CSE 575
Statistical Machine Learning

Lecture 5

YooJung Choi

Fall 2022

Thumbtack problem

- If I toss a thumbtack, what's the probability it will land with the nail up?
- Toss it a few times:

  Down, Down, **Up**, Down, **Up**
- Probability is…? And Why?
- 2/5. MLE!

[Slides adapted from David Sontag]

Thumbtack problem

- $p(\text{Head}) = \theta,\ p(\text{Tail}) = 1 - \theta$

  "Bernoulli distribution"

- Flips are i.i.d. (independently and identically distributed according to Bernoulli)
- Likelihood of a sequence $\mathcal{D}$ of α_H Heads and α_T Tails:

$$p(\mathcal{D}|\theta) = \theta^{\alpha_H} \cdot (1-\theta)^{\alpha_T}$$

- MLE: $\text{argmax}_\theta\, p(\mathcal{D}|\theta) = \text{argmax}_\theta \ln p(\mathcal{D}|\theta)$
- Set the derivative to zero and solve: $\theta_{ML} = \frac{\alpha_H}{\alpha_H + \alpha_T}$

Bernoulli distribution

- Binary random variable $x \in \{0,1\}$, $\quad p(x=1|\mu)=\mu$
- $\mathrm{Bern}(x|\mu)=\mu^x(1-\mu)^{1-x}$
- Expectation and variance:

$$E[x]=\mu, \qquad Var[x]=\mu(1-\mu)$$

- Given a dataset $\mathcal{D}=\{x_1,\ldots,x_N\}$, the likelihood is:

$$p(\mathcal{D}|\theta)=\prod_{n=1}^{N} p(x_n|\theta)=\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

- MLE: $\mu_{ML}=\frac{1}{N}\sum_{n=1}^{N} x_n$ *Sufficient statistics*

Binomial distribution

- Let m be the number of observations where $x = 1$
- The distribution of m is the *binomial* distribution:

$$\text{Bin}(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

where $\binom{N}{m} = \frac{N!}{(N-m)!m!}$

- MLE: $\mu_{ML} = m/N$
- Expectation and variance:

$$E[x] = N\mu, \qquad Var[x] = N\mu(1-\mu)$$

Thumbtack problem

- If I toss a thumbtack, what's the probability it will land with the nail up?
- What if the trial results were:

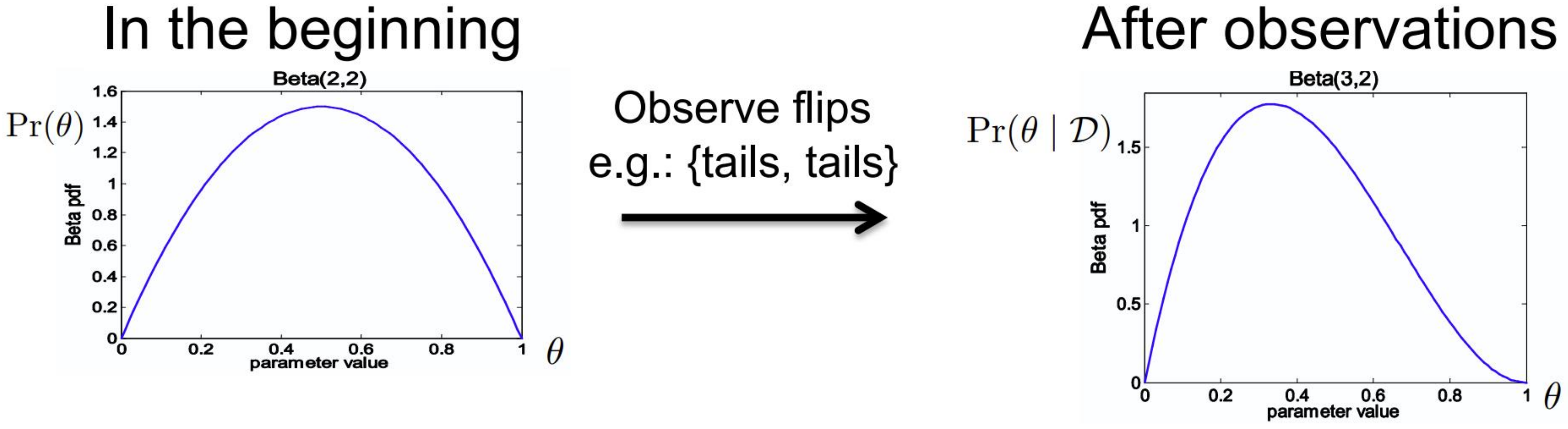
  Up, **Up**, **Up**

- Maximum likelihood estimate:

$$\mu_{ML} = \frac{3}{3} = 1$$

Thumbtack problem

- What if I tell you that the thumbtack is “close” to 50-50?
- Bayesian approach: rather than estimating a single θ, obtain a distribution over possible values of θ



Bayesian learning

- Recall from Bayes' Theorem: $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta) \cdot p(\theta)$
- For *uniform priors*, reduces to MLE!

$$p(\theta) \propto 1 \quad \Rightarrow \quad p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)$$

- What should the prior be?
  - Represent expert knowledge
  - Simple posterior form
- A prior is called a *conjugate prior* for the likelihood function $p(\mathcal{D}|\theta)$ if the posterior $p(\theta|\mathcal{D})$ has the same form as the prior

Conjugate prior for Binomial

- Likelihood:

$$p(\mathcal{D}|\theta) = \binom{N}{m} \theta^m (1-\theta)^{N-m}$$

- If prior is of the following form:

$$p(\theta) = C \cdot \theta^\alpha (1-\theta)^\beta$$

- Then the posterior will also look like:

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta) \cdot p(\theta) = C' \cdot \theta^{\alpha'} (1-\theta)^{\beta'}$$

Gamma function & Beta distribution

- The Gamma function is defined by the integration:

$$\Gamma(x) \equiv \int_0^\infty u^{x-1} e^{-u} \, \mathrm{d}u.$$

- Some properties:
  - $\Gamma(x+1) = x\Gamma(x)$ for $x > 0$
  - $\Gamma(1) = 1$
  - $\Gamma(x+1) = x!$ when x is a positive integer

Gamma function & Beta distribution

- Beta distribution is given by: $\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$

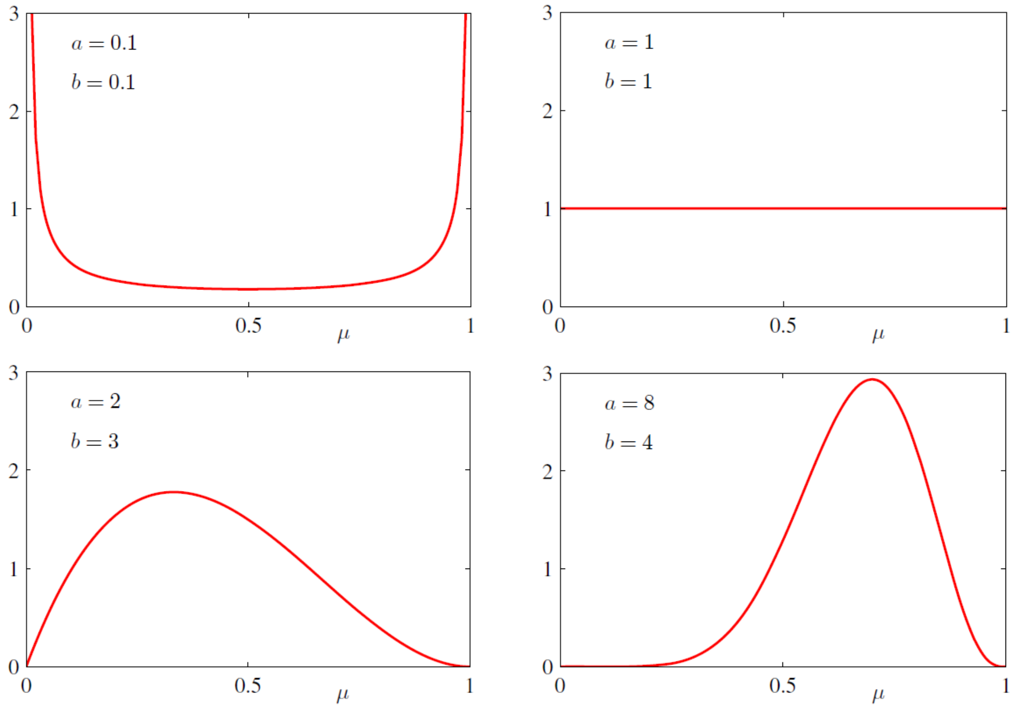


Figure 2.2 Plots of the beta distribution $\mathrm{Beta}(\mu|a,b)$ given by (2.13) as a function of μ for various values of the hyperparameters a and b.

Gamma function & Beta distribution

- Beta distribution is given by: $\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$
- Expectation and variance:

$$\begin{aligned} \mathbb{E}[\mu] &= \frac{a}{a+b} \\ \mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)}. \end{aligned}$$

- The posterior distribution of a Beta distribution is also Beta:

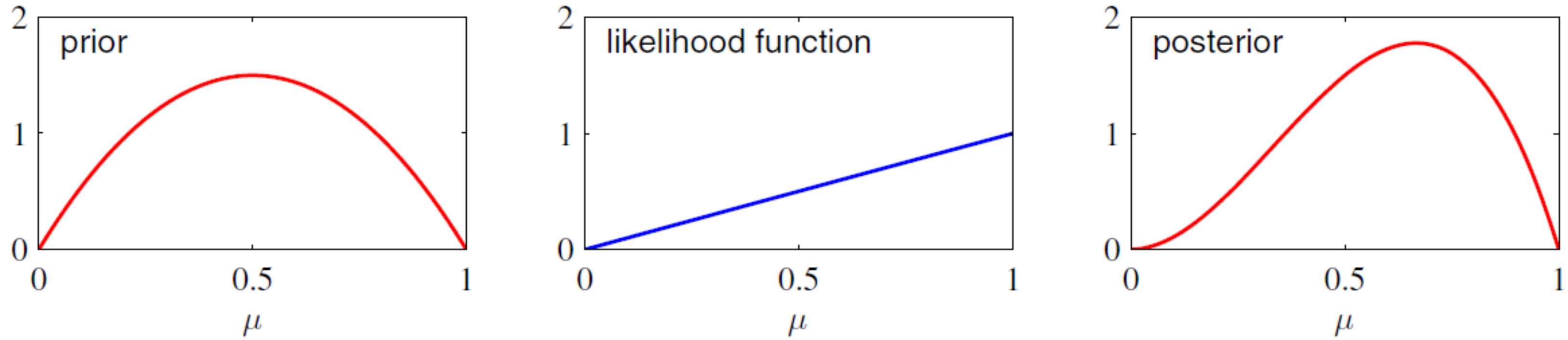
$$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}. \qquad (l = N - m)$$

Gamma function & Beta distribution

- Beta distribution is the *conjugate prior* for *Binomial*
- Suppose the prior is given by a beta distribution with a=2, b=2, and the likelihood function given by (N=m=1):

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

- The posterior distribution is a beta distribution with parameters a=3, b=2.



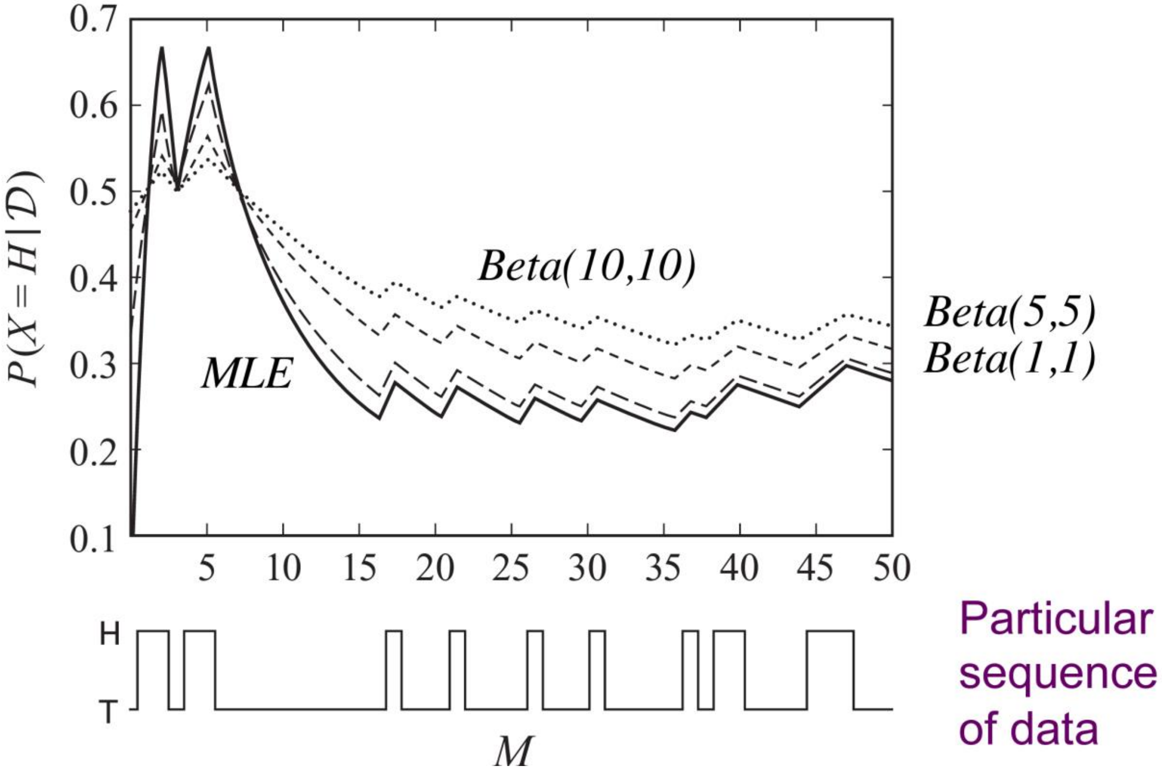
Gamma function & Beta distribution

- Prior: $\mathrm{Beta}(\mu|a,b)$
- Likelihood of m “heads” and $l = N - m$ tails: $\mathrm{Bin}(m|N,\mu)$
- Posterior: $\mathrm{Beta}(\mu|m+a,l+b)$

- MAP estimate: $\mu_{MAP} = \max_{\mu} \mathrm{Beta}(\mu|m+a,l+b) = \frac{m+a-1}{m+a+l+b-2}$

equivalent / imaginary sample size

Effect of different priors



Smoother estimates with higher equivalent sample size

[Slide adapted from Sargur Srihari]

Some observations

- The posterior after N=m+l observations is:

$$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}.$$

- When $N = 0$, it reduces to the prior
- As $N \to \infty$, it converges to the MLE

- Variance: $Var[\mu] = \frac{(m+a)(l+b)}{(m+a+l+b)^2(m+a+l+b+1)}$

  converges to 0 as $N \to \infty$

Sequential learning

- We can perform sequential learning

$$\mathrm{Beta}(\mu|a,b)$$

m heads and l tails

$$\mathrm{Beta}(\mu|m+a,l+b)$$

m' heads and l' tails

$$\mathrm{Beta}(\mu|m'+m+a,l'+l+b)$$

- If our goal is to predict the outcome of the next trial, then we must evaluate the predictive distribution of x, given the observed data set D. Then we have

$$p(x=1|\mathcal{D}) = \frac{m+a}{m+a+l+b}$$

Multinomial variables

- Discrete variable that can take one of K possible values
- We can use a K-dimensional vector x s.t. one of x_k equals 1.

$$\sum_{k=1}^{K} x_k = 1.$$

- E.g. dice roll outcome 2: $x = (0,1,0,0,0,0)^T$ for $K = 6$
- Probability of $x_k = 1$ denoted by μ_k. Then:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

- Note that μ_k are non-negative and sum to 1.

Multinomial variables

- Expectation: $$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_M)^{\mathrm{T}} = \boldsymbol{\mu}.$$

- Consider a set D of N independent observations $x_1, \ldots, x_N$.
  Likelihood: $$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k}$$

  where $$m_k = \sum_n x_{nk}$$

- MLE (need to solve constrained optimization): $\mu_K^{ML} = \frac{m_k}{N}$

Multinomial distribution

- Let $m_1, \dots, m_K$ be the number of observations for each $x_k = 1$ where $\sum_{k=1}^{K} m_k = N$
- Their joint distribution is called the *multinomial* distribution:

$$\mathrm{Mult}(m_1, \dots, m_K | N, \boldsymbol{\mu}) = \binom{N}{m_1 \dots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

where $\binom{N}{m_1 \dots m_K} = \frac{N!}{m_1! \cdots m_K!}$

- The conjugate prior would take the form: $p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$

Dirichlet distribution

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k.$$

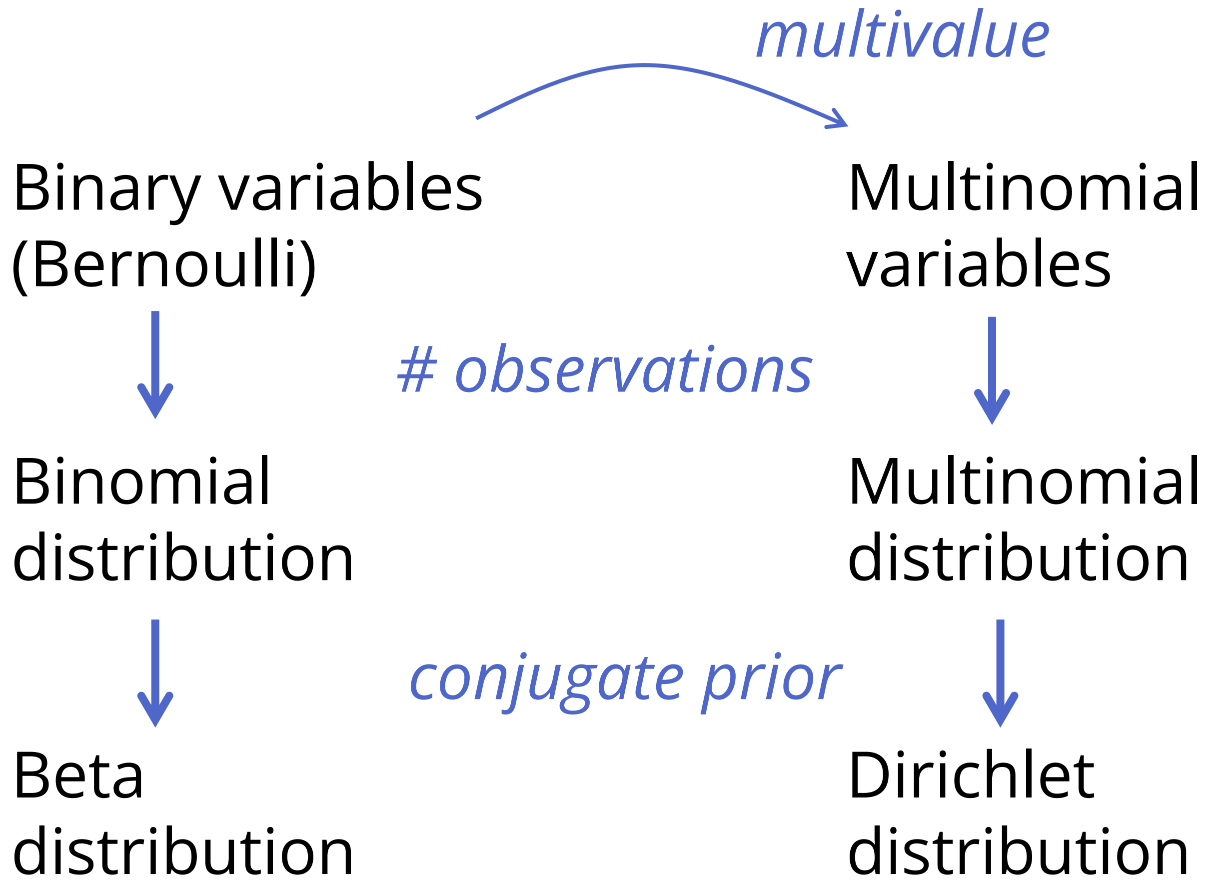
- Prior: $\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \dfrac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$

- Likelihood: $\mathrm{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) = \begin{pmatrix} N \\ m_1 m_2 \ldots m_K \end{pmatrix} \prod_{k=1}^{K} \mu_k^{m_k}$

- Posterior: $$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) &= \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m}) \\ &= \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1)\cdots\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \end{aligned}$$

- Similar observations as Beta distribution

Discrete probability distributions



MLE for the Gaussian

- MLE parameters: $$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$

- *Sufficient statistics* $$\sum_{n=1}^{N}\mathbf{x}_n, \qquad \sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}$$

Bayesian inference for the Gaussian

- Assume σ^2 is known, infer μ, after N observations
- Likelihood: $$p(\mathbf{X}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$
- Prior: $p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right)$
- Posterior: $p(\mu|\mathbf{X}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$

where

$$\begin{aligned} \mu_N &= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}} \\ \frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \end{aligned}$$

Sequential estimation

$$\prod_{n=1}^{N} p(x_n|\mu)$$

- Express the posterior as $p(\boldsymbol{\mu}|D) \propto \left[p(\boldsymbol{\mu}) \boxed{\prod_{n=1}^{N-1} p(\mathbf{x}_n|\boldsymbol{\mu}) \right] p(\mathbf{x}_N|\boldsymbol{\mu})}$

- i.e. [posterior after N-1 observations] x [likelihood of Nth data point]
- We can perform sequential estimation:

$$\begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \\ &= \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1} \mathbf{x}_n \\ &= \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} \\ &= \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}) \end{aligned}$$

move a small amount of the estimation after N-1 observations towards the last observation.