10-301/10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

PAC Learning

+

MLE/MAP

Matt Gormley
Lecture 16
Mar. 15, 2023

Reminders

- Homework 5: Neural Networks
  - Out: Sun, Feb 26
  - Due: Fri, Mar 17 at 11:59pm
- Peer Tutoring
- Homework 6: Learning Theory / Generative Models
  - Out: Fri, Mar. 17
  - Due: Fri, Mar. 24 at 11:59pm
  - IMPORTANT: only 2 grace/late days permitted
- Exam 2 (Thu, Mar 30)
- Exam 3 (Tue, May 2)

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about…

	Realizable	Agnostic
Finite $\|\mathcal{H}\|$	**Thm** bele… prob… have $R(h) \leq \epsilon$.	… have that $\|R(h) - \hat{R}(h)\| \leq \epsilon$.
Infinite $\|\mathcal{H}\|$		

For these two cases, we will use a new definition for the "complexity" of a Hypothesis space called **VC Dimension**

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about…

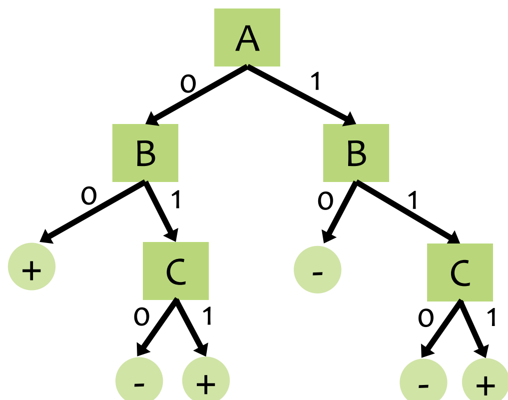
	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.
Infinite $\lvert\mathcal{H}\rvert$	**Thm. 3** $N = O(\frac{1}{\epsilon}\left[\mathrm{VC}(\mathcal{H})\log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 4** $N = O(\frac{1}{\epsilon^2}\left[\mathrm{VC}(\mathcal{H}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.

VC-DIMENSION

Finite vs. Infinite |H|

Finite |H|

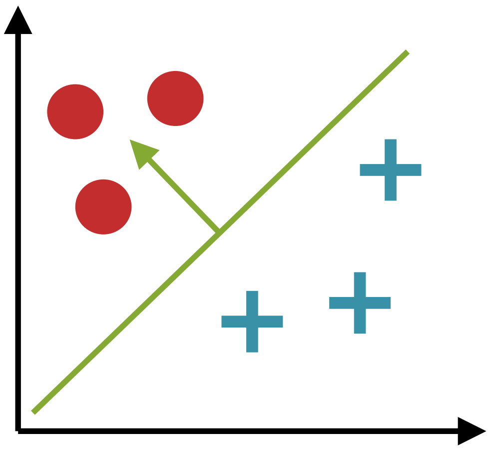
- *Example*: H = the set of all decision trees of depth D over binary feature vectors of length M



- *Example*: H = the set of all conjunctions over binary feature vectors of length M

Infinite |H|

- *Example*: H = the set of all linear decision boundaries in M dimensions



- *Example*: H = the set of all neural networks with 1-hidden layer with length M inputs

Labelings & Shattering

Def: A hypothesis h applied to some dataset S generates a **labeling** of S.

Def: Let $\mathcal{H}[S]$ be the set of all (distinct) labelings of S generated by hypotheses $h \in \mathcal{H}$.

$\mathcal{H}$ **shatters** S if $|\mathcal{H}[S]| = 2^{|S|}$

Equivalently, the hypotheses in $\mathcal{H}$ can generate every possible labeling of S.

Labelings & Shattering

Whiteboard:

– Shattering example: binary classification

VC-dimension

Def: The **VC-dimension** (or Vaporik-Chervonenkis dimension) of $\mathcal{H}$ is the cardinality of the largest set S such that $\mathcal{H}$ can shatter S.

Special Case: If $\mathcal{H}$ can shatter arbitrarily large finite sets, then the VC-dimension of $\mathcal{H}$ is infinity

Notation: We write $\mathrm{VC}(\mathcal{H}) = d$ to say the VC-Dimension of a hypothesis space $\mathcal{H}$ is d

VC-dimension Proof

Proof Technique: To **prove** that $\mathrm{VC}(\mathcal{H}) = d$ there are two steps:

1. show that there exists a set of d points that can be shattered by $\mathcal{H}$
   ➔ $\mathrm{VC}(\mathcal{H}) \geq d$
2. show that there does NOT exist a set of $d + 1$ points that can be shattered by $\mathcal{H}$
   ➔ $\mathrm{VC}(\mathcal{H}) < d + 1$

VC-dimension

Whiteboard:

- VC-dimension Example: linear separators
- Proof sketch of VC-dimension for linear separators in 2D

∃ vs. ∀

VC-dimension

– Proving **VC-dimension** requires us to show that **there exists** (∃) a dataset of size d that can be shattered and that **there does not exist** (∄) a dataset of size d+1 that can be shattered

Shattering

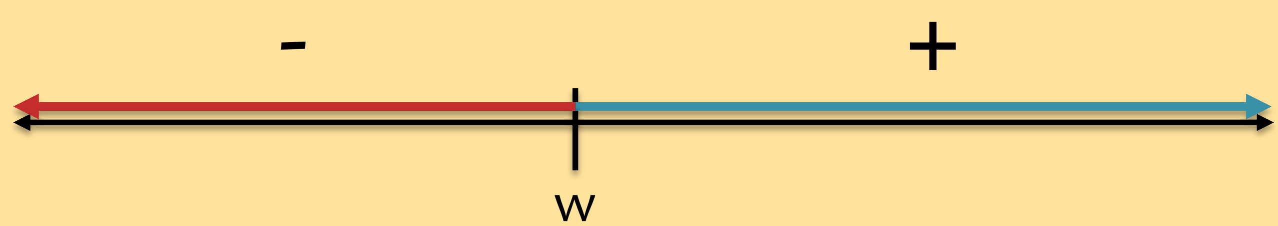
– Proving that a particular dataset can be **shattered** requires us to show that **for all** (∀) labelings of the dataset, our hypothesis class contains a hypothesis that can correctly classify it

VC-dimension Examples

- *Definition*: If VC(H) = d, then **there exists** (∃) a dataset of size d that can be shattered and that **there does not exist** (∄) a dataset of size d+1 that can be shattered

Question:

What is the VC-dimension of H = **1D positive rays.** That is for a threshold w, everything to the right of w is labeled as +1, everything else is labeled -1.



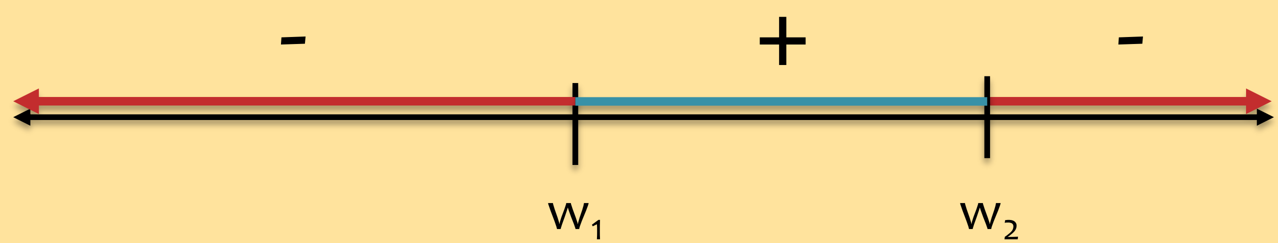
Answer:

VC-dimension Examples

- *Definition*: If VC(H) = d, then **there exists** (∃) a dataset of size d that can be shattered and that **there does not exist** (∄) a dataset of size d+1 that can be shattered

Question:

What is the VC-dimension of H = **1D positive intervals.** That is for an interval (w_1, w_2), everything inside the interval is labeled as +1, everything else is labeled -1.



Answer:

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about…

	Realizable	Agnostic
Finite $\|\mathcal{H}\|$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\|\mathcal{H}\|) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\|\mathcal{H}\|) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\|R(h) - \hat{R}(h)\| \leq \epsilon$.
Infinite $\|\mathcal{H}\|$	**Thm. 3** $N = O(\frac{1}{\epsilon}\left[\text{VC}(\mathcal{H}) \log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 4** $N = O(\frac{1}{\epsilon^2}\left[\text{VC}(\mathcal{H}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\|R(h) - \hat{R}(h)\| \leq \epsilon$.

SLT-STYLE COROLLARIES

SLT-style Corollaries

Thm. 1 $N \geq \frac{1}{\epsilon}\left[\log(|\mathcal{H}|) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.

Solve the inequality in Thm.1 for epsilon to obtain Corollary 1

Corollary 1 (Realizable, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1 - \delta)$, for any h in $\mathcal{H}$ consistent with the training data (i.e. $\hat{R}(h) = 0$),

$$R(h) \leq \frac{1}{N}\left[\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right]$$

We can obtain similar corollaries for each of the theorems…

SLT-style Corollaries

Corollary 1 (Realizable, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for any h in $\mathcal{H}$ consistent with the training data (i.e. $\hat{R}(h) = 0$),

$$R(h) \leq \frac{1}{N}\left[\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right]$$

Corollary 2 (Agnostic, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for all hypotheses h in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2N}\left[\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right]}$$

SLT-style Corollaries

Corollary 3 (Realizable, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for any hypothesis h in $\mathcal{H}$ consistent with the data (i.e. with $\hat{R}(h) = 0$),

$$R(h) \leq O\left(\frac{1}{N}\left[\mathsf{VC}(\mathcal{H}) \ln\left(\frac{N}{\mathsf{VC}(\mathcal{H})}\right) + \ln\left(\frac{1}{\delta}\right)\right]\right) \tag{1}$$

Corollary 4 (Agnostic, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for all hypotheses h in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N}\left[\mathsf{VC}(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right]}\right) \tag{2}$$

SLT-style Corollaries

Corollary 3 (Realizable, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for any hypothesis h in $\mathcal{H}$ consistent with the data (i.e. with $\hat{R}(h) = 0$),

$$R(h) \leq O\left(\frac{1}{N}\left[\text{VC}(\mathcal{H}) \ln\left(\frac{N}{\text{VC}(\mathcal{H})}\right) + \ln\left(\frac{1}{\delta}\right)\right]\right) \tag{1}$$

Corollary 4 (Agnostic, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for all hypotheses h in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N}\left[\text{VC}(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right]}\right) \tag{2}$$

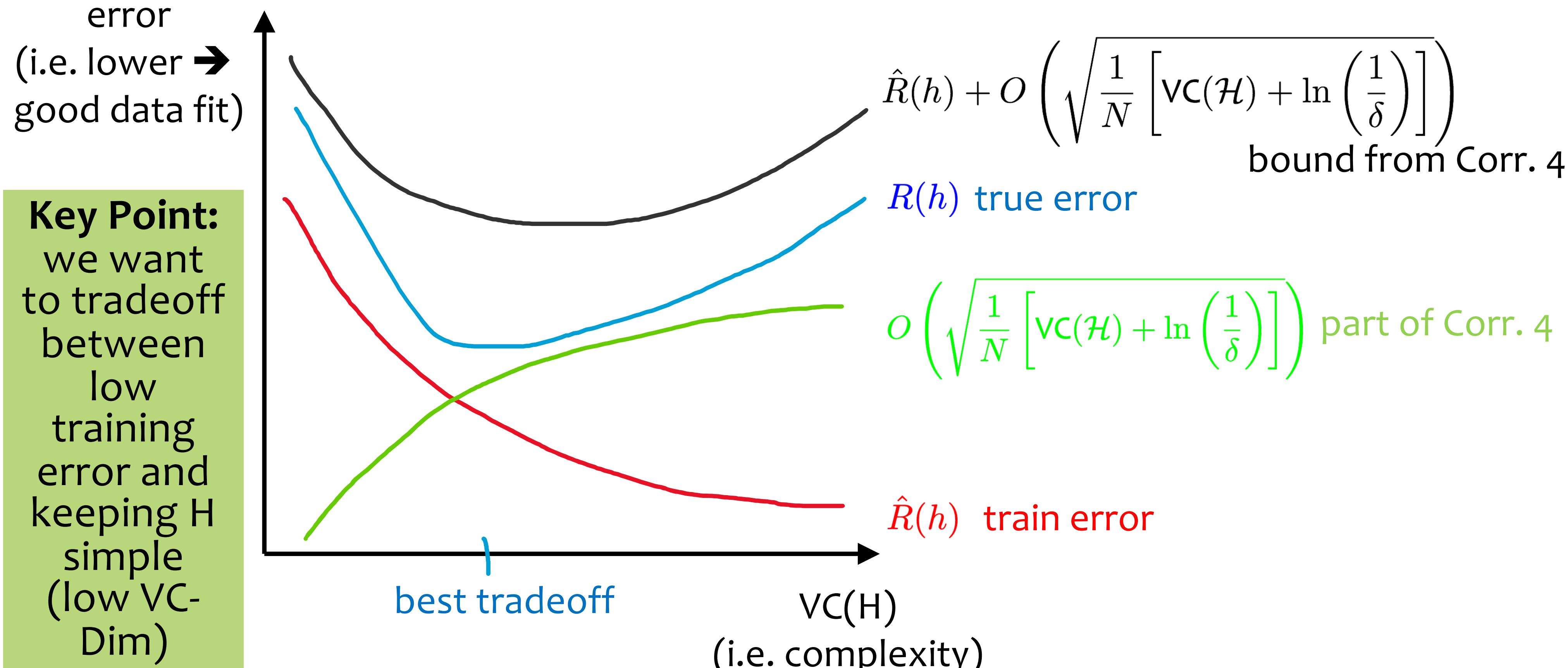
Should these corollaries inform how we do model selection?

Learning Theory & Model Selection

Key Point: we want to tradeoff between low training error and keeping H simple (low VC-Dim)

Q: Is Corollary 4 useful?
A: Yes!

Learning Theory & Model Selection



Key Point: we want to tradeoff between low training error and keeping H simple (low VC-Dim)

Q: Is Corollary 4 useful?
A: Yes!

Ex: H = Linear Separators in R^M

VC(H) = M+1

Q: In practice, how do we tradeoff between error and VC(H)?

A: Use a regularizer! That is, reducing the number of (effective) features reduces the VC dimension. More features usually leads to a better fit to the data.

Questions For Today

1. Given a classifier with zero training error, what can we say about generalization error?
   (Sample Complexity, Realizable Case)

2. Given a classifier with low training error, what can we say about generalization error?
   (Sample Complexity, Agnostic Case)

3. Is there a theoretical justification for regularization to avoid overfitting?
   (Structural Risk Minimization)

Learning Theory Objectives

You should be able to…

- Identify the properties of a learning setting and assumptions required to ensure low generalization error
- Distinguish true error, train error, test error
- Define PAC and explain what it means to be approximately correct and what occurs with high probability
- Apply sample complexity bounds to real-world learning examples
- Distinguish between a large sample and a finite sample analysis
- Theoretically motivate regularization

PROBABILITY

Random Variables: Definitions

Discrete Random Variable	X	Random variable whose values come from a countable set (e.g. the natural numbers or {True, False})
Probability mass function (pmf)	$p(x)$	Function giving the probability that discrete r.v. X takes value x. $$p(x) := P(X = x)$$

Random Variables: Definitions

Continuous Random Variable	X	Random variable whose values come from an interval or collection of intervals (e.g. the real numbers or the range (3, 5))
Probability density function (pdf)	$f(x)$	Function the returns a nonnegative real indicating the relative likelihood that a continuous r.v. X takes value x

- For any continuous random variable: $P(X = x) = 0$
- Non-zero probabilities are only available to intervals:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Random Variables: Definitions

Cumulative distribution function	$F(x)$	Function that returns the probability that a random variable X is less than or equal to x: $$F(x) = P(X \leq x)$$

- For **discrete** random variables:

$$F(x) = P(X \leq x) = \sum_{x' < x} P(X = x') = \sum_{x' < x} p(x')$$

- For **continuous** random variables:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x')dx'$$

Notational Shortcuts

A convenient shorthand:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

$\Rightarrow$ For all values of a and b:

$$P(A = a|B = b) = \frac{P(A = a, B = b)}{P(B = b)}$$

Notational Shortcuts

But then how do we tell $P(E)$ apart from $P(X)$?

Event

Random Variable

Instead of writing:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

We should write:

$$P_{A|B}(A|B) = \frac{P_{A,B}(A,B)}{P_B(B)}$$

…but only probability theory textbooks go to such lengths.

COMMON PROBABILITY DISTRIBUTIONS

Common Probability Distributions

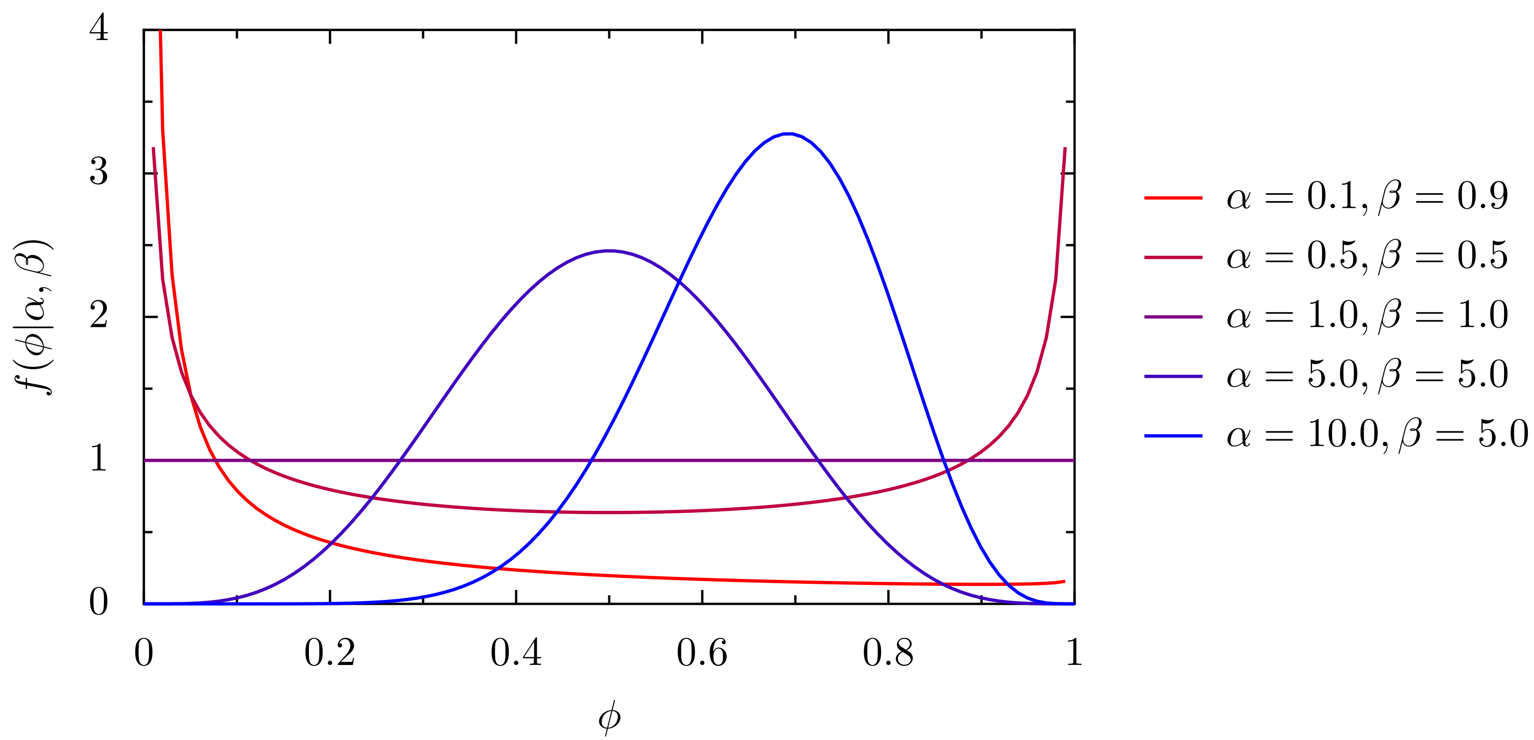
- For Discrete Random Variables:
  - Bernoulli
  - Binomial
  - Multinomial
  - Categorical
  - Poisson
- For Continuous Random Variables:
  - Exponential
  - Gamma
  - Beta
  - Dirichlet
  - Laplace
  - Gaussian (1D)
  - Multivariate Gaussian

Common Probability Distributions

Beta Distribution

probability density function:

$$f(\phi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

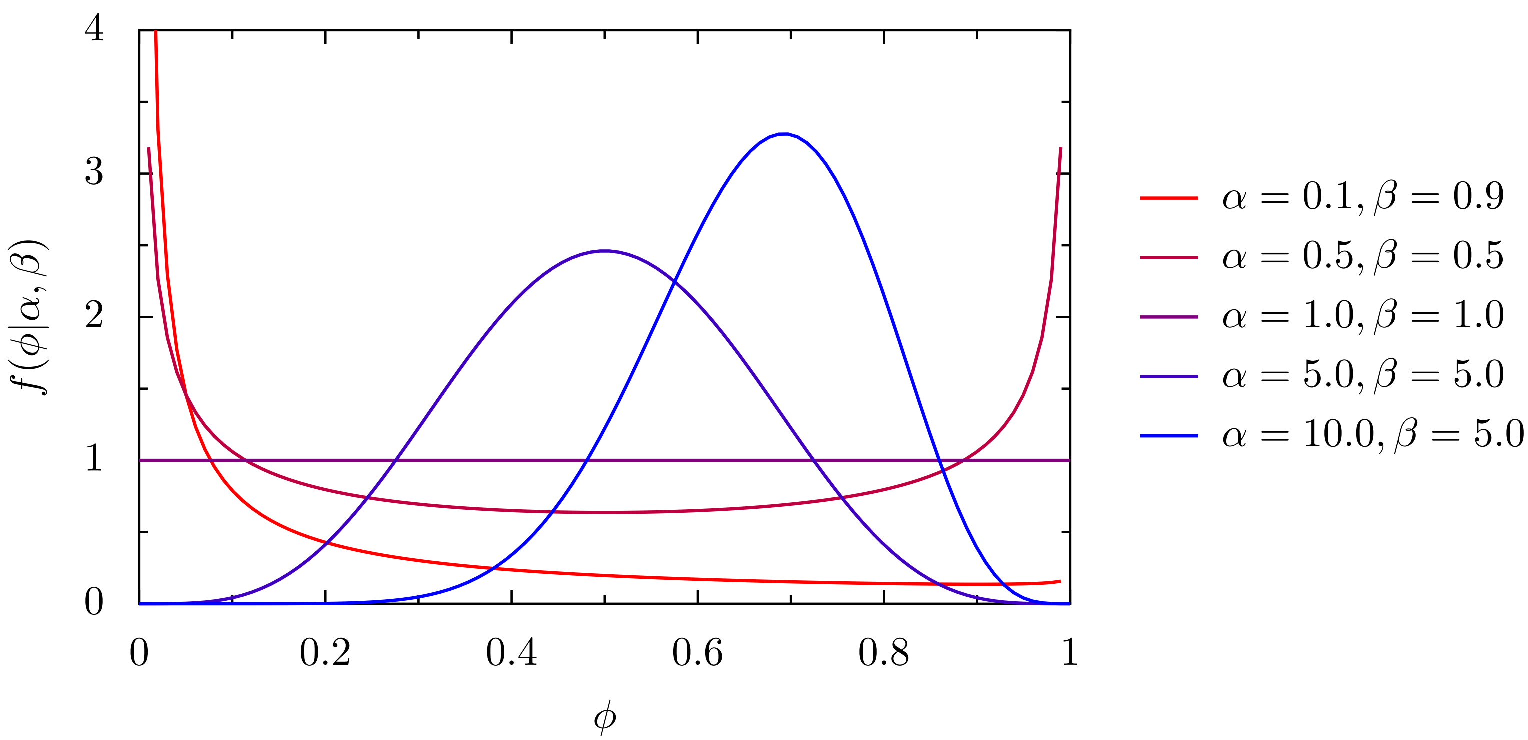


Common Probability Distributions

Dirichlet Distribution

probability density function:

$$f(\phi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

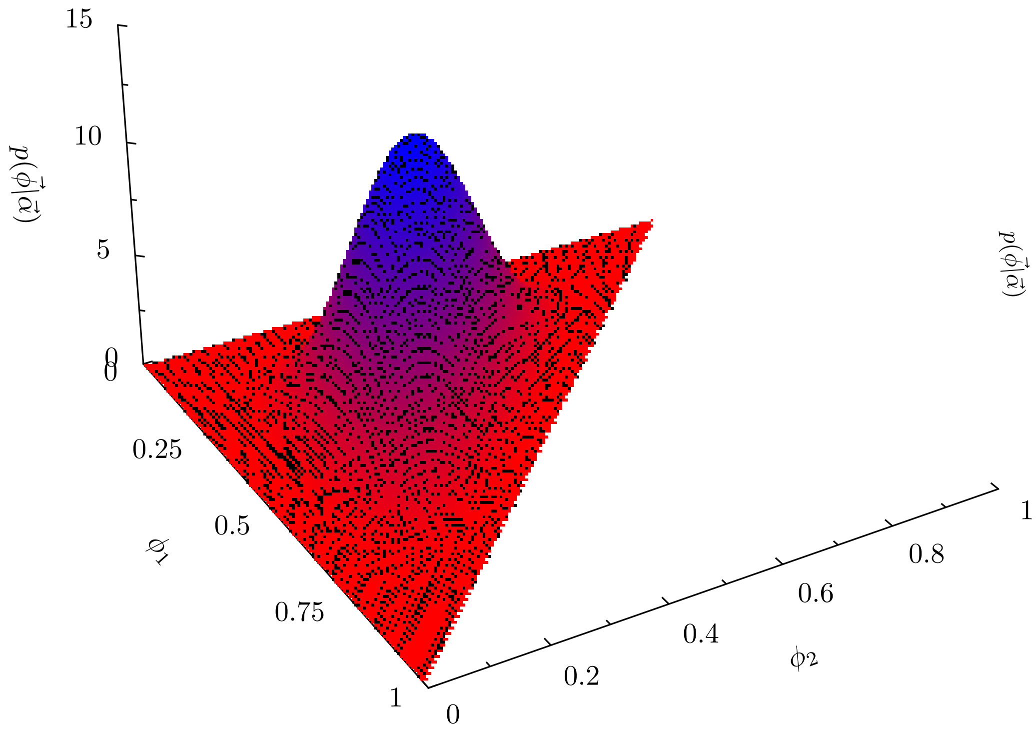


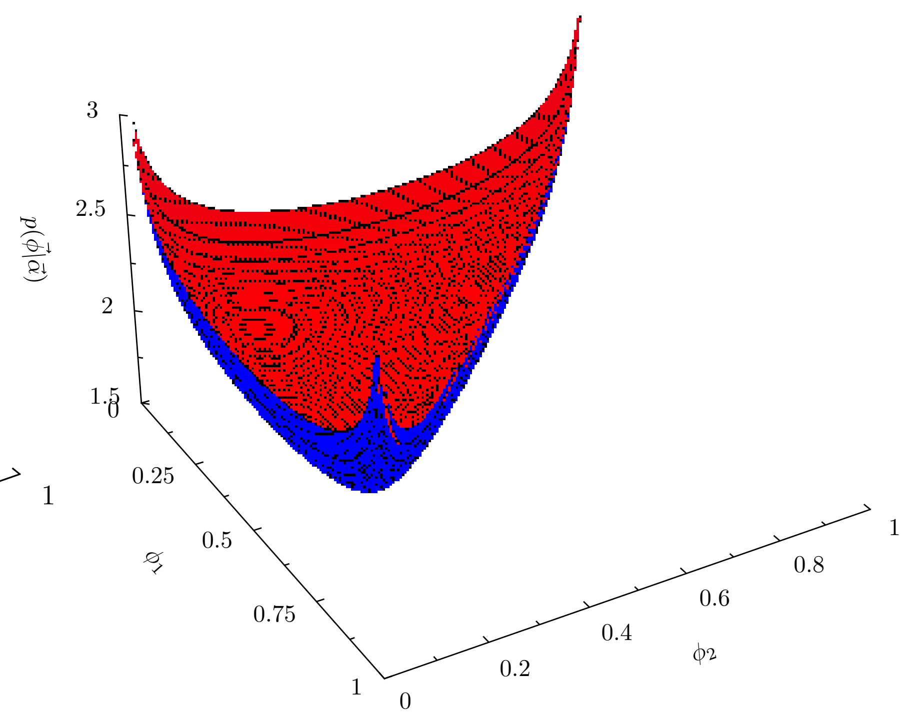
Common Probability Distributions

Dirichlet Distribution

probability density function:

$$p(\vec{\phi}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K}\phi_k^{\alpha_k - 1} \quad \text{where } B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)}$$





EXPECTATION AND VARIANCE

Expectation and Variance

The **expected value** of X is $E[X]$. Also called the mean.

- Discrete random variables:

  Suppose X can take any value in the set $\mathcal{X}$.

  $$E[X] = \sum_{x \in \mathcal{X}} x p(x)$$

- Continuous random variables:

  $$E[X] = \int_{-\infty}^{+\infty} x f(x) dx$$

Expectation and Variance

The **variance** of X is $Var(X)$.

$$Var(X) = E[(X - E[X])^2]$$

$$\mu = E[X]$$

- Discrete random variables:

$$Var(X) = \sum_{x \in \mathcal{X}} (x - \mu)^2 p(x)$$

- Continuous random variables:

$$Var(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

Joint probability

Marginal probability

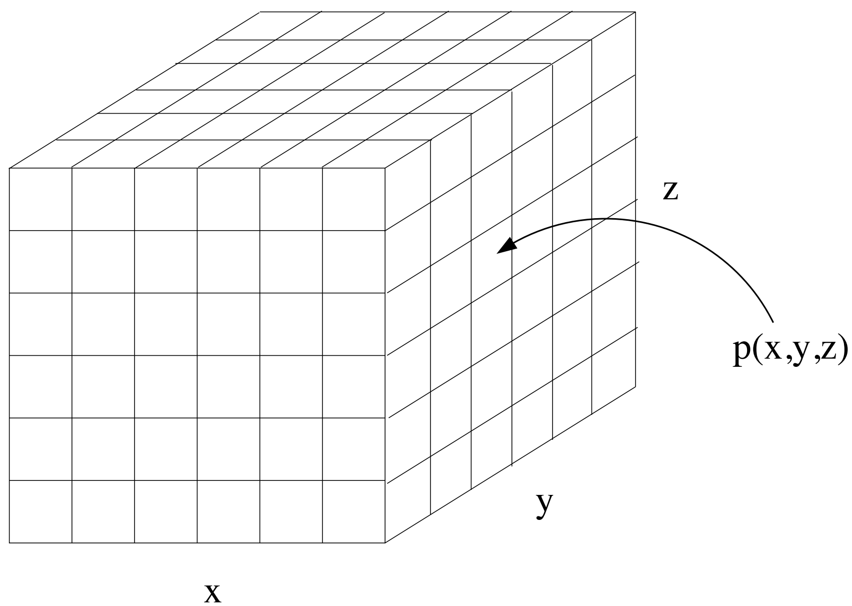
Conditional probability

MULTIPLE RANDOM VARIABLES

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

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



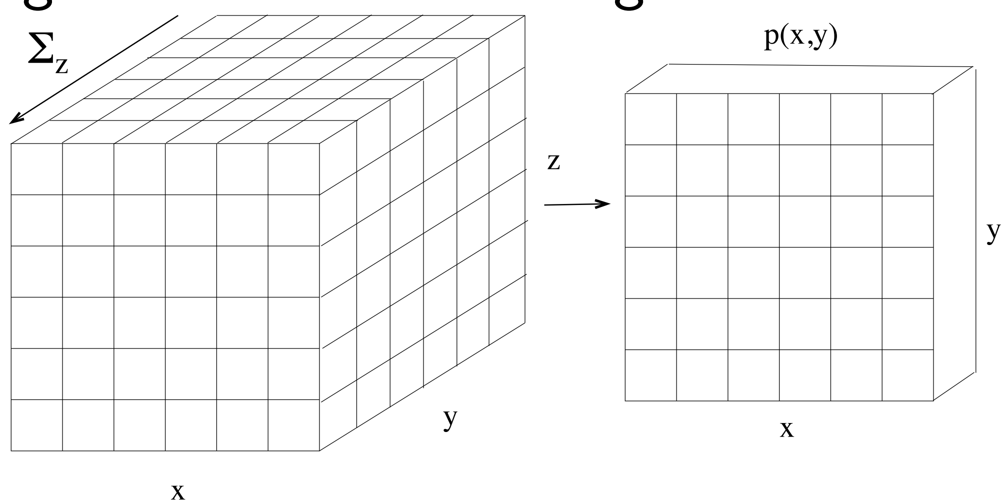
Slide from Sam Roweis (MLSS, 2005)

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.



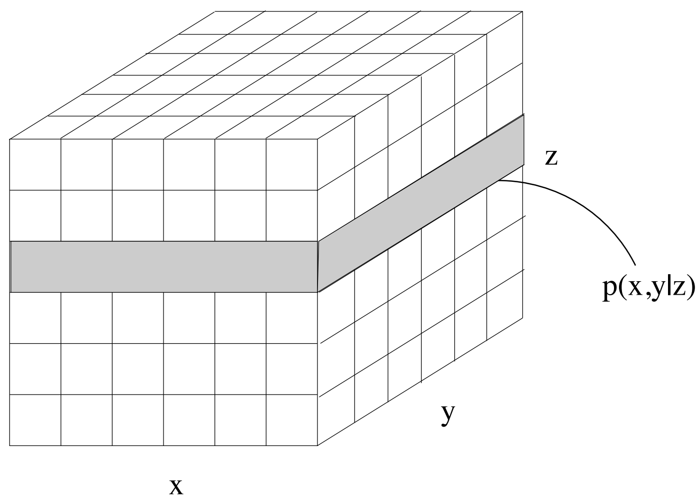
- Another equivalent definition: $p(x) = \sum_y p(x|y)p(y)$.

Slide from Sam Roweis (MLSS, 2005)

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.

$$\boxed{p(x|y) = p(x,y)/p(y)}$$

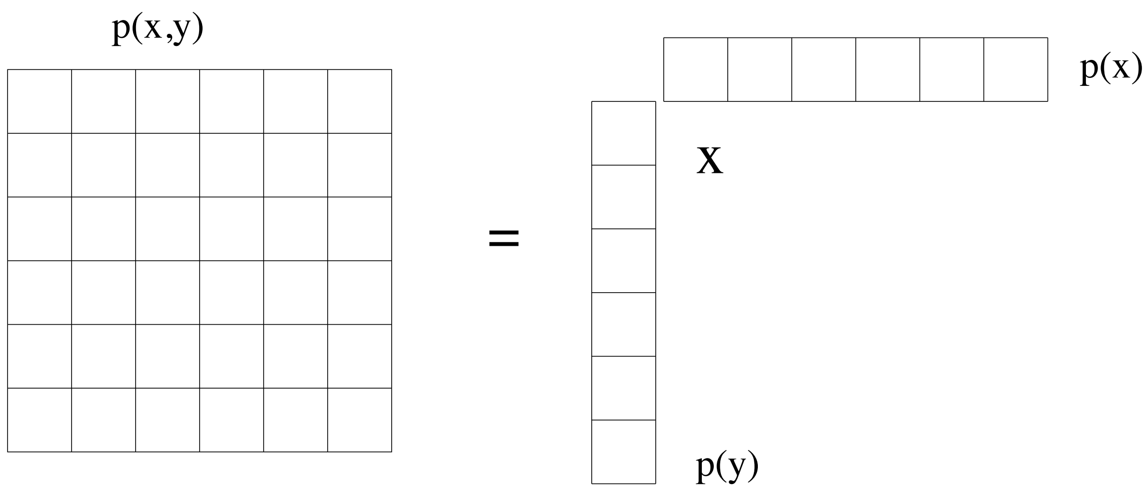


Slide from Sam Roweis (MLSS, 2005)

Independence and 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:

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

Slide from Sam Roweis (MLSS, 2005)

PROBABILISTIC LEARNING

Probabilistic Learning

Function Approximation

Previously, we assumed that our output was generated using a **deterministic target function:**

$$\mathbf{x}^{(i)} \sim p^*(\cdot)$$

$$y^{(i)} = c^*(\mathbf{x}^{(i)})$$

Our goal was to learn a hypothesis h(**x**) that best approximates c*(**x**)

Probabilistic Learning

Today, we assume that our output is **sampled** from a conditional **probability distribution:**

$$\mathbf{x}^{(i)} \sim p^*(\cdot)$$

$$y^{(i)} \sim p^*(\cdot|\mathbf{x}^{(i)})$$

Our goal is to learn a probability distribution p(y|**x**) that best approximates p*(y|**x**)

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

Likelihood Function

One R.V.

- Given N **independent, identically distributed (iid)** samples $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$ from a **random variable** X …

- The **likelihood** function is
  - Case 1: X is **discrete** with probability mass function (*pmf*) $p(x|\theta)$
    $$L(\theta) = p(x^{(1)}|\theta)\, p(x^{(2)}|\theta) \ldots\ p(x^{(N)}|\theta)$$
  - Case 2: X is **continuous** with probability density function (pdf) $f(x|\theta)$
    $$L(\theta) = f(x^{(1)}|\theta)\, f(x^{(2)}|\theta) \ldots\ f(x^{(N)}|\theta)$$

The **likelihood** tells us how likely one sample is relative to another

- The **log-likelihood** function is
  - Case 1: X is **discrete** with probability mass function (*pmf*) $p(x|\theta)$
    $$\ell(\theta) = \log p(x^{(1)}|\theta) + \ldots + \log p(x^{(N)}|\theta)$$
  - Case 2: X is **continuous** with probability density function (pdf) $f(x|\theta)$
    $$\ell(\theta) = \log f(x^{(1)}|\theta) + \ldots + \log f(x^{(N)}|\theta)$$

Likelihood Function

Two R.V.s

- Given N **iid** samples $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ from a pair of **random variables** X, Y

- The **conditional likelihood** function:
  - Case 1: Y is **discrete** with *pmf* $p(y \mid x, \theta)$

    $L(\theta) = p(y^{(1)} \mid x^{(1)}, \theta) \ldots p(y^{(N)} \mid x^{(N)}, \theta)$
  - Case 2: Y is **continuous** with *pdf* $f(y \mid x, \theta)$

    $L(\theta) = f(y^{(1)} \mid x^{(1)}, \theta) \ldots f(y^{(N)} \mid x^{(N)}, \theta)$

- The **joint likelihood** function:
  - Case 1: X and Y are **discrete** with *pmf* $p(x,y|\theta)$

    $L(\theta) = p(x^{(1)}, y^{(1)}|\theta) \ldots p(x^{(N)}, y^{(N)}|\theta)$
  - Case 2: X and Y are **continuous** with *pdf* $f(x,y|\theta)$

    $L(\theta) = f(x^{(1)}, y^{(1)}|\theta) \ldots f(x^{(N)}, y^{(N)}|\theta)$

Likelihood Function

Two R.V.s

- Given N **iid** samples $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ from a pair of **random variables** X, Y

- The **joint likelihood** function:
  - Case 1: X and Y are **discrete** with *pmf* $p(x,y|\theta)$
    $L(\theta) = p(x^{(1)}, y^{(1)}|\theta) \ldots p(x^{(N)}, y^{(N)}|\theta)$
  - Case 2: X and Y are **continuous** with *pdf* $f(x,y|\theta)$
    $L(\theta) = f(x^{(1)}, y^{(1)}|\theta) \ldots f(x^{(N)}, y^{(N)}|\theta)$
  - Case 3: Y is **discrete** with *pmf* $p(y|\beta)$ and
    X is **continuous** with *pdf* $f(x|y,\alpha)$
    $L(\alpha, \beta) = f(x^{(1)}| y^{(1)}, \alpha)\, p(y^{(1)}|\beta) \ldots f(x^{(N)}| y^{(N)}, \alpha)\, p(y^{(N)}|\beta)$
  - Case 4: Y is **continuous** with *pdf* $f(y|\beta)$ and
    X is **discrete** with *pmf* $p(x|y,\alpha)$
    $L(\alpha, \beta) = p(x^{(1)}| y^{(1)}, \alpha)\, f(y^{(1)}|\beta) \ldots p(x^{(N)}| y^{(N)}, \alpha)\, f(y^{(N)}|\beta)$

Mixed discrete/ continuous!

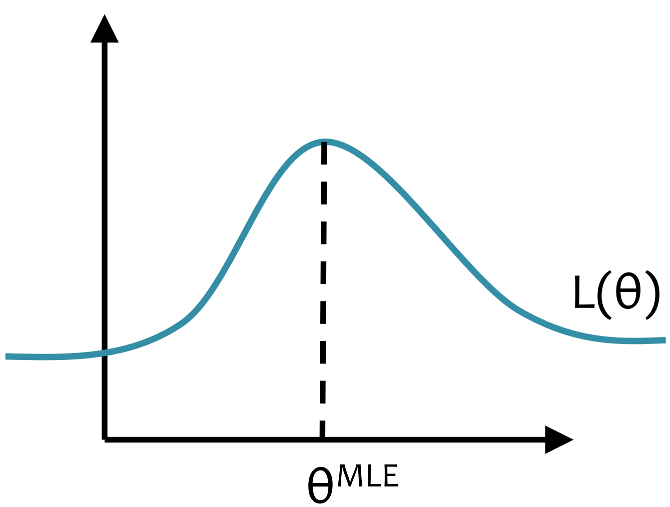
MLE

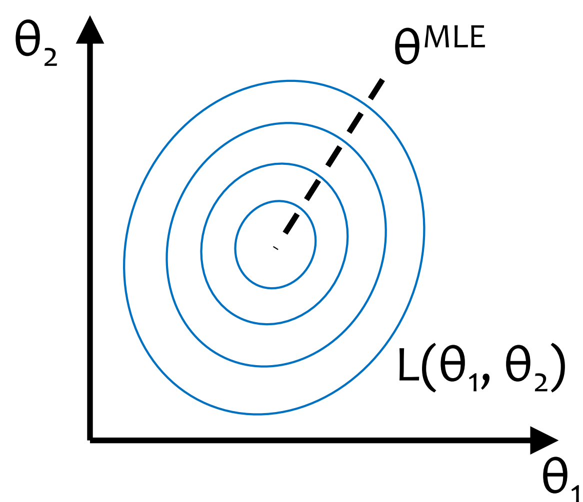
Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

Principle of Maximum Likelihood Estimation:
Choose the parameters that maximize the likelihood of the data.

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)} | \boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)





MLE

What does maximizing likelihood accomplish?

- There is only a finite amount of probability mass (i.e. sum-to-one constraint)
- MLE tries to allocate **as much** probability mass **as possible** to the things we have observed…

  … **at the expense** of the things we have **not** observed

Recipe for Closed-form MLE

1. Assume data was generated iid from some model, i.e., write the *generative story*
   $x^{(i)} \sim p(x|\boldsymbol{\theta})$
2. Write the log-likelihood
   $\ell(\boldsymbol{\theta}) = \log p(x^{(1)}|\boldsymbol{\theta}) + \ldots + \log p(x^{(N)}|\boldsymbol{\theta})$
3. Compute partial derivatives, i.e., the gradient
   $\partial\ell(\boldsymbol{\theta})/\partial\theta_1 = \ldots$
   $\ldots$
   $\partial\ell(\boldsymbol{\theta})/\partial\theta_M = \ldots$
4. Set derivatives equal to zero and solve for $\boldsymbol{\theta}$
   $\partial\ell(\boldsymbol{\theta})/\partial\theta_m = 0$ for all $m \in \{1, \ldots, M\}$
   $\boldsymbol{\theta}^{MLE}$ = solution to system of M equations and M variables
5. Compute the second derivative and check that $\ell(\boldsymbol{\theta})$ is concave down at $\boldsymbol{\theta}^{MLE}$

What we earlier called “Closed Form Solution for Linear Regression”

EXAMPLE:
MLE FOR LINEAR REGRESSION

Linear Regression as Function Approximation

$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$
where $\mathbf{x} \in \mathbb{R}^M$ and $y \in \mathbb{R}$

1. Assume $\mathcal{D}$ generated as:

$$\mathbf{x}^{(i)} \sim p^*(\cdot)$$
$$y^{(i)} = h^*(\mathbf{x}^{(i)})$$

2. Choose hypothesis space, $\mathcal{H}$:
all linear functions in M-dimensional space

$$\mathcal{H} = \{h_{\boldsymbol{\theta}} : h_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x}, \boldsymbol{\theta} \in \mathbb{R}^M\}$$

3. Choose an objective function:
mean squared error (MSE)

$$\begin{aligned} J(\boldsymbol{\theta}) &= \frac{1}{N} \sum_{i=1}^N e_i^2 \\ &= \frac{1}{N} \sum_{i=1}^N \left(y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) \right)^2 \\ &= \frac{1}{N} \sum_{i=1}^N \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)} \right)^2 \end{aligned}$$

4. Solve the unconstrained optimization problem via favorite method:
   - *gradient descent*
   - *closed form*
   - *stochastic gradient descent*
   - . . .

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, J(\boldsymbol{\theta})$$

5. Test time: given a new $\mathbf{x}$, make prediction $\hat{y}$

$$\hat{y} = h_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) = \hat{\boldsymbol{\theta}}^T \mathbf{x}$$

Linear Regression: Closed Form

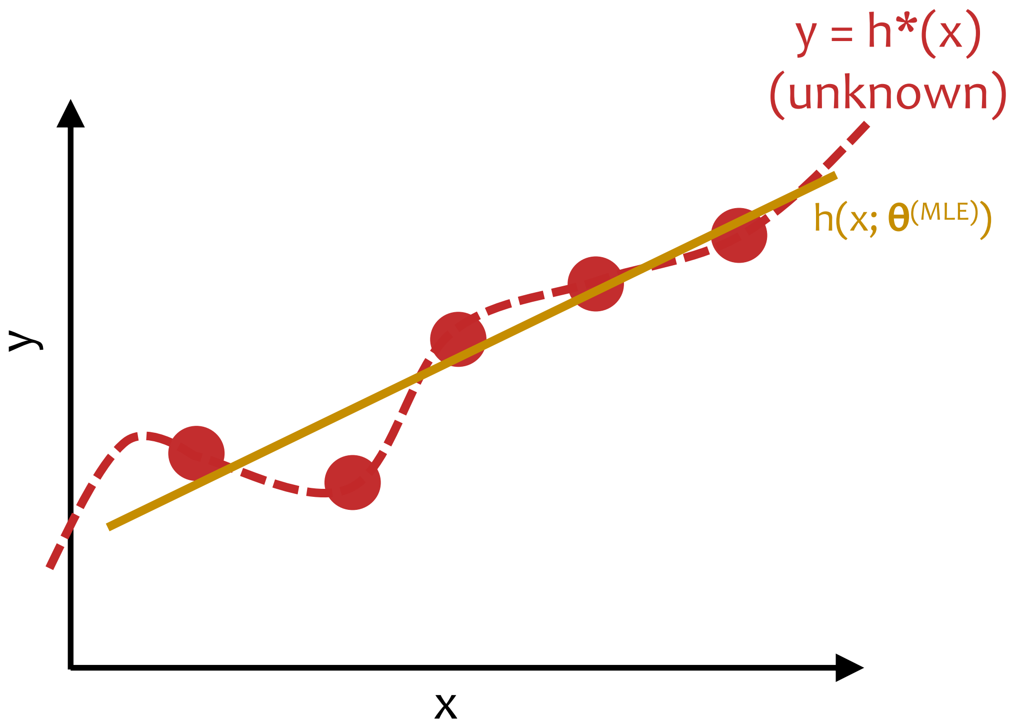
Optimization Method #2: Closed Form

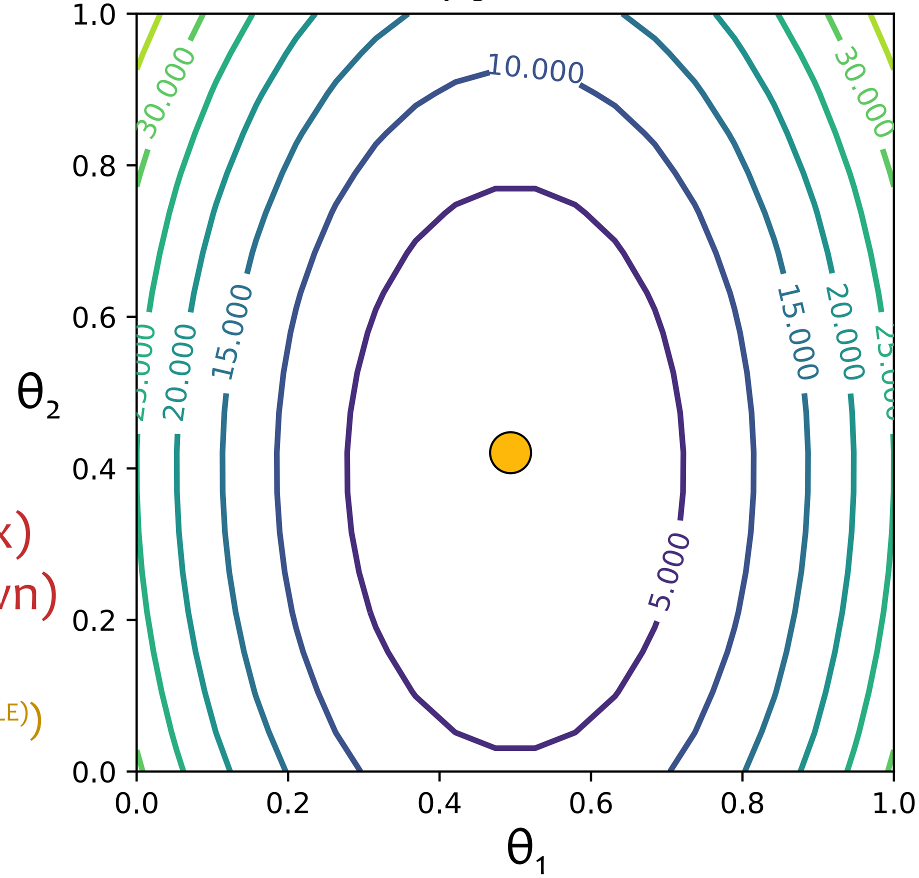
1. Evaluate

$$\boldsymbol{\theta}^{\text{MLE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

2. Return $\boldsymbol{\theta}^{\text{MLE}}$

$$J(\boldsymbol{\theta}) = J(\theta_1, \theta_2) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)^2$$





t	θ_1	θ_2	$J(\theta_1, \theta_2)$
MLE	0.59	0.43	0.2

MLE for Linear Regression

You'll work through the view of linear regression as a probabilistic model in the homework!

MLE EXAMPLES

MLE of Exponential Distribution

Goal:

- pdf of Exponential(λ): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim$ Exponential(λ) for $1 \leq i \leq N$.
- Find MLE for data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

Steps:

- First write down log-likelihood of sample.
- Compute first derivative, set to zero, solve for λ.
- Compute second derivative and check that it is concave down at λ^{MLE}.

MLE of Exponential Distribution

- pdf of Exponential(λ): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim$ Exponential(λ) for $1 \leq i \leq N$.
- Find MLE for data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

- First write down log-likelihood of sample.

$$\ell(\lambda) = \sum_{i=1}^{N} \log f(x^{(i)}) \tag{1}$$

$$= \sum_{i=1}^{N} \log(\lambda \exp(-\lambda x^{(i)})) \tag{2}$$

$$= \sum_{i=1}^{N} \log(\lambda) + -\lambda x^{(i)} \tag{3}$$

$$= N \log(\lambda) - \lambda \sum_{i=1}^{N} x^{(i)} \tag{4}$$

MLE of Exponential Distribution

- pdf of Exponential(λ): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim$ Exponential(λ) for $1 \leq i \leq N$.
- Find MLE for data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

- Compute first derivative, set to zero, solve for λ.

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{d}{d\lambda} N \log(\lambda) - \lambda \sum_{i=1}^{N} x^{(i)} \quad (1)$$

$$= \frac{N}{\lambda} - \sum_{i=1}^{N} x^{(i)} = 0 \quad (2)$$

$$\boxed{\Rightarrow \lambda^{\text{MLE}} = \frac{N}{\sum_{i=1}^{N} x^{(i)}}} \quad (3)$$

MLE of Bernoulli

In-Class Exercise

Show that the MLE of parameter ϕ for N samples drawn from Bernoulli(ϕ) is:

$$\phi_{MLE} = \frac{\text{Number of } x_i = 1}{N}$$

Steps to answer:

1. Write log-likelihood of sample
2. Compute derivative w.r.t. ϕ
3. Set derivative to zero and solve for ϕ

MLE of Bernoulli

Question:

Assume we have N iid samples $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ drawn from a Bernoulli(ϕ).

Step 1: What is the **log-likelihood** of the data $\ell(\phi)$?

Assume $N_1 = \# \text{ of } (x^{(i)} = 1)$

$N_0 = \# \text{ of } (x^{(i)} = 0)$

Answer:

A. $l(\phi) = N_1 \log(\phi) + N_0 (1 - \log(\phi))$

B. $l(\phi) = N_1 \log(\phi) + N_0 \log(1-\phi)$

C. $l(\phi) = \log(\phi)^{N_1} + (1 - \log(\phi))^{N_0}$

D. $l(\phi) = \log(\phi)^{N_1} + \log(1-\phi)^{N_0}$

E. $l(\phi) = N_0 \log(\phi) + N_1 (1 - \log(\phi))$

F. $l(\phi) = N_0 \log(\phi) + N_1 \log(1-\phi)$

G. $l(\phi) = \log(\phi)^{N_0} + (1 - \log(\phi))^{N_1}$

H. $l(\phi) = \log(\phi)^{N_0} + \log(1-\phi)^{N_1}$

I. $l(\phi) = N_0 + N_1$

MLE of Bernoulli

Question:

Assume we have N iid samples $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ drawn from a Bernoulli(ϕ).

Step 2: What is the **derivative** of the log-likelihood $\partial\ell(\boldsymbol{\theta})/\partial\theta$?

Assume $N_1 = \#\ of\ (x^{(i)} = 1)$

$N_0 = \#\ of\ (x^{(i)} = 0)$

Answer:

A. $\partial\ell(\boldsymbol{\theta})/\partial\theta = \phi^{N_1} - (1 - \phi)^{N_0}$

B. $\partial\ell(\boldsymbol{\theta})/\partial\theta = \phi / N_1 - (1 - \phi) / N_0$

C. $\partial\ell(\boldsymbol{\theta})/\partial\theta = N_1 / \phi - N_0 / (1 - \phi)$

D. $\partial\ell(\boldsymbol{\theta})/\partial\theta = \log(\phi) / N_1 - \log(1 - \phi) / N_0$

E. $\partial\ell(\boldsymbol{\theta})/\partial\theta = N_1 / \log(\phi) - N_0 / \log(1 - \phi)$

F. $\partial\ell(\boldsymbol{\theta})/\partial\theta = 0$

MLE of Bernoulli

Whiteboard

– Example: MLE of Bernoulli