Linear Algebra Review
(with a Small Dose of Optimization)

Hristo Paskov
CS246

Outline

- Basic definitions
- Subspaces and Dimensionality
- Matrix functions: inverses and eigenvalue decompositions
- Convex optimization

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

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

$$\boldsymbol{r}_{\color{red}i} = [M_{{\color{red}i}1} \quad \dots \quad M_{{\color{red}i}n}]^T \quad \boldsymbol{c}_{\color{red}i} = [M_{1{\color{red}i}} \quad \dots \quad M_{m{\color{red}i}}]^T$$

Multiplication

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

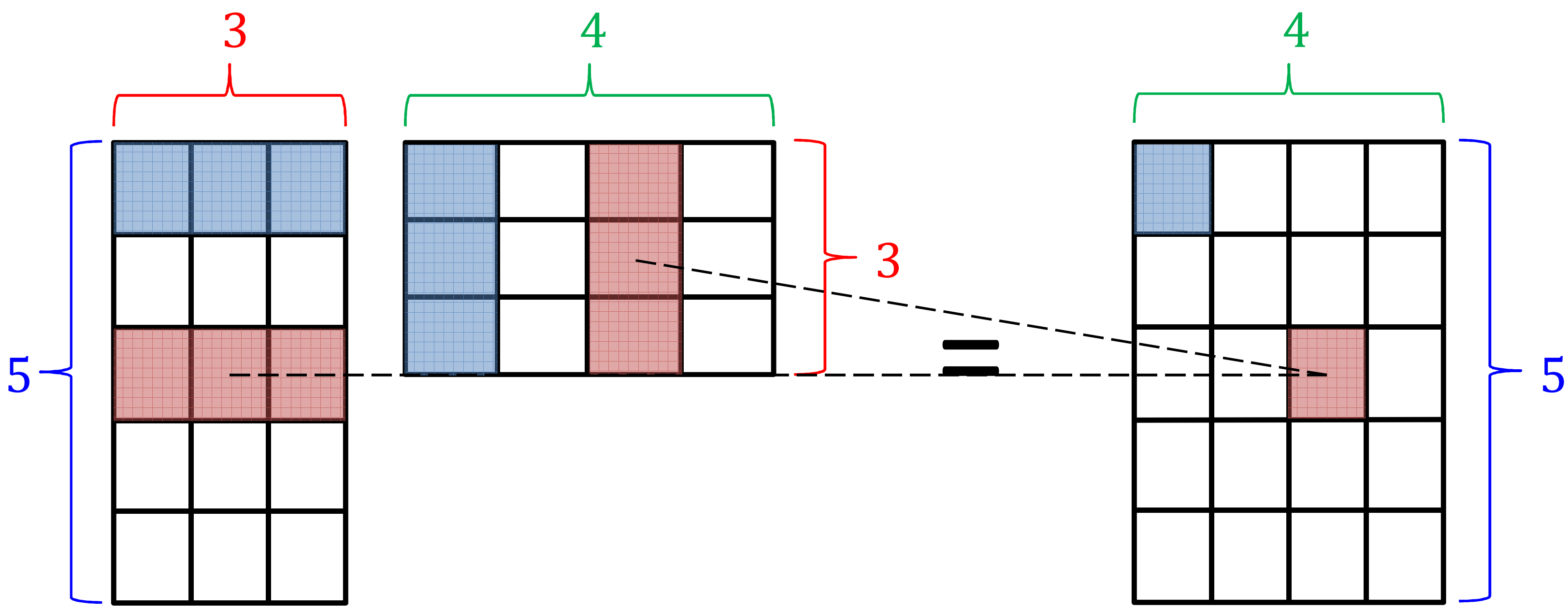
$$x^T y = \sum_{i=1}^{d} x_i y_i$$

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

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

Multiplication

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



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

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

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

Multiplication Properties

- Associative

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

- Distributive

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

- NOT commutative

$$AB \neq BA$$

  – Dimensions may not even be conformable

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 $A \in \mathbb{R}^{m \times m}$

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

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 $L \in \mathbb{R}^{m \times m}$

$$A = LL^T$$

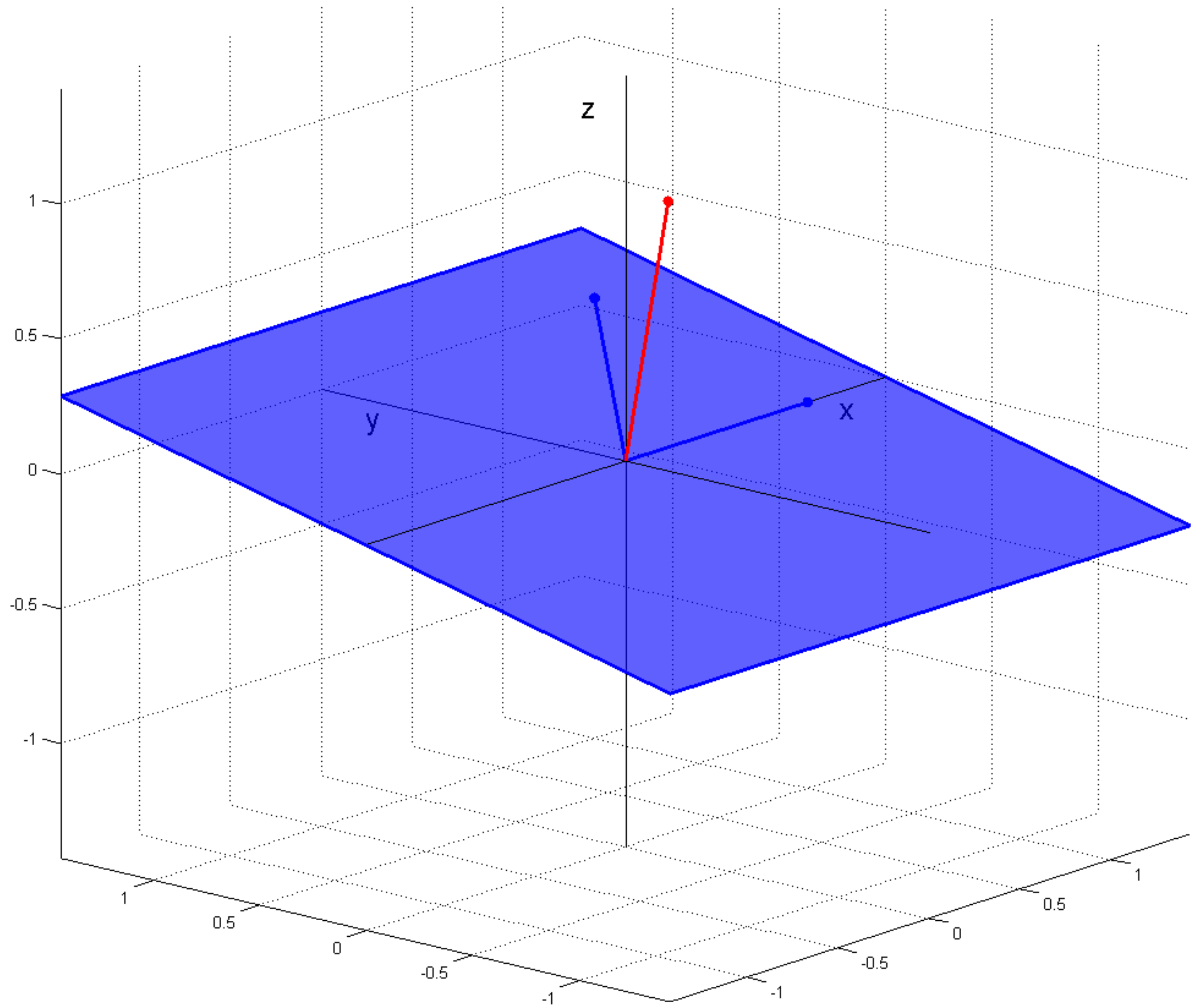
Outline

- Basic definitions
- Subspaces and Dimensionality
- Matrix functions: inverses and eigenvalue decompositions
- Convex optimization

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



Linear Subspaces

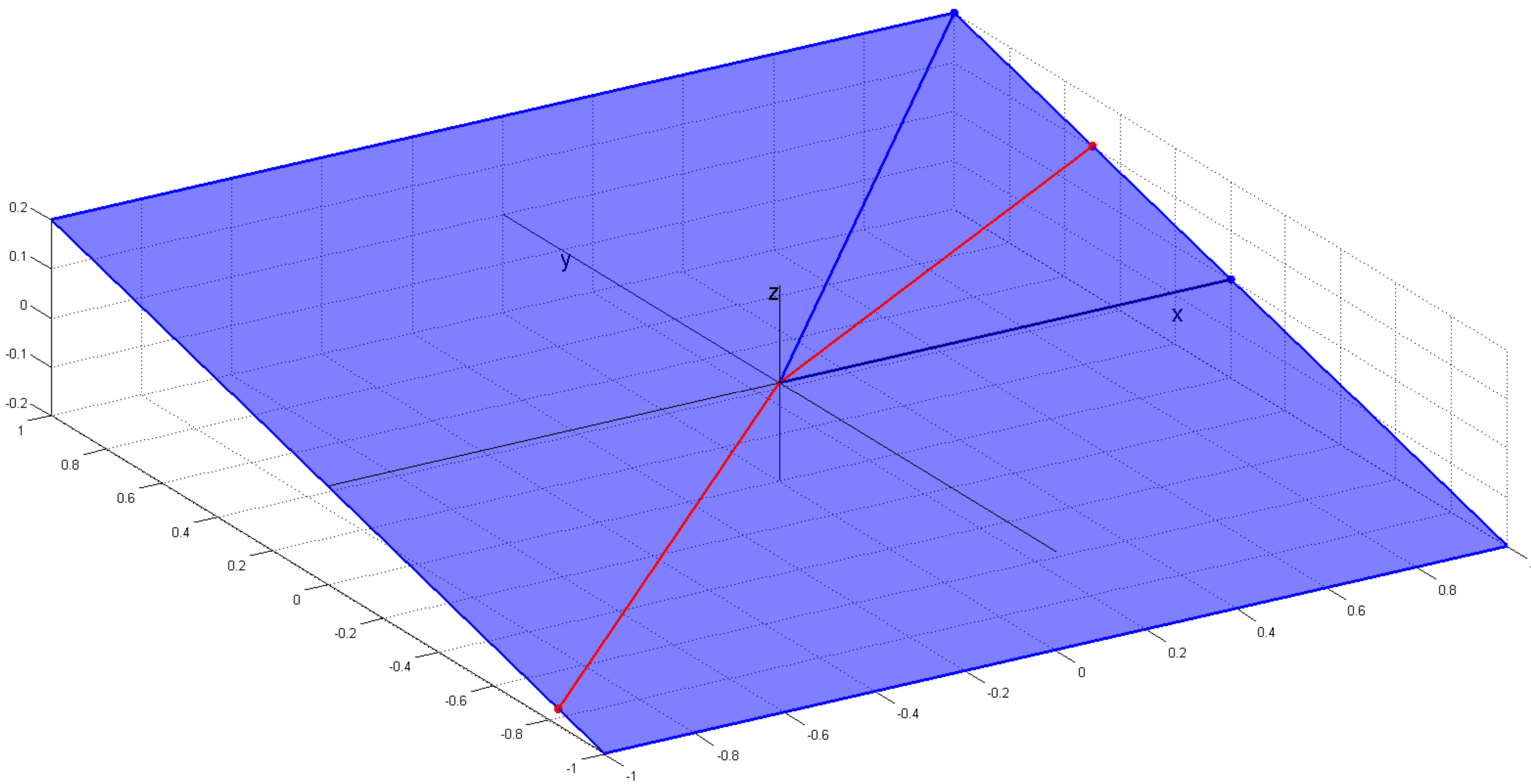
- Subspace $\mathcal{V} \subset \mathbb{R}^n$ satisfies
  1. $0 \in \mathcal{V}$
  2. If $x, y \in \mathcal{V}$ and $c \in \mathbb{R}$, then $c(x + y) \in \mathcal{V}$
- Vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$ *span* $\mathcal{V}$ if

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

Linear Independence and Dimension

- Vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ 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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ span $\mathcal{V}$ and are linearly independent
  - If $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ span $\mathcal{V}$ then
    - $k \geq m$
    - If $k > m$ then $\boldsymbol{y}_i$ are NOT linearly independent

Linear Independence and Dimension

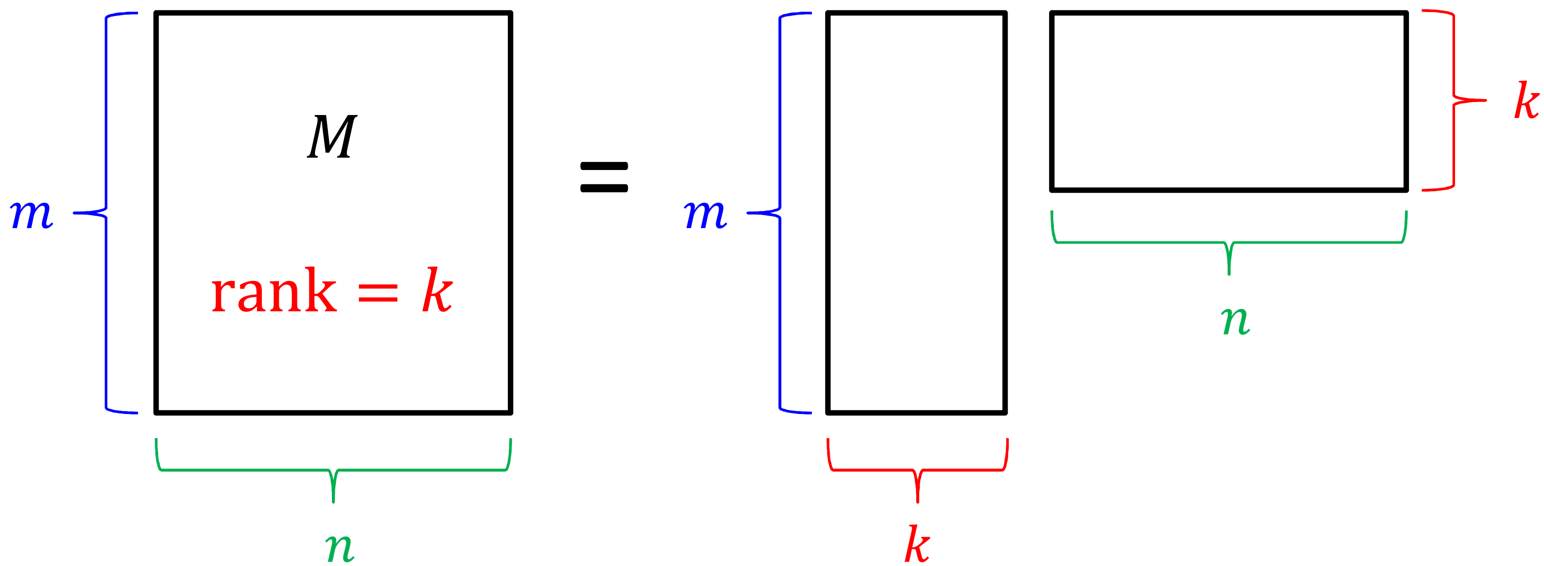


Matrix Subspaces

- 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

Matrix Rank

- $\text{rank}(M)$ gives dimensionality of row and 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. $\text{rank}(AB) \leq \min\big(\text{rank}(A), \text{rank}(B)\big)$
  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 $n > \text{rank}(A)$

Outline

- Basic definitions
- Subspaces and Dimensionality
- Matrix functions: inverses and eigenvalue decompositions
- Convex optimization

Matrix Inverse

- $M \in \mathbb{R}^{m \times m}$ is invertible iff $\text{rank}(M) = m$
- 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

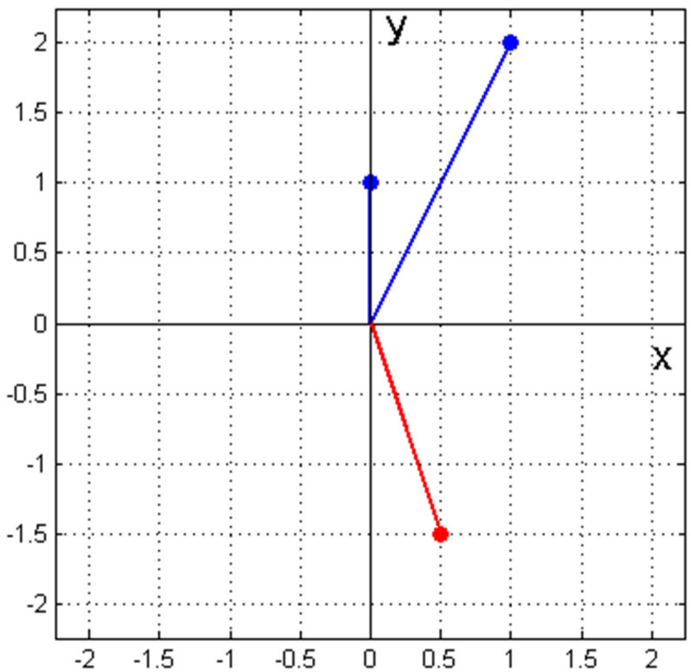
- Given $M \in \mathbb{R}^{m \times n}, y \in \mathbb{R}^m$ wish to solve
$$Mx = y$$
  - Exists only if $y \in \mathrm{col}(M)$
    - Possibly infinite number of solutions
- If M is invertible then $x = M^{-1}y$
  - Notational device, <u>do not</u> actually invert matrices
  - Computationally, use solving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

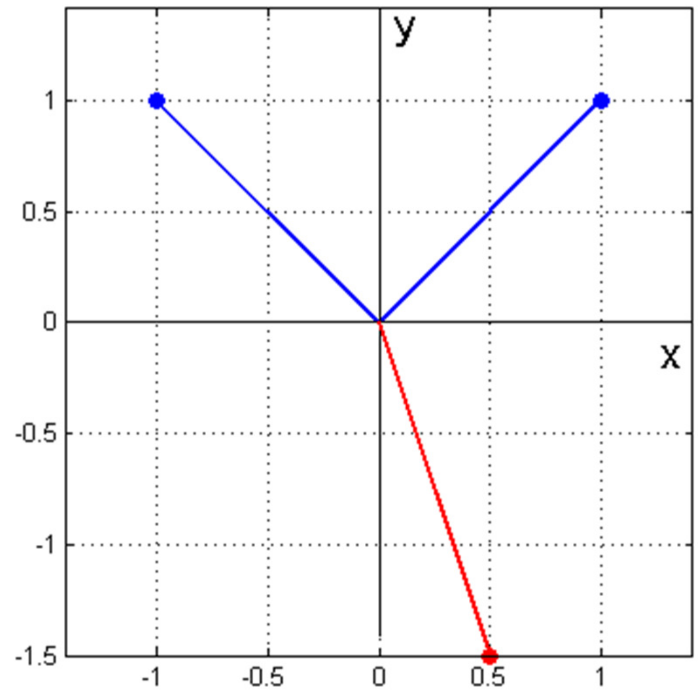
Systems of Equations

- 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 $\mathrm{rank}(M) = n < m$

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

Systems of Equations





$$\begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} .5 \\ -2.5 \end{bmatrix} = \begin{bmatrix} .5 \\ -1.5 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ -.5 \end{bmatrix} = \begin{bmatrix} .5 \\ -1.5 \end{bmatrix}$$

Eigenvalue Decomposition

- Eigenvalue decomposition of symmetric $M \in \mathbb{R}^{m \times m}$ is

$$M = Q\Sigma Q^T = \sum_{i=1}^{m} \lambda_i \boldsymbol{q}_i \boldsymbol{q}_i^T$$

  - $\Sigma = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ contains eigenvalues of M
  - Q is orthogonal and contains eigenvectors $\boldsymbol{q}_i$ of M
- If M is not symmetric but *diagonalizable*

$$M = Q\Sigma Q^{-1}$$

  - Σ is diagonal by possibly complex
  - Q not necessarily orthogonal

Characterizations of Eigenvalues

- Traditional formulation

$$Mx = \lambda x$$

  - Leads to characteristic polynomial

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

- Rayleigh quotient (symmetric M)

$$\max_x \frac{x^T M x}{\|x\|_2^2}$$

Eigenvalue Properties

- For $M \in \mathbb{R}^{m \times m}$ with eigenvalues λ_i
  1. $\operatorname{tr}(M) = \sum_{i=1}^{m} \lambda_i$
  2. $\det(M) = \lambda_1 \lambda_2 \ldots \lambda_m$
  3. $\operatorname{rank}(M) = \#\lambda_i \neq 0$
- When M is symmetric
  - Eigenvalue decomposition is singular value decomposition
  - Eigenvectors for nonzero eigenvalues give orthogonal basis for $\operatorname{row}(M) = \operatorname{col}(M)$

Simple Eigenvalue Proof

- Why $\det(M - \lambda I) = 0$?
- Assume M is symmetric and full rank

1. $M = Q\Sigma Q^T$
2. $M - \lambda I = Q\Sigma Q^T - \lambda I = Q(\Sigma - \lambda I)Q^T$ ← $QQ^T = I$
3. If $\lambda = \lambda_i$, i^{th} eigenvalue of $M - \lambda I$ is 0
4. Since $\det(M - \lambda I)$ is product of eigenvalues, one of the terms is 0, so product is 0

Outline

- Basic definitions
- Subspaces and Dimensionality
- Matrix functions: inverses and eigenvalue decompositions
- Convex optimization

Convex Optimization

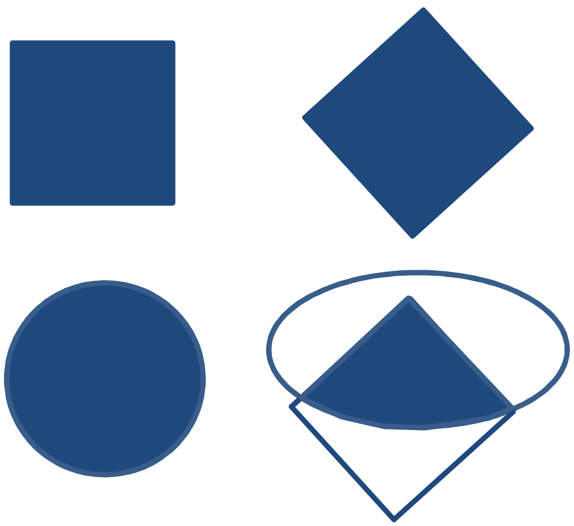
- Find minimum of a function subject to solution constraints
- Business/economics/ game theory
  - Resource allocation
  - Optimal planning and strategies
- Statistics and Machine Learning
  - All forms of regression and classification
  - Unsupervised learning
- Control theory
  - Keeping planes in the air!

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

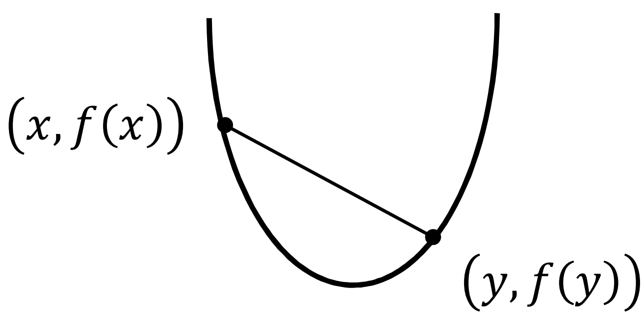


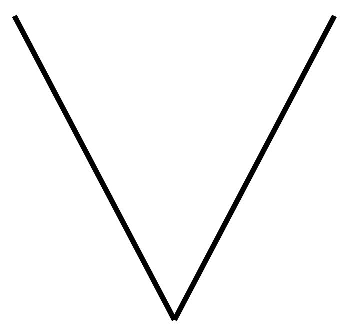
Convex Functions

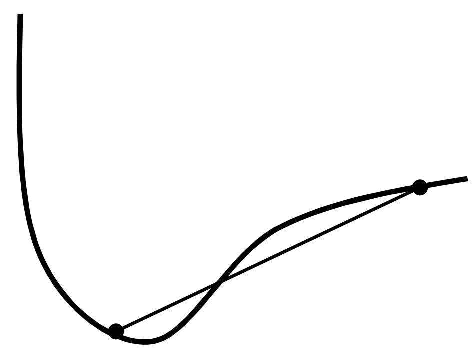
- A real-valued function f is convex if $\mathrm{dom} f$ *is* convex and $\forall x, y \in \mathrm{dom} f$ and $\forall \alpha \in [0,1]$

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

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



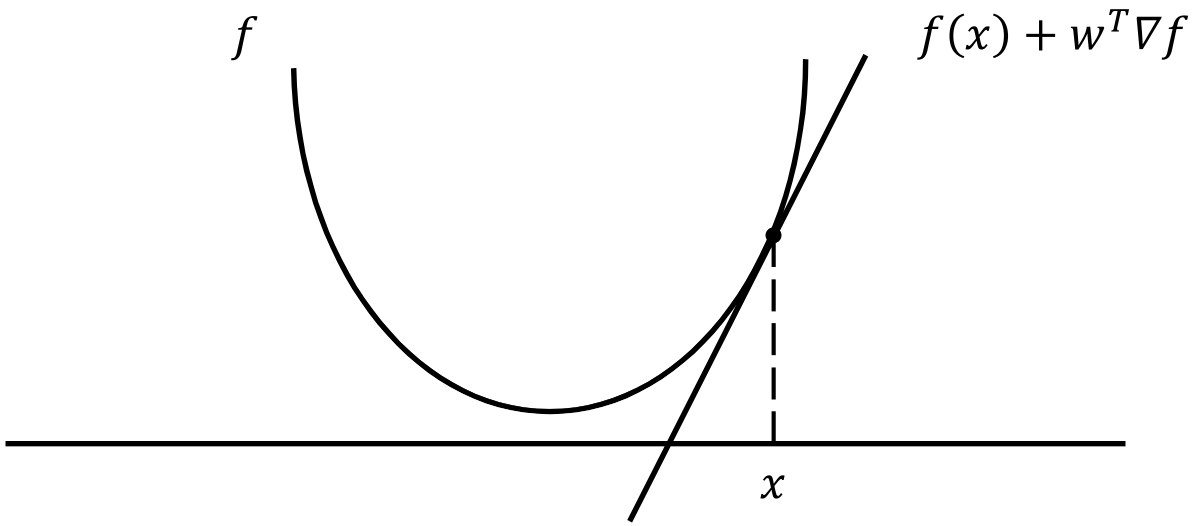




Gradients

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

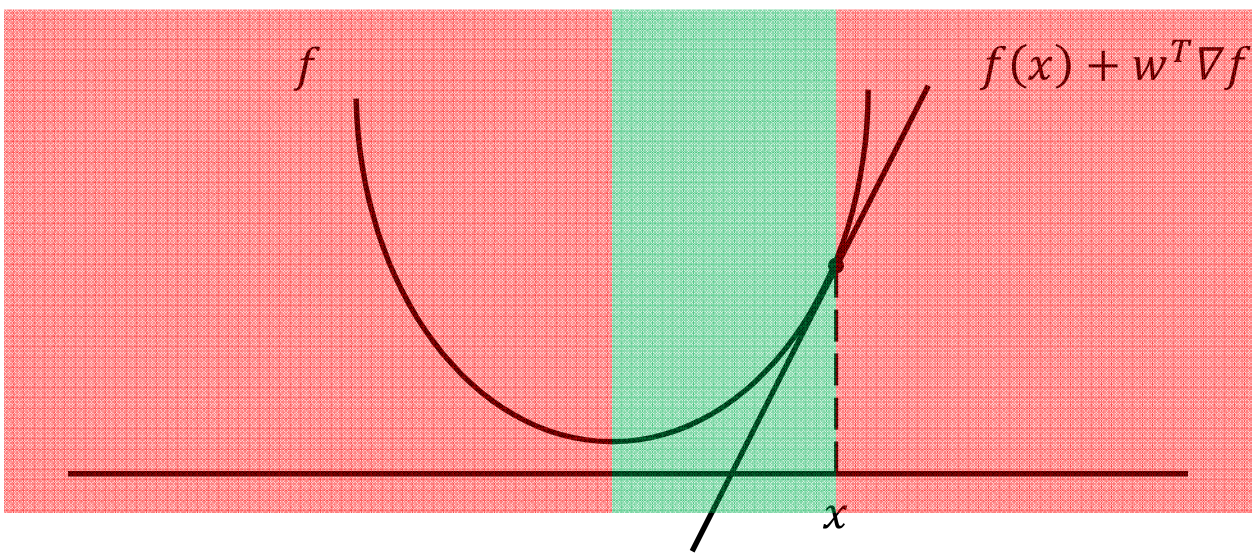
$$\nabla f = \left[\frac{\delta f}{\delta x_1} \quad \cdots \quad \frac{\delta f}{\delta x_d}\right]^T$$



Gradients

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

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



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 $\nabla f = 0$

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

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

Stochastic Gradient Descent

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

  $\theta = \theta_0$

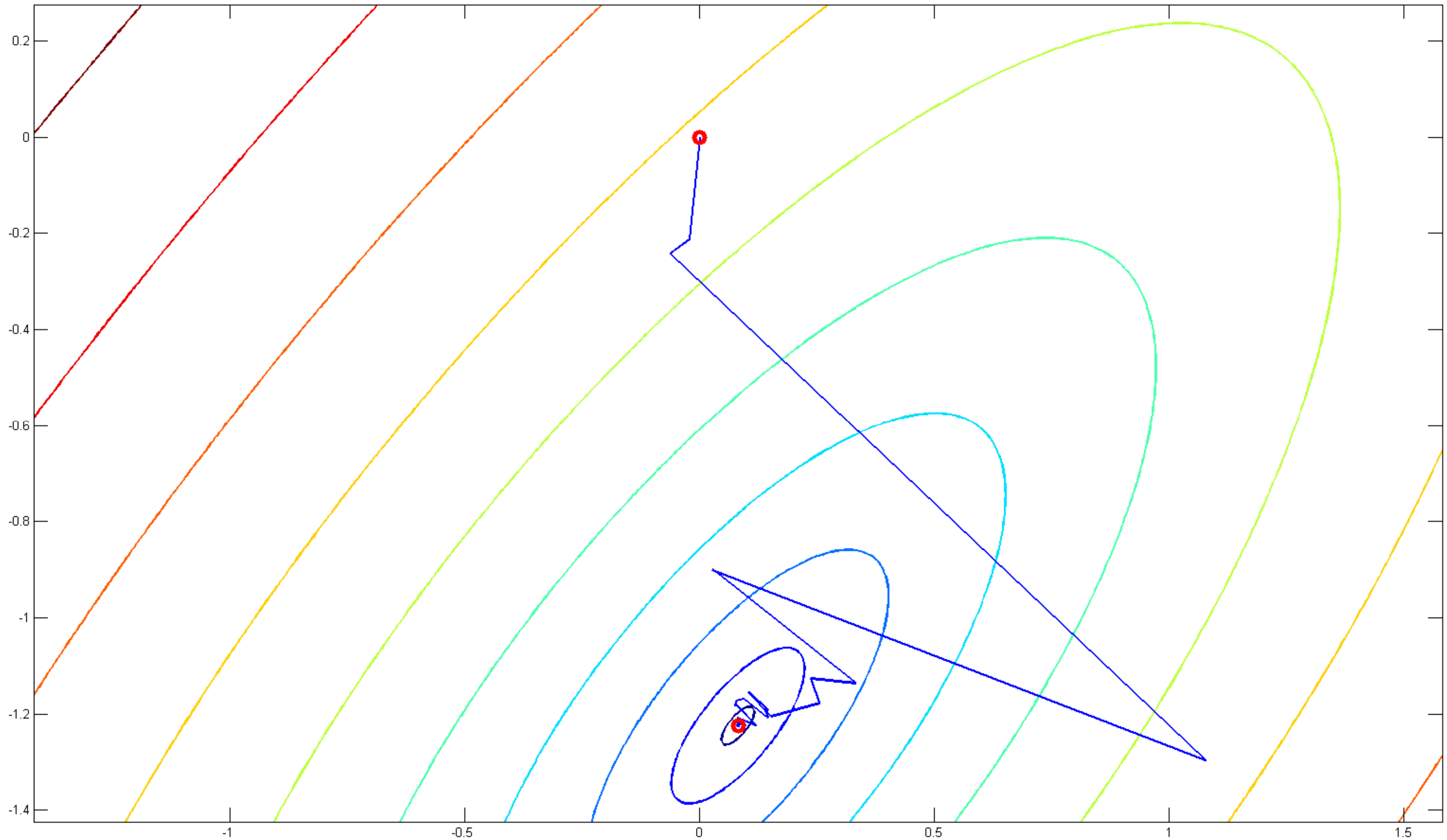
  Do until $f(\theta)$ nearly optimal

  For $i = 1$ to n <u>in random order</u>

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

- Finds nearly optimal θ

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



Learning Parameter