Introduction to Advanced Probability for Graphical Models

CSC 412

Elliot Creager

Thursday January 11, 2018

*Many slides based on Kaustav Kundu's, Kevin Swersky's, Inmar Givoni's, Danny Tarlow's, Jasper Snoek's slides, Sam Roweis 's review of probability, Bishop's book, and some images from Wikipedia

Outline

- Basics
- Probability rules
- Exponential families
- Maximum likelihood
- Bayesian inference
  - Conjugate priors (time permitting)

Why Represent Uncertainty?

- The world is full of uncertainty
  - “What will the weather be like today?”
  - “Will I like this movie?”
  - “Is there a person in this image?”
- We’re trying to build systems that understand and (possibly) interact with the real world
- We often can’t *prove* something is true, but we can still ask how likely different outcomes are or ask for the most likely explanationx

Why Use Probability to Represent Uncertainty?

- Write down simple, reasonable criteria that you'd want from a system of uncertainty (common sense stuff), and you always get probability.
- Cox Axioms (Cox 1946); See Bishop, Section 1.2.3
- We will restrict ourselves to a relatively informal discussion of probability theory.

Notation

- A **random variable X** represents outcomes or states of the world.
- We will write p(x) to mean Probability(X = x)
- **Sample space**: the space of all possible outcomes (may be discrete, continuous, or mixed)
- p(x) is the **probability mass (density) function**
  - Assigns a number to each point in sample space
  - Non-negative, sums (integrates) to 1
  - Intuitively: how often does x occur, how much do we believe in x.

Joint Probability Distribution

- Prob(X=x, Y=y)
  - “Probability of X=x and Y=y”
  - p(x, y)

Conditional Probability Distribution

- Prob(X=x|Y=y)
  - “Probability of X=x given Y=y”

The Rules of Probability

- Sum Rule (marginalization/summing out):

$$p(x) = \sum_{y} p(x, y)$$

$$p(x_1) = \sum_{x_2} \sum_{x_3} \dots \sum_{x_N} p(x_1, x_2, \dots, x_N)$$

- Product/Chain Rule:

$$p(x, y) = p(y \mid x)\, p(x)$$

$$p(x_1, \dots, x_N) = p(x_1)\, p(x_2 \mid x_1) \dots p(x_N \mid x_1, \dots, x_{N-1})$$

Bayes' Rule

- One of the most important formulas in probability theory

$$p(x \mid y) = \frac{p(x, y)}{p(y)} = \frac{p(x, y)}{\sum_{x'} p(y \mid x')\, p(x')}$$

- This gives us a way of "reversing" conditional probabilities

Independence

- Two random variables are said to be **independent** iff their joint distribution factors

$$p(x, y) = p(x)\,p(y)$$

- Two random variables are **conditionally independent** given a third if they are independent after conditioning on the third

$$p(x, y \mid z) = p(y \mid x, z)\,p(x \mid z) = p(y \mid z)\,p(x \mid z) \quad \forall z$$

- i.i.d. observations are an example of conditional independence

$$p(x_1 \ldots x_n \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$$

Continuous Random Variables

- Outcomes are real values. Probability density functions define distributions.
  - E.g.,

$$P(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}$$

- Continuous joint distributions: replace sums with integrals, and everything holds
  - E.g., Marginalization and conditional probability

$$P(x,z) = \int_y P(x,y,z) = \int_y P(x,z \mid y)P(y)$$

Summarizing Probability Distributions

- It is often useful to give summaries of distributions without defining the whole distribution (E.g., mean and variance)
- Mean: $$E[x] = \langle x \rangle = \int_x x \cdot p(x)\, dx$$
- Variance: $$\mathrm{var}(x) = E[(x - E[x])^2]$$

$$= \int_x (x - E[x])^2 \cdot p(x)\, dx = E[x^2] - E[x]^2$$

Exponential Families

- Families of distributions with nice properties
  - *e.g.* They have a conjugate prior (we'll get to that later. Important for Bayesian statistics)
- Includes many standard distributions
  - Bernoulli, binomial/multinomial, Poisson, Normal (Gaussian), beta/Dirichlet,…
- But not all
  - Uniform, Cauchy

Definition

- All exponential families of distributions over x, given parameter η (eta) take the form

$$p(x|\eta) = h(x)g(\eta)\exp\{\eta^T u(x)\}$$

- To get a particular family of dist'ns, we specify
  - x-scalar/vector, discrete/continuous
  - η – 'natural parameters'
  - u(x) – some function of x (natural statistic)
  - h(x) – base measure (often constant)
  - g(η) – normalizer

Why "exponential"? Note:

$$p(x|\eta) = \exp\{\eta^T u(x) + \ln h(x) + \ln g(\eta)\}$$

$$1 = \int g(\eta)h(x)\exp\{\eta^T u(x)\}dx = g(\eta)\int h(x)\exp\{\eta^T u(x)\}dx$$

Sufficient Statistics

- Vague definition: called so because they completely summarize a distribution.
- Less vague: they are the only part of the distribution that interacts with the parameters and are therefore sufficient to estimate the parameters.
- $\mu(x)$, the natural statistic in an exponential family, is also *sufficient*

Example 1: Bernoulli

- Binary random variable - $X \in \{0,1\}$
- p(heads) = μ $\mu \in [0,1]$
- Coin toss

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x}$$

Example 1: Bernoulli

$$p(x \mid \eta) = h(x) g(\eta) \exp\{\eta^T u(x)\}$$

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x}$$

$$= \exp\{\ln\{\mu^x (1-\mu)^{1-x}\}\}$$

$$= \exp\{x \ln \mu + (1-x)\ln(1-\mu)\}$$

$$= (1-\mu)\exp\{\ln\left(\frac{\mu}{1-u}\right)x\}$$

$$p(x \mid \eta) = \sigma(-\eta)\exp(\eta x)$$

$$h(x) = 1$$

$$u(x) = x$$

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right) \Rightarrow \mu = \sigma(\eta) = \frac{1}{1+e^{-\eta}}$$

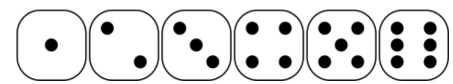
$$g(\eta) = \sigma(-\eta)$$

Example 2: Multinomial

- p(value k) = μ_k

$$\mu_k \in [0,1], \sum_{k=1}^{M} \mu_k = 1$$

- For a single observation – e.g., die toss
  - Sometimes called Categorical

$x_i \in$ 

- For multiple observations
  - integer counts on N trials

$$\sum_{k=1}^{M} x_k = N$$

  - Prob(1 came out 3 times, 2 came out once,…,6 came out 7 times if I tossed a die 20 times)

= Prob(x_1=3, x_2=1,…,x_6=7) with N=20

$$P(x_1,\ldots,x_M \mid \mu) = \frac{N!}{\prod_k x_k!} \prod_{k=1}^{M} \mu_k^{x_k}$$

Example 2: Multinomial (1 observation)

$$p(x \mid \eta) = h(x) g(\eta) \exp\{\eta^T u(x)\}$$

$$P(x_1, \ldots, x_M \mid \mu) = \prod_{k=1}^{M} \mu_k^{x_k}$$

$$= \exp\{\sum_{k=1}^{M} x_k \ln \mu_k\}$$

$$p(\mathbf{x} \mid \eta) = \exp(\eta^T \mathbf{x})$$

$$h(\mathbf{x}) = 1$$

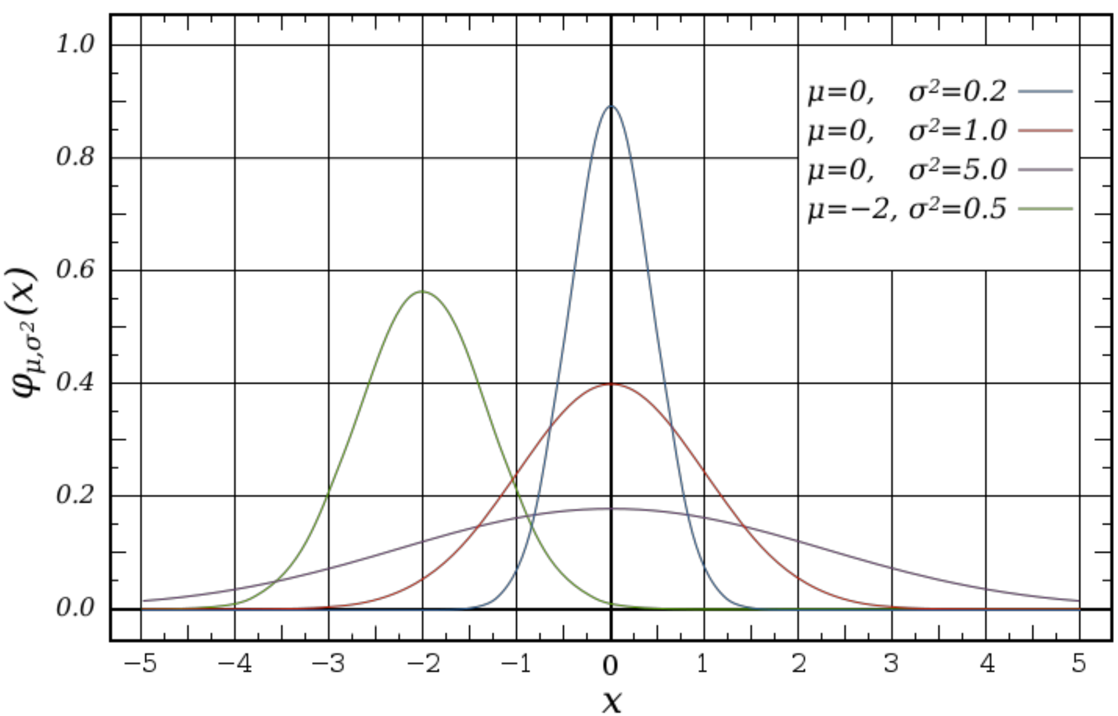
$$u(\mathbf{x}) = \mathbf{x}$$

Parameters are not independent due to constraint of summing to 1, there's a slightly more involved notation to address that, see Bishop 2.4

Example 3: Univariate Gaussian

- Gaussian (Normal)

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}$$



Example 3: Univariate Gaussian

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}$$

- μ is the mean
- σ^2 is the variance
- Can verify these by computing integrals. E.g.,

$$\int_{x \to -\infty}^{x \to \infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} dx = \mu$$

Univariate Gaussian as Exp. Family

$$p(x \mid \eta) = h(x) g(\eta) \exp\{\eta^T u(x)\}$$

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}$$

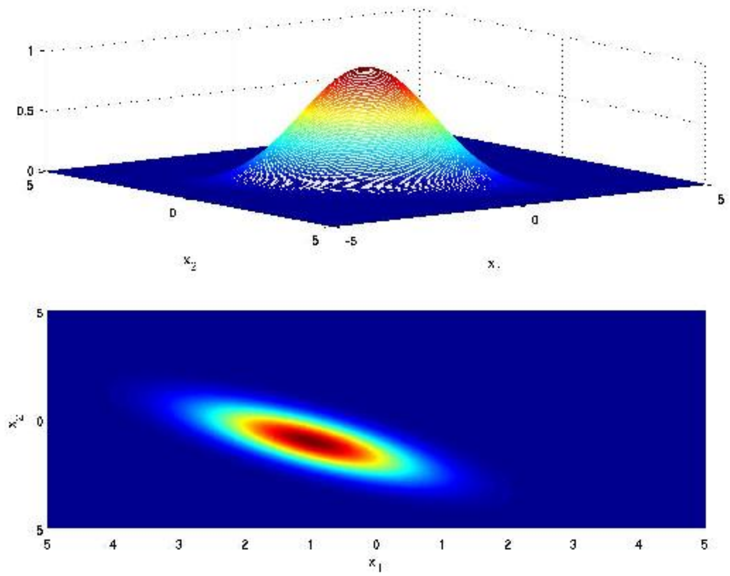
$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\{\frac{-1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x + \frac{-1}{2\sigma^2} \mu^2\} =$$

$$= \underbrace{(2\pi)^{-\frac{1}{2}}}_{h(x)} \underbrace{(-2\eta_2)^{\frac{1}{2}} \exp(\frac{\eta_1^2}{4\eta_2})}_{g(\eta)} \exp\{\underbrace{\begin{bmatrix} \frac{\mu}{\sigma^2} & \frac{-1}{2\sigma^2} \end{bmatrix}}_{\eta^T} \underbrace{\begin{bmatrix} x \\ x^2 \end{bmatrix}}_{u(x)}\}$$

Example 4: Multivariate Gaussian

- Multivariate Gaussian

$$P(x \mid \mu, \Sigma) = \left|2\pi \Sigma\right|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$



Example 4: Multivariate Gaussian

- Multivariate Gaussian

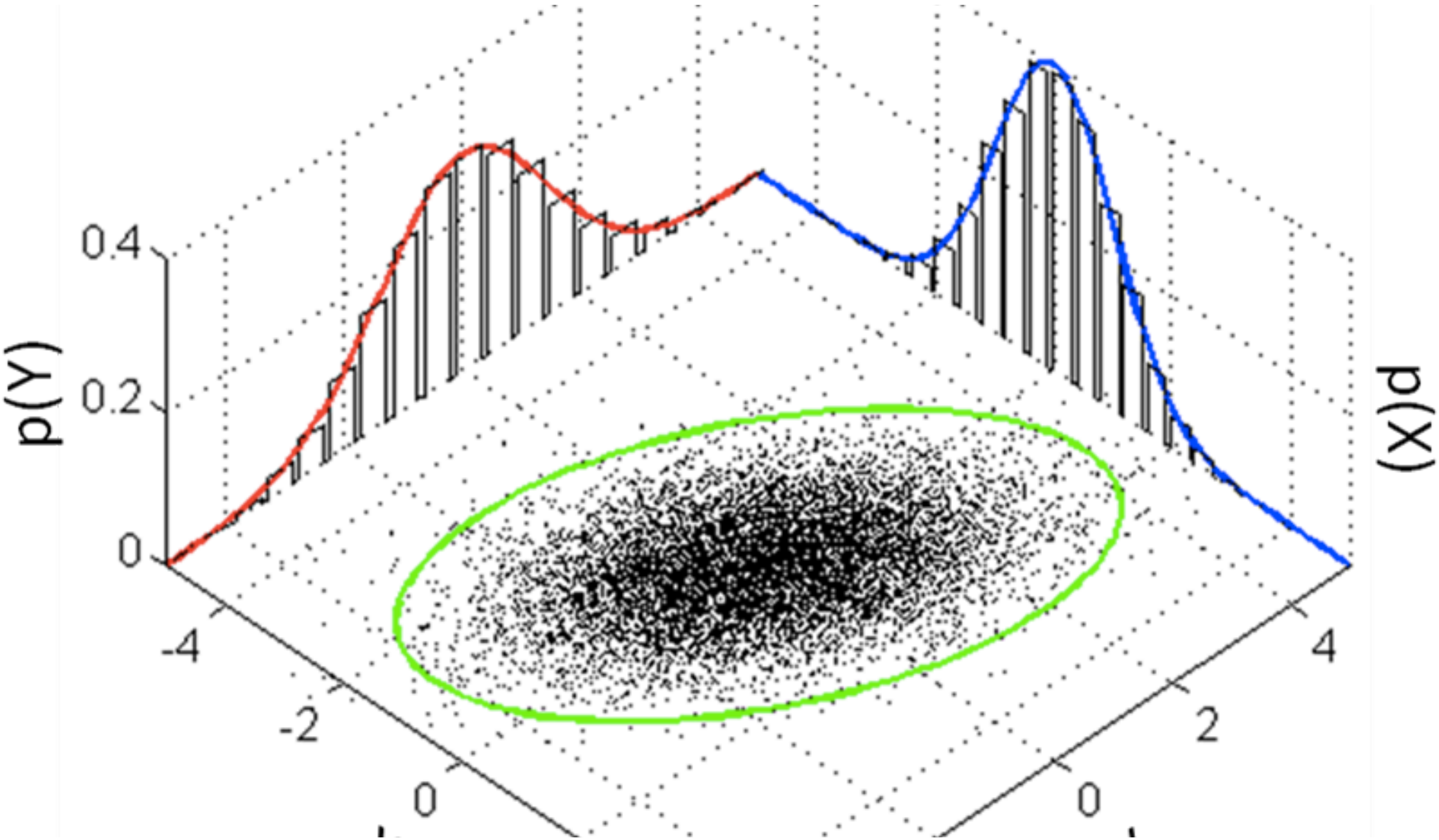
$$p(x \mid \mu, \Sigma) = \left|2\pi \Sigma\right|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

- x is now a vector
- μ is the **mean vector**
- Σ is the **covariance matrix**

Important Properties of Gaussians

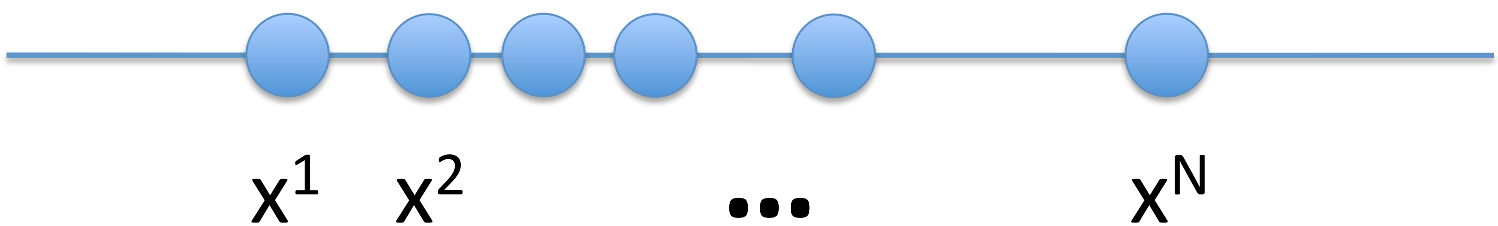
- All marginals of a Gaussian are again Gaussian
- Any conditional of a Gaussian is Gaussian
- The product of two Gaussians is again Gaussian
- Even the sum of two independent Gaussian RVs is a Gaussian.
- Beyond the scope of this tutorial, but **very** important: marginalization and conditioning rules for multivariate Gaussians.

Gaussian marginalization visualization



Example 5: Maximum Likelihood For a 1D Gaussian

- Suppose we are given a data set of samples of a Gaussian random variable X, D={x^1,..., x^N} and told that the variance of the data is σ^2



What is our best guess of μ?

*Need to assume data is independent and identically distributed (i.i.d.)

Example 5: Maximum Likelihood For a 1D Gaussian

What is our best guess of μ?

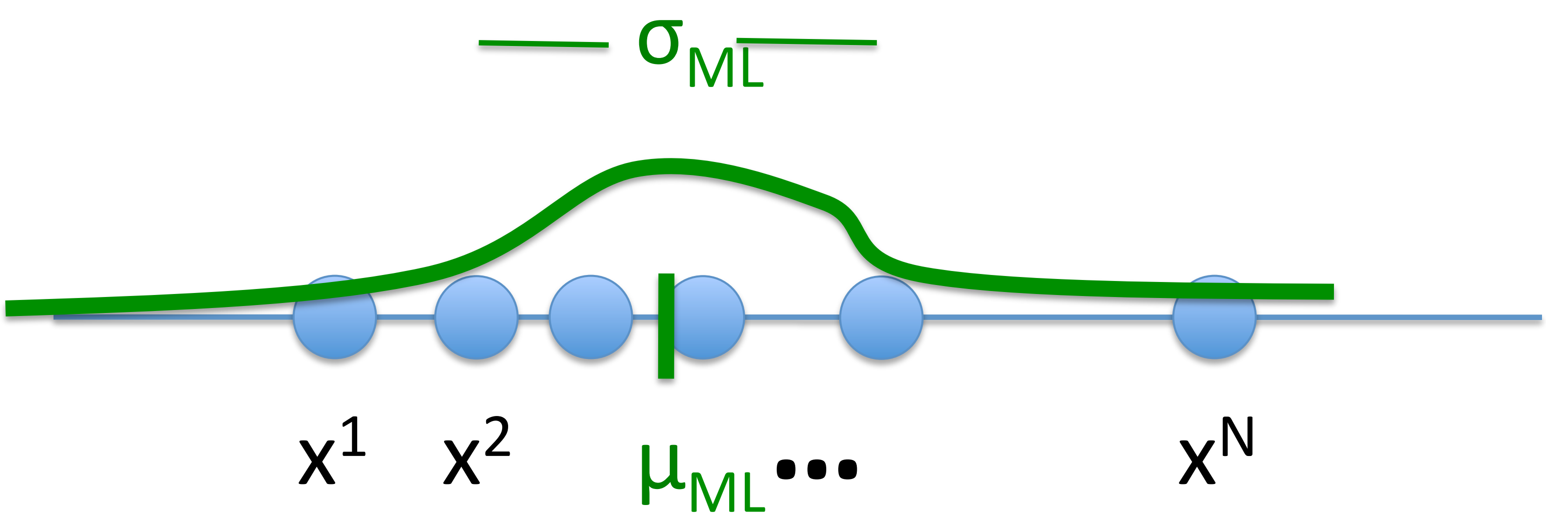
- We can write down the **likelihood function**:

$$p(d\,|\,\mu) = \prod_{i=1}^{N} p(x^i \,|\, \mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2\sigma^2}(x^i - \mu)^2\right\}$$

- We want to choose the μ that maximizes this expression
  - Take log, then basic calculus: differentiate w.r.t. μ, set derivative to 0, solve for μ to get **sample mean**

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Example 5: Maximum Likelihood For a 1D Gaussian



Maximum Likelihood

ML estimation of model parameters for Exponential Family

$$p(D \mid \eta) = p(x_1, \ldots, x_N) = \left(\prod h(x_n)\right) g(\eta)^N \exp\{\eta^T \sum_n u(x_n)\}$$

$$\partial \frac{\ln(p(D \mid \eta))}{\partial \eta} = \ldots, \text{set to } 0, \text{ solve for } \nabla \mathrm{g}(\eta)$$

$$-\nabla \ln g(\eta_{ML}) = \frac{1}{N} \sum_{n=1}^{N} u(x_n)$$

- Can in principle be solved to get estimate for eta.
- The solution for the ML estimator depends on the data only through sum over u, which is therefore called **sufficient statistic**
- What we need to store in order to estimate parameters.

Bayesian Probabilities

$$p(\theta \mid d) = \frac{p(d \mid \theta)\,p(\theta)}{p(d)}$$

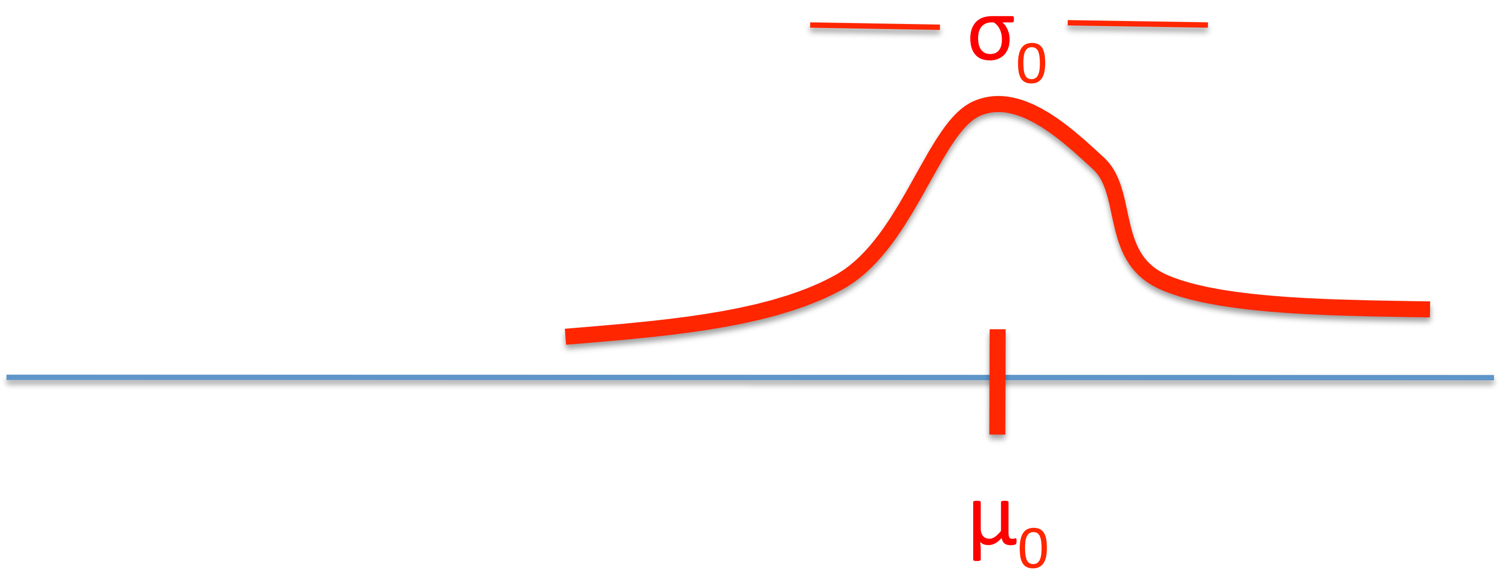
- $p(d \mid \theta)$ is the **likelihood function**
- $p(\theta)$ is the **prior probability** of (or our **prior belief** over) θ
  - our beliefs over what models are likely or not *before seeing any data*
- $p(d) = \int p(d \mid \theta) P(\theta) d\theta$ is the **normalization constant** or **partition function**
- $p(\theta \mid d)$ is the **posterior distribution**
  - **Readjustment of our prior beliefs in the face of data**

Example 6: Bayesian Inference For a 1D Gaussian

- Suppose we have a prior belief that the mean of some random variable X is μ_0 and the variance of our belief is σ_0^2
- We are then given a data set of samples of X, $d=\{x^1,\ldots, x^N\}$ and somehow know that the variance of the data is σ^2

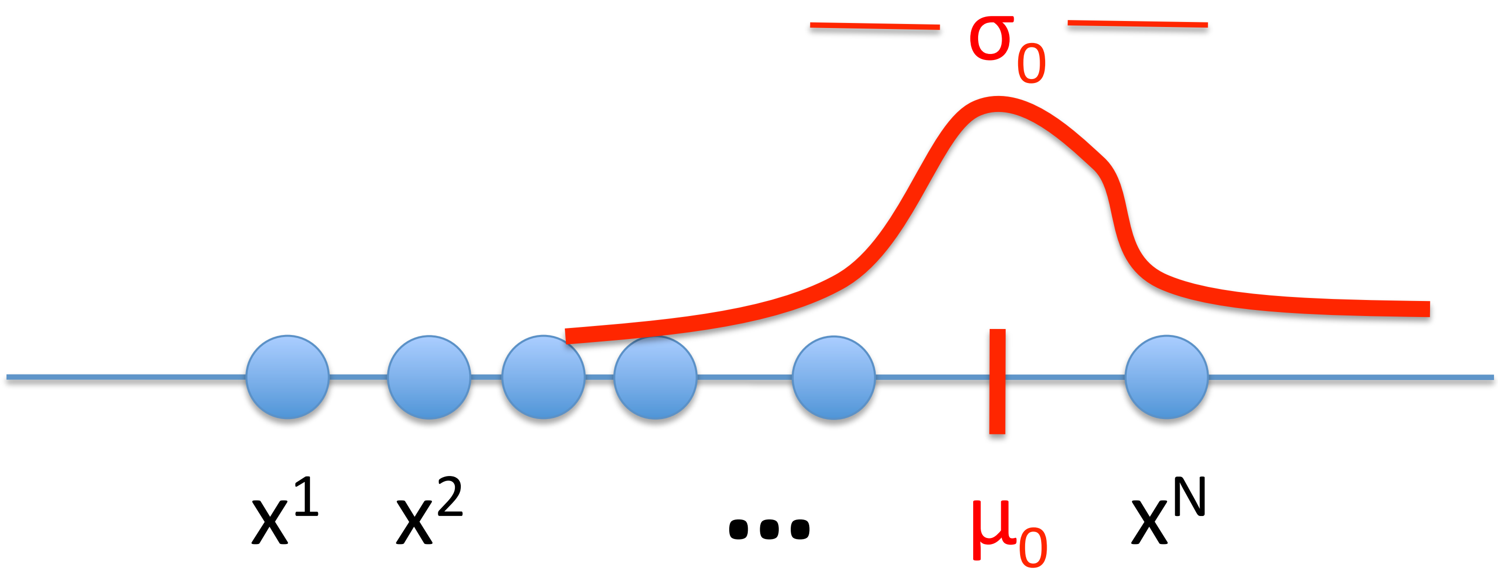
What is the posterior distribution over (our belief about the value of) μ?

Example 6: Bayesian Inference For a 1D Gaussian



Prior belief

Example 6: Bayesian Inference For a 1D Gaussian



Prior belief

Example 6: Bayesian Inference For a 1D Gaussian

- Remember from earlier $p(\mu \mid d) = \dfrac{p(d \mid \mu)\, p(\mu)}{p(d)}$

- $p(d \mid \mu)$ is the **likelihood function**

$$p(d \mid \mu) = \prod_{i=1}^{N} P(x^i \mid \mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x^i - \mu)^2 \right\}$$

- $p(\mu)$ is the **prior probability** of (or our **prior belief** over) μ

$$p(\mu \mid \mu_0, \sigma_0) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{ -\frac{1}{2{\sigma_0}^2}(\mu - \mu_0)^2 \right\}$$

Example 6: Bayesian Inference For a 1D Gaussian

$$p(\mu \mid D) \propto p(D \mid \mu)\, p(\mu)$$

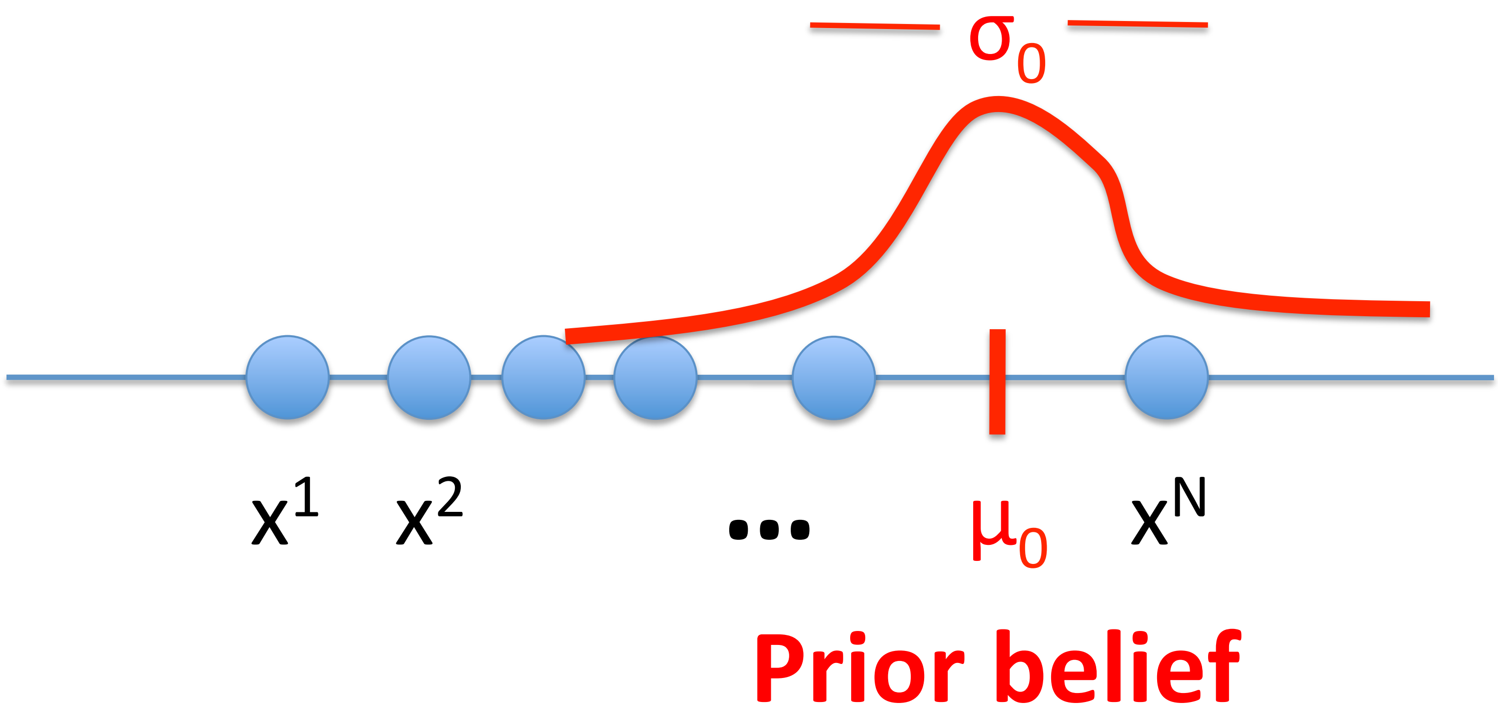
$$p(\mu \mid D) = \mathbf{Normal}(\mu \mid \mu_N, \sigma_N)$$

where

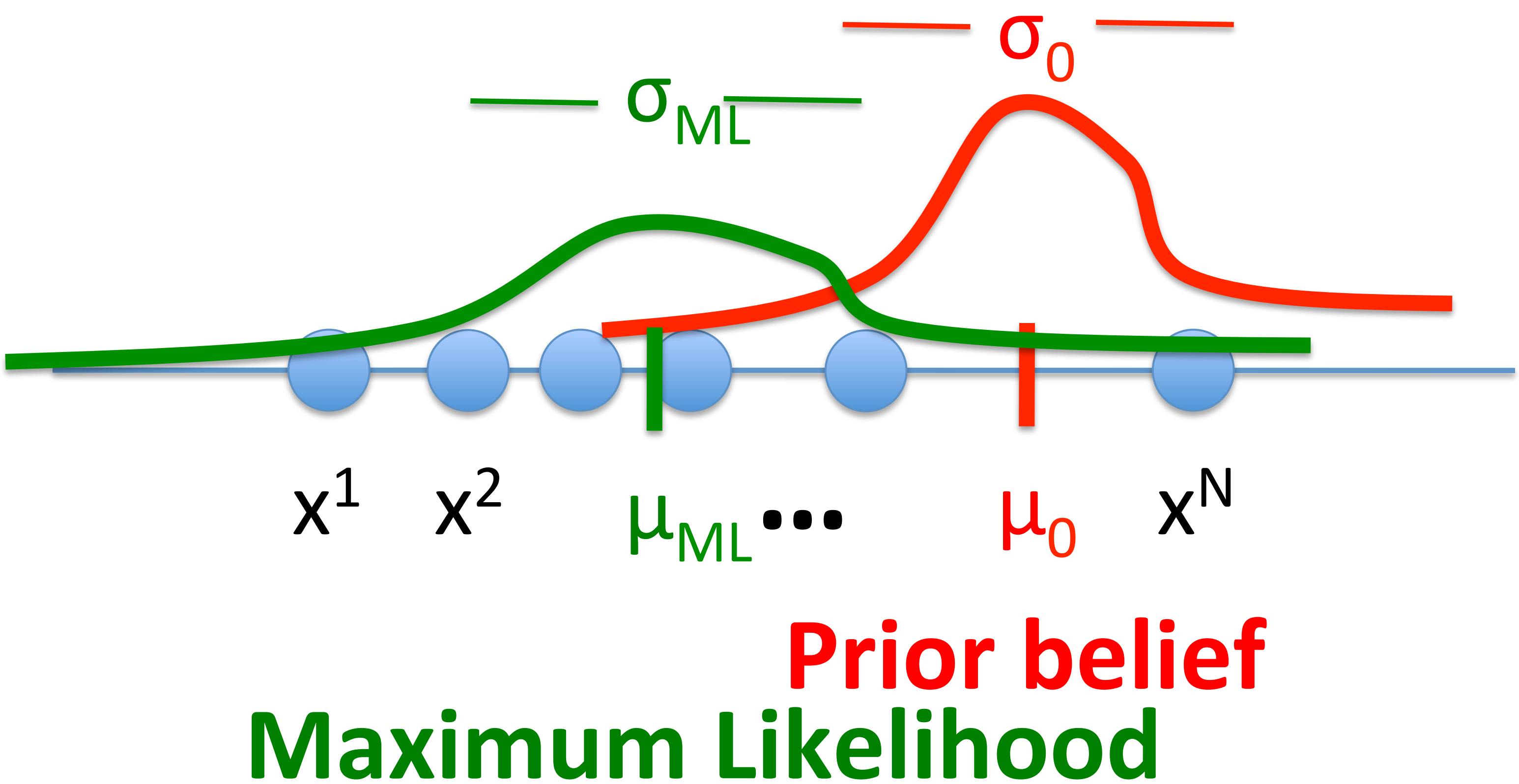
$$\mu_N = \frac{\sigma^2}{N{\sigma_0}^2 + \sigma^2}\mu_0 + \frac{N{\sigma_0}^2}{N{\sigma_0}^2 + \sigma^2}\mu_{ML}$$

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

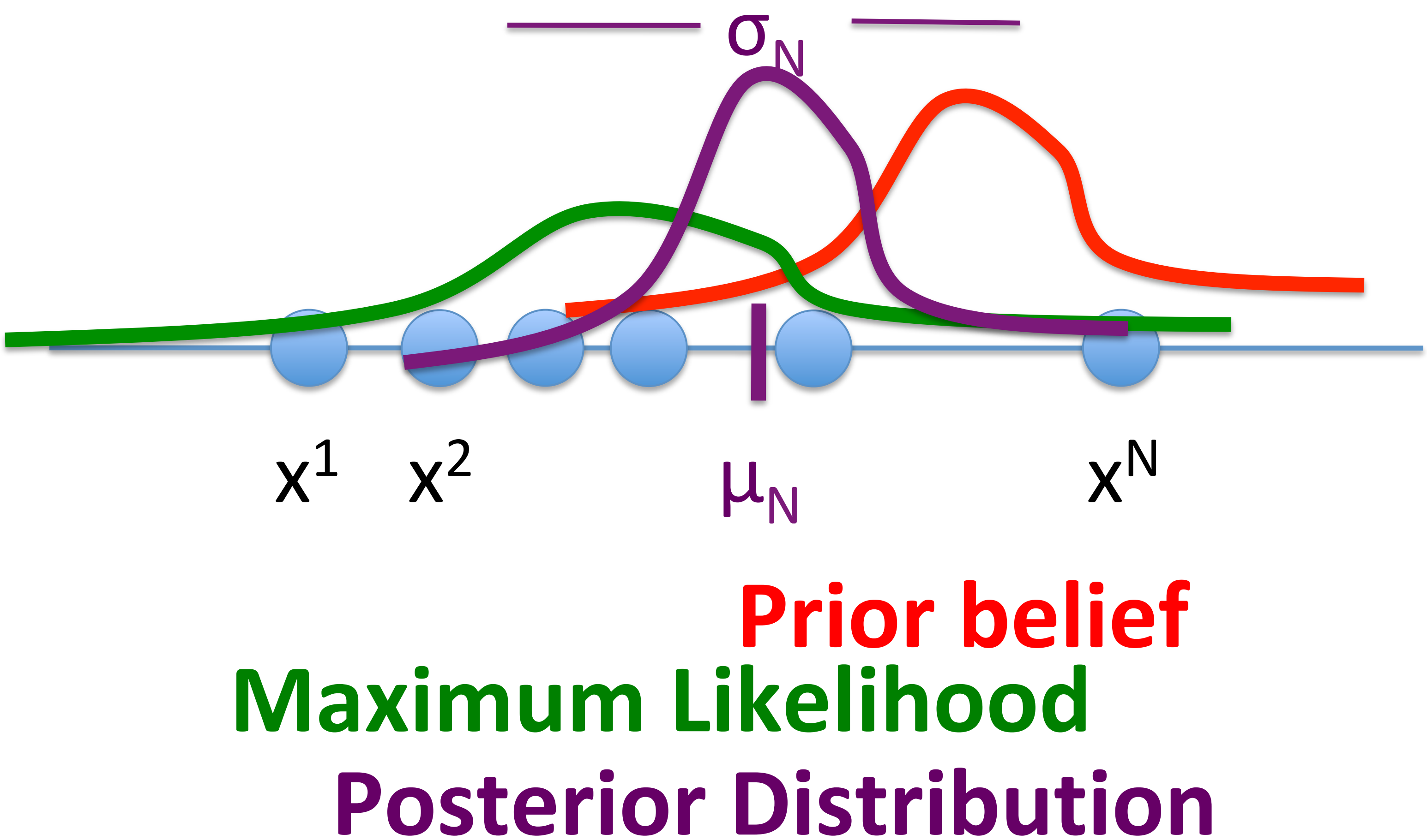
Example 6: Bayesian Inference For a 1D Gaussian



Example 6: Bayesian Inference For a 1D Gaussian



Example 6: Bayesian Inference For a 1D Gaussian



Conjugate Priors

- In the previous example, the prior, likelihood and posterior were all Gaussian. The Gaussian dist'n is *self-conjugate*.
- More generally, *conjugate priors* induce a *posterior* of the same form (the likelihood need not take this form).
  - E.g., Gamma is the conjugate prior of a Poisson likelihood
- For any member of the exponential family there exists a conjugate prior that can be written like

$$p(\eta \mid \chi, \nu) = f(\chi, \nu) g(\eta)^{\nu} \exp\{\nu \eta^T \chi\}$$

- Multiply by likelihood to obtain posterior (up to normalization) of the form

$$p(\eta \mid D, \chi, \nu) \propto g(\eta)^{\nu + N} \exp\{\eta^T (\nu \chi + \sum_{n=1}^{N} u(x_n))\}$$

- Notice the addition to the sufficient statistic
- ν is the effective number of pseudo-observations.

Conjugate Priors - Examples

- Beta for Bernoulli/binomial
- Dirichlet for categorical/multinomial
- Normal for mean of Normal
- And many more...