HTF: … 2 …
B: Ch 2
RN: Ch 13

Probability 101

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?

Binomial Distribution

- Model:
  - P(Heads) = θ, P(Tails) = $1-\theta$
  - Flips are i.i.d.:
    - Independent events
    - Identically distributed according to distribution

- P(H,H,T,T,H) = $\theta\,\theta\,(1-\theta)\,(1-\theta)\,\theta = \theta^3(1-\theta)^2$
- Sequence *D* of α_H Heads and α_T Tails:

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Maximum Likelihood Estimation

- **Data:** Observed set *D* of
  α_H Heads and α_T Tails
- **Hypothesis Space:** Binomial distributions

- Learning "best" θ is an *optimization problem*
  - What's the objective function?

- **MLE**: Choose θ that maximizes the probability of observed data:

$$\widehat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$
$$= \arg\max_{\theta} \ln P(\mathcal{D} \mid \theta)$$

Simple "Learning" Algorithm

$$\widehat{\theta} = \arg\max_{\theta} \ln P(\mathcal{D} \mid \theta)$$
$$= \arg\max_{\theta} \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

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

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

$$\frac{h}{\theta} + \frac{-t}{(1-\theta)} = 0 \Rightarrow \hat{\theta} = \frac{t}{t+h}$$

So just average!!!

How many flips are "needed"?

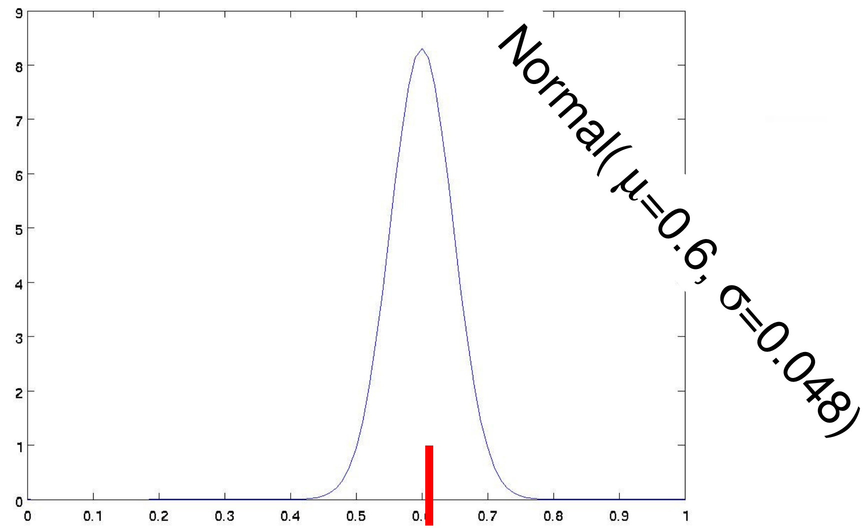
$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

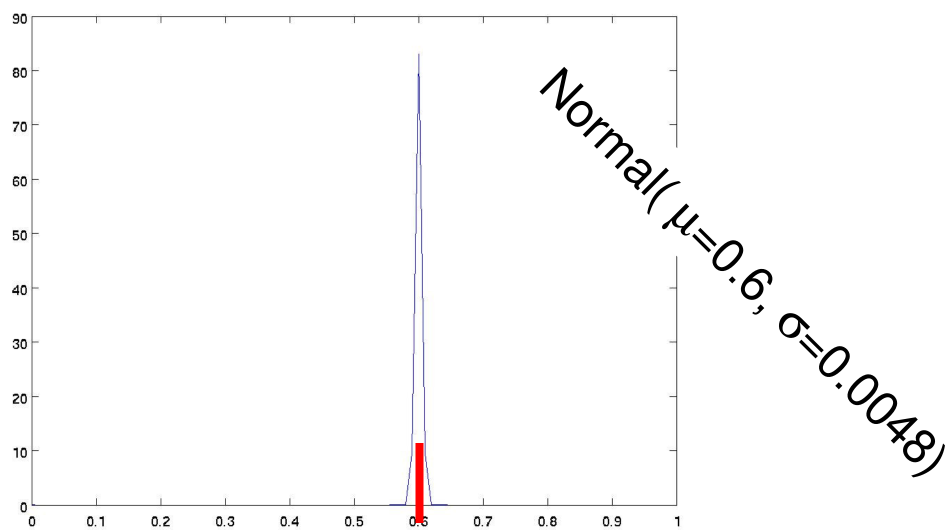
- Given 3 heads and 2 tails, θ_{MLE} = 3/5 = 0.6
- But…
  Given 30 heads and 20 tails, θ_{MLE} = 0.6
- **SAME!!!**
  Which is better? … more precise?

Using Variance

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

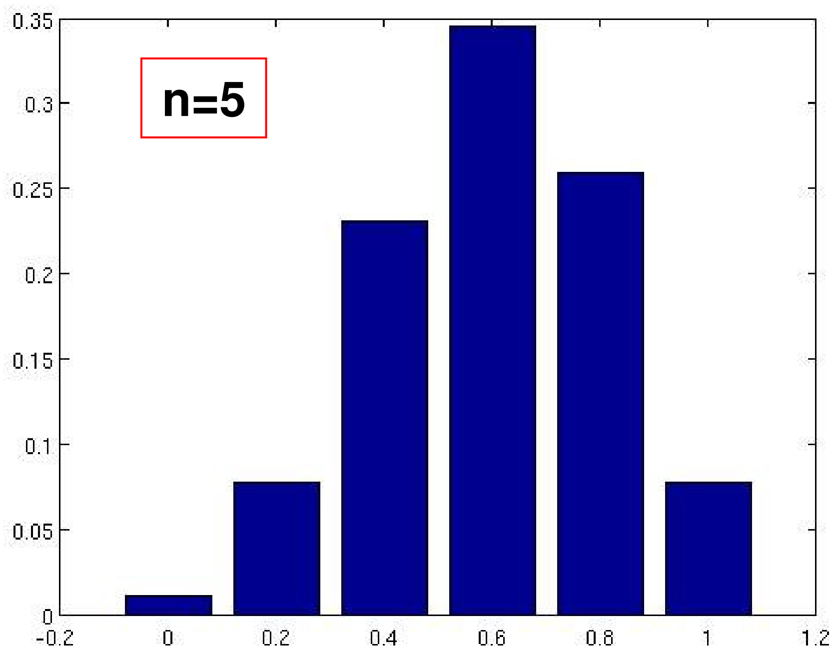
- Binomial(3H, 2T)
  $\mu=0.6 \quad \sigma=0.048$
- Binomial(30H, 20T)
  $\mu=0.6 \quad \sigma=0.0048$

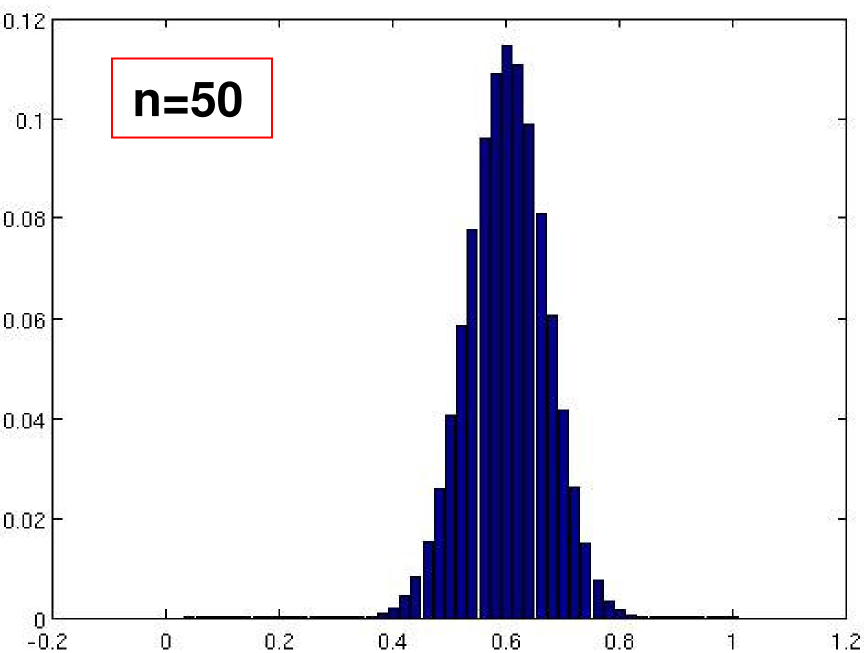




Binomial Distribution

P(D | θ) for fixed θ=0.6

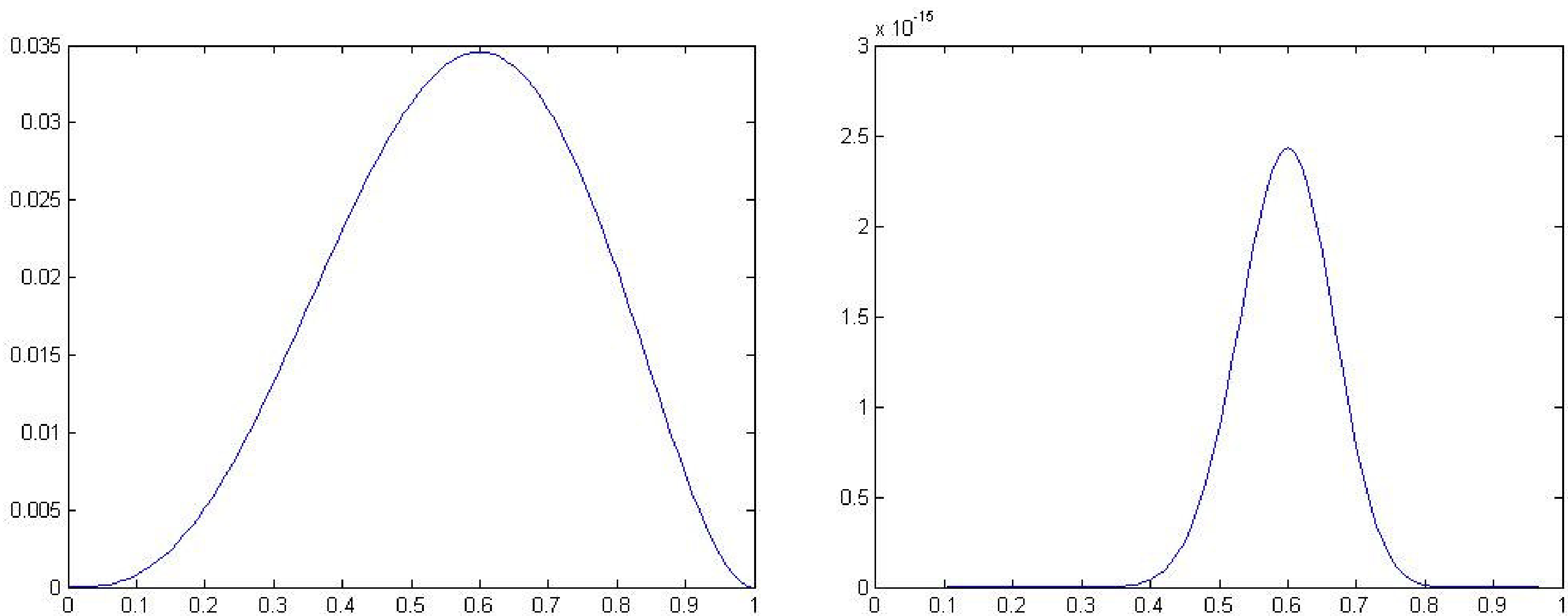




Prob that p=0.6 coin generates k/n heads, in n flips

Probability Functions

P(D | θ) for fixed D

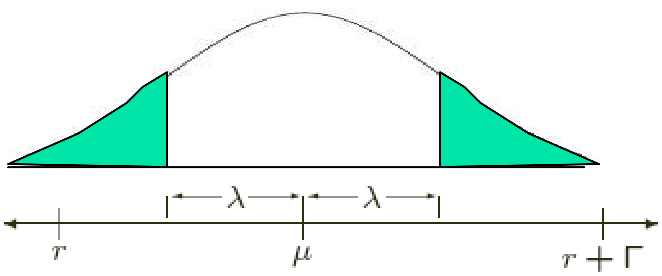


Prob that p=θ coin generates h heads, t tails

Hoeffding's Equality

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

- $P[S_m > \mu + \lambda\ ] < e^{-2m\lambda\wedge 2}$



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

- 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 $m = \alpha_H + \alpha_T$
- Sample average = $\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$
- Let θ^* be the true parameter

For any $\varepsilon > 0$:

$$P(|\,\widehat{\theta} - \theta^*\,| \geq \epsilon) \;\leq\; 2e^{-2N\epsilon^2}$$

PAC Learning

- PAC: Probably Approximate Correct

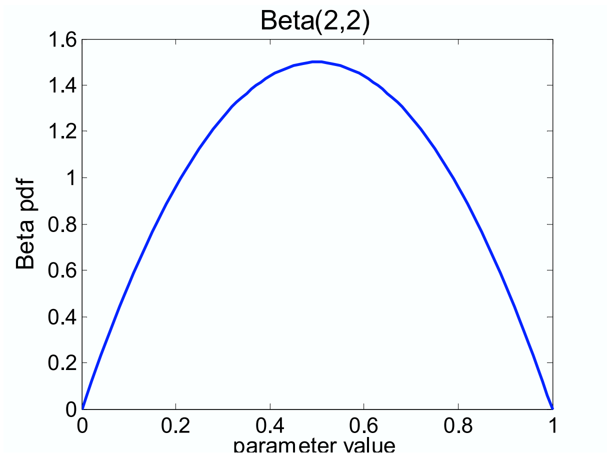
$$P(|\,\widehat{\theta} - \theta^* \,| \geq \epsilon) \;\leq\; 2e^{-2N\epsilon^2}$$

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

require #flips $m > (\ln 2/\delta)/\, 2\varepsilon^2$ ≈ 460.2

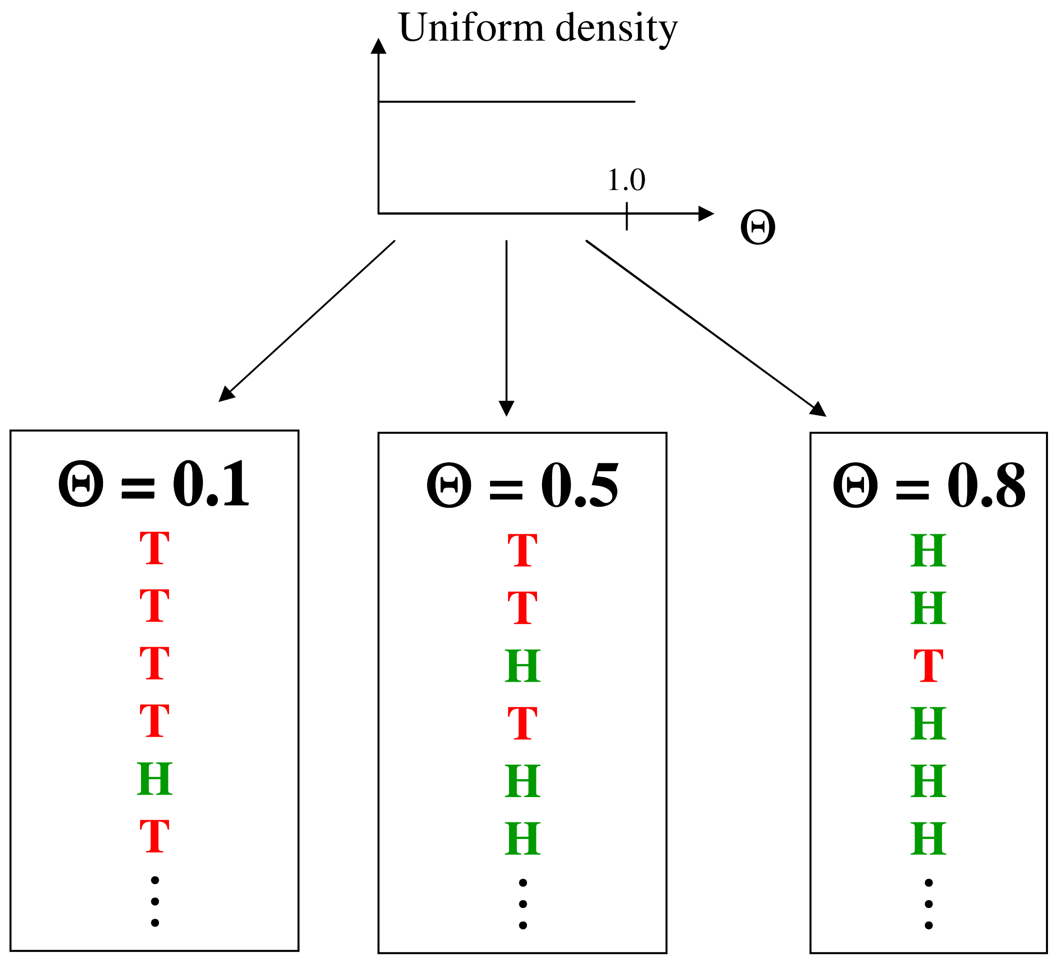
What about prior knowledge?

- 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

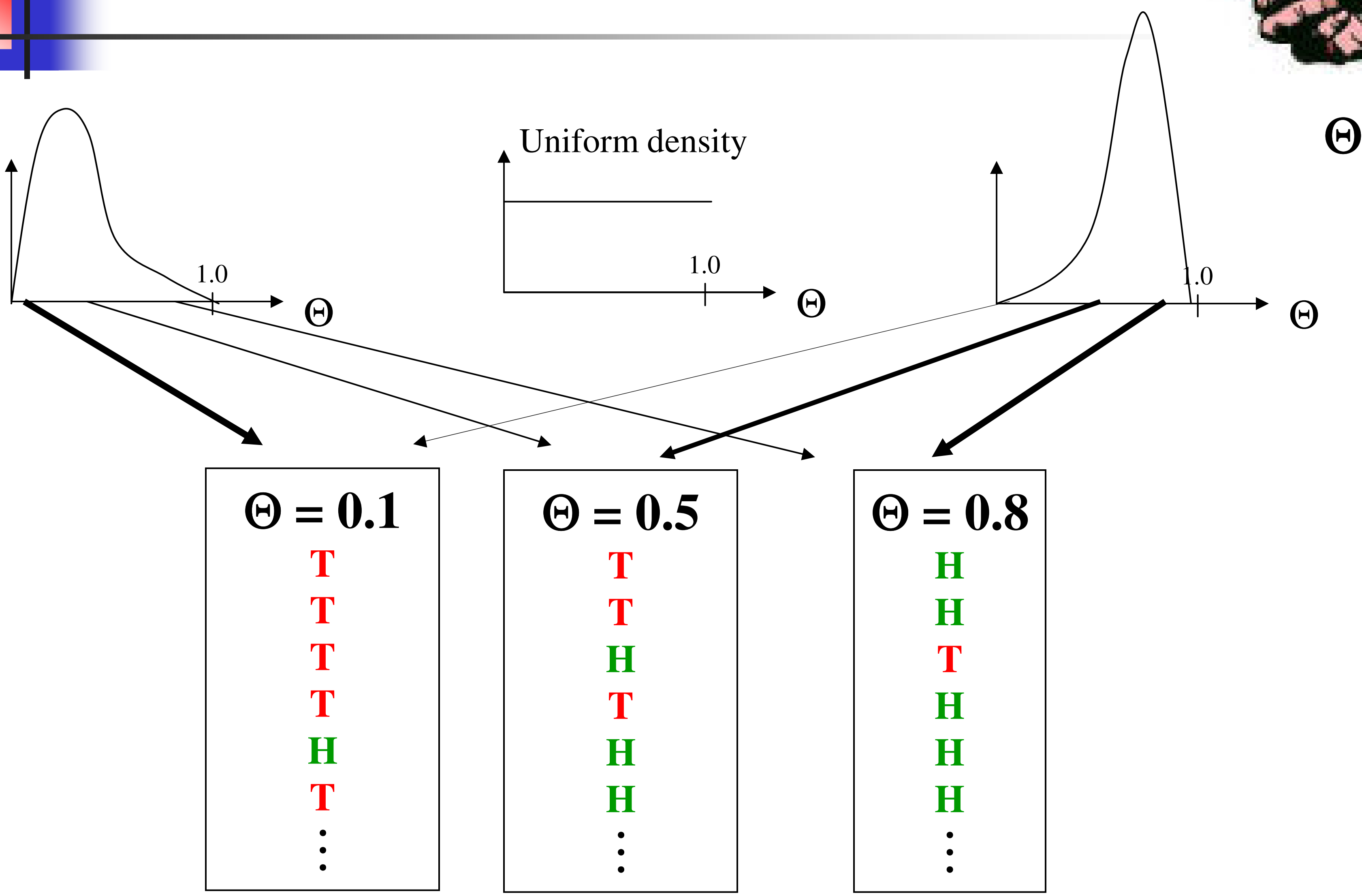


Two (related) Distributions: Parameter, Instances

Θ



Two (related) Distributions: Parameter, Instances



Bayesian Learning

- Use Bayes rule:

likelihood, prior, posterior

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

- Or equivalently (wrt argmax_θ P(θ|D))

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

Bayesian Learning for Thumbtack

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

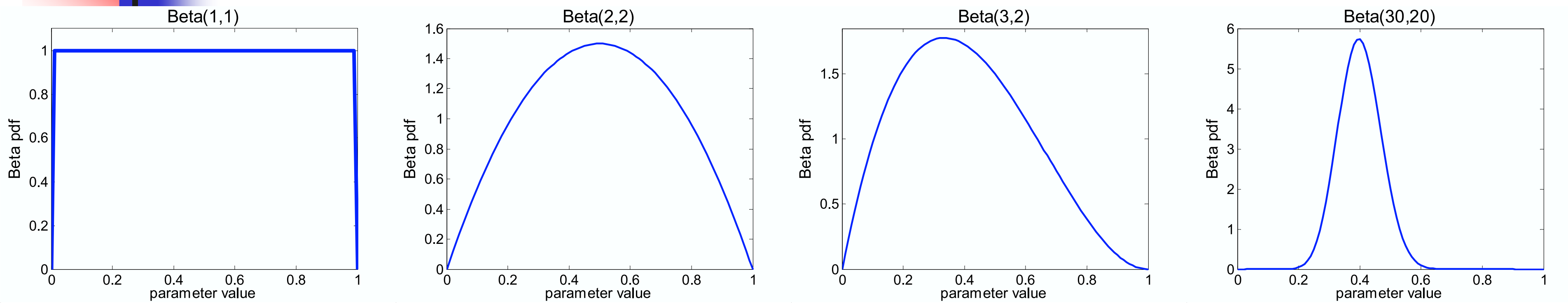
posterior *likelihood* *prior*

- Likelihood function is simply Binomial:

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

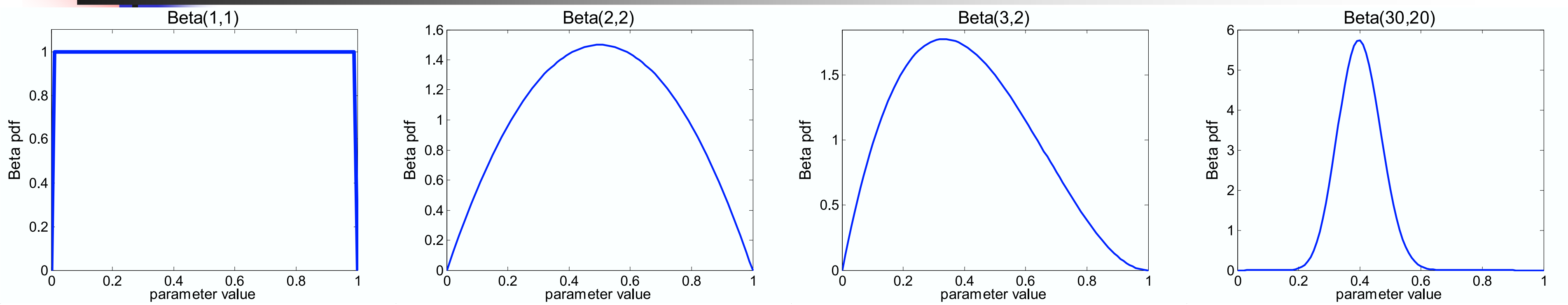
- What about prior?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors:
  - Closed-form representation of posterior (more details soon)
  - **For Binomial, conjugate prior is Beta distribution**

Beta prior distribution – P(θ)



- Prior: $$P(\theta) = \frac{\theta^{\alpha_H - 1}(1-\theta)^{\alpha_T - 1}}{B(\alpha_H, \alpha_T)} \sim Beta(\alpha_H, \alpha_T)$$
- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{m_H}(1-\theta)^{m_T}$
- Given X ~ Beta(a, b):
  - Mean: a/(a + b)
  - Unimodal if a,b>1… here mode: (a–1) / (a+b–2)
  - Variance: a b / $(a+b)^2$ (a+b–1)

Posterior distribution… from Beta



Prior P(θ)

Likelihood P(D|θ)

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

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

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

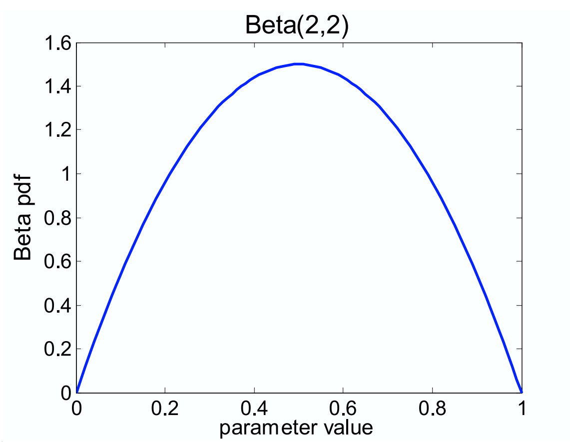
$$\sim \text{Beta}(\alpha_M + m_H,\ \alpha_T + m_T)$$

So Posterior is same form as Prior!! Conjugate!

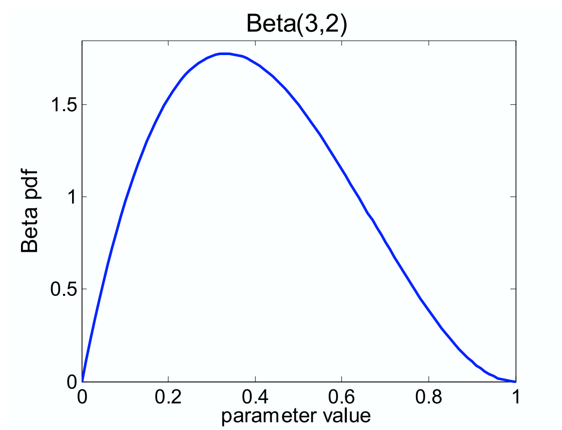
Posterior Distribution

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

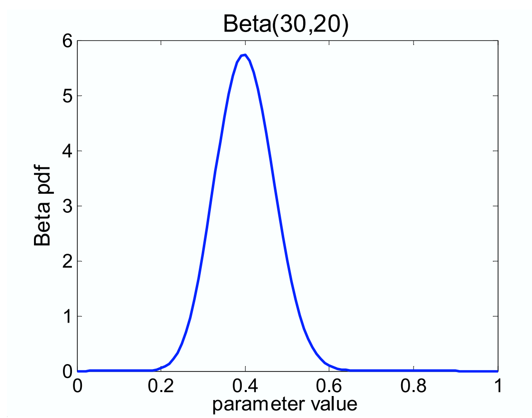
- Posterior distribution:
  $\theta \mid \mathcal{D} \sim Beta(m_H + \alpha_H, m_T + \alpha_T)$



Prior



+ observe 1 head



+ observe 27 more heads; 18 tails

Conjugate Prior

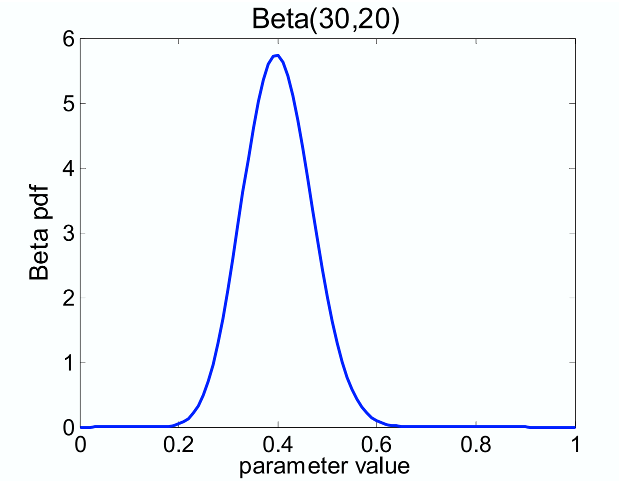
- Given
  - Prior: $\Theta \sim \mathrm{Beta}(\alpha_H, \alpha_T)$
  - Data: $\mathcal{D}$ with m_H heads and m_T tails (binomial likelihood)
- Posterior distribution:

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

- (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'

Using Bayesian Posterior



- Posterior distribution:

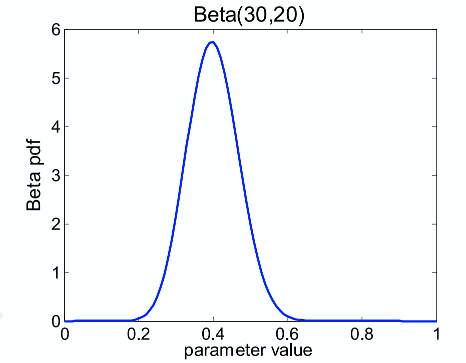
$$P(\theta \mid \mathcal{D}) \sim Beta(m_H + \alpha_H, m_T + \alpha_T)$$

- Bayesian inference … want f(θ)
  - No longer single parameter
  - Can use Expected value:

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

… but integral is often hard to compute

MAP: Maximum a posteriori approximation



$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

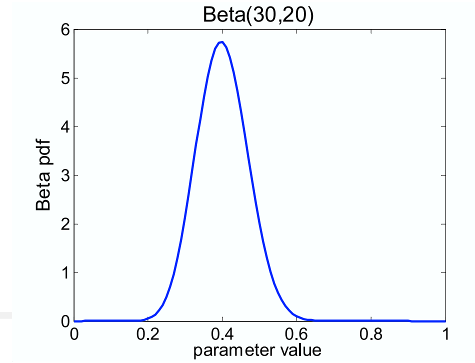
- As more data is observed, dist. is more peaked… more of distribution is at MAP:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

  - Like MLE = $\mathrm{argmax}_\theta P(D \mid \theta)$ but after "observing" prior ≈ (β_H-1 , β_T-1) extra flips
- MAP: use most likely parameter:

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta \approx f(\hat{\theta}_{MAP})$$

MAP for Beta distribution



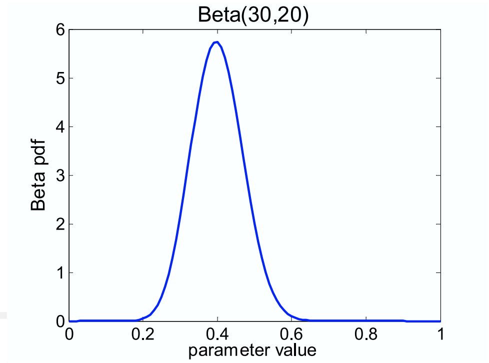
$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H,\beta_T+\alpha_T)} \sim Beta(\beta_H+\alpha_H,\beta_T+\alpha_T)$$

- MAP: use most likely parameter:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D) = \frac{\alpha_H+\beta_H-1}{\alpha_H+\beta_H+\alpha_T+\beta_T-2}$$

- Beta prior equivalent to extra thumbtack flips
- As $N \rightarrow \infty$, prior is "forgotten"
- **For small sample size, prior is important!**

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+1st) flip is heads?

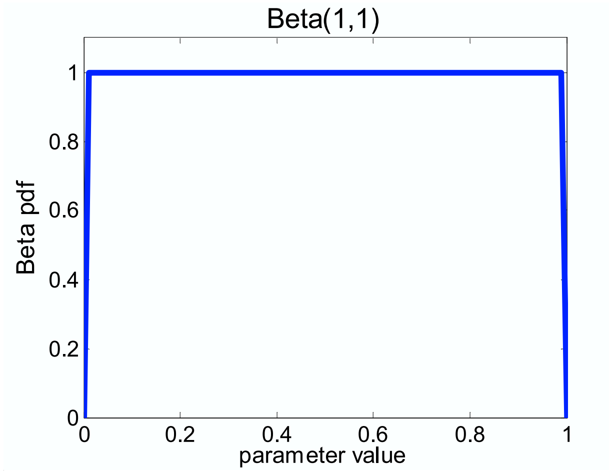
$$P(X_{m+1}=H \mid D) = \int_0^1 P(X_{m+1}=H \mid \Theta, D) \times P(\Theta \mid D)\, d\Theta$$

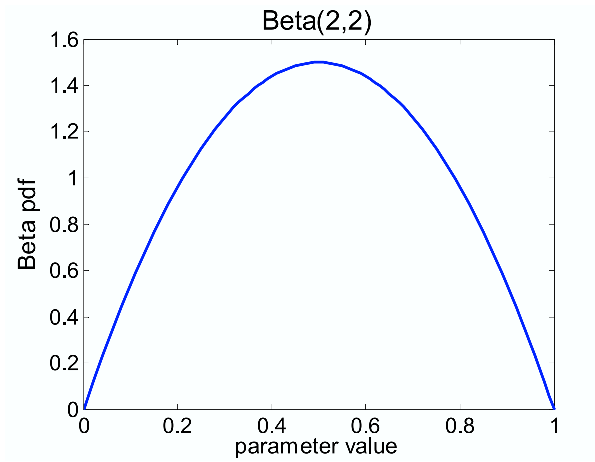
$$= \int_0^1 \Theta \times Beta(\Theta : \alpha_H + m_H, \alpha_T + m_T)\, d\Theta$$

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

Alternative "Encoding"

- Beta(a, b) ≡ B(m, μ) where
  - m = (a+b) … effective sample size
  - μ = 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)
  - …



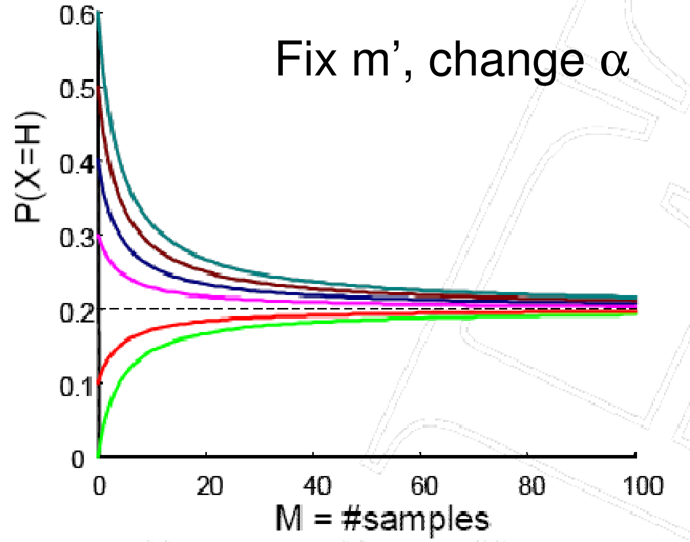


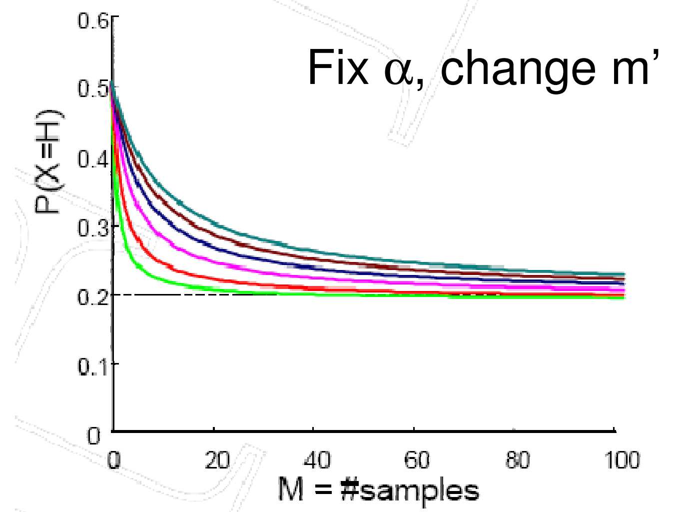
Asymptotic behavior and equivalent sample size

- Beta prior equivalent to extra flips:
  - $$E[\theta] = \frac{m_H + \alpha_H}{m_H + \alpha_H + m_T + \alpha_T}$$
- As $m \rightarrow \infty$, prior is "forgotten"
- **But, for small sample size, prior is important!**

$$E[\theta] = \frac{m_H + \alpha m'}{m_H + m_T + m'}$$

- **Equivalent sample size**:
  - Prior parameterized by α_H, α_T, or
  - m' (equivalent sample size) and α





Bayesian learning ≈ Smoothing

$$E[\Theta] = \frac{\alpha_H + m_H}{\alpha_H + m_H + \alpha_T + m_T}$$

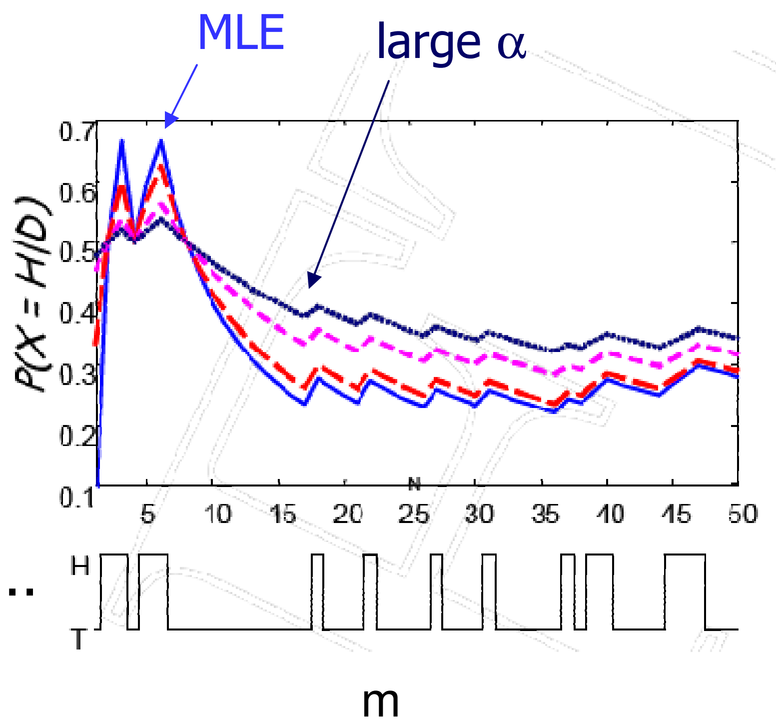
m = m_H + m_T … α = α_H + α_T
… equivalent sample size

$$= \frac{\alpha_H}{m+\alpha} + \frac{m_H}{m+\alpha}$$

$$= \frac{\alpha_H}{\alpha}\left[\frac{\alpha}{m+\alpha}\right] + \frac{m_H}{m}\left[\frac{m}{m+\alpha}\right]$$

prior

θ_{MLE}



- MLE estimate, biased towards prior…
  - m=0 ⇒ prior parameter
  - m→∞ ⇒ MLE

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**:
  - $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k) \sim \prod_i \theta_i^{\alpha_i - 1}$
- **Observe** *m* data points, m_i from assignment i, **posterior**:
  - Dirichlet($\alpha_1 + m_i, \ldots, \alpha_k + m_k$)
- **Prediction**: $P(X_{m+1} = i \mid D) = \frac{\alpha_i + m_i}{\sum_j (\alpha_j + m_j)}$

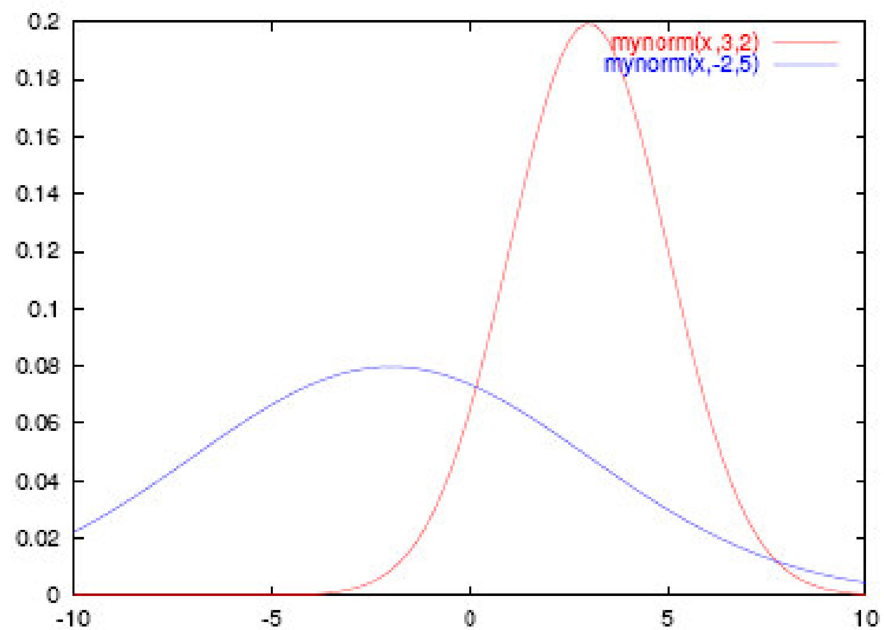
Outline

- 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

Multivariate Normal Distributions: A tutorial

- **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)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \Rightarrow Y \sim N(a\mu+b, a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim N(\mu_X, \sigma^2_X)$
  - $Y \sim N(\mu_Y, \sigma^2_Y)$
  - $Z = X+Y \Rightarrow Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

MVG = MultiVariate Gaussian
= Gaussian over many variables…

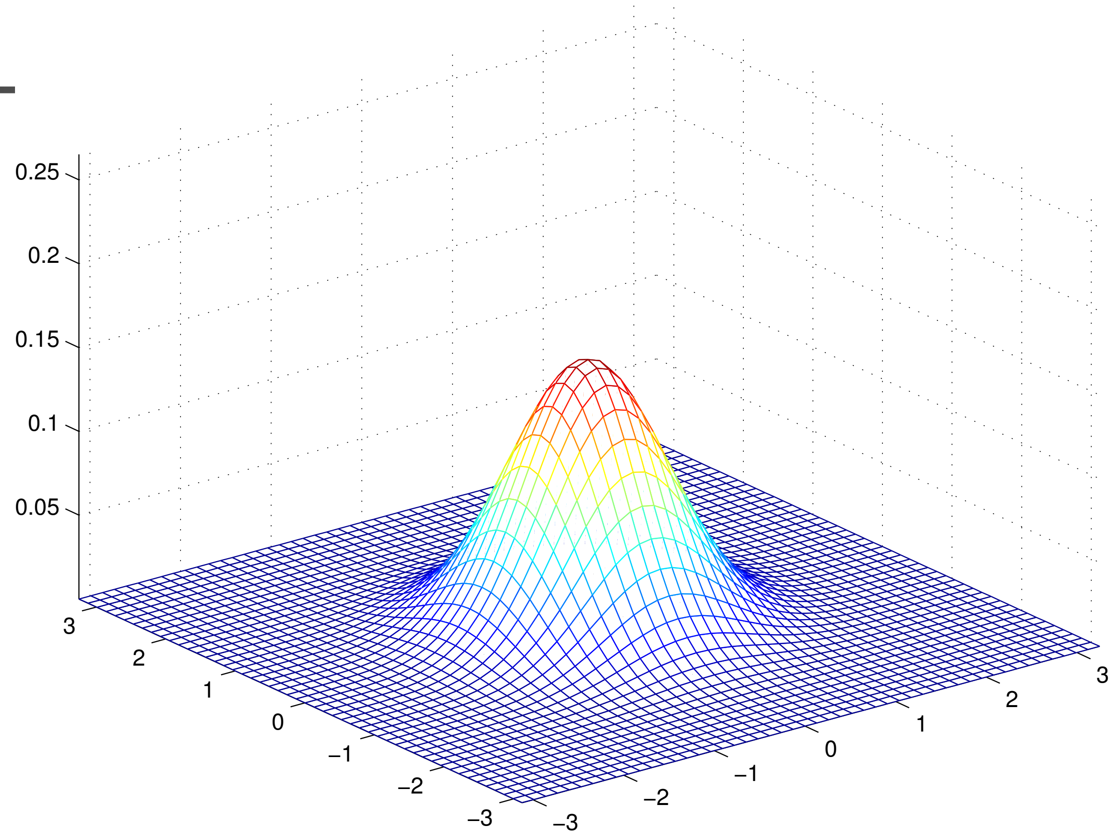
The Multivariate Gaussian

- 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": $\forall x$: $x^T \Sigma x \geq 0$

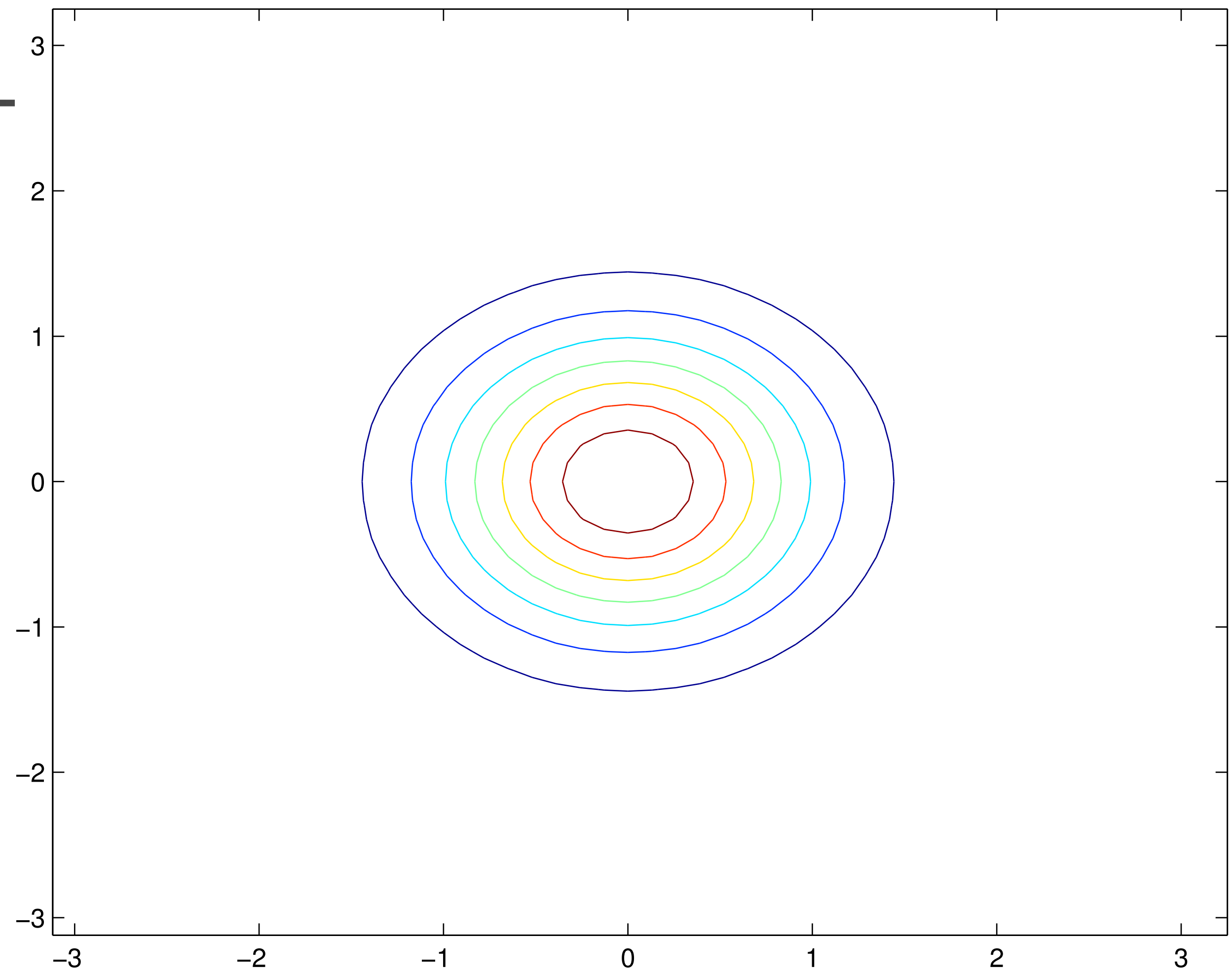
Standard Normal Distribution



- Standard normal for
  - Σ = the identity matrix
  - $\mu = (0,0)$

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

MVG examples – contour plots



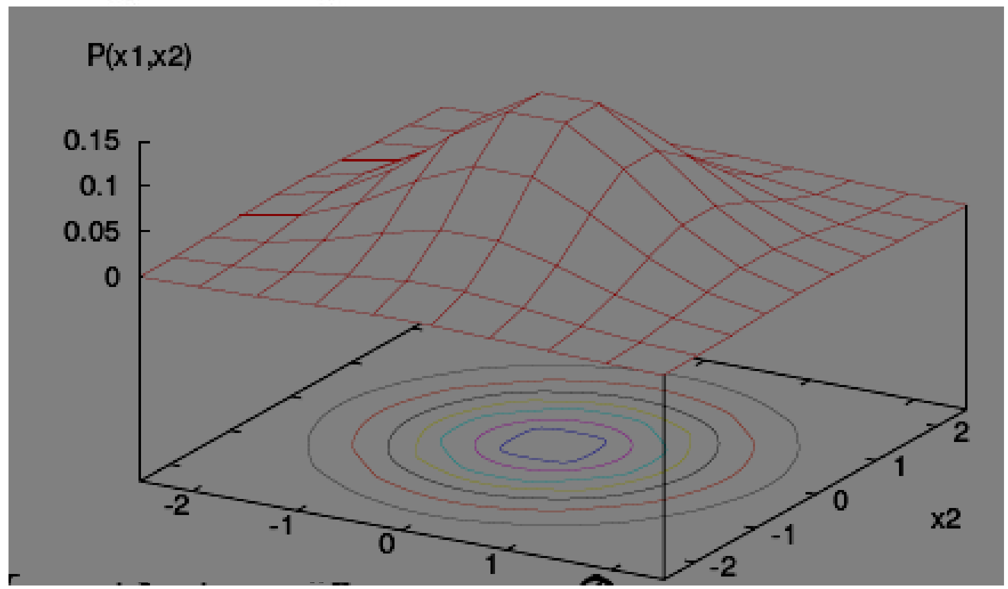
$\mu = (0,0)$

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

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

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

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

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

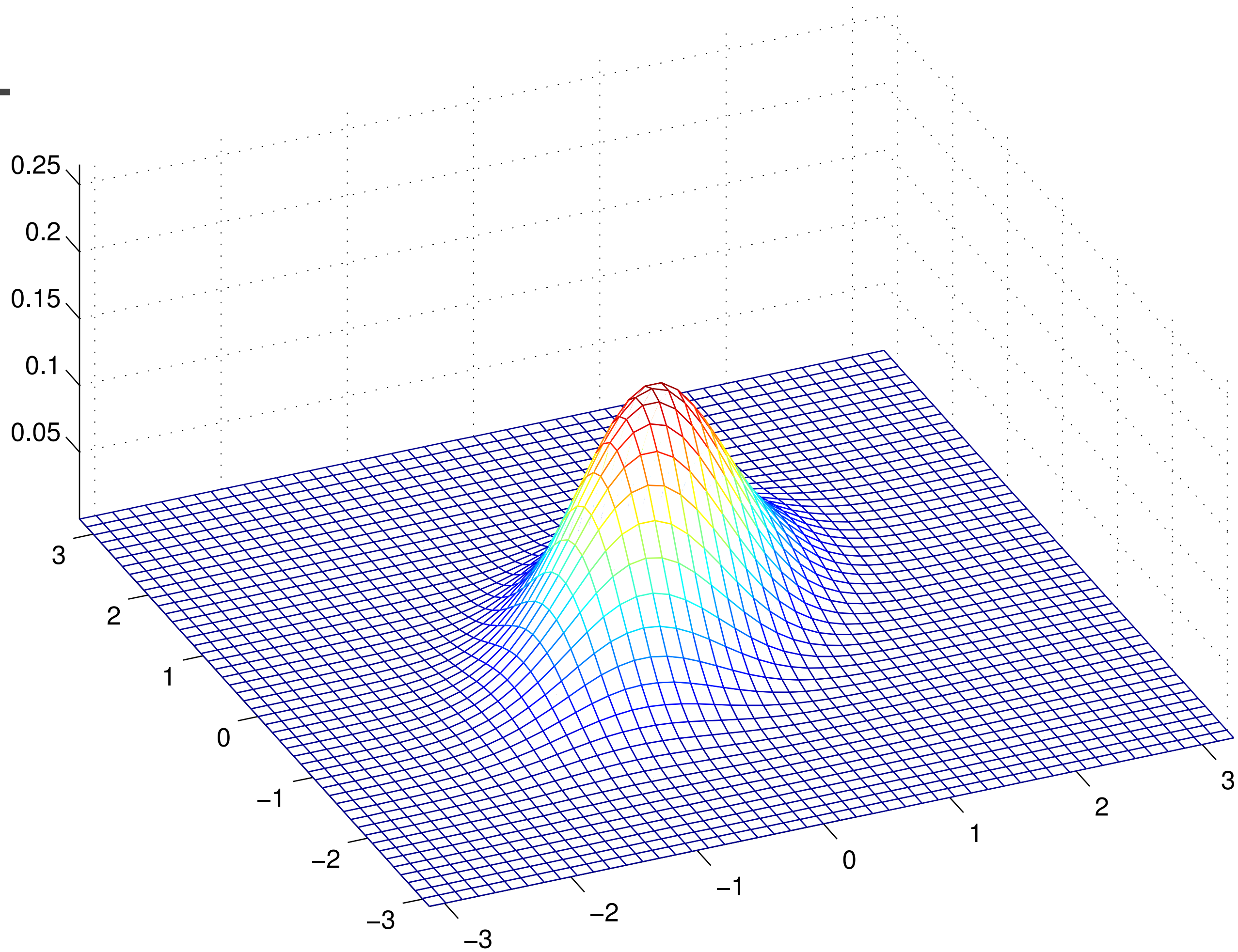
- $(\langle 3,-2\rangle - \langle 0,0\rangle)^\top I_2(\langle 3,-2\rangle - \langle 0,0\rangle)$

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

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

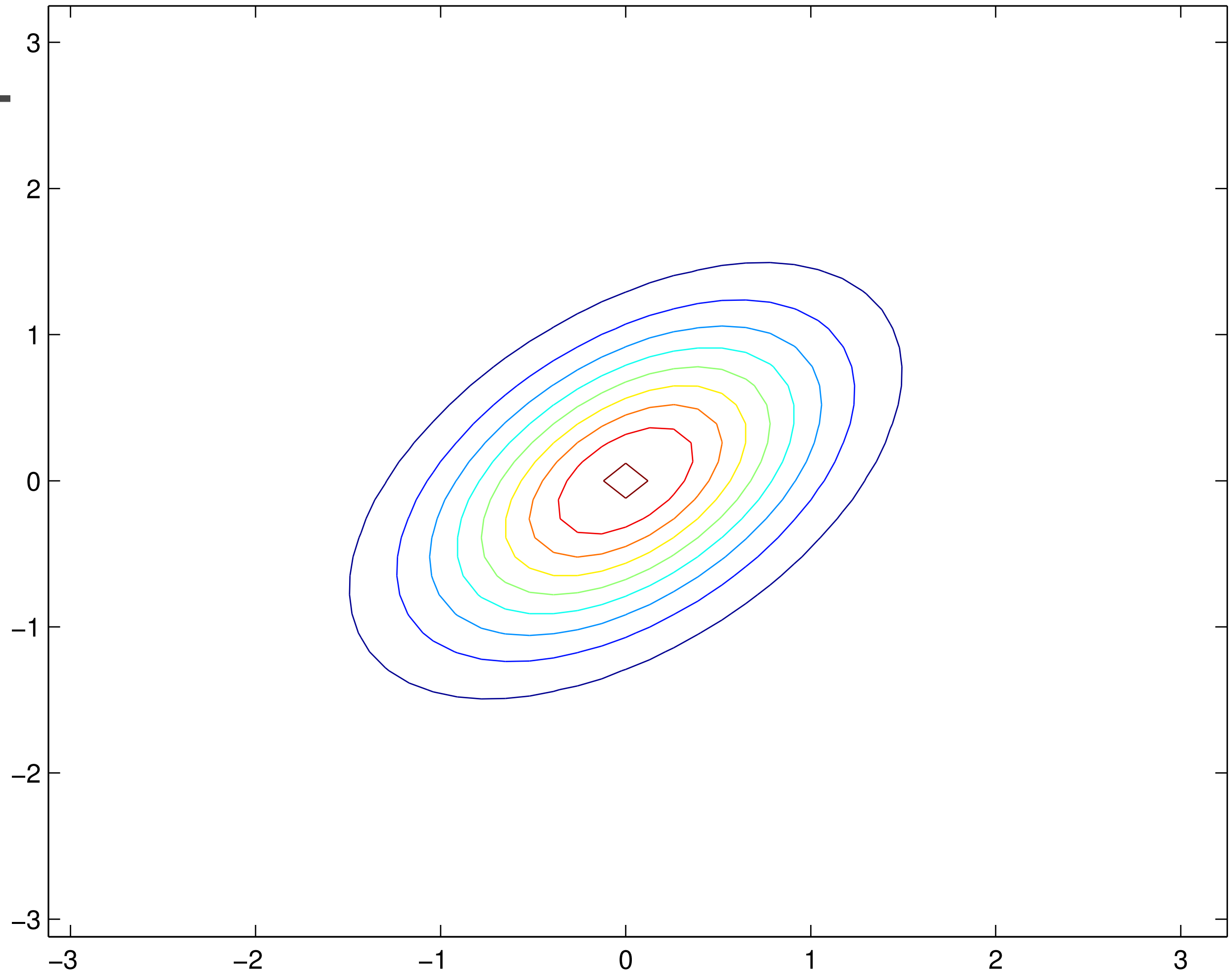
So $P(\,\langle -3,2\rangle \mid \ldots) = \frac{1}{(2\pi)} \exp\left[-\tfrac{1}{2}13\right] = \ldots$

MVG examples



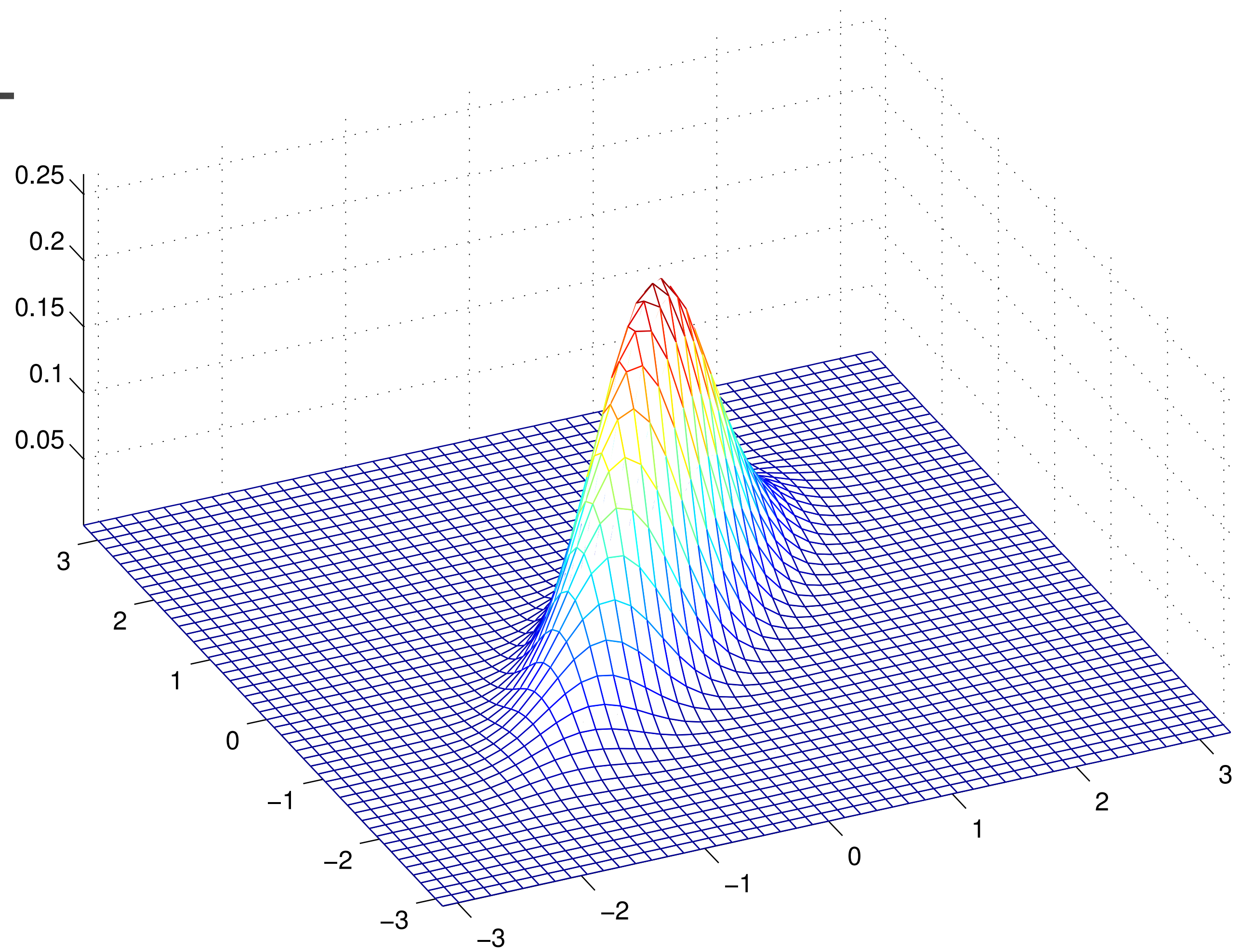
μ = (0,0) $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

MVG examples



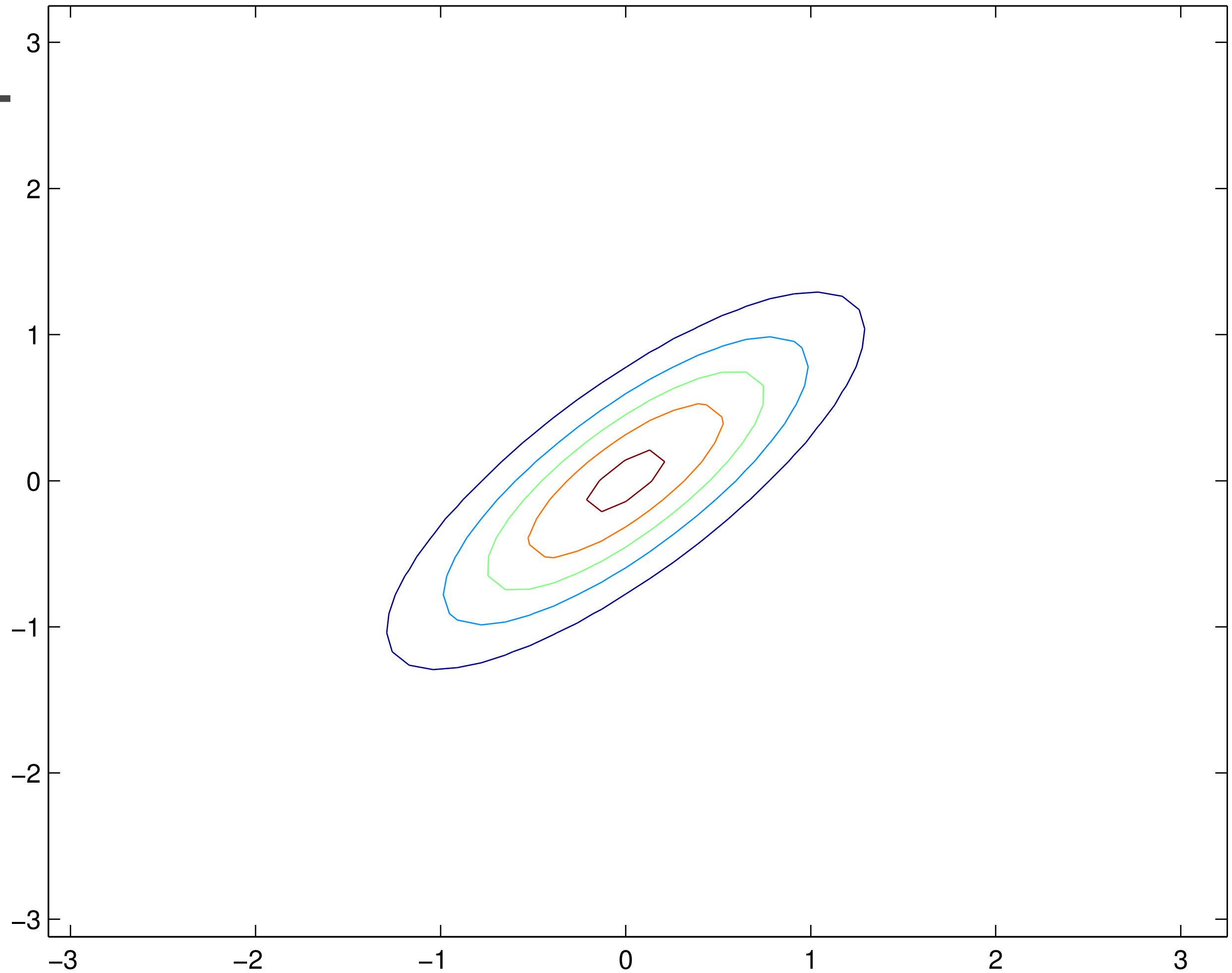
$\mu = (0,0)$ $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

MVG examples



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

MVG examples



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

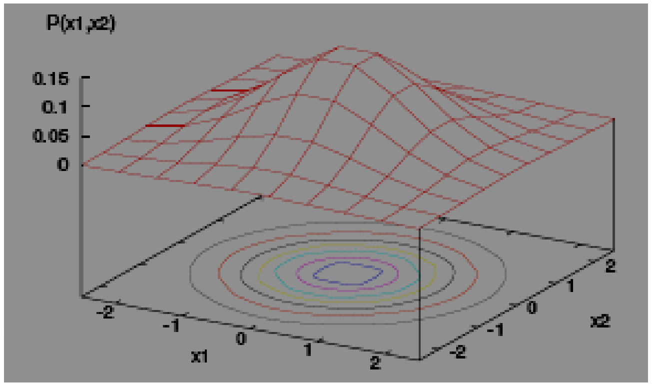
Independent Variables

- Variables independent $\equiv$
  Covariance matrix is Diagonal

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

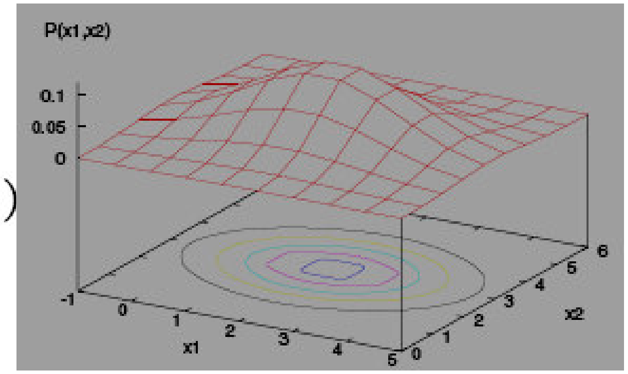
- $P(\langle x, y\rangle = \langle 3, -2\rangle \mid \langle x, y\rangle \sim \mathcal{N}(\langle 0,0\rangle, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}))$

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



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

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

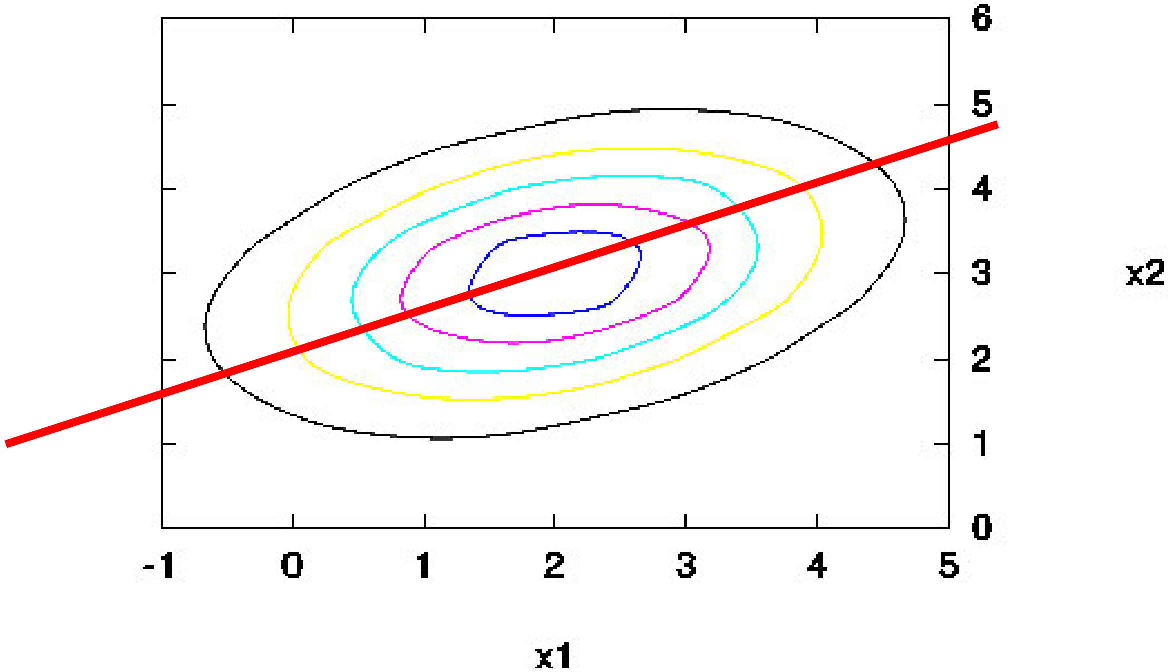


The Multivariate Gaussian: Ex 3

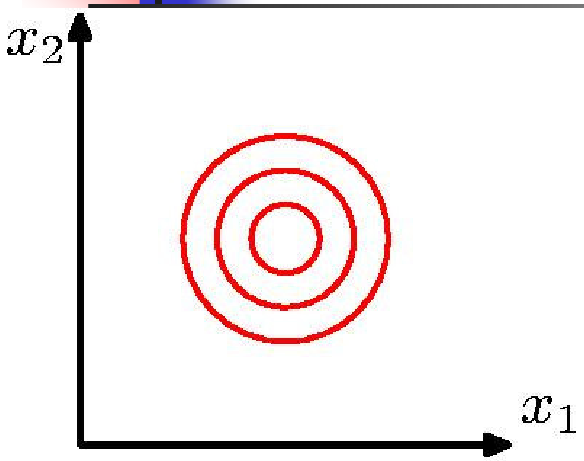
- If Σ is arbitrary,
  then x_1 and x_2 are dependent

Lines of equal probability are
"tilted" ellipses

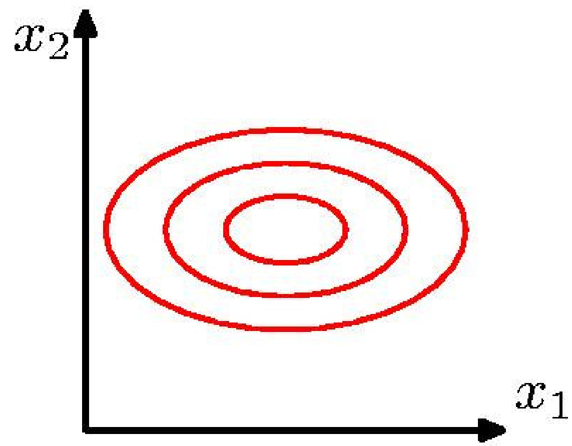
Eg For $\mu = \langle 2, 3 \rangle$ and $\Sigma = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}$:



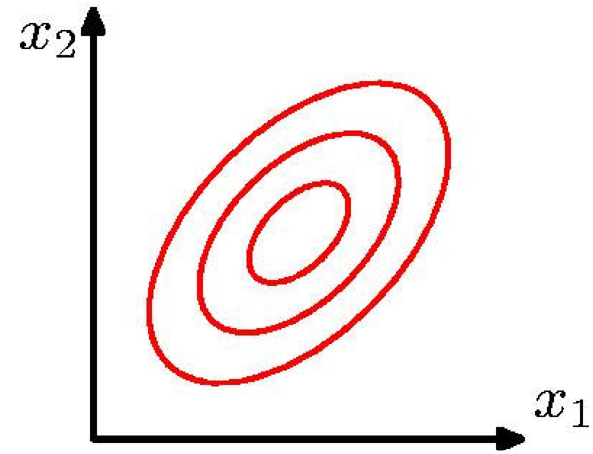
Examples of Gaussians



$\Sigma = \alpha I$
$= \text{Diag}(\alpha, \dots \alpha)$

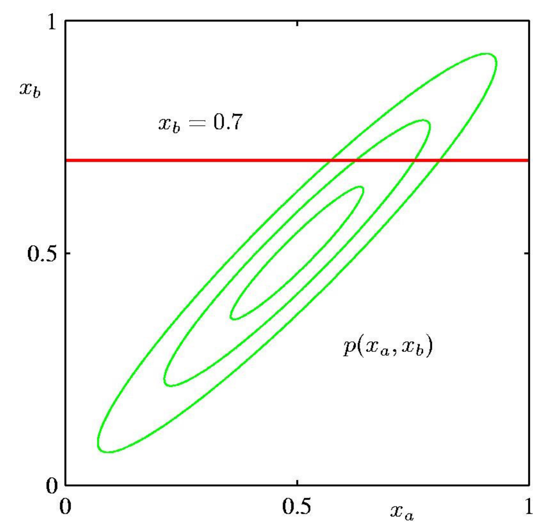


$\Sigma = \text{Diag}(\alpha_1, \dots \alpha_k)$



General Σ

Marginal…



Useful Properties of Gaussians I

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

- Every general Gaussian ≡
  a standard Gaussian that has undergone an affine transformation

Useful Properties of Gaussians II

- A Gaussian distribution is completely specific by
  - a vector of means
  - a covariance matrix
- Requires $O(n^2)$ space
- Requires $O(n^3)$ time to manipulate
- Bad 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:

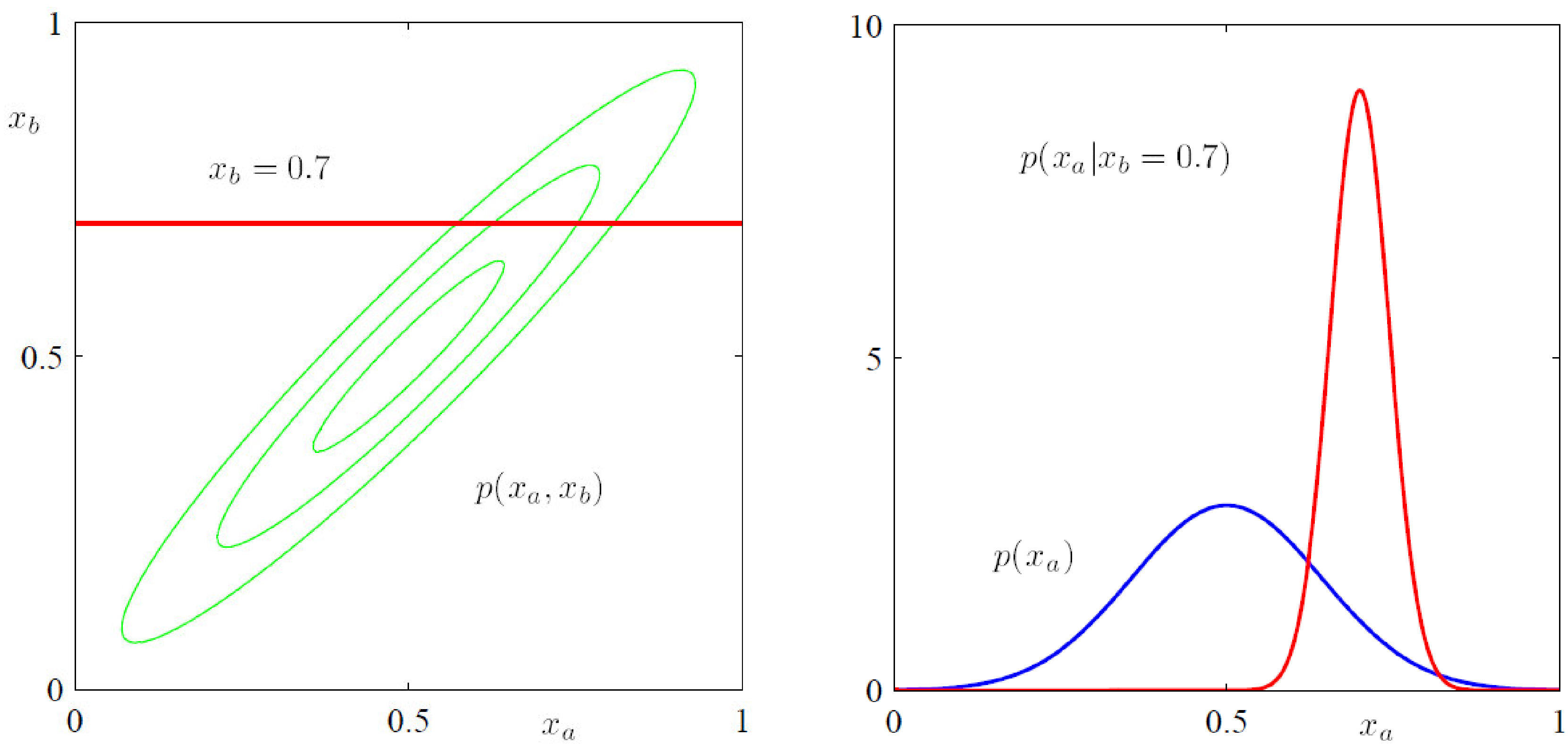
$$\Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

- Conditional Distribution:

$$p(x_a \mid x_b) = N(x_a \mid \mu_{a|b}, \Lambda_{aa}^{-1})$$

$$\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab} (x_b - \mu_a)$$

Visualizing Marginalization & Conditioning



Useful Properties of Gaussians V

- Affine transformations of Gaussian variables are Gaussian
  - Suppose x is Gaussian
  - y=Ax+b is Gaussian
- Uses:
  - Compute distribution on Y from distribution on x
  - Compute posterior on x after observing y

Useful Properties of Gaussians

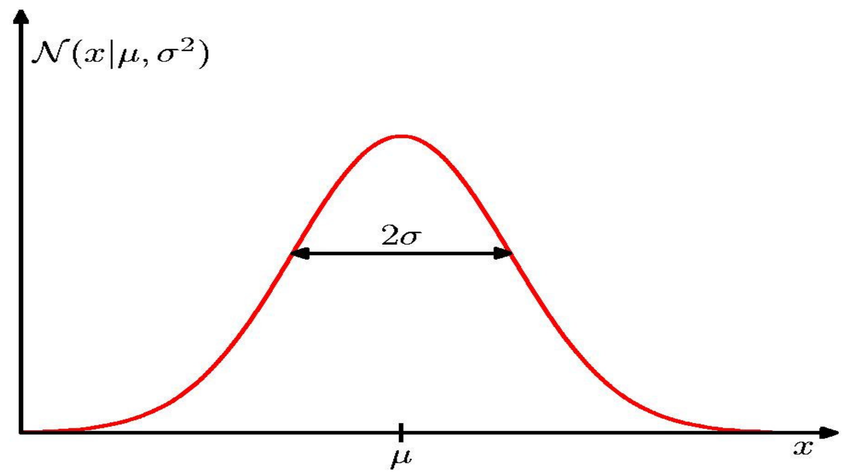
- Lots of things can (arguably) be approximated well by Gaussians
- Central Limit Theorem:
  The sum of IID variables with finite variances will tend towards a Gaussian distribution

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

Learning a Gaussian

- Collect a set of data, D
  of real-valued i.i.d. instances
  - e.g., exam scores

99
75
82
…
93
:



- 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

- 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

- 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

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

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

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

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

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:

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

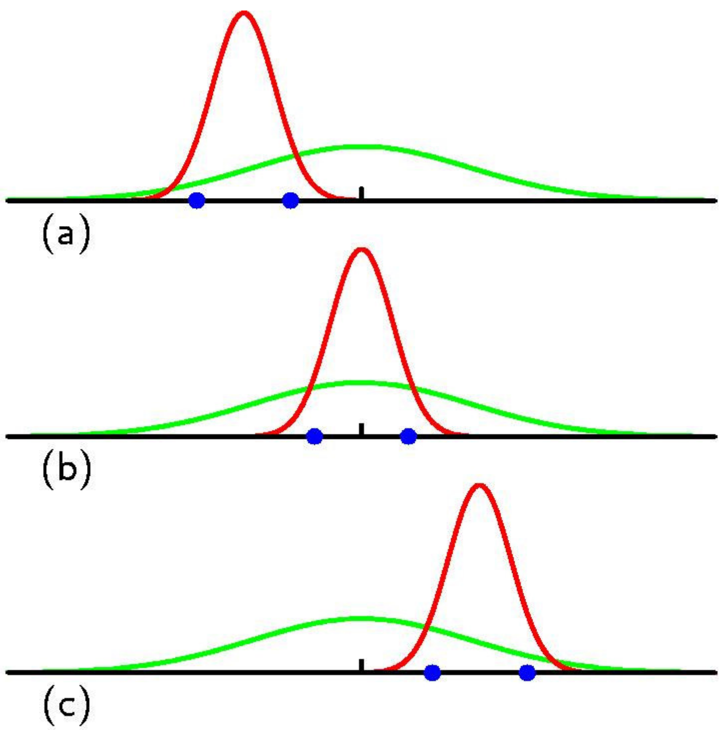
Homework#1 !!

Why is it Biased?

- Bias is wrt Mean; MLE is wrt Mode

  … Mean ≠ Mode

- Consider…



Estimating a Multivariate Gaussian

- Given data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, MLE is…

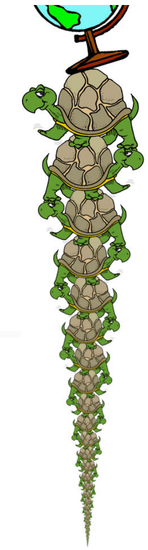
$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_i x_i$$

$$\hat{\Sigma}_{MLE} = \frac{1}{N}\sum_i (x_i - \hat{\mu}) \cdot (x_i - \hat{\mu})^T$$

- Recall…

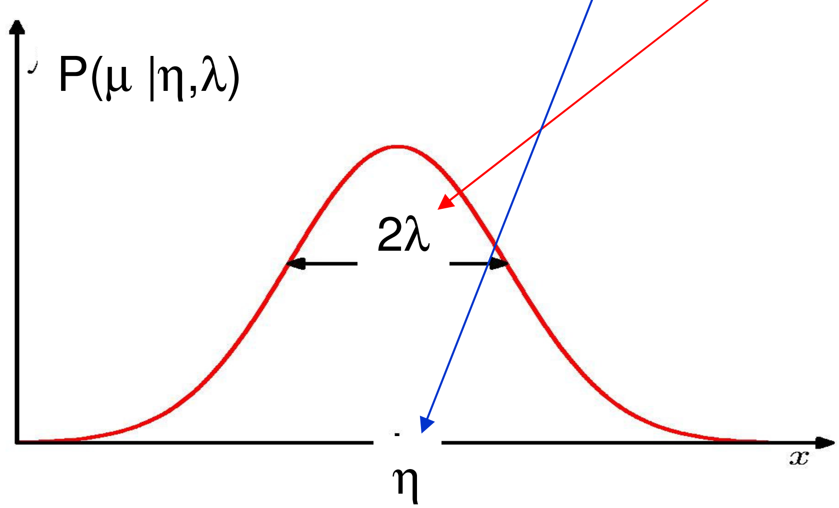
$$\mathbf{x} \cdot \mathbf{y}^\top = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \cdot [y_1\ y_2\ y_3] = \begin{bmatrix} x_1y_1 & x_1y_2 & x_1y_3 \\ x_2y_1 & x_2y_2 & x_2y_3 \\ x_3y_1 & x_3y_2 & x_3y_3 \end{bmatrix}$$

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

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

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

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

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

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

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, $\lambda^2 \to \infty$

  $\Rightarrow$ 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

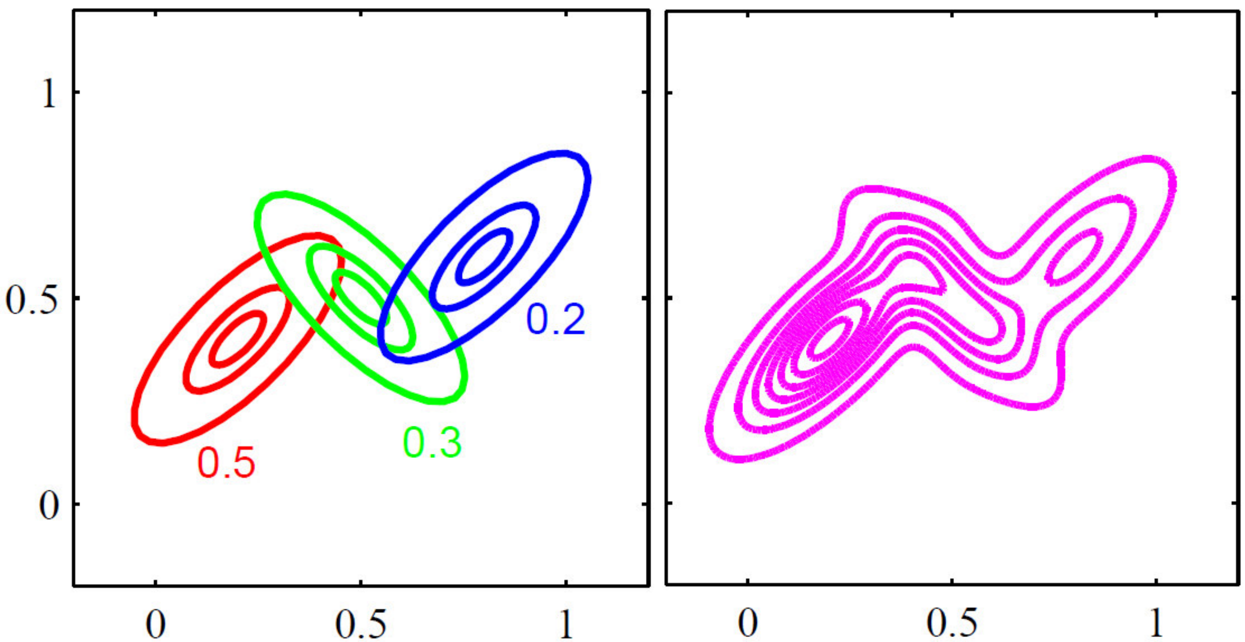
- Want to approximate distribution that is not unimodal?
- Density is weighted combination of Gaussians

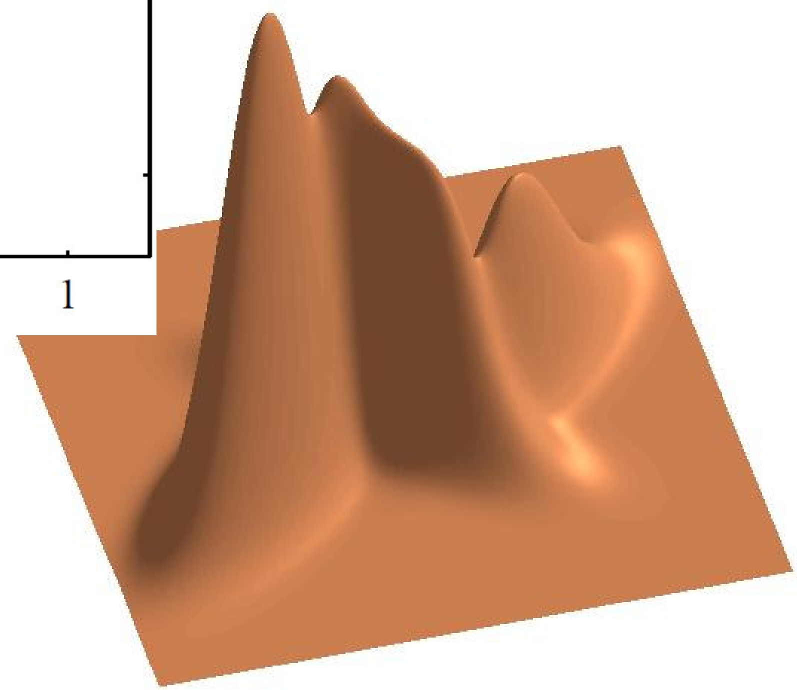
$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$

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

- Idea: Flip coin (roll dice) to select Gaussian, then sample from the Gaussian
- Can be arbitrarily expressive with enough Gaussians

Mixture of Gaussians Example





What you need to know

- Probability 101
- Point Estimation
  - MLE
  - Hoeffding inequality (PAC)
  - Bayesian learning
    - Beta, Dirichlet distributions
    - Gaussian, …
  - MAP

Factoids…

- $\ln a^b = b \ln a$
- $\ln (a * b) = \ln a + \ln b$

$$\frac{\partial}{\partial \theta} \ln \theta = \frac{1}{\theta}$$

$$\frac{\partial}{\partial \theta} \ln (1-\theta) = \frac{-1}{(1-\theta)}$$

Basic concepts for random variables

- Atomic outcome: assignment $x_1,\ldots,x_n$ to $X_1,\ldots,X_n$
- Conditional probability: $P(X,Y) = P(X)\,P(Y|X)$
- Bayes rule: $P(X|Y) = P(Y|X)\,P(X) / P(Y)$
- Chain rule:
  $P(X_1,\ldots,X_n) =$
  $P(X_1)\,P(X_2|X_1)\ldots\,P(X_k|X_1,\ldots,X_{k-1})\ \ldots\ P(X_n|X_1,\ldots,X_{n-1})$

Chebyshev's Inequality

- X with finite mean, variance

$$P(|X - E(X)| \geq c) \leq \frac{Var(X)}{c^2}$$

- Variance governs chance of missing mean

Convergence of Sample Mean

- Apply Chebyshev's Inequality to sample mean:

$$P(|X - E(X)| \geq c) \leq \frac{Var(X)}{c^2}$$

$$Var(\overline{X}) = Var\left(\sum_i \frac{X_i}{n}\right) = \sum_i \frac{1}{n^2} Var(X_i) = \frac{Var(X)}{n}$$

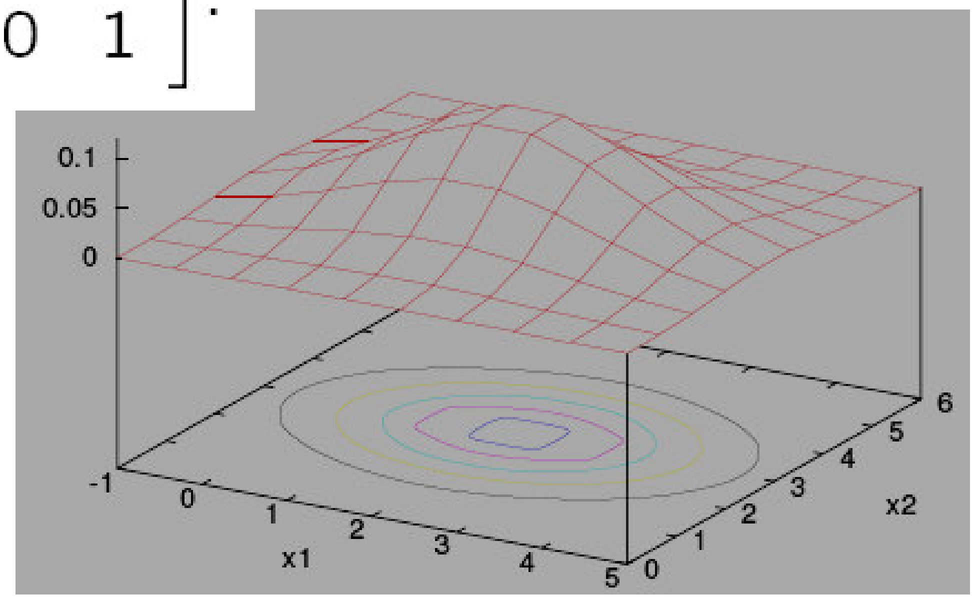
$$\lim_{n\to\infty} P(|X - E(X)| \geq c) \ \leq\ \lim_{n\to\infty} \frac{Var(X)}{nc^2} \ =\ 0$$

Random Variable

- Events are complicated – we think about attributes
  - Age, Grade, HairColor
- Random variables formalize attributes:
  - Grade=A shorthand for event $\{\omega \in \Omega: f_{Grade}(\omega) = A\}$
- Properties of random vars, X:
  - Val(X) = possible values of random var X
  - For discrete (categorical): $\sum_{i=1 \ldots |Val(X)|} P(X=x_i) = 1$
  - For continuous: $\int_x p(X=x)\, dx = 1$

The Multivariate Gaussian: Ex 2

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



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

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

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

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