EECS 127/227A – Final Review Summary

This note was summarized by Yatong Bai and Sam Pfrommer from UC Berkeley's Fall 2024 EECS 127/227A lecture notes created by Somayeh Sojoudi.

1 Linear Algebra – Vectors

1.1 Vector Space, Subspace, Affine Set, and Basis

- $\mathbb{R}^n$ is the space of vectors with n elements.
- Vectors $x^{(1)}, \ldots, x^m \in \mathbb{R}^n$ are linearly dependent if there is a non-trivial linear combination $\sum_i \alpha_i x^{(i)}$ which is the zero vector. Otherwise, they are linearly independent.
- A non-empty set $\mathcal{S} \subseteq \mathbb{R}^n$ is a subspace if for all $x, y \in \mathcal{S}$ and scalars α, β we have $\alpha x + \beta y \in \mathcal{S}$.
- For m vectors $x^{(1)}, \ldots, x^m \in \mathbb{R}^n$, we define span$(x^{(1)}, \ldots, x^m)$ as the set of all linear combinations of $x^{(1)}, \ldots, x^m$. This set is a subspace.
- A set of vectors $x^{(1)}, \ldots, x^{(d)}$ is a basis for a subspace $\mathcal{S}$ if
  - $x^{(1)}, \ldots, x^{(d)}$ are linearly independent;
  - For all $x \in \mathcal{S}$, there exist scalars $\alpha_1, \ldots, \alpha_d$ such that $x = \sum_i \alpha_i x^{(i)}$.
- For a subspace $\mathcal{S}$, the basis is not unique, but all bases have the same number of vectors, d. This number d is the dimension of the subspace $\mathcal{S}$.
- A set $\mathcal{X} \subseteq \mathbb{R}^n$ is affine if there is a subspace $\mathcal{S} \subseteq \mathbb{R}^n$ and a vector $x^{(0)} \in \mathbb{R}^n$ such that $\mathcal{X} = x^{(0)} + \mathcal{S}$ (adding $x^{(0)}$ to all vectors in $\mathcal{S}$). To prove a set to be affine, first find $x^{(0)}$ and then show that $\mathcal{X} - x^{(0)}$ is a subspace.

1.2 Inner Product and Orthogonal Vectors

- For a pair of vectors $x, y \in \mathbb{R}^n$, the standard inner product (dot product) is $\langle x, y \rangle = x^\top y = y^\top x = x_1 y_1 + \cdots + x_n y_n$.
- It holds that $\langle x, y \rangle = \|x\|_2 \|y\|_2 \cos(\theta)$, where θ is the angle between x and y.
- Two vectors are orthogonal if $\langle x, y \rangle = 0$ (denoted $x \perp y$).
- d vectors $x^{(1)}, \ldots, x^{(d)}$ are mutually orthogonal if $x^{(i)} \perp x^{(j)}$ for all $i \neq j$. This guarantees that $x^{(1)}, \ldots, x^{(d)}$ are linearly independent.
- We say $x^{(1)}, \ldots, x^{(d)}$ are orthonormal if they are mutually orthogonal and have norm one. I.e., $\|x^{(i)}\|_2^2 = \langle x^{(i)}, x^{(i)} \rangle = 1$ for all $i = 1, \ldots, d$ and $\langle x^{(i)}, x^{(j)} \rangle = 0$ for all $i \neq j$.

1.3 Vector Norms

- A function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is a norm if
  1. $\|x\| \geq 0$ for all $x \in \mathbb{R}^n$ and $\|x\| = 0$ if and only if $x = 0$;
  2. $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$;
  3. $\|\alpha x\| = |\alpha| \|x\|$ for all $\alpha \in \mathbb{R}, \forall x \in \mathbb{R}^n$.
- An ℓ_p norm, for $1 \leq p < \infty$, is of the form $\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$.
- We define $\|x\|_0$ to be the number of non-zero elements in x. This is not a true norm but appears frequently.
- For an arbitrary vector $x \in \mathbb{R}^n$, it holds that $\|x\|_2^2 = x^\top x$.

1.4 Linear Functions and Affine Functions

- A function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is linear if $f(ax + by) = af(x) + bf(y)$ for all $x, y \in \mathbb{R}^n$ and scalars a, b.
- If $f(x)$ is linear, there exists an $a \in \mathbb{R}^n$ such that $f(x) = a^\top x$.
- A function $f(x)$ is affine if $f(x) - f(0)$ is linear. This means $f(x) = a^\top x + b$ for some $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

1.5 Hyperplanes

- A hyperplane in $\mathbb{R}^n$ is a $(n-1)$ 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 called the normal vector of the hyperplane. I.e., for any two vectors $z^1, z^2 \in H$, we have that $a \perp (z^1 - z^2)$.
- Hyperplanes divide $\mathbb{R}^n$ into half spaces $H_- = \{x \mid a^\top x \leq b\}$ and $H_+ = \{x \mid a^\top x \geq b\}$.

1.6 Projections

- Let $\mathcal{S}$ be a subspace of a space $\mathcal{X}$. The projection of a point $x \in \mathcal{X}$ onto $\mathcal{S}$ is $\Pi_{\mathcal{S}}(x) = \arg\min_{y \in \mathcal{S}} \|y - x\|$.
- The minimizer $y^\star = \Pi_{\mathcal{S}}(x)$ exists and is unique. Furthermore, $y^\star = \Pi_{\mathcal{S}}(x)$ if and only if $(x - y^\star) \perp \mathcal{S}$.
  I.e., $(x - y^\star)$ is orthogonal to every vector in $\mathcal{S}$.
- For projection onto an affine space, this condition becomes $(x - y^\star) \perp (y - y^\star)$ for all $y \in \mathcal{S}$.
  - Suppose that $y^{(1)}, \ldots, y^{(d)}$ form a basis for the affine space $\mathcal{S}$.
    We can find $y^\star$ by solving for the set of equations $y^\star \in \mathcal{S}$ and $y - y^\star \perp y^{(i)}$ for $i = 1, \ldots, d$.
- For projection onto a 1-dimensional subspace $\mathcal{S} = \text{span}(v)$, we have the formula $\Pi_{\mathcal{S}}(x) = \frac{\langle x, v \rangle}{\|v\|^2} v$.
- Now generalize projection onto a subspace $\mathcal{S} = \text{span}(x^{(1)}, \ldots, x^{(d)})$, where $x^{(1)}, \ldots, x^{(d)}$ are an orthonormal basis: $\Pi_{\mathcal{S}}(x) = \sum_i \langle x, x^{(i)} \rangle x^{(i)}$.

2 Linear Algebra – Matrices

2.1 Range, Nullspace, and Rank

- The range of A is the set of all linear combinations of A's columns: $\mathcal{R}(A) = \{Ax \mid x \in \mathbb{R}^n\}$.
- $\mathcal{R}(A)$ is a subspace, and its dimension is $\text{Rank}(A)$, which is equal to the number of linearly independent columns of A, and equal to the number of linearly independent rows.
- The nullspace of A is $\mathcal{N}(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}$.
- The nullspace is also a subspace. The fundamental theorem of linear algebra relates the null space and the range:
  1. $\mathcal{N}(A) \perp \mathcal{R}(A^\top)$;
  2. $\mathcal{N}(A) \oplus \mathcal{R}(A^\top) = \mathbb{R}^n$, where $\oplus$ denotes "direct sum". I.e., any vector in $\mathbb{R}^n$ can be decomposed into a sum of a vector from the null space of A and a vector from the column space of $A^\top$;
  3. $\dim(\mathcal{N}(A)) + \text{Rank}(A) = n$.

2.2 Eigenvalues and Eigenvectors

- Consider a square matrix $A \in \mathbb{R}^{n \times n}$. If there exists a scalar λ and a vector v such that $Av = \lambda v$, then we say that λ is an eigenvalue of A and v is the corresponding eigenvector.
- To find the eigenvalues of A, we solve for λ that makes $\det(A - \lambda I) = 0$.
  Then, for each eigenvalue λ_i, we can solve $Av^{(i)} = \lambda_i v^{(i)}$ to find the corresponding eigenvector $v^{(i)}$.

- If A is rank-deficient (not full rank, i.e., there are linear dependent rows/columns), then its determinant is 0 and at least one of its eigenvalues is 0.
- $AA^\top$ and $A^\top A$ share the same non-zero eigenvalues.
- A's trace (sum of the diagonal entries) is equal to the sum of its eigenvalues.

2.3 Symmetric Matrices and Positive/Negative (Semi)Definite Matrices

- A square matrix $A \in \mathbb{R}^{n\times n}$ is symmetric if $A = A^\top$. We denote the set of all $n \times n$ symmetric matrices as $\mathbb{S}^n$.
- The eigenvalues of a symmetric matrix are all real.
- A symmetric matrix $A \in \mathbb{S}^n$ is positive semidefinite (PSD, denoted as $A \succeq 0$ or $A \succcurlyeq 0$) if all eigenvalues are non-negative, i.e., $\lambda_1(A), \ldots, \lambda_n(A) \geq 0$. We denote the set of $n \times n$ PSD matrices as $\mathbb{S}^n_+$.
- An alternative PSD definition: A matrix $A \in \mathbb{S}^n$ is PSD if the scalar $x^\top Ax$ is non-negative for all $x \in \mathbb{R}^n$.
- Note: showing that all elements of a matrix are non-negative does NOT prove PSD.
- A symmetric matrix $A \in \mathbb{S}^n$ is positive definite (PD, denoted as $A \succ 0$) if all eigenvalues are strictly positive, i.e., $\lambda_1(A), \ldots, \lambda_n(A) > 0$. We denote the set of $n \times n$ PD matrices as $\mathbb{S}^n_{++}$.
  Alternatively, A is PD if $x^\top Ax > 0$ for all $x \neq 0$.
- An easier way to check whether a matrix is PD without calculating eigenvalues: A symmetric matrix A is PD if and only if all of its leading principal minors are strictly positive.
- A symmetric matrix A is negative semidefinite (NSD) if $\lambda_1(A), \ldots, \lambda_n(A) \leq 0$ or $x^\top Ax \leq 0$ for all $x \in \mathbb{R}^n$.
- A symmetric matrix 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 and all ND matrices are NSD.
- A matrix neither PSD nor NSD is called sign indefinite. It has at least one positive and one negative eigenvalue.

2.4 Orthogonal Matrices

- A square matrix $U \in \mathbb{R}^{n\times n}$ with columns $u^{(1)}, \ldots, u^{(n)}$ is called orthogonal if its columns are orthonormal to each other. I.e., the columns are mutually orthogonal and have norm 1.
  I.e., for arbitrary pairs of $i, j \in \{1, \ldots, n\}$, we have $\langle u^{(i)}, u^{(j)}\rangle$ is 1 if $i = j$ and 0 if $i \neq j$.
- A matrix U is orthogonal if and only if $U^\top U = I_n$, where I_n denotes the $n\times n$ identity matrix. I.e., $U^\top = U^{-1}$.
- An identity matrix is orthogonal. It is also diagonal and full-rank.

2.5 Eigenvalue Decomposition and Spectral Theorem

- Consider $A \in \mathbb{R}^{m\times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$. Let $u^{(1)}, \ldots, u^{(n)}$ be arbitrary eigenvectors each associated with one eigenvalue.
- Assume $u^{(1)}, \ldots, u^{(n)}$ are linearly independent. Then, A can be decomposed as $U\Lambda U^{-1}$,
  where $U = \begin{bmatrix} u^{(1)} & \ldots & u^{(n)} \end{bmatrix}$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. We say A is a diagonalizable matrix.
- If $\lambda_1, \ldots, \lambda_n$ are all distinct, A is always diagonalizable. If A has repeated eigenvalues, Theorem 3.4 of our textbook *Optimization Model. G.C. Calafiore and L. El Ghaoui* explains when linearly independent eigenvectors exist.
- Spectral theorem: Consider a *symmetric* matrix $A \in \mathbb{S}^n$. For each eigenvalue λ_i, select an eigenvector $u^{(i)}$ with length 1 to assemble the matrix U. Then, it holds that $A = U\Lambda U^\top$, i.e., U is an orthogonal matrix.
- Symmetric matrices are always diagonalizable.

2.6 Singular Value Decomposition (SVD)

- SVD Theorem: Given an arbitrary (not necessarily square) matrix $A \in \mathbb{R}^{m \times n}$, there exist 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:

  1. $A = U\Sigma V^\top$.
  2. U and V are each orthogonal matrices, i.e., $U^\top U = I_m$ and $V^\top V = I_n$.
  3. Σ is a "rectangular diagonal matrix" in the form of $\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m & 0 & \cdots & 0 \end{bmatrix}$ if $n \geq m$ and in the form of $\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$ if $n \leq m$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$.

- $\sigma_1, \sigma_2, \ldots$ are called the singular values of A.
- Let r be the number of non-zero singular values of A, i.e., $\underbrace{\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r}_{\text{non-zero}} > \sigma_{r+1} = \sigma_{r+2} = \ldots = 0$. It holds that $r = \text{Rank}(A)$.
- If A is symmetric and PSD, then its eigenvalues and singular values are the same, and its eigenvalue decomposition $A = U\Lambda U^\top$ is a valid SVD. However, eigenvalues and singular values are different in general.
- Finding SVD by hand:
  The non-zero singular values of A are the square root of the non-zero eigenvalues of $AA^\top$ or $A^\top A$.
  The columns of U (called the left singular vectors) are the eigenvectors of $AA^\top$.
  The columns of V (called the right singular vectors) are the eigenvectors of $A^\top A$.
- If αA, where α is some non-negative real scalar, is an orthogonal matrix, then one possible SVD for A is $A = I_n \frac{I_n}{\alpha}(\alpha A)$.

2.7 Matrix Pseudo-Inverse

- The pseudo-inverse (or Moore-Penrose inverse) of a matrix $A = U\Sigma V^\top$ is $A^\dagger = V\Sigma^\dagger U^\top$, where

$$\Sigma^\dagger = \begin{bmatrix} 1/\sigma_1 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_r & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} \text{ if } n \leq m \text{ and } \Sigma^\dagger = \begin{bmatrix} 1/\sigma_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_r & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \text{ if } n \geq m,$$

  i.e., we take the inverse of the positive singular values and fill the rest with zero.
- If A is invertible, then $A^\dagger = A^{-1}$ and therefore $AA^\dagger = I_n$. 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

Consider a matrix $A \in \mathbb{R}^{m \times n}$.

- Frobenius norm: $\|A\|_F \coloneqq \|\text{vec}(A)\|_2$, where the vector $\text{vec}(A) \in \mathbb{R}^{mn}$ is a concatenation of all columns of A. I.e., `A_frob = (A ** 2).sum().sqrt()` with Python-like pseudo code.
- It holds that $\|A\|_F^2$ is equal to the sum of the squared singular values of A, i.e., $\|A\|_F^2 = \sum_{i=1}^{r} \sigma_i^2(A)$.
- ℓ_p-induced norm: $\|A\|_p \coloneqq \max_{z \in \mathbb{R}^n, z \neq 0} \frac{\|Az\|_p}{\|z\|_p} = \max_{\|w\|_p = 1} \|Aw\|_p$.
- One example of the ℓ_p-induced norm is the *spectral norm* for $p = 2$.
  - It holds that $\|A\|_2 = \sigma_1(A) = \sqrt{\lambda_{\max}(A^\top A)}$, where $\sigma_1(A)$ is the largest singular value of A and $\lambda_{\max}(A^\top A)$ is the largest eigenvalue of $A^\top A$.

3 Set Theory

3.1 Basic Set Theory

- A set $\mathcal{S} \subseteq \mathbb{R}^n$ is open if for every $x \in \mathcal{S}$, there exists $\epsilon > 0$ such that $B_\epsilon(x) \subset S$, where $B_\epsilon(x)$ is a ball centered at x with radius ϵ.
- A set $\mathcal{S} \subseteq \mathbb{R}^n$ is closed if its complement $\mathbb{R}^n \setminus S$ is open.
- A set $\mathcal{S} \subseteq \mathbb{R}^n$ is bounded if there exists an $r > 0$ such that $S \subseteq B_r(0)$.
- A set is compact if it is closed and bounded.
- Given a set $\mathcal{S} \subseteq \mathbb{R}^n$, a point $x \in \mathcal{S}$ is said to belong to the *interior* of $\mathcal{S}$ is we can draw a ball in $\mathbb{R}^n$ centered at x and of non-zero radius that belongs to $\mathcal{S}$. The interior is denoted as $\text{int}\, \mathcal{S}$.
- The closure of a set is defined as:

$$\text{cls}(\mathcal{S}) = \left\{ z \in \mathbb{R}^n \;\middle|\; z = \lim_{k \to \infty} x^{(k)} \text{ where } x^{(k)} \in \mathcal{S},\ \forall k \right\}$$

- Boundary of a set is $\partial \mathcal{S} = \text{cls}(\mathcal{S}) \setminus \text{int}(\mathcal{S})$.

3.2 Affine Sets, Convex Sets, Dimension, and Relative Interior

- Affine combination of $x_1, \ldots, x_k \in \mathbb{R}^n$: $\left\{ \sum_{i=1}^k \alpha_i x_i \;\middle|\; \sum_{i=1}^k \alpha_i = 1 \right\}$.
- Convex combination $x_1, \ldots, x_k \in \mathbb{R}^n$: $\left\{ \sum_{i=1}^k \alpha_i x_i \;\middle|\; \sum_{i=1}^k \alpha_i = 1,\ \alpha_i \geq 0,\ \forall i \right\}$.
- A set $\mathcal{S}$ is affine if for $x, y \in \mathcal{S}$ and $t \in \mathbb{R}$, the affine combination $tx + (1-t)y$ is in $\mathcal{S}$ (also recall that Section 1.1 introduced affine sets based on subspaces).
  - A hyperplane is an affine set, but a half-space is not.
- A set $\mathcal{S}$ 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 $\left\{ x \;\middle|\; a_i^\top x \leq b_i,\ c_j^\top x = d,\ \forall i, j \right\}$ is a convex set.
  - Norm balls and half-spaces are convex.
  - The set of PD matrices is convex, and the set of PSD matrices is also convex.
- The affine hull of a set is the smallest affine set containing the set. It is the set of affine combinations of any k points in the set.

- The convex hull of a set is the smallest convex set containing the set. It is the set of convex combinations of any k points in the set.
- Operations that preserve set convexity:
  - The intersection of convex sets is convex (note that the union of convex sets may not be convex).
  - Affine transformation: consider a convex set $\mathcal{S}$ and an affine function $f : \mathbb{R}^n \to \mathbb{R}^m$. The set $\bar{\mathcal{S}} := \{f(x) : x \in \mathcal{S}\}$ is convex.
  - Projections of a convex set are convex.
- Dimension of a set $S \subseteq \mathbb{R}^n$:
  1. If S is a subspace, dimension is minimum number of spanning vectors.
  2. If S is affine, then $S = x_0 + V$, where V is a subspace, and the dimension of S is the dimension of V.
  3. If S is convex, dimension of S is defined as the dimension of the affine hull of S.
- Given a convex set $\mathcal{S} \subseteq \mathbb{R}^n$, a point $x \in \mathcal{S}$ is said to belong to the *relative interior* of $\mathcal{S}$ is we can draw a ball in the affine hull of $\mathcal{S}$ centered at x and of non-zero radius that belongs to $\mathcal{S}$. The relative interior is denoted as relint $\mathcal{S}$.

3.3 Separating Hyperplane

- Supporting hyperplane theorem: for a convex set C and a boundary point $z \in \partial C$, we can always find a *supporting hyperplane* $H = \{x \in \mathbb{R}^n \mid a^\top x = b\}$ which is defined as satisfying:
  1. $z \in H$
  2. $C \subseteq H_-$, where $H_- = \{x \in \mathbb{R}^n \mid a^\top x \leq b\}$
- Separating hyperplane: for two convex sets $C_1, C_2 \subseteq \mathbb{R}^n$, a hyperplane $H = \{x \in \mathbb{R}^n \mid a^\top x = b\}$ is said to separate 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. Then we have 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 and Basic Properties

- Consider functions $f_i : \mathbb{R}^n \mapsto \mathbb{R}$ for $i = 0, \ldots, n$. The standard form of optimization problems is
$$\min_{x \in \mathbb{R}^n} \quad f_0(x) \quad \text{subject to} \quad f_i(x) \leq 0, \ \ \forall i = 1, \ldots, m. \tag{1}$$
- Equality constraints can be converted into inequality constraints.
For some function $h : \mathbb{R}^n \mapsto \mathbb{R}$, it holds that $h(x) = 0 \iff h(x) \leq 0$ and $-h(x) \leq 0$.
- Consider the optimization problem (1). A point $y \in \mathbb{R}^n$ is called feasible if $f_i(y) \leq 0$ for all $i \in 1, \ldots, m$.
Furthermore, the feasible set $\mathcal{X}$ is the set of all feasible points: $\mathcal{X} = \{x \in \mathbb{R}^n \mid f_i(x) \leq 0, \ \forall i \in 1, \ldots, m\}$.
- A point $x^\star \in \mathbb{R}^n$ is a global minimum if $f_0(x^\star) \leq f_0(x)$ for all $x \in \mathcal{X}$.
- Suppose that some x is the optimal solution to $\min_x f(x)$ where $f(x)$ is an arbitrary function, then it is also optimal for $\max_x -f(x)$ and $\min_x \alpha f(x)$, where $\alpha > 0$ is any positive scalar.

4.2 Optimization Problem Solution Types

- Infeasible: There is no input that satisfies all the constraints. E.g., we have constraints $x > 1$ and $x < 0$.
- Unbounded: The optimal objective value of the minimization problem is $-\infty$. E.g., minimize x without constraints.
- Unattainable: There is no finite solution. E.g., minimize $\frac{1}{x}$ subject to $x > 0$ (we can always improve the solution by increasing x).
- Tractable: There is an algorithm to solve it efficiently (polynomial time). Otherwise, the problem is intractable.
- For minimization problems, the optimal objective value is $+\infty$ if infeasible, $-\infty$ if unbounded from below, and finite otherwise ($x^\star$ may or may not be attainable). Maximization problems see the opposite.

4.3 Coercive Functions and Finite Solutions

- Optimization problems might not have a finite solution (e.g., $\min_x e^x$).
- Need a definition: a function $f : \mathbb{R}^n \to \mathbb{R}$ is *coercive* if $\lim_{\|x\|\to\infty} f(x) = \infty$.
  - Note that $f(x)$ must tend to $+\infty$ along all directions when $\|x\| \to \infty$ to be coercive.
  - Conversely, to prove a function is not coercive, we 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 S (convex or non-convex) is closed, the optimization problem has a finite solution.
  2. Consider the optimization problem

$$\min \quad f_0(x) \quad \text{subject to} \quad \begin{aligned} f_i(x) &\le 0 \qquad i = 1, \dots, m \\ h_j(x) &= 0 \qquad j = 1, \dots, k, \end{aligned} \tag{2}$$

     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.
- What if we don't have coerciveness? Then there is an alternative theorem.
- Weierstrass theorem: consider $\min f(x)$ s.t. $x \in \mathcal{S}$, where $f : \mathbb{R}^n \to \mathbb{R}$ is continuous. If S is compact, then the optimization has a finite solution.
  - So for an optimization of the form (2), as long as f_0 is continuous and the feasible set is bounded, we have a finite solution.

4.4 Convex Functions

- A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if its domain is a convex set and $f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha) f(y)$ for all $x, y \in \text{dom} f$ 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 $\le$ with $<$ gives the definition of strict convexity.

  - Thus, the set defined by $\{x : f(x) \leq 0\}$, where f is a convex function, is a convex set.

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

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

- Some 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 $\tilde{x}$ of $\text{dom} f$. f can still be convex if it "jumps up" at $\tilde{x}$.
  - The discontinuity should happen only on the boundaries.

- Operations that produce convex functions:
  - The point-wise maximum of a set of convex functions is convex.
    The point-wise minimum of a set of concave functions is concave.
  - A summation of convex functions $f(x) := \sum_{i=1}^{k} \alpha_i f_i(x)$ for $a_i \geq 0$ is convex if f_i is convex for all i.
  - If $f(x)$ is convex, then the affine transformation $g(x) = f(Ax + b)$ is also convex.
  - If $f(x)$ is convex and $g(x)$ is convex and non-decreasing, then the composite function $g \circ f(x)$ is convex.
  - **Compositions of convex functions are not convex in general.**

4.5 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 (see Section 4.4);
  - $\mathcal{X}$ is a convex set (see Section 3.2).

- Consider an optimization problem $\min_x f(x)$ subject to $g_i(x) \leq 0$ for all i and $h_j(x) = 0$ for all j.
  This problem is convex when
  - f 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:

- All local solutions are global.
- The feasible set is a convex set.
- The set of all global minima is a convex set.
- If the objective is strictly convex, then there is either no solution or a unique solution.

4.6 Linear Programming (LP)

- Linear program can be written as:

$$\min_x \quad a_0^\top x \quad \text{subject to} \quad a_i^\top x - b_i = 0 \quad i = 1, \ldots, m,$$
$$c_j^\top x - d_j \leq 0 \quad j = 1, \ldots, k.$$

- Rewritten in matrix form:

$$\min_x \quad a_0^\top x \quad \text{subject to} \quad Ax = b, \quad Cx \leq d. \tag{3}$$

- Rewritten in standard form:

$$\min_x \quad a_0^\top x \quad \text{subject to} \quad Ax = b, \quad x \geq 0. \tag{4}$$

  - If an LP is reformulated from the form (3) into the standard form (4), A and b in the standard form can be different from the ones the original form.
  - To convert an affine inequality constraint $Cx \leq d$ into the 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 optimization 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).

4.6.1 LP Solutions are at Feasible Set Vertices

- Definition: for a convex set 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.

4.6.2 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 $z^\star$ to the linear system $A_i^{\text{sub}} z = b$ is feasible (i.e., non-negative).

4.6.3 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:

$$\min_{x \in \mathbb{R}^n} \quad f(x) \quad \text{subject to} \quad x \in \mathcal{S}.$$

- Reformulate using a slack variable t:

$$\min_{x \in \mathbb{R}^n, t \in \mathbb{R}} t \quad \text{subject to} \quad x \in \mathcal{S}, \quad f(x) \leq t.$$

- 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.

4.7 Quadratic Programming (QP)

- QP includes a quadratic term in the objective, where $P_0 \succeq 0$:

$$\min_{x \in \mathbb{R}^n} \quad x^\top P_0 x + q_0^\top x + r_0 \quad \text{subject to} \quad Ax = b, \quad Cx \leq d.$$

4.8 Quadratically Constrained Quadratic Program (QCQP)

- A QCQP can be written in the form of

$$\min_{x \in \mathbb{R}^n} \quad x^\top P_0 x + q_0^\top x + r_0 \quad \text{subject to} \quad \begin{aligned} &Ax = b, \\ &x^\top P_j x + q_j^\top x + r_j \leq 0 \ \text{ for } \ j = 1, \ldots, k, \end{aligned}$$

  where $P_j \succeq 0$ for $j = 0, \ldots, k$.

- Hierarchy of optimization problems: LP $\subset$ QP $\subset$ QCQP $\subset$ convex optimization.

4.9 Convex Relaxations

- Consider optimization problem with $f(x)$ convex but S non-convex:

$$\min_{x \in \mathbb{R}^n} \quad f(x) \quad \text{subject to} \quad x \in \mathcal{S}$$

- If we replace S with a convex $\tilde{S}$ such that $S \subset \tilde{S}$, we get a *convex relaxation*:

$$\min_{x \in \mathbb{R}^n} \quad f(x) \quad \text{subject to} \quad x \in \tilde{S}$$

- Let x^* and $\tilde{x}$ be global minima of the original and relaxed optimizations, respectively. Then
  - $f(\tilde{x}) \leq f(x^*)$
  - If $\tilde{x} \in \mathcal{S}$, then $\tilde{x}$ is a global min for original optimization problem.

4.10 Integer Programming (IP)

- An IP is just an LP with a constraint that all elements of x are integers:

$$\min_x \quad a_0^\top x \quad \text{subject to} \quad \begin{aligned} &Ax = b, \\ &x \geq 0, \\ &x_i \ \text{ are integers for } \ i = 1, \ldots, n. \end{aligned}$$

- 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 and Hessian

Consider a function $f(x) : \mathbb{R}^n \mapsto \mathbb{R}$ 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 $\nabla f(x) \coloneqq \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right)$.
- The Hessian is an $n \times n$ symmetric matrix $\nabla^2 f(x) \coloneqq \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}$.
- 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 $\phi(x) \coloneqq f(g(x))$. Then

$$\underbrace{\nabla \phi(x)}_{n\text{-dimensional vector}} = \underbrace{\begin{bmatrix} \nabla g_1(x) & \ldots & \nabla g_m(x) \end{bmatrix}}_{n \times m \text{ matrix}} \times \underbrace{\nabla f(z)|_{z=g(x)}}_{m\text{-dimensional vector}}.$$

- 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 Optimality Conditions for Unconstrained Optimization Problems

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

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

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 $y \in \text{relint } \mathcal{D}$ 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 Optimality Conditions for Constrained Optimization Problems

Again, consider the optimization problem

$$\min_x f_0(x) \quad \text{subject to} \quad \begin{aligned} f_i(x) &= 0,\ i = 1, \ldots, k, \\ h_j(x) &\leq 0,\ j = 1, \ldots, m. \end{aligned}$$

Denote the dual variables associated with the equality constraints 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

$$L(x, \lambda, \mu) \coloneqq f(x) + \sum_{i=1}^{k} \mu_i f_i(x) + \sum_{j=1}^{m} \lambda_j h_j(x).$$

- **First-Order Necessary Condition (FOC):** If $x^\star$ is a regular point and a local minimum, then there exist Lagrangian multipliers $\lambda_1^\star, \ldots, \lambda_m^\star$ and $\mu_1^\star, \ldots, \mu_l^\star$ that makes the following holds true:
  1. **Primal Feasibility**: $f_i(x^\star) = 0$ for all $i = 1, \ldots, k$ and $h_j(x^\star) \leq 0$ for all $j = 1, \ldots, m$;
  2. **Dual Feasibility**: $\lambda_j^\star \geq 0$ for all $j = 1, \ldots m$;
  3. **Lagrangian Stationarity**: $\nabla f(x^\star) + \sum_{i=1}^{k} \mu_i^\star \nabla f_i(x^\star) + \sum_{j=1}^{m} \lambda_j^\star \nabla h_j(x^\star) = 0$;
  4. **Complementary Slackness**: $\lambda_j^\star \cdot h_j(x^\star) = 0$ for all $j = 1, \ldots m$.

  These conditions are called the Karush–Kuhn–Tucker (**KKT**) conditions.
- For convex optimization problems, the KKT conditions are sufficient can be used to find global optima.
- 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 $P_0 \in \mathbb{S}_+^n$, $A \in \mathbb{R}^{m \times n}$, $q_0 \in \mathbb{R}^n$, and $r_0 \in \mathbb{R}$.

  Suppose that Slater's condition holds, i.e., $Ax = b$ admits one or more solutions. Then, the optimal primal-dual solution $(x^\star, \mu^\star)$ satisfies $\begin{bmatrix} A & \mathbf{0}_{m \times m} \\ 2P_0 & A^\top \end{bmatrix} \begin{bmatrix} x^\star \\ \mu^\star \end{bmatrix} = \begin{bmatrix} b \\ -q_0 \end{bmatrix}$.

6 Linear Systems, Least Squares, and 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 $x^\star$ produces a space of solutions $x^\star + z$ 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 $y \in \mathcal{R}(A)$.
- Fat matrix: now assume $n > m$, 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:

  $$\min_{x \in \mathbb{R}^n} \|x\| \quad \text{subject to} \quad Ax = y.$$

  The minimum-norm solution can be derived as $x^\star = A^\top (AA^\top)^{-1} y = A^\dagger y$.
- 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 $x^\star$. Note that $x^\star$ also solves $\min_{x \in \mathbb{R}^n} \|Ax - y\|_2^2$.
- The set of solutions for the LS problem is $\mathcal{S} \coloneqq \{x^\star \mid A^\top A x^\star = A^\top y\}$. 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 Relationships between Least Squares and Projection

- Geometrically, the LS problem finds the projection of y onto $\mathcal{R}(A)$, the range of A.
- The projection result $y^\star \coloneqq Ax^\star = \Pi_{\mathcal{R}(A)} y$ exists and is unique.
- Theorem on projection: $y - y^\star \perp \mathcal{R}(A)$. I.e., $\langle y - y^\star, v \rangle = 0$ for all $v \in \mathcal{R}(A)$.
- We can find $y^\star$ by solving for the vector that simultaneously satisfies $y^\star \in \mathcal{R}(A)$ and $y - y^\star \perp \mathcal{R}(A)$.

6.4 Minimum-Norm Solution to Least Squares

- To find the minimum-norm solution, solve $\min_{x \in \mathcal{S}} \|x\|_2$. 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, $x^\star = A^\dagger y$ 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 $x^\star = (A^\top A + \alpha I_n)^{-1} A^\top y$.

6.6 Sparsity and 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:

$$\min_{x,t \in \mathbb{R}^n} x^\top P_0 x + q_0^\top x + r_0 + \alpha \sum_{i=1}^{n} t_i \quad \text{subject to} \quad -t_i \leq x_i \leq t_i \ \text{ for } \ i = 1, \ldots, n,$$

  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.
- $x^\star$ is a solution to LASSO if and only if $2P_0 x^\star + q_0 + \lambda^\star = 0$, where each entry of $\lambda^\star$ satisfies:
  $\lambda_i^\star = \alpha$ if $x_i^\star \geq 0$, $\lambda_i^\star = -\alpha$ if $x_i^\star \leq 0$, and $\lambda_i^\star \in [-\alpha, \alpha]$ if $x_i^\star = 0$.
  Furthermore, it holds that $|x_i^\star| = t_i^\star$ 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 $x + \Delta x$:

$$A(x + \Delta x) = y + \Delta y \quad \text{and} \quad Ax = y \quad \Longrightarrow \quad \Delta x = A^{-1} \Delta y.$$

- 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 – Least Squares

- 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:

$$E = \{x \in \mathbb{R}^n \mid x = By, \|y\|_2 \leq 1\};$$
$$E = \{x \in \mathbb{R}^n \mid x^\top P^{-1} x \leq 1\} \quad \text{where } P = BB^\top \text{ is PSD.}$$

  - 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 $\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0$ 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

As in Section 5.4, Consider the (potentially non-convex) optimization problem

$$\min_x f_0(x) \quad \text{subject to} \quad f_i(x) = 0,\ i = 1, \ldots, k,$$
$$h_j(x) \leq 0,\ j = 1, \ldots, m.$$

Denote the global optimal solution to this problem as $x^\star$. Also denote the dual variables associated with the equality constraints as $\mu_1, \ldots, \mu_k$. Similarly, denote the dual variables associated with the inequality constraints as $\lambda_1, \ldots, \lambda_m$.

Once again, the Lagrangian of this problem is

$$L(x, \lambda, \mu) \coloneqq f(x) + \sum_{i=1}^{k} \mu_i f_i(x) + \sum_{j=1}^{m} \lambda_j h_j(x).$$

7.1 Weak Duality

- In the context of duality, the original problem is called the *primal problem*. We call its optimal objective $p^\star \coloneqq f_0(x^\star)$ the *primal solution*.
- Consider arbitrary $\mu \in \mathbb{R}^k$ and $\lambda \in \mathbb{R}^m$ where $\lambda \geq 0$. It holds that $\min_x L(x, \lambda, \mu) \leq p^\star$.
- Hence, to find a meaningful lower bound to $p^\star$, we can solve

$$\max_{\mu \in \mathbb{R}^k, \lambda \in \mathbb{R}^m} \min_x L(x, \lambda, \mu) \quad \text{subject to} \quad \lambda \geq 0. \tag{5}$$

- We define $d(\lambda, \mu) \coloneqq \min_x L(x, \lambda, \mu)$ as the *dual function*.

  We can then reformulate the lower bound optimization problem (5) as the maximization problem

$$d^\star \coloneqq \max_{\mu \in \mathbb{R}^k, \lambda \in \mathbb{R}^m} d(\lambda, \mu) \quad \text{subject to} \quad \lambda \geq 0, \tag{6}$$

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

- It holds that $d^\star \leq p^\star$. The value of $p^\star - d^\star$ 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 (6) 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 $p^\star = d^\star$, 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., $x^\star$ solves the primal problem and $(\lambda^\star, \mu^\star)$ solves the dual problem.
  - If $x^\star$ is an arbitrary optimal solution to the primal problem and $(\lambda^\star, \mu^\star)$ is an arbitrary optimal solution to the dual problem, then $(x^\star, \lambda^\star, \mu^\star)$ 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

  $$\min_{x\in\mathbb{R}^n} \quad a_0^\top x \quad \text{subject to} \quad Ax \leq b, \quad Cx = d.$$

  The dual problem is

  $$\max_{\lambda\in\mathbb{R}^m, \mu\in\mathbb{R}^k} -\lambda^\top b - \mu^\top d \quad \text{subject to} \quad a_0 + A^\top\lambda + C^\top\mu = 0, \quad \lambda \geq 0.$$

- **The dual of a QP is also a QP.**

  Specifically, for some $P_0 \in \mathbb{S}^n_{++}$, $q_0 \in \mathbb{R}^n$, $r_0 \in \mathbb{R}$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $C \in \mathbb{R}^{k\times n}$, $d \in \mathbb{R}^k$, consider the QP

  $$\min_{x\in\mathbb{R}^n} \quad x^\top P_0 x + q_0^\top x + r_0 \quad \text{subject to} \quad Ax \leq b, \quad Cx = d.$$

  The dual problem is

  $$\max_{\lambda\in\mathbb{R}^m, \mu\in\mathbb{R}^k} -\frac{1}{4}\left(q_0 + A^\top\lambda + C^\top\mu\right)^\top P_0^{-1}\left(q_0 + A^\top\lambda + C^\top\mu\right) - \lambda^\top b - u^\top d \quad \text{subject to} \quad \lambda \geq 0.$$

- 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 $\max_\mu -\frac{1}{4}\mu^\top AA^\top\mu - \mu^\top b$.

  By Lagrangian stationarity, $x^\star = -\frac{1}{2}A^\top\mu^\star$. Setting the gradient of the dual problem objective to zero gives $\mu^\star = -2(AA^\top)^{-1}b$.

7.4 Certificate and Farkas' Lemma

- Suppose that we want to show that the set $\left\{ x \in \mathbb{R}^n \;\middle|\; \begin{array}{l} f_i(x) \leq 0, \;\; i = 1, \ldots, k \\ h_j(x) = 0, \; j = 1, \ldots, m \end{array} \right\}$ is empty.

  We can consider the optimization problem $\min_x 0$ subject to $f_i(x) \leq 0$ for all i and $h_j(x) = 0$ for all j.

  Next, we find the dual function of this optimization problem $d(\lambda, \mu)$. 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 $x \geq 0$ have no solutions if and only if there is a solution μ to $A^\top\mu \leq 0$ and $b^\top\mu < 0$.

8 Constraint Sensitivity Analysis

8.1 Constraint Elimination

- An inequality constraint $f_i(x) \leq 0$ is called active or binding at $x^\star$ if $f_i(x^\star) = 0$. It is called inactive or non-binding if $f_i(x^\star) < 0$.
- Suppose that $x^\star$ is an optimal solution to some problem. If we remove all non-binding solutions, then $x^\star$ is also a solution to the updated problem. However, the set of optimal solutions may change.
- Suppose that Slater's condition holds. Due to complementary slackness, the Lagrangian multipliers associated with the non-binding constraints are zero at $x^\star$.

8.2 Sensitivity Analysis

We are interested in comparing the optimization problem

$$\min_x f_0(x) \quad \text{subject to} \quad f_i(x) \leq 0 \text{ for } i = 1, \ldots, k, \quad h_j(x) = 0 \text{ for } j = 1, \ldots, m \tag{7}$$

with the problem that has perturbed constraints

$$\min_x f_0(x) \quad \text{subject to} \quad f_i(x) \leq v_i \text{ for } i = 1, \ldots, k, \quad h_j(x) = w_j \text{ for } j = 1, \ldots, m, \tag{8}$$

where each v_i and w_j is some scalar.

Denote the optimal objective value of the perturbed problem (8) as $p^\star(v, w)$. The optimal objective of the original problem (7) is $p^\star(0, 0)$. If the problem is infeasible for some (v, w), then $p^\star(v, w) = +\infty$.

We then have the following properties.

- $p^\star(v, w)$ is a convex function of v and w.
- Assume Slater's condition holds. If $p^\star(v, w)$ is differentiable at $(0, 0)$, then the Lagrangian multipliers $(\lambda^\star, \mu^\star)$ of the original problem (7) satisfies $\lambda_i^\star = -\frac{\partial p^\star(0,0)}{\partial v_i}$ for all i and $\mu_j^\star = -\frac{\partial p^\star(0,0)}{\partial w_j}$ for all j.
- As a result, it holds that $p^\star(v, w) \approx p^\star(0, 0) - \sum_i \lambda_i^\star v_i - \sum_j \mu_j^\star w_j$.

  This is the first-order Taylor's approximation for $p^\star(v, w)$. Given $x^\star$ (which can be used to compute $p^\star(0, 0)$), $\lambda^\star$, and $\mu^\star$ of the original unperturbed problem, this approximation can be computed efficiently.

  - If $\lambda_i^\star = 0$ for some i or $\mu_j^\star = 0$ for some j, then changing the corresponding constraint a little does not affect the optimal objective. Hence, those constraints can be eliminated.
  - If $\lambda_i^\star$ or $\mu_j^\star$ is small, then the optimization problem is not sensitive to the associated constraints.
  - If $\lambda_i^\star$ or $\mu_j^\star$ is large, then the optimization problem is highly sensitive to the associated constraints.

9 Numerical Optimization Algorithms

9.1 Gradient and 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 $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$ in a way that $f(x^{(k+1)}) < f(x^{(k)})$ for $k = 0, 1, 2, \ldots$.
- **Descent direction:** At a point $\bar{x} \in \mathbb{R}^n$, Δx is a descent direction if $\nabla f(\bar{x})^\top \Delta x < 0$.

- A family of optimization algorithms can be designed with descent directions: starting from $x^{(0)}$ as the initial guess, the k^{th} 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 k^{th} iteration.
  - Using descent directions guarantees that $f(x^{(k+1)}) < f(x^{(k)})$ for all small enough step sizes $s^{(k)}$.
- **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:** $x^{(k+1)} \leftarrow x^{(k)} - s^{(k)} \left(\nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)})$.
  Here, we use $-\left(\nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)})$, which is another descent direction when $\nabla f(x^{(k)}) \neq 0$ and $\nabla^2 f(x^{(k)}) \succ 0$, 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, if $\|x^{(k)} - x^\star\|$ is no greater than some positive threshold at some k, the algorithm terminates and we accept $x^{(k)}$ as a solution. However, since the true $x^\star$ is unknown, we need to estimate $\|x^{(k)} - x^\star\|$.

9.2 Analysis on Gradient Algorithm

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

10 Applications

10.1 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 $\min_{B \in \mathbb{R}^{m \times n}} \|A - B\|_{2 \text{ or } F}$ subject to $\text{Rank}(B) \leq k$.
- Eckart-Young-Mirsky theorem:
  - For a given $k \leq \min(m, n)$, define $A_k \coloneqq \sum_{i=1}^{k} \sigma_i u^{(i)} v^{(i)\top}$ constructed with the top k singular values of A and the corresponding 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 that $k < \text{Rank}(A)$. 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 $e_k \coloneqq \frac{\|A - A_k\|_F^2}{\|A\|_F^2}$ is equal to $\frac{\sigma_{k+1}^2 + \ldots + \sigma_r^2}{\sigma_1^2 + \ldots + \sigma_r^2}$, where $r = \text{Rank}(A)$.
- The relative ℓ_2-induced norm approximation error $\frac{\|A - A_k\|_2}{\|A\|_2}$ is equal to $\frac{\sigma_{k+1}}{\sigma_1}$.

10.2 Principal Component Analysis (PCA)

- Given points $x^1, \ldots, x^m$, first center data points to $\tilde{x}^1, \ldots, \tilde{x}^m$ by subtracting $\sum_{i=1}^m x^i$.
- 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.

10.3 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

$$\min_{X,Z \in \mathbb{R}^{m \times n}} \text{Rank}(X) + \lambda \, \text{Card}(Z) \quad \text{subject to} \quad Y = X + Z,$$

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:

$$\min_{X,Z \in \mathbb{R}^{m \times n}} \|X\|^\star + \lambda \sum_{i=1}^m \sum_{j=1}^n |Z_{ij}| \quad \text{subject to} \quad Y = X + Z.$$

10.3.1 Video Processing

Consider the problem of identifying moving objects in a video.

- Suppose that the video has q frames, $V^{(1)}, \ldots, V^{(q)}$, where the (i,j) entry $V^{(k)}$ is the color value of the (i,j) pixel at frame k for $k = 1, \ldots, q$.
- We compose each frame $V^{(k)}$ into the foreground $V_f^{(k)}$ (moving objects) and the background $V_b^{(k)}$ (changes slowly). Assuming that the moving object occupies a small portion of the video, then $V_f^{(k)}$ is sparse.
- Now we vectorize each $V^{(k)}$ and concatenate them into a single matrix as $Y = \begin{bmatrix} \text{vec}(V^{(1)}) & \ldots & \text{vec}(V^{(k)}) \end{bmatrix}$. We can then decompose the known matrix Y as $Y = X + Z$, where X is low-rank and Z is sparse.
  - X takes on the role of $\begin{bmatrix} \text{vec}(V_b^{(1)}) & \ldots & \text{vec}(V_b^{(k)}) \end{bmatrix}$.
  - Z takes on the role of $\begin{bmatrix} \text{vec}(V_f^{(1)}) & \ldots & \text{vec}(V_f^{(k)}) \end{bmatrix}$.
- We can then solve this decomposition via robust PCA.

10.4 Matrix Completion

- Consider a matrix $X^\star = \mathbb{R}^{m \times n}$ whose entries are unknown but is known to be low rank. Assume that we measure the entries $X_{ij}^\star$ only when (i,j) belongs to some given set $\mathcal{S}$.
- To estimate $X^\star$ using the measurements, we can find the lowest-rank X whose (i,j) entries match the measurements by solving for the optimization problem

$$\min_{X \in \mathbb{R}^{m \times n}} \text{Rank}(X) \quad \text{subject to} \quad X_{ij} = X_{ij}^\star, \ \forall (i,j) \in \mathcal{S}.$$

- This problem is non-convex due to the discrete rank function in the objective.
  Over the restricted space $\{X \in \mathbb{R}^{m \times n} \mid \|X\|_2 \leq 1\}$, a convex relaxation is

$$\min_{X \in \mathbb{R}^{m \times n}} \|X\|_* \quad \text{subject to} \quad X_{ij} = X_{ij}^\star, \ \forall (i,j) \in \mathcal{S}.$$

10.5 Compressed Sensing

- Let $x^\star \in \mathbb{R}^n$ denote some states of some system. We want to know $x^\star$ but can only measure $b \coloneqq Ax^\star \in \mathbb{R}^m$ for some $m \times n$ matrix A. When $m < n$, the linear system is underdetermined.
- Suppose that $x^\star$ is known to be sparse. Then $x^\star$ can be estimated via the optimization problem

$$\min_x \|x\|_0 \quad \text{subject to} \quad Ax = b.$$

- This problem is non-convex, but can be approximated with its convex relaxation over the restricted space of $-1 \leq x \leq 1$:

$$\min_x \|x\|_1 \quad \text{subject to} \quad Ax = b,$$

  which can be reformulated as an LP

$$\min_{x,t \in \mathbb{R}^n} \mathbf{1}_n^\top t \quad \text{subject to} \quad Ax = b,\ -t \leq x \leq t,$$

  where $\mathbf{1}_n$ denotes the n-dimensional all-one column vector.
- Suppose that our measurements are noisy, i.e., $\hat{b} = Ax + w$ where w is random (often Gaussian). Then, the problem we should solve is

$$\min_{x,w} \|w\|_2^2 + \lambda \|x\|_1 \quad \text{subject to} \quad Ax + w = b,$$

  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).

10.6 Power Network Identification with Adversarial Sensor Injections

- Power networks can be represented as a graph, where each node n_i is a generator or a load and each edge e_{ij} connects node n_i to node n_j.
- Each node n_i has phase θ_i and each edge e_{ij} has a reactance $z_{ij} \in \mathbb{R}$.
- The power flow on each edge is $p_{ij} = \frac{\theta_i - \theta_j}{z_{ij}}$.
- Due to the graph structure, $p_{ij} = -p_{ji}$. Furthermore, the power at each node n_i is $p_i = \sum_{j \in \mathbb{N}_i} p_{ij}$, where $\mathbb{N}_i$ is the set of indices of all nodes n_j that are connected to n_i via an edge e_{ij}.
- We have node measurements $\hat{p}_i$ for all $i \in \hat{\mathbb{N}}$ where $\hat{\mathbb{N}}$ is the set of measured node indices. Here, Here, $\hat{p}_i = p_i + w_i + v_i$, where w_i is sensor noise (usually random, dense, but small) and v_i is potential adversarial injection (usually large but sparse).
- We also have edge measurements $\hat{p}_{ij}$ for all $(i,j) \in \hat{\mathbb{E}}$ where $\hat{\mathbb{E}}$ is the set of measured edge indices. Here, $\hat{p}_{ij} = p_{ij} + w_{ij} + v_{ij}$, where w_{ij} is sensor noise and v_{ij} is potential adversarial injection.
- We assume that only a small number of nodes and edges are subject to adversarial injection.
- Given the (noisy and potentially adversarially perturbed) sensor measurements, we want to estimate the power network status (i.e., the true p_i and p_{ij}). To achieve so, we can solve the LASSO problem

$$\begin{aligned} \min_{w,v,\theta} \quad \|w\|_2^2 + \lambda \|v\|_1 \quad \text{subject to} \quad & \hat{p_i} = \sum_{j \in N_i} \frac{\theta_i - \theta_j}{z_{ij}} + w_i + v_i, \\ & \hat{p_{ij}} = \frac{\theta_i - \theta_j}{z_{ij}} + w_{ij} + v_{ij}, \\ & \theta_1 = 0. \end{aligned}$$

  Here, each $\hat{p}_i$ and each $\hat{p}_{ij}$ is measured, each z_{ij} is known, w and v are vectors that contains all w_i, w_{ij} and v_i, v_{ij} respectively, θ is a vector that contains all node phases, and $\lambda > 0$ is a hyperparameter that adjusts the sparsity of the solution.

