UCLA Samueli Computer Science

THE UNIVERSITY OF CALIFORNIA · LET THERE BE LIGHT · UCLA

CS145 Discussion: Week 1

Math Prep Collection

Junheng Hao

Friday, 10/09/2020

Roadmap

- Math review
  - Probability
  - Linear algebra
  - Optimization
  - Matrix calculus

Math Review

- Slides reference
  - Jeff Howbert, https://courses.washington.edu/css490/2012.Winter/lecture_slides/02_math_essentials.pdf
  - Xinkun Nie, http://cs229.stanford.edu/notes2020fall/notes2020fall/TA-slides1.pdf
  - Hristo Paskov, http://snap.stanford.edu/class/cs246-2014/slides/LinAlgSession.pdf

Probability

Probability spaces

A *probability space* is a *random process* or *experiment* with three components:

- Ω, the set of possible *outcomes* O
  - number of possible outcomes = $| \Omega | = N$
- F, the set of possible *events* E
  - an event comprises 0 to N outcomes
  - number of possible events = $| F | = 2^N$
- P, the *probability distribution*
  - function mapping each outcome and event to real number between 0 and 1 (the *probability* of O or E)
  - probability of an event is *sum* of probabilities of possible outcomes in event

Probability

Axioms of probability

1. Non-negativity:

   for any event $E \in F$, $p(E) \geq 0$

2. All possible outcomes:

   $p(\Omega) = 1$

3. Additivity of disjoint events:

   for all events $E, E' \in F$ where $E \cap E' = \varnothing$,
   $p(E \cup E') = p(E) + p(E')$

Probability

Types of probability spaces

Define | Ω | = number of possible outcomes

- Discrete space | Ω | is finite
  - Analysis involves *summations* (Σ)
- Continuous space | Ω | is infinite
  - Analysis involves *integrals* ($\int$)

Probability

Example of discrete probability space

Single roll of a six-sided die

- 6 possible outcomes: O = 1, 2, 3, 4, 5, or 6
- $2^6 = 64$ possible events
  - example: $E = (O \in \{ 1, 3, 5 \})$, i.e. outcome is odd
- If die is fair, then probabilities of outcomes are equal

  $p(1) = p(2) = p(3) =$
  $p(4) = p(5) = p(6) = 1 / 6$

  - example: probability of event E = (outcome is odd) is
    $p(1) + p(3) + p(5) = 1 / 2$

Probability

Example of discrete probability space

Three consecutive flips of a coin

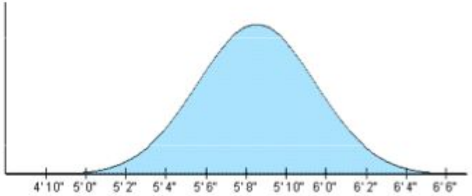
- 8 possible outcomes: O = HHH, HHT, HTH, HTT, THH, THT, TTH, TTT
- $2^8 = 256$ possible events
  - example: E = ($O \in$ { HHT, HTH, THH }), i.e. exactly two flips are heads
  - example: E = ($O \in$ { THT, TTT }), i.e. the first and third flips are tails
- If coin is fair, then probabilities of outcomes are equal
  p(HHH) = p(HHT) = p(HTH) = p(HTT) = p(THH) = p(THT) = p(TTH) = p(TTT) = 1 / 8
  - example: probability of event E = (exactly two heads) is
    p(HHT) + p(HTH) + p(THH) = 3 / 8

Probability

Example of continuous probability space

Height of a randomly chosen American male

- Infinite number of possible outcomes: *O* has some single value in range 2 feet to 8 feet
- Infinite number of possible events
  - example: $E = (O \mid O < 5.5 \text{ feet})$, i.e. individual chosen is less than 5.5 feet tall
- Probabilities of outcomes are not equal, and are described by a continuous function, $p(O)$

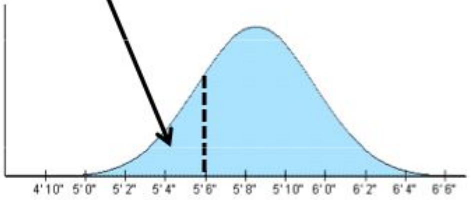


Probability

Example of continuous probability space

Height of a randomly chosen American male

- Probabilities of outcomes O are not equal, and are described by a continuous function, $p(O)$
- $p(O)$ is a *relative*, not an *absolute* probability
  - $p(O)$ for any particular O is zero
  - $\int p(O)$ from $O = -\infty$ to ∞ (i.e. area under curve) is 1
  - example: $p(O = 5'8'') > p(O = 6'2'')$
  - example: $p(O < 5'6'') = (\int p(O)$ from $O = -\infty$ to $5'6'') \approx 0.25$



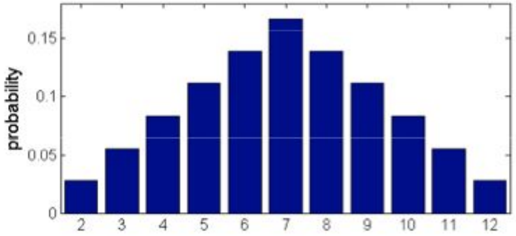
UCLA Engineer Change.

Probability

Probability distributions

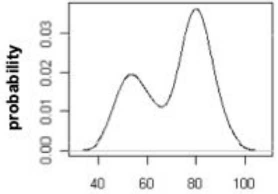
- Discrete: *probability mass function* (pmf)

  example:
  sum of two
  fair dice



- Continuous: *probability density function* (pdf)

  example:
  waiting time between
  eruptions of Old Faithful
  (minutes)



Probability

Random variables

- A random variable X is a function that associates a number x with each outcome O of a process
  - Common notation: $X(O) = x$, or just $X = x$
- Basically a way to redefine (usually simplify) a probability space to a new probability space
  - X must obey axioms of probability (over the possible values of x)
  - X can be discrete or continuous
- Example: X = number of heads in three flips of a coin
  - Possible values of X are 0, 1, 2, 3
  - $p(X = 0) = p(X = 3) = 1 / 8$ $p(X = 1) = p(X = 2) = 3 / 8$
  - Size of space (number of "outcomes") reduced from 8 to 4
- Example: X = average height of five randomly chosen American men
  - Size of space unchanged (X can range from 2 feet to 8 feet), but pdf of X different than for single man

Probability

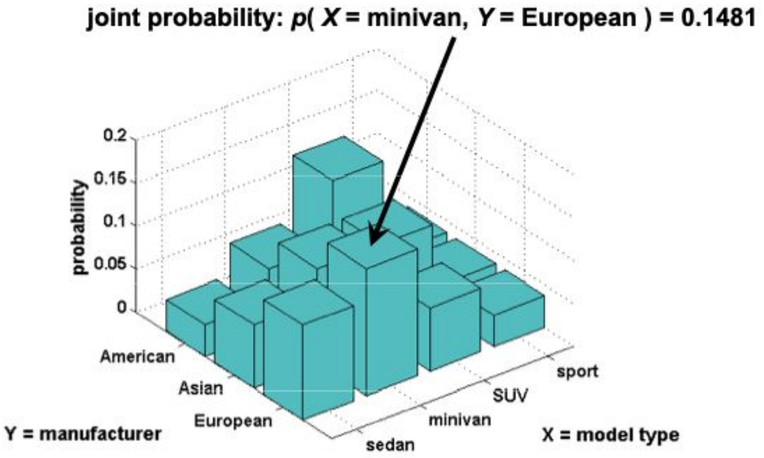
Multivariate probability distributions

- Scenario
  - Several random processes occur (doesn't matter whether in parallel or in sequence)
  - Want to know probabilities for each possible combination of outcomes
- Can describe as *joint probability* of several random variables
  - Example: two processes whose outcomes are represented by random variables X and Y. Probability that process X has outcome x and process Y has outcome y is denoted as:

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

Probability

Example of multivariate distribution



UCLA Engineer Change.

Probability

Multivariate probability distributions

- *Marginal* probability
  - Probability distribution of a single variable in a joint distribution
  - Example: two random variables *X* and *Y*:

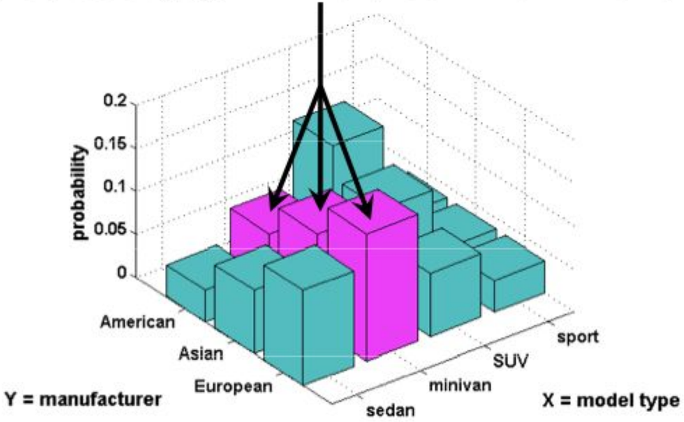
    $p(X = x) = \sum_{b=\text{all values of } Y} p(X = x, Y = b)$
- *Conditional* probability
  - Probability distribution of one variable *given* that another variable takes a certain value
  - Example: two random variables *X* and *Y*:

    $p(X = x \mid Y = y) = p(X = x, Y = y) / p(Y = y)$

Probability

Example of marginal probability

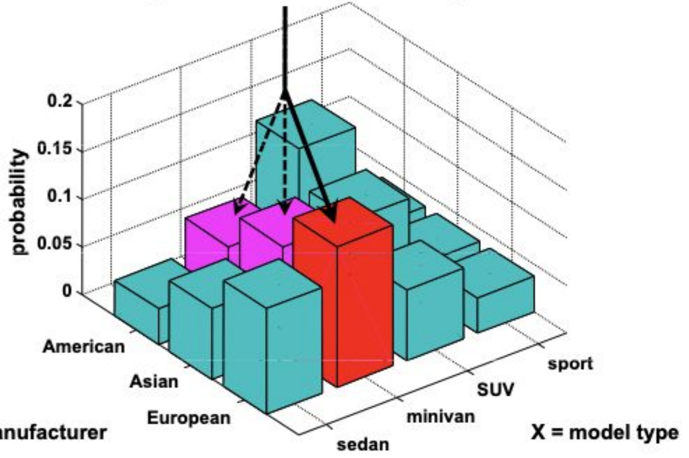
marginal probability: $p(X = \text{minivan}) = 0.0741 + 0.1111 + 0.1481 = 0.3333$



Probability

Example of conditional probability

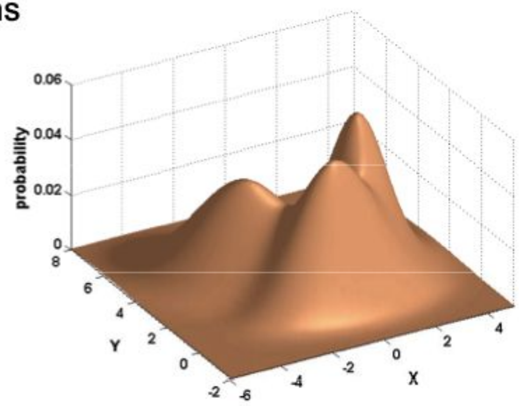
conditional probability: $p(\ Y = \text{European} \mid X = \text{minivan}\ ) =$
$0.1481 / (\ 0.0741 + 0.1111 + 0.1481\ ) = 0.4433$



Probability

Continuous multivariate distribution

- Same concepts of joint, marginal, and conditional probabilities apply (except use integrals)
- Example: three-component Gaussian mixture in two dimensions



Probability

Expected value

Given:

- A discrete random variable X, with possible values $x = x_1, x_2, \ldots x_n$
- Probabilities $p(X = x_i)$ that X takes on the various values of x_i
- A function $y_i = f(x_i)$ defined on X

The *expected value* of f is the probability-weighted "average" value of $f(x_i)$:

$$E(f) = \Sigma_i \, p(x_i) \cdot f(x_i)$$

Probability

Example of expected value

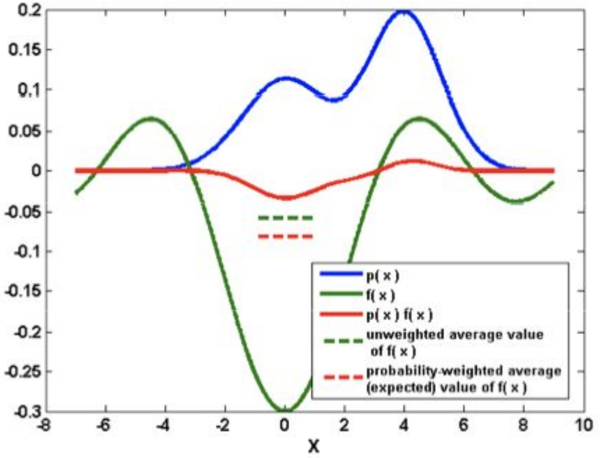
- Process: game where one card is drawn from the deck
  - If face card, dealer pays you $10
  - If not a face card, you pay dealer $4
- Random variable X = { face card, not face card }
  - p(face card) = 3/13
  - p(not face card) = 10/13
- Function $f(X)$ is payout to you
  - f(face card) = 10
  - f(not face card) = -4
- *Expected value* of payout is:

$$E(f) = \Sigma_i \, p(x_i) \cdot f(x_i) = 3/13 \cdot 10 + 10/13 \cdot -4 = -0.77$$

Probability

Expected value in continuous spaces

$$E(f) = \int_{x=a \to b} p(x) \cdot f(x)$$



Probability

Common forms of expected value (1)

- Mean (μ)

  $f(x_i) = x_i \Rightarrow \quad \mu = E(f) = \sum_i p(x_i) \cdot x_i$

  - Average value of $X = x_i$, taking into account probability of the various x_i
  - Most common measure of "center" of a distribution

- Compare to formula for mean of an actual sample

  $$\mu = \frac{1}{N}\sum_{i=1}^{n} x_i$$

Probability

Common forms of expected value (2)

- Variance (σ^2)

  $f(x_i) = (x_i - \mu) \quad \Rightarrow \quad \sigma^2 = \sum_i p(x_i) \cdot (x_i - \mu)^2$

  - Average value of squared deviation of $X = x_i$ from mean μ, taking into account probability of the various x_i
  - Most common measure of "spread" of a distribution
  - σ is the *standard deviation*

- Compare to formula for variance of an actual sample

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

Probability

Common forms of expected value (3)

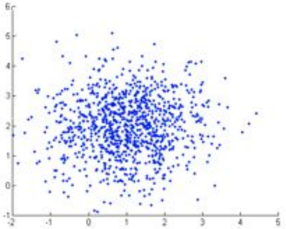
- Covariance

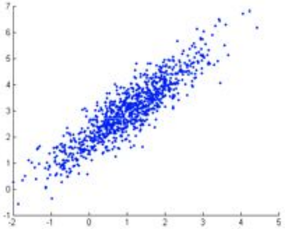
  $f(x_i) = (x_i - \mu_x)$, $g(y_i) = (y_i - \mu_y)$ $\Rightarrow$

  $\text{cov}(x, y) = \sum_i p(x_i, y_i) \cdot (x_i - \mu_x) \cdot (y_i - \mu_y)$

  – Measures tendency for *x* and *y* to deviate from their means in same (or opposite) directions at same time

no covariance





high (positive) covariance

- Compare to formula for covariance of actual samples

$$\text{cov}(x, y) = \frac{1}{N-1} \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)$$

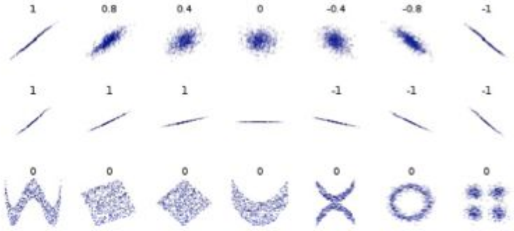
Probability

Correlation

- Pearson's correlation coefficient is covariance normalized by the standard deviations of the two variables

$$\mathrm{corr}(x,y) = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}$$

  - Always lies in range -1 to 1
  - Only reflects *linear dependence* between variables



Linear dependence with noise

Linear dependence without noise

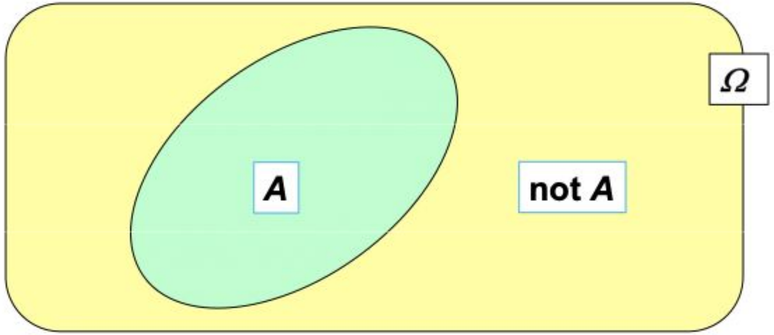
Various nonlinear dependencies

Probability

Complement rule

Given: event *A*, which can occur or not

$$p(\text{ not } A\ ) = 1 - p(\ A\ )$$



areas represent relative probabilities

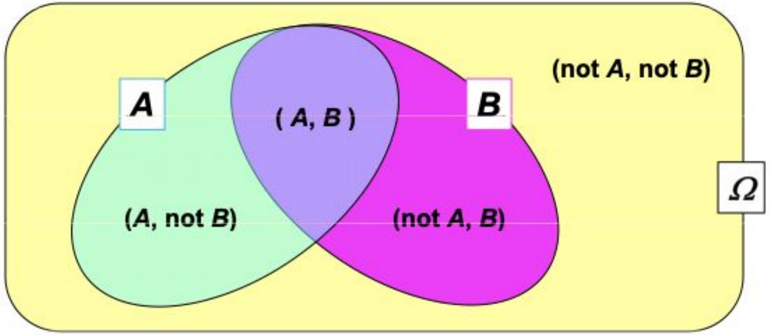
Probability

Product rule

Given: events A and B, which can co-occur (or not)

$$p(A, B) = p(A \mid B) \cdot p(B)$$

(same expression given previously to define conditional probability)



areas represent relative probabilities

Probability

Example of product rule

- Probability that a man has white hair (event A) and is over 65 (event B)
  - $p(B) = 0.18$
  - $p(A \mid B) = 0.78$
  - $p(A, B) = p(A \mid B) \cdot p(B) =$
    $0.78 \cdot 0.18 =$
    0.14

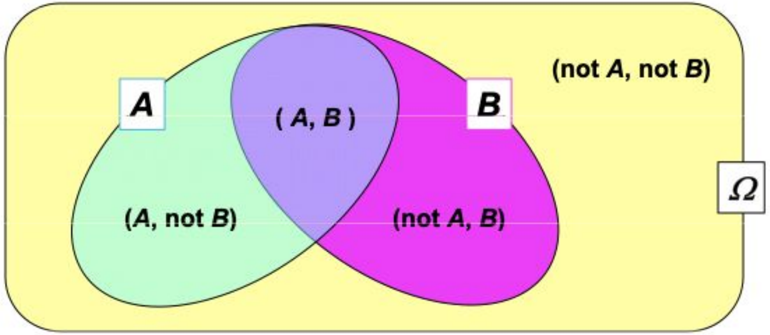
Probability

Rule of total probability

Given: events *A* and *B*, which can co-occur (or not)

$$p(A) = p(A, B) + p(A, \text{not } B)$$

(same expression given previously to define marginal probability)



areas represent relative probabilities

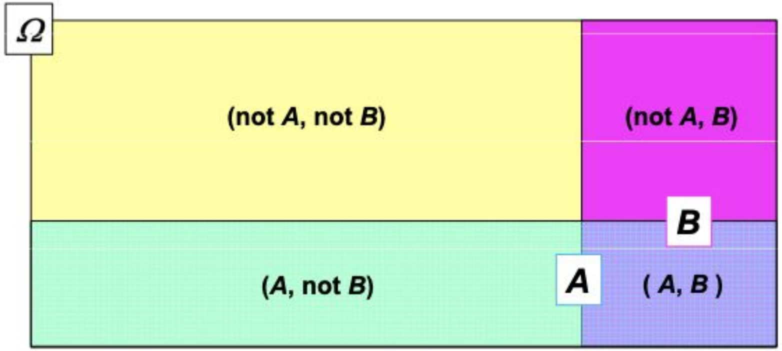
UCLA
Engineer Change.

Probability

Independence

Given: events *A* and *B*, which can co-occur (or not)

$$p(A \mid B) = p(A) \quad \text{or} \quad p(A, B) = p(A) \cdot p(B)$$



areas represent relative probabilities

Probability

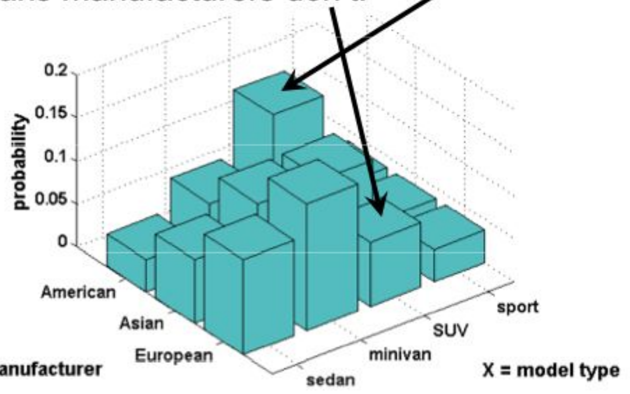
Examples of independence / dependence

- Independence:
  - Outcomes on multiple rolls of a die
  - Outcomes on multiple flips of a coin
  - Height of two unrelated individuals
  - Probability of getting a king on successive draws from a deck, if card from each draw is *replaced*
- Dependence:
  - Height of two related individuals
  - Duration of successive eruptions of Old Faithful
  - Probability of getting a king on successive draws from a deck, if card from each draw is *not replaced*

Probability

Example of independence vs. dependence

- Independence: All manufacturers have identical product mix. $p(X = x \mid Y = y) = p(X = x)$.
- Dependence: American manufacturers love SUVs, Europeans manufacturers don't.



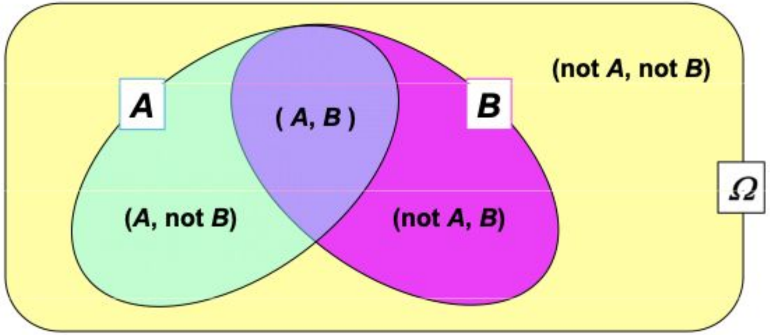
Probability

Bayes rule

A way to find conditional probabilities for one variable when conditional probabilities for another variable are known.

$$p(B \mid A) = p(A \mid B) \cdot p(B) \,/\, p(A)$$

where $p(A) = p(A, B) + p(A, \text{not } B)$

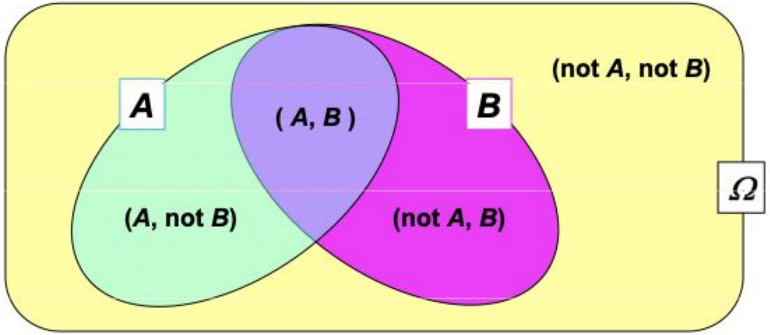


Probability

Bayes rule

posterior probability $\propto$ likelihood $\times$ prior probability

$$p(B \mid A) = p(A \mid B) \cdot p(B) \,/\, p(A)$$



Probability

Example of Bayes rule

- Marie is getting married tomorrow at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman is forecasting rain for tomorrow. When it actually rains, the weatherman has forecast rain 90% of the time. When it doesn't rain, he has forecast rain 10% of the time. What is the probability it will rain on the day of Marie's wedding?
- Event *A*: The weatherman has forecast rain.
- Event *B*: It rains.
- We know:
  - $p(B) = 5 / 365 = 0.0137$ [It rains 5 days out of the year.]
  - $p(\text{not } B) = 360 / 365 = 0.9863$
  - $p(A \mid B) = 0.9$ [When it rains, the weatherman has forecast rain 90% of the time.]
  - $p(A \mid \text{not } B) = 0.1$ [When it does not rain, the weatherman has forecast rain 10% of the time.]

Probability

Example of Bayes rule, cont'd.

- We want to know $p(B \mid A)$, the probability it will rain on the day of Marie's wedding, given a forecast for rain by the weatherman. The answer can be determined from Bayes rule:

1. $p(B \mid A) = p(A \mid B) \cdot p(B) / p(A)$
2. $p(A) = p(A \mid B) \cdot p(B) + p(A \mid \text{not } B) \cdot p(\text{not } B) =$ $(0.9)(0.014) + (0.1)(0.986) = 0.111$
3. $p(B \mid A) = (0.9)(0.0137) / 0.111 = 0.111$

- The result seems unintuitive but is correct. Even when the weatherman predicts rain, it only rains only about 11% of the time. Despite the weatherman's gloomy prediction, it is unlikely Marie will get rained on at her wedding.

Probability

Probabilities: when to add, when to multiply

- **ADD**: When you want to allow for occurrence of any of several possible outcomes of a *single* process. Comparable to logical OR.

- **MULTIPLY**: When you want to allow for simultaneous occurrence of *particular* outcomes from *more than one* process. Comparable to logical AND.
  - But only if the processes are *independent*.

Probability

Sample variance vs Variance (Why N-1?)

Proof & Explanation:https://en.wikipedia.org/wiki/Bessel%27s_correction

Vectors and Matrices

- Vector $x \in \mathbb{R}^d$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

- May also write

$$x = [x_1 \quad x_2 \quad \dots \quad x_d]^T$$

Vectors and Matrices

- Matrix $M \in \mathbb{R}^{m \times n}$

$$M = \begin{bmatrix} M_{11} & \cdots & M_{1n} \\ \vdots & \ddots & \vdots \\ M_{m1} & \cdots & M_{mn} \end{bmatrix}$$

- Written in terms of rows or columns

$$M = \begin{bmatrix} \boldsymbol{r}_1^T \\ \vdots \\ \boldsymbol{r}_m^T \end{bmatrix} = [\boldsymbol{c}_1 \quad \dots \quad \boldsymbol{c}_n]$$

$$\boldsymbol{r}_i = [M_{i1} \quad \dots \quad M_{in}]^T \quad \boldsymbol{c}_i = [M_{1i} \quad \dots \quad M_{mi}]^T$$

Multiplication

- Vector-vector: $x, y \in \mathbb{R}^d \rightarrow \mathbb{R}$

$$x^T y = \sum_{i=1}^{d} x_i y_i$$

- Matrix-vector: $x \in \mathbb{R}^n, M \in \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^m$

$$Mx = \begin{bmatrix} \boldsymbol{r}_1^T \\ \vdots \\ \boldsymbol{r}_m^T \end{bmatrix} x = \begin{bmatrix} \boldsymbol{r}_1^T x \\ \vdots \\ \boldsymbol{r}_m^T x \end{bmatrix}$$

Multiplication

- Vector-vector: $x, y \in \mathbb{R}^d \to \mathbb{R}$

$$x^T y = \sum_{i=1}^{d} x_i y_i$$

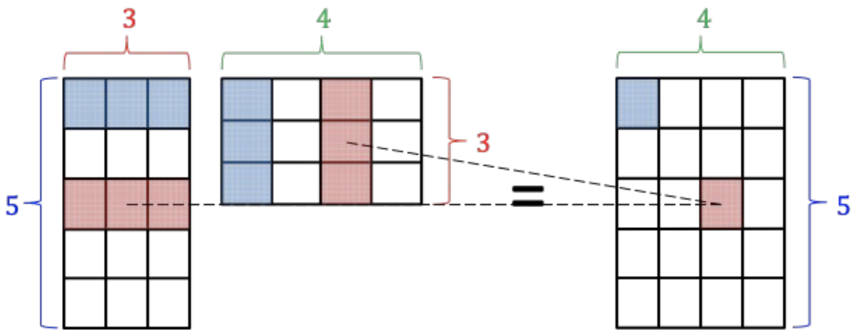
- Matrix-vector: $x \in \mathbb{R}^n, M \in \mathbb{R}^{m \times n} \to \mathbb{R}^m$

$$Mx = \begin{bmatrix} \boldsymbol{r}_1^T \\ \vdots \\ \boldsymbol{r}_m^T \end{bmatrix} x = \begin{bmatrix} \boldsymbol{r}_1^T x \\ \vdots \\ \boldsymbol{r}_m^T x \end{bmatrix}$$

Linear algebra

Multiplication

- Matrix-matrix: $A \in \mathbb{R}^{m \times k}, B \in \mathbb{R}^{k \times n} \rightarrow \mathbb{R}^{m \times n}$



Multiplication

- Matrix-matrix: $A \in \mathbb{R}^{m\times k}, B \in \mathbb{R}^{k\times n} \to \mathbb{R}^{m\times n}$
  - $\boldsymbol{a}_i$ rows of A, $\boldsymbol{b}_j$ cols of B

$$AB = [A\boldsymbol{b}_1 \quad \dots \quad A\boldsymbol{b}_n] = \begin{bmatrix} \boldsymbol{a}_1^T B \\ \vdots \\ \boldsymbol{a}_m^T B \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{a}_1^T \boldsymbol{b}_1 & \cdots & \boldsymbol{a}_1^T \boldsymbol{b}_n \\ \vdots & \boldsymbol{a}_i^T \boldsymbol{b}_j & \vdots \\ \boldsymbol{a}_m^T \boldsymbol{b}_1 & \cdots & \boldsymbol{a}_m^T \boldsymbol{b}_n \end{bmatrix}$$

Multiplication Properties

- Associative

$$(AB)C = A(BC)$$

- Distributive

$$A(B + C) = AB + BC$$

- NOT commutative

$$AB \neq BA$$

  – Dimensions may not even be conformable

Linear algebra

Useful Matrices

- Identity matrix $I \in \mathbb{R}^{m \times m}$
  - $AI = A, IA = A$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad I_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

- Diagonal matrix $A \in \mathbb{R}^{m \times m}$

$$A = \mathrm{diag}(a_1, \dots, a_m) = \begin{bmatrix} a_1 & \cdots & 0 \\ \vdots & a_i & \vdots \\ 0 & \cdots & a_m \end{bmatrix}$$

Linear algebra

Useful Matrices

- Symmetric $A \in \mathbb{R}^{m \times m}$: $A = A^T$
- Orthogonal $U \in \mathbb{R}^{m \times m}$:
$$U^T U = U U^T = I$$
  - Columns/ rows are orthonormal
- Positive semidefinite $A \in \mathbb{R}^{m \times m}$:
$$x^T A x \geq 0 \quad \text{for all } x \in \mathbb{R}^m$$
  - Equivalently, there exists $L \in \mathbb{R}^{m \times m}$
$$A = L L^T$$

Norms

- Quantify "size" of a vector
- Given $x \in \mathbb{R}^n$, a norm satisfies
  1. $\|cx\| = |c|\|x\|$
  2. $\|x\| = 0 \Leftrightarrow x = 0$
  3. $\|x + y\| \leq \|x\| + \|y\|$
- Common norms:
  1. Euclidean L_2-norm: $\|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$
  2. L_1-norm: $\|x\|_1 = |x_1| + \cdots + |x_n|$
  3. L_∞-norm: $\|x\|_\infty = \max_i |x_i|$

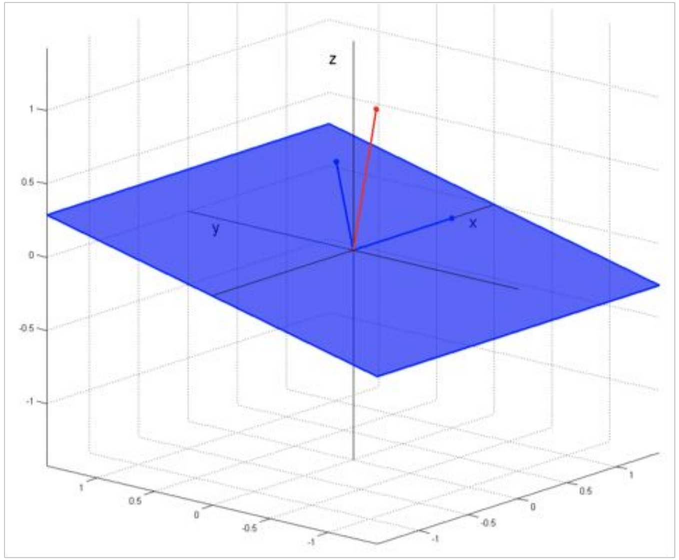
Linear Subspaces

- Subspace $\mathcal{V} \subset \mathbb{R}^n$ satisfies
  1. $0 \in \mathcal{V}$
  2. If $x, y \in \mathcal{V}$ and $c \in \mathbb{R}$, then $c(x+y) \in \mathcal{V}$
- Vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$ *span* $\mathcal{V}$ if

$$\mathcal{V} = \left\{ \sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i \,\middle|\, \alpha \in \mathbb{R}^m \right\}$$

Linear algebra

Linear Subspaces



Linear algebra

Linear Independence and Dimension

- Vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$ are *linearly independent* if

$$\sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i = 0 \Leftrightarrow \alpha = 0$$

  – Every linear combination of the $\boldsymbol{x}_i$ is unique

- $\mathrm{Dim}(\mathcal{V}) = m$ if $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$ span $\mathcal{V}$ and are linearly independent
  – If $\boldsymbol{y}_1, \dots, \boldsymbol{y}_k$ span $\mathcal{V}$ then
    - $k \geq m$
    - If $k > m$ then $\boldsymbol{y}_i$ are NOT linearly independent

Linear algebra

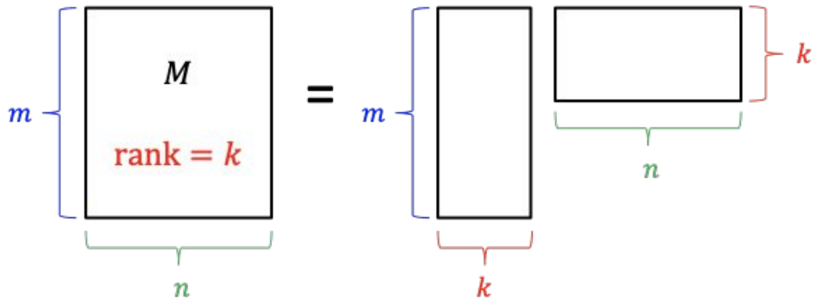
Matrix Subspaces

- Matrix $M \in \mathbb{R}^{m \times n}$ defines two subspaces
  - Column space $\mathrm{col}(M) = \{M\alpha | \alpha \in \mathbb{R}^n\} \subset \mathbb{R}^m$
  - Row space $\mathrm{row}(M) = \{M^T\beta | \beta \in \mathbb{R}^m\} \subset \mathbb{R}^n$
- Nullspace of M: $\mathrm{null}(M) = \{x \in \mathbb{R}^n | Mx = 0\}$
  - $\mathrm{null}(M) \perp \mathrm{row}(M)$
  - $\dim\big(\mathrm{null}(M)\big) + \dim\big(\mathrm{row}(M)\big) = n$
  - Analog for column space

Linear algebra

Matrix Rank

- $\text{rank}(M)$ gives dimensionality of row <u>and</u> column spaces
- If $M \in \mathbb{R}^{m \times n}$ has rank k, can decompose into product of $m \times k$ and $k \times n$ matrices



Properties of Rank

- For $A, B \in \mathbb{R}^{m \times n}$
  1. $\text{rank}(A) \leq \min(m, n)$
  2. $\text{rank}(A) = \text{rank}(A^T)$
  3. $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$
  4. $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$
- A has *full rank* if $\text{rank}(A) = \min(m, n)$
- If $m > \text{rank}(A)$ rows not linearly independent
  – Same for columns if $n > \text{rank}(A)$

Matrix Inverse

- $M \in \mathbb{R}^{m \times m}$ is invertible iff $\mathrm{rank}(M) = m$
- Inverse is unique and satisfies
  1. $M^{-1}M = MM^{-1} = I$
  2. $(M^{-1})^{-1} = M$
  3. $(M^T)^{-1} = (M^{-1})^T$
  4. If A is invertible then MA is invertible and $(MA)^{-1} = A^{-1}M^{-1}$

Systems of Equations

- Given $M \in \mathbb{R}^{m \times n}, y \in \mathbb{R}^m$ wish to solve
$$Mx = y$$
  - Exists only if $y \in \text{col}(M)$
    - Possibly infinite number of solutions
- If M is invertible then $x = M^{-1}y$
  - Notational device, do not actually invert matrices
  - Computationally, use solving routines like Gaussian elimination

Linear algebra

Systems of Equations

- What if $y \notin \mathrm{col}(M)$?
- Find x that gives $\hat{y} = Mx$ *closest to* y
  - $\hat{y}$ is projection of y onto $\mathrm{col}(M)$
  - Also known as regression
- Assume $\mathrm{rank}(M) = n < m$

$$x = \underbrace{(M^T M)}_{\text{Invertible}}{}^{-1} M^T y \qquad \hat{y} = \underbrace{M(M^T M)^{-1} M^T}_{\text{Projection matrix}} y$$

Characterizations of Eigenvalues

- Traditional formulation

$$Mx = \lambda x$$

  - Leads to characteristic polynomial

$$\det(M - \lambda I) = 0$$

Linear algebra

Eigenvalue Properties

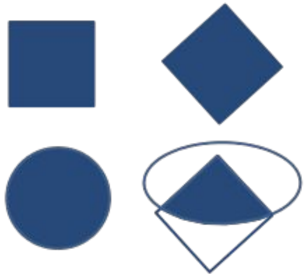
- For $M \in \mathbb{R}^{m \times m}$ with eigenvalues λ_i
  1. $\mathrm{tr}(M) = \sum_{i=1}^{m} \lambda_i$
  2. $\det(M) = \lambda_1 \lambda_2 \dots \lambda_m$
  3. $\mathrm{rank}(M) = \#\lambda_i \neq 0$

Convex Sets

- A set C is convex if $\forall x, y \in C$ and $\forall \alpha \in [0,1]$

$$\alpha x + (1-\alpha) y \in C$$

  - Line segment between points in C also lies in C
- Ex
  - Intersection of halfspaces
  - L_p balls
  - Intersection of convex sets



Convex Functions

- A real-valued function f is convex if $\mathrm{dom} f$ *is* convex and $\forall x, y \in \mathrm{dom} f$ and $\forall \alpha \in [0,1]$

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

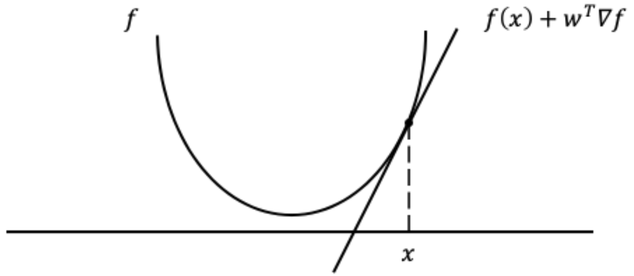
  - Graph of f upper bounded by line segment between points on graph

Optimization

Gradients

- Differentiable convex f with $\mathrm{dom} f = \mathbb{R}^d$
- Gradient ∇f at x gives linear approximation

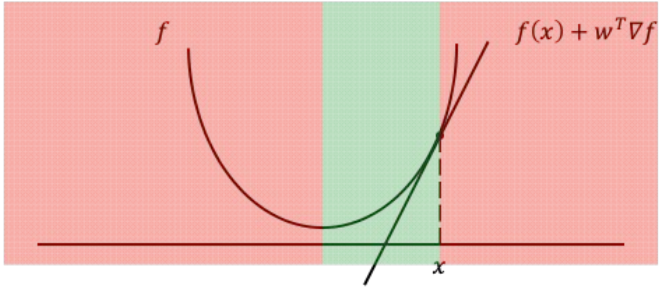
$$\nabla f = \left[\frac{\delta f}{\delta x_1} \quad \dots \quad \frac{\delta f}{\delta x_d}\right]^T$$



Gradients

- Differentiable convex f with $\mathrm{dom} f = \mathbb{R}^d$
- Gradient ∇f at x gives linear approximation

$$\nabla f = \left[\frac{\delta f}{\delta x_1} \quad \dots \quad \frac{\delta f}{\delta x_d}\right]^T$$



Optimization

Gradient Descent

- To minimize f move down gradient
  - But not too far!
  - Optimum when $\nabla f = 0$
- Given f, learning rate α, starting point x_0

  $x = x_0$

  Do until $\nabla f = 0$

  $$x = x - \alpha \nabla f$$

Optimization

Stochastic Gradient Descent

- Many learning problems have extra structure

$$f(\theta) = \sum_{i=1}^{n} L(\theta; \boldsymbol{x}_i)$$

- Computing gradient requires iterating over all points, can be too costly
- Instead, compute gradient at single training example

Optimization

Stochastic Gradient Descent

- Given $f(\theta) = \sum_{i=1}^{n} L(\theta; \boldsymbol{x}_i)$, learning rate α, starting point θ_0

  $\theta = \theta_0$

  Do until $f(\theta)$ nearly optimal

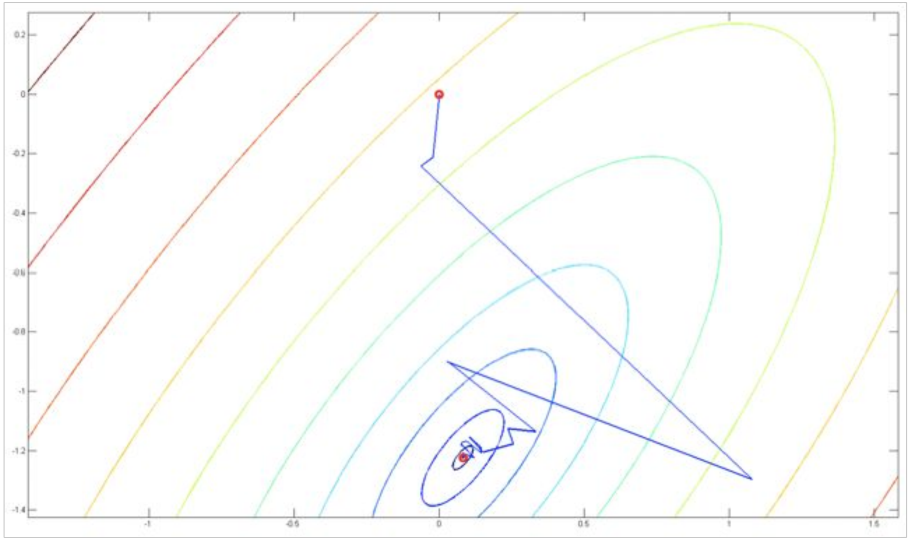
  For $i = 1$ to n in random order

  $$\theta = \theta - \alpha \nabla L(\theta; \boldsymbol{x}_i)$$

- Finds nearly optimal θ

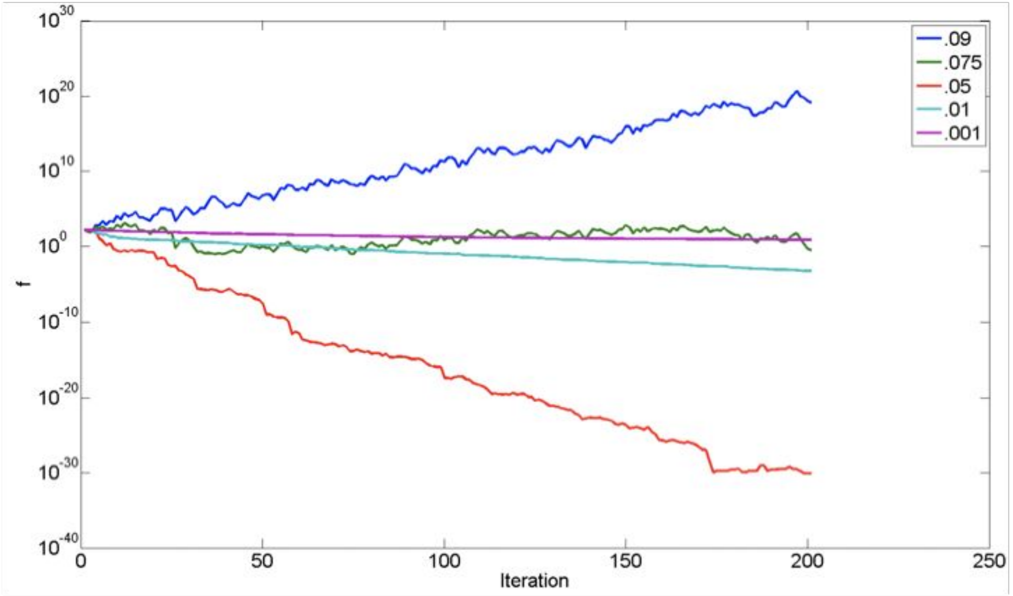
Optimization

Minimize $\sum_{i=1}^{n}(y_i - \theta^T \boldsymbol{x}_i)^2$



Optimization

Learning Parameter



Matrix Calculus

The Gradient

Suppose that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a function that takes as input a matrix A of size $m \times n$ and returns a real value. Then the ***gradient*** of f (with respect to $A \in \mathbb{R}^{m\times n}$) is the matrix of partial derivatives, defined as:

$$\nabla_A f(A) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

i.e., an $m \times n$ matrix with

$$(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}.$$

The Gradient

Note that the size of $\nabla_A f(A)$ is always the same as the size of A. So if, in particular, A is just a vector $x \in \mathbb{R}^n$,

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

It follows directly from the equivalent properties of partial derivatives that:

- $\nabla_x(f(x) + g(x)) = \nabla_x f(x) + \nabla_x g(x)$.
- For $t \in \mathbb{R}$, $\nabla_x(t\, f(x)) = t\nabla_x f(x)$.

Matrix Calculus

The Hessian

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in $\mathbb{R}^n$ and returns a real number. Then the ***Hessian*** matrix with respect to x, written $\nabla_x^2 f(x)$ or simply as H is the $n \times n$ matrix of partial derivatives,

$$\nabla_x^2 f(x) \in \mathbb{R}^{n\times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}.$$

In other words, $\nabla_x^2 f(x) \in \mathbb{R}^{n\times n}$, with

$$(\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}.$$

Note that the Hessian is always symmetric, since

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}.$$

Gradients of Linear Functions

For $x \in \mathbb{R}^n$, let $f(x) = b^T x$ for some known vector $b \in \mathbb{R}^n$. Then

$$f(x) = \sum_{i=1}^{n} b_i x_i$$

so

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_k.$$

From this we can easily see that $\nabla_x b^T x = b$. This should be compared to the analogous situation in single variable calculus, where $\partial/(\partial x)\ ax = a$.

Gradients of Quadratic Function

Now consider the quadratic function $f(x) = x^T Ax$ for $A \in \mathbb{S}^n$. Remember that

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j.$$

To take the partial derivative, we'll consider the terms including x_k and x_k^2 factors separately:

$$\begin{aligned}
\frac{\partial f(x)}{\partial x_k} &= \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j \\
&= \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right] \\
&= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2 A_{kk} x_k \\
&= \sum_{i=1}^{n} A_{ik} x_i + \sum_{j=1}^{n} A_{kj} x_j = 2 \sum_{i=1}^{n} A_{ki} x_i,
\end{aligned}$$

Hessian of Quadratic Functions

Finally, let's look at the Hessian of the quadratic function $f(x) = x^T Ax$
In this case,

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_\ell} = \frac{\partial}{\partial x_k}\left[\frac{\partial f(x)}{\partial x_\ell}\right] = \frac{\partial}{\partial x_k}\left[2\sum_{i=1}^{n} A_{\ell i} x_i\right] = 2A_{\ell k} = 2A_{k\ell}.$$

Therefore, it should be clear that $\nabla_x^2 x^T Ax = 2A$, which should be entirely expected (and again analogous to the single-variable fact that $\partial^2/(\partial x^2)\ ax^2 = 2a$).

Matrix Calculus Example: Least Squares

- Given a full rank matrices $A \in \mathbb{R}^{m \times n}$, and a vector $b \in \mathbb{R}^m$ such that $b \notin \mathcal{R}(A)$, we want to find a vector x such that Ax is as close as possible to b, as measured by the square of the Euclidean norm $\|Ax - b\|_2^2$.

- Using the fact that $\|x\|_2^2 = x^T x$, we have

$$\|Ax - b\|_2^2 = (Ax - b)^T (Ax - b) = x^T A^T Ax - 2b^T Ax + b^T b$$

- Taking the gradient with respect to x we have:

$$\begin{aligned} \nabla_x (x^T A^T Ax - 2b^T Ax + b^T b) &= \nabla_x x^T A^T Ax - \nabla_x 2b^T Ax + \nabla_x b^T b \\ &= 2A^T Ax - 2A^T b \end{aligned}$$

- Setting this last expression equal to zero and solving for x gives the normal equations

$$x = (A^T A)^{-1} A^T b$$

Thank you!

Q & A