EE 127

Convex Optimization

Notes

Druv Pai

Contents

1 Logistics and Overview

Office hours are 6:30-7:30 after class on Tuesday. Homeworks and self-grades are due on Friday at 11 PM.

Example 1. Say that you have 10^5 barrels of crude oil. You can process each barrel of crude oil into a barrel of jet fuel, which costs 10\$ or a barrel of gasoline, which costs 20\$. In an ideal world, you build all the gasoline possible. But we don't live in such a world, so we add some constraints, for example that we need to produce at least 10^3 barrels of jet fuel and $2 \cdot 10^3$ barrels of gasoline. Further, we can transport at most $1.8 \cdot 10^6$ barrel-miles. The gasoline distributor is 30 miles away, and the jet fuel distributor is 10 miles away. How do we maximize profit?

Solution. Let x_1 be the number of gallons of jet fuel and x_2 be the number of gallons of gasoline. The task resolves to

$$\begin{aligned} \text{maximize } & 10x_1 + 20x_2 \\ \text{subject to } & x_1 \geq 10^3 \\ & x_2 \geq 2 \cdot 10^3 \\ & 10x_1 + 30x_2 \leq 1.8 \cdot 10^6 \end{aligned}$$

This is a linear program, which is one of the problems that we will solve a lot during the latter part of the course. □

A minimization program, in general, has an optimization variable $x \in \mathbb{R}^n$ with optimal value x^* among the feasible set, or set of values for x that satisfy the m constraints placed upon x. The program is of the form

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{subject to } f_i(x) \leq b_i, \quad i \in \{1, \ldots, m\} \end{aligned}$$

Example 2. Say that the i^{th} class has value α_i and workload β_i. Let $x_i = \mathbb{1}\left(\text{you take } i^{\text{th}} \text{ class}\right)$. Then if the total workload you can handle is b, the integer linear program (linear program where the decision variable takes only integral values) for the classes you should take is given by

$$\begin{aligned} &\text{maximize } \alpha^\mathsf{T} x \\ &\text{subject to } \beta^\mathsf{T} x \leq b \end{aligned}$$

and solving this gives you a class schedule.

Example 3. Say that the i^{th} class has size x_i, credits c_i, and resources r_i. Then to maximize credit hours for students subject to a resource budget b, the integer linear program is

$$\begin{aligned} &\text{maximize } c^\mathsf{T} x \\ &\text{subject to } r^\mathsf{T} x \leq b \end{aligned}$$

These are examples of generic programs that we will solve by the end of the course.

2 Least Squares

The least squares problem is one of the most basic optimization problems. Given a matrix A and vector b (or appropriately sized tensors), the least squares problem finds the projection of b onto the span of the

columns of A:

$$x^* = \underset{x}{\operatorname{argmin}} \|Ax - b\|^2$$

Example 4. Say we have a set of data points $\{(x_i, y_i)\}_{i=1}^m$, where the noiseless relationship is $y = \beta_1 x + \beta_2$ and, we wish to find β. In matrix-vector form, we have

$$\begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

More simply, if x is a vector of the x_i, and y is a vector of the y_i,

$$\begin{bmatrix} x & 1 \end{bmatrix} \beta = y$$

This leads to a quadratic cost, minimizable in β:

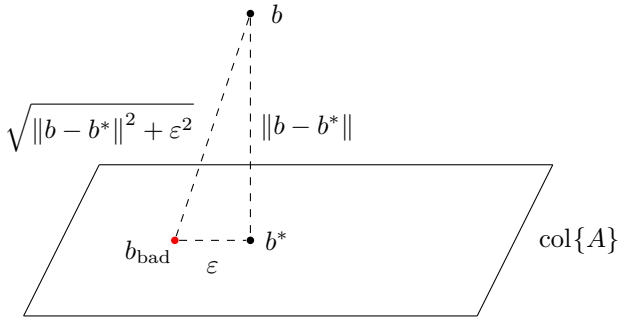
$$\beta^* = \underset{\beta_1, \beta_2}{\operatorname{argmin}} \sum_{i=1}^m [y_i - (\beta_1 x_i + \beta_2)]^2$$

which we can minimize in closed form, or numerical methods if m is large.

How do we actually find the quantity

$$x^* = \underset{x}{\operatorname{argmin}} \|Ax - b\|^2$$

One way is to solve it geometrically. Consider the subspace $\operatorname{col}(A)$; then b may be on $\operatorname{col}(A)$, or it may not be. If $b \in \operatorname{col}(A)$, then we are done; if not, then the closest point is $b^* = \operatorname{proj}_{\operatorname{col}(A)}(b)$, and other points are off by an appropriate amount.



By triangle geometry, we require $b - b^* = b - Ax^*$ to be orthogonal to the columns of A, so we require $Ax^* - b$ to be orthogonal to the columns of A, so $A^\mathsf{T}(Ax^* - b) = 0$. Then

$$\begin{aligned} 0 &= A^\mathsf{T}(Ax^* - b) \\ &= A^\mathsf{T}Ax^* - A^\mathsf{T}b \\ x^* &= \left(A^\mathsf{T}A\right)^{-1} A^\mathsf{T}b \end{aligned}$$

Of course, this requires $A^\mathsf{T}A$ to be invertible, so A must be full rank. If not, then the relevant equation is $A^\mathsf{T}Ax^* = A^\mathsf{T}b$; one can show by rank-nullity that $A^\mathsf{T}b \in \operatorname{range}\left(A^\mathsf{T}A\right)$, so computing $A^\mathsf{T}b$ and then row-reducing finds the solution.

The cost function for the least squares problem is quadratic; quadratic functions are convex, and thus all of their local minima are global minima. This is very nice because we immediately know how to minimize

such functions extremely well, often in closed form.

3 Vector Calculus

We begin with scalar-valued Taylor's theorem.

Theorem 5 (Taylor's Theorem in $\mathbb{R} \to \mathbb{R}$ Case)
Let $f \colon \mathbb{R} \to \mathbb{R}$. Then

$$f(x+\varepsilon) = \sum_{n=0}^{\infty} \frac{\mathrm{d}^n f(x)}{\mathrm{d}x^n} \frac{\varepsilon^n}{n!} = f(x) + \frac{\mathrm{d}f(x)}{\mathrm{d}x}\varepsilon + \frac{1}{2}\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2}\varepsilon^2 + \cdots$$

in the case that f is smooth; there exists a remainder term $R_k(x)$ such that $|R_k(x)| \leq \left|\frac{\varepsilon^{k+1}}{(k+1)!}\frac{\mathrm{d}^{k+1} f(x)}{\mathrm{d}x^{k+1}}\right|$ otherwise.

Now we cover the multivariable case. First, some notation. Formally, if $f \colon \mathbb{R}^n \to \mathbb{R}$, then $\frac{\partial f(x)}{\partial x} = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} & \cdots & \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$ and $\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$. If $f(x)$ is scalar valued, then $\nabla_x f(x)$ has the same shape as x.

Now, we cover the second derivative. In the case that $f \colon \mathbb{R}^n \to \mathbb{R}$, the second derivative $\nabla_x^2 f(x)$ is called the Hessian. In particular, $\left[\nabla_x^2 f(x)\right]_{i,j} = \frac{\partial}{\partial x_j}\left(\nabla_x f(x)\right)_i = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$.

Theorem 6 (Taylor's Theorem in $\mathbb{R}^n \to \mathbb{R}$ Case)
Let $f \colon \mathbb{R}^n \to \mathbb{R}$. Then

$$\begin{aligned} f(x+\varepsilon) &= f(x) + \frac{\partial f}{\partial x}\varepsilon + \frac{1}{2}\varepsilon^{\mathsf{T}} \nabla_x^2 f(x)\varepsilon + \cdots \\ &= f(x) + \left(\nabla_x f(x)\right)^{\mathsf{T}} \varepsilon + \frac{1}{2}\varepsilon^{\mathsf{T}} \nabla_x^2 f(x)\varepsilon + \cdots \\ &= f(x) + \langle \nabla_x f(x) | \varepsilon \rangle + \frac{1}{2}\varepsilon^{\mathsf{T}} \nabla_x^2 f(x)\varepsilon + \cdots \end{aligned}$$

Derivatives at order $n \geq 3$ or more involve higher-order tensors and we won't cover those now.

Definition 7 (Level Set). Define the c-level set of f as $L_c(f) = \{x \mid f(x) = c\}$.

Since the gradient is sometimes interpreted as the direction of fastest movement or steepest ascent, we have that (by the definition of the directional derivative as $D_a f(x) = \langle \nabla_x f(x) | a \rangle$, then for all c $[\nabla_x f(x)]_{x \colon f(x)=c} \perp L_c(f)$.

Example 8. Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be defined as $f(x) = x^{\mathsf{T}} x$. We have $\nabla_x f(x) = 2x$ manually, and $\nabla_x^2 f(x) = 2I$. Now we can compute it via Taylor's theorem:

$$\begin{aligned} f(x+\varepsilon) &= (x+\varepsilon)^{\mathsf{T}} (x+\varepsilon) \\ &= x^{\mathsf{T}} x + x^{\mathsf{T}} \varepsilon + \varepsilon^{\mathsf{T}} x + \varepsilon^{\mathsf{T}} \varepsilon \end{aligned}$$

$$= x^\mathsf{T} x + (2x)^\mathsf{T} \varepsilon + \frac{1}{2}\left(2\varepsilon^\mathsf{T}\varepsilon\right)$$

We have that $\nabla_x f(x) = 2x$ and $\nabla_x^2 = 2I$ by pattern matching.

Example 9. If $f(x) = x^\mathsf{T} a$, where $x, a \in \mathbb{R}^n$, then

$$f(x+\varepsilon) = (x+\varepsilon)^\mathsf{T} a = x^\mathsf{T} a + \varepsilon^\mathsf{T} a = f(x) + a^\mathsf{T}\varepsilon$$

so $\nabla_x f(x) = a$.

Example 10. If $f(x) = x^\mathsf{T} A x$ then

$$\begin{aligned} f(x+\varepsilon) &= (x+\varepsilon)^\mathsf{T} A (x+\varepsilon) \\ &= \left(x^\mathsf{T} + \varepsilon^\mathsf{T}\right)(Ax + A\varepsilon) \\ &= x^\mathsf{T} A x + \varepsilon^\mathsf{T} A x + x^\mathsf{T} A \varepsilon + \varepsilon^\mathsf{T} A \varepsilon \\ &= x^\mathsf{T} A x + x^\mathsf{T} A^\mathsf{T} \varepsilon + x^\mathsf{T} A \varepsilon + \varepsilon^\mathsf{T} A \varepsilon \\ &= x^\mathsf{T} A x + x^\mathsf{T}\left(A^\mathsf{T} + A\right)\varepsilon + \varepsilon^\mathsf{T} A \varepsilon \end{aligned}$$

Therefore we have that $\nabla_x f(x) = \left(A + A^\mathsf{T}\right) x$ and $\nabla_x^2 f(x) = 2A$.

Example 11 (Least Squares). We want to find $x^* = \operatorname{argmin}_x \|Ax - b\|_2^2$. Let

$$f(x) = \|Ax - b\|_2^2 = (Ax-b)^\mathsf{T}(Ax-b) = x^\mathsf{T} A^\mathsf{T} A x - x^\mathsf{T} A^\mathsf{T} b - b^\mathsf{T} A x + b^\mathsf{T} b.$$

Using the previous example, $\nabla_x f(x) = 2\left(A^\mathsf{T} A x - A^\mathsf{T} b\right)$. Setting this to 0, we have $x^* = \left(A^\mathsf{T} A\right)^{-1} A^\mathsf{T} b$ which agrees.

4 Linear Algebra Review

Definition 12 (Vector). A vector x is an element of a vector space $X(\mathbb{F})$, which is a group that is closed under an addition operation, as well as another operation which multiplies a vector x by a member a of its component field $\mathbb{F}$.

Norms

We now discuss norms. Let X be a normed vector space, which is a vector space equipped with a norm.

Definition 13 (Norm). A function $f\colon X \to \mathbb{F}$ is a norm if

- $f(x) \geq 0$ for all $x \in X$
- $f(x) = 0$ if and only if $x = 0$
- $f(ax) = af(x)$ for all $x \in X$ and $a \in \mathbb{F}$.
- $f(x+y) \leq f(x) + f(y)$ for all $x, y \in X$ (triangle inequality).

One norm is the 2-norm on real sequences $x = (x_i)_{i=1}^n$, which is defined as

$$\|x\|_2 \stackrel{\text{def}}{=} \left(\sum_{i=1}^n |x_i|^2\right)^{\frac{1}{2}}$$

Definition 14. The ℓ^p norm on real sequences, defined as

$$\|x\|_p \stackrel{\text{def}}{=} \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$$

For $p \in [1, \infty)$, the p-norm is a proper norm (for $p \in [0, 1)$, the p-norm is what's called a seminorm).

When $p = 2$, we recover the Euclidean norm. When $p = 1$, we obtain $\|x\|_1 = \sum_{i=1}^n |x_i|$; when $p = 0$, we obtain $\|x\|_0 = n$; when $p \to \infty$, we obtain $\|x\|_\infty = \max_i x_i$.

Theorem 15 (Cauchy Schwarz Inequality)
Let X be a normed inner product space. Then for any $x, y \in X$, we have

$$|\langle x|y\rangle| \leq \|x\|_2 \|y\|_2$$

Proof. We have that

$$|\langle x|y\rangle| = |\|x\|_2\|y\|_2 \cos(\theta)| = \|x\|_2\|y\|_2|\cos(\theta)| \leq \|x\|_2\|y\|_2$$

as desired. □

There is a generalization of this for arbitrary norms:

Theorem 16 (Holder's Inequality)
Let X be a normed inner product space. Then for any p and q such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$|\langle x|y\rangle| \leq \sum_{i=1}^{n} |x_i y_i| \leq \|x\|_p \|y\|_q$$

Proof. The proof will be covered once we have developed the tools of convexity.

Note that gradient proofs don't work because the functions on the left hand side are not differentiable. □

Example 17. Say that we have some $y \in \mathbb{R}^n$. We want to find

$$\max_{\|x\|_p \leq 1} \langle x|y\rangle$$

In the $p = 2$ case, we obtain $x^* = \frac{y}{\|y\|_2}$; by Theorem 15, we have that this is optimal.

In the $p \to \infty$ case, we pick $x^* = \text{sign}(y)$, so $\max_{\|x\|_p \leq 1} \langle x|y\rangle = \sum_i^n |y_i| = \|y\|_1$.

In the $p = 1$ case, we pick x^* such that $x_i^* = \mathbb{1}(|y_i| = \max_j |y_j|)$, so $\max_{\|x\|_p \leq 1} \langle x|y\rangle = \max_j |y_j| = \|y\|_\infty$.

It turns out that $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are dual norms, and $\|\cdot\|_2$ is a self-dual norm.

Orthogonality and Orthonormalization

The Gram-Schmidt process converts a basis for a vector space into an orthonormal basis for the same vector space.

Algorithm 1 The Gram-Schmidt process.

Input: A basis $\{a_1, \ldots, a_n\}$ for a vector space V.
Output: An orthonormal basis $\{q_1, \ldots, q_n\}$ for V.

$q_1 \leftarrow \frac{a_1}{\|a_1\|}$ ▷ Normalize a_1.
for $i \in \{2, \ldots, n-1\}$ **do**
  $p_i \leftarrow \sum_{j=1}^{i-1} q_j \langle a_i | q_j \rangle$ ▷ Project a_i onto $\text{span}(q_1 < \ldots, q_{i-1})$.
  $s_i \leftarrow a_i - p_i$ ▷ Subtract out the projection.
  $q_i \leftarrow \frac{s_i}{\|s_i\|_2}$ ▷ Normalize p_i.
return $\{q_1, \ldots, q_n\}$

We introduce the QR decomposition, where any full-rank matrix A can be expressed as $A = QR$, where Q is an orthogonal matrix and R is an upper-triangular matrix. The construction is

$$Q_{i,j} = q_{i,j}, \quad R_{i,j} = \langle q_i | a_j \rangle \text{ for } i \geq j \text{ or } \|s_i\|_2 \text{ for } i = j$$

One may also write $A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$ since only a few components of the orthogonal matrix are used.

Claim 18. Let X be an inner product space and S be a subspace of V. Let $x \in X$. Then x can be written uniquely as the sum of $s \in S$ and $s' \in S^\perp$, where $S^\perp$ is the orthogonal complement of S, i.e. $S^\perp = \{s' \colon \langle s' | s \rangle = 0 \ \forall s \in S\}$. Another way to write this is that $X = S \oplus S^\perp$.

Proof. We want to show that $S \oplus S^\perp = X$. Clearly $S \cap S^\perp = \{0\}$, and $S + S^\perp \subseteq X$. Assume that $W \subset X$ for the sake of contradiction. Let $\{w_n\}$ be an orthonormal basis for W, and extend this basis to X (say perhaps via Gram-Schmidt, or in the infinite dimensional case using Hahn-Banach extension theorem). Then there exists $x \notin W$ a new basis vector, and by Gram-Schmidt $x \perp W$, so $x \perp S$, and $x \perp S^\perp$, a contradiction.

It remains to show uniqueness. Consider $x_1, x_2 \in S$ and $y_1, y_2 \in S^\perp$, where $x_1 + y_1 = x_2 + y_2$ but $x_1 \neq x_2$ and $y_1 \neq y_2$. Then $x_1 - x_2 = y_2 - y_1$. Clearly $x_1 - x_2 \in S$ and $y_2 - y_1 \in S^\perp$, so $x_1 - x_2 = y_2 - y_1 = 0$, a contradiction. □

Theorem 19 (Fundamental Theorem of Linear Algebra)
We have that

$$\text{range}(A) = \text{null}\left(A^\mathsf{T}\right)^\perp$$

and

$$\dim(\text{range}(A)) + \dim(\text{null}(A)) = \text{rank}(A) + \text{nullity}(A) = n$$

for $A \in \mathbb{R}^{m \times n}$.

Principal Component Analysis

Principal component analysis is a technique for dimensionality reduction. Say that we have unlabeled data points $\{x_i\}_{i=1}^n \in V$, where V is a normed inner product space, and has $\dim(V) = p$ which is considered "high-dimensional". Further assume that the x_i are drawn from a distribution with 0 mean, that is, $x_1, \ldots, x_n \sim p_x(\cdot)$ where $\text{Ex}[x] = 0$. Our goal is to find projections of x_i onto some subspace $W \subset V$, where

W is a normed inner product space, $\dim(W) = k$ where $k < p$, where $\text{proj}_W(x_i)$ is as close to x_i as possible, across all the x_i.

Let $k = 1$. The goal is to determine a w with $\|w\|_2^2 = 1$ such that the projections onto $\text{span}(w)$ are as close to the original vectors as possible. These projections are $\{\langle x_i|w\rangle\}_{i=1}^n$, with error norms $e_i^2 = \|x_i - \langle w|x_i\rangle\|^2$, and average projection error $\frac{1}{n}\sum_{i=1}^n e_i^2$. Noting that $\text{Ex}[x] = 0$, we simplify the projection error to become

$$\begin{aligned}\|x_i - \langle w|x_i\rangle\, w\|_2^2 &= (x_i - \langle w|x_i\rangle\, w)^\mathsf{T}(x_i - \langle w|x_i\rangle\, w)\\ &= \|x_i\|_2^2 - 2\langle w|x_i\rangle^2 + \langle w|x_i\rangle^2\|w\|_2^2\\ &= \|x_i\|_2^2 - \langle w|x_i\rangle^2\\ e_i^2 &= \|x_i\|_2^2 - \langle w|x_i\rangle^2\end{aligned}$$

The mean square projection error is

$$\text{MSE}(w) = \frac{1}{n}\sum_{i=1}^n \|x_i\|_2^2 - \frac{1}{n}\sum_{i=1}^n \langle w|x_i\rangle^2$$

The goal is to minimize the mean square projection error, and we can see that in expectation this is equivalent to the problem:

$$\begin{aligned}\text{Ex}\left[\min_w \text{MSE}(w)\right] &= \min_w \text{Ex}[\text{MSE}(w)]\\ &= \max_w \text{Ex}\left[\frac{1}{n}\sum_{i=1}^n \langle w|x_i\rangle^2\right]\\ &= \max_w \text{Ex}\left[\langle w|x\rangle^2\right]\\ &= \max_w \text{Ex}[\langle w|x\rangle]^2 + \text{Var}[\langle w|x\rangle]\\ &= \max_w \text{Var}[\langle w|x\rangle]\end{aligned}$$

Thus the principal component analysis maximizes the variance of w with the distribution p_X.

Another way to look at principal component analysis (and indeed, the way to consider an instantiation of principal component analysis, where we cannot work in expectation) is to note that, if $X = \begin{bmatrix} x_1^\mathsf{T} \\ \vdots \\ x_n^\mathsf{T}\end{bmatrix}$ is the data matrix, we are looking for

$$\begin{aligned}\min_w \text{MSE}(w) &= \max_w \frac{1}{n}\sum_{i=1}^n \langle w|x_i\rangle^2\\ &= \frac{1}{n}\|Xw\|_2^2\\ &= \frac{1}{n}(Xw)^\mathsf{T}(Xw)\\ &= \frac{1}{n}w^\mathsf{T}X^\mathsf{T}Xw\\ &= w^\mathsf{T}\left(\frac{X^\mathsf{T}X}{n}\right)w\end{aligned}$$

where $C = \left(\frac{X^\mathsf{T}X}{n}\right)$ is the coveriance matrix of the data. Then the problem is to find $\max_{\|w\|_2=1} w^\mathsf{T}Cw$.

This motivates the discussion of symmetric matrices.

Symmetric Matrices and Positive Semidefiniteness

Definition 20 (Symmetric Matrix). A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is a matrix A such that $A^\mathsf{T} = A$, or $a_{i,j} = a_{j,i}$ for each pair (i, j).

The set of symmetric matrices is sometimes called $\mathcal{S}$.

Definition 21 (Diagonalizability). Let A be square and have characteristic polynomial $p(\lambda) = \det(A - \lambda I) = \prod_{i=1}^{n} (\lambda - \lambda_i)^{n_i}$. Then for each λ_i, we have $n_i = \dim(\text{null}(A - \lambda_i I))$.

Theorem 22 (Spectral Theorem for Real Symmetric Matrices)
Let $A \in \mathbb{R}^{n \times n}$ be a real symmetric matrix, with eigenvalues $\lambda_1, \dots, \lambda_k$ which themselves have algebraic multiplicities $\mu_1, \dots, \mu_k$. Further, define $\phi_i = \text{null}(A - \lambda_i I)$. Then for all i and $j \neq i$, we have

- $\lambda_i \in \mathbb{R}$,
- $\phi_i \perp \phi_j$,
- $\dim(\phi_i) = \mu_i$

Therefore, there exists orthogonal U and diagonal Λ such that $A = U \Lambda U^\mathsf{T}$.

Proof. The first two claims are trivial, so we prove the third. Let $U = \begin{bmatrix} u & U_1 \end{bmatrix}$, where we may obtain U by Gram-Schmidt. Then $U^\mathsf{T} A U = \begin{bmatrix} u^\mathsf{T} \\ U_1^\mathsf{T} \end{bmatrix} A \begin{bmatrix} u & U_1 \end{bmatrix} = \begin{bmatrix} u^\mathsf{T} \\ U_1^\mathsf{T} \end{bmatrix} \begin{bmatrix} \lambda u & AU_1 \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & U_1^\mathsf{T} A U_1 \end{bmatrix}$ If $B = U_1^\mathsf{T} A U_1$, then B is symmetric, and we are done by induction. □

Definition 23 (Rayleigh Quotient). For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, we define the Rayleigh quotient with respect to x as

$$\frac{x^\mathsf{T} A x}{x^\mathsf{T} x}$$

Theorem 24
We have that

$$\lambda_{\min}(A) \leq \frac{x^\mathsf{T} A x}{x^\mathsf{T} x} \leq \lambda_{\max}(A)$$

for all $x \in \mathbb{R}^n \setminus \{0\}$.

Proof. Write $A = U_A \Lambda_A U_A^\mathsf{T}$. Then

$$x^\mathsf{T} A x = x^\mathsf{T} U_A \Lambda_A U_A^\mathsf{T} x$$

Write $y = U_A^\mathsf{T} x$. Then

$$\begin{aligned} x^\mathsf{T} A x &= y^\mathsf{T} \Lambda_A y \\ &= \sum_{i=1}^{n} \lambda_i(A) y_i^2 \\ &\leq \lambda_{\max}(A) \sum_{i=1}^{n} y_i^2 \end{aligned}$$

$$\geq \lambda_{\min}(A) \sum_{i=1}^{n} y_i^2$$

Also note that $\sum_{i=1}^{n} y_i^2 = \|y\|_2^2 = \left\|U_A^\mathsf{T} x\right\|_2^2 = \|x\|_2^2$ since U_A is orthonormal. □

Corollary 25. We have

$$\lambda_{\max}(A) = \max_{\|x\|_2=1} x^\mathsf{T} A x$$

and

$$\lambda_{\min}(A) = \min_{\|x\|_2=1} x^\mathsf{T} A x.$$

Proof. Let $x = v_{\max}$ (the eigenvector associated with $\lambda_{\max}(A)$) to get the first inequality, and $x = v_{\min}$ to get the second inequality; these clearly work when substitued into the proof. □

Definition 26 (Positive Semidefinite). A matrix $A \in \mathbb{R}^{n \times n}$ is positive semidefinite (PSD) if for all $x \in \mathbb{R}^n$, we have $x^\mathsf{T} A x \geq 0$. We also say $A \succeq 0$ or $A > 0$.

Corollary 27. If A is PSD then $\lambda_i(A) \geq 0$.

Proof. Directly follows from Theorem 24. □

Definition 28 (Matrix Square Root). If A is PSD then there exists B such that $A = B^\mathsf{T} B$. Indeed, if $A = U_A \Lambda_A U_A^\mathsf{T}$ then $B = U_A \Lambda_A^{1/2}$ (where $\Lambda_{A;i,j}^{1/2} = \sqrt{\Lambda_{A;i,j}}$) suffices, but is not symmetric; one can show that there is a symmetric matrix C such that $A = C^\mathsf{T} C$, but this itself is not unique in the case of A having repeated eigenvalues.

Back to PCA. We showed that our problem was equivalent to

$$w^* = \operatorname*{argmax}_{\|w\|_2=1} w^\mathsf{T} C w$$

From 25, we have that w^* is the eigenvector corresponding to $\lambda_{\max}(C)$. Then w is the first principal component.

Singular Value Decomposition

Definition 29 (Singular Value Decomposition). Let $A \in \mathbb{R}^{m \times n}$ with $\operatorname{rank}(A) = r$. Then the singular value decomposition of A is an expression of A in the form

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\mathsf{T}$$

with the convention $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$, $u_1, \ldots, u_r \in \mathbb{R}^m$ are orthonormal, and $v_1, \ldots, v_r \in \mathbb{R}^n$ are also orthonormal.

We also write $A = U_r \Sigma_r V_r^\mathsf{T}$, where $U_r = \begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix} \in \mathbb{R}^{m \times r}$, $\Sigma_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r \times r}$, and $V_r^\mathsf{T} = \begin{bmatrix} v_1^\mathsf{T} \\ \vdots \\ v_r^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{r \times n}$. This is sometimes called the "compact" SVD.

We can compute the singular value decomposition via the following algorithm:

Algorithm 2 Computes the singular value decomposition.

Input: $A \in \mathbb{R}^{m \times n}$,
Output: $U_r \in \mathbb{R}^{m \times r}$ (orthonormal), $\Sigma_r \in \mathbb{R}^{r \times r}$ (diagonal), $V_r^\mathsf{T} \in \mathbb{R}^{r \times n}$ (orthonormal) such that $A = U_r \Sigma_r V_r^\mathsf{T}$.
Compute $A^\mathsf{T} A$. ▷ $A^\mathsf{T} A$ is symmetric.
$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0 \leftarrow$ eigenvalues of $A^\mathsf{T} A$.
$v_1, \ldots, v_r \leftarrow$ orthonormal eigenvectors of $A^\mathsf{T} A$ (such that $A^\mathsf{T} A v_i = \lambda v_i$ for each i).
$\sigma_i \leftarrow \sqrt{\lambda_i}$ for each i.
$u_1, \ldots, u_r \leftarrow$ defined by $A v_i = \sigma_i u_i$ for all i.
return $\begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix}, \operatorname{diag}(\sigma_1, \ldots, \sigma_r), \begin{bmatrix} v_1 & \cdots & v_r \end{bmatrix}^\mathsf{T}$

Claim 30. The u_i are orthonormal.

Proof. First, we claim that $\|u_i\|_2^2 = 1$ for all i. We have

$$\begin{aligned} \|\sigma_i u_i\|_2^2 &= (\sigma_i u_i)^\mathsf{T} (\sigma_i u_i) \\ &= (A v_i)^\mathsf{T} (A v_i) \\ &= v_i^\mathsf{T} A^\mathsf{T} A v_i \\ &= v_i^\mathsf{T} \lambda_i v_i \\ &= \lambda_i v_i^\mathsf{T} v_i \\ &= \lambda_i \|v_i\|_2^2 \\ \sigma_i^2 \|u_i\|_2^2 &= \lambda_i \|v_i\|_2^2 \\ \lambda_i \|u_i\|_2^2 &= \lambda_i \|v_i\|_2^2 \\ \|u_i\|_2^2 &= \|v_i\|_2^2 \\ &= 1 \end{aligned}$$

so $\|u_i\|_2^2 = 1$ as desired. Now we claim that $\langle u_i | u_j \rangle = 0$ for all $i \neq j$. We have

$$\begin{aligned} \langle \sigma_i u_i | \sigma_i u_i \rangle &= (\sigma_i u_i)^\mathsf{T} (\sigma_j u_j) \\ &= \sigma_i \sigma_j u_j^\mathsf{T} u_i \\ \langle \sigma_i u_i | \sigma_j u_j \rangle &= \langle A v_i | A v_j \rangle \\ &= v_j^\mathsf{T} A^\mathsf{T} A v_i \\ &= v_j^\mathsf{T} \lambda_i v_i \\ &= \lambda_i v_j^\mathsf{T} v_i \\ \sigma_i \sigma_j u_j^\mathsf{T} u_i &= \lambda_i v_j^\mathsf{T} v_i \\ &= 0 \\ u_j^\mathsf{T} u_i &= 0 \\ \langle u_i | u_j \rangle &= 0 \end{aligned}$$

Therefore, the u_i are orthonormal. □

This works because $\lambda_i \geq 0$, so we prove this.

Claim 31. The $\lambda_i \geq 0$.

Proof. If $A^\mathsf{T}Ax = \lambda x$ then $x^\mathsf{T}A^\mathsf{T}Ax = x^\mathsf{T}\lambda x$ so $\|Ax\|_2^2 = \lambda\|x\|_2^2$, so $\lambda \geq 0$ since $\|\cdot\|_2^2 \geq 0$. □

One may note that we have $AV_r = U_r\Sigma_r$ by construction, so then $AV_rV_r^\mathsf{T} = U_r\Sigma_rV_r^\mathsf{T}$. However, $\operatorname{rank}\left(V_rV_r^\mathsf{T}\right) = r \neq n$, so $AV_rV_r^\mathsf{T} \neq I$. This is when we extend $V_r \in \mathbb{R}^{n\times r}$ to $V_n \in \mathbb{R}^{n\times n}$ via Gram Schmidt, and thereby develop the non-compact SVD.

Claim 32. A key fact we will use is that $\operatorname{null}(A) = \operatorname{null}\left(A^\mathsf{T}A\right)$.

Proof. Take a vector $v \in \operatorname{null}(A)$; then $A^\mathsf{T}Av = A^\mathsf{T}0 = 0$, so $v \in \operatorname{null}\left(A^\mathsf{T}A\right)$, so $\operatorname{null}(A) \subseteq \operatorname{null}\left(A^\mathsf{T}A\right)$. Now take a vector $v \in \operatorname{null}\left(A^\mathsf{T}A\right)$; then $0 = v^\mathsf{T}A^\mathsf{T}Av = (Av)^\mathsf{T}(Av) = \|Av\|_2^2$, so $Av = 0$, so $v \in \operatorname{null}(A)$, so $\operatorname{null}\left(A^\mathsf{T}A\right) \subseteq \operatorname{null}(A)$. Therefore $\operatorname{null}(A) = \operatorname{null}\left(A^\mathsf{T}A\right)$. □

Now we are ready to construct the non-compact SVD.

Let $V_n = \begin{bmatrix} V_r & V_e \end{bmatrix}$, where $V_e \in \mathbb{R}^{n\times(n-r)}$ is the collection of orthonormal vectors that span $\mathbb{R}^n/\operatorname{range}(V_r)$, generated by Gram-Schmidt. Then

$$V_nV_n^\mathsf{T} = \begin{bmatrix} V_r & V_g \end{bmatrix}\begin{bmatrix} V_r^\mathsf{T} \\ V_g^\mathsf{T} \end{bmatrix} = V_rV_r^\mathsf{T} + V_gV_g^\mathsf{T}$$

Therefore,

$$AV_nV_n^\mathsf{T} = AV_rV_r^\mathsf{T} + AV_gV_g^\mathsf{T}$$

Claim 33. $\operatorname{range}(V_g) \subseteq \operatorname{null}\left(A^\mathsf{T}A\right)$.

Proof. First, we claim that $\operatorname{range}(V_r) \oplus \operatorname{null}\left(A^\mathsf{T}A\right) = \mathbb{R}^n$. We start with the almost tautological equation $\operatorname{range}\left(A^\mathsf{T}A\right) \oplus \operatorname{range}\left(A^\mathsf{T}A\right)^\perp = \mathbb{R}^n$. Then, we rewrite $\operatorname{range}\left(A^\mathsf{T}A\right) = \operatorname{range}(V_r)$ since V_r has columns that are orthonormal basis vectors for $\operatorname{range}\left(A^\mathsf{T}A\right)$. Then, we write that $\operatorname{range}\left(A^\mathsf{T}A\right)^\perp = \operatorname{null}\left(\left(A^\mathsf{T}A\right)^\mathsf{T}\right) = \operatorname{null}\left(A^\mathsf{T}A\right)$. This yields that $\operatorname{range}(V_r) \oplus \operatorname{null}\left(A^\mathsf{T}A\right) = \mathbb{R}^n$.

Now, we have that $\operatorname{range}(V_g) \cap \operatorname{range}(V_r) = \emptyset$. This follows from construction of V_r. To see why, assume otherwise; then there exists nonzero $v \in \operatorname{range}(V_g) \cap \operatorname{range}(V_r)$, so there exists a linear combination of basis vectors, without all coefficients from one space being 0, in $\operatorname{range}(V_g)$ and $\operatorname{range}(V_r)$ that equals v. But since V_r has a full complement of r basis vectors for its rank r subspace, this means that if $v \in \operatorname{range}(V_r)$, then there would be no need to extend $\operatorname{range}(V_r)$ to $\operatorname{span}(v)$ by Gram-Schmidt so $v \notin \operatorname{range}(V_g)$,a contradiction. □

Since $\operatorname{range}(V_g) \subseteq \operatorname{null}\left(A^\mathsf{T}A\right)$, $\operatorname{range}(V_g) \in \operatorname{null}(A)$ by Claim 32. Hence $AV_gV_g^\mathsf{T} = 0V_g^\mathsf{T} = 0$, so $AV_rV_r^\mathsf{T} = AV_nV_n^\mathsf{T}$. But then V_n has rank n and is orthonormal, so $V_nV_n^\mathsf{T} = I$.

Extending U_r to U_n in a similar manner gives the non-compact SVD.

We consider the geometry of the SVD. In general, if S is the unit circle, then if U is an orthonormal linear transformation, $US = S$. However, if T is a general linear transformation, TS is an ellipse, or a generalization of an ellipse to $\text{range}(A)$ dimensions. We attempt to find the longest axis direction:

$$x^* = \underset{\|x\|_2=1}{\text{argmax}} \|Ax\|_2 = \underset{\|x\|_2=1}{\text{argmax}} x^\mathsf{T} A^\mathsf{T} Ax = v_{\max}\left(A^\mathsf{T} A\right)$$

with the longest squared axis direction being $\lambda_{\max}\left(A^\mathsf{T} A\right)$, so the longest axis direction is $\sigma_1\left(A^\mathsf{T} A\right)$.

Least Norm Solution to Linear System, Pseudoinverses

We now consider the minimum-norm problem, which runs sort of counter to the least squares problem. The least squares problem has a matrix $A \in \mathbb{R}^{m\times n}$, where $m > n$, and we want to approximate the best inverse to A, to solve in x the overdetermined system $Ax = b$. The minimum norm problem has a matrix $A \in \mathbb{R}^{m\times n}$, where $m < n$, and there can be infinitely many solutions to $Ax = b$. We wish to find $\min_x \|x\|_2$ subject to the constraint $Ax = b$. Recall that from Theorem 19, that $\text{null}(A) \oplus \text{range}\left(A^\mathsf{T}\right) = \mathbb{R}^n$. Consider a decomposition of a solution x into $x = x_n + x_r$, where $x_n \in \text{null}(A)$, so $Ax_n = 0$, and $x_r \in \text{range}\left(A^\mathsf{T}\right)$, so $A^\mathsf{T} y = x_r$. Then

$$Ax = A\left(x_n + x_r\right) = A\left(x_n + A^\mathsf{T} y\right) = AA^\mathsf{T} z.$$

Therefore

$$\begin{aligned}
\|x\|_2^2 &= \|x_n + x_r\|_2^2 \\
&= \left\|x_n + A^\mathsf{T} y\right\|_2^2 \\
&= \left(x_n + A^\mathsf{T} y\right)^\mathsf{T}\left(x_n + A^\mathsf{T} y\right) \\
&= \|x_n\|_2^2 + 2\left\langle x_n \middle| A^\mathsf{T} y\right\rangle + \left\|A^\mathsf{T} y\right\|_2^2 \\
&= \|x_n\|_2^2 + \left\|A^\mathsf{T} y\right\|_2^2 && \text{(By Theorem 19)}
\end{aligned}$$

To minimize $\|x\|_2$, we can pick $x_n = 0$ since $Ax = AA^\mathsf{T} y$ is not a function of x_n. Therefore the chosen solution is

$$\begin{aligned}
Ax &= AA^\mathsf{T} y \\
&\overset{\text{set}}{=} b \\
y &= \left(AA^\mathsf{T}\right)^{-1} b
\end{aligned}$$

We have $x = x_n + x_r = x_n + A^\mathsf{T} y$. Since in the minimum norm case $x_n = 0$, then we choose $x = A^\mathsf{T} z = A^\mathsf{T}\left(AA^\mathsf{T}\right)^{-1} b$. Note that we want A to be full rank, or else we can reduce A until it has linearly independent columns.

Consider the compact singular value decomposition of $A = U_r \Sigma_r V_r^\mathsf{T}$, with U_r and V_r orthonormal and Σ_r square and diagonal. Then the formula for x is

$$\begin{aligned}
x &= A^\mathsf{T}\left(AA^\mathsf{T}\right)^{-1} b \\
&= \left(V_r \Sigma_r^\mathsf{T} U_r^\mathsf{T}\right)\left[\left(U_r \Sigma_r V_r^\mathsf{T}\right)\left(V_r \Sigma_r^\mathsf{T} U_r^\mathsf{T}\right)\right]^{-1} b
\end{aligned}$$

$$\begin{aligned}
&= \left(V_r\Sigma_r U_r^\mathsf{T}\right)\left(U_r\Sigma_r V_r^\mathsf{T} V_r\Sigma_r U_r^\mathsf{T}\right)^{-1} b \\
&= V_r\Sigma_r U_r^\mathsf{T}\left(U_r\Sigma_r^2 U_r^\mathsf{T}\right)^{-1} b \\
&= V_r\Sigma_r U_r^\mathsf{T} U_r\Sigma_r^{-2} U_r^\mathsf{T} \\
&= V_r\Sigma_r^{-1} U_r^\mathsf{T} b
\end{aligned}$$

We can take this inverse because

$$\begin{aligned}
\left(U_r\Sigma_r^2 U_r^\mathsf{T}\right)^{-1}\left(U_r\Sigma_r^2 U_r^\mathsf{T}\right) &= U_r\Sigma_r^{-2} U_r^\mathsf{T} U_r\Sigma_r^2 U_r^\mathsf{T} \\
&= U_r U_r^\mathsf{T} \\
&= \begin{bmatrix} I_r & 0_{m-r} \\ 0_{m-r} & 0_{(m-r)^2} \end{bmatrix}
\end{aligned}$$

If A is full row rank (as assumed), then $m = r$ and so $U_r U_r^\mathsf{T} = I_r$.

Definition 34 (Pseudoinverse). If $A = U_r\Sigma_r V_r^\mathsf{T}$ is the compact SVD of A, then $A^\dagger = V_r\Sigma_r^{-1} U_r^\mathsf{T}$ is the (Moore-Penrose) psuedoinverse of A.

Some properties of the psuedoinverse are:

- $AA^\dagger A = A$
- $AA^\dagger = U_r U_r^\mathsf{T}$
- $A^\dagger A = V_r V_R^\mathsf{T}$
- $A^\dagger AA^\dagger = A^\dagger$

Writing $A^\dagger$ in terms of A gives

- If A is invertible, then $A^{-1} = A^\dagger$.
- If A is full row rank, then $A^\dagger = A^\mathsf{T}\left(AA^\mathsf{T}\right)^{-1}$ (the minimum-norm solution, also the right inverse).
- If A is full column rank, then $A^\dagger = \left(A^\mathsf{T}A\right)^{-1} A^\mathsf{T}$ (the least-squares solution, also the left inverse).

Matrix Norms, Low Rank Approximation

We now discuss matrix norms.

Definition 35 (Frobenius Norm). The Frobenius norm is the induced ℓ_2-norm:

$$\|A\|_F = \|A\|_2 = \left(\sum_{i=1}^m\sum_{j=1}^n |a_{i,j}|^2\right)^{1/2} = \operatorname{trace}\left(A^\mathsf{T}A\right)^{1/2}$$

Since it is the induced ℓ_2-norm, it has a nice property.

Claim 36. If U and V are orthonormal matrices, then $\|AV\|_F = \|UA\|_F = \|A\|_F$.

Proof. Pick U and V orthonormal. Then

$$\|UA\|_F^2 = \operatorname{trace}\left((UA)^\mathsf{T}(UA)\right) = \operatorname{trace}\left(A^\mathsf{T}U^\mathsf{T}UA\right) = \operatorname{trace}\left(A^\mathsf{T}A\right) = \|A\|_F^2$$

Also,

$$\|AV\|_F^2 = \operatorname{trace}\left((AV)^\mathsf{T}(AV)\right) = \operatorname{trace}\left(V^\mathsf{T}A^\mathsf{T}AV\right) = \operatorname{trace}\left(A^\mathsf{T}AVV^\mathsf{T}\right) = \operatorname{trace}\left(A^\mathsf{T}A\right) = \|A\|_F^2$$

as desired. □

Definition 37 (Operator Norm). The spectral norm, or operator norm, also called the ℓ_2-norm, gives

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|_2=1} \left(x^\mathsf{T}A^\mathsf{T}Ax\right)^{1/2} = \lambda_{\max}\left(A^\mathsf{T}A\right)^{1/2} = \sigma_{\max}(A)$$

We now discuss the low rank approximation of matrices.

Theorem 38 (Eckart-Young-Mirsky Theorem)
Let $A \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(A) = r$ and $A = U_A\Sigma_A V_A^\mathsf{T}$ be the full SVD of A. Let $A_k = \sum_{i=1}^k \sigma_i(A)u_i(A)v_i(A)^\mathsf{T}$, with $\sigma_1(A) \geq \cdots \geq \sigma_n(A) > 0$. Then

$$A_k = \operatorname*{argmin}_{\substack{B\in\mathbb{R}^{m\times n}\\ \operatorname{rank}(B)=k}} \|A-B\|_2 = \operatorname*{argmin}_{\substack{B\in\mathbb{R}^{m\times n}\\ \operatorname{rank}(B)=k}} \|A-B\|_F.$$

Proof for Spectral Norm. We want to show that $A_k = \sum_{i=1}^k \sigma_i(A)u_i(A)v_i(A)^\mathsf{T} = \operatorname{argmin}_{\substack{B\in\mathbb{R}^{m\times n}\\ \operatorname{rank}(B)=k}} \|A-B\|_2$.

Consider

$$\begin{aligned}\|A-A_k\|_2 &= \left\|\sum_{i=k+1}^n \sigma_i(A)u_i(A)v_i(A)^\mathsf{T}\right\|_2 \\ &= \sigma_{k+1}(A) \qquad \text{(Spectral re-composition of } \textstyle\sum_{i=k+1}^n \sigma_i u_i(A)v_i(A)^\mathsf{T}\text{, then Definition 37.)}\end{aligned}$$

We want to show that for any other matrix B with $\operatorname{rank}(B) = k$, that $\|A-B\|_2 \geq \sigma_{k+1}(A)$. Since $\operatorname{rank}(B) = k$, then $\dim(\operatorname{null}(B)) = n-k > 0$. We have that since $\|A\|_2 = \max_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2}$, then $\|A\|_2 \geq \frac{\|Ax\|_2}{\|x\|_2}$ for any x, so $\|A\|_2\|x\|_2 \geq \|Ax\|_2$. Also, define $V_{k+1} = \begin{bmatrix} v_1 & \cdots & v_{k+1}\end{bmatrix}$, and $\operatorname{range}(V_{k+1}) = k+1$. Therefore by Pidgeonhole principle $\operatorname{null}(B) \cap \operatorname{range}(V_{k+1}) \neq \emptyset$. Therefore

$$\begin{aligned}\|A-B\|_2^2 &\geq \|(A-B)w\|_2^2 && (\|w\|_2 = 1.)\\ &= \|Aw\|_2^2 && (\text{Pick } w \in \operatorname{null}(B).)\\ &= \left\|A\sum_{i=1}^{k+1}\alpha_i v_i(A)\right\|_2^2 && (\text{Pick } w \in \operatorname{null}(B)\cap\operatorname{range}(V_{k+1}).)\\ &= \left\|U_A\Sigma_A V_A^\mathsf{T}\sum_{i=1}^{k+1}\alpha_i v_i(A)\right\|_2^2 \\ &= \left\|U_A\Sigma_A V_A^\mathsf{T}V\alpha\right\|_2^2 \\ &= \|U_A\Sigma_A\alpha\|_2^2 \\ &= \|\Sigma_A\alpha\|_2^2 \\ &= \sum_{i=1}^{k+1}\alpha_i^2\sigma_i(A)^2\end{aligned}$$

$$\geq \sigma_{k+1}(A)^2 \left(\sum_{i=1}^{n} \alpha_i^2\right)$$
$$\geq \sigma_{k+1}(A)^2 \qquad \text{(Since } \|w\|_2 = 1, \textstyle\sum_{i=1}^{k+1} \alpha_i^2 = 1.\text{)}$$

as desired. □

Proof for Frobenius Norm. We want to show that $A_k = \sum_{i=1}^{k} \sigma_i(A) u_i(A) v_i(A)^\mathsf{T} = \operatorname{argmin}_{\substack{B \in \mathbb{R}^{m \times n} \\ \operatorname{rank}(B) = k}} \|A - B\|_F$.

Consider

$$\|A - A_k\|_F^2 = \left\| \sum_{i=1}^{n} \sigma_i(A) u_i(A) v_i(A)^\mathsf{T} - \sum_{i=1}^{k} \sigma_i(A) u_i(A) v_i(A)^\mathsf{T} \right\|_F^2$$
$$= \left\| \sum_{i=k+1}^{n} \sigma_i(A) u_i(A) v_i(A)^\mathsf{T} \right\|_F^2$$

Recall that the Frobenius norm is invariant to an orthonormal transformation. In particular, if U and W are orthonormal matrices, $\|A\|_F = \|UA\|_F = \|AW\|_F$. So then if $A = U_A \Sigma_A V_A^\mathsf{T}$ is the singular value decomposition of A,

$$\|A\|_F^2 = \left\| U_A \Sigma_A V_A^\mathsf{T} \right\|_F^2 = \|\Sigma_A\|_F^2 = \sum_{i=1}^{n} \sigma_i(A)^2.$$

Therefore

$$\|A - A_k\|_F^2 = \left\| \sum_{i=k+1}^{n} \sigma_i(A) u_i(A) v_i(A)^\mathsf{T} \right\|_F^2$$
$$= \sum_{i=k+1}^{n} \sigma_i(A)^2$$

Now we want to show that for any other matrix B with $\operatorname{rank}(B) = k$, that $\|A - B\|_F^2 \geq \sum_{i=k+1}^{n} \sigma_i(A)^2$. In particular, we want to show that

$$\|A - B\|_F^2 = \sum_{i=1}^{n} \sigma_i(A - B)^2 \geq \sum_{i=k+1}^{n} \sigma_i(A)^2 = \|A - A_k\|_F^2$$

In particular, it's sufficient to show that $\sigma_i(A - B) \geq \sigma_{k+i}(A)$.

Define $A - B = C$ and $C_j = \sum_{i=1}^{j} \sigma_i(C) u_i(C) v_i(C)^\mathsf{T}$, then $\sigma_i(C) = \|C - C_{i-1}\|_2$. Also we write $\sigma_{k+i}(A) = \|A - A_{k+i-1}\|_2 = \|B + C - A_{k+i-1}\|_2$. Then

$$\sigma_i(A - B) \geq \|C - C_i\|_2$$
$$\geq \|A - B - C_i\|_2$$
$$\geq \|A - A_{k+i}\|_2 \qquad \text{(}\operatorname{rank}(B + C_i) \leq \operatorname{rank}(B) + \operatorname{rank}(C_i) = k + i.\text{)}$$
$$\geq \sigma_{k+i}(A)$$

which completes the result. □

5 Dealing With Noise

We've already covered least squares as a way to deal with noise. The methods we discuss will cover additional methods of dealing with noise.

Let $Ax = y$, where $A \in \mathbb{R}^{n\times n}$ is invertible. If $y \to y + \Delta_y$, and because of this $x \to x + \Delta_x$, we want to find the relative change in Δ_x i.e. $\frac{\|\Delta_x\|_2}{\|x\|_2}$. We have

$$\begin{aligned} A(x + \Delta_x) &= y + \Delta_y \\ A\Delta_x &= \Delta_y \\ \Delta_x &= A^{-1}\Delta_y \\ \|\Delta_x\|_2 &= \left\|A^{-1}\Delta_y\right\|_2 \\ &\le \left\|A^{-1}\right\|_2 \|\Delta_y\|_2 && \text{(Definition of spectral norm.)} \end{aligned}$$

We also desire to bound $\|x\|_2$.

$$\begin{aligned} Ax &= y \\ \|Ax\|_2 &= \|y\|_2 \\ &\le \|A\|_2\|x\|_2 \end{aligned}$$

Therefore

$$\begin{aligned} \frac{\|\Delta_x\|}{\|x\|} &= \frac{\left\|A^{-1}\right\|_2\|\Delta_y\|_2}{\|x\|_2} \\ &\le \frac{\left\|A^{-1}\right\|_2\|\Delta_y\|_2}{\|y\|_2/\|A\|_2} \\ &\le \|A\|_2\left\|A^{-1}\right\|_2\left(\frac{\|\Delta_y\|_2}{\|y\|_2}\right) \end{aligned}$$

We know that $\|A\|_2 = \sigma_{\max}(A)$ and $\left\|A^{-1}\right\|_2 = \sigma_{\min}(A)^{-1}$, so

$$\frac{\|\Delta_x\|_2}{\|x\|_2} \le \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \cdot \frac{\|\Delta_y\|_2}{\|y\|_2}$$

Definition 39 (Condition Number). The *condition number* of A is the quantity $K(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$.

The idea is that a lower condition number means a system given by $Ax = b$ is more resistant to noise.

Example 40. Let's look at least squares in terms of noise resistance. The full solution is that $x^* = \left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T}b$, and the normal equation is $A^\mathsf{T}Ax^* = A^\mathsf{T}b$. We care about the stability of this system. We have

$$K\left(A^\mathsf{T}A\right) = \frac{\sigma_{\max}\left(A^\mathsf{T}A\right)}{\sigma_{\min}(A^\mathsf{T}A)} = \frac{\sigma_{\max}(A)^2}{\sigma_{\min}(A)^2} = \left(\frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}\right)^2 = \kappa(A)^2$$

due to the definition of singular values. We observe that this can create instability due to noise very fast (quadratically) compared to the underlying system $Ax = b$.

This is important because it motivates *ridge regression*. We wish to change the singular values of $A^\mathsf{T}A$ so that we get a better condition vector.

Claim 41. If $A \in \mathbb{R}^{n \times n}$, then $\lambda_i(A + \lambda I) = \lambda_i(A) + \lambda$.

Proof. Let $v_i(A)$ be such that $Av_i(A) = \lambda_i(A)v_i(A)$. Then

$$(A + \lambda I)\, v_i(A) = Av_i(A) + \lambda I v_i(A) = \lambda_i(A)v_i(A) + \lambda v_i(A) = (\lambda_i(A) + \lambda)\, v_i(A)$$

as desired. □

Notice that the least squares problem is the optimization problem

$$\min_x \|Ax - b\|_2^2$$

We consider the new *regularized* optimization problem

$$\min_x \|Ax - b\|_2^2 + \underbrace{\lambda^2 \|x\|_2^2}_{\text{penalty term}}$$

The λ parameter trades off how much we care about $\|x\|_2^2$ being small against our accuracy. The penalty term is also called a *regularizer*.

This is a quadratic convex problem. Let $f(x) = \|Ax - b\|_2^2 + \lambda^2 \|x\|_2^2$. Then expanding we have

$$\begin{aligned}
f(x) &= \|Ax - b\|_2^2 + \lambda^2 \|x\|_2^2 \\
&= (Ax - b)^\mathsf{T} (Ax - b) + \lambda^2 x^\mathsf{T} x \\
&= x^\mathsf{T} A^\mathsf{T} A x - x^\mathsf{T} A^\mathsf{T} b - b^\mathsf{T} A x + b^\mathsf{T} b + \lambda^2 x^\mathsf{T} x \\
\boldsymbol{\nabla}_x f(x) &= \boldsymbol{\nabla}_x \left(x^\mathsf{T} A^\mathsf{T} A x \right) - 2 \boldsymbol{\nabla}_x \left(b^\mathsf{T} A x \right) + \boldsymbol{\nabla}_x \left(b^\mathsf{T} b \right) + \boldsymbol{\nabla}_x \left(\lambda^2 x^\mathsf{T} x \right) \\
&= 2A^\mathsf{T} A x - 2A^\mathsf{T} b + 2\lambda^2 x \\
&\overset{\text{set}}{=} 0 \\
2A^\mathsf{T} A x + 2\lambda^2 x &= 2A^\mathsf{T} b \\
\left(A^\mathsf{T} A + \lambda^2 I \right) x^* &= A^\mathsf{T} b \\
x^* &= \left(A^\mathsf{T} A + \lambda^2 I \right)^{-1} A^\mathsf{T} b
\end{aligned}$$

This is the solution to the ridge regression problem, or ℓ_2 regularized least squares. One can see that $K\left(A^\mathsf{T} A + \lambda^2 I\right) < K\left(A^\mathsf{T} A\right)$, showing the resistance to noise of this system.

Another interpretation of the ridge regression is that we can encode prior information on the norm of x. If we have information that the i^{th} coordinate of x has $\lambda x_i \approx 0$ for some $\lambda \neq 0$, then we can augment A to obtain

$$\underbrace{\begin{bmatrix} A \\ \lambda I \end{bmatrix}}_{\widetilde{A}} x = \underbrace{\begin{bmatrix} b \\ 0 \end{bmatrix}}_{\widetilde{b}}$$

The least squares formulation of this matrix has

$$x^* = \left(\widetilde{A}^\mathsf{T} \widetilde{A} \right)^{-1} \widetilde{A}^\mathsf{T} b = \begin{bmatrix} A^\mathsf{T} & \lambda I \end{bmatrix} \begin{bmatrix} A \\ \lambda I \end{bmatrix} \begin{bmatrix} A^\mathsf{T} & \lambda I \end{bmatrix} \begin{bmatrix} b \\ 0 \end{bmatrix} = \left(A^\mathsf{T} A + \lambda^2 I \right)^{-1} A^\mathsf{T} b,$$

What if we had some information that the i^{th} coordinate of x has $\lambda x_i \approx x_i'$? Then we can generalize, and

obtain the matrix equation

$$\begin{bmatrix} A \\ \lambda I \end{bmatrix} x = \begin{bmatrix} b \\ x_0 \end{bmatrix}$$

We can also weight the different rows of the data matrix.

Definition 42 (Tikhonov Regularization). Let W_1, W_2 be weight matrices, usually diagonal. Then the *Tikhonov regularization* problem is

$$\min_x \|W_1 (Ax - b)\|_2^2 + \|W_2 (x - x')\|_2^2$$

for the side information vector x'.

These two different perspectives are first, to improve the condition number of A, and second, to induce prior information into our system.

The next goal is to use a probabilistic model to choose our hyperparameter for i.e. ridge regression. The motivation for this is that we want to include probabilistic information into the model.

Let's say we have $\{(x_i, y_i)\}_{i=1}^m$ be the data points, where $y_i = g(x_i) + Z_i$, where $Z_i \sim \text{Normal}(0, \sigma_i^2)$, and $p_{Z_i}(z) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{-z_i^2}{2\sigma_i^2}\right)$. Dealing with general $g \in L^2[x]$ is difficult, so we restrict g to be linear. In particular, let $g(x) = w^\mathsf{T} x$, and w is the set of model parameters (what we want to learn), so that $y_i = w^\mathsf{T} x_i + Z_i$.

Least squares finds w such that $\sum_{i=1}^m \left(w^\mathsf{T} x_i - y_i\right)^2$ is minimized. We incorporate the probabilistic information into this model using *maximum likelihood estimation*. We find the w that maximizes the probability density of the data:

$$\text{MLE}(w \mid x, y) = \operatorname*{argmax}_w p_{X,Y|w}(x, y \mid w)$$

In this particular case, X is deterministic, so

$$\begin{aligned}
\text{MLE}(w \mid x, y) &= \text{MLE}(w \mid y) \\
&= \operatorname*{argmax}_w p_{Y|w}(y \mid w) \\
&= \operatorname*{argmax}_w \prod_{i=1}^m p_{Y_i|w}(y_i \mid w) \\
&= \operatorname*{argmax}_w \prod_{i=1}^m p_{Z_i|w}\left(y_i - w^\mathsf{T} x_i \,\middle|\, w\right) \\
&= \operatorname*{argmax}_w \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\left(y_i - w^\mathsf{T} x_i\right)^2}{2\sigma_i^2}\right) \\
&= \operatorname*{argmax}_w \frac{1}{\left(\sqrt{2\pi}\right)^n} \frac{1}{\prod_{i=1}^m \sigma_i} \exp\left(-\sum_{i=1}^m \frac{\left(y_i - w^\mathsf{T} x_i\right)^2}{2\sigma_i^2}\right) \\
&= \operatorname*{argmin}_w \sum_{i=1}^m \frac{\left(y_i - w^\mathsf{T} x_i\right)^2}{2\sigma_i^2} \\
&= \operatorname*{argmin}_w \|S (y - xw)\|_2^2
\end{aligned}$$

where $S = \operatorname{diag}\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}, \dots, \frac{1}{\sqrt{2\pi\sigma_m^2}}\right)$. This is related to Tikhonov regularization and weighted least squares, and suggests that a way of interpreting least squares is to see which components or features matter more to the final prediction.

If we have a prior on w, we can use *maximum a posteriori estimation*:

$$\operatorname{MAP}(w \mid x, y) = \operatorname*{argmax}_w p_{w|X,Y}(w \mid x, y) = \operatorname*{argmax}_w p_{X,Y|w}(x, y \mid w) p_w(w)$$

In this case, assume that $w \sim \operatorname{Normal}(\mu, \Sigma_w)$ is the prior distribution on w. Then

$$\begin{aligned}
\operatorname{MAP}(w \mid x, y) &= \operatorname*{argmax}_w p_{Y|w}(y \mid w) p_w(w) \\
&= \operatorname*{argmax}_w p_w \prod_{i=1}^m p_{Y|w}(w \mid w) \\
&= \operatorname*{argmax}_w \left[\prod_{i=1}^m \frac{\exp\left(-\frac{\left(y_i - w^\mathsf{T} x_i\right)^2}{2\sigma_i^2}\right)}{\sqrt{2\pi\sigma_i}}\right] \left[\frac{\exp\left(-\left(w-\mu\right)^\mathsf{T} \Sigma_w^{-1} \left(w-\mu\right)\right)}{\sqrt{(2\pi)^m \det(\Sigma_w)}}\right] \\
&= \operatorname*{argmin}_w \exp\left(\sum_{i=1}^m \frac{\left(y_i - w^\mathsf{T} x_i\right)^2}{2\sigma_i^2} + \left(w-\mu\right)^\mathsf{T} \Sigma_w^{-1} \left(w-\mu\right)\right) \\
&= \operatorname*{argmin}_w \|S\left(xw - y\right)\|_2^2 + \left\|\Sigma_2^{-1/2}\left(w-\mu\right)\right\|_2^2
\end{aligned}$$

The $\Sigma_w^{-1/2}$ is similar to the regularization term, and enforces the viewpoint that the regularizer enforces prior beliefs and certainty of those beliefs.

We use this to introduce *principal component regression*, which is the technique of projecting onto low dimensional principal components and doing regression from there.

The problem is to minimize $\|Xw - y\|_2^2$. If $X = U_X \Sigma_X V_X^\mathsf{T}$, then the least squares solution is $w^* = V\Sigma^\dagger U^\mathsf{T} y$, where $\Sigma_{i,j}^\dagger = \frac{\delta_{i,j}}{\Sigma_{i,i}}$. For principal components regression, we do the regression with only the top k principal components, i.e., $w^* = V\Sigma^{\text{PCR}} U^\mathsf{T} y$ where $\Sigma^{\text{PCR}} = \frac{\mathbb{1}(i=j \wedge i \le k)}{\Sigma_{i,i}}$.

There's yet another perspective on ridge regression, which is motivated by this. In particular, ridge regression is a soft form of principal components analysis. Starting from the ridge regression standpoint, we have

$$\begin{aligned}
w^* &= \operatorname*{argmin}_w \|Xw - y\|_2^2 + \lambda \|w\|_2^2 \\
&= \operatorname*{argmin}_{w = V_X z} \|XV_X z - y\|_2^2 + \lambda \|V_X z\|_2^2 && (X = U_X \Sigma_X V_X^\mathsf{T}) \\
&= \operatorname*{argmin}_z \|XVz - y\|_2^2 + \lambda \|z\|_2^2 \\
z^* &= \left[\left(XV_X\right)^\mathsf{T} \left(XV_X\right) + \lambda I\right]^{-1} \left(XV_X\right)^\mathsf{T} y \\
&= \left(V_X^\mathsf{T} X^\mathsf{T} X V_X + \lambda I\right)^{-1} \left(XV_X\right)^\mathsf{T} y \\
&= \left[V_X^\mathsf{T} \left(U_X \Sigma_X V_X^\mathsf{T}\right) V_X + \lambda I\right]^{-1} \left(XV_X\right)^{-1} y \\
&= \left(\Sigma_X^\mathsf{T} \Sigma_X + \lambda I\right)^{-1} \Sigma_X^\mathsf{T} U^\mathsf{T} y
\end{aligned}$$

$$= \begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\lambda} & & \\ & \ddots & \\ & & \frac{\sigma_n}{\sigma_n^2+\lambda} \end{bmatrix} U^\mathsf{T} y$$

and one can read off the asymptotic behavior of z^* as a function of λ.

6 Convexity

Today we discuss convex sets and convex functions. Convex problems are particularly nice because there exists a unique optimum point, which can be computed in closed form if the objective function is differentiable.

Definition 43 (Covex Combination). A linear combination $\sum_{i=1}^n a_i x_i$ is a *convex combination* if $\sum_{i=1}^n a_i = 1$ and $a_i \geq 0$ for all i.

Definition 44 (Convex Set). A set C is *convex* if for any set of points $S = \{x_1, \ldots, x_n\}$ any convex combination of S is in C. Equivalently, C is convex if for any $x, y \in C$, then $\alpha x + (1-\alpha) y \in C$, for $\alpha \in [0, 1]$.

Pictorially, C is convex if any line between two points in C consists only of points in C.

Definition 45 (Convex Hull). The *convex hull* of a set of points $S = \{x_1, \ldots, x_n\}$ is the minimal-measure convex set that has S as a subset.

Example 46. Let $C = \left\{x \,\middle|\, a^\mathsf{T} x = b\right\}$. Then we claim C is convex. Take $x_1, x_2 \in C$. Let $\alpha \in [0, 1]$ and $x_3 = \alpha x_1 + (1-\alpha) x_2$. Then $a^\mathsf{T} x_3 = \alpha a^\mathsf{T} x_1 + a^\mathsf{T} x_2 - \alpha a^\mathsf{T} x_2 = \alpha a^\mathsf{T} x_1 - a^\mathsf{T} x_2 + a^\mathsf{T} x_2 = a^\mathsf{T} x_2 = b$, so $x_3 \in C$, so C is convex.

Definition 47 (Half Space). A *half space* corresponding to vectors a and b is the set $\left\{x \,\middle|\, a^\mathsf{T} x \geq b\right\}$ or $\left\{x \,\middle|\, a^\mathsf{T} x \leq b\right\}$.

Example 48. We claim $\mathbb{S}_{\geq 0}^n$ (the set of positive semidefinite matrices) is convex. Take $\alpha \in [0, 1]$. Let $A_1, A_2 \in \mathbb{S}_{\geq 0}^n$ and $A_3 = \alpha A_1 + (1-\alpha) A_2$. Then for any x we have $\alpha x^\mathsf{T} A_1 x + (1-\alpha) x^\mathsf{T} A_2 x \geq \alpha \cdot 0 + (1-\alpha) \cdot 0 = 0$, so $A_3 \in \mathbb{S}_{\geq 0}^n$, so $\mathbb{S}_{\geq 0}^n$ is convex.

Theorem 49 (Separating Hyperplane Theorem)
Let $C, D \subseteq \mathbb{R}^n$ be two convex, compact sets with $C \cap D = \emptyset$. Then there exists a hyperplane $a^\mathsf{T} x = b$, such that

- for all $x \in C$ we have $a^\mathsf{T} x \geq b$.
- for all $x \in D$ we have $a^\mathsf{T} x \leq b$.

Proof. Define $d(C, D) = \inf \{\|c - d\|_2 \,|\, c \in C, d \in D\}$. Let c and d be the minimizing points; we know that such a pair exists since C and D are compact. Then the hyperplane with $a = d - c$ as the normal vector and that passes through $x_0 = \text{midpoint}(c, d) = \frac{c+d}{2}$ is given by

$$f(x) = a^\mathsf{T} (x - x_0)$$

$$= (d-c)^{\mathsf{T}}\left(x - \frac{c+d}{2}\right)$$

We have that $f(d) = (d-c)^{\mathsf{T}}\left(d - \frac{c+d}{2}\right) = \frac{1}{2}\|d-c\|_2^2 \geq 0$. Similarly $f(c) = -\frac{1}{2}\|d-c\|_2^2$.

We want to show that for all $x \in D$ that $f(x) \geq 0$. By symmetry this shows that for all $x \in C$ that $f(x) \leq 0$.

Assume for the sake of contradiction that $u \in D$ such that $f(u) < 0$. Then

$$\begin{aligned}
f(u) &= (d-c)^{\mathsf{T}}\left(u - \frac{c+d}{2}\right) \\
&= (d-c)\left(u - d + d - \frac{c+d}{2}\right) \\
&= (d-c)^{\mathsf{T}}\left[(u-d) + \frac{d-c}{2}\right] \\
&= (d-c)^{\mathsf{T}}(u-d) + \frac{1}{2}\|d-c\|_2^2 \\
&= \langle d-c|u-d\rangle + \frac{1}{2}\|d-c\|_2^2 \\
&< 0
\end{aligned}$$

Therefore $\langle d-c|u-d\rangle < 0$.

Let $t \in [0,1]$; then define $p = d + t(u-d)$. Clearly $p \in D$. Then

$$\begin{aligned}
\|c-p\|_2^2 &= \|c-d-t(u-d)\|_2^2 \\
&= [(c-d)-t(u-d)]^{\mathsf{T}}[(c-d)-t(u-d)] \\
&= \|c-d\|_2^2 + t^2\|u-d\|_2^2 - 2\langle c-d|t(u-d)\rangle \\
&= \|c-d\|_2^2 + t^2\|u-d\|_2^2 - 2t\langle c-d|c-d\rangle
\end{aligned}$$

We want $t^2\|u-d\|_2^2 - 2t\langle c-d|(u-d)\rangle < 0$, so we want $2t\langle d-c|u-d\rangle + t^2\|u-d\|_2^2 < 0$, so we want $2\langle d-c|u-d\rangle + t\|u-d\|_2^2 < 0$. Since we can choose t and $\langle d-c|u-d\rangle < 0$, choosing $t < 2\left|\frac{\langle d-c|u-d\rangle}{\|u-d\|_2^2}\right|$ shows that $\|c-p\|_2^2 \leq 0$. But then $\|c-p\|_2^2 < \|c-d\|_2^2$, which is a contradiction with the definition of c and d. □

Definition 50 (Convex Function). A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex if for all $\alpha \in [0,1]$, $x, y \in \mathbb{R}^n$, we have

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

Definition 51 (Epigraph). The epigraph of f is the set

$$\operatorname{epi}(f) = \{(x,t)\colon x \in \operatorname{domain}(f), f(x) < t\}$$

f is a convex function if and only if $\operatorname{epi}(f)$ is a convex set.

Theorem 52 (First-Order Conditions)
let $f \in C^1(\mathbb{R}^n, \mathbb{R})$. Then f is convex if and only if $\operatorname{domain}(f)$ is convex and for each $x, y \in \mathbb{R}^n$

$$f(y) \geq f(x) + \nabla_x f(x)^{\mathsf{T}}(y-x)$$

This is what gives convexity its power. The derivative (gradient) of f is a local property at x; however, we can find global properties of f if f is convex. In particular, if f is convex and $\nabla_x f(x^*) = 0$ then $f(y) \geq f(x^*) + 0$ for all x, so $f(x^*)$ is a global minimum.

Theorem 53 (Second-Order Conditions)
Let $f \in C^2(\mathbb{R}^n, \mathbb{R})$. Then f is convex if and only if $\text{domain}(f)$ is convex and $\nabla_x^2 f(x) \succeq 0$ (is positive semidefinite).

Theorem 54 (Jensen's Inequality for Probability)
Let f be a convex function and X a random variable. Then

$$f(\text{Ex}[X]) \leq \text{Ex}[f(X)]$$

In the discrete non-probabilistic case, we have

Theorem 55 (Jensen's Inequality)
Let f be a convex function and $\lambda_1, \dots, \lambda_n$ be such that $\sum_{i=1}^n \lambda_i = 1$. Then for all $x_1, \dots, x_n \in \mathbb{R}$, we have

$$f\left(\sum_{i=1}^n \lambda_i x_i\right) \leq \sum_{i=1}^n \lambda_i f(x_i)$$

The proof is by definitions of convexity.

A general optimization problem is

$$\begin{aligned} x^* &= \operatorname*{argmin}_x f(x) \\ \text{s.t. } f_i(x) &\leq 0, \quad i \in [1, m] \\ g_i(x) &= 0, \quad i \in [1, \ell] \end{aligned}$$

Definition 56 (Convex Problem). A *convex problem* is an optimization problem where $f(x)$ and $f_i(x)$ and $g_i(x)$ are all convex, and the feasible region $D = \text{domain}(f) \cap \bigcap_{i=1} \text{domain}(f_i) \cap \bigcap_{i=1}^{\ell} \text{domain}(g_i)$ is convex.

Example 57 (Linear Program). A linear program

$$\begin{aligned} x^* &= \operatorname*{argmin}_x c^\mathsf{T} x \\ \text{s.t. } Ax &= b \end{aligned}$$

is a convex problem.

Example 58. The optimization problem

$$\begin{aligned} x^* &= \operatorname*{argmin}_x x^2 \\ \text{s.t. } x &\geq 1 \\ x &\leq 2 \end{aligned}$$

is a convex problem; in this case the constraint $x \geq 1$ is called a *active constraint*.

In general, if $f(x)$ is convex, then to find the optimum point x^* we set $[\nabla_x f(x)]_{x=x*} = 0$.

Example 59. The optimization problem

$$\begin{aligned} x^* &= \operatorname*{argmin}_x x^2 \\ \text{s.t. } & x \geq 1 \\ & x \leq -2 \end{aligned}$$

is *infeasible*; the feasible region $D = \emptyset$.

Example 60. The optimization problem

$$\begin{aligned} x^* &= \operatorname*{argmin}_x x_1 + x_2 \\ \text{s.t. } & x_1^2 \geq 2 \\ & x_2^2 \leq 1 \end{aligned}$$

is a convex problem; taking the gradient gives $\nabla_x f(x) = 1$ (the ones vector), but by geometry or observation one sees that $x^* = \begin{bmatrix} -\sqrt{2} \\ -1 \end{bmatrix}$ works. Both constraints are active.

Example 61. The optimization problem

$$\begin{aligned} x^* &= \operatorname*{argmin}_x x_1 \\ \text{s.t. } & x_1 + x_2 \geq 0 \end{aligned}$$

is *unbounded*; we can take x_1 arbitrarily low and consider $x_2 = -x_1$.

At this point it's useful to consider inf vs min; $\inf S$ need not be attained by any value in S.

Theorem 62
Let X be a convex and closed set. Then if $x^* = \operatorname{argmin}_{x \in X} c^\mathsf{T} x$ then $x \in \overline{X} - X^o$, the set of limit points of X.

Proof. Assume for the sake of contradiction that $x^* \in X^o$, the interior of X. Then there exists r such that there exists some ball $\{z \in X : \|x^* - x\|_2 < r\} = B_r(x^*) \subseteq X^o$. We have that $\nabla_x f(x) = c$. Consider $z = \frac{rc}{\|c\|_2} \in B_r(x^*)$, so $z \in X^o$. Then

$$\begin{aligned} f(x^* - z) &= c^\mathsf{T} x^* - c^\mathsf{T} z \\ &= c^\mathsf{T} x^* - \frac{r}{\|c\|_2} c^\mathsf{T} c \\ &= c^\mathsf{T} x^* - r\|c\|_2 \\ &\leq c^\mathsf{T} x^* \end{aligned}$$

which has a better optimum, a contradiction. □

We define the *epigraph reformulation* as

$$\min_{x \in X} f(x) = \min_{f(x) \leq t} t$$

where t is known as a *slack variable*.

Definition 63 (LASSO Regression). We now introduce LASSO regression, or ℓ^1 normalization. We have the optimization problem

$$x^* = \operatorname*{argmin}_x \|Ax - y\|_2^2 + \|x\|_1$$

Adding the epigraph reformulation, we have

$$\begin{aligned} x^* = \operatorname*{argmin}_{x,t} \|Ax - y\|_2^2 + \sum_{i=1}^n t_i \\ \text{s.t.} \quad -t_i \le x_i \le t_i, \quad i \in [1,n] \\ t_i \ge 0, \quad i \in [1,n] \end{aligned}$$

Definition 64 (Monotone Transformation). If $\phi(x)$ is continuous and strictly increasing, then over a domain D we have that $\operatorname{argmin}_{x\in D} f(x) = \operatorname{argmin}_{x\in D} \phi(f(x))$, and similarly for monotone decreasing functions.

Definition 65 (Logistic Regression). Say that we have the data points $X_1, \dots, X_m$ and corresponding labels $y_1, \dots, y_m \in \{-1, 1\}$. We wish to predict $p_{Y|X}(1 \mid X_i)$. We achieve this by learning w and β such that $p_{Y|X}(1 \mid X_i) = \sigma\Big(w^\mathsf{T} X_i + \beta\Big)$ where $\sigma(x) = \frac{1}{1+\mathrm{e}^{-x}}$. In particular, $p_{Y|X}(y_i \mid X_i) = \sigma\Big(-y_i\Big(w^\mathsf{T} X_i + \beta\Big)\Big)$.

How do we solve for w and β? We use maximum likelihood estimation. We have

$$\begin{aligned} (w^*, \beta^*) &= \operatorname*{argmax}_{w,\beta} \prod_{i=1}^m p_{Y|X}(y_i \mid X_i) \\ &= \operatorname*{argmax}_{w,\beta} \prod_{i=1}^n \sigma\Big(-y_i\Big(w^\mathsf{T} X_i + \beta\Big)\Big) \\ &= \operatorname*{argmax}_{w,\beta} \sum_{i=1}^m \log\Big(\sigma\Big(-y_i\Big(w^\mathsf{T} X_i + \beta\Big)\Big)\Big) \end{aligned}$$

and the rest is computation.

7 Optimization Algorithms

Gradient Descent

For the purposes of this lecture, we consider unconstrained optimization problems of the form

$$x^* = \min_{x\in\mathbb{R}^n} f(x)$$

For most functions f, even if f is convex, there is no closed-form solution. To optimize globally these functions, we use gradient descent. This is a consequence of Taylor's theorem, which states that

$$\begin{aligned} f(x + \Delta_x) &= f(x) + \boldsymbol{\nabla}_x f(x)^\mathsf{T} \Delta_x + o\Big(\|\Delta_x\|_2^3\Big) \\ f(x + su) &= f(x) + \boldsymbol{\nabla}_x f(x)^\mathsf{T} su \\ &= f(x) + s\,\langle \boldsymbol{\nabla}_x f(x) | u \rangle \end{aligned}$$

We want to find the minimum, so we want $\langle \boldsymbol{\nabla}_x f(x) | u \rangle < 0$. By Cauchy-Schwarz, we see that the minimizing direction is $u^* = -\boldsymbol{\nabla}_x f(x)$.

This yields an algorithm:

Algorithm 3 Gradient descent algorithm.

Input: Initial vector $x^{(0)}$, differentiable function $f \in C^1(\mathbb{R}^n, \mathbb{R})$, learning rate η.
Output: Local minimum of f.

```
k ← 0
while ∇_{x^(k)} f(x^(k)) ≠ 0 do
    x^(k+1) ← x^(k) − η ∇_{x^(k)} f(x^(k))
    k ← k + 1
return x^(k)
```

The interesting questions are how and when gradient descent converges to the minimal point, i.e. when $x^* = \textsc{GradientDescent}\left(x^{(0)}, f, \eta\right)$.

Example 66 (Least Squares Gradient Descent). Let $f(x) = \|Ax - b\|_2^2$, the least squares objective. Then $\nabla_x f(x) = 2A^\mathsf{T}Ax - 2A^\mathsf{T}b$. Then the gradient update rule is

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \eta \nabla_{x^{(k)}} f\left(x^{(k)}\right) \\ &= x^{(k)} - \eta\left(2A^\mathsf{T}Ax^{(k)} - 2A^\mathsf{T}b\right) \\ &= \left(I - 2\eta A^\mathsf{T}A\right)x^{(k)} + 2\eta A^\mathsf{T}b \end{aligned}$$

Recalling the analysis of stability, we want $\left|\lambda_i\left(I - 2\eta A^\mathsf{T}A\right)\right| < 1$.

If A is full column rank, then

$$\begin{aligned} x^{k+1} - x^* &= x^{(k+1)} - \left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T}b \\ &= \left(I - 2\eta A^\mathsf{T}A\right)x^{(k)} + 2\eta A^\mathsf{T}b - \left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T}b \\ &= \left(I - 2\eta A^\mathsf{T}A\right)x^{(k)} + 2\eta\left(A^\mathsf{T}A\right)\left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T}b - \left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T}b \\ &= \left(I - 2\eta A^\mathsf{T}A\right)x^{(k)} + \left(2\eta A^\mathsf{T}A - I\right)\left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T}b \\ &= \left(I - 2\eta A^\mathsf{T}A\right)\left[x^{(k)} - \left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T}b\right] \\ &= \left(I - 2\eta A^\mathsf{T}A\right)^{k+1}\left[x^{(0)} - \left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T}b\right] \end{aligned}$$

We want to figure out for which η that $\left|\lambda_i\left(I - 2\eta A^\mathsf{T}A\right)\right| \leq 1$; we'll do this later.

For very large A, computing gradient descent is lower cost computationally than computing the closed form iterates.

We want to generalize the idea of least squares to smooth, strongly convex functions.

Definition 67 (μ-Strongly Convex). A function f is μ-strongly convex on a domain D if for all $x, y \in D$

$$f(y) \geq f(x) + \nabla_x f(x)^\mathsf{T}(y - x) + \frac{\mu}{2}\|y - x\|_2^2.$$

Definition 68 (L-Smoooth). A function f is L-smooth on a domain D if for all $x, y \in D$

$$f(y) \leq f(x) + \nabla_x f(x)^\mathsf{T}(y-x) + \frac{L}{2}\|y-x\|_2^2.$$

Theorem 69
Let $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ be a sequence of gradient descent iterates for a μ-strongly convex, L-smooth function f. Then there exists C such that

$$\left\|x^{(k+1)} - x^*\right\|_2^2 \leq C^{k+1}\left\|x^{(0)} - x^*\right\|_2^2.$$

Proof. We require a small claim on smoothness:

Claim. If f is L-smooth, then

$$\|\nabla_x f(x)\|_2^2 \leq 2L\left(f(x) - f(x^*)\right).$$

Proof. We have that

$$\begin{aligned}
f(x^*) &\leq f(x) \\
&\leq f\left(x - \frac{\nabla_x f(x)}{L}\right) \\
f\left(x - \frac{\nabla_x f(x)}{L}\right) &\leq f(x) + \left[\nabla_{x'} f(x')\right]_{x'=x}^\mathsf{T}\left(-L\left[\nabla_{x'} f(x')\right]_{x'=x}\right) + \frac{L}{2}\left\|\frac{\left[\nabla_{x'} f(x')\right]_{x'=x}}{L}\right\|_2^2 \\
&= f(x) - \frac{1}{L}\left\|\left[\nabla_{x'} f(x')\right]_{x'=x}\right\|_2^2 + \frac{1}{2L}\left\|\left[\nabla_{x'} f(x')\right]_{x'=x}\right\|_2^2 \\
&= f(x) - \frac{\left\|\left[\nabla_{x'} f(x')\right]_{x'=x}\right\|_2^2}{2L} \\
f(x^*) &\leq f(x) - \frac{\left\|\left[\nabla_{x'} f(x')\right]_{x'=x}\right\|_2^2}{2L} \\
\left\|\left[\nabla_{x'} f(x')\right]_{x'=x}\right\|_2^2 &\leq 2L\left(f(x) - f(x^*)\right)
\end{aligned}$$

as desired. ∎

By the definition of strong convexity,

$$\begin{aligned}
f(x^*) &\geq f(x) + \nabla_x f(x)^\mathsf{T}(x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2 \\
\nabla_x f(x)^\mathsf{T}(x - x^*) &\geq f(x) - f(x^*) = \frac{\mu}{2}\|x^* - x\|_2^{[}2]
\end{aligned}$$

Now we want to bound the difference $\left\|x^{(k+1)} - x^*\right\|_2^2$:

$$\begin{aligned}
\left\|x^{(k+1)} - x^*\right\|_2^2 &= \left\|x^{(k)} - \eta\nabla_{x^{(k)}} f\left(x^{(k)}\right) - x^*\right\|_2^2 \\
&= \left\|\left(x^{(k)} - x^*\right) - \eta\nabla_{x^{(k)}} f\left(x^{(k)}\right)\right\|_2^2 \\
&= \left\|x^{(k)} - x^*\right\|_2^2 + \eta^2\left\|\nabla_{x^{(k)}} f\left(x^{(k)}\right)\right\|_2^2 - 2\eta\left\langle\nabla_{x^{(k)}} f\left(x^{(k)}\right)\middle| x^{(k)} - x^*\right\rangle \\
&\leq \left\|x^{(k)} - x^*\right\|_2^2 + 2\eta^2 L\left(f\left(x^{(k)}\right) - f(x^*)\right) - 2\eta\left(f\left(x^{(k)}\right) - f(x^*) + \frac{\mu}{2}\left\|x^{(k)} - x^*\right\|_2^2\right) \\
&\leq (1 - \eta\mu)\left\|x^{(k)} - x^*\right\|_2^2 + \left(2\eta^2 L - 2\eta\right)\left(f\left(x^{(k)}\right) - f(x^*)\right)
\end{aligned}$$

$$\leq \left(1 - \frac{\mu}{L}\right) \left\|x^{(k)} - x^*\right\|_2^2 \qquad \text{(Choose } \eta = \tfrac{1}{L}\text{.)}$$

$$\left\|x^{(k+1)} - x^*\right\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^{k+1} \left\|x^{(0)} - x^*\right\|_2^2$$

Therefore, one may write $C = 1 - \frac{\mu}{L}$ and prove the claim. □

This is exponential convergence, but since C is a linear function, sometimes it's called linear convergence.

Let f be differentiable.

- If f is μ-strongly convex and L-smooth, then convergence is $\mathcal{O}(\exp(-t))$.
- If f is μ-strongly convex and Lipschitz-continuous, or convex and smooth, then convergence is $\mathcal{O}\left(\frac{1}{t}\right)$.
- If f is convex and Lipschitz continuous, then convergence is $\mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$.

A function f is K-Lipschitz-continuous over D if for all $x, y \in D$, $\|f(x) - f(y)\|_2 \leq K\|x - y\|_2$, where $\|\cdot\|$ can be replaced in a general metric space by a distance function.

8 Duality

Unconstrained optimization of a differentiable function f can be solved by gradient descent, or solving for x in the differential equation $\nabla_x f(x) - 0$. Constrained optimization problems can be turned into unconstrained optimization problems in the following way. Let $f(x)$ be the objective, $g(x) \leq 0$ be the set of inequality constraints put into vector form, and $h(x) = 0$ be the set of equality constraints put into vector form:

$$\begin{aligned} p^* &= \min_x f(x) \\ \text{s.t. } & g_i(x) \leq 0, i \in \{1, \ldots, n_{\leq}\} \\ & h_i(x) = 0, i \in \{1, \ldots, n_{=}\} \end{aligned}$$

We define the *Lagrangian* as

$$L(x, \lambda, \mu) = f(x) + \lambda^\mathsf{T} g(x) + \mu^\mathsf{T} h(x)$$

where $\lambda \geq 0$ (as in, $\lambda_i \geq 0$). The λ and μ are *dual variables* or *Lagrange multipliers*. Define $\min_x \mathcal{L}(x, \lambda, \mu) = \ell(\lambda, \mu)$.

It's easy to show that the pointwise maximum of convex functions is convex; a corollary is that the pointwise minimum of concave functions is concave. Therefore $\ell(\lambda, \mu)$ is concave, and the concavity does not depend on $f(x)$.

Claim 70. $\ell(\lambda, \mu) \leq p^*$, for all $\lambda \geq 0$ (elementwise) and μ.

Proof. Say that $\widetilde{x}$ is a feasible point (fulfills constraints) for the primal problem (the one that defines p^*). Then $g(\widetilde{x}) \leq 0$ elementwise, and $h(\widetilde{x}) = 0$ elementwise. Therefore for any positive $\lambda \geq 0$ we have $\lambda^\mathsf{T} g(\widetilde{x}) \leq 0$ and $\mu^\mathsf{T} h(\widetilde{x}) = 0$, since $\lambda \geq 0$ componentwise. Then

$$\begin{aligned} \mathcal{L}(\widetilde{x}, \lambda, \mu) &= f(\widetilde{x}) + \lambda^\mathsf{T} g(\widetilde{x}) + \mu^\mathsf{T} h(\widetilde{x}) \\ &\leq f(\widetilde{x}) \end{aligned}$$

Since $\ell(\lambda, \mu) = \inf_x \mathcal{L}(x, \lambda, \mu)$, it must be true that $\ell(\lambda, \mu) \leq \mathcal{L}(\widetilde{x}, \lambda, \mu) = p^*$ as desired. □

An interpretation of this result is that we can do an unconstrained optimization of the objective function $f(x) + \sum_{i=1}^{n_\leq} g_i(x) (\infty \cdot \mathbb{1}(g_i(x) > 0)) + \sum_{i=1}^{n_=} h_i(x) (\infty \cdot \mathbb{1}(h_i(x) \neq 0))$ and obtain the minimum every time if we have a minimizer that can deal with such penalties. But this too harsh of a penalty to be practical, and it's not smooth; the Lagrange multipliers give a linear penalty on violating the constraints.

Example 71 (Minimum Norm Solution). We use Lagrange multipliers to solve the optimization problem

$$\begin{aligned} p^* &= \min_x x^\mathsf{T} x \\ \text{s.t. } & Ax = b \end{aligned}$$

The Lagrangian is $\mathcal{L}(x) = x^\mathsf{T} x + \mu^\mathsf{T} (Ax - b)$; define $\ell(\mu) = \min_x \mathcal{L}(x, \mu)$. We have $\boldsymbol{\nabla}_x \mathcal{L}(x) = 2x + A^\mathsf{T} \mu \overset{\text{set}}{=} 0$, so $x^* = -\frac{1}{2} A^\mathsf{T} \mu$. Then

$$\begin{aligned} \ell(\mu) &= \mathcal{L}(x^*, \mu) \\ &= \mathcal{L}\left(-\frac{1}{2} A^\mathsf{T} \mu, \mu\right) \\ &= \frac{1}{4} \mu^\mathsf{T} A A^\mathsf{T} \mu + \mu^\mathsf{T} \left[A \left(-\frac{1}{2} A^\mathsf{T} \mu\right) - b\right] \\ &= -\frac{1}{4} \mu^\mathsf{T} A A^\mathsf{T} \mu - \mu^\mathsf{T} b \end{aligned}$$

Since $p^* \geq \ell(\mu)$, we wish to maximize $\ell(\mu)$ over all μ. We have

$$\begin{aligned} \boldsymbol{\nabla}_\mu \ell(\mu) &= \boldsymbol{\nabla}_\mu \left(-\frac{1}{4} \mu^\mathsf{T} A A^\mathsf{T} \mu - \mu^\mathsf{T} b\right) \\ &= -\frac{1}{4} \left(2 A A^\mathsf{T}\right) \mu - b \\ &\overset{\text{set}}{=} 0 \\ \mu^* &= -2 \left(A A^\mathsf{T}\right)^{-1} b \\ x^* &= -\frac{1}{2} A^\mathsf{T} \left(-2 \left(A A^\mathsf{T}\right)^{-1} b\right) \\ &= A^\mathsf{T} \left(A A^\mathsf{T}\right)^{-1} b \end{aligned}$$

This motivates the construction of the *Lagrangian dual problem*:

$$d^* = \max_{\lambda \geq 0} \ell(\lambda, \mu)$$

This is the case even when the problem is not convex.

Example 72 (Partitioning Problem). Consider the problem

$$p^* = \min_{x_i^2 = 1} x^\mathsf{T} W x$$

where $W \in \mathbb{S}^n$. The Lagrangian is

$$\begin{aligned} \mathcal{L}(x, \mu) &= x^\mathsf{T} W x + \sum_{i=1}^n \mu_i \left(x_i^2 - 1\right) \\ &= x^\mathsf{T} \left(W + \operatorname{diag}(\mu)\right) x - \sum_{i=1}^n \mu_i \end{aligned}$$

If $W + \operatorname{diag}(\mu)$ is not positive semidefinite, then $\inf_x \mathcal{L}(x, \mu) = -\infty$. Now assume that $W + \operatorname{diag}(\mu)$ is positive semidefinite, so we can pick x such that $x^\mathsf{T} (W + \operatorname{diag}(\mu)) x = 0$, so $\ell(\mu) = \inf_x \mathcal{L}(x, \mu) = -\sum_{i=1}^n \mu_i$. The dual problem is $\max_{W+\operatorname{diag}(\mu)\succeq 0} \ell(\mu)$. A lower bound is $\mu = -\lambda_{\min}(W)1$, giving $d^* \geq n_= \lambda_{\min}(W)$.

Formally, for a Lagrangian $\mathcal{L}(x, \lambda, \mu)$ and $\ell(\lambda, \mu) = \min_x \mathcal{L}(x, \lambda, \mu)$, the primal problem is

$$\begin{aligned} p^* &= \min_x f(x) \\ \text{s.t. } & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

and the dual problem is

$$\begin{aligned} d^* &= \max_{\lambda,\mu} \ell(x) \\ \text{s.t. } & \lambda \geq 0 \end{aligned}$$

We know that $d^* \leq p^*$; this is called *weak duality*. Under some conditions $d^* = p^*$; this is called *strong duality*. The difference $p^* - d^*$ is the *duality gap*.

Theorem 73 (Min-Max Theorem)
For any sets X, Y, and any function $f \colon X \times Y \to \mathbb{R}$,

$$\min_{x\in X} \max_{y\in Y} f(x, y) \geq \max_{y\in Y} \min_{x\in X} f(x, y).$$

Proof. Fix $x_0 \in X$ and $y_0 \in Y$. Define $h(y) = \min_{x\in X} f(x, y)$, and similarly define $g(x) = \max_{y\in Y} f(x, y)$. Then

$$\begin{aligned} h(y_0) &= \min_{x\in X} f(x, y_0) \\ &\leq f(x_0, y_0) \\ &\leq \max_{y\in Y} f(x_0, y) \\ &\leq g(x_0) \end{aligned}$$

Therefore for each x and y, $h(y) \leq g(x)$, so

$$\begin{aligned} \max_{y\in Y} h(y) &\leq \min_{x\in X} g(x) \\ \max_{y\in Y} \min_{x\in X} f(x, y) &\leq \min_{x\in X} \max_{y\in Y} f(x, y) \end{aligned}$$

as desired. □

Clearly this is relevant in terms of duality. Consider for simplicity the problem

$$\begin{aligned} p^* &= \min_{\substack{x \\ g(x)\leq 0}} f(x) \\ \mathcal{L}(x, \lambda) &= f(x) + \lambda^\mathsf{T} g(x) \\ \max_{\lambda\geq 0} \mathcal{L}(x, \lambda) &= \max_{\lambda\geq 0} \Big(f(x) + \lambda^\mathsf{T} g(x) \Big) \\ &= \begin{cases} \infty, & g(x) \neq 0 \\ f(x), & g(x) \leq 0 \end{cases} \end{aligned}$$

$$\begin{aligned} p^* &= \min_x \max_{\lambda \geq 0} \mathcal{L}(x, \lambda) \\ d^* &= \max_{\lambda \geq 0} \ell(\lambda) \\ &= \max_{\lambda \geq 0} \min_x \mathcal{L}(x, \lambda) \\ p^* &\geq d^* \end{aligned}$$

which proves weak duality from the min-max theorem.

We want to find where strong duality holds, i.e. where $p^* = d^*$, which is true sometimes.

Theorem 74 (Slader's Condition)
For a convex problem

$$\begin{aligned} p^* = \min_x \ & f(x) \\ \text{s.t. } & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

where there exists a point x_0 such that $g(x) < 0$ (a strictly feasible region), then strong duality holds.

A comment on strict feasibility: this means that, if the feasible region is X, the condition implies that $\overline{X} - X \neq \emptyset$, that is, the set is not only its limit points.

Theorem 75 (Refined Slader's Condition)
For a convex problem

$$\begin{aligned} p^* = \min_x \ & f(x) \\ \text{s.t. } & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

where there exists x_0 such that for all affine constraints g_i we have $g_i(x) \leq 0$ and all other constraints we have $g_i(x) < 0$, strong duality holds.

Duality is of special interest in linear programs. Suppose we have the linear program

$$\begin{aligned} p^* = \min_x \ & c^\mathsf{T} x \\ \text{s.t. } & Ax \leq b \end{aligned}$$

Clearly, each constraint is linear or affine, so strong duality always holds (as long as the linear program has a feasible region).

The Lagrangian is

$$\begin{aligned} \mathcal{L}(x, \lambda) &= c^\mathsf{T} x + \lambda^\mathsf{T} (Ax - b) \\ &= \left(A^\mathsf{T} \lambda + c\right)^\mathsf{T} x - \lambda^\mathsf{T} b \\ \ell(\lambda) &= \min_x \mathcal{L}(x, \lambda) \\ &= \min_x \left(A^\mathsf{T} \lambda + c\right)^\mathsf{T} x - \lambda^\mathsf{T} b \end{aligned}$$

$$= \begin{cases} -\infty, & A^\mathsf{T}\lambda + c \neq 0 \\ -b^\mathsf{T}\lambda, & A^\mathsf{T}\lambda + c = 0 \end{cases}$$

The dual is then

$$\begin{aligned} d^* &= \max_{\lambda \geq 0} \ell(\lambda) \\ &= \max_{\substack{\lambda \geq 0 \\ A^\mathsf{T}\lambda + c = 0}} -b^\mathsf{T}\lambda \end{aligned}$$

One can see that strong duality holds here, as long as the feasible region is nonempty.

We finish with some intuition about Slader's condition. Consider the optimization

$$\begin{aligned} p^* = \min_x f(x) \\ \text{s.t. } g(x) \leq 0 \\ h(x) = 0 \end{aligned}$$

and consider the set of ordered tuples $G = \{g(x), f(x)\}$. Then $p^* = \min\{t \,|\, (u,t) \in G, u \leq 0\}$. The Lagrangian is $\mathcal{L}(u,t,\lambda) = t + \lambda^\mathsf{T} u$, an affine hyperplane. In fact these hyperplanes are *supporting hyperplanes* which have at least one point on the curve, and separate it from another half-plane.