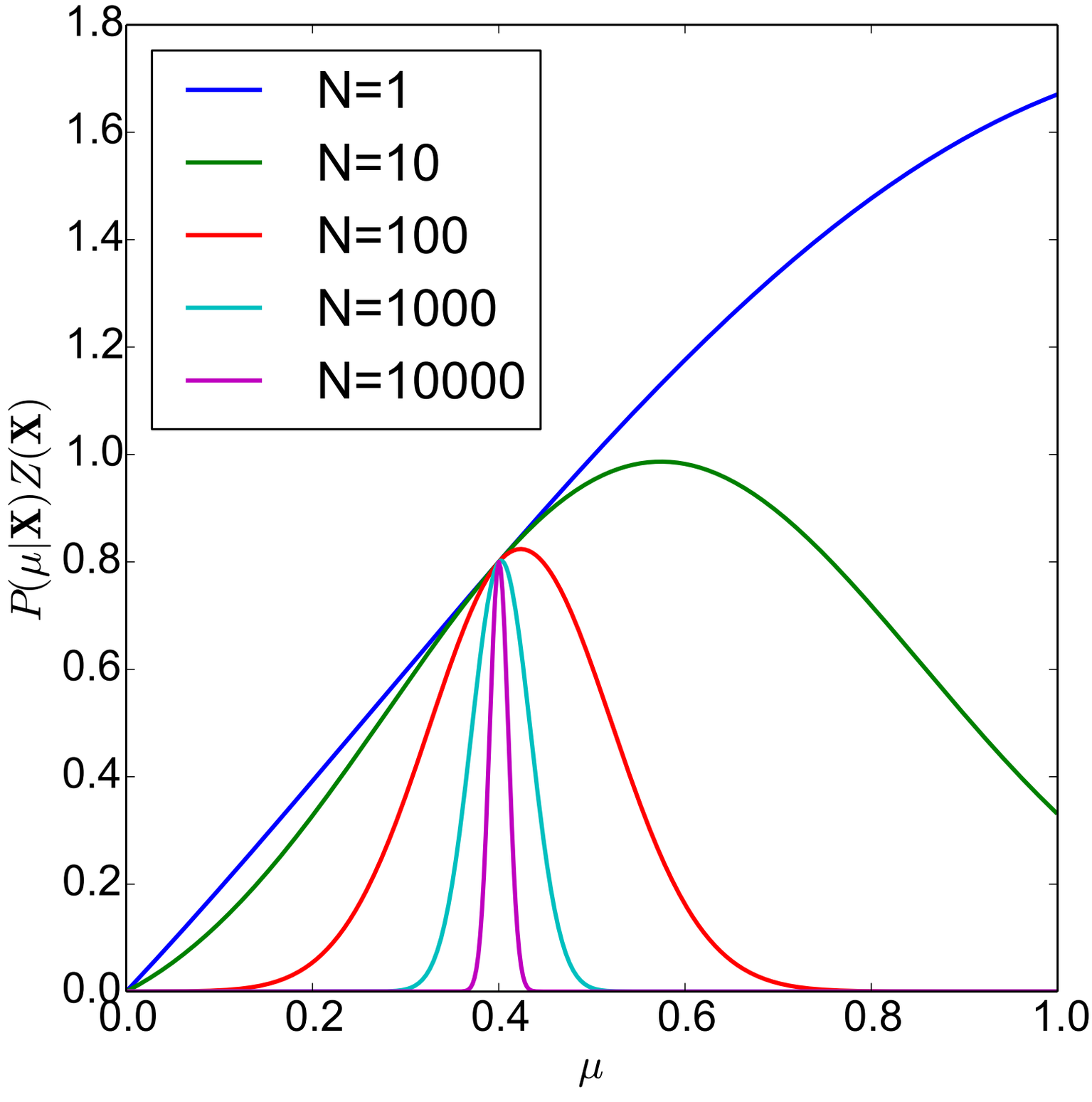

2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
$\beta(x|a,b)$
x
$a=1,\ b=1$
$a=2,\ b=2$
$a=b=0.5$
$a=5,\ b=1$
$a=1,\ b=3$
$a=2,\ b=5$

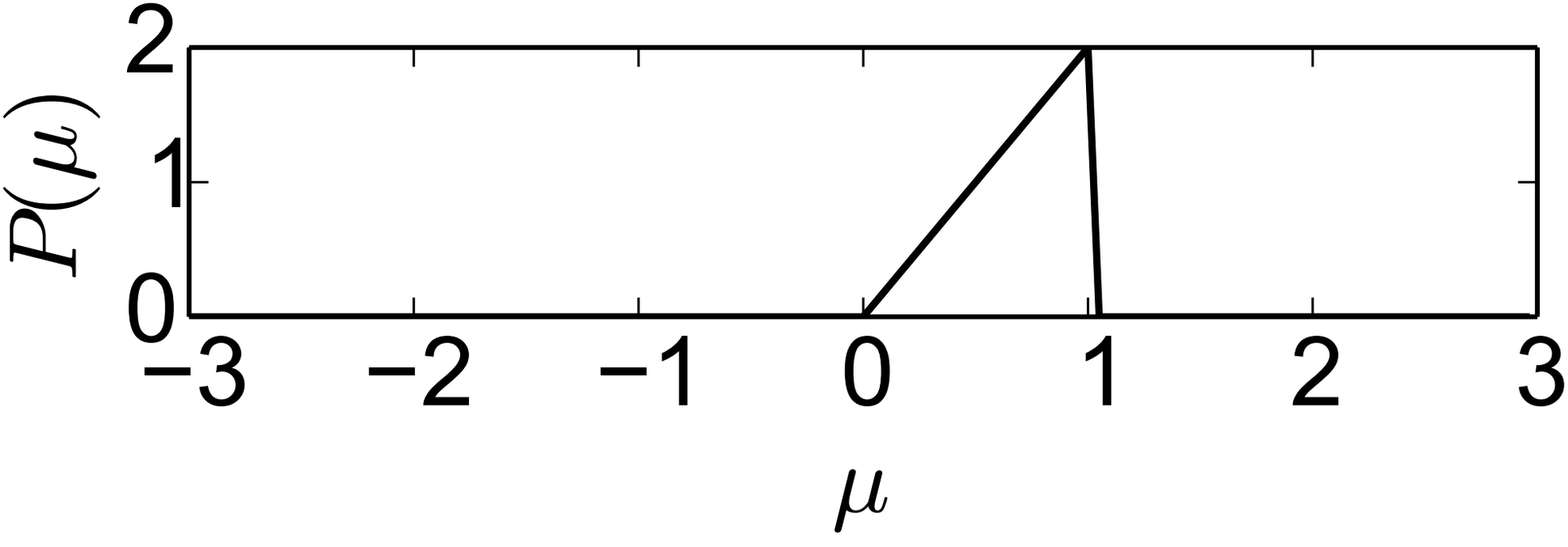

likelihood $P(R,N|\pi)$
prior $P(\pi)$
posterior $P(\pi|R,N)$
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
π

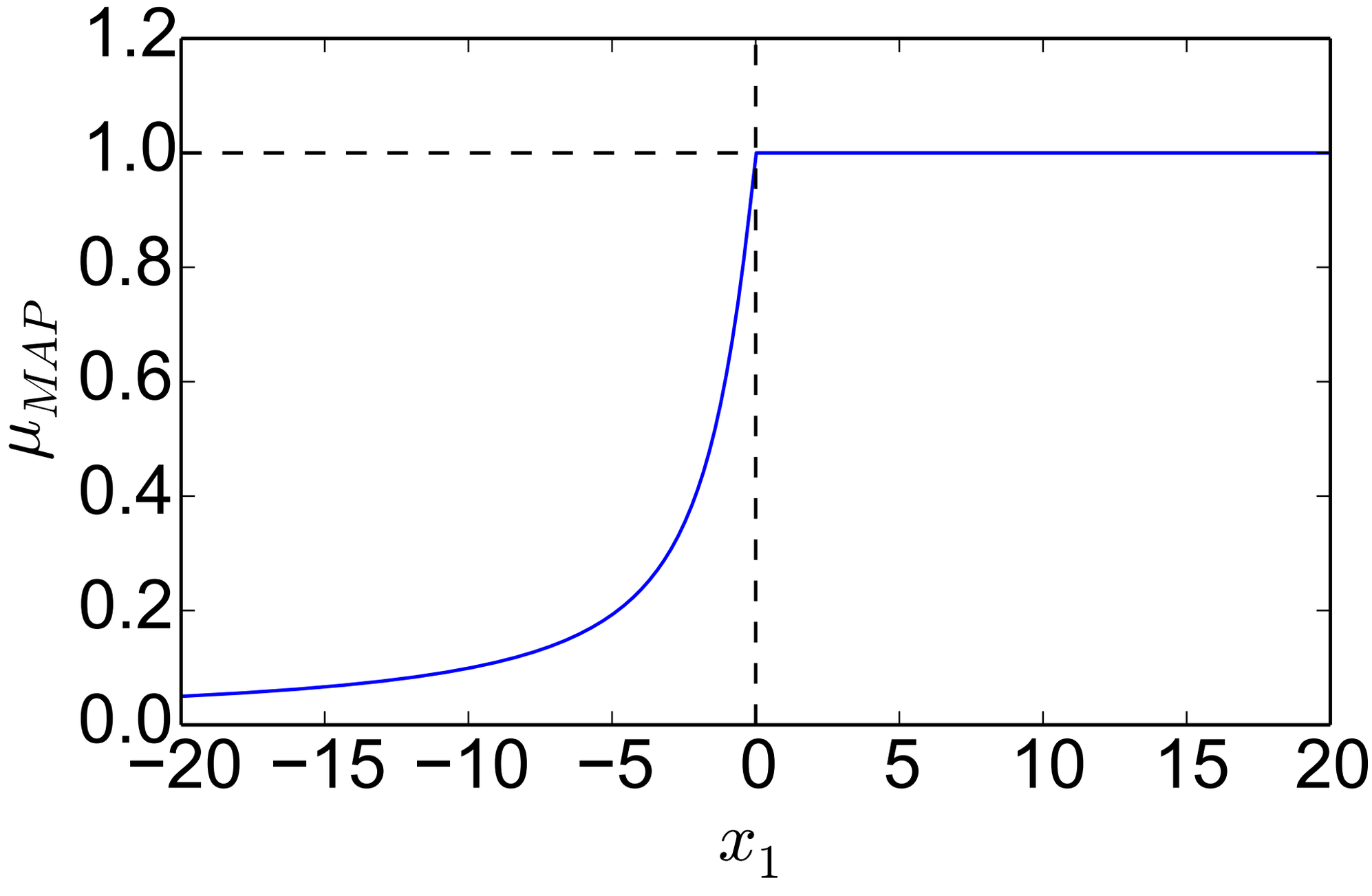

likelihood $P(R,N|\pi)$ ×20
prior $P(\pi)$
posterior $P(\pi|R,N)$ ×0.1
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
π


prior
N=1
N=10
N=100
N=1000
$P(\mu|\mathcal{T})$
μ
(scaled to the same peak level)
$P(\mu|\mathcal{T})Z(\mathcal{T})$
μ


2
1
0
$P(\mu)$
−3
−2
−1
0
1
2
3
μ


N=1
N=10
N=100
N=1000
N=10000
$P(\mu|\mathbf{X})Z(\mathbf{X})$
μ
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8


2
1
0
$P(\mu)$
−3
−2
−1
0
1
2
3
μ


1.2
1.0
0.8
0.6
0.4
0.2
0.0
μ_{MAP}
−20
−15
−10
−5
0
5
10
15
20
x_1