Cmput466/551

Probability 101b (con't)

Covering…
HTF Ch2 (kinda)…
\+ Review of Probability Theory
\+ …

R Greiner
Department of Computing Science
University of Alberta

Thanks to R Parr, C Guesterin

Outline

- Foundations
  - Bayes Theorem
  - (Conditional) Independence
  - Dutch Book Theorem
  - Moments: Mean, Variance
- Estimation
  - MLE (Binomial)
  - Bayesian model
- Gaussian (Normal)

Learning involves Estimation

- Consider flipping a Thumbtack.
  What is the probability it will land with the nail up?

- Try flipping it a few times…
  observe H,H,T,T,H

- What is your BEST GUESS?

Jump

Simple "Learning" Algorithm

$$\hat{\theta} = \arg\max_{\theta} \ \ln P(D \mid \theta)$$
$$= \arg\max_{\theta} \ \ln \theta^{h} (1 - \theta)^{t}$$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0}$

$$\frac{\partial}{\partial\theta} \ln[\theta^{h}(1-\theta)^{t}] = \frac{\partial}{\partial\theta}[h \ln\theta + t \ln(1-\theta)] = \frac{h}{\theta} + \frac{-t}{(1-\theta)}$$

$$\frac{h}{\theta} + \frac{-t}{1-\theta} = 0 \quad \Rightarrow \quad \hat{\theta} = \frac{\text{h}}{\text{t+h}}$$

So just average!!!

Jump

How many flips are "needed"?

$$\boxed{\hat{\theta}_{MLE} = \frac{\#H}{\#H + \#T}}$$

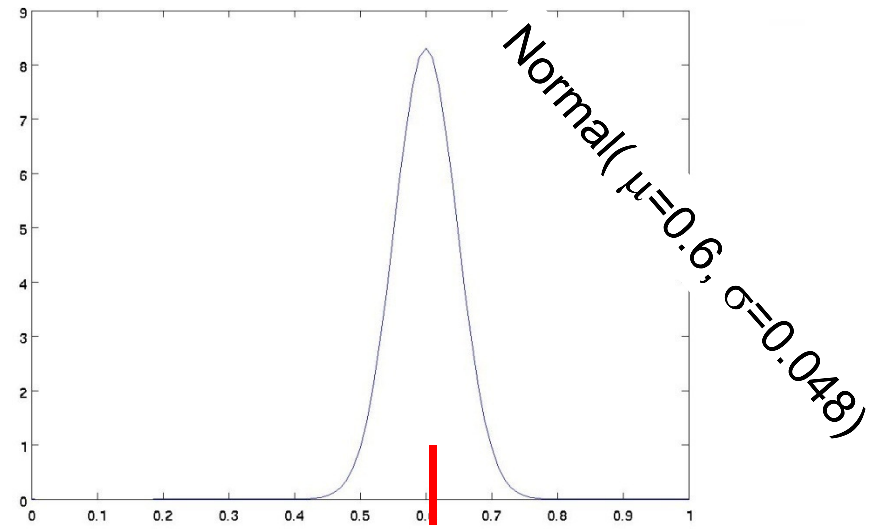
- Given 3 heads and 2 tails, $\theta_{MLE} = \frac{3}{5} = 0.6$
- But…
  Given 30 heads and 20 tails, $\theta_{MLE} = \frac{30}{50} = 0.6$
- SAME!!!
  Which is better? … more precise?

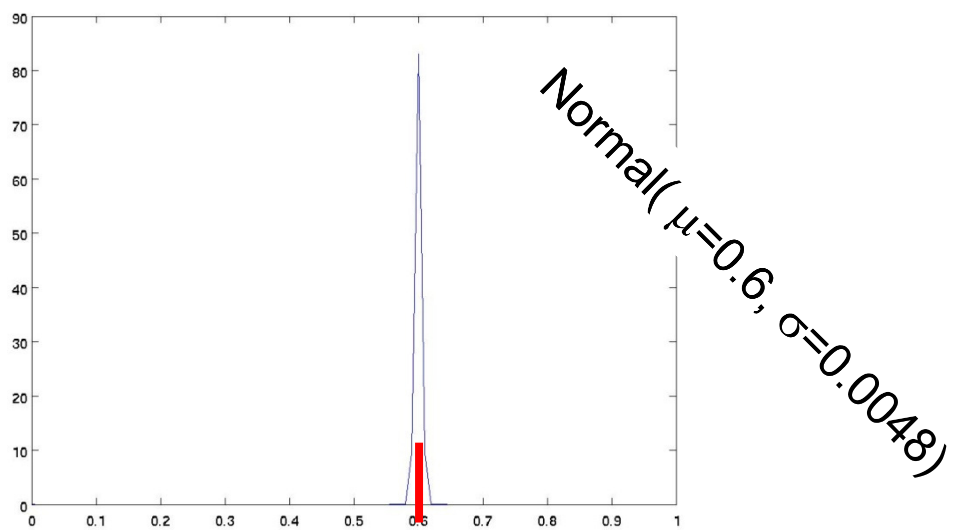
Jump

Using Variance

- Variance measures "spread" around mean
- For Binomial(h, t)
  - Mean: $\mu = \frac{h}{h+t}$
  - Variance: $\sigma = \frac{\mu(1-\mu)}{h+t}$

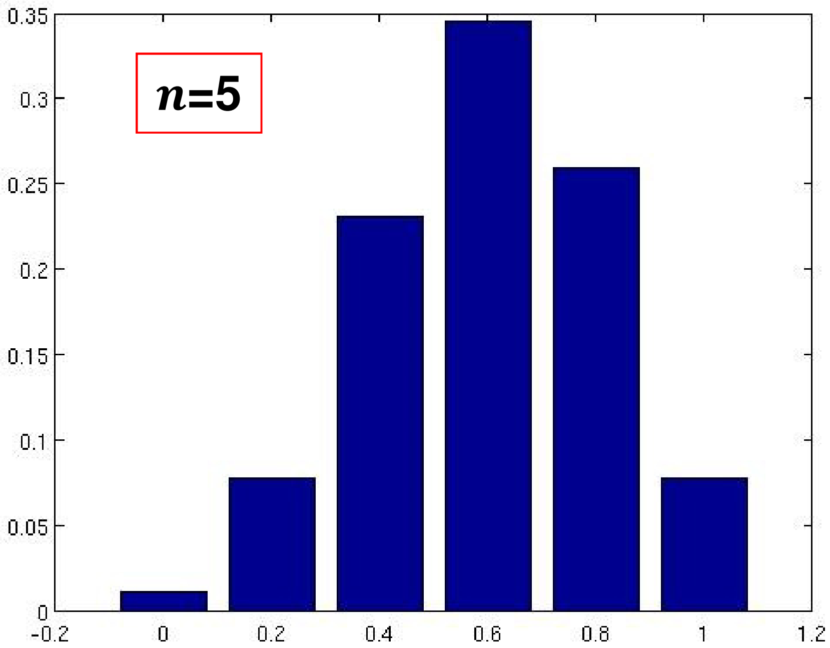
- Binomial(**3**H, **2**T)
  μ=0.6 σ=0.048
- Binomial(**30**H, **20**T)
  μ=0.6 σ=0.0048

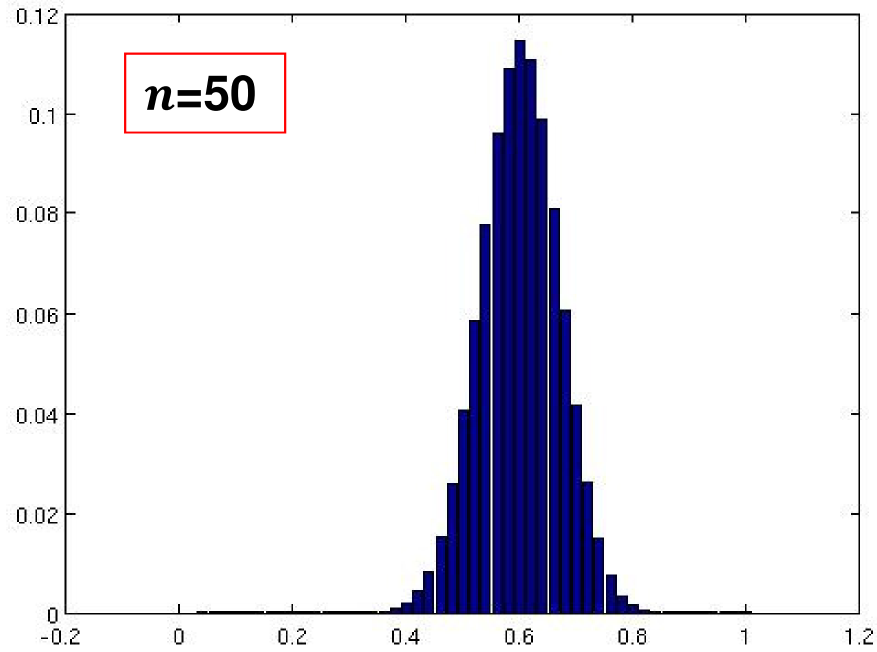




Binomial Distribution

P(D | θ)   for fixed θ=0.6

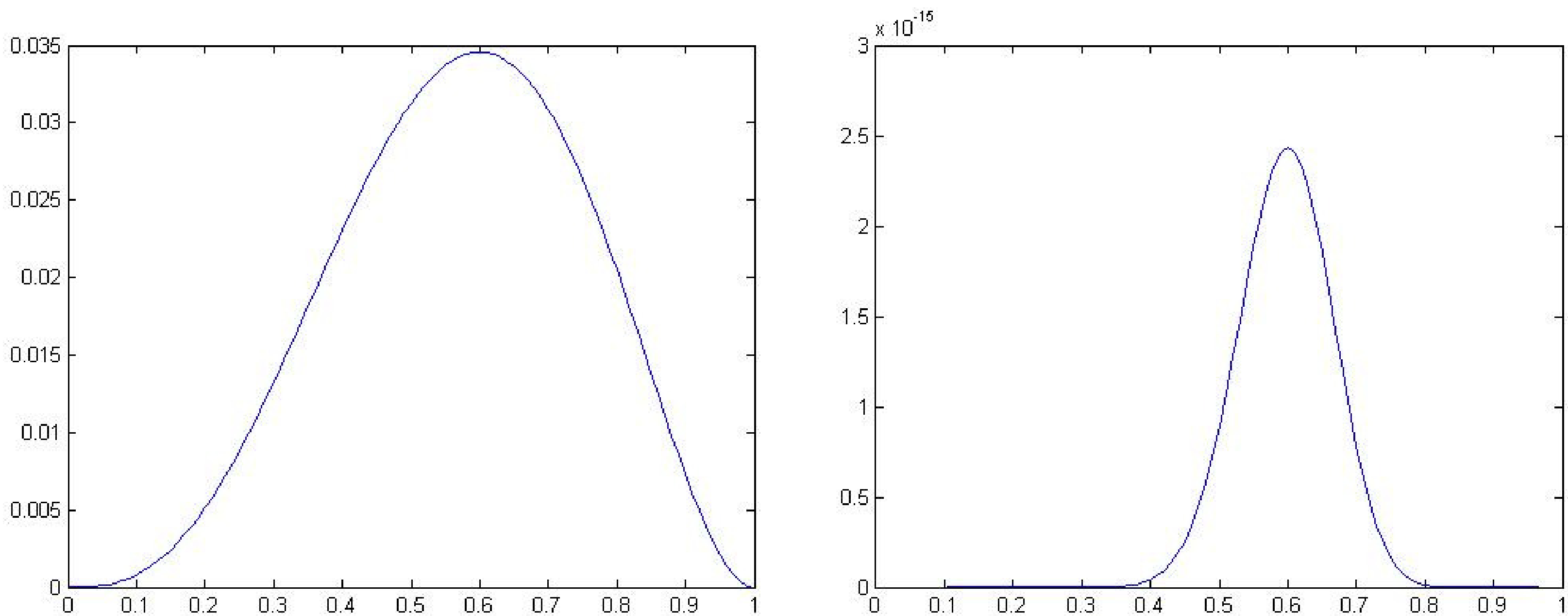




Prob that p=0.6 coin generates $\frac{k}{n}$ heads, in n flips

Probability Functions

P(D | θ) for fixed D



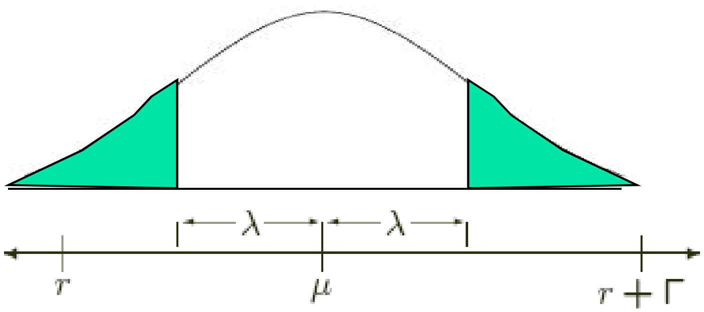
Prob that p=θ coin generates $\frac{h}{N}$ heads

($1-\frac{h}{N}$ tails)

Hoeffding's Inequality

Defn: $S_m = \frac{1}{m}\sum_{i=1}^{m} X_i$ observed average over m r.v.s in {0,1}

- $P(\, S_m > \mu + \lambda \,) \; < \; e^{-2\, m\, \lambda^2}$



$\Pr[\ |S_m - \mu\ | < \lambda\ ] \geq 1 - 2e^{-2m\,(\lambda/\Gamma)^2}$

- Holds ∀ (bounded) distributions ... not just Bernoulli...
- Sample average likely to be close to true value as #samples (*m*) increases…

Simple bound (using Hoeffding's Inequality)

Here…

- #flips $m = m_H + m_T$
- Sample average = $\hat{\theta}^{(m)} = \frac{m_H}{m_H + m_T}$
- Let θ^* be the true parameter

For any m, ε>0:

$$\boxed{P(\,|\hat{\theta}^{(m)} - \theta^*| > \epsilon\,) \;<\; 2\,e^{-2\,m\,\epsilon^2}}$$

Using Hoeffding's Inequality

$$P(\,|\hat{\theta} - \theta^*| > \epsilon\,) < 2\,e^{-2\,m\,\epsilon^2}$$

- To estimate the thumbtack parameter θ,
  - within $\varepsilon = 0.1$,
  - with probability $\geq 1 - \delta = 0.95$

require #flips $m > \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$ ≈ 460.2

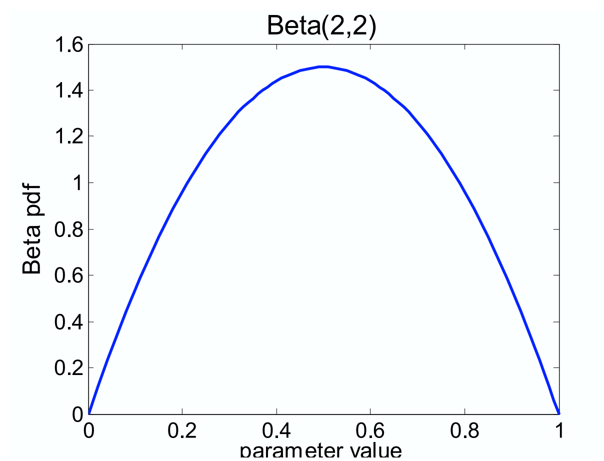
Problems with MLE

- Do you really believe 0% if 0 / 0+2 ?
- 0/0 issues
- Which is better?
  - 3 heads, 2 tails $\theta = \frac{3}{3+2} = 0.6$
  - 30 heads, 20 tails $\theta = \frac{30}{30+20} = 0.6$
  - 3E23 heads, 2E23 tails $\theta = \frac{3E23}{3E3+2E23} = 0.6$
- What if you already know SOMETHING about the variable…

≈ 50/50 …

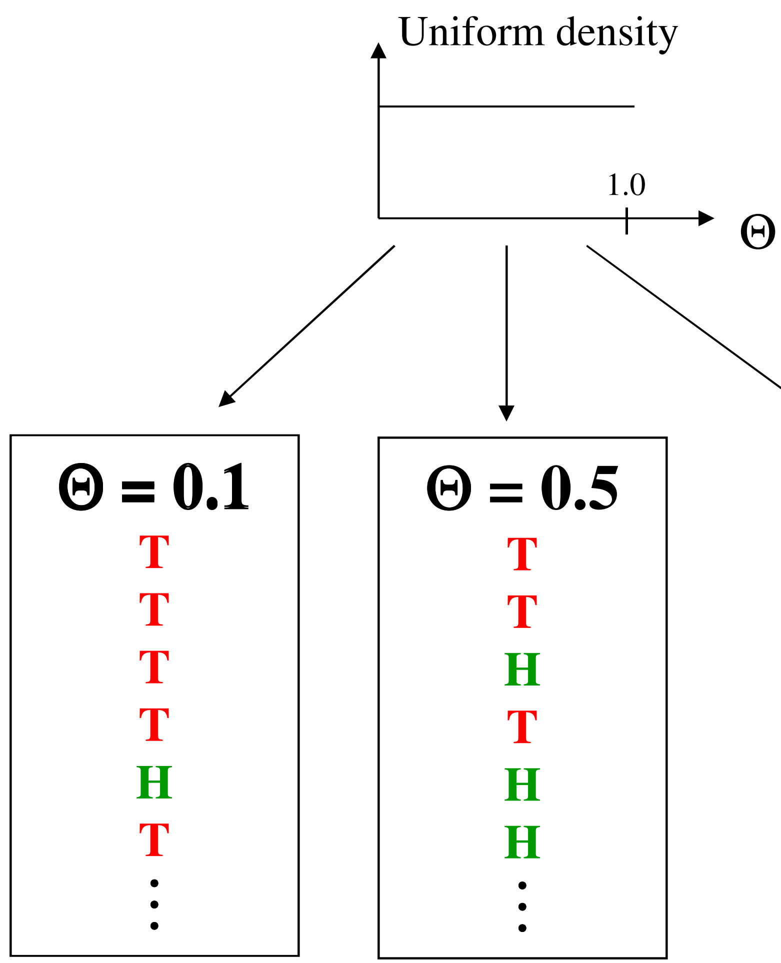
What about prior knowledge?

- Spse you *know* the thumbtack θ is "close" to 50-50
- **You can estimate it the Bayesian way…**
- Rather than estimate a single θ, obtain a *distrib'n* over possible values of θ

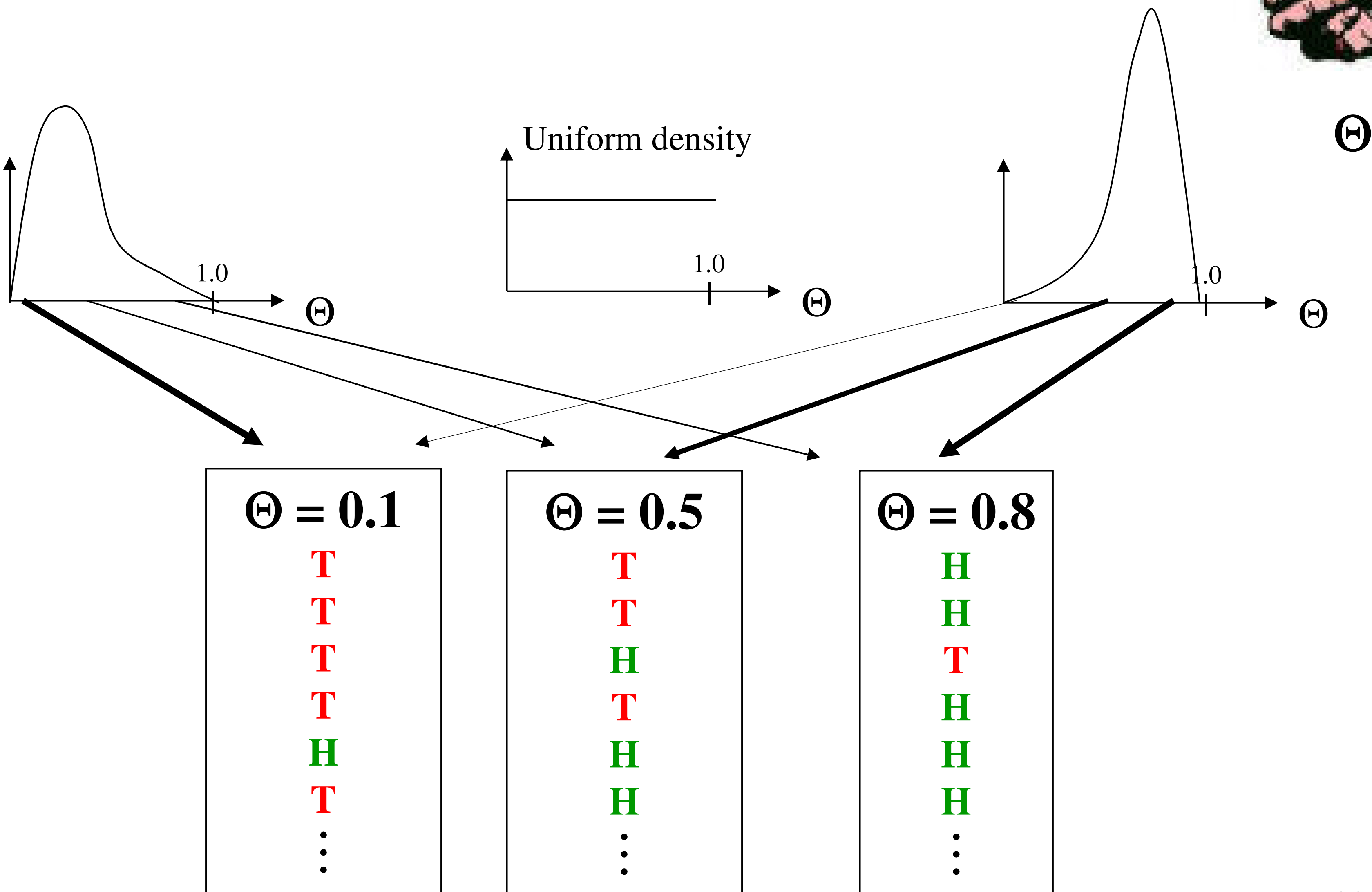


Two (related) Distributions: Parameter, Instances

Θ



Two (related) Distributions: Parameter, Instances



Jump

Bayesian Learning

- Use Bayes rule:

$$\underbrace{P(\theta \mid D)}_{\textit{posterior}} = \frac{\overbrace{P(D \mid \theta)}^{\textit{likelihood}}\,\overbrace{P(\theta)}^{\textit{prior}}}{P(D)}$$

- Or equivalently (wrt $\operatorname{argmax}_\theta P(\theta|D)$)

$$\boxed{P(\theta \mid D) \propto P(D \mid \theta)\, P(\theta)}$$

Bayesian Learning for Thumbtack

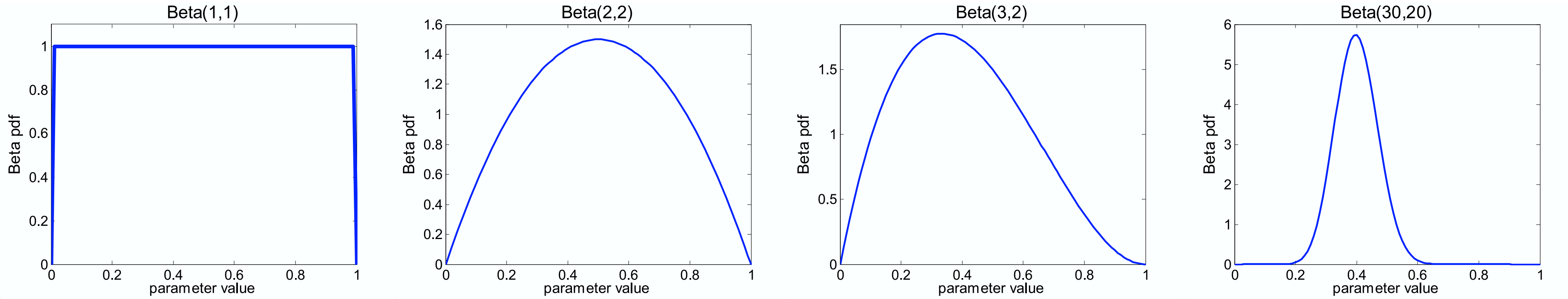
$$\underbrace{\mathrm{P}(\theta \mid D)}_{\textit{posterior}} \propto \underbrace{P(D \mid \theta)}_{\textit{likelihood}} \underbrace{P(\theta)}_{\textit{prior}}$$

- Likelihood function is simply Binomial:

$$P(\mathcal{D} \mid \theta) = \theta^{m_H}(1-\theta)^{m_T}$$

- What about prior, P(θ) ?
  - Represent expert knowledge
  - Simple posterior form

Beta prior distribution – P(θ)

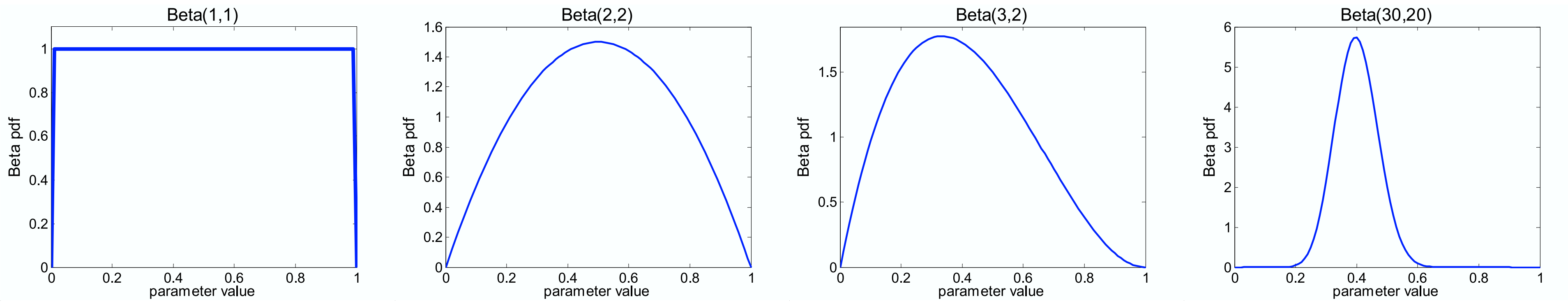


For $\theta \sim \mathrm{Beta}(a, b)$:

- PDF: $P(\theta) = \dfrac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$
- Mean: $E[\theta] = \dfrac{a}{a+b}$
- Variance: $\mathrm{Var}[\theta] = \dfrac{a\,b}{(a+b)^2(a+b+1)} = \dfrac{a}{a+b}\dfrac{b}{a+b}\dfrac{1}{a+b+1}$
- Unimodal if $a,b>1$
- Likelihood function:

$$P(h \text{ “+”s}, t \text{ “–”s} \mid \theta) = \theta^h (1-\theta)^t$$

Posterior distribution… from Beta



Prior $P(\theta)$

Likelihood $P(D|\theta)$

$$P(\theta \mid \mathcal{D}) \propto P(\theta)\ P(\mathcal{D} \mid \theta)$$

$$= \boxed{\Theta^{\alpha_H - 1}(1-\Theta)^{\alpha_T - 1}} \times \boxed{\Theta^{m_H}(1-\Theta)^{m_T}}$$

$$= \theta^{\alpha_H + m_H - 1}\ (1-\theta)^{\alpha_T + m_T - 1}$$

$$\sim \text{Beta}(\ \alpha_H + m_H,\ \alpha_T + m_T\ )$$

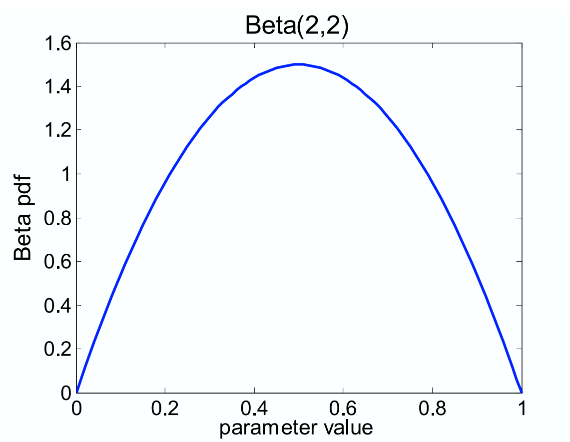
So Posterior is same form as Prior!! Conjugate!

Posterior Distribution

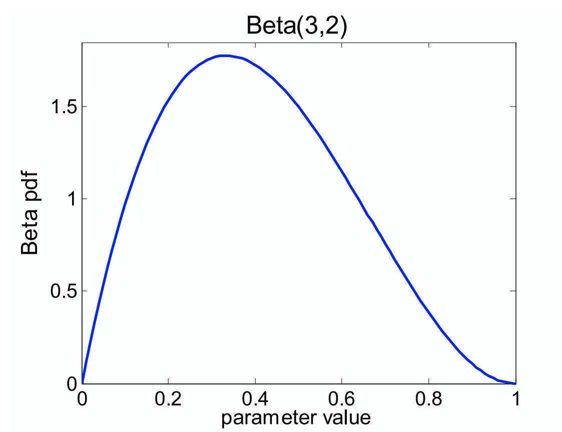
- Prior: $\theta \sim \text{Beta}(\alpha_H, \alpha_T)$
- Data $\mathcal{D}$: m_H heads, m_T tails

$\Rightarrow$ Posterior distribution:

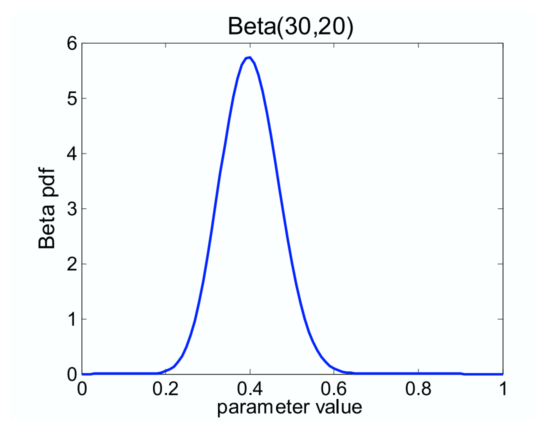
$$\theta \mid \mathcal{D} \sim \text{Beta}(\alpha_H + m_H, \alpha_T + m_T)$$



Prior



+ observe 1 head



+ observe
27 more heads;
18 tails

Conjugate Prior

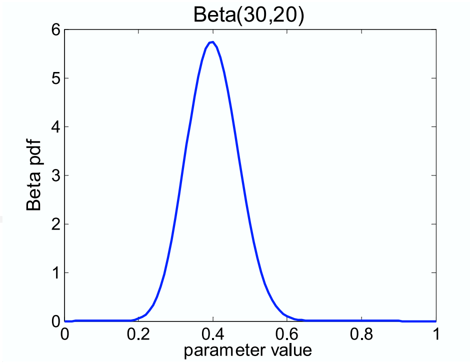
- Given
  - Prior: $\Theta \sim \text{Beta}(\alpha_H, \alpha_T)$
  - Data: $\mathcal{D}$ with m_H heads and m_T tails (binomial likelihood)
- Posterior distribution:

$$\Theta|\mathcal{D} \sim \text{Beta}(\alpha_H + m_H, \alpha_T + m_T)$$

- (Parametric) prior $P(\theta|\alpha)$ is **conjugate** to likelihood function if **posterior is of the same parametric family**, and can be written as:

  $P(\theta|\alpha')$ for some new set of parameters α'

Bayesian Prediction of a New Coin Flip



- Prior: $\Theta \sim Beta(\alpha_H, \alpha_T)$
- Observed m_H heads, m_T tails
- What is probability that next ($m+1^{st}$) flip is heads?

$$P(X_{m+1} = H \mid D) = \int_0^1 P(X_{m+1} = H \mid \Theta, D) \times P(\Theta \mid D) \; d\Theta$$

$$= \int_0^1 \Theta \times \boxed{Beta\,(\Theta : \alpha_H + m_H, \alpha_T + m_T)} \; d\Theta$$

$$= E_{\theta \sim Beta(\alpha_H + m_H, \alpha_T + m_T)}[\theta] = \frac{\alpha_H + m_H}{\alpha_H + m_H + \alpha_T + m_T}$$

Bayesian learning ≈ Smoothing

- Spse $\theta \sim$ Beta(1,4)
  Then see D = {+−++− − − −+−} = 4 +'s, 6 −'s
- Initially: E[θ] = $\frac{1}{5}$

  … MLE is $\frac{4}{4+6}$ = 0.4
- θ|D ~ Beta(1+4, 4+6) = Beta(5, 10)
  What is *Mean a posteri* ?

$$E[\Theta \mid D] = \frac{\alpha_H + m_H}{\alpha_H + m_H + \alpha_T + m_T} = \frac{\alpha_H}{m+\alpha} + \frac{m_H}{m+\alpha}$$

$$= \left[\frac{\alpha}{m+\alpha}\right]\underbrace{\boxed{\frac{\alpha_H}{\alpha}}}_{\text{prior}} + \left[\frac{m}{m+\alpha}\right]\underbrace{\boxed{\frac{m_H}{m}}}_{\theta_{MLE}}$$

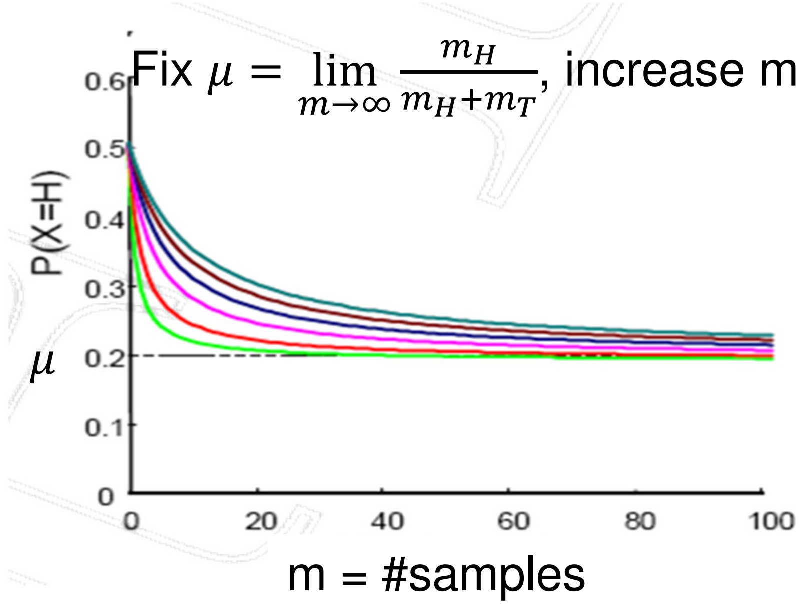
$m = m_H + m_T$ $\alpha = \alpha_H + \alpha_T$
… equivalent sample size

Bayesian learning ≈ Smoothing

- Spse $\theta \sim \text{Beta}(1,4)$
  Then see $D = \{+-++-\,-\,-\,-+-\}$ = 4 +'s, 6 –'s
- Initially: $E[\theta] = \frac{1}{5}$
  … MLE is $\frac{4}{4+6} = 0.4$
- $\theta|D \sim \text{Beta}(1+4,\ 4+6) = \text{Beta}(5,\ 10)$
  What is *Mean a posteri*?
  $$E[\,\theta\,|D\,] = \frac{1 \quad + \quad 4}{1+4 \quad + \quad 4+6} = \frac{5}{15}$$
- Note $E[\theta\ |\ D]$ is BLUR between $E[\theta]$ and MLE
  - … weighted by $\frac{5}{5+10}$ and $\frac{10}{5+10}$
  - Equivalent sample sizes:
    $\alpha = \alpha_H + \alpha_T = 5 \qquad m = m_H + m_T = 10$

Asymptotic behavior

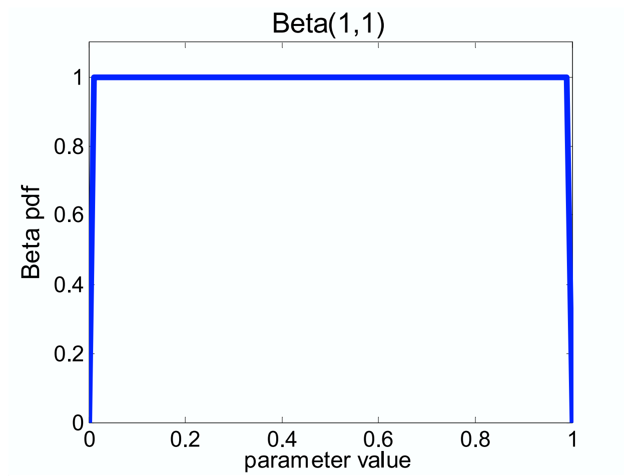
- $E[\theta\,] = \left[\frac{\alpha}{m+\alpha}\right]\frac{\alpha_H}{\alpha} + \left[\frac{m}{m+\alpha}\right]\frac{m_H}{m}$

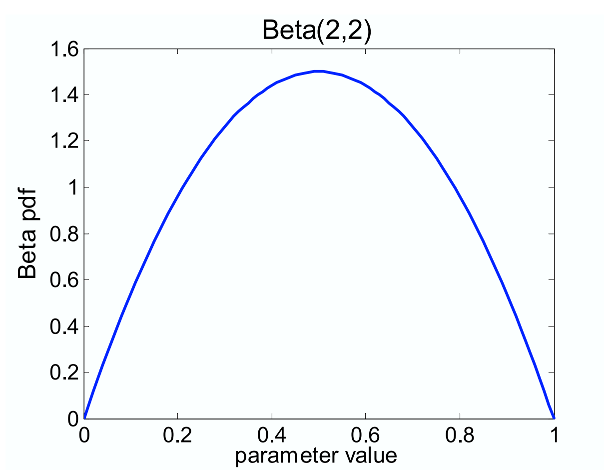


- For small sample size $m \approx 0$, prior $\frac{\alpha_H}{\alpha}$ is important
- As $m = m_T + m_H \to \infty$, prior is "forgotten…

Alternative "Encoding"

- Beta(a, b) ≡ B'(m, μ)
  where
  - m = (a+b)
    … effective sample size
  - $\mu = \frac{a}{a+b}$
- Eg
  - Beta(1, 1) = B'(2, 0.5)
  - Beta(10,10) = B'(20, 0.5)
  - Beta(7, 3) = B'(10, 0.7)
  - …





Bayesian learning for *Multi* nomial

- What if you have a k-sided thumbtack???
  - … still just ONE thumbtack (so just one event)
- Likelihood function if **multinomial**:
  - $P(X = i) = \theta_i \quad i = 1..k$
  - $\sum_i \theta_i = 1 \quad \theta_i \geq 0$
- **Conjugate** prior for multinomial is **Dirichlet**:
  - $\theta \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_k) \sim \prod_i \theta_i^{\alpha_i - 1}$
- **Observe** *m* data points, m_i from assignment i, **posterior**:
  - $\mathrm{Dirichlet}(\alpha_1 + m_i, \ldots, \alpha_k + m_k)$
- **Prediction**: $$P(X_{m+1} = i \mid D) = \frac{\alpha_i + m_i}{\sum_j (\alpha_j + m_j)}$$

Outline

- Foundations
  - Bayes Theorem
  - (Conditional) Independence
  - Dutch Book Theorem
  - Moments: Mean, Variance
- Estimation
  - MLE (Binomial)
  - Bayesian model
- Gaussian (Normal)
  - Properties of Gaussians
  - Learning Parameters of Gaussians

Overlap with Earlier Lecture

- Many of the slides in this section appeared in 1b-Foundation.
- Feel free to glance over…

Repeat

Useful Properties of Gaussians

- Lots of things can (arguably) be approximated well by Gaussians

- Central Limit Theorem:

The sum of IID variables with finite variances will tend towards a Gaussian distribution

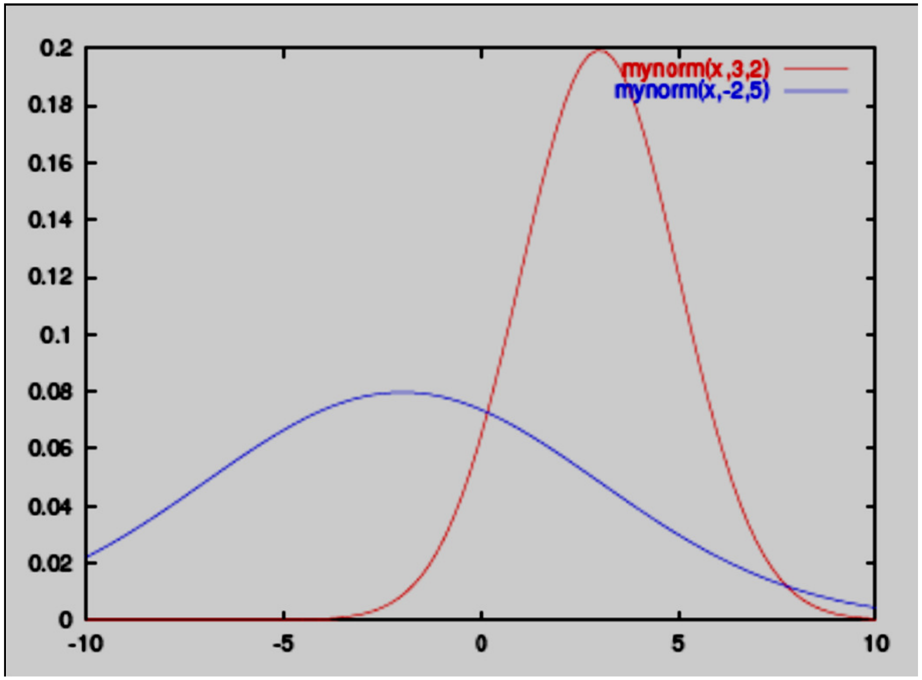
- CLT often used as a hand-waving argument to justify using the Gaussian distribution for almost anything

Multivariate Normal Distributions: A tutorial

Repeat

- **univariate normal** (Gaussian), with mean μ; variance σ^2
- PDF (probability distribution function)

$$p(x) = \frac{1}{(2\pi)^{1/2}\ \sigma} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$



Some Properties of Gaussians

- Affine transformation

  (multiplying by scalar and adding a constant)

  of Gaussian variables are Gaussian
  - $X \sim \mathcal{N}(\mu, \sigma^2)$
  - $Y = aX + b \Rightarrow Y \sim \mathcal{N}(a\mu+b, a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$
  - $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$
  - $Z = X+Y \Rightarrow Z \sim \mathcal{N}(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2)$

MVG = MultiVariate Gaussian
= Gaussian over many variables…

The Multivariate Gaussian

Repeat

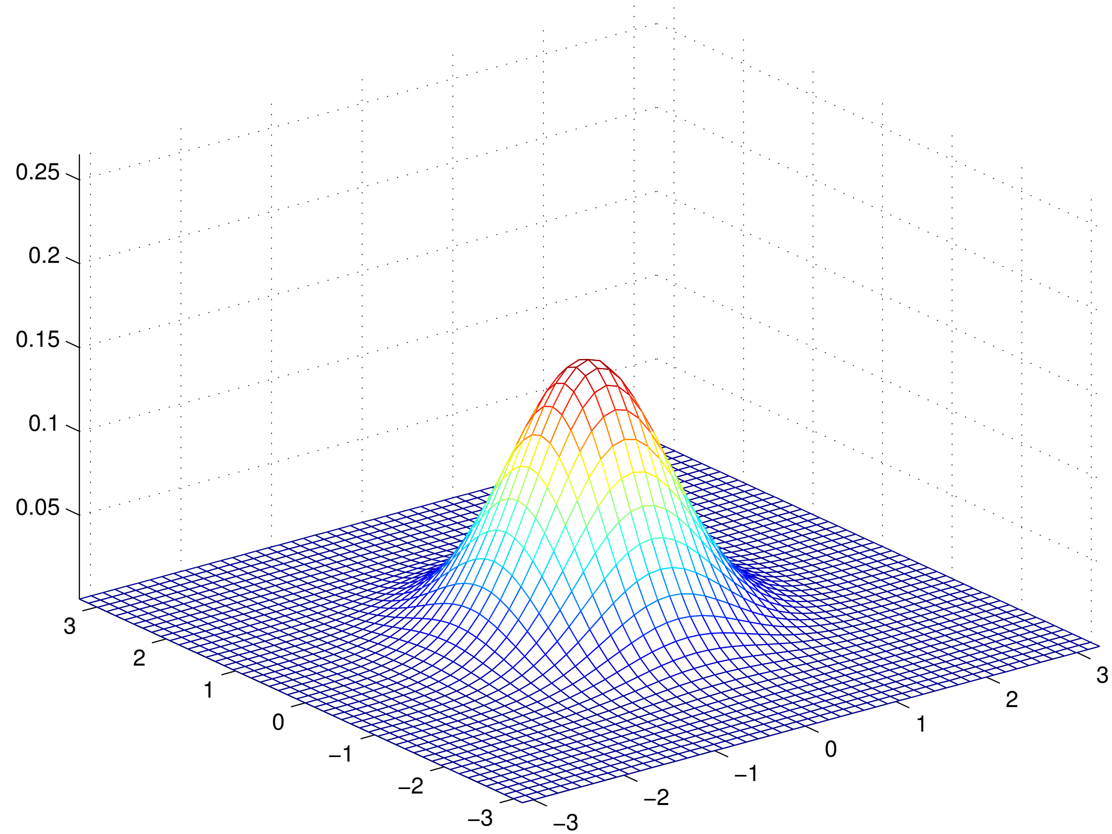
- A 2-dimensional Gaussian is defined by
  - a mean vector $\mu = [\ \mu_1, \mu_2\ ]$
  - a covariance matrix: $\Sigma = \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{2,1}^2 \\ \sigma_{1,2}^2 & \sigma_{2,2}^2 \end{bmatrix}$

  where $\sigma_{i,j}^2 = E[\ (x_i - \mu_i)\ (x_j - \mu_j)\ ]$
  is (co)variance

- Note: Σ is symmetric,
  "positive semi-definite": $\forall x$: $x^T \Sigma x \geq 0$

Jump

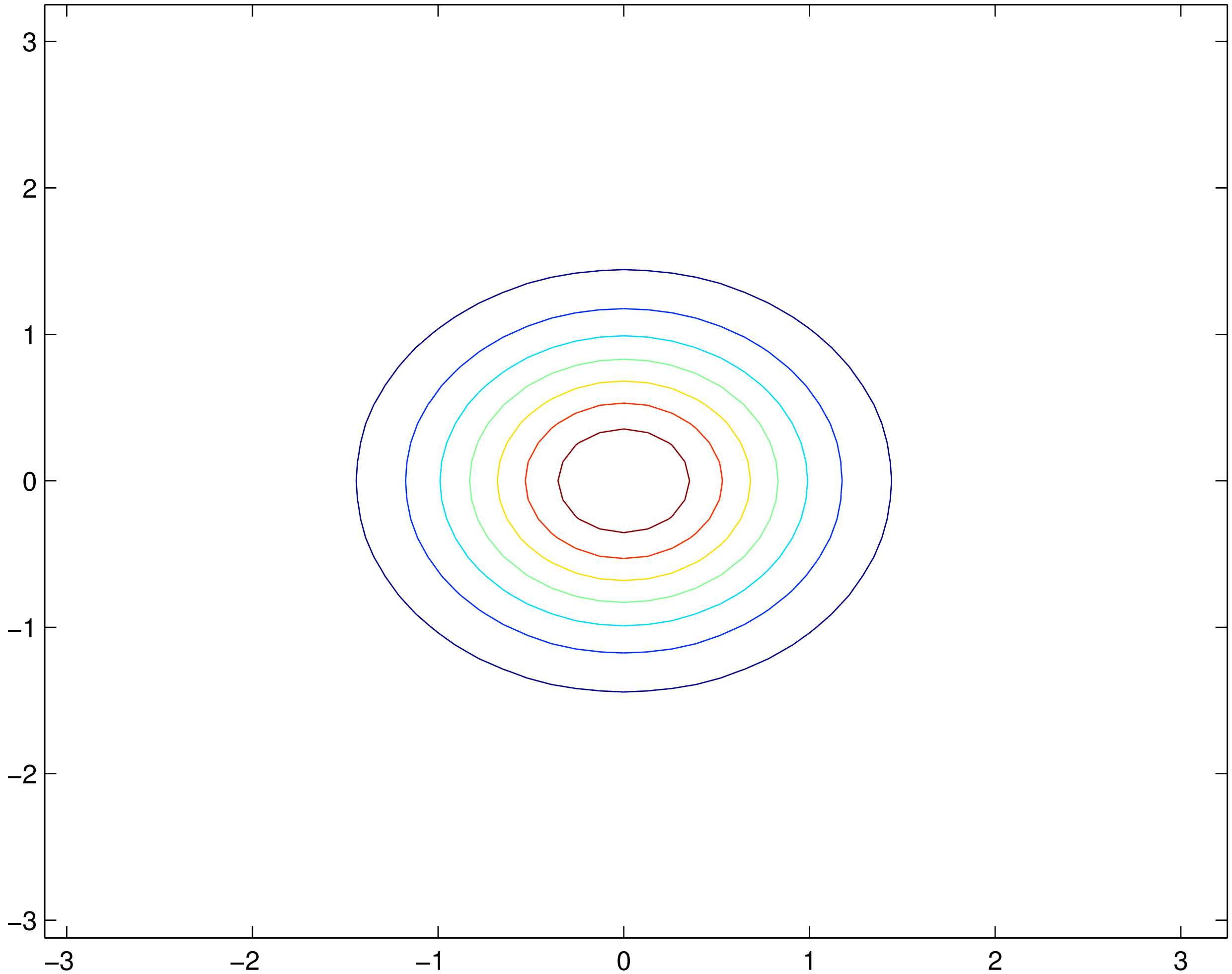
Standard Normal Distribution



- Standard normal for
  - Σ = the identity matrix
  - μ = (0,0)

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

MVG examples – contour plots



$\mu = (0,0)$ $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Standard Independent Gaussian

- Standard independent normal:

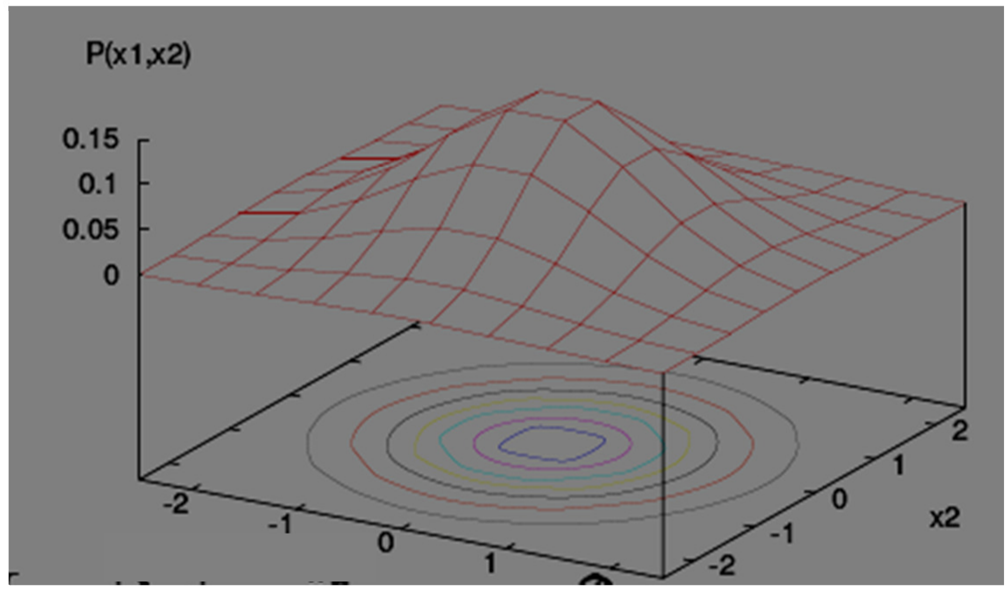
$$\mu = \langle 0, 0 \rangle \text{ and } \Sigma = I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Here: $\Sigma^{-1} = I_2$, $|\Sigma| = 1$; $n = 2$

$$P(\langle 3, -2 \rangle \mid \mathcal{N}(\langle 0,0 \rangle, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}))$$

$$= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\tfrac{1}{2}(\mathrm{x}-\mu)^{\top}\Sigma^{-1}(\mathrm{x}-\mu)\right]$$

$$= \frac{1}{(2\pi)^{2/2}\ 1^{1/2}} \exp\left[-\tfrac{1}{2}(\langle 3,-2 \rangle - \langle 0,0 \rangle)^{\top} I_2 (\langle 3,-2 \rangle - \langle 0,0 \rangle)\right]$$



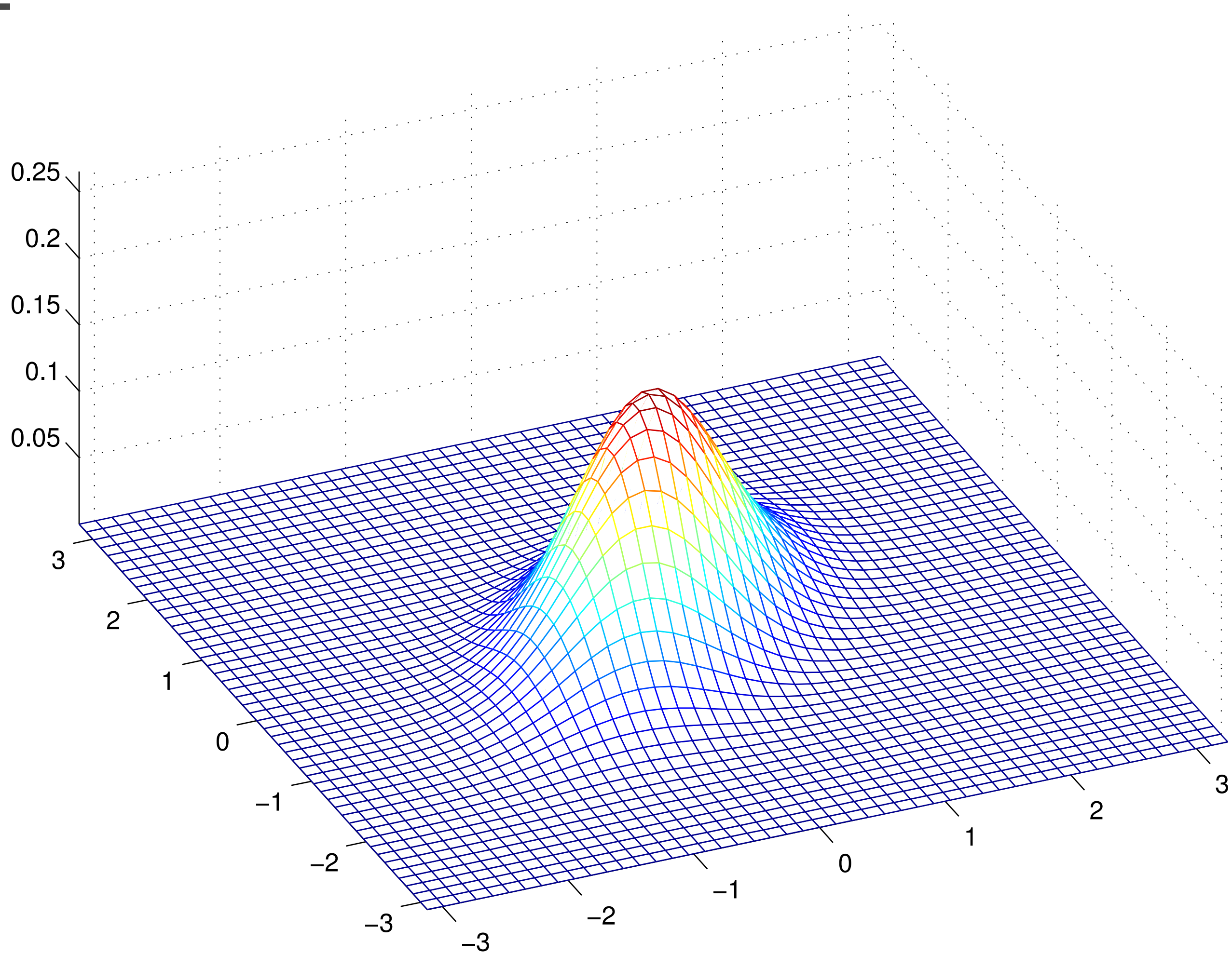
- $(\langle 3,-2 \rangle - \langle 0,0 \rangle)^{\top} I_2 (\langle 3,-2 \rangle - \langle 0,0 \rangle)$

$$= [3, -2] \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix}$$

$$= (3 \times 3) + (-2 \times -2) = 13$$

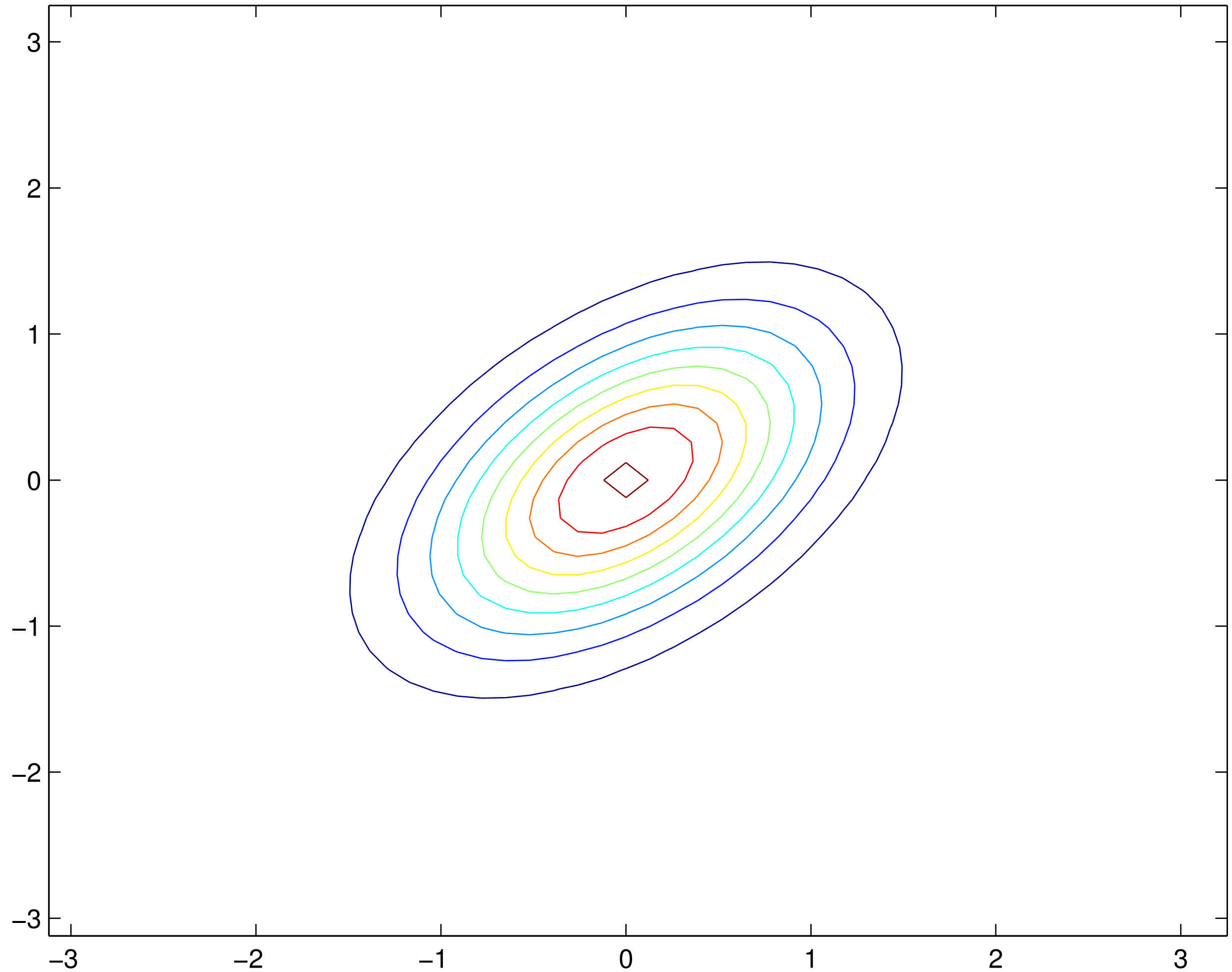
So $P(\langle -3, 2 \rangle \mid \ldots) = \frac{1}{(2\pi)} \exp\left[-\tfrac{1}{2}13\right] = \ldots$

MVG examples



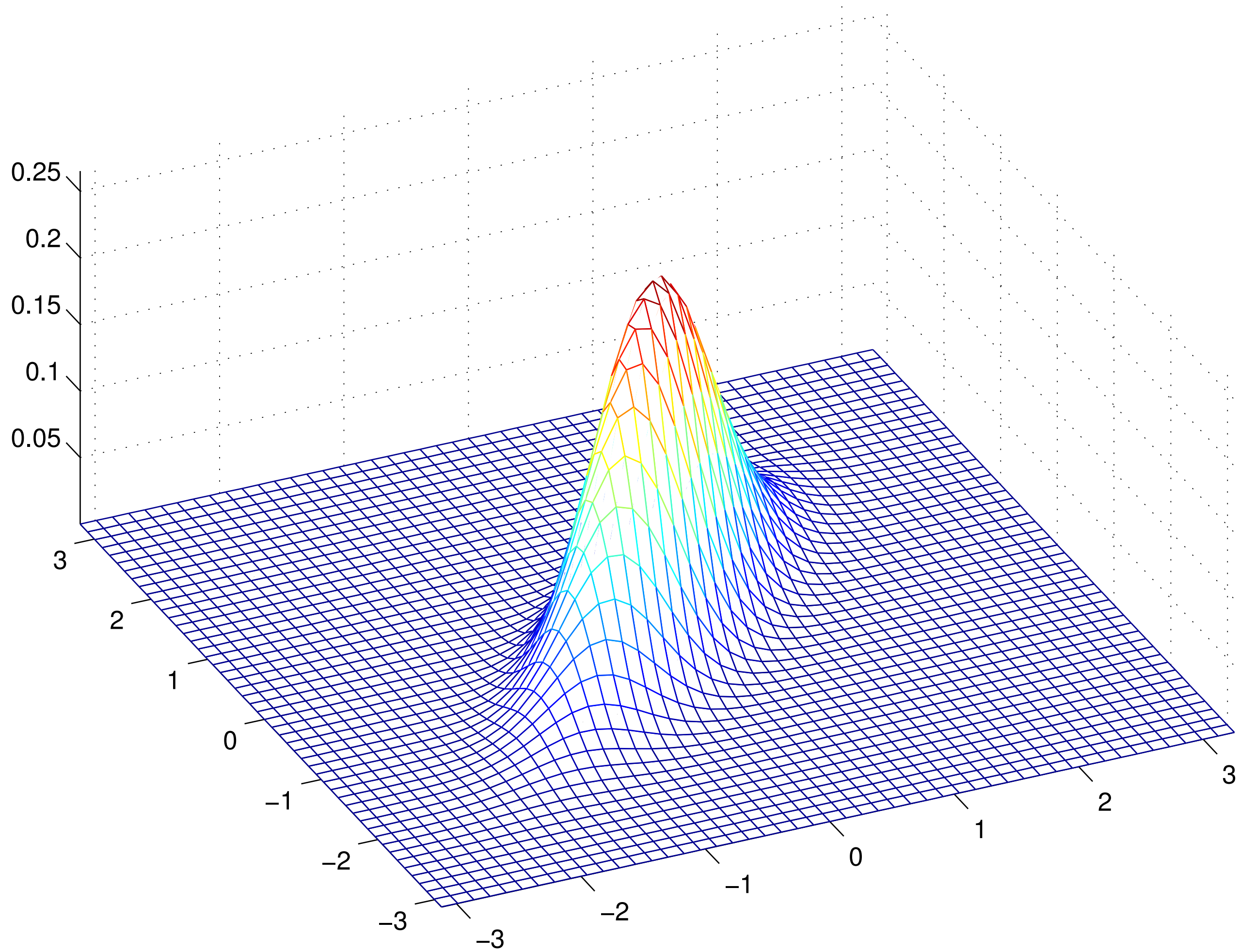
μ = (0,0) $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

MVG examples



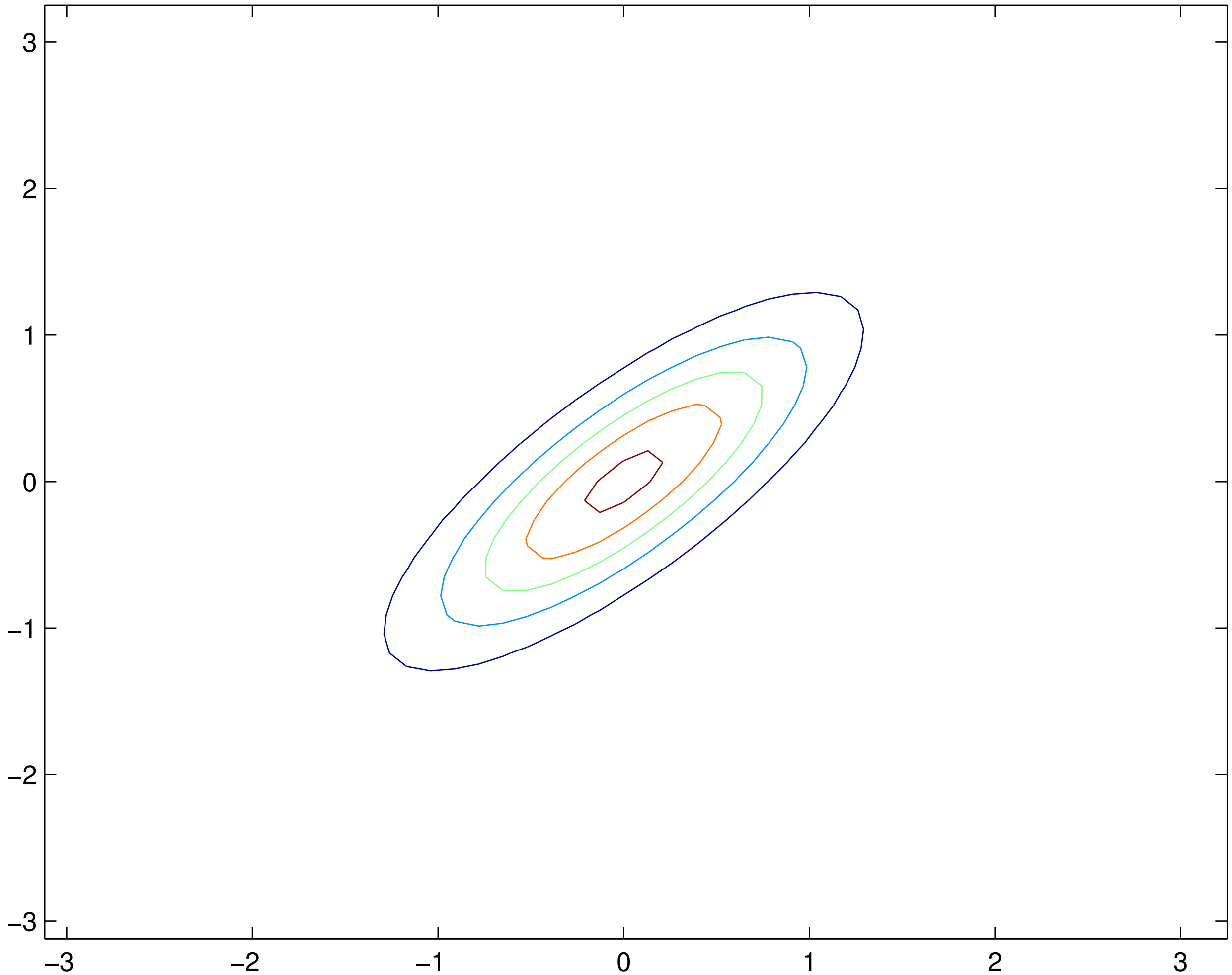
$\mu = (0,0)$ $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

MVG examples



$\mu = (0,0)$ $\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$

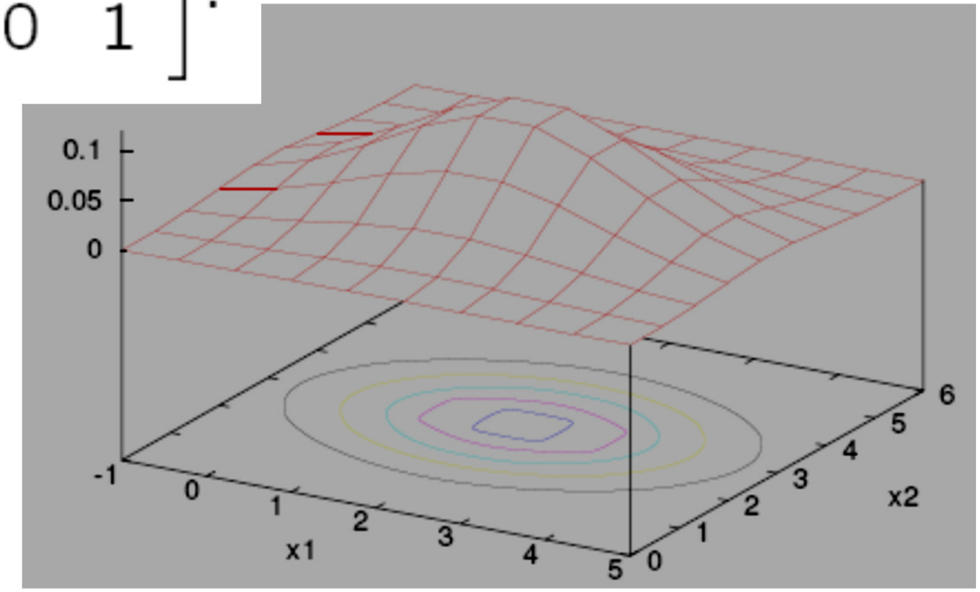
MVG examples



$\mu = (0,0)$ $\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$

The Multivariate Gaussian: Ex 2

$$Eg \quad \mu = \langle 2, 3 \rangle \quad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}:$$



- $P(\langle 3, -2 \rangle \mid \mathcal{N}(\langle 2, 3 \rangle, \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}))$

$$= \frac{1}{(2\pi)^{2/2} 2^{1/2}} \exp\left[-\frac{1}{2}(\langle 3, -2 \rangle - \langle 2, 3 \rangle)^\top \Sigma^{-1} (\langle 3, -2 \rangle - \langle 2, 3 \rangle)\right]$$

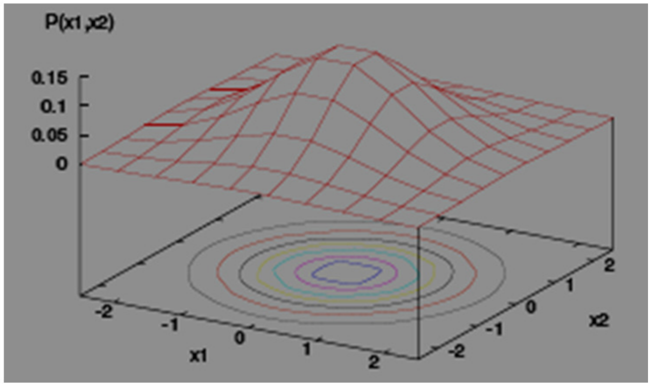
$$= \frac{1}{(2\pi) 2^{1/2}} \exp\left(-\frac{1}{2} \begin{bmatrix} 1 \\ -5 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} [1, -5]\right)$$

$$= \frac{1}{\alpha} \exp\left(-\frac{1}{2}[\frac{1}{2} \times 1^2 + 1 \times (-5)^2]\right)$$

Independent Variables

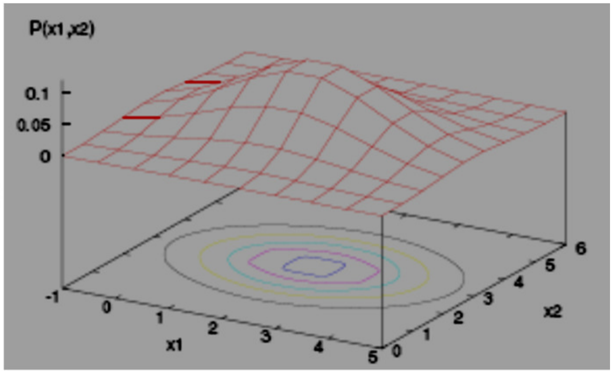
- Variables independent $\equiv$
  Covariance matrix is Diagonal

  Lines of equal probability $\equiv$ ellipses parallel to axes



- $P(\langle x,y\rangle = \langle 3,-2\rangle \mid \langle x,y\rangle \sim \mathcal{N}(\langle 0,0\rangle, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}))$

  $= P(x=3 \mid x \sim \mathcal{N}(0,1)) \times P(y=-2 \mid y \sim \mathcal{N}(0,1))$



- $P(\langle x,y\rangle = \langle 3,-2\rangle \mid \langle x,y\rangle \sim \mathcal{N}(\langle 2,3\rangle, \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}))$

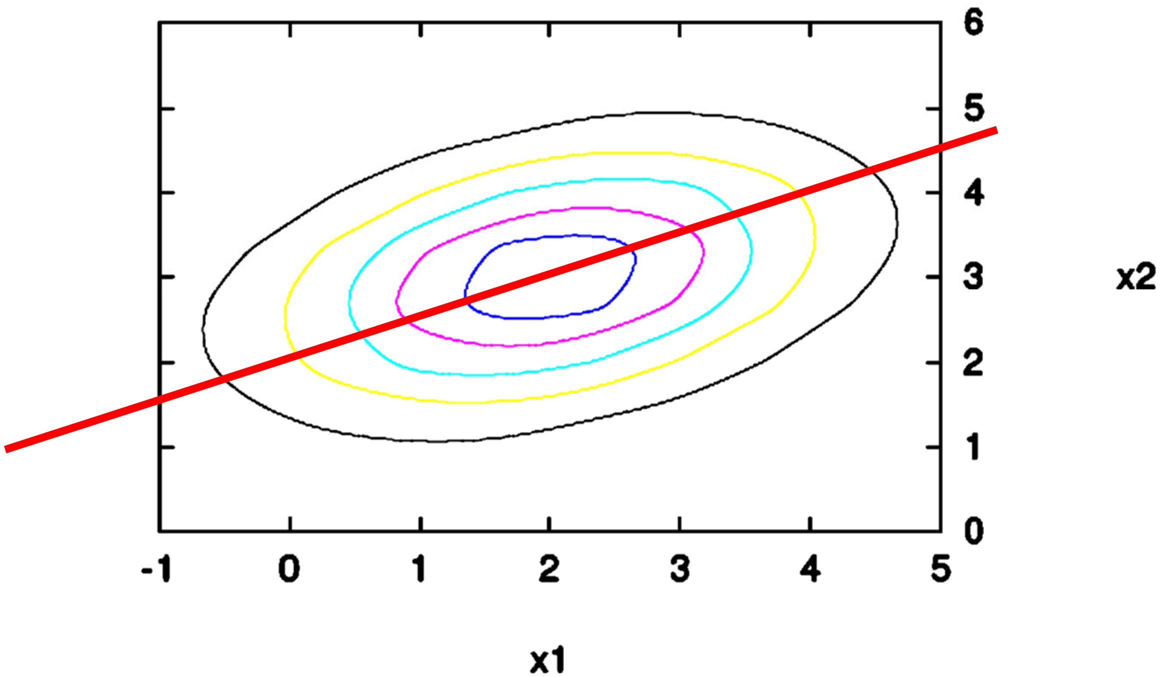
  $= P(x=3 \mid x \sim \mathcal{N}(2,2)) \times P(y=-2 \mid y \sim \mathcal{N}(3,1))$

The Multivariate Gaussian: Ex 3

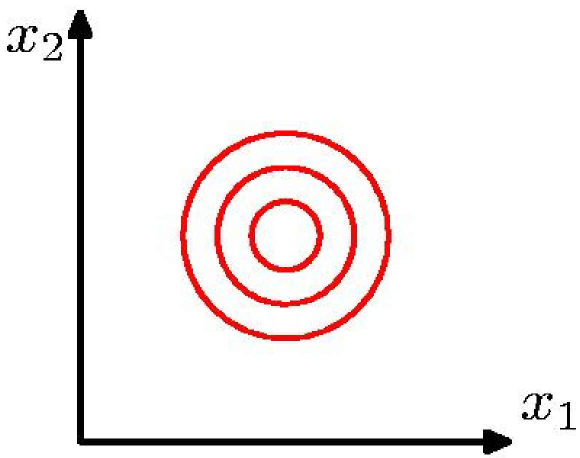
- If Σ is arbitrary,
  then x_1 and x_2 are dependent

  Lines of equal probability are
  "tilted" ellipses

  Eg For $\mu = \langle 2, 3 \rangle$ and $\Sigma = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}$:

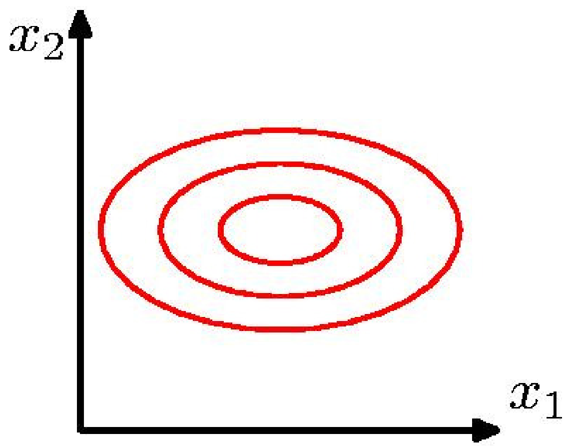


Examples of Gaussians

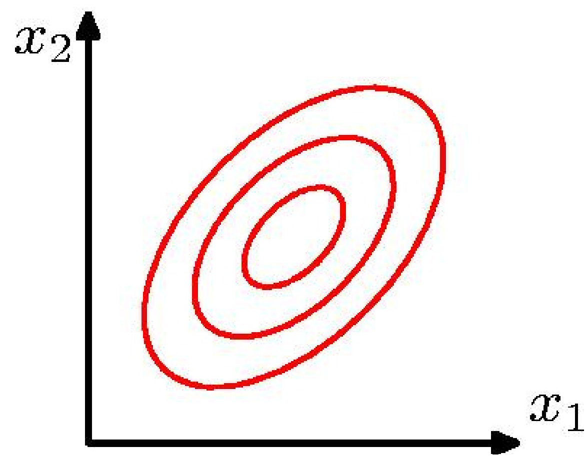


$\Sigma = \alpha I$

$= \mathrm{Diag}(\alpha, \ldots \alpha)$

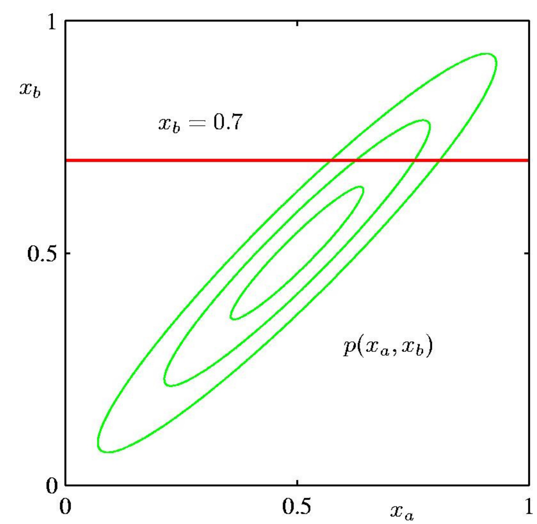


$\Sigma = \mathrm{Diag}(\alpha_1, \ldots \alpha_k)$



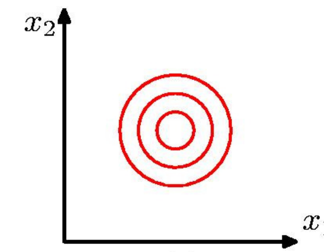
General Σ

Marginal…



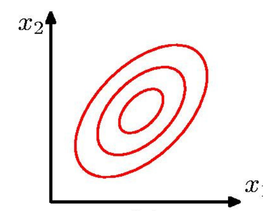
Useful Properties of Gaussians I

- Surfaces of equal probability ... (Mahalanobis curve)
  - for standard (mean 0, covariance I) Gaussians: spheroids



(c)

  - general Gaussians: ellipsoids



(a)

- Every general Gaussian ≡
  a standard Gaussian N(0,1) that has undergone an affine transformation

Useful Properties of Gaussians II

- A Gaussian distribution is completely specified by
  - a vector of means
  - a covariance matrix
- Requires $O(n^2)$ space
- Requires $O(n^3)$ time to manipulate
- Not great but… a joint distribution over n binary variables requires $O(2^n)$ space

Useful Properties of Gaussians III

- Marginals of Gaussians are Gaussian
- Given:

$$x = (x_a, x_b), \mu = (\mu_a, \mu_b)$$

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

- Marginal Distribution:

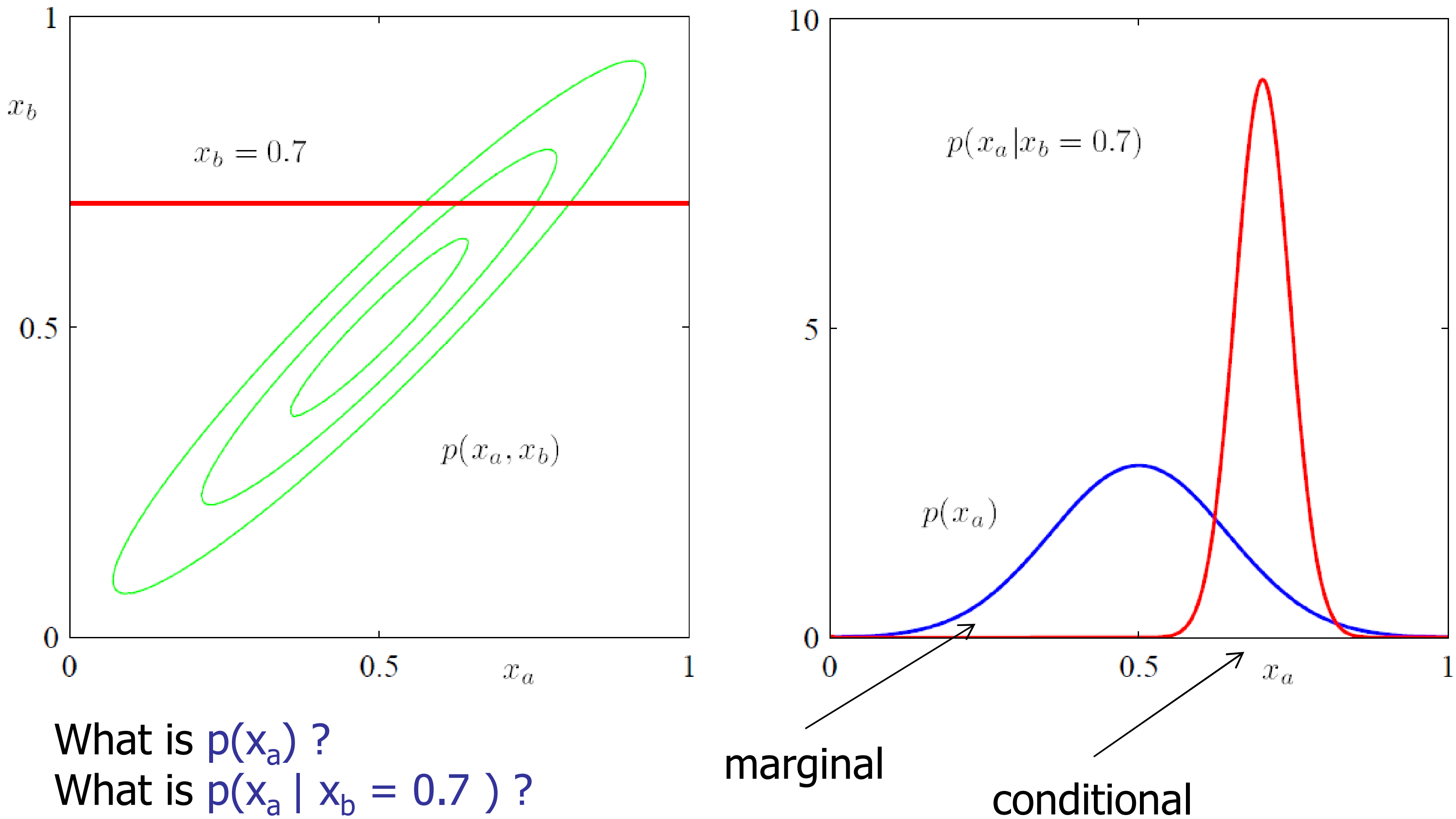
$$p(x_a) = N(x_a \mid \mu_a, \Sigma_{aa})$$

  - (Marginalize by ignoring)

Useful Properties of Gaussians IV

- Conditionals of Gaussians are Gaussian
- Notation:
  - $\Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$
  - "Precision matrix"
- Conditional Distribution:
  - $p(x_a|x_b) = N(x_a | \mu_{a|b}, \Lambda_{aa}^{-1})$
  - $\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab}(x_b - \mu_a)$

Visualizing Marginalization & Conditioning



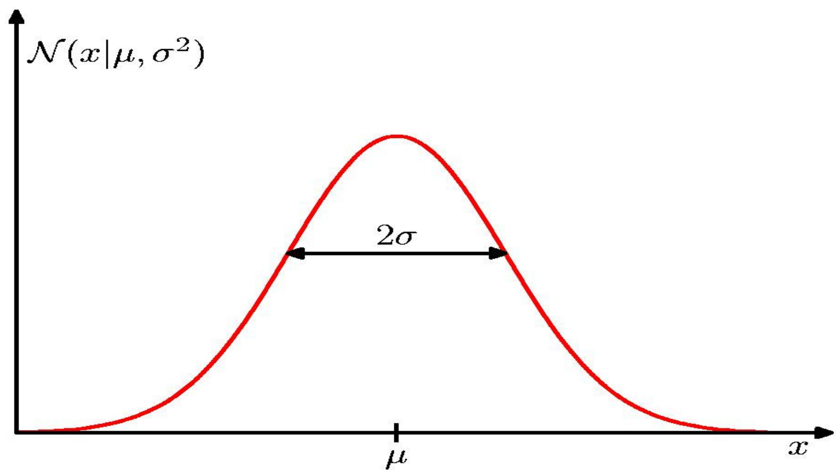
What is $p(x_a)$?
What is $p(x_a \mid x_b = 0.7)$?

Learning a Gaussian

Repeat

99
75
82
…
93
:

- Collect a set of data, D
  of real-valued i.i.d. instances
  - e.g., exam scores
- Learn parameters
  - Mean, μ
  - Variance, σ



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

MLE for Gaussian

Repeat

- Prob. of i.i.d. instances $D = \{x_1,\ldots,x_N\}$:

$$P(D \mid \mu,\sigma) = \prod_{i=1}^{N} P(x_i \mid \mu,\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- Log-likelihood of data:

$$\begin{aligned} \ln P(\mathcal{D} \mid \mu,\sigma) &= \ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right] \\ &= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

MLE for mean of a Gaussian

Repeat

- What is ML estimate $\hat{\mu}_{MLE}$ for mean μ?

$$\frac{d}{d\mu}\ln P(\mathcal{D}\mid\mu,\sigma) = \frac{d}{d\mu}\left[-N\ln\sigma\sqrt{2\pi} - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= -\sum_{i=1}^{N}\frac{d}{d\mu}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right] = \frac{1}{2\sigma^2}\sum_{i=1}^{N}2(x_i-\mu) = \frac{1}{\sigma^2}\left[\sum_{i=1}^{N}x_i - N\mu\right]$$

$$\frac{d}{d\mu}\ln P(D\mid\mu,\sigma) = 0 \Rightarrow \left[\sum_{i=1}^{N}x_i - N\mu\right] = 0$$

$$\Rightarrow \hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}x_i$$

Just empirical mean!!

MLE for Variance

- Again, set derivative to zero:

$$\frac{d}{d\sigma}\ln P(\mathcal{D}\mid\mu,\sigma) = \frac{d}{d\sigma}\left[-N\ln\sigma\sqrt{2\pi} - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{d}{d\sigma}\left[-N\ln\sigma\sqrt{2\pi}\right] - \sum_{i=1}^{N}\frac{d}{d\sigma}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{-N}{\sigma} - \sum_i \frac{-2(x_i-\mu)^2}{2\sigma^3}$$

$$\ldots = 0 \quad \Rightarrow \quad \boxed{\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_i (x_i-\mu)^2}$$

Just empirical variance!!

$\hat{\mu}_{MLE}$ is unbiased

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Estimator ŷ of y is unbiased *iff* E[ŷ] = y
- Observe { x_1, …, x_n }
  - drawn iid (independent and identically distributed)
  - … with common mean $E[x_i] = \mu$

$$E[\hat{\mu}_{MLE}] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu$$

Learning Gaussian parameters

- MLE:

$$\widehat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

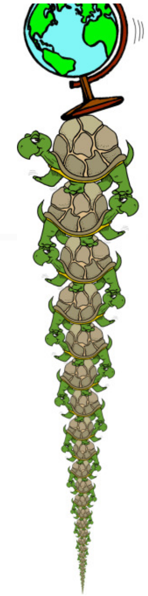
$$\widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{\mu})^2$$

- But… MLE for Gaussian variance is **biased**
  - Expected result of estimation ≠ true parameter!
  - Unbiased variance estimator:

Homework#3 !!

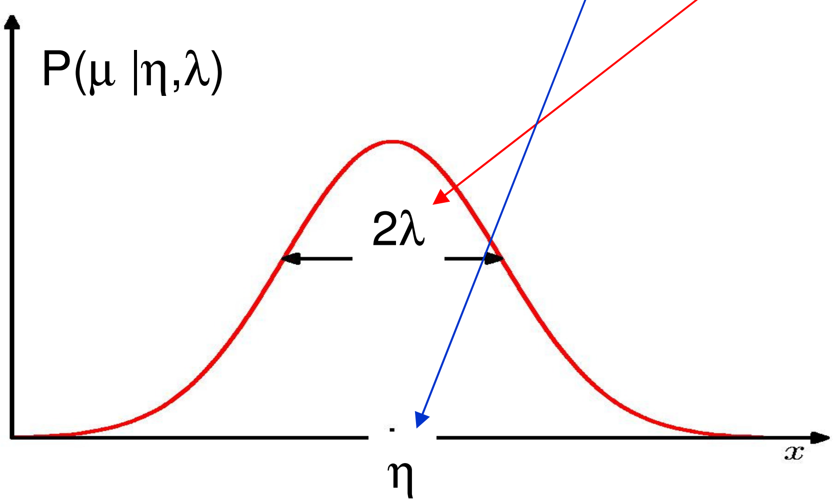
$$\widehat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \widehat{\mu})^2$$

Bayesian learning of Gaussian parameters



- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution
- Prior for mean:

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}}$$



MAP for mean of Gaussian

$$\boxed{P(\mu \mid D,\sigma,\eta,\lambda) \propto P(D \mid \mu,\sigma)\, P(\mu \mid \eta,\lambda)}$$

$$P(\mathcal{D} \mid \mu,\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \qquad P(\mu \mid \eta,\lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}}$$

$$\frac{d}{d\mu}\ln P(D \mid \mu)P(\mu) = \frac{d}{d\mu}\ln P(D \mid \mu) + \frac{d}{d\mu}\ln P(\mu)$$

$$= -\sum_i \frac{(\mu - x_i)}{\sigma^2} - \frac{(\mu-\eta)}{\lambda^2}$$

$$\ldots = 0 \quad \Rightarrow \quad \hat{\mu}_{MAP} = \left[\left(\sum_i \frac{x_i}{\sigma^2}\right) + \frac{\eta}{\lambda^2}\right] \Big/ \left[\frac{N}{\sigma^2} + \frac{1}{\lambda^2}\right]$$

MAP for mean of Gaussian

$$\hat{\mu}_{MAP} = \left[\left(\sum_i \frac{x_i}{\sigma^2}\right) + \frac{\eta}{\lambda^2}\right] \Big/ \left[\frac{N}{\sigma^2} + \frac{1}{\lambda^2}\right]$$

- If know nothing about mean μ, $\lambda^2 \to \infty$
  $\Rightarrow$ MAP estimate is same as MLE!
- But if $\lambda^2 < \infty$,
  then MAP is WEIGHTed AVERAGE of
  MLE and "prior" η

Limitations of Gaussians

- Gaussians are unimodal
  - single peak at mean
- $O(n^2)$ and $O(n^3)$ can get expensive
- Definite integrals of Gaussian distributions do not have a closed form solution (somewhat inconvenient)
  - Must approximate, use lookup tables, etc.
  - Sampling from Gaussian is inelegant

Mixtures of Gaussians

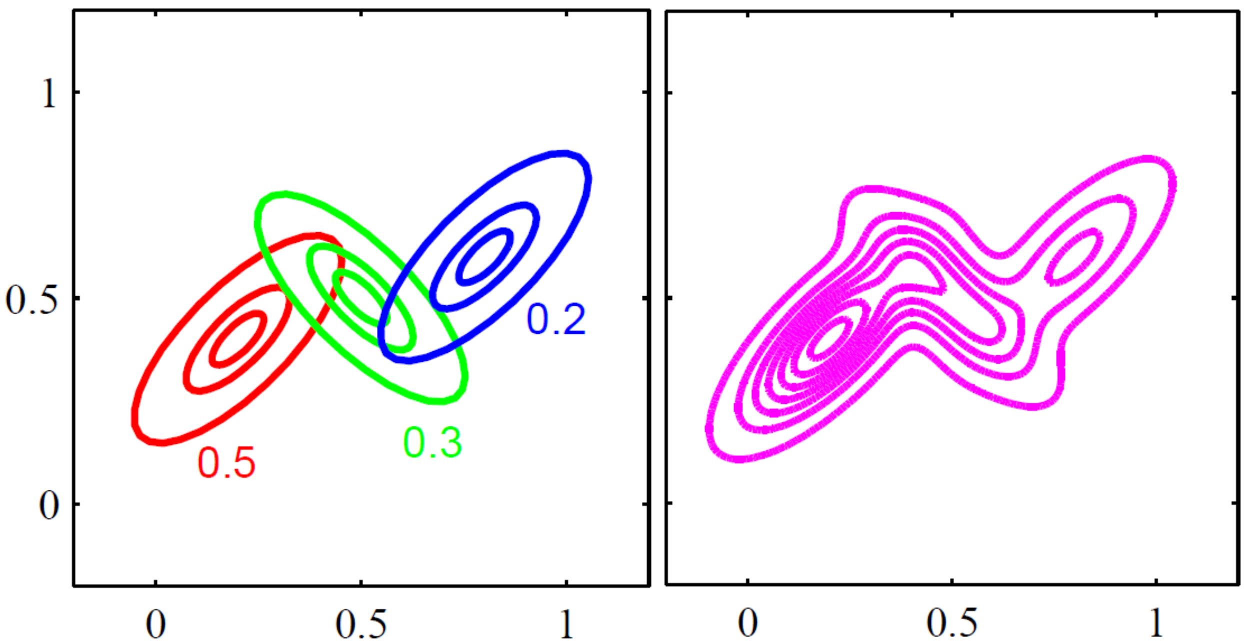
- Want to approximate distribution that is not unimodal…
- Density is weighted combination of Gaussians

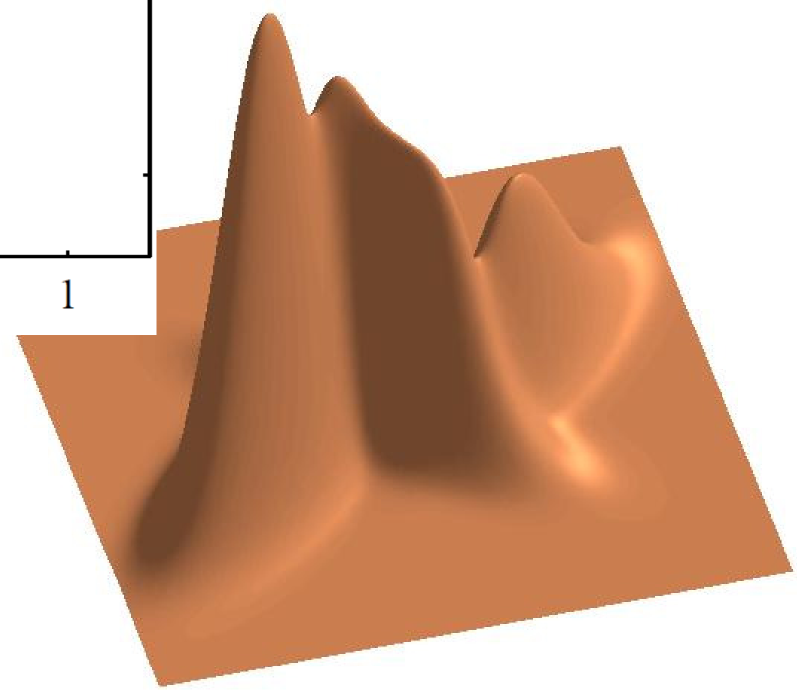
$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$

$$\sum_{k=1}^{K} \pi_k = 1$$

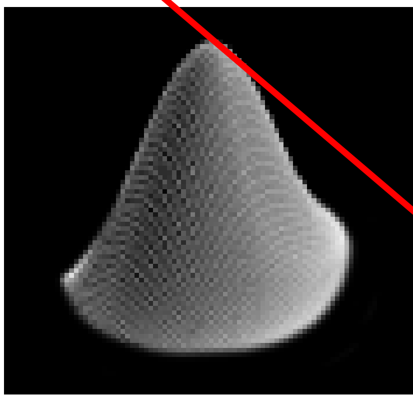
- Idea: Roll dice to select one of the Gaussians G_i, then draw sample from that Gaussian G_i
- Can be arbitrarily expressive with enough Gaussians

Mixture of Gaussians Example

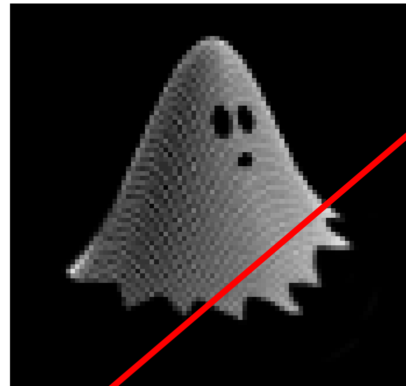




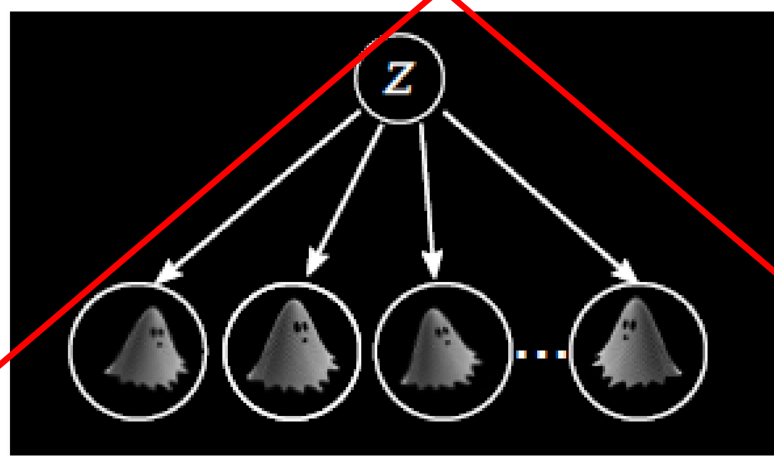
Types of Normal Distr’ns



Normal Distribution



ParaNormal Distribution



Mixtures of paranormal distributions with occult variables

D Maturana, A Spectral Approach to Ghost Detection, 2013

What you need to know

- Probability 101
- Point Estimation
  - MLE
  - Hoeffding inequality (PAC)
  - Bayesian learning
    - Beta, Dirichlet distributions
    - Gaussian, …
  - MAP (Maximum A Postiori) estimation
- ~~ParaNormal Distributions~~