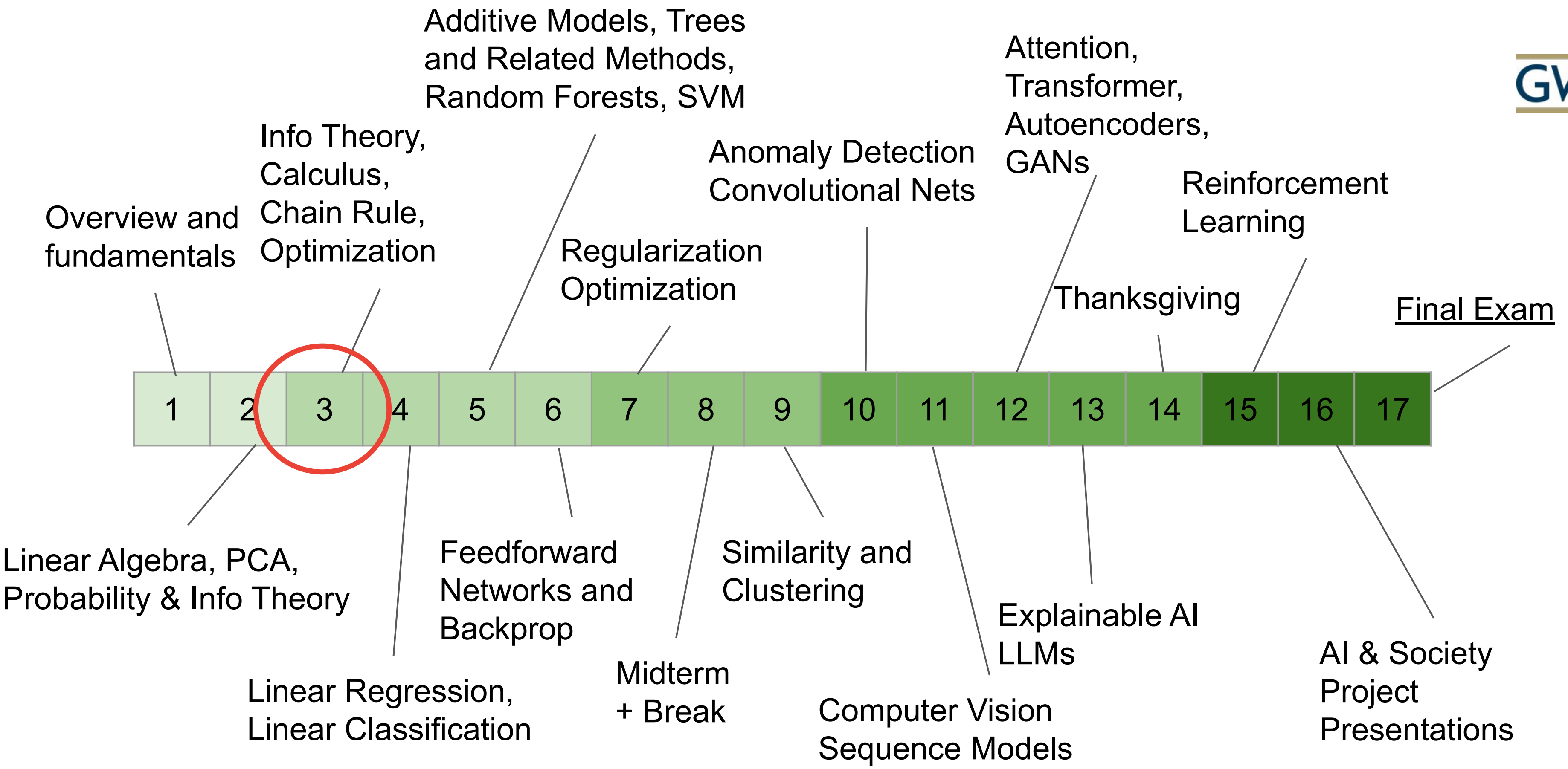
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

CS 4364/6364

Machine Learning

Fall Semester 9/5/2023
Lecture 4.
From Coin Flips to KL Divergence

John Sipple
jsipple@gwu.edu


Overview and fundamentals
Linear Algebra, PCA, Probability & Info Theory
Info Theory, Calculus, Chain Rule, Optimization
Linear Regression, Linear Classification
Additive Models, Trees and Related Methods, Random Forests, SVM
Feedforward Networks and Backprop
Regularization Optimization
Midterm + Break
Similarity and Clustering
Anomaly Detection Convolutional Nets
Computer Vision Sequence Models
Attention, Transformer, Autoencoders, GANs
Explainable AI LLMs
Thanksgiving
Reinforcement Learning
AI & Society Project Presentations
Final Exam
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

Transforming probability distributions

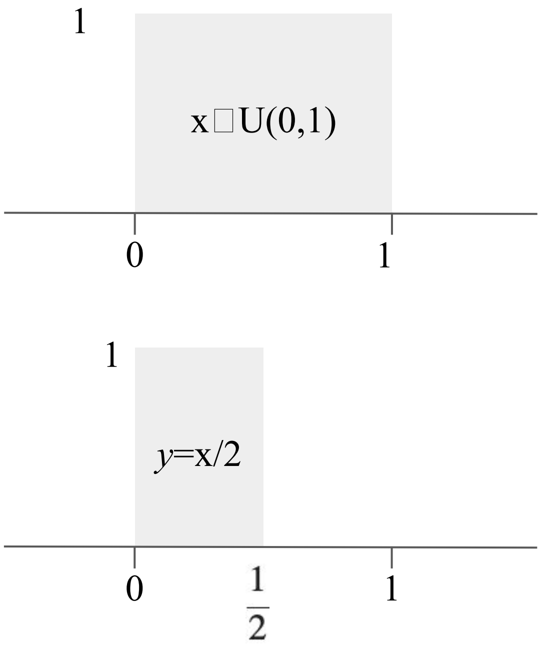
Suppose we have: $y = g(x) = \frac{x}{2}$ and $x \sim U(0, 1)$

And we apply the fuls $p_y(y) = p_x(g^{-1}(x))$

Then we violate the rules of probability:

$$\int p_y(y)dy = \frac{1}{2}$$

We need to be cautious about applying functions to probability distributions!



Not a probability distribution!

Change of Variables on PDs requires rescaling

$$y = g(x)$$

$$p_y(y) \neq p_x(g^{-1}(y))$$

We need to preserve the property:

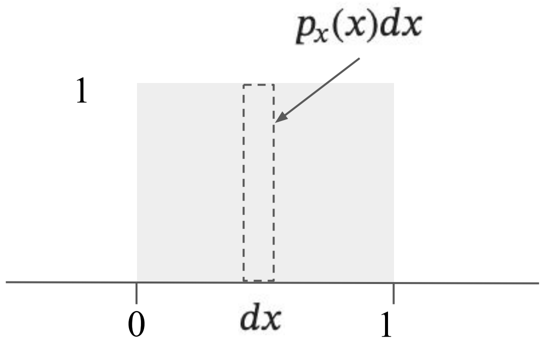
$$|p_y(g(x))dy| = |p_x(x)dx|$$

Can be rewritten as:

$$p_x(x) = p_y(g(x)) \left| \frac{\partial g(x)}{\partial x} \right|$$

And in higher dimension:

$$p_x(\boldsymbol{x}) = p_y(g(\boldsymbol{x})) \left| \det \left(\frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}} \right) \right|$$



Binomial Random Variable

Each event takes on one of two possible outcomes.

$$E = \{0, 1\}$$

for example, 0 = Tails and 1 = Heads

Specified with a parameter p, where:

$$p = P(E = 1)$$

and

$$1 - p = P(E = 0)$$

i.i.d.

Assume the events are **i.i.d.**:

- **independent**:

$$P(E_1 = 1, E_2 = 1) = P(E_1 = 1)P(E_2 = 1)$$

  and

- **identically distributed**: p is constant

$$p_{E_1} = p_{E_2} = \ldots$$

Bernoulli Distribution

Describes the probability of seeing s sucesses in N trials, given a parameter p.

For example, what's the probaility of seeing 2 heads in 3 trials with a fair coin, $p = 0.5$?

$$b(s, N; p) =$$

Bernoulli Distribution

Let's enumerate al possible, and exclusive outcomes:

Outcome a. $1, 1, 0$: $P(E = 1) \times P(E = 1) \times P(E = 0) = p^2(1 - p)$

Outcome b. $0, 1, 1$: $P(E = 0) \times P(E = 1) \times P(E = 1) = p^2(1 - p)$

Outcome c. $1, 0, 1$: $P(E = 1) \times P(E = 0) \times P(E = 1) = p^2(1 - p)$

Since the outcomes are exclusive (i.e., a or b or c) could occur, we can add them:

$$P(a \cup b \cup c) = P(a) + P(b) + P(c)$$

Then

$$b(2, 3; 0.5) = 3 \times p^2(1 - p) = 3 \times 0.5^3 = 0.375$$

Bernoulli Distribution (general form)

In general, we'll just count all possible permutations of s sucesses in N **exclusive** events.

$\binom{N}{s}$: The number of possible permutations of having s sucesses in N trials.

$$\binom{N}{s} = \frac{N!}{(N-s)!s!}$$

Back to the example:

$$\binom{3}{2} = \frac{3!}{(3-2)!2!} = \frac{1 \times 2 \times 3}{1 \times 1 \times 2} = 3$$

In general, the Binomial distribution is:

$$b(s, N; p) = \binom{N}{s} p^s (1-p)^{N-s}$$

Bernoulli Distribution with multiple levels

Given p that you'll be exposed to the Coronavirus and catch Covid-19 on a Metro ride.

One might ask, what's the probability that you'll be exposed once **or more** in $N = 6$ trips?

$$B(s > 1, N = 6; p) = b(s = 1, N; p) + b(s = 2, N; p) + \cdots = \sum_{s=1}^{6} b(s, N; p)$$

Alternatively:

$$B(s > 1, N = 6; p) = 1 - b(s = 0, N; p) = 1 - (1 - p)^6$$

In other words, we can restate the question as, 1 minus the probability of zero exposures in 6 trips.

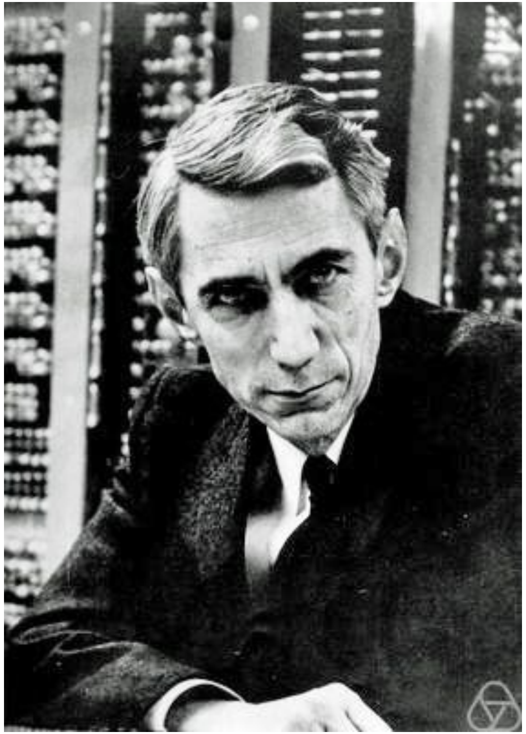
Information Theory

The study of the transmission, processing, extraction, and utilization of information.

Formalized in 1948 by Claude Shannon: [A Mathematical Theory of Communication]

- Reconstruct messages accurately despite a noisy channel

Foundational to data compression and digital communication



Information Theory

Basic Idea: Learning that an unlikely event occurred is more informative than learning that a likely event occurred.

Information of a single event: $I(x) = -\log P(\mathrm{x} = x)$

Entropy of a distribution:

$$H(\mathrm{x}) = \mathbb{E}_{x \sim P}[I(x)] = -\mathbb{E}_{x \sim P}[\log P(x)]$$

$$H(\mathrm{x}) = -\sum_{x} P(\mathrm{x} = x) \log P(\mathrm{x} = x)$$

Entropy

The universal measure of **randomness**.

For a discrete random variable x, entropy is defined as:

$$H(\mathrm{x}) = -\sum_x P(\mathrm{x}=x)\log P(\mathrm{x}=x)$$
$$= \mathbb{E}\left[-\log P(\mathrm{x}=x)\right]$$

Note on $\log$ **bases**: Where the log base can be $e, 2, 10$ etc., but if it's $|\mathrm{x}|$, then maximum entropy, $\max H(\mathrm{x}) = 1$. (The shape of H is the same for all log bases.)

Entropy

For two outcomes (i.e., coin flips):

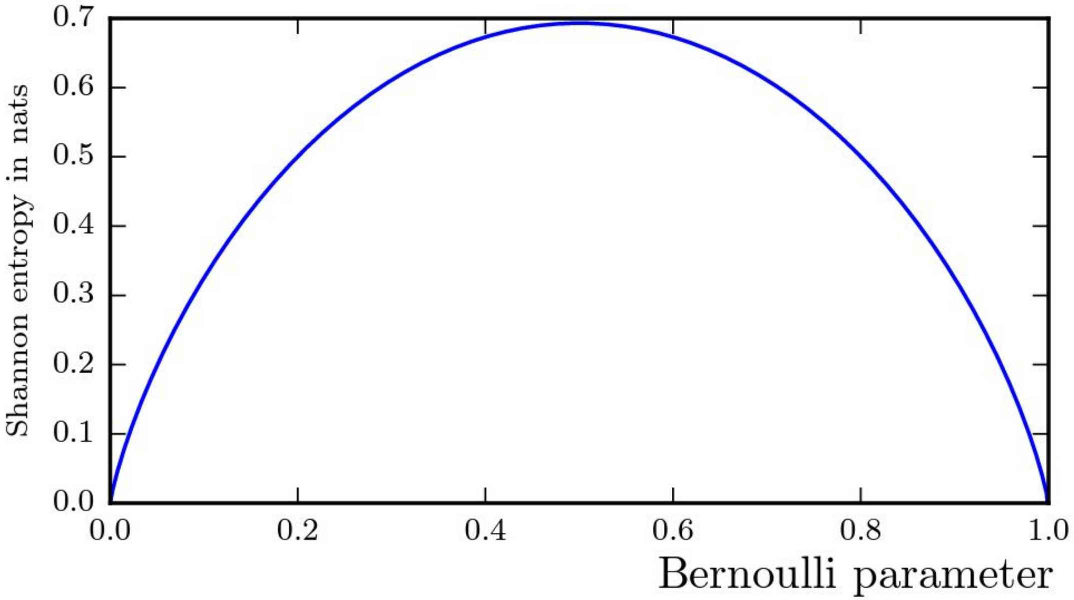
$$H(p) = -p \log p - (1 - p) \log(1 - p)$$

For example, $p = 0.4$:

$$H(p = 0.4) = -0.4 \log 0.4 - 0.6 \log 0.6 = 0.673$$

How about multiple outcomes (e.g, die)?

Entropy of a Coin Flip



$$H(p) = -p \log p - (1-p)\log(1-p)$$

Cross-Entropy of two probability distributions

In Machine Learning we are often interested in approximating a true, but unknown RV (p) with an estimator function (q)

Cross-entropy can be a good measure for how well q approximates p.

$$H(\mathrm{p},\mathrm{q}) = -\sum_i \mathrm{p}\log\mathrm{q}$$

For the coin-flip case:

$$H(p,q) = -p\log q - (1-p)\log(1-q)$$

Cross-Entropy of two probability distributions

For example:

- **True probability** $p = 0.4$
- **Estimated probability** $q = 0.5$

$$H(p = 0.4, q = 0.5) = -0.4 \log 0.5 - 0.6 \log 0.5 = 0.693$$

Which is larger than $H(p = 0.4) = 0.673$ and is minimized when $q = p$.

Kullback-Leibler (KL) Divergence

A measure of how one probability distribution q is different from another probability distrbution p.

$$D_{KL}\,(\mathrm{p}||\mathrm{q}) = H(\mathrm{p},\mathrm{q}) - H(\mathrm{p})$$

$$= -\sum_i \mathrm{p}\log\mathrm{q} - \left(-\sum_i \mathrm{p}\log\mathrm{p}\right)$$

$$= -\sum_i \mathrm{p}\log\frac{\mathrm{q}}{\mathrm{p}}$$

Note:

$$D_{KL}\,(\mathrm{p}||\mathrm{p}) = 0$$

$$D_{KL}\,(\mathrm{p}||\mathrm{q}) \neq D_{KL}\,(\mathrm{q}||\mathrm{p})$$

Kullback-Leibler (KL) Divergence

Back to the example, $p = 0.4$ and $q = 0.5$:

$$D_{KL}\,(p = 0.4||q = 0.5) = 0.693 - 0.673 = 0.02$$

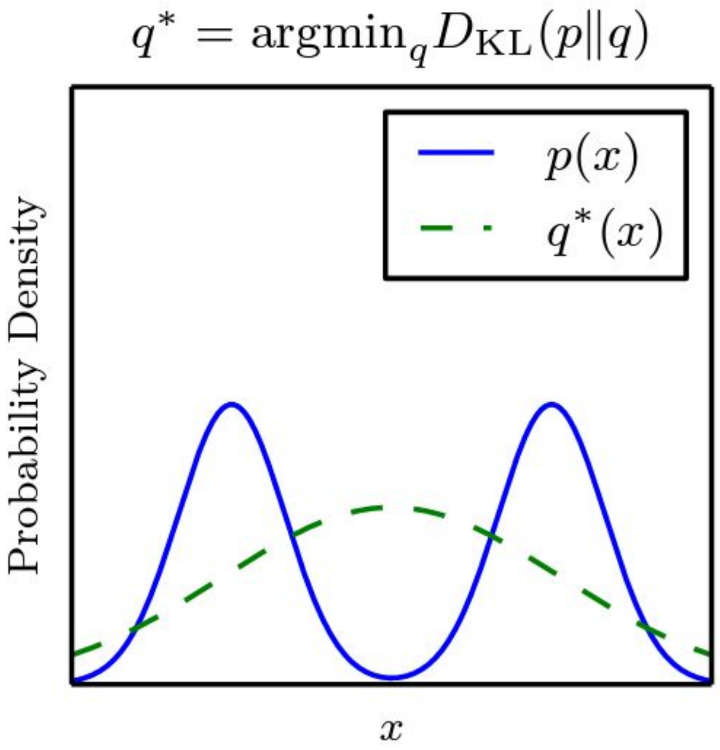
Kullback-Leibler (KL) Divergence

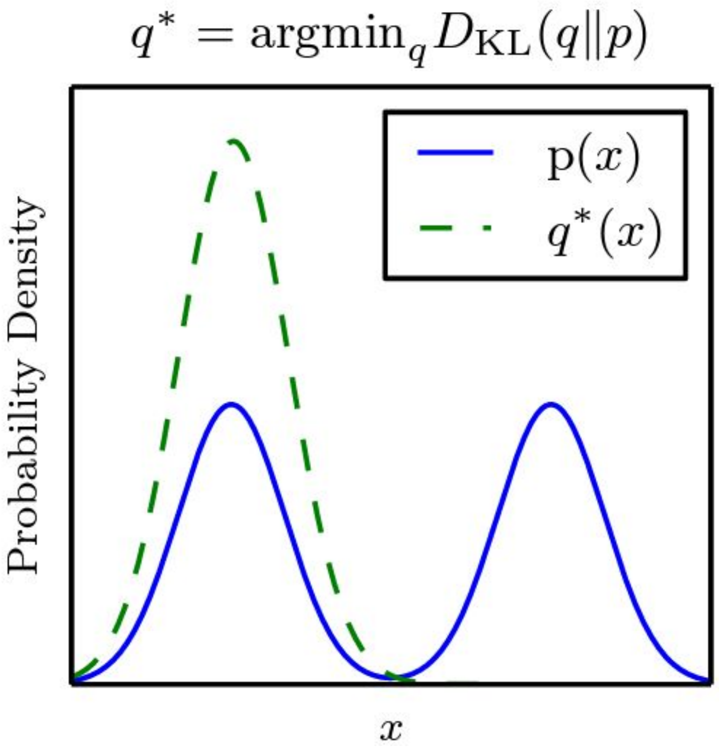
Measures the difference between two distributions from the "perspective" of one distribution

$$D_{KL}(P||Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - logQ(x)\right]$$

$$D_{KL}(P||Q) = \sum_{x} P(\mathrm{x}=x) \log \frac{P(\mathrm{x}=x)}{Q(\mathrm{x}=x)}$$

KL Divergence is Asymmetric





Binary Cross-Entropy Loss

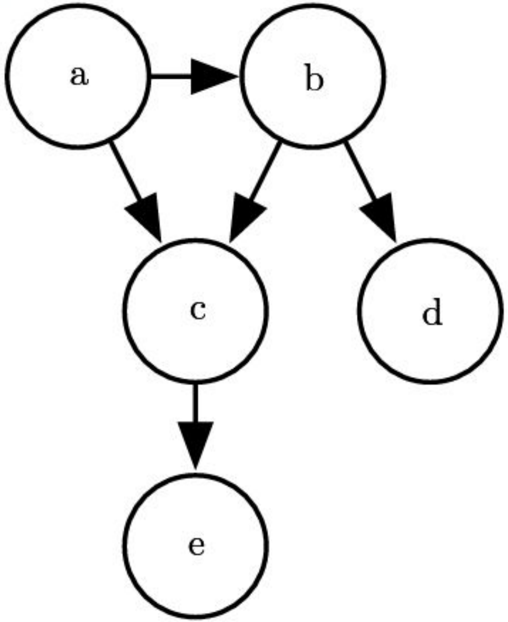
Suppose you want to train a binary classifier and you need a loss function that you want to minimize.

- **Test Data**: $p = [1, 1, 0, 1, 0, 1, \ldots]$
- **Classifier predictions**: $q = [0.8, 0.9, 0.3, 0.7, 0.1, 0.9, \ldots]$

$$\mathcal{L}(p, q) = \frac{1}{N} \sum_{i}^{N} -p_i \log q_i - (1 - p_i) \log(1 - q_i)$$

Directed Structured Model

Figure 3.7



$$p(\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d},\mathrm{e}) = p(\mathrm{a})p(\mathrm{b} \mid \mathrm{a})p(\mathrm{c} \mid \mathrm{a},\mathrm{b})p(\mathrm{d} \mid \mathrm{b})p(\mathrm{e} \mid \mathrm{c}).$$

Undirected Structured Model

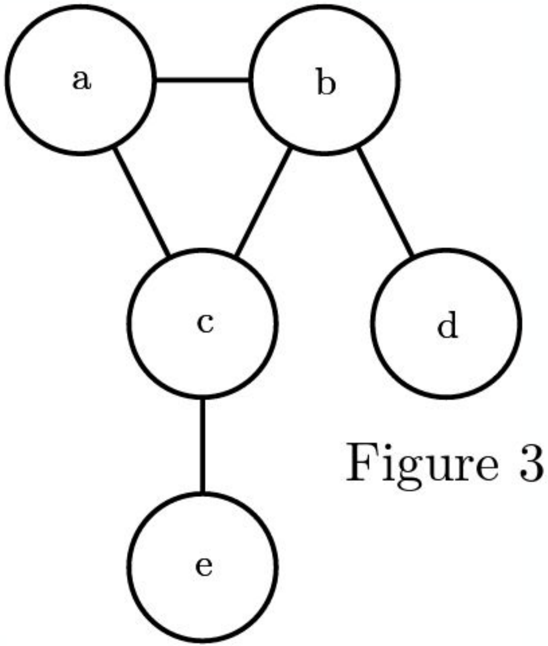


Figure 3.8

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = \frac{1}{Z}\phi^{(1)}(\mathrm{a}, \mathrm{b}, \mathrm{c})\phi^{(2)}(\mathrm{b}, \mathrm{d})\phi^{(3)}(\mathrm{c}, \mathrm{e}).$$