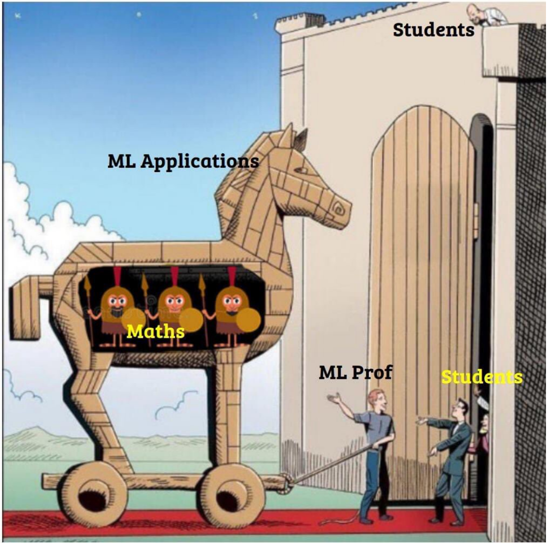
Section 1: Math Reviews/Previews

Ruofan Ma

Gov2018 2024 Spring

January 31, 2024

Now that you're here...



Road Map for Today

- ~~Introduction~~
- MATH!
  - Linear Algebra: vectors, matrices, and projections
  - Calculus: derivatives, multivariate calculus, and optimizations
  - Statistics: probability, inference, and computation

Linear Algebra: Basic Ideas

- Let $\mathbf{A} = (a_{ij})_{p \times p}$ denote a $p \times p$ matrix with its (i, j) th entry being a_{ij}, and let $\mathbf{x} = (x_1, \ldots, x_p)^\top$ be a p-dim (column) vector.

Linear Algebra: Basic Ideas

- Let $\mathbf{A} = (a_{ij})_{p \times p}$ denote a $p \times p$ matrix with its (i, j) th entry being a_{ij}, and let $\mathbf{x} = (x_1, \ldots, x_p)^\top$ be a p-dim (column) vector.
  - A vector can be viewed as a function of indexes (input index i, output $\mathbf{x}[i] = x_i$).

Linear Algebra: Basic Ideas

- Let $\mathbf{A} = (a_{ij})_{p \times p}$ denote a $p \times p$ matrix with its (i, j) th entry being a_{ij}, and let $\mathbf{x} = (x_1, \ldots, x_p)^\top$ be a p-dim (column) vector.
  - A vector can be viewed as a function of indexes (input index i, output $\mathbf{x}[i] = x_i$).
  - Solving a system of linear equations: $\mathbf{Ax} = \mathbf{b}$, which has the solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ when $\mathbf{A}$ is invertible.

Linear Algebra: Basic Ideas

- Let $\mathbf{A} = (a_{ij})_{p\times p}$ denote a $p \times p$ matrix with its (i, j) th entry being a_{ij}, and let $\mathbf{x} = (x_1, \ldots, x_p)^\top$ be a p-dim (column) vector.
  - A vector can be viewed as a function of indexes (input index i, output $\mathbf{x}[i] = x_i$).
  - Solving a system of linear equations: $\mathbf{Ax} = \mathbf{b}$, which has the solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ when $\mathbf{A}$ is invertible.
  - Linear dependence and inverse of a matrix: for a matrix to have its inverse, it has to be a square matrix, and its columns are linearly independent.

Linear Algebra: Basic Ideas

- Let $\mathbf{A} = (a_{ij})_{p \times p}$ denote a $p \times p$ matrix with its (i, j) th entry being a_{ij}, and let $\mathbf{x} = (x_1, \ldots, x_p)^\top$ be a p-dim (column) vector.
  - A vector can be viewed as a function of indexes (input index i, output $\mathbf{x}[i] = x_i$).
  - Solving a system of linear equations: $\mathbf{Ax} = \mathbf{b}$, which has the solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ when $\mathbf{A}$ is invertible.
  - Linear dependence and inverse of a matrix: for a matrix to have its inverse, it has to be a square matrix, and its columns are linearly independent.
  - Linear transformation of a vector: $\mathbf{x}' = \mathbf{Ax}$; entry $\mathbf{x}'[i] = \sum_{j=1}^{p} a_{ij}x_j$. It can also be written as $\mathbf{x}' = \sum_{j=1}^{p} x_j \mathbf{A}_{\cdot j}$, where $\mathbf{A}_{\cdot j}$, denote the j th column of $\mathbf{A}$.

Linear Algebra: Vectors

- Things that should sound familiar to you: vector space, subspace, span, basis, dimension . . .

Linear Algebra: Vectors

- Things that should sound familiar to you: vector space, subspace, span, basis, dimension ...
- L_p-norm: $\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$

Linear Algebra: Vectors

- Things that should sound familiar to you: vector space, subspace, span, basis, dimension ...
- L_p-norm: $\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$
  - L_2 and L_1 norms are most frequently seen.

Linear Algebra: Vectors

- Things that should sound familiar to you: vector space, subspace, span, basis, dimension ...
- L_p-norm: $\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$
  - L_2 and L_1 norms are most frequently seen.
  - L_∞-norm is defined as $\|\mathbf{x}\|_\infty = \max_i |x_i|$; and L_0 norm is defined as $\|\mathbf{x}\|_0 = \sum_{i=1}^{p} I_{\{x_i \neq 0\}}$, i.e., the number of non-zero entries in $\mathbf{x}$.

Linear Algebra: Vectors

- Things that should sound familiar to you: vector space, subspace, span, basis, dimension ...
- L_p-norm: $\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$
  - L_2 and L_1 norms are most frequently seen.
  - L_∞-norm is defined as $\|\mathbf{x}\|_\infty = \max_i |x_i|$; and L_0 norm is defined as $\|\mathbf{x}\|_0 = \sum_{i=1}^{p} I_{\{x_i \neq 0\}}$, i.e., the number of non-zero entries in $\mathbf{x}$.
  - Frobenius norm of a matrix: $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$, analogous to the L_2 norm.

Linear Algebra: Vectors

- Things that should sound familiar to you: vector space, subspace, span, basis, dimension ...
- L_p-norm: $\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$
  - L_2 and L_1 norms are most frequently seen.
  - L_∞-norm is defined as $\|\mathbf{x}\|_\infty = \max_i |x_i|$; and L_0 norm is defined as $\|\mathbf{x}\|_0 = \sum_{i=1}^{p} I_{\{x_i \neq 0\}}$, i.e., the number of non-zero entries in $\mathbf{x}$.
  - Frobenius norm of a matrix: $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$, analogous to the L_2 norm.
- Inner product (aka "dot product") of two vectors:

Linear Algebra: Vectors

- Things that should sound familiar to you: vector space, subspace, span, basis, dimension ...
- L_p-norm: $\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$
  - L_2 and L_1 norms are most frequently seen.
  - L_∞-norm is defined as $\|\mathbf{x}\|_\infty = \max_i |x_i|$; and L_0 norm is defined as $\|\mathbf{x}\|_0 = \sum_{i=1}^p I_{\{x_i \neq 0\}}$, i.e., the number of non-zero entries in $\mathbf{x}$.
  - Frobenius norm of a matrix: $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$, analogous to the L_2 norm.
- Inner product (aka "dot product") of two vectors:
  - $$\mathbf{x} \cdot \mathbf{y} \equiv \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x} = \sum_{i=1}^p x_i y_i = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos(\theta)$$

Linear Algebra: Vectors

- Things that should sound familiar to you: vector space, subspace, span, basis, dimension ...
- L_p-norm: $\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$
  - L_2 and L_1 norms are most frequently seen.
  - L_∞-norm is defined as $\|\mathbf{x}\|_\infty = \max_i |x_i|$; and L_0 norm is defined as $\|\mathbf{x}\|_0 = \sum_{i=1}^{p} I_{\{x_i \neq 0\}}$, i.e., the number of non-zero entries in $\mathbf{x}$.
  - Frobenius norm of a matrix: $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$, analogous to the L_2 norm.
- Inner product (aka "dot product") of two vectors:
  - $$\mathbf{x} \cdot \mathbf{y} \equiv \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x} = \sum_{i=1}^{p} x_i y_i = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos(\theta)$$
  - Thus, $\mathbf{x} \cdot \mathbf{x} = \mathbf{x}^\top \mathbf{x} = \|\mathbf{x}\|_2^2$

Linear Algebra: Matrix Multiplication

- Matrix multiplication: $\mathbf{AB}$ is a valid matrix product if $\mathbf{A}$ is $p \times q$ and $\mathbf{B}$ is $q \times r$. The standard matrix product is defined as follows:

Linear Algebra: Matrix Multiplication

- Matrix multiplication: $\mathbf{AB}$ is a valid matrix product if $\mathbf{A}$ is $p \times q$ and $\mathbf{B}$ is $q \times r$. The standard matrix product is defined as follows:

$$(\mathbf{AB})_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{iq}b_{qj} = \sum_{k=1}^{q} a_{ik}b_{kj}$$

Linear Algebra: Matrix Multiplication

- Matrix multiplication: **AB** is a valid matrix product if **A** is $p \times q$ and **B** is $q \times r$. The standard matrix product is defined as follows:

$$(\mathbf{AB})_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{iq}b_{qj} = \sum_{k=1}^{q} a_{ik}b_{kj}$$

, where $i = 1, \ldots, p$ and $j = 1, \ldots, r$. In other words, $(\mathbf{AB})_{ij}$ is the dot product of the i th row of **A** with the j th column of **B**.

Linear Algebra: Matrix Multiplication

- Matrix multiplication: $\mathbf{AB}$ is a valid matrix product if $\mathbf{A}$ is $p \times q$ and $\mathbf{B}$ is $q \times r$. The standard matrix product is defined as follows:

$$(\mathbf{AB})_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{iq}b_{qj} = \sum_{k=1}^{q} a_{ik}b_{kj}$$

  , where $i = 1, \ldots, p$ and $j = 1, \ldots, r$. In other words, $(\mathbf{AB})_{ij}$ is the dot product of the i th row of $\mathbf{A}$ with the j th column of $\mathbf{B}$.
- Properties of matrix multiplication:

Linear Algebra: Matrix Multiplication

- Matrix multiplication: $\mathbf{AB}$ is a valid matrix product if $\mathbf{A}$ is $p \times q$ and $\mathbf{B}$ is $q \times r$. The standard matrix product is defined as follows:

$$(\mathbf{AB})_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{iq}b_{qj} = \sum_{k=1}^{q} a_{ik}b_{kj}$$

, where $i = 1, \ldots, p$ and $j = 1, \ldots, r$. In other words, $(\mathbf{AB})_{ij}$ is the dot product of the i th row of $\mathbf{A}$ with the j th column of $\mathbf{B}$.

- Properties of matrix multiplication:
  - Generally not commutative: $\mathbf{AB} \neq \mathbf{BA}$

Linear Algebra: Matrix Multiplication

- Matrix multiplication: $\mathbf{AB}$ is a valid matrix product if $\mathbf{A}$ is $p \times q$ and $\mathbf{B}$ is $q \times r$. The standard matrix product is defined as follows:

$$(\mathbf{AB})_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{iq}b_{qj} = \sum_{k=1}^{q} a_{ik}b_{kj}$$

  , where $i = 1, \ldots, p$ and $j = 1, \ldots, r$. In other words, $(\mathbf{AB})_{ij}$ is the dot product of the i th row of $\mathbf{A}$ with the j th column of $\mathbf{B}$.

- Properties of matrix multiplication:
  - Generally not commutative: $\mathbf{AB} \neq \mathbf{BA}$
  - Distributive over addition: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$. $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$.

Linear Algebra: Matrix Multiplication

- Matrix multiplication: $\mathbf{AB}$ is a valid matrix product if $\mathbf{A}$ is $p \times q$ and $\mathbf{B}$ is $q \times r$. The standard matrix product is defined as follows:

$$(\mathbf{AB})_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{iq}b_{qj} = \sum_{k=1}^{q} a_{ik}b_{kj}$$

  , where $i = 1, \ldots, p$ and $j = 1, \ldots, r$. In other words, $(\mathbf{AB})_{ij}$ is the dot product of the i th row of $\mathbf{A}$ with the j th column of $\mathbf{B}$.
- Properties of matrix multiplication:
  - Generally not commutative: $\mathbf{AB} \neq \mathbf{BA}$
  - Distributive over addition: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$. $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$.
  - Scalable: $\lambda(\mathbf{AB}) = (\lambda\mathbf{A})\mathbf{B} = (\mathbf{AB})\lambda = \mathbf{A}(\mathbf{B}\lambda)$

Linear Algebra: Matrix Multiplication

- Matrix multiplication: $\mathbf{AB}$ is a valid matrix product if $\mathbf{A}$ is $p \times q$ and $\mathbf{B}$ is $q \times r$. The standard matrix product is defined as follows:

$$(\mathbf{AB})_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{iq}b_{qj} = \sum_{k=1}^{q} a_{ik}b_{kj}$$

, where $i = 1, \ldots, p$ and $j = 1, \ldots, r$. In other words, $(\mathbf{AB})_{ij}$ is the dot product of the i th row of $\mathbf{A}$ with the j th column of $\mathbf{B}$.

- Properties of matrix multiplication:
  - Generally not commutative: $\mathbf{AB} \neq \mathbf{BA}$
  - Distributive over addition: $\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB} + \mathbf{AC}$. $(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$.
  - Scalable: $\lambda(\mathbf{AB}) = (\lambda\mathbf{A})\mathbf{B} = (\mathbf{AB})\lambda = \mathbf{A}(\mathbf{B}\lambda)$
  - Transpose of product: $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$

Linear Algebra: Matrices

- Things that should sound familiar to you: rank, trace, determinant, inverse . . .

Linear Algebra: Matrices

- Things that should sound familiar to you: rank, trace, determinant, inverse . . .
- Matrix Properties

Linear Algebra: Matrices

- Things that should sound familiar to you: rank, trace, determinant, inverse ...
- Matrix Properties
  - $\mathbf{A}^\top$ is the transpose of $\mathbf{A}$ and has $A_{ji}^\top = A_{ij}$. This is just like flipping the two dimensions of your matrix.

Linear Algebra: Matrices

- Things that should sound familiar to you: rank, trace, determinant, inverse . . .
- Matrix Properties
  - $\mathbf{A}^\top$ is the transpose of $\mathbf{A}$ and has $A^\top_{ji} = A_{ij}$. This is just like flipping the two dimensions of your matrix.
  - $\mathbf{A}$ is symmetric if $A_{ij} = A_{ji}$. That is, $\mathbf{A} = \mathbf{A}^\top$. Only square matrices can be symmetric.

Linear Algebra: Matrices

- Things that should sound familiar to you: rank, trace, determinant, inverse ...
- Matrix Properties
  - $\mathbf{A}^\top$ is the transpose of $\mathbf{A}$ and has $A_{ji}^\top = A_{ij}$. This is just like flipping the two dimensions of your matrix.
  - $\mathbf{A}$ is symmetric if $A_{ij} = A_{ji}$. That is, $\mathbf{A} = \mathbf{A}^\top$. Only square matrices can be symmetric.
  - $\mathbf{A}$ is orthogonal if its rows and its columns are orthogonal unit vectors: $\mathbf{A}^\top \mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}$. For an orthogonal matrix $\mathbf{A}$ we have $\mathbf{A}^\top = \mathbf{A}^{-1}$.

Linear Algebra: Matrices

- Things that should sound familiar to you: rank, trace, determinant, inverse ...
- Matrix Properties
  - $\mathbf{A}^\top$ is the transpose of $\mathbf{A}$ and has $A^\top_{ji} = A_{ij}$. This is just like flipping the two dimensions of your matrix.
  - $\mathbf{A}$ is symmetric if $A_{ij} = A_{ji}$. That is, $\mathbf{A} = \mathbf{A}^\top$. Only square matrices can be symmetric.
  - $\mathbf{A}$ is orthogonal if its rows and its columns are orthogonal unit vectors: $\mathbf{A}^\top\mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}$. For an orthogonal matrix $\mathbf{A}$ we have $\mathbf{A}^\top = \mathbf{A}^{-1}$.
  - Diagonal matrices have non-zero values on the main diagonal and zeros elsewhere. Diagonal matrices are easy to take powers of because you just take the powers of the diagonal entries.

Linear Algebra: Matrices

- Things that should sound familiar to you: rank, trace, determinant, inverse ...
- Matrix Properties
  - $\mathbf{A}^\top$ is the transpose of $\mathbf{A}$ and has $A^\top_{ji} = A_{ij}$. This is just like flipping the two dimensions of your matrix.
  - $\mathbf{A}$ is symmetric if $A_{ij} = A_{ji}$. That is, $\mathbf{A} = \mathbf{A}^\top$. Only square matrices can be symmetric.
  - $\mathbf{A}$ is orthogonal if its rows and its columns are orthogonal unit vectors: $\mathbf{A}^\top\mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}$. For an orthogonal matrix $\mathbf{A}$ we have $\mathbf{A}^\top = \mathbf{A}^{-1}$.
  - Diagonal matrices have non-zero values on the main diagonal and zeros elsewhere. Diagonal matrices are easy to take powers of because you just take the powers of the diagonal entries.
  - Eigen-everything: $\mathbf{Av} = \lambda\mathbf{v} \Rightarrow$ eigenvalue decomposition, single value decomposition, etc.

Linear Algebra: Projection

- Project vector $\mathbf{x}$ to the direction of vector $\mathbf{v} \neq 0$ (define $\text{proj}_0(\mathbf{x}) \equiv 0$):

Linear Algebra: Projection

- Project vector $\mathbf{x}$ to the direction of vector $\mathbf{v} \neq 0$ (define $\mathrm{proj}_0(\mathbf{x}) \equiv 0$):

$$\mathrm{proj}_{\mathbf{v}}(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \mathbf{v} \equiv \left[\mathbf{v} \left(\mathbf{v}^\top \mathbf{v} \right)^{-1} \mathbf{v}^\top \right] \mathbf{x}$$

Linear Algebra: Projection

- Project vector $\mathbf{x}$ to the direction of vector $\mathbf{v} \neq 0$ (define $\text{proj}_0(\mathbf{x}) \equiv 0$):

$$\text{proj}_{\mathbf{v}}(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle}\mathbf{v} \equiv \left[\mathbf{v}\left(\mathbf{v}^\top \mathbf{v}\right)^{-1}\mathbf{v}^\top\right]\mathbf{x}$$

  - Thus, $\mathbf{x} - \text{proj}_{\mathbf{v}}(\mathbf{x}) = \left[\mathbf{I} - \mathbf{v}\left(\mathbf{v}^\top \mathbf{v}\right)^{-1}\mathbf{v}^\top\right]\mathbf{x} \stackrel{\text{def}}{=} \text{orth}_{\mathbf{v}}(\mathbf{x})$ is orthogonal to $\mathbf{v}$.

Linear Algebra: Projection

- Project vector $\mathbf{x}$ to the direction of vector $\mathbf{v} \neq 0$ (define $\text{proj}_0(\mathbf{x}) \equiv 0$):

$$\text{proj}_{\mathbf{v}}(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \mathbf{v} \equiv \left[\mathbf{v} \left(\mathbf{v}^\top \mathbf{v} \right)^{-1} \mathbf{v}^\top \right] \mathbf{x}$$

  - Thus, $\mathbf{x} - \text{proj}_{\mathbf{v}}(\mathbf{x}) = \left[\mathbf{I} - \mathbf{v} \left(\mathbf{v}^\top \mathbf{v} \right)^{-1} \mathbf{v}^\top \right] \mathbf{x} \stackrel{\text{def}}{=} \text{orth}_{\mathbf{v}}(\mathbf{x})$ is orthogonal to $\mathbf{v}$.
  - By Pythagorean theorem: $\|\mathbf{x}\|_2^2 = \|\text{proj}_{\mathbf{v}}(\mathbf{x})\|_2^2 + \|\text{orth}_{\mathbf{v}}(\mathbf{x})\|_2^2$

Linear Algebra: Projection

- Project vector $\mathbf{x}$ to the direction of vector $\mathbf{v} \neq 0$ (define $\text{proj}_0(\mathbf{x}) \equiv 0$):

$$\text{proj}_{\mathbf{v}}(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \mathbf{v} \equiv \left[\mathbf{v} \left(\mathbf{v}^\top \mathbf{v} \right)^{-1} \mathbf{v}^\top \right] \mathbf{x}$$

  - Thus, $\mathbf{x} - \text{proj}_{\mathbf{v}}(\mathbf{x}) = \left[\mathbf{I} - \mathbf{v} \left(\mathbf{v}^\top \mathbf{v} \right)^{-1} \mathbf{v}^\top \right] \mathbf{x} \overset{\text{def}}{=} \text{orth}_{\mathbf{v}}(\mathbf{x})$ is orthogonal to $\mathbf{v}$.
  - By Pythagorean theorem: $\|\mathbf{x}\|_2^2 = \|\text{proj}_{\mathbf{v}}(\mathbf{x})\|_2^2 + \|\text{orth}_{\mathbf{v}}(\mathbf{x})\|_2^2$
- Projection to a subspace:
  - $\text{proj}_{\mathbf{V}}(\mathbf{x}) = \left[\mathbf{V} \left(\mathbf{V}^\top \mathbf{V} \right)^{-1} \mathbf{V}^\top \right] \mathbf{x}$, $\text{orth}_V(\mathbf{x}) = \left[\mathbf{I} - \mathbf{V} \left(\mathbf{V}^\top \mathbf{V} \right)^{-1} \mathbf{V}^\top \right] \mathbf{x}$

Linear Algebra: Projection

- Project vector $\mathbf{x}$ to the direction of vector $\mathbf{v} \neq 0$ (define $\text{proj}_0(\mathbf{x}) \equiv 0$):

$$\text{proj}_{\mathbf{v}}(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle}\mathbf{v} \equiv \left[\mathbf{v}\left(\mathbf{v}^\top \mathbf{v}\right)^{-1}\mathbf{v}^\top\right]\mathbf{x}$$

  - Thus, $\mathbf{x} - \text{proj}_{\mathbf{v}}(\mathbf{x}) = \left[\mathbf{I} - \mathbf{v}\left(\mathbf{v}^\top \mathbf{v}\right)^{-1}\mathbf{v}^\top\right]\mathbf{x} \stackrel{\text{def}}{=} \text{orth}_{\mathbf{v}}(\mathbf{x})$ is orthogonal to $\mathbf{v}$.
  - By Pythagorean theorem: $\|\mathbf{x}\|_2^2 = \|\text{proj}_{\mathbf{v}}(\mathbf{x})\|_2^2 + \|\text{orth}_{\mathbf{v}}(\mathbf{x})\|_2^2$
- Projection to a subspace:
  - $\text{proj}_{\mathbf{V}}(\mathbf{x}) = \left[\mathbf{V}\left(\mathbf{V}^\top \mathbf{V}\right)^{-1}\mathbf{V}^\top\right]\mathbf{x}$, $\text{orth}_V(\mathbf{x}) = \left[\mathbf{I} - \mathbf{V}\left(\mathbf{V}^\top \mathbf{V}\right)^{-1}\mathbf{V}^\top\right]\mathbf{x}$
  - $\mathbf{P} = \mathbf{V}\left(\mathbf{V}^\top \mathbf{V}\right)^{-1}\mathbf{V}^\top$ (hat/projection matrix);

Linear Algebra: Projection

- Project vector $\mathbf{x}$ to the direction of vector $\mathbf{v} \neq 0$ (define $\text{proj}_0(\mathbf{x}) \equiv 0$):

$$\text{proj}_{\mathbf{v}}(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle}\mathbf{v} \equiv \left[\mathbf{v}\left(\mathbf{v}^\top \mathbf{v}\right)^{-1}\mathbf{v}^\top\right]\mathbf{x}$$

  - Thus, $\mathbf{x} - \text{proj}_{\mathbf{v}}(\mathbf{x}) = \left[\mathbf{I} - \mathbf{v}\left(\mathbf{v}^\top \mathbf{v}\right)^{-1}\mathbf{v}^\top\right]\mathbf{x} \stackrel{\text{def}}{=} \text{orth}_{\mathbf{v}}(\mathbf{x})$ is orthogonal to $\mathbf{v}$.
  - By Pythagorean theorem: $\|\mathbf{x}\|_2^2 = \|\text{proj}_{\mathbf{v}}(\mathbf{x})\|_2^2 + \|\text{orth}_{\mathbf{v}}(\mathbf{x})\|_2^2$
- Projection to a subspace:
  - $\text{proj}_{\mathbf{V}}(\mathbf{x}) = \left[\mathbf{V}\left(\mathbf{V}^\top \mathbf{V}\right)^{-1}\mathbf{V}^\top\right]\mathbf{x}$, $\text{orth}_{V}(\mathbf{x}) = \left[\mathbf{I} - \mathbf{V}\left(\mathbf{V}^\top \mathbf{V}\right)^{-1}\mathbf{V}^\top\right]\mathbf{x}$
  - $\mathbf{P} = \mathbf{V}\left(\mathbf{V}^\top \mathbf{V}\right)^{-1}\mathbf{V}^\top$ (hat/projection matrix); $\mathbf{I} - \mathbf{P}$ (orthogonalization/annihilation matrix)

Calculus: Differentiation

- Things that should sound familiar to you: product rule, quotient rule, chain rule, increasing/decreasing, concave/convex . . .

Calculus: Differentiation

- Things that should sound familiar to you: product rule, quotient rule, chain rule, increasing/decreasing, concave/convex ...
- Derivative as "rate-of-change":

$$f'(x) \equiv \frac{df(x)}{dx} \equiv \frac{\partial f(x)}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Calculus: Differentiation

- Things that should sound familiar to you: product rule, quotient rule, chain rule, increasing/decreasing, concave/convex ...
- Derivative as "rate-of-change":

$$f'(x) \equiv \frac{df(x)}{dx} \equiv \frac{\partial f(x)}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

- If $\mathbf{x} = (x_1, \ldots, x_p)^\top$ is multi-dimensional, we have the gradient:

Calculus: Differentiation

- Things that should sound familiar to you: product rule, quotient rule, chain rule, increasing/decreasing, concave/convex ...
- Derivative as "rate-of-change":

$$f'(x) \equiv \frac{df(x)}{dx} \equiv \frac{\partial f(x)}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

- If $\mathbf{x} = (x_1, \ldots, x_p)^\top$ is multi-dimensional, we have the gradient:

$$\nabla f(\mathbf{x}) = \frac{df(\mathbf{x})}{d\mathbf{x}} = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_p} \right)^\top$$

Calculus: Differentiation

- Things that should sound familiar to you: product rule, quotient rule, chain rule, increasing/decreasing, concave/convex ...
- Derivative as "rate-of-change":

$$f'(x) \equiv \frac{df(x)}{dx} \equiv \frac{\partial f(x)}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

- If $\mathbf{x} = (x_1, \ldots, x_p)^\top$ is multi-dimensional, we have the gradient:

$$\nabla f(\mathbf{x}) = \frac{df(\mathbf{x})}{d\mathbf{x}} = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_p} \right)^\top$$

  - The gradient vector points in the direction of steepest ascent in $f(\mathbf{x})$. This is useful for optimization.

Calculus: Differentiation

- Things that should sound familiar to you: product rule, quotient rule, chain rule, increasing/decreasing, concave/convex ...
- Derivative as "rate-of-change":

$$f'(x) \equiv \frac{df(x)}{dx} \equiv \frac{\partial f(x)}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

- If $\mathbf{x} = (x_1, \ldots, x_p)^\top$ is multi-dimensional, we have the gradient:

$$\nabla f(\mathbf{x}) = \frac{df(\mathbf{x})}{d\mathbf{x}} = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_p} \right)^\top$$

  - The gradient vector points in the direction of steepest ascent in $f(\mathbf{x})$. This is useful for optimization.
- If $\mathbf{f}$ has multiple outputs, we have the **Jacobian**

Calculus: Differentiation

- Things that should sound familiar to you: product rule, quotient rule, chain rule, increasing/decreasing, concave/convex ...
- Derivative as "rate-of-change":

$$f'(x) \equiv \frac{df(x)}{dx} \equiv \frac{\partial f(x)}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

- If $\mathbf{x} = (x_1, \ldots, x_p)^\top$ is multi-dimensional, we have the gradient:

$$\nabla f(\mathbf{x}) = \frac{df(\mathbf{x})}{d\mathbf{x}} = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_p} \right)^\top$$

  - The gradient vector points in the direction of steepest ascent in $f(\mathbf{x})$. This is useful for optimization.
- If $\mathbf{f}$ has multiple outputs, we have the **Jacobian**
- The **Hessian** matrix is like the Jacobian but with second-order derivatives

Calculus: Multivariate Calculus

Calculus: Multivariate Calculus

- Univariate normal: $\mathcal{N}\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$

Calculus: Multivariate Calculus

- Univariate normal: $\mathcal{N}\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$
  - Properties of Gaussians:
  - If X, Y are independent normals then $X+Y \sim \mathcal{N}\left(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2\right)$
  - Any PDF proportional to $\exp\left(ax^2+bx+c\right)$ must be a Gaussian PDF.

Calculus: Multivariate Calculus

- Univariate normal: $\mathcal{N}\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$
  - Properties of Gaussians:
  - If X, Y are independent normals then $X+Y \sim \mathcal{N}\left(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2\right)$
  - Any PDF proportional to $\exp\left(ax^2+bx+c\right)$ must be a Gaussian PDF.
- Multivariate normal:
  $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\det(2\pi\boldsymbol{\Sigma})^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$

Calculus: Multivariate Calculus

- Univariate normal: $\mathcal{N}\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$
  - Properties of Gaussians:
  - If X, Y are independent normals then $X + Y \sim \mathcal{N}\left(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2\right)$
  - Any PDF proportional to $\exp\left(ax^2 + bx + c\right)$ must be a Gaussian PDF.
- Multivariate normal:
  $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\det(2\pi\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$
- Matrix differentiation: Generally analogous to univariate differentiation. But pay attention to the dimensions! For example:

Calculus: Multivariate Calculus

- Univariate normal: $\mathcal{N}\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$
  - Properties of Gaussians:
  - If X, Y are independent normals then $X+Y \sim \mathcal{N}\left(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2\right)$
  - Any PDF proportional to $\exp\left(ax^2+bx+c\right)$ must be a Gaussian PDF.
- Multivariate normal:
  $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\det(2\pi\boldsymbol{\Sigma})^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$
- Matrix differentiation: Generally analogous to univariate differentiation. But pay attention to the dimensions! For example:

$$\frac{d\mathbf{x}^\top\mathbf{a}}{d\mathbf{x}} = \frac{d\mathbf{a}^\top\mathbf{x}}{d\mathbf{x}} = \mathbf{a}\ , \ \frac{d\mathbf{a}^\top\mathbf{X}\mathbf{b}}{d\mathbf{X}} = \mathbf{a}\mathbf{b}^\top$$

$$\frac{d\mathbf{a}^\top\mathbf{X}^\top\mathbf{b}}{d\mathbf{X}} = \mathbf{b}\mathbf{a}^\top\ , \ \frac{d\mathbf{a}^\top\mathbf{X}\mathbf{a}}{d\mathbf{X}} = \frac{d\mathbf{a}^\top\mathbf{X}^\top\mathbf{a}}{d\mathbf{X}} = \mathbf{a}\mathbf{a}^\top$$

Calculus: Optimization in high-dimension

- Local Extrema: Recall that the local extrema of a single-variable function can be found by setting its derivative to 0. The same is true in multivariate case, using the condition $\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \mathbf{0}$. However, this equation is often intractable. We can also search for local minima numerically using gradient-based methods.

Calculus: Optimization in high-dimension

- Local Extrema: Recall that the local extrema of a single-variable function can be found by setting its derivative to 0. The same is true in multivariate case, using the condition $\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \mathbf{0}$. However, this equation is often intractable. We can also search for local minima numerically using gradient-based methods.
- Gradient Descent (finding minima): We start with an initial guess at a useful value for $\mathbf{x}$: $\mathbf{x}_0$. Then at each step i we update our guess by going *against* the direction of the gradient vector:

Calculus: Optimization in high-dimension

- Local Extrema: Recall that the local extrema of a single-variable function can be found by setting its derivative to 0. The same is true in multivariate case, using the condition $\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \mathbf{0}$. However, this equation is often intractable. We can also search for local minima numerically using gradient-based methods.
- Gradient Descent (finding minima): We start with an initial guess at a useful value for $\mathbf{x}$: $\mathbf{x}_0$. Then at each step i we update our guess by going *against* the direction of the gradient vector:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \eta \nabla f(\mathbf{x}_i)$$

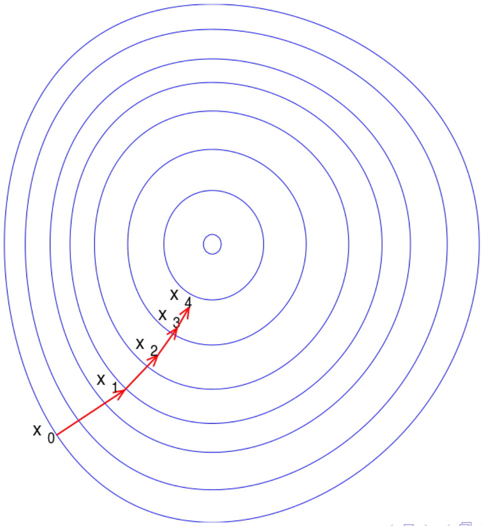
Calculus: Optimization in high-dimension

- Local Extrema: Recall that the local extrema of a single-variable function can be found by setting its derivative to 0. The same is true in multivariate case, using the condition $\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \mathbf{0}$. However, this equation is often intractable. We can also search for local minima numerically using gradient-based methods.
- Gradient Descent (finding minima): We start with an initial guess at a useful value for $\mathbf{x}$: $\mathbf{x}_0$. Then at each step i we update our guess by going *against* the direction of the gradient vector:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \eta \nabla f(\mathbf{x}_i)$$

where $\eta > 0$ is the step size. We stop updating $\mathbf{x}_i$ when the value of the gradient is close to 0

Visualization of gradient descent



Calculus: Optimization in high-dimension

- Lagrange Multipliers: This technique is used to optimize a function $f(\mathbf{x})$ given some constraint $g(\mathbf{x}) = 0$. First, construct what is called the Lagrangian function $L(\mathbf{x}, \lambda)$:

Calculus: Optimization in high-dimension

- Lagrange Multipliers: This technique is used to optimize a function $f(\mathbf{x})$ given some constraint $g(\mathbf{x}) = 0$. First, construct what is called the Lagrangian function $L(\mathbf{x}, \lambda)$:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

Calculus: Optimization in high-dimension

- Lagrange Multipliers: This technique is used to optimize a function $f(\mathbf{x})$ given some constraint $g(\mathbf{x}) = 0$. First, construct what is called the Lagrangian function $L(\mathbf{x}, \lambda)$:

  $$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

  Then, set the derivative of L with respect to both $\mathbf{x}$ and λ equal to 0:

Calculus: Optimization in high-dimension

- Lagrange Multipliers: This technique is used to optimize a function $f(\mathbf{x})$ given some constraint $g(\mathbf{x}) = 0$. First, construct what is called the Lagrangian function $L(\mathbf{x}, \lambda)$:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

Then, set the derivative of L with respect to both $\mathbf{x}$ and λ equal to 0:

$$\nabla L_{\mathbf{x}} = \nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = 0, \quad \frac{\partial L}{\partial \lambda} = g(\mathbf{x}) = 0$$

Calculus: Optimization in high-dimension

- Lagrange Multipliers: This technique is used to optimize a function $f(\mathbf{x})$ given some constraint $g(\mathbf{x}) = 0$. First, construct what is called the Lagrangian function $L(\mathbf{x}, \lambda)$:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

Then, set the derivative of L with respect to both $\mathbf{x}$ and λ equal to 0:

$$\nabla L_{\mathbf{x}} = \nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = 0, \quad \frac{\partial L}{\partial \lambda} = g(\mathbf{x}) = 0$$

If $\mathbf{x}$ is d-dimensional, this will give you a system of $d + 1$ equations. In this way, you can solve analytically for $\mathbf{x}$ to find the optimal value of $f(\mathbf{x})$ subject to the constraint $g(\mathbf{x})$.

Calculus: Optimization in high-dimension

- Lagrange Multipliers: This technique is used to optimize a function $f(\mathbf{x})$ given some constraint $g(\mathbf{x}) = 0$. First, construct what is called the Lagrangian function $L(\mathbf{x}, \lambda)$:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

Then, set the derivative of L with respect to both $\mathbf{x}$ and λ equal to 0:

$$\nabla L_{\mathbf{x}} = \nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = 0, \quad \frac{\partial L}{\partial \lambda} = g(\mathbf{x}) = 0$$

If $\mathbf{x}$ is d-dimensional, this will give you a system of $d + 1$ equations. In this way, you can solve analytically for $\mathbf{x}$ to find the optimal value of $f(\mathbf{x})$ subject to the constraint $g(\mathbf{x})$. As with unconstrained optimization, this too becomes intractable as the dimension increases and gradient descent is used to make progress.

Statistics: Probability and Inference

- Things that should sound familiar to you:

Statistics: Probability and Inference

- Things that should sound familiar to you:
    - PDF/PMF, CDF/CMF, conditional/marginal/joint
    - Expectations, Variance, Covariance
    - LOTP, LOTE/Adam's Law, Eve's Law, Bayes' Theorem
    - LLN, CLT
    - Likelihood function, MLE, prior/posterior

Statistics: Probability and Inference

- Things that should sound familiar to you:
  - PDF/PMF, CDF/CMF, conditional/marginal/joint
  - Expectations, Variance, Covariance
  - LOTP, LOTE/Adam's Law, Eve's Law, Bayes' Theorem
  - LLN, CLT
  - Likelihood function, MLE, prior/posterior
- Useful facts:
  - Triangle Inequality: $|\operatorname{Cov}(X, Y)| \leq \sqrt{\operatorname{Var}(X) \operatorname{Var}(Y)}$

Statistics: Probability and Inference

- Things that should sound familiar to you:
  - PDF/PMF, CDF/CMF, conditional/marginal/joint
  - Expectations, Variance, Covariance
  - LOTP, LOTE/Adam's Law, Eve's Law, Bayes' Theorem
  - LLN, CLT
  - Likelihood function, MLE, prior/posterior
- Useful facts:
  - Triangle Inequality: $|\operatorname{Cov}(X, Y)| \leq \sqrt{\operatorname{Var}(X) \operatorname{Var}(Y)}$
  - Cauchy-Schwarz Inequality: $|E(XY)| \leq \sqrt{E\left(X^2\right) E\left(Y^2\right)}$
  - Markov Inequality: for any $a > 0$, $P(|Y| \geq a) \leq \frac{E|Y|}{a}$

Statistics: Probability and Inference

- Things that should sound familiar to you:
  - PDF/PMF, CDF/CMF, conditional/marginal/joint
  - Expectations, Variance, Covariance
  - LOTP, LOTE/Adam's Law, Eve's Law, Bayes' Theorem
  - LLN, CLT
  - Likelihood function, MLE, prior/posterior
- Useful facts:
  - Triangle Inequality: $|\operatorname{Cov}(X, Y)| \leq \sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}$
  - Cauchy-Schwarz Inequality: $|E(XY)| \leq \sqrt{E\left(X^2\right)E\left(Y^2\right)}$
  - Markov Inequality: for any $a > 0$, $P(|Y| \geq a) \leq \frac{E|Y|}{a}$
  - Chebyshev Inequality : For any Y with finite variance and $\epsilon > 0$, $P(|Y - \mu| \geq \epsilon) \leq \sigma^2/\epsilon^2$

Statistics: Probability and Inference

- Things that should sound familiar to you:
  - PDF/PMF, CDF/CMF, conditional/marginal/joint
  - Expectations, Variance, Covariance
  - LOTP, LOTE/Adam's Law, Eve's Law, Bayes' Theorem
  - LLN, CLT
  - Likelihood function, MLE, prior/posterior
- Useful facts:
  - Triangle Inequality: $|\operatorname{Cov}(X, Y)| \leq \sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}$
  - Cauchy-Schwarz Inequality: $|E(XY)| \leq \sqrt{E\left(X^2\right)E\left(Y^2\right)}$
  - Markov Inequality: for any $a > 0$, $P(|Y| \geq a) \leq \frac{E|Y|}{a}$
  - Chebyshev Inequality : For any Y with finite variance and $\epsilon > 0$, $P(|Y - \mu| \geq \epsilon) \leq \sigma^2/\epsilon^2$
  - Jensen's Inequality: $Eg(Y) \geq g(EY)$ for g convex

Statistics: Markov Chain

Statistics: Markov Chain

- A sequence of random variables $X_1, X_2, \ldots$ is said to be a Markov chain if it satisfies the Markov property:

Statistics: Markov Chain

- A sequence of random variables $X_1, X_2, \ldots$ is said to be a Markov chain if it satisfies the Markov property:

$$X_{n+1} \,|X_1, \ldots, X_n \sim X_{n+1}|\, X_n$$

Statistics: Markov Chain

- A sequence of random variables $X_1, X_2, \ldots$ is said to be a Markov chain if it satisfies the Markov property:

$$X_{n+1} \,|X_1, \ldots, X_n \sim X_{n+1}|\, X_n$$

i.e., knowing the value of X_n tells you the same amount of information about X_{n+1} as knowing all of $X_1, \ldots, X_n$.

Statistics: Markov Chain

- A sequence of random variables $X_1, X_2, \ldots$ is said to be a Markov chain if it satisfies the Markov property:

$$X_{n+1} \,|X_1, \ldots, X_n \sim X_{n+1}|\, X_n$$

i.e., knowing the value of X_n tells you the same amount of information about X_{n+1} as knowing all of $X_1, \ldots, X_n$. If the X_i 's are a discrete distribution the Markov property can be written as:

Statistics: Markov Chain

- A sequence of random variables $X_1, X_2, \ldots$ is said to be a Markov chain if it satisfies the Markov property:

$$X_{n+1} \,|X_1, \ldots, X_n \sim X_{n+1}|\, X_n$$

i.e., knowing the value of X_n tells you the same amount of information about X_{n+1} as knowing all of $X_1, \ldots, X_n$. If the X_i 's are a discrete distribution the Markov property can be written as:

$$\mathbb{P}\left(X_{n+1} = j_{n+1} \mid X_n = j_n, \ldots, X_1 = j_1\right) = \mathbb{P}\left(X_{n+1} = j_{n+1} \mid X_n = j_n\right)$$

Statistics: Markov Chain

- A sequence of random variables $X_1, X_2, \ldots$ is said to be a Markov chain if it satisfies the Markov property:

$$X_{n+1} \,| X_1, \ldots, X_n \sim X_{n+1} | \, X_n$$

  i.e., knowing the value of X_n tells you the same amount of information about X_{n+1} as knowing all of $X_1, \ldots, X_n$. If the X_i 's are a discrete distribution the Markov property can be written as:

$$\mathbb{P}\left(X_{n+1} = j_{n+1} \mid X_n = j_n, \ldots, X_1 = j_1\right) = \mathbb{P}\left(X_{n+1} = j_{n+1} \mid X_n = j_n\right)$$

- Application: latent Dirichlet allocation, Viterbi algorithm, EM algorithm, missing data, etc.

Statistics: Computational Thinking

- Solutions to all practical problems **need and (for the most part) only need to be** "computable"

Statistics: Computational Thinking

- Solutions to all practical problems **need and (for the most part) only need to be** "computable"
- Overflow and Underflow

Statistics: Computational Thinking

- Solutions to all practical problems **need and (for the most part) only need to be** "computable"
- Overflow and Underflow
  - Never multiply many probabilities or density values literally

Statistics: Computational Thinking

- Solutions to all practical problems **need and (for the most part) only need to be** "computable"
- Overflow and Underflow
  - Never multiply many probabilities or density values literally
  - Operate at the logarithmic scale if you can: For example, when computing the summation of many small (or huge) numbers, it is better to do them properly via logarithm.

Statistics: Computational Thinking

- Solutions to all practical problems **need and (for the most part) only need to be** "computable"
- Overflow and Underflow
  - Never multiply many probabilities or density values literally
  - Operate at the logarithmic scale if you can: For example, when computing the summation of many small (or huge) numbers, it is better to do them properly via logarithm.
  - Example: softmax function. $\text{softmax}(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{k} \exp(x_j)}$

Statistics: Computational Thinking

- Solutions to all practical problems **need and (for the most part) only need to be** "computable"
- Overflow and Underflow
  - Never multiply many probabilities or density values literally
  - Operate at the logarithmic scale if you can: For example, when computing the summation of many small (or huge) numbers, it is better to do them properly via logarithm.
  - Example: softmax function. $\text{softmax}(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{k} \exp(x_j)}$
- Approximation

Statistics: Computational Thinking

- Solutions to all practical problems **need and (for the most part) only need to be** "computable"
- Overflow and Underflow
  - Never multiply many probabilities or density values literally
  - Operate at the logarithmic scale if you can: For example, when computing the summation of many small (or huge) numbers, it is better to do them properly via logarithm.
  - Example: softmax function. $\text{softmax}(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{k} \exp(x_j)}$
- Approximation
  - Taylor expansion: $f(x+\epsilon) = f(x) + f'(x)\epsilon + \cdots + \frac{f^{(k)}(x)}{k!}\epsilon^k + \ldots$

Statistics: Computational Thinking

- Solutions to all practical problems **need and (for the most part) only need to be** "computable"
- Overflow and Underflow
  - Never multiply many probabilities or density values literally
  - Operate at the logarithmic scale if you can: For example, when computing the summation of many small (or huge) numbers, it is better to do them properly via logarithm.
  - Example: softmax function. $\text{softmax}(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{k} \exp(x_j)}$
- Approximation
  - Taylor expansion: $f(x+\epsilon) = f(x) + f'(x)\epsilon + \cdots + \frac{f^{(k)}(x)}{k!}\epsilon^k + \ldots$

$$\Rightarrow f(\mathbf{x}+\epsilon) = f(\mathbf{x}) + [\nabla f(\mathbf{x})]^\top \epsilon + \frac{1}{2}\epsilon^T H(\mathbf{x})\epsilon + o\left(\|\epsilon\|^2\right)$$