M3. Density Estimation

Manikandan Narayanan

Week 3 (Aug 11-, 2025)

PRML Jul-Nov 2025 (Grads section)

Acknowledgment of Sources

- Slides based on content from related
  - Courses:
    - IITM – Profs. Arun/Harish[HR]/Chandra[CC]/Prashanth's PRML offerings (slides, quizzes, notes, etc.), Prof. Ravi's "Intro to ML" slides – cited (e.g., [HR]/[HG]) in the bottom right of a slide.
    - India – NPTEL PR course by IISc Prof. PS. Sastry (slides, etc.) – cited as [PSS] in the bottom right of a slide.
  - Books:
    - PRML by **Bishop**. (content, figures, slides, etc.) – cited as **[CMB]**
    - Pattern Classification by Duda, Hart and Stork. (content, figures, etc.) – [DHS]
    - Mathematics for ML by Deisenroth, Faisal and Ong. (content, figures, etc.) – [DFO]
    - Foundations of ML by Mohri, Rostamizadeh, and Talwalkar (content, figures, slides by Mohri, etc.). – [MRT]
    - Information Theory, Inference and Learning Algorithms by **MacKay** – [**DJM**]

Outline for Module M3

- M3. Density Estimation
  - M3.0 Introduction/Warmup
  - M3.1 Parametric methods
  - M3.2 Nonparametric methods (introduced, but not covered)

Outline for Module M3 (detailed)

- M3. Density Estimation
  - M3.0 Introduction/Warmup
    - M3.0.0 What it means to “learn” from data?
    - M3.0.1 Intuitive warmup to ML (Estimation)
  - M3.1 Parametric methods

    (aka parameter learning of probabilistic models)
    - M3.1.1 Maximum Likelihood Estimation (MLE)

      (for continuous/discrete densities, incl. mixture densities (brief mention))
    - M3.1.2 Bayesian Inference(/estimation)
  - M3.2 Nonparametric methods (not covered)
    - M3.2.0 General idea
    - M3.2.1 K-nearest neighbors

Outline for Module M3

- M3. Density Estimation
  - **M3.0 Introduction/Warmup**
    - **M3.0.0 What it means to “learn” from data?**
    - **M3.0.1 Intuitive warmup to ML (Estimation)**
  - M3.1 Parametric methods
  - M3.2 Nonparametric methods (not covered)

Introduction to Density estimation

- So far: Decision theory (incl. Bayes classifiers)
  - Two steps in a generative or discriminative model setting: Inference vs. Decision steps
  - But how to do inference, i.e., how to “learn” a Bayes classifier from data???
    - estimate the joint (class prior and class conditional) **p(x,t)** or posterior density **p(t|x)**.
    - So density estimation needed in both generative/discriminative model settings.

[CMB]

Inference & Decision (steps in detail for a generative classification model):

Setup (training data):

Learning the Bayes classifier from data:

- $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ with $x_i \in \mathcal{X}$, $y_i \in \pm 1$

Inference (density est.):

- Assume $P_{X|Y}(x|1)$ and $P_{X|Y}(x|-1)$ are from P.
- Estimate $P_{X|Y}(x|1)$ from positively labelled samples and $P_{X|Y}(x|-1)$ from negatively labelled samples using Maximum Likelihood.
- Estimate $P(Y=1)$ & $P(Y=-1)$
- Use Bayes Rule to make an estimate of the posterior probability $P(Y=1|X=x) = \eta(x)$

Decision:

- Compute $h^*(x) = \text{sign}(2\eta(x)-1)$

[HR]

Inference & Decision (steps in detail for a discriminative regression model):

A: **P(t|x)** captures the input-output map. Steps involved are:

(1) Model/estimate **P(t|x)**

(how? *Density Estimation*; MLE/Bayesian Inference)

(2) Predict **t** for a new **x** from estimated P(t|x)

(how? *Decision Theory*; e.g., $y(\mathbf{x}_{\boldsymbol{new}}) = E[\boldsymbol{t}|\mathbf{x} = \mathbf{x}_{\boldsymbol{new}}]$)

Going forward…

…let’s focus on a single r.v. X and estimate its density!

(it can be used to estimate p(x), p(t), p(x|t), p(t|x), etc. as we will see through the course)

Introduction to Density estimation

- So far: Decision theory (incl. Bayes classifiers)
  - Two steps in a generative or discriminative model setting: Inference vs. Decision steps
  - But how to do inference, i.e., how to "learn" a Bayes classifier from data???
    - estimate the joint (class prior and class conditional) **p(x,t)** or posterior density **p(t|x)**.
    - So density estimation needed in both generative/discriminative model settings.

- Density estimation (informally aka learning the data distbn.):
  - Addresses a fundamental question of what it means to learn from data.
    - be it supervised (**p(x,t)** or **p(t|x)**) or unsupervised (**p(x)**) learning!
  - Relies heavily on assumptions made in model selection step – otherwise, an ill-posed problem!!

[CMB]

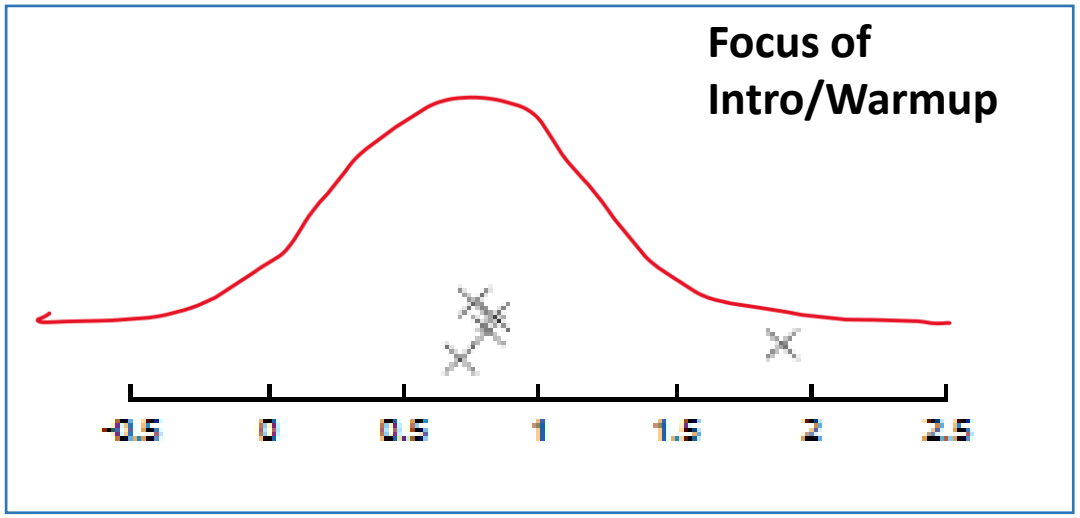
Density Estimation: Problem Statement & Notations

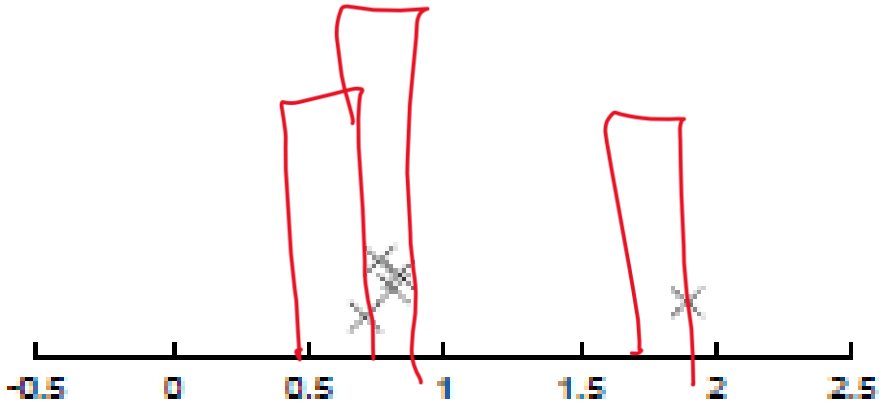
- **Problem:** "Learn (a model) from data" == "Estimate a density/distribution $\mathbb{D}$ from independent observations (i.e., iid samples drawn from $\mathbb{D}$)"

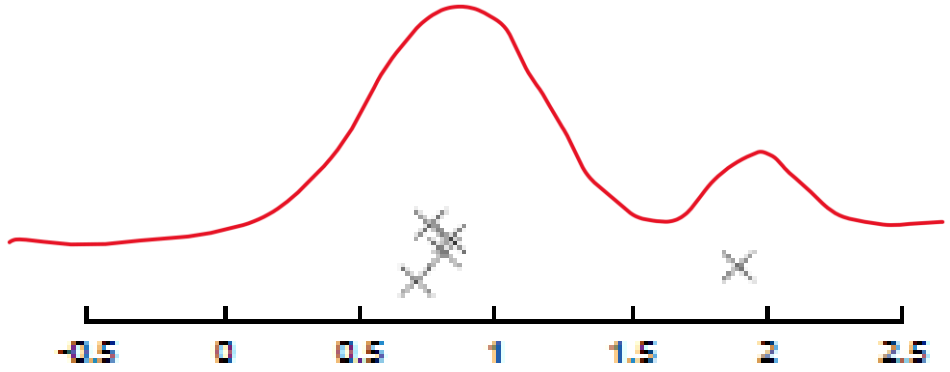
- **Input:** N data points $(\mathrm{x}_1, \ldots, \mathrm{x}_N)^T$ assumed to be iid samples from an unknown probability distribution $\mathbb{D}$
  - $\mathrm{x}_n \sim_{iid} \mathbb{D}$ for all $n = 1, \ldots, N$.
  - $\mathrm{x}_n \in \mathbb{R}^d$

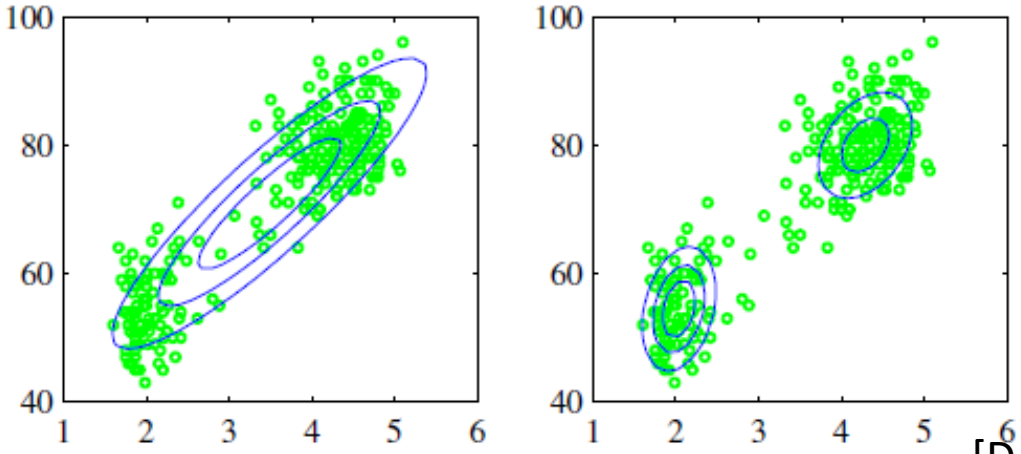
- **Output:** Probability density/distribution $\mathbb{D}$ that "best fits" the data
  - Univariate distbn. if d=1, and Multivariate/Joint distbn. if multiple (d>1) r.v.s are to be modelled (e.g.., fish length, width and color).
  - Family/Form of distributions fixed at "model selection" step to get a well-posed problem.

Density estimation (intuitively in pictures)









[DJM,CMB]

Approaches to Density estimation

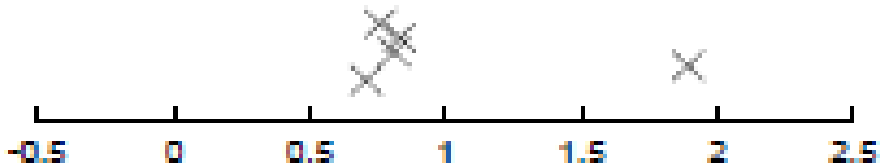
- Parametric approach:
  - some *functional form* of probability distribution D assumed for the data points
  - family of models parameterized by θ i.e., $p(x|\theta)$ or $f(x|\theta)$, with each family member specified by a particular value of the parameter vector θ.
  - Distribution could be simple (e.g., unimodal density) or complex (e.g., multi-modal density, incl. mixture density for mixture models)

- Nonparametric approach:
  - distribution not assumed to be of a functional form specified by a few parameters; instead form of distribution typically depends on the size of the dataset.
  - Still have some "parameters" but they control model complexity (more so than specifying the exact functional form of the distribution)

[CMB]

Warmup: Intuitive depiction of density estimation example

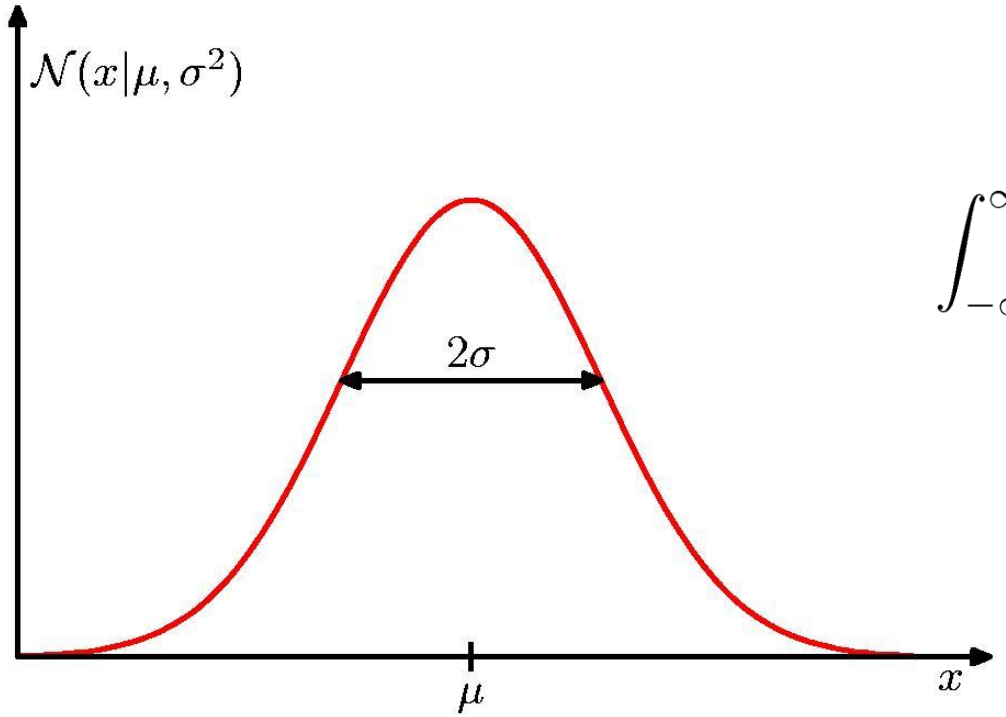
Warmup: Parametric approach on a toy dataset



[DJM]

Recap: The (1D) Gaussian/Normal Distribution

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

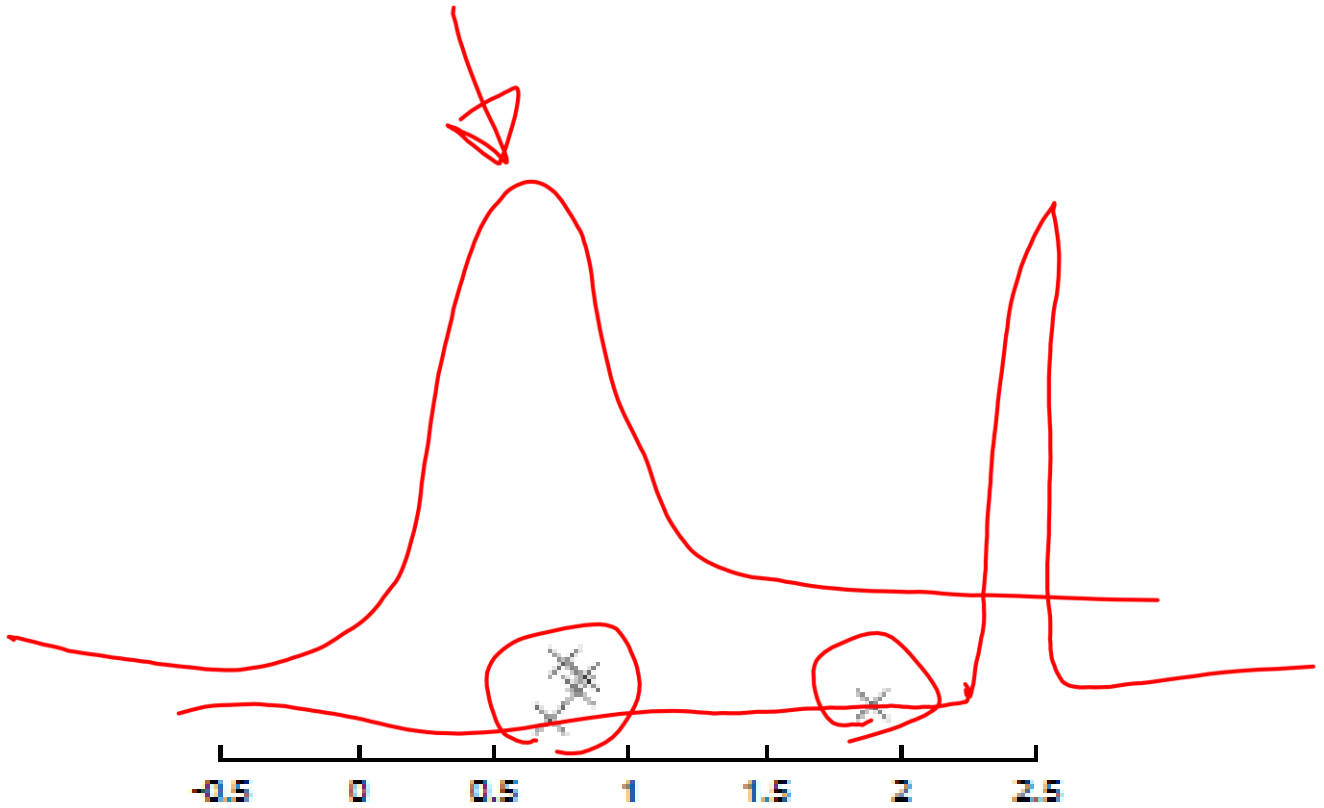


$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

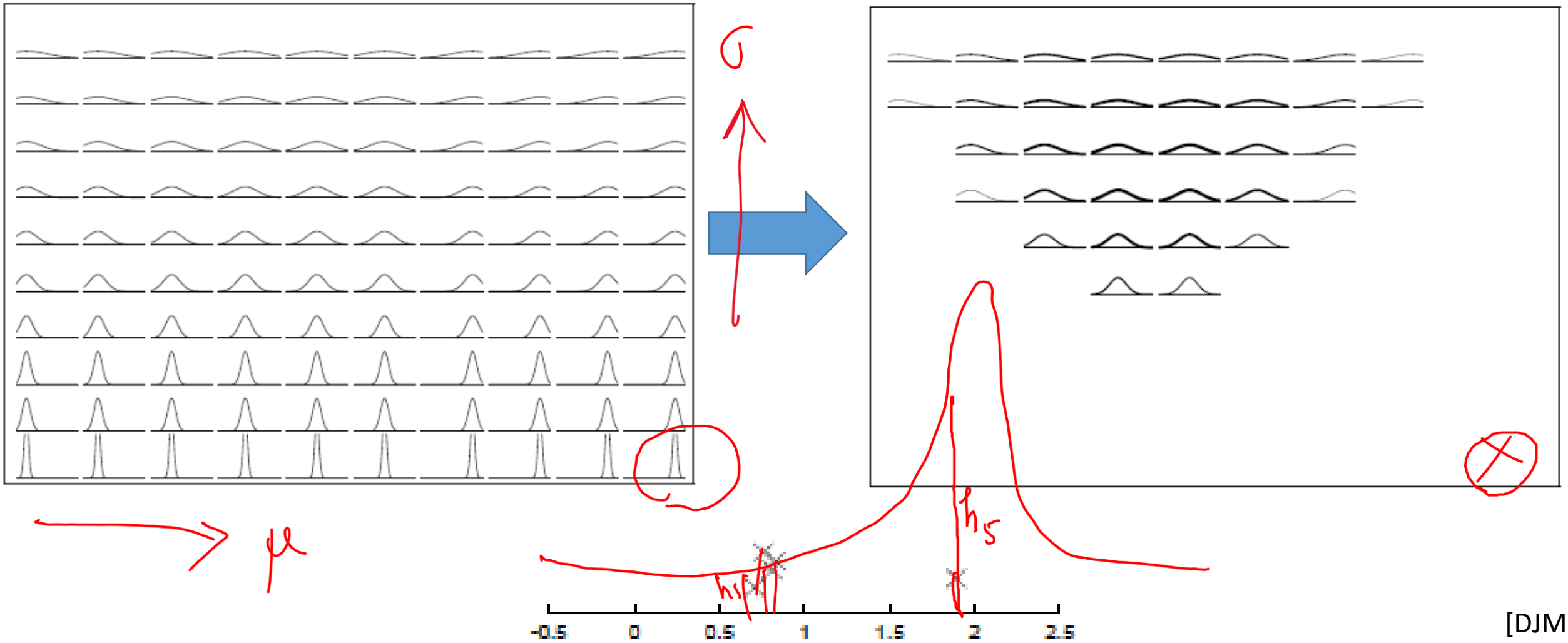
[CMB]

Warmup: How to fit a 1D Gaussian to this data? – Intuition



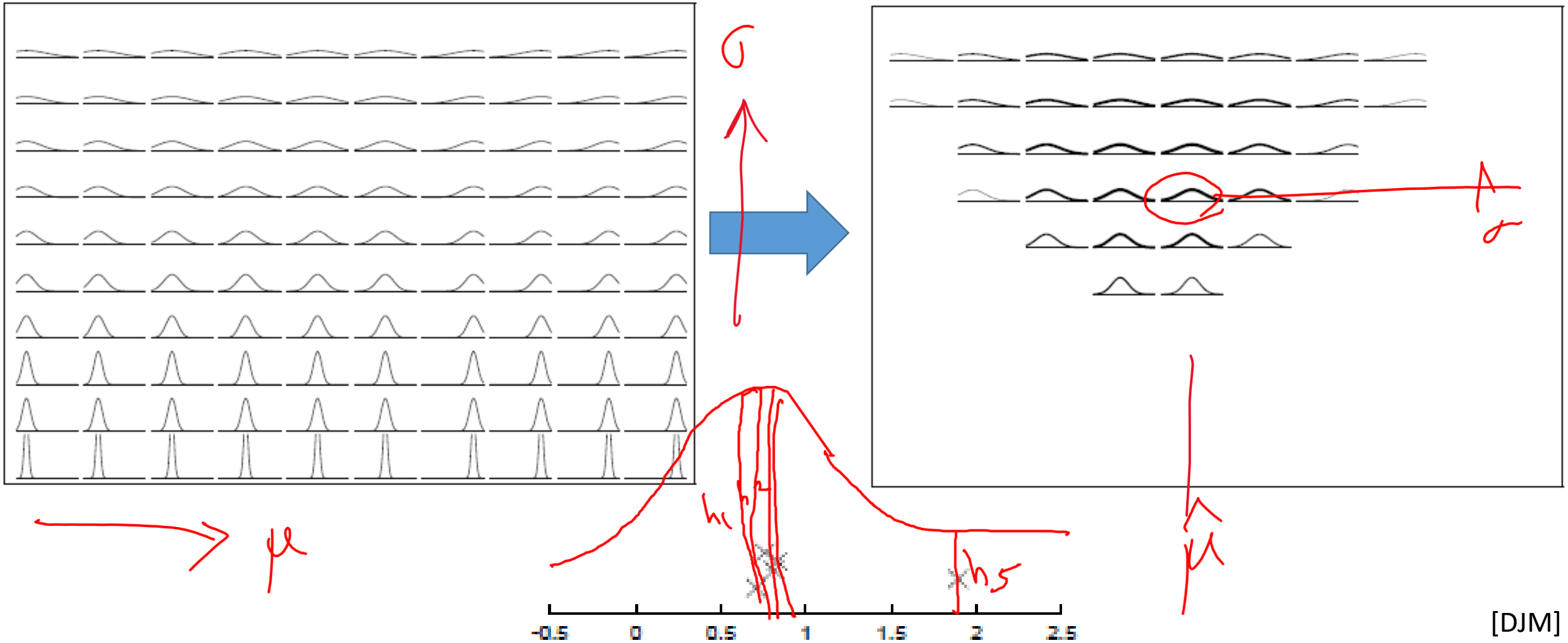
[DJM]

Warmup: How to fit a 1D Gaussian to this data? – “Visual” MLE



[DJM]

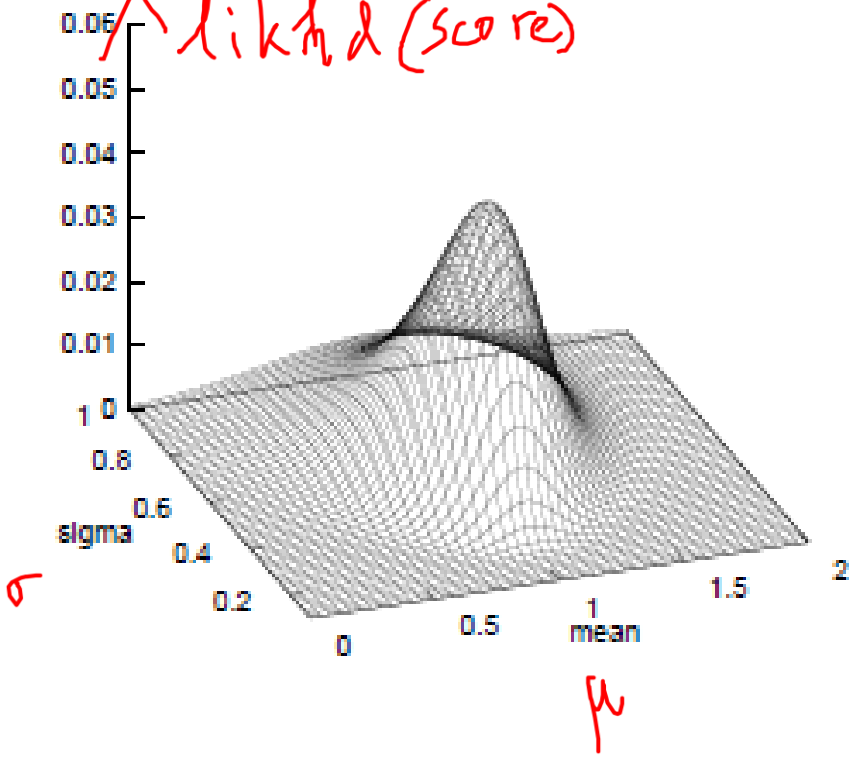
Warmup: How to fit a 1D Gaussian to this data? (contd.)

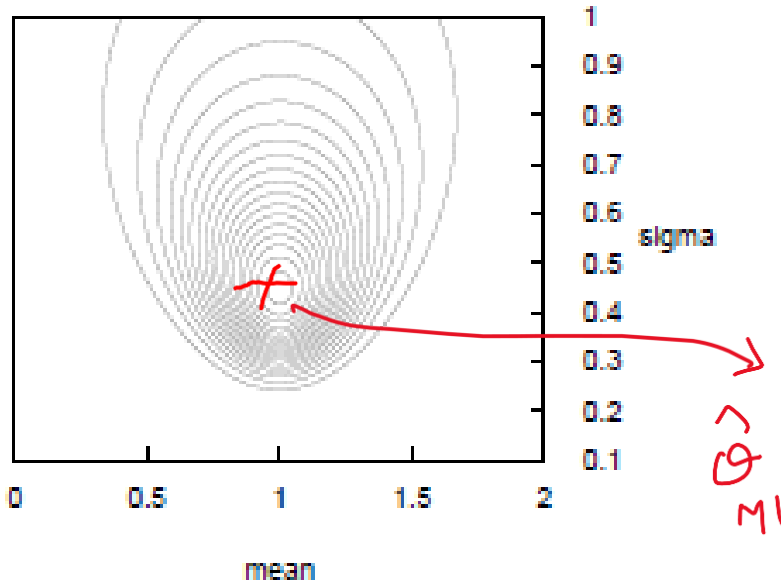


[DJM]

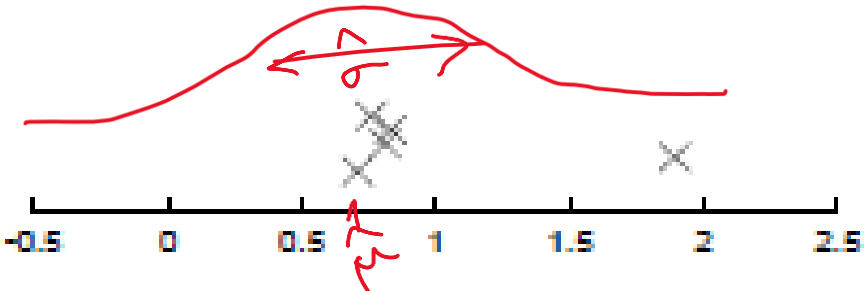
Warmup: MLE for 1D Gaussian (the need for "continuous optimization")

$\theta = (\mu, \sigma)$ $\quad L(\theta) = L(\mu, \sigma)$





$\hat{\theta}_{MLE} = (\hat{\mu}, \hat{\sigma})$



[DJM]

MLE for one 1D Gaussian (closed-form solution)

- Log likelihood:

$$\mathcal{LL}(\mu, \sigma \mid D_N) = \ln P\left(\left\{x^{(n)}\right\}_{n=1}^{N} \mid \mu, \sigma\right) = -N \ln(\sqrt{2\pi}\sigma) - \sum_n \left(x^{(n)} - \mu\right)^2 / (2\sigma^2)$$

- MLE estimates:

$$\hat{\mu} = \sum_{n=1}^{N} \frac{x^{(n)}}{N}, \qquad \hat{\sigma}_N^2 = \frac{\sum_{n=1}^{N} \left(x^{(n)} - \hat{\mu}\right)^2}{N}$$

why not $N-1$?

$x^{(n)} := x_n$

[DJM]

Let’s take a ~~brief~~ detour...

..to “Calculus/Optimization Background”,

and additionally

..to “Linear Algebra Background”, and

..to “MVG Background”.

(above completes all the prereqs./background we need!)

Outline for Module M3

- M3. Density Estimation
  - M3.0 Introduction/Background
  - **M3.1 Parametric methods**
    - **M3.1.1 Maximum Likelihood Estimation (MLE)**

      (**for continuous/discrete densities**, incl. mixture densities (brief mention))
    - M3.1.2 Bayesian Inference(/estimation)
  - M3.2 Nonparametric methods (not covered)

MLE approach

- Dataset D or $D_N = \{x_1, \ldots, x_N\}$ (iid samples from $p(x|\theta)$; p denotes pmf or pdf)
- Likelihood (function of parameters, given the data, is used as the score function):

$$\mathcal{L}(\theta; D_N) = \; p(\{x_1, \ldots, x_N\} \mid \theta) = \prod_{n=1,\ldots,N} p(x_n|\theta)$$

- ML Estimate
  (opt. problem, solved analytically or numerically):

$$\hat{\theta}_N = \arg\max_{\theta} \mathcal{L}(\theta; D_N)$$

- Has desirable properties, mainly consistency (for "most" densities).
  MLE converges in probab. to the true parameter(s):

$$\text{Let } P(|\hat{\theta}_N - \theta| > \varepsilon) = 0 \quad \forall \varepsilon > 0$$
$$N \to \infty \quad D_N$$

[PSS]

Examples we will see:

1) Gaussian (uni- and multi-variate)
2) Bernoulli
3) Categorical/Multinoulli

MLE for 1D Gaussian (general N datapoints)

$D(x_1, \ldots, x_N) \overset{\text{i.i.d}}{\sim} p(\text{1D Gauss})$.

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

$[\mu, \sigma]$

$$\mathcal{LL}(\theta; D) = \sum_{n=1}^{N} \log \mathcal{N}(x_n | \mu, \sigma^2)$$

$$= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_n ()(n - \mu)^2$$

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

Rough space for illustrations

Let $\lambda = \frac{1}{\sigma^2}$. $\quad \mathcal{LL}(\theta; D_N) = \mathcal{LL}(\mu, \lambda) = -\frac{N}{2}\log\left(\frac{2\pi}{\lambda}\right) - \frac{\lambda}{2}\sum_n (x_n - \mu)^2$

$\lambda > 0$ (precision)

$\mathcal{LL}(\lambda, \mu)$ not jointly concave in λ, μ

but concave separately in λ (or μ)

Set $\frac{\partial \mathcal{LL}}{\partial \mu} = 0$

$\Rightarrow +\frac{\lambda}{2} 2 \sum_n (x_n - \mu) = 0$

$\Rightarrow \sum_n x_n = N\mu \Rightarrow \hat{\mu}_{ML} = \frac{1}{N}\sum_n x_n$

[First, $\mathcal{LL}$ is concave in μ.

Next, $\mathcal{LL}$ is also concave in λ (holds for $\mu = \hat{\mu}_{ML}$)

(⊗ indept. of $\hat{\lambda}_{ML}$!!)

MLE for one 1D Gaussian

- Log likelihood:

$$\mathcal{LL}(\mu, \sigma \mid D_N) = \ln P\left(\left\{x^{(n)}\right\}_{n=1}^{N} \mid \mu, \sigma\right) = -N \ln(\sqrt{2\pi}\sigma) - \sum_n \left(x^{(n)} - \mu\right)^2 / (2\sigma^2)$$

- MLE estimates:

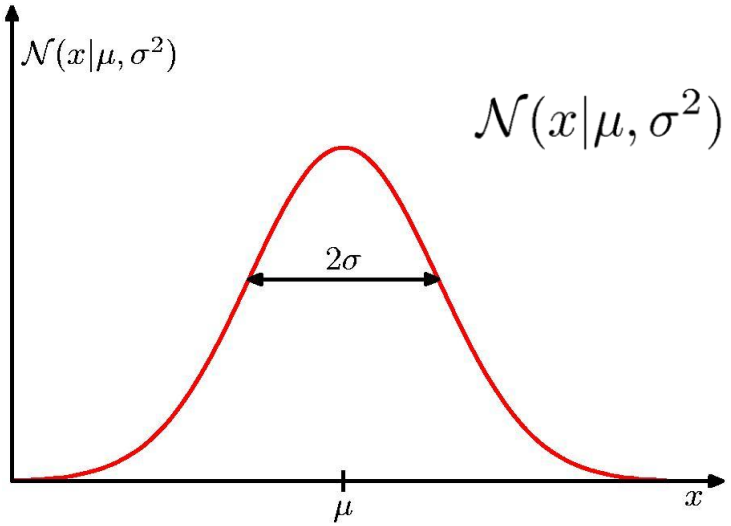
$$\hat{\mu} = \sum_{n=1}^{N} \frac{x^{(n)}}{N}, \qquad \hat{\sigma}_N^2 = \frac{\sum_{n=1}^{N} \left(x^{(n)} - \hat{\mu}\right)^2}{N}$$

why not $N-1$?

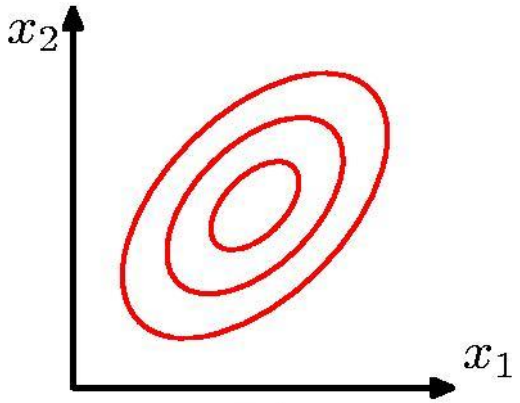
$x^{(n)} := x_n$

[DJM]

From uni- to multi-variate Gaussian



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



$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

[CMB]

Matrix/Vector Derivative Info Needed!

- Defn. of Partial derivatives of a real-valued function $f(x, A)$ wrt vector x or matrix A:
  - $\left(\frac{\partial f}{\partial x}\right)_i = \frac{\partial f}{\partial x_i}$ (also known as gradient of f as a function of x)
  - $\left(\frac{\partial f}{\partial A}\right)_{ij} = \frac{\partial f}{\partial a_{ij}}$

- Facts on Partial derivatives (wrt vector or matrix of parameters):
  - $\frac{\partial}{\partial x} x^T A\, x = A^T x + Ax$ (or $2Ax$ if A is symmetric)
  - $\frac{\partial}{\partial A} x^T A\, x = xx^T$ (outer-product)
  - $\frac{\partial}{\partial A} \log|A| = A^{-T}$ (derivation more involved)

Maximum Likelihood for the Gaussian (1)

- Given i.i.d. data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$, the log likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

[CMB]

Maximum Likelihood for the Gaussian (2)

- Set the gradient of the log likelihood function to zero,

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0$$

- and solve to obtain

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n.$$

- Similarly

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

[CMB]

Derivation of MLE of Multi-variate Gaussian

- Recall: Facts on Partial derivatives (wrt vector or matrix of parameters):
  - $\frac{\partial}{\partial x} x^T A\, x = A^T x + Ax$ (or $2Ax$ if A is symmetric)
  - $\frac{\partial}{\partial A} x^T A\, x = xx^T$ (outer-product)
  - $\frac{\partial}{\partial A} \log |A| = A^{-T}$

- Gradient of $LL(\mu, \Lambda) := LL(\mu, \Sigma^{-1})$

$$l(\Sigma \mid \mathcal{D}) = -\frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_n (x_n - \mu)^T \Sigma^{-1} (x_n - \mu) \qquad (13.48)$$

$$\frac{\partial l}{\partial \Sigma^{-1}} = \frac{N}{2} \Sigma - \frac{1}{2} \sum_n (x_n - \mu)(x_n - \mu)^T. \qquad (13.51)$$

[From Secn. 13.5 of The Multivariate Gaussian from https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/other-readings/chapter13.pdf]

Outline for Module M3

- M3. Density Estimation
  - M3.0 Introduction/Background
  - **M3.1 Parametric methods**
    - **M3.1.1 Maximum Likelihood Estimation (MLE)**

      (for continuous/**discrete densities, incl. mixture densities (brief mention)**)
    - M3.1.2 Bayesian Inference(/estimation)
  - M3.2 Nonparametric methods (not covered)

Example 2: Bernoulli/Binary RVs

- Coin flipping: heads=1, tails=0

$$p(x = 1|\mu) = \mu$$

- Bernoulli Distribution

$$\begin{aligned} \text{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \text{var}[x] &= \mu(1-\mu) \end{aligned}$$

[CMB]

(Parametric) Density Estimation / Parameter Estimation / Parameter learning

- ML for Bernoulli
- Given: $\mathcal{D} = \{x_1, \ldots, x_N\},\ m$ heads $(1),\ N - m$ tails (0)
-

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

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n)\ln(1-\mu)\}$$

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N}$$

[CMB]

Example 3: From Bernoulli to Multinoulli

Categorical (Multinoulli) Variables

1-of-K coding scheme: $\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}}$

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

$$\forall k : \mu_k \geqslant 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1$$

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

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

[CMB]

ML Parameter estimation

- Given: $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\left(\sum_n x_{nk}\right)} = \prod_{k=1}^{K} \mu_k^{m_k}$$

- Ensure $\sum_k \mu_k = 1$, use a Lagrange multiplier.

$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

$$\mu_k = -m_k/\lambda \qquad \mu_k^{\mathrm{ML}} = \frac{m_k}{N}$$

[CMB]

Aside in Appendix

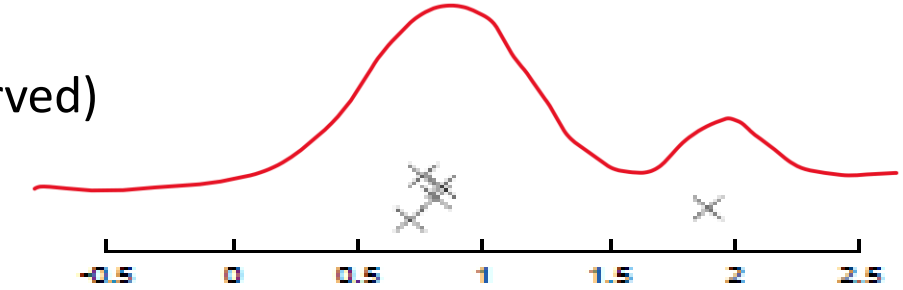
- An Aside: Relation between Bernoulli and Binomial distribution
- An Aside: Relation between Categorical and Multinomial Distribution

Example mixture density (very brief mention)

Mixture Model or Latent Variable Model (LVM)

$Z \sim Bernoulli(\pi)$ (latent variable, not observed)

$(X|Z = z) \sim \mathcal{N}(\mu_z, \sigma_z^2)$ (X is observed)



MLE of $\theta = (\boldsymbol{\pi},\ \mu_1, \sigma_1,\ \mu_2, \sigma_2)$ based on:

$$\mathcal{L}(\theta|D_N) = \prod_{n=1}^{N}(\boldsymbol{\pi}\, \mathcal{N}(x_n|\mu_1, \sigma_1^2) + (\mathbf{1} - \boldsymbol{\pi})\, \mathcal{N}(x_n|\mu_2, \sigma_2^2))$$

$\hat{\theta}_{MLE} = \text{argmax}_\theta \mathcal{L}(\theta|D_N)$ (using numerical methods like Newton-Raphson, or Expectation-Maximization (EM) algorithm)

Putting it together: Density Estimation + Decision Theory to build a Bayes classifier $h(x)$

Focus on a generative classification model: $p(x, y) = p(y)p(x|y)$. Building a Bayes classifier involves the following steps:

- **Inference:** Learn $p(y), p(x|y)$ from training data.
  - **Sometime, Naïve** Bayes Assumption used, i.e., conditional indep. of features given class label assumed, i.e.,
    $$p(x|y) = \prod_{i=1}^{p} p\left(x^{(i)}|y\right)$$
  - Use Bayes thm. to derive posterior $p(y|x)$.

- **Decision:** For a new datapoint x_{new}, calculate the posterior $p(y|x_{new})$ and use it
  - with a standard 0-1 loss function (or general loss matrix) to predict its class label (make a decision).
  - Write down the decision regions/boundaries of the learnt classifier $h(x)$.

Let's look at an example similar to Assignment 1 "Putting it together" question.

Example

- What is the Bayes classifier to predict Y given X, given these training data points?

x	1	5	2	3	10	6	11
y	1	-1	1	1	-1	1	-1

Outline for Module M3

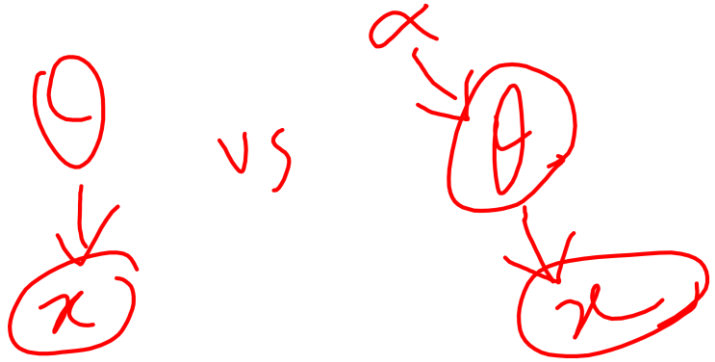
- M3. Density Estimation
  - M3.0 Introduction/Background
  - **M3.1 Parametric methods**
    - M3.1.1 Maximum Likelihood Estimation (MLE)
    - **M3.1.2 Bayesian Inference(/estimation)**
  - M3.2 Nonparametric methods (not covered)

Motivation: Why go from MLE to Bayesian inference?

- Small sample sizes - overfitting to training data $\mathcal{D}$
  - $\mathcal{D} = \{1, 1, 1\} \rightarrow \mu_{\mathrm{ML}} = \frac{3}{3} = 1$ ⇒ Prediction: *all* future tosses will land heads up
  - Laplace's sunrise problem: What is the probability that the sun will rise tomorrow? [https://en.wikipedia.org/wiki/Sunrise_problem]
- Prior information
  - MLE cannot use additional information we may have about the parameter!
- Richer (compound or hierarchical) distbns. to fit the data, and robustness to outliers
  - Treating parameters as r.v.s with their own distributions can offer a "natural" plug-and-play hierarchical modelling framework to construct complex distbns. (marginal distbns. with heavy tails or overdispersion, etc.) that fit the data better.

[CMB]

Bayesian approach



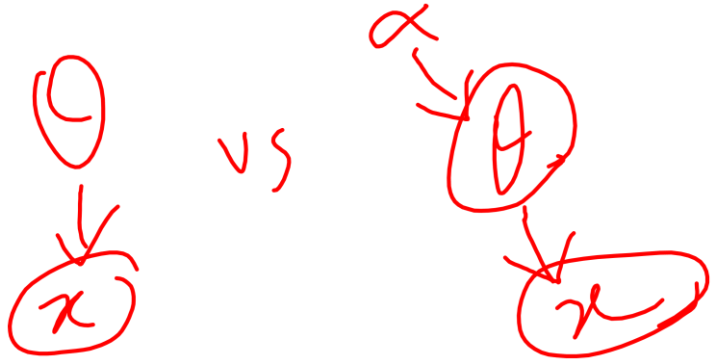
- Bayesian approach in theory: View parameter θ as a r.v. and not as a fixed constant as in MLE
  - Why Bayesian? ML (frequentist/Fisherian) approach gives useful/consistent estimators for many distbns., but fails for small sample sizes and doesn't permit incorporation of additional info. about the parameter!
  - Information about the r.v. before seeing the data is encoded as a prior distribution $P(\theta)$
  - Use Bayes rule to get posterior that captures your degree of belief/uncertainty about θ after seeing the data:

$$P(\theta|D_N) \propto P(\theta)\, P(D_N|\theta)$$

posterior $\propto$ prior x likelihood

[PSS,CMB]

Bayesian approach



- Bayesian approach in theory: View parameter θ as a r.v. and not as a fixed constant as in MLE
  - Why Bayesian? ML (frequentist/Fisherian) approach gives useful/consistent estimators for many distbns., but fails for small sample sizes and doesn't permit incorporation of additional info. about the parameter!
  - Information about the r.v. before seeing the data is encoded as a prior distribution $P(\theta)$
  - Use Bayes rule to get posterior that captures your degree of belief/uncertainty about θ after seeing the data:

$$P(\theta|D_N) \propto P(\theta)\, P(D_N|\theta)$$

posterior $\propto$ prior x likelihood

- Bayesian approach in practice:
  - Conjugate priors make calcn./interpretn. easy by ensuring posterior & prior follow same distbn.
    - But may not be applicable always (use approximate inference such as MCMC/Gibbs sampling for more complex priors)
    - What about that pesky hyperparameter (i.e., pseudocounts for beta distbn.)?
  - Full (posterior) distribution vs. a point estimate?
    - Posterior mode (MAP) or Posterior mean – a practical resort
    - an ideal Bayesian can integrate over uncertainty around the parameter - posterior predictive distbn.

[PSS,CMB]

Three examples again:

Bayesian inference for:

Example 2: Bernoulli

Example 3: Categorial/Multinoulli

Example 1: Gaussian (mostly 1D, optionally multi-variate in Appendix)

Example 2: Bayesian inference for Bernoulli

What is a good prior?

What is a good prior? Beta Distribution

- Distribution over $\mu \in [0, 1]$

$$\begin{aligned} \text{Beta}(\mu|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\ \mathbb{E}[\mu] &= \frac{a}{a+b} \\ \text{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

$\propto \mu^{a-1}(1-\mu)^{b-1}$

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$$

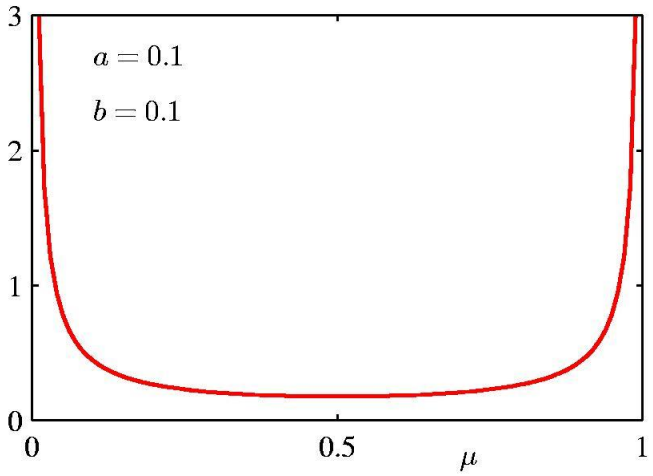
(converges for z > 0)

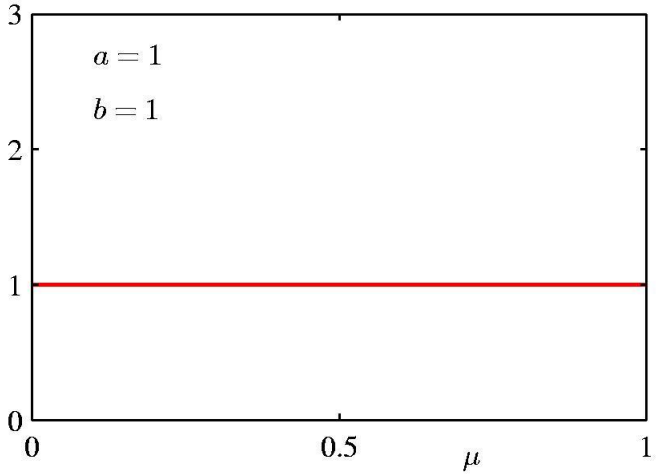
$\Gamma(1) = 1$

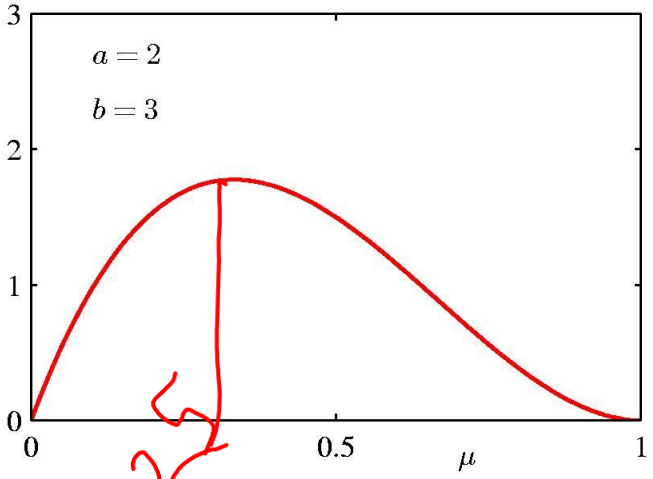
$\Gamma(z+1) = z\Gamma(z)$; $\Gamma(z+1) = z!$

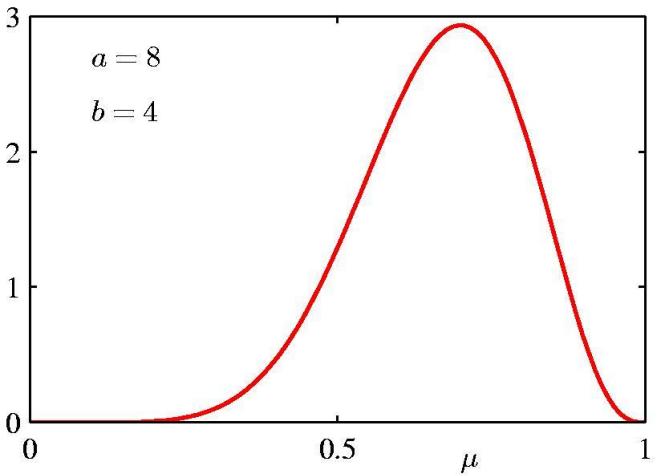
[CMB]

Beta Distribution









$$\frac{a-1}{a+b-2} \quad \text{(mode)}$$

[CMB]

Bayesian Bernoulli

$$
\begin{aligned}
p(\mu|a_0, b_0, \mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0) \\
&= \left(\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}\right) \mathrm{Beta}(\mu|a_0, b_0) \\
&\propto \mu^{m+a_0-1}(1-\mu)^{(N-m)+b_0-1} \\
&\propto \mathrm{Beta}(\mu|a_N, b_N)
\end{aligned}
$$

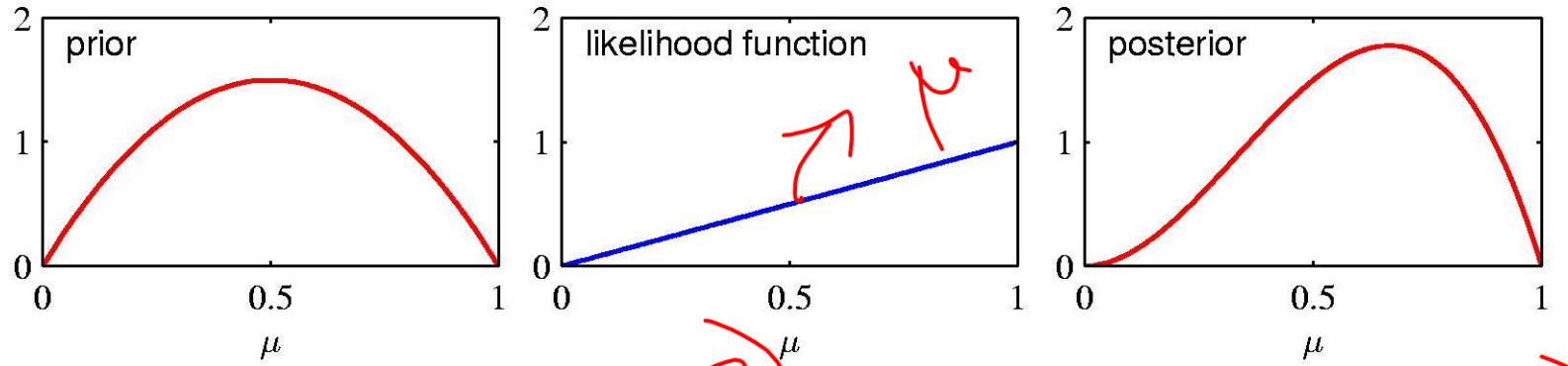
$$a_N = a_0 + m \qquad b_N = b_0 + (N - m)$$

Beta distribution is the *conjugate* prior for the parameter of the Bernoulli distribution (or Bernoulli likelihood fn.).

post ∝ prior × likelihd
(Beta) (Beta) (Bern.)

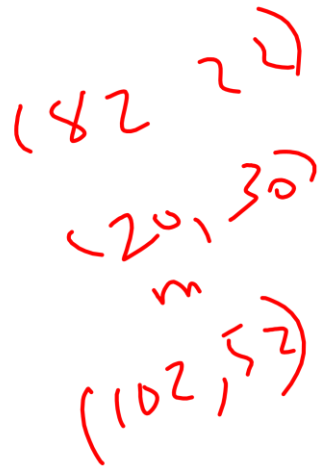
[CMB]

Bayesian inference in action: Beta-Bernoulli (Prior · Likelihood = Posterior)



[CMB]

Pseudocounts, and updating these counts with new data - example

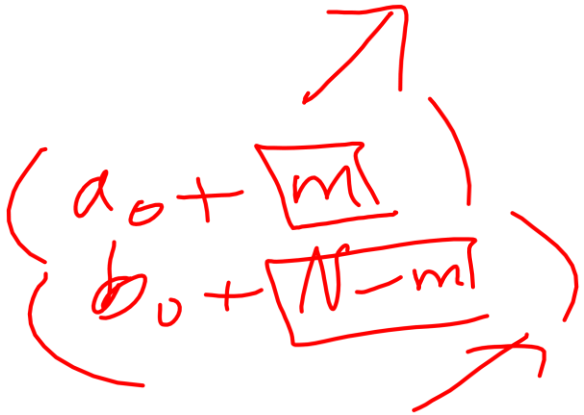


[CMB]

Properties of the Posterior

As the size of the data set, N, increase

$$
\begin{aligned}
a_N &\to m \\
b_N &\to N - m \\
\mathbb{E}[\mu] &= \frac{a_N}{a_N + b_N} \to \frac{m}{N} = \mu_{\mathrm{ML}} \\
\mathrm{var}[\mu] &= \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \to 0
\end{aligned}
$$



Under certain assumptions (cf. Bernstein–von Mises thm)

Lt $N \to \infty$ | post mean → MLE or post mode (MAP)

[CMB]

Point estimate vs. using the full posterior: Prediction under the (full) posterior

What is the probability that the next coin toss will land heads up?

Posterior predictive

[CMB]

Point estimate vs. using the full posterior: Prediction under the (full) posterior

What is the probability that the next coin toss will land heads up?

$$
\begin{aligned}
p(x = 1|a_0, b_0, \mathcal{D}) &= \int_0^1 p(x = 1|\mu) p(\mu|a_0, b_0, \mathcal{D}) \, \mathrm{d}\mu \\
&= \int_0^1 \mu p(\mu|a_0, b_0, \mathcal{D}) \, \mathrm{d}\mu \\
&= \mathbb{E}[\mu|a_0, b_0, \mathcal{D}] = \frac{a_N}{b_N}
\end{aligned}
$$

$$\frac{a_N}{a_N + b_N}$$

Link back to density estmn. $p(x|\mathcal{D}_N)$

[CMB]

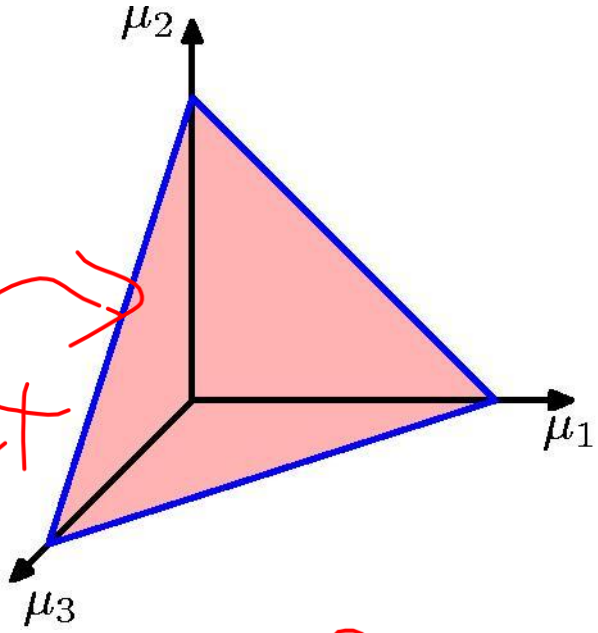
Example 3: Bayesian inference for Categorical/Multinoulli

Dirichlet Distribution for the Prior

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

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

Conjugate prior for the categorical likelihood fn. i.e., Dirichlet distbn. is conjugate prior for the parameters of the Categorical distbn.



Post ∝ prior × lik

Dir. Dir. cat.

simplex

$\{\mu \in \mathbb{R}^3_{\geq 0} : \sum_i \mu_i = 1\}$

[CMB]

Bayesian Categorical

$(m_1, m_2, \ldots, m_k, \ldots, m_K)$

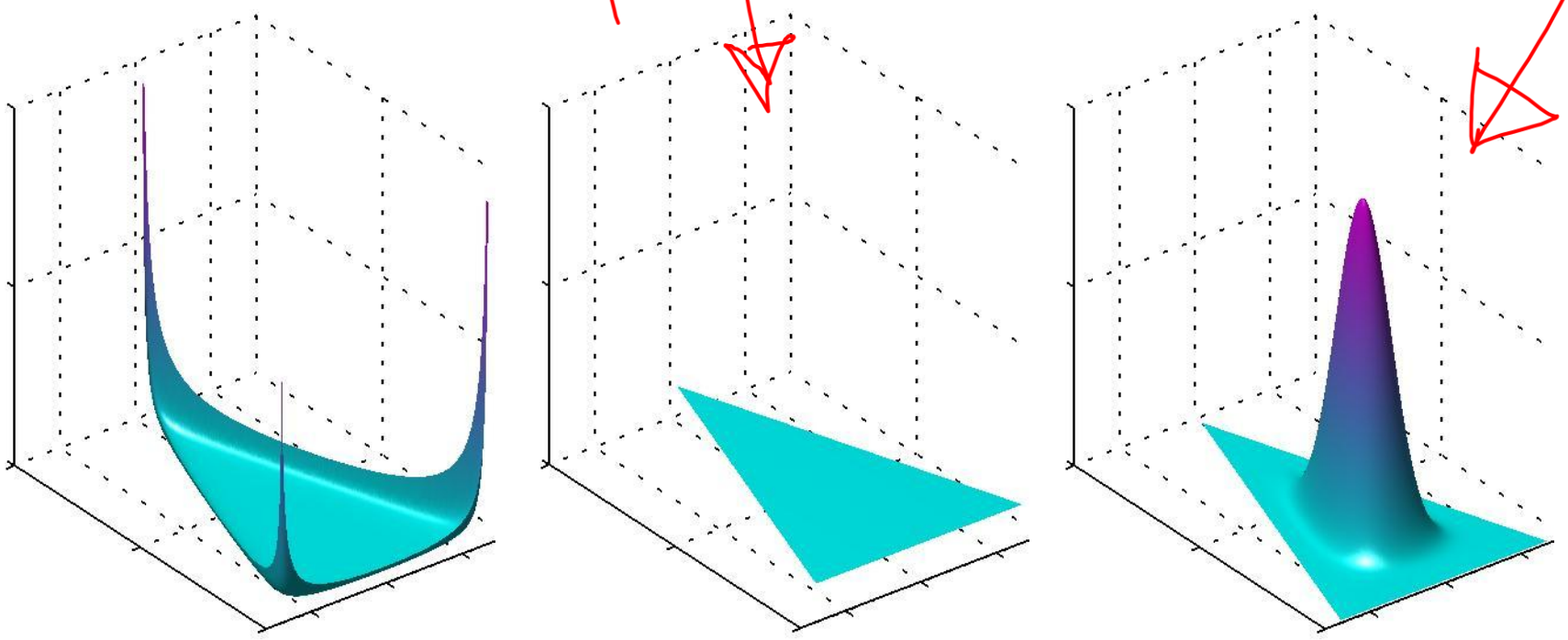
$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu}) p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$$

post ∝ cat · Dir

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

[CMB]

Bayesian Categorical



$\alpha_k = 10^{-1}$ $\alpha_k = 10^{0}$ $\alpha_k = 10^{1}$

[CMB]

Example 1: Bayesian inference for 1D Gaussian?

Bayesian Inference for the Gaussian (1)

- Assume σ^2 is known. Given i.i.d. data $\mathbf{x} = \{x_1, \ldots, x_N\}$, the likelihood function for μ is given by

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

- This has a Gaussian shape as a function of μ (but it is *not* a distribution over μ).

Bayesian Inference for the Gaussian (2)

- Combined with a Gaussian prior over μ,

$$p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right).$$

- this gives the posterior

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu).$$

- Completing the square over μ, we see that

$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$

Bayesian Inference for the Gaussian (3)

- … where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

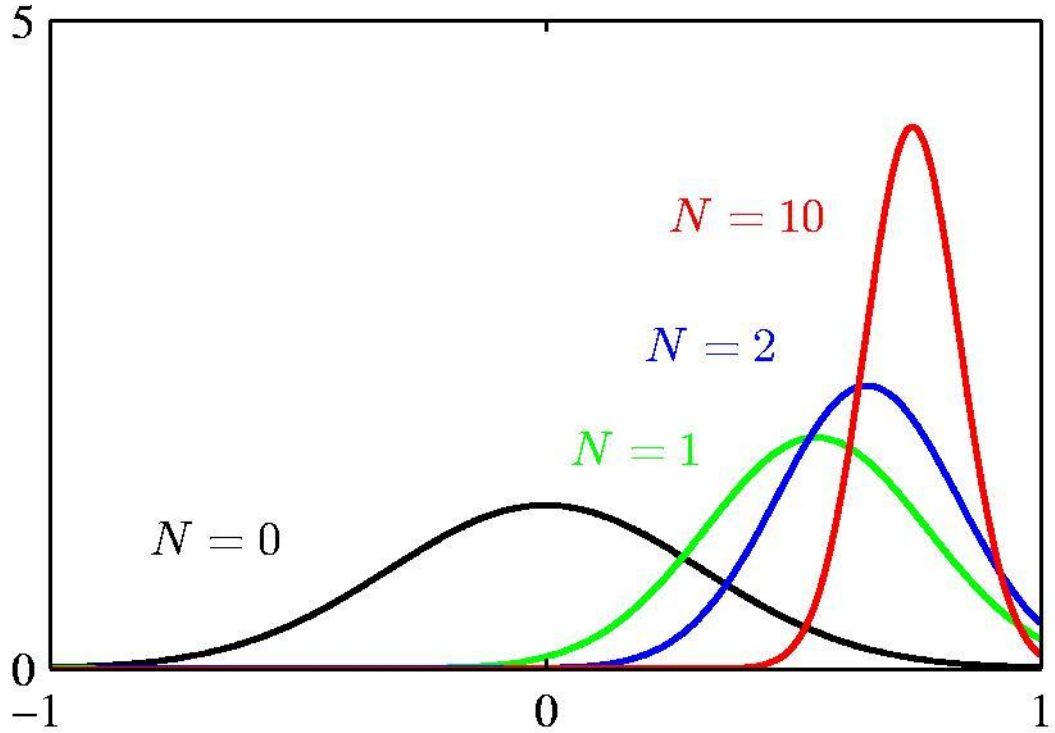
$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

- Note:

	$N = 0$	$N \to \infty$
μ_N	μ_0	μ_{ML}
σ_N^2	σ_0^2	0

Bayesian Inference for the Gaussian (4)

- Example: $p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$ for N = 0, 1, 2 and 10.



Bayesian Inference for the Gaussian (5)

- Sequential Estimation

$$
\begin{aligned}
p(\mu|\mathbf{x}) &\propto p(\mu)p(\mathbf{x}|\mu) \\
&= \left[p(\mu)\prod_{n=1}^{N-1} p(x_n|\mu)\right] p(x_N|\mu) \\
&\propto \mathcal{N}\left(\mu|\mu_{N-1},\sigma^2_{N-1}\right) p(x_N|\mu)
\end{aligned}
$$

- The posterior obtained after observing N-1 data points becomes the prior when we observe the N^{th} data point.

Bayesian Inference for the Gaussian (6)

- Now assume μ is known. The likelihood function for $\lambda = 1/\sigma^2$ is given by

$$p(\mathbf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

- This has a Gamma shape as a function of λ.

- (cf. Appendix for more on Bayesian inference of Gaussian)

Outline for Module M3

- M3. Density Estimation
  - M3.0 Introduction/Background
  - M3.1 Parametric methods
  - **M3.2 Nonparametric methods (not covered)**
    - **M3.2.0 General idea**
    - M3.2.1 K-Nearest Neighbors

Nonparametric Methods – Why?

- Parametric distribution models are restricted to specific forms, which may not always be suitable; for example, consider modelling a multimodal distribution with a single, unimodal model.

- Nonparametric approaches make few assumptions about the overall shape of the distribution being modelled.

- Model family is specified by:
  - Finite # of params. -> parametric
  - Everything else (infinite # of params.) -> nonparametric
    - Typically, has flexible # of params. that grows with sample size, with a smoothing param. that controls model complexity

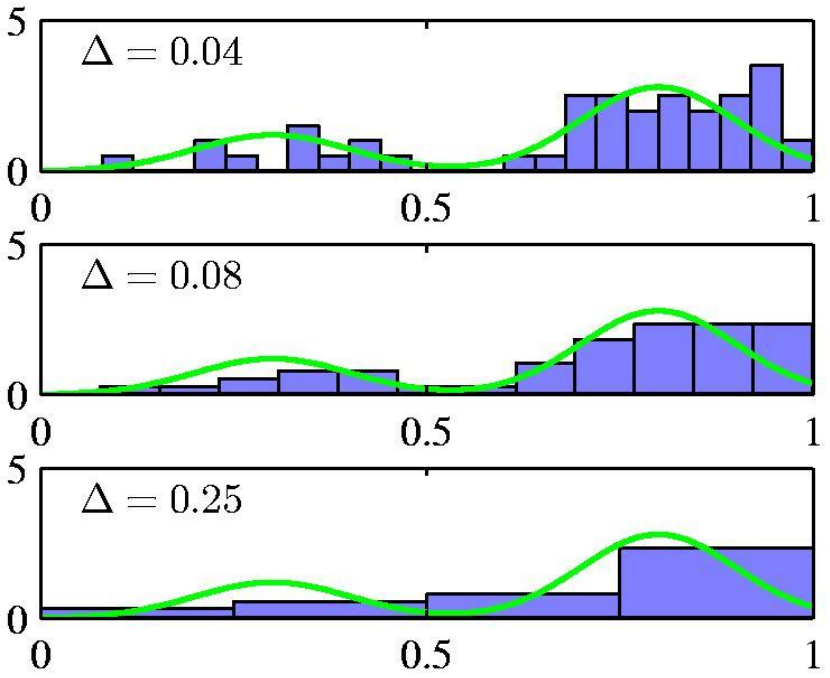
[CMB]

Nonparametric Methods - Warmup

- **Histogram methods** partition the data space into distinct bins with widths Δ_i and count the number of observations, n_i, in each bin.

$$p_i = \frac{n_i}{N\Delta_i}$$

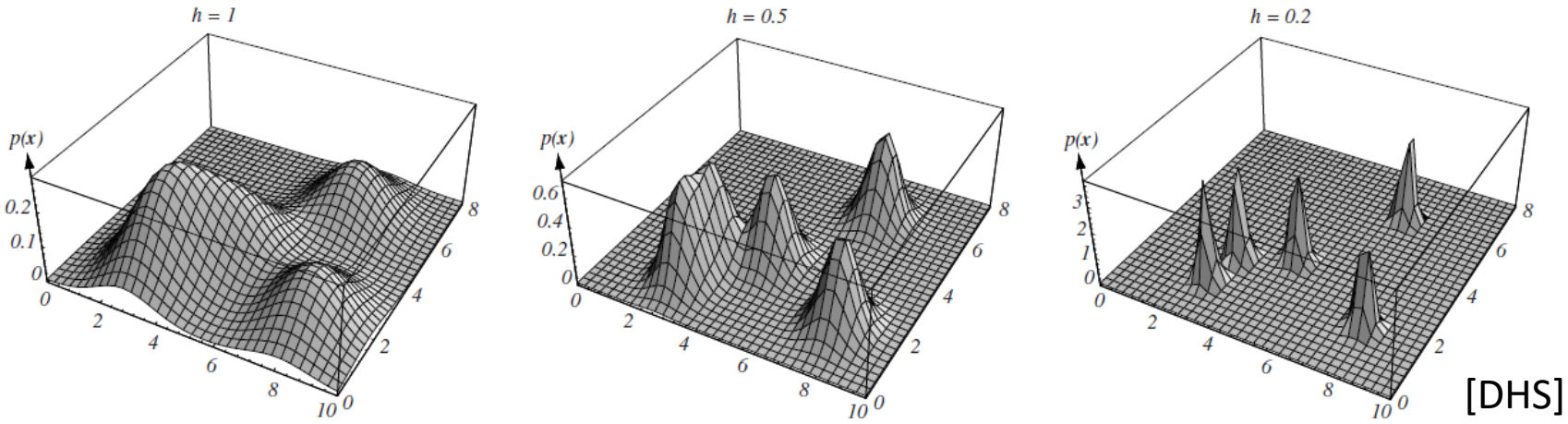
- Often, the same width is used for all bins, $\Delta_i = \Delta$.
- Δ acts as a smoothing parameter.



[CMB]

Problems with the histogram method

- In a D-dimensional space, using M bins in each dimension will require M^D bins! Are there other methods that can scale better with number of dimensions?
- Are some optimal choices of binwidths possible?



[DHS]

[CMB]

Extending the “local region around x” idea…

•Assume observations drawn from a density p(x) and consider a small region R containing x such that

$$P = \int_{\mathcal{R}} p(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

•The probability that K out of N observations lie inside R is

Binom(K | N,P) and if N is large

$$K \simeq NP.$$

If V := volume(R) is sufficiently small, p(x) is approximately constant over R and

$$P \simeq p(\mathbf{x})V$$

Thus

$$p(\mathbf{x}) = \frac{K}{NV}.$$

V small: to get K>0, need large N.
V large: p(x) constant approx. fails!

[CMB]

Two class of “local region around x” methods

- fix V, estimate K from the data – **Parzen window (aka Parzen or more generally, Kernel density)** estimation
- fix K, estimate V from the data – **K-Nearest Neighbors (k-NN)** method

- Both kernel and k-NN density estimators are consistent i.e., converge to the true probability density as $N \rightarrow \infty$, provided V shrinks suitably with N and K grows with N.

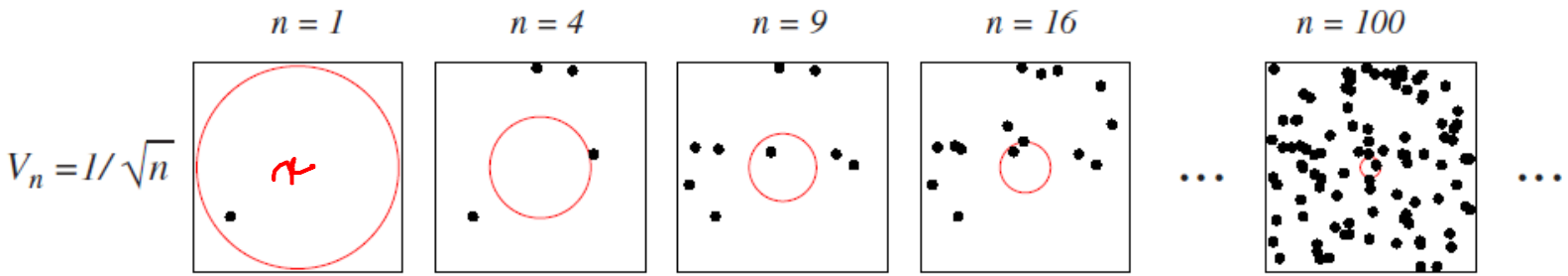
[CMB]

Outline for Module M3

- M3. Density Estimation
  - M3.0 Introduction/Background
  - M3.1 Parametric methods
  - **M3.2 Nonparametric methods (not covered)**
    - M3.2.0 General Idea
    - **M3.2.1 K-Nearest Neighbors**

Recall: “local region around x” two approaches (**fix K** for K-NN)

$$p(\mathbf{x}) = \frac{K}{NV}.$$

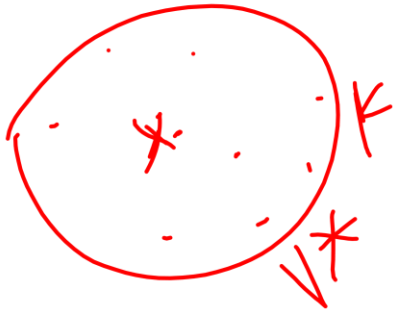


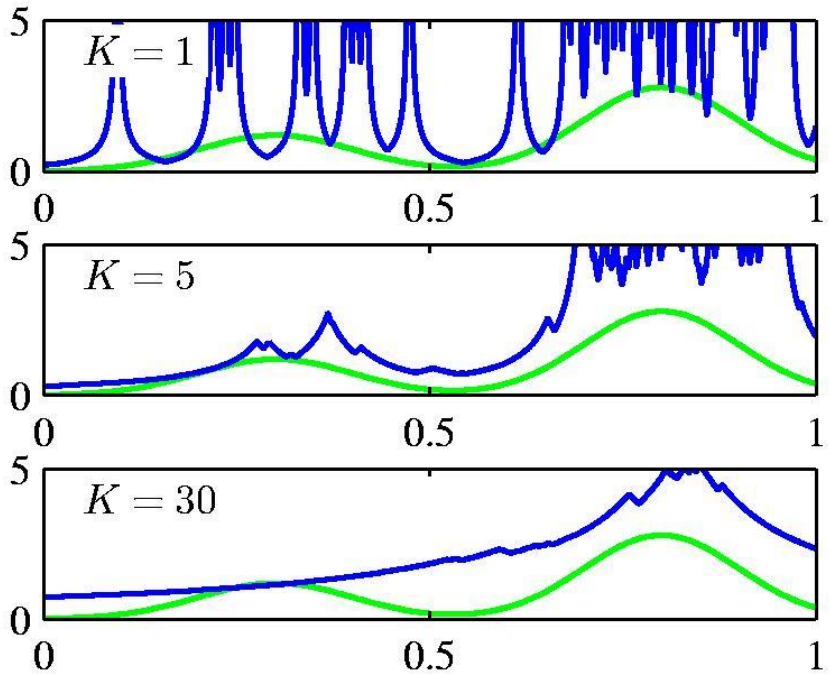
[DHS]

K-NN density estimation

- fix K, estimate V from the data.
- Consider a small hypersphere centred on x and let it grow to a volume V^* that includes precisely K of the given N data points. Then

$$p(\mathbf{x}) \simeq \frac{K}{NV^\star}.$$





K acts as a smoother.

[CMB]

K-NN density estimation can be used to…

…build a k-NN classifier. See Worksheet.

Pros/cons of nonparametric vs parametric…

- Nonparametric models (exclude histograms) requires storing and computing with the entire data set; however can model complex distributions.
  - “For every complex question, there is a simple answer… and it is wrong.” by HL. Mencken

- Parametric models, once fitted, are much more efficient in terms of storage and computation; but assume a relatively simple functional form.
  - “All models are wrong, but some are useful.” by George Box

Summary (of Density Estimation)

- Density estimation using frequentist/Fisherian (MLE) and Bayesian approaches for parametric models.
  - Simple models (Bernoulli, Categorical, 1D Gaussian)
  - Complex models later if time permits (Mixture of 1D Gaussians or GMMs using EM algo.)

- Density estimation using frequentist (Parzen/kernel-density and k-NN) for non-parametric methods.
  - Bayesian nonparametric methods not covered (e.g., Dirichlet process).

- A look back and a look ahead:
  - Density estimation helps us implement the **Bayes classifiers** we saw earlier as part of Decision Theory (by allowing us to learn these classifiers i.e., learn class prior and class conditional densities from data).
  - Density estimation has many appns.; e.g., linear regression is simply density est. of P(t | x), k-means clustering involves mixture density estimation, etc.

Thank you!
(appendix follows)

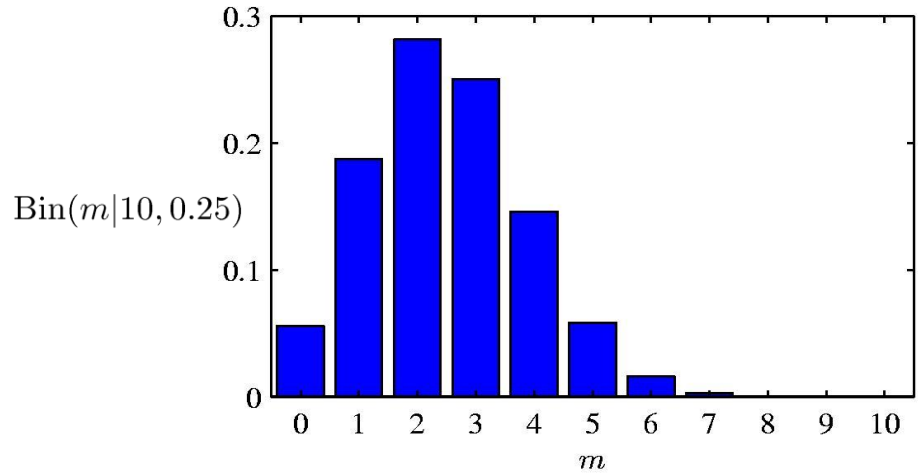
Appendix on

- Aside on Bernoulli-Binomial and Categorical-Multinomial relationship
- Bayesian inference of Gaussian (contd.) - optional

An Aside: Relation between Bernoulli and Binomial distribution

- N coin flips:

$$p(m \text{ heads}|N, \mu)$$



- Binomial Distribution

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

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\text{Bin}(m|N, \mu) = N\mu$$

$$\text{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \text{Bin}(m|N, \mu) = N\mu(1-\mu)$$

[CMB]

An Aside: Relation between Categorical and Multinomial Distribution

$$
\begin{aligned}
\mathrm{Mult}(m_1, m_2, \ldots, m_K | \boldsymbol{\mu}, N) &= \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k} \\
\mathbb{E}[m_k] &= N\mu_k \\
\mathrm{var}[m_k] &= N\mu_k(1-\mu_k) \\
\mathrm{cov}[m_j m_k] &= -N\mu_j\mu_k
\end{aligned}
$$

[CMB]

Bayesian inference of Gaussian (contd.) – optional

Bayesian Inference for the Gaussian (6)

- Now assume μ is known. The likelihood function for $\lambda = 1/\sigma^2$ is given by

$$p(\mathbf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

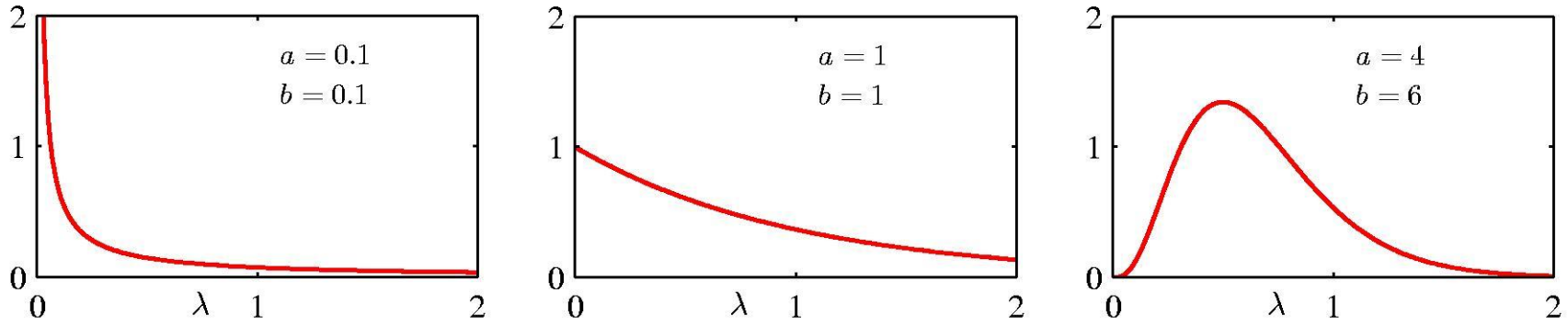
- This has a Gamma shape as a function of λ.

Bayesian Inference for the Gaussian (7)

- The Gamma distribution

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

$$\mathbb{E}[\lambda] = \frac{a}{b} \qquad \mathrm{var}[\lambda] = \frac{a}{b^2}$$



Bayesian Inference for the Gaussian (8)

- Now we combine a Gamma prior, $\mathrm{Gam}(\lambda|a_0, b_0)$ with the likelihood function for λ to obtain

$$p(\lambda|\mathbf{x}) \propto \lambda^{a_0-1}\lambda^{N/2}\exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

- which we recognize as $\mathrm{Gam}(\lambda|a_N, b_N)$ with

$$\begin{aligned} a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma^2_{\mathrm{ML}}. \end{aligned}$$

Bayesian Inference for the Gaussian (9)

- If both μ and λ are unknown, the joint likelihood function is given by

$$p(\mathbf{x}|\mu,\lambda) = \prod_{n=1}^{N}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left\{-\frac{\lambda}{2}(x_n-\mu)^2\right\}$$

$$\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}.$$

- We need a prior with the same functional dependence on λ and σ.

Bayesian Inference for the Gaussian (10)

- The Gaussian-gamma distribution

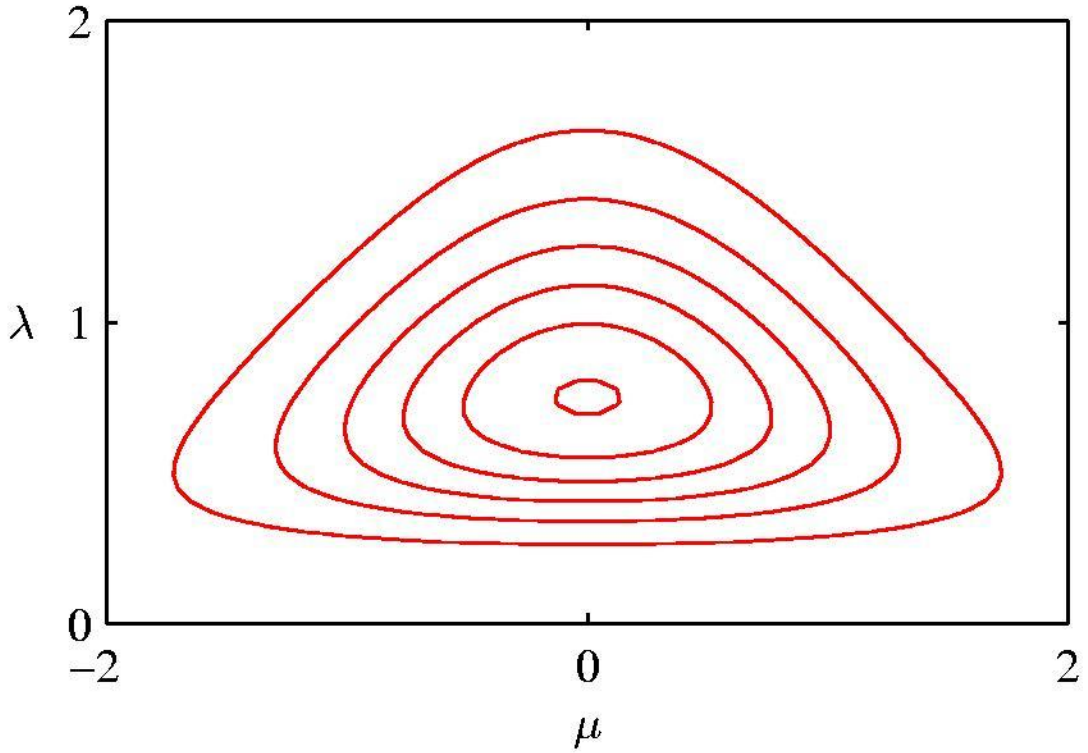
$$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a, b)$$

$$\propto \quad \exp\left\{-\frac{\beta\lambda}{2}(\mu - \mu_0)^2\right\} \lambda^{a-1} \exp\{-b\lambda\}$$

- Quadratic in μ.
- Linear in λ.
- Gamma distribution over λ.
- Independent of μ.

Bayesian Inference for the Gaussian (11)

- The Gaussian-gamma distribution



Bayesian Inference for the Gaussian (12)

- Multivariate conjugate priors
- $\boldsymbol{\mu}$ unknown, $\boldsymbol{\Lambda}$ known: p($\boldsymbol{\mu}$) Gaussian.
- $\boldsymbol{\Lambda}$ unknown, $\boldsymbol{\mu}$ known: p($\boldsymbol{\Lambda}$) Wishart,

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu) = B|\boldsymbol{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right).$$

- $\boldsymbol{\Lambda}$ and $\boldsymbol{\mu}$ unknown: p($\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$) Gaussian-Wishart,

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{\mu}_0,\beta,\mathbf{W},\nu) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0,(\beta\boldsymbol{\Lambda})^{-1})\,\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu)$$