Discovering Models / Theories

Domain Theories

- **Agent** :

  given precept history $p \in P$,
  select decision from set of choices $a \in A$

  so as to meet a goal g (performance) –

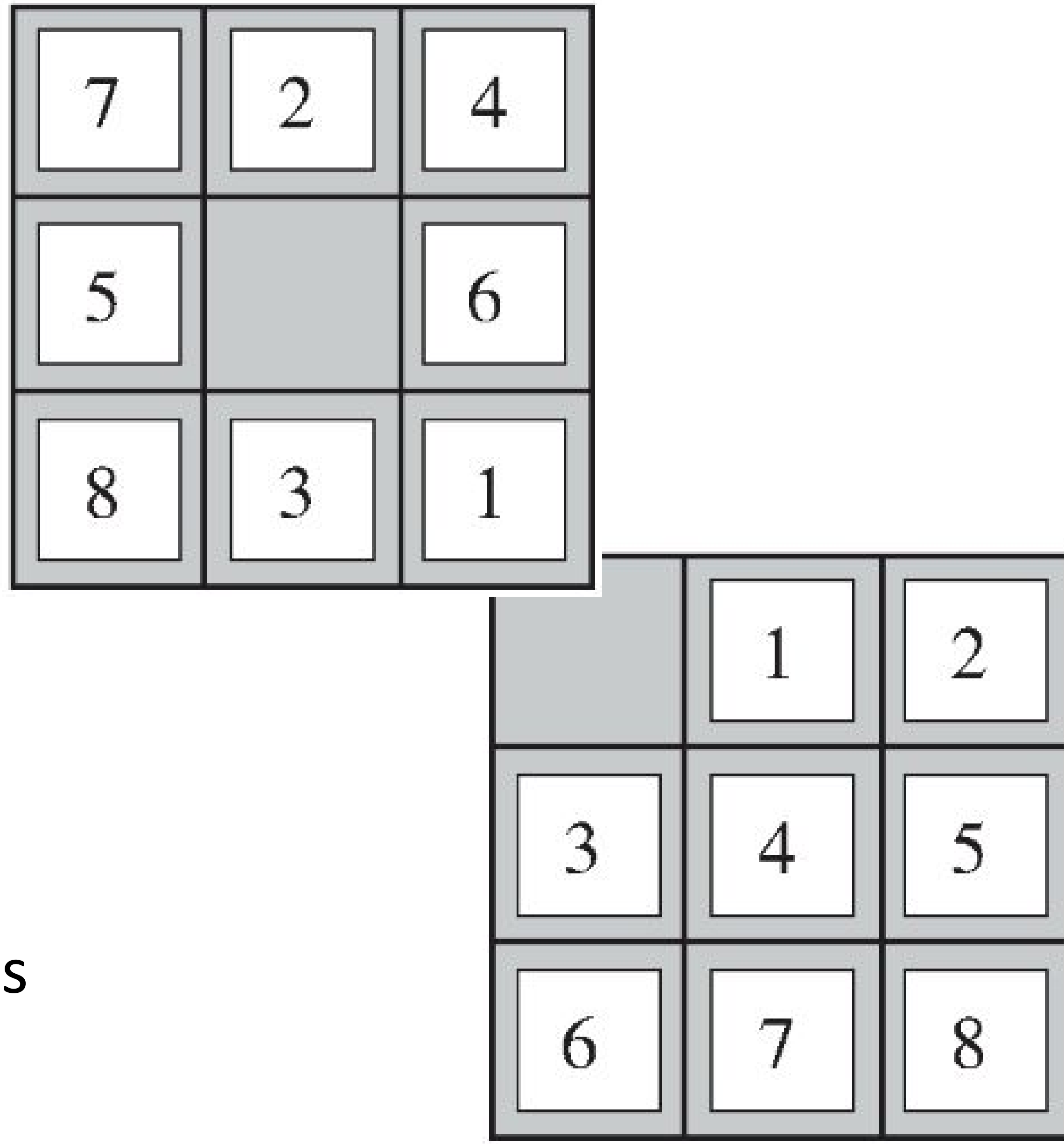
  maximize utility function U()

- Requires knowledge of how actions under different precepts affect the goal

  → Model or **Theory**

- Task domains: a) **8-puzzle,** [detrmnstc] b) **Soccer** [stochastic]

8-puzzle

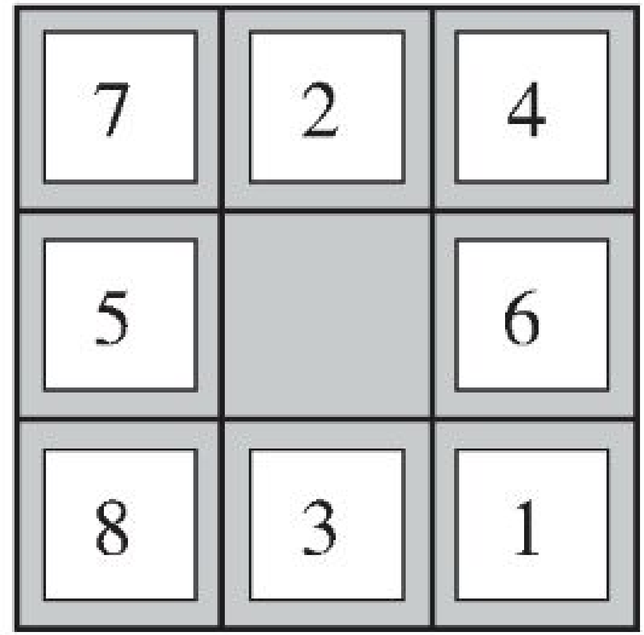


- Precept = state
- Actions = move
- Goal : T/F
- Utility : num moves

8-puzzle

- State = [7,2,4,5,B,6,8,3,1]
- Actions = L,R, U,D

  State + Action
  → new State

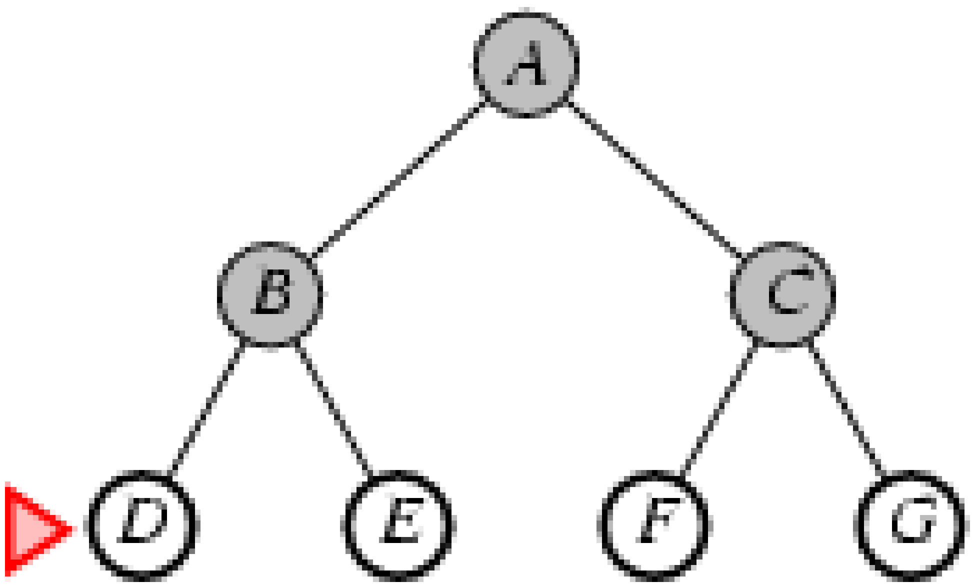


- Decision: based on Search
  - [Informed / Uninformed]

Breadth-first search

- Expand shallowest unexpanded node

- Fringe: FIFO queue new successors go at end

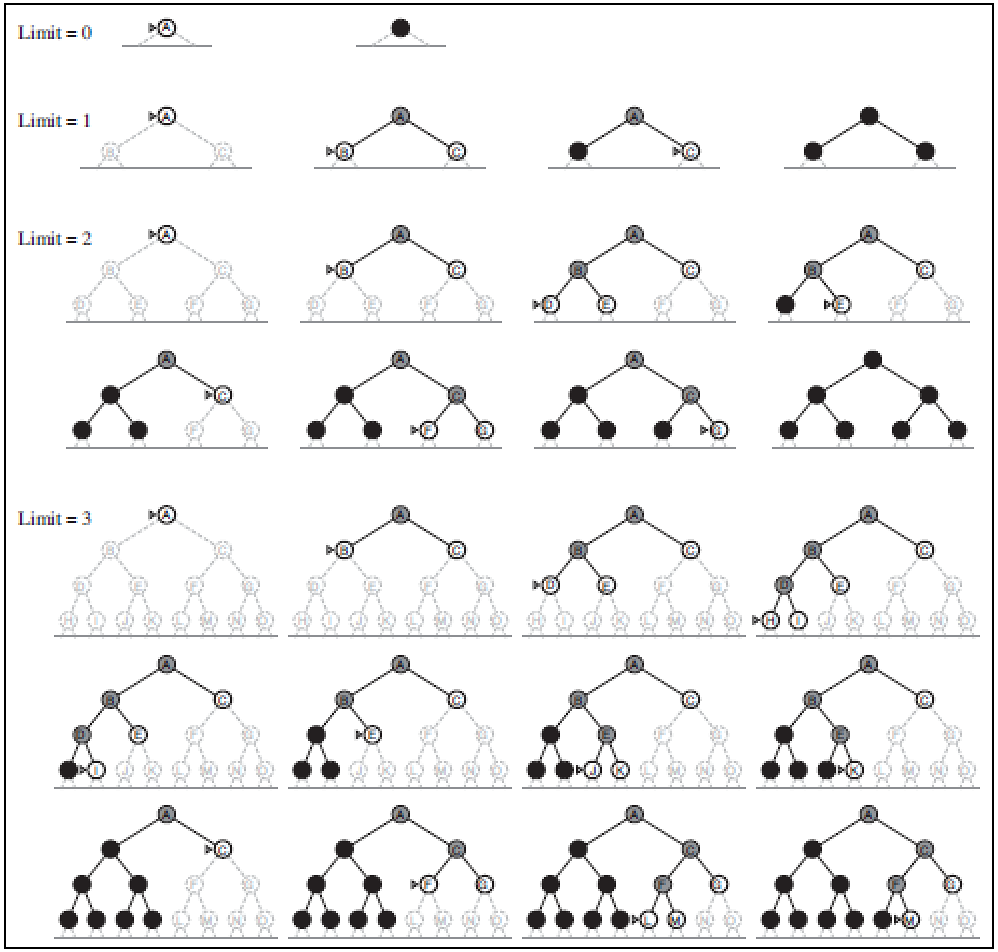


$O(b^{1+d})$

Properties of breadth-first search

- Complete? Yes (if b is finite)
- Time? $1+b+b^2+b^3+\ldots+b^d + b(b^d-1) = O(b^{d+1})$
- Space? $O(b^{d+1})$ (keeps every node in memory)
- Optimal? Yes (if cost = 1 per step)

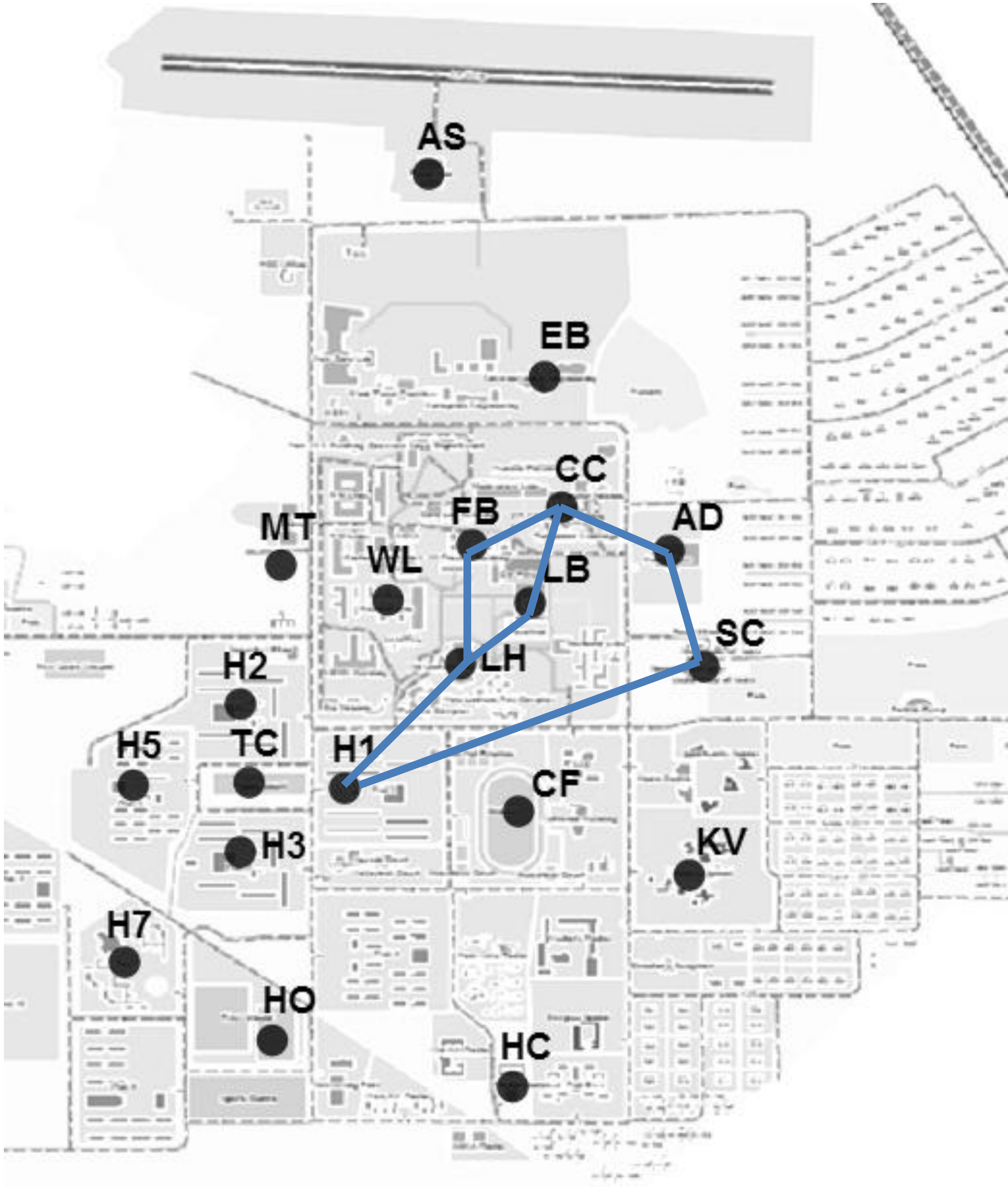
Iterative-Deepening search



Cost-based search

- edges don't have equal cost
- Breadth-first = first search lower costs from START
- Fringe: FIFO

$O(b^{1+C/\varepsilon})$



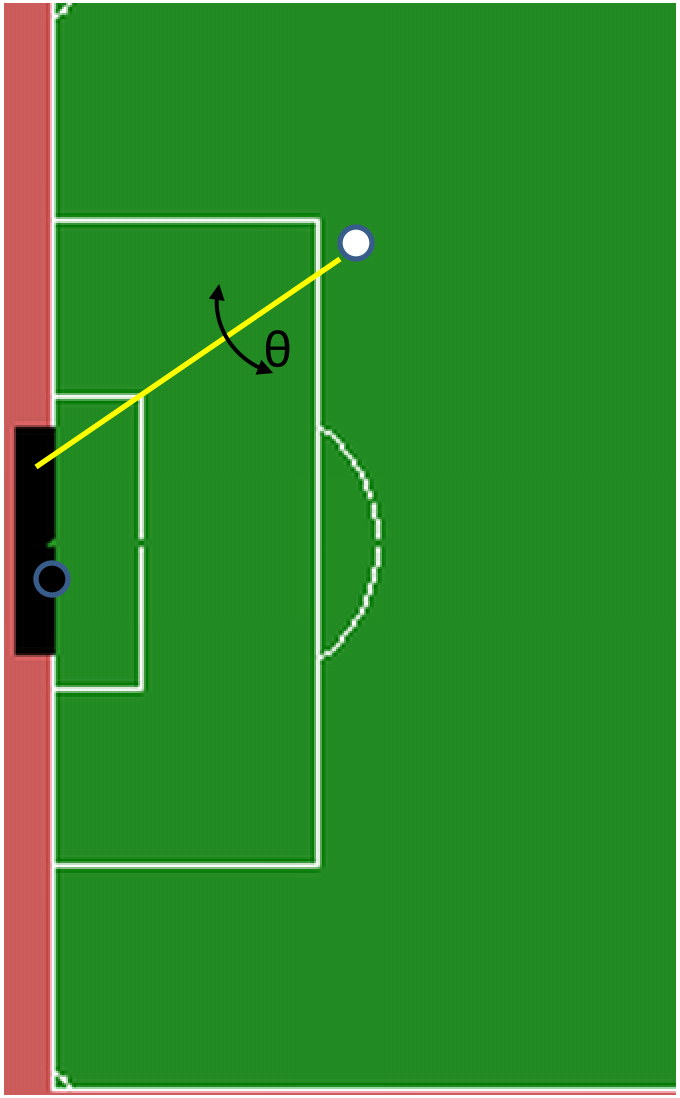
Soccer

- Precept = goalie, self, ball

  + wind, opponents, teammates…

- Actions = kick (angle, speed, swing)

- Utility : goal probability



Discrete-Deterministic Spaces:

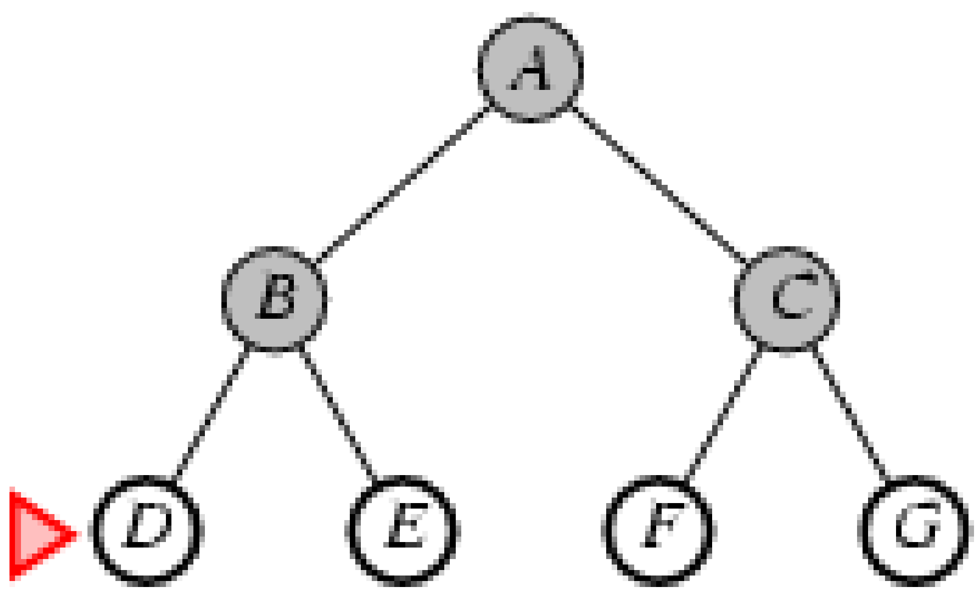
Search

Uninformed search strategies

- Uninformed search strategies use only the information available in the problem definitio
- Breadth-first search
- Uniform-cost search
- Depth-first search
- Depth-limited search
- Iterative deepening search

Breadth-first search

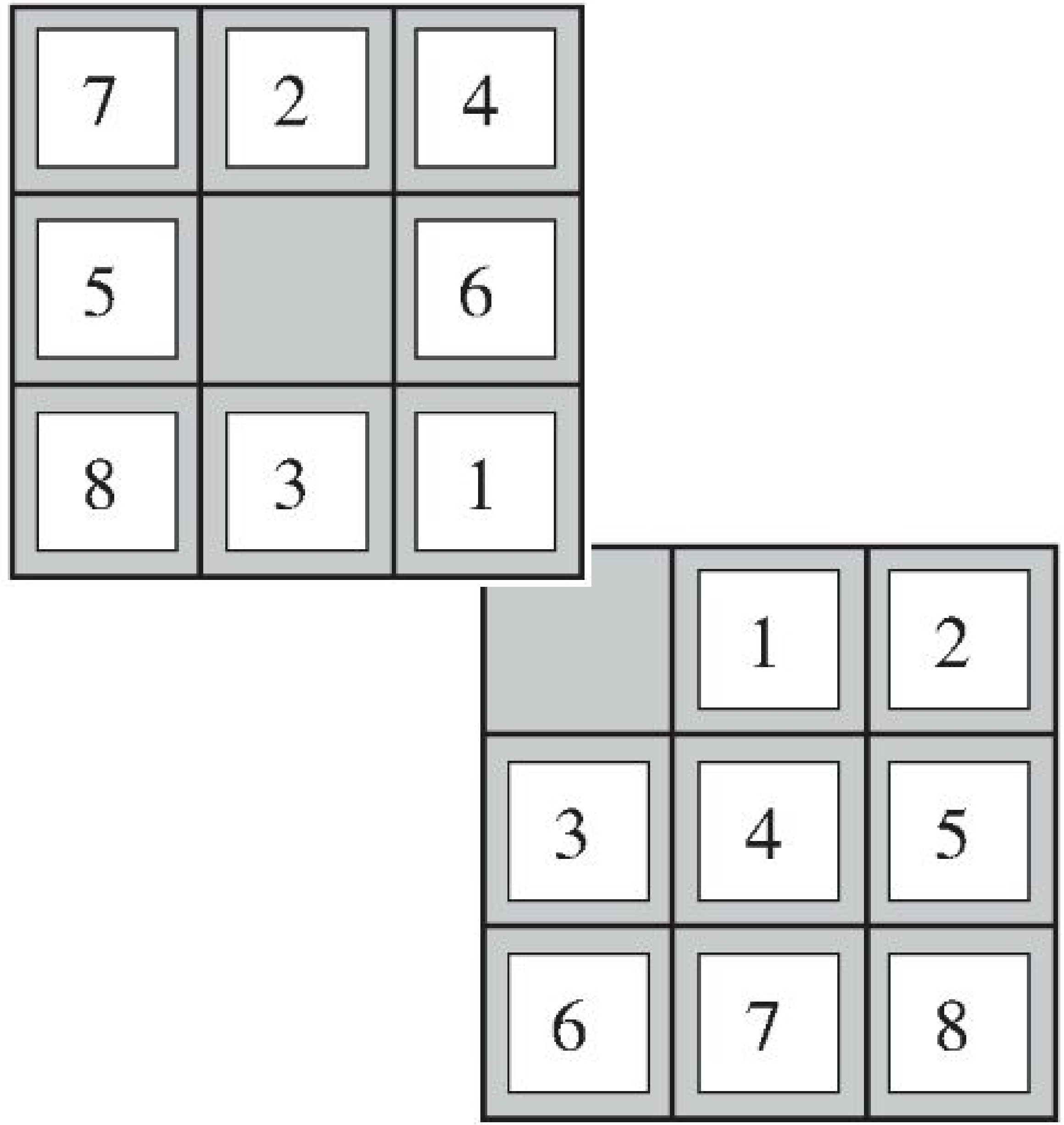
- Expand shallowest unexpanded node
- Fringe: FIFO queue new successors go at end



Properties of breadth-first search

- Complete? Yes (if *b* is finite)
- Time? $1+b+b^2+b^3+\ldots+b^d + b(b^d-1) = O(b^{d+1})$
- Space? $O(b^{d+1})$ (keeps every node in memory)
- Optimal? Yes (if cost = 1 per step)

Representing the state space



1. States:
2. Actions :
3. Goal test:
4. Cost:

8-puzzle heuristics

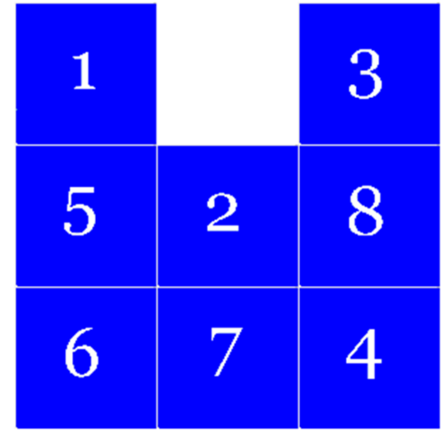
Admissible:

- h1 : Number of misplaced tiles

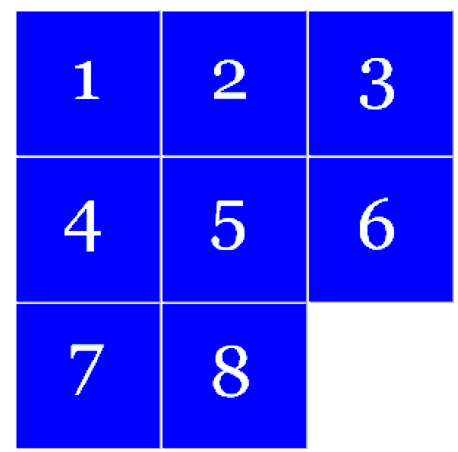
  = 6

- h2: Sum of Manhattan distances of the tiles from their goal positions

  = 0+0+1+1+2+3+1+3=11



goal:



8-puzzle heuristics

Nilsson's Sequence

$$Score(n) = P(n) + 3\,S(n)$$

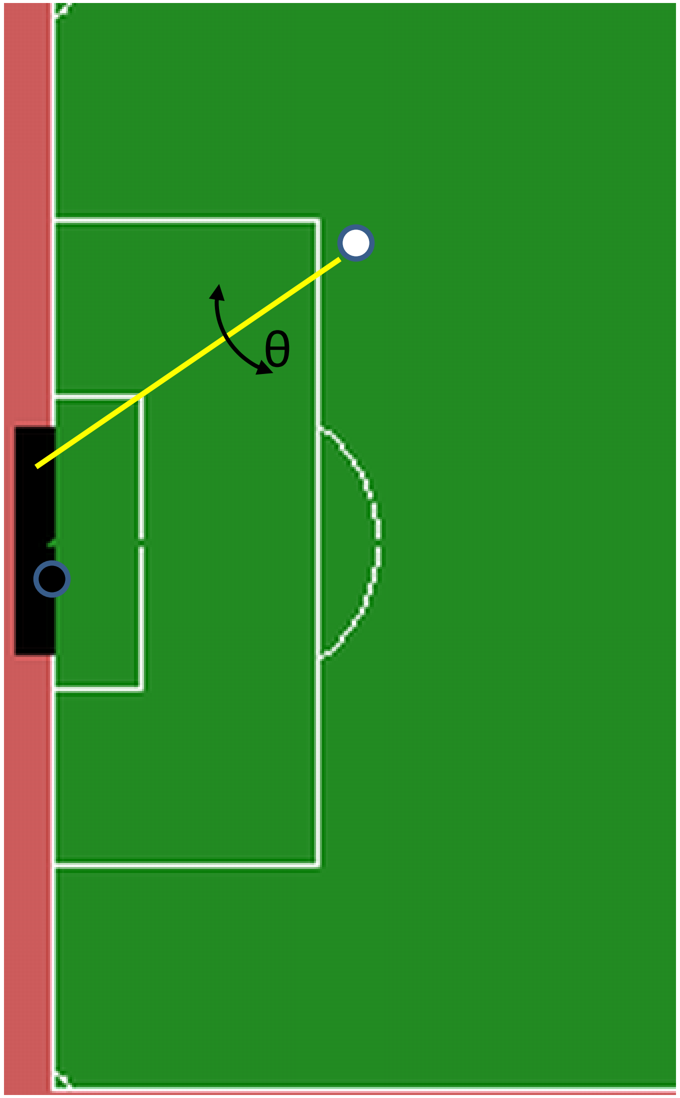
P(n) : Sum of Manhattan distances of each tile from its proper position

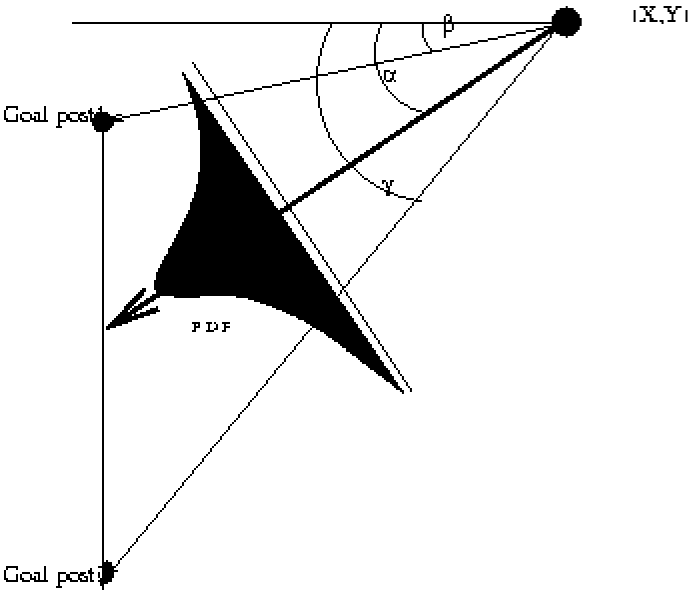
S(n), sequence score : check around the non-central squares:

- +2 for every tile not followed by successor
- 0 for every other tile.
- piece in center = +1

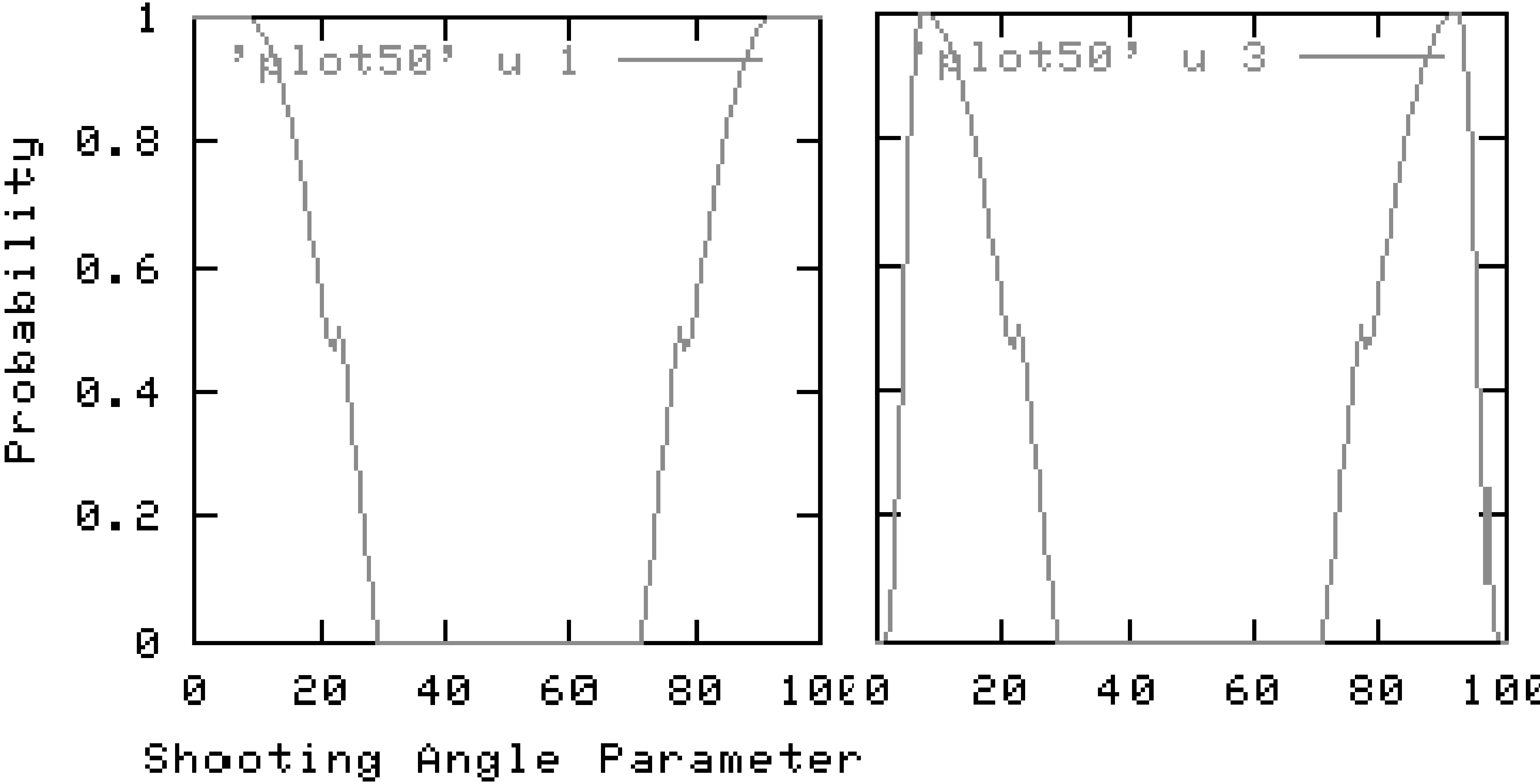
Stochastic Spaces

Soccer



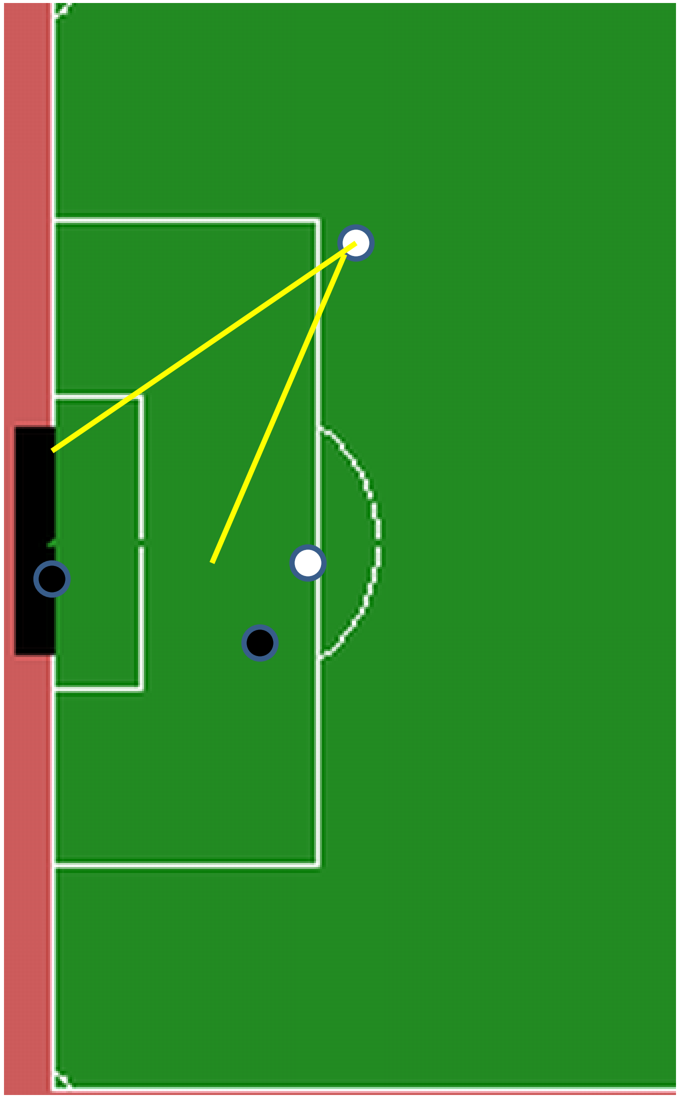


Soccer : Shooting at goal

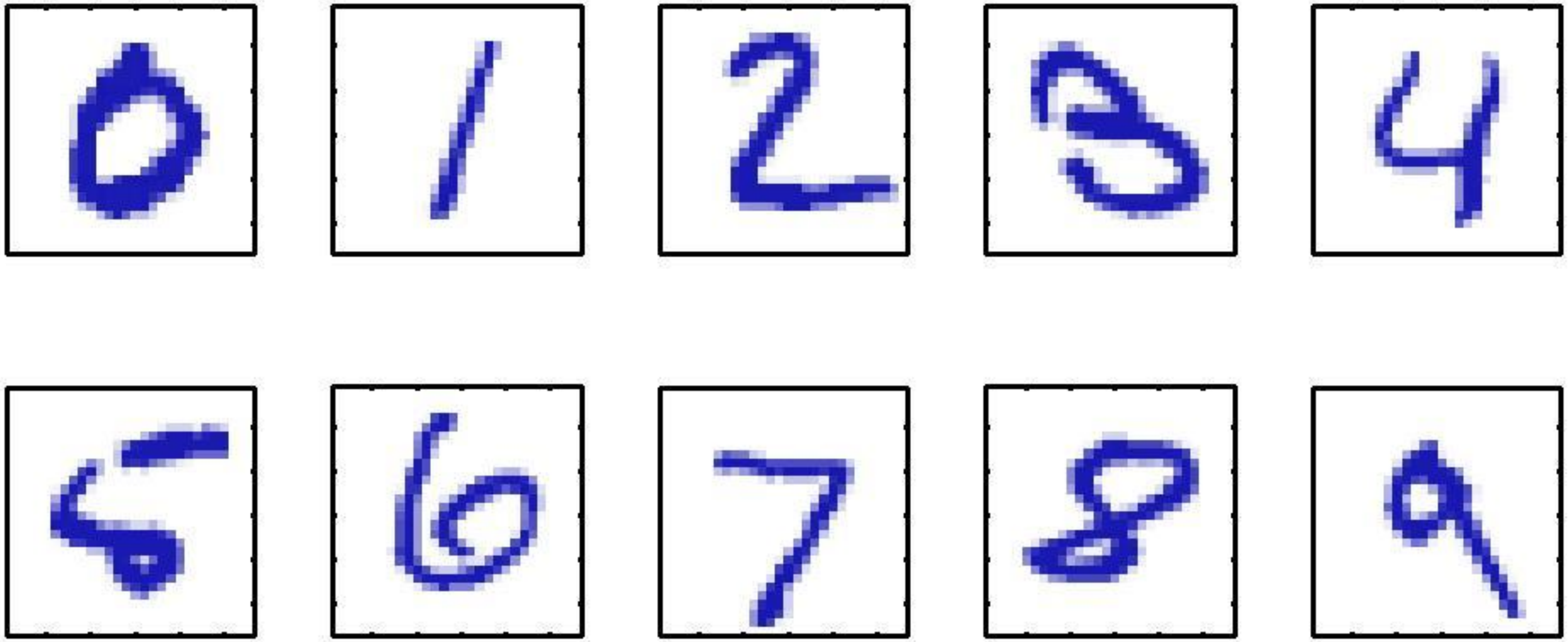


[acharya mukerjee 01]

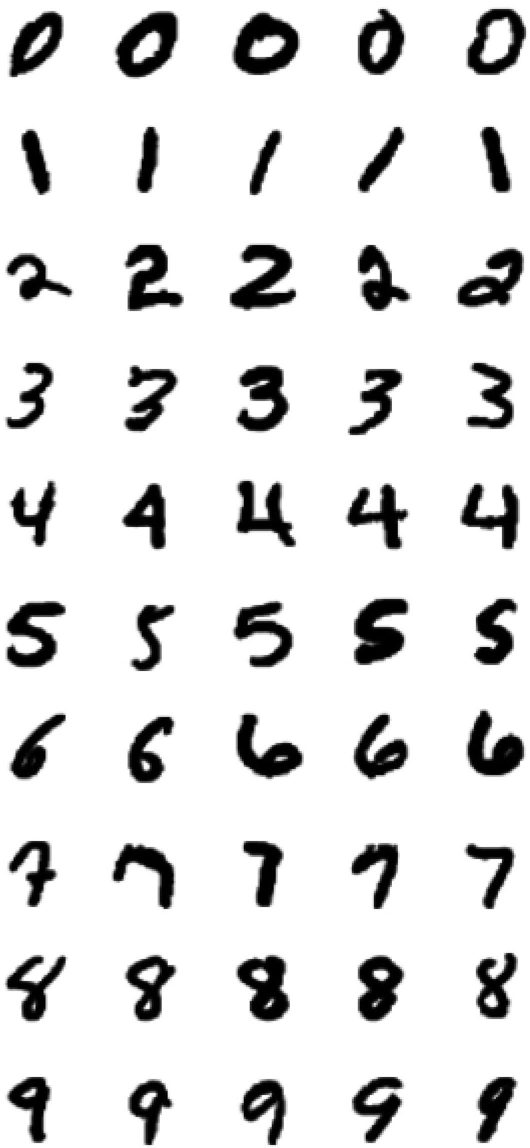
Soccer : Shoot, Pass, dribble, or ... ?



Handwritten digits - MNIST



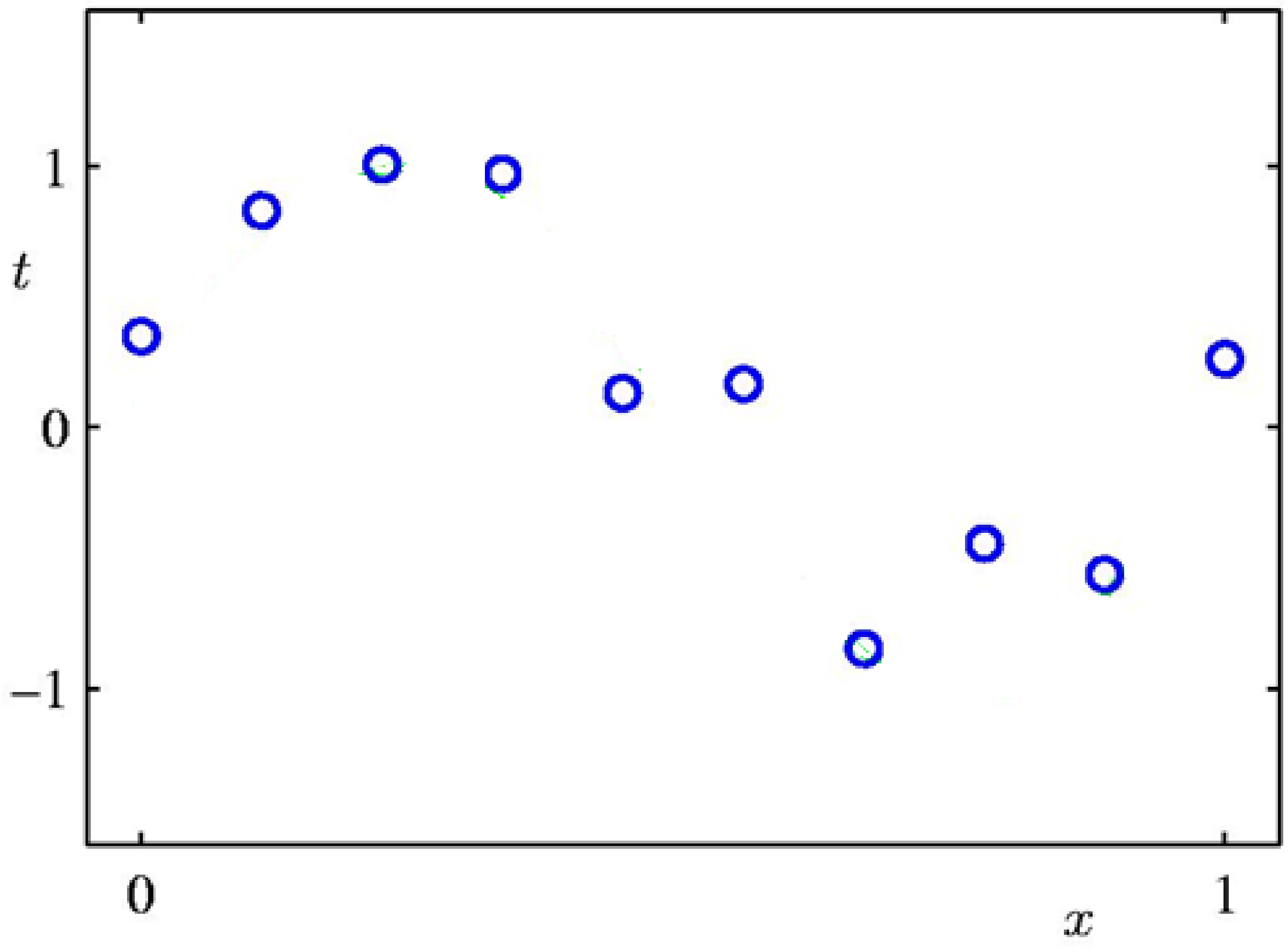
Confusion matrix



stimulus class \ response class	0	1	2	3	4	5	6	7	8	9
9	9	4	2	6	8	9	2	5	8	897
8	3		5	9	1	12	7	3	888	22
7		2	15	7	6			881	6	33
6	19	1	3		2	3	914	1	5	2
5	3		10	23	54	840	3	2	10	5
4	3				890		13	6	4	34
3			15	908		19		2	6	
2	3		916	16		5	2	1	6	1
1		937	3	1	1		1	5		2
0	937		2	1		1	6			3

Discovering theories

Continuous Data



Discrete Attribute data

- Examples described by attribute values (Boolean, discrete, continuous)
- E.g., situations where I will/won't wait at a restaurant:

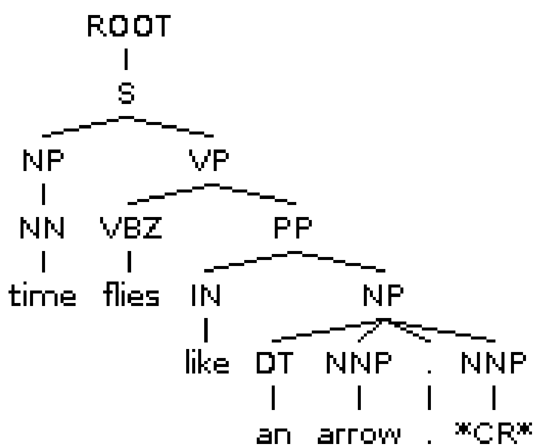
Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*Wait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

- Classification of examples is positive (T) or negative (F)

Discrete Features

- Parse the sentence: “Time flies like an arrow”

```
(ROOT
  (S
    (NP (NN time))
    (VP (VBZ flies)
      (PP (IN like)
        (NP (DT an) (NNP arrow) (. .) (NNP *CR*))))))
```



May have many parses.
How to rank the choices?

Regression

Modelling as Regression

Given a set of decisions y_i based on observations x_i,

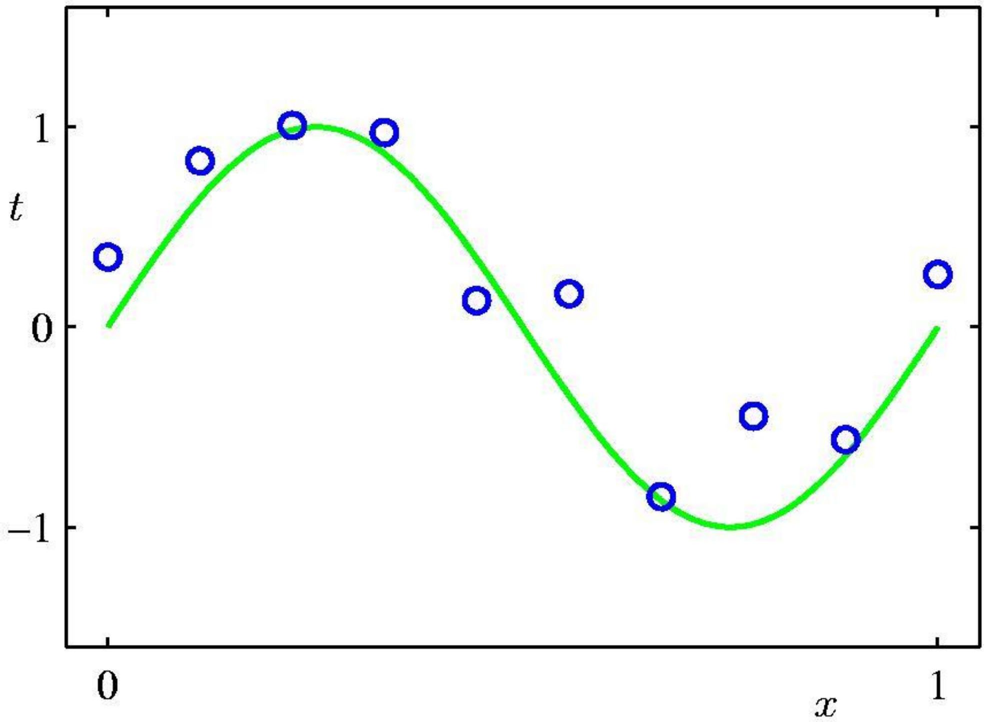
- derived from unknown function $\boldsymbol{y = f(x)}$
- with noise

Try to find a model or theory:

$$\boldsymbol{y = h(x) \approx f(x)}$$

where h() is drawn from the hypothesis space – e.g. the space of radial basis functions, or polynomials, etc.

Polynomial Curve Fitting



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

[Bishop 06] ch.1

Linear Regression

$$\mathbf{y = f(x) = \Sigma_i}\ w_i\ \mathbf{.\ \phi_i(x)}$$

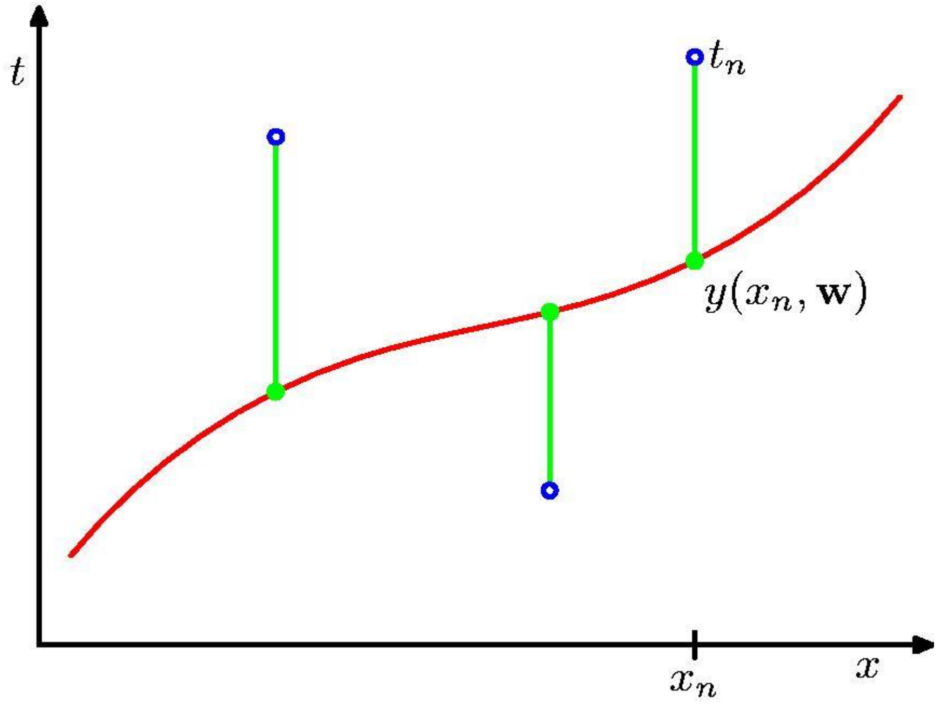
$\boldsymbol{\phi_i(x)}$: basis function

w_i : weights

Linear : function is linear in the weights

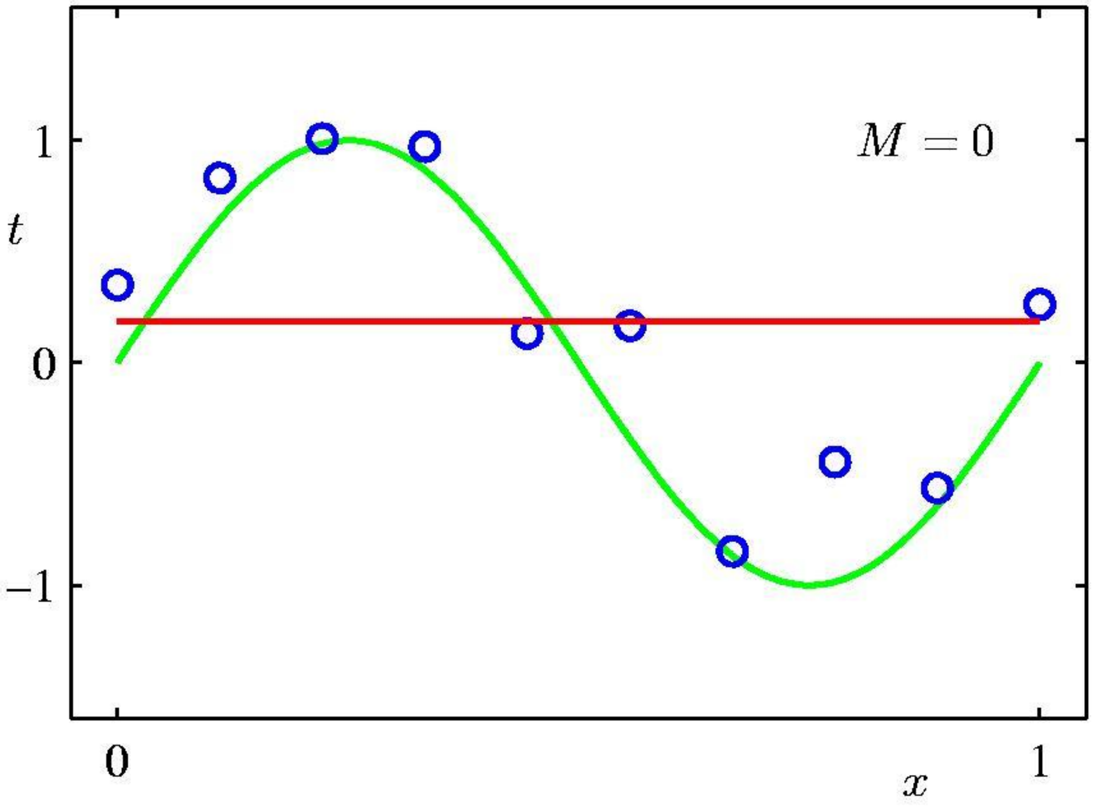
Quadratic error function --> derivative is linear in **w**

Sum-of-Squares Error Function

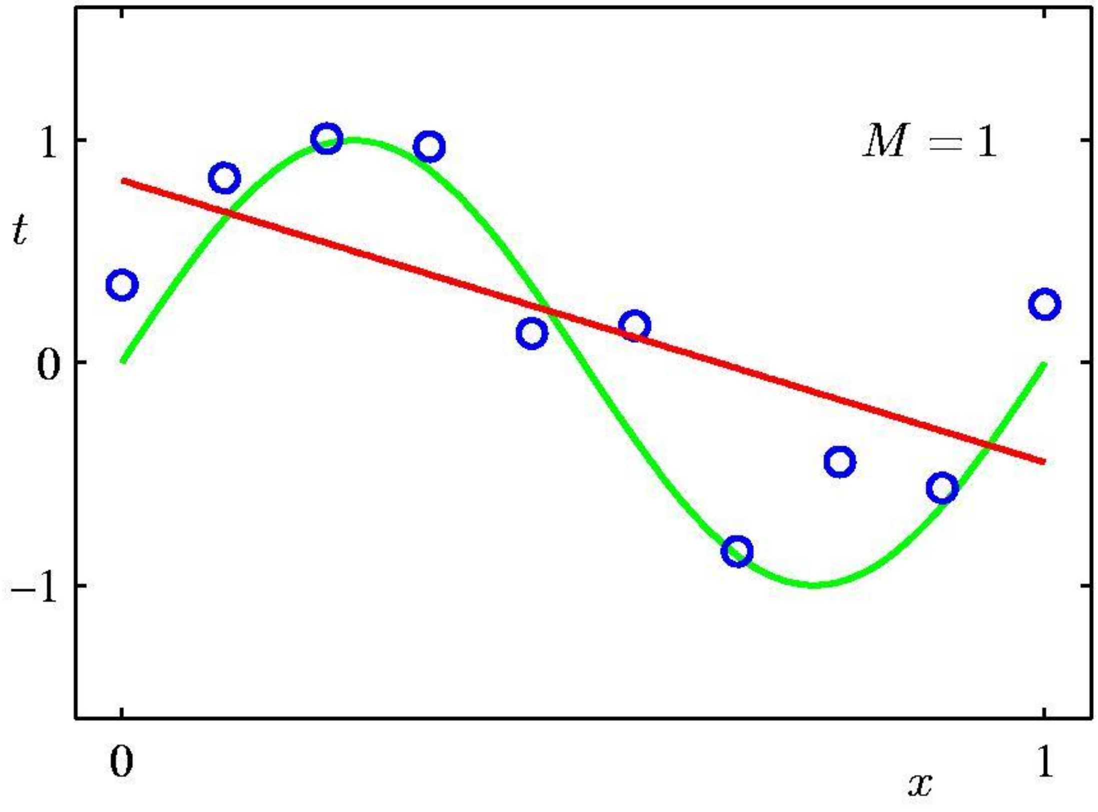


$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

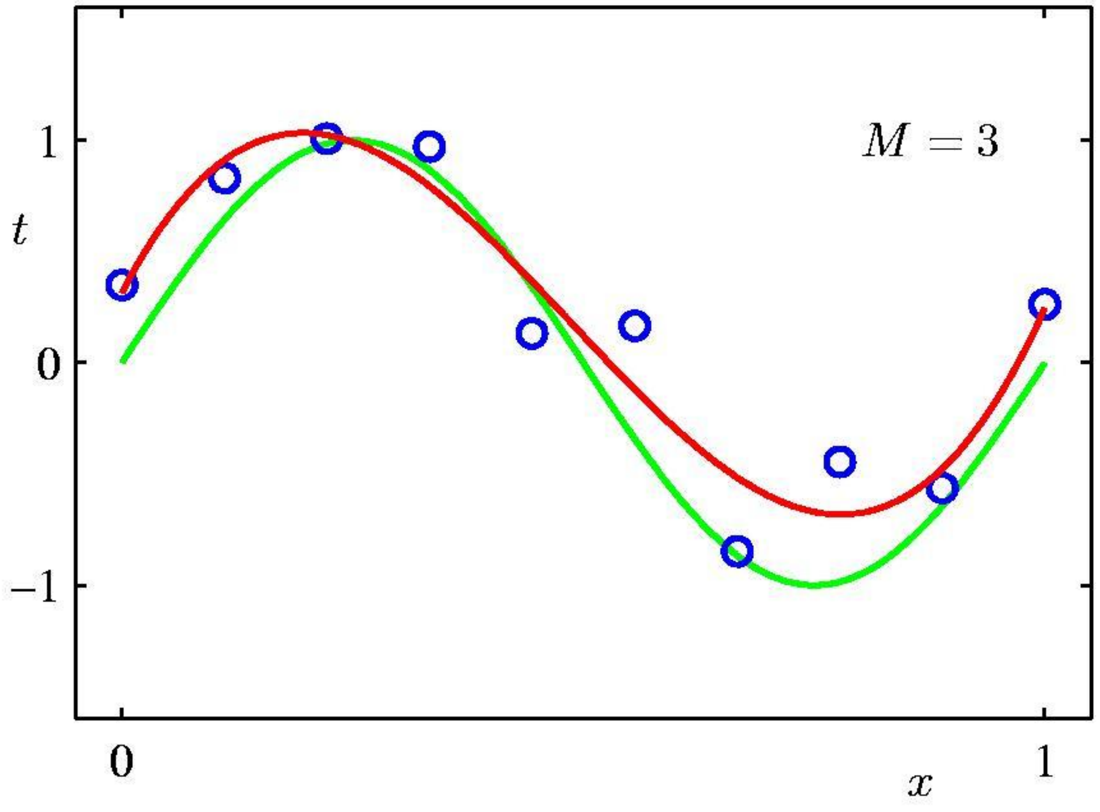
0^{th} Order Polynomial



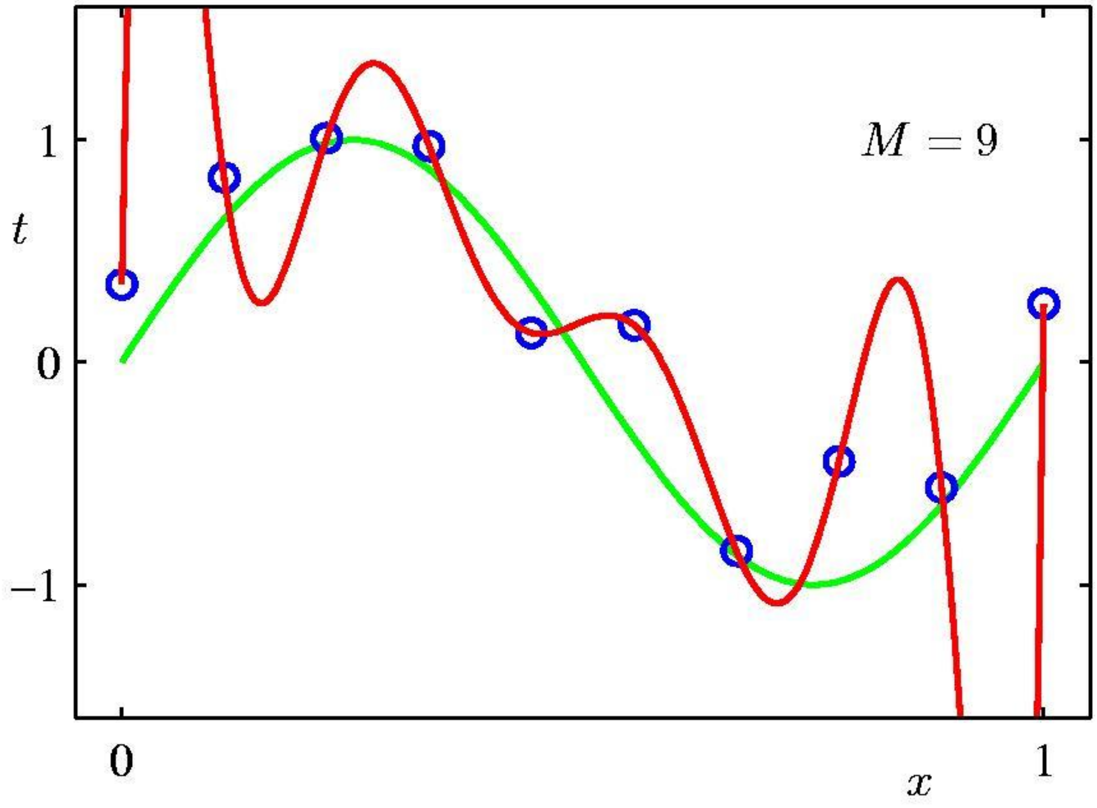
1^{st} Order Polynomial



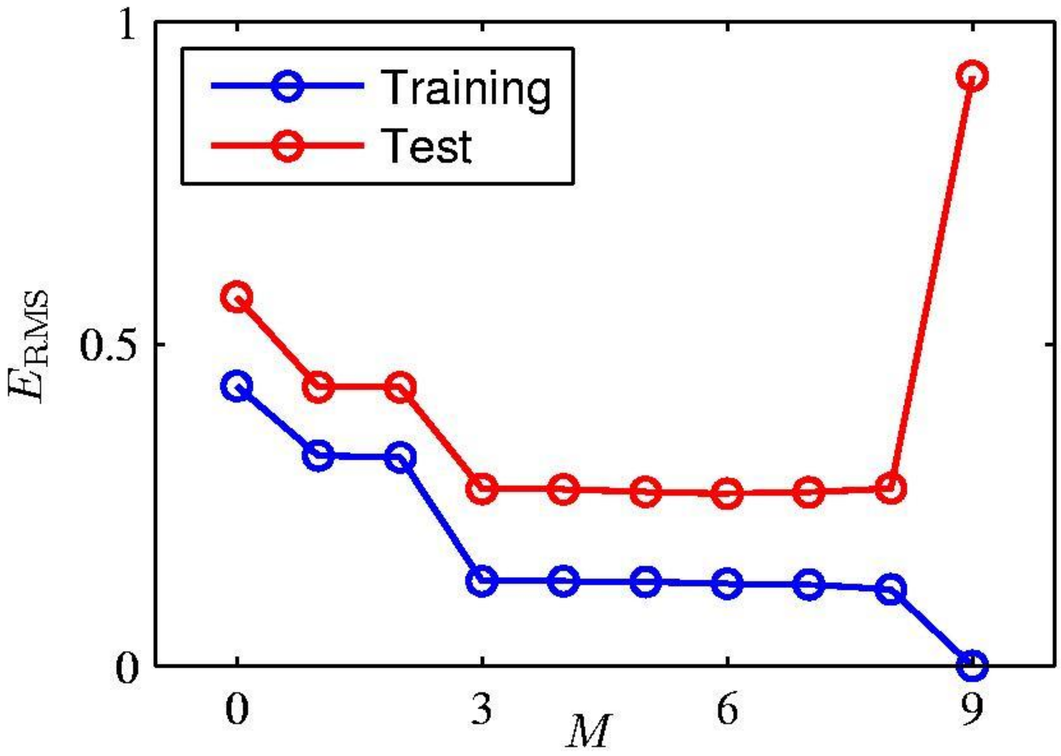
3rd Order Polynomial



9th Order Polynomial



Over-fitting

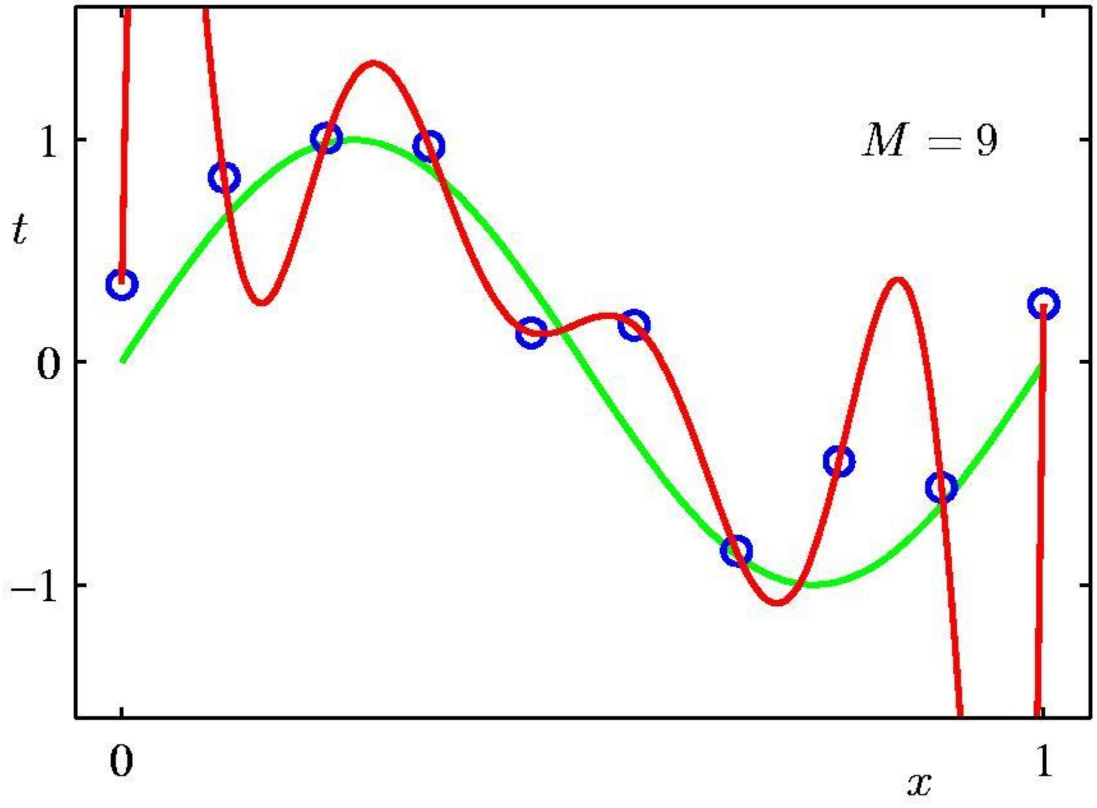


Root-Mean-Square (RMS) Error: $E_{\text{RMS}} = \sqrt{2E(\mathbf{w}^{\star})/N}$

Polynomial Coefficients

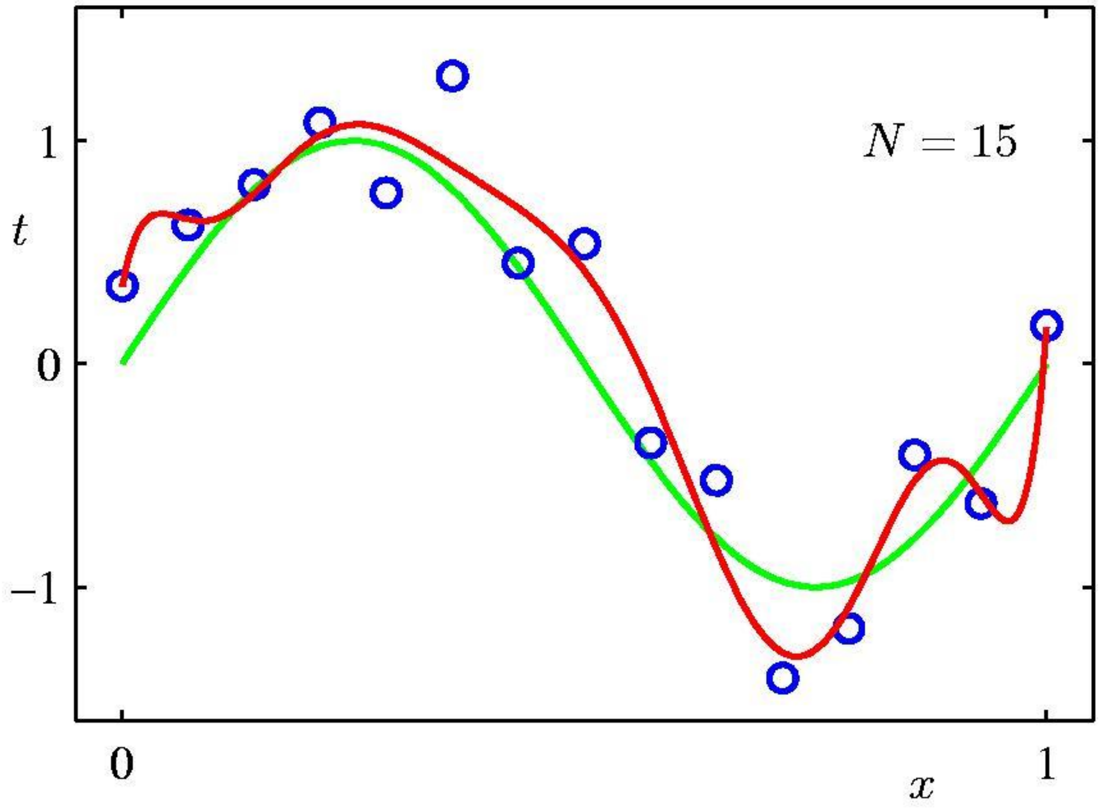
	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

9th Order Polynomial



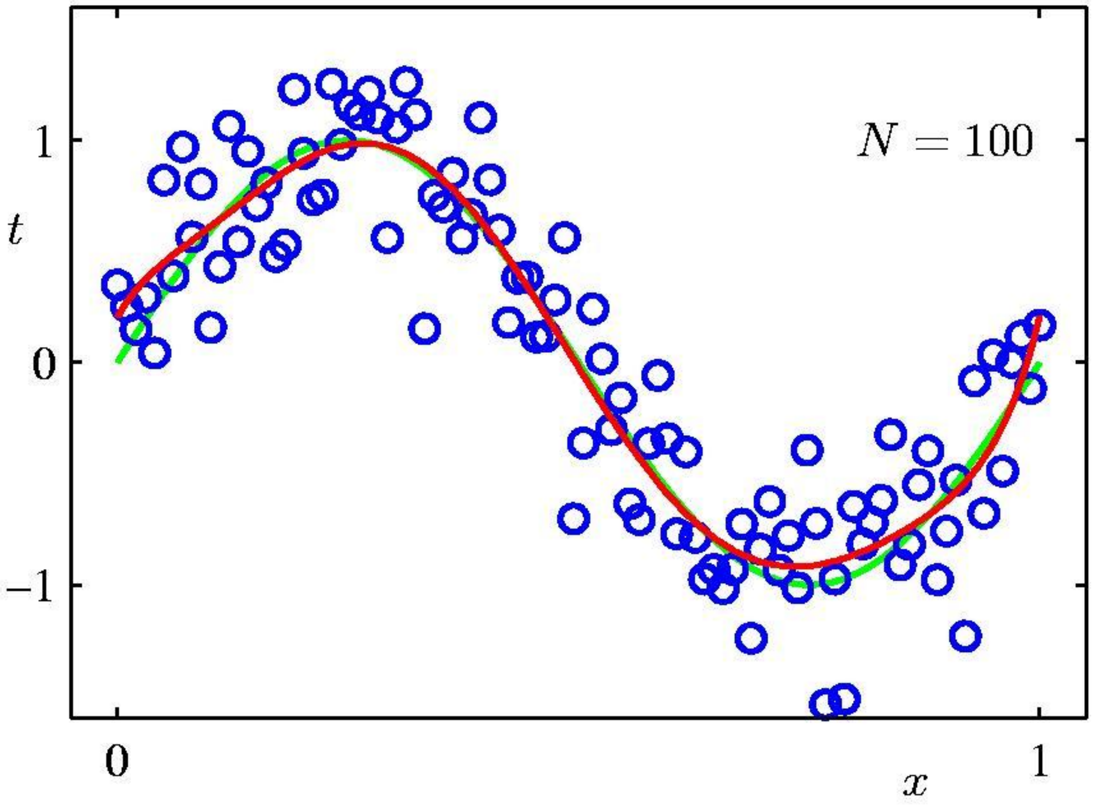
Data Set Size: $N = 15$

9^{th} Order Polynomial



Data Set Size: $N = 100$

9th Order Polynomial

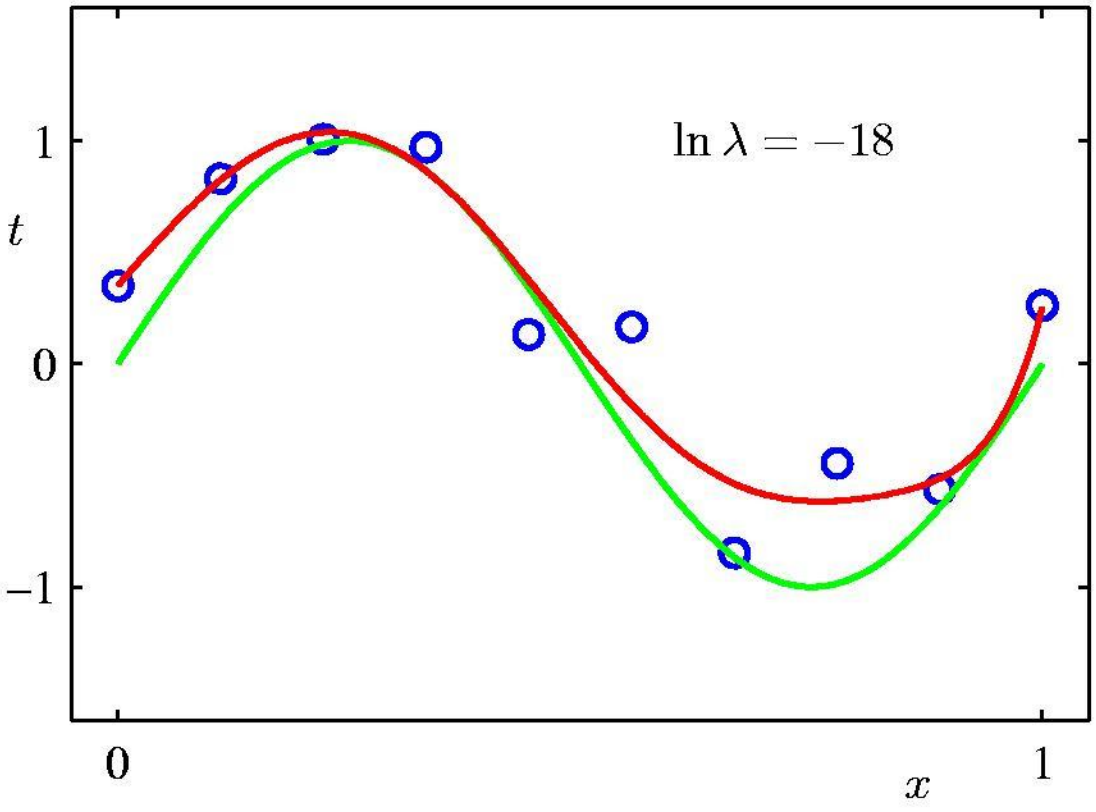


Regularization

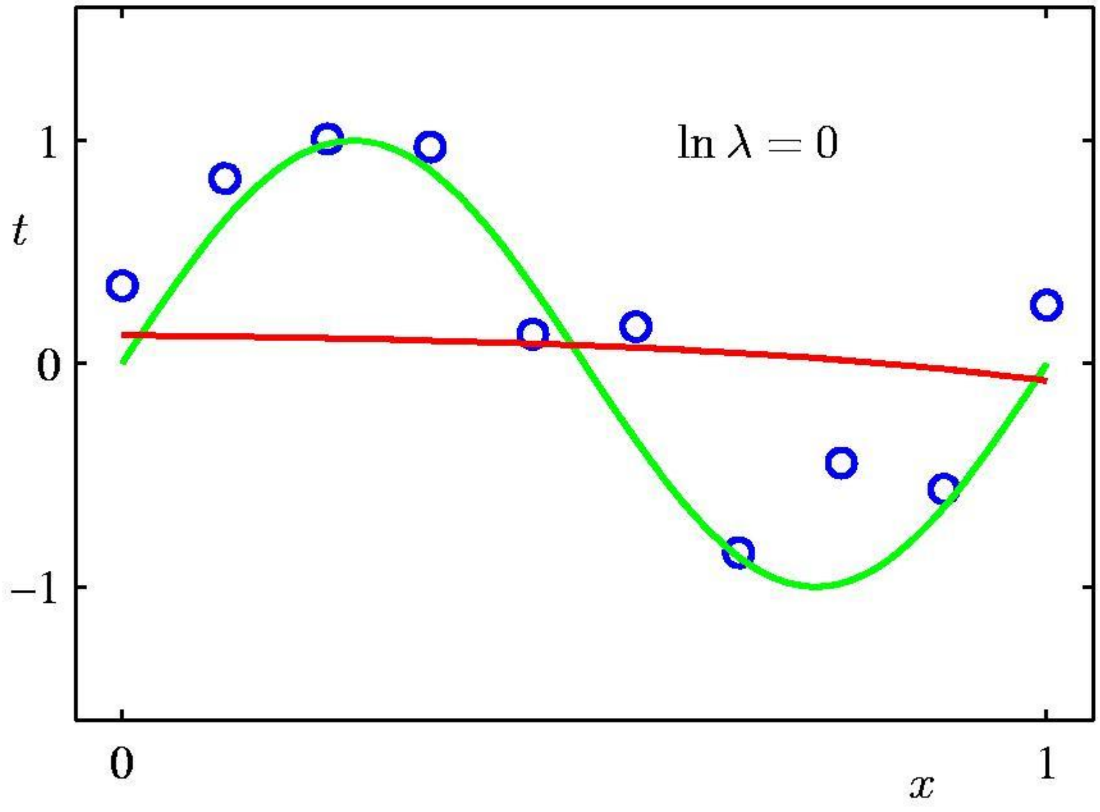
Penalize large coefficient values

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

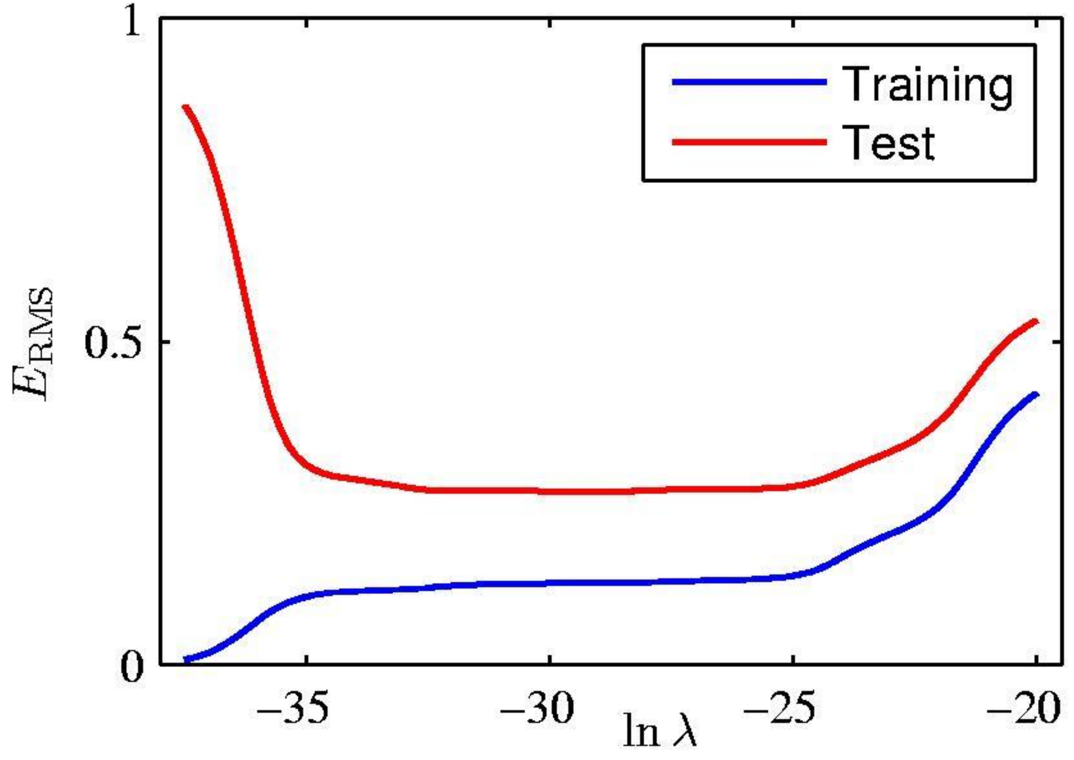
Regularization: $\ln \lambda = -18$



Regularization: $\ln \lambda = 0$



Regularization: E_{RMS} vs. $\ln \lambda$



Polynomial Coefficients

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

Probability Theory

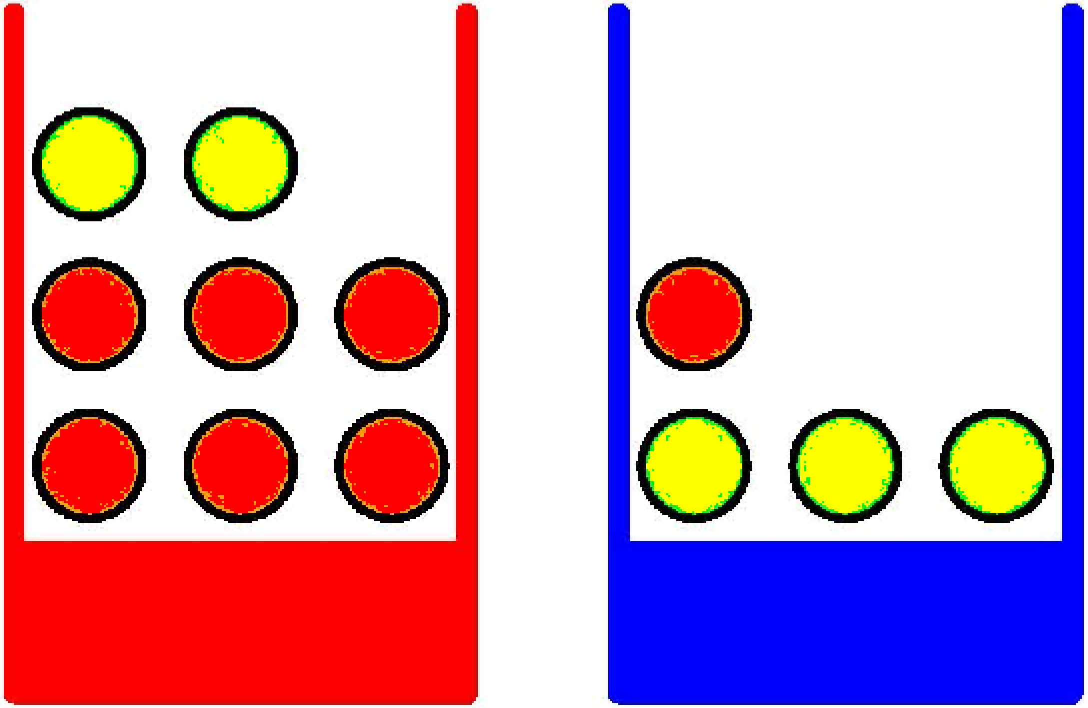
Learning = discovering regularities

- **Regularity** : repeated experiments:
  outcome not be fully predictable

  outcome = “possible world”
  set of all possible worlds = Ω

Probability Theory

Apples and Oranges



Sample Space

Sample ω = Pick two fruits,

e.g. Apple, then Orange

Sample Space $\Omega = \{(A,A), (A,O), (O,A),(O,O)\}$

= all possible worlds

Event e = set of possible worlds, $e \subseteq \Omega$

- e.g. second one picked is an apple

Learning = discovering regularities

- **Regularity** : repeated experiments:
  outcome not be fully predictable

- **Probability** p(e) : "the fraction of possible worlds in
  which e is true" i.e. outcome is event e

- **Frequentist** view : p(e) = limit as $N \to \infty$

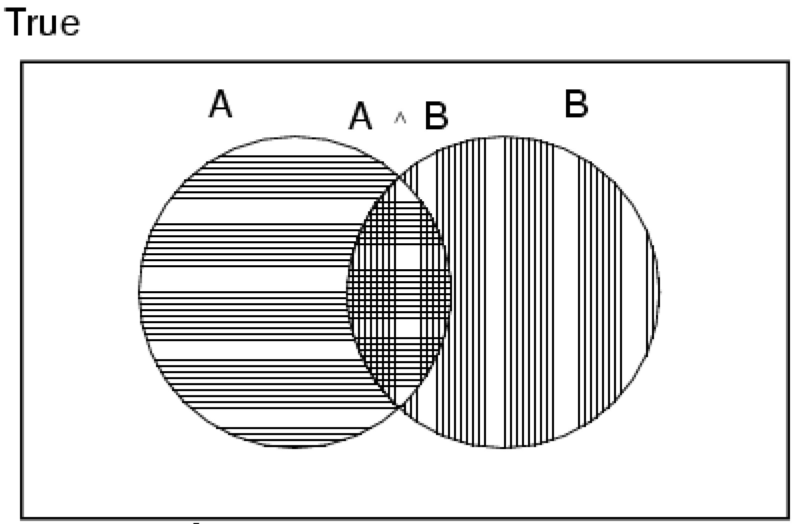
- **Belief** view: in wager : equivalent odds
  (1-p):p that outcome is in e, or vice versa

Axioms of Probability

- **non-negative** : $p(e) \geq 0$

- **unit sum** $p(\Omega) = 1$

i.e. no outcomes outside sar



- **additive** : if e1, e2 are disjoint events (no common outcome):

$$p(e1) + p(e2) = p(e1 \cup e2)$$

ALT:

$$p(e1 \vee e2) = p(e1) + p(e2) - p(e1 \wedge e2)$$

Why probability theory?

different methodologies attempted for uncertainty:

- Fuzzy logic
- Multi-valued logic
- Non-monotonic reasoning

But **unique property** of probability theory:

If you gamble using probabilities you have the best chance in a wager. [de Finetti 1931]

=> if opponent uses some other system, he's more likely to lose

Ramsay-diFinetti theorem (1931)

If agent X's degrees of belief are rational, then X 's degrees of belief function defined by fair betting rates is (formally) a probability function

Fair betting rates: opponent decides which side one bets on

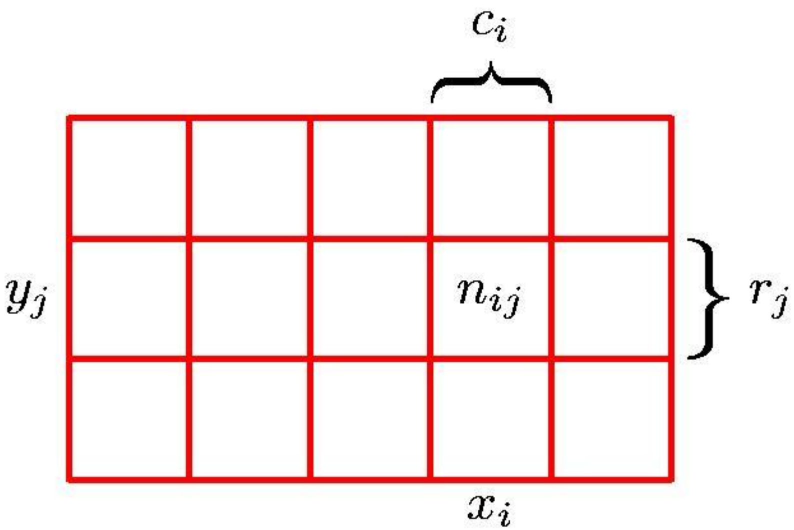
Proof: fair odds result in a function pr () that satisifies the Kolmogrov axioms:

Normality : pr(S) >=0

Certainty : pr(T)=1

Additivity : pr (S1 v S2 v..)= Σ(Si)

Joint vs. conditional probability



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

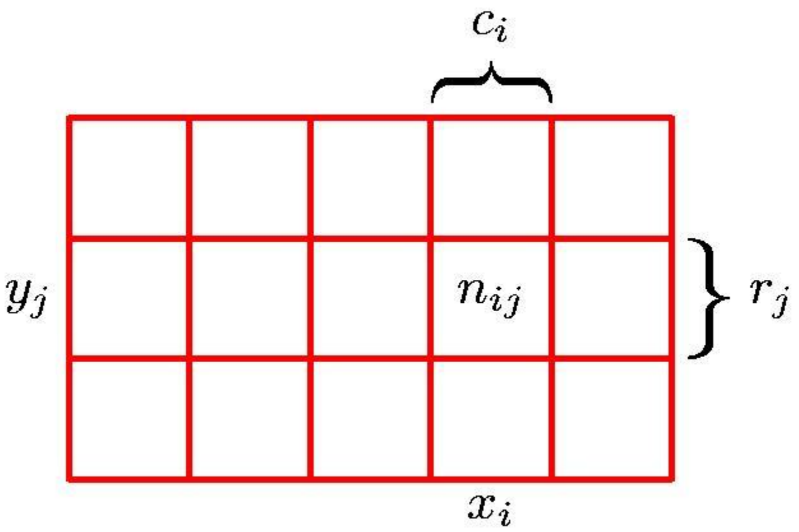
Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Probability Theory



Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned}$$

Rules of Probability

Sum Rule $$p(X) = \sum_Y p(X,Y)$$

Product Rule $$p(X,Y) = p(Y|X)p(X)$$

Example

A disease *d* occurs in 0.05% of population. A test is 99% effective in detecting the disease, but 5% of the cases test positive in absence of *d.*

10000 people are tested. How many are expected to test positive?

$p(d) = 0.0005$; $p(t/d) = 0.99$; $p(t/\sim d) = 0.05$

$p(t) = p(t,d) + p(t,\sim d)$ [Sum Rule]

$= p(t/d)p(d) + p(t/\sim d)p(\sim d)$ [Product Rule]

$= 0.99*0.0005 + 0.05 * 0.9995 = 0.0505$ → **505** +ve

Bayes' Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_{Y} p(X|Y)p(Y)$$

posterior $\propto$ likelihood × prior

Bayes' Theorem

Thomas Bayes (c.1750):

how can we infer causes from effects?

How can one learn the probability of a future event if one knew only

how many times it had (or had not) occurred in the past?

as new evidence comes in --> prob knowledge improves.

e.g. throw a die. guess is poor (1/6)

throw die again. is it > or < than prev? Can improve guess.

throw die repeatedly. can improve prob of guess quite a lot.

Hence: initial estimate (*prior* belief $P(h)$, not well formulated)

\+ new evidence (support) – compute likelihood $P(data|h)$

→ improved estimate (*posterior* $P(h|data)$)

Example

A disease *d* occurs in 0.05% of population. A test is 99% effective in detecting the disease, but 5% of the cases test positive in absence of *d.*

If you are tested +ve, what is the probability you have the disease?

p(d/t) = p(d) . p(t/d) / p(t) ; p(t) = 0.0505

p(d/t) = 0.0005 * 0.99 / 0.0505 = 0.0098 (about **1%**)

if 10K people take the test, E(d) = 5
FPs = 0.05 * 9995 = 500
TPs = 0.99 * 5 = 5. ➔ only **5/505** have *d*

Bayesian Inference

Testing for hypothesis H given evidence E

- **Evidence** : based on new observation E
- **Prior** : Earlier evaluation about the probability of H
- **Likelihood** : probability of evidence given hypothesis

$$P(E|H)$$

normalization(
(marginal lklihood)

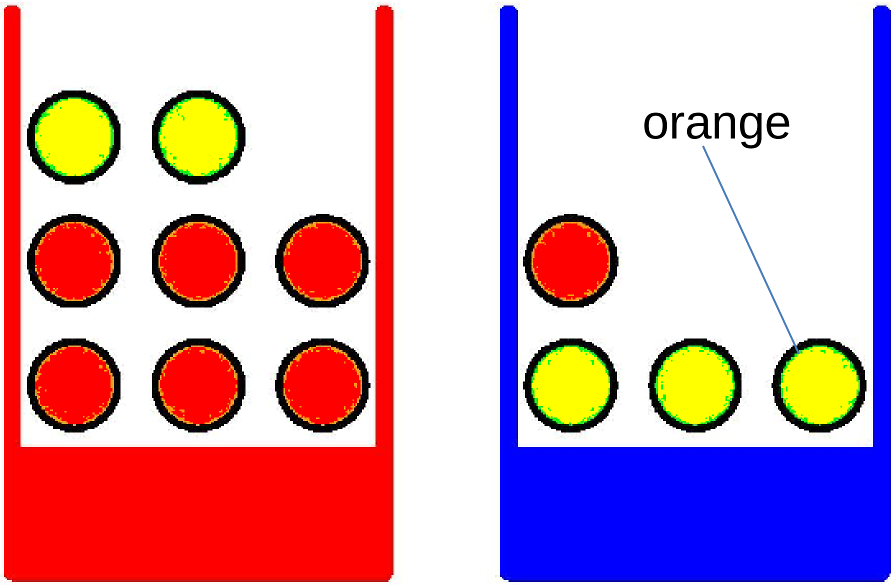
Bayesian inference:

$$P(H|E) = P(E|H)\, P(H) / P(E)$$

Posterior probability

Bayesian Inference

The fruit picked is an orange (o). What is the probability that it's from the blue box (B)?



P(B|o) =
P(o|B)p(B) / P(o)

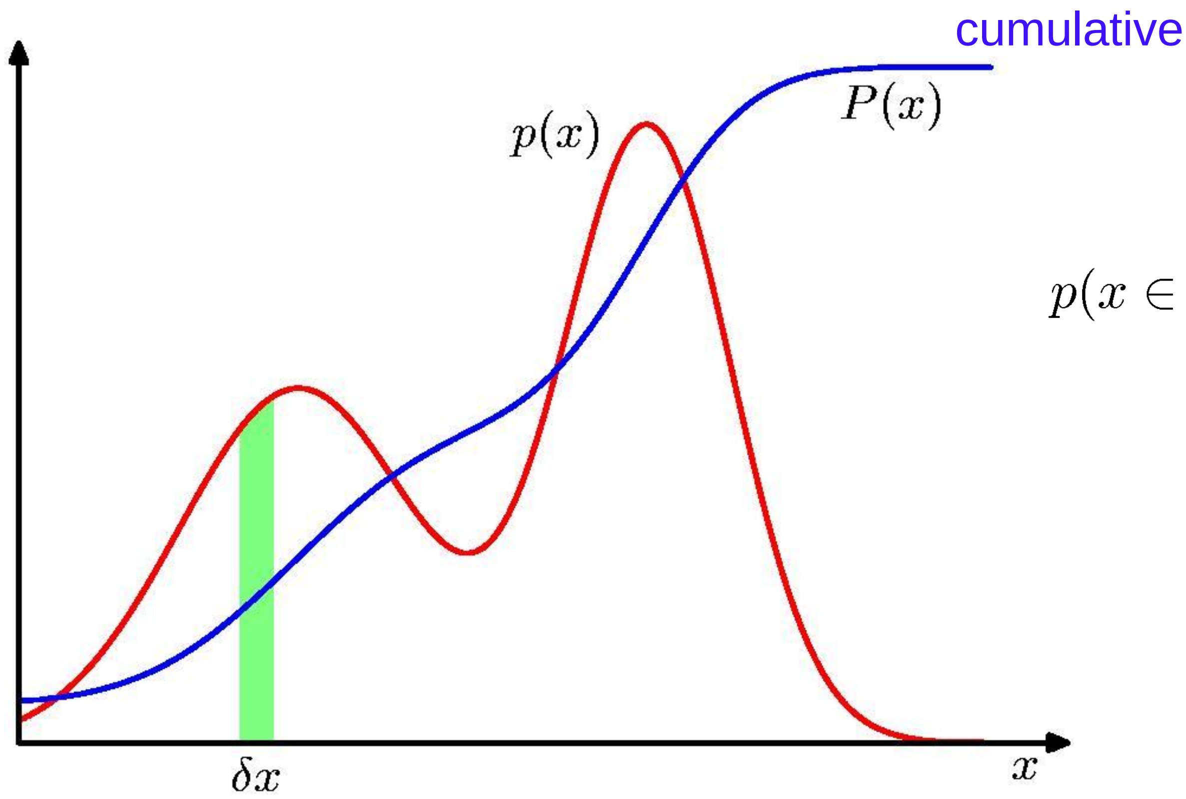
Given: red box is picked 40% → p(B) = 0.6

P(o) = (¾*.6 + 1/3*0.4) = 11/20

P(B|o) = ¾ * .6 * 20/11 = 9/11

Continuous variables: Probability Densities

Probability Densities



$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

$$p(x) \geqslant 0$$

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Expectations

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

discrete x

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x$$

continuous x

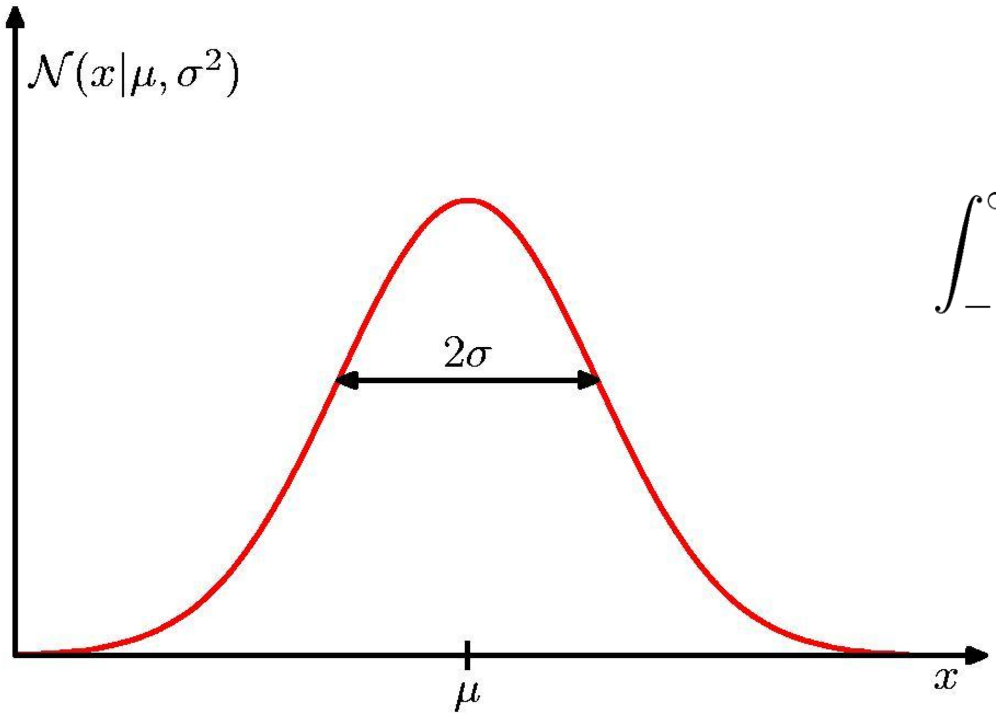
Frequentist approximation w unbiased sample

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

(both discrete / continuous)

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

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

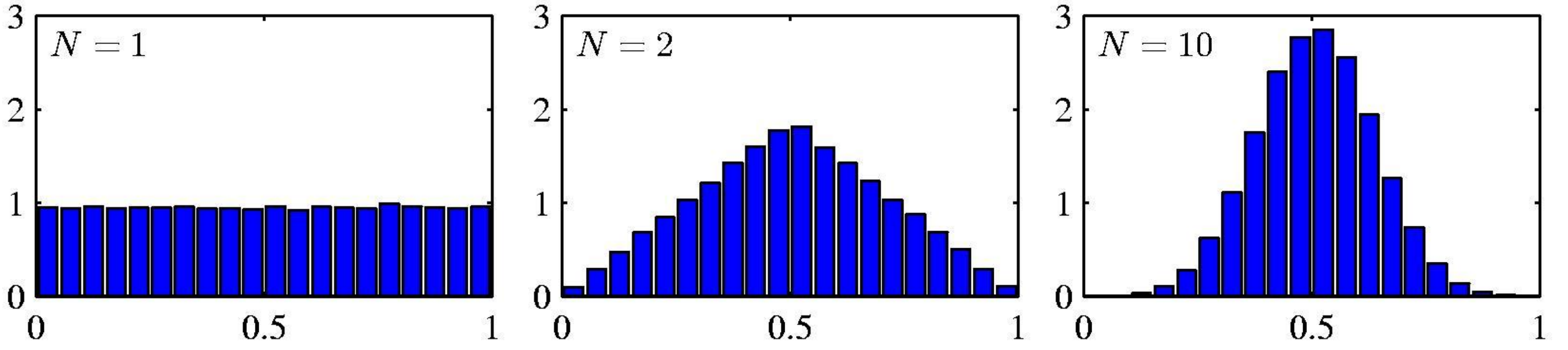
$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

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

Central Limit Theorem

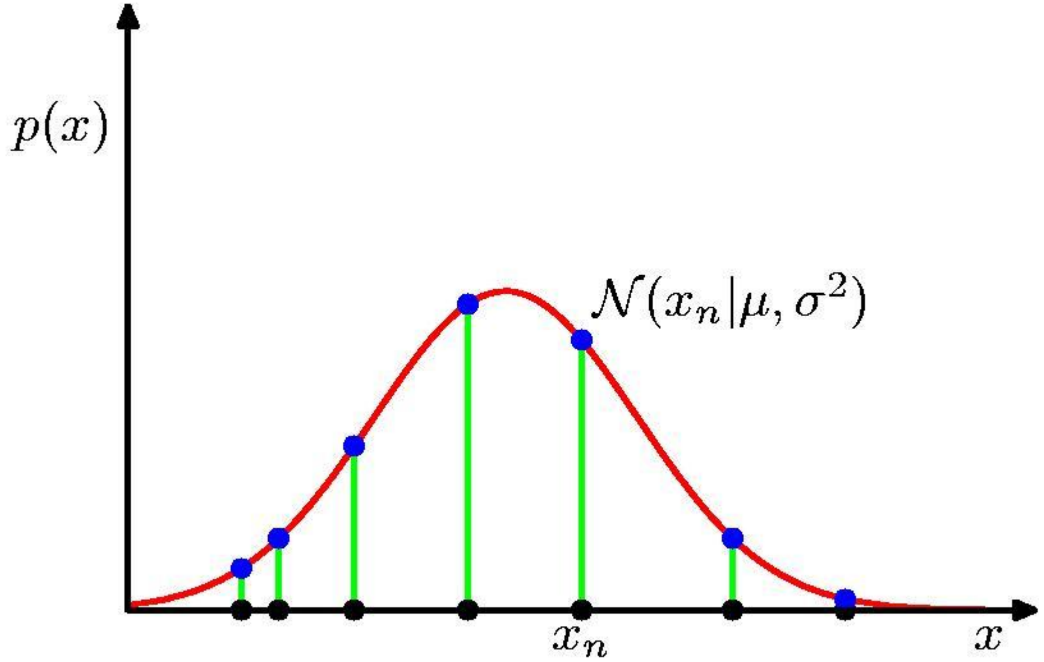
Distribution of sum of N i.i.d. random variables becomes increasingly Gaussian for larger N.

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



Gaussian Parameter Estimation

Observations assumed to be indpendently drawn from same distribution (i.i.d)



Likelihood function

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

Maximum (Log) Likelihood

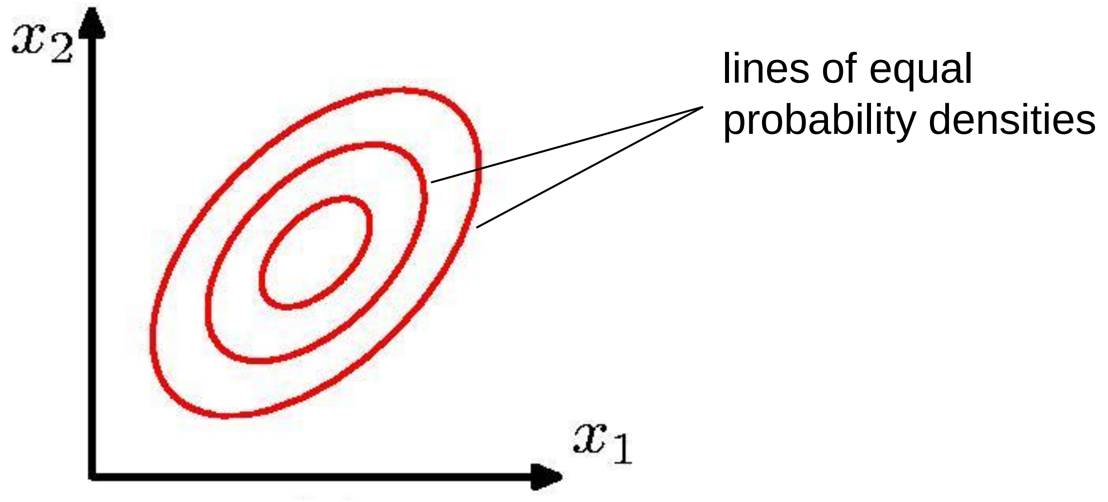
$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

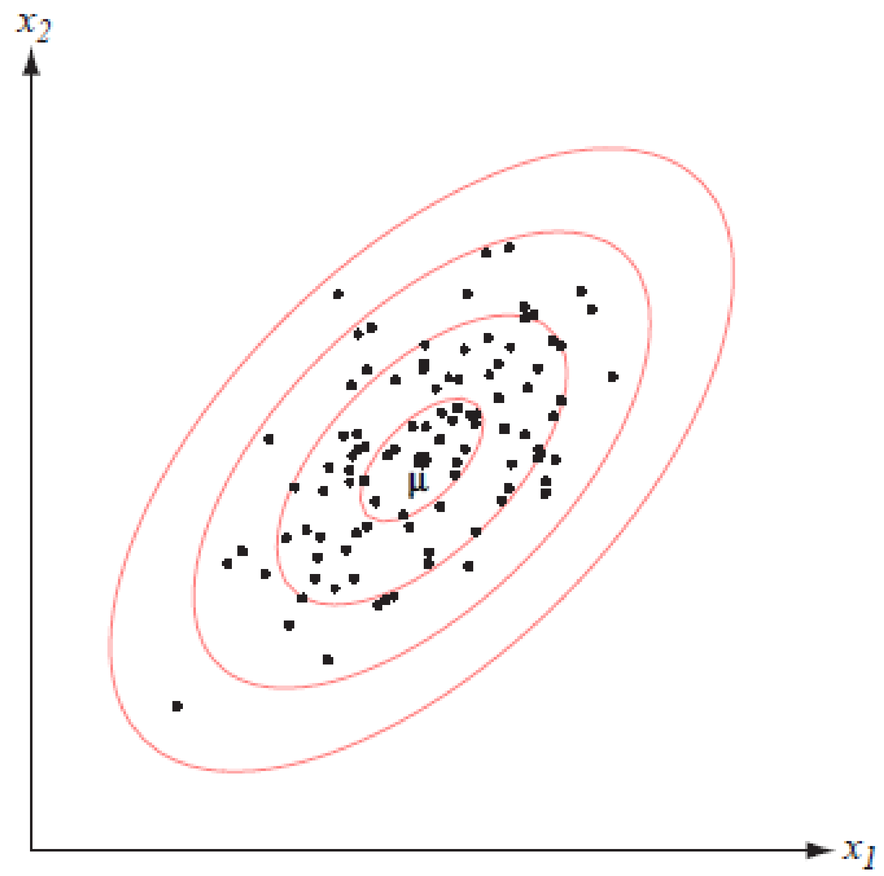
$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

Distributions over Multi-dimensional spaces

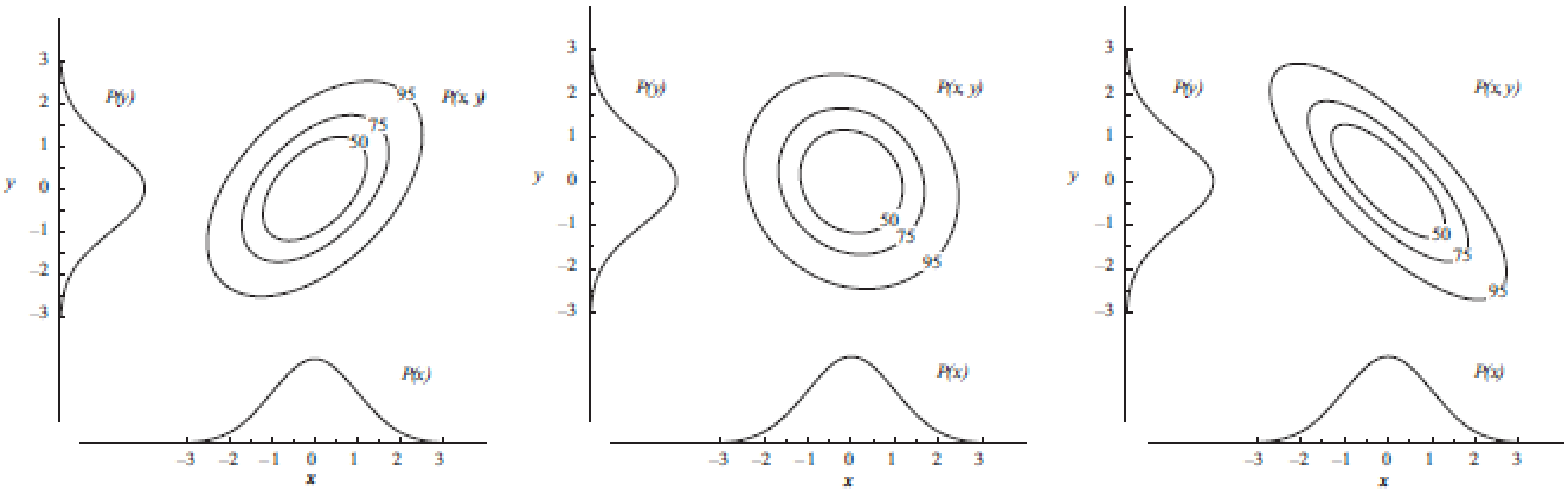
The Multivariate Gaussian

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




x2
x1
μ

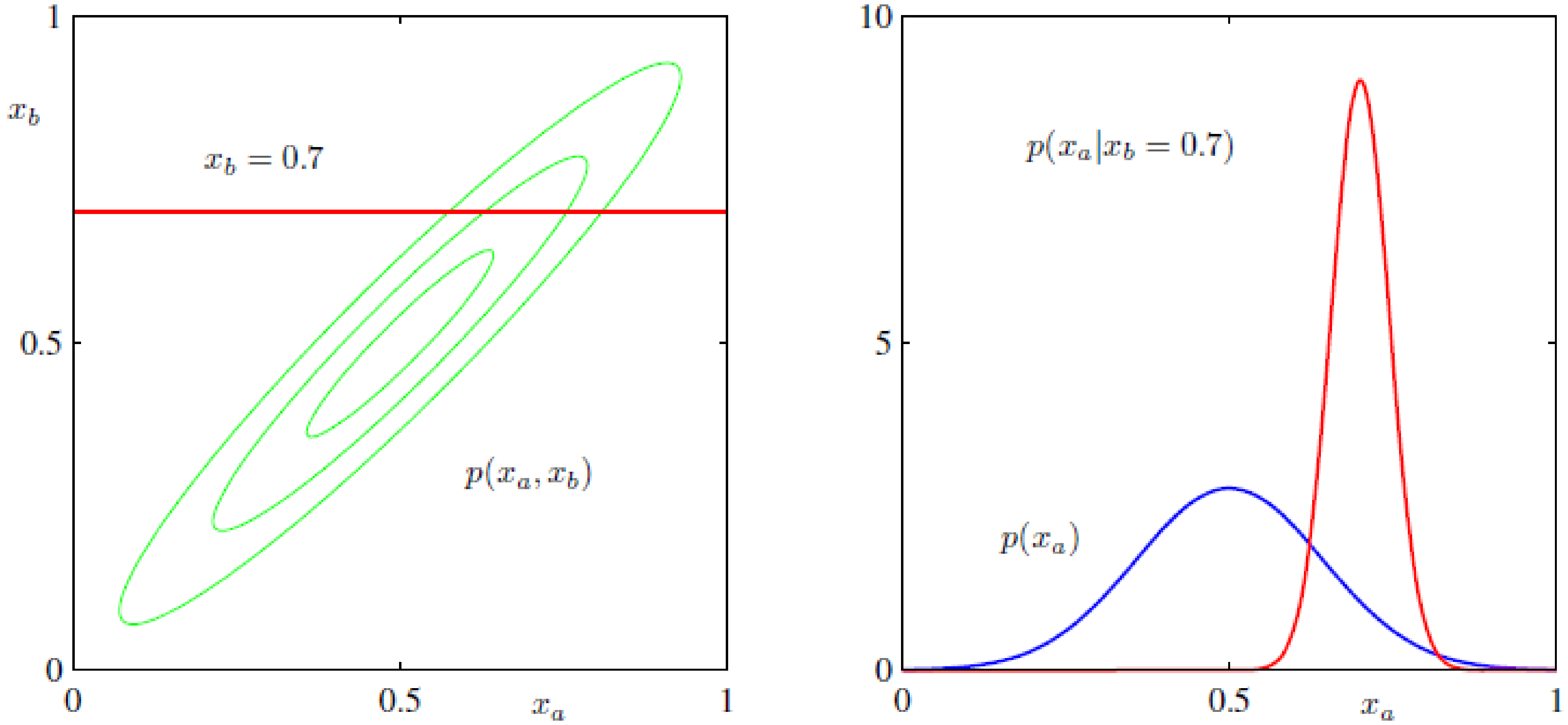
Multivariate distribution



joint distribution P(x,y) varies considerably
though marginals P(x), P(y) are identical

estimating the joint distribution requires
much larger sample: $O(n^k)$ vs nk

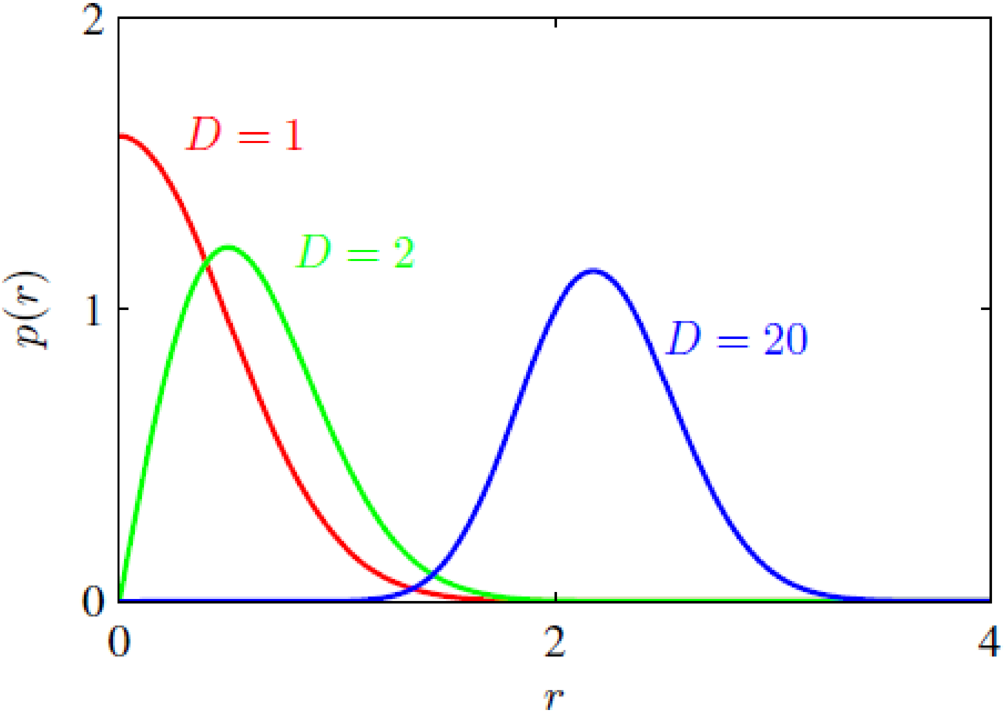
Marginals and Conditionals



marginals P(x), P(y) are gaussian
conditional P(x|y) is also gaussian

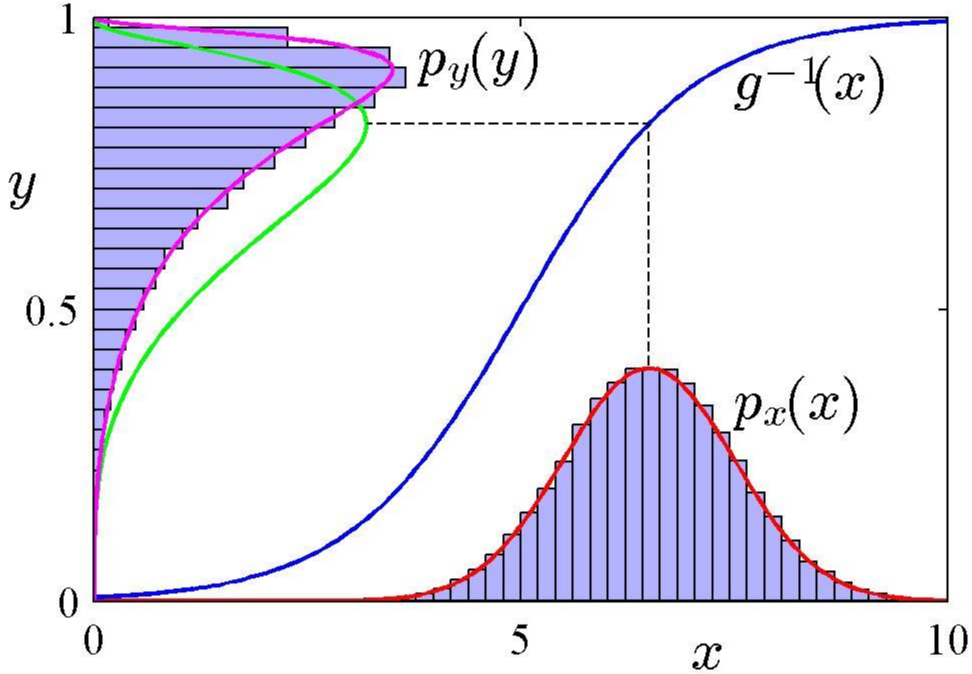
Non-intuitive in high dimensions

As dimensionality increases, bulk of data moves away from center



Gaussian in polar coordinates;
p(r)δr : prob. mass inside annulus *δr* at *r*.

Change of variable x=g(y)



$$p_y(y) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = p_x(g(y))\,|g'(y)|$$

Bernoulli Process

Successive Trials – e.g. Toss a coin three times:

HHH, HHT, HTH, HTT, THH, THT, TTH, TTT

Probability of k Heads:

k	0	1	2	3
P(k)	1/8	3/8	3/8	1/8

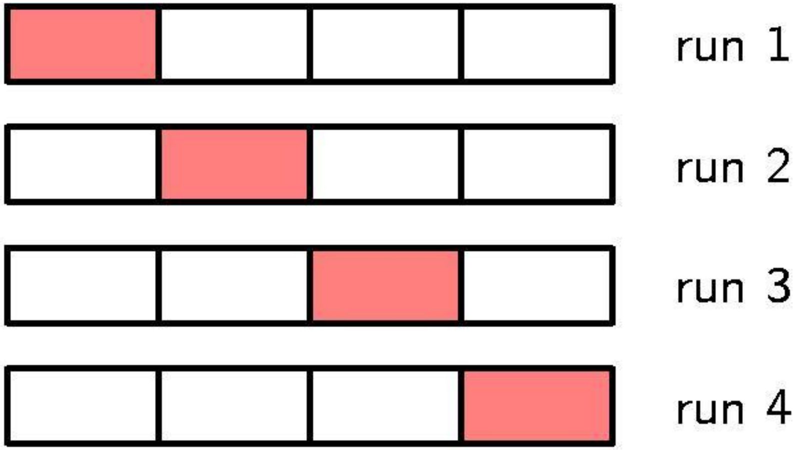
Probability of success: p, failure q, then

$$P(k) = \binom{n}{k} p^k q^{n-k}$$

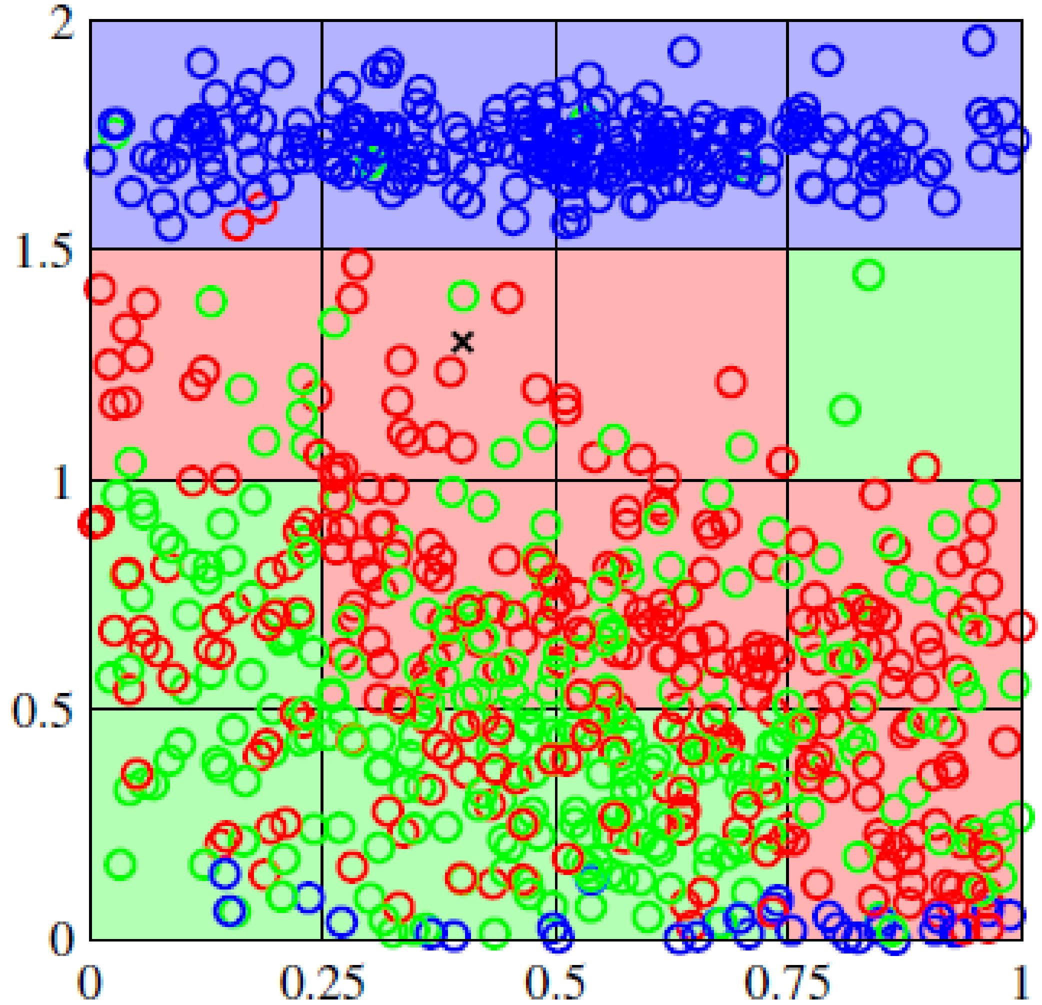
Model Selection

Model Selection

Cross-Validation



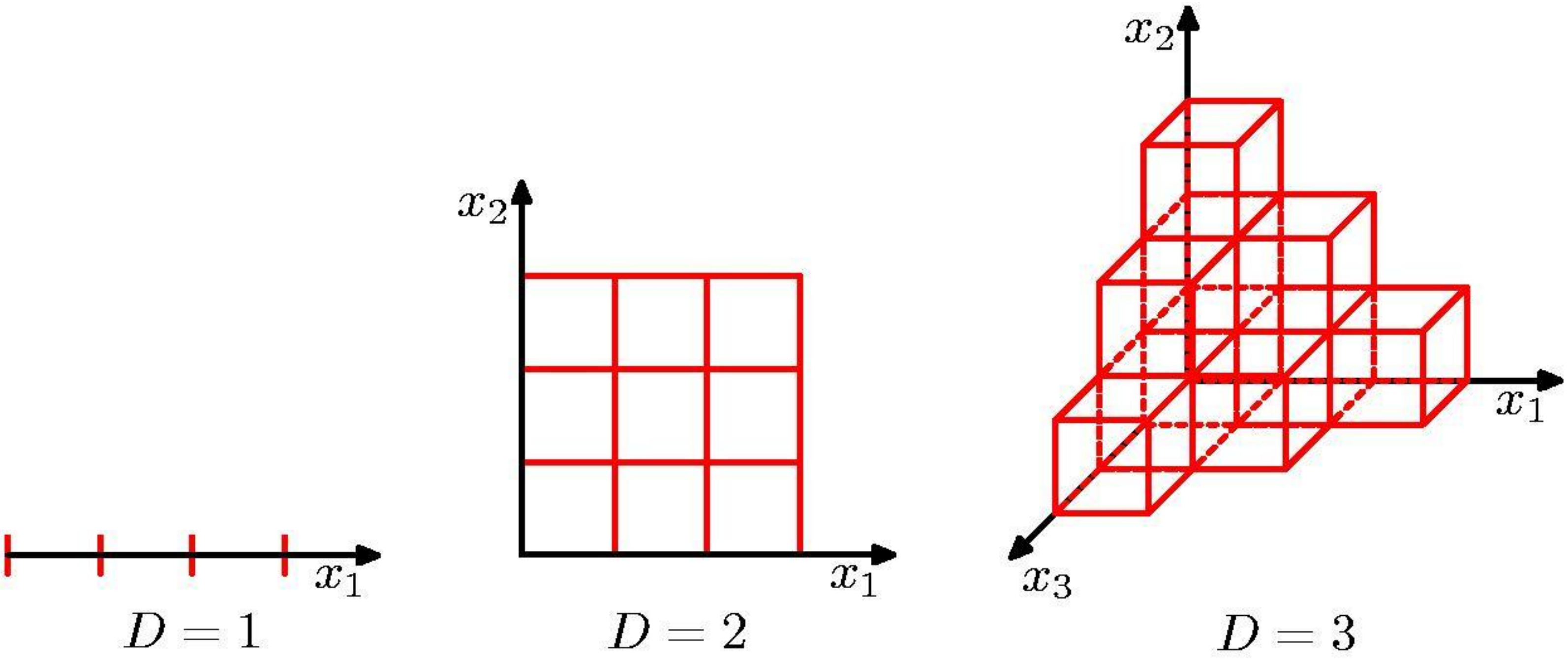
Quantized-Cell Classification



flow data

red: ‘homogenous’,
green : ‘annular’,
blue : ‘laminar’.

Curse of Dimensionality

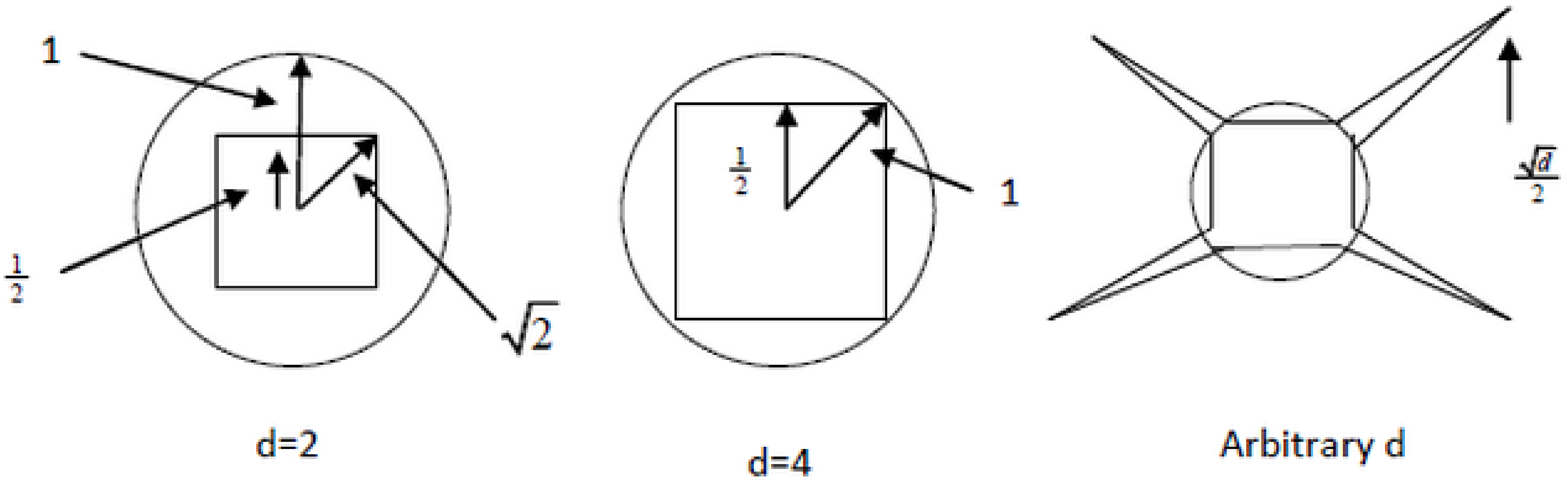


general cubic **polynomial** for D dimensions : *O(D³)* parameters

$$w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

Curse of Dimensionality

The unit hyper cube and unit sphere in high dimensions



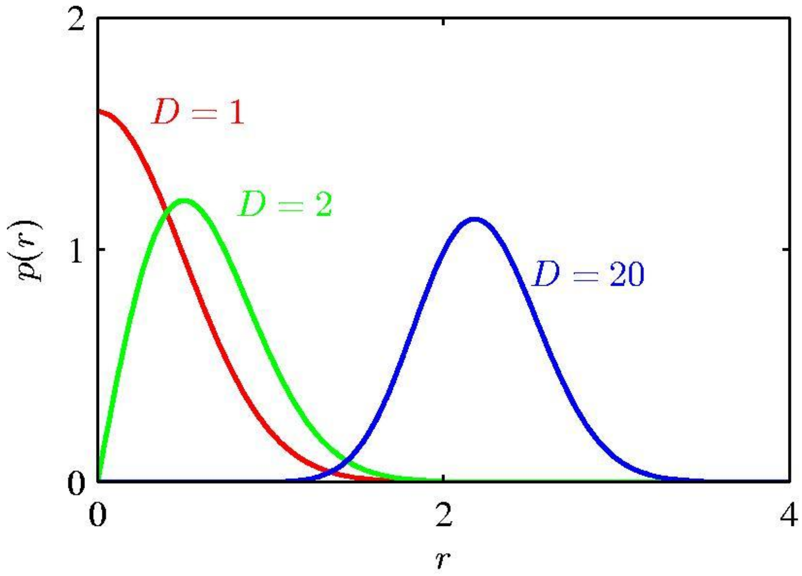
At higher dim, vol(sphere) / vol(hypercube) → 0

Curse of Dimensionality

Polynomial curve fitting, M = 3

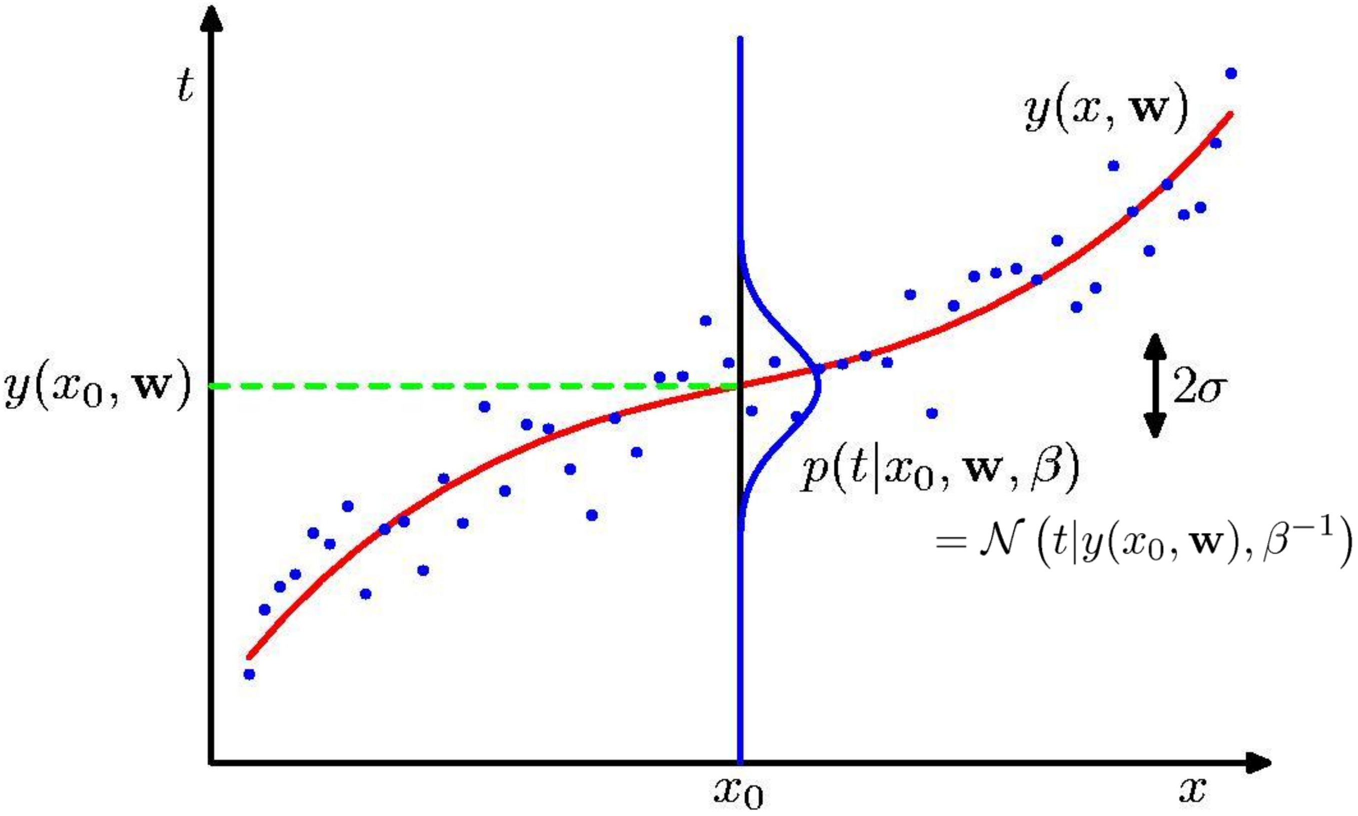
$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

Gaussian Densities in higher dimensions



Regression with Polynomials

Curve Fitting Re-visited



Bayesian Inference

Testing for hypothesis H given evidence E

likelihood

Bayesian inference:

$$P(H|E) = P(E|H)\,P(H) / P(E)$$

prior

posterior

Maximum Likelihood

Evidence = ***t***; Hypothesis = *poly(**x**,**w**)*

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

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

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

Maximum Likelihood

Evidence = ***t***; Hypothesis = *poly(**x**,**w**)*

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

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

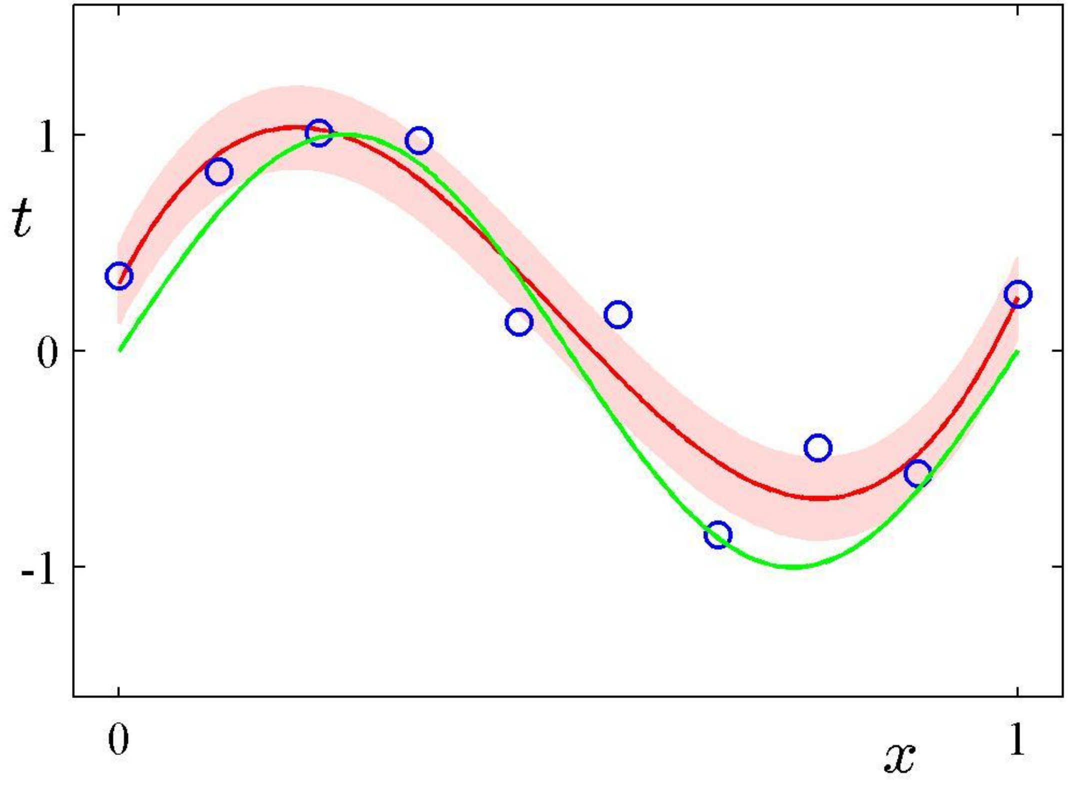
Determine $\mathbf{w}_{\mathrm{ML}}$ by minimizing sum-of-squares error, $E(\mathbf{w})$

.

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

Predictive Distribution

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



MAP: A Step towards Bayes

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

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

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

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

MAP = Maximum Posterior

Bayesian Curve Fitting

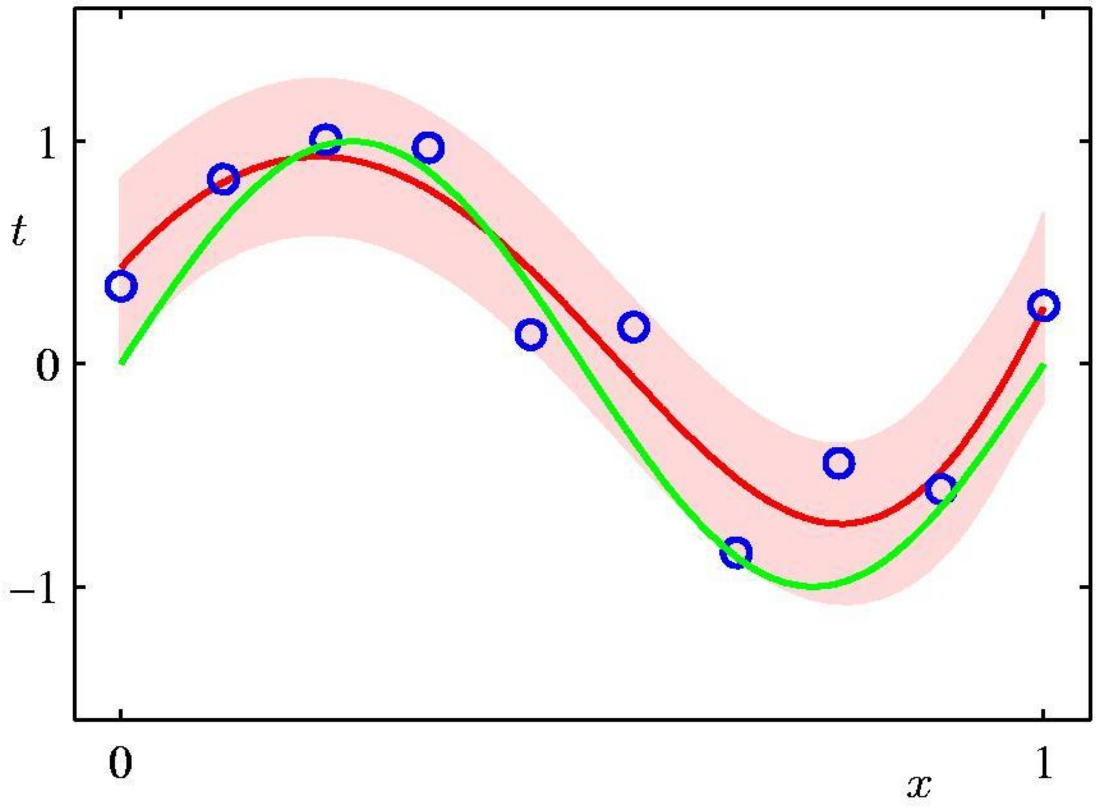
$$p(t|x,\mathbf{x},\mathbf{t}) = \int p(t|x,\mathbf{w})p(\mathbf{w}|\mathbf{x},\mathbf{t})\,\mathrm{d}\mathbf{w} = \mathcal{N}\left(t|m(x),s^2(x)\right)$$

$$m(x) = \beta\boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)t_n \qquad s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\boldsymbol{\phi}(x)$$

$$\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^{\mathrm{T}} \qquad \boldsymbol{\phi}(x_n) = \left(x_n^0,\ldots,x_n^M\right)^{\mathrm{T}}$$

Bayesian Predictive Distribution

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right)$$



Information Theory

Twenty Questions

Knower: thinks of object (point in a probability space)
Guesser: asks knower to evaluate random variables

Stupid approach:

Guesser: Is it my left big toe?
Knower: No.

Guesser: Is it Valmiki?
Knower: No.

Guesser: Is it Aunt Lakshmi?
...

Expectations & Surprisal

Turn the key: expectation: lock will open

Exam paper showing: could be 100, could be zero.
random variable: function from set of marks
to real interval [0,1]

Interestingness $\propto$ unpredictability

$$\textit{surprisal}\ (r.v. = x) = -\log_2 p(x)$$

$= 0$ when $p(x) = 1$
$= 1$ when $p(x) = \frac{1}{2}$
$= \infty$ when $p(x) = 0$

Expectations in data

A: 00010001000100010001. . . 000100010001000100010001

B: 01110100110100100110. . . 1010111010111011000101100010

C: 00011000001010100000. . . 001000100001000000001000110000

Structure in data → easy to remember

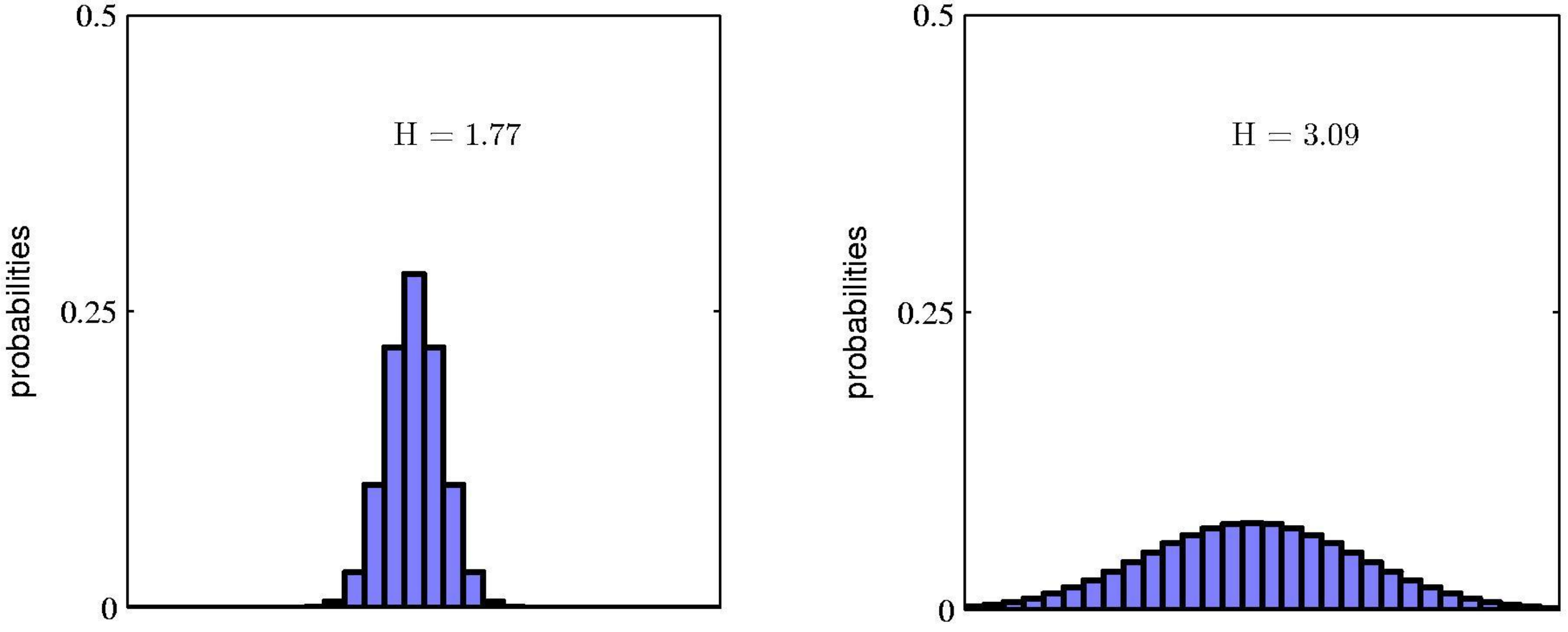
Entropy

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

Used in

- coding theory
- statistical physics
- machine learning

Entropy



Entropy

In how many ways can N identical objects be allocated M bins?

$$W = \frac{N!}{\prod_i n_i!}$$

$$\mathrm{H} = \frac{1}{N} \ln W \simeq - \lim_{N \to \infty} \sum_i \left(\frac{n_i}{N} \right) \ln \left(\frac{n_i}{N} \right) = - \sum_i p_i \ln p_i$$

Entropy maximized when $\forall i : p_i = \frac{1}{M}$

Entropy in Coding theory

x discrete with 8 possible states; how many bits to transmit the state of x?

All states equally likely

$$\mathrm{H}[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits.}$$

Coding theory

x	a	b	c	d	e	f	g	h
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$
code	0	10	110	1110	111100	111101	111110	111111

$$\begin{aligned} \mathrm{H}[x] &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

$$\begin{aligned} \text{average code length} &= \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}\times 6 \\ &= 2 \text{ bits} \end{aligned}$$

Entropy in Twenty Questions

Intuitively : try to ask q whose answer is 50-50

Is the first letter between A and M?

question entropy = p(Y)logp(Y) + p(N)logP(N)

For both answers equiprobable:

entropy = - ½ * $\log_2$(½) - ½ * $\log_2$(½) = 1.0

For P(Y)=1/1028

entropy = - 1/1028 * -10 - eps = 0.01