1 Linear Algebra – Vectors

1.1 Vector Spaces & Subspaces

$\mathbb{R}^n$: space of vectors with n elements. Vectors $v^{(1)}, \ldots, v^{(m)} \in \mathbb{R}^n$ are **linearly independent** if $\sum_i \alpha_i v^{(i)} = 0$ implies all $\alpha_i = 0$.

Subspace $\mathcal{S} \subseteq \mathbb{R}^n$: for all $x, y \in \mathcal{S}$ and scalars α, β, have $\alpha x + \beta y \in \mathcal{S}$.

Span: $\text{span}(v^{(1)}, \ldots, v^{(m)})$ = all linear combos of $v^{(1)}, \ldots, v^{(m)}$.

Basis: $v^{(1)}, \ldots, v^{(d)}$ is a basis for $\mathcal{S}$ if (1) linearly independent, (2) for all $x \in \mathcal{S}$, $\exists$ scalars α_i s.t. $x = \sum_i \alpha_i v^{(i)}$.

Dimension: number of vectors in basis = d.

Affine set: $\mathcal{X} \subseteq \mathbb{R}^n$ is affine if $\exists$ subspace $\mathcal{S}$ and vector $v^{(0)}$ s.t. $\mathcal{X} = v^{(0)} + \mathcal{S}$ (add $v^{(0)}$ to all vectors in $\mathcal{S}$). To prove affine, show $x^{(0)} \perp \mathcal{S}$ is a subspace.

1.2 Inner Product & Orthogonality

Inner product: $\langle x, y \rangle = x^\top y = y^\top x = x_1 y_1 + \cdots + x_n y_n$.

$\langle x, y \rangle = \|x\|_2 \|y\|_2 \cos(\theta)$, where θ is angle between x and y.

Orthogonal: $x \perp y$ if $\langle x, y \rangle = 0$.

d vectors $x^{(1)}, \ldots, x^{(d)}$ are **mutually orthogonal** if $\langle x^{(i)}, x^{(j)} \rangle = 0$ for all $i \neq j$ (guarantees linear independence).

Orthonormal: mutually orthogonal and $\|x^{(i)}\|_2 = 1$ for all i.

1.3 Vector Norms

A function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is a **norm** if:

- $\|x\| \geq 0$ for all x and $\|x\| = 0$ iff $x = 0$
- $\|x + y\| \leq \|x\| + \|y\|$ for all x, y
- $\|\alpha x\| = |\alpha| \|x\|$ for all $\alpha \in \mathbb{R}$, $x \in \mathbb{R}^n$

ℓ_p **norm** ($1 \leq p < \infty$): $\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$.

Special cases: $\|x\|_1 = |x_1| + \cdots + |x_n|$, $\|x\|_2 = \sqrt{x^\top x}$.

ℓ_0 "norm": $\|x\|_0 = \#$ of non-zero elements (not a true norm).

For arbitrary $x \in \mathbb{R}^n$: $\|x\|_2^2 = x^\top x$.

1.4 Projections

Projection of x onto subspace $\mathcal{S}$: $\Pi_{\mathcal{S}}(x) = \arg\min_{y \in \mathcal{S}} \|y - x\|$.

Unique solution $y^* = \Pi_{\mathcal{S}}(x)$ exists and is or $(x - y^*) \perp \mathcal{S}$.

For projection onto affine space: $(x - y^*) \perp (y - y^*)$ for all $y \in \mathcal{S}$.

Projection onto 1-D subspace $\mathcal{S} = \text{span}(v)$: $\Pi_{\mathcal{S}}(x) = \frac{\langle x, v \rangle}{\|v\|^2} v$.

Projection onto subspace with orthonormal basis $x^{(1)}, \ldots, x^{(d)}$: $\Pi_{\mathcal{S}}(x) = \sum_{i=1}^d \langle x, x^{(i)} \rangle x^{(i)}$.

2 Linear Algebra – Matrices

2.1 Range, Nullspace, Rank

For $A \in \mathbb{R}^{m \times n}$:

Range (column space): $\mathcal{R}(A) = \{Ax \mid x \in \mathbb{R}^n\}$.

$\mathcal{R}(A)$ is a subspace. Rank(A) = dimension of $\mathcal{R}(A)$ = # linearly independent columns = # linearly independent rows.

Nullspace: $\mathcal{N}(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}$.

$\mathcal{N}(A)$ is a subspace. Key relationships:

- $\mathcal{N}(A) \perp \mathcal{R}(A^\top)$
- $\mathcal{N}(A) \oplus \mathcal{R}(A^\top) = \mathbb{R}^n$ (any $v \in \mathbb{R}^n$ decomposes into sum from $\mathcal{N}(A)$ and $\mathcal{R}(A^\top)$)
- $\dim(\mathcal{N}(A)) + \text{Rank}(A) = n$

2.2 Eigenvalues & Eigenvectors

For square $A \in \mathbb{R}^{n \times n}$: $Av = \lambda v$ means λ is **eigenvalue** and v is **eigenvector**.

Find eigenvalues: solve $\det(A - \lambda I) = 0$. Then solve $(A - \lambda I)v = 0$ for eigenvector v.

If A is rank-deficient, then $\det(A) = 0$ and at least one eigenvalue is 0.

$AA^\top$ and $A^\top A$ share same non-zero eigenvalues.

Tr(A) (sum of diagonal entries) = sum of eigenvalues.

2.3 Symmetric Matrices & PSD/PD

$A \in \mathbb{R}^{n \times n}$ is **symmetric** if $A = A^\top$. Set of $n \times n$ symmetric matrices: $\mathbb{S}^n$.

Symmetric matrices have all real eigenvalues.

$A \in \mathbb{S}^n$ is **positive semidefinite (PSD)** (denoted $A \succeq 0$) if all eigenvalues are non-negative, i.e., $\lambda_1(A), \ldots, \lambda_n(A) \geq 0$. Set of $n \times n$ PSD matrices: $\mathbb{S}^n_+$.

Alternative PSD definition: $A \in \mathbb{S}^n$ is PSD if $x^\top Ax \geq 0$ for all $x \in \mathbb{R}^n$.

Note: showing all elements non-negative does NOT prove PSD.

$A \in \mathbb{S}^n$ is **positive definite (PD)** (denoted $A \succ 0$) if all eigenvalues strictly positive. Set of $n \times n$ PD matrices: $\mathbb{S}^n_{++}$.

Alternative: $x^\top Ax > 0$ for all $x \neq 0$.

Check PD easily: all leading principal minors strictly positive.

A is **negative semidefinite (NSD)** if $\lambda_1(A), \ldots, \lambda_n(A) \leq 0$ or $x^\top Ax \leq 0$ for all x.

A is **negative definite (ND)** if $\lambda_1(A), \ldots, \lambda_n(A) < 0$ or $x^\top Ax < 0$ for all $x \neq 0$.

All PD matrices are PSD. All ND matrices are NSD.

Sign indefinite: has at least one positive and one negative eigenvalue.

2.4 Orthogonal Matrices

$U \in \mathbb{R}^{n \times n}$ with columns $u^{(1)}, \ldots, u^{(n)}$ is **orthogonal** if columns are orthonormal, i.e., $\langle u^{(i)}, u^{(j)} \rangle$ is 1 if $i = j$ and 0 if $i \neq j$.

Equivalently: $U^\top U = I_n$ (where I_n is $n \times n$ identity matrix), i.e., $U^\top = U^{-1}$.

Identity matrix is orthogonal. Also diagonal and full-rank.

2.5 Eigenvalue Decomposition

Consider $A \in \mathbb{R}^{n \times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ and eigenvectors $v^{(1)}, \ldots, v^{(n)}$ (each associated with one eigenvalue).

If $v^{(1)}, \ldots, v^{(n)}$ are linearly independent, then $A = U \Lambda U^{-1}$, where $U = [v^{(1)} \cdots v^{(n)}]$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. A is **diagonalizable**.

If $\lambda_1, \ldots, \lambda_n$ are all distinct, A is always diagonalizable.

Spectral theorem: For symmetric $A \in \mathbb{S}^n$, select eigenvector $v^{(i)}$ with length 1 for each eigenvalue λ_i. Then $A = U \Lambda U^\top$, i.e., U is orthogonal.

Symmetric matrices are always diagonalizable.

2.6 Singular Value Decomposition (SVD)

For arbitrary $A \in \mathbb{R}^{m \times n}$, $\exists$ matrices $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$, and $\Sigma \in \mathbb{R}^{m \times n}$ such that:

$A = U \Sigma V^\top$.

U and V are orthogonal matrices.

Σ is rectangular diagonal matrix: if $n \geq m$, $\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m & 0 & \cdots & 0 \end{bmatrix}$;

if $n \leq m$, $\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$,

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$.

$\sigma_1, \sigma_2, \ldots$ are **singular values** of A.

Let $r = \#$ of non-zero singular values, i.e., $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \sigma_{r+2} = \ldots = 0$. Then $r = \text{Rank}(A)$.

For symmetric PSD, eigenvalues and singular values are the same, and eigenvalue decomposition $A = U \Lambda U^\top$ is a valid SVD. However, eigenvalues and singular values differ in general.

Finding SVD: The non-zero singular values of A are the square root of the non-zero eigenvalues of $AA^\top$ or $A^\top A$. Columns of U (left singular vectors) are eigenvectors of $AA^\top$. Columns of V (right singular vectors) are eigenvectors of $A^\top A$.

If αA, where α is non-negative scalar, is an orthogonal matrix, then one possible SVD for A is $A = I_n \frac{1}{\alpha} (\alpha A)$.

2.7 Matrix Pseudo-Inverse

Pseudo-inverse (Moore-Penrose inverse) of $A = U \Sigma V^\top$ is $A^\dagger = V \Sigma^\dagger U^\top$, where we take the inverse of positive singular values and fill rest with zero.

If A is invertible, then $A^\dagger = A^{-1}$ and $AA^\dagger = I_n$. However, $AA^\dagger$ does not produce I_n in general.

If $A \in \mathbb{R}^{m \times n}$ has linearly independent rows, i.e., $n \geq m = \text{Rank}(A)$, then $A^\dagger = A^\top (AA^\top)^{-1}$.

If $A \in \mathbb{R}^{m \times n}$ has linearly independent columns, i.e., $m \geq n = \text{Rank}(A)$, then $A^\dagger = (A^\top A)^{-1} A^\top$.

2.8 Matrix Norms

Frobenius norm: $\|A\|_F = \|\text{vec}(A)\|_2$, where $\text{vec}(A) \in \mathbb{R}^{mn}$ concatenates all columns of A. Equivalently, $\|A\|_F^2 =$

$\sum_{i=1}^{r} \sigma_i^2(A)$.

ℓ_p-induced norm: $\|A\|_p = \max_{z\in\mathbb{R}^n, z\neq 0} \frac{\|Az\|_p}{\|z\|_p} = \max_{\|x\|_p=1} \|Ax\|_p$.

Spectral norm ($p = 2$): $\|A\|_2 = \sigma_1(A) = \sqrt{\lambda_{\max}(A^\top A)}$, where $\sigma_1(A)$ is largest singular value of A and $\lambda_{\max}(A^\top A)$ is largest eigenvalue of $A^\top A$.

3 Set Theory

3.1 Basic Set Properties

A set $\mathcal{S} \subseteq \mathbb{R}^n$ is **open** if for every $x \in \mathcal{S}$, $\exists \epsilon > 0$ s.t. $B_\epsilon(x) \subset \mathcal{S}$, where $B_\epsilon(x)$ is a ball centered at x with radius ϵ.

$\mathcal{S} \subseteq \mathbb{R}^n$ is **closed** if its complement $\mathbb{R}^n \setminus \mathcal{S}$ is open.

$\mathcal{S} \subseteq \mathbb{R}^n$ is **bounded** if $\exists r > 0$ s.t. $\mathcal{S} \subseteq B_r(0)$.

A set is **compact** if it is closed and bounded.

Interior of $\mathcal{S}$: points $x \in \mathcal{S}$ s.t. we can draw a ball in $\mathbb{R}^n$ centered at x of non-zero radius that belongs to $\mathcal{S}$. Denoted as $\operatorname{int}\mathcal{S}$.

Closure: $\operatorname{cls}(\mathcal{S}) = \{z \in \mathbb{R}^n \mid z = \lim_{k\to\infty} x^{(k)} \text{ where } x^{(k)} \in \mathcal{S}, \forall k\}$.

Boundary: $\partial\mathcal{S} = \operatorname{cls}(\mathcal{S}) \setminus \operatorname{int}(\mathcal{S})$.

3.2 Affine & Convex Sets

Affine combination of $x_1, \ldots, x_k \in \mathbb{R}^n$: $\{\sum_{i=1}^k \alpha_i x_i \mid \sum_{i=1}^k \alpha_i = 1\}$.

Convex combination: $\{\sum_{i=1}^k \alpha_i x_i \mid \sum_{i=1}^k \alpha_i = 1, \alpha_i \geq 0, \forall i\}$.

$\mathcal{S}$ is **affine** if for all $x, y \in \mathcal{S}$ and $t \in \mathbb{R}$, the affine combination $tx + (1-t)y$ is in $\mathcal{S}$ (affine sets are based on subspaces). A hyperplane is an affine set, but a half-space is not.

$\mathcal{S}$ is **convex** if for all $x, y \in \mathcal{S}$ and $t \in [0, 1]$, the convex combination $tx + (1-t)y$ is in $\mathcal{S}$.

A polyhedron $\{x \mid a_i^\top x \leq b_i, c_j x = d, \forall i, j\}$ is convex. Norm balls and half-spaces are convex. The set of PD matrices is convex, and the set of PSD matrices is also convex.

Affine hull of a set: smallest affine set containing the set. It is the set of affine combinations of any k points in the set.

Convex hull of a set: smallest convex set containing the set. It is the set of convex combinations of any k points in the set.

Operations preserving convexity: (1) Intersection of convex sets is convex (note union may not be convex). (2) Affine transformation: $\mathcal{S} = \{f(x) : x \in \mathcal{S}\}$ is convex if $f : \mathbb{R}^n \to \mathbb{R}^m$ is affine and $\mathcal{S}$ is convex. (3) Projections of convex sets are convex.

3.3 Dimension & Relative Interior

Dimension of a set $\mathcal{S} \subseteq \mathbb{R}^n$: If $\mathcal{S}$ is a subspace, dimension is minimum number of spanning vectors. If $\mathcal{S}$ is affine, $\mathcal{S} = x_0 + V$ where V is a subspace, and the dimension is the dimension of V. If $\mathcal{S}$ is convex, dimension is defined as the dimension of the affine hull of $\mathcal{S}$.

Relative interior of convex set $\mathcal{S} \subseteq \mathbb{R}^n$: a point $x \in \mathcal{S}$ is in the relative interior if we can draw a ball in the affine hull of $\mathcal{S}$ centered at x of non-zero radius that belongs to $\mathcal{S}$. Denoted as $\operatorname{relint}\mathcal{S}$.

3.4 Separating Hyperplane

Hyperplane: $(n-1)$ dimensional affine set, can be written as $H = \{z \in \mathbb{R}^n \mid a^\top z = b\}$ for a non-zero vector $a \in \mathbb{R}^n$ and scalar b.

a is the **normal vector** of the hyperplane. For any two vectors $z^1, z^2 \in H$, we have $a \perp (z^1 - z^2)$.

Hyperplanes divide $\mathbb{R}^n$ into half-spaces: $H_- = \{x \mid a^\top x \leq b\}$ and $H_+ = \{x \mid a^\top x \geq b\}$.

Supporting hyperplane theorem: For a convex set C and boundary point $z \in \partial C$, we can always find a supporting hyperplane $H = \{x \in \mathbb{R}^n \mid a^\top x = b\}$ satisfying: (1) $z \in H$, (2) $C \subseteq H_-$, where $H_- = \{x \in \mathbb{R}^n \mid a^\top x \leq b\}$.

Separating hyperplane: A hyperplane $H = \{x \in \mathbb{R}^n \mid a^\top x = b\}$ separates C_1 and C_2 if (1) $C_1 \subseteq H_-$, where $H_- = \{x \mid a^\top x \leq b\}$, (2) $C_2 \subseteq H_+$, where $H_+ = \{x \mid a^\top x \geq b\}$.

If $H \cap C_1 = H \cap C_2 = \emptyset$, then H strictly separates C_1 and C_2.

Separating hyperplane theorem: Assume C_1, C_2 are convex. Two statements: (1) If $C_1 \cap C_2 = \emptyset$, then a separating hyperplane exists. (2) If $C_1 \cap C_2 = \emptyset$, C_1 and C_2 are closed, and either C_1 or C_2 are bounded, then a strictly separating hyperplane exists.

4 Optimization Problems

4.1 Standard Form & Solution Types

Standard form: $\min_x f_0(x)$ subject to $f_i(x) \leq 0$ for $i = 1, \ldots, m$.

Equality constraints can be converted to inequality: $h(x) = 0 \iff h(x) \leq 0$ and $-h(x) \leq 0$.

Point $y \in \mathbb{R}^n$ is **feasible** if $f_i(y) \leq 0$ for all $i \in 1, \ldots, m$.

Feasible set $\mathcal{X} = \{x \in \mathbb{R}^n \mid f_i(x) \leq 0, \forall i \in 1, \ldots, m\}$.

Point $x^* \in \mathbb{R}^n$ is **global minimum** if $f_0(x^*) \leq f_0(x)$ for all $x \in \mathcal{X}$.

If some x is the optimal solution to $\min_x f(x)$, then x is also optimal for $\max_x -f(x)$ and $\min_x \alpha f(x)$ where $\alpha > 0$.

Solution types: *Infeasible*: no input satisfies all constraints (e.g., $x > 1$ and $x < 0$). *Unbounded*: optimal objective value is $-\infty$ (e.g., minimize x without constraints). *Unattainable*: no finite solution (e.g., minimize $\frac{1}{x}$ subject to $x > 0$). *Tractable*: algorithm to solve efficiently (polynomial time).

For minimization: optimal objective is $+\infty$ if infeasible, $-\infty$ if unbounded from below, and finite otherwise (x^* may or may not be attainable). Max problems see opposite.

4.2 Coercive Functions & Finite Solutions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is **coercive** if $\lim_{\|x\|\to\infty} f(x) = \infty$. Note: $f(x)$ must tend to $+\infty$ along all directions when $\|x\| \to \infty$ to be coercive. Conversely, to prove not coercive, just need to find one direction along which $f(x)$ does not go to $+\infty$ when $\|x\| \to \infty$.

Theorem (unconstrained): Consider $f : \mathbb{R}^n \to \mathbb{R}$ with domain $\mathbb{R}^n$ (either convex or non-convex). Then if f is continuous and coercive, $\min f(x)$ has a finite solution.

Theorem (constrained): (1) Consider $\min f(x)$ subject to $x \in \mathcal{S}$. Suppose that f (convex or non-convex but with domain $\mathbb{R}^n$) is coercive and continuous. If $\mathcal{S}$ (convex or non-convex) is closed, the optimization problem has a finite solution. (2) Consider $\min f_0(x)$ subject to $f_i(x) \leq 0$ for $i = 1, \ldots, m$ and $h_j(x) = 0$ for $j = 1, \ldots, k$, where f_0, f_i's, and h_j's are arbitrary but continuous with domain $\mathbb{R}^n$. Then if f_0 is coercive, the optimization has a finite solution.

Weierstrass theorem: consider $\min f(x)$ s.t. $x \in \mathcal{S}$, where $f : \mathbb{R}^n \to \mathbb{R}$ is continuous. If $\mathcal{S}$ is compact, then the optimization has a finite solution. So for optimization of form $\min f_0(x)$ s.t. $f_i(x) \leq 0$, as long as f_0 is continuous and the feasible set is bounded, we have a finite solution.

4.3 Convex Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if and only if its domain is a convex set and $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$ for all $x, y \in \operatorname{dom} f$ and $\alpha \in [0, 1]$. This is the **zeroth-order condition** for convexity.

Geometric intuition: the graph of the function must entirely lie below the line segment that connects two arbitrary points on the graph. Replacing $\leq$ with $<$ gives **strict convexity**. Thus, the set $\{x : f(x) \leq 0\}$ is a convex set if f is a convex function.

First-order convexity condition: $f(y) + \nabla f(y)^\top (x - y) \leq f(x)$ for all $x, y \in \operatorname{dom} f$ (replace $\leq$ with $<$ for strict convexity). Geometric: graph must entirely lie above the tangent line at arbitrary point on graph.

Second-order convexity condition: f is convex if and only if $\nabla^2 f(x) \succeq 0$ for all $x \in \operatorname{dom} f$. If $\nabla^2 f(x) \succ 0$ for all $x \in \operatorname{dom} f$, then f is strictly convex. Reverse may not hold (e.g., $f(x) = x^4$). Geometric: graph must be "bowl-shaped" everywhere.

Example convex functions: $f(x) = e^{ax}$, $f(x) = x^a$ where $a \geq 1$ or $a \leq 0$ on $\mathbb{R}_{++}$, $f(x) = -\log(x)$ on $\mathbb{R}_{++}$, any ℓ_p form function $f(x) = \|x\|_p$, quadratic functions $f(x) = x^\top P x + q^\top x + r$ where P is symmetric and $P \succeq 0$. If $P \succ 0$, f is strictly convex.

A function f is called **concave** if $-f$ is convex.

Affine functions are simultaneously convex and concave.

Convexity does not imply continuity. Example: consider an end point $\bar{x}$ of $\operatorname{dom} f$. f can still be convex if it "jumps up" at $\bar{x}$. Discontinuity should happen only on the boundaries.

Operations producing convex functions: (1) Point-wise maximum of a set of convex functions is convex. Point-wise minimum of a set of concave functions is concave. (2) A summation of convex functions $f(x) := \sum_{i=1}^k \alpha_i f_i(x)$ for $\alpha_i \geq 0$ is convex if f_i is convex for all i. (3) If $f(x)$ is convex, then the affine transformation $g(x) = f(Ax + b)$ is also convex. (4) If $f(x)$ is convex and $g(x)$ is convex and non-decreasing, then

the composite function $g \circ f(x)$ is convex. (5) Compositions of convex functions are not convex in general.

4.4 Convex Optimization Problems

Consider an optimization problem $\min_x f(x)$ subject to $x \in \mathcal{X}$. This problem is **convex** when f is a convex function and $\mathcal{X}$ is a convex set.

Consider $\min_x f_0(x)$ subject to $g_i(x) \leq 0$ for all i and $h_j(x) = 0$ for all j. This problem is convex when f_0 is a convex function, g_i is a convex function for each i, h_j is an affine function for each j.

For a convex optimization problem: (1) All local solutions are global. (2) The feasible set is a convex set. (3) The set of all global minima is a convex set. (4) If the objective is strictly convex, then there is either no solution or a unique solution.

4.5 Linear Programming (LP)

An **LP** can be written as: $\min_x a_0^\top x$ subject to $Ax = b$ and $Cx \leq d$.

Rewritten in standard form: $\min_x a_0^\top x$ subject to $Ax = b$ and $x \geq 0$.

If an LP is reformulated from the form $Ax = b$ and $Cx \leq d$ into the standard form $Ax = b$ and $x \geq 0$, A and b in the standard form can be different from those in the original problem. To convert an affine inequality constraint $Cx \leq d$ into standard form, we can introduce a slack variable s (same shape as d) and rewrite the constraint as $Cx + s = d$ and $s \geq 0$. The constraint $x \geq 0$ must apply to all variables. If some variables are not constrained to be non-negative in the original problem (say x_i is one of such variables), we can "split" it into $x_{i+} \geq 0$ and $x_{i-} \geq 0$ and represent x_i as $x_{i+} - x_{i-}$.

Algorithms to solve LPs: *Simplex*: start at an arbitrary vertex and repeatedly go to a neighbor vertex with lower objective value. *Interior point*: start in the interior of polyhedron and move towards optimal solution (stays in the interior as opposed to moving on the boundary).

LP solutions at vertices: For a convex set $\mathcal{S}$, a point $y \in \mathcal{S}$ is an **extreme point** if there do not exists points $u, v \in \mathcal{S}$ such that $y = \alpha u + (1-\alpha)v$ for some $0 < \alpha < 1$. Extreme points of a polyhedron are called **vertices**.

Theorem: assume LP has a solution. Then one of its feasible set vertices is a solution (could have other solutions as well).

Theorem: if an LP's feasible set is bounded, then a solution exists.

Finding all vertices: Consider a feasible set for x defined by $Ax = b$ and $x \geq 0$, where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$. Assume that $m \leq n$, i.e., A is wide. An algorithm that finds all vertices of the feasible set is as follows. Find all possible combinations of m columns of A and denote the resulting square sub-matrices formed by these columns as $A_i^{\text{sub}} \in \mathbb{R}^{m \times m}$ for $i = 1, \ldots, \binom{n}{m}$, where $\binom{n}{m}$ denotes n-choose-m and is equal to $\frac{n!}{m!(n-m)!}$. Then, the number of vertices is the number of A_i^{sub} matrices that satisfies: (1) A_i^{sub} is invertible; (2) The solution z^* to the linear system $A_i^{\text{sub}} z = b$ is feasible (i.e., non-negative).

Converting to LP via epigraph formulation: Sometimes, can convert an optimization problem into an LP via an epigraph formulation. Start with general optimization problem where $\mathcal{S}$ is some feasible set: $\min_{x \in \mathbb{R}^n} f(x)$ subject to $x \in \mathcal{S}$. Reformulate using a slack variable t: $\min_{x \in \mathbb{R}^n, t \in \mathbb{R}} t$ subject to $x \in \mathcal{S}$, $f(x) \leq t$.

For example, $\min_{x \in \mathbb{R}^n} \|x\|_\infty$ subject to $x \in \mathcal{S}$ can be converted to $\min_{x \in \mathbb{R}^n, t \in \mathbb{R}} t$ subject to $x \in \mathcal{S}$ and $\|x\|_\infty \leq t$, where $\|x\|_\infty \leq t \iff |x_i| \leq t$ for all i.

4.6 Quadratic Programming (QP)

QP includes a quadratic term in the objective, where $P_0 \succeq 0$: $\min_x x^\top P_0 x + q_0^\top x + r_0$ subject to $Ax = b$ and $Cx \leq d$.

4.7 QCQP & Convex Relaxations

QCQP can be written in the form of $\min_x x^\top P_0 x + q_0^\top x + r_0$ subject to $Ax = b$ and $x^\top P_j x + q_j^\top x + r_j \leq 0$ for $j = 1, \ldots, k$, where $P_j \succeq 0$ for $j = 0, \ldots, k$.

Hierarchy: LP $\subseteq$ QP $\subseteq$ QCQP $\subseteq$ convex optimization.

Convex relaxations: Consider optimization problem with $f(x)$ convex but $\mathcal{S}$ non-convex: $\min_x f(x)$ subject to $x \in \mathcal{S}$. If we replace $\mathcal{S}$ with a convex $\hat{\mathcal{S}}$ such that $\mathcal{S} \subset \hat{\mathcal{S}}$, we get a *convex relaxation*: $\min_x f(x)$ subject to $x \in \hat{\mathcal{S}}$. Let x^* and $\hat{x}$ be global minima of the original and relaxed optimizations, respectively. Then $f(\hat{x}) \leq f(x^*)$. If $\hat{x} \in \mathcal{S}$, then $\hat{x}$ is a global min for original optimization problem.

4.8 Integer Programming (IP)

An **IP** is just an LP with a constraint that all elements of x are integers: $\min_x a_0^\top x$ subject to $Ax = b$, $x \geq 0$, and x_i are integers for $i = 1, \ldots, n$. IPs are non-convex!

Can form a convex relaxation by dropping the integer constraint. Let P_1 be the above IP, and let P_2 be the corresponding relaxed LP dropping the integer constraint. **Theorem**: if all vertices of the feasible set of P_2 are integral, then the convex relaxation is exact, and the optimal objectives of P_1 and P_2 are equal. This is the case for assignment / transport problems (see Lecture 19)!

5 Optimality Conditions

5.1 Gradient & Hessian

Consider a function $f(x) : \mathbb{R}^n \to \mathbb{R}$ and assume $f(x)$ is twice continuously differentiable. Let x_i denote the i-th entry of x for $i = 1, \ldots, n$.

The **gradient** is an n-dimensional vector $\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right)$.

The **Hessian** is an $n \times n$ symmetric matrix $\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}$.

If $n = 1$, then the gradient is the first-order derivative and the Hessian is the second-order derivative.

Suppose that $f(x)$ is quadratic, i.e., $f(x) = x^\top P x + q^\top x + r$ for some $P \in \mathbb{S}^n$, $q \in \mathbb{R}^n$, and $r \in \mathbb{R}$. Then, it holds that $\nabla f(x) = 2Px + q$ and $\nabla^2 f(x) = 2P$.

Gradient chain rule: Consider functions $f : \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$. Define $\phi(x) := f(g(x))$. Then $\nabla \phi(x) = \begin{bmatrix} \nabla g_1(x) & \ldots & \nabla g_m(x) \end{bmatrix} \times \nabla f(z)|_{z=g(x)}$.

Taylor series approximation: given a function $f(x) : \mathbb{R}^n \to \mathbb{R}$ that is differentiable at $x_0 \in \mathbb{R}^n$, it can be approximated by an affine function in a neighborhood of x_0: $f(x) = f(x_0) + \nabla f(x_0)^\top (x - x_0) + \epsilon(x)$, where $\epsilon(x)$ goes to zero faster than first order, i.e., $\lim_{x \to x_0} \frac{\epsilon(x)}{\|x - x_0\|} = 0$. So, to the first order we have the approximation: $f(x) \approx f(x_0) + \nabla f(x_0)^\top (x - x_0)$.

5.2 Unconstrained Optimality Conditions

Consider the optimization problem $\min_{x \in \mathbb{R}^n} f(x)$, where f is differentiable.

First-order necessary condition: If x^* is a local minimum, then $\nabla f(x^*) = 0$.

Suppose that $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^n$, i.e., the problem is convex. Then:

All local minima are global minima.

x^* is a global minimum (and a local minimum) if and only if $\nabla f(x^*) = 0$.

5.3 Slater's Condition

Slater's condition is a widely used regularity condition.

Consider a convex problem $\min_x f_0(x)$ subject to $f_i(x) = 0$ for $i = 1, \ldots, k$ and $h_j(x) \leq 0$ for $j = 1, \ldots, m$. Denote the intersection of each f_i and each h_j's domain as $\mathcal{D}$.

Slater's condition holds if there exists a point $y \in \operatorname{relint} \mathcal{D}$ such that

$f_i(y) = 0$ for $i = 1, \ldots, k$.

$h_j(y) \leq 0$ for all affine h_j.

$h_j(y) < 0$ for all non-affine f_j.

y is not unique in general.

When there are no constraints, Slater's condition holds by convention.

When all constraints are affine, e.g., LP or QP, Slater's condition is equivalent to feasibility. However, **Slater's condition is stricter than feasibility in general**.

5.4 Constrained Optimality (KKT)

Again, consider the optimization problem $\min_x f_0(x)$ subject to $f_i(x) = 0$ for $i = 1, \ldots, k$ and $h_j(x) \leq 0$ for $j = 1, \ldots, m$. Denote the dual variables associated with the equality con-

straints as $\mu_1, \ldots, \mu_k$. Similarly, denote the dual variables associated with the inequality constraints as $\lambda_1, \ldots, \lambda_m$. The Lagrangian of this problem is then $L(x, \lambda, \mu) := f(x) + \sum_{i=1}^{k} \mu_i f_i(x) + \sum_{j=1}^{m} \lambda_j h_j(x)$.

Karush–Kuhn–Tucker (KKT) conditions: Consider Lagrangian multipliers $\lambda_1^*, \ldots, \lambda_m^*$ and $\mu_1^*, \ldots, \mu_k^*$.

1. **Primal Feasibility**: $f_i(x^*) = 0$ for all $i = 1, \ldots, k$ and $h_j(x^*) \leq 0$ for all $j = 1, \ldots, m$;
2. **Dual Feasibility**: $\lambda_j^* \geq 0$ for all $j = 1, \ldots, m$;
3. **Lagrangian Stationarity**: $\nabla f(x^*) + \sum_{i=1}^{k} \mu_i^* \nabla f_i(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla h_j(x^*) = 0$;
4. **Complementary Slackness**: $\lambda_j^* \cdot h_j(x^*) = 0$ for all $j = 1, \ldots, m$.

For convex optimization problems that satisfy Slater's condition, the KKT conditions are sufficient and necessary can be used to find global optima.

Example problem: Consider a quadratic optimization with equality constraints in the form of $\min_{x \in \mathbb{R}^n} x^\top P_0 x + q_0^\top x + r_0$ subject to $Ax = b$ where $P_0 \in \mathbb{S}_{++}^n$, $A \in \mathbb{R}^{m \times n}$, $a_0 \in \mathbb{R}^n$, and $r_0 \in \mathbb{R}$. Suppose that Slater's condition holds, i.e., $Ax = b$ admits one or more solutions. Then, using the KKT condition, we can show that the optimal primal-dual solution (x^*, μ^*) satisfies $\begin{bmatrix} A & 0_{m \times m} \\ 2P_0 & A^\top \end{bmatrix} \begin{bmatrix} x^* \\ \mu^* \end{bmatrix} = \begin{bmatrix} b \\ -q_0 \end{bmatrix}$.

6 Linear Systems, LS, & Regression

6.1 Solving Linear Systems

Consider solving a system of linear equations $Ax = y$.

$Ax = y$ has a unique solution if and only if $y \in \mathcal{R}(A)$ and $\mathcal{N}(A) = \{0\}$.

If A's nullspace satisfies $\mathcal{N}(A) \neq \{0\}$, any solution x^* produces a space of solutions $x^* + z$ where $z \in \mathcal{N}(A)$.

Tall matrix: if $A \in \mathbb{R}^{m \times n}$, where $m > n$, then we have an overdetermined case, and there is likely no solution unless we are lucky and $y \in \mathbb{R}(A)$.

Fat matrix: now assume $m > n$, and our rows are linearly independent. Now we have an underdetermined case, and the solution space is $\bar{x} + \mathcal{N}(A)$ where $\bar{x}$ is an arbitrary solution. For many applications, the "best" solution is the one with minimum norm: $\min_{x \in \mathbb{R}^n} \|x\|$ subject to $Ax = y$.

The minimum-norm solution can be derived as $x^* = A^\top (AA^\top)^{-1} y = A^\dagger y$.

If A is square and full-rank (invertible), we can solve directly $x = A^{-1} y$.

6.2 Least Squares (LS)

What if we are in the overdetermined case and y is not in the range of A? We need to minimize how much we violate the equation $Ax = y$, instead of solving it exactly.

Given a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $y \in \mathbb{R}^m$, we aim to solve the problem $\min_{x \in \mathbb{R}^n} \|Ax - y\|_2$.

Denote the optimal solution as x^*. Note that x^* also solves $\min_{x \in \mathbb{R}^n} \|Ax - y\|_2^2$.

The set of solutions for the LS problem is $\mathcal{S} := \{x^* \mid A^\top A x^* = A^\top y\}$. Proof: optimality conditions.

It holds that $\mathcal{S} = A^\dagger y + \mathcal{N}(A)$, where $A^\dagger$ is the pseudo-inverse of A as defined above.

6.3 LS & Projection

Geometrically, the LS problem finds the projection of y onto $\mathcal{R}(A)$, the range of A.

The projection result $y^* = Ax^* = \Pi_{\mathcal{R}(A)} y$ exists and is unique.

Theorem on projection: $y = y^* \perp \mathcal{R}(A)$. I.e., $(y - y^*, v) = 0$ for all $v \in \mathcal{R}(A)$.

We can find y^* by solving for the vector that simultaneously satisfies $y^* \in \mathcal{R}(A)$ and $y - y^* \perp \mathcal{R}(A)$.

6.4 Minimum-Norm LS Solution

To find the minimum-norm solution, solve $\min_{x \in \mathbb{R}^n} \|x\|_2$. I.e., $\min_{x \in \mathbb{R}^n} \|x\|_2$ subject to $A^\top A x = A^\top y$.

The minimum-norm LS solution is unique and equal to $A^\dagger y = (A^\top A)^{-1} A^\top y$.

If A has full column rank, i.e., $m \geq n = \text{Rank}(A)$, then $A^\top A$ is invertible and $\mathcal{N}(A) = \{0\}$. In this case, $x^* = A^\dagger y$ is the unique LS solution.

6.5 Ridge Regression

An ℓ_2-regularized LS problem: $\min_{x \in \mathbb{R}^n} \|Ax - y\|_2^2 + \alpha \|x\|_2^2$ where α is a non-negative scalar.

The matrix $A^\top A + \alpha I_n$ is invertible, and the unique solution to the ridge regression problem is $x^* = (A^\top A + \alpha I_n)^{-1} A^\top y$.

6.6 Sparsity & LASSO Regression

$x \in \mathbb{R}^n$ is called **sparse** if many of its entries are zero. Otherwise it is called dense.

The number of non-zero entries of x is called its *cardinality*, denoted as $\|x\|_0$. When all entries of x are within $[-1, 1]$, it holds that $\|x\|_1 \leq \|x\|_0$.

LASSO is an ℓ_1-regularized LS problem that promotes solution sparsity: $\min_{x \in \mathbb{R}^n} \|Ax - y\|_2^2 + \alpha \|x\|_1$, where α is a non-negative scalar.

LASSO's objective function is not always differentiable. However, it can be reformulated as a QP via the epigraph method: $\min_{x \in \mathbb{R}^n, t \in \mathbb{R}^n} x^\top P_0 x + q_0^\top x + r_0 + \alpha \sum_{i=1}^{n} t_i$ subject to $-t_i \leq x_i \leq t_i$ for $i = 1, \ldots, n$,

where $P_0 \in \mathbb{S}_+^n$, $q_0 \in \mathbb{R}^n$, and $r_0 \in \mathbb{R}$ are expressions of A and y.

x^* is a solution to LASSO if and only if $2P_0 x^* + q_0 + \lambda^* = 0$, where each entry of λ^* satisfies: $\lambda_i^* = \alpha$ if $x_i^* > 0$, $\lambda_i^* = -\alpha$ if $x_i^* < 0$, and $\lambda_i^* \in [-\alpha, \alpha]$ if $x_i^* = 0$. Furthermore, it holds that $|x_i^*| = t_i^*$ for all i.

6.7 Sensitivity Analysis – Linear Systems

Consider system of linear equations with $A \in \mathbb{R}^{n \times n}$ invertible and $y \in \mathbb{R}^n$ given; we want to find $x : Ax = y$.

Due to invertibility solution is given by $A^{-1} y$.

What if y changes to $y + \Delta y$ due to measurement noise?

Consider solution change to $x + \Delta x$: $A(x + \Delta x) = y + \Delta y$ and $Ax = y \implies \Delta x = A^{-1} \Delta y$.

Lemma: for matrix B and vector y: $\|By\|_2 \leq \|B\|_2 \|y\|_2$.

So we have $\|\Delta x\|_2 \leq \|A^{-1}\|_2 \|\Delta y\|_2$ and $\|y\|_2 \leq \|A\|_2 \|x\|_2$.

Combining these two yields that $\frac{\|\Delta x\|_2}{\|x\|_2} \leq \|A\|_2 \|A^{-1}\|_2 \frac{\|\Delta y\|_2}{\|y\|_2}$.

Define **condition number** $\kappa(A) = \|A\|_2 \|A^{-1}\|_2$.

Theorem: the relative change in x with regard to a relative change in y, when solving $y = Ax$ for A invertible, is given by $\frac{\|\Delta x\|_2}{\|x\|_2} \leq \kappa(A) \frac{\|\Delta y\|_2}{\|y\|_2}$.

Recall that $\|A\|_2 = \sigma_1$ is the largest singular value and $\|A^{-1}\|_2 = \frac{1}{\sigma_n}$ is the largest singular value of A^{-1}.

If $\kappa(A)$ is close to 1, then A is called well conditioned; if $\kappa(A)$ is large, then A is ill conditioned.

Similar bound if we perturb A to $A + \Delta A$: $\frac{\|\Delta x\|_2}{\|x\|_2} \leq \kappa(A) \frac{\|\Delta A\|_2}{\|A\|_2}$.

6.8 Sensitivity Analysis – LS

Let's consider least-square problem $\min_x \|Ax - y\|_2$, where y is a measurement vector with noise.

How does perturbing y to $y + \Delta y$ affect solutions?

Recall we can define an ellipse in two equivalent forms: $E = \{x \in \mathbb{R}^n \mid x = By, \|y\|_2 \leq 1\}$; $E = \{x \in \mathbb{R}^n \mid x^\top P^{-1} x \leq 1\}$ where $P = BB^\top$ is PSD.

Let $v^1, \ldots, v^n$ be eigenvectors of P with associated eigenvalues $\lambda_1, \ldots, \lambda_n$. The the ellipse has semi-axes is the directions $v^1, \ldots, v^n$ with lengths $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}$.

Recall that $x^* = A^\dagger y$ solves least squares; consider $x^* + \Delta x = A^\dagger (y + \Delta y)$.

Theorem: for an uncertainty ball on the measurement $\|\Delta y\| \leq 1$, we get an ellipsoidal uncertainty set on the solution changes: $E = \{\Delta x \in \mathbb{R}^n \mid \Delta x = A^\dagger \Delta y, \|\Delta y\| \leq 1\}$.

E is an ellipse with semi-axes $v^1, \ldots, v^n$ and lengths $\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_n}, 0, \ldots, 0$ from the SVD $A = U \Sigma V^\top$ (this is because $A = U \Sigma V^\top \implies A^\dagger = V \Sigma^\dagger U^\top$).

7 Duality

7.1 Weak Duality

In the context of duality, the original problem is called the *primal problem*. We call its optimal objective $p^* := f_0(x^*)$ the *primal solution*.

Consider arbitrary $\mu \in \mathbb{R}^k$ and $\lambda \in \mathbb{R}^m$ where $\lambda \geq 0$. It holds that $\min_x L(x, \lambda, \mu) \leq p^*$.

Hence, to find a meaningful lower bound to p^*, we can solve $\max_{\mu \in \mathbb{R}^k, \lambda \in \mathbb{R}^m} \min_x L(x, \lambda, \mu)$ subject to $\lambda \geq 0$.

We define $d(\lambda, \mu) = \min_x L(x, \lambda, \mu)$ as the **dual function**. We can then reformulate the lower bound optimization problem as the maximization problem $d^* := \max_{\mu \in \mathbb{R}^k, \lambda \in \mathbb{R}^m} d(\lambda, \mu)$ subject to $\lambda \geq 0$,

which we refer to as the *dual problem*. Its optimal objective d^* is called the *dual solution*.

It holds that $d^* \leq p^*$. The value of $p^* - d^*$ is called the **duality gap**.

Since $d(\lambda, \mu)$ is a point-wise minimum of affine functions, it is concave no matter whether the primal problem is convex or not, and therefore the dual problem is always a convex optimization problem.
Hence, leveraging weak duality, we can use convex optimization to obtain a lower bound to a hard, potentially non-convex problem.

7.2 Strong Duality

If it holds that $p^* = d^*$, i.e., duality gap is zero, then *strong duality* holds.
If the primal problem is convex and Slater's condition holds, then
Strong duality holds.
The KKT conditions of the primal problem simultaneously solve the primal problem and the dual problem. I.e., x^* solves the primal problem and (λ^*, μ^*) solves the dual problem.
If x^* is an arbitrary optimal solution to the primal problem and (λ^*, μ^*) is an arbitrary optimal solution to the dual problem, then (x^*, λ^*, μ^*) satisfies the primal problem's KKT conditions.

7.3 Dual of LP and QP

The dual of an LP is also an LP. Specifically, for some $a_0 \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $C \in \mathbb{R}^{k\times n}$, and $d \in \mathbb{R}^k$, consider the LP $\min_{x\in\mathbb{R}^n} a_0^\top x$ subject to $Ax \le b$ and $Cx = d$. The dual problem is $\max_{\lambda\in\mathbb{R}^m,\mu\in\mathbb{R}^k} -\lambda^\top b - \mu^\top d$ subject to $a_0 + A^\top\lambda + C^\top\mu = 0$ and $\lambda \ge 0$.
The dual of a QP is also a QP. Specifically, for some $P_0 \in \mathbb{S}^n_{++}$, $q_0 \in \mathbb{R}^n$, $r_0 \in \mathbb{R}$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $C \in \mathbb{R}^{k\times n}$, and $d \in \mathbb{R}^k$, consider the QP $\min_{x\in\mathbb{R}^n} x^\top P_0 x + q_0^\top x + r_0$ subject to $Ax \le b$ and $Cx = d$.
The dual problem is $\max_{\lambda\in\mathbb{R}^m,\mu\in\mathbb{R}^k} -\frac{1}{4}(q_0 + A^\top\lambda + C^\top\mu)^\top P_0^{-1}(q_0 + A^\top\lambda + C^\top\mu) - \lambda^\top b - \mu^\top d$ subject to $\lambda \ge 0$.
As a special case for QP, consider the problem of finding the minimum-norm solution of a system of equations, i.e., $\min_x \|x\|_2^2$ subject to $Ax = b$ (A has full row rank). The dual problem is $\max_\mu -\frac{1}{4}\mu^\top AA^\top\mu - b^\top\mu$. By Lagrangian stationarity, $x^* = -\frac{1}{2}A^\top\mu^*$. Setting the gradient of the dual problem objective to zero gives $\mu^* = -2(AA^\top)^{-1}b$.

7.4 Farkas' Lemma

Suppose that we want to show that the set $\left\{x \in \mathbb{R}^n \mid \begin{array}{ll} f_i(x) \le 0, & i = 1,\dots,k \\ h_j(x) = 0, & j = 1,\dots,m \end{array}\right\}$ is empty.
We can consider the optimization problem $\min_x 0$ subject to $f_i(x) \le 0$ for all i and $h_j(x) = 0$ for all j. Next, we find the dual function of this optimization problem $d(\lambda, \bar\mu)$. Suppose that we can find some $(\hat\lambda, \hat\mu)$ such that $d(\hat\lambda, \hat\mu) > 0$, then the optimal objective of the primal problem is $+\infty$, and hence the set of interest is empty.
For linear case we have **Farkas' Lemma** as following. Equations $Ax = b$ and $x \ge 0$ have no solutions if and only if there is a solution μ to $A^\top\mu \le 0$ and $b^\top\mu < 0$.

7.5 Constraint Sensitivity Analysis

We are interested in comparing the optimization problem $\min f_0(x)$ subject to $f_i(x) \le 0$ for $i = 1,\dots,k$ and $h_j(x) = 0$ for $j = 1,\dots,m$
with the problem that has perturbed constraints $\min f_0(x)$ subject to $f_i(x) \le v_i$ for $i = 1,\dots,k$ and $h_j(x) = w_j$ for $j = 1,\dots,m$,
where each v_i and w_j is some scalar.
Denote the optimal objective value of the perturbed problem as $p^*(v, w)$. The optimal objective of the original problem is $p^*(0, 0)$. If the problem is infeasible for some (v, w), then $p^*(v, w) = +\infty$.
We then have the following properties.
$p^*(v, w)$ is a convex function of v and w.
Assume Slater's condition holds. If $p^*(v, w)$ is differentiable at $(0, 0)$, then the Lagrangian multipliers (λ^*, μ^*) of the original problem satisfies $\lambda_i^* = -\frac{\partial p^*(0,0)}{\partial v_i}$ for all i and $\mu_j^* = -\frac{\partial p^*(0,0)}{\partial w_j}$ for all j.
As a result, it holds that $p^*(v, w) \approx p^*(0, 0) - \sum_i \lambda_i^* v_i - \sum_j \mu_j^* w_j$.
This is the first-order Taylor's approximation for $p^*(v, w)$. Given x^* (which can be used to compute $p^*(0, 0)$), λ^* and μ^* of the original unperturbed problem, this approximation can be computed efficiently.
If $\lambda_i^* = 0$ for some i or $\mu_j^* = 0$ for some j, then changing the corresponding constraint a little does not affect the optimal objective. Hence, those constraints can be eliminated.
If λ_i^* or μ_j^* is small, then the optimization problem is not sensitive to the associated constraints.
If λ_i^* or μ_j^* is large, then the optimization problem is highly sensitive to the associated constraints.

8 Numerical Algorithms & Applications

8.1 Gradient & Newton's Methods

The gradient method is a first-order method, whereas Newton's method is second-order. They apply to uni-variate and multi-variate optimization problems. Specifically, consider the problem $\min_{x\in\mathbb{R}^n} f(x)$.
Descent algorithm: An iterative algorithm that generates a sequence $x^{(0)}, x^{(1)}, x^{(2)}, \dots$ in a way that $f(x^{(k+1)}) < f(x^{(k)})$ for $k = 0, 1, 2, \dots$
Descent direction: At a point $x \in \mathbb{R}^n$, Δx is a descent direction if $\nabla f(x)^\top \Delta x < 0$.
Using descent directions guarantees that $f(x^{(k-1)}) < f(x^{(k)})$ for all small enough step sizes $s^{(k)}$.
A family of optimization algorithms can be designed with descent directions: starting from $x^{(0)}$ as the initial guess, the kth iteration is $x^{(k+1)} \leftarrow x^{(k)} - s^{(k)}\Delta x^{(k)}$ (this is called the update rule), where $\Delta x^{(k)}$ is a descent direction w.r.t. $x^{(k)}$, and $s^{(k)}$ is the step size for the kth iteration.
Gradient method: $x^{(k+1)} \leftarrow x^{(k)} - s^{(k)}\nabla f(x^{(k)})$. Here, we use $-\nabla f(x^{(k)})$, which is a descent direction when $\nabla f(x^{(k)}) \neq 0$, as $\Delta x^{(k)}$.
Newton's method: $x^{(k+1)} \leftarrow x^{(k)} - s^{(k)}(\nabla^2 f(x^{(k)}))^{-1}\nabla f(x^{(k)})$. Here, we use $-(\nabla^2 f(x^{(k)}))^{-1}\nabla f(x^{(k)})$, which is another descent direction when $\nabla^2 f(x^{(k)}) \succ 0$, as $\Delta x^{(k)}$.
If $\nabla f(x^{(k)})$ is zero, then $x^{(k)}$ is a stationary point and we stop the algorithm.
Why gradient/Newton? The gradient direction minimizes a local first-order Taylor approximation of the objective function. Similarly, the Newton direction minimizes a second-order approximation, and therefore Newton's method can solve certain quadratic problems in one iteration with $s^{(k)} = 1$.
Newton's method converges faster than the gradient method, but each iteration takes longer.
For an iterative optimization algorithm with step sizes $s^{(0)}, s^{(1)}, \dots$, if $\|x^{(k)} - x^*\|$ is greater than some positive threshold at some k, the algorithm terminates and we accept $x^{(k)}$ as a solution. However, since the true x^* is unknown, we need to estimate $\|x^{(k)} - x^*\|$.

8.2 Analysis on Gradient Algorithm

Given an initial guess $x^{(0)}$, define the set $\mathcal{S} := \{x \in \mathbb{R}^n | f(x) \le f(x^{(0)})\}$. It is said that ∇f is Lipschitz continuous with constant $L > 0$ if $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$ for all $x, y \in \mathcal{S}$. If f is twice continuous differentiable and $\mathcal{S}$ is compact, then L exists.
Suppose that L exists. For the gradient algorithm, consider an arbitrary $\epsilon > 0$. If the step size $s^{(0)}, s^{(1)}, \dots$ are chosen in the interval $(\frac{\epsilon}{L}, \frac{2-\epsilon}{L})$, then $\|\nabla f(x^{(k)})\| \to 0$ as $k \to \infty$.
This means that the gradient algorithm converges to a stationary point (which can be a local minimum, a local maximum, or a saddle point).
If f is convex, $\nabla f(x^*) = 0$ iff x^* is the global minimum. Hence, gradient algorithm always converges to a global minimum of a convex function if $s^{(k)}$ is small for all k.

8.3 Low-Rank Matrix Approximation

Given a matrix $A \in \mathbb{R}^{m\times n}$, consider the problem of finding a low-rank matrix $B \in \mathbb{R}^{m\times n}$ that best approximates A.
This problem can be formulated as $\min_{B\in\mathbb{R}^{m\times n}} \|A - B\|_2$ or $\|A - B\|_F$ subject to $\mathrm{Rank}(B) \le k$.
Eckart-Young-Mirsky theorem:
For a given $k \le \min(m, n)$, define $A_k := \sum_{i=1}^k \sigma_i u^{(i)} v^{(i)\top}$ constructed with the top k singular values of A and the left/right singular vectors. A_k has rank at most k. Intuitively, we "chop off" the smaller singular values starting from the $k+1$-th largest.
$B = A_k$ is an optimal solution to both optimization problems (Frobenius or ℓ_2-induced norm).
Suppose $k < \mathrm{Rank}(A)$. The optimal solution is unique if and

only if $\sigma_k \neq \sigma_{k+1}$, i.e., the k-th largest singular value of A is not equal to the $k+1$.

The relative Frobenius norm approximation error $\frac{\|A-A_k\|_F}{\|A\|_F^2}$ is equal to $\frac{\sigma_{k+1}^2+\cdots+\sigma_r^2}{\sigma_1^2+\cdots+\sigma_r^2}$, where $r = \text{Rank}(A)$.

The relative ℓ_2-induced norm approximation error $\frac{\|A-A_k\|_2}{\|A\|_2}$ is equal to $\frac{\sigma_{k+1}}{\sigma_1}$.

8.4 Principal Component Analysis (PCA)

Given points $x^1, \ldots, x^m \in \mathbb{R}^n$, first center data points to $\bar{x}^1, \ldots, \bar{x}^m$ by subtracting $\frac{1}{m}\sum_{i=1}^m x^i$.

Compute the left singular vectors $v^1, \ldots, v^m$.

Most variation is along v^1 (explains $\sigma_1^2/\sum_i \sigma_i^2$), second most along v^2, etc.

8.5 Robust PCA

We aim to decompose $Y \in \mathbb{R}^{m\times n}$ as the sum of a low-rank matrix $X \in \mathbb{R}^{m\times n}$ and a sparse (most entries are zero) matrix $Z \in \mathbb{R}^{m\times n}$. To achieve this, we can solve the optimization problem $\min_{X\in\mathbb{R}^{m\times n}, Z\in\mathbb{R}^{m\times n}} \text{Rank}(X) + \lambda\,\text{Card}(Z)$ subject to $Y = X + Z$, where $\text{Card}(Z)$ is the number of non-zero entries in Z and $\lambda > 0$ is a regularization coefficient.

The above problem is non-convex. To this end, we can solve the following convex problem as a surrogate: $\min_{X\in\mathbb{R}^{m\times n}, Z\in\mathbb{R}^{m\times n}} \|X\|_* + \lambda\sum_{i=1}^m\sum_{j=1}^n |Z_{ij}|$ subject to $Y = X + Z$.

8.6 Matrix Completion

Consider a matrix $X^* = \mathbb{R}^{m\times n}$ whose entries are unknown but is known to be low rank. Assume that we measure the entries X_{ij}^* only when (i,j) belongs to some given set $\mathcal{S}$.

To estimate X^* using the measurements, we can find the lowest-rank X whose (i,j) entries match the measurements by solving for the optimization problem $\min_{X\in\mathbb{R}^{m\times n}} \text{Rank}(X)$ subject to $X_{ij} = X_{ij}^*, \forall (i,j) \in \mathcal{S}$.

This problem is non-convex due to the discrete rank function in the objective. Over the restricted space $\{X \in \mathbb{R}^{m\times n} | \|X\|_2 \leq 1\}$, a convex relaxation is $\min_{X\in\mathbb{R}^{m\times n}} \|X\|_*$ subject to $X_{ij} = X_{ij}^*, \forall (i,j) \in \mathcal{S}$.

8.7 Compressed Sensing

Let $x^* \in \mathbb{R}^n$ denote some states of some system. We want to know x^* but can only measure $b := Ax^* \in \mathbb{R}^m$ for some $m \times n$ matrix A. When $m < n$, the linear system is underdetermined.

Suppose that x^* is known to be sparse. Then x^* can be estimated via the optimization problem $\min_x \|x\|_0$ subject to $Ax = b$.

This problem is non-convex, but can be approximated with its convex relaxation over the restricted space of $-1 \leq x \leq 1$: $\min_x \|x\|_1$ subject to $Ax = b$,

which can be reformulated as an LP $\min_{x\in\mathbb{R}^n, t\in\mathbb{R}^n} \mathbb{1}_n^\top t$ subject to $Ax = b$ and $-t \leq x \leq t$, where $\mathbb{1}_n$ denotes the n-dimensional all-one column vector.

Suppose that our measurements are noisy, i.e., $b = Ax + w$ where w is random (often Gaussian). Then, the problem we should solve is $\min_w \|w\|_2^2 + \lambda\|x\|_1$ subject to $Ax + w = b$, where $\lambda > 0$ is a user-defined balancing constant. This is a constrained LASSO problem that can be reformulated as a QP (see Section 6.6).