Bayesian Estimation/MAP

Example: MLE Math

dendrogram.

Suppose that n samples $x_1, \ldots, x_n$ are drawn independently according to $p(x \mid \theta)$, where

$$p(x \mid \theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Show that the maximum likelihood estimate for θ is given by $\hat{\theta} = \dfrac{1}{\frac{1}{n}\sum_{k=1}^{n} x^k}$

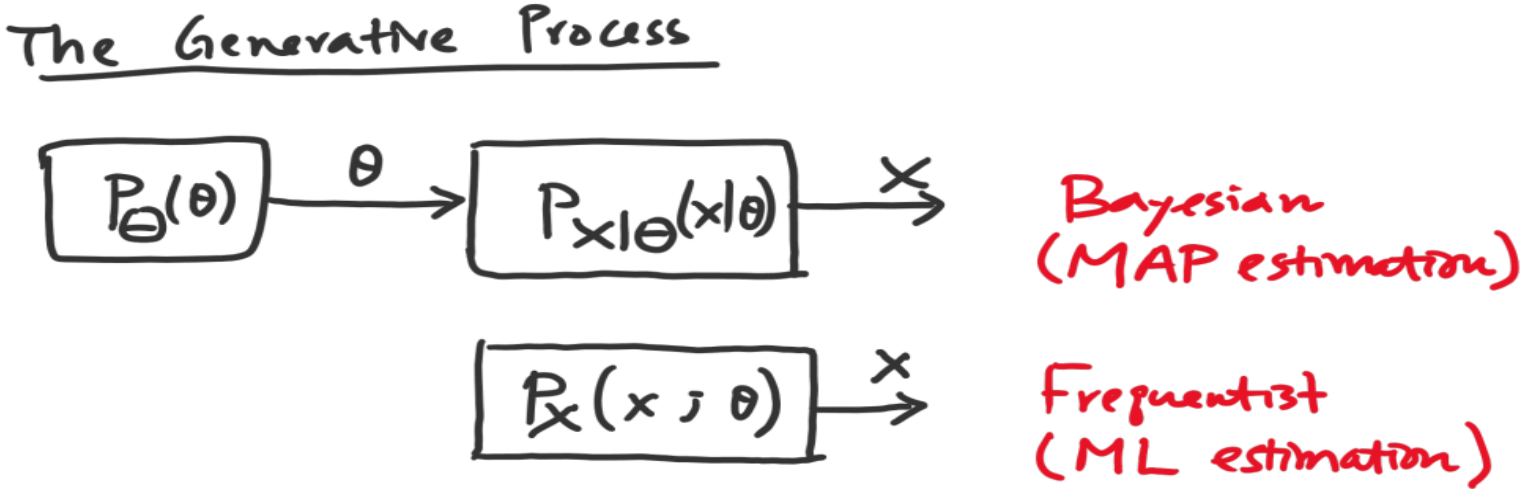
Limitations of MLE (*Example)*

- Consider a language model for sentences based on the bag-of-words assumption. In such a model, the probability of a sentence can be factored as the probability of the words appearing in the sentence.
- For simplicity, assume that our language corpus consists of a single sentence, "Probabilistic graphical models are fun. They are also powerful."
- We can estimate the probability of each of the individual words based on the counts. Our corpus contains 10 words with each word appearing once, and hence, each word in the corpus is assigned a probability of 0.1.
- Now, while testing the generalization of our model to the English language, we observe another sentence, "Probabilistic graphical models are hard."
- The probability of the sentence under our model is $0.1 \times 0.1 \times 0.1 \times 0.1 \times 0 = 00.1 \times 0.1 \times 0.1 \times 0.1 \times 0 = 0$.
- We did not observe one of the words ("hard") during training which made our language model infer the sentence as impossible, even though it is a perfectly plausible sentence.

- Out-of-vocabulary words are a common phenomena even for language models trained on large corpus.
- One of the simplest ways to handle these words is to assign a prior probability of observing an out-of-vocabulary word such that the model will assign a low, but non-zero probability to test sentences containing such words.
- This mechanism of incorporating prior knowledge is a practical application of Bayesian learning

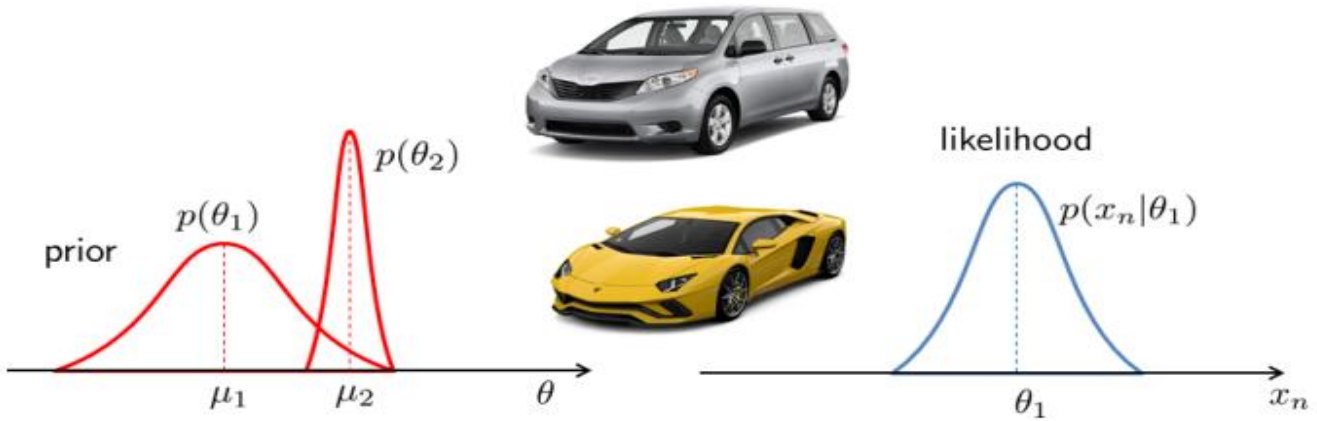
MLE and MAP

There are two typical ways of estimating parameters.



- Maximum-likelihood estimation (MLE): $\boldsymbol{\theta}$ is deterministic.
- Maximum-a-posteriori estimation (MAP): $\boldsymbol{\theta}$ is random and has a prior distribution.

Moving from MLE to MAP



- Likelihood:

$$p(x_n|\theta_1) = \mathcal{N}(x_n|\theta_1, \sigma_1^2), \quad \text{and} \quad p(x_n|\theta_2) = \mathcal{N}(x_n|\theta_2, \sigma_2^2).$$

- Maximum-likelihood: You know nothing about θ_1 and θ_2. So you need to take measurements to estimate θ_1 and θ_2.
- Maximum-a-Posteriori: You know something about θ_1 and θ_2.
- Prior

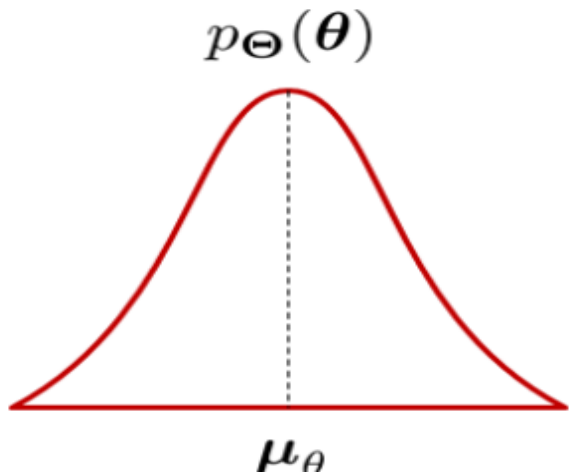
$$p(\theta_1) = \mathcal{N}(\mu_1|\gamma_1^2), \quad \text{and} \quad p(\theta_2) = \mathcal{N}(\mu_2|\gamma_2^2).$$

Moving from MLE to MAP

- In MLE, the parameter θ is **deterministic**.
- What if we assume θ has a distribution?
- This makes θ **probabilistic**.
- So make $\boldsymbol{\Theta}$ as a random variable, and θ a state of $\boldsymbol{\Theta}$.
- Distribution of $\boldsymbol{\Theta}$:

$$p_{\boldsymbol{\Theta}}(\theta)$$

- $p_{\boldsymbol{\Theta}}(\theta)$ is the distribution of the parameter $\boldsymbol{\Theta}$.
- $\boldsymbol{\Theta}$ has its own mean and own variance.



Maximum-a-Posteriori

By Bayes Theorem again:

$$p_{\Theta|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}_n) = \frac{p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}_n|\boldsymbol{\theta})p_{\Theta}(\boldsymbol{\theta})}{p_{\boldsymbol{X}}(\boldsymbol{x}_n)}.$$

- To maximize the posterior distribution

$$\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} \; p_{\Theta|\boldsymbol{X}}(\boldsymbol{\theta}|\mathcal{D}) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \; \prod_{n=1}^{N} p_{\Theta|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}_n) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \; \prod_{n=1}^{N} \frac{p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}_n|\boldsymbol{\theta})p_{\Theta}(\boldsymbol{\theta})}{\cancel{p_{\boldsymbol{X}}(\boldsymbol{x}_n)}} \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} \; -\sum_{n=1}^{N} \Big\{ \log p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}_n|\boldsymbol{\theta}) + \log p_{\Theta}(\boldsymbol{\theta}) \Big\}
\end{aligned}$$

Illustration: 1D Example

Suppose that:

$$p_{X|\Theta}(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}$$

$$p_{\Theta}(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\}.$$

When $N = 1$. The MAP problem is simply

$$\begin{aligned}
\widehat{\theta} &= \underset{\theta}{\text{argmax}} \;\; p_{X|\Theta}(x|\theta)p_{\Theta}(\theta) \\
&= \underset{\theta}{\text{argmax}} \;\; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\} \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\} \\
&= \underset{\theta}{\text{argmax}} \;\; -\frac{(x-\theta)^2}{2\sigma^2} - \frac{(\theta-\theta_0)^2}{2\sigma_0^2}
\end{aligned}$$

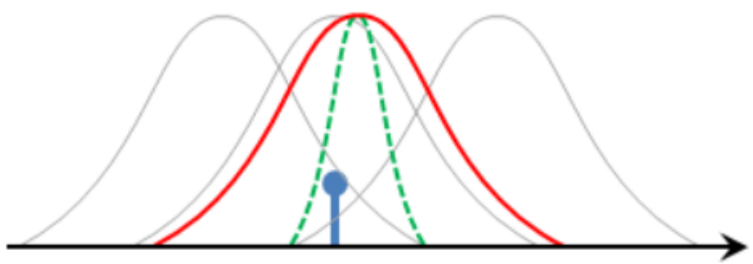
Illustration: 1D Example

Taking derivatives:

$$\begin{aligned}
& \frac{d}{d\theta}\left\{-\frac{(x-\theta)^2}{2\sigma^2}-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\} && = 0 \\
\Rightarrow \quad & \frac{(x-\theta)}{\sigma^2}-\frac{(\theta-\theta_0)}{\sigma_0^2} && = 0 \\
\Rightarrow \quad & \sigma_0^2(x-\theta) && = \sigma^2(\theta-\theta_0) \\
\Rightarrow \quad & \sigma_0^2 x+\sigma^2\theta_0 && = (\sigma_0^2+\sigma^2)\theta
\end{aligned}$$

Therefore, the solution is

$$\theta = \frac{\sigma_0^2 x+\sigma^2\theta_0}{\sigma_0^2+\sigma^2}.$$



Interpreting the Result

Let us interpret the result

$$\theta = \frac{\sigma_0^2 x + \sigma^2 \theta_0}{\sigma_0^2 + \sigma^2}.$$

Does it make sense?

- If $\sigma_0 = 0$, then $\theta = \frac{\cancel{\sigma_0^2} x + \sigma^2 \theta_0}{\cancel{\sigma_0^2} + \sigma^2} = \theta_0$.
- This means: No uncertainty. Absolutely sure that $\theta = \theta_0$.
- $p_\Theta(\theta) = \delta(\theta - \theta_0)$

The other extreme

- If $\sigma_0 = \infty$, then $\theta = \frac{\sigma_0^2 x + \cancel{\sigma^2 \theta_0}}{\sigma_0^2 + \cancel{\sigma^2}} = x$.
- This means: I don't trust my prior at all. Use data.
- $p_\Theta = \frac{1}{|\Omega|}$, for all $\theta \in \Omega$.

Therefore, MAP solution gives you a trade-off between data and prior.

Class-Conditional Densities

- **Posterior probabilities, $P(\omega_i|\mathbf{x})$, are central to Bayesian classification.**
- **Bayes formula allows us to compute $P(\omega_i|\mathbf{x})$ from the priors, $P(\omega_i)$, and the likelihood, $p(\mathbf{x}|\omega_i)$.**
- **But what If the priors and class-conditional densities are unknown?**
- **The answer is that we can compute the posterior, $P(\omega_i|\mathbf{x})$, using all of the information at our disposal (e.g., training data).**
- **For a training set, *D*, Bayes formula becomes:**

$$P(\omega_i \mid \mathbf{x}, D) = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} = \frac{p(\mathbf{x}|\omega_i, D)P(\omega_i|D)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j, D)P(\omega_j|D)}$$

- **We assume priors are known: $P(\omega_i|D) = P(\omega_i)$.**
- **Also, assume functional independence:**

  D_i have no influence on $p(\mathbf{x}|\omega_j, D) \quad if \ i \neq j$

  This gives: $P(\omega_i \mid \mathbf{x}, D) = \frac{p(\mathbf{x}|\omega_i, D_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j, D_j)P(\omega_j)}$

The Parameter Distribution

- **Assume the parametric form of the evidence, p(x), is known: p(x|θ).**
- **Any information we have about θ prior to collecting samples is contained in a known prior density p(θ).**
- **Observation of samples converts this to a posterior, p(θ|*D*), which we hope is peaked around the true value of θ.**
- **Our goal is to estimate a parameter vector:**

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

- **We can write the joint distribution as a product:**

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

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

because the samples are drawn independently.

- **This equation links the class-conditional density** $p(\mathbf{x}|D)$

to the posterior, $p(\theta|D)$ **. But numerical solutions are typically required!**

Univariate Gaussian Case

- **Case: only mean unknown**

$$p(x|\mu) \approx N(\mu, \sigma^2)$$

- **Known prior density:**

$$p(\mu) \approx N(\mu_0, \sigma_0^2)$$

- **Using Bayes formula:**

$$p(\mu|D)\,p(D) = p(D|\mu)\,p(\mu)$$

- **Rationale: Once a value of μ is known, the density for x is completely known. α is a normalization factor that depends on the data, *D*.**

$$p(\mu|D) = \frac{p(D|\mu)\,p(\mu)}{p(D)}$$

$$= \frac{p(D|\mu)\,p(\mu)}{\int p(D|\mu)\,p(\mu)d\mu}$$

$$= \alpha[\,p(D|\mu)\,p(\mu)]$$

$$= \alpha \prod_{k=1}^{n} p(x_k|\mu)\,p(\mu)$$

Univariate Gaussian Case

- **Applying our Gaussian assumptions:**

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

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

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

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

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

Univariate Gaussian Case (Cont.)

- **Now we need to work this into a simpler form:**

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

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

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

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

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

$$\text{where} \quad \hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$$

Univariate Gaussian Case (Cont.)

- p(μ|*D*) **is an exponential of a quadratic function, which makes it a normal distribution. Because this is true for any** n**, it is referred to as a *reproducing density*.**

- p(μ) **is referred to as a *conjugate prior*.**

- **Write** p(μ|*D*) **~** $N(\mu_n,\sigma_n^2)$**:** $p(\mu|D) = \frac{1}{\sqrt{2\pi}\sigma_n}\exp[-\frac{1}{2}(\frac{\mu-\mu_n}{\sigma_n})^2]$

- **Expand the quadratic term:**

$$p(\mu|D) = \frac{1}{\sqrt{2\pi}\sigma_n}\exp[-\frac{1}{2}(\frac{\mu-\mu_n}{\sigma_n})^2] = \frac{1}{\sqrt{2\pi}\sigma_n}\exp[-\frac{1}{2}(\frac{\mu^2-2\mu\mu_n+\mu_n^2}{\sigma_n^2})]$$

- **Equate coefficients of our two functions:**

$$\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left(-\frac{1}{2}\left[\frac{\mu^2-2\mu\mu_n+\mu_n^2}{\sigma_n^2}\right]\right) =$$

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

Equating coefficients

- In mathematics, the method of **equating the coefficients** is a way of solving a functional equation of two expressions such as polynomials for a number of unknown parameters. It relies on the fact that two expressions are identical precisely when corresponding coefficients are equal for each different type of term. The method is used to bring formulas into a desired form.

Conjugate prior: P(θ) is the conjugate prior for likelihood function P(data | θ) if the forms of P(θ) and P(θ | data) are the same.

Univariate Gaussian Case (Cont.)

- **Rearrange terms so that the dependencies on μ are clear:**

$$\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left(-\frac{1}{2}\left(\frac{\mu_n^2}{\sigma_n^2}\right)\right)\exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_n^2}\mu^2-2\frac{\mu_n}{\sigma_n^2}\mu\right)\right)=$$

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

- **Associate terms related to σ^2 and μ:**

$$\sigma_n^2 \Leftrightarrow \sigma^2 : \frac{1}{\sigma_n^2}=\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}$$

$$\mu_n \Leftrightarrow \mu : \quad \frac{\mu_n}{\sigma_n^2}=\frac{n}{\sigma^2}\hat{\mu}_n+\frac{\mu_0}{\sigma_0^2}$$

- **There is actually a third equation involving terms not related to μ:**

$$\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left(-\frac{1}{2}\left[\frac{\mu_n^2}{\sigma_n^2}\right]\right)=\alpha''-or-\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left(-\frac{1}{2}\left[\frac{\mu_n^2}{\sigma_n^2}\right]\right)=\alpha\frac{1}{\sqrt{2\pi}\sigma_0}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n$$

but we can ignore this since it is not a function of μ and is a complicated equation to solve.

Univariate Gaussian Case (Cont.)

- **Two equations and two unknowns. Solve for μ_n and σ_n^2. First, solve for μ_n^2 :**

$$\sigma_n^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

- **Next, solve for μ_n:**

$$\mu_n = \hat{\mu}_n\left(\frac{n\sigma_n^2}{\sigma^2}\right) + \mu_0\left(\frac{\sigma_n^2}{\sigma_0^2}\right)$$

$$= \hat{\mu}_n\left(\frac{n}{\sigma^2}\right)\left(\frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}\right) + \mu_0\left(\frac{1}{\sigma_0^2}\right)\left(\frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}\right)$$

$$= \hat{\mu}_n\left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right) + \mu_0\left(\frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\right)$$

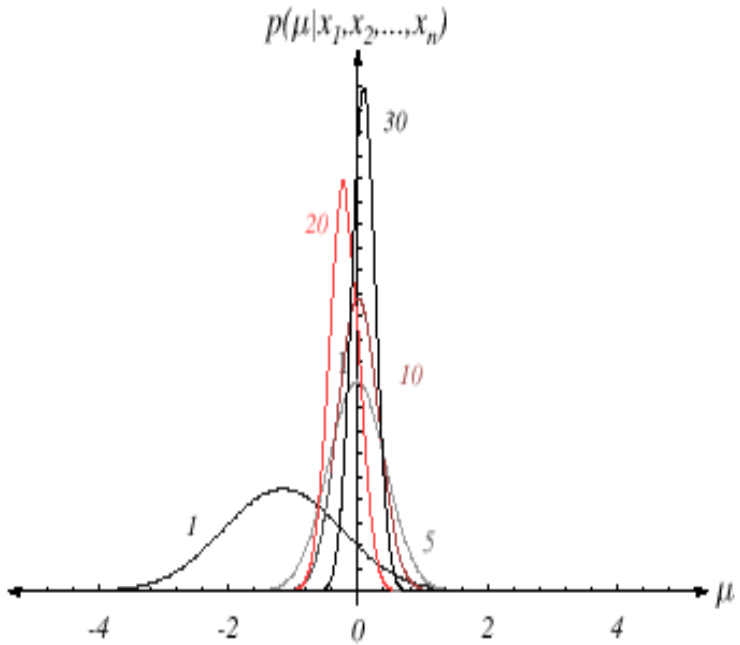
- **Summarizing:**

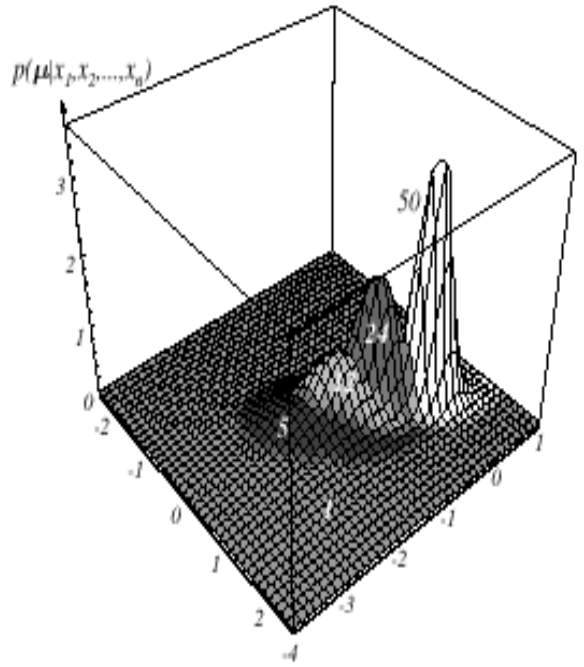
$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \left(\frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\right)\mu_0$$

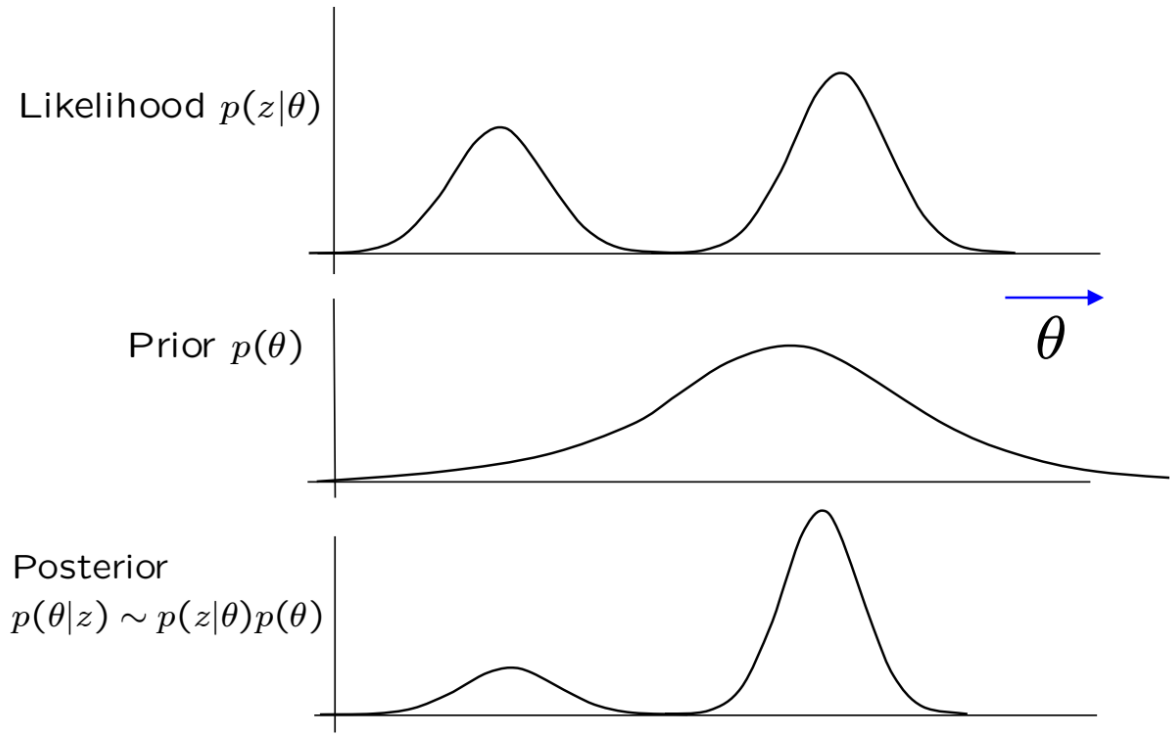
$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

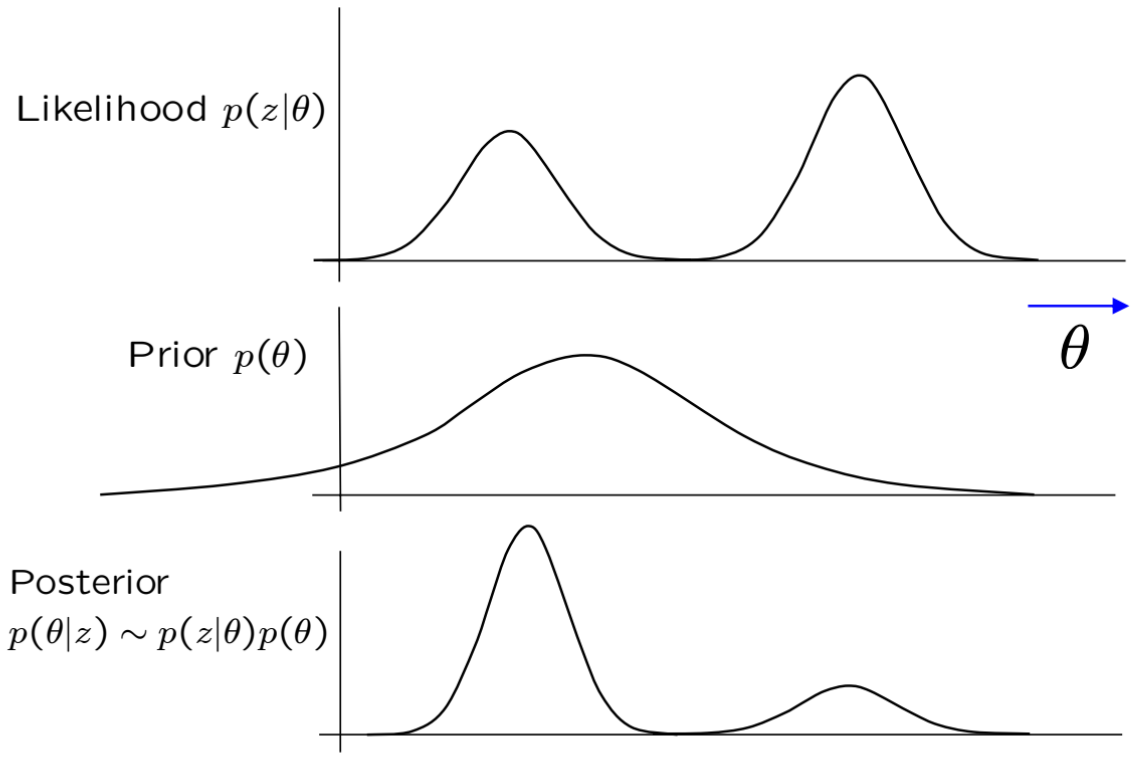
Bayesian Learning

- μ_n **represents our best guess after n samples.**
- σ_n^2 **represents our uncertainty about this guess.**
- σ_n^2 **approaches** σ^2/n **for large** n **– each additional observation decreases our uncertainty.**
- **The posterior,** $p(\mu|D)$**, becomes more sharply peaked as n grows large. This is known as *Bayesian learning*.**






Likelihood $p(z|\theta)$
Prior $p(\theta)$
θ
Posterior
$p(\theta|z) \sim p(z|\theta)p(\theta)$


Likelihood $p(z|\theta)$
Prior $p(\theta)$
θ
Posterior
$p(\theta|z) \sim p(z|\theta)p(\theta)$

MAP Estimator

$$\hat{\theta} = \arg\max_{\theta} p(\theta|\mathbf{z})$$