EE 127

Optimization Models in Engineering

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

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

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:

$$x^* = \underset{x}{\operatorname{arginf}} \left\| Ax - b \right\|^2$$

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:

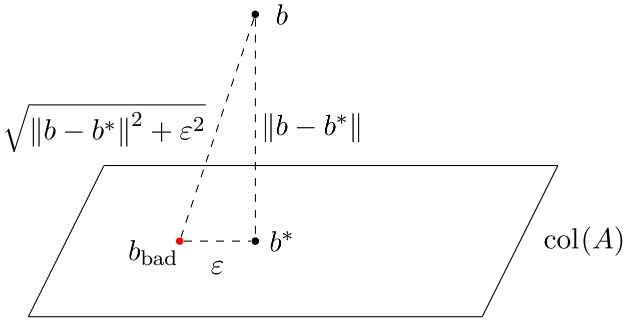
$$\beta^* = \operatorname*{arginf}_{\beta_1, \beta_2} \sum_{i=1}^{m} \left[y_i - (\beta_1 x_i + \beta_2) \right]^2$$

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

How do we actually find the quantity

$$x^* = \operatorname*{arginf}_{x} \|Ax - b\|^2$$

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 $A^\mathsf{T}b \in \operatorname{image}\left(A^\mathsf{T}A\right)$, 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 $\boldsymbol{\nabla}_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$. If $f(x)$ is scalar valued, then $\boldsymbol{\nabla}_x f(x)$ 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 $\boldsymbol{\nabla}_x^2 f(x)$ is called the Hessian. In particular, $\left[\boldsymbol{\nabla}_x^2 f(x)\right]_{i,j} = \frac{\partial}{\partial x_j}\left(\boldsymbol{\nabla}_x f(x)\right)_i = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$.

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

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

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 $D_a f(x) = \langle \boldsymbol{\nabla}_x f(x) | a\rangle$, then for all c $\left[\boldsymbol{\nabla}_x f(x)\right]_{x\colon f(x)=c} \perp L_c(f)$.

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

$$\begin{aligned} f(x+\varepsilon) &= (x+\varepsilon)^\mathsf{T}(x+\varepsilon) \\ &= x^\mathsf{T}x + x^\mathsf{T}\varepsilon + \varepsilon^\mathsf{T}x + \varepsilon^\mathsf{T}\varepsilon \\ &= x^\mathsf{T}x + (2x)^\mathsf{T}\varepsilon + \frac{1}{2}\left(2\varepsilon^\mathsf{T}\varepsilon\right) \end{aligned}$$

We have that $\boldsymbol{\nabla}_x f(x) = 2x$ and $\boldsymbol{\nabla}_x^2 = 2I$ 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 $\boldsymbol{\nabla}_x f(x) = a$.

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

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

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

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

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

Using the previous example, $\boldsymbol{\nabla}_x f(x) = 2\left(A^\mathsf{T} Ax - A^\mathsf{T} b\right)$. 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

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

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

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

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 $\|x\|_\infty = \sup_i x_i$.

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

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

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

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

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 $x^* = \operatorname{sign}(y)$, so $\sup_{\|x\|_p \leq 1} \langle x|y \rangle = \sum_i^n |y_i| = \|y\|_1$.

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

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 $\operatorname{span}(q_1 < \ldots, q_{i-1})$.
    $s_i \leftarrow a_i - p_i$                                       ▷ 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

$$\operatorname{image}(A) = \operatorname{kernel}\left(A^\mathsf{T}\right)^\perp$$

and

$$\dim\Big(\text{image}(A)\Big) + \dim\Big(\text{kernel}(A)\Big) = \text{rank}(A) + \text{nullity}(A) = n$$

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:

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

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

$$\begin{aligned}
\inf_w \text{MSE}(w) &= \sup_w \frac{1}{n}\sum_{i=1}^n \langle w|x_i\rangle^2 \\
&= \frac{1}{n}\|Xw\|_2^2 \\
&= \frac{1}{n}\,(Xw)^\mathsf{T}\,(Xw)
\end{aligned}$$

$$\begin{aligned} &= \frac{1}{n} w^\mathsf{T} X^\mathsf{T} X w \\ &= w^\mathsf{T} \left(\frac{X^\mathsf{T} X}{n} \right) w \end{aligned}$$

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

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 $\det(A - \lambda I) = \prod_{i=1}^{n} (\lambda - \lambda_i)^{\mu_i}$. Then for each λ_i, we have $\mu_i = \dim(\text{kernel}(A - \lambda_i I))$.

Theorem 22 (Spectral Theorem for Real Symmetric Matrices). Let $A \in \mathbb{R}^{n \times n}$ be a real symmetric matrix, with eigenvalues $\lambda_1, \ldots, \lambda_k$ which themselves have algebraic multiplicities $\mu_1, \ldots, \mu_k$. Further, define $\phi_i = \text{kernel}(A - \lambda_i I)$. 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 $U^\mathsf{T} A U = \begin{bmatrix} u^\mathsf{T} \\ U_1^\mathsf{T} \end{bmatrix} A \begin{bmatrix} u & U_1 \end{bmatrix} = \begin{bmatrix} u^\mathsf{T} \\ U_1^\mathsf{T} \end{bmatrix} \begin{bmatrix} \lambda u & A U_1 \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & U_1^\mathsf{T} A U_1 \end{bmatrix}$ 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

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

and

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

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

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

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

Claim 30. The u_i are orthonormal.

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

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

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

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

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 $A^\top A x = \lambda x$ then $x^\top A^\top A x = x^\top \lambda x$ 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 $AV_rV_r^\top = U_r\Sigma_rV_r^\top$. However, $\operatorname{rank}\left(V_rV_r^\top\right) = r \neq n$, so $AV_rV_r^\top \neq I$. 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 $\operatorname{kernel}(A) = \operatorname{kernel}\left(A^\top A\right)$.

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

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 $\mathbb{R}^n/\operatorname{image}(V_r)$, generated by Gram-Schmidt. Then

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

Therefore,

$$AV_nV_n^\top = AV_rV_r^\top + AV_gV_g^\top$$

Claim 33. $\operatorname{image}(V_g) \subseteq \operatorname{kernel}\left(A^\top A\right)$.

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

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

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

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 $\operatorname{range}(A)$ dimensions. We attempt to find the longest axis direction:

$$x^* = \underset{\|x\|_2=1}{\operatorname{argsup}} \|Ax\|_2 = \underset{\|x\|_2=1}{\operatorname{argsup}}\, x^\top A^\top Ax = v_{\max}\left(A^\top A\right)$$

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

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 $\inf_x \|x\|_2$ subject to the constraint $Ax = b$. Recall that from Theorem 19, that $\operatorname{kernel}(A) \oplus \operatorname{image}\left(A^\top\right) = \mathbb{R}^n$. Consider a decomposition of a solution x into $x = x_n + x_r$, where $x_n \in \operatorname{kernel}(A)$, so $Ax_n = 0$, and $x_r \in \operatorname{image}\left(A^\top\right)$, so $A^\top y = x_r$. Then

$$Ax = A\left(x_n + x_r\right) = A\left(x_n + A^\top y\right) = AA^\top z.$$

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

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

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:

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

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 $A^\dagger = \left(A^\top A\right)^{-1} A^\top$ (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:

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

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

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

Also,

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

as desired. □

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

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

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 $A = U_A\Sigma_A V_A^\top$ be the full SVD of A. Let A_k be defined as $A_k = \sum_{i=1}^{k} \sigma_i(A)u_i(A)v_i(A)^\top$, with $\sigma_1(A) \geq \cdots \geq \sigma_n(A) > 0$. Then

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

Proof for Spectral Norm. We want to show $A_k = \operatorname*{arginf}_{\substack{B\in\mathbb{R}^{m\times n}\\ \operatorname{rank}(B)\leq k}} \|A - B\|_2$. First, we compute $\|A - A_k\|_2$, then we show that $\|A - A_k\|_2 \leq \|A - B\|_2$ for any rank k matrix B.

The theorem is trivial when $n = k$, so now assume $n > k$.

First, we compute $\|A - A_k\|_2$. We have

$$\begin{aligned}\|A - A_k\|_2 &= \left\|\sum_{i=1}^{n} \sigma_i(A)u_i(A)v_i(A)^\top - \sum_{i=1}^{k} \sigma_i(A)u_i(A)v_i(A)^\top\right\|_2 \\ &= \left\|\sum_{i=k+1}^{n} \sigma_i(A)u_i(A)v_i(A)^\top\right\|_2 \\ &= \sigma_{k+1}(A)\end{aligned}$$

Now we want to show that $B \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(B) = k$ has $\|A - B\|_2 \geq \sigma_{k+1}(A)$. Since $\operatorname{rank}(B) = k$, by rank-nullity we have $\dim(\operatorname{kernel}(B)) = n - k > 0$. Define $V_{k+1} = \begin{bmatrix} v_1(A) & \cdots & v_{k+1}(A)\end{bmatrix}$. Then since the $v_i(A)$ are pairwise orthogonal by the singular value decomposition, $\operatorname{rank}(V_{k+1}) = k + 1$. By the Pigeonhole Principle,

$\dim(\text{kernel}(B)) + \text{rank}(V_{k+1}) = n - k + k + 1 = n + 1 > n$, so there exists $v \neq 0$ such that $v \in \text{kernel}(B) \cap \text{image}(V_{k+1})$. Pick $w \in \text{kernel}(B) \cap \text{image}(V_{k+1})$ such that $\|w\|_2 = 1$. Therefore

$$\begin{aligned}
\|A - B\|_2^2 &\geq \|(A - B)\, w\|_2^2 = \|Aw\|_2^2 \\
&\geq \left\| A \sum_{i=1}^{k+1} \langle w | v_i(A) \rangle \, v_i(A) \right\|_2^2 \\
&\geq \left\| A \sum_{i=1}^{k+1} \alpha_i v_i(A) \right\|_2^2 && (\alpha_i = \langle w | v_i(A) \rangle.) \\
&\geq \left\| U_A \Sigma_A V_A^\mathsf{T} \sum_{i=1}^{k+1} \alpha_i v_i(A) \right\|_2^2 \\
&\geq \left\| U_A \Sigma_A V_A^\mathsf{T} V_A \alpha \right\|_2^2 = \|U_A \Sigma_A \alpha\|_2^2 = \|\Sigma_A \alpha\|_2^2 \\
&\geq \sum_{i=1}^{k+1} \alpha_i^2 \sigma_i(A)^2 \geq \sigma_{k+1}(A)^2 \left(\sum_{i=1}^{k+1} \alpha_i^2 \right) \geq \sigma_{k+1}(A)^2 && (\textstyle\sum_{i=1}^{k+1} \alpha_i = \alpha^\mathsf{T} \alpha = w^\mathsf{T} w = 1.) \\
\|A - B\|_2 &\geq \sigma_{k+1}(A)
\end{aligned}$$

as desired. □

Proof for Frobenius Norm. We want to show $A_k = \underset{\substack{B \in \mathbb{R}^{m \times n} \\ \text{rank}(B) \leq k}}{\text{arginf}} \|A - B\|_F$. First, we compute $\|A - A_k\|_F$, then we show that $\|A - A_k\|_F \leq \|A - B\|_F$ for any rank k matrix B.

The theorem is trivial when $n = k$, so now assume $n > k$.

First, we compute $\|A - A_k\|_F$. We have

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

Now we want to show that $B \in \mathbb{R}^{m \times n}$ with $\text{rank}(B) = k$ has $\|A - B\|_F^2 \geq \sum_{i=k+1}^{n} \sigma_i(A)^2$. In particular, we want to show that

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

It's sufficient to show that $\sigma_i(A - B) \geq \sigma_{k+i}(A)$, since it trivially leads to the following inequality chain, that proves the claim:

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

We show that $\sigma_i(A - B) \geq \sigma_{k+i}(A)$. Define $A - B = C$ and $C_j = \sum_{i=1}^{j} \sigma_i(C) u_i(C) v_i(C)^\mathsf{T}$ is the best rank j approximation of C. Then $\sigma_i(C) = \|C - C_{i-1}\|_2$. We have

$$\begin{aligned}
\sigma_i(A - B) &= \sigma_i(C) \\
&\geq \|C - C_{i-1}\|_2 \\
&\geq \|A - B - C_{i-1}\|_2
\end{aligned}$$

We now claim that $\|A - B - C_{i-1}\|_2 \geq \|A - A_{k+i-1}\|_2$. We have that $\operatorname{rank}(B + C_{i-1}) \leq \operatorname{rank}(B) + \operatorname{rank}(C_{i-1}) = k + i - 1$, so $B + C_{i-1}$ is an at-most-rank $k + i - 1$ matrix, so by applying the Eckart-Young-Mirsky theorem for spectral norms we obtain $\|A - B - C_{i-1}\|_2 \geq \|A - A_{k+i-1}\|_2$. Therefore

$$\sigma_i(A - B) \geq \|A - B - C_i\|_2 \geq \|A - A_{k+i-1}\|_2 = \sigma_{k+i}(A)$$

so $\sigma_i(A - B) \geq \sigma_{k+i}(A)$, completing the proof. □

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

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

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

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

Therefore

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

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

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

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 $A^\mathsf{T} A x^* = A^\mathsf{T} b$. We care about the stability of this system. We have

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

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

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

We consider the new **regularized** optimization problem

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

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

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

This is the solution to the ridge regression problem, or ℓ_2 regularized least squares. One can see that $K\left(A^\top A + \lambda^2 I\right) < K\left(A^\top A\right)$, 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

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

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

$$\inf_{x} \left\|W_1\left(Ax-b\right)\right\|_2^2+\left\|W_2\left(x-x'\right)\right\|_2^2$$

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 $Z_i \sim \mathrm{Normal}\left(0, \sigma_i^2\right)$, 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:

$$\mathrm{MLE}(w \mid x, y) = \operatorname*{argsup}_{w} p_{X,Y\mid w}(x, y \mid w)$$

In this particular case, X is deterministic, so

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

where $S = \mathrm{diag}\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}, \ldots, \frac{1}{\sqrt{2\pi\sigma_m^2}}\right)$. 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**:

$$\mathrm{MAP}(w \mid x, y) = \operatorname*{argsup}_{w} p_{w\mid X,Y}(w \mid x, y) = \operatorname*{argsup}_{w} p_{X,Y\mid w}(x, y \mid w)p_w(w)$$

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

$$\begin{aligned}
\mathrm{MAP}(w \mid x, y) &= \operatorname*{argsup}_{w} p_{Y\mid w}(y \mid w)p_w(w)\\
&= \operatorname*{argsup}_{w} p_w \prod_{i=1}^{m} p_{Y\mid w}(w \mid w)
\end{aligned}$$

$$
\begin{aligned}
&= \operatorname*{argsup}_w \left[\prod_{i=1}^{m} \frac{\exp\left(-\frac{\left(y_i - w^\top x_i\right)^2}{2\sigma_i^2}\right)}{\sqrt{2\pi\sigma_i}}\right] \left[\frac{\exp\left(-\left(w-\mu\right)^\top \Sigma_w^{-1}\left(w-\mu\right)\right)}{\sqrt{\left(2\pi\right)^m \det\left(\Sigma_w\right)}}\right] \\
&= \operatorname*{arginf}_w \exp\left(\sum_{i=1}^{m} \frac{\left(y_i - w^\top x_i\right)^2}{2\sigma_i^2} + \left(w-\mu\right)^\top \Sigma_w^{-1}\left(w-\mu\right)\right) \\
&= \operatorname*{arginf}_w \left\|S\left(xw-y\right)\right\|_2^2 + \left\|\Sigma_2^{-1/2}\left(w-\mu\right)\right\|_2^2
\end{aligned}
$$

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 $X = U_X\Sigma_X V_X^\top$, then the least squares solution is $w^* = V\Sigma^\dagger U^\top y$, 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., $w^* = V\Sigma^{\text{PCR}}U^\top y$ where $\Sigma^{\text{PCR}} = \frac{\mathbb{1}(i=j \wedge i \leq k)}{\Sigma_{i,i}}$.

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

$$
\begin{aligned}
w^* &= \operatorname*{arginf}_w \|Xw-y\|_2^2 + \lambda\|w\|_2^2 \\
&= \operatorname*{arginf}_{w=V_Xz} \|XV_Xz-y\|_2^2 + \lambda\|V_Xz\|_2^2 && (X = U_X\Sigma_X V_X^\top) \\
&= \operatorname*{arginf}_z \|XVz-y\|_2^2 + \lambda\|z\|_2^2 \\
z^* &= \left[\left(XV_X\right)^\top\left(XV_X\right) + \lambda I\right]^{-1}\left(XV_X\right)^\top y \\
&= \left(V_X^\top X^\top X V_X + \lambda I\right)^{-1}\left(XV_X\right)^\top y \\
&= \left[V_X^\top\left(U_X\Sigma_X V_X^\top\right)V_X + \lambda I\right]^{-1}\left(XV_X\right)^{-1} y \\
&= \left(\Sigma_X^\top\Sigma_X + \lambda I\right)^{-1}\Sigma_X^\top U^\top y \\
&= \begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\lambda} & & \\ & \ddots & \\ & & \frac{\sigma_n}{\sigma_n^2+\lambda} \end{bmatrix} U^\top y
\end{aligned}
$$

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.

Convex Sets

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 $C = \left\{x \,\middle|\, a^\top x = b\right\}$. 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 $a^\top x_3 = \alpha a^\top x_1 + a^\top x_2 - \alpha a^\top x_2 = \alpha a^\top x_1 - a^\top x_2 + a^\top x_2 = a^\top x_2 = b$, 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 $\left\{x \,\middle|\, a^\top x \geq b\right\}$ or $\left\{x \,\middle|\, a^\top x \leq b\right\}$.

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

- for all $x \in C$ we have $a^\top x \geq b$.
- for all $x \in D$ we have $a^\top x \leq b$.

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

$$\begin{aligned} f(x) &= a^\top (x - x_0) \\ &= (d - c)^\top \left(x - \frac{c + d}{2}\right) \end{aligned}$$

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

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

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

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

We want $t^2\|u-d\|_2^2 - 2t\left\langle c-d|(u-d)\right\rangle < 0$, so we want $2t\left\langle d-c|u-d\right\rangle + t^2\|u-d\|_2^2 < 0$, so we want $2\left\langle d-c|u-d\right\rangle + t\|u-d\|_2^2 < 0$. Since we can choose t and $\left\langle d-c|u-d\right\rangle < 0$, 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. □

Convex Functions

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

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

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 $x, y \in \mathbb{R}^n$

$$f(y) \geq f(x) + \boldsymbol{\nabla}_x f(x)^{\mathsf{T}}\,(y-x)$$

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 $\boldsymbol{\nabla}_x f(x^*) = 0$ 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 $\operatorname{domain}(f)$ is convex and $\boldsymbol{\nabla}_x^2 f(x) \succeq 0$ (is positive semidefinite).

In the discrete non-probabilistic case, we have

Theorem 54 (Jensen's Inequality). Let f be a convex function and $\lambda_1, \ldots, \lambda_n$ be such that $\sum_{i=1}^n \lambda_i = 1$. Then for all $x_1, \ldots, x_n \in \mathbb{R}$, 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 direct from the definitions of convexity (and induction on n, *thanks for your extremely valuable contribution, Rahul.*).

We can now introduce the useful notion of convexity into our optimization problems. Assume our optimization problems take the form

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

We can make strong claims about certain (convex) classes of these problems.

Convex Problems

Definition 55 (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 $D = \operatorname{domain}(f) \cap \bigcap_{i=1} \operatorname{domain}(f_i) \cap \bigcap_{i=1}^{\ell} \operatorname{domain}(g_i)$ is convex.

Example 56 (Linear Program). A linear program

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

is a convex problem.

Example 57. The optimization problem

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

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 $[\boldsymbol{\nabla}_x f(x)]_{x=x*} = 0$.

Example 58. The optimization problem

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

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

Example 59. The optimization problem

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

is a convex problem; taking the gradient gives $\boldsymbol{\nabla}_x f(x) = 1$ (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 60. The optimization problem

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

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 inf; $\inf S$ need not be attained by any value in S.

Theorem 61 (Maximum Theorem for Linear Functions). Let X be a convex and closed set. Then if $x^* = \operatorname{arginf}_{x \in X} c^\mathsf{T} x$ 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 $\boldsymbol{\nabla}_x f(x) = c$. Consider $z = \frac{rc}{\|c\|_2}$, so $x^* - z \in B_r(x^*)$, so $x^* - z \in X^o$. Then

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

which has a better optimum, achievable at $x^* - z$, a contradiction. □

We define the **epigraph reformulation** as

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

where t is known as a **slack variable**.

Definition 62 (LASSO Regression). We now introduce LASSO regression, or ℓ^1 normalization. We have the optimization problem

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

Adding the epigraph reformulation, we have

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

Definition 63 (Monotone Transformation). If $\phi(x)$ is continuous and strictly increasing, then over a domain D we have that $\operatorname{arginf}_{x \in D} f(x) = \operatorname{arginf}_{x \in D} \phi(f(x))$, and similarly for monotone decreasing functions.

Definition 64 (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 $p_{Y \mid X}(1 \mid X_i)$. We achieve this by learning w and β such that $p_{Y \mid X}(1 \mid X_i) = \sigma\left(w^\top X_i + \beta\right)$ where $\sigma(x) = \frac{1}{1+e^{-x}}$. In particular, $p_{Y \mid X}(y_i \mid X_i) = \sigma\left(-y_i\left(w^\top X_i + \beta\right)\right)$.

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

$$\begin{aligned} (w^*, \beta^*) &= \operatorname*{argsup}_{w,\beta} \prod_{i=1}^{m} p_{Y \mid X}(y_i \mid X_i) \\ &= \operatorname*{argsup}_{w,\beta} \prod_{i=1}^{n} \sigma\left(-y_i\left(w^\top X_i + \beta\right)\right) \\ &= \operatorname*{argsup}_{w,\beta} \sum_{i=1}^{m} \log\left(\sigma\left(-y_i\left(w^\top X_i + \beta\right)\right)\right) \end{aligned}$$

and the rest is computation.

7 Gradient Descent

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

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

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

$$\begin{aligned} f(x + \Delta_x) &= f(x) + \nabla_x f(x)^\top \Delta_x + o\left(\|\Delta_x\|_2^3\right) \\ f(x + su) &= f(x) + \nabla_x f(x)^\top su \\ &= f(x) + s \langle \nabla_x f(x) | u \rangle \end{aligned}$$

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

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 $\eta$.
Output: Local minimum of $f$.
  $k \leftarrow 0$
  while $\nabla_{x^{(k)}} f\big(x^{(k)}\big) \neq 0$ do
      $x^{(k+1)} \leftarrow x^{(k)} - \eta \nabla_{x^{(k)}} f\big(x^{(k)}\big)$
      $k \leftarrow k + 1$
  return $x^{(k)}$
```

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

Example 65 (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 $\left|\lambda_i\big(I - 2\eta A^\mathsf{T}A\big)\right| < 1$.

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 $\left|\lambda_i\big(I - 2\eta A^\mathsf{T}A\big)\right| \leq 1$; we'll do this later (i.e. homework).

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 66 (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 67 (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 68. Let $\big(x^{(k)}\big)_{k\in\mathbb{N}}$ 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

$$\begin{aligned}
f(x^*) &\leq f(x) \\
&\leq f\left(x - \frac{\nabla_x f(x)}{L}\right) \\
f\left(x - \frac{\nabla_x f(x)}{L}\right) &\leq f(x) + \nabla_x f(x)^\mathsf{T}\left(-\frac{\nabla_x f(x)}{L}\right) + \frac{L}{2}\left\|-\frac{\nabla_x f(x)}{L}\right\|_2^2 \\
&= f(x) - \frac{1}{L}\|\nabla_x f(x)\|_2^2 + \frac{1}{2L}\|\nabla_x f(x)\|_2^2 \\
&= f(x) - \frac{\|\nabla_x f(x)\|_2^2}{2L} \\
f(x^*) &\leq f(x) - \frac{\|\nabla_x f(x)\|_2^2}{2L} \\
\|\nabla_x f(x)\|_2^2 &\leq 2L\left(f(x) - f(x^*)\right)
\end{aligned}$$

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 $\left\|x^{(k+1)} - x^*\right\|_2^2$:

$$\begin{aligned}
\left\|x^{(k+1)} - x^*\right\|_2^2 &= \left\|x^{(k)} - \eta\nabla_{x^{(k)}} f\left(x^{(k)}\right) - x^*\right\|_2^2 \\
&= \left\|\left(x^{(k)} - x^*\right) - \eta\nabla_{x^{(k)}} f\left(x^{(k)}\right)\right\|_2^2 \\
&= \left\|x^{(k)} - x^*\right\|_2^2 + \eta^2\left\|\nabla_{x^{(k)}} f\left(x^{(k)}\right)\right\|_2^2 - 2\eta\left\langle \nabla_{x^{(k)}} f\left(x^{(k)}\right) \middle| x^{(k)} - x^*\right\rangle \\
&\leq \left\|x^{(k)} - x^*\right\|_2^2 + 2\eta^2 L\left(f\left(x^{(k)}\right) - f(x^*)\right) - 2\eta\left(f\left(x^{(k)}\right) - f(x^*) + \frac{\mu}{2}\left\|x^{(k)} - x^*\right\|_2^2\right) \\
&\leq (1 - \eta\mu)\left\|x^{(k)} - x^*\right\|_2^2 + \left(2\eta^2 L - 2\eta\right)\left(f\left(x^{(k)}\right) - f(x^*)\right) \\
&\leq \left(1 - \frac{\mu}{L}\right)\left\|x^{(k)} - x^*\right\|_2^2 && \text{(Choose } \eta = \tfrac{1}{L}.\text{)} \\
\left\|x^{(k+1)} - x^*\right\|_2^2 &\leq \left(1 - \frac{\mu}{L}\right)^{k+1}\left\|x^{(0)} - x^*\right\|_2^2
\end{aligned}$$

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

Lagrangian 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:

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

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 $\inf_x \mathcal{L}(x, \lambda, \mu) = \ell(\lambda, \mu)$.

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 69. $\ell(\lambda, \mu) \leq p^*$, 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 $f(x) + \sum_{i=1}^{n_\leq} g_i(x) \left(\infty \cdot \mathbb{1}(g_i(x) > 0)\right) + \sum_{i=1}^{n_=} h_i(x) \left(\infty \cdot \mathbb{1}(h_i(x) \neq 0)\right)$ 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 70 (Minimum Norm Solution). We use Lagrange multipliers to solve the optimization problem

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

The Lagrangian is $\mathcal{L}(x) = x^\mathsf{T} x + \mu^\mathsf{T} (Ax - b)$; define $\ell(\mu) = \inf_x \mathcal{L}(x, \mu)$. We have $\nabla_x \mathcal{L}(x) = 2x + A^\mathsf{T}\mu \overset{\text{set}}{=} 0$, 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

$$\nabla_\mu \ell(\mu) = \nabla_\mu \left(-\frac{1}{4} \mu^\mathsf{T} A A^\mathsf{T} \mu - \mu^\mathsf{T} b\right)$$

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

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

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

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

Example 71 (Partitioning Problem). Consider the problem

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

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

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

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 $x^{\mathsf{T}}\left(W + \operatorname{diag}(\mu)\right)x = 0$, so $\ell(\mu) = \inf_x \mathcal{L}(x, \mu) = -\sum_{i=1}^{n} \mu_i$. The dual problem is $\sup_{W + \operatorname{diag}(\mu) \succeq 0} \ell(\mu)$. 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 $\ell(\lambda, \mu) = \inf_x \mathcal{L}(x, \lambda, \mu)$, the primal problem is

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

and the dual problem is

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

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

$$
\inf_{x \in X} \sup_{y \in Y} f(x, y) \geq \sup_{y \in Y} \inf_{x \in X} f(x, y).
$$

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

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

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

$$\sup_{y \in Y} h(y) \leq \inf_{x \in X} g(x)$$
$$\sup_{y \in Y} \inf_{x \in X} f(x, y) \leq \inf_{x \in X} \sup_{y \in Y} f(x, y)$$

as desired. □

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

$$\begin{aligned} p^* &= \inf_{\substack{x \\ g(x) \leq 0}} \; f(x) \\ \mathcal{L}(x, \lambda) &= f(x) + \lambda^\mathsf{T} g(x) \\ \sup_{\lambda \geq 0} \mathcal{L}(x, \lambda) &= \sup_{\lambda \geq 0} \left(f(x) + \lambda^\mathsf{T} g(x) \right) \\ &= \begin{cases} \infty, & g(x) \neq 0 \\ f(x), & g(x) \leq 0 \end{cases} \\ p^* &= \inf_x \sup_{\lambda \geq 0} \mathcal{L}(x, \lambda) \\ d^* &= \sup_{\lambda \geq 0} \ell(\lambda) \\ &= \sup_{\lambda \geq 0} \inf_x \mathcal{L}(x, \lambda) \\ p^* &\geq d^* \end{aligned}$$

which proves weak duality from the min-max theorem.

Conditions for Strong Duality

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

Theorem 73 (Slater's Condition). For a convex problem

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

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 74 (Refined Slater's Condition). For a convex problem

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

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

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

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

$$\begin{aligned}\mathcal{L}(x,\lambda) &= c^{\mathsf{T}}x + \lambda^{\mathsf{T}}(Ax-b)\\ &= \left(A^{\mathsf{T}}\lambda + c\right)^{\mathsf{T}}x - \lambda^{\mathsf{T}}b\\ \ell(\lambda) &= \inf_x \mathcal{L}(x,\lambda)\\ &= \inf_x \left(A^{\mathsf{T}}\lambda + c\right)^{\mathsf{T}}x - \lambda^{\mathsf{T}}b\\ &= \begin{cases}-\infty, & A^{\mathsf{T}}\lambda + c \neq 0\\ -b^{\mathsf{T}}\lambda, & A^{\mathsf{T}}\lambda + c = 0\end{cases}\end{aligned}$$

The dual is then

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

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

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

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

and consider the set of ordered tuples $G = \{g(x), f(x)\}$. Then $p^* = \inf\{t \mid (u,t) \in G, u \leq 0\}$. 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.

We'll now cover some interesting examples that will hopefully solidify concepts. These examples are:

- the dual problem of logistic regression, and
- total least squares.

Example 75 (Logistic Regression Dual Problem). Let $(X_i, Y_i)_{i=1}^m$ be a set of data points, where $X_i \in \mathbb{R}^n$ and $Y_i \in \{-1, 1\}$. We want a model with parameters w and β such that $w^{\mathsf{T}}x + \beta = \log\left(\frac{p_{C\mid X}(1\mid x)}{p_{C\mid X}(-1\mid x)}\right)$. We derive w and β through maximum likelihood estimation.

$$\begin{aligned}\mathrm{MLE}(w,\beta \mid (X_i,Y_i)_{i=1}^m) &= \operatorname*{argsup}_{w,\beta}\prod_{i=1}^m p_{X_i,Y_i\mid w,\beta}(X_i, Y_i \mid w, \beta)\\ &= \operatorname*{argsup}_{w,\beta}\prod_{i=1}^m \frac{\exp\left(Y_i\left(w^{\mathsf{T}}X_i+\beta\right)\right)}{1+\exp(Y_i\left(w^{\mathsf{T}}X_i+\beta\right))}\\ &= \operatorname*{argsup}_{w,\beta}\sum_{i=1}^m \log\left(\frac{1}{1+\exp(-Y_i\left(w^{\mathsf{T}}X_i+\beta\right))}\right)\\ &= \operatorname*{argsup}_{w,\beta} -\sum_{i=1}^m \log\left(1+\exp\left(-Y_i\left(w^{\mathsf{T}}X_i+\beta\right)\right)\right)\\ &= \operatorname*{arginf}_{w,\beta}\sum_{i=1}^m \log\left(1+\exp\left(-Y_i\left(w^{\mathsf{T}}X_i+\beta\right)\right)\right)\end{aligned}$$

For simplicity, set $\beta = 0$ (equivalent to adding an extra dimension to X_i and w). Then we have the optimization problem

$$p^* = \inf_w \sum_{i=1}^m \log\Big(1 + \exp\big(-Y_i w^\top X_i\big)\Big)$$

We want to understand a dual problem, so we need to add some constraints. Define $X = \begin{bmatrix} X_1^\top \\ \vdots \\ X_m^\top \end{bmatrix}$ and $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_m \end{bmatrix}$, and define $A = \begin{bmatrix} Y_1 X_1 & \cdots & Y_m X_m \end{bmatrix}$. Then define $v = A^\top w$. This yields the optimization problem

$$\begin{aligned} p^* = \inf_{v,w} & \sum_{i=1}^n \log\Big(1 + \exp(-v_i)\Big) \\ \text{s.t. } & v = A^\top w \end{aligned}$$

Defining $f(x) = \log(1 + \exp(-x))$, we obtain the equivalent optimization problem

$$\begin{aligned} p^* = \inf_{v,w} & \sum_{i=1}^n f(v_i) \\ \text{s.t. } & v = A^\top w \end{aligned}$$

This is a convex problem, since the function f is convex, so the objective function is convex, and the constraint is clearly linear. Slater's condition says that strong duality holds since picking any w produces an applicable v. This means that if we don't feel like computing p^*, we can solve for d^* in the dual problem, since they are the same.

The Lagrangian is

$$\mathcal{L}(v, w, \nu) = \sum_{i=1}^m f(v_i) + \nu^\top \left(v - A^\top w\right)$$

We want to minimize $\mathcal{L}$ over v and w. Therefore

$$\begin{aligned} g(\nu) &= \inf_{v,w} \mathcal{L}(v, w, \nu) \\ &= \inf_{v,w} \sum_{i=1}^m f(v_i) + \nu^\top \left(v - A^\top w\right) \\ &= \inf_v \inf_w \sum_{i=1}^m f(v_i) + \nu^\top \left(v - A^\top w\right) \end{aligned}$$

The term that contains w is $-\nu^\top A^\top w$; in the case that any coefficient $\left(\nu^\top A^\top\right)_i \neq 0$, we must have that $g(\nu) = -\infty$. If $A\nu = 0$ then we recover the addtiional terms.

$$\begin{aligned} g(\nu) &= \inf_v \inf_w \sum_{i=1}^m f(v_i) + \nu^\top \left(v - A^\top w\right) \\ &= \inf_v \begin{cases} -\infty, & A\nu \neq 0 \\ \sum_{i=1}^m f(v_i) + \nu^\top v, & A\nu = 0 \end{cases} \end{aligned}$$

Define $h(t) = \log(1 + \mathrm{e}^{-t}) + \nu_i t$, so

$$\begin{aligned} g(\nu) &= \inf_v \begin{cases} -\infty, & A\nu \neq 0 \\ \sum_{i=1}^m f(v_i) + \nu^\top v, & A\nu = 0 \end{cases} \\ &= \inf_v \begin{cases} -\infty, & A\nu \neq 0 \\ \sum_{i=1}^m h(v_i), & A\nu = 0 \end{cases} \end{aligned}$$

Clearly $h(t)$ is convex; we have $\frac{\mathrm{d}h}{\mathrm{d}t} = \frac{1}{1+\mathrm{e}^{-t}} \cdot \mathrm{e}^{-t} \cdot (-1) + \nu_i \overset{\text{set}}{=} 0$, so $\mathrm{e}^{t^*} = \frac{1-\nu_i}{\nu_i}$, so $t^* = \log\left(\frac{1-\nu_i}{\nu_i}\right)$. If $\nu_i < 0$, then $\lim_{t\to\infty} h(t) = -\infty$; if $\nu_i > 1$, then $\lim_{t\to\infty} h(t) = -\infty$ as well (one can show this via L'Hopital's rule or otherwise).

So the interesting range for ν is $\nu_i \in [0,1]^n$. In this range we have

$$\begin{aligned} h(t^*) &= \log\left(1 + \exp\left(-\log\left(\frac{1-\nu_i}{\nu_i}\right)\right)\right) + \nu_i \log\left(\frac{1-\nu_i}{\nu_i}\right) \\ &= \log\left(1 + \frac{\nu_i}{1-\nu_i}\right) + \nu_i \log\left(\frac{1-\nu_i}{\nu_i}\right) \\ &= -\nu_i \log(\nu_i) - (1-\nu_i)\log(1-\nu_i) \end{aligned}$$

where the last term is the **binary entropy function** of a Bernoulli variable. Therefore

$$g(\nu) = \begin{cases} \sum_{i=1}^m -\nu_i \log(\nu_i) - (1-\nu_i)\log(1-\nu_i), & A\nu = 0, \nu \in [0,1]^n \\ -\infty, & \text{otherwise} \end{cases}$$

Termwise, we have $\frac{1}{p_{Y \mid X}(Y_i \mid X_i)} = \frac{1}{1-\nu_i}$, so $\nu_i = p_{Y \mid X}(\neg Y_i \mid X_i) = 1 - p_{Y \mid X}(Y_i \mid X_i)$. The final dual optimization problem is

$$\begin{aligned} d^* = \sup_{\nu \geq 0} \ & g(\nu) \\ \text{s.t. } & A\nu = 0 \\ & \nu \in [0,1]^n \end{aligned}$$

and plugging in obtains

$$d^* = \sup_{\nu \geq 0} \begin{cases} \sum_{i=1}^m -\nu_i \log(\nu_i) - (1-\nu_i)\log(1-\nu_i), & A\nu = 0, v \in [0,1]^n, \\ -\infty, & \text{otherwise} \end{cases}$$

Example 76 (Total Least Squares). Normally, our least squares problem assumes a model $Ax = b$ where b is noisy (i.e. $b = \hat{b} + \widetilde{)}$. If A and b both have noise, then we obtain some model $\left(\hat{A} + \widetilde{A}\right) x = b + \widetilde{b}$. We want to find minimal perturbations $\widetilde{A}$ and $\widetilde{b}$ (in some sense) such that $\left(\hat{A} + \widetilde{A}\right) x = \hat{b} + \widetilde{b}$. Actually, since we are looking at entrywise perturbations, we use the Frobenius norm as a measure of distance. Therefore the objective function is $\left\|\widetilde{A}\right\|_F^2 + \left\|\widetilde{b}\right\|_2^2 = \left\|\widetilde{A} \mid \widetilde{b}\right\|_F^2$ (the augmentation of $\widetilde{A}$ by $\widetilde{b}$). The problem is then

$$\begin{aligned} p^* = \inf_{\widetilde{A}, \widetilde{b}} \ & \left\|\widetilde{A} \mid \widetilde{b}\right\|_F^2 \\ \text{s.t. } & \left(\hat{A} + \widetilde{A}\right) x = \hat{b} + b \end{aligned}$$

Therefore

$$\left[\hat{A} + \widetilde{A} \mid \hat{b} + \widetilde{b}\right] \begin{bmatrix} x \\ -1 \end{bmatrix} = 0$$

so we want $\begin{bmatrix} x \\ -1 \end{bmatrix} \in \text{kernel}\left(\left[\hat{A} + \widetilde{A} \mid \hat{b} + \widetilde{b}\right]\right)$, so the matrix $\left[\hat{A} + \widetilde{A} \mid \hat{b} + \widetilde{b}\right]$ must not have full rank (since otherwise the kernel is trivial by rank-nullity).

Define $\left[\hat{A} + \widetilde{A} \mid \hat{b} + \widetilde{b}\right] = \widetilde{Z}$ and $\left[\hat{A} \mid \hat{b}\right] = Z$. Then the problem becomes

$$\begin{aligned} p^* = \inf_{\widetilde{Z}} \ & \left\|\hat{Z} - \widetilde{Z}\right\|_F^2 \\ \text{s.t. } & \text{rank}\left(\widetilde{Z}\right) \leq n \end{aligned}$$

and the solution is given by Eckart-Young Theorem and the Singular Value Decomposition.

Optimality from Dual Problem

The dual feasible solutions can be used as certificates of optimality. Recall that the primal problem is

$$\begin{aligned} p^* = \inf_x\ & f(x) \\ \text{s.t. } & g(x) \leq 0 \\ & h(x) = 0 \\ p^* = \inf_x\ & \sup_{\lambda, \nu}\ \mathcal{L}(x, \lambda, \nu) \end{aligned}$$

and the dual problem is

$$\begin{aligned} d^* &= \sup_{\lambda \geq 0, \nu}\ \ell(\lambda, \nu) \\ &= \sup_{\lambda, \nu} \inf_x \mathcal{L}(x, \lambda, \nu) \end{aligned}$$

Example 77. Say that we are working with the following primal linear problem:

$$\begin{aligned} p^* = \inf_x\ & c^\mathsf{T} x \\ \text{s.t. } & Ax - b \leq 0 \end{aligned}$$

Then

$$\begin{aligned} \mathcal{L}(x, \lambda) &= c^\mathsf{T} x + \lambda^\mathsf{T} (Ax - b) \\ &= -b^\mathsf{T} \lambda + \left(A^\mathsf{T} \lambda + c\right)^\mathsf{T} x \\ \ell(\lambda) &= \inf_x \mathcal{L}(x, \lambda) \\ &= \begin{cases} -b^\mathsf{T} \lambda, & A^\mathsf{T} \lambda + c = 0, \\ -\infty, & A^\mathsf{T} \lambda + c \neq 0 \end{cases} \end{aligned}$$

and so the dual is

$$\begin{aligned} d^* = \sup_{\lambda \geq 0} & -b^\mathsf{T} \lambda \\ \text{s.t. } & A^\mathsf{T} \lambda + c = 0 \end{aligned}$$

as desired.

If we can find some λ' and ν' that are dual-feasible, that is, satisfy the constraints of the dual, then we can say $p^* \geq d^* = \sup_{\lambda \geq 0, \nu} \ell(\lambda, \nu) \geq \ell(\lambda', \nu')$. If x' is a primal feasible point, then we can say $f(x') - p \leq f(x') - g(\lambda', \nu')$, and so if $f(x') - g(\lambda', \nu') = \varepsilon$ for any feasible x', λ', ν', then $f(x') - p^* \leq \varepsilon$.

Conditions and Consequences of Strong Duality

We have one cheap way to produce an optimal solution where strong duality holds:

Theorem 78 (Complementary Slackness). Consider the primal program

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

with dual

$$d^* = \inf_{\lambda \geq 0, \nu}\ \ell(\lambda, \nu).$$

If strong duality holds, i.e., $p^* = d^*$, then for each i either $g_i(x^*) = 0$ or $\lambda_i^* = 0$.

Proof. We have that

$$\begin{aligned} p^* &= d^* \\ f(x^*) &= g(\lambda^*, \nu^*) \\ &= \inf_x \mathcal{L}(x, \lambda^*, \nu^*) \\ &\leq \mathcal{L}(x^*, \lambda^*, \nu^*) \\ &\leq f(x^*) + \lambda^{*\mathsf{T}} g(x) + \nu^{*\mathsf{T}} h(x^*) \\ &\leq f(x^*) + \underbrace{0}_{\substack{g(x^*)\leq 0 \\ \lambda \geq 0}} + \underbrace{0}_{h(x^*)=0} \\ &\leq f(x^*) \end{aligned}$$

But these quantities must be equal, so all of the inequalities are actually equalities. Therefore $\lambda^{*\mathsf{T}} g(x^*) = 0$, and since $\lambda_i^* g_i(x^*) \leq 0$, we must have that $\lambda_i^* g_i(x^*) = 0$, proving the claim. □

We now cover some conditions for strong duality to hold. These conditions are known as the **Karush-Kuhn-Tucker** (KKT) conditions.

Definition 79 (KKT Conditions). Let the primal problem

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

with dual

$$d^* = \sup_{\lambda \geq 0, \nu} \inf_x \left(f(x) + \lambda^{\mathsf{T}} g(x) + \nu^{\mathsf{T}} h(x)\right)$$

be a not-necessarily-convex problem, where f, g, and h are differentiable. Then if x^* is the optimal primal point and (λ^*, ν^*) is the dual optimal point, the KKT conditions are defined as:

- $g(x^*) \leq 0$ (definition of primal problem).
- $h(x^*) = 0$ (definition of primal problem).
- $\lambda^* \geq 0$ (definition of dual problem).
- $\lambda_i^* g_i(x^*) = 0$ for each i (complementary slackness).
- $\boldsymbol{\nabla}_{x^*} \mathcal{L}(x^*, \lambda^*, \nu^*) = \boldsymbol{\nabla}_{x^*} f(x^*) + (\boldsymbol{\nabla}_{x^*} g(x^*))^{\mathsf{T}} \lambda^* + (\boldsymbol{\nabla}_{x^*} h(x^*))^{\mathsf{T}} \nu^* = 0$ (x^* minimizes $\mathcal{L}(x, \lambda^*, \nu^*)$).

In particular, we say that for the given problem, x^* and (λ^*, ν^*) satisfy the KKT conditions if the above five (in)equalities are true.

This gives a clean interpretation of necessary and sufficient conditions for strong duality.

Theorem 80 (KKT Necessary Conditions). If $p^* = d^*$, i.e., strong duality holds, the KKT conditions must hold for the given problem at x^* and (λ^*, ν^*).

It is important to note that for non-convex problems, these are necessary but not sufficient conditions for optimality, i.e. the optimal $x^*, (\lambda^*, \nu^*)$ satisfy the KKT conditions, but there may be non-optimal points $x, (\lambda, \nu)$ that satisfy the KKT conditions as well.

Theorem 81 (KKT Sufficient Conditions). If for the problem above each $g_i(x)$ is convex and each $h_i(x)$ is affine, then any set of points $\widetilde{x}$, $\left(\widetilde{\lambda}, \widetilde{\nu}\right)$ that satisfy the KKT conditions is optimal.

In total, $\widetilde{x}$, $\left(\widetilde{\lambda}, \widetilde{\nu}\right)$ satisfy the KKT conditions, each $g_i(x)$ is convex, each $h_i(x)$ is affine (and Slater's condition holds, so $p^* = d^*$) if and only if $\widetilde{x}$, $\left(\widetilde{\lambda}, \widetilde{\nu}\right)$ are the primal-optimal and dual-optimal points. In this way the KKT conditions are necessary and sufficient conditions for optimality in some cases.

Proof. We claim that if $\widetilde{x}$, $\left(\widetilde{\lambda}, \widetilde{\nu}\right)$ satisfy the KKT conditions, each $g_i(x)$ is convex, each $h_i(x)$ are affine, then $\widetilde{x}$, $\left(\widetilde{\lambda}, \widetilde{\nu}\right)$ are primal-optimal and dual-optimal points.

Consider the Lagrangian $\mathcal{L}(x, \lambda, \nu) = f(x) + \lambda^\mathsf{T} g(x) + \nu^\mathsf{T} h(x)$. Then consider $\widetilde{x}$, $\left(\widetilde{\lambda}, \widetilde{\nu}\right)$ defined as above, and consider the function of x that is $\mathcal{L}\left(x, \widetilde{\lambda}, \widetilde{\nu}\right)$. Since $\widetilde{\lambda} \geq 0$ and $g_i(x)$ is convex in x, $\mathcal{L}\left(x, \widetilde{\lambda}, \widetilde{\nu}\right)$ is convex in x. If $\boldsymbol{\nabla}_{\widetilde{x}} \mathcal{L}\left(\widetilde{x}, \widetilde{\lambda}, \widetilde{\nu}\right) = 0$, then $\widetilde{x}$ is a local minimum of $\mathcal{L}\left(x, \widetilde{\lambda}, \widetilde{\nu}\right)$, and since convexity holds $\widetilde{x}$ is a global minimum.

Then

$$\begin{aligned}
\ell\left(\widetilde{\lambda}, \widetilde{\nu}\right) &= \inf_x \mathcal{L}\left(x, \widetilde{\lambda}, \widetilde{\nu}\right) \\
&= \mathcal{L}\left(\widetilde{x}, \widetilde{\lambda}, \widetilde{\nu}\right) \\
&= f(\widetilde{x}) + \widetilde{\lambda}^\mathsf{T} g(\widetilde{x}) + \widetilde{\nu}^\mathsf{T} h(\widetilde{x}) \\
&= f(\widetilde{x}) + \underbrace{0}_{\text{complementary slackness, KKT}} + \underbrace{0}_{h\left(\widetilde{x}\right)=0} \\
&= f(\widetilde{x})
\end{aligned}$$

Since $g(\lambda, \nu) \leq f(x)$ for any x, (λ, ν), if $g\left(\widetilde{\lambda}, \widetilde{\nu}\right) = f(\widetilde{x})$, then $\widetilde{x}$ must be the primal-optimal point, and correspondingly $\left(\widetilde{\lambda}, \widetilde{\nu}\right)$ is the dual-optimal point. □

9 Types of Optimization Programs

Linear Program

The general form of a linear program is

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

The standard form for a linear program is slightly different:

$$\begin{aligned}
p^* = \inf_x \; & c^\mathsf{T} x \\
\text{s.t. } & Ax = b \\
& x \geq 0
\end{aligned}$$

Let's discuss methods to turn any arbitrary linear program into this form.

- Any equality constraint $a^\mathsf{T} x = b$ can be turned into inequality constraints $a^\mathsf{T} x \leq b$ and $a^\mathsf{T} x \geq b$.
- A "greater than" inequality constraint $a^\mathsf{T} x \geq b$ can be turned into a "less than" inequality constraint $-a^\mathsf{T} x \leq -b$.
- Inequalities can be turned into equalities: $a^\mathsf{T} x < b$ turns into $a^\mathsf{T} x + s = b$ for $s \geq 0$.
- If x_j doesn't have to be in $\mathbb{R}_{\geq 0}$, then we introduce $x_j^+, x_j^- \geq 0$, and write the constraint $x_j = x_j^+ - x_j^-$.

We have $Ax + s = b$ by the inequality-equality conversion, so we can replace $Ax = b$ with $A'x' = b$, where $A' = \begin{bmatrix} A & I \end{bmatrix}$ and $x' = \begin{bmatrix} x \\ s \end{bmatrix}$.

Example 82. Consider the linear program

$$\begin{aligned} p^* = \inf_x \; & 2x_1 + 4x_2 \\ \text{s.t. } & x_1 + x_2 \geq 3 \\ & 3x_1 + 2x_2 = 14 \\ & x_1 \geq 0 \end{aligned}$$

Write $x_2 = x_2^+ - x_2^-$, with constraints $x_2^+, x_2^- \geq 0$. Introduce a slack variable x_3, writing the first constraint as $x_1 + x_2 - x_3 = 3$. Now the program is

$$\begin{aligned} p^* = \inf_x \; & 2x_1 + 4x_2^+ - 4x_2^- \\ \text{s.t. } & x_1 + x_2^+ - x_2^- - x_3 = 3 \\ & 3x_1 + 2x_2 = 14 \\ & x_1, x_2^+, x_2^-, x_3 \geq 0 \end{aligned}$$

which is in standard form.

We now cover the simplex method for solving linear programs. By the maximum principle, a linear program takes its optimum values at vertices of the polytope that constitutes its feasible region.

Consider a general linear program

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

As before, the maximum principle states that the linear program takes its optimum value at a vertex of the polytope $\mathcal{F} = \{x \mid Ax \leq b\}$, a path connected region. This provides an intuition; we want to traverse the edges in the direction of greatest decrease in $c^\mathsf{T} x$, as in gradient descent. Eventually we hit the point where $c^\mathsf{T} x$ is minimal.

We'll now rigorously define some of these terms.

Definition 83 (Polyhedron). A polyhedron is $\{x \mid Ax \geq b\}$ for A a matrix and b a vector.

Definition 84 (Extreme Point). Let $\mathcal{P}$ be a polyhedron. Then $x \in \mathcal{P}$ is an extreme point (vertex) of $\mathcal{P}$ if we cannot find two vectors $y, z \in \mathcal{P} \setminus \{x\}$ such that

$$x = \lambda y + (1 - \lambda) z$$

for any $\lambda \in [0, 1]$.

Alternatively, $x \in \mathcal{P}$ is an extreme point of $\mathcal{P}$ if there exists c such that $\langle c|x \rangle < \langle c|y \rangle$ for all $y \in \mathcal{P} \setminus \{x\}$, that is, all of $\mathcal{P}$ is on one side of the hyperplane $\mathcal{H} = \{y \mid \langle c|x \rangle = \langle c|y \rangle\}$, and $\mathcal{P} \cap \mathcal{H} = \{x\}$.

Fact 85. Let $\mathcal{P}$ be a polyhedron in $\mathbb{R}^n$. Then $\mathcal{P}$ has an extreme point if and only if $\mathcal{P}$ does not contain a line.

Theorem 86. Consider a linear program of the form $\min_{x \in \mathcal{P}} c^\mathsf{T} x$ where $\mathcal{P} = \{x \mid Ax \leq b\}$ is a polyhedron. Suppose $\mathcal{P}$ has at least one extreme point and a finite optimal solution exists. Then there exists an optimal solution that is an extreme point of $\mathcal{P}$.

Proof. Let $\mathcal{Q}$ be the set of all optimum points: $\mathcal{Q} = \left\{x \,\middle|\, Ax \leq b, c^\mathsf{T} x = p^*\right\}$; by presumption $\mathcal{Q}$ is nonempty. Since the new constraint $c^\mathsf{T} x = p^*$ is linear and $\mathcal{Q} \subseteq \mathcal{P}$, $\mathcal{Q}$ is also a polyhedron. Since $\mathcal{P}$ has an extreme point, $\mathcal{P}$ contains no lines. Therefore $\mathcal{Q}$ contains no lines, so $\mathcal{Q}$ has an extreme point x^*.

We wish to show that x^* is also an extreme point of $\mathcal{P}$. Assume for sake of contradiction that x^* is not an extreme point of $\mathcal{P}$. Then there exist $y, z \in \mathcal{P} \setminus \{x^*\}$ such that $x^* = \lambda y + (1 - \lambda) z$. Then

$$\begin{aligned} p^* &= c^\mathsf{T} x \\ &= c^\mathsf{T} [\lambda y + (1 - \lambda) z] \\ &= \lambda c^\mathsf{T} y + (1 - \lambda) c^\mathsf{T} z \end{aligned}$$

Since $\lambda \in [0,1]$, $c^{\mathsf{T}}y \geq p^*$ and $c^{\mathsf{T}}z \geq p^*$, so $c^{\mathsf{T}}y = p^*$ and $c^{\mathsf{T}}z = p^*$. Then $y, z \in \mathcal{Q}$, so $x^* \notin \mathcal{Q}$, a contradiction. □

The implication of this theorem is that to solve a linear program we can simply iterate over all vertices in $\mathcal{P}$ in a greedy fashion, which is the basis for the simplex algorithm.

Quadratic Programs

We begin with a motivating example. We've discussed the least-ℓ^2-norm problem

$$\begin{aligned} p^* = \inf_{x} \ & \|x\|_2 \\ \text{s.t. } & Ax = b \end{aligned}$$

which is used when we have many solutions x to $Ax = b$; this is one way to choose the "best" solution. Another way to choose the "best" solution is to take the least-ℓ^1-norm problem

$$\begin{aligned} p^* = \inf_{x} \ & \|x\|_1 \\ \text{s.t. } & Ax = b \end{aligned}$$

For safety, we assume that A is wide (so $n > m$) and full row rank (so $\operatorname{rank}(A) = m$). Recall that $\|x\|_1 = \sum_{i=1}^{n} |x_i|$. Consider the transformation $x = x^+ - x^-$ where $x^+ = \max(x, 0)$ and $x^- = -\min(x, 0)$. Then $|x| = x^+ + x^-$. Therefore

$$\begin{aligned} p^* = \inf_{x} \ & \|x\|_1 \\ \text{s.t. } & Ax = b \end{aligned}$$

turns into

$$\begin{aligned} p^* = \inf_{x} \ & 1^{\mathsf{T}}\left(x^+ + x^-\right) \\ \text{s.t. } & A\left(x^+ - x^-\right) = b \\ & x^+ \geq 0 \\ & x^- \geq 0 \end{aligned}$$

and, one step further, define $y = \begin{bmatrix} x^+ \\ x^- \end{bmatrix}$. Then

$$\begin{aligned} p^* = \inf_{x} \ & 1^{\mathsf{T}}\begin{bmatrix} x^+ \\ x^- \end{bmatrix} \\ \text{s.t. } & \begin{bmatrix} A & -A \end{bmatrix}\begin{bmatrix} x^+ \\ x^- \end{bmatrix} = b \\ & \begin{bmatrix} x^+ \\ x^- \end{bmatrix} \geq 0 \end{aligned}$$

We want to find this in terms of x^+ and x^-, not x. It turns out that we can minimize over both independently, so the program is

$$\begin{aligned} p^* = \inf_{x^+, x^-} \ & 1^{\mathsf{T}}\begin{bmatrix} x^+ \\ x^- \end{bmatrix} \\ \text{s.t. } & \begin{bmatrix} A & -A \end{bmatrix}\begin{bmatrix} x^+ \\ x^- \end{bmatrix} = b \\ & \begin{bmatrix} x^+ \\ x^- \end{bmatrix} \geq 0 \end{aligned}$$

The constraint that $x_i^+ = 0$ or $x_i^- = 0$ for each i is necessary for our construction of x_i^+ and x_i^- out of x.

It turns out that this constraint is implicitly encoded, in that every optimal value for x^+ and x^- can be turned into a pair of optimal values for x^+ and x^- that do fulfill this condition with the same objective value. Indeed, if $x_i^+ x_i^- > 0$

then we do $\left(x_i^+, x_i^-\right) \leftarrow \left(\max(x_i^+, x_i^-) - \min(x_i^+, x_i^-), 0\right)$. Clearly the new x_i^+ and x_i^- fulfill the constraints and have the same objective value.

We can turn this into the parallel to the least squares problem, which is

$$p^* = \inf_x \ \|Ax - b\|_2$$

The parallel looks like

$$p^* = \inf_x \ \|Ax - b\|_1$$

which can be expressed in the previous verbiage as

$$\begin{aligned} p^* = \inf_x \ & \|e\|_1 \\ \text{s.t. } & Ax - b = e \end{aligned}$$

which can be converted to a linear program.

The ℓ^2-norm can be thought of as a generalization of the mean. Consider some sorted list $(b_i)_{i=1}^n \in \mathbb{R}^n$, then

$$\begin{aligned} \mu &= \operatorname*{arginf}_x \sum_{i=1}^n (x - b_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^n b_i \end{aligned}$$

Correspondingly, the ℓ^1-norm is a generalization of the median.

$$m = \operatorname*{arginf}_x \sum_{i=1}^n |x - b_i|$$

Note that

$$|x - b_i| = \begin{cases} x - b_i, & x > b_i \\ b_i - x, & x \leq b_i \end{cases}$$

is not differentiable everywhere, but

$$\frac{\mathrm{d}}{\mathrm{d}x}|x - b_i| = \begin{cases} 1, & x > b_i \\ -1, & x < b_i \\ \text{DNE}, & x = b_i \end{cases}$$

Then $\frac{\mathrm{d}}{\mathrm{d}x} m = |\{i \,|\, b_i > x\}| - |\{i \,|\, b_i < x\}|$ for $x \notin (b_i)_{i=1}^n$. Since we can write

$$m = \operatorname*{arginf}_x \sum_{i=1}^n |x - b_i| = \operatorname*{arginf}_x \left(|x - b_i| + \sum_{j \neq i} |x - b_j| \right)$$

then m is minimized, looking componentwise, when $x \in \left[b_{\lfloor (n+1)/2 \rfloor}, b_{\lceil (n+1)/2 \rceil}\right]$. The median is robust to outliers, and thus the ℓ^1-norm is also.

Ridge regression has the problem

$$p^* = \inf_x \left(\|Ax - b\|_2^2 + \underbrace{\lambda \|x\|_2^2}_{\text{regularizer}} \right)$$

The corresponding regularized least squares for ℓ^1-regularization, LASSO, is

$$p^* = \inf_x \left(\|Ax - b\|_2^2 + \lambda \|x\|_1 \right)$$

LASSO tends to give sparse solutions, and many components of x^* will be equal to 0.

The ridge regression problem

$$p^* = \inf_x \left(\|Ax - b\|_2^2 + \lambda \|x\|_2^2 \right)$$

is actually the Lagrangian of the constrained optimization problem

$$p^* = \inf_x \|Ax - b\|_2^2$$
$$\text{s.t. } \|x\|_2 \leq t$$

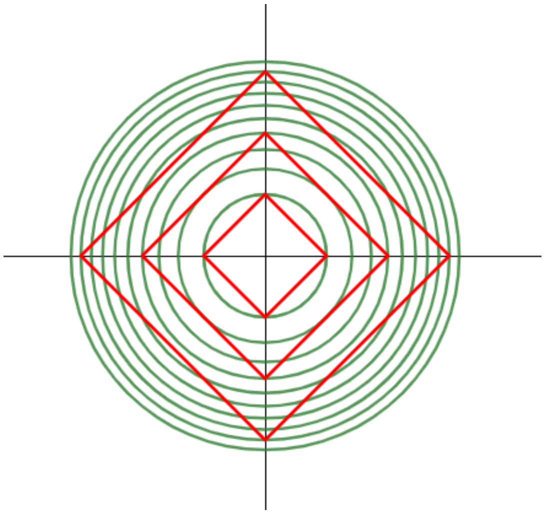
Correspondingly, the LASSO problem

$$p^* = \inf_x \left(\|Ax - b\|_2^2 + \lambda\|x\|_1\right)$$

is the Lagrangian of the constrained optimization problem

$$p^* = \inf_x \|Ax - b\|_2^2$$
$$\text{s.t. } \|x\|_1 \leq t$$

The level sets of these constraints – $\|x\|_2 \leq t$ and $\|x\|_1 \leq t$ – in $\mathbb{R}^2$, look like



Similarly, the level sets of the objective function $\|Ax - b\|_2^2$ are ellipses centered at b. The optimal ridge regression solution will have the level set $\|x\|_2 = t$ tangent at exactly one point to $\|Ax - b\|_2^2 = s$, for some values of s and t. It's highly likely that the optimal LASSO solution will intersect the level set $\|x\|_1 = t$ on a side rather than the corner, which ensures that the optimal coordinates the side connects the axes of are set to 0.

LASSO is usually solved via coordinate descent.

Let's look at how to solve the 1-dimensional LASSO problem, and then we will introduce the coordinate descent algorithm.

The 1-dimensional LASSO problem is

$$p^* = \inf_x \left\{\left[\frac{1}{2}\sum_{i=1}^{n}(a_i x - b_i)^2\right] + \lambda|x|\right\}$$

The objective function $f(x)$ is written as

$$f(x) = \frac{1}{2}\sum_{i=1}^{n}(a_i x - b_i)^2 + \begin{cases} \lambda x, & x > 0 \\ -\lambda x, & x < 0 \\ 0, & x = 0 \end{cases}$$

Then

$$\boldsymbol{\nabla}_x f(x) = \sum_{i=1}^{n} a_i (a_i x - b_i) + \begin{cases} \lambda, & x > 0 \\ -\lambda, & x < 0 \\ \text{DNE}, & x = 0 \end{cases}$$

so $\boldsymbol{\nabla}_x f(x)$ at $x = 0$ is undefined.

Case (Case 1). If $x > 0$, then

$$\begin{aligned} 0 &= \left(\sum_{i=1}^n a_i^2\right) x^* - \left(\sum_{i=1}^n a_i b_i\right) + \lambda \\ x^* &= \frac{\left(\sum_{i=1}^n a_i b_i\right) - \lambda}{\sum_{i=1}^n a_i^2} \end{aligned}$$

Therefore $x > 0$ if and only if $\lambda > \sum_{i=1}^n a_i b_i$.

Case (Case 2). If $x < 0$, then

$$\begin{aligned} 0 &= \left(\sum_{i=1}^n a_i^2\right) x^* - \left(\sum_{i=1}^n a_i b_i\right) - \lambda \\ x^* &= \frac{\left(\sum_{i=1}^n a_i b_i\right) + \lambda}{\sum_{i=1}^n a_i^2} \end{aligned}$$

Therefore $x < 0$ if and only if $\lambda < -\sum_{i=1}^n a_i b_i$.

Case (Case 3). $\lambda \in (-\sum_{i=1}^n a_i b_i, \sum_{i=1}^n a_i b_i)$ if and only if $x = 0$.

This is soft thresholding.

This is the scalar case, but in the vector case the same quantities are at play; the only difference is that we compute coordinatewise. This is the essential idea of coordinate descent.

Consider the problem

$$p^* = \inf_x \; f(x)$$

where f is not necessarily differentiable. Consider $x(0) = \begin{bmatrix} x_1(0) & \cdots & x_n(0) \end{bmatrix}^\mathsf{T}$ a fixed vector. Then for a fixed number of iterations in k (or until convergence), we iterate $j \in [1, n]$ with the update rule $x_j(k) = \operatorname{arginf}_x \; \begin{bmatrix} x_1(k) & \cdots & x_{j-1}(k) & x & x_j$

Quadratic Program

LASSO is an example of what is claled a quadratic program. A quadratic program is a problem of the form

$$\begin{aligned} p^* = \inf_x \; & \left(\frac{1}{2} x^\mathsf{T} H x + c^\mathsf{T} x + d\right) \\ \text{s.t. } & Ax \leq b \\ & Cx = e \end{aligned}$$

where H is a symmetric PSD/PD matrix.

Take the equivalent case

$$\begin{aligned} p^* = \inf_x \; & \left(\frac{1}{2} x^\mathsf{T} H x + c^\mathsf{T} x + d\right) \\ \text{s.t. } & Cx = e \end{aligned}$$

Then a feasible x can be generated by solutions x_0 to $Cx = e$. In particular, let $x = x_0 + \eta x_n$, where $x_n \in \operatorname{kernel}(C)$. Plugging this into the objective obtains an unconstrained quadratic in η. Minima of unconstrained quadratic programs can be:

- Unique point
- Multiple points
- Infinitely many points

- $-\infty$.

Example 87 (Linear Quadratic Regulator). Say that we have time-evolving $\left(x^{(t)}\right)_{t\in\mathbb{N}}$ subject to the equation

$$x^{(t+1)} = Ax^{(t)} + Bu^{(t)}$$

where u_t is a user-provided control. We want to choose $\left(u^{(t)}\right)_{t\in\mathbb{N}}$ such that $x^{(t)} \to x_0$, a "goal" value.

By recursion,

$$x^{(t)} = A^t x^{(0)} + \sum_{i=0}^{t-1} A^{t-1-i}Bu^{(i)}$$

Say that the maximum control time is T. Then one optimization problem is

$$\begin{aligned} p^* = \inf_{x,u} \left\|x^{(T)} - x_0\right\|_2^2 + \sum_{t=0}^{T} \left\|u^{(t)}\right\|_2^2 \\ \text{s.t. } x^{(t)} = A^t x^{(0)} + \sum_{i=0}^{t-1} A^{t-i-1}Bu^{(i)} \end{aligned}$$

which is a quadratic objective function with linear constraints.

Example 88 (Piecewise Constant Function). Suppose we have a sequence $\left(x^{(t)}, y^{(t)}\right)_t$ where y is our observation of the signal x. Suppose we have side information that the signal is piecewise constant. We wish to find $\left(\hat{x}^{(t)}\right)$ that does not change as much as possible. Consider z componentwise where $z_i = \hat{x}^{(i+1)} - \hat{x}^{(i)}$. Then $z = Dx$ where $D_{i,j}$ is -1 if $i = j$, $+1$ if $j = i + 1 \pmod T$, and 0 otherwise. Then we want to solve the problem

$$\begin{aligned} p^* = \inf_{\hat{x}} \|y - \hat{x}\|_2^2 \\ \text{s.t. } \|Dx\|_0 \leq k \end{aligned}$$

which can be relaxed to the problem

$$\begin{aligned} p^* = \inf_{\hat{x}} \|y - \hat{x}\|_2^2 \\ \text{s.t. } \|Dx\|_1 \leq k \end{aligned}$$

which can be solved via LASSO.

Second Order Conic Programs

Definition 89 (Cone). A cone is a set of points $C \subseteq \mathbb{R}^n$ such that if $x \in C$ then for all $\alpha > 0$, $\alpha x \in C$. C is a convex cone if for all $x, y \in C$, $x + y \in C$.

Most cones we deal with are convex cones.

Examples of convex cones are $\{(x, y) \mid |x| \leq y\}$, the polyhedral cone $\{(x, t) \mid Ax = bt, t \geq 0\} = \mathcal{P}$ for a fixed A and b, and ellipsoidal cone $\{(x, t) \mid \|Ax + bt\|_2 \leq ct\} = \mathcal{E}$ for fixed A, b, c. This last one is of particular importance. First we will show that the definition is of a cone. Indeed, consider some pair $(x, t) \in \mathcal{E}$. Then $\|\alpha Ax + \alpha bt\|_2 = \alpha\|Ax + bt\|_2 \leq \alpha ct$, so $(\alpha x, \alpha t) \in \mathcal{E}$, so $\mathcal{E}$ is a cone.

Definition 90 (Second Order Cone). The second order cone in $\mathbb{R}^n$ is defined as $\mathcal{K}_n = \{(x, t) \mid \|x\|_2 = t\}$.

The first cone presented as an example is K_2.

The last and critical example of a cone is $\left\{(x, t) \,\middle|\, \|Ax + bt\|_2 \leq c^\mathsf{T} x + dt\right\}$. One sees that this is a convex cone via i.e. triangle inequality.

The second order cone program (SOCP) is therefore defined to be

$$p^* = \inf_x c^\mathsf{T} x$$

$$\text{s.t. } \|A_i x + b_i\|_2 \leq c_i^\top x + d_i, \quad \forall i \in [m]$$

The constraints are second order cone constraints, which constrain x to slices at $t = 1$ of second order cones.

A LP

$$\begin{aligned} p^* = \inf_x \ & c^\top x \\ \text{s.t. } & a_i^\top x \leq b_i, \quad \forall i \in [m] \end{aligned}$$

can be framed as a SOCP, by setting $(A_i, b_i, c_i, d_i) = (0, 0, -a_i, b_i)$, for $i \in [m]$.

A QP with $Q \succeq 0$

$$\begin{aligned} p^* = \inf_x \ & \left(x^\top Q x + c^\top x\right) \\ \text{s.t. } & a_i^\top x \leq b_i, \quad \forall i \in [m] \end{aligned}$$

can be written as a SOCP, by making the objective $y + c^\top x$ where $x^\top Q x = y$, and turning this latter constraint into $\left\| \begin{bmatrix} 2Q^{1/2}x \\ y - 1 \end{bmatrix} \right\|_2 \leq y + 1$; then, turning the pre-existing linear constraints $a_i^\top x \leq b_i$ into second order cone constraints suffices to create the SOCP.

Example 91. The optimization problem

$$p^* = \inf_x \sum_{i=1}^{m} \|A_i x + b_i\|_2$$

reduces to

$$\begin{aligned} p^* = \inf_{x,y} \ & \sum_{i=1}^{m} y_i \\ \text{s.t. } & \|A_i x + b_i\|_2 \leq y_i \end{aligned}$$

Correspondingly,

$$p^* = \inf_x \sup_{i \in [m]} \|A_i x + b_i\|_2$$

reduces to

$$\begin{aligned} p^* = \inf_{x,y} \ & y \\ \text{s.t. } & \|A_i x + b_i\|_2 \leq y, \quad \forall i \in [m] \end{aligned}$$

The quadratic optimization method to solve these problems are Newton's method, following this loop:

$$x^{(k+1)} - x^{(k)} - \left(\boldsymbol{\nabla}^2_{x^{(k)}} f\left(x^{(k)}\right)\right)^{-1} \left(\boldsymbol{\nabla}_{x^{(k)}} f\left(x^{(k)}\right)\right)$$

If f is quadratic, then Newton's method will obtain the minimum in one step.

10 Applications

Linear Quadratic Regulator

Say we have the optimization problem

$$\begin{aligned} p^* = \inf_{x,u} \ & \frac{1}{2} \left\{ \left[\sum_{t=0}^{N-1} \left(x_t^\top Q x_t + u_t^\top R u_t\right) \right] + \left(x_N^\top Q_f x_N\right) \right\} \\ \text{s.t. } & x_{t+1} = A x_t + B u_t \\ & x_0 = x_{\text{init}} \end{aligned}$$

The x_t is the state at time t, the control at time t is u_t; correspondingly the penalty on the state at time t is $x_t^\mathsf{T} Q x_t$ and the penalty on the control at time t is $u_t^\mathsf{T} R u_t$. The system dynamics equation is $x_{t+1} = Ax_t + Bu_t$. Time N is the terminal time.

$$\begin{aligned} p^* = \inf_{x,u} \ & \frac{1}{2} \left\{ \left[\sum_{t=0}^{N-1} \left(\underbrace{x_t^\mathsf{T} Q \underbrace{x_t}_{\text{state}}}_{\text{state penalty}} + \underbrace{u_t^\mathsf{T} R \underbrace{u_t}_{\text{control}}}_{\text{control penalty}} \right) \right] + \left(\underbrace{x_N^\mathsf{T} Q_f x_N}_{\text{terminal cost}} \right) \right\} \\ \text{s.t.} \ & \underbrace{x_{t+1} = Ax_t + Bu_t}_{\text{system dynamics}} \\ & x_0 = x_{\text{init}} \end{aligned}$$

We can solve this problem by converting it into least squares problem on the u_t, but this is conceptually unwieldy. So we tae the dual. The Lagrangian is

$$\mathcal{L}(x, u, \mu) = \frac{1}{2} \left[\sum_{t=0}^{N-1} \left(x_t Q x_t + u_t^\mathsf{T} R u_t \right) + x_N^\mathsf{T} Q_f x_N \right] + \left[\sum_{t=0}^{N} \nu_{t+1}^\mathsf{T} \left(Ax_t + Bu_t - x_{t+1} \right) \right]$$

The only interesting condition is that $\boldsymbol{\nabla}_{x,u} \mathcal{L}(x, u, \nu) = 0$. In particular we have

$$\begin{aligned} \boldsymbol{\nabla}_{u_t} \mathcal{L}(x, u, \nu) &= Ru_t + B^\mathsf{T} \nu_{t+1} = 0, \quad \forall t \in \{0, \ldots, N-1\} \\ \boldsymbol{\nabla}_{x_t} \mathcal{L}(x, u, \nu) &= Qx_t + A^\mathsf{T} \nu_{t+1} - \nu_t = 0, \quad t \in \{0, \ldots, N-1\} \\ \boldsymbol{\nabla}_{x_N} \mathcal{L}(x, u, \nu) &= Q_f x_N - \nu_N = 0 \end{aligned}$$

From the second constraint, we obtain

$$\begin{aligned} \nu_t &= A^\mathsf{T} \nu_{t+1} + Qx_t \\ \nu_N &= Q_f x_N \end{aligned}$$

Rewriting the first constraint, we get

$$u_t = -R^{-1} B^\mathsf{T} \nu_{t+1}$$

As a reminder, we want optimal x, u.

Claim 92. For each $t \in \{0, \ldots, N\}$, $\nu_t = P_t x_t$ (ν_t and x_t are linearly related).

Proof. We have that $\nu_N = Q_f x_N$ where $Q_f = P_N$. Now assume that $\nu_{t+1} = P_{t+1} x_{t+1}$ (by induction). Then

$$\begin{aligned} \nu_{t+1} &= P_{t+1} x_{t+1} \\ &= P_{t+1} \left(Ax_t + Bu_t \right) \\ &= P_{t+1} \left[Ax_t + B \left(-R^{-1} B^\mathsf{T} \nu_{t+1} \right) \right] \\ &= P_{t+1} \left(Ax_t - BR^{-1} B^\mathsf{T} \nu_{t+1} \right) \\ P_{t+1} A x_t &= \left(I + P_{t+1} B R^{-1} B^\mathsf{T} \right) \nu_{t+1} \\ \nu_{t+1} &= \left(I + P_{t+1} B R^{-1} B^\mathsf{T} \right)^{-1} P_{t+1} A x_t \end{aligned}$$

Recall that

$$\begin{aligned} \nu_t &= A^\mathsf{T} \nu_{t+1} + Qx_t \\ &= A^\mathsf{T} \left(I + P_{t+1} B R^{-1} B^\mathsf{T} \right)^{-1} P_{t+1} A x_t + Qx_t \\ &= \underbrace{\left[A^\mathsf{T} \left(I + P_{t+1} B R^{-1} B^\mathsf{T} \right)^{-1} P_{t+1} A + Q \right]}_{P_t} x_t \end{aligned}$$

We run backwards from N to get all $P_0, \cdots, P_N$, starting with $P_N = Q_f$. This solution system is called the Ricatti equations. □

So the way we solve the linear quadratic regulator is to solve for all $P_0, \dots, P_N$ using the Ricatti equation. Then we write

$$\begin{aligned} u_t &= -R^{-1}B^{\mathsf{T}}\lambda_{t+1} \\ &= -R^{-1}B^{\mathsf{T}}\left(I + P_{t+1}BR^{-1}B^{\mathsf{T}}\right)^{-1} P_{t+1}Ax_t \end{aligned}$$

which gives a relationship between current state and current control.

Support Vector Machines

We want to find the maximum-margin linear classifier, i.e., given data points $(X_i, Y_i)_{i=1}^m$ with $Y_i \in \{-1, 1\}$, we want to solve the program

$$p^* = \sup_{w,b} \inf_{i \in [m]} d\big(X_i, \big\{X \,\big|\, X^{\mathsf{T}}w = b\big\}\big)$$

assuming that the data is linearly separable, i.e., $p^* > 0$. In particular, if $\mathcal{H}_{w,b}$ is a hyperplane described by $\mathcal{H}_{w,b} = \big\{x \,\big|\, w^{\mathsf{T}}x - b = 0\big\}$, we want $\mathcal{H}_{\widetilde{w},\widetilde{b}}$ to neatly separate the data points into their two classes, one per half-space. In particular, if $Y_i = -1$ then we want $X_i^{\mathsf{T}}w - b < 0$, and if $Y_i = 1$ then we want $X_i^{\mathsf{T}}w - b > 0$. In particular we want $Y_i = \operatorname{sign}\big(X_i^{\mathsf{T}}w - b\big)$ so we want $Y_i\left(X_i^{\mathsf{T}}w - b\right) > 0$.

By Lagrange multipliers or projective linear algebra, we obtain that the signed distance $d_s(X, \mathcal{H}_{w,b}) = \frac{X^{\mathsf{T}}w - b}{\|w\|_2}$ and so the unsigned distance $d(X, \mathcal{H}_{w,b}) = \frac{|X^{\mathsf{T}}w - b|}{\|w\|_2}$; since we assume that the data is linearly separable, we can impose the additional constraint that $Y_i = \operatorname{sign}\big(X_i^{\mathsf{T}}w - b\big)$, so $d(X, \mathcal{H}_{w,b}) = \frac{Y_i\left(X_i^{\mathsf{T}}w - b\right)}{\|w\|_2}$. Therefore the program becomes

$$\begin{aligned} p^* = \sup_{w,b,a} \;& a \\ \text{s.t. } & \frac{Y_i\left(X_i^{\mathsf{T}}w - b\right)}{\|w\|_2} \geq a, \quad \forall i \in [m] \end{aligned}$$

In particular, a is the margin of the classifier.

We can make one more optimization. If (s, w, b) is a solution to the optimization, then $(s, \alpha w, \alpha b)$ is also a solution to the optimization, since for any $\alpha > 0$,

$$\frac{Y_i\left(\alpha X_i^{\mathsf{T}}w - \alpha b\right)}{\|\alpha w\|_2} = \frac{\alpha Y_i\left(X_i^{\mathsf{T}}w - b\right)}{\alpha\|w\|_2} = \frac{Y_i\left(X_i^{\mathsf{T}}w - b\right)}{\|w\|_2} \geq s.$$

Since we can choose any $\alpha > 0$ we can choose any $s > 0$, we can pick $s = \frac{1}{\|w\|_2}$. Then rewriting the constraint, we obtain $Y_i\left(X_i^{\mathsf{T}}w - b\right) \geq 1$. This gives the program

$$\begin{aligned} p^* = \sup_{w,b} \;& s \\ \text{s.t. } & Y_i\left(X_i^{\mathsf{T}}w - b\right) \geq 1, \quad \forall i \in [m] \\ = \sup_{w,b} \;& \frac{1}{\|w\|_2} \\ \text{s.t. } & Y_i\left(X_i^{\mathsf{T}}w - b\right) \geq 1, \quad \forall i \in [m] \\ = \inf_{w,b} \;& \|w\|_2 \\ \text{s.t. } & Y_i\left(X_i^{\mathsf{T}}w - b\right) \geq 1, \quad \forall i \in [m] \end{aligned}$$

To make the problem a quadratic program, we can simply square the objective:

$$\begin{aligned} p^* = \sup_{w,b} \;& \frac{1}{2}\|w\|_2^2 \\ \text{s.t. } & Y_i\left(X_i^{\mathsf{T}}w - b\right) \geq 1, \quad \forall i \in [m] \end{aligned}$$

The distance from the closest point to the hyperplane $\left\{x \mid x^{\mathsf{T}}w^{*} = b\right\}$ is $\frac{1}{\|w\|_2}$; there is a slab of length $\frac{2}{\|w\|_2}$ separating points on opposite sides of the hyperplane.

This can be generalized in cases where the data are not linearly separable. We add slack variables ξ_i which allow us to penalize terms that are on the opposite side of the classifier. How much we penalize violations is left up to hyperparameter C. The new soft-margin program is

$$\begin{aligned} p^{*} = \sup_{w,b,\xi} &\left(\frac{1}{2}\|w\|_2^2 + C1^{\mathsf{T}}\xi\right) \\ \text{s.t. } & Y_i\left(X_i^{\mathsf{T}}w - b\right) \geq 1 - \xi_i, \quad \forall i \in [m] \\ & \xi_i \geq 0, \quad \forall i \in [m] \end{aligned}$$

The variable C is a hyperparameter, where large C are more sensitive to violations of the margin.

We will cover the hinge-loss interpretation of the soft-margin support vector machine. The zero-one loss function is $L_{0-1}(f(X_i), Y_i) = \begin{cases} 0, & Y_i f(X_i) > 0 \\ 1, & Y_i f(X_i) < 0 \end{cases}$. Then the minimization of the average loss with respect to this loss and the function $f_{w,b}(x) = x^{\mathsf{T}}w - b$ is

$$\begin{aligned} p^{*} &= \inf_{w,b} \frac{1}{m}\sum_{i=1}^{m} L_{0-1}(f_{w,b}(X_i), Y_i) \\ &= \inf_{w,b} \frac{1}{m}\sum_{i=1}^{m} L_{0-1}\left(X_i^{\mathsf{T}}w - b, Y_i\right) \end{aligned}$$

This is a non-convex problem. So we'll relax the zero-one loss function.

One key fact is that since $\xi_i \geq 0$ and $\xi_i \geq 1 - Y_i\left(X_i^{\mathsf{T}}w - b\right)$, then $\xi_i \geq \sup\left(1 - Y_i\left(X_i^{\mathsf{T}}w - b\right), 0\right)$. Since if $\xi_i = \sup\left(1 - Y_i\left(X_i^{\mathsf{T}}w - b\right), 0\right) + \varepsilon$, we fulfill all constraints and incur $C\varepsilon > 0$ extra cost, the minimal $\xi_i^{*} = \sup\left(1 - Y_i\left(X_i^{\mathsf{T}}w - b\right), 0\right)$, so we can encode this into our formulation.

Define $L_{\text{hinge}}(f(X), Y) = \sup(0, 1 - Y_i f(X_i))$. Finally, the regularized problem for some $\lambda > 0$ becomes

$$\begin{aligned} p^{*} &= \inf_{w,b}\left[\frac{1}{m}\left(\sum_{i=1}^{m} L_{\text{hinge}}(f_{w,b}(X_i), Y_i)\right) + \lambda\|w\|_2^2\right] \\ &= \inf_{w,b}\left[\frac{1}{m}\left(\sum_{i=1}^{m} L_{\text{hinge}}\left(X_i^{\mathsf{T}}w - b, Y_i\right)\right) + \lambda\|w\|_2^2\right] \\ &= \inf_{w,b}\left\{\frac{1}{m}\left[\sum_{i=1}^{m}\sup\left(1 - Y_i\left(X_i^{\mathsf{T}}w - b\right), 0\right)\right] + \lambda\|w\|_2^2\right\} \\ &= \inf_{w,b,\xi}\left[\frac{1}{m}\left(\sum_{i=1}^{m}\xi_i\right) + \lambda\|w\|_2^2\right] \\ &\quad \text{s.t. } \xi_i \geq \sup\left(1 - Y_i\left(X_i^{\mathsf{T}}w - b\right), 0\right) \\ &= \inf_{w,b,\xi}\left[\frac{1}{2}\|w\|_2^2 + \frac{1}{2\lambda m}\left(\sum_{i=1}^{m}\xi_i\right)\right] \\ &\quad \text{s.t. } \xi_i \geq 0 \\ &\quad\quad\ \ \xi_i \geq 1 - Y_i\left(X_i^{\mathsf{T}}w - b\right) \end{aligned}$$

which is the soft-margin support vector machine.

It's important to consider which points actually matter in the support vector machines.

We can also frame the hard-margin support vector machine in this way. Define $L_{0-\infty}(a,b) = \begin{cases} 0, & 1-ab<0 \\ \infty, & 1-ab \geq 0 \end{cases}$. Then the regularized problem is

$$p^* = \inf_{w,b}\left[\frac{1}{m}\left(\sum_{i=1}^{m} L_{0-\infty}(f_{w,b}(X_i), Y_i)\right) + \lambda\|w\|_2^2\right]$$

and if one works it out in a similar way as before, they obtain the hard-margin support vector machine.

We observe that both of these problems penalize errors in different ways, giving different models in the end. A different way to penalize misclassification is logistic regression, which solves the problem

$$p^* = \inf_{w,b}\left[\frac{1}{n}\sum_{i=1}^{m}\log\left(1 - \mathrm{e}^{-Y_i\left(X_i^\mathsf{T} w - b\right)}\right)\right]$$

and if $Y_i\left(X_i^\mathsf{T} w - b\right) > 0$ then the point is correctly classified.

We now look at the dual of the soft-margin support vector machine. Recall that the problem is

$$\begin{aligned} p^* = \inf_{w,b,\xi} &\ \left(\frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m}\xi_i\right) \\ \text{s.t.} &\ -\xi_i \leq 0 \\ 1 - Y_i\left(X_i^\mathsf{T} w - b\right) - \xi_i \leq 0 & \end{aligned}$$

This has Lagrangian

$$\begin{aligned} \mathcal{L}(w,b,\xi,\alpha,\beta) &= f(w,b,\xi) + \alpha^\mathsf{T} g(w,b,\xi) + \beta^\mathsf{T} h(w,b,\xi) \\ &= \frac{1}{2}\|w\|_2^2 + C\left(\sum_{i=1}^{n}\xi_i\right) + \left\{\sum_{i=1}^{n}\alpha_i\left[(1-\xi_i) - Y_i\left(X_i^\mathsf{T} w - b\right)\right]\right\} - \left(\sum_{i=1}^{n}\beta_i\xi_i\right) \\ &= \frac{1}{2}\|w\|_2^2 - \left[\sum_{i=1}^{n}\alpha_i Y_i\left(X_i^\mathsf{T} w - b\right)\right] + \left(\sum_{i=1}^{n}\alpha_i\right) + \left[\sum_{i=1}^{n}(C-\alpha_i-\beta_i)\,\xi_i\right] \end{aligned}$$

This problem has convex objective and affine constraints, so strong duality holds. Therefore the KKT conditions hold.

Therefore

$$\begin{aligned} \nabla_w\mathcal{L}(w,b,\xi,\alpha,\beta) &= \nabla_w\left\{\frac{1}{2}\|w\|_2^2 - \left[\sum_{i=1}^{n}\alpha_i Y_i\left(X_i^\mathsf{T} w - b\right)\right] + \left(\sum_{i=1}^{n}\alpha_i\right) + \left[\sum_{i=1}^{n}(C-\alpha_i-\beta_i)\,\xi_i\right]\right\} \\ &= w - \sum_{i=1}^{n}\alpha_i Y_i X_i \\ &\overset{\text{set}}{=} 0 \\ w^* &= \sum_{i=1}^{n}\alpha_i Y_i X_i \end{aligned}$$

Then the optimal w can be expressed in terms of the optimal dual variables and training data.

Also,

$$\begin{aligned} \nabla_b\mathcal{L}(w,b,\xi,\alpha,\beta) &= \nabla_b\left\{\frac{1}{2}\|w\|_2^2 - \left[\sum_{i=1}^{n}\alpha_i Y_i\left(X_i^\mathsf{T} w - b\right)\right] + \left(\sum_{i=1}^{n}\alpha_i\right) + \left[\sum_{i=1}^{n}(C-\alpha_i-\beta_i)\,\xi_i\right]\right\} \\ &= \alpha_i y_i \\ &\overset{\text{set}}{=} 0 \end{aligned}$$

Finally,

$$\begin{aligned}
\boldsymbol{\nabla}_\xi \mathcal{L}(w,b,\xi,\alpha,\beta) &= \boldsymbol{\nabla}_\xi \left\{\frac{1}{2}\|w\|_2^2 - \left[\sum_{i=1}^n \alpha_i Y_i \left(X_i^\mathsf{T} w - b\right)\right] + \left(\sum_{i=1}^n \alpha_i\right) + \left[\sum_{i=1}^n \left(C - \alpha_i - \beta_i\right)\xi_i\right]\right\} \\
&= C1 - \alpha - \beta && \overset{\text{set}}{=} 0
\end{aligned}$$

Substituting this back into the Lagrangian, we obtain

$$\begin{aligned}
\mathcal{L}(w,b,\xi,\alpha,\beta) &= \frac{1}{2}\|w\|_2^2 - \left[\sum_{i=1}^n \alpha_i Y_i \left(X_i^\mathsf{T} w - b\right)\right] + \left(\sum_{i=1}^n \alpha_i\right) + \left[\sum_{i=1}^n \underbrace{\left(C - \alpha_i - \beta_i\right)}_{0}\xi_i\right] \\
&= \frac{1}{2}\|w\|_2^2 - \left(\sum_{i=1}^n \alpha_i Y_i (X_i^\mathsf{T} w)\right) + \underbrace{b \cdot \left(\sum_{i=1}^n \alpha_i Y_i\right)}_{0} + \left(\sum_{i=1}^n \alpha_i\right) \\
&= \frac{1}{2} w^\mathsf{T} w - \left(\sum_{i=1}^n \alpha_i Y_i X_i^\mathsf{T} w\right) + \left(\sum_{i=1}^n \alpha_i\right) \\
&= \left(\frac{1}{2} w^\mathsf{T} - \sum_{i=1}^n \alpha_i Y_i X_i^\mathsf{T}\right) w + \left(\sum_{i=1}^n \alpha_i\right) \\
&= -\frac{1}{2} w^\mathsf{T} w + \sum_{i=1}^n \alpha_i && \left(w^* = \textstyle\sum_{i=1}^n \alpha_i Y_i X_i\right) \\
&= -\frac{1}{2}\left(\sum_{i=1}^n \alpha_i Y_i X_i\right)^\mathsf{T} \left(\sum_{i=1}^n \alpha_i Y_i X_i\right) + \sum_{i=1}^n \alpha_i \\
&= -\frac{1}{2}\alpha^\mathsf{T} \underbrace{\operatorname{diag}(Y) X X^\mathsf{T} \operatorname{diag}(Y)}_{Q} \alpha + \sum_{i=1}^n \alpha_i \\
&= -\frac{1}{2}\alpha^\mathsf{T} Q \alpha + \sum_{i=1}^n \alpha_i \\
d^* &= \sup_{\alpha,\beta} \inf_{w,b} \mathcal{L}(w,b,\xi,\alpha,\beta) \\
&= \sup_{\alpha,\beta} \left(\alpha^\mathsf{T} Q \alpha + \sum_{i=1}^n \alpha_i\right) \\
&\quad \text{s.t. } \alpha^\mathsf{T} y = 0 \\
&\qquad\quad C1 - \alpha - \beta = 0 \\
&\qquad\quad \alpha \geq 0 \\
&\qquad\quad \beta \geq 0
\end{aligned}$$

The traditional formulation of the dual eliminates β and obtains

$$\begin{aligned}
d^* = \sup_{\alpha} & \left(-\frac{1}{2}\alpha^\mathsf{T} Q \alpha + \sum_{i=1}^n \alpha_i\right) \\
\text{s.t. } & \alpha^\mathsf{T} y = 0 \\
& 0 \leq \alpha \leq C1
\end{aligned}$$

This dual will help us pick our support vectors.

Complementary slackness gives $\alpha_i\left[(1 - \xi_i) - Y_i\left(X_i^\mathsf{T} w - b\right)\right] = 0$ and $\beta_i \xi_i = 0$. Recall that $C - \alpha_i - \beta_i = 0$. This yields some cases:

- If $\alpha_i = 0$ then $C = \beta_i$ so $\beta_i \neq 0$ so $\xi_i = 0$. Then the i^{th} point has no margin violation.
- If $\alpha_i \neq 0$ then there are two cases:
  - If $\alpha_i = C$ then $\beta_i = 0$, so we can't say anything about ξ_i. But then $br1 - \varepsilon_i - Y_i \left(X_i^\mathsf{T} w - b\right) = 0$, so $Y_i \left(X_i^\mathsf{T} w - b\right) = 1 - \xi_i \leq 1$. Then the i^{th} point is either on the margin or it violates the margin.
  - If $\alpha_i \in (0, C)$ then $\beta_i \neq 0$ so $\xi_i = 0$. Then the constraint gives $Y_i \left(X_i^\mathsf{T} w - b\right) \geq 1 - \xi_i = 1$, so there is no margin violation. In particular, since $\alpha > 0$, we must have $1 - \xi_i - Y_i \left(X_i^\mathsf{T} w - b\right) = 0$, so $1 - \xi_i = Y_i \left(X_i^\mathsf{T} w - b\right)$, so the i^{th} point is exactly on the margin. Then (X_i, Y_i) is a support vector.

In particular, nonzero α_i implies (X_i, Y_i) are support vectors. Then we can write $w^* = \sum_{i=1}^{n} \alpha_i Y_i X_i$, and compute b^* by finding i such that $\alpha_i \in (0, C)$ and using $Y_i \left(X_i^\mathsf{T} w - b\right) = 1$.

Finally, we observe that $Q = \operatorname{diag}(Y) X X^\mathsf{T} \operatorname{diag}(Y)$. The matrix XX^T is a Gram matrix whose $(i,j)^{\text{th}}$ entry is $\langle X_i | X_j \rangle$. In this way we can make a feature map $\Phi \colon \mathbb{R}^n \to \mathbb{R}^m$ where $m > n$ that makes computing the inner product $\langle \Phi(X_i) | \Phi(X_j) \rangle$ particularly easy, and makes the features more meaningful. So then we can replace XX^T by $\Phi(X)\Phi(X)^\mathsf{T}$ with very little cost for high expressiveness.