Probability

Introduction to Graphical Models

Prof. Alexander Ihler

BREN:ICS INFORMATION AND COMPUTER SCIENCES

UNIVERSITY of CALIFORNIA IRVINE

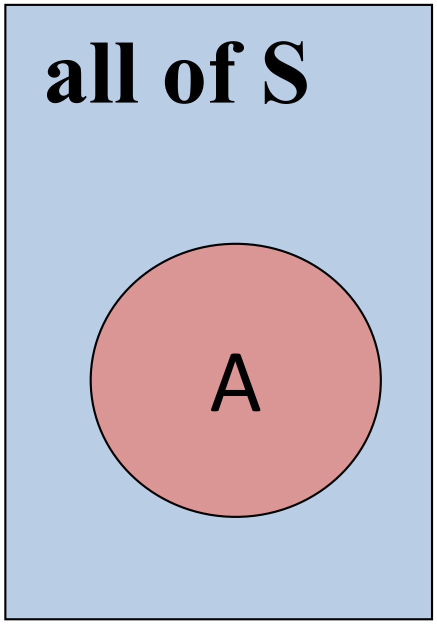
Uncertainty in the world

- Uncertainty due to
  - Randomness
  - Overwhelming complexity
  - Lack of knowledge
  - …
- Example: time to the airport
- Without representing & communicating uncertainty, it’s easy to make and compound mistakes
- Probability gives
  - natural way to describe our assumptions
  - rules for how to combine information

Probability

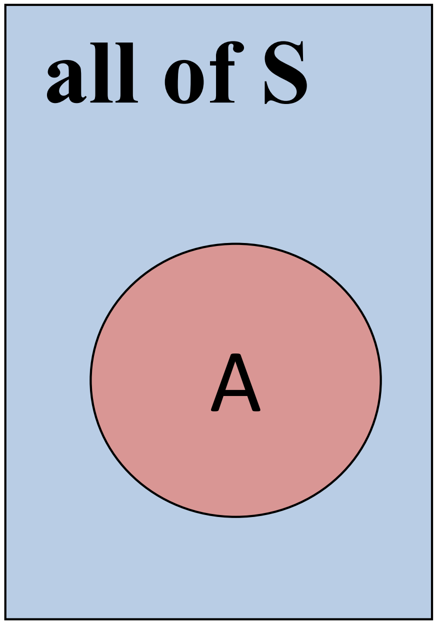
- Event “A” in event space “S”
  - Ex: “I have a headache”
  - Ex: “I have the flu”
  - Ex: “I have Ebola”

- Probability Pr[A]
  - Think of e.g. “# of worlds in which A happens”
  - This is a measure, like area
  - Can think of it in those terms



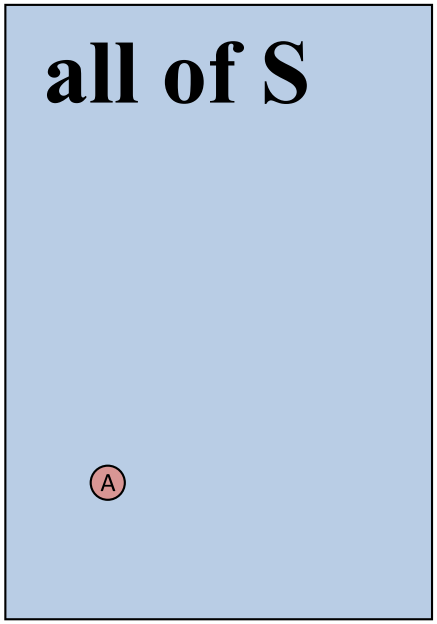
Probability

- Event "A" in event space "S"
- Probability Pr[A]
- Axioms of probability
  - $0 \le Pr[A] \le 1$
  - $Pr[S] = 1$
  - $Pr [\emptyset] = 0$
  - $Pr[A \cup B] = Pr[A] + Pr[B] - Pr[A \cap B]$



Probability

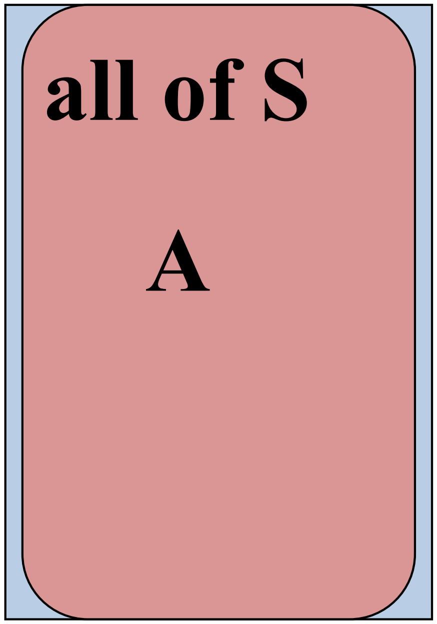
- Event "A" in event space "S"
- Probability Pr[A]
- Axioms of probability
  - $0 \leq \Pr[A] \leq 1$
  - $\Pr[\, S \,] = 1$
  - $\Pr[\, \emptyset \,] = 0$
  - $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$



"A" can't get any smaller
than size zero…
No worlds in which "A" is true

Probability

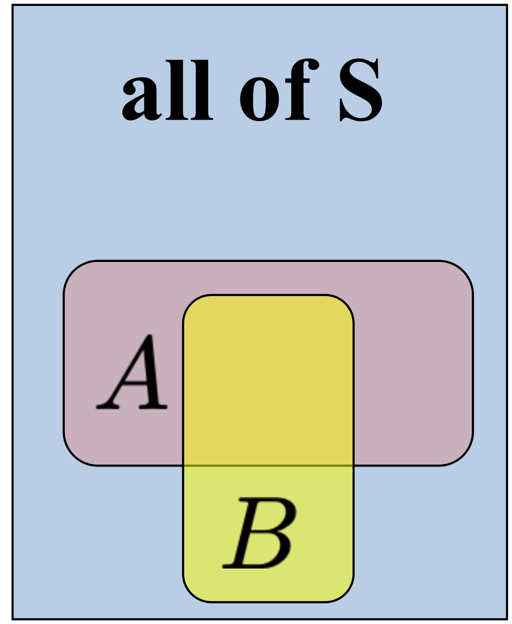
- Event “A” in event space “S”
- Probability Pr[A]
- Axioms of probability
  - $0 \leq \Pr[A] \leq 1$
  - $\Pr[\,S\,] = 1$
  - $\Pr[\,\emptyset\,] = 0$
  - $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$

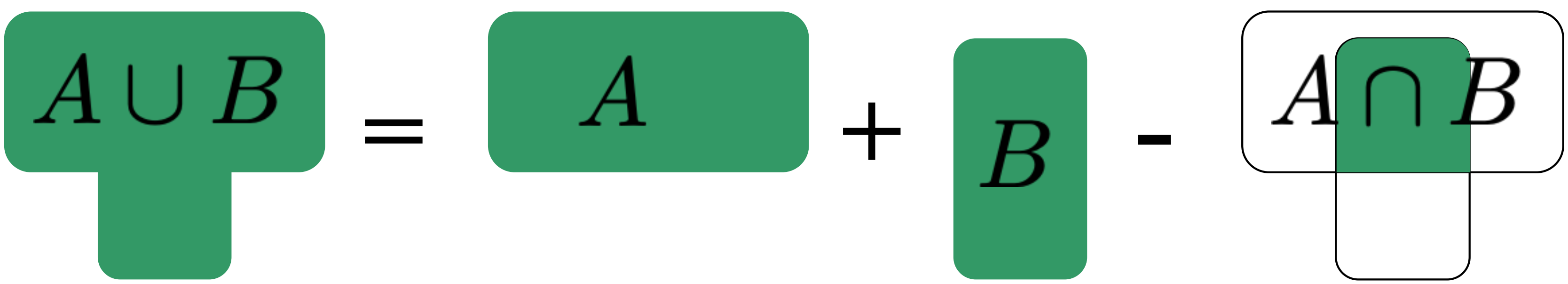


“A” can’t get any larger than all worlds: 100% of worlds have “A” true

Probability

- Event "A" in event space "S"
- Probability Pr[A]
- Axioms of probability
  - $0 \leq \Pr[A] \leq 1$
  - $\Pr[S] = 1$
  - $\Pr[\emptyset] = 0$
  - $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$

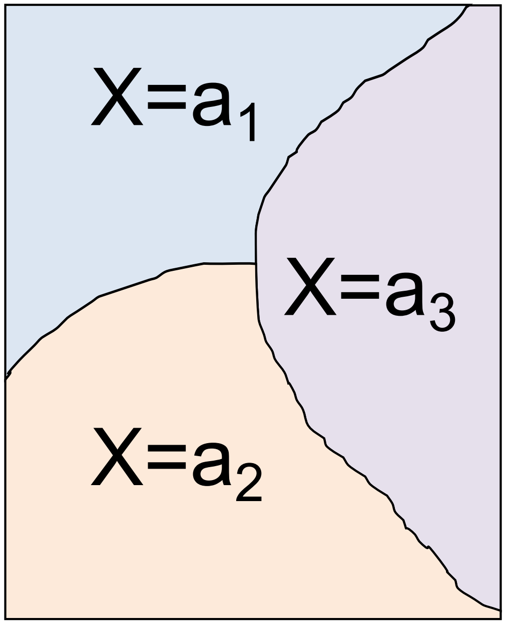




Discrete random variables

- X takes on finite set of values $S=\{a_1 \ldots a_d\}$
  - *Disjoint* and *Exhaustive*
- Probability mass functions (pmfs)
  - Define a measure on subsets of S
- $\Pr[X=a_i]$ defined for each value a_i

$$\Pr[X \in A \subseteq S] = \sum_{a_i \in A} Pr[X = a_i]$$



- Constraints:

$$0 \leq \Pr[X = a_i] \leq 1 \qquad \sum_i \Pr[X = a_i] = 1$$

Examples

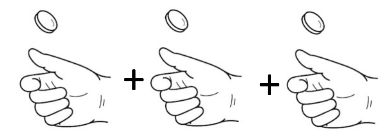
Bernoulli RV: "coin toss"

$$X \in \{0, 1\}$$

$$\Pr[X = 1] = \rho$$
$$\Pr[X = 0] = 1 - \rho$$

Binomial(p,n): toss the coin n times & count

$$Y = \sum_{i=1}^{n} X_i$$



Discrete(d): d-sided die roll

$$X \in \{1, \ldots, d\}$$

$$\Pr[X = 1] = \rho_1$$
$$\vdots$$
$$\Pr[X = d] = \rho_d$$

$$\sum_i \rho_i = 1$$

Multinomial(d,n): roll the die n times and count outcomes

$$Y = \left[\#\{X_i = 1\}, \ldots, \#\{X_i = d\}\right]$$

Probability distributions

- Discrete random variables
  - Typically represent as a table
  - But, useful to express analytically
    - Later: take derivatives, fit to data, etc.

- Ex: Bernoulli, X = 0 or 1

"Exponential family" form

$$p(x) = (\rho)^x \cdot (1-\rho)^{(1-x)}$$
$$= \begin{cases} (\rho)^1 \cdot (1-\rho)^0 = \rho & \text{if } x = 1 \\ (\rho)^0 \cdot (1-\rho)^1 = (1-\rho) & \text{if } x = 0 \end{cases}$$

"Network polynomial" form

$$p(x) = (\rho) \cdot (x) + (1-\rho) \cdot (1-x)$$
$$= \begin{cases} \rho + 0 & \text{if } x = 1 \\ 0 + (1-\rho) & \text{if } x = 0 \end{cases}$$

Joint distributions

- Often, we want to reason about multiple variables
- Example: dentist
  - T: have a toothache
  - D: dental probe catches
  - C: have a cavity

T	D	C	p(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

- Joint distribution
  - Assigns each event (T=t, D=d, C=c) a probability
  - Probabilities sum to 1.0
- Law of total probability:

$$p(C=1) = \sum_{t,d} p(T=t, D=d, C=1)$$

= 0.008 + 0.072 + 0.012 + 0.108 = 0.20

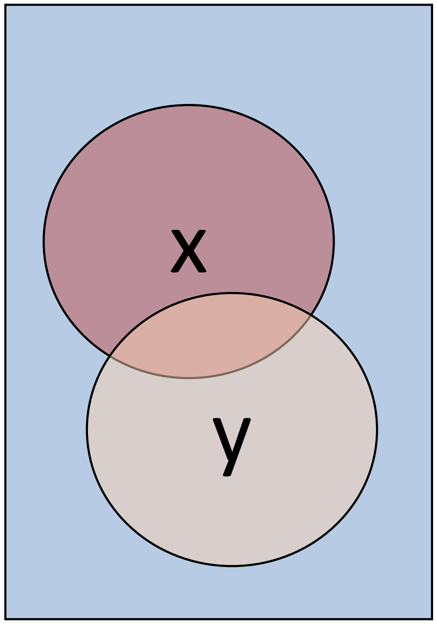
  - *Some* value of (T,D) must occur; values disjoint
  - "Marginal probability" of C; "marginalize" or "sum over" T,D

Conditional probability

- Chain rule:

$$p(X=x, Y=y) = p(X=x)p(Y=y|X=x)$$

  - p(X=x,Y=y) : probability that both X=x and Y=y
  - p(X=x) : probability that X=x (and some Y)
  - P(Y=y|X=x): probability that Y=y given X=x already
  - If p(X) > 0 : $p(Y|X) = \frac{p(X,Y)}{p(X)}$



- More generally:

$$p(X,Y,Z) = p(X)\;p(Y|X)\;p(Z|X,Y)$$

$$p(W,X,Y,Z) = p(X)\;p(Y|X)\;p(Z|X,Y)\;p(W|X,Y,Z)$$

(can apply using any order of expansion; each conditional depends on previous variables in order)

Example from Russell & Norvig

The effect of evidence

- Example: dentist
  - T: have a toothache
  - D: dental probe catches
  - C: have a cavity

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

- Recall p(C=1) = 0.20
- Suppose we observe D=0, T=0?

$$p(C=1|D=0,T=0) = \frac{p(C=1,D=0,T=0)}{p(D=0,T=0)}$$

$$= \frac{0.008}{0.576 + 0.008} = 0.012$$

- Observe D=1, T=1?

$$= \frac{0.108}{0.016 + 0.108} = 0.871$$

Called ***posterior probabilities***
(posterior = after observing)

Example from Russell & Norvig

The effect of evidence

- Example: dentist
  - T: have a toothache
  - D: dental probe catches
  - C: have a cavity

- Combining these rules:

$$p(C=1|T=1) = \frac{p(C=1,T=1)}{p(T=1)}$$

$$= \frac{0.012 + 0.108}{0.064 + 0.012 + 0.016 + 0.108} = 0.60$$

$$p(T=1) = 0.20$$

Called the ***probability of evidence***

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

Computing posteriors

- Sometimes easiest to normalize last

$$p(C|T=1) \;=\; \frac{1}{p(T=1)}\, p(C,T=1) \;\propto\; p(C,T=1) \;=\; \sum_d p(C,d,T=1)$$

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

Assign T=1

D	C	F(D,C)
0	0	0.064
0	1	0.012
1	0	0.016
1	1	0.108

Sum over D

C	G(C)
0	0.08
1	0.120

Normalize

C	P(C|T=1)
0	0.40
1	0.60

```
P = gm.Factor( [T,D,C] )
P[ {T:0,D:0,C:0} ] = 0.576
...  # define joint distribution
```

```
F = P.condition( {T:1} )      # assign T=1
G = F.sum( [D] )              # sum over D
H = G / G.sum()               # normalize
```

Bayes rule

- Lets us calculate posterior given evidence

$$p(Y|X)\ p(X) = p(X,Y) = p(X|Y)\ p(Y)$$

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

"Bayes rule"

- Example: flu
  - P(F), P(H|F)
  - P(F=1 | H=1) = ?

F	P(F)
0	0.95
1	0.05

F	H	P(H|F)
0	0	0.80
0	1	0.20
1	0	0.50
1	1	0.50

$$= \frac{0.50 * 0.05}{0.50 * 0.05 + 0.20 * 0.95} = 0.116$$

Independence

- X, Y independent:
  - p(X=x,Y=y) = p(X=x) p(Y=y)   for all x,y
  - Shorthand: p(X,Y) = P(X) P(Y)
  - Equivalent: p(X|Y) = p(X) or p(Y|X) = p(Y)   (if p(Y), p(X) > 0)
  - Intuition: knowing X has no information about Y (or vice versa)

Independent probability distributions:

A	P(A)
0	0.4
1	0.6

B	P(B)
0	0.7
1	0.3

C	P(C)
0	0.1
1	0.9

This reduces representation size!

Joint:

A	B	C	P(A,B,C)
0	0	0	.4 * .7 * .1
0	0	1	.4 * .7 * .9
0	1	0	.4 * .3 * .1
0	1	1	…
1	0	0	
1	0	1	
1	1	0	
1	1	1	

Independence

- X, Y independent:
  - p(X=x,Y=y) = p(X=x) p(Y=y)   for all x,y
  - Shorthand: p(X,Y) = P(X) P(Y)
  - Equivalent: p(X|Y) = p(X)   or p(Y|X) = p(Y)   (if p(Y), p(X) > 0)
  - Intuition: knowing X has no information about Y (or vice versa)

Independent probability distributions:

A	P(A)
0	0.4
1	0.6

B	P(B)
0	0.7
1	0.3

C	P(C)
0	0.1
1	0.9

Joint:

A	B	C	P(A,B,C)
0	0	0	0.028
0	0	1	0.252
0	1	0	0.012
0	1	1	0.108
1	0	0	0.042
1	0	1	0.378
1	1	0	0.018
1	1	1	0.162

This reduces representation size!

Note: it is hard to "read" independence from the joint distribution.
We can "test" for it, however.

Conditional Independence

- X, Y independent given Z
  - $p(X=x,Y=y|Z=z) = p(X=x|Z=z)\, p(Y=y|Z=z)$ for all x,y,z
  - Equivalent: $p(X|Y,Z) = p(X|Z)$ or $p(Y|X,Z) = p(Y|Z)$ (if all > 0)
  - Intuition: X has no additional info about Y beyond Z's

- Example

X = height — p(height|reading, age) = p(height|age)

Y = reading ability — p(reading|height, age) = p(reading|age)

Z = age

Height and reading ability are dependent (not independent), but are conditionally independent given age

Conditional Independence

- X, Y independent given Z
  - $p(X=x,Y=y|Z=z) = p(X=x|Z=z)\ p(Y=y|Z=z)$ for all x,y,z
  - Equivalent: $p(X|Y,Z) = p(X|Z)$ or $p(Y|X,Z) = p(Y|Z)$
  - Intuition: X has no additional info about Y beyond Z's

- Example: Dentist

Again, hard to "read" from the joint probabilities; only from the conditional probabilities.

Like independence, reduces representation size!

Joint prob:

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

Conditional prob:

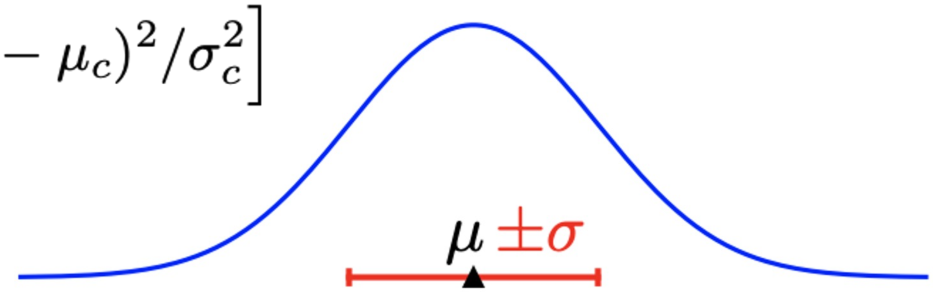
T	D	C	P(T|D,C)
0	0	0	0.90
0	0	1	0.40
0	1	0	0.90
0	1	1	0.40
1	0	0	0.10
1	0	1	0.60
1	1	0	0.10
1	1	1	0.60

Continuous random variables

- Definition
  - Cumulative distribution function, Pr[X < x] = P(x)
  - Probability density function p(x) = (d/dx) P(x)
  - Now, $0 \le P(x) \le 1$, but $p(x) \ge 0$.
- Uniform distribution on [0,T]
  - Density p(x) = 1/T if x in [0,T] and 0 otherwise

- Gaussian distribution
  - Classical probability distribution over continuous values
  - Density function:

$$\mathcal{N}(x\,;\,\mu_c,\sigma_c^2) = \left(2\pi\sigma_c^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\mu_c)^2/\sigma_c^2\right]$$



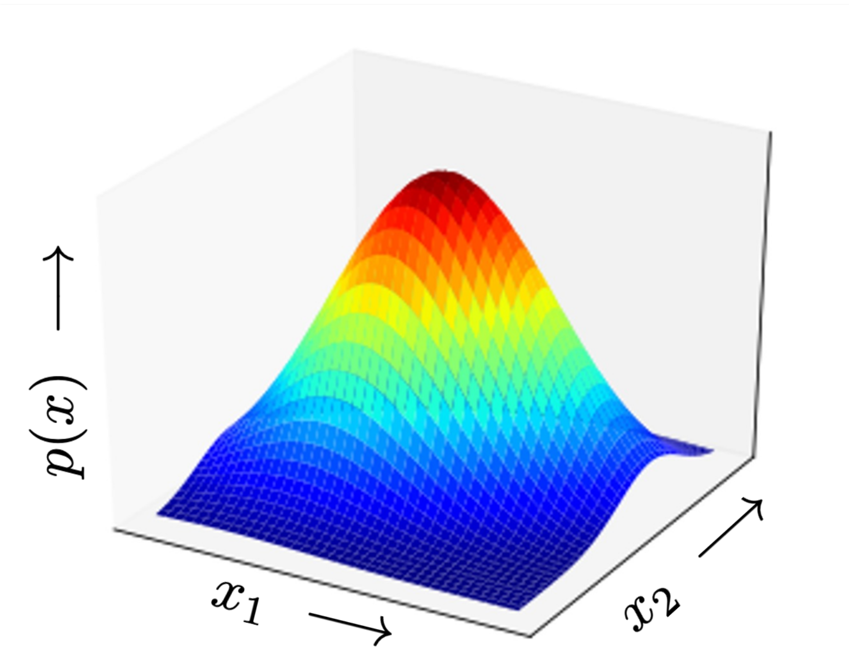
Multivariate Gaussian models

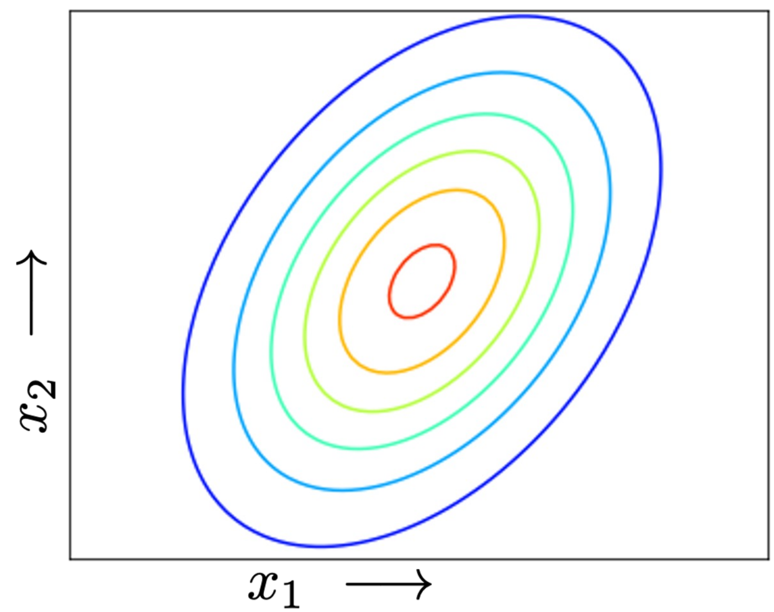
- Similar to univariate case

$$p(x) = \mathcal{N}(x; \mu, \Sigma) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)^T \right]$$

μ $1 \times n$ mean vector

Σ $n \times n$ covariance matrix





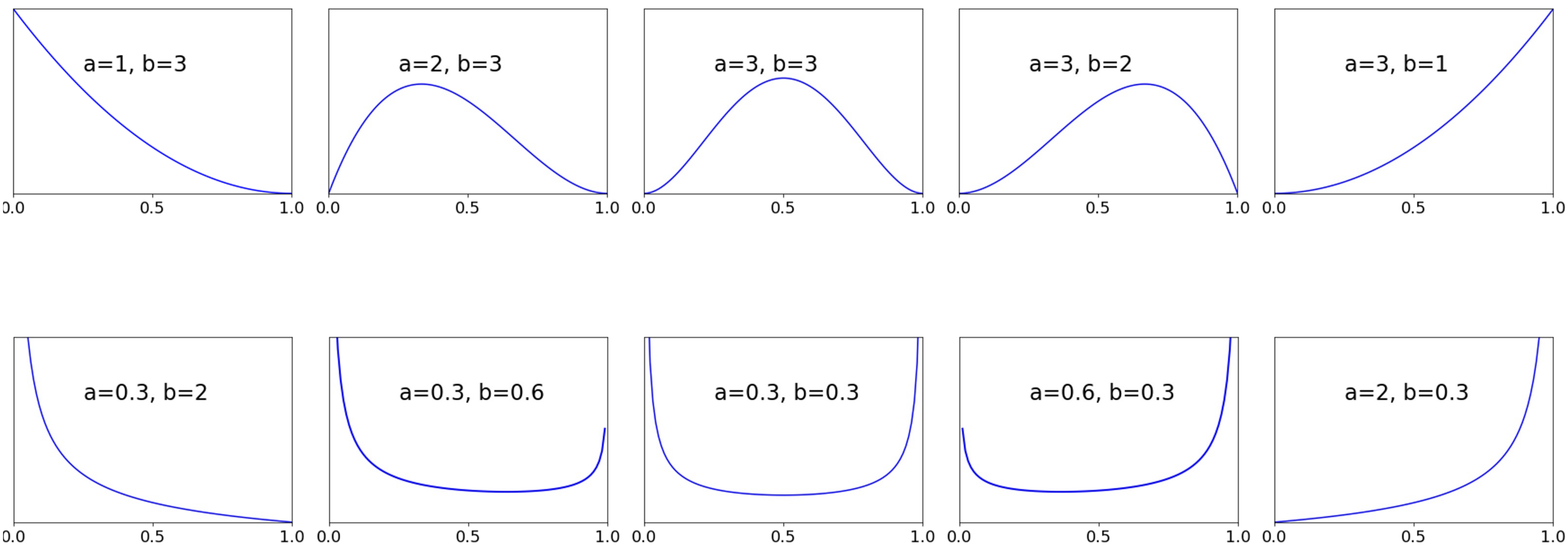
Beta distributions

- Distribution on continuous X in range [0,1]

$$p(x) = \mathrm{Beta}(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}\,(1-x)^{b-1}$$

(where a, b > 0)

Examples:

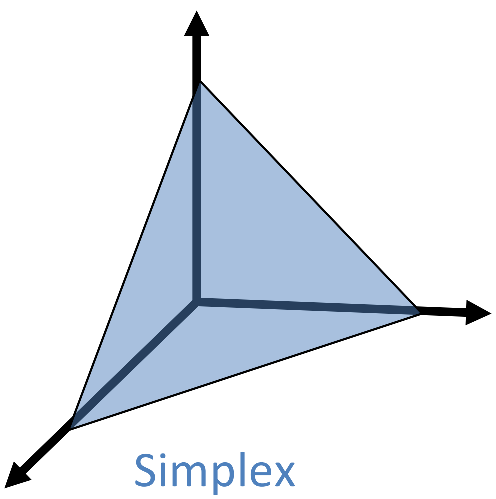


Dirichlet distributions

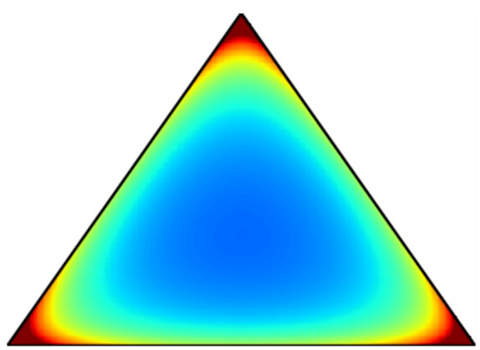
- Generalizes the Beta distribution to vectors

$$p(x) = \mathrm{Dir}(x;\alpha) = \frac{\Gamma\left(\sum_{j=1}^{n}\alpha_j\right)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_n)} \prod_{j=1}^{n} x_j^{\alpha_j - 1} \qquad \left(\text{if } \sum_{j=1}^{n} x_j = 1\right)$$

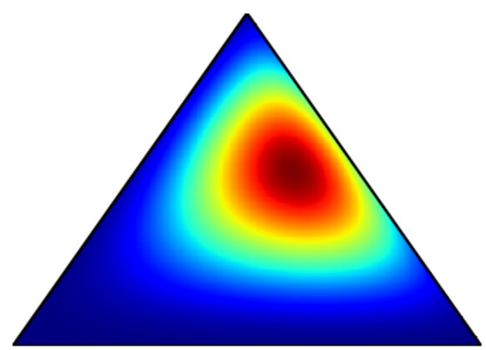
- Distribution over simplex
  - vectors that sum to one (pmfs)



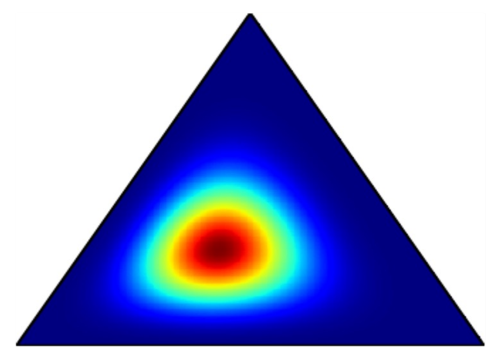
α = [.3, .3, .3]



α = [2, 3, 4]



α = [8, 6, 6]



The exponential family

- General class that includes many common distributions

$$p(x;\theta) = \frac{1}{Z(\theta)} \exp\big(\eta(\theta)\cdot\phi(x)\big)\, h(x)$$

(parameter-only transform) (data-only transform: "features")

(dot product between vectors = linear function)

Ex: Bernoulli distribution

$$\rho^X (1-\rho)^{(1-X)} = \exp\Big(\log(\rho)X + \log(1-\rho)(1-X)\Big) = (1-\rho)\exp\Big(\log\big(\frac{\rho}{1-\rho}\big)X\Big)$$

$$\eta(\rho) = [\log(\rho) \quad \log(1-\rho)]$$
$$\phi(x) = [\ x \qquad (1-x)\ \ ]$$

$$\eta(\rho) = [\ \log\ \ \rho/(1-\rho)\ ]$$
$$\phi(x) = [\qquad x \qquad]$$
$$Z(\rho) = (1-\rho)^{-1}$$

"Natural parameters":

$$p(X;\eta) = \frac{1}{1+\exp(\eta)}\ \exp(\eta\, X)$$

Pyro

- A library for “probabilistic programming”

```
import numpy as np                       # linear algebra library from CS178
import matplotlib.pyplot as plt          # plotting library from CS178
```

```
import torch                             # like numpy, but with extra features
import pyro                              # PyPI package "pyro-ppl"; uses torch
import pyro.distributions as dist
```

- Define a random variable & give it a distribution

```
x = pyro.sample('X', dist.Bernoulli(0.33)) # define & sample var "X"
x
```

```
tensor(1.)
```

  - Can also sample without naming the variable:

```
pX = dist.Bernoulli(1./3)                # the distribution p(X)
print( pX.sample([20]) )                 # draw 20 samples
```

```
tensor([1., 0., 0., 1., 0., 0., 0., 1., 1., 1., 0., 0., 0., 0.,
        0., 0., 0., 0., 1., 0.])
```

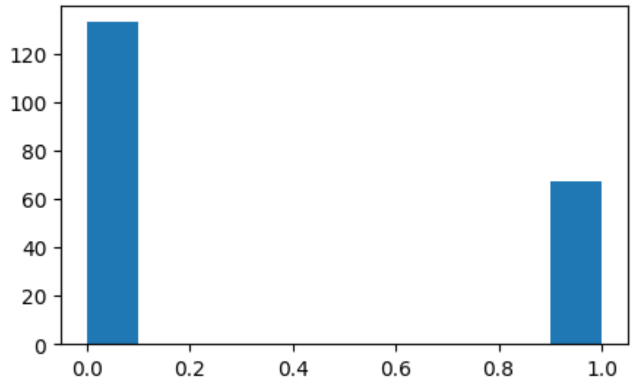
Pyro

- Visualizing our samples: histograms

```
samples = pX.sample([200])
print(f'Average number of ones: {samples.mean()}')

plt.figure(figsize = (5,3))    # set the size of the figure we'll plot on
plt.hist(samples);             # display the histogram of our samples
```

```
Average number of ones: 0.33500000834465027
```



Pyro

- Visualizing our samples: scatterplots

```
pZ = dist.MultivariateNormal( torch.zeros(2), torch.eye(2) )
pZ.sample()
```

```
tensor([ 1.3664, -0.9620])
```

```
Z = pZ.sample([1000]).numpy().   # numpy is more convenient for plotting

print(Z.shape)
```

```
(1000, 2)
```

```
plt.plot(Z[:,0],Z[:,1], 'b.');   # axis 0 vs axis 1, using blue dots
plt.axis([-4,4,-4,4]);
```