Maximum a posteriori estimation (**MAP**)

MAP

- A maximum a posteriori probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution.

- The MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data.

- It is closely related to the method of maximum likelihood **(ML**) estimation.

- The difference lies as it employs an augmented optimization objective which incorporates a prior distribution (that quantifies the additional information available through prior knowledge of a related event) over the quantity one wants to estimate.

- MAP estimation can therefore be seen as a regularization of maximum likelihood estimation.

- Assume that we want to estimate an unobserved population parameter θ on the basis of observations x.

- Let f be the sampling distribution of x, so that f(x| θ) is the probability of x when the underlying population parameter is θ . Then the function:

$\theta \mapsto f(x \mid \theta)$ → is known as the likelihood function

- The estimate:

$\hat{\theta}_{\mathrm{MLE}}(x) = \underset{\theta}{\arg\max}\, f(x \mid \theta)$ → is the maximum likelihood estimate θ

- Now assume that a prior distribution g over exists. This allows us to treat as a random variable as in Bayesian statistics. We can calculate the posterior distribution of using Bayes' theorem:

$$\theta \mapsto f(\theta \mid x) = \frac{f(x \mid \theta)\, g(\theta)}{\displaystyle\int_{\Theta} f(x \mid \vartheta)\, g(\vartheta)\, d\vartheta}$$

where g is density function of θ, Θ is the domain of g.

- The method of maximum a posteriori estimation then estimates as the mode of the posterior distribution of this random variable:

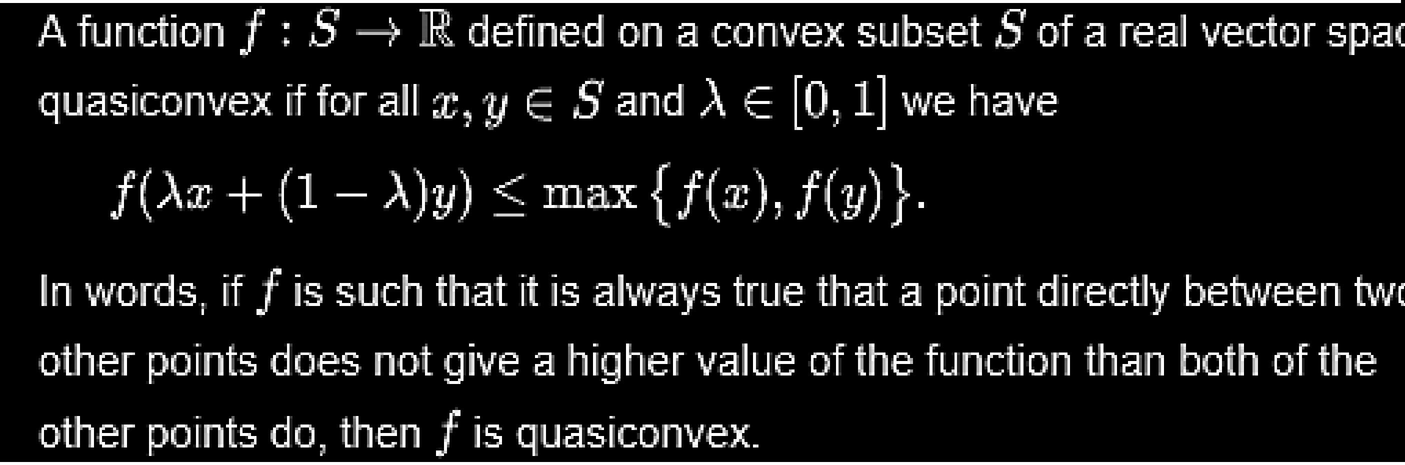
$$\boxed{\begin{aligned}\hat{\theta}_{\mathrm{MAP}}(x) &= \underset{\theta}{\arg\max}\ f(\theta \mid x) \\ &= \underset{\theta}{\arg\max}\ \frac{f(x \mid \theta)\, g(\theta)}{\displaystyle\int_{\Theta} f(x \mid \vartheta)\, g(\vartheta)\, d\vartheta} \\ &= \underset{\theta}{\arg\max}\ f(x \mid \theta)\, g(\theta).\end{aligned}}$$

- The denominator of the posterior distribution (so-called marginal likelihood) is always positive and does not depend on θ and therefore plays no role in the optimization.
- Observe that the MAP estimate of θ coincides with the ML estimate when the prior g is uniform (i.e., g is a constant function).

- When the loss function is of the form:

$$\boxed{L(\theta, a) = \begin{cases} 0, & \text{if } |a - \theta| < c, \\ 1, & \text{otherwise,} \end{cases}}$$

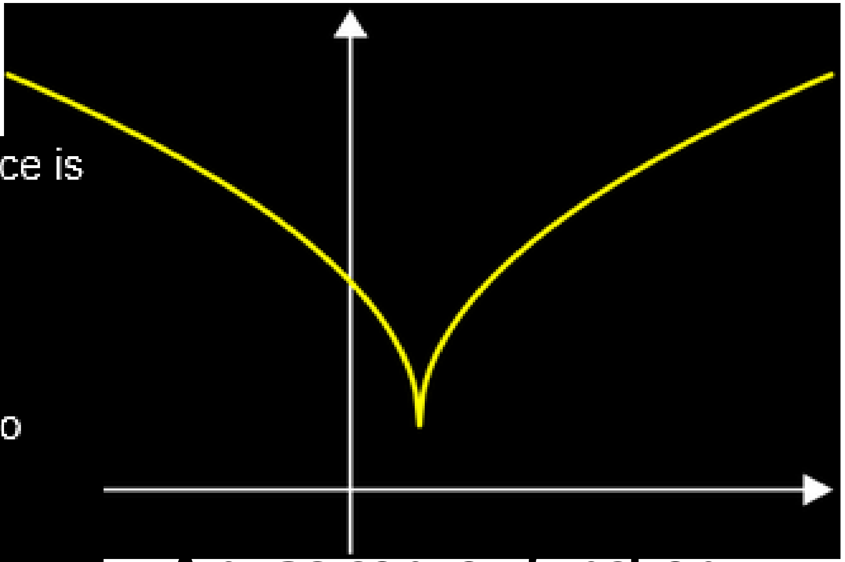
as c goes to 0, the Bayes estimator approaches the MAP estimator, provided that the **distribution of θ is quasi-concave**. But generally a MAP estimator θ is not a Bayes estimator unless is discrete.



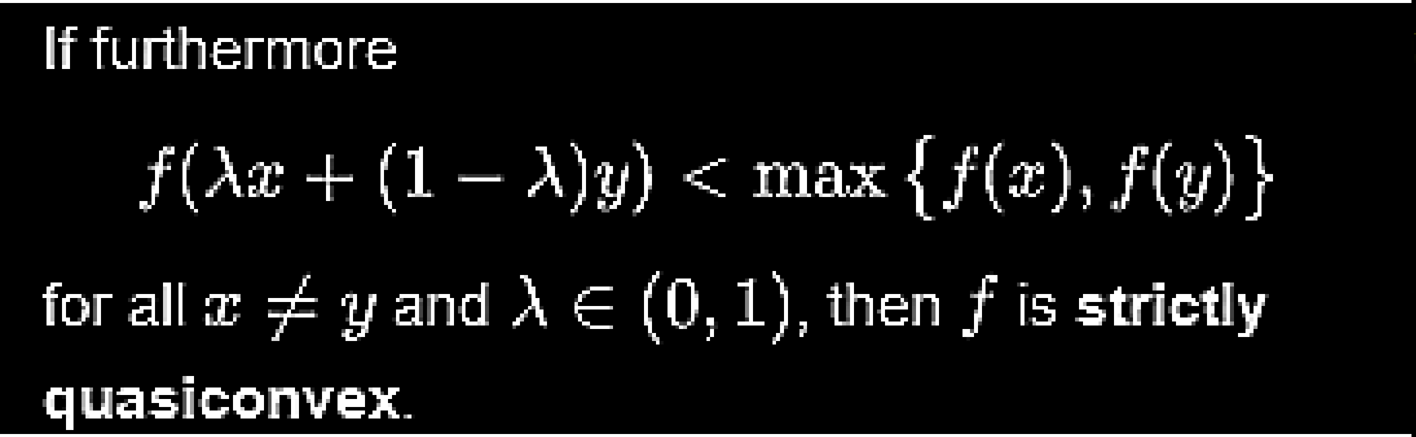
A function $f : S \to \mathbb{R}$ defined on a convex subset S of a real vector space is quasiconvex if for all $x, y \in S$ and $\lambda \in [0,1]$ we have

$$f(\lambda x + (1-\lambda)y) \leq \max\{f(x), f(y)\}.$$

In words, if f is such that it is always true that a point directly between two other points does not give a higher value of the function than both of the other points do, then f is quasiconvex.



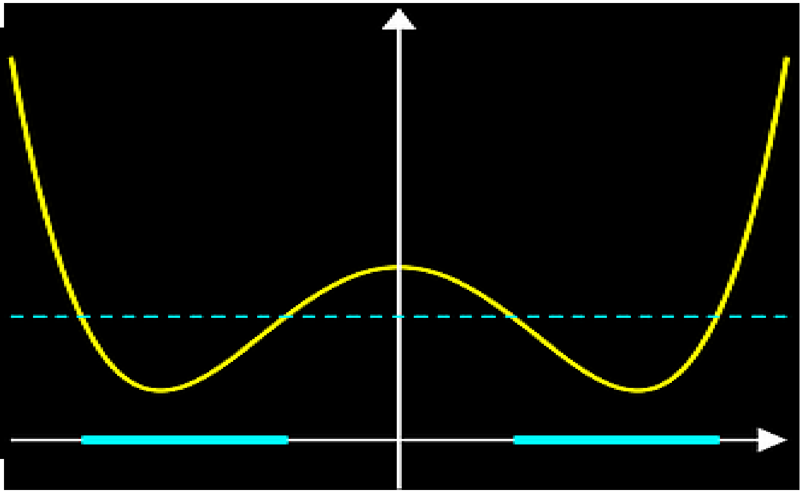
A quasiconvex function that is not convex



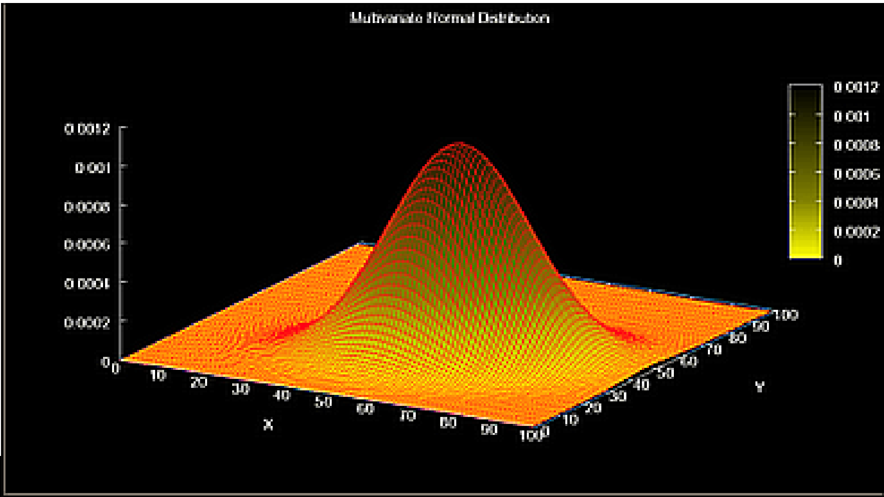
If furthermore

$$f(\lambda x + (1-\lambda)y) < \max\{f(x), f(y)\}$$

for all $x \neq y$ and $\lambda \in (0,1)$, then f is **strictly quasiconvex**.



A function that is not quasiconvex



The bivariate normal joint density is quasiconcave.

A **quasiconcave function** is a function whose negative is quasiconvex, and a **strictly quasiconcave function** is a function whose negative is strictly quasiconvex. Equivalently a function f is quasiconcave if

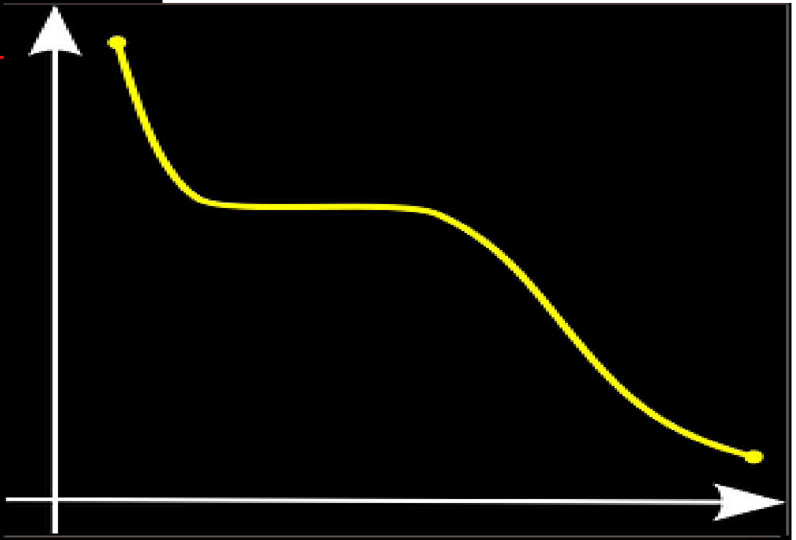
$$f(\lambda x + (1-\lambda)y) \geq \min\{f(x), f(y)\}.$$

and strictly quasiconcave if

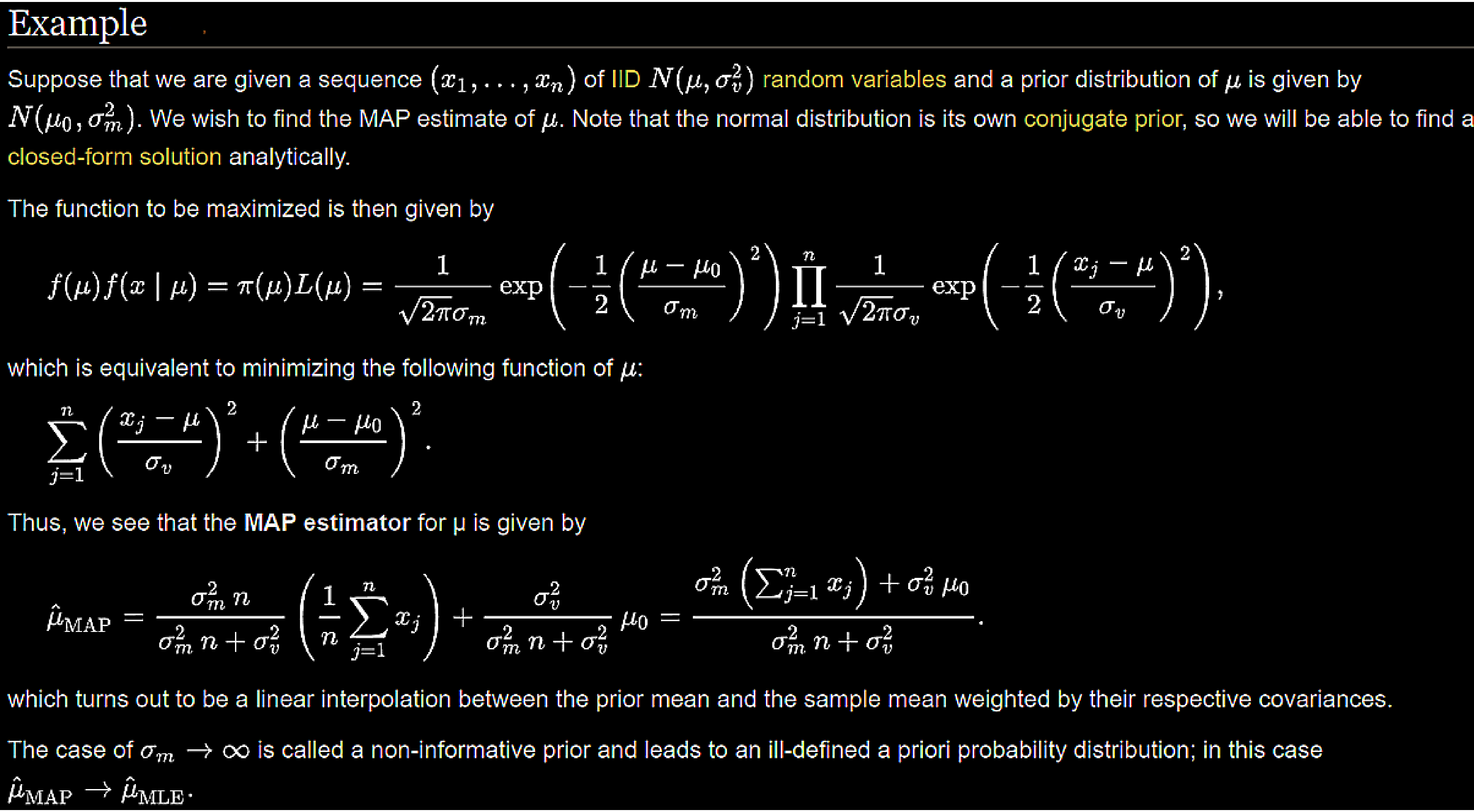
$$f(\lambda x + (1-\lambda)y) > \min\{f(x), f(y)\}$$

A (strictly) quasiconvex function has (strictly) convex lower contour sets, while a (strictly) quasiconcave function has (strictly) convex upper contour sets.

A function that is both quasiconvex and quasiconcave is **quasilinear**.



A quasilinear function is both quasiconvex and quasiconcave.



Example

Suppose that we are given a sequence $(x_1, \dots, x_n)$ of IID $N(\mu, \sigma_v^2)$ random variables and a prior distribution of μ is given by $N(\mu_0, \sigma_m^2)$. We wish to find the MAP estimate of μ. Note that the normal distribution is its own conjugate prior, so we will be able to find a closed-form solution analytically.

The function to be maximized is then given by

$$f(\mu)f(x \mid \mu) = \pi(\mu)L(\mu) = \frac{1}{\sqrt{2\pi}\sigma_m} \exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_m}\right)^2\right) \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left(-\frac{1}{2}\left(\frac{x_j-\mu}{\sigma_v}\right)^2\right),$$

which is equivalent to minimizing the following function of μ:

$$\sum_{j=1}^{n} \left(\frac{x_j-\mu}{\sigma_v}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_m}\right)^2.$$

Thus, we see that the **MAP estimator** for μ is given by

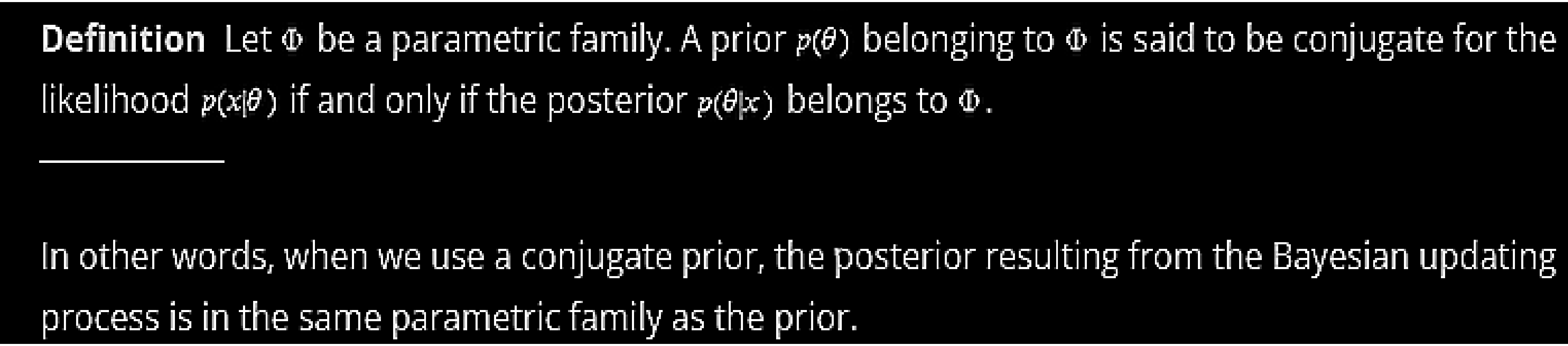
$$\hat{\mu}_{\mathrm{MAP}} = \frac{\sigma_m^2\, n}{\sigma_m^2\, n + \sigma_v^2} \left(\frac{1}{n}\sum_{j=1}^{n} x_j\right) + \frac{\sigma_v^2}{\sigma_m^2\, n + \sigma_v^2}\, \mu_0 = \frac{\sigma_m^2 \left(\sum_{j=1}^{n} x_j\right) + \sigma_v^2\, \mu_0}{\sigma_m^2\, n + \sigma_v^2}.$$

which turns out to be a linear interpolation between the prior mean and the sample mean weighted by their respective covariances.

The case of $\sigma_m \to \infty$ is called a non-informative prior and leads to an ill-defined a priori probability distribution; in this case $\hat{\mu}_{\mathrm{MAP}} \to \hat{\mu}_{\mathrm{MLE}}.$

In Bayesian inference, the prior distribution of a parameter and the likelihood of the observed data are combined to obtain the posterior distribution of the parameter.

If the prior and the posterior belong to the same parametric family, then the prior is said to be conjugate for the likelihood.



Definition Let Φ be a parametric family. A prior $p(\theta)$ belonging to Φ is said to be conjugate for the likelihood $p(x|\theta)$ if and only if the posterior $p(\theta|x)$ belongs to Φ.

In other words, when we use a conjugate prior, the posterior resulting from the Bayesian updating process is in the same parametric family as the prior.

In case of Bayesian inference suing Normal distribution, both the prior and the posterior distribution of the parameters are normal. Hence, the prior and the posterior belong to the same parametric family of normal distributions, and the prior is conjugate with the likelihood

References:

1. https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation#Example

2. https://www.probabilitycourse.com/chapter9/9_1_2_MAP_estimation.php