

REVIEW:  
PROBABILITY AND STATISTICS

Sam Roweis

October 2, 2002

---

RANDOM VARIABLES AND DENSITIES

- Random variables  $X$  represents outcomes or states of world.  
Instantiations of variables usually in lower case:  $x$   
We will write  $p(x)$  to mean probability( $X = x$ ).
- Sample Space: the space of all possible outcomes/states.  
(May be discrete or continuous or mixed.)
- Probability mass (density) function  $p(x) \geq 0$   
Assigns a non-negative number to each point in sample space.  
Sums (integrates) to unity:  $\sum_x p(x) = 1$  or  $\int_x p(x)dx = 1$ .  
Intuitively: how often does  $x$  occur, how much do we believe in  $x$ .
- Ensemble: random variable + sample space+ probability function

---

PROBABILITY

- We use probabilities  $p(x)$  to represent our beliefs  $B(x)$  about the states  $x$  of the world.
- There is a formal calculus for manipulating uncertainties represented by probabilities.
- Any consistent set of beliefs obeying the *Cox Axioms* can be mapped into probabilities.
  1. Rationally ordered degrees of belief:  
if  $B(x) > B(y)$  and  $B(y) > B(z)$  then  $B(x) > B(z)$
  2. Belief in  $x$  and its negation  $\bar{x}$  are related:  $B(x) = f[B(\bar{x})]$
  3. Belief in conjunction depends only on conditionals:  
 $B(x \text{ and } y) = g[B(x), B(y|x)] = g[B(y), B(x|y)]$

---

EXPECTATIONS, MOMENTS

- Expectation of a function  $a(x)$  is written  $E[a]$  or  $\langle a \rangle$

$$E[a] = \langle a \rangle = \sum_x p(x)a(x)$$

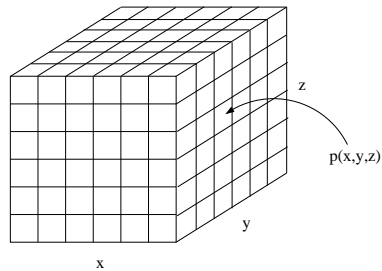
e.g. mean =  $\sum_x xp(x)$ , variance =  $\sum_x (x - E[x])^2 p(x)$

- Moments are expectations of higher order powers.  
(Mean is first moment. Autocorrelation is second moment.)
- Centralized moments have lower moments subtracted away  
(e.g. variance, skew, kurtosis).
- Deep fact: Knowledge of all orders of moments completely defines the entire distribution.

## JOINT PROBABILITY

- Key concept: two or more random variables may interact.  
Thus, the probability of one taking on a certain value depends on which value(s) the others are taking.
- We call this a joint ensemble and write  

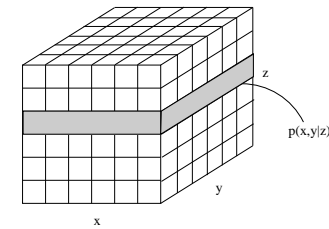
$$p(x, y) = \text{prob}(X = x \text{ and } Y = y)$$



## CONDITIONAL PROBABILITY

- If we know that some event has occurred, it changes our belief about the probability of other events.
- This is like taking a "slice" through the joint table.

$$p(x|y) = p(x, y) / p(y)$$

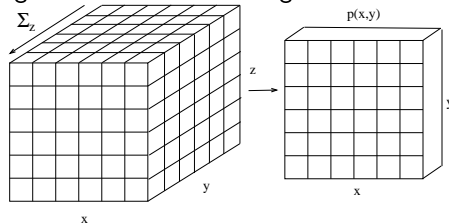


## MARGINAL PROBABILITIES

- We can "sum out" part of a joint distribution to get the *marginal distribution* of a subset of variables:

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

- This is like adding slices of the table together.



- Another equivalent definition:  $p(x) = \sum_y p(x|y)p(y)$ .

## BAYES' RULE

- Manipulating the basic definition of conditional probability gives one of the most important formulas in probability theory:

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

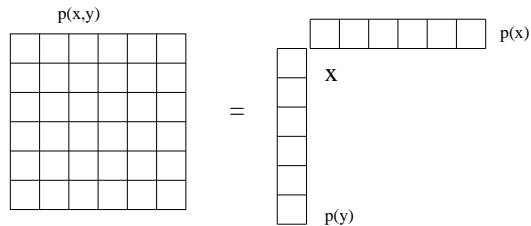
- This gives us a way of "reversing" conditional probabilities.
- Thus, all joint probabilities can be factored by selecting an ordering for the random variables and using the "chain rule":

$$p(x, y, z, \dots) = p(x)p(y|x)p(z|x, y)p(\dots|x, y, z)$$

## INDEPENDENCE & CONDITIONAL INDEPENDENCE

- Two variables are independent iff their joint factors:

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



- Two variables are conditionally independent given a third one if for all values of the conditioning variable, the resulting slice factors:

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

## ENTROPY

- Measures the amount of ambiguity or uncertainty in a distribution:

$$H(p) = - \sum_x p(x) \log p(x)$$

- Expected value of  $-\log p(x)$  (a function which depends on  $p(x)$ !).
- $H(p) > 0$  unless only one possible outcome in which case  $H(p) = 0$ .
- Maximal value when  $p$  is uniform.
- Tells you the expected "cost" if each event costs  $-\log p(\text{event})$

## BE CAREFUL!

- Watch the context:  
e.g. Simpson's paradox
- Define random variables and sample spaces carefully:  
e.g. Prisoner's paradox

## CROSS ENTROPY (KL DIVERGENCE)

- An asymmetric measure of the distance between two distributions:

$$KL[p||q] = \sum_x p(x) [\log p(x) - \log q(x)]$$

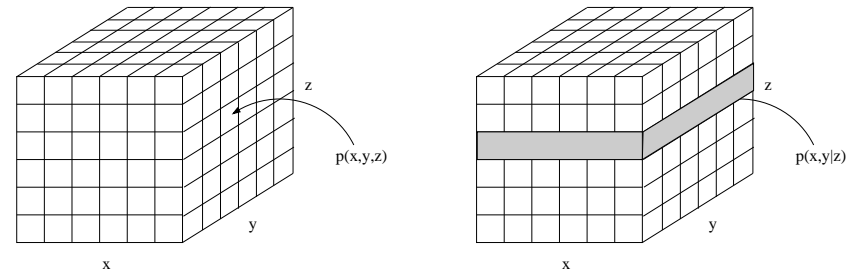
- $KL > 0$  unless  $p = q$  then  $KL = 0$
- Tells you the extra cost if events were generated by  $p(x)$  but instead of charging under  $p(x)$  you charged under  $q(x)$ .

## STATISTICS

- Probability: inferring probabilistic quantities for data given fixed models (e.g. prob. of events, marginals, conditionals, etc).
- Statistics: inferring a model given fixed data observations (e.g. clustering, classification, regression).
- Many approaches to statistics:  
*frequentist, Bayesian, decision theory, ...*

## (CONDITIONAL) PROBABILITY TABLES

- For discrete (categorical) quantities, the most basic parametrization is the probability table which lists  $p(x_i = k^{th} \text{ value})$ .
- Since PTs must be nonnegative and sum to 1, for  $k$ -ary variables there are  $k - 1$  free parameters.
- If a discrete variable is conditioned on the values of some other discrete variables we make one table for each possible setting of the parents: these are called *conditional probability tables* or CPTs.



## SOME (CONDITIONAL) PROBABILITY FUNCTIONS

- Probability density functions  $p(x)$  (for continuous variables) or probability mass functions  $p(x = k)$  (for discrete variables) tell us how likely it is to get a particular value for a random variable (possibly conditioned on the values of some other variables.)
- We can consider various types of variables: binary/discrete (categorical), continuous, interval, and integer counts.
- For each type we'll see some basic *probability models* which are parametrized families of distributions.

## EXPONENTIAL FAMILY

- For (continuous or discrete) random variable  $\mathbf{x}$

$$\begin{aligned} p(\mathbf{x}|\eta) &= h(\mathbf{x}) \exp\{\eta^\top T(\mathbf{x}) - A(\eta)\} \\ &= \frac{1}{Z(\eta)} h(\mathbf{x}) \exp\{\eta^\top T(\mathbf{x})\} \end{aligned}$$

is an exponential family distribution with *natural parameter*  $\eta$ .

- Function  $T(\mathbf{x})$  is a *sufficient statistic*.
- Function  $A(\eta) = \log Z(\eta)$  is the log normalizer.
- Key idea: all you need to know about the data is captured in the summarizing function  $T(\mathbf{x})$ .

## BERNOULLI

---

- For a binary random variable with  $p(\text{heads})=\pi$ :

$$\begin{aligned} p(x|\pi) &= \pi^x(1-\pi)^{1-x} \\ &= \exp \left\{ \log \left( \frac{\pi}{1-\pi} \right) x + \log(1-\pi) \right\} \end{aligned}$$

- Exponential family with:

$$\begin{aligned} \eta &= \log \frac{\pi}{1-\pi} \\ T(x) &= x \\ A(\eta) &= -\log(1-\pi) = \log(1+e^\eta) \\ h(x) &= 1 \end{aligned}$$

- The logistic function relates the natural parameter and the chance of heads

$$\pi = \frac{1}{1 + e^{-\eta}}$$

## MULTINOMIAL

---

- For a set of integer counts on  $k$  trials

$$p(\mathbf{x}|\pi) = \frac{k!}{x_1!x_2!\cdots x_n!} \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_n^{x_n} = h(\mathbf{x}) \exp \left\{ \sum_i x_i \log \pi_i \right\}$$

- But the parameters are constrained:  $\sum_i \pi_i = 1$ .  
So we define the last one  $\pi_n = 1 - \sum_{i=1}^{n-1} \pi_i$ .

$$p(\mathbf{x}|\pi) = h(\mathbf{x}) \exp \left\{ \sum_{i=1}^{n-1} \log \left( \frac{\pi_i}{\pi_n} \right) x_i + k \log \pi_n \right\}$$

- Exponential family with:

$$\begin{aligned} \eta_i &= \log \pi_i - \log \pi_n \\ T(x_i) &= x_i \\ A(\eta) &= -k \log \pi_n = k \log \sum_i e^{\eta_i} \\ h(\mathbf{x}) &= k! / x_1! x_2! \cdots x_n! \end{aligned}$$

## POISSON

---

- For an integer count variable with rate  $\lambda$ :

$$\begin{aligned} p(x|\lambda) &= \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \frac{1}{x!} \exp \{ x \log \lambda - \lambda \} \end{aligned}$$

- Exponential family with:

$$\begin{aligned} \eta &= \log \lambda \\ T(x) &= x \\ A(\eta) &= \lambda = e^\eta \\ h(x) &= \frac{1}{x!} \end{aligned}$$

- e.g. number of photons  $\mathbf{x}$  that arrive at a pixel during a fixed interval given mean intensity  $\lambda$
- Other count densities: binomial, exponential.

- The *softmax* function relates the basic and natural parameters:

$$\pi_i = \frac{e^{\eta_i}}{\sum_j e^{\eta_j}}$$

## GAUSSIAN (NORMAL)

- For a continuous univariate random variable:

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

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

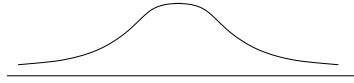
- Exponential family with:

$$\eta = [\mu/\sigma^2; -1/2\sigma^2]$$

$$T(x) = [x; x^2]$$

$$A(\eta) = \log \sigma + \mu/2\sigma^2$$

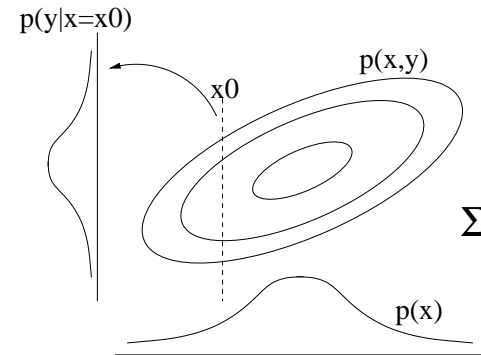
$$h(x) = 1/\sqrt{2\pi}$$



- Note: a univariate Gaussian is a two-parameter distribution with a two-component vector of sufficient statistics.

## IMPORTANT GAUSSIAN FACTS

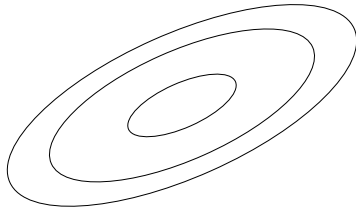
- All marginals of a Gaussian are again Gaussian.  
Any conditional of a Gaussian is again Gaussian.



## MULTIVARIATE GAUSSIAN DISTRIBUTION

- For a continuous vector random variable:

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



- Exponential family with:

$$\eta = [\Sigma^{-1}\mu; -1/2\Sigma^{-1}]$$

$$T(\mathbf{x}) = [\mathbf{x}; \mathbf{x}\mathbf{x}^\top]$$

$$A(\eta) = \log |\Sigma|/2 + \mu^\top \Sigma^{-1} \mu/2$$

$$h(\mathbf{x}) = (2\pi)^{-n/2}$$

- Sufficient statistics: mean vector and correlation matrix.
- Other densities: Student-t, Laplacian.
- For non-negative values use exponential, Gamma, log-normal.

## GAUSSIAN MARGINALS/CONDITIONALS

- To find these parameters is mostly linear algebra:  
Let  $\mathbf{z} = [\mathbf{x}^\top \mathbf{y}^\top]^\top$  be normally distributed according to:

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}; \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{B} \end{bmatrix} \right)$$

where  $\mathbf{C}$  is the (non-symmetric) cross-covariance matrix between  $\mathbf{x}$  and  $\mathbf{y}$  which has as many rows as the size of  $\mathbf{x}$  and as many columns as the size of  $\mathbf{y}$ .

The marginal distributions are:

$$\mathbf{x} \sim \mathcal{N}(\mathbf{a}; \mathbf{A})$$

$$\mathbf{y} \sim \mathcal{N}(\mathbf{b}; \mathbf{B})$$

and the conditional distributions are:

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\mathbf{a} + \mathbf{C}\mathbf{B}^{-1}(\mathbf{y} - \mathbf{b}); \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\top)$$

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{b} + \mathbf{C}^\top \mathbf{A}^{-1}(\mathbf{x} - \mathbf{a}); \mathbf{B} - \mathbf{C}^\top \mathbf{A}^{-1}\mathbf{C})$$

## MOMENTS

- For continuous variables, moment calculations are important.
- We can easily compute moments of any exponential family distribution by taking the derivatives of the log normalizer  $A(\eta)$ .
- The  $q^{th}$  derivative gives the  $q^{th}$  centred moment.

$$\begin{aligned}\frac{dA(\eta)}{d\eta} &= \text{mean} \\ \frac{d^2A(\eta)}{d\eta^2} &= \text{variance} \\ &\dots\end{aligned}$$

- When the sufficient statistic is a vector, partial derivatives need to be considered.

## GENERALIZED LINEAR MODELS (GLMs)

- Generalized Linear Models:  $p(\mathbf{y}|\mathbf{x})$  is exponential family with conditional mean  $\mu = f(\theta^\top \mathbf{x})$ .
- The function  $f$  is called the *response function*.
- If we chose  $f$  to be the inverse of the mapping b/w conditional mean and natural parameters then it is called the *canonical response function*.

$$\begin{aligned}\eta &= \psi(\mu) \\ f(\cdot) &= \psi^{-1}(\cdot)\end{aligned}$$

## PARAMETERIZING CONDITIONALS

- When the variable(s) being conditioned on (parents) are discrete, we just have one density for each possible setting of the parents. e.g. a table of natural parameters in exponential models or a table of tables for discrete models.
- When the conditioned variable is continuous, its value sets some of the parameters for the other variables.
- A very common instance of this for regression is the “linear-Gaussian”:  $p(\mathbf{y}|\mathbf{x}) = \text{gauss}(\theta^\top \mathbf{x}; \Sigma)$ .
- For discrete children and continuous parents, we often use a Bernoulli/multinomial whose parameters are some function  $f(\theta^\top \mathbf{x})$ .

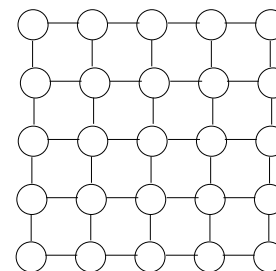
## POTENTIAL FUNCTIONS

- We can be even more general and define distributions by arbitrary *energy* functions proportional to the log probability.

$$p(\mathbf{x}) \propto \exp\left\{-\sum_k H_k(\mathbf{x})\right\}$$

- A common choice is to use pairwise terms in the energy:

$$H(\mathbf{x}) = \sum_i a_i x_i + \sum_{\text{pairs } ij} w_{ij} x_i x_j$$



### SPECIAL VARIABLES

---

- If certain variables are *always observed* we may not want to model their density. For example inputs in regression or classification. This leads to conditional density estimation.
- If certain variables are *always unobserved*, they are called *hidden* or *latent* variables. They can always be marginalized out, but can make the density modeling of the observed variables easier. (We'll see more on this later.)

### LIKELIHOOD FUNCTION

---

- So far we have focused on the (log) probability function  $p(\mathbf{x}|\theta)$  which assigns a probability (density) to any joint configuration of variables  $\mathbf{x}$  given fixed parameters  $\theta$ .
- But in learning we turn this on its head: we have some fixed data and we want to find parameters.
- Think of  $p(\mathbf{x}|\theta)$  as a function of  $\theta$  for fixed  $\mathbf{x}$ :

$$L(\theta; \mathbf{x}) = p(\mathbf{x}|\theta)$$

$$\ell(\theta; \mathbf{x}) = \log p(\mathbf{x}|\theta)$$

This function is called the (log) “likelihood”.

- Chose  $\theta$  to maximize some cost function  $c(\theta)$  which includes  $\ell(\theta)$ :

$$c(\theta) = \ell(\theta; \mathcal{D}) \quad \text{maximum likelihood (ML)}$$

$$c(\theta) = \ell(\theta; \mathcal{D}) + r(\theta) \quad \text{maximum a posteriori (MAP)/penalized ML}$$

(also cross-validation, Bayesian estimators, BIC, AIC, ...)

### MULTIPLE OBSERVATIONS, COMPLETE DATA, IID SAMPLING

---

- A single observation of the data  $\mathbf{X}$  is rarely useful on its own.
- Generally we have data including many observations, which creates a *set of random variables*:  $\mathcal{D} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^M\}$
- Two very common assumptions:
  1. Observations are independently and identically distributed according to joint distribution of graphical model: IID samples.
  2. We observe all random variables in the domain on each observation: complete data.

### MAXIMUM LIKELIHOOD

---

- For IID data:

$$p(\mathcal{D}|\theta) = \prod_m p(\mathbf{x}^m|\theta)$$

$$\ell(\theta; \mathcal{D}) = \sum_m \log p(\mathbf{x}^m|\theta)$$

- Idea of maximum likelihood estimation (MLE): pick the setting of parameters most likely to have generated the data we saw:

$$\theta_{\text{ML}}^* = \operatorname{argmax}_{\theta} \ell(\theta; \mathcal{D})$$

- Very commonly used in statistics.  
Often leads to “intuitive”, “appealing”, or “natural” estimators.



### EXAMPLE: BERNOULLI TRIALS

- We observe  $M$  iid coin flips:  $\mathcal{D}=\text{H,H,T,H},\dots$
- Model:  $p(H) = \theta \quad p(T) = (1 - \theta)$
- Likelihood:

$$\begin{aligned}\ell(\theta; \mathcal{D}) &= \log p(\mathcal{D}|\theta) \\ &= \log \prod_m \theta^{\mathbf{x}^m} (1 - \theta)^{1-\mathbf{x}^m} \\ &= \log \theta \sum_m \mathbf{x}^m + \log(1 - \theta) \sum_m (1 - \mathbf{x}^m) \\ &= \log \theta N_H + \log(1 - \theta) N_T\end{aligned}$$

- Take derivatives and set to zero:

$$\begin{aligned}\frac{\partial \ell}{\partial \theta} &= \frac{N_H}{\theta} - \frac{N_T}{1 - \theta} \\ \Rightarrow \theta_{\text{ML}}^* &= \frac{N_H}{N_H + N_T}\end{aligned}$$

### EXAMPLE: UNIVARIATE NORMAL

- We observe  $M$  iid real samples:  $\mathcal{D}=1.18,-.25,.78,\dots$
- Model:  $p(x) = (2\pi\sigma^2)^{-1/2} \exp\{-(x - \mu)^2/2\sigma^2\}$
- Likelihood (using probability density):

$$\begin{aligned}\ell(\theta; \mathcal{D}) &= \log p(\mathcal{D}|\theta) \\ &= -\frac{M}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_m \frac{(x^m - \mu)^2}{\sigma^2}\end{aligned}$$

- Take derivatives and set to zero:

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= (1/\sigma^2) \sum_m (x_m - \mu) \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{M}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_m (x_m - \mu)^2 \\ \Rightarrow \mu_{\text{ML}} &= (1/M) \sum_m x_m \\ \sigma_{\text{ML}}^2 &= (1/M) \sum_m x_m^2 - \mu_{\text{ML}}^2\end{aligned}$$

### EXAMPLE: MULTINOMIAL

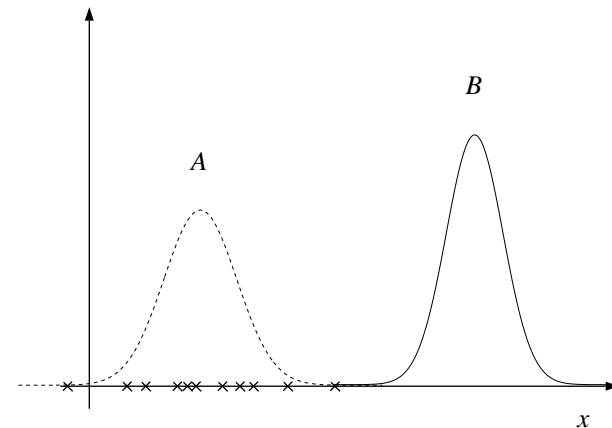
- We observe  $M$  iid die rolls (K-sided):  $\mathcal{D}=3,1,K,2,\dots$
- Model:  $p(k) = \theta_k \quad \sum_k \theta_k = 1$
- Likelihood (for binary indicators  $[\mathbf{x}^m = k]$ ):

$$\begin{aligned}\ell(\theta; \mathcal{D}) &= \log p(\mathcal{D}|\theta) \\ &= \log \prod_m \theta_{\mathbf{x}^m} = \log \prod_m \theta_1^{[\mathbf{x}^m=1]} \dots \theta_k^{[\mathbf{x}^m=k]} \\ &= \sum_k \log \theta_k \sum_m [\mathbf{x}^m = k] = \sum_k N_k \log \theta_k\end{aligned}$$

- Take derivatives and set to zero (enforcing  $\sum_k \theta_k = 1$ ):

$$\begin{aligned}\frac{\partial \ell}{\partial \theta_k} &= \frac{N_k}{\theta_k} - M \\ \Rightarrow \theta_k^* &= \frac{N_k}{M}\end{aligned}$$

### EXAMPLE: UNIVARIATE NORMAL



### EXAMPLE: LINEAR REGRESSION

- In linear regression, some inputs (covariates, parents) and all outputs (responses, children) are continuous valued variables.
- For each child and setting of discrete parents we use the model:

$$p(y|\mathbf{x}, \theta) = \text{gauss}(y|\theta^\top \mathbf{x}, \sigma^2)$$

- The likelihood is the familiar “squared error” cost:

$$\ell(\theta; \mathcal{D}) = -\frac{1}{2\sigma^2} \sum_m (y^m - \theta^\top \mathbf{x}^m)^2$$

- The ML parameters can be solved for using linear least-squares:

$$\begin{aligned} \frac{\partial \ell}{\partial \theta} &= -\sum_m (y^m - \theta^\top \mathbf{x}^m) \mathbf{x}^m \\ \Rightarrow \theta_{\text{ML}}^* &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \end{aligned}$$

### SUFFICIENT STATISTICS

- A statistic is a function of a random variable.

- $T(\mathbf{X})$  is a “sufficient statistic” for  $\mathbf{X}$  if

$$T(\mathbf{x}^1) = T(\mathbf{x}^2) \Rightarrow L(\theta; \mathbf{x}^1) = L(\theta; \mathbf{x}^2) \quad \forall \theta$$

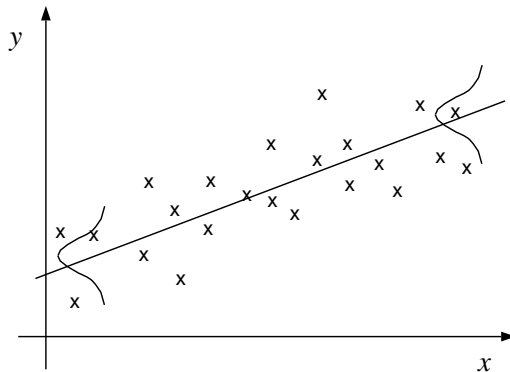
- Equivalently (by the Neyman factorization theorem) we can write:

$$p(\mathbf{x}|\theta) = h(\mathbf{x}, T(\mathbf{x})) g(T(\mathbf{x}), \theta)$$

- Example: exponential family models:

$$p(\mathbf{x}|\theta) = h(\mathbf{x}) \exp\{\eta^\top T(\mathbf{x}) - A(\eta)\}$$

### EXAMPLE: LINEAR REGRESSION



### SUFFICIENT STATISTICS ARE SUMS

- In the examples above, the sufficient statistics were merely sums (counts) of the data:

Bernoulli: # of heads, tails

Multinomial: # of each type

Gaussian: mean, mean-square

Regression: correlations

- As we will see, this is true for all exponential family models: sufficient statistics are average natural parameters.
- Only exponential family models have simple sufficient statistics.

## MLE FOR EXPONENTIAL FAMILY MODELS

- Recall the probability function for exponential models:

$$p(\mathbf{x}|\theta) = h(\mathbf{x}) \exp\{\eta^\top T(\mathbf{x}) - A(\eta)\}$$

- For iid data, sufficient statistic is  $\sum_m T(\mathbf{x}^m)$ :

$$\ell(\eta; \mathcal{D}) = \log p(\mathcal{D}|\eta) = \left( \sum_m \log h(\mathbf{x}^m) \right) - M A(\eta) + \left( \eta^\top \sum_m T(\mathbf{x}^m) \right)$$

- Take derivatives and set to zero:

$$\begin{aligned} \frac{\partial \ell}{\partial \eta} &= \sum_m T(\mathbf{x}^m) - M \frac{\partial A(\eta)}{\partial \eta} \\ \Rightarrow \frac{\partial A(\eta)}{\partial \eta} &= \frac{1}{M} \sum_m T(\mathbf{x}^m) \\ \eta_{\text{ML}} &= \frac{1}{M} \sum_m T(\mathbf{x}^m) \end{aligned}$$

recalling that the natural moments of an exponential distribution are the derivatives of the log normalizer.

## FUNDAMENTAL OPERATIONS WITH DISTRIBUTIONS

- Generate data*: draw samples from the distribution. This often involves generating a uniformly distributed variable in the range  $[0,1]$  and transforming it. For more complex distributions it may involve an iterative procedure that takes a long time to produce a single sample (e.g. Gibbs sampling, MCMC).
- Compute log probabilities*.  
When all variables are either observed or marginalized the result is a single number which is the log prob of the configuration.
- Inference*: Compute expectations of some variables given others which are observed or marginalized.
- Learning*.  
Set the parameters of the density functions given some (partially) observed data to maximize likelihood or penalized likelihood.

## BASIC STATISTICAL PROBLEMS

- Let's remind ourselves of the basic problems we discussed on the first day: *density estimation*, *clustering*, *classification* and *regression*.
- Density estimation is hardest. If we can do joint density estimation then we can always condition to get what we want:  
Regression:  $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}, \mathbf{x})/p(\mathbf{x})$   
Classification:  $p(c|\mathbf{x}) = p(c, \mathbf{x})/p(\mathbf{x})$   
Clustering:  $p(c|\mathbf{x}) = p(c, \mathbf{x})/p(\mathbf{x})$   $c$  unobserved

## LEARNING

- In AI the bottleneck is often knowledge acquisition.
- Human experts are rare, expensive, unreliable, slow.
- But we have lots of data.
- Want to build systems automatically based on data and a small amount of prior information (from experts).

## KNOWN MODELS

---

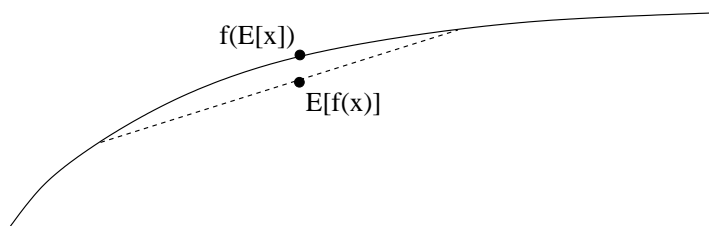
- Many systems we build will be essentially probability models.
- Assume the prior information we have specifies type & structure of the model, as well as the form of the (conditional) distributions or potentials.
- In this case learning  $\equiv$  setting parameters.
- Also possible to do “structure learning” to learn model.

## JENSEN'S INEQUALITY

---

- For any concave function  $f()$  and any distribution on  $x$ ,

$$E[f(x)] \leq f(E[x])$$



- e.g.  $\log()$  and  $\sqrt{\quad}$  are concave
- This allows us to bound expressions like  $\log p(x) = \log \sum_z p(x, z)$