EECS 127/227A – Midterm Review

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 $a \in \mathbb{R}^n$ s.t. $f(x) = a^\top x$.
- 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) if all eigenvalues are non-negative.
  I.e., $\lambda_1(A), \ldots, \lambda_n(A) \geq 0$. The corresponding notation is $A \succeq 0$ or $A \succcurlyeq 0$.
- 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) if all eigenvalues are strictly positive.
  I.e., $\lambda_1(A), \ldots, \lambda_n(A) > 0$. The corresponding notation is $A \succ 0$.
  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: $\|A\|_F := \|\text{vec}(A)\|_2$, 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: $\|A\|_p := \max_{z \in \mathbb{R}^n, z \neq 0} \frac{\|Az\|_p}{\|z\|_p} = \max_{\|w\|_p = 1} \|Aw\|_p$.
- 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 Optimization Problems

3.1 Standard Form and Constraints

- 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 $h(x) = 0 \iff \{h(x) \leq 0, -h(x) \leq 0\}$.
- 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 $x \in \mathcal{X}$.
- Consider an arbitrary function $f(x)$. Suppose that some x is the optimal solution to $\min_x f(x)$, then it is also optimal for $\max_x -f(x)$ and $\min_x \alpha f(x)$, where $\alpha > 0$ is any positive scalar.

3.2 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 negative infinity. 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.
- Optimal objective value is $+\infty$ if infeasible, $-\infty$ if unbounded from below, and finite otherwise ($x^\star$ may or may not be attainable).

4 Optimality Conditions

4.1 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 first order we have the approximation: $f(x) \approx f(x_0) + \nabla f(x_0)^\top (x - x_0)$.

4.2 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* $x \in \mathbb{R}^n$. 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 Linear Systems and Least Squares

5.1 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 $x = A^{-1} y$.

5.2 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.

5.3 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 $y - y^\star \perp \mathcal{R}(A)$.

5.4 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.

5.5 Ridge Regression

- A 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 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 $\frac{\sigma_{k+1}}{\sigma_1}$.

