**Piazza** started

# Announcements

**Matlab Grader homework,** email Friday,
2 (of 9) homeworks Due 21 April, Binary graded.

**Jupyter homework?:** translate matlab to Jupiter, TA Harshul h6gupta@eng.ucsd.edu or me
I would like this to happen.

“GPU” homework. NOAA climate data in Jupyter on the datahub.ucsd.edu, 15 April.

Projects: Any language

**Podcast** might work eventually.

**Today:**

- Stanford CNN
- Bernoulli
- Gaussian 1.2
- Gaussian 2.3
- Decision theory 1.5
- Information theory 1.6

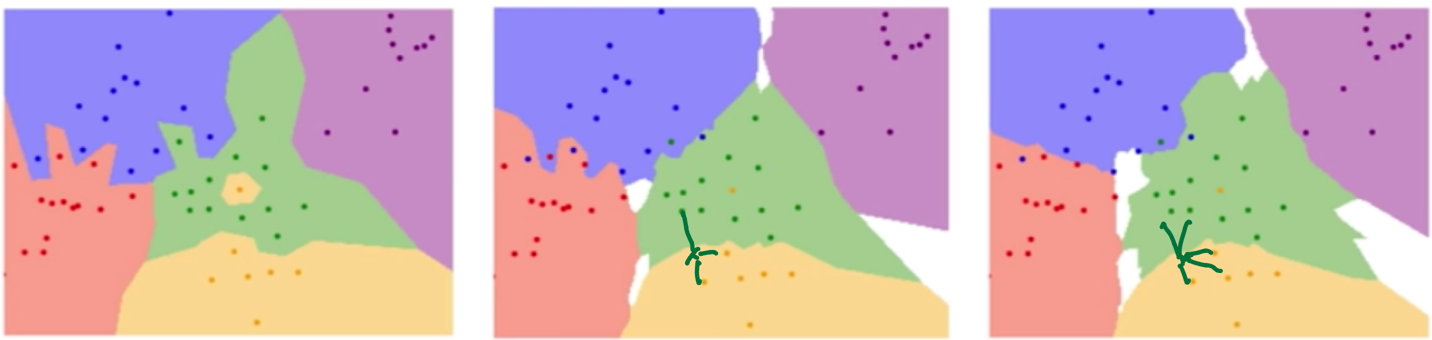
Monday
Stanford CNN, Linear models for regression 3

# Non-parametric method

K- means

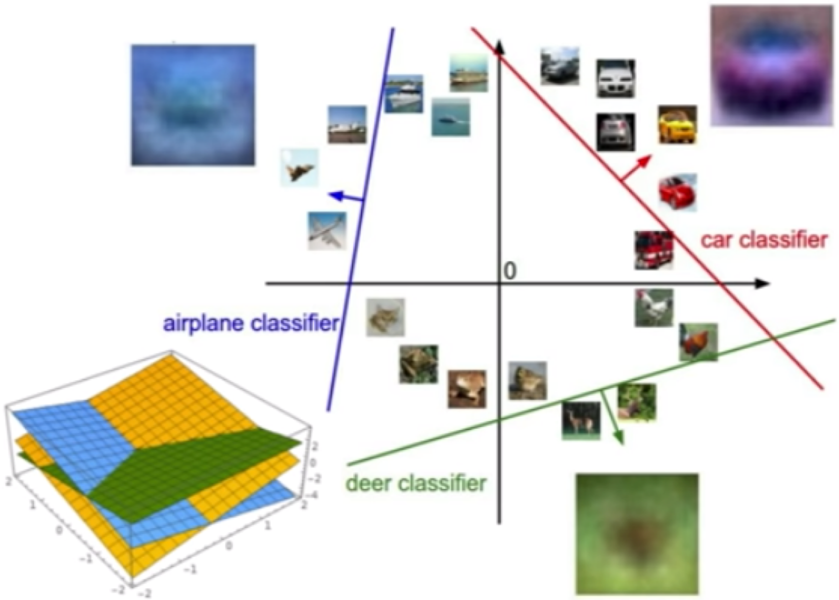
## K-Nearest Neighbors

Instead of copying label from nearest neighbor,
take **majority vote** from K closest points



K = 1 K = 3 K = 5

# Interpreting a Linear Classifier




$$f(x,W) = Wx + b$$

Array of **32x32x3** numbers
(3072 numbers total)

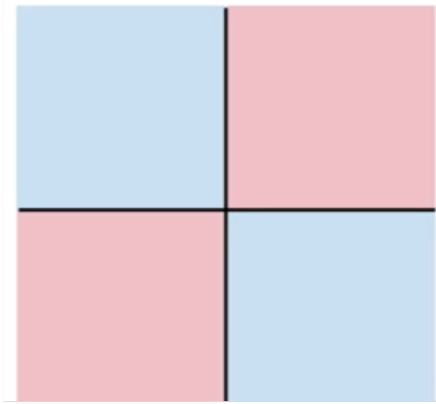
Plot created using Wolfram Cloud

CC-BY

# Hard cases for a linear classifier

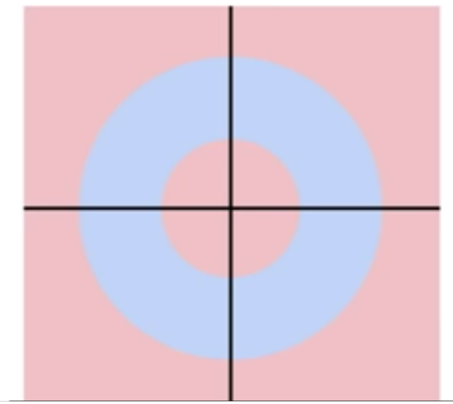
**Class 1**:
number of pixels > 0 odd

**Class 2**:
number of pixels > 0 even



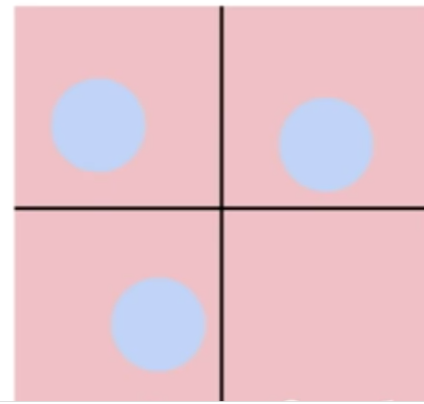
**Class 1**:
1 <= L2 norm <= 2

**Class 2**:
Everything else



**Class 1**:
Three modes

**Class 2**:
Everything else



# Coin estimate (Bishop 2.1)

- Binary variables x={0,1}

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

- Bernoulli distributed

$$\mathrm{Bern}(x|\mu)=\mu^x(1-\mu)^{1-x} \tag{2.2}$$

$$\mathbb{E}[x] = \mu$$
$$\mathrm{var}[x] = \mu(1-\mu).$$

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

$p(x=0)=p(T)=1-\mu$

$E(x)=\int x\, p(x)\, dx$

$=1\cdot\mu+0\cdot(1-\mu=\mu$

$Var(x)=E[(x-\bar{x})^2]$

$=(1-\mu)^2\cdot\mu+(0-\mu)(1-\mu)$

- N observations, Likelihood:

$$p(\mathcal{D}|\mu)=\prod_{n=1}^{N}p(x_n|\mu)=\prod_{n=1}^{N}\mu^{x_n}(1-\mu)^{1-x_n}. \tag{2.5}$$

$$\ell = \ln p(\mathcal{D}|\mu)=\sum_{n=1}^{N}\ln p(x_n|\mu)=\sum_{n=1}^{N}\{x_n\ln\mu+(1-x_n)\ln(1-\mu)\}. \tag{2.6}$$

- Max likelihood

$$\frac{\partial \ell}{\partial \mu}=\sum\frac{x_n}{\mu}+\frac{1-x_n}{1-\mu}$$

$$\mu_{\mathrm{ML}}=\frac{1}{N}\sum_{n=1}^{N}x_n$$

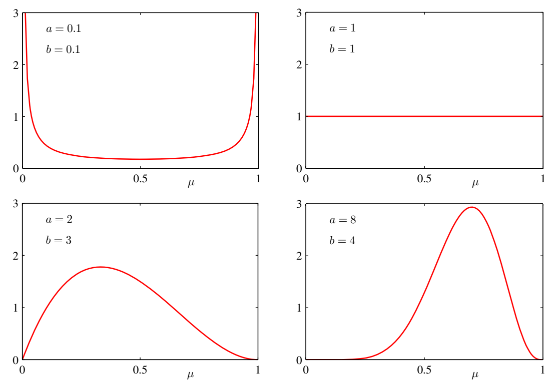
# Coin estimate (Bishop 2.1)

post. like prior

- Bayes p(x|y)=p(y|x)p(x)

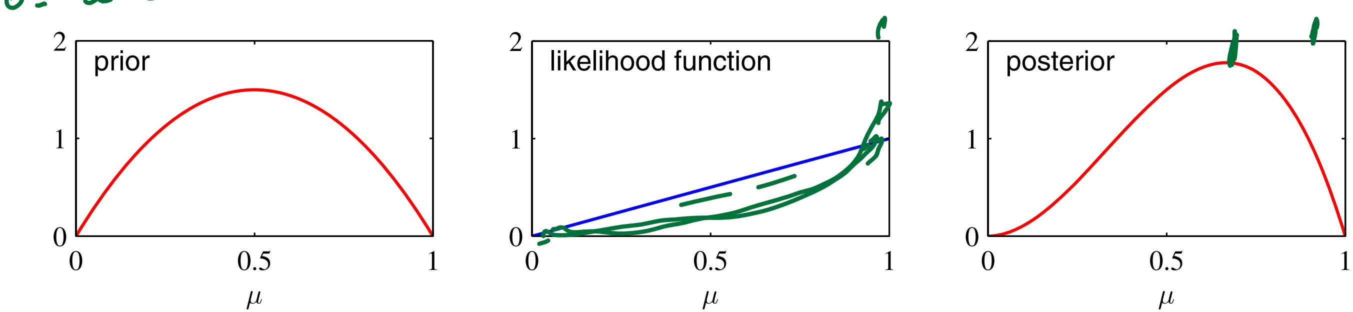
- Conjugate prior

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




Bayes:

b= a=2

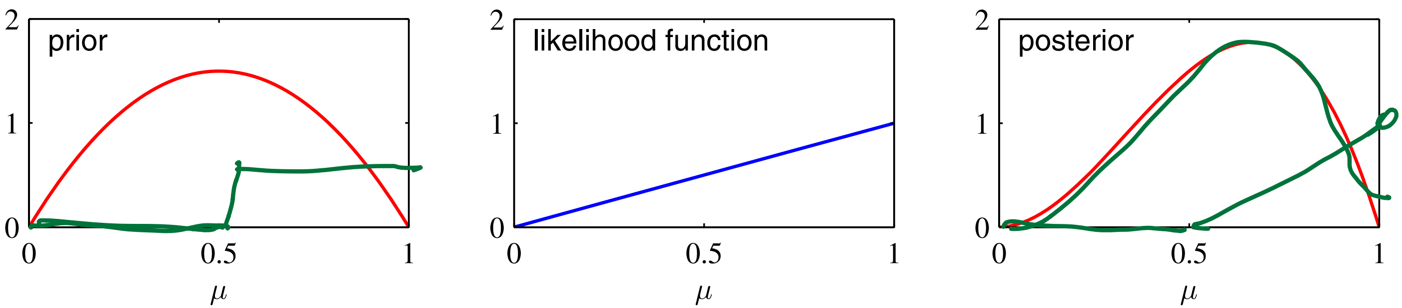


# ML MAP BAYES

- ML point estimate

- MAP point estimate (often in literature ML=MAP)

- Bayes => probability =>From which all information can be obtained
  - MAP, median, error estimates
  - Further analysis as sequential
  - Disadvantage… not a point estimate.



# Bayes Rule

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})}$$

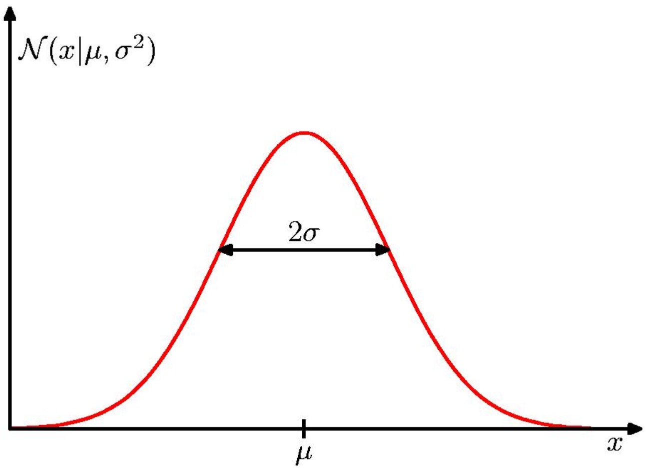
Rev'd Thomas Bayes (1702–1761)

- Bayes rule tells us how to do inference about hypotheses from data.

- Learning and prediction can be seen as forms of inference.

# The Gaussian Distribution

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




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

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

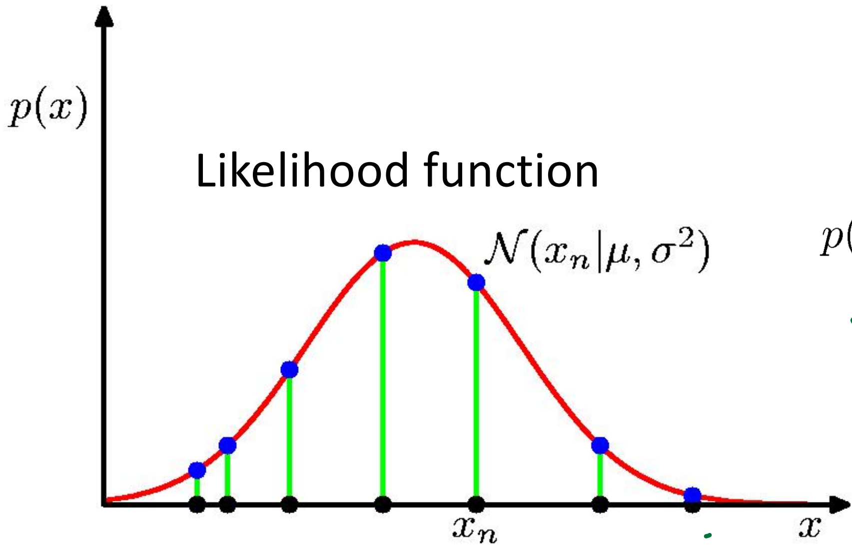
## Gaussian Mean and Variance

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

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

# Gaussian Parameter Estimation




$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

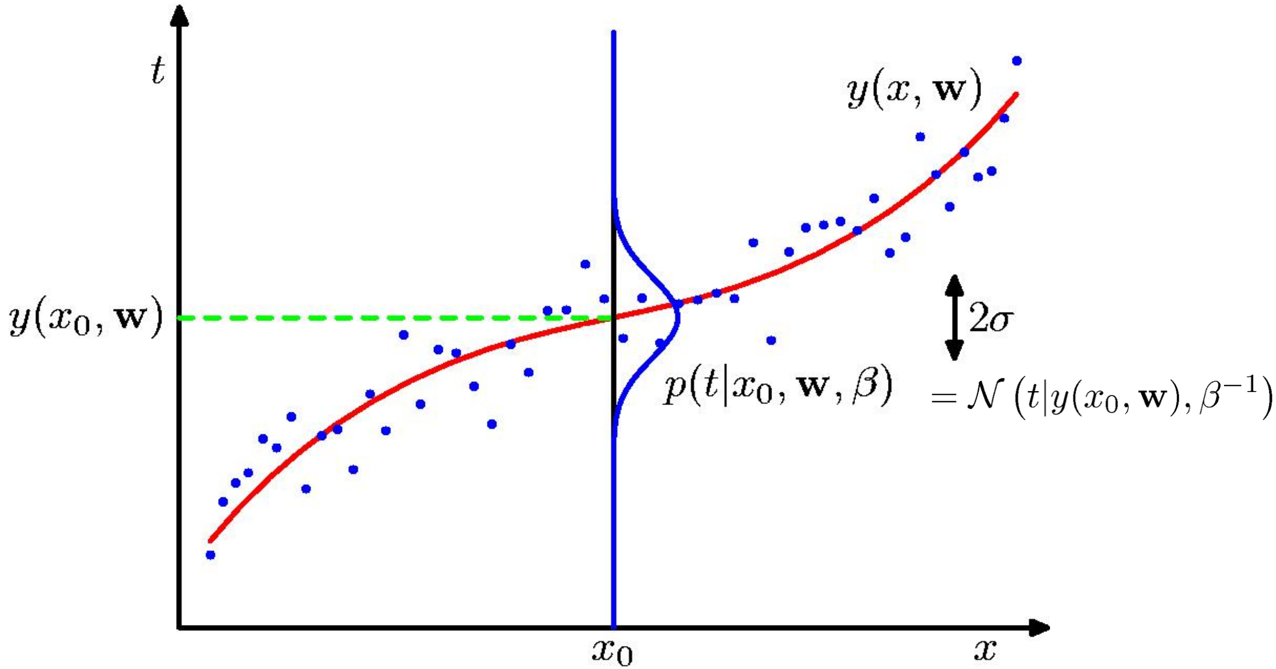
$$\frac{\partial l}{\partial \mu} = \frac{1}{2\sigma^2}\sum_{n}^{N}(x_n - \mu) = 0$$

## Maximum (Log) Likelihood

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

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2$$

# Curve Fitting Re-visited, Bishop1.2.5



# Maximum Likelihood

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \beta^{-1}\right). \tag{1.61}$$

As we did in the case of the simple Gaussian distribution earlier, it is convenient to maximize the logarithm of the likelihood function. Substituting for the form of the Gaussian distribution, given by (1.46), we obtain the log likelihood function in the form

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi). \tag{1.62}$$

$\sigma^2 =$

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2. \tag{1.63}$$

Giving estimates of W and beta, we can predict

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right). \tag{1.64}$$

# MAP: A Step towards Bayes 1.2.5

prior

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

$$\mathcal{N}(w^T x, \beta^{-1})\,\mathcal{N}(0, \alpha)$$

$$-\ln(p(w|t)) = \frac{\beta}{2}\sum_{n}^{N}(w^T x_n - t_n)^2 + \frac{\alpha}{2} w^T w + \text{cost}$$

$\|w\|_1$

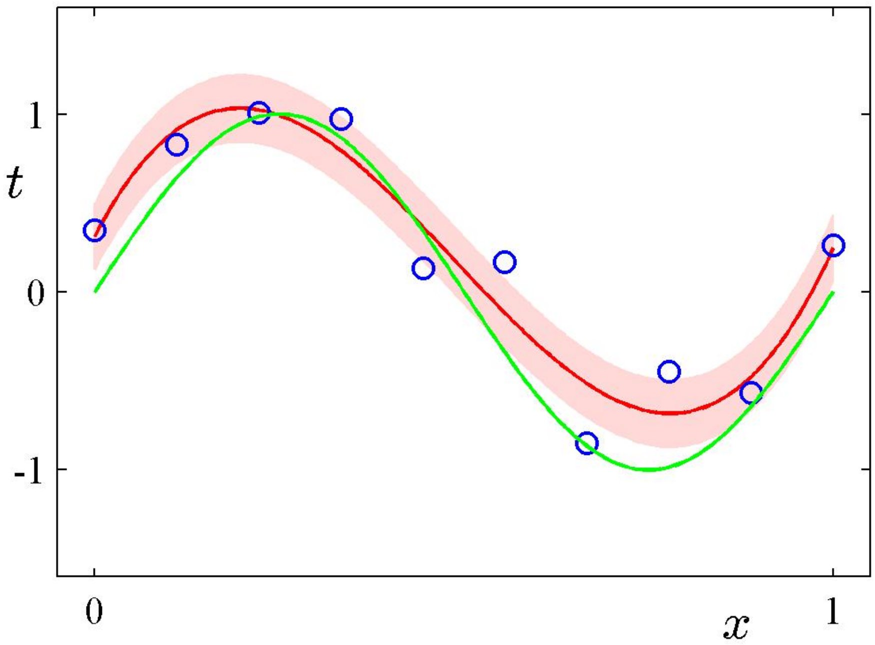
$$\rightarrow \quad \beta\widetilde{E}(\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

Determine $\mathbf{w}_{\mathrm{MAP}}$ by minimizing regularized sum-of-squares error, $\widetilde{E}(\mathbf{w})$.

Regularized sum of squares $\Leftarrow$

# Predictive Distribution

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right)$$




True data

Estimated
+/- std dev

# Parametric Distributions

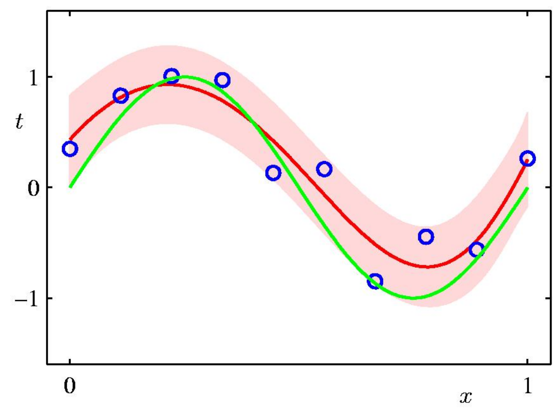
Basic building blocks: $p(\mathbf{x}|\boldsymbol{\theta})$

Need to determine $\boldsymbol{\theta}$ given $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

Representation: $\boldsymbol{\theta}^\star$ or $p(\boldsymbol{\theta})$?

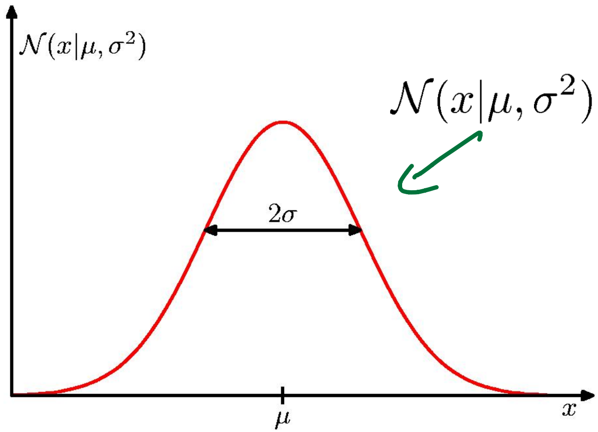
Recall Curve Fitting

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})\, \mathrm{d}\mathbf{w}$$

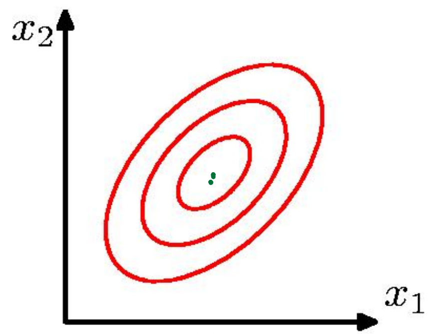



**We focus on Gaussians!**

# The Gaussian Distribution




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

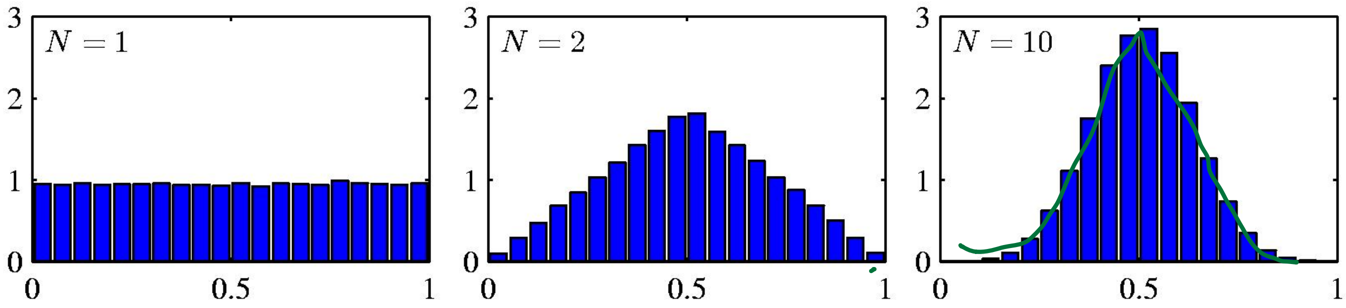



$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

# Central Limit Theorem

•The distribution of the sum of N i.i.d. random variables becomes increasingly Gaussian as N grows.

•Example: N uniform [0,1] random variables.



# Geometry of the Multivariate Gaussian

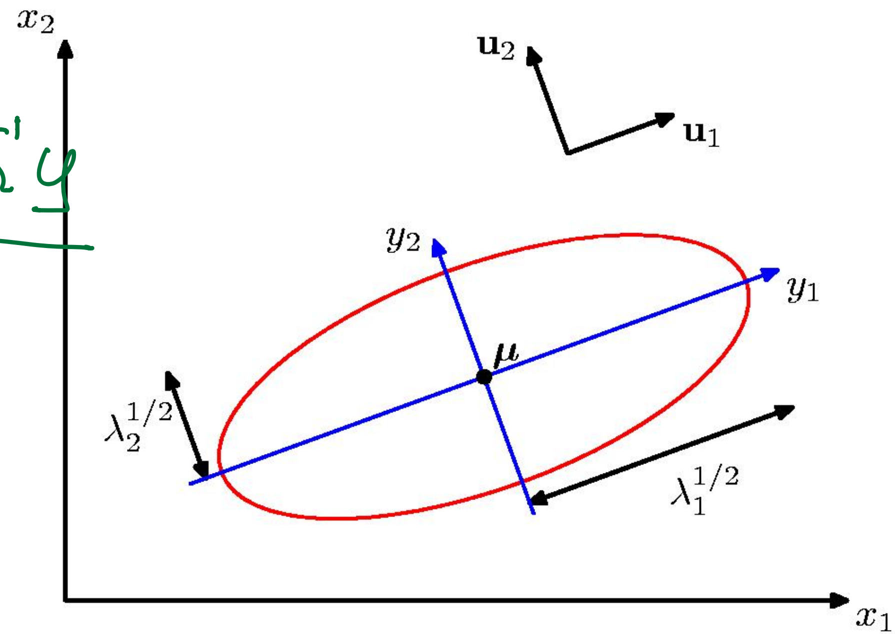
$e^{-(x-\mu)^T \Sigma^{-1} x-\mu}$

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})$$

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}$$

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} = y^T \Lambda^{-1} y$$

$$y_i = \mathbf{u}_i^{\mathrm{T}} (\mathbf{x}-\boldsymbol{\mu})$$



# Moments of the Multivariate Gaussian (2)

$E(x) = \mu$

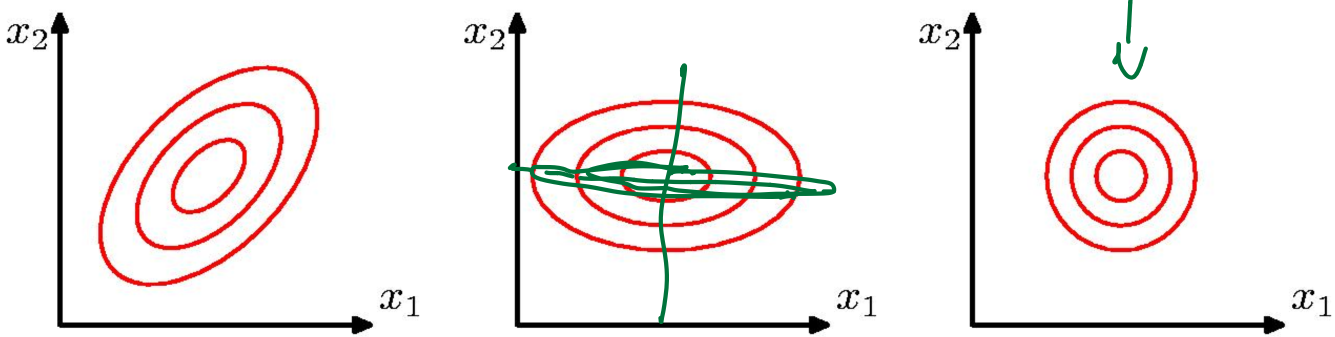
$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] = \underline{\boldsymbol{\Sigma}} = \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots \\ & & & \sigma_N^2 \end{bmatrix}$$

$= \sigma^2 I$

A Gaussian requires D*(D-1)/2 +D parameters.
Often we use D +D or
Just D+1 parameters.

$\sigma_2 = 0$



(a) (b) (c)

# Partitioned Conditionals and Marginals, page 89

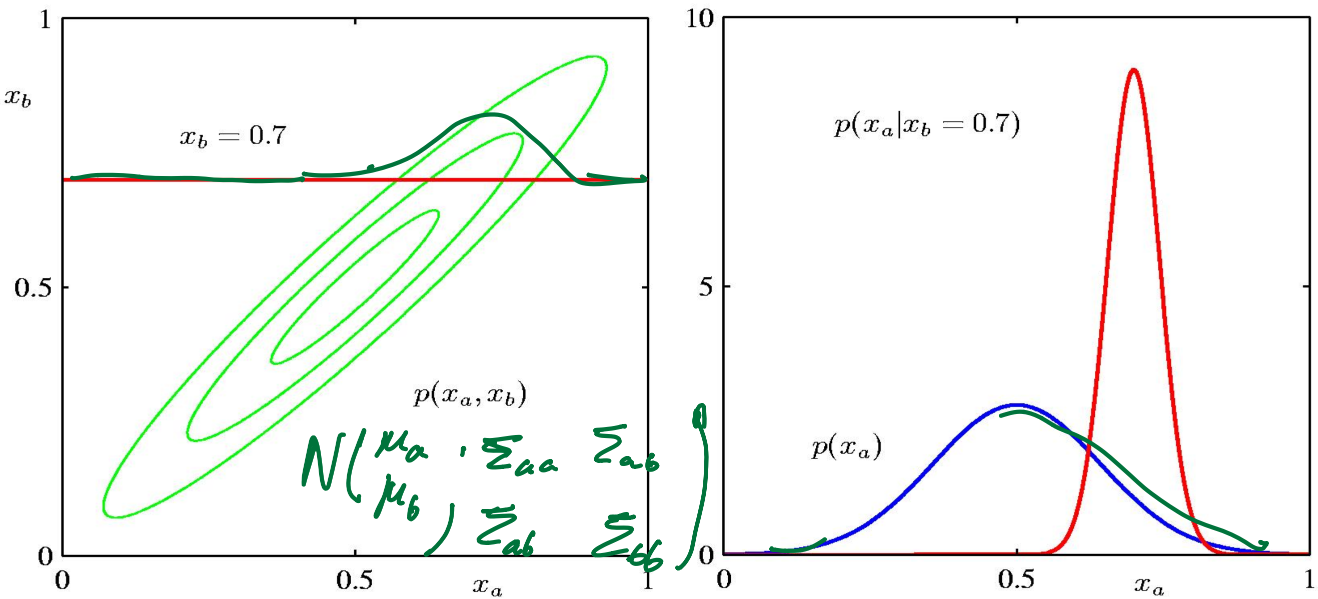
Conditional

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

$$\begin{aligned}\boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)\end{aligned}$$

marginal

$$\begin{aligned}\underline{p(\mathbf{x}_a)} &= \int p(\mathbf{x}_a, \mathbf{x}_b)\,\mathrm{d}\mathbf{x}_b \\ &= \underline{\mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})}\end{aligned}$$



# ML for the Gaussian (1) Bisphop 2.3.4

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

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

$$\ell = -\ln p$$

$$= \frac{N}{2}\ln|\Sigma| + \frac{1}{2}\mathrm{Tr}\left(\sum_n (x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$$= \frac{N}{2}\ln|\Sigma| + \frac{1}{2}\mathrm{tr}\left(\sum_n (x-\mu)(x-\mu)^T \Sigma^{-1}\right)$$

$$= \frac{N}{2}\left[\ln|\Sigma| + \mathrm{tr}(S_y \Sigma^{-1})\right]$$

$$S_y = \frac{1}{N}\sum (x-\mu)(x-\mu)^T$$

$$\frac{\partial \ell}{\partial \mu}$$

$$\mathrm{tr}(ABC) = \mathrm{tr}(CAB)$$

$$\frac{\partial \ell}{\partial \Sigma} = \Sigma^{-1} + S_y \Sigma^{-1} \Sigma^{-1} = 0$$

$$\Sigma = S_\mu$$

$$\frac{\partial}{\partial \mathbf{A}} \ln|\mathbf{A}| = \left(\mathbf{A}^{-1}\right)^{\mathrm{T}} \tag{C.28}$$

$$\frac{\partial}{\partial \mathbf{A}} \mathrm{Tr}(\mathbf{AB}) = \mathbf{B}^{\mathrm{T}}. \tag{C.24}$$

$$\frac{\partial}{\partial x}\left(\mathbf{A}^{-1}\right) = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1} \tag{C.21}$$

# Maximum Likelihood for the Gaussian

- Set the derivative 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

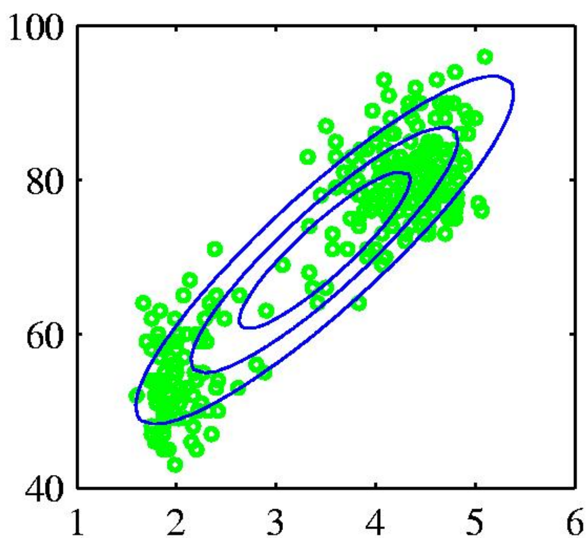
$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n.$$

- Similarly

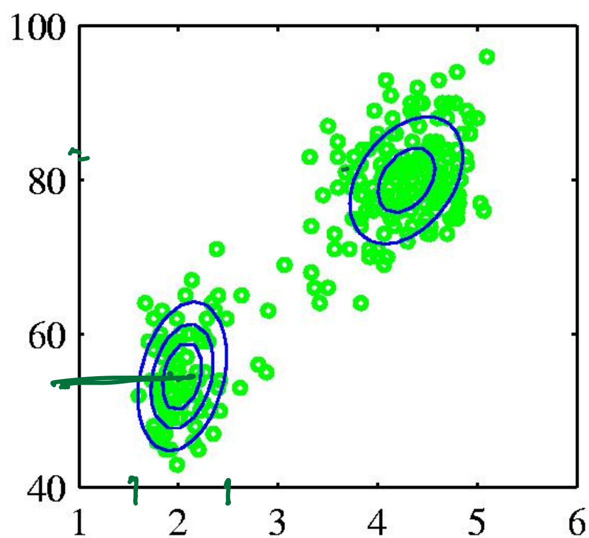
$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

# Mixtures of Gaussians (Bishop 2.3.9)

**Old Faithful geyser:**
The time between eruptions has a bimodal distribution, with the mean interval being either 65 or 91 minutes, and is dependent on the length of the prior eruption. Within a margin of error of ±10 minutes, Old Faithful will erupt either 65 minutes after an eruption lasting less than $2\,{}^{1}\!/_{2}$ minutes, or 91 minutes after an eruption lasting more than $2\,{}^{1}\!/_{2}$ minutes.




Single Gaussian



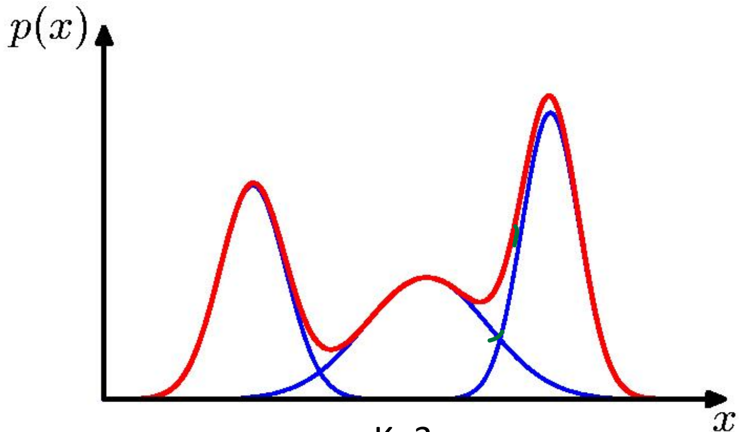

Mixture of two Gaussians

# Mixtures of Gaussians (Bishop 2.3.9)

- Combine simple models into a complex model:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \underbrace{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Component}}$$

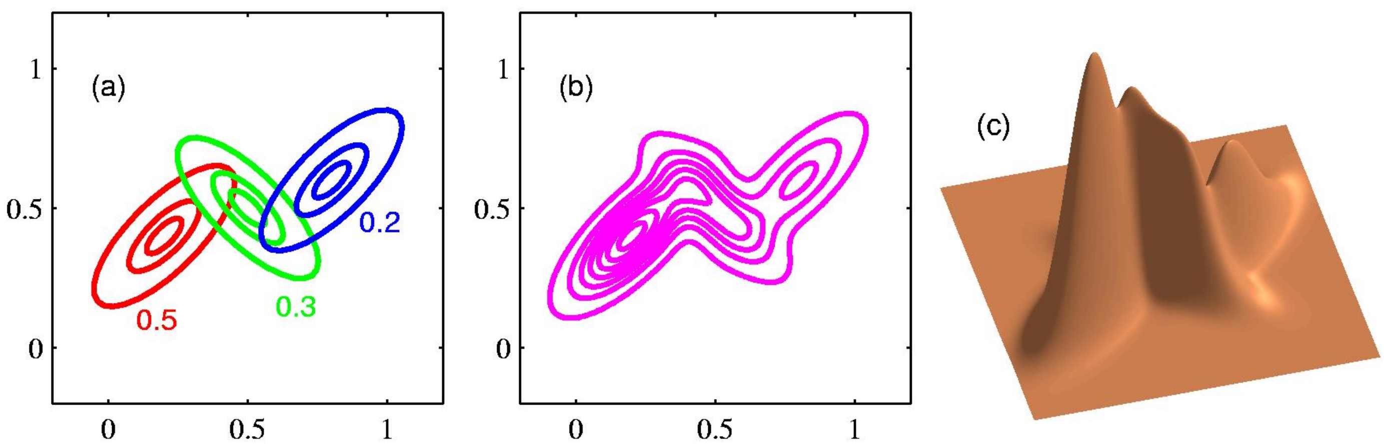
Mixing coefficient




K=3

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

# Mixtures of Gaussians (Bishop 2.3.9)



# Mixtures of Gaussians (Bishop 2.3.9)

- Determining parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ using maximum log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \underbrace{\ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}}_{\text{Log of a sum; no closed form maximum.}}$$

- Solution: use standard, iterative, numeric optimization methods or the *expectation maximization* algorithm (Chapter 9).

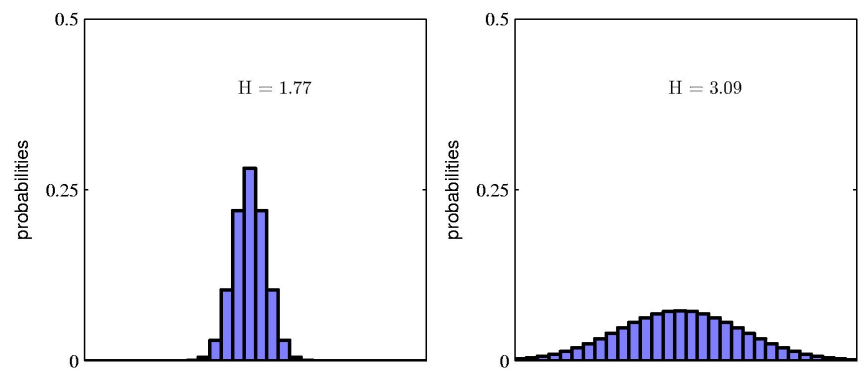
EM

# Entropy 1.6

$$\mathrm{H}[x] = -\sum_{x} p(x) \log_2 p(x)$$

Important quantity in

- coding theory
- statistical physics
- machine learning



# Differential Entropy

Put bins of width ¢ along the real line

$$\lim_{\Delta \to 0} \left\{ - \sum_i p(x_i) \Delta \ln p(x_i) \right\} = - \int p(x) \ln p(x) \, \mathrm{d}x$$

For fixed $\sigma^2$ differential entropy maximized when

$$p(x) = \mathcal{N}(x|\mu, \sigma^2)$$

in which case

$$\mathrm{H}[x] = \frac{1}{2} \left\{ 1 + \ln(2\pi\sigma^2) \right\}.$$

# The Kullback-Leibler Divergence

**P true distribution, q is approximating distribution**

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} \mathrm{d}\mathbf{x}\end{aligned}$$

$$\mathrm{KL}(p\|q) \simeq \frac{1}{N}\sum_{n=1}^{N}\{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}$$

$$\mathrm{KL}(p\|q) \geqslant 0 \qquad \mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p)$$