Lec 1

1.1 Prerequisites

Allowed coding languages: Matlab, R, Python, Julia.

1.2 Optimization Problems

1.2.1 Unconstrained Optimization

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

where $\boldsymbol{x}$ is the unknown *variable* (also called the *decision variable*) and

$$f : \mathbb{R}^n \to \mathbb{R}$$

is called the *objective function*.

1.2.2 Constrained Optimization

$$\min_{\boldsymbol{x} \in C} f(\boldsymbol{x})$$

where $C \subseteq \mathbb{R}^n$ is called the *feasible region*.

Commonly, C is specified by inequality and equality constraints:

$$C = \Big\{ \boldsymbol{x} \in \mathbb{R}^n : g_1(\boldsymbol{x}) \geq 0, \ldots, g_k(\boldsymbol{x}) \geq 0;\ h_1(\boldsymbol{x}) = 0, \ldots, h_j(\boldsymbol{x}) = 0 \Big\}.$$

1.3 Supervised Learning

1.3.1 Definitions

Training data. We are given labeled examples, typically as pairs

$$(\boldsymbol{a}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{a}_m, \boldsymbol{y}_m),$$

where

$$\boldsymbol{a}_i \in \mathbb{R}^d, \quad i = 1, \ldots, m \quad (\text{“feature vectors”}),$$
$$\boldsymbol{y}_i \in \mathbb{R}^e, \quad i = 1, \ldots, m \quad (\text{“labels”}).$$

We seek a function Φ such that

$$\Phi(\boldsymbol{a}_i) \approx \boldsymbol{y}_i, \quad i = 1, \ldots, m.$$

Why find such a function Φ?

- Reason: To apply it to future *unseen* feature vectors $\boldsymbol{a} \in \mathbb{R}^d$.

1.3.2 Usual Optimization-Based Approach

Assume that Φ comes from a parametric family:

$$\Phi(\boldsymbol{a}) \equiv f(\boldsymbol{a}, \boldsymbol{x}),$$

where f is a fixed known function, and $\boldsymbol{x}$ is a vector of (say) p parameters, initially unknown.

Loss function. Define a *loss function* l to measure the discrepancy between the prediction $f(\boldsymbol{a}_i, \boldsymbol{x})$ and the label $\boldsymbol{y}_i$.

Example

Least squares. A common choice is the squared ℓ_2 loss:

$$l(\boldsymbol{a}, \boldsymbol{y}, \boldsymbol{x}) = \frac{1}{2}\|f(\boldsymbol{a}, \boldsymbol{x}) - \boldsymbol{y}\|^2.$$

Let the training set be

$$D = \{(\boldsymbol{a}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{a}_m, \boldsymbol{y}_m)\}.$$

Define the *empirical loss*:

$$L_D(\boldsymbol{x}) = \sum_{i=1}^{m} l(\boldsymbol{a}_i, \boldsymbol{y}_i, \boldsymbol{x}).$$

Then the learning problem becomes the optimization problem

$$\min_{\boldsymbol{x}} \; L_D(\boldsymbol{x}).$$

How to find an optimal x?
This is the role of optimization *algorithms*, which will be introduced later in the course.

Generalization and test set.

- Question: How do we know that $f(\boldsymbol{a}, \boldsymbol{x})$ is a good predictor for $\boldsymbol{y}$ when $\boldsymbol{a} \notin \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m\}$?

Hold aside additional labeled data items

$$(\boldsymbol{a}_{m+1}, \boldsymbol{y}_{m+1}), \ldots, (\boldsymbol{a}_{m'}, \boldsymbol{y}_{m'}),$$

called the *test set*. We evaluate the learned parameter vector $\boldsymbol{x}$ by computing the *test loss* on this set:

$$\sum_{i=m+1}^{m'} l(\boldsymbol{a}_i, \boldsymbol{y}_i, \boldsymbol{x}).$$

1.4 Unsupervised Learning

1.4.1 Clustering

Classic version: Clustering.
Given data points

$$\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in \mathbb{R}^d,$$

we want to partition them into k clusters. This can be expressed by selecting labels

$$y_1, \dots, y_m \quad \text{such that} \quad y_i \in \{1, \dots, k\}, \ \forall i = 1, \dots, m,$$

and enforcing that

$$\|\boldsymbol{a}_i - \boldsymbol{a}_j\| \quad \text{is} \quad \begin{cases} \text{small}, & y_i = y_j, \\ \text{large}, & y_i \neq y_j. \end{cases}$$

1.4.2 K-means Objective Function

Introduce additional variables (cluster centers)

$$\boldsymbol{c}_1, \dots, \boldsymbol{c}_k \in \mathbb{R}^d.$$

Example

K-means objective. The K-means clustering problem can be written as

$$\min_{\{y_1,\dots,y_m\}} \ \min_{\{\boldsymbol{c}_1,\dots,\boldsymbol{c}_k\}} \ \sum_{i=1}^{m} \|\boldsymbol{a}_i - \boldsymbol{c}_{y_i}\|^2.$$

- This is a *partially discrete* optimization problem (e.g., K-means): the labels y_i are discrete, while the centers $\boldsymbol{c}_j$ are continuous.
- In contrast, an example of a purely *continuous* optimization problem is linear least squares.

1.5 Linear Least Squares

Consider data points

$$(\boldsymbol{a}_1, y_1), \dots, (\boldsymbol{a}_m, y_m), \quad \boldsymbol{a}_i \in \mathbb{R}^n, \ y_i \in \mathbb{R}.$$

Linear model hypothesis. Assume there exists $\boldsymbol{x} \in \mathbb{R}^n$ such that

$$y_i \approx \boldsymbol{a}_i^\top \boldsymbol{x}, \quad \forall i = 1, 2, \dots, m.$$

Inner product and transpose.
For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$,

$$\boldsymbol{a}^\top \boldsymbol{b} = a_1 b_1 + \cdots + a_n b_n$$

is the inner product. The superscript $^\top$ denotes transpose.

Example

Least-squares fit as supervised learning.
In the earlier supervised-learning notation, we have

$$\Phi(\boldsymbol{a}) \equiv f(\boldsymbol{a}, \boldsymbol{x}) \equiv \boldsymbol{a}^\top \boldsymbol{x}.$$

Use the loss

$$l(\boldsymbol{a}, \boldsymbol{y}, \boldsymbol{x}) = \frac{1}{2}(\boldsymbol{a}^\top \boldsymbol{x} - \boldsymbol{y})^2,$$

so that

$$L_D(\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{a}_i^\top \boldsymbol{x} - y_i)^2.$$

Now define the data matrix

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_1^\top \\ \vdots \\ \boldsymbol{a}_m^\top \end{pmatrix} \in \mathbb{R}^{m\times n}.$$

Usually, vectors are treated as column vectors, which means $\boldsymbol{x}^\top$ is the corresponding row vector with the same entries.

Fact

Least squares in matrix form.

$$\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 = \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})_i^2 = \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{a}_i^\top \boldsymbol{x} - y_i)^2 = L_D(\boldsymbol{x}),$$

where $\boldsymbol{A}$ is an $m \times n$ matrix, $\boldsymbol{x}$ is n-dimensional, and $\boldsymbol{y}$ is m-dimensional.

1.6 Linear Algebra Review

Vectors are (by default) *column* vectors in $\mathbb{R}^n$.

Special vectors.

$\mathbf{0}$	all-zero vector,
$\boldsymbol{e}$	all-one vector,
$\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$	standard basis vectors.

1.6.1 Linear Dependence

Say $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in \mathbb{R}^n$ are **dependent** if there exist scalars $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$, not all 0, such that

$$\alpha_1 \boldsymbol{v}_1 + \cdots + \alpha_k \boldsymbol{v}_k = \mathbf{0}.$$

The left-hand side is called a *linear combination* of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$.

If the *only* solution is $\alpha_1 = \cdots = \alpha_k = 0$, then $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ are **independent**.

1.6.2 Subspaces and Span

A set $S \subseteq \mathbb{R}^n$ is a **subspace** if:

$$\mathbf{0} \in S,$$

and for all $\boldsymbol{x}, \boldsymbol{y} \in S$ and all $\alpha, \beta \in \mathbb{R}$,

$$\alpha\boldsymbol{x} + \beta\boldsymbol{y} \in S.$$

Two extreme examples of subspaces are $\{\mathbf{0}\}$ and $\mathbb{R}^n$.

Given k vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in \mathbb{R}^n$, the set of all linear combinations is their *span*:

$$\operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}.$$

Fact

- A span is always a subspace.
- If $S \subseteq \mathbb{R}^n$ is a subspace, then there exist $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in S$ such that

  $$S = \operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\},$$

  and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ are independent. The set $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ is called a *basis* of S, where $k = \dim(S)$.

  Note that

  $$\operatorname{span}\{\} = \{\mathbf{0}\}.$$

- For a subspace S, all bases of S have the same cardinality. This cardinality is called the *dimension* of S and is denoted $\dim(S)$.
- If $S, T \subseteq \mathbb{R}^n$ are subspaces with $S \subseteq T$, then

  $$\dim(S) \leq \dim(T).$$

  If $\dim(S) = \dim(T)$, then $S = T$.

  Furthermore, if $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ is a basis of S with $k = \dim(S)$, then there exist vectors

  $$\boldsymbol{v}_{k+1}, \ldots, \boldsymbol{v}_{\dim(T)} \in T$$

  such that $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{\dim(T)}\}$ is a basis of T. This is the *Basis Extension Theorem*.

1.6.3 Special Matrices

Special matrices.

- $\mathbf{0}$: all-zero matrix.
- $\boldsymbol{I}$: identity matrix.

Lec 2

2.1 Matrix Concepts

2.1.1 Matrix–Vector Product

Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{x} \in \mathbb{R}^n$. Define $\boldsymbol{y} \in \mathbb{R}^m$ by

$$y_i = \sum_{j=1}^{n} A(i,j)\, x_j, \quad i = 1, \ldots, m.$$

Then $\boldsymbol{y}$ is called the *matrix–vector product* of $\boldsymbol{A}$ and $\boldsymbol{x}$, and we write

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}.$$

Fact

For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and all scalars $\alpha \in \mathbb{R}$:

- $\boldsymbol{A}(\boldsymbol{x} + \boldsymbol{y}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}\boldsymbol{y}$,
- $\boldsymbol{A}(\alpha\boldsymbol{x}) = \alpha(\boldsymbol{A}\boldsymbol{x})$.

A mapping $T : \mathbb{R}^n \to \mathbb{R}^m$ is called a **linear transformation** if and only if it satisfies

$$T(\boldsymbol{x} + \boldsymbol{y}) = T(\boldsymbol{x}) + T(\boldsymbol{y}), \quad T(\alpha\boldsymbol{x}) = \alpha\, T(\boldsymbol{x})$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and all $\alpha \in \mathbb{R}$.

Fact

For any linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$, there exists a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ such that

$$T(\boldsymbol{x}) \equiv \boldsymbol{A}\boldsymbol{x}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Interpretations of $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \iff \boldsymbol{y} = \boldsymbol{A}(:,1)\, x_1 + \cdots + \boldsymbol{A}(:,n)\, x_n \iff \boldsymbol{y} = \begin{pmatrix} \boldsymbol{A}(1,:)\, \boldsymbol{x} \\ \vdots \\ \boldsymbol{A}(m,:)\, \boldsymbol{x} \end{pmatrix},$$

where $\boldsymbol{A}(:,j)$ denotes the j-th column of $\boldsymbol{A}$, and $\boldsymbol{A}(i,:)$ denotes the i-th row.

Example

Identity matrix. For the identity matrix $\boldsymbol{I} \in \mathbb{R}^{n\times n}$,

$$\boldsymbol{I}\boldsymbol{x} = \boldsymbol{x}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

2.1.2 Transpose

For $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, its **transpose** $\boldsymbol{A}^\top \in \mathbb{R}^{n\times m}$ is defined by

$$\boldsymbol{A}^\top(j,i) = A(i,j), \quad \forall i = 1,\dots,m,\ \forall j = 1,\dots,n.$$

For a vector $\boldsymbol{x} \in \mathbb{R}^n$, the transpose $\boldsymbol{x}^\top$ is a row vector with the same entries as $\boldsymbol{x}$.

2.1.3 Inner Product

For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the **inner product** is

$$\boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i.$$

2.1.4 Range and Nullspace

For any $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, define:

(i) $\text{Range}(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$,

(ii) $\text{Null}(\boldsymbol{A}) = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{y} = \boldsymbol{0}\} \subseteq \mathbb{R}^n$.

Fact

$\text{Range}(\boldsymbol{A})$ and $\text{Null}(\boldsymbol{A})$ are subspaces.

Intuitively:

- $\text{Range}(\boldsymbol{A})$ is the *output space* (all vectors that can appear as $\boldsymbol{A}\boldsymbol{x}$),
- $\text{Null}(\boldsymbol{A})$ is the space of vectors that *disappear* under the transformation (they are mapped to $\boldsymbol{0}$).

The **rank** of $\boldsymbol{A}$ is

$$\text{rank}(\boldsymbol{A}) = \dim(\text{Range}(\boldsymbol{A})).$$

Fact

$$\text{rank}(\boldsymbol{A}) \le \min(m,n).$$

We can express the range as

$$\text{Range}(\boldsymbol{A}) = \text{span}\{\boldsymbol{A}(:,1),\dots,\boldsymbol{A}(:,n)\}.$$

Fact

$$\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A}^\top).$$

Example

If $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is **diagonal** (i.e., $A(i,j) = 0$ whenever $i \neq j$), then

$$\operatorname{rank}(\boldsymbol{A}) = \text{number of nonzero diagonal entries of } \boldsymbol{A}.$$

2.1.5 Matrix Multiplication

Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times p}$. Their product $\boldsymbol{C} = \boldsymbol{AB} \in \mathbb{R}^{m\times p}$ is defined by

$$C(i,j) = \sum_{k=1}^{n} A(i,k)\, B(k,j), \quad \forall i = 1,\dots,m,\ \forall j = 1,\dots,p.$$

Fact

Matrix multiplication corresponds to composition of linear transformations:

$$\forall \boldsymbol{x} \in \mathbb{R}^p, \quad \boldsymbol{A}(\boldsymbol{Bx}) = (\boldsymbol{AB})\boldsymbol{x}.$$

Fact

Matrix multiplication is associative:

$$\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C}.$$

In general, matrix multiplication is *not* commutative:

$$\boldsymbol{AB} \neq \boldsymbol{BA}.$$

2.1.6 Square Matrices and Invertibility

Fact

Given $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, the following are equivalent:

(i) There exists a matrix $\boldsymbol{A}^{-1} \in \mathbb{R}^{n\times n}$ such that

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$$

(then $\boldsymbol{A}^{-1}$ is called the *inverse* of $\boldsymbol{A}$).

(ii) $\operatorname{rank}(\boldsymbol{A}) = n \iff \operatorname{Range}(\boldsymbol{A}) = \mathbb{R}^n$.

(iii) $\operatorname{Null}(\boldsymbol{A}) = \{\boldsymbol{0}\}$.

(iv) The columns of $\boldsymbol{A}$ are linearly independent.

If any (and hence all) of these hold, we say $\boldsymbol{A}$ is **invertible** or **nonsingular**.

2.1.7 Properties of the Inverse

Fact

Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$.

(i) If $\boldsymbol{A}$ is invertible, then $\boldsymbol{A}^{-1}$ is uniquely determined and

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}.$$

(ii) If $\boldsymbol{A}$ is invertible, then

$$(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}.$$

(iii) If $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times n}$ are invertible, then $\boldsymbol{AB}$ is invertible and

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}.$$

Similarly for transposes:

$$(\boldsymbol{AB})^{\top} = \boldsymbol{B}^{\top}\boldsymbol{A}^{\top}.$$

(iv) If $\boldsymbol{A}$ is invertible, then $\boldsymbol{A}^{\top}$ is invertible and

$$(\boldsymbol{A}^{-1})^{\top} = (\boldsymbol{A}^{\top})^{-1}.$$

Notation:

$$\boldsymbol{A}^{-\top} := (\boldsymbol{A}^{-1})^{\top}.$$

2.2 Quadratic Functions

2.2.1 Definition of Quadratic Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **quadratic** if it is a sum of terms, each of which is

- constant (e.g. 7),
- linear (e.g. $b_i x_i$),
- or quadratic (e.g. $7x_i^2$ or $8x_i x_j$).

Fact

If $f : \mathbb{R}^n \to \mathbb{R}$ is quadratic, then it can be written in the **standard form**

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\top}\boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^{\top}\boldsymbol{x} + d,$$

where

- $\boldsymbol{H} \in \mathbb{S}^n$ is a **symmetric** matrix:

$$\boldsymbol{H} = \boldsymbol{H}^{\top}, \quad H(i,j) = H(j,i), \quad i,j = 1,\ldots,n,$$

- $\boldsymbol{g} \in \mathbb{R}^n$ is a vector;
- $d \in \mathbb{R}$ is a scalar.

Example

Let

$$f(\boldsymbol{x}) = 2x_1^2 - 7x_1x_2 - 3x_1 - 9.$$

Then we can write

$$f(\boldsymbol{x}) = \frac{1}{2}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^\top \begin{pmatrix} 4 & -7 \\ -7 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} -3 \\ 0 \end{pmatrix}^\top \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 9.$$

Note that $\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x}$ is a scalar.

2.2.2 Positive (Semi)definite Matrices

Let $\boldsymbol{A} \in \mathbb{S}^n$ (symmetric).

- $\boldsymbol{A}$ is **positive semidefinite** (psd) if

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0.$$

  We write $\boldsymbol{A} \succeq \boldsymbol{0}$.

- $\boldsymbol{A}$ is **positive definite** (pd) if

$$\forall \boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}, \quad \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0.$$

  We write $\boldsymbol{A} \succ \boldsymbol{0}$.

Example

$$\boldsymbol{A} \text{ pd} \Rightarrow \boldsymbol{A} \text{ psd},$$

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ is pd},$$

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \text{ is psd, not pd},$$

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \text{ is pd},$$

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \text{ is psd, not pd},$$

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \text{ is not psd},$$

$$\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \text{ is pd}.$$

2.3 Theorem on Quadratic Minimization

Theorem

Consider the quadratic function

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^\top \boldsymbol{x} + d,$$

where $\boldsymbol{H} \in \mathbb{S}^n$. Then f is **bounded below** and has a **minimizer** if and only if

- $\boldsymbol{H}$ is positive semidefinite, and
- $\boldsymbol{g} \in \text{Range}(\boldsymbol{H})$.

In this case, any solution to

$$\boldsymbol{H}\boldsymbol{x} = -\boldsymbol{g}$$

is a minimizer of f.

If $\boldsymbol{H}$ is positive definite, then all the above conditions hold automatically:

$$\boldsymbol{H} \text{ pd } \Rightarrow \boldsymbol{H} \text{ psd},$$

$$\boldsymbol{H} \text{ pd } \Rightarrow \boldsymbol{H} \text{ nonsingular } \Rightarrow \text{Range}(\boldsymbol{H}) = \mathbb{R}^n \Rightarrow \boldsymbol{g} \in \text{Range}(\boldsymbol{H}).$$

Why a positive definite matrix is nonsingular?

If $\boldsymbol{H}$ were singular, then there would exist $\boldsymbol{x} \in \text{Null}(\boldsymbol{H}) \setminus \{\boldsymbol{0}\}$ with $\boldsymbol{H}\boldsymbol{x} = \boldsymbol{0}$. Then

$$\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{0} = 0,$$

contradicting the condition for positive definiteness $\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$.

Fact

If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and f has a minimizer, then f is bounded below. (The converse is not always true.)

2.3.1 Proof of Theorem (Forward Direction)

Proof

Assume $\boldsymbol{H}$ is positive semidefinite and $\boldsymbol{g} \in \text{Range}(\boldsymbol{H})$. Then $-\boldsymbol{g} \in \text{Range}(\boldsymbol{H})$, so there exists $\boldsymbol{w} \in \mathbb{R}^n$ such that

$$\boldsymbol{H}\boldsymbol{w} = -\boldsymbol{g}.$$

Let $\boldsymbol{x} \in \mathbb{R}^n$ be arbitrary. Then

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^\top \boldsymbol{x} + d \\ &= \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} - \boldsymbol{w}^\top \boldsymbol{H}\boldsymbol{x} + d \quad (\text{since } \boldsymbol{g} = -\boldsymbol{H}\boldsymbol{w}) \\ &= \frac{1}{2}(\boldsymbol{x} - \boldsymbol{w})^\top \boldsymbol{H}(\boldsymbol{x} - \boldsymbol{w}) - \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{H}\boldsymbol{w} + d \quad (\text{completing the square}). \end{aligned}$$

Note that $\boldsymbol{w}^\top \boldsymbol{H}\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{w}$ since both are scalars.
Because $\boldsymbol{H}$ is positive semidefinite,

$$\frac{1}{2}(\boldsymbol{x}-\boldsymbol{w})^\top \boldsymbol{H}(\boldsymbol{x}-\boldsymbol{w}) \geq 0 \quad \text{for all } \boldsymbol{x},$$

and the minimum value 0 is attained at $\boldsymbol{x} = \boldsymbol{w}$.
The remaining term $-\frac{1}{2}\boldsymbol{w}^\top \boldsymbol{H}\boldsymbol{w} + d$ is constant, so f is bounded below and attains its minimum at $\boldsymbol{x} = \boldsymbol{w}$.

2.3.2 Tools for the Converse Direction (Unfinished)

Let $S \subseteq \mathbb{R}^n$ be a subspace. Its **orthogonal complement** is

$$S^\perp = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x}^\top \boldsymbol{y} = 0,\ \forall\, \boldsymbol{y} \in S\}.$$

If $\boldsymbol{x}^\top \boldsymbol{y} = 0$, we say that $\boldsymbol{x}$ and $\boldsymbol{y}$ are **orthogonal**.

Fact

For a subspace $S \subseteq \mathbb{R}^n$:

- $S^\perp$ is a subspace.
- $(S^\perp)^\perp = S$.
- $\dim(S) + \dim(S^\perp) = n$.
- For every $\boldsymbol{w} \in \mathbb{R}^n$, there exist unique $\boldsymbol{x} \in S$ and $\boldsymbol{y} \in S^\perp$ such that

$$\boldsymbol{w} = \boldsymbol{x} + \boldsymbol{y}.$$

Fact

Fundamental Theorem of Linear Algebra. For all $\boldsymbol{A} \in \mathbb{R}^{m\times n}$,

$$\text{Range}(\boldsymbol{A})^\perp = \text{Null}(\boldsymbol{A}^\top).$$

Contrapositive statement (for the theorem).
If $\boldsymbol{H}$ is not positive semidefinite, *or* $\boldsymbol{H}$ is positive semidefinite and $\boldsymbol{g} \notin \text{Range}(\boldsymbol{H})$, then the quadratic function

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^\top \boldsymbol{x} + d$$

is **unbounded below**. (Proof of this converse direction is finished in next lecture.)

Lec 3

3.1 Quadratic Functions

3.1.1 The Minimizer

Given $\boldsymbol{H} \in \mathbb{S}^n$, $\boldsymbol{g} \in \mathbb{R}^n$, $d \in \mathbb{R}$, consider

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^\top \boldsymbol{x} + d.$$

Theorem

The function f is bounded below and has a minimizer if and only if:

a) $\boldsymbol{H}$ is positive semidefinite, and

b) $\boldsymbol{g} \in \text{Range}(\boldsymbol{H})$.

In this case, any solution of

$$\boldsymbol{H}\boldsymbol{x} = -\boldsymbol{g}$$

is a minimizer.

Fact

Contrapositive Theorem. If $\boldsymbol{H}$ is not positive semidefinite, or $\boldsymbol{H}$ is positive semidefinite and $\boldsymbol{g} \notin \text{Range}(\boldsymbol{H})$, then f is unbounded below.

3.1.2 Proof of the Contrapositive

Proof

Case 1: (a) fails: $\boldsymbol{H}$ is not positive semidefinite.
Then there exists $\boldsymbol{z} \in \mathbb{R}^n$ such that

$$\boldsymbol{z}^\top \boldsymbol{H}\boldsymbol{z} < 0.$$

Define the univariate function

$$q(t) = f(t\boldsymbol{z}) = \frac{1}{2}t^2 \boldsymbol{z}^\top \boldsymbol{H}\boldsymbol{z} + t\,\boldsymbol{g}^\top \boldsymbol{z} + d.$$

This is a quadratic function in t with *negative* leading coefficient $\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{H}\boldsymbol{z} < 0$, so $q(t)$ attains arbitrarily negative values as $t \to \pm\infty$. Therefore f has no lower bound and has no minimizer.

Case 2: (a) holds but (b) fails: $\boldsymbol{H}$ is positive semidefinite and $\boldsymbol{g} \notin \text{Range}(\boldsymbol{H})$.
By the Fundamental Theorem of Linear Algebra for symmetric $\boldsymbol{H}$,

$$\mathbb{R}^n = \text{Range}(\boldsymbol{H}) \oplus \text{Null}(\boldsymbol{H}),$$

with $\text{Range}(\boldsymbol{H}) \perp \text{Null}(\boldsymbol{H})$. Thus there exist

$$\boldsymbol{g}_1 \in \text{Range}(\boldsymbol{H}), \quad \boldsymbol{g}_2 \in \text{Null}(\boldsymbol{H})$$

such that

$$\boldsymbol{g} = \boldsymbol{g}_1 + \boldsymbol{g}_2, \quad \boldsymbol{g}_1^\top \boldsymbol{g}_2 = 0.$$

Define

$$q(t) := f(t\boldsymbol{g}_2).$$

Since $\boldsymbol{H}\boldsymbol{g}_2 = \boldsymbol{0}$ (because $\boldsymbol{g}_2 \in \text{Null}(\boldsymbol{H})$), we have

$$\begin{aligned} q(t) &= \frac{1}{2}t^2 \boldsymbol{g}_2^\top \boldsymbol{H}\boldsymbol{g}_2 + t\,\boldsymbol{g}^\top \boldsymbol{g}_2 + d \\ &= t\,\boldsymbol{g}^\top \boldsymbol{g}_2 + d. \end{aligned}$$

Next,

$$\boldsymbol{g}^\top \boldsymbol{g}_2 = (\boldsymbol{g}_1 + \boldsymbol{g}_2)^\top \boldsymbol{g}_2 = \boldsymbol{g}_2^\top \boldsymbol{g}_2 > 0,$$

because $\boldsymbol{g} \notin \text{Range}(\boldsymbol{H}) \Rightarrow \boldsymbol{g} \neq \boldsymbol{g}_1 \Rightarrow \boldsymbol{g}_2 \neq \boldsymbol{0}$.
Hence $q(t)$ is a linear function of t with positive slope. Therefore

$$q(t) \to -\infty \quad \text{as} \quad t \to -\infty,$$

so f is unbounded below.

3.2 Linear Least Squares

3.2.1 Norms

For $\boldsymbol{x} \in \mathbb{R}^n$, the Euclidean norm (or 2-norm) is

$$\|\boldsymbol{x}\| = \sqrt{x(1)^2 + \cdots + x(n)^2}.$$

A **norm** $\|\cdot\|$ on $\mathbb{R}^n$ satisfies, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and all $\lambda \in \mathbb{R}$:

(i) Positivity:

$$\|\boldsymbol{x}\| > 0 \text{ for } \boldsymbol{x} \neq \boldsymbol{0}, \quad \text{and} \quad \|\boldsymbol{x}\| = 0 \Rightarrow \boldsymbol{x} = \boldsymbol{0}.$$

(ii) Positive 1-homogeneity:

$$\|\lambda \boldsymbol{x}\| = |\lambda| \cdot \|\boldsymbol{x}\|.$$

(iii) Triangle inequality:

$$\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|.$$

3.2.2 Linear Least Squares (LLS)

Given $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{y} \in \mathbb{R}^m$, we want $\boldsymbol{x} \in \mathbb{R}^n$ to solve

$$\min_{\boldsymbol{x}} \ \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|.$$

This is equivalent to

$$\min_{\boldsymbol{x}} \ \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 \quad \text{or} \quad \min_{\boldsymbol{x}} \ \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2.$$

Compute:

$$\begin{aligned}\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|^2 &= \frac{1}{2}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y})^\top(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y})\\ &= \frac{1}{2}(\boldsymbol{x}^\top\boldsymbol{A}^\top-\boldsymbol{y}^\top)(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y})\\ &= \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x}-\frac{1}{2}\boldsymbol{y}^\top\boldsymbol{A}\boldsymbol{x}-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{y}+\frac{1}{2}\boldsymbol{y}^\top\boldsymbol{y}\\ &= \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}^\top\boldsymbol{A}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}^\top\boldsymbol{y},\end{aligned}$$

since $\boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{y}=\boldsymbol{y}^\top\boldsymbol{A}\boldsymbol{x}$ (both scalars).

This is a quadratic function in $\boldsymbol{x}$ with

$$\boldsymbol{H}:=\boldsymbol{A}^\top\boldsymbol{A},\quad \boldsymbol{g}:=-\boldsymbol{A}^\top\boldsymbol{y},\quad d:=\frac{1}{2}\boldsymbol{y}^\top\boldsymbol{y}.$$

Theorem

For all $\boldsymbol{A}\in\mathbb{R}^{m\times n}$, the matrix $\boldsymbol{A}^\top\boldsymbol{A}$ is positive semidefinite.

Proof

$$\boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x}=(\boldsymbol{A}\boldsymbol{x})^\top\boldsymbol{A}\boldsymbol{x}=\|\boldsymbol{A}\boldsymbol{x}\|^2\geq 0,\quad \forall\boldsymbol{x}\in\mathbb{R}^n.$$

Theorem

For all $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ and all $\boldsymbol{y}\in\mathbb{R}^m$,

$$\boldsymbol{A}^\top\boldsymbol{y}\in\text{Range}(\boldsymbol{A}^\top\boldsymbol{A}).$$

Proof

(Idea.) Otherwise, by the quadratic theorem for

$$\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|^2=\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}^\top\boldsymbol{A}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}^\top\boldsymbol{y},$$

the objective would be unbounded below, contradicting the fact that squared norms are always nonnegative. Therefore a minimizer exists and must satisfy

$$\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x}=\boldsymbol{A}^\top\boldsymbol{y}.$$

This system is called the **system of normal equations**. The solution is unique if $\boldsymbol{A}^\top\boldsymbol{A}$ is positive definite.

Theorem

For $\boldsymbol{A}\in\mathbb{R}^{m\times n}$, the following are equivalent:

$$\boldsymbol{A}^\top\boldsymbol{A}\text{ is positive definite}\iff\boldsymbol{A}^\top\boldsymbol{A}\text{ is nonsingular}\iff\text{rank}(\boldsymbol{A})=n.$$

Proof

(Sketch.) If $\operatorname{rank}(\boldsymbol{A}) = n$, the columns of $\boldsymbol{A}$ are independent, so

$$\boldsymbol{A}\boldsymbol{x} \neq \boldsymbol{0} \quad \forall \boldsymbol{x} \neq \boldsymbol{0}.$$

Hence

$$\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} = \|\boldsymbol{A}\boldsymbol{x}\|^2 > 0 \quad \forall \boldsymbol{x} \neq \boldsymbol{0},$$

so $\boldsymbol{A}^\top \boldsymbol{A}$ is positive definite and thus nonsingular. The other implications follow from standard linear algebra facts about positive definite and nonsingular matrices.

3.3 Orthogonal Matrices

Recall: if $\boldsymbol{x}^\top \boldsymbol{y} = 0$, we say $\boldsymbol{x}$ and $\boldsymbol{y}$ are **orthogonal**.

If $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ is an orthonormal set with

$$\boldsymbol{x}_i^\top \boldsymbol{x}_j = \delta_{ij},$$

and $\boldsymbol{U}$ is the matrix having these vectors as columns, then

$$\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}, \quad \operatorname{rank}(\boldsymbol{U}) = k \leq n.$$

If $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is square and satisfies

$$\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I},$$

then $\boldsymbol{Q}$ is an **orthogonal matrix**. In this case,

$$\boldsymbol{Q}^{-1} = \boldsymbol{Q}^\top,$$

and the rows of $\boldsymbol{Q}$ are also orthonormal.

Fact

- If $\boldsymbol{Q}$ is orthogonal, then $\boldsymbol{Q}^\top$ is also orthogonal.
- If $\boldsymbol{Q}_1, \boldsymbol{Q}_2$ are orthogonal, then $\boldsymbol{Q}_1\boldsymbol{Q}_2$ is orthogonal as well. (The set

  $$O(n) = \{\boldsymbol{Q} \in \mathbb{R}^{n \times n} : \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}\}$$

  is a *Lie group*.)
- For any orthogonal $\boldsymbol{Q}$ and any $\boldsymbol{x} \in \mathbb{R}^n$,

  $$\|\boldsymbol{Q}\boldsymbol{x}\| = \|\boldsymbol{x}\|.$$

Orthogonal matrices represent rotations or reflections.

3.4 Eigendecomposition

3.4.1 Spectral Theorem

Theorem

For any $\boldsymbol{A} \in \mathbb{S}^n$, there exists an orthogonal matrix $\boldsymbol{Q}$ and a diagonal matrix $\boldsymbol{D}$ such that

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top.$$

The diagonal entries of $\boldsymbol{D}$ are the eigenvalues of $\boldsymbol{A}$ (unique up to ordering), and the columns of $\boldsymbol{Q}$ are the corresponding eigenvectors.
Equivalently,

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top \iff \boldsymbol{A}\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{D}.$$

3.4.2 Characterizations of Positive Semidefiniteness

Theorem

For $\boldsymbol{A} \in \mathbb{S}^n$, the following are equivalent:

(i) $\boldsymbol{A}$ is positive semidefinite.

(ii) All eigenvalues of $\boldsymbol{A}$ are ≥ 0.

(iii) There exists a matrix $\boldsymbol{G}$ such that $\boldsymbol{A} = \boldsymbol{G}\boldsymbol{G}^\top$.

3.4.3 Characterizations of Positive Definiteness

Theorem

For $\boldsymbol{A} \in \mathbb{S}^n$, the following are equivalent:

(i) $\boldsymbol{A}$ is positive definite.

(ii) All eigenvalues of $\boldsymbol{A}$ are > 0.

(iii) There exists a matrix $\boldsymbol{G}$ such that $\boldsymbol{A} = \boldsymbol{G}\boldsymbol{G}^\top$ and $\operatorname{rank}(\boldsymbol{G}) = n$.

Notation.

- $\boldsymbol{A} \succeq \boldsymbol{B}$ means $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite.
- $\boldsymbol{A} \succ \boldsymbol{B}$ means $\boldsymbol{A} - \boldsymbol{B}$ is positive definite.
- $\mathbb{S}^n_+$ denotes the set of positive semidefinite matrices.
- $\mathbb{S}^n_{++}$ denotes the set of positive definite matrices.

3.5 Topology Review

Point-set topology of $\mathbb{R}^n$.

The **open ball** of radius $r > 0$ around $\boldsymbol{x} \in \mathbb{R}^n$ is

$$\mathbb{B}(\boldsymbol{x}, r) = \{\boldsymbol{y} \in \mathbb{R}^n : \|\boldsymbol{y} - \boldsymbol{x}\| < r\}.$$

The **closed ball** of radius $r > 0$ around $\boldsymbol{x} \in \mathbb{R}^n$ is

$$\overline{\mathbb{B}}(\boldsymbol{x}, r) = \{\boldsymbol{y} \in \mathbb{R}^n : \|\boldsymbol{y} - \boldsymbol{x}\| \leq r\}.$$

Fact

- A set $S \subseteq \mathbb{R}^n$ is **open** if for every $\boldsymbol{x} \in S$ there exists $r > 0$ such that
$$\mathbb{B}(\boldsymbol{x}, r) \subseteq S.$$
- A set $S \subseteq \mathbb{R}^n$ is **closed** if every convergent sequence in S has its limit in S.

A set that contains only part of its boundary is, in general, neither open nor closed.

Lec 4

4.1 Topology Review

A set $S \subseteq \mathbb{R}^n$ is **compact** if it is closed and bounded (there is a more general notion of compactness in pure math).

A set $S \subseteq \mathbb{R}^n$ is **bounded** if $\exists r > 0$ such that

$$S \subseteq \mathbb{B}(\mathbf{0}, r).$$

Theorem

If $S \subseteq \mathbb{R}^n$ is compact, nonempty, and $f : S \to \mathbb{R}$ is continuous, then there exist $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S$ such that

$$f(\boldsymbol{x}_1) \leq f(\boldsymbol{x}) \leq f(\boldsymbol{x}_2), \quad \forall \boldsymbol{x} \in S.$$

(That is, f attains both its minimum and maximum on S.)

Theorem

If $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ is an arbitrary infinite sequence in a compact set S, then there exists a subsequence

$$\boldsymbol{x}_{k_1}, \boldsymbol{x}_{k_2}, \ldots \quad \text{with} \quad k_1 < k_2 < \cdots$$

that converges to some $\boldsymbol{x} \in S$.

4.2 Basics of Continuous Optimization

4.2.1 Minimizers

Let $\Omega \subseteq \mathbb{R}^n$, and let $f : \Omega \to \mathbb{R}$ (we will also write $\mathrm{dom}(f) = \Omega$).

We say $\boldsymbol{x}^* \in \mathrm{dom}(f)$ is a **local minimizer** of f if there exists $r > 0$ such that

$$f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathrm{dom}(f) \cap \mathbb{B}(\boldsymbol{x}^*, r).$$

We say $\boldsymbol{x}^*$ is a **global minimizer** if

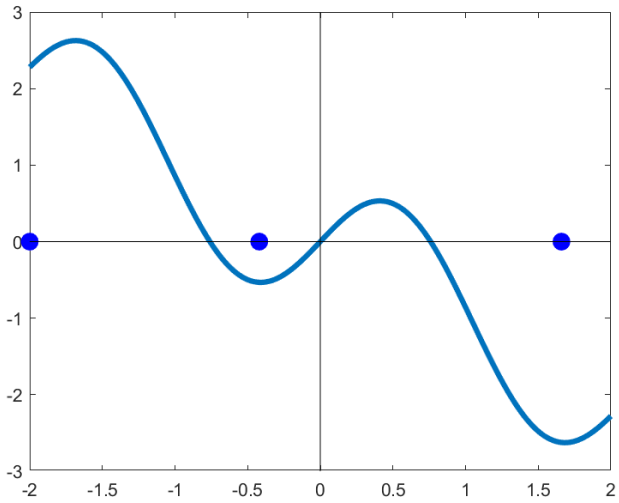
$$f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathrm{dom}(f).$$

Example

Let $n = 1$ and $\mathrm{dom}(f) = [a, b]$. In the picture below, the circled blue points are local minimizers; the third local minimizer is also the global minimizer.

We say $\boldsymbol{x}^* \in \Omega$ is a **strict local minimizer** if there exists $r > 0$ such that

$$f(\boldsymbol{x}^*) < f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in (\mathrm{dom}(f) \cap \mathbb{B}(\boldsymbol{x}^*, r)) \setminus \{\boldsymbol{x}^*\}.$$



Remarks:

- $\Omega (= \mathrm{dom}(f))$ could be all of $\mathbb{R}^n$.
- Ω could be specified by constraints.
- Ω could be the set where f is naturally defined (for example, if $f(x) = x - \ln x$, then $\mathrm{dom}(f) = \{x : x > 0\}$).

4.2.2 Derivatives

Let $f : \Omega \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$, and let $\boldsymbol{x} \in \Omega$.

We say f is **differentiable** at $\boldsymbol{x}$ if:

a) $\boldsymbol{x} \in \mathrm{int}(\Omega)$, i.e., there exists $r > 0$ such that $\mathbb{B}(\boldsymbol{x}, r) \subseteq \Omega$, and

b) there exists $\boldsymbol{g} \in \mathbb{R}^n$ (the *gradient* or *derivative* of f at $\boldsymbol{x}$) such that

$$\lim_{\boldsymbol{h} \to \boldsymbol{0},\ \boldsymbol{h} \neq \boldsymbol{0}} \frac{f(\boldsymbol{x} + \boldsymbol{h}) - f(\boldsymbol{x}) - \boldsymbol{g}^\top \boldsymbol{h}}{\|\boldsymbol{h}\|} = 0.$$

Equivalently, differentiability means that there exists a function $\Phi_{\boldsymbol{x}}(\boldsymbol{h})$ such that

$$f(\boldsymbol{x} + \boldsymbol{h}) = f(\boldsymbol{x}) + \boldsymbol{g}^\top \boldsymbol{h} + \Phi_{\boldsymbol{x}}(\boldsymbol{h}),$$

with

$$\lim_{\boldsymbol{h} \to \boldsymbol{0},\ \boldsymbol{h} \neq \boldsymbol{0}} \frac{\Phi_{\boldsymbol{x}}(\boldsymbol{h})}{\|\boldsymbol{h}\|} = 0.$$

Fact

If f is differentiable at $\boldsymbol{x}$, then the derivative is uniquely determined; we denote it by $\nabla f(\boldsymbol{x})$.

Given $\boldsymbol{x} \in \operatorname{int}(\Omega)$, the **partial derivatives** are defined by

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}) = \lim_{h \to 0,\ h \neq 0} \frac{f(\boldsymbol{x} + h\boldsymbol{e}_i) - f(\boldsymbol{x})}{h}, \quad i = 1, \ldots, n.$$

We say $f : \Omega \to \mathbb{R}$ is $\mathcal{C}^1$ or **continuously differentiable** at $\boldsymbol{x}$ if all partial derivatives

$$\frac{\partial f}{\partial x_i}(\boldsymbol{y})$$

exist and are continuous

$$\forall i = 1, \ldots, n, \quad \forall \boldsymbol{y} \in \mathbb{B}(\boldsymbol{x}, r) \text{ for some } r > 0.$$

A similar correction applies to the definition of $\mathcal{C}^2$.

Theorem

If f is $\mathcal{C}^1$ at $\boldsymbol{x}$, then

$$\nabla f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{pmatrix}.$$

Example

(Quadratic.)

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^\top \boldsymbol{x} + d,$$

where $\boldsymbol{H} \in \mathbb{S}^n$, $\boldsymbol{g} \in \mathbb{R}^n$, $d \in \mathbb{R}$.
Equivalently,

$$f(\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n H(i,j)\, x(i)\, x(j) + \sum_{i=1}^n g(i)\, x(i) + d,$$

so

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}) = \sum_{k=1}^n H(i,k)\, x(k) + g(i),$$

and

$$\nabla f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{pmatrix} = \begin{pmatrix} H(1,:)\boldsymbol{x} + g(1) \\ \vdots \\ H(n,:)\boldsymbol{x} + g(n) \end{pmatrix} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}.$$

4.2.3 Second Derivatives and Hessian

Given $\Omega \subseteq \mathbb{R}^n$, we say f is **twice differentiable** at $\boldsymbol{x} \in \mathbb{R}^n$ if:

a) $\boldsymbol{x} \in \text{int}(\Omega)$ (the interior condition ensures that $\boldsymbol{h}$ can come from every direction), and

b) there exist $\boldsymbol{g} \in \mathbb{R}^n$ (the gradient) and $\boldsymbol{H} \in \mathbb{R}^{n \times n}$ (the **Hessian** or **second derivative**) such that

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \boldsymbol{g}^\top \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^\top \boldsymbol{H}\boldsymbol{h} + \psi_{\boldsymbol{x}}(\boldsymbol{h}),$$

where

$$\lim_{\boldsymbol{h}\to\boldsymbol{0},\ \boldsymbol{h}\neq\boldsymbol{0}} \frac{\psi_{\boldsymbol{x}}(\boldsymbol{h})}{\|\boldsymbol{h}\|^2} = 0.$$

If $\boldsymbol{H}$ exists, it is uniquely determined; we denote it by $\nabla^2 f(\boldsymbol{x})$.

We say f is $\mathcal{C}^2$ at $\boldsymbol{x}$ if $\nabla^2 f(\boldsymbol{y})$ is continuous for $\boldsymbol{y}$ in some open ball around $\boldsymbol{x}$.

Theorem

If f is $\mathcal{C}^2$ at $\boldsymbol{x}$, then

$$\nabla^2 f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\boldsymbol{x}) \end{pmatrix},$$

and this matrix is symmetric:

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\boldsymbol{x}), \quad \forall i, j = 1, \ldots, n.$$

Example

(Quadratic, Hessian.)

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^\top \boldsymbol{x} + d.$$

Then

$$\nabla f(\boldsymbol{x}) = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}, \qquad \nabla^2 f(\boldsymbol{x}) = \boldsymbol{H}.$$

4.2.4 Taylor's Theorem Variants

Let $f : \Omega \to \mathbb{R}$ be $\mathcal{C}^1$, and let $\boldsymbol{x} \in \text{int}(\Omega)$. By the **Fundamental Theorem of Calculus**,

$$f(\boldsymbol{x}+\boldsymbol{p}) = f(\boldsymbol{x}) + \int_0^1 \nabla f(\boldsymbol{x}+\gamma\boldsymbol{p})^\top \boldsymbol{p}\, \mathrm{d}\gamma.$$

Applying the **Mean Value Theorem for integrals**, we obtain

$$f(\boldsymbol{x}+\boldsymbol{p}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}+\gamma\boldsymbol{p})^\top \boldsymbol{p}$$

for some $\gamma \in [0,1]$.

(Here we are implicitly assuming that $\boldsymbol{x} \in \text{int}(\Omega)$, $\boldsymbol{x}+\boldsymbol{p} \in \text{int}(\Omega)$, and $\boldsymbol{x}+\gamma\boldsymbol{p} \in \text{int}(\Omega)$ for all $\gamma \in [0,1]$.)

These results require that $\boldsymbol{x}$, $\boldsymbol{x}+\boldsymbol{p}$, and $\boldsymbol{x}+\gamma\boldsymbol{p}$ all lie in $\text{int}(\Omega)$ for every $\gamma \in [0,1]$.

4.2.5 Convexity

A set $C \subseteq \mathbb{R}^n$ is **convex** if

$$\forall \boldsymbol{x} \in C,\ \forall \boldsymbol{p} \text{ with } \boldsymbol{x}+\boldsymbol{p} \in C,\ \text{we also have } \boldsymbol{x}+\gamma\boldsymbol{p} \in C \quad \forall \gamma \in [0,1].$$

Equivalently, C is convex if and only if

$$\forall \boldsymbol{x}, \boldsymbol{y} \in C,\ \forall \gamma \in [0,1], \quad (1-\gamma)\boldsymbol{x}+\gamma\boldsymbol{y} \in C.$$

Example

(i) Some trivial examples: $\varnothing$, $\{\boldsymbol{x}_0\}$, $\mathbb{R}^n$.

(ii) **Affine set.** Given $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$,

$$C = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$$

is an affine set.

An equivalent definition: C is **affine** if and only if

$$\forall \boldsymbol{x}, \boldsymbol{y} \in C,\ \forall \gamma \in \mathbb{R}, \quad (1-\gamma)\boldsymbol{x}+\gamma\boldsymbol{y} \in C,$$

where γ is allowed to be any real number (not just in $[0,1]$).

Note: Affine sets are always convex.

(iii) Open or closed ball in any norm:

$$\mathbb{B}_{\square}(\boldsymbol{x}, r) = \{\boldsymbol{y} \in \mathbb{R}^n : \|\boldsymbol{y}-\boldsymbol{x}\|_{\square} < r\}, \quad \overline{\mathbb{B}}_{\square}(\boldsymbol{x}, r) = \{\boldsymbol{y} \in \mathbb{R}^n : \|\boldsymbol{y}-\boldsymbol{x}\|_{\square} \le r\}.$$

(iv) **Halfspace.** Given $\boldsymbol{A} \in \mathbb{R}^n$ and $b \in \mathbb{R}$,

$$H = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}^\top \boldsymbol{x} \le b\}.$$

(v) **Polyhedron.** A polyhedron is the intersection of a finite number of halfspaces.

Given $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m \in \mathbb{R}^n$ and $b_1, \ldots, b_m \in \mathbb{R}$, define

$$C = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}_1^\top \boldsymbol{x} \le b_1, \ldots, \boldsymbol{A}_m^\top \boldsymbol{x} \le b_m\}.$$

Let

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_1^\top \\ \vdots \\ \boldsymbol{A}_m^\top \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}.$$

Then equivalently,

$$C = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}\},$$

where the inequality is interpreted componentwise.

Note: For $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$, we write $\boldsymbol{u} \leq \boldsymbol{v}$ if

$$u(i) \leq v(i), \quad \forall i = 1, \ldots, m.$$

The same componentwise interpretation holds for $<$, $>$, and $\geq$.

Note: Affine sets are a special case of polyhedra. Given $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$,

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \Longleftrightarrow \begin{pmatrix} \boldsymbol{A} \\ -\boldsymbol{A} \end{pmatrix} \boldsymbol{x} \leq \begin{pmatrix} \boldsymbol{b} \\ -\boldsymbol{b} \end{pmatrix}.$$

4.2.6 Lipschitz Continuity

Let $\boldsymbol{f} : \Omega \to \mathbb{R}^m$, $\Omega \subseteq \mathbb{R}^n$. We say $\boldsymbol{f}$ is **Lipschitz continuous** with modulus L if

$$\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \Omega.$$

Lec 5

5.1 L-smooth Functions

Given $\boldsymbol{f} : \Omega \to \mathbb{R}^m$, $\Omega \subseteq \mathbb{R}^n$, we say $\boldsymbol{f}$ is **Lipschitz continuous** with modulus L if

$$\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \Omega.$$

Given $f : \Omega \to \mathbb{R}$, we say f is **L-smooth** (smooth with modulus L) if f is differentiable and its gradient is L-Lipschitz continuous, i.e.,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \Omega.$$

Lemma

If f is L-smooth and $\mathrm{dom}(f)$ is open and convex, then

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f).$$

(This looks like a Taylor expansion with a quadratic remainder term.)

Proof

$$\begin{aligned}
&f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \\
&\qquad = \left(\int_0^1 \nabla f(\boldsymbol{x} + \gamma(\boldsymbol{y} - \boldsymbol{x}))^\top (\boldsymbol{y} - \boldsymbol{x}) \, \mathrm{d}\gamma \right) - \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \\
&\qquad = \int_0^1 \left[\nabla f(\boldsymbol{x} + \gamma(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}) \right]^\top (\boldsymbol{y} - \boldsymbol{x}) \, \mathrm{d}\gamma \\
&\qquad \leq \int_0^1 \|\nabla f(\boldsymbol{x} + \gamma(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\| \, \mathrm{d}\gamma \quad \text{(Cauchy–Schwarz)} \\
&\qquad \leq \int_0^1 L\|\boldsymbol{x} + \gamma(\boldsymbol{y} - \boldsymbol{x}) - \boldsymbol{x}\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\| \, \mathrm{d}\gamma \quad (L\text{-smoothness}) \\
&\qquad = \int_0^1 L\gamma \|\boldsymbol{y} - \boldsymbol{x}\|^2 \, \mathrm{d}\gamma \\
&\qquad = L\|\boldsymbol{y} - \boldsymbol{x}\|^2 \int_0^1 \gamma \, \mathrm{d}\gamma \\
&\qquad = \frac{1}{2} L\|\boldsymbol{y} - \boldsymbol{x}\|^2.
\end{aligned}$$

Rearranging gives the desired inequality.

5.2 Necessary and Sufficient Conditions for Optimality

5.2.1 First-order Necessary Condition

Theorem

Let $f : \Omega \to \mathbb{R}$ be differentiable at a point $\boldsymbol{x}^* \in \Omega$, and assume that $\boldsymbol{x}^*$ is an interior point of Ω and a local minimizer of f. Then

$$\nabla f(\boldsymbol{x}^*) = \mathbf{0}.$$

Proof

(Contrapositive.) Assume $\nabla f(\boldsymbol{x}^*) \neq \mathbf{0}$. Let

$$\boldsymbol{p} := -\nabla f(\boldsymbol{x}^*).$$

By the definition of differentiability, there exists a function $\phi_{\boldsymbol{x}^*}$ such that, for all $|\alpha|$ sufficiently small,

$$f(\boldsymbol{x}^* + \alpha \boldsymbol{p}) = f(\boldsymbol{x}^*) + \alpha \nabla f(\boldsymbol{x}^*)^\top \boldsymbol{p} + \phi_{\boldsymbol{x}^*}(\alpha \boldsymbol{p}),$$

and

$$\frac{\phi_{\boldsymbol{x}^*}(\alpha \boldsymbol{p})}{\|\alpha \boldsymbol{p}\|} \to 0 \quad \Leftrightarrow \quad \frac{\phi_{\boldsymbol{x}^*}(\alpha \boldsymbol{p})}{|\alpha|} \to 0 \quad \text{as } \alpha \to 0.$$

Since $\boldsymbol{p} = -\nabla f(\boldsymbol{x}^*)$,

$$f(\boldsymbol{x}^* + \alpha \boldsymbol{p}) = f(\boldsymbol{x}^*) - \alpha \|\nabla f(\boldsymbol{x}^*)\|^2 + \phi_{\boldsymbol{x}^*}(\alpha \boldsymbol{p}).$$

Because $\|\boldsymbol{p}\| > 0$ and $\phi_{\boldsymbol{x}^*}(\alpha \boldsymbol{p})/|\alpha| \to 0$, there exists $\bar{\alpha} > 0$ such that

$$\frac{\|\phi_{\boldsymbol{x}^*}(\alpha \boldsymbol{p})\|}{|\alpha| \|\boldsymbol{p}\|} \leq \frac{\|\boldsymbol{p}\|}{2}, \quad \forall |\alpha| \leq \bar{\alpha}.$$

For such $\alpha \in (0, \bar{\alpha}]$,

$$\begin{aligned} f(\boldsymbol{x}^* + \alpha \boldsymbol{p}) &\leq f(\boldsymbol{x}^*) - \alpha \|\nabla f(\boldsymbol{x}^*)\|^2 + \frac{\|\boldsymbol{p}\|}{2} |\alpha| \, \|\boldsymbol{p}\| \\ &= f(\boldsymbol{x}^*) - \frac{\alpha}{2} \|\boldsymbol{p}\|^2 \\ &< f(\boldsymbol{x}^*). \end{aligned}$$

Thus f strictly decreases in direction $\boldsymbol{p}$ away from $\boldsymbol{x}^*$, so $\boldsymbol{x}^*$ cannot be a local minimizer.

5.2.2 Second-order Necessary Condition

Theorem

Suppose f is $\mathcal{C}^2$ at $\boldsymbol{x}^* \in \Omega$, and $\boldsymbol{x}^*$ is a local minimizer. Then

$$\nabla f(\boldsymbol{x}^*) = \mathbf{0} \quad \text{and} \quad \nabla^2 f(\boldsymbol{x}^*) \succeq \mathbf{0} \text{ (positive semidefinite).}$$

Proof

(Contrapositive idea.) We already know from Theorem 1 that a local minimizer must satisfy $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$. Now assume $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ but $\nabla^2 f(\boldsymbol{x}^*)$ is not positive semidefinite. Then there exists $\boldsymbol{p}$ such that

$$\boldsymbol{p}^\top \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{p} < 0.$$

By the second-order Taylor expansion,

$$f(\boldsymbol{x}^* + \alpha \boldsymbol{p}) = f(\boldsymbol{x}^*) + \alpha \nabla f(\boldsymbol{x}^*)^\top \boldsymbol{p} + \frac{1}{2}\alpha^2 \boldsymbol{p}^\top \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{p} + \psi_{\boldsymbol{x}^*}(\alpha \boldsymbol{p}),$$

and we know

$$\frac{\psi_{\boldsymbol{x}^*}(\alpha \boldsymbol{p})}{\alpha^2} \to 0 \quad \text{as } \alpha \to 0.$$

Choose $\bar{\alpha} > 0$ sufficiently small such that

$$\frac{|\psi_{\boldsymbol{x}^*}(\alpha \boldsymbol{p})|}{\alpha^2} \leq \frac{1}{4} |\boldsymbol{p}^\top \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{p}|, \quad \forall \alpha \in (0, \bar{\alpha}].$$

Using $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$, we get for all $\alpha \in (0, \bar{\alpha}]$,

$$\begin{aligned} f(\boldsymbol{x}^* + \alpha \boldsymbol{p}) &\leq f(\boldsymbol{x}^*) + \frac{1}{2}\alpha^2 \boldsymbol{p}^\top \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{p} + \frac{1}{4}\alpha^2 |\boldsymbol{p}^\top \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{p}| \\ &= f(\boldsymbol{x}^*) - \frac{1}{2}\alpha^2 |\boldsymbol{p}^\top \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{p}| + \frac{1}{4}\alpha^2 |\boldsymbol{p}^\top \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{p}| \\ &= f(\boldsymbol{x}^*) - \frac{\alpha^2}{4} |\boldsymbol{p}^\top \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{p}| \\ &< f(\boldsymbol{x}^*), \end{aligned}$$

so $\boldsymbol{x}^*$ is not a local minimizer.

Example

(Quadratic and Least Squares.)
Let

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{H} \boldsymbol{x} + \boldsymbol{g}^\top \boldsymbol{x} + d.$$

Then

$$\nabla f(\boldsymbol{x}) = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g},$$

and

$$\nabla f(\boldsymbol{x}) = \mathbf{0} \iff \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g} = \mathbf{0} \iff \boldsymbol{H}\boldsymbol{x} = -\boldsymbol{g}.$$

Also,

$$\nabla^2 f(\boldsymbol{x}) = \boldsymbol{H},$$

so

$$\nabla^2 f(\boldsymbol{x}) \text{ is positive semidefinite} \iff \boldsymbol{H} \text{ is positive semidefinite.}$$

For quadratics, these conditions are **necessary and sufficient** for optimality.
For linear least squares

$$f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2,$$

we have

$$\nabla f(\boldsymbol{x}) = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}), \quad \nabla^2 f(\boldsymbol{x}) = \boldsymbol{A}^\top \boldsymbol{A}.$$

5.2.3 Second-order Sufficient Condition

Theorem

Let $f : \Omega \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$. Suppose $\boldsymbol{x}^* \in \text{int}(\Omega)$, f is $\mathcal{C}^2$ at $\boldsymbol{x}^*$, $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$, and

$$\nabla^2 f(\boldsymbol{x}^*) \succ \boldsymbol{0}.$$

Then $\boldsymbol{x}^*$ is a **strict local minimizer** of f.

Proof

Lemma

If $\boldsymbol{U} \succ \boldsymbol{0}$, then there exists $r > 0$ such that

$$\boldsymbol{V} \succ \boldsymbol{0}, \quad \forall \boldsymbol{V} \in \mathbb{B}(\boldsymbol{U}, r),$$

where $\mathbb{B}(\boldsymbol{U}, r) \subseteq \mathbb{S}^n$ is the Frobenius-norm ball:

$$\|\boldsymbol{X}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n X(i,j)^2}.$$

Equivalently, the set of positive definite matrices is an open subset of $\mathbb{S}^n$.

Proof

Proof of Lemma.

- The eigenvalues are continuous functions of the matrix entries.
- If $\boldsymbol{U} \succ \mathbf{0}$, then $\lambda_{\min}(\boldsymbol{U}) > 0$, where

$$\lambda_{\min}(\boldsymbol{U}) = \min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{x}.$$

These imply that for sufficiently small $\boldsymbol{\Delta} \in \mathbb{S}^n$,

$$\begin{aligned}
\lambda_{\min}(\boldsymbol{U}+\boldsymbol{\Delta}) &= \min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^\top(\boldsymbol{U}+\boldsymbol{\Delta})\boldsymbol{x} \\
&\geq \min_{\|\boldsymbol{x}\|=1} \left(\boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{x} - |\boldsymbol{x}^\top \boldsymbol{\Delta}\boldsymbol{x}|\right) \\
&\geq \min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{x} - \min_{\|\boldsymbol{x}\|=1} \left(\|\boldsymbol{x}\| \, \|\boldsymbol{\Delta}\|_F \, \|\boldsymbol{x}\|\right) \\
&= \lambda_{\min}(\boldsymbol{U}) - \|\boldsymbol{\Delta}\|_F \\
&\geq \frac{\lambda_{\min}(\boldsymbol{U})}{2} \quad \text{provided } \|\boldsymbol{\Delta}\|_F \leq \frac{\lambda_{\min}(\boldsymbol{U})}{2}.
\end{aligned}$$

Hence $\lambda_{\min}(\boldsymbol{U}+\boldsymbol{\Delta}) > 0$, so $\boldsymbol{U}+\boldsymbol{\Delta} \succ \mathbf{0}$.

Proof of Theorem. Because f is $\mathcal{C}^2$ at $\boldsymbol{x}^*$ and $\nabla^2 f(\boldsymbol{x}^*) \succ \mathbf{0}$, by the lemma there exists $\rho > 0$ such that

$$\mathbb{B}(\boldsymbol{x}^*, \rho) \subseteq \Omega \quad \text{and} \quad \nabla^2 f(\boldsymbol{x}) \succ \mathbf{0}, \quad \forall \boldsymbol{x} \in \mathbb{B}(\boldsymbol{x}^*, \rho).$$

Choose any $\boldsymbol{y} \in \mathbb{B}(\boldsymbol{x}^*, \rho) \setminus \{\boldsymbol{x}^*\}$. By the second-order Taylor expansion, there exists $\gamma \in [0,1]$ such that

$$f(\boldsymbol{y}) = f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{y}-\boldsymbol{x}^*) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x}^*)^\top \nabla^2 f(\boldsymbol{x}^* + \gamma(\boldsymbol{y}-\boldsymbol{x}^*))(\boldsymbol{y}-\boldsymbol{x}^*).$$

Since $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ and

$$\nabla^2 f(\boldsymbol{x}^* + \gamma(\boldsymbol{y}-\boldsymbol{x}^*)) \succ \mathbf{0} \quad (\text{as } \boldsymbol{x}^* + \gamma(\boldsymbol{y}-\boldsymbol{x}^*) \in \mathbb{B}(\boldsymbol{x}^*, \rho)),$$

we have

$$f(\boldsymbol{y}) > f(\boldsymbol{x}^*).$$

Thus $\boldsymbol{x}^*$ is a strict local minimizer.

5.2.4 Summary of First- and Second-order Conditions

Summary

Let $f : \Omega \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$, and $\boldsymbol{x}^* \in \Omega$. Suppose f is $\mathcal{C}^2$ at $\boldsymbol{x}^*$. Then we have the chain:

$$\begin{aligned} &\nabla f(\boldsymbol{x}^*) = \mathbf{0}, \quad \nabla^2 f(\boldsymbol{x}^*) \succ \mathbf{0} \;\Rightarrow\; \boldsymbol{x}^* \text{ is a strict local minimizer} \\ &\Rightarrow\; \boldsymbol{x}^* \text{ is a local minimizer} \\ &\Rightarrow\; \nabla f(\boldsymbol{x}^*) = \mathbf{0}, \quad \nabla^2 f(\boldsymbol{x}^*) \succeq \mathbf{0}. \end{aligned}$$

This chain of implications cannot be reversed in general.

Example

- $f(x) = x^4$:

  $x^* = 0$ is a strict local minimizer, but $f''(0) = 0$ (not > 0).

- $f(x) = 1$:

  $x^* = 0$ is a local minimizer (in fact, every point is), but not a strict local minimizer.

- $f(x) = x^3$:

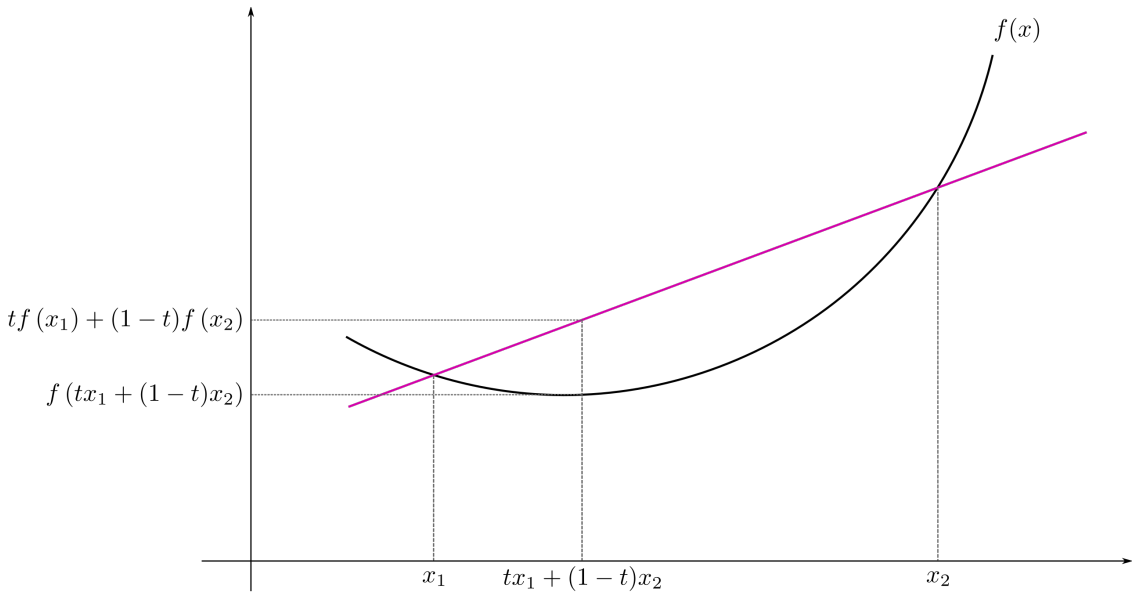
  $f'(0) = 0,\ f''(0) = 0,$ but $x^* = 0$ is not a local minimizer.

5.3 Convex Functions

Let $f : \Omega \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$, and assume $\Omega = \mathrm{dom}(f)$ is a convex set. We say f is a **convex function** if

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega, \ \forall \lambda \in [0,1],$$

$$f\big((1-\lambda)\boldsymbol{x} + \lambda \boldsymbol{y}\big) \leq (1-\lambda) f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}).$$



Lec 6

6.1 Convex Functions

6.1.1 Definitions

Let $f : \Omega \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$, and $\Omega = \mathrm{dom}(f)$. Assume Ω is convex. We say f is a **convex function** if

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega,\ \forall \lambda \in [0,1], \quad f\big((1-\lambda)\boldsymbol{x} + \lambda \boldsymbol{y}\big) \le (1-\lambda) f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}).$$

Extended-value notation. If $f : \Omega \to \mathbb{R}$ is convex, we extend it by setting

$$f(\boldsymbol{x}) = \infty \quad \text{for } \boldsymbol{x} \notin \Omega.$$

Thus we can regard

$$f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\},$$

and define

$$\mathrm{dom}(f) = \{\boldsymbol{x} \in \mathbb{R}^n : f(\boldsymbol{x}) < \infty\}.$$

The convexity inequality above is then interpreted for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$; it enforces that $\mathrm{dom}(f)$ is convex.

6.1.2 Examples

Example

1. $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^\top \boldsymbol{x} + d, \quad \boldsymbol{H} \succeq \boldsymbol{0}.$
2. $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_\square$ (any norm).
3. $f(x) = \mathrm{e}^x$.
4. $f(x) = \begin{cases} -\ln x, & x > 0, \\ \infty, & x \le 0, \end{cases}$
5. $f(x) = \begin{cases} \dfrac{1}{x}, & x > 0, \\ \infty, & x \le 0, \end{cases}$
6. If $g_1, \ldots, g_m : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are convex and $\alpha_1, \ldots, \alpha_m \ge 0$, then $\alpha_1 g_1 + \cdots + \alpha_m g_m$ is convex.
7. Under the same assumption as in 6, $\max\{\alpha_1 g_1, \ldots, \alpha_m g_m\}$ is convex.
   Example: $f(x) = \max(x, -x) = |x|$ is convex.

Theorem

If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and $\boldsymbol{x}^* \in \mathrm{dom}(f)$ is a local minimizer, then $\boldsymbol{x}^*$ is a global minimizer.

Proof

Assume $\boldsymbol{x}^*$ is a local minimizer and let $\boldsymbol{y} \in \mathbb{R}^n$ be arbitrary. Define

$$\boldsymbol{y}_i := \left(1 - \frac{1}{i}\right)\boldsymbol{x}^* + \frac{1}{i}\boldsymbol{y}, \quad i = 1, 2, 3, \ldots$$

Then $\boldsymbol{y}_i \to \boldsymbol{x}^*$ as $i \to \infty$. Since $\boldsymbol{x}^*$ is a local minimizer, there exists k such that

$$f(\boldsymbol{y}_i) \geq f(\boldsymbol{x}^*), \quad \forall i \geq k.$$

In particular,

$$f\left(\left(1 - \frac{1}{k}\right)\boldsymbol{x}^* + \frac{1}{k}\boldsymbol{y}\right) \geq f(\boldsymbol{x}^*).$$

By convexity,

$$f\left(\left(1 - \frac{1}{k}\right)\boldsymbol{x}^* + \frac{1}{k}\boldsymbol{y}\right) \leq \left(1 - \frac{1}{k}\right) f(\boldsymbol{x}^*) + \frac{1}{k} f(\boldsymbol{y}).$$

Combining,

$$\left(1 - \frac{1}{k}\right) f(\boldsymbol{x}^*) + \frac{1}{k} f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*).$$

Rearranging gives

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*).$$

Since $\boldsymbol{y}$ was arbitrary, $\boldsymbol{x}^*$ is a global minimizer.

6.2 Differentiable Convex Functions

Suppose $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and differentiable at $\boldsymbol{x} \in \mathrm{intdom}(f)$. Then for all $\boldsymbol{y} \in \mathbb{R}^n$ and all $\alpha \in [0, 1]$,

$$f((1-\alpha)\boldsymbol{x} + \alpha\boldsymbol{y}) \leq (1-\alpha) f(\boldsymbol{x}) + \alpha f(\boldsymbol{y}).$$

Write the left-hand side using differentiability at $\boldsymbol{x}$:

$$\begin{aligned} f((1-\alpha)\boldsymbol{x} + \alpha\boldsymbol{y}) &= f(\boldsymbol{x} + \alpha(\boldsymbol{y} - \boldsymbol{x})) \\ &= f(\boldsymbol{x}) + \alpha \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \phi(\alpha(\boldsymbol{y} - \boldsymbol{x})), \end{aligned}$$

for some remainder term ϕ with

$$\lim_{\boldsymbol{h} \to \boldsymbol{0},\ \boldsymbol{h} \neq \boldsymbol{0}} \frac{\phi(\boldsymbol{h})}{\|\boldsymbol{h}\|} = 0.$$

Combining with convexity,

$$f(\boldsymbol{x}) + \alpha \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \phi(\alpha(\boldsymbol{y} - \boldsymbol{x})) \leq (1-\alpha) f(\boldsymbol{x}) + \alpha f(\boldsymbol{y}).$$

Assuming $\alpha > 0$, subtract $f(\boldsymbol{x})$ and divide by α:

$$\nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{\phi(\alpha(\boldsymbol{y} - \boldsymbol{x}))}{\alpha} \leq f(\boldsymbol{y}) - f(\boldsymbol{x}).$$

Taking the limit $\alpha \downarrow 0$, the remainder term vanishes and we obtain the

Fact

Subgradient Inequality. For convex differentiable f and any $\boldsymbol{x} \in \text{intdom}(f)$,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}), \quad \forall \boldsymbol{y} \in \mathbb{R}^n.$$

The converse also holds: if this inequality is satisfied for all $\boldsymbol{x}, \boldsymbol{y}$, then f is convex.

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex. Suppose $\boldsymbol{x}^* \in \text{intdom}(f)$, f is differentiable at $\boldsymbol{x}^*$, and $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$. Then $\boldsymbol{x}^*$ is a global minimizer of f.

Proof

Applying the subgradient inequality at $\boldsymbol{x}^*$,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{y} - \boldsymbol{x}^*) = f(\boldsymbol{x}^*), \quad \forall \boldsymbol{y} \in \mathbb{R}^n,$$

since $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$. Hence $\boldsymbol{x}^*$ is a global minimizer.

6.3 Descent Methods

6.3.1 Descent Directions

Let $f : \Omega \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$, and $\boldsymbol{x} \in \Omega$. We say $\boldsymbol{d} \in \mathbb{R}^n$ is a **descent direction** at $\boldsymbol{x}$ if there exists $\bar{t} > 0$ such that

$$\boldsymbol{x} + t\boldsymbol{d} \in \Omega, \ \forall t \in [0, \bar{t}], \quad \text{and} \quad f(\boldsymbol{x} + t\boldsymbol{d}) < f(\boldsymbol{x}), \ \forall t \in (0, \bar{t}].$$

Theorem

Let $f : \Omega \to \mathbb{R}$ with $\Omega \subseteq \mathbb{R}^n$. Assume $\boldsymbol{x} \in \text{int}(\Omega)$ and f is differentiable at $\boldsymbol{x}$. Then any $\boldsymbol{d}$ satisfying

$$\boldsymbol{d}^\top \nabla f(\boldsymbol{x}) < 0$$

is a descent direction at $\boldsymbol{x}$.

Proof

Assume $\nabla f(\boldsymbol{x}) \neq \boldsymbol{0}$ and $\boldsymbol{d} \neq \boldsymbol{0}$ (otherwise the statement is vacuous). Differentiability at $\boldsymbol{x}$ means

$$\lim_{\boldsymbol{h} \to \boldsymbol{0}, \ \boldsymbol{h} \neq \boldsymbol{0}} \frac{f(\boldsymbol{x} + \boldsymbol{h}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top \boldsymbol{h}}{\|\boldsymbol{h}\|} = 0.$$

Take $\boldsymbol{h} = t\boldsymbol{d}$ with $t \downarrow 0$:

$$\lim_{t \downarrow 0} \frac{f(\boldsymbol{x} + t\boldsymbol{d}) - f(\boldsymbol{x}) - t \nabla f(\boldsymbol{x})^\top \boldsymbol{d}}{t} = 0,$$

or equivalently,

$$\lim_{t \downarrow 0} \left(\frac{f(\boldsymbol{x} + t\boldsymbol{d}) - f(\boldsymbol{x})}{t} - \nabla f(\boldsymbol{x})^\top \boldsymbol{d} \right) = 0.$$

By assumption, $\nabla f(\boldsymbol{x})^\top \boldsymbol{d} < 0$. Therefore, for sufficiently small $t > 0$,

$$\frac{f(\boldsymbol{x} + t\boldsymbol{d}) - f(\boldsymbol{x})}{t} < 0,$$

which implies $f(\boldsymbol{x} + t\boldsymbol{d}) < f(\boldsymbol{x})$ for all sufficiently small $t > 0$. Thus $\boldsymbol{d}$ is a descent direction.

Steepest descent direction. Assume $\nabla f(\boldsymbol{x}) \neq \boldsymbol{0}$. Among all unit vectors, the direction that minimizes the directional derivative is

$$\arg\min\{\boldsymbol{d}^\top \nabla f(\boldsymbol{x}) : \|\boldsymbol{d}\| = 1\} = -\frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|}.$$

This follows from the strong form of the Cauchy–Schwarz inequality:

$$\boldsymbol{u}^\top \boldsymbol{v} \leq \|\boldsymbol{u}\| \, \|\boldsymbol{v}\|$$

with equality if and only if either $\boldsymbol{u} = \boldsymbol{0}$ or $\boldsymbol{v} = \boldsymbol{0}$, or there exists $\lambda \geq 0$ such that $\boldsymbol{u} = \lambda \boldsymbol{v}$.

6.3.2 Gradient Descent Algorithm (Steepest Descent)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable, and choose an initial point $\boldsymbol{x}^0 \in \mathbb{R}^n$.

For $k = 0, 1, 2, \ldots$:

- select a step size $\alpha_k > 0$,
- update

$$\boldsymbol{x}^{k+1} := \boldsymbol{x}^k - \alpha_k \nabla f(\boldsymbol{x}^k).$$

End.

Here α_k is the **step size** (learning rate). The procedure used to choose α_k is called a **line search**.

If f is L-smooth, then the constant choice $\alpha_k \equiv \frac{1}{L}$ always works. Any fixed $\alpha_k \in (0, \frac{1}{L})$ also leads to convergence but typically more slowly. A common stopping rule is $\|\nabla f(\boldsymbol{x}^k)\| \leq$ tolerance.

Theorem

(**Convex case of Gradient Descent**) Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L-smooth. Suppose there exists $\boldsymbol{x}^*$ such that

$$\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}.$$

Run gradient descent from $\boldsymbol{x}^0$ with step sizes $\alpha_k \equiv \frac{1}{L}$. Then

$$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*) \leq \frac{L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{2k}, \quad \forall k = 1, 2, \ldots$$

Proof

From L-smoothness,

$$\begin{aligned} f(\boldsymbol{x}^{k+1}) &\leq f(\boldsymbol{x}^k) + \nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) + \frac{L}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 \\ &= f(\boldsymbol{x}^k) + \nabla f(\boldsymbol{x}^k)^\top \left(-\frac{1}{L}\nabla f(\boldsymbol{x}^k)\right) + \frac{L}{2}\left\|-\frac{1}{L}\nabla f(\boldsymbol{x}^k)\right\|^2 \\ &= f(\boldsymbol{x}^k) - \frac{1}{L}\|\nabla f(\boldsymbol{x}^k)\|^2 + \frac{1}{2L}\|\nabla f(\boldsymbol{x}^k)\|^2 \\ &= f(\boldsymbol{x}^k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}^k)\|^2. \end{aligned}$$

From the subgradient inequality for convex differentiable f,

$$f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}^k) + \nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{x}^* - \boldsymbol{x}^k),$$

so

$$f(\boldsymbol{x}^k) \leq f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{x}^k - \boldsymbol{x}^*).$$

Combining,

$$f(\boldsymbol{x}^{k+1}) \leq f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{x}^k - \boldsymbol{x}^*) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}^k)\|^2.$$

Using the update $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \frac{1}{L}\nabla f(\boldsymbol{x}^k)$, we compute

$$\begin{aligned} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2 &= \left\|\boldsymbol{x}^k - \frac{1}{L}\nabla f(\boldsymbol{x}^k) - \boldsymbol{x}^*\right\|^2 \\ &= \|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \frac{2}{L}\nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{x}^k - \boldsymbol{x}^*) + \frac{1}{L^2}\|\nabla f(\boldsymbol{x}^k)\|^2. \end{aligned}$$

Rearranging,

$$\nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{x}^k - \boldsymbol{x}^*) = \frac{L}{2}(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2) + \frac{1}{2L}\|\nabla f(\boldsymbol{x}^k)\|^2.$$

Substitute into the bound for $f(\boldsymbol{x}^{k+1})$:

$$\begin{aligned} f(\boldsymbol{x}^{k+1}) &\leq f(\boldsymbol{x}^*) + \frac{L}{2}(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2) + \frac{1}{2L}\|\nabla f(\boldsymbol{x}^k)\|^2 - \frac{1}{2L}\|\nabla f(\boldsymbol{x}^k)\|^2 \\ &= f(\boldsymbol{x}^*) + \frac{L}{2}(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2). \end{aligned}$$

Thus

$$f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^*) \leq \frac{L}{2}(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2).$$

Summing from $k = 0$ to $k = \ell - 1$,

$$\sum_{k=0}^{\ell-1} (f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^*)) \leq \frac{L}{2}(\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^\ell - \boldsymbol{x}^*\|^2) \leq \frac{L}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2.$$

Since $f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^*) \geq 0$ and is nonincreasing, we have

$$\sum_{k=0}^{\ell-1} (f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^*)) \geq \ell(f(\boldsymbol{x}^\ell) - f(\boldsymbol{x}^*)).$$

Therefore,

$$\ell(f(\boldsymbol{x}^\ell) - f(\boldsymbol{x}^*)) \leq \frac{L}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2,$$

which gives

$$f(\boldsymbol{x}^\ell) - f(\boldsymbol{x}^*) \leq \frac{L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{2\ell}.$$

Renaming ℓ as k yields the stated result.

Lec 7

7.1 Descent Methods

7.1.1 Analysis of Gradient Descent for L-smooth Convex Case

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L-smooth, and suppose there exists $\boldsymbol{x}^*$ such that

$$\nabla f(\boldsymbol{x}^*) = \mathbf{0}$$

(so $\boldsymbol{x}^*$ is a global minimizer). If gradient descent is initiated at $\boldsymbol{x}^0$ with step sizes

$$\alpha_k \equiv \frac{1}{L},$$

then

$$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*) \leq \frac{L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{2k}, \quad \forall k = 1, 2, \ldots.$$

Aside:

$$f(\boldsymbol{x}^l) \leq f(\boldsymbol{x}^{l-1}) \leq \cdots \leq f(\boldsymbol{x}^0).$$

Any method with this monotone decrease property is called a **descent method**.

The proof was given in the previous lecture.

7.1.2 Analysis of Gradient Descent for L-smooth Nonconvex Case

For minimizing nonconvex functions, global minimizers are usually hard to find, even for $n = 1$. In higher dimensions, even local minimizers can be difficult.

So we often settle for **stationary points**, i.e. seek $\boldsymbol{x}^*$ such that

$$\nabla f(\boldsymbol{x}^*) = \mathbf{0},$$

(although even this may be impossible if f has no stationary point).

Example

$$f(x) = \sqrt{x^2 + 1} + x$$

is smooth, convex, and L-smooth with $L = 1$, but it has no stationary point.
If gradient descent is applied to this f, the sequence x^k tends to $-\infty$ and $f'(x^k) \to 0$.

Theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth (not necessarily convex) and bounded below by $f_{\min}$. Apply gradient descent with step size $1/L$ starting at $\boldsymbol{x}^0$. Then:

(i)

$$\min_{0\le j\le k-1} \|\nabla f(\boldsymbol{x}^j)\| \;\le\; \sqrt{\frac{2L(f(\boldsymbol{x}^0)-f_{\min})}{k}},$$

(ii)

$$\nabla f(\boldsymbol{x}^j)\to \mathbf{0} \quad \text{as } j\to\infty.$$

Proof

Recall from L-smoothness and the *GD* update with step size $1/L$:

$$f(\boldsymbol{x}^{j+1}) \le f(\boldsymbol{x}^j) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}^j)\|^2.$$

Summing from $j=0$ to $k-1$,

$$f(\boldsymbol{x}^k) \le f(\boldsymbol{x}^0) - \frac{1}{2L}\sum_{j=0}^{k-1}\|\nabla f(\boldsymbol{x}^j)\|^2.$$

Since $f(\boldsymbol{x}^k)\ge f_{\min}$, we obtain

$$f_{\min} - f(\boldsymbol{x}^0) \le -\frac{1}{2L}\sum_{j=0}^{k-1}\|\nabla f(\boldsymbol{x}^j)\|^2,$$

or equivalently

$$\frac{1}{2L}\sum_{j=0}^{k-1}\|\nabla f(\boldsymbol{x}^j)\|^2 \le f(\boldsymbol{x}^0) - f_{\min}.$$

Taking $k\to\infty$ shows

$$\sum_{j=0}^{\infty}\|\nabla f(\boldsymbol{x}^j)\|^2 < \infty,$$

so $\|\nabla f(\boldsymbol{x}^j)\|\to 0$ as $j\to\infty$, proving (2).
For (1), note

$$\sum_{j=0}^{k-1}\|\nabla f(\boldsymbol{x}^j)\|^2 \;\ge\; k\cdot \min_{0\le j\le k-1}\|\nabla f(\boldsymbol{x}^j)\|^2.$$

Thus

$$\frac{k}{2L}\min_{0\le j\le k-1}\|\nabla f(\boldsymbol{x}^j)\|^2 \le f(\boldsymbol{x}^0) - f_{\min},$$

which implies

$$\min_{0\le j\le k-1}\|\nabla f(\boldsymbol{x}^j)\| \le \sqrt{\frac{2L(f(\boldsymbol{x}^0)-f_{\min})}{k}}.$$

7.2 Strongly Convex Functions

7.2.1 Definitions

Say $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is m-**strongly convex** with modulus $m \geq 0$ if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and all $\lambda \in [0, 1]$,

$$f((1-\lambda)\boldsymbol{x} + \lambda \boldsymbol{y}) \leq (1-\lambda) f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}) - \frac{1}{2}\lambda(1-\lambda) m \|\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{1}$$

Notes:

- $m = 0$ is exactly the usual notion of convexity.
- "Strongly convex" (without specifying m) means the above holds for some $m > 0$.

7.2.2 Differentiable Strongly Convex Functions

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be differentiable on $\operatorname{dom}(f)$, and assume $\operatorname{dom}(f)$ is open and convex. Then f is strongly convex with modulus m if and only if

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{2}$$

Note that the last term strengthens the usual subgradient inequality.

Proof

Forward direction $\Rightarrow$.
By differentiability, for any $\boldsymbol{x}, \boldsymbol{y}$ and $\lambda \in [0, 1]$,

$$f((1-\lambda)\boldsymbol{x} + \lambda \boldsymbol{y}) = f(\boldsymbol{x}) + \lambda \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \phi_{\boldsymbol{x}}(\lambda(\boldsymbol{y} - \boldsymbol{x})), \tag{3}$$

where

$$\lim_{\boldsymbol{h} \to \boldsymbol{0},\ \boldsymbol{h} \neq \boldsymbol{0}} \frac{\phi_{\boldsymbol{x}}(\boldsymbol{h})}{\|\boldsymbol{h}\|} = 0.$$

Combine strong convexity (1) with (3):

$$\begin{aligned} &f(\boldsymbol{x}) + \lambda \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \phi_{\boldsymbol{x}}(\lambda(\boldsymbol{y} - \boldsymbol{x})) \\ &\leq (1-\lambda) f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}) - \frac{1}{2} m \lambda (1-\lambda) \|\boldsymbol{y} - \boldsymbol{x}\|^2. \end{aligned}$$

Subtract $f(\boldsymbol{x})$ and divide by $\lambda > 0$:

$$\nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{\phi_{\boldsymbol{x}}(\lambda(\boldsymbol{y} - \boldsymbol{x}))}{\lambda} \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \frac{1}{2} m (1-\lambda) \|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

Let $\lambda \downarrow 0$. The remainder term vanishes and $(1 - \lambda) \to 1$, yielding

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2,$$

which is (2).

Backward direction $\Leftarrow$.
Assume (2). Take $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f)$ and $\lambda \in [0,1]$. Let

$$\boldsymbol{z} := (1-\lambda)\boldsymbol{x} + \lambda \boldsymbol{y}.$$

Apply (2) twice, once with $(\boldsymbol{z}, \boldsymbol{x})$ and once with $(\boldsymbol{z}, \boldsymbol{y})$:

$$\text{a)} \quad f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + \nabla f(\boldsymbol{z})^\top (\boldsymbol{x} - \boldsymbol{z}) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2,$$

$$\text{b)} \quad f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \nabla f(\boldsymbol{z})^\top (\boldsymbol{y} - \boldsymbol{z}) + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{z}\|^2.$$

Multiply (a) by $(1-\lambda)$ and (b) by λ, then add:

$$\begin{aligned}
&(1-\lambda) f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}) \\
&\geq (1-\lambda)\Big[f(\boldsymbol{z}) + \nabla f(\boldsymbol{z})^\top (\boldsymbol{x} - \boldsymbol{z})\Big] + \lambda\Big[f(\boldsymbol{z}) + \nabla f(\boldsymbol{z})^\top (\boldsymbol{y} - \boldsymbol{z})\Big] \\
&\quad + \frac{m}{2}\Big[(1-\lambda)\|\boldsymbol{x} - \boldsymbol{z}\|^2 + \lambda \|\boldsymbol{y} - \boldsymbol{z}\|^2\Big].
\end{aligned}$$

Note that $(1-\lambda)(\boldsymbol{x} - \boldsymbol{z}) + \lambda(\boldsymbol{y} - \boldsymbol{z}) = \boldsymbol{0}$, and

$$(1-\lambda)\|\boldsymbol{x} - \boldsymbol{z}\|^2 + \lambda\|\boldsymbol{y} - \boldsymbol{z}\|^2 = \lambda(1-\lambda)\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

Hence

$$(1-\lambda) f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \frac{m}{2}\lambda(1-\lambda)\|\boldsymbol{x} - \boldsymbol{y}\|^2,$$

which is exactly the strong convexity inequality (1).

7.2.3 Second Derivatives and Convexity

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be $\mathcal{C}^2$ on $\mathrm{dom}(f)$, and assume $\mathrm{dom}(f)$ is open and convex. Then f is strongly convex with modulus m on $\mathrm{dom}(f)$ if and only if

$$\nabla^2 f(\boldsymbol{x}) \succeq m\boldsymbol{I}, \quad \forall \boldsymbol{x} \in \mathrm{dom}(f).$$

In particular,

$$f \text{ convex} \iff \nabla^2 f(\boldsymbol{x}) \text{ is positive semidefinite for all } \boldsymbol{x}.$$

Proof

Fix $\boldsymbol{x} \in \mathrm{dom}(f)$ and $\boldsymbol{u} \in \mathbb{R}^n$. Choose $\alpha \geq 0$ small enough that $\boldsymbol{x} + \alpha\boldsymbol{u} \in \mathrm{dom}(f)$. By the second-order Taylor expansion, there exists $\gamma \in [0,1]$ such that

$$f(\boldsymbol{x} + \alpha\boldsymbol{u}) = f(\boldsymbol{x}) + \alpha\, \nabla f(\boldsymbol{x})^\top \boldsymbol{u} + \frac{1}{2}\alpha^2\, \boldsymbol{u}^\top \nabla^2 f(\boldsymbol{x} + \gamma\alpha\boldsymbol{u})\, \boldsymbol{u}. \tag{4}$$

Forward direction $\Rightarrow$. Assume f is strongly convex with modulus m. Then by the characterization (2),

$$f(\boldsymbol{x}+\alpha\boldsymbol{u}) \geq f(\boldsymbol{x}) + \alpha\,\nabla f(\boldsymbol{x})^\top \boldsymbol{u} + \frac{m}{2}\alpha^2\|\boldsymbol{u}\|^2.$$

Subtract (4) and cancel the linear terms:

$$\frac{1}{2}\alpha^2\,\boldsymbol{u}^\top \nabla^2 f(\boldsymbol{x}+\gamma\alpha\boldsymbol{u})\,\boldsymbol{u} \geq \frac{m}{2}\alpha^2\|\boldsymbol{u}\|^2.$$

Dividing by $\frac{1}{2}\alpha^2$,

$$\boldsymbol{u}^\top \nabla^2 f(\boldsymbol{x}+\gamma\alpha\boldsymbol{u})\,\boldsymbol{u} \geq m\|\boldsymbol{u}\|^2.$$

Let $\alpha \downarrow 0$. By continuity of the Hessian,

$$\boldsymbol{u}^\top \nabla^2 f(\boldsymbol{x})\,\boldsymbol{u} \geq m\|\boldsymbol{u}\|^2, \quad \forall \boldsymbol{u},$$

i.e. $\nabla^2 f(\boldsymbol{x}) - m\boldsymbol{I} \succeq 0$ for all $\boldsymbol{x}$.

Backward direction $\Leftarrow$. Assume $\nabla^2 f(\boldsymbol{x}) \succeq m\boldsymbol{I}$ for all $\boldsymbol{x}$. Then for any $\boldsymbol{x}, \boldsymbol{z} \in \mathrm{dom}(f)$, apply (4) with $\alpha\boldsymbol{u} := \boldsymbol{z} - \boldsymbol{x}$:

$$f(\boldsymbol{z}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{z}-\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{x})^\top \nabla^2 f(\boldsymbol{x}+\gamma(\boldsymbol{z}-\boldsymbol{x}))(\boldsymbol{z}-\boldsymbol{x}).$$

Using $\nabla^2 f(\cdot) \succeq m\boldsymbol{I}$,

$$\frac{1}{2}(\boldsymbol{z}-\boldsymbol{x})^\top \nabla^2 f(\boldsymbol{x}+\gamma(\boldsymbol{z}-\boldsymbol{x}))(\boldsymbol{z}-\boldsymbol{x}) \geq \frac{m}{2}\|\boldsymbol{z}-\boldsymbol{x}\|^2.$$

Hence

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{z}-\boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{z}-\boldsymbol{x}\|^2,$$

which is the strong convexity condition (2). Thus f is strongly convex with modulus m.

Example

- Quadratic:

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^\top \boldsymbol{x} + d, \quad \nabla f(\boldsymbol{x}) = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}, \quad \nabla^2 f(\boldsymbol{x}) = \boldsymbol{H}.$$

  Then

$$f \text{ convex} \iff \boldsymbol{H} \succeq \boldsymbol{0}, \quad f \text{ strongly convex (modulus } m) \iff \boldsymbol{H} \succeq m\boldsymbol{I}.$$

- Exponential:

$$f(x) = e^x, \quad f''(x) = e^x > 0 \;\Rightarrow\; f \text{ convex}.$$

- Reciprocal:

$$f(x) = \begin{cases} \dfrac{1}{x}, & x > 0, \\ \infty, & x \leq 0, \end{cases} \qquad f''(x) = \frac{2}{x^3} > 0 \;\Rightarrow\; f \text{ convex}.$$

- Negative log:

$$f(x) = \begin{cases} -\ln x, & x > 0, \\ \infty, & x \le 0, \end{cases} \quad f''(x) = \frac{1}{x^2} > 0 \Rightarrow f \text{ convex.}$$

- Non-differentiable example: $f(x) = \|x\|_\square$ is convex, but the second-derivative test does not apply directly.

7.2.4 Norm of Matrices

Given $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, the operator 2-norm is defined as

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})}.$$

Fact

- $\|\boldsymbol{A}\|_2 \le \|\boldsymbol{A}\|_F \le \sqrt{\min(m,n)}\, \|\boldsymbol{A}\|_2$, where

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A(i,j)^2}$$

  is the Frobenius norm.

- For a $\mathcal{C}^2$ function f, f is L-smooth if and only if

$$\|\nabla^2 f(\boldsymbol{x})\|_2 \le L \quad \forall \boldsymbol{x}.$$

- If $\boldsymbol{A} \in \mathbb{S}^n$ (symmetric), then

$$\|\boldsymbol{A}\|_2 = \max\{|\lambda_{\min}(\boldsymbol{A})|, |\lambda_{\max}(\boldsymbol{A})|\}.$$

Lec 8

8.1 Linear Algebra Facts

1. For a matrix $\boldsymbol{A}$, define the operator 2-norm:

$$\|\boldsymbol{A}\|_2 := \left(\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})\right)^{\frac{1}{2}}.$$

For $\boldsymbol{A} \in \mathbb{S}^n$,

$$\|\boldsymbol{A}\|_2 = \max\left(|\lambda_{\min}(\boldsymbol{A})|, |\lambda_{\max}(\boldsymbol{A})|\right).$$

For $\boldsymbol{A} \in \mathbb{S}^n_+$,

$$\|\boldsymbol{A}\|_2 = \lambda_{\max}(\boldsymbol{A}).$$

2. Recall: eigenvalues of $\boldsymbol{A} - \lambda \boldsymbol{I}$ are the eigenvalues of $\boldsymbol{A}$ shifted by $-\lambda$.

For $\boldsymbol{A} \in \mathbb{S}^n$,

$$\begin{aligned} \boldsymbol{A} \succeq t\boldsymbol{I} &\Leftrightarrow \lambda_i(\boldsymbol{A}) \geq t, \quad \forall i = 1, \ldots, n, \\ \boldsymbol{A} \preceq t\boldsymbol{I} &\Leftrightarrow \lambda_i(\boldsymbol{A}) \leq t, \quad \forall i = 1, \ldots, n. \end{aligned}$$

8.2 m-strongly Convex Functions

8.2.1 Two Lemmas

Given $f : \mathbb{R}^n \to \mathbb{R}$, with $f \in \mathcal{C}^2$,

$$\begin{aligned} f \text{ is } L\text{-smooth} &\Leftrightarrow -L\boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}, && \forall \boldsymbol{x} \in \mathbb{R}^n, \\ f \text{ is convex} &\Leftrightarrow \nabla^2 f(\boldsymbol{x}) \succeq \boldsymbol{0}, && \forall \boldsymbol{x} \in \mathbb{R}^n, \\ f \text{ is } m\text{-strongly convex} &\Leftrightarrow \nabla^2 f(\boldsymbol{x}) \succeq m\boldsymbol{I}, && \forall \boldsymbol{x} \in \mathbb{R}^n, \\ f \text{ is } L\text{-smooth, } m\text{-strongly convex} &\Leftrightarrow m\boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}, && \forall \boldsymbol{x} \in \mathbb{R}^n. \end{aligned}$$

Example

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^\top \boldsymbol{x} + d, \qquad \boldsymbol{H} \in \mathbb{S}^n.$$

Then

$$\begin{aligned} f \text{ is } L\text{-smooth} &\Leftrightarrow -L\boldsymbol{I} \preceq \boldsymbol{H} \preceq L\boldsymbol{I}, \\ f \text{ is convex} &\Leftrightarrow \boldsymbol{H} \succeq \boldsymbol{0}, \\ f \text{ is } m\text{-strongly convex} &\Leftrightarrow \boldsymbol{H} \succeq m\boldsymbol{I}. \end{aligned}$$

Lemma

$f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is m-strongly convex $\Leftrightarrow$ $g(\boldsymbol{x}) := f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|^2$ is convex.

Proof

Algebra shows $\forall \lambda \in \mathbb{R},\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$\begin{aligned} &(1-\lambda)g(\boldsymbol{x}) + \lambda g(\boldsymbol{y}) - g((1-\lambda)\boldsymbol{x} + \lambda \boldsymbol{y}) \\ &= (1-\lambda)f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}) - f((1-\lambda)\boldsymbol{x} + \lambda \boldsymbol{y}) - \frac{m}{2}\lambda(1-\lambda)\|\boldsymbol{x}-\boldsymbol{y}\|^2. \end{aligned}$$

Lemma

If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex, then $\exists\, \boldsymbol{a} \in \mathbb{R}^n,\ b \in \mathbb{R}$ such that

$$f(\boldsymbol{x}) \geq \underbrace{\boldsymbol{a}^\top \boldsymbol{x} + b}_{\text{Affine Underestimator}}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Proof

If f is differentiable, this is an immediate consequence of the subgradient inequality.

8.2.2 Two Theorems

Say $f : \mathbb{R}^n \to \mathbb{R}$ is *coercive* if

$$\forall s \in \mathbb{R},\ \exists r \in \mathbb{R} \text{ such that } \|\boldsymbol{x}\| \geq r \Rightarrow f(\boldsymbol{x}) \geq s.$$

Intuitively, $f(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$.

Facts

- If $f, g : \mathbb{R}^n \to \mathbb{R}$, $f \geq g$, and g is coercive, then f is coercive.
- $q(t) = \frac{1}{2}at^2 + bt + c : \mathbb{R} \to \mathbb{R}$ is coercive whenever $a > 0$.

Theorem

If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and coercive, then f has a minimizer.

Proof

Choose r such that $f(\boldsymbol{x}) \geq f(\boldsymbol{0})$ whenever $\|\boldsymbol{x}\| \geq r$ (by coercivity). Observe that f restricted to $\overline{\mathrm{B}}(\boldsymbol{0}, r)$ has a minimizer by compactness. Call it $\boldsymbol{x}^*$. Then $\boldsymbol{x}^*$ is a minimizer of f over $\mathbb{R}^n$ since for $\boldsymbol{x} \notin \overline{\mathrm{B}}(\boldsymbol{0}, r)$,

$$f(\boldsymbol{x}) \geq f(\boldsymbol{0}) \geq f(\boldsymbol{x}^*).$$

Theorem

If $f : \mathbb{R}^n \to \mathbb{R}$ is m-strongly convex for $m > 0$, then f has a unique minimizer.

Proof

Uniqueness. Suppose $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ are both minimizers. Then

$$f\left(\frac{\boldsymbol{x}_1+\boldsymbol{x}_2}{2}\right) \leq \frac{1}{2}(f(\boldsymbol{x}_1)+f(\boldsymbol{x}_2)) - \underbrace{\frac{1}{8}m\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^2}_{>0},$$

contradicting minimality of $\boldsymbol{x}_1, \boldsymbol{x}_2$.

Existence. Let $g(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|^2$, which is convex by Lemma 1. Let $\boldsymbol{a}^\top \boldsymbol{x} + b$ be an affine underestimate of g by Lemma 2. Define $h(\boldsymbol{x}) := \boldsymbol{a}^\top \boldsymbol{x} + b + \frac{m}{2}\|\boldsymbol{x}\|^2$, an underestimate of $f(\boldsymbol{x})$. We will show the function h is coercive, which implies f is coercive, so f has a minimizer. Rewrite h by completing the square:

$$\begin{aligned} h(\boldsymbol{x}) &= \frac{m}{2}\left\|\boldsymbol{x}+\frac{\boldsymbol{a}}{m}\right\|^2 - \frac{1}{2m}\boldsymbol{a}^\top\boldsymbol{a} + b, \\ h(\boldsymbol{x}) - s &= \frac{m}{2}\left\|\boldsymbol{x}+\frac{\boldsymbol{a}}{m}\right\|^2 - \frac{1}{2m}\boldsymbol{a}^\top\boldsymbol{a} + b - s, \quad \forall s \in \mathbb{R} \\ &\geq \frac{m}{2}\left(\|\boldsymbol{x}\| - \frac{\|\boldsymbol{a}\|}{m}\right)^2 - \frac{1}{2m}\boldsymbol{a}^\top\boldsymbol{a} + b - s. \end{aligned}$$

This is a univariate quadratic in $\|\boldsymbol{x}\|$ with positive leading coefficient, hence coercive. Thus

$$\exists r \text{ such that } \|\boldsymbol{x}\| \geq r \quad (r>0) \quad \Rightarrow \quad h(\boldsymbol{x}) \geq s.$$

8.3 Analysis of *GD* for m-strongly Convex Case

Theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is m-strongly convex, $\boldsymbol{x}^* \in \text{intdom}(f)$, f is differentiable at $\boldsymbol{x}^*$, and $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$ (so $\boldsymbol{x}^*$ is a minimizer; we cannot deduce the converse since f may not be differentiable at $\boldsymbol{x}^*$). Then for all $\boldsymbol{x} \in \text{intdom}(f)$ where $\nabla f(\boldsymbol{x})$ exists,

$$\text{a)} \quad f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \leq \frac{\|\nabla f(\boldsymbol{x})\|^2}{2m},$$
$$\text{b)} \quad \|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \frac{2}{m}\|\nabla f(\boldsymbol{x})\|.$$

Proof

For all $\boldsymbol{y} \in \mathbb{R}^n$,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2.$$

Both sides are functions of $\boldsymbol{y}$ (with $\boldsymbol{x}$ fixed). In general, if $\phi, \psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ satisfy $\phi(\boldsymbol{y}) \geq \psi(\boldsymbol{y})$ for all $\boldsymbol{y}$, then

$$\inf_{\boldsymbol{y}} \phi(\boldsymbol{y}) \geq \inf_{\boldsymbol{y}} \psi(\boldsymbol{y}).$$

Apply this to the preceding inequality:

$$f(\boldsymbol{x}^*) \geq \min_{\boldsymbol{y}} \left[f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{m}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \right].$$

The RHS is quadratic in $\boldsymbol{y}$, with quadratic coefficient $\frac{m}{2}\boldsymbol{I}$ and linear coefficient $\nabla f(\boldsymbol{x}) - m\boldsymbol{x}$, so its minimizer $\boldsymbol{y}^*$ is

$$\boldsymbol{y}^* = -\frac{\nabla f(\boldsymbol{x}) - m\boldsymbol{x}}{m}.$$

(We regard $\boldsymbol{a}^\top \boldsymbol{x}$ as having coefficient $\boldsymbol{a}$, not $\boldsymbol{a}^\top$.)
The minimal value is

$$\min_{\boldsymbol{y}} \text{RHS} = f(\boldsymbol{x}) - \frac{1}{2m} \|\nabla f(\boldsymbol{x})\|^2,$$

which rearranges to yield (a).
For (b),

$$\begin{aligned} f(\boldsymbol{x}^*) &\geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{x}^* - \boldsymbol{x}) + \frac{m}{2} \|\boldsymbol{x}^* - \boldsymbol{x}\|^2 \\ &\geq f(\boldsymbol{x}) - \|\nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{x}^* - \boldsymbol{x}\| + \frac{m}{2} \|\boldsymbol{x}^* - \boldsymbol{x}\|^2 \quad \text{(Cauchy–Schwarz)} \\ 0 \leq f(\boldsymbol{x}) - f(\boldsymbol{x}^*) &\leq \|\nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{x}^* - \boldsymbol{x}\| - \frac{m}{2} \|\boldsymbol{x}^* - \boldsymbol{x}\|^2 \\ \frac{m}{2} \|\boldsymbol{x}^* - \boldsymbol{x}\|^2 &\leq \|\nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{x}^* - \boldsymbol{x}\|. \end{aligned}$$

Divide by $\|\boldsymbol{x}^* - \boldsymbol{x}\|$ to obtain (b).

Theorem

Gradient Descent (GD) with stepsizes $\alpha_k = \frac{1}{L}$ applied to $f : \mathbb{R}^n \to \mathbb{R}$ differentiable and m-strongly convex satisfies

$$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*) \leq \left(1 - \frac{m}{L}\right)^k \cdot (f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)), \quad \forall k = 1, 2, \ldots$$

Proof

$$\begin{aligned} f(\boldsymbol{x}^{k+1}) &\leq f(\boldsymbol{x}^k) - \frac{1}{2L} \|\nabla f(\boldsymbol{x}^k)\|^2 \\ &\leq f(\boldsymbol{x}^k) - \frac{m}{L} (f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)). \end{aligned}$$

Subtract $f(\boldsymbol{x}^*)$ from both sides and rearrange:

$$f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^*) \leq \left(1 - \frac{m}{L}\right) (f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)).$$

Induction on k proves the theorem:

$$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*) \leq \left(1 - \frac{m}{L}\right)^k (f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)).$$

This is exponential in k, called *Linear Convergence*. Compared to the convex (not strongly convex) case:

$$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*) \leq \frac{L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{2k},$$

which is called *Sublinear Convergence*, worse than for m-strongly convex functions.

8.4 Accelerated Gradient Descent and Momentum

GD is a *first-order* method, i.e., each iteration requires first derivatives of the objective.

Nemirovsky and Yudin (1982) found a first-order method such that

$$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*) \leq \mathcal{O}\left(\frac{1}{k^2}\right).$$

Furthermore, under some assumptions, $\mathcal{O}\left(\frac{1}{k^2}\right)$ is the best possible.

The *N–Y algorithm* is difficult to implement. Nesterov (1983) proposed *Accelerated Gradient Descent (AGD)* which attains $\mathcal{O}\left(\frac{1}{k^2}\right)$ (*Nesterov's Fast Gradient Method*).

Lec 9

9.1 Algorithm

An L-smooth convex function can be minimized with accuracy

$$\mathcal{O}\left(\frac{1}{k^2}\right)$$

after k iterations.

Choose an initial point $\boldsymbol{y}^0 \in \mathbb{R}^n$ arbitrarily. For $k = 0, 1, 2, \ldots$, update as follows:

$$\begin{aligned}\boldsymbol{x}^{k+1} &= \boldsymbol{y}^k - \alpha_k \nabla f(\boldsymbol{y}^k),\\ \boldsymbol{y}^{k+1} &= \boldsymbol{x}^{k+1} + \beta_{k+1}(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k).\end{aligned}$$

Here we fix $\alpha_k \equiv \dfrac{1}{L}$, while $\{\beta_k\}$ is a sequence to be determined.

9.2 Convergence Analysis

Assume f is L-smooth and convex, and that f has a minimizer $\boldsymbol{x}^*$. Denote $f_{\min} := f(\boldsymbol{x}^*)$. Define

$$v_k := f(\boldsymbol{x}^k) - f_{\min} + \frac{L}{2}\left\|\boldsymbol{x}^k - \boldsymbol{x}^* - \rho_{k-1}^2(\boldsymbol{x}^{k-1} - \boldsymbol{x}^*)\right\|^2.$$

The sequence $\{\rho_k\}$ is to be determined. We take $\boldsymbol{x}^0 := \boldsymbol{y}^0$.

Observe that $v_k \geq 0$; moreover, if $v_k = 0$ then $\boldsymbol{x}^k$ is optimal.

From the definition of v_k,

$$v_{k+1} = f(\boldsymbol{x}^{k+1}) - f_{\min} + \frac{L}{2}\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^* - \rho_k^2(\boldsymbol{x}^k - \boldsymbol{x}^*)\right\|^2.$$

Fact

Recall. For gradient descent with step size $1/L$ applied at $\boldsymbol{y}^k$,

$$f(\boldsymbol{x}^{k+1}) \leq f(\boldsymbol{y}^k) - \frac{1}{2L}\|\nabla f(\boldsymbol{y}^k)\|^2.$$

Reason.

$$\begin{aligned}f(\boldsymbol{x}^{k+1}) &\leq f(\boldsymbol{y}^k) + \nabla f(\boldsymbol{y}^k)^\top(\boldsymbol{x}^{k+1} - \boldsymbol{y}^k) + \frac{L}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{y}^k\|^2\\ &= f(\boldsymbol{y}^k) - \frac{1}{L}\|\nabla f(\boldsymbol{y}^k)\|^2 + \frac{L}{2}\left\|\frac{1}{L}\nabla f(\boldsymbol{y}^k)\right\|^2\\ &= f(\boldsymbol{y}^k) - \frac{1}{2L}\|\nabla f(\boldsymbol{y}^k)\|^2.\end{aligned}$$

The orange inequality is exactly the L-smoothness condition, which is equivalent to the gradient Lipschitz property

$$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \leq L\|\boldsymbol{y} - \boldsymbol{x}\|.$$

Intuitively, this restricts the second-order growth of f. If f is twice differentiable, it corresponds to bounding the Hessian $\nabla^2 f$ in operator norm, and the inequality can be seen as a Taylor expansion with a quadratic remainder controlled by L.

Substituting this into v_{k+1}, we obtain

$$\begin{aligned} v_{k+1} &\leq f(\boldsymbol{y}^k) - f_{\min} - \underbrace{\frac{1}{2L}\|\nabla f(\boldsymbol{y}^k)\|^2}_{\mathcal{A}} + \underbrace{\frac{L}{2}\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^* - \rho_k^2(\boldsymbol{x}^k - \boldsymbol{x}^*)\right\|^2}_{\mathcal{B}} \\ &= f(\boldsymbol{y}^k) - f_{\min} - \mathcal{A} + \mathcal{B}. \end{aligned}$$

We use an uncommon splitting trick:

$$f(\boldsymbol{y}^k) - f_{\min} = \rho_k^2(f(\boldsymbol{y}^k) - f_{\min}) + (1 - \rho_k^2)(f(\boldsymbol{y}^k) - f_{\min}).$$

Substituting back,

$$v_{k+1} \leq \rho_k^2(f(\boldsymbol{y}^k) - f_{\min}) + (1 - \rho_k^2)(f(\boldsymbol{y}^k) - f_{\min}) - \mathcal{A} + \mathcal{B}.$$

Fact

Recall (subgradient inequality). For convex differentiable f,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y} - \boldsymbol{x}).$$

Reason. By convexity, from a geometric point of view the graph of f lies above all its tangents.

Assume $0 \leq \rho_k \leq 1$ for all k. By the subgradient inequality, we have

$$\begin{aligned} f(\boldsymbol{x}^k) &\geq f(\boldsymbol{y}^k) + \nabla f(\boldsymbol{y}^k)^\top(\boldsymbol{x}^k - \boldsymbol{y}^k), \\ f_{\min} = f(\boldsymbol{x}^*) &\geq f(\boldsymbol{y}^k) + \nabla f(\boldsymbol{y}^k)^\top(\boldsymbol{x}^* - \boldsymbol{y}^k). \end{aligned}$$

Plugging these inequalities into the previous relation yields

$$\begin{aligned}
v_{k+1} &\le \rho_k^2\Big(f(\boldsymbol{x}^k) - f_{\min} - \nabla f(\boldsymbol{y}^k)^\top(\boldsymbol{x}^k - \boldsymbol{y}^k)\Big) + (1-\rho_k^2)(\nabla f(\boldsymbol{y}^k)^\top(\boldsymbol{y}^k - \boldsymbol{x}^*)) - \mathcal{A} + \mathcal{B} \\
&= \rho_k^2\Big(\underbrace{f(\boldsymbol{x}^k) - f_{\min} + \frac{L}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^* - \rho_{k-1}^2(\boldsymbol{x}^{k-1} - \boldsymbol{x}^*)\|^2}_{v_k}\Big) - \rho_k^2\nabla f(\boldsymbol{y}^k)^\top(\boldsymbol{x}^k - \boldsymbol{y}^k) \\
&\qquad + (1-\rho_k^2)(\nabla f(\boldsymbol{y}^k)^\top(\boldsymbol{y}^k - \boldsymbol{x}^*)) - \rho_k^2 \cdot \frac{L}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^* - \rho_{k-1}^2(\boldsymbol{x}^{k-1} - \boldsymbol{x}^*)\|^2 - \mathcal{A} + \mathcal{B} \\
&= \rho_k^2 v_k - \rho_k^2\nabla f(\boldsymbol{y}^k)^\top(\boldsymbol{x}^k - \boldsymbol{y}^k) + (1-\rho_k^2)(\nabla f(\boldsymbol{y}^k)^\top(\boldsymbol{y}^k - \boldsymbol{x}^*)) \\
&\qquad - \rho_k^2 \cdot \frac{L}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^* - \rho_{k-1}^2(\boldsymbol{x}^{k-1} - \boldsymbol{x}^*)\|^2 - \mathcal{A} + \mathcal{B} \\
&= \rho_k^2 v_k + \underbrace{\nabla f(\boldsymbol{y}^k)^\top(\boldsymbol{y}^k - \rho_k^2\boldsymbol{x}^k - (1-\rho_k^2)\boldsymbol{x}^*)}_{\mathcal{C}} \\
&\qquad - \rho_k^2 \cdot \frac{L}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^* - \rho_{k-1}^2(\boldsymbol{x}^{k-1} - \boldsymbol{x}^*)\|^2 - \mathcal{A} + \mathcal{B} \\
&= \rho_k^2 v_k - \rho_k^2 \cdot \frac{L}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^* - \rho_{k-1}^2(\boldsymbol{x}^{k-1} - \boldsymbol{x}^*)\|^2 - \mathcal{A} + \mathcal{B} + \mathcal{C}.
\end{aligned}$$

Here

$$\mathcal{A} = \frac{1}{2L}\|\nabla f(\boldsymbol{y}^k)\|^2, \qquad \mathcal{B} = \frac{L}{2}\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^* - \rho_k^2(\boldsymbol{x}^k - \boldsymbol{x}^*)\right\|^2,$$
$$\mathcal{C} = \nabla f(\boldsymbol{y}^k)^\top(\boldsymbol{y}^k - \rho_k^2\boldsymbol{x}^k - (1-\rho_k^2)\boldsymbol{x}^*).$$

Using the update

$$\boldsymbol{x}^{k+1} := \boldsymbol{y}^k - \frac{1}{L}\nabla f(\boldsymbol{y}^k),$$

and the identity $\|\boldsymbol{a} - \boldsymbol{b}\|^2 = \|\boldsymbol{a}\|^2 - 2\boldsymbol{a}^\top\boldsymbol{b} + \|\boldsymbol{b}\|^2$, we have

$$\begin{aligned}
\underbrace{\frac{L}{2}\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^* - \rho_k^2(\boldsymbol{x}^k - \boldsymbol{x}^*)\right\|^2}_{\mathcal{B}} &= \frac{L}{2}\left\|\boldsymbol{y}^k - \frac{1}{L}\nabla f(\boldsymbol{y}^k) - \boldsymbol{x}^* - \rho_k^2(\boldsymbol{x}^k - \boldsymbol{x}^*)\right\|^2 \\
&= \frac{L}{2}\left\|\boldsymbol{y}^k - \boldsymbol{x}^* - \rho_k^2(\boldsymbol{x}^k - \boldsymbol{x}^*)\right\|^2 \\
&\quad - \underbrace{\nabla f(\boldsymbol{y}^k)^\top(\boldsymbol{y}^k - \boldsymbol{x}^* - \rho_k^2(\boldsymbol{x}^k - \boldsymbol{x}^*))}_{\mathcal{C}} + \underbrace{\frac{1}{2L}\|\nabla f(\boldsymbol{y}^k)\|^2}_{\mathcal{A}}.
\end{aligned}$$

Thus

$$-\mathcal{A} + \mathcal{B} + \mathcal{C} = \frac{L}{2}\left\|\boldsymbol{y}^k - \boldsymbol{x}^* - \rho_k^2(\boldsymbol{x}^k - \boldsymbol{x}^*)\right\|^2.$$

Therefore,

$$\begin{aligned} v_{k+1} &\le \rho_k^2 v_k + \frac{L}{2}\left\|\boldsymbol{y}^k - \boldsymbol{x}^* - \rho_k^2(\boldsymbol{x}^k - \boldsymbol{x}^*)\right\|^2 \\ &\quad - \rho_k^2 \cdot \frac{L}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^* - \rho_{k-1}^2(\boldsymbol{x}^{k-1} - \boldsymbol{x}^*)\|^2 \\ &= \rho_k^2 v_k + \frac{L}{2}(\underbrace{\|\boldsymbol{y}^k - \boldsymbol{x}^* - \rho_k^2(\boldsymbol{x}^k - \boldsymbol{x}^*)\|^2 - \rho_k^2\|\boldsymbol{x}^k - \boldsymbol{x}^* - \rho_{k-1}^2(\boldsymbol{x}^{k-1} - \boldsymbol{x}^*)\|^2}_{\mathcal{S}}). \end{aligned}$$

We force $\mathcal{S} = 0$ to specify ρ_k. Then $v_{k+1} \le \rho_k^2 v_k$, and

$$\boldsymbol{y}^k - \boldsymbol{x}^* - \rho_k^2(\boldsymbol{x}^k - \boldsymbol{x}^*) = \rho_k(\boldsymbol{x}^k - \boldsymbol{x}^* - \rho_{k-1}^2(\boldsymbol{x}^{k-1} - \boldsymbol{x}^*)).$$

Recalling $\boldsymbol{y}^k = (1+\beta_k)\boldsymbol{x}^k - \beta_k \boldsymbol{x}^{k-1}$, we obtain

$$(1+\beta_k)\boldsymbol{x}^k - \beta_k\boldsymbol{x}^{k-1} - \boldsymbol{x}^* - \rho_k^2(\boldsymbol{x}^k - \boldsymbol{x}^*) = \rho_k(\boldsymbol{x}^k - \boldsymbol{x}^*) - \rho_k\rho_{k-1}^2(\boldsymbol{x}^{k-1} - \boldsymbol{x}^*).$$

Rearranging,

$$(1+\beta_k - \rho_k^2)(\boldsymbol{x}^k - \boldsymbol{x}^*) - \beta_k(\boldsymbol{x}^{k-1} - \boldsymbol{x}^*) = \rho_k(\boldsymbol{x}^k - \boldsymbol{x}^*) - \rho_k\rho_{k-1}^2(\boldsymbol{x}^{k-1} - \boldsymbol{x}^*).$$

Comparing coefficients, we deduce

$$1 + \beta_k - \rho_k^2 = \rho_k, \qquad \beta_k = \rho_k\rho_{k-1}^2.$$

Substituting β_k,

$$1 + \rho_k\rho_{k-1}^2 - \rho_k^2 = \rho_k,$$

so ρ_k is a root of the quadratic

$$\rho_k^2 + (1 - \rho_{k-1}^2)\rho_k - 1 = 0.$$

Hence the two roots satisfy

$$\rho_{k_1} + \rho_{k_2} = -(1 - \rho_{k-1}^2), \qquad \rho_{k_1}\rho_{k_2} = -1.$$

Since $v_{k+1} \le \rho_k^2 v_k$, by induction we may assume $\rho_{k-1} \in [0,1]$. Thus the positive root lies in $[0,1]$, while the negative root is ≤ -1. Therefore, choosing ρ_k as the positive root gives $\rho_k \in [0,1]$.

An equivalent way to rewrite the recursion is

$$(1 - \rho_{k-1}^2)\rho_k = 1 - \rho_k^2,$$

that is,

$$\rho_k = \frac{1 - \rho_k^2}{1 - \rho_{k-1}^2}.$$

Since $v_{k+1} \le \rho_k^2 v_k$, by induction

$$v_k \le \rho_{k-1}^2\rho_{k-2}^2 \cdots \rho_1^2\, v_1.$$

A telescoping argument on the right-hand side yields

$$v_k \leq \frac{(1-\rho_{k-1}^2)^2}{(1-\rho_0^2)^2}\, v_1.$$

Setting $\rho_0 := 0$ gives

$$v_k \leq (1-\rho_{k-1}^2)^2 v_1.$$

Claim

$$1-\rho_k^2 \leq \frac{2}{k+2}.$$

Proof

We proceed by induction on k.
Base case ($k = 0$). The inequality holds trivially.
Induction step ($k \geq 1$). From the recurrence relation,

$$\rho_k = \frac{1-\rho_k^2}{1-\rho_{k-1}^2} \;\geq\; \frac{1-\rho_k^2}{\frac{2}{k+1}} = \frac{k+1}{2}(1-\rho_k^2).$$

Therefore

$$1-\rho_k^2 \leq \frac{2}{k+1}\,\rho_k.$$

Let $x := 1-\rho_k^2$. Then

$$x - \frac{2}{k+1}\sqrt{1-x} \leq 0,$$

and the left-hand side is an increasing function of x.
Test $x = \frac{2}{k+2}$:

$$\frac{2}{k+2} \;\overset{?}{\leq}\; \frac{2}{k+1}\sqrt{1-\frac{2}{k+2}}.$$

Squaring both sides and dividing by 4 gives

$$\frac{1}{(k+2)^2} \;\overset{?}{\leq}\; \frac{1}{(k+1)^2}\left(1-\frac{2}{k+2}\right),$$

equivalently,

$$(k+1)^2 \;\overset{?}{\leq}\; (k+2)k.$$

This inequality does not hold, so the condition implies

$$x < \frac{2}{k+2}.$$

Thus

$$1-\rho_k^2 < \frac{2}{k+2},$$

which completes the induction.

Since $v_k \le (1-\rho_{k-1}^2)^2 v_1$, the claim implies

$$v_k \le \frac{4}{(k+1)^2} v_1.$$

Because $v_k \ge f(\boldsymbol{x}^k) - f_{\min}$, we obtain

$$f(\boldsymbol{x}^k) - f_{\min} \le \frac{4}{(k+1)^2} v_1,$$

which is the $\mathcal{O}\left(\frac{1}{k^2}\right)$ convergence rate.

It remains to bound v_1 in terms of the initial point:

$$v_1 = f(\boldsymbol{x}^1) - f(\boldsymbol{x}^*) + \frac{L}{2}\|\boldsymbol{x}^1 - \boldsymbol{x}^*\|^2.$$

Note that $\boldsymbol{x}^1$ is obtained from $\boldsymbol{x}^0$ by a plain gradient descent step.

Fact

Recall. For gradient descent with step size $1/L$,

$$f(\boldsymbol{x}^{k+1}) \le f(\boldsymbol{x}^*) + \frac{L}{2}(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2).$$

Reason.

$$\begin{aligned} f(\boldsymbol{x}^k) &\le f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{x}^k - \boldsymbol{x}^*) \\ f(\boldsymbol{x}^{k+1}) &\le f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{x}^k - \boldsymbol{x}^*) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}^k)\|^2 \\ &\le f(\boldsymbol{x}^*) + \frac{L}{2}(2(\boldsymbol{x}^k - \boldsymbol{x}^*)^\top \frac{1}{L}\nabla f(\boldsymbol{x}^k) - \|\frac{1}{L}\nabla f(\boldsymbol{x}^k)\|^2) \\ &\le f(\boldsymbol{x}^*) + \frac{L}{2}(2(\boldsymbol{x}^k - \boldsymbol{x}^*)^\top (\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2). \end{aligned}$$

The purple inequality is in the form of the law of cosines

$$\|\boldsymbol{b} - \boldsymbol{c}\|^2 = \|\boldsymbol{b}\|^2 + \|\boldsymbol{c}\|^2 - 2\boldsymbol{b}^\top \boldsymbol{c}.$$

Rearranging yields

$$f(\boldsymbol{x}^{k+1}) \le f(\boldsymbol{x}^*) + \frac{L}{2}(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2).$$

Substituting $k = 0$ and rearranging, we obtain

$$v_1 \le \frac{L}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2.$$

Therefore,

$$f(\boldsymbol{x}^k) - f_{\min} \le \frac{2L}{(k+1)^2}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2,$$

which is the standard bound for Accelerated Gradient Descent *AGD*.

9.3 Implementation

A practical implementation of *AGD* is:

$$
\begin{aligned}
&\rho_0 = 0,\\
&\boldsymbol{y}^0 = \text{arbitrary},\\
&\text{for } k = 0, 1, 2, \dots \text{ do}\\
&\quad \boldsymbol{x}^{k+1} := \boldsymbol{y}^k - \frac{1}{L}\nabla f(\boldsymbol{y}^k),\\
&\quad \rho_{k+1} := \text{positive root of } \rho_{k+1}^2 + (1-\rho_k^2)\rho_{k+1} - 1 = 0,\\
&\quad \beta_{k+1} := \rho_{k+1}\rho_k^2,\\
&\quad \boldsymbol{y}^{k+1} := \boldsymbol{x}^{k+1} + \beta_{k+1}(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k).
\end{aligned}
$$

Lec 10

10.1 *AGD* in Strongly Convex Cases

AGD for L-smooth convex functions was discussed in the last lecture.

Here we consider *AGD* for L-smooth and m-strongly convex functions.

We use constant parameters

$$\alpha_k = \frac{1}{L}, \qquad \beta_k = \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}, \quad \forall k.$$

This implicitly assumes $L > m$, where L is the upper bound on the curvature and m is the lower bound.

Theorem

For an L-smooth and m-strongly convex function f with minimizer $\boldsymbol{x}^*$, accelerated gradient descent with

$$\alpha_k = \frac{1}{L}, \qquad \beta_k = \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}$$

satisfies

$$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*) \le \left(1-\sqrt{\frac{m}{L}}\right)^k \left[f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*) + \frac{m}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2\right].$$

In particular, the method has **linear convergence**.

This rate is faster than standard *GD*, whose linear rate is

$$\left(1-\frac{m}{L}\right)^k.$$

Hence, in the strongly convex case, *AGD* provides a significant improvement over *GD*.

When $\frac{m}{L} \ll 1$, the *convex* (non-strongly-convex) version of *AGD* may actually perform better for moderate k. A version of *AGD* that simultaneously attains both the strongly-convex and merely-convex bounds is well known and can be found in Nesterov's book.

Note. Conjugate Gradient is not covered in this lecture.

10.2 Aside on Programming

From Problem Set 2:

```
function huber_minimize(A, y, tol, x0, L, tau)

  while ...
    % ← Gradient Descent main loop
    evaluate huber gradient

end
```

We would like to separate the Huber-specific code from the general *GD* algorithm.

Tendency.

- The `gradfunc()` represents the gradient of the objective, and it internally depends on `A`, `y`, and `tau`.

Example

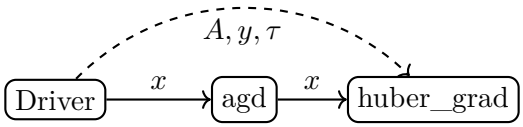
(Huber gradient wrapper.)

```
function g = huber_grad(x)
    % Evaluate Huber gradient
end
```

Then, we pass `huber_grad()` as the first argument of the optimizer:

```
function xopt = agd(gradfunc, x0, tol, L)
```

Now we have a call chain:



However, `huber_grad()` needs access to `A`, `y`, and `tau`, which are defined in `driver()` and not directly passed to it.

Old solutions.

(i) **Global variables.**
Store `A`, `y`, and `tau` in global variables known to both `driver()` and `huber_grad()`.
Risk: name clashes, hard to debug.

(ii) **Opaque parameter struct.**
Pass an additional argument `params` to `agd()`. The `driver()` packs `A`, `y`, and `tau` inside, and `huber_grad()` unpacks them.
Risk: troublesome packing and unpacking, error-prone.

Modern solutions.

(i) **Inner function.**

```
function driver
    A = ...;
    y = ...;
    tau = ...;

    function g = huber_grad(x)
        % uses A, y, tau
    end
```

```
    xopt = agd(@huber_grad, x0, tol, L);
end
```

(ii) **Anonymous function (closure / lambda).**

```
function g = huber_grad(x, A, y, tau)
    % Evaluate Huber gradient
end

function driver
    A = ...;
    y = ...;
    tau = ...;

    gradfunc = @(x) huber_grad(x, A, y, tau);
    xopt = agd(gradfunc, x0, tol, L);
end
```

Note. For Problem Set 3, use one of these two modern methods (inner or anonymous functions). Or, besides the two modern methods, can also use *objects* to represent functions in Problem Set 3.

10.3 Binary Classification

Given N data points $(\boldsymbol{a}_1, y_1), \ldots, (\boldsymbol{a}_N, y_N)$ with $y_i \in \{+1, -1\}$ and $\boldsymbol{a}_i \in \mathbb{R}^n$ for all i, we want to learn a vector $\boldsymbol{x}$ such that

$$\boldsymbol{a}_i^\top \boldsymbol{x} > 0 \text{ for } y_i = +1, \qquad \boldsymbol{a}_i^\top \boldsymbol{x} < 0 \text{ for } y_i = -1.$$

Such an $\boldsymbol{x}$ defines a **linear classifier**.

We can extend this idea to nonlinear classifiers by enriching the feature representation of each $\boldsymbol{a}_i$.

Example

(Feature mapping.) One approach is to use monomial features (as in Problem Set 1) to construct a higher-dimensional feature vector for each $\boldsymbol{a}_i$, on which a linear classifier is then learned.

To allow the classifier to represent a hyperplane not passing through the origin, we can append a constant 1 to each $\boldsymbol{a}_i$, which is known as the **padding trick**.

10.3.1 Logistic Regression

We say the data is **linearly separable** if there exists $\boldsymbol{x}$ such that $\boldsymbol{a}_i^\top \boldsymbol{x} > 0$ for $y_i = +1$ and $\boldsymbol{a}_i^\top \boldsymbol{x} < 0$ for $y_i = -1$.

Define

$$\varphi(t) := \begin{cases} 1, & t > 0, \\ -1, & t < 0. \end{cases}$$

Define

$$L := \prod_{i:y_i=1} \frac{\varphi(\boldsymbol{a}_i^\top \boldsymbol{x}) + 1}{2} \prod_{i:y_i=-1} \frac{-\varphi(\boldsymbol{a}_i^\top \boldsymbol{x}) + 1}{2}.$$

Then $L = 1$ if all points are correctly classified, and $L = 0$ otherwise.

One idea is to find the classifier by maximizing L as a function of $\boldsymbol{x}$. This fails if the data points are not linearly separable, since L then cannot achieve 1.

To fix this, redefine φ (since the original one is discontinuous):

$$\varphi(t) := \frac{-1 + \mathrm{e}^t}{1 + \mathrm{e}^t}.$$

Use the same formula for L, but with this new φ. (It is easier to work with $\ln L$.)

$$\max_{\boldsymbol{x}} \left[\sum_{i:y_i=1} \ln(\varphi(\boldsymbol{a}_i^\top \boldsymbol{x}) + 1) + \sum_{i:y_i=-1} \ln(-\varphi(\boldsymbol{a}_i^\top \boldsymbol{x}) + 1) \right].$$

This is called **Logistic Regression**.

Compute the logarithmic terms:

$$\ln(\varphi(t) + 1) = \ln\left(\frac{-1 + \mathrm{e}^t + 1 + \mathrm{e}^t}{1 + \mathrm{e}^t}\right) = \ln\left(\frac{2\mathrm{e}^t}{1 + \mathrm{e}^t}\right) = \ln\left(\frac{2}{1 + \mathrm{e}^{-t}}\right) = \ln 2 - \ln(1 + \mathrm{e}^{-t}),$$

$$\ln(-\varphi(t) + 1) = \ln 2 - \ln(1 + \mathrm{e}^t).$$

Dropping additive constants, logistic regression becomes

$$\min_{\boldsymbol{x}} \left[\sum_{i:y_i=1} \ln(1 + \mathrm{e}^{-\boldsymbol{a}_i^\top \boldsymbol{x}}) + \sum_{i:y_i=-1} \ln(1 + \mathrm{e}^{\boldsymbol{a}_i^\top \boldsymbol{x}}) \right].$$

Fact

The logistic regression objective

$$\boldsymbol{x} \mapsto \sum_{i:y_i=1} \ln(1 + \mathrm{e}^{-\boldsymbol{a}_i^\top \boldsymbol{x}}) + \sum_{i:y_i=-1} \ln(1 + \mathrm{e}^{\boldsymbol{a}_i^\top \boldsymbol{x}})$$

is convex and L-smooth (for some L depending on the data).

However, if the data is linearly separable, then the problem has *no* minimizer, since any separating $\boldsymbol{x}$ can be scaled arbitrarily large. Gradient descent then drives $\|\boldsymbol{x}\| \to \infty$.

To limit this, add a **regularizing term**:

$$\min_{\boldsymbol{x}} \frac{1}{N} \left[\sum_{i:y_i=1} \cdots + \sum_{i:y_i=-1} \cdots \right] + \frac{\gamma}{2} \|\boldsymbol{x}\|^2.$$

Here $\gamma > 0$ is the **regularization parameter**, and this is known as ℓ_2 regularization.

Later we will cover ℓ_1 regularization.

Now we can minimize this regularized objective using *GD* or *AGD*.

But what if $N = 10^9$? Gradient evaluation is very expensive, and not all data points are needed to obtain a good $\boldsymbol{x}$.

10.3.2 Stochastic Gradient Descent *SGD*

Rewrite the regularized objective as

$$\mathbb{E}[l(\boldsymbol{a}, y; \boldsymbol{x})] + \frac{\gamma}{2}\|\boldsymbol{x}\|^2,$$

where the expectation is over a random data pair $(\boldsymbol{a}, y)$ drawn uniformly from the dataset, and l is the logistic loss for a single sample.

In *SGD*, instead of using the full gradient of the expectation, we randomly select one data pair $(\boldsymbol{a}_i, y_i)$ and compute the stochastic gradient

$$\nabla_{\boldsymbol{x}} \left[l(\boldsymbol{a}_i, y_i; \boldsymbol{x}) + \frac{\gamma}{2}\|\boldsymbol{x}\|^2 \right].$$

At each iteration, a new random pair is chosen. This yields much cheaper iterations and scales to very large datasets.

Lec 11

11.1 Stochastic Gradient Descent

Given a dataset $(\boldsymbol{a}_i, y_i)$ for $i = 1, \ldots, N$, we seek $\boldsymbol{x}$ such that

$$\boldsymbol{a}_i^\top \boldsymbol{x} > 0 \quad \text{when } y_i = 1, \qquad \boldsymbol{a}_i^\top \boldsymbol{x} < 0 \quad \text{when } y_i = -1.$$

Consider the regularized empirical risk

$$F(\boldsymbol{x}) = \frac{1}{N} \sum_{i=1}^{N} \ell(\boldsymbol{a}_i, y_i, \boldsymbol{x}) + \frac{\gamma}{2} \|\boldsymbol{x}\|^2.$$

***SGD* algorithm (logistic regression setting).**

$\boldsymbol{x}^0$ arbitrary,
for $k = 0, 1, 2, \ldots$
  Choose $i_k \in \{1, \ldots, N\}$ uniformly at random,
  $\boldsymbol{x}^{k+1} := \boldsymbol{x}^k - \alpha_k \nabla_{\boldsymbol{x}} \Big(\ell(\boldsymbol{a}_{i_k}, y_{i_k}, \boldsymbol{x}) + \frac{\gamma}{2} \|\boldsymbol{x}\|^2 \Big)\big|_{\boldsymbol{x}=\boldsymbol{x}^k}$.

Finite-sum *SGD* viewpoint. Suppose

$$F(\boldsymbol{x}) = \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}; \boldsymbol{\xi}_i), \qquad \boldsymbol{\xi}_i := (\boldsymbol{a}_i, y_i).$$

Let

$$g(\boldsymbol{x}; \boldsymbol{\xi}_i) := \nabla_{\boldsymbol{x}} f(\boldsymbol{x}; \boldsymbol{\xi}_i).$$

Then

$$\nabla F(\boldsymbol{x}) = \frac{1}{N} \sum_{i=1}^{N} g(\boldsymbol{x}; \boldsymbol{\xi}_i).$$

***SGD* algorithm.**

$\boldsymbol{x}^0$ arbitrary,
for $k = 0, 1, 2, \ldots$
  Choose $i_k \in \{1, \ldots, N\}$ uniformly at random,
  $\boldsymbol{x}^{k+1} := \boldsymbol{x}^k - \alpha_k \, g(\boldsymbol{x}^k; \boldsymbol{\xi}_{i_k})$.

Fully general expectation form. In the most general setting,

$$F(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x}; \boldsymbol{\xi})],$$

where $\boldsymbol{\xi}$ is drawn from some distribution. On each iteration, we sample $\boldsymbol{\xi}$ from this distribution and define

$$g(\boldsymbol{x}; \boldsymbol{\xi}) := \nabla_{\boldsymbol{x}} f(\boldsymbol{x}; \boldsymbol{\xi}).$$

Then

$$\nabla F(\boldsymbol{x}) = \mathbb{E}[g(\boldsymbol{x}; \boldsymbol{\xi})].$$

Here $g(\boldsymbol{x}; \boldsymbol{\xi})$ is called a **stochastic gradient**.

11.1.1 Convergence Theorem for *SGD*

Theorem

(Bubeck, Theorem 6.3, special case.)
Let $F : \mathbb{R}^n \to \mathbb{R}$ be convex and L-smooth (with F the expectation or sum of the $f(\cdot;\boldsymbol{\xi})$). Assume:

- **Unbiasedness:**
$$\mathbb{E}[g(\boldsymbol{x};\boldsymbol{\xi})] = \nabla F(\boldsymbol{x}), \quad \forall \boldsymbol{x}.$$
- **Variance bound:**
$$\mathbb{E}[\|g(\boldsymbol{x};\boldsymbol{\xi}) - \nabla F(\boldsymbol{x})\|^2] \le \sigma^2, \quad \forall \boldsymbol{x}.$$

Let $\boldsymbol{x}^*$ be a minimizer of F, and let

$$R := \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|, \qquad \eta := \frac{R}{\sigma\sqrt{\ell}}.$$

Run *SGD* for $k = 0, 1, \ldots, \ell - 1$ with the constant step size

$$\alpha_k \equiv \frac{1}{L + \frac{1}{\eta}}.$$

Then

$$\mathbb{E}\left[F\left(\frac{1}{\ell}\sum_{k=1}^{\ell} \boldsymbol{x}^k\right)\right] - F(\boldsymbol{x}^*) \le \underbrace{\frac{R\sigma}{\sqrt{\ell}}}_{\text{stochastic part}} + \underbrace{\frac{LR^2}{2\ell}}_{\text{GD-like part (decays faster in } \ell)}, \qquad \forall\ \ell = 1, 2, \ldots.$$

Note. The step size α_k depends on the total number of iterations ℓ.

11.1.2 Jensen's Inequality

Fact

Let $F : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex. Suppose $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{R}^n$ and $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ satisfy

$$\lambda_i \geq 0, \quad \sum_{i=1}^{k} \lambda_i = 1.$$

Then

$$F\left(\sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i\right) \leq \sum_{i=1}^{k} \lambda_i F(\boldsymbol{x}_i).$$

This can be proved by induction on k; the case $k = 2$ is exactly the definition of convexity.

11.1.3 Proof of the *SGD* Theorem

Proof

We start from L-smoothness of F, which gives

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \leq \nabla F(\boldsymbol{x}^k)^\top (\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) + \frac{L}{2} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2.$$

Introduce the notation

$$\boldsymbol{g}_k := g(\boldsymbol{x}^k; \boldsymbol{\xi}_{i_k}), \qquad \boldsymbol{x}^{k+1} - \boldsymbol{x}^k = -\frac{1}{L + \frac{1}{\eta}} \boldsymbol{g}_k.$$

Then

$$\begin{aligned} F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) &= \nabla F(\boldsymbol{x}^k)^\top (\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) + \frac{L}{2} \left(\frac{1}{L + \frac{1}{\eta}}\right)^2 \|\boldsymbol{g}_k\|^2 \\ &= \boldsymbol{g}_k^\top (\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) + \left(\nabla F(\boldsymbol{x}^k) - \boldsymbol{g}_k\right)^\top (\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) \\ &\quad + \frac{L}{2} \left(\frac{1}{L + \frac{1}{\eta}}\right)^2 \|\boldsymbol{g}_k\|^2. \end{aligned}$$

Applying Cauchy–Schwarz gives

$$\begin{aligned} F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) &\leq \boldsymbol{g}_k^\top (\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) + \|\nabla F(\boldsymbol{x}^k) - \boldsymbol{g}_k\| \, \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \\ &\quad + \frac{L}{2} \left(\frac{1}{L + \frac{1}{\eta}}\right)^2 \|\boldsymbol{g}_k\|^2. \end{aligned}$$

For any $\eta > 0$ and $a, b \in \mathbb{R}$ we have

$$\frac{1}{2}\left(\sqrt{\eta}\, a - \tfrac{1}{\sqrt{\eta}} b\right)^2 \geq 0 \quad \Rightarrow \quad ab \leq \frac{\eta}{2} a^2 + \frac{1}{2\eta} b^2.$$

Applying this to $a = \|\nabla F(\boldsymbol{x}^k) - \boldsymbol{g}_k\|$ and $b = \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|$, we obtain

$$\begin{aligned} F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) &\le \boldsymbol{g}_k^\top(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) + \frac{\eta}{2}\|\nabla F(\boldsymbol{x}^k) - \boldsymbol{g}_k\|^2 + \frac{1}{2\eta}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 \\ &\quad + \frac{L}{2}\left(\frac{1}{L + \frac{1}{\eta}}\right)^2 \|\boldsymbol{g}_k\|^2 \\ &= \boldsymbol{g}_k^\top(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) + \frac{\eta}{2}\|\nabla F(\boldsymbol{x}^k) - \boldsymbol{g}_k\|^2 \\ &\quad + \frac{1}{2}\Big(L + \tfrac{1}{\eta}\Big)\left(\frac{1}{L + \frac{1}{\eta}}\right)^2 \|\boldsymbol{g}_k\|^2 \\ &= \boldsymbol{g}_k^\top(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) + \frac{\eta}{2}\|\nabla F(\boldsymbol{x}^k) - \boldsymbol{g}_k\|^2 + \frac{1}{2(L + \frac{1}{\eta})}\|\boldsymbol{g}_k\|^2. \end{aligned}$$

Identity step. Use the vector identity

$$2(b-a)^\top(a-c) = \|b-c\|^2 - \|a-c\|^2 - \|b-a\|^2.$$

Take

$$b := \boldsymbol{x}^k, \quad a := \boldsymbol{x}^{k+1}, \quad c := \boldsymbol{x}^*,$$

so that

$$\boldsymbol{x}^k - \boldsymbol{x}^{k+1} = \frac{1}{L + \frac{1}{\eta}}\,\boldsymbol{g}_k.$$

Then the left-hand side is

$$2 \cdot \frac{1}{L + \frac{1}{\eta}}\,\boldsymbol{g}_k^\top(\boldsymbol{x}^{k+1} - \boldsymbol{x}^*),$$

while the right-hand side is

$$\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2 - \left(\frac{1}{L + \frac{1}{\eta}}\right)^2 \|\boldsymbol{g}_k\|^2.$$

Multiplying both sides by $\frac{L + \frac{1}{\eta}}{2}$ and rearranging yields

$$\frac{1}{2(L + \frac{1}{\eta})}\|\boldsymbol{g}_k\|^2 = -\boldsymbol{g}_k^\top(\boldsymbol{x}^{k+1} - \boldsymbol{x}^*) + \frac{L + \frac{1}{\eta}}{2}\Big(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2\Big).$$

Substitute this into the previous inequality to eliminate the quadratic term:

$$\begin{aligned} F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) &\le \boldsymbol{g}_k^\top(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) + \frac{\eta}{2}\|\nabla F(\boldsymbol{x}^k) - \boldsymbol{g}_k\|^2 \\ &\quad - \boldsymbol{g}_k^\top(\boldsymbol{x}^{k+1} - \boldsymbol{x}^*) + \frac{L + \frac{1}{\eta}}{2}\Big(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2\Big) \\ &= \boldsymbol{g}_k^\top(\boldsymbol{x}^* - \boldsymbol{x}^k) + \frac{\eta}{2}\|\nabla F(\boldsymbol{x}^k) - \boldsymbol{g}_k\|^2 \\ &\quad + \frac{L + \frac{1}{\eta}}{2}\Big(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2\Big). \end{aligned}$$

Write

$$\boldsymbol{g}_k^\top(\boldsymbol{x}^* - \boldsymbol{x}^k) = \nabla F(\boldsymbol{x}^k)^\top(\boldsymbol{x}^* - \boldsymbol{x}^k) + (\boldsymbol{g}_k - \nabla F(\boldsymbol{x}^k))^\top(\boldsymbol{x}^* - \boldsymbol{x}^k),$$

so that

$$\begin{aligned} F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \le{}& \nabla F(\boldsymbol{x}^k)^\top(\boldsymbol{x}^* - \boldsymbol{x}^k) + (\boldsymbol{g}_k - \nabla F(\boldsymbol{x}^k))^\top(\boldsymbol{x}^* - \boldsymbol{x}^k) \\ &+ \frac{\eta}{2}\|\nabla F(\boldsymbol{x}^k) - \boldsymbol{g}_k\|^2 \\ &+ \frac{L + \frac{1}{\eta}}{2}\Big(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2\Big). \end{aligned}$$

By convexity of F,

$$\nabla F(\boldsymbol{x}^k)^\top(\boldsymbol{x}^* - \boldsymbol{x}^k) \le F(\boldsymbol{x}^*) - F(\boldsymbol{x}^k),$$

so

$$\begin{aligned} F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \le{}& F(\boldsymbol{x}^*) - F(\boldsymbol{x}^k) + (\boldsymbol{g}_k - \nabla F(\boldsymbol{x}^k))^\top(\boldsymbol{x}^* - \boldsymbol{x}^k) \\ &+ \frac{\eta}{2}\|\nabla F(\boldsymbol{x}^k) - \boldsymbol{g}_k\|^2 \\ &+ \frac{L + \frac{1}{\eta}}{2}\Big(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2\Big). \end{aligned}$$

Rearranging,

$$\begin{aligned} F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^*) \le{}& (\boldsymbol{g}_k - \nabla F(\boldsymbol{x}^k))^\top(\boldsymbol{x}^* - \boldsymbol{x}^k) + \frac{\eta}{2}\|\nabla F(\boldsymbol{x}^k) - \boldsymbol{g}_k\|^2 \\ &+ \frac{L + \frac{1}{\eta}}{2}\Big(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2\Big). \end{aligned}$$

Taking expectations. Take conditional expectation with respect to the randomness at step k, conditioning on the history so that $\boldsymbol{x}^k$ is fixed. By the unbiasedness and variance assumptions,

$$\mathbb{E}[\boldsymbol{g}_k \mid \text{history}] = \nabla F(\boldsymbol{x}^k), \qquad \mathbb{E}[\|\nabla F(\boldsymbol{x}^k) - \boldsymbol{g}_k\|^2 \mid \text{history}] \le \sigma^2.$$

Thus

$$\mathbb{E}\big[(\boldsymbol{g}_k - \nabla F(\boldsymbol{x}^k))^\top(\boldsymbol{x}^* - \boldsymbol{x}^k) \mid \text{history}\big] = 0,$$

and we obtain

$$\mathbb{E}\big[F(\boldsymbol{x}^{k+1}) \mid \text{history}\big] - F(\boldsymbol{x}^*) \le \frac{\eta}{2}\sigma^2 + \frac{L + \frac{1}{\eta}}{2}\Big(\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \mathbb{E}\big[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2 \mid \text{history}\big]\Big).$$

Now take full expectation and let $R^2 := \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2$:

$$\mathbb{E}[F(\boldsymbol{x}^{k+1})] - F(\boldsymbol{x}^*) \le \frac{\eta}{2}\sigma^2 + \frac{L + \frac{1}{\eta}}{2}\Big(\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \mathbb{E}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2\Big).$$

Average this inequality over $k = 0, 1, \ldots, \ell - 1$:

$$\begin{aligned} \frac{1}{\ell}\sum_{k=0}^{\ell-1}\big(\mathbb{E}[F(\boldsymbol{x}^{k+1})] - F(\boldsymbol{x}^*)\big) &\le \frac{\eta}{2}\sigma^2 + \frac{L + \frac{1}{\eta}}{2\ell}\Big(\mathbb{E}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2 - \mathbb{E}\|\boldsymbol{x}^\ell - \boldsymbol{x}^*\|^2\Big) \\ &\le \frac{\eta}{2}\sigma^2 + \frac{L + \frac{1}{\eta}}{2\ell}R^2. \end{aligned}$$

By convexity of F and Jensen's inequality,

$$\mathbb{E}\left[F\left(\frac{1}{\ell}\sum_{k=1}^{\ell}\boldsymbol{x}^k\right)\right] \le \frac{1}{\ell}\sum_{k=1}^{\ell}\mathbb{E}[F(\boldsymbol{x}^k)],$$

so

$$\mathbb{E}\left[F\left(\frac{1}{\ell}\sum_{k=1}^{\ell}\boldsymbol{x}^k\right)\right] - F(\boldsymbol{x}^*) \le \frac{\eta}{2}\sigma^2 + \frac{L+\frac{1}{\eta}}{2\ell}R^2.$$

Finally, substitute $\eta = \dfrac{R}{\sigma\sqrt{\ell}}$:

$$\begin{aligned}\frac{\eta}{2}\sigma^2 + \frac{L+\frac{1}{\eta}}{2\ell}R^2 &= \frac{1}{2}\cdot\frac{R}{\sigma\sqrt{\ell}}\sigma^2 + \frac{1}{2\ell}\Big(L + \frac{\sigma\sqrt{\ell}}{R}\Big)R^2 \\ &= \frac{R\sigma}{2\sqrt{\ell}} + \frac{LR^2}{2\ell} + \frac{R\sigma}{2\sqrt{\ell}} \\ &= \frac{R\sigma}{\sqrt{\ell}} + \frac{LR^2}{2\ell},\end{aligned}$$

which gives the desired bound.

Remarks.

- The step size α_k depends on ℓ, the total number of iterations. Why not keep going after iteration ℓ? The convergence effectively *stalls* after ℓ because each new step introduces a stochastic error of size on the order of

$$\frac{\eta}{2}\sigma^2 = \frac{1}{2}\cdot\frac{R}{\sigma\sqrt{\ell}}\sigma^2 = \frac{R\sigma}{2\sqrt{\ell}}.$$

- This is an *ergodic* bound (on the average iterate $\dfrac{1}{\ell}\sum_{k=1}^{\ell}\boldsymbol{x}^k$), rather than a *last-iterate* bound. In practice, people often just keep the last iterate, but averaging is important for the theoretical guarantee.
- By increasing ℓ, we can make the bound smaller; however, the variance term decays only as $1/\sqrt{\ell}$, unlike the $1/\ell$ decay seen in deterministic gradient descent.

Lec 12

12.1 Review of *SGD*

From last lecture: *SGD* convergence.

- *Ergodic* convergence vs. *last-iterate* convergence.
- Step size $\alpha = \dfrac{1}{L + \frac{1}{\eta}}$ depends on the number of steps ℓ (this implies α cannot be arbitrarily large, otherwise the method diverges; the upper bound is $\dfrac{1}{L}$).
- Larger α means faster convergence to a poorer solution.
- Decrease step size: one can use, for example, $\alpha_k = \dfrac{\text{const}}{\text{const} + k}$.
- In practice: use a *stepsize schedule* (e.g. α_k is a staircase function of iteration k).

SGD and its variants are the main optimization algorithms used for training modern machine-learning models (including ChatGPT, Deepseek, ...).

12.2 *SGD* Variants

12.2.1 *SGD* Variants in Theory

- SVRG,
- SAGA.

Both of these do variance reduction, mainly for convex problems.

12.2.2 *SGD* Variants in Practice

- **Minibatching:** e.g. select 8 random samples instead of 1 and compute the gradient of the sum of 8 terms.
- **Momentum** (similar to *AGD* momentum).
- **Coordinatewise scale**
$$\boldsymbol{x}^{k+1} := \boldsymbol{x}^k - \boldsymbol{D}^k \boldsymbol{g}^k,$$
where $\boldsymbol{D}^k$ is a diagonal matrix adaptively updated. In ordinary *SGD*, $\boldsymbol{D}^k = \alpha_k \boldsymbol{I}$.
An extreme case is
$$\boldsymbol{x}^{k+1} := \boldsymbol{x}^k - \alpha_k \cdot \operatorname{sign}(\boldsymbol{g}^k).$$
- A well-known *SGD* variant is called **ADAM**, which combines momentum and coordinatewise scaling.

12.3 Constrained and Nonsmooth Analysis

12.3.1 Motivating Examples

ℓ_1-regularized least squares. Given $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{y} \in \mathbb{R}^m$, $\gamma > 0$,

$$\ell_1\text{LS}: \quad \min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 + \gamma\|\boldsymbol{x}\|_1,$$

where $\|\boldsymbol{x}\|_1 = |\boldsymbol{x}_1| + \cdots + |\boldsymbol{x}_n|$. One can check the three properties of a norm are satisfied.

Compare to ridge regression:

$$\text{RR}: \quad \min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 + \frac{\gamma}{2}\|\boldsymbol{x}\|_2^2.$$

(i) Both ℓ_1LS and RR encourage shrinkage, i.e. large $\boldsymbol{x}$ is penalized.

(ii) ℓ_1LS has a second effect: *selection*.

A solution $\boldsymbol{x}$ to ℓ_1LS has many entries identically equal to 0. Thus it identifies features (columns of $\boldsymbol{A}$) that are irrelevant to explaining $\boldsymbol{y}$. Closely related to ℓ_1LS is Lasso regression:

$$\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 \quad \text{s.t.} \quad \|\boldsymbol{x}\|_1 \le \rho.$$

Observe that

$$g(\boldsymbol{x}) := \|\boldsymbol{x}\|_1$$

is a convex function, but not differentiable. So ℓ_1LS is convex and nonsmooth.

Therefore plain *GD* and *AGD* (which need gradients) are not directly applicable.

We can rewrite ℓ_1LS as a smooth constrained problem by introducing auxiliary variables:

$$\min_{\boldsymbol{x},\boldsymbol{t}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 + \gamma\sum_{i=1}^n t_i \quad \text{s.t.} \quad t_i \ge x_i,\ \ t_i \ge -x_i,\ \ \forall i = 1,\ldots,n.$$

These constraints are equivalent to $t_i \ge |x_i|$, but this latter form is nonsmooth.

To exactly reproduce ℓ_1LS, we should constrain $t_i = |x_i|$. It is nevertheless OK to use $t_i \ge |x_i|$, because no optimizer of the above problem would have $t_i > |x_i|$, due to the form of the objective function.

Support vector machines. Binary classification: given $(\boldsymbol{a}_1, y_1), \ldots, (\boldsymbol{a}_N, y_N)$, with $\boldsymbol{a}_i \in \mathbb{R}^n$, $y_i \in \{1, -1\}$, seek $\boldsymbol{x} \in \mathbb{R}^n$, $\xi \in \mathbb{R}$ such that

$$\begin{aligned} \boldsymbol{a}_i^\top \boldsymbol{x} + \xi > 0 \quad &\text{if } y_i = 1, \\ \boldsymbol{a}_i^\top \boldsymbol{x} + \xi < 0 \quad &\text{if } y_i = -1. \end{aligned}$$

These are strict linear inequalities on $(\boldsymbol{x}, \xi)$. The solution set of a system of (non-strict) linear inequalities is called a *polyhedron* (generalization of polygons in 2D).

Idea of SVM: find $(\boldsymbol{x}, \xi)$ to maximize the width between the $+1$ and -1 classes.

Theorem

Assume $\|\boldsymbol{x}\| = 1$. Given $\boldsymbol{a} \in \mathbb{R}^n$, the distance from $\boldsymbol{a}$ to the hyperplane

$$\{\boldsymbol{v} : \boldsymbol{v}^\top \boldsymbol{x} + \xi = 0\}$$

is $|\boldsymbol{a}^\top \boldsymbol{x} + \xi|$.

Proof

Claim that the closest point on the plane to $\boldsymbol{a}$ is

$$\boldsymbol{v}^* := \boldsymbol{a} - (\boldsymbol{a}^\top \boldsymbol{x} + \xi)\boldsymbol{x}.$$

Check that $\boldsymbol{v}^*$ lies on the plane.

$$(\boldsymbol{v}^*)^\top \boldsymbol{x} + \xi = \boldsymbol{a}^\top \boldsymbol{x} - (\boldsymbol{a}^\top \boldsymbol{x} + \xi)\, \boldsymbol{x}^\top \boldsymbol{x} + \xi = \boldsymbol{a}^\top \boldsymbol{x} - (\boldsymbol{a}^\top \boldsymbol{x} + \xi) \cdot 1 + \xi = 0.$$

Check that the distance from $\boldsymbol{a}$ to $\boldsymbol{v}^*$ is as claimed:

$$\|\boldsymbol{a} - \boldsymbol{v}^*\| = \|\boldsymbol{a} - \boldsymbol{a} + (\boldsymbol{a}^\top \boldsymbol{x} + \xi)\boldsymbol{x}\| = |\boldsymbol{a}^\top \boldsymbol{x} + \xi| \cdot \|\boldsymbol{x}\| = |\boldsymbol{a}^\top \boldsymbol{x} + \xi|$$

since $\|\boldsymbol{x}\| = 1$.
Check that no other point on the plane is closer to $\boldsymbol{a}$ than $\boldsymbol{v}^*$.
Let $\boldsymbol{v}^1$ also lie on the plane, so $(\boldsymbol{v}^1)^\top \boldsymbol{x} + \xi = 0$. Then

$$(\boldsymbol{v}^1 - \boldsymbol{v}^*)^\top \boldsymbol{x} = 0.$$

Let $\boldsymbol{p} := \boldsymbol{v}^1 - \boldsymbol{v}^*$. Then

$$\begin{aligned} \|\boldsymbol{a} - \boldsymbol{v}^1\|^2 &= \|\boldsymbol{a} - \boldsymbol{a} + (\boldsymbol{a}^\top \boldsymbol{x} + \xi)\boldsymbol{x} - \boldsymbol{p}\|^2 \\ &= |\boldsymbol{a}^\top \boldsymbol{x} + \xi|^2 \|\boldsymbol{x}\|^2 - 2(\boldsymbol{a}^\top \boldsymbol{x} + \xi)\, \boldsymbol{x}^\top \boldsymbol{p} + \|\boldsymbol{p}\|^2 \\ &= |\boldsymbol{a}^\top \boldsymbol{x} + \xi|^2 + \|\boldsymbol{p}\|^2 \end{aligned}$$

because $\boldsymbol{x}^\top \boldsymbol{p} = 0$. So the distance is minimized when $\boldsymbol{p} = \boldsymbol{0}$, i.e. when $\boldsymbol{v}^1 = \boldsymbol{v}^*$.

Thus we can write the SVM optimization problem as

$$\begin{aligned} \text{(SVM-1)} \quad & \max_{\boldsymbol{x},\xi} \min_{i=1,\dots,N} |\boldsymbol{a}_i^\top \boldsymbol{x} + \xi| \\ \text{s.t.} \quad & \boldsymbol{a}_i^\top \boldsymbol{x} + \xi \geq 0 \quad (y_i = 1), \\ & \boldsymbol{a}_i^\top \boldsymbol{x} + \xi \leq 0 \quad (y_i = -1), \\ & \|\boldsymbol{x}\| = 1. \end{aligned}$$

Assume for now that there exists a separating plane $(\boldsymbol{x}, \xi)$ for which the above objective is positive. This implies the objective function is positive at the optimizer.

Introduce variables $\boldsymbol{t}$ to remove the absolute value:

$$\begin{aligned} \text{(SVM-2)} \quad & \max_{\boldsymbol{x},\xi,\boldsymbol{t}} \ \min_{i=1,\dots,N} t_i \\ \text{s.t.} \quad & \boldsymbol{a}_i^\top \boldsymbol{x} + \xi = \ \ t_i \quad (y_i = 1), \\ & \boldsymbol{a}_i^\top \boldsymbol{x} + \xi = -t_i \quad (y_i = -1), \\ & \|\boldsymbol{x}\| = 1, \quad \boldsymbol{t} \geq \boldsymbol{0}. \end{aligned}$$

This formulation gets rid of the absolute value in SVM-1.

Claim

We obtain the same optimizer if we replace $\|\boldsymbol{x}\| = 1$ with $\|\boldsymbol{x}\| \leq 1$:

$$\begin{aligned} \text{(SVM-3)} \quad & \max_{\boldsymbol{x},\xi,\boldsymbol{t}} \ \min_{i=1,\dots,N} t_i \\ \text{s.t.} \quad & \boldsymbol{a}_i^\top \boldsymbol{x} + \xi = \ \ t_i \quad (y_i = 1), \\ & \boldsymbol{a}_i^\top \boldsymbol{x} + \xi = -t_i \quad (y_i = -1), \\ & \|\boldsymbol{x}\| \leq 1, \quad \boldsymbol{t} \geq \boldsymbol{0}. \end{aligned}$$

Proof

Suppose $(\boldsymbol{x}, \xi, \boldsymbol{t})$ is feasible for SVM-3 with $\|\boldsymbol{x}\| < 1$.
Assume $\boldsymbol{x} \neq \boldsymbol{0}$ (indeed, if $\boldsymbol{x} = \boldsymbol{0}$, then the constraints force $\xi = 0$ and hence the objective $\min_i t_i = 0$, which cannot be optimal since we assumed the optimal value is positive).
Define the rescaled variables

$$\boldsymbol{x}^1 := \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, \qquad \xi^1 := \frac{\xi}{\|\boldsymbol{x}\|}, \qquad \boldsymbol{t}^1 := \frac{\boldsymbol{t}}{\|\boldsymbol{x}\|}.$$

One checks that $(\boldsymbol{x}^1, \xi^1, \boldsymbol{t}^1)$ is feasible and

$$\min_i t_i^1 = \frac{1}{\|\boldsymbol{x}\|} \min_i t_i > \min_i t_i,$$

since $\|\boldsymbol{x}\| < 1$. Thus we obtain a strictly higher objective value, contradicting optimality of a solution with $\|\boldsymbol{x}\| < 1$. Therefore any optimizer of SVM-3 must satisfy $\|\boldsymbol{x}\| = 1$, and the optimizers of SVM-2 and SVM-3 coincide.

Lec 13

13.1 Constrained and Nonsmooth Analysis

13.1.1 SVM Example

Binary classification: given $(\boldsymbol{a}_1, y_1), \ldots, (\boldsymbol{a}_N, y_N)$ with

$$\boldsymbol{a}_i \in \mathbb{R}^n, \quad y_i \in \{1, -1\},$$

seek $\boldsymbol{x} \in \mathbb{R}^n$, $\xi \in \mathbb{R}$ such that

$$\begin{aligned} \boldsymbol{a}_i^\top \boldsymbol{x} + \xi > 0 \quad &\text{if } y_i = 1, \\ \boldsymbol{a}_i^\top \boldsymbol{x} + \xi < 0 \quad &\text{if } y_i = -1. \end{aligned}$$

$$\text{(SVM-1)} \quad \max_{\boldsymbol{x},\xi} \min_{i=1,\ldots,N} |\boldsymbol{a}_i^\top \boldsymbol{x} + \xi| \quad \text{s.t.} \quad \boldsymbol{a}_i^\top \boldsymbol{x} + \xi \geq 0 \ (y_i = 1),\ \boldsymbol{a}_i^\top \boldsymbol{x} + \xi \leq 0 \ (y_i = -1),\ \|\boldsymbol{x}\| = 1.$$

$$\text{(SVM-2)} \quad \max_{\boldsymbol{x},\xi,\boldsymbol{t}} \min_{i=1,\ldots,N} t_i \quad \text{s.t.} \quad \boldsymbol{a}_i^\top \boldsymbol{x} + \xi = t_i \ (y_i = 1),\ \boldsymbol{a}_i^\top \boldsymbol{x} + \xi = -t_i \ (y_i = -1),\ \|\boldsymbol{x}\| = 1,\ \boldsymbol{t} \geq \boldsymbol{0}.$$

$$\text{(SVM-3)} \quad \max_{\boldsymbol{x},\xi,\boldsymbol{t}} \min_{i=1,\ldots,N} t_i \quad \text{s.t.} \quad \boldsymbol{a}_i^\top \boldsymbol{x} + \xi = t_i \ (y_i = 1),\ \boldsymbol{a}_i^\top \boldsymbol{x} + \xi = -t_i \ (y_i = -1),\ \|\boldsymbol{x}\| \leq 1,\ \boldsymbol{t} \geq \boldsymbol{0}.$$

SVM-2 is nonconvex; SVM-3 is convex.

Note: $\min(x, y)$ is not differentiable when $x = y$.

Let $\tilde{t} = \min(t_1, \ldots, t_N)$. Then we can write

$$\text{(SVM-4)} \quad \max_{\boldsymbol{x},\xi,\tilde{t}} \tilde{t} \quad \text{s.t.} \quad \boldsymbol{a}_i^\top \boldsymbol{x} + \xi \geq \tilde{t} \ (y_i = 1),\ \boldsymbol{a}_i^\top \boldsymbol{x} + \xi \leq -\tilde{t} \ (y_i = -1),\ \|\boldsymbol{x}\| \leq 1,\ \tilde{t} \geq 0.$$

This is equivalent to SVM-3: given a feasible point for SVM-4, define

$$t_i = \begin{cases} \boldsymbol{a}_i^\top \boldsymbol{x} + \xi, & y_i = 1, \\ -\boldsymbol{a}_i^\top \boldsymbol{x} - \xi, & y_i = -1. \end{cases}$$

Any $\tilde{t} \leq \min(t_1, \ldots, t_N)$ is feasible for SVM-4. Among these choices of $\tilde{t}$, SVM-4 selects the largest one, i.e. $\tilde{t} = \min(t_1, \ldots, t_N)$.

Move the quadratic constraint into the objective: change variables

$$\tilde{\boldsymbol{x}} := \frac{\boldsymbol{x}}{\tilde{t}}, \qquad \tilde{\xi} := \frac{\xi}{\tilde{t}},$$

which is valid since $\tilde{t} > 0$ at the minimizer of SVM-4. Divide all constraints by $\tilde{t}$ to obtain SVM-5:

$$\text{(SVM-5)} \quad \max_{\tilde{\boldsymbol{x}},\tilde{\xi},\tilde{t}} \tilde{t} \quad \text{s.t.} \quad \boldsymbol{a}_i^\top \tilde{\boldsymbol{x}} + \tilde{\xi} \geq 1 \ (y_i = 1),\ \boldsymbol{a}_i^\top \tilde{\boldsymbol{x}} + \tilde{\xi} \leq -1 \ (y_i = -1),\ \|\tilde{t}\,\tilde{\boldsymbol{x}}\| \leq 1,\ \tilde{t} \geq 0.$$

Observe that

$$\max \tilde{t} \text{ s.t. } \|\tilde{t}\,\tilde{\boldsymbol{x}}\| \le 1,\ \tilde{t} \ge 0$$

has the same optimizer (in $\tilde{\boldsymbol{x}}$) as

$$\min \|\tilde{\boldsymbol{x}}\| \text{ s.t. } \|\tilde{t}\,\tilde{\boldsymbol{x}}\| \le 1,\ \tilde{t} \ge 0.$$

Thus we arrive at

$$\text{(SVM-6)} \quad \min_{\tilde{\boldsymbol{x}},\tilde{\xi}} \ \|\tilde{\boldsymbol{x}}\|^2 \quad \text{s.t.} \quad \boldsymbol{a}_i^\top \tilde{\boldsymbol{x}} + \tilde{\xi} \ge 1 \ (y_i = 1),\ \boldsymbol{a}_i^\top \tilde{\boldsymbol{x}} + \tilde{\xi} \le -1 \ (y_i = -1).$$

Both SVM-6 and the smooth reformulation of ℓ_1LS (with constraints) are examples of *quadratic programming* (QP).

Fact

QP refers to minimizing a quadratic objective function subject to linear equality and/or inequality constraints.

SVM-6 is called **hard-margin SVM**; it requires that the data have an affine linear separator (otherwise the problem is infeasible).

For nonseparable data, we use SVM-7, the **soft-margin SVM**, where $\gamma > 0$ is a penalty parameter:

$$\text{(SVM-7)} \quad \min_{\boldsymbol{x},\xi,\boldsymbol{s}} \ \frac{1}{2}\|\tilde{\boldsymbol{x}}\|^2 + \gamma \sum_{i=1}^{N} s_i$$

$$\text{s.t.} \quad \boldsymbol{a}_i^\top \boldsymbol{x} + \xi \ge 1 - s_i \ (y_i = 1), \quad \boldsymbol{a}_i^\top \boldsymbol{x} + \xi \le -1 + s_i \ (y_i = -1), \quad \boldsymbol{s} \ge \boldsymbol{0}.$$

If $\boldsymbol{s} = \boldsymbol{0}$, we recover SVM-6. If $s_i > 0$ for some i, then $\boldsymbol{a}_i$ is misclassified. Thus the second term of the objective penalizes misclassification; $\gamma \to \infty$ forces $s_i = 0$ and recovers SVM-6. When $\gamma \to 0$, misclassifications are penalized very little, so the solution prioritizes a wider margin.

Many of the s_i are expected to be exactly zero at a typical solution. We can eliminate the s_i to obtain a nonsmooth equivalent problem.

If $\boldsymbol{x}, \xi$ are fixed, the best choice for s_i when $y_i = 1$ is either

$$s_i = 1 - \boldsymbol{a}_i^\top \boldsymbol{x} - \xi \quad \text{or} \quad s_i = 0,$$

so the optimal choice is

$$s_i = \max(0,\ 1 - \boldsymbol{a}_i^\top \boldsymbol{x} - \xi).$$

Write this as $s_i = \Phi(\boldsymbol{a}_i^\top \boldsymbol{x} + \xi)$ where

$$\Phi(t) = \max(0, 1 - t).$$

For $y_i = -1$, the best choice is $s_i = \Phi\big(-(\boldsymbol{a}_i^\top \boldsymbol{x} + \xi)\big)$. With this notation we obtain SVM-8:

$$\text{(SVM-8)} \quad \min_{\boldsymbol{x},\xi} \ \frac{1}{2}\|\tilde{\boldsymbol{x}}\|^2 + \gamma \sum_{y_i=1} \Phi(\boldsymbol{a}_i^\top \boldsymbol{x} + \xi) + \gamma \sum_{y_i=-1} \Phi\big(-(\boldsymbol{a}_i^\top \boldsymbol{x} + \xi)\big).$$

Observe that Φ is convex, so SVM-8 is a convex, nonsmooth, unconstrained problem. The function Φ is called the *hinge loss*.

Fact

The logistic loss $\ln(1+e^{-t})$ from logistic regression is a smooth approximation of the hinge loss.

Differences between SVM-8 and logistic regression with ℓ_2 regularization:

(i) In logistic regression the penalty parameter multiplies the quadratic term (a notational difference).

(ii) Logistic regression uses a smooth loss; the hinge loss is nonsmooth.

(iii) The quadratic term in logistic regression is $\frac{1}{2}\gamma\|(\boldsymbol{x},\xi)\|^2$, whereas in SVM-8 the quadratic term is independent of ξ.

13.1.2 General Case

Consider the constrained optimization problem

$$\min f(\boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{x}\in\Omega,\ \Omega\subseteq\mathbb{R}^n.$$

Assume Ω is closed and convex.

The **normal cone** to Ω at $\boldsymbol{x}\in\Omega$ is

$$N_\Omega(\boldsymbol{x}) = \left\{\boldsymbol{v}\in\mathbb{R}^n : \boldsymbol{v}^\top(\boldsymbol{y}-\boldsymbol{x})\le 0,\ \forall \boldsymbol{y}\in\Omega\right\}.$$

Geometrically, if $\boldsymbol{x}$ lies on the boundary $\partial\Omega$, the normal cone $N_\Omega(\boldsymbol{x})$ consists of all vectors that point "outward" and are normal to Ω at $\boldsymbol{x}$. For an interior point, $N_\Omega(\boldsymbol{x})=\{\boldsymbol{0}\}$. For a nonsmooth boundary point, the normal cone is generated by all limiting normals coming from adjacent smooth patches.

A **closed convex cone** $C\subseteq\mathbb{R}^n$ satisfies:

- C is closed;
- $\boldsymbol{0}\in C$;
- if $\boldsymbol{x}\in C$ and $\lambda\ge 0$, then $\lambda\boldsymbol{x}\in C$;
- if $\boldsymbol{x},\boldsymbol{y}\in C$, then $\boldsymbol{x}+\boldsymbol{y}\in C$.

Fact

- Any closed convex cone is a convex set.
- If Ω is a closed convex set and $\boldsymbol{x}\in\Omega$, then $N_\Omega(\boldsymbol{x})$ is a closed convex cone.

Theorem

Suppose $\Omega\subseteq\mathbb{R}^n$ is closed, nonempty, and convex, and $f:\mathbb{R}^n\to\mathbb{R}\cup\{\infty\}$ is convex and differentiable, with $\Omega\subseteq\operatorname{int}(\operatorname{dom} f)$. Then

$$\boldsymbol{x}^*\in\arg\min\{f(\boldsymbol{x}):\boldsymbol{x}\in\Omega\} \iff \boldsymbol{x}^*\in\Omega \text{ and } -\nabla f(\boldsymbol{x}^*)\in N_\Omega(\boldsymbol{x}^*).$$

Proof

Backward direction ($\Leftarrow$).
Assume $\boldsymbol{x}^* \in \Omega$ and $-\nabla f(\boldsymbol{x}^*) \in N_\Omega(\boldsymbol{x}^*)$. For any $\boldsymbol{z} \in \Omega$, the subgradient (gradient) inequality for convex f gives

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{z} - \boldsymbol{x}^*).$$

Since $-\nabla f(\boldsymbol{x}^*) \in N_\Omega(\boldsymbol{x}^*)$ and $\boldsymbol{z} \in \Omega$, we have

$$-\nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{z} - \boldsymbol{x}^*) \leq 0,$$

so

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}^*).$$

Thus $\boldsymbol{x}^*$ is a minimizer over Ω.

Forward direction ($\Rightarrow$).
Assume $\boldsymbol{x}^*$ is a minimizer. Take any $\boldsymbol{z} \in \Omega$. For all $\alpha \in [0,1]$, convexity of Ω implies $(1-\alpha)\boldsymbol{x}^* + \alpha \boldsymbol{z} \in \Omega$, and by optimality,

$$f(\boldsymbol{x}^*) \leq f\big((1-\alpha)\boldsymbol{x}^* + \alpha \boldsymbol{z}\big) = f\big(\boldsymbol{x}^* + \alpha(\boldsymbol{z} - \boldsymbol{x}^*)\big).$$

By differentiability, write a first-order expansion with remainder:

$$f(\boldsymbol{x}^* + \alpha(\boldsymbol{z} - \boldsymbol{x}^*)) = f(\boldsymbol{x}^*) + \alpha \nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{z} - \boldsymbol{x}^*) + \phi(\alpha(\boldsymbol{z} - \boldsymbol{x}^*)),$$

where

$$\frac{\phi(\alpha(\boldsymbol{z} - \boldsymbol{x}^*))}{\|\alpha(\boldsymbol{z} - \boldsymbol{x}^*)\|} \to 0 \quad \text{as } \alpha \to 0.$$

Thus

$$0 \leq \alpha \nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{z} - \boldsymbol{x}^*) + \phi(\alpha(\boldsymbol{z} - \boldsymbol{x}^*)).$$

Divide by $\alpha > 0$:

$$0 \leq \nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{z} - \boldsymbol{x}^*) + \frac{\phi(\alpha(\boldsymbol{z} - \boldsymbol{x}^*))}{\alpha}.$$

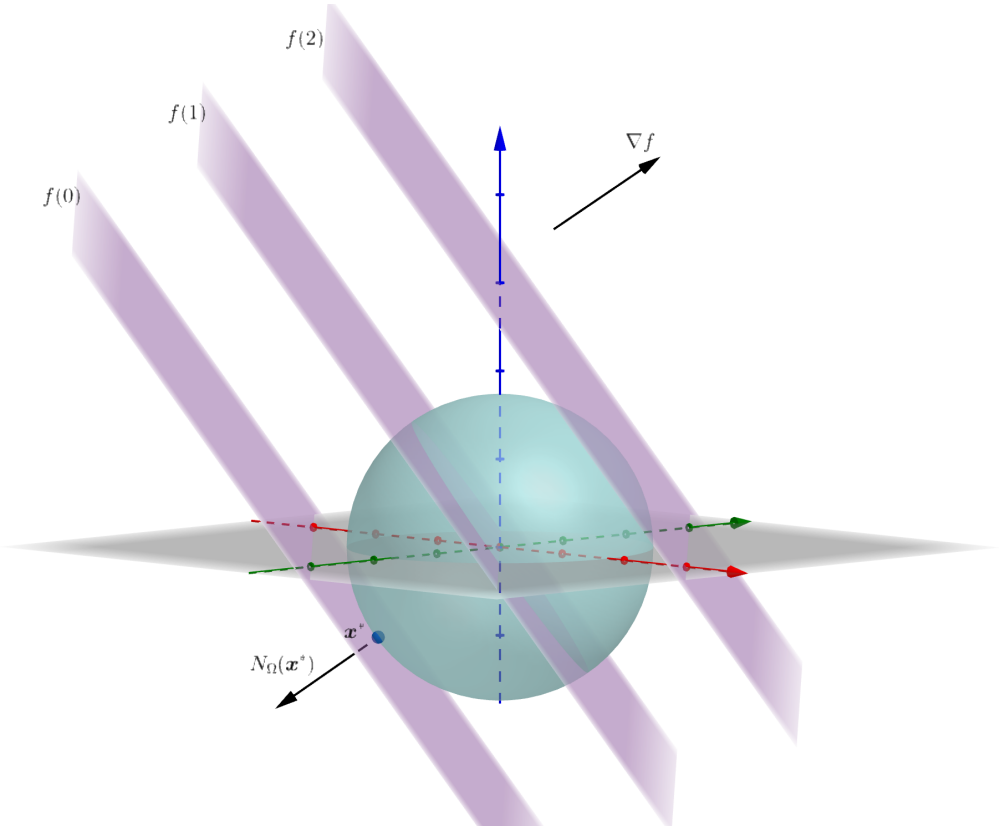
Let $\alpha \downarrow 0$; the second term tends to 0, so

$$\nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{z} - \boldsymbol{x}^*) \geq 0, \quad \forall \boldsymbol{z} \in \Omega.$$

Equivalently,

$$-\nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{z} - \boldsymbol{x}^*) \leq 0, \quad \forall \boldsymbol{z} \in \Omega,$$

which means $-\nabla f(\boldsymbol{x}^*) \in N_\Omega(\boldsymbol{x}^*)$.
Moreover, if $\lambda \geq 0$, then $-\lambda \nabla f(\boldsymbol{x}^*) \in N_\Omega(\boldsymbol{x}^*)$ as well, showing that $N_\Omega(\boldsymbol{x}^*)$ is a cone.



(This has a nice geometric interpretation and explains why the Lagrange multiplier method makes sense.)

Lec 14

14.1 Theorems with Normal Cones

Theorem

(From last lecture). Suppose $\Omega \subseteq \mathbb{R}^n$ is closed, nonempty, and convex, and $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and differentiable, with $\Omega \subseteq \text{intdom}(f)$. Then

$$\boldsymbol{x}^* \in \arg\min\{f(\boldsymbol{x}) : \boldsymbol{x} \in \Omega\} \quad \iff \quad \boldsymbol{x}^* \in \Omega \text{ and } -\nabla f(\boldsymbol{x}^*) \in N_\Omega(\boldsymbol{x}^*).$$

Theorem

Let $\Omega = \Omega_1 \cap \cdots \cap \Omega_m$, where each Ω_i is closed and convex (so Ω is closed and convex). Then, for all $\boldsymbol{x} \in \Omega$,

$$N_{\Omega_1}(\boldsymbol{x}) + \cdots + N_{\Omega_m}(\boldsymbol{x}) \ \subseteq \ N_\Omega(\boldsymbol{x}).$$

Here $+$ denotes the **Minkowski sum**: given $A, B \subseteq \mathbb{R}^n$,

$$A + B = \{\boldsymbol{a} + \boldsymbol{b} : \boldsymbol{a} \in A, \ \boldsymbol{b} \in B\}.$$

Proof

Let $\boldsymbol{v} = \boldsymbol{v}_1 + \cdots + \boldsymbol{v}_m$ with $\boldsymbol{v}_i \in N_{\Omega_i}(\boldsymbol{x})$ for all $i = 1, \ldots, m$.
By definition of the normal cone,

$$\boldsymbol{v}_i^\top (\boldsymbol{z} - \boldsymbol{x}) \le 0, \quad \forall \boldsymbol{z} \in \Omega, \ \forall i = 1, \ldots, m.$$

Summing these inequalities over i gives

$$\boldsymbol{v}^\top (\boldsymbol{z} - \boldsymbol{x}) = \sum_{i=1}^m \boldsymbol{v}_i^\top (\boldsymbol{z} - \boldsymbol{x}) \le 0, \quad \forall \boldsymbol{z} \in \Omega.$$

Hence $\boldsymbol{v} \in N_\Omega(\boldsymbol{x})$.

In many generic situations, the inclusion in the theorem is actually an equality:

$$N_{\Omega_1}(\boldsymbol{x}) + \cdots + N_{\Omega_m}(\boldsymbol{x}) = N_\Omega(\boldsymbol{x}).$$

A condition that guarantees this equality is called a **constraint qualification** (CQ).

Fact

Constraint qualifications ensuring equality.

- (CQ1) If $\Omega_1, \ldots, \Omega_m$ are all polyhedral sets, then

$$N_{\Omega_1}(\boldsymbol{x}) + \cdots + N_{\Omega_m}(\boldsymbol{x}) = N_\Omega(\boldsymbol{x}).$$

- (CQ2, Slater condition) If $\text{int}(\Omega) \neq \varnothing$, then the same equality holds.

Example

Example where equality fails.
Let $n = 2$, $m = 2$, and

$$\Omega_1 = \{\boldsymbol{x} : x_2 \leq 0\}, \qquad \Omega_2 = \{\boldsymbol{x} : x_2 \geq x_1^2\},$$

and set $\Omega := \Omega_1 \cap \Omega_2 = \{\mathbf{0}\}$.
Then

$$N_\Omega(\mathbf{0}) = \mathbb{R}^2.$$

However,

$$N_{\Omega_1}(\mathbf{0}) = \{\lambda(0,1)^\top : \lambda \geq 0\}, \qquad N_{\Omega_2}(\mathbf{0}) = \{\lambda(0,1)^\top : \lambda \leq 0\}.$$

Thus

$$N_{\Omega_1}(\mathbf{0}) + N_{\Omega_2}(\mathbf{0}) = \{\lambda(0,1)^\top : \lambda \in \mathbb{R}\} \subsetneq \mathbb{R}^2 = N_\Omega(\mathbf{0}).$$

Neither CQ1 nor CQ2 holds in this example. Intuitively, the normal cone summarizes the local behavior of Ω via linear approximation; here the linear approximations of Ω_1 and Ω_2 fail to capture the "sharp" intersection at the origin.

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex, and fix $\beta \in \mathbb{R}$. Define the level set

$$S = \{\boldsymbol{y} : f(\boldsymbol{y}) \leq \beta\}.$$

For any $\boldsymbol{x} \in S$:

a) S is convex.

b) If $f(\boldsymbol{x}) < \beta$, then $N_S(\boldsymbol{x}) = \{\mathbf{0}\}$.

c) If $f(\boldsymbol{x}) = \beta$ and $\nabla f(\boldsymbol{x}) \neq \mathbf{0}$, then

$$N_S(\boldsymbol{x}) = \{\lambda \nabla f(\boldsymbol{x}) : \lambda \geq 0\}.$$

Proof

(a) Convexity of S.
Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S$ and $\lambda \in [0,1]$. By convexity of f,

$$f\big((1-\lambda)\boldsymbol{x}_1 + \lambda\boldsymbol{x}_2\big) \leq (1-\lambda)f(\boldsymbol{x}_1) + \lambda f(\boldsymbol{x}_2) \leq (1-\lambda)\beta + \lambda\beta = \beta.$$

Hence $(1-\lambda)\boldsymbol{x}_1 + \lambda\boldsymbol{x}_2 \in S$, so S is convex.

(b) Interior point: $f(\boldsymbol{x}) < \beta \Rightarrow N_S(\boldsymbol{x}) = \{\mathbf{0}\}$.
Since f is convex, it is continuous on the interior of its domain. Thus there exists $r > 0$ such that

$$f(\boldsymbol{y}) \leq \beta, \quad \forall \boldsymbol{y} \in \overline{\mathbb{B}}(\boldsymbol{x}, r),$$

so $\overline{\mathbb{B}}(\boldsymbol{x}, r) \subseteq S$.
Assume, to get a contradiction, that $\boldsymbol{v} \in N_S(\boldsymbol{x}) \setminus \{\mathbf{0}\}$. Then, by definition of the normal cone,

$$\boldsymbol{v}^\top\big(\boldsymbol{x} + r\,\tfrac{\boldsymbol{v}}{\|\boldsymbol{v}\|} - \boldsymbol{x}\big) \leq 0,$$

i.e.

$$\frac{r\,\boldsymbol{v}^\top \boldsymbol{v}}{\|\boldsymbol{v}\|} = \frac{r\|\boldsymbol{v}\|^2}{\|\boldsymbol{v}\|} = r\|\boldsymbol{v}\| \le 0,$$

which is impossible since $r > 0$ and $\boldsymbol{v} \neq \boldsymbol{0}$. Hence $N_S(\boldsymbol{x}) = \{\boldsymbol{0}\}$.

(c) Boundary point: $f(\boldsymbol{x}) = \beta$, $\nabla f(\boldsymbol{x}) \neq 0$.
First we show

$$\{\lambda \nabla f(\boldsymbol{x}) : \lambda \ge 0\} \subseteq N_S(\boldsymbol{x}).$$

Let $\boldsymbol{y} \in S$, so $f(\boldsymbol{y}) \le \beta = f(\boldsymbol{x})$. By convexity and differentiability,

$$f(\boldsymbol{y}) \ge f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}).$$

Hence

$$0 \ge f(\boldsymbol{y}) - f(\boldsymbol{x}) \ge \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}),$$

so $\nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \le 0$ for all $\boldsymbol{y} \in S$, which means $\nabla f(\boldsymbol{x}) \in N_S(\boldsymbol{x})$, and therefore $\lambda \nabla f(\boldsymbol{x}) \in N_S(\boldsymbol{x})$ for all $\lambda \ge 0$.
The converse inclusion

$$N_S(\boldsymbol{x}) \subseteq \{\lambda \nabla f(\boldsymbol{x}) : \lambda \ge 0\}$$

is more delicate and is omitted here.

Theorem

Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$, and define

$$\Omega = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}.$$

Then for every $\boldsymbol{x} \in \Omega$,

$$N_\Omega(\boldsymbol{x}) = \mathrm{Range}(\boldsymbol{A}^\top).$$

Example

Geometric intuition.
Fix a point $\boldsymbol{x} \in \Omega$. Any feasible direction $\boldsymbol{d}$ at $\boldsymbol{x}$ must satisfy

$$\boldsymbol{A}(\boldsymbol{x} + \boldsymbol{d}) = \boldsymbol{b} \quad \Longrightarrow \quad \boldsymbol{A}\boldsymbol{d} = \boldsymbol{0},$$

so $\boldsymbol{d} \in \mathrm{Null}(\boldsymbol{A})$. By the fundamental theorem of linear algebra,

$$\mathrm{Null}(\boldsymbol{A})^\perp = \mathrm{Range}(\boldsymbol{A}^\top).$$

On the other hand, the normal cone $N_\Omega(\boldsymbol{x})$ consists of all vectors orthogonal to every feasible direction $\boldsymbol{d}$, i.e.

$$N_\Omega(\boldsymbol{x}) = \mathrm{Null}(\boldsymbol{A})^\perp = \mathrm{Range}(\boldsymbol{A}^\top).$$

Proof

We prove both inclusions.

($\subseteq$ direction). We first show

$$\text{Range}(\boldsymbol{A}^\top) \subseteq N_\Omega(\boldsymbol{x}).$$

Let $\boldsymbol{p} \in \text{Range}(\boldsymbol{A}^\top)$, so $\boldsymbol{p} = \boldsymbol{A}^\top \boldsymbol{v}$ for some $\boldsymbol{v} \in \mathbb{R}^m$. For any $\boldsymbol{w} \in \Omega$,

$$\boldsymbol{p}^\top(\boldsymbol{w} - \boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{A}(\boldsymbol{w} - \boldsymbol{x}) = \boldsymbol{v}^\top(\boldsymbol{b} - \boldsymbol{b}) = 0.$$

Thus $\boldsymbol{p}^\top(\boldsymbol{w} - \boldsymbol{x}) \leq 0$ for all $\boldsymbol{w} \in \Omega$, so $\boldsymbol{p} \in N_\Omega(\boldsymbol{x})$.

($\supseteq$ direction). We next show

$$N_\Omega(\boldsymbol{x}) \subseteq \text{Range}(\boldsymbol{A}^\top).$$

Equivalently, we show

$$\mathbb{R}^n \setminus \text{Range}(\boldsymbol{A}^\top) \subseteq \mathbb{R}^n \setminus N_\Omega(\boldsymbol{x}).$$

Take $\boldsymbol{p} \in \mathbb{R}^n \setminus \text{Range}(\boldsymbol{A}^\top)$. By the fundamental theorem of linear algebra, we can write

$$\boldsymbol{p} = \boldsymbol{p}_1 + \boldsymbol{p}_2,$$

where

$$\boldsymbol{p}_1 \in \text{Range}(\boldsymbol{A}^\top), \quad \boldsymbol{p}_2 \in \text{Null}(\boldsymbol{A}), \quad \boldsymbol{p}_1^\top \boldsymbol{p}_2 = 0.$$

Since $\boldsymbol{p} \notin \text{Range}(\boldsymbol{A}^\top)$, we must have $\boldsymbol{p}_2 \neq \boldsymbol{0}$.
Note that $\boldsymbol{x} + \boldsymbol{p}_2 \in \Omega$ because

$$\boldsymbol{A}(\boldsymbol{x} + \boldsymbol{p}_2) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}\boldsymbol{p}_2 = \boldsymbol{b} + \boldsymbol{0} = \boldsymbol{b}.$$

Then

$$\boldsymbol{p}^\top((\boldsymbol{x} + \boldsymbol{p}_2) - \boldsymbol{x}) = \boldsymbol{p}^\top \boldsymbol{p}_2 = (\boldsymbol{p}_1 + \boldsymbol{p}_2)^\top \boldsymbol{p}_2 = \boldsymbol{p}_2^\top \boldsymbol{p}_2 > 0.$$

Hence $\boldsymbol{p}^\top(\boldsymbol{w} - \boldsymbol{x}) > 0$ for the feasible point $\boldsymbol{w} = \boldsymbol{x} + \boldsymbol{p}_2$, so $\boldsymbol{p} \notin N_\Omega(\boldsymbol{x})$. Thus $N_\Omega(\boldsymbol{x}) \subseteq \text{Range}(\boldsymbol{A}^\top)$.

14.2 Convex Programming

A **convex program (CP)** has the form

$$\min_{\boldsymbol{x}} f_0(\boldsymbol{x}) \quad \text{s.t.} \quad f_1(\boldsymbol{x}) \leq 0, \ \ldots, \ f_m(\boldsymbol{x}) \leq 0, \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

where $f_0, \ldots, f_m$ are convex functions.

Example

Examples of convex programming.

- Least squares (LS).
- ℓ_1-regularized least squares (ℓ_1LS).
- The SVM formulations SVM-3, SVM-4, SVM-6, SVM-7, SVM-8.

Let

$$\Omega = \{\boldsymbol{x} : f_1(\boldsymbol{x}) \le 0, \ldots, f_m(\boldsymbol{x}) \le 0,\ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\},$$

the **feasible region** of convex programming.

We say the problem is **smooth** if $f_0, \ldots, f_m$ are all differentiable on an open set containing Ω. (Examples: SVM-4, SVM-6, SVM-7.)

Recall SVM-3:

$$\max_{\boldsymbol{x},\xi,\boldsymbol{t}} \min_{i=1,\ldots,N} t_i \quad \text{s.t.} \quad \boldsymbol{a}_i^\top \boldsymbol{x} + \xi = t_i \ (y_i = 1), \quad \boldsymbol{a}_i^\top \boldsymbol{x} + \xi = -t_i \ (y_i = -1), \quad \|\boldsymbol{x}\| \le 1,\ \boldsymbol{t} \ge \boldsymbol{0}.$$

Equivalently,

$$\min_{\boldsymbol{x},\xi,\boldsymbol{t}} - \min_{i=1,\ldots,N} t_i \quad \text{s.t. the same constraints,}$$

or

$$\min_{\boldsymbol{x},\xi,\boldsymbol{t}} \max_{i=1,\ldots,N} (-t_i) \quad \text{s.t. the same constraints.}$$

Here the function $f(s,t) = \max(s,t)$ is convex but not differentiable when $s = t$.

14.2.1 KKT Conditions

Let $\widehat{\boldsymbol{x}} \in \Omega$ and assume

$$\nabla f_1(\widehat{\boldsymbol{x}}) \ne \boldsymbol{0}, \ldots, \nabla f_m(\widehat{\boldsymbol{x}}) \ne \boldsymbol{0},$$

and that a suitable constraint qualification (polyhedrality, Slater, or similar) holds.

Then:

$$\widehat{\boldsymbol{x}} \text{ is optimal for CP} \quad \Longleftrightarrow \quad -\nabla f_0(\widehat{\boldsymbol{x}}) \in N_\Omega(\widehat{\boldsymbol{x}}).$$

Using the normal cone decomposition,

$$N_\Omega(\widehat{\boldsymbol{x}}) = N_{\Omega_1}(\widehat{\boldsymbol{x}}) + \cdots + N_{\Omega_{m+1}}(\widehat{\boldsymbol{x}}),$$

where

$$\Omega_i := \{\boldsymbol{x} : f_i(\boldsymbol{x}) \le 0\},\ i = 1, \ldots, m, \quad \Omega_{m+1} := \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}.$$

Let

$$C := \{i \in \{1, \ldots, m\} : f_i(\widehat{\boldsymbol{x}}) = 0\}$$

be the set of **active constraints** at $\widehat{\boldsymbol{x}}$, and

$$I := \{1, \ldots, m\} \setminus C$$

the set of **inactive constraints**.

Then the optimality condition becomes

$$-\nabla f_0(\widehat{\boldsymbol{x}}) \in \sum_{i \in C} N_{\Omega_i}(\widehat{\boldsymbol{x}}) + N_{\Omega_{m+1}}(\widehat{\boldsymbol{x}}).$$

Using the previous theorem about level sets and the equality-constraint normal cone, this is equivalent to the existence of multipliers $\lambda_i \ge 0$ for $i \in C$ and $\boldsymbol{v} \in \mathbb{R}^m$ such that

$$-\nabla f_0(\widehat{\boldsymbol{x}}) = \sum_{i \in C} \lambda_i \nabla f_i(\widehat{\boldsymbol{x}}) + \boldsymbol{A}^\top \boldsymbol{v}.$$

These are the (primal) **KKT conditions** in this smooth convex setting.

Lec 15

15.1 KKT Conditions

Recall convex programming (CP):

$$\min_{\boldsymbol{x}} f_0(\boldsymbol{x}) \quad \text{s.t.} \quad f_i(\boldsymbol{x}) \le 0,\ i = 1, \dots, m, \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

with

$$\Omega := \{\boldsymbol{x} : f_1(\boldsymbol{x}) \le 0, \dots, f_m(\boldsymbol{x}) \le 0,\ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}.$$

We have the normal-cone optimality condition:

$$\begin{aligned} \hat{\boldsymbol{x}} \text{ is optimal for CP} &\iff -\nabla f_0(\hat{\boldsymbol{x}}) \in N_\Omega(\hat{\boldsymbol{x}}),\ \hat{\boldsymbol{x}} \in \Omega \\ &\iff -\nabla f_0(\hat{\boldsymbol{x}}) \in N_{\Omega_1}(\hat{\boldsymbol{x}}) + \cdots + N_{\Omega_{m+1}}(\hat{\boldsymbol{x}}),\ \hat{\boldsymbol{x}} \in \Omega, \end{aligned}$$

where

$$\Omega_i := \{\boldsymbol{x} : f_i(\boldsymbol{x}) \le 0\},\ i = 1, \dots, m, \qquad \Omega_{m+1} := \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\},$$

and we assume a suitable **constraint qualification** (CQ).

Let

$$C \subseteq \{1, \dots, m\}$$

be the set of **active constraints** at $\hat{\boldsymbol{x}}$:

$$i \in C \iff f_i(\hat{\boldsymbol{x}}) = 0.$$

Assume also

$$\nabla f_i(\hat{\boldsymbol{x}}) \neq \boldsymbol{0}, \quad i = 1, \dots, m.$$

Then

$$-\nabla f_0(\hat{\boldsymbol{x}}) \in \sum_{i \in C} N_{\Omega_i}(\hat{\boldsymbol{x}}) + N_{\Omega_{m+1}}(\hat{\boldsymbol{x}})$$

is equivalent to the existence of multipliers $\lambda_i \ge 0$ $(i \in C)$ and $\boldsymbol{v} \in \mathbb{R}^p$ such that

$$-\nabla f_0(\hat{\boldsymbol{x}}) = \sum_{i \in C} \lambda_i \nabla f_i(\hat{\boldsymbol{x}}) + \boldsymbol{A}^\top \boldsymbol{v}.$$

The condition above is called the **KKT condition** (Karush–Kuhn–Tucker) for convex programming.

15.2 Subdifferential

15.2.1 Definition

Recall the (gradient) subgradient inequality: if $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and differentiable at $\boldsymbol{x} \in \operatorname{dom}(f)$, then for all $\boldsymbol{y} \in \mathbb{R}^n$,

$$f(\boldsymbol{y}) \ge f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}).$$

For a nonsmooth convex function (e.g. $f(x) = \max(x, 0)$ at $x = 0$), the gradient may not exist but we can still have affine lower bounds.

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex and $\boldsymbol{x} \in \mathrm{dom}(f)$. A vector $\boldsymbol{g} \in \mathbb{R}^n$ is called a **subgradient** of f at $\boldsymbol{x}$ if

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^\top(\boldsymbol{y} - \boldsymbol{x}), \quad \forall\, \boldsymbol{y} \in \mathbb{R}^n.$$

The set of all subgradients at $\boldsymbol{x}$ is the **subdifferential** of f at $\boldsymbol{x}$, denoted

$$\partial f(\boldsymbol{x}).$$

15.2.2 Relative Interior

For a convex set $C \subseteq \mathbb{R}^n$, the (usual) interior is

$$\mathrm{int}(C) = \{\boldsymbol{x} \in C : \exists r > 0 \text{ s.t. } \mathbb{B}(\boldsymbol{x}, r) \subseteq C\}.$$

Given $S \subseteq \mathbb{R}^n$, its **affine hull** is

$$\mathrm{aff}(S) = \left\{ \boldsymbol{z} : \exists\, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_m \in S,\ \exists\, \lambda_1, \ldots, \lambda_m \text{ with } \sum_{i=1}^m \lambda_i = 1,\ \boldsymbol{z} = \sum_{i=1}^m \lambda_i \boldsymbol{s}_i \right\}.$$

Moreover, there exist $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$ such that

$$\mathrm{aff}(S) = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}.$$

For a convex set $S \subseteq \mathbb{R}^n$, the **relative interior** is

$$\mathrm{relint}(S) = \{\boldsymbol{x} \in S : \exists r > 0 \text{ s.t. } \mathbb{B}(\boldsymbol{x}, r) \cap \mathrm{aff}(S) \subseteq S\}.$$

Intuitively, this is the interior of S relative to the (possibly lower-dimensional) affine space $\mathrm{aff}(S)$.

For any nonempty convex set $S \subseteq \mathbb{R}^n$,

$$\mathrm{relint}(S) \neq \varnothing.$$

Fact

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex. Then:

(i) For all $\boldsymbol{x} \in \mathrm{dom}(f)$, the subdifferential $\partial f(\boldsymbol{x})$ is a closed, convex set.

(ii) For all $\boldsymbol{x} \in \mathrm{relint}(\mathrm{dom}(f))$, the subdifferential is nonempty:

$$\partial f(\boldsymbol{x}) \neq \varnothing.$$

(iii) If $\boldsymbol{x} \in \mathrm{int}(\mathrm{dom}(f))$, then $\partial f(\boldsymbol{x})$ is bounded; hence it is convex, compact, and nonempty.

(iv) If $\boldsymbol{x} \in \mathrm{int}(\mathrm{dom}(f))$, then f is differentiable at $\boldsymbol{x}$ if and only if

$$\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}.$$

Example

Why use relative interior?
Consider

$$f(x) = \begin{cases} -\sqrt{1-x^2}, & x \in [-1,1], \\ \infty, & \text{otherwise.} \end{cases}$$

Then $\partial f(x) = \varnothing$ at the endpoints $x = \pm 1$, even though f is convex. These points lie on the boundary of $\mathrm{dom}(f)$, outside the relative interior.

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex and $\boldsymbol{x} \in \mathrm{dom}(f)$. Then

$$\boldsymbol{x} \in \arg\min f \quad \Longleftrightarrow \quad \mathbf{0} \in \partial f(\boldsymbol{x}).$$

Proof

We have

$$\boldsymbol{x} \text{ is a global minimizer} \quad \Longleftrightarrow \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}),\ \forall \boldsymbol{y} \in \mathbb{R}^n.$$

This is equivalent to

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \mathbf{0}^\top (\boldsymbol{y} - \boldsymbol{x}), \quad \forall \boldsymbol{y} \in \mathbb{R}^n,$$

which is exactly the condition $\mathbf{0} \in \partial f(\boldsymbol{x})$.

Terminology. A convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is called *proper* if $\mathrm{dom}(f) \neq \varnothing$.

15.3 Epigraphs

For $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, the **epigraph** of f is

$$\mathrm{epi}(f) = \{(\boldsymbol{x}, y) \in \mathbb{R}^n \times \mathbb{R} : y \geq f(\boldsymbol{x})\}.$$

We say that f is **closed** if $\mathrm{epi}(f)$ is a closed set.

Example

Let

$$f(x) = \begin{cases} -\ln x, & x > 0, \\ \infty, & x \leq 0. \end{cases}$$

Here $\mathrm{dom}(f) = (0, \infty)$ is not closed, but $\mathrm{epi}(f)$ is closed. Conversely, a closed domain does not guarantee a closed epigraph.

Fact

A proper convex function with closed epigraph is often called **lower semicontinuous (lsc)** or a *closed convex function*. If $f : \mathbb{R}^n \to \mathbb{R}$ is (finite-valued) convex, then f is automatically

proper and closed.

Theorem

A convex function f is closed if and only if each level set

$$\{\boldsymbol{x} : f(\boldsymbol{x}) \leq \beta\}, \quad \beta \in \mathbb{R},$$

is a closed set.

Theorem

If $f, g : \mathbb{R}^n \to \mathbb{R}$ are closed (finite-valued) convex functions, then their sum $f + g$ is also closed.

Theorem

Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex, closed, and proper. Assume

$$\operatorname{relint}(\operatorname{dom}(f_1)) \cap \operatorname{relint}(\operatorname{dom}(f_2)) \neq \varnothing.$$

Then for all $\boldsymbol{x} \in \operatorname{dom}(f_1) \cap \operatorname{dom}(f_2)$,

$$\partial f_1(\boldsymbol{x}) + \partial f_2(\boldsymbol{x}) = \partial(f_1 + f_2)(\boldsymbol{x}),$$

where $+$ denotes the Minkowski sum of sets.

Fact

The relative-interior intersection condition above acts as a type of **constraint qualification** for subdifferentials of sums.

15.4 Indicator Function

Example

Indicator of an interval.
Let

$$f(x) = \begin{cases} 0, & x \in [0,1], \\ \infty, & x \notin [0,1]. \end{cases}$$

This is the **indicator function** of $[0,1]$, denoted $\mathbb{I}_{[0,1]}(x)$.

Claim

For the interval $[0,1]$,

$$\partial \mathbb{I}_{[0,1]}(1) = [0, \infty).$$

Theorem

Let $\Omega \subseteq \mathbb{R}^n$ be nonempty, closed, and convex. The indicator function

$$\mathbb{I}_\Omega(\boldsymbol{x}) = \begin{cases} 0, & \boldsymbol{x} \in \Omega, \\ \infty, & \boldsymbol{x} \notin \Omega, \end{cases}$$

is proper, closed, and convex. Moreover, for all $\boldsymbol{x} \in \Omega$,

$$\partial \mathbb{I}_\Omega(\boldsymbol{x}) = N_\Omega(\boldsymbol{x}),$$

the normal cone of Ω at $\boldsymbol{x}$.

Proof

Proper: Since $\Omega \neq \varnothing$, we have $\operatorname{dom}(\mathbb{I}_\Omega) = \Omega \neq \varnothing$.

Convex: For any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$ and $\lambda \in [0,1]$, convexity follows by checking cases depending on whether $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ lie in Ω.

Closed: The epigraph is

$$\operatorname{epi}(\mathbb{I}_\Omega) = \{(\boldsymbol{x}, y) : \boldsymbol{x} \in \Omega,\ y \geq 0\} = \Omega \times [0, \infty),$$

which is closed because Ω is closed.

Subdifferential equals normal cone.
First inclusion: $\partial \mathbb{I}_\Omega(\boldsymbol{x}) \subseteq N_\Omega(\boldsymbol{x})$.
Let $\boldsymbol{g} \in \partial \mathbb{I}_\Omega(\boldsymbol{x})$. Then, by definition,

$$\mathbb{I}_\Omega(\boldsymbol{y}) \geq \mathbb{I}_\Omega(\boldsymbol{x}) + \boldsymbol{g}^\top(\boldsymbol{y} - \boldsymbol{x}), \quad \forall\, \boldsymbol{y} \in \mathbb{R}^n.$$

For $\boldsymbol{y} \in \Omega$, both sides are finite and

$$0 = \mathbb{I}_\Omega(\boldsymbol{y}) \geq 0 + \boldsymbol{g}^\top(\boldsymbol{y} - \boldsymbol{x}),$$

so $\boldsymbol{g}^\top(\boldsymbol{y} - \boldsymbol{x}) \leq 0$ for all $\boldsymbol{y} \in \Omega$, which means $\boldsymbol{g} \in N_\Omega(\boldsymbol{x})$.

Second inclusion: $N_\Omega(\boldsymbol{x}) \subseteq \partial \mathbb{I}_\Omega(\boldsymbol{x})$.
Take $\boldsymbol{v} \in N_\Omega(\boldsymbol{x})$, so

$$\boldsymbol{v}^\top(\boldsymbol{y} - \boldsymbol{x}) \leq 0, \quad \forall\, \boldsymbol{y} \in \Omega.$$

For $\boldsymbol{y} \in \Omega$,

$$\mathbb{I}_\Omega(\boldsymbol{y}) = 0 \geq 0 + \boldsymbol{v}^\top(\boldsymbol{y} - \boldsymbol{x}) = \mathbb{I}_\Omega(\boldsymbol{x}) + \boldsymbol{v}^\top(\boldsymbol{y} - \boldsymbol{x}).$$

For $\boldsymbol{y} \notin \Omega$, we have $\mathbb{I}_\Omega(\boldsymbol{y}) = \infty$, so the inequality

$$\mathbb{I}_\Omega(\boldsymbol{y}) \geq \mathbb{I}_\Omega(\boldsymbol{x}) + \boldsymbol{v}^\top(\boldsymbol{y} - \boldsymbol{x})$$

holds trivially. Thus $\boldsymbol{v} \in \partial \mathbb{I}_\Omega(\boldsymbol{x})$.
Combining both inclusions gives $\partial \mathbb{I}_\Omega(\boldsymbol{x}) = N_\Omega(\boldsymbol{x})$.

15.5 Rederiving the Normal-Cone Optimality Condition

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be differentiable on a neighborhood of Ω, where $\Omega \subseteq \mathbb{R}^n$ is closed and convex. Then

$$\boldsymbol{x} \in \arg\min\{f(\boldsymbol{x}) : \boldsymbol{x} \in \Omega\}$$

if and only if

$$-\nabla f(\boldsymbol{x}) \in N_\Omega(\boldsymbol{x}).$$

Proof

We can rewrite the constrained problem as unconstrained:

$$\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x}) \quad \Longleftrightarrow \quad \min_{\boldsymbol{x}} \big(f(\boldsymbol{x}) + \mathbb{I}_\Omega(\boldsymbol{x})\big).$$

Thus

$$\boldsymbol{x} \in \arg\min\{f(\boldsymbol{x}) : \boldsymbol{x} \in \Omega\} \quad \Longleftrightarrow \quad \boldsymbol{x} \in \arg\min\{f(\boldsymbol{x}) + \mathbb{I}_\Omega(\boldsymbol{x})\}.$$

By the subgradient optimality condition,

$$\boldsymbol{x} \text{ minimizer} \quad \Longleftrightarrow \quad \mathbf{0} \in \partial(f + \mathbb{I}_\Omega)(\boldsymbol{x}).$$

Assuming

$$\operatorname{relint}(\operatorname{dom} f) \cap \operatorname{relint}(\operatorname{dom} \mathbb{I}_\Omega) \neq \varnothing,$$

we can apply the subdifferential sum rule:

$$\partial(f + \mathbb{I}_\Omega)(\boldsymbol{x}) = \partial f(\boldsymbol{x}) + \partial \mathbb{I}_\Omega(\boldsymbol{x}).$$

Since f is differentiable at $\boldsymbol{x}$,

$$\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\},$$

and by the indicator theorem,

$$\partial \mathbb{I}_\Omega(\boldsymbol{x}) = N_\Omega(\boldsymbol{x}).$$

Thus

$$\mathbf{0} \in \partial(f + \mathbb{I}_\Omega)(\boldsymbol{x}) \quad \Longleftrightarrow \quad \mathbf{0} \in \{\nabla f(\boldsymbol{x})\} + N_\Omega(\boldsymbol{x}) \quad \Longleftrightarrow \quad -\nabla f(\boldsymbol{x}) \in N_\Omega(\boldsymbol{x}).$$

Lec 16

16.1 Composite Minimization

We consider the problem

$$\min_{\boldsymbol{x}} F(\boldsymbol{x}), \qquad F(\boldsymbol{x}) := f(\boldsymbol{x}) + \psi(\boldsymbol{x}),$$

where

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-smooth;
- $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is proper, closed, and convex.

Example

- SVM-8 (hinge loss + quadratic regularizer).
- ℓ_1-regularized least squares:

$$\ell_1\text{LS}: \quad \min_{\boldsymbol{x}} \underbrace{\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2}_{f(\boldsymbol{x})} + \underbrace{\gamma\|\boldsymbol{x}\|_1}_{\psi(\boldsymbol{x})}.$$

Theorem

Optimality condition for composite minimization.
For $F(\boldsymbol{x}) = f(\boldsymbol{x}) + \psi(\boldsymbol{x})$ as above,

$$\boldsymbol{x} \in \arg\min F \quad \iff \quad -\nabla f(\boldsymbol{x}) \in \partial\psi(\boldsymbol{x}).$$

Proof

Since $F = f + \psi$ with f differentiable,

$$\partial F(\boldsymbol{x}) = \partial f(\boldsymbol{x}) + \partial\psi(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\} + \partial\psi(\boldsymbol{x}).$$

By the subgradient optimality condition,

$$\boldsymbol{x} \in \arg\min F \quad \iff \quad \mathbf{0} \in \partial F(\boldsymbol{x}) \quad \iff \quad \mathbf{0} \in \{\nabla f(\boldsymbol{x})\} + \partial\psi(\boldsymbol{x}),$$

which is equivalent to $-\nabla f(\boldsymbol{x}) \in \partial\psi(\boldsymbol{x})$.

Theorem

Uniqueness under strong convexity.
Suppose, in addition, that f is m-strongly convex with $m > 0$. Then $F = f + \psi$ has a unique minimizer.

Proof

A sum of a strongly convex function and a convex function is strongly convex. Thus F is strongly convex, so it is strictly convex and hence has at most one minimizer. Since F is proper, closed, and coercive (cf. the level-set argument below), a minimizer exists and must be unique.

Claim

For any $\boldsymbol{x}_0 \in \operatorname{dom}(\psi)$, the level set

$$S := \{\boldsymbol{x} \in \mathbb{R}^n : F(\boldsymbol{x}) \le F(\boldsymbol{x}_0)\}$$

is closed.

Proof

Both f and ψ are closed, so their sum $F = f + \psi$ is closed. For a closed function, all sublevel sets $\{\boldsymbol{x} : F(\boldsymbol{x}) \le \beta\}$ are closed. Taking $\beta = F(\boldsymbol{x}_0)$ gives that S is closed.

Claim

Choose $\boldsymbol{x}_0 \in \operatorname{ri}(\operatorname{dom}(\psi))$ and let S be the level set above. Then S is bounded.

Proof

Because $\boldsymbol{x}_0 \in \operatorname{ri}(\operatorname{dom}(\psi))$ and ψ is convex, we have $\partial\psi(\boldsymbol{x}_0) \neq \varnothing$. Choose some $\boldsymbol{g} \in \partial\psi(\boldsymbol{x}_0)$. Then for all $\boldsymbol{x}$,

$$\psi(\boldsymbol{x}) \ge \psi(\boldsymbol{x}_0) + \boldsymbol{g}^\top(\boldsymbol{x} - \boldsymbol{x}_0).$$

Since f is m-strongly convex,

$$f(\boldsymbol{x}) \ge f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^\top(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|^2.$$

Adding gives

$$F(\boldsymbol{x}) \ge F(\boldsymbol{x}_0) + (\nabla f(\boldsymbol{x}_0) + \boldsymbol{g})^\top(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|^2.$$

For $\boldsymbol{x} \in S$, we have $F(\boldsymbol{x}) - F(\boldsymbol{x}_0) \le 0$, hence

$$0 \ge F(\boldsymbol{x}) - F(\boldsymbol{x}_0) \ge (\nabla f(\boldsymbol{x}_0) + \boldsymbol{g})^\top(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|^2.$$

Applying Cauchy–Schwarz,

$$(\nabla f(\boldsymbol{x}_0) + \boldsymbol{g})^\top(\boldsymbol{x} - \boldsymbol{x}_0) \ge -\|\nabla f(\boldsymbol{x}_0) + \boldsymbol{g}\|\,\|\boldsymbol{x} - \boldsymbol{x}_0\|.$$

Thus

$$0 \ge -\|\nabla f(\boldsymbol{x}_0) + \boldsymbol{g}\|\,\|\boldsymbol{x} - \boldsymbol{x}_0\| + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|^2.$$

Rearranging (and assuming $\boldsymbol{x} \neq \boldsymbol{x}_0$ so the norm is nonzero) gives

$$\frac{m}{2}\,\|\boldsymbol{x} - \boldsymbol{x}_0\| \le \|\nabla f(\boldsymbol{x}_0) + \boldsymbol{g}\|.$$

Hence $\|\boldsymbol{x} - \boldsymbol{x}_0\|$ is uniformly bounded for all $\boldsymbol{x} \in S$. Therefore S is bounded.

Claim

F is bounded below on S and attains a minimum over S (and hence over $\mathbb{R}^n$).

Proof

For any $\boldsymbol{x} \in S$, using the same inequality as above but dropping the quadratic term yields

$$F(\boldsymbol{x}) \geq F(\boldsymbol{x}_0) + (\nabla f(\boldsymbol{x}_0) + \boldsymbol{g})^\top (\boldsymbol{x} - \boldsymbol{x}_0) \geq F(\boldsymbol{x}_0) - \|\nabla f(\boldsymbol{x}_0) + \boldsymbol{g}\| \, \|\boldsymbol{x} - \boldsymbol{x}_0\|.$$

Using the bound on $\|\boldsymbol{x} - \boldsymbol{x}_0\|$ from Claim 2, we obtain a uniform lower bound:

$$F(\boldsymbol{x}) \geq F_{\min} \quad \text{for all } \boldsymbol{x} \in S$$

for some finite $F_{\min}$.
Since S is closed and bounded, it is compact. The epigraph

$$\operatorname{epi}(F) = \{(\boldsymbol{x}, y) : y \geq F(\boldsymbol{x})\}$$

is closed, so

$$\operatorname{epi}(F) \cap (S \times [F_{\min}, \infty))$$

is a nonempty compact set. The projection map $\phi(\boldsymbol{x}, y) = y$ attains its minimum on this set at some $(\boldsymbol{x}^*, F(\boldsymbol{x}^*))$. Thus F attains its minimum on S at $\boldsymbol{x}^*$.
For any $\boldsymbol{x} \notin S$, we have $F(\boldsymbol{x}) > F(\boldsymbol{x}_0) \geq F(\boldsymbol{x}^*)$, so $\boldsymbol{x}^*$ is in fact a global minimizer of F over $\mathbb{R}^n$.

16.2 Proximal Operator

Let $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be proper, closed, and convex. The **proximal operator** of ψ is defined by

$$\operatorname{prox}_\psi(\boldsymbol{x}) := \arg\min_{\boldsymbol{z}} \Big\{ \psi(\boldsymbol{z}) + \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2 \Big\}.$$

Intuitively, $\operatorname{prox}_\psi(\boldsymbol{x})$ finds a point $\boldsymbol{z}$ that balances two goals: staying close to the current point $\boldsymbol{x}$, and having a small value of the regularizer ψ.

Theorem

Existence and uniqueness of the proximal point.
For each $\boldsymbol{x} \in \mathbb{R}^n$, the function

$$\boldsymbol{z} \mapsto \psi(\boldsymbol{z}) + \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2$$

has a unique minimizer, i.e. $\operatorname{prox}_\psi(\boldsymbol{x})$ is well-defined and single-valued.

Proof

The function $\boldsymbol{z} \mapsto \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2$ is strongly convex. Adding the convex function ψ preserves strong convexity, so $\psi(\boldsymbol{z}) + \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2$ is strongly convex and proper, closed. Hence it has a unique minimizer.

We say that ψ is *proximable* if there exists an efficient algorithm to compute $\text{prox}_{\psi}(\boldsymbol{x})$.

Example

Example 1: Scalar ℓ_1-prox.
Let $n = 1$ and

$$\psi(x) = t|x|, \quad t > 0.$$

Define

$$q(z) := t|z| + \frac{1}{2}(z - x)^2.$$

Then $\text{prox}_{t|\cdot|}(x)$ is the unique minimizer of q. We have

$$\partial q(z) = t\,\partial|z| + (z - x).$$

Recall

$$\partial|z| = \begin{cases} \{-1\}, & z < 0, \\ [-1, 1], & z = 0, \\ \{1\}, & z > 0. \end{cases}$$

The optimality condition

$$0 \in \partial q(z) \quad \Longleftrightarrow \quad 0 \in t\,\partial|z| + z - x$$

leads to three cases:

$$x - z \in \begin{cases} \{-t\}, & z < 0, \\ [-t, t], & z = 0, \\ \{t\}, & z > 0. \end{cases}$$

- **Case $z < 0$:** $x - z = -t \Rightarrow z = x + t$. Validity requires $z < 0 \Rightarrow x < -t$.
- **Case $z = 0$:** $x - z \in [-t, t] \Rightarrow x \in [-t, t]$.
- **Case $z > 0$:** $x - z = t \Rightarrow z = x - t$. Validity requires $z > 0 \Rightarrow x > t$.

Summarizing,

$$\text{prox}_{t|\cdot|}(x) = \begin{cases} x + t, & x < -t, \\ 0, & x \in [-t, t], \\ x - t, & x > t. \end{cases}$$

This is the usual scalar *soft-thresholding* operator.

Example

Example 2: Vector ℓ_1-prox.
For $\boldsymbol{x} \in \mathbb{R}^n$ and $t > 0$,

$$\text{prox}_{t\|\cdot\|_1}(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \{t\|\boldsymbol{z}\|_1 + \tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2\}.$$

Because the function is separable across coordinates, the solution is obtained by applying the

scalar soft thresholding coordinatewise:

$$\operatorname{prox}_{t\|\cdot\|_1}(\boldsymbol{x}) = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \quad \text{with} \quad u_i = \begin{cases} x_i + t, & x_i < -t, \\ 0, & x_i \in [-t, t], \\ x_i - t, & x_i > t, \end{cases} \quad i = 1, \ldots, n.$$

Example

Example 3: Prox of the Euclidean norm.
Let

$$\psi(\boldsymbol{x}) = t\|\boldsymbol{x}\|, \quad t > 0.$$

We know

$$\partial\|\boldsymbol{z}\| = \begin{cases} \left\{ \dfrac{\boldsymbol{z}}{\|\boldsymbol{z}\|} \right\}, & \boldsymbol{z} \neq \boldsymbol{0}, \\ \overline{\mathbb{B}}(\boldsymbol{0}, 1), & \boldsymbol{z} = \boldsymbol{0}, \end{cases}$$

where $\overline{\mathbb{B}}(\boldsymbol{0}, 1)$ is the closed unit ball.
Consider

$$q(\boldsymbol{z}) := t\|\boldsymbol{z}\| + \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2.$$

Then

$$\boldsymbol{0} \in \partial q(\boldsymbol{z}) \quad \Longleftrightarrow \quad \boldsymbol{0} \in t\,\partial\|\boldsymbol{z}\| + (\boldsymbol{z} - \boldsymbol{x}),$$

i.e.

$$\boldsymbol{x} - \boldsymbol{z} \in \begin{cases} \overline{\mathbb{B}}(\boldsymbol{0}, t), & \boldsymbol{z} = \boldsymbol{0} \quad \text{(a)}, \\ t\,\dfrac{\boldsymbol{z}}{\|\boldsymbol{z}\|}, & \boldsymbol{z} \neq \boldsymbol{0} \quad \text{(b)}. \end{cases}$$

Case (a): $\boldsymbol{z} = \boldsymbol{0}$ is valid if and only if $\boldsymbol{x} \in \overline{\mathbb{B}}(\boldsymbol{0}, t)$, i.e. $\|\boldsymbol{x}\| \leq t$.
Case (b): Assume $\boldsymbol{z} \neq \boldsymbol{0}$ and write $\boldsymbol{z} = \theta\boldsymbol{x}$ for some scalar θ. Then

$$\boldsymbol{x} - \theta\boldsymbol{x} = t\,\frac{\theta\boldsymbol{x}}{\|\theta\boldsymbol{x}\|} = t\,\frac{\theta}{|\theta|}\,\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}.$$

Assuming $\theta > 0$ (we expect $\boldsymbol{z}$ to be in the same direction as $\boldsymbol{x}$), we get

$$1 - \theta = \frac{t}{\|\boldsymbol{x}\|} \quad \Rightarrow \quad \theta = 1 - \frac{t}{\|\boldsymbol{x}\|}.$$

Thus

$$\boldsymbol{z} = \left(1 - \frac{t}{\|\boldsymbol{x}\|}\right)\boldsymbol{x},$$

which is valid only when $\boldsymbol{z} \neq \boldsymbol{0}$, i.e. $\|\boldsymbol{x}\| > t$.
Combining both cases:

$$\operatorname{prox}_{t\|\cdot\|}(\boldsymbol{x}) = \begin{cases} \boldsymbol{0}, & \|\boldsymbol{x}\| \leq t, \\ \left(1 - \dfrac{t}{\|\boldsymbol{x}\|}\right)\boldsymbol{x}, & \|\boldsymbol{x}\| > t. \end{cases}$$

This is often called the *vector soft-thresholding* or *shrinkage* operator.

Fact

For any closed, proper, convex function ψ, the proximal map $\operatorname{prox}_\psi : \mathbb{R}^n \to \mathbb{R}^n$ is 1-Lipschitz, i.e.

$$\|\operatorname{prox}_\psi(\boldsymbol{x}) - \operatorname{prox}_\psi(\boldsymbol{y})\| \le \|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

In particular, the proximal map is continuous.

16.3 Proximal Gradient

We return to the composite setup

$$\min_{\boldsymbol{x}} F(\boldsymbol{x}), \qquad F(\boldsymbol{x}) = f(\boldsymbol{x}) + \psi(\boldsymbol{x}),$$

with f convex and L-smooth and ψ proper, closed, convex.

Proximal gradient.
The *proximal gradient* of F (with parameter L) is

$$\mathrm{G}_L(\boldsymbol{x}) := L\left(\boldsymbol{x} - \operatorname{prox}_{\psi/L}\left(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})\right)\right).$$

Intuitively, the mapping

$$\boldsymbol{x} \mapsto \operatorname{prox}_{\psi/L}\left(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})\right)$$

performs one gradient step on the smooth part f followed by one prox step on the (possibly nonsmooth) part ψ. The vector $\mathrm{G}_L(\boldsymbol{x})$ is the scaled difference between the current point and this update, and plays the role of a generalized gradient.

Example

Special case $\psi \equiv 0$.
If $\psi \equiv 0$, then

$$\operatorname{prox}_0(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2 = \boldsymbol{x}.$$

Hence

$$\mathrm{G}_L(\boldsymbol{x}) = L\left(\boldsymbol{x} - \left(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})\right)\right) = \nabla f(\boldsymbol{x}),$$

so the proximal gradient reduces to the usual gradient.

Lec 17

17.1 Proximal Gradient

17.1.1 Set-up

Recall the composite problem

$$\min_{\boldsymbol{x}} F(\boldsymbol{x}), \qquad F(\boldsymbol{x}) = f(\boldsymbol{x}) + \psi(\boldsymbol{x}),$$

where

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-smooth,
- $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex, closed, proper, and proximable.

The *proximal gradient* is defined as

$$\mathrm{G}_L(\boldsymbol{x}) := L\left(\boldsymbol{x} - \mathrm{prox}_{\psi/L}\left(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})\right)\right).$$

Theorem

For F and G_L as above,

$$\boldsymbol{x} \in \arg\min F \iff \mathrm{G}_L(\boldsymbol{x}) = \mathbf{0}.$$

Proof

$$\begin{aligned}
\mathrm{G}_L(\boldsymbol{x}) = \mathbf{0} &\iff \boldsymbol{x} = \mathrm{prox}_{\psi/L}\left(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})\right) \\
&\iff \boldsymbol{x} = \arg\min_{\boldsymbol{z}} \left\{\frac{1}{L}\psi(\boldsymbol{z}) + \frac{1}{2}\left\|\boldsymbol{z} - \boldsymbol{x} + \frac{1}{L}\nabla f(\boldsymbol{x})\right\|^2\right\} \\
&\iff \mathbf{0} \in \left[\frac{1}{L}\partial\psi(\boldsymbol{z}) + \left\{\boldsymbol{z} - \boldsymbol{x} + \tfrac{1}{L}\nabla f(\boldsymbol{x})\right\}\right]_{\boldsymbol{z}=\boldsymbol{x}} \\
&\iff \mathbf{0} \in \frac{1}{L}\partial\psi(\boldsymbol{x}) + \frac{1}{L}\{\nabla f(\boldsymbol{x})\} \\
&\iff \mathbf{0} \in \frac{1}{L}(\partial\psi(\boldsymbol{x}) + \{\nabla f(\boldsymbol{x})\}) \\
&\iff \mathbf{0} \in \partial\psi(\boldsymbol{x}) + \{\nabla f(\boldsymbol{x})\} = \partial F(\boldsymbol{x}) \\
&\iff \boldsymbol{x} \in \arg\min F.
\end{aligned}$$

17.1.2 Proximal Gradient Descent

The **Proximal Gradient Descent (*PGD*)** algorithm is

$$\begin{aligned}
\boldsymbol{x}^0 &= \text{arbitrary}, \\
\boldsymbol{x}^{k+1} &= \boldsymbol{x}^k - \frac{1}{L}\mathrm{G}_L(\boldsymbol{x}^k).
\end{aligned}$$

Under our assumptions, this iteration converges to a minimizer of F (assuming a minimizer exists). The proof is omitted in class.

Accelerated Proximal Gradient Descent (*APGD*).

- Nesterov (2004),
- Beck & Teboulle (2013): **FISTA**.

APGD is obtained from *AGD*by replacing the gradient with the proximal gradient operator.

- *AGD*update:
$$\boldsymbol{x}^{k+1} = \boldsymbol{y}^k - \frac{1}{L}\nabla f(\boldsymbol{y}^k).$$
- *APGD* update:
$$\boldsymbol{x}^{k+1} = \boldsymbol{y}^k - \frac{1}{L}\mathrm{G}_L(\boldsymbol{y}^k).$$

More explicitly,

$$\begin{aligned}
\boldsymbol{x}^{k+1} &= \boldsymbol{y}^k - \frac{1}{L}\mathrm{G}_L(\boldsymbol{y}^k) \\
&= \boldsymbol{y}^k - \frac{1}{L}\left[L\left(\boldsymbol{y}^k - \mathrm{prox}_{\psi/L}\left(\boldsymbol{y}^k - \frac{1}{L}\nabla f(\boldsymbol{y}^k)\right)\right)\right] \\
&= \boldsymbol{y}^k - \boldsymbol{y}^k + \mathrm{prox}_{\psi/L}\left(\boldsymbol{y}^k - \frac{1}{L}\nabla f(\boldsymbol{y}^k)\right) \\
&= \mathrm{prox}_{\psi/L}\left(\boldsymbol{y}^k - \frac{1}{L}\nabla f(\boldsymbol{y}^k)\right).
\end{aligned}$$

This view is called **operator splitting**, or a **forward–backward step**:

- forward step: $\boldsymbol{y}^k - \frac{1}{L}\nabla f(\boldsymbol{y}^k)$ (gradient step on f),
- backward step: $\mathrm{prox}_{\psi/L}(\cdot)$ (prox step on ψ).

The terminology comes from an analogy with forward/backward Euler methods in ODEs.

Theorem

Convergence rates for *PGD* & *APGD*.
Let $F = f + \psi$ with f convex and L-smooth, and ψ convex, closed, proper.

- ***PGD* (sublinear rate).** For a minimizer $\boldsymbol{x}^*$ of F,
$$F(\boldsymbol{x}^k) - F(\boldsymbol{x}^*) \;\le\; \frac{L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{2k}.$$
- If f is additionally m-strongly convex with $m > 0$ (so $\boldsymbol{x}^*$ is unique), then *PGD* has a **linear** rate:
$$\|\boldsymbol{x}^k - \boldsymbol{x}^*\| \;\le\; \left(\frac{L-m}{L+m}\right)^k \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|.$$
- ***APGD* (accelerated rate).** Without strong convexity, *APGD* attains
$$F(\boldsymbol{x}^k) - F(\boldsymbol{x}^*) = \mathcal{O}\left(\frac{1}{k^2}\right).$$

- If f is also m-strongly convex, *APGD* enjoys a linear rate with factor

$$\left(\frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}\right)^k.$$

For details, see Beck, *First-Order Methods in Optimization* (2017).

A common termination criterion for *PGD* / *APGD* is

$$\|\mathrm{G}_L(\boldsymbol{x}^k)\| \le \text{tol}.$$

Alternative view of one *PGD* step.

We can write

$$\begin{aligned}\boldsymbol{x}^{k+1} &= \mathrm{prox}_{\psi/L}\left(\boldsymbol{x}^k - \frac{1}{L}\nabla f(\boldsymbol{x}^k)\right)\\ &= \arg\min_{\boldsymbol{z}}\left\{\frac{1}{L}\psi(\boldsymbol{z}) + \frac{1}{2}\left\|\boldsymbol{x}^k - \frac{1}{L}\nabla f(\boldsymbol{x}^k) - \boldsymbol{z}\right\|^2\right\}.\end{aligned}$$

Expanding the square,

$$\begin{aligned}\boldsymbol{x}^{k+1} &= \arg\min_{\boldsymbol{z}}\left\{\frac{1}{L}\psi(\boldsymbol{z}) + \frac{1}{2}\|\boldsymbol{x}^k - \boldsymbol{z}\|^2 - \frac{1}{L}\nabla f(\boldsymbol{x}^k)^\top(\boldsymbol{x}^k - \boldsymbol{z}) + \frac{1}{2L^2}\|\nabla f(\boldsymbol{x}^k)\|^2\right\}\\ &= \arg\min_{\boldsymbol{z}}\left\{\psi(\boldsymbol{z}) + \frac{L}{2}\|\boldsymbol{x}^k - \boldsymbol{z}\|^2 - \nabla f(\boldsymbol{x}^k)^\top(\boldsymbol{x}^k - \boldsymbol{z})\right\}\\ &= \arg\min_{\boldsymbol{z}}\left\{\psi(\boldsymbol{z}) + \underbrace{f(\boldsymbol{x}^k) + \nabla f(\boldsymbol{x}^k)^\top(\boldsymbol{z} - \boldsymbol{x}^k) + \frac{L}{2}\|\boldsymbol{z} - \boldsymbol{x}^k\|^2}_{\tilde{f}(\boldsymbol{z})}\right\}.\end{aligned}$$

Here $\tilde{f}(\boldsymbol{z})$ is a quadratic upper model of f at $\boldsymbol{x}^k$: we replace $f(\boldsymbol{z})$ by its first-order Taylor expansion at $\boldsymbol{x}^k$ plus a quadratic term $\frac{L}{2}\|\boldsymbol{z} - \boldsymbol{x}^k\|^2$.

Theorem

For all $\boldsymbol{z} \in \mathbb{R}^n$,

$$\tilde{f}(\boldsymbol{z}) \;\ge\; f(\boldsymbol{z}).$$

This inequality is simply the L-smoothness (descent lemma) of f.

Why not use the *true* Hessian $\nabla^2 f(\boldsymbol{x}^k)$ as the quadratic term? In general, the resulting subproblem

$$\min_{\boldsymbol{z}} \psi(\boldsymbol{z}) + f(\boldsymbol{x}^k) + \nabla f(\boldsymbol{x}^k)^\top(\boldsymbol{z} - \boldsymbol{x}^k) + \tfrac{1}{2}(\boldsymbol{z} - \boldsymbol{x}^k)^\top\nabla^2 f(\boldsymbol{x}^k)(\boldsymbol{z} - \boldsymbol{x}^k)$$

does not admit an efficient closed-form solution, whereas the $\frac{L}{2}\|\boldsymbol{z} - \boldsymbol{x}^k\|^2$ model does for many popular regularizers ψ.

17.2 Projected Gradient

17.2.1 Definition

Projected Gradient is the special case of *PGD* where we enforce a simple constraint $\boldsymbol{x} \in \Omega$ via an indicator function.

Let $\Omega \subseteq \mathbb{R}^n$ be closed, convex, nonempty, and define

$$\psi(\boldsymbol{x}) = \mathbb{I}_\Omega(\boldsymbol{x}) = \begin{cases} 0, & \boldsymbol{x} \in \Omega, \\ \infty, & \boldsymbol{x} \notin \Omega. \end{cases}$$

Then

$$\operatorname{prox}_{t\mathbb{I}_\Omega}(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \left\{ t\mathbb{I}_\Omega(\boldsymbol{z}) + \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2 \right\} = \arg\min_{\boldsymbol{z}\in\Omega} \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2.$$

The *projection* onto Ω is

$$\operatorname{proj}_\Omega(\boldsymbol{x}) := \arg\min_{\boldsymbol{z}\in\Omega} \|\boldsymbol{z} - \boldsymbol{x}\|.$$

In this setting, *PGD* is simply gradient descent followed by projection, and *APGD* also applies provided we can compute $\operatorname{proj}_\Omega$ efficiently.

17.2.2 Examples

Example

Example 1: Projection onto a ball.
Let $\overline{\mathbb{B}}(\mathbf{0}, r) = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \le r\}$. Then

$$\operatorname{proj}_{\overline{\mathbb{B}}(\mathbf{0},r)}(\boldsymbol{x}) = \begin{cases} \boldsymbol{x}, & \|\boldsymbol{x}\| \le r, \\ r\,\dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, & \|\boldsymbol{x}\| > r. \end{cases}$$

Example

Example 2: Projection onto an affine subspace.
Let

$$\Omega = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}, \quad \boldsymbol{A} \in \mathbb{R}^{m\times n},\ \boldsymbol{b} \in \mathbb{R}^m.$$

Precomputation.

- Find some $\boldsymbol{x}_0$ such that $\boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{b}$.
- Compute a matrix $\boldsymbol{U} \in \mathbb{R}^{n\times p}$ whose columns form an orthonormal basis of $\operatorname{Null}(\boldsymbol{A})$, so $\boldsymbol{U}^\top\boldsymbol{U} = I_p$.

If $\operatorname{rank}(\boldsymbol{A}) = m$ (full row rank), we may use a **QR factorization** of $\boldsymbol{A}^\top$; then $p = n - m$ by the rank–nullity theorem. If $\operatorname{rank}(\boldsymbol{A}) < m$, one can use an **SVD** of $\boldsymbol{A}$; then $p > n - m$. We assume $\boldsymbol{b} \in \operatorname{Range}(\boldsymbol{A})$ so that $\Omega \neq \varnothing$.

Note that

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \iff \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x}_0 \iff \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_0) = \boldsymbol{0} \iff \boldsymbol{x} - \boldsymbol{x}_0 \in \mathrm{Null}(\boldsymbol{A}) \iff \boldsymbol{x} - \boldsymbol{x}_0 = \boldsymbol{U}\boldsymbol{w}$$

for some $\boldsymbol{w} \in \mathbb{R}^p$.

Theorem

With $\Omega, \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{U}, \boldsymbol{x}_0$ as above,

$$\mathrm{proj}_\Omega(\boldsymbol{y}) = \boldsymbol{x}_0 + \boldsymbol{U}\boldsymbol{U}^\top(\boldsymbol{y} - \boldsymbol{x}_0).$$

Proof

We want

$$\min_{\boldsymbol{x}} \left\{ \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \right\}.$$

Using the parameterization $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{U}\boldsymbol{w}$, this is equivalent to

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{x}_0 + \boldsymbol{U}\boldsymbol{w} - \boldsymbol{y}\|^2.$$

Expanding,

$$\begin{aligned}\frac{1}{2}\|\boldsymbol{x}_0 + \boldsymbol{U}\boldsymbol{w} - \boldsymbol{y}\|^2 &= \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{w} + (\boldsymbol{x}_0 - \boldsymbol{y})^\top\boldsymbol{U}\boldsymbol{w} + C_1 \\ &= \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + (\boldsymbol{x}_0 - \boldsymbol{y})^\top\boldsymbol{U}\boldsymbol{w} + C_1 \\ &= \frac{1}{2}\|\boldsymbol{w} + \boldsymbol{U}^\top(\boldsymbol{x}_0 - \boldsymbol{y})\|^2 + C_2,\end{aligned}$$

where C_1, C_2 are constants independent of $\boldsymbol{w}$.
This is minimized by

$$\boldsymbol{w}^* = \boldsymbol{U}^\top(\boldsymbol{y} - \boldsymbol{x}_0).$$

Hence

$$\boldsymbol{x}^* = \boldsymbol{x}_0 + \boldsymbol{U}\boldsymbol{w}^* = \boldsymbol{x}_0 + \boldsymbol{U}\boldsymbol{U}^\top(\boldsymbol{y} - \boldsymbol{x}_0),$$

which is exactly $\mathrm{proj}_\Omega(\boldsymbol{y})$.

Lec 18

18.1 Nonlinear Programming

Consider general nonlinear programming (NLP):

$$\min_{\boldsymbol{x}} f_0(\boldsymbol{x}) \quad \text{S.T.} \quad f_1(\boldsymbol{x}) \le 0, \dots, f_m(\boldsymbol{x}) \le 0,\ h_1(\boldsymbol{x}) = 0, \dots, h_p(\boldsymbol{x}) = 0,$$

where $f_0, \dots, f_m, h_1, \dots, h_p : \mathbb{R}^n \to \mathbb{R}$.

Let the **feasible region** be

$$\Omega := \{\boldsymbol{x} : f_i(\boldsymbol{x}) \le 0,\ i = 1, \dots, m;\ h_i(\boldsymbol{x}) = 0,\ i = 1, \dots, p\}.$$

Define the **Lagrangian**

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{i=1}^{p} \mu_i h_i(\boldsymbol{x}),$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m$, $\boldsymbol{\mu} \in \mathbb{R}^p$ are the **multipliers**.

Theorem

$$\inf_{\boldsymbol{x}} \{f_0(\boldsymbol{x}) : \boldsymbol{x} \in \Omega\} = \inf_{\boldsymbol{x}} \sup_{\boldsymbol{\lambda} \ge \boldsymbol{0},\, \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}).$$

Note. inf is like min, except the minimum need not be attained (e.g. $\inf\{e^x : x \in \mathbb{R}\} = 0$, and $\inf \varnothing = \infty$). Similarly, sup is analogous to max.

Proof

Fix $\boldsymbol{x} \in \mathbb{R}^n$.
Case 1: $\boldsymbol{x} \notin \Omega$ (infeasible).
Then either $f_i(\boldsymbol{x}) > 0$ for some i, or $h_i(\boldsymbol{x}) \ne 0$ for some i (or both).
If $f_i(\boldsymbol{x}) > 0$ for some i, set all multipliers to 0 except $\lambda_i \to \infty$. Then

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \to \infty,$$

so

$$\sup_{\boldsymbol{\lambda} \ge \boldsymbol{0},\, \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \infty,$$

which can never be the value of the outer infimum.
Similarly, if $h_i(\boldsymbol{x}) \ne 0$ for some i, fix all multipliers at 0 except $\mu_i = t \cdot \operatorname{sgn}(h_i(\boldsymbol{x}))$, $t \to \infty$.
Again $\sup_{\boldsymbol{\lambda} \ge \boldsymbol{0}, \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \infty$.
So for $\boldsymbol{x} \notin \Omega$, the inner sup is ∞ and cannot contribute to the outer inf.

Case 2: $\boldsymbol{x} \in \Omega$ (feasible).
Then $f_i(\boldsymbol{x}) \le 0$ for all $i = 1, \dots, m$ and $h_i(\boldsymbol{x}) = 0$ for all $i = 1, \dots, p$. For any $\boldsymbol{\lambda} \ge \boldsymbol{0}$, any $\boldsymbol{\mu}$,

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{i=1}^{p} \mu_i h_i(\boldsymbol{x}) \le f_0(\boldsymbol{x}),$$

since $\sum \lambda_i f_i(\boldsymbol{x}) \le 0$ and $\sum \mu_i h_i(\boldsymbol{x}) = 0$. In fact,

$$\sup_{\boldsymbol{\lambda} \ge \mathbf{0},\, \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f_0(\boldsymbol{x}),$$

attained by taking $\boldsymbol{\lambda} = \mathbf{0}$ (and any $\boldsymbol{\mu}$).

Therefore, when we take the outer infimum over $\boldsymbol{x}$, we see:

$$\sup_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \begin{cases} \infty, & \boldsymbol{x} \notin \Omega, \\ f_0(\boldsymbol{x}), & \boldsymbol{x} \in \Omega. \end{cases}$$

Thus

$$\inf_{\boldsymbol{x}} \sup_{\boldsymbol{\lambda} \ge \mathbf{0}, \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{x} \in \Omega} f_0(\boldsymbol{x}),$$

as claimed.

Theorem

(Weak Duality.)

$$\inf_{\boldsymbol{x}} \sup_{\boldsymbol{\lambda} \ge \mathbf{0}, \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \;\ge\; \sup_{\boldsymbol{\lambda} \ge \mathbf{0}, \boldsymbol{\mu}} \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}).$$

Proof

Fix $\boldsymbol{x}_1 \in \mathbb{R}^n$ and $\boldsymbol{\lambda}_1 \ge \mathbf{0}$, $\boldsymbol{\mu}_1$. Then

$$\mathcal{L}(\boldsymbol{x}_1, \boldsymbol{\lambda}_1, \boldsymbol{\mu}_1) \;\ge\; \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}_1, \boldsymbol{\mu}_1).$$

Since $\boldsymbol{\lambda}_1, \boldsymbol{\mu}_1$ are arbitrary, we can take sup over them:

$$\sup_{\boldsymbol{\lambda}_1 \ge \mathbf{0}, \boldsymbol{\mu}_1} \mathcal{L}(\boldsymbol{x}_1, \boldsymbol{\lambda}_1, \boldsymbol{\mu}_1) \;\ge\; \sup_{\boldsymbol{\lambda}_1 \ge \mathbf{0}, \boldsymbol{\mu}_1} \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}_1, \boldsymbol{\mu}_1).$$

The left-hand side holds for all $\boldsymbol{x}_1$, so taking inf over $\boldsymbol{x}_1$ preserves the inequality:

$$\inf_{\boldsymbol{x}_1} \sup_{\boldsymbol{\lambda}_1 \ge \mathbf{0}, \boldsymbol{\mu}_1} \mathcal{L}(\boldsymbol{x}_1, \boldsymbol{\lambda}_1, \boldsymbol{\mu}_1) \;\ge\; \sup_{\boldsymbol{\lambda}_1 \ge \mathbf{0}, \boldsymbol{\mu}_1} \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}_1, \boldsymbol{\mu}_1).$$

Game interpretation. Think of a two-player game:

- the inf-player chooses $\boldsymbol{x}$ (tries to minimize the payoff),
- the sup-player chooses $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ (tries to maximize the payoff).

Weak duality says: it is better for the inf-player (i.e. lower cost) if the sup-player must announce their strategy first.

18.1.1 Dual Objective and Dual Problem

Define the **dual objective**

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) := \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}).$$

The **dual optimization problem** is

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \; g(\boldsymbol{\lambda},\boldsymbol{\mu}) \quad \text{S.T.} \quad \boldsymbol{\lambda} \geq \mathbf{0}.$$

Weak duality can be rewritten as

$$\sup_{\boldsymbol{\lambda}\geq\mathbf{0},\boldsymbol{\mu}} \; g(\boldsymbol{\lambda},\boldsymbol{\mu}) \;\leq\; \inf_{\boldsymbol{x}}\{f_0(\boldsymbol{x}) : \boldsymbol{x}\in\Omega\}.$$

Equivalently:

- For any primal feasible $\boldsymbol{x}\in\Omega$,

$$\sup_{\boldsymbol{\lambda}\geq\mathbf{0},\boldsymbol{\mu}} \; g(\boldsymbol{\lambda},\boldsymbol{\mu}) \;\leq\; f_0(\boldsymbol{x}).$$

- For any dual feasible $(\boldsymbol{\lambda},\boldsymbol{\mu})$ with $\boldsymbol{\lambda}\geq\mathbf{0}$,

$$\inf_{\boldsymbol{x}}\{f_0(\boldsymbol{x}) : \boldsymbol{x}\in\Omega\} \;\geq\; g(\boldsymbol{\lambda},\boldsymbol{\mu}).$$

Weak duality requires no assumptions, so it is very general but often too weak by itself.

Theorem

(Strong Duality.)

$$\sup_{\boldsymbol{\lambda}\geq\mathbf{0},\boldsymbol{\mu}} \; g(\boldsymbol{\lambda},\boldsymbol{\mu}) \;=\; \inf_{\boldsymbol{x}}\{f_0(\boldsymbol{x}) : \boldsymbol{x}\in\Omega\}.$$

Equivalently,

$$\inf_{\boldsymbol{x}} \sup_{\boldsymbol{\lambda}\geq\mathbf{0},\boldsymbol{\mu}} \mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \;=\; \sup_{\boldsymbol{\lambda}\geq\mathbf{0},\boldsymbol{\mu}} \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}).$$

Note. The equality above does *not* require that the infimum/supremum are attained. Some authors define strong duality to include the existence of optimizers:

$$\exists\, \boldsymbol{x}^*\in\Omega,\; \boldsymbol{\lambda}^*\geq\mathbf{0},\; \boldsymbol{\mu}^* \quad \text{s.t.} \quad g(\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*) = f_0(\boldsymbol{x}^*).$$

Theorem

(KKT-type characterization.)
Strong duality is attained at some $\boldsymbol{x}^*\in\Omega$, $\boldsymbol{\lambda}^*\geq\mathbf{0}$, $\boldsymbol{\mu}^*$ **if and only if** all four conditions hold:

(i) **Primal feasibility:** $\boldsymbol{x}^*\in\Omega$;

(ii) **Dual feasibility:** $\boldsymbol{\lambda}^*\geq\mathbf{0}$;

(iii) **Complementarity:**

$$\lambda_i^*\, f_i(\boldsymbol{x}^*) = 0, \quad i=1,\ldots,m;$$

(iv) **Stationarity:**

$$\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*).$$

These are exactly the (convex) KKT conditions.

Proof

($\Leftarrow$) Assume (1)–(4) hold.
Using (1) and (3),

$$\sum_{i=1}^{m} \lambda_i^* f_i(\boldsymbol{x}^*) = 0, \quad \sum_{i=1}^{p} \mu_i^* h_i(\boldsymbol{x}^*) = 0.$$

Thus

$$\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = f_0(\boldsymbol{x}^*) + 0 + 0 = f_0(\boldsymbol{x}^*).$$

By (4),

$$\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = g(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*).$$

Hence $g(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = f_0(\boldsymbol{x}^*)$, so strong duality holds and is attained at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$.

($\Rightarrow$) Suppose strong duality holds and is attained at $\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*$, i.e.

$$g(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = f_0(\boldsymbol{x}^*), \quad \boldsymbol{x}^* \in \Omega,\ \boldsymbol{\lambda}^* \geq \mathbf{0}.$$

Then (1) and (2) hold by assumption.
From the strong duality equality and weak duality chain,

$$\begin{aligned} f_0(\boldsymbol{x}^*) &= \inf_{\boldsymbol{x}} \sup_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ &\geq \sup_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\mu}} \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ &= g(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \\ &\leq \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \\ &\leq \sup_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ &\leq f_0(\boldsymbol{x}^*), \end{aligned}$$

so all inequalities are in fact equalities. In particular,

$$f_0(\boldsymbol{x}^*) = \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = f_0(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i^* f_i(\boldsymbol{x}^*) + \sum_{i=1}^{p} \mu_i^* h_i(\boldsymbol{x}^*).$$

Hence

$$\sum_{i=1}^{m} \lambda_i^* f_i(\boldsymbol{x}^*) + \sum_{i=1}^{p} \mu_i^* h_i(\boldsymbol{x}^*) = 0.$$

But $f_i(\boldsymbol{x}^*) \leq 0$ and $\lambda_i^* \geq 0$; therefore each term $\lambda_i^* f_i(\boldsymbol{x}^*) \leq 0$, and the sum being zero forces

$$\lambda_i^* f_i(\boldsymbol{x}^*) = 0, \quad i = 1, \ldots, m$$

(complementarity, (3)). Also,

$$f_0(\boldsymbol{x}^*) = \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*),$$

so $\boldsymbol{x}^*$ is a minimizer of $\mathcal{L}(\cdot, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$, i.e. (4) holds.

Convex specialization.

Suppose $f_0, \ldots, f_m$ are convex, and the equality constraints are affine:

$$\begin{pmatrix} h_1(\boldsymbol{x}) \\ \vdots \\ h_p(\boldsymbol{x}) \end{pmatrix} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}.$$

Then for fixed $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ with $\boldsymbol{\lambda}^* \geq \boldsymbol{0}$, the Lagrangian $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ is convex in $\boldsymbol{x}$.

Condition (4) is equivalent to

$$\boldsymbol{0} \in \partial_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*).$$

If moreover $f_0, \ldots, f_m$ are differentiable, then

$$\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \nabla f_0(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(\boldsymbol{x}^*) + \boldsymbol{A}^\top \boldsymbol{\mu}^*,$$

since

$$\nabla_{\boldsymbol{x}} (\boldsymbol{\mu}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})) = \boldsymbol{A}^\top \boldsymbol{\mu}.$$

In earlier KKT formulations, the sum over i was only over *active* constraints. Here we have a multiplier λ_i^* for every inequality constraint, but complementarity (3) gives $\lambda_i^* = 0$ whenever $f_i(\boldsymbol{x}^*) < 0$ (inactive).

18.2 Kernel Machines

Consider **ridge regression**:

$$\min_{\boldsymbol{x}} \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 + \frac{1}{2} \gamma \|\boldsymbol{x}\|^2,$$

with $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\gamma > 0$.

The closed-form solution is

$$\boldsymbol{x}^* = (\boldsymbol{A}^\top \boldsymbol{A} + \gamma \boldsymbol{I})^{-1} \boldsymbol{A}^\top \boldsymbol{y}.$$

Often one wants to enrich the feature set: a linear model may not fit well (e.g. a parabola is needed). Starting from p original features, including all quadratic monomials would yield $\frac{p(p+1)}{2}$ additional feature columns.

In general, we might want N new features with $N \gg m$. This leads naturally to kernel methods, where the feature space can be extremely high-dimensional (or infinite), but the optimization is done via inner products (kernels) rather than explicit feature vectors.

Lec 19

19.1 Kernel Ridge Regression

19.1.1 Ridge Regression

Given $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, consider

$$\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 + \frac{1}{2}\gamma\|\boldsymbol{x}\|^2.$$

19.1.2 Kernel Ridge Regression

Now "lift" to a higher-dimensional feature matrix $\tilde{\boldsymbol{A}} \in \mathbb{R}^{m \times N}$ with $N \gg \max(m, n)$ and solve

$$\tilde{\boldsymbol{x}} := (\tilde{\boldsymbol{A}}^\top \tilde{\boldsymbol{A}} + \gamma \boldsymbol{I}_N)^{-1} \tilde{\boldsymbol{A}}^\top \boldsymbol{y}.$$

19.2 Kernel Trick

19.2.1 Part 1

Theorem

$$\underbrace{(\tilde{\boldsymbol{A}}^\top \tilde{\boldsymbol{A}} + \gamma \boldsymbol{I}_N)}_{N \times N}^{-1} \tilde{\boldsymbol{A}}^\top = \tilde{\boldsymbol{A}}^\top \underbrace{(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\top + \gamma \boldsymbol{I}_m)}_{m \times m}^{-1}.$$

Proof

$$\begin{aligned}(\tilde{\boldsymbol{A}}^\top \tilde{\boldsymbol{A}} + \gamma \boldsymbol{I}_N)\, \tilde{\boldsymbol{A}}^\top (\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\top + \gamma \boldsymbol{I}_m)^{-1} &= (\tilde{\boldsymbol{A}}^\top \tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\top + \gamma \tilde{\boldsymbol{A}}^\top)(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\top + \gamma \boldsymbol{I}_m)^{-1} \\ &= \tilde{\boldsymbol{A}}^\top (\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\top + \gamma \boldsymbol{I}_m)(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\top + \gamma \boldsymbol{I}_m)^{-1} \\ &= \tilde{\boldsymbol{A}}^\top.\end{aligned}$$

Thus the claimed identity holds.

19.2.2 Part 2

We do not need to store $\tilde{\boldsymbol{x}} \in \mathbb{R}^N$ explicitly. By the theorem,

$$\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{A}}^\top (\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\top + \gamma \boldsymbol{I}_m)^{-1} \boldsymbol{y} = \tilde{\boldsymbol{A}}^\top \boldsymbol{w},$$

where

$$\boldsymbol{w} := (\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\top + \gamma \boldsymbol{I}_m)^{-1} \boldsymbol{y} \in \mathbb{R}^m.$$

So we store only $\boldsymbol{w}$. For a new instance $\tilde{\boldsymbol{a}} \in \mathbb{R}^N$,

$$\tilde{\boldsymbol{a}}^\top \tilde{\boldsymbol{x}} = \tilde{\boldsymbol{a}}^\top \tilde{\boldsymbol{A}}^\top \boldsymbol{w},$$

so classification can be done using matrix–vector products involving $\tilde{\boldsymbol{A}}$ and the new $\tilde{\boldsymbol{a}}$, and $\boldsymbol{w}$.

19.2.3 Part 3

A **kernel** is the matrix

$$\boldsymbol{K} := \tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\top \in \mathbb{S}^m,$$

often called the Gram matrix or inner-product matrix.

For some feature families it is possible to compute an approximation $\hat{\boldsymbol{K}}$ much faster than explicitly forming $\tilde{\boldsymbol{A}}$ and then $\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\top$.

Example: all monomials up to degree d in the original features.

Example

(Polynomial features.)
Let $n = 2$, degree $d = 2$. Then $N = 6$ features:

$$1,\ x,\ y,\ x^2,\ y^2,\ xy.$$

For $\boldsymbol{K}(i,j)$, suppose row i of $\boldsymbol{A}$ is (v_1, v_2) and row j is (w_1, w_2). Then the corresponding rows of $\tilde{\boldsymbol{A}}$ are

$$\text{row } i \mapsto (1,\ v_1,\ v_2,\ v_1^2,\ v_1v_2,\ v_2^2),$$
$$\text{row } j \mapsto (1,\ w_1,\ w_2,\ w_1^2,\ w_1w_2,\ w_2^2).$$

Then $\boldsymbol{K}(i,j)$ is the inner product of these two 6-vectors.
For a more convenient kernel, consider

$$(1 + v_1w_1 + v_2w_2)^2.$$

Expanding gives

$$1 + 2v_1w_1 + 2v_2w_2 + v_1^2w_1^2 + v_2^2w_2^2 + 2v_1w_1v_2w_2,$$

which is the inner product of

$$(1,\ \sqrt{2}v_1,\ \sqrt{2}v_2,\ v_1^2,\ \sqrt{2}v_1v_2,\ v_2^2)$$

and

$$(1,\ \sqrt{2}w_1,\ \sqrt{2}w_2,\ w_1^2,\ \sqrt{2}w_1w_2,\ w_2^2).$$

Call this inner product $\hat{\boldsymbol{K}}(i,j)$. Then $\hat{\boldsymbol{K}}$ behaves like $\boldsymbol{K}$ (up to scaling factors of $\sqrt{2}$, which only rescale $\tilde{\boldsymbol{x}}$) but can be computed without explicitly forming the rows of $\tilde{\boldsymbol{A}}$.

Conclusion. The kernel trick lets us effectively add many columns (features) without working explicitly with an $m \times N$ matrix.

Part 3 also applies to forming $\tilde{\boldsymbol{A}}\tilde{\boldsymbol{a}}$ for a new data point $\tilde{\boldsymbol{a}}$: kernels can be defined for many feature families beyond polynomials.

19.3 Kernel Machines

Let the i-th row of $\boldsymbol{A}$ be $\boldsymbol{\xi}_i^\top \in \mathbb{R}^n$. After lifting by a feature map

$$\boldsymbol{\Phi} : \mathbb{R}^n \to \mathbb{R}^N, \quad \boldsymbol{\Phi}(\boldsymbol{\xi}_i) = \begin{bmatrix} \Phi_1(\boldsymbol{\xi}_i) \\ \vdots \\ \Phi_N(\boldsymbol{\xi}_i) \end{bmatrix},$$

we have

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{\xi}_1^\top \\ \vdots \\ \boldsymbol{\xi}_m^\top \end{bmatrix} \xrightarrow{\text{Map}} \tilde{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{\Phi}^\top(\boldsymbol{\xi}_1) \\ \vdots \\ \boldsymbol{\Phi}^\top(\boldsymbol{\xi}_m) \end{bmatrix} = \begin{bmatrix} \Phi_1(\boldsymbol{\xi}_1) & \cdots & \Phi_N(\boldsymbol{\xi}_1) \\ \vdots & \ddots & \vdots \\ \Phi_1(\boldsymbol{\xi}_m) & \cdots & \Phi_N(\boldsymbol{\xi}_m) \end{bmatrix}.$$

Then the kernel (Gram matrix) $\boldsymbol{K}$ is

$$\boldsymbol{K} = \begin{bmatrix} \boldsymbol{\Phi}^\top(\boldsymbol{\xi}_1)\boldsymbol{\Phi}(\boldsymbol{\xi}_1) & \cdots & \boldsymbol{\Phi}^\top(\boldsymbol{\xi}_1)\boldsymbol{\Phi}(\boldsymbol{\xi}_m) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Phi}^\top(\boldsymbol{\xi}_m)\boldsymbol{\Phi}(\boldsymbol{\xi}_1) & \cdots & \boldsymbol{\Phi}^\top(\boldsymbol{\xi}_m)\boldsymbol{\Phi}(\boldsymbol{\xi}_m) \end{bmatrix} \in \mathbb{S}^m.$$

$\boldsymbol{K}$ is also called a *Gram matrix* or *inner-product matrix*.

Example

(Nonlinear separator via kernel SVM.)

Recall soft-margin SVM (SVM-7):

$$\min_{\boldsymbol{x},\xi,\boldsymbol{s}} \frac{1}{2}\|\boldsymbol{x}\|^2 + \gamma \sum_{i=1}^m s_i \quad \text{S.T.} \quad \begin{cases} \boldsymbol{a}_i^\top \boldsymbol{x} + \xi \geq 1 - s_i, & y_i = 1, \\ \boldsymbol{a}_i^\top \boldsymbol{x} + \xi \leq -1 + s_i, & y_i = -1, \\ \boldsymbol{s} \geq \boldsymbol{0}, \end{cases}$$

with $\boldsymbol{a}_i \in \mathbb{R}^n$.

After mapping $\boldsymbol{a}_i \mapsto \boldsymbol{\Phi}(\boldsymbol{a}_i) \in \mathbb{R}^N$, the *kernel SVM* seeks $\tilde{\boldsymbol{x}} \in \mathbb{R}^N$, $\xi \in \mathbb{R}$, $\boldsymbol{s} \in \mathbb{R}^m$ solving

$$\min_{\tilde{\boldsymbol{x}},\xi,\boldsymbol{s}} \frac{1}{2}\|\tilde{\boldsymbol{x}}\|^2 + \gamma \sum_{i=1}^m s_i \quad \text{S.T.} \quad \begin{cases} \boldsymbol{\Phi}^\top(\boldsymbol{a}_i)\tilde{\boldsymbol{x}} + \xi \geq 1 - s_i, & y_i = 1, \\ \boldsymbol{\Phi}^\top(\boldsymbol{a}_i)\tilde{\boldsymbol{x}} + \xi \leq -1 + s_i, & y_i = -1, \\ \boldsymbol{s} \geq \boldsymbol{0}. \end{cases}$$

Lagrangian formulation.
Introduce multipliers $\boldsymbol{\lambda} \geq \boldsymbol{0}$ for $y_i = 1$ constraints, $\boldsymbol{\lambda}' \geq \boldsymbol{0}$ for $y_i = -1$ constraints, and $\boldsymbol{\pi} \geq \boldsymbol{0}$ for $s_i \geq 0$. The Lagrangian is

$$\min_{\tilde{\boldsymbol{x}},\xi,\boldsymbol{s}} \max_{\substack{\boldsymbol{\lambda}\geq\boldsymbol{0} \\ \boldsymbol{\lambda}'\geq\boldsymbol{0} \\ \boldsymbol{\pi}\geq\boldsymbol{0}}} \mathcal{L}(\tilde{\boldsymbol{x}}, \xi, \boldsymbol{s}; \boldsymbol{\lambda}, \boldsymbol{\lambda}', \boldsymbol{\pi}),$$

where

$$\mathcal{L}(\tilde{\boldsymbol{x}}, \xi, \boldsymbol{s}; \boldsymbol{\lambda}, \boldsymbol{\lambda}', \boldsymbol{\pi}) = \underbrace{\frac{1}{2}\|\tilde{\boldsymbol{x}}\|^2 + \gamma \sum_{i=1}^{m} s_i}_{f_0(\tilde{\boldsymbol{x}}, \boldsymbol{s})} + \sum_{y_i=1} \lambda_i (1 - s_i - \boldsymbol{\Phi}^\top(\boldsymbol{a}_i)\tilde{\boldsymbol{x}} - \xi) + \sum_{y_i=-1} \lambda_i' (1 - s_i + \boldsymbol{\Phi}^\top(\boldsymbol{a}_i)\tilde{\boldsymbol{x}} + \xi) + \sum_{i=1}^{m} (-s_i)\pi_i.$$

Quadratic programming (QP).

A QP is an optimization problem of the form

$$\min_{\boldsymbol{x}} q(\boldsymbol{x}) \quad \text{S.T.} \quad A_1\boldsymbol{x} = \boldsymbol{b}_1, \quad A_2\boldsymbol{x} \le \boldsymbol{b}_2,$$

where q is quadratic with positive semidefinite Hessian. Soft-margin SVM (SVM-7) and ℓ_1-LS are QPs.

Fact

If a QP has at least one feasible point and is bounded below, then it has an optimizer and strong duality holds.

For SVM-7 the objective is bounded below by 0, and there is an obvious feasible point (e.g. $\tilde{\boldsymbol{x}} = \boldsymbol{0}$, $\xi = 0$, $\boldsymbol{s} = \boldsymbol{e}$), so strong duality applies.

Thus the dual problem is

$$\max_{\substack{\boldsymbol{\lambda} \ge \boldsymbol{0} \\ \boldsymbol{\lambda}' \ge \boldsymbol{0} \\ \boldsymbol{\pi} \ge \boldsymbol{0}}} \min_{\tilde{\boldsymbol{x}}, \xi, \boldsymbol{s}} \mathcal{L}(\tilde{\boldsymbol{x}}, \xi, \boldsymbol{s}; \boldsymbol{\lambda}, \boldsymbol{\lambda}', \boldsymbol{\pi}).$$

The inner minimization in each s_i is

$$\min_{s_i} \gamma s_i - \begin{cases} \lambda_i s_i, & y_i = 1, \\ \lambda_i' s_i, & y_i = -1 \end{cases} - \pi_i s_i,$$

which is $-\infty$ unless

$$\gamma - \begin{cases} \lambda_i, \\ \lambda_i' \end{cases} - \pi_i = 0.$$

These are **hidden constraints** on dual variables; they must hold at any optimal dual solution. Under these, the s_i disappear.

We then minimize over ξ; the Lagrangian is linear in ξ, so the minimum is finite only if

$$\sum_{y_i=1} \lambda_i = \sum_{y_i=-1} \lambda_i',$$

another hidden constraint. With both hidden constraints satisfied, the inner problem reduces to

$$\min_{\tilde{\boldsymbol{x}}} \frac{1}{2}\|\tilde{\boldsymbol{x}}\|^2 + \boldsymbol{p}^\top \tilde{\boldsymbol{x}} + c,$$

where

$$\boldsymbol{p} = \sum_{y_i=-1} \lambda_i' \boldsymbol{\Phi}(\boldsymbol{a}_i) - \sum_{y_i=1} \lambda_i \boldsymbol{\Phi}(\boldsymbol{a}_i), \qquad c = \sum_{y_i=1} \lambda_i + \sum_{y_i=-1} \lambda_i'.$$

The minimizer is

$$\tilde{\boldsymbol{x}}^* = -\boldsymbol{p},$$

and the optimal inner value is

$$-\frac{1}{2}\|\boldsymbol{p}\|^2 + c.$$

Thus the dual becomes

$$\begin{aligned} \max_{\substack{\boldsymbol{\lambda}\geq\mathbf{0}\\ \boldsymbol{\lambda}'\geq\mathbf{0}\\ \boldsymbol{\pi}\geq\mathbf{0}}} \quad & -\frac{1}{2}\Big\|\sum_{y_i=1} \lambda_i \boldsymbol{\Phi}(\boldsymbol{a}_i) - \sum_{y_i=-1} \lambda_i' \boldsymbol{\Phi}(\boldsymbol{a}_i)\Big\|^2 + \sum_{y_i=1} \lambda_i + \sum_{y_i=-1} \lambda_i' \\ \text{s.t.} \quad & \begin{cases} \boldsymbol{\lambda} \geq \mathbf{0},\ \boldsymbol{\lambda}' \geq \mathbf{0},\ \boldsymbol{\pi} \geq \mathbf{0}, \\ \lambda_i + \pi_i = \gamma, \quad \lambda_i' + \pi_i = \gamma,\ \forall i, \\ \displaystyle\sum_{y_i=1} \lambda_i = \sum_{y_i=-1} \lambda_i'. \end{cases} \end{aligned}$$

Eliminating $\boldsymbol{\pi}$ via $\pi_i = \gamma - \lambda_i = \gamma - \lambda_i'$ gives the simpler dual:

$$\begin{aligned} \max_{\boldsymbol{\lambda},\boldsymbol{\lambda}'} \quad & -\frac{1}{2}\Big\|\sum_{y_i=1} \lambda_i \boldsymbol{\Phi}(\boldsymbol{a}_i) - \sum_{y_i=-1} \lambda_i' \boldsymbol{\Phi}(\boldsymbol{a}_i)\Big\|^2 + \sum_{y_i=1} \lambda_i + \sum_{y_i=-1} \lambda_i' \\ \text{s.t.} \quad & \boldsymbol{\lambda} \geq \mathbf{0}, \quad \boldsymbol{\lambda}' \geq \mathbf{0}, \quad \boldsymbol{\lambda} \leq \gamma \boldsymbol{e}, \quad \boldsymbol{\lambda}' \leq \gamma \boldsymbol{e}, \\ & \sum_{y_i=1} \lambda_i = \sum_{y_i=-1} \lambda_i'. \end{aligned}$$

Let $\boldsymbol{K}$ denote the quadratic coefficients in $(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$ in the dual objective. In terms of feature vectors,

$$\boldsymbol{K} = \begin{bmatrix} \boldsymbol{\Phi}^\top(\boldsymbol{a}_1)\boldsymbol{\Phi}(\boldsymbol{a}_1) & \cdots & \pm\boldsymbol{\Phi}^\top(\boldsymbol{a}_1)\boldsymbol{\Phi}(\boldsymbol{a}_m) \\ \vdots & \ddots & \vdots \\ \pm\boldsymbol{\Phi}^\top(\boldsymbol{a}_m)\boldsymbol{\Phi}(\boldsymbol{a}_1) & \cdots & \boldsymbol{\Phi}^\top(\boldsymbol{a}_m)\boldsymbol{\Phi}(\boldsymbol{a}_m) \end{bmatrix},$$

with signs depending on the labels y_i (same labels $\Rightarrow +$, different labels $\Rightarrow -$).

Lec 20

20.1 Singular Value Decomposition

Given $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we want a matrix $\tilde{\boldsymbol{A}}$ such that $\tilde{\boldsymbol{A}} \approx \boldsymbol{A}$ and $\operatorname{rank}(\tilde{\boldsymbol{A}}) = r \ll \min(m,n)$.

Theorem

(SVD.)
For any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, there exist orthogonal $\boldsymbol{U} \in \mathbb{R}^{m\times m}$, $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ and a diagonal $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ such that

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top.$$

The diagonal entries of $\boldsymbol{\Sigma}$, denoted $\sigma_1, \ldots, \sigma_{\min(m,n)}$, are the **singular values** of $\boldsymbol{A}$ and satisfy

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0.$$

The singular values are uniquely determined by $\boldsymbol{A}$.

Properties of the SVD:

1. If $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$ and $\sigma_{r+1} = \cdots = \sigma_{\min(m,n)} = 0$, then $\operatorname{rank}(\boldsymbol{A}) = r$.
2. $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \iff \boldsymbol{A}^\top = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top$.
3. If $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, then
$$\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\boldsymbol{V}^\top,$$
where $\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}$ is square diagonal. Thus the singular values of $\boldsymbol{A}$ are the square roots of the eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$.
4. If $\operatorname{rank}(\boldsymbol{A}) = r$, then
$$\boldsymbol{A} = \boldsymbol{U}(:, 1\!:\!r)\,\boldsymbol{\Sigma}(1\!:\!r, 1\!:\!r)\,\boldsymbol{V}(:, 1\!:\!r)^\top = \sum_{i=1}^{r} \boldsymbol{U}(:,i)\,\sigma_i\,\boldsymbol{V}(:,i)^\top,$$
a sum of r rank–1 matrices.
5. If $\boldsymbol{B} := \boldsymbol{Q}\boldsymbol{A}\boldsymbol{Z}^\top$ with $\boldsymbol{Q}, \boldsymbol{Z}$ orthogonal and $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is an SVD, then
$$\boldsymbol{B} = (\boldsymbol{Q}\boldsymbol{U})\boldsymbol{\Sigma}(\boldsymbol{V}\boldsymbol{Z})^\top$$
is an SVD of $\boldsymbol{B}$. Hence $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same singular values.
6. If $\boldsymbol{D}$ is diagonal, then its singular values are the absolute values of its diagonal entries, sorted in nonincreasing order.

Theorem

(Eckart–Young.)
Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $0 \leq \rho \leq \min(m,n)$. Then

$$\min\{\|\boldsymbol{A} - \boldsymbol{X}\|_{\mathrm{F}} : \boldsymbol{X} \in \mathbb{R}^{m\times n},\ \operatorname{rank}(\boldsymbol{X}) \leq \rho\} = \sqrt{\sigma_{\rho+1}^2 + \cdots + \sigma_{\min(m,n)}^2},$$

and a minimizer is

$$\boldsymbol{X} = \boldsymbol{U}(:, 1:\rho)\, \boldsymbol{\Sigma}(1:\rho, 1:\rho)\, \boldsymbol{V}(:, 1:\rho)^\top,$$

where $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is an SVD.

20.2 Latent Semantic Indexing

(Deerwester et al., 1990)

Given a corpus of documents, form the **term–document matrix** $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, where m is the number of terms in the language and n is the number of documents. The entry $\boldsymbol{A}(i, j)$ is the number of occurrences of term i in document j (after normalization).

Compute the SVD of $\boldsymbol{A}$ and keep the rank–ρ approximation

$$\hat{\boldsymbol{A}} = \boldsymbol{U}(:, 1:\rho)\, \boldsymbol{\Sigma}(1:\rho, 1:\rho)\, \boldsymbol{V}(:, 1:\rho)^\top,$$

where ρ is an estimate of the number of topics in the corpus.

Given a query vector $\boldsymbol{q} \in \mathbb{R}^m$, the vector

$$\boldsymbol{q}^\top \hat{\boldsymbol{A}} \in \mathbb{R}^n$$

gives scores that can be interpreted as the relevance of each document to the query.

20.3 Matrix Completion Problem

Let $\boldsymbol{M} \in (\mathbb{R} \cup \{?\})^{m\times n}$ be a partially observed matrix. We want to fill in the missing entries "?" with numbers.

Application: recommender systems.

Rows of $\boldsymbol{M}$ correspond to users, columns to products, and $\boldsymbol{M}(i, j)$ is the rating user i gives to product j. Most entries are missing; the vendor wants to predict missing entries to recommend products.

A reasonable model: assume the completed $\boldsymbol{M}$ has low rank.

Intuition: each user can be modeled as a linear combination of a small number of *prototypical users* (unknown). If ratings are linear combinations of the ratings given by prototypical users, then the completed $\boldsymbol{M}$ will have rank equal to the number of prototypical users.

Mathematical formulation:

Find $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ such that $\boldsymbol{X}(\Omega) = \boldsymbol{M}(\Omega)$, where

$$\Omega = \{(i, j) : \boldsymbol{M}(i, j) \neq ?\},$$

and subject to this constraint, $\operatorname{rank}(\boldsymbol{X})$ is minimized. This problem is **NP-hard**.

Candès and Recht showed that under certain assumptions on $\boldsymbol{M}$ and Ω, the matrix $\boldsymbol{X}$ solving this NP-hard problem is also a solution of the convex optimization problem

$$\min \|\boldsymbol{X}\|_* \quad \text{s.t.} \quad \boldsymbol{X}(\Omega) = \boldsymbol{M}(\Omega),$$

where $\|\boldsymbol{X}\|_*$ is the **nuclear norm** of $\boldsymbol{X}$,

$$\|\boldsymbol{X}\|_* = \sigma_1(\boldsymbol{X}) + \cdots + \sigma_{\min(m,n)}(\boldsymbol{X}).$$

We check that this is (indeed) a norm.

A norm $\|\cdot\|$ must satisfy:

1. $\|\boldsymbol{X}\| \geq 0$ and $\|\boldsymbol{X}\| = 0 \iff \boldsymbol{X} = \boldsymbol{0}$.
2. $\|\lambda \boldsymbol{X}\| = |\lambda| \, \|\boldsymbol{X}\|$ for all scalars λ.
3. $\|\boldsymbol{X} + \boldsymbol{Y}\| \leq \|\boldsymbol{X}\| + \|\boldsymbol{Y}\|$ (triangle inequality).

Theorem

Let $\|\cdot\|_\square$ be a norm on $\mathbb{R}^m$ and $\|\cdot\|_\triangle$ a norm on $\mathbb{R}^n$. The $\triangle \to \square$ **operator norm** on $\mathbb{R}^{m\times n}$ is defined by

$$\|\boldsymbol{A}\|_{\triangle\to\square} := \sup\{\|\boldsymbol{A}\boldsymbol{v}\|_\square : \|\boldsymbol{v}\|_\triangle \leq 1\}.$$

This is a norm on $\mathbb{R}^{m\times n}$ (an **induced norm**).

Theorem

For $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, the largest singular value $\sigma_1(\boldsymbol{A})$ is the operator norm induced by the Euclidean norm $\|\cdot\|_2$, i.e.

$$\|\boldsymbol{A}\|_{2\to 2} = \sigma_1(\boldsymbol{A}).$$

Proof

We show that

$$\sup_{\|\boldsymbol{x}\|_2 \leq 1} \|\boldsymbol{A}\boldsymbol{x}\|_2 = \sigma_1(\boldsymbol{A}).$$

First, $\|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \sigma_1(\boldsymbol{A})$ for all $\|\boldsymbol{x}\|_2 \leq 1$.
Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be an SVD. For any $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_2 \leq 1$,

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 = \|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{x}\|_2 = \|\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{x}\|_2,$$

since orthogonal matrices preserve the 2-norm. Let $\boldsymbol{w} = \boldsymbol{V}^\top\boldsymbol{x}$. Then $\|\boldsymbol{w}\|_2 = \|\boldsymbol{x}\|_2 \leq 1$, and if $m \geq n$,

$$\|\boldsymbol{\Sigma}\boldsymbol{w}\|_2 = \left(\sigma_1^2 w_1^2 + \cdots + \sigma_n^2 w_n^2\right)^{1/2} \leq \sigma_1\left(w_1^2 + \cdots + w_n^2\right)^{1/2} \leq \sigma_1.$$

Thus $\|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \sigma_1$ for all $\|\boldsymbol{x}\|_2 \leq 1$, so $\|\boldsymbol{A}\|_2 \leq \sigma_1(\boldsymbol{A})$.
For the reverse inequality, take

$$\boldsymbol{x} = \boldsymbol{V}\boldsymbol{e}_1, \quad \boldsymbol{e}_1 = (1, 0, \ldots, 0)^\top.$$

Then $\|\boldsymbol{x}\|_2 = 1$ and

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 = \|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{e}_1\|_2 = \|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{e}_1\|_2 = \|\boldsymbol{\Sigma}\boldsymbol{e}_1\|_2 = \|\,(\sigma_1, 0, \ldots, 0)^\top\|_2 = \sigma_1.$$

Hence $\|\boldsymbol{A}\|_2 \geq \sigma_1(\boldsymbol{A})$, and combining both directions gives $\|\boldsymbol{A}\|_2 = \sigma_1(\boldsymbol{A})$.

Theorem

Let $\|\cdot\|_\square$ be a norm on $\mathbb{R}^m$ (or on $\mathbb{R}^{m\times n}$). The function

$$\boldsymbol{x} \mapsto \sup\{\langle \boldsymbol{x}, \boldsymbol{y}\rangle : \|\boldsymbol{y}\|_\square \le 1\}$$

is the **dual norm** of $\|\cdot\|_\square$ and is itself a norm.

Proof

This is the operator norm of the linear map $\boldsymbol{y} \mapsto \langle \boldsymbol{x}, \boldsymbol{y}\rangle$, so it satisfies the norm axioms.

For vectors:

1. The dual norm of the Euclidean norm is itself (Cauchy–Schwarz inequality).
2. The dual norm of $\|\boldsymbol{x}\|_1$ is $\|\boldsymbol{x}\|_\infty$.

Theorem

For matrices, the dual norm of $\|\cdot\|_2$ (spectral norm) is the nuclear norm $\|\cdot\|_*$.

Proof

See Problem Set 6.

Back to Candès and Recht: why does

$$\min \|\boldsymbol{X}\|_* \quad \text{s.t.} \quad \boldsymbol{X}(\Omega) = \boldsymbol{M}(\Omega)$$

typically yield a low-rank $\boldsymbol{X}$?

Theorem

Let $\boldsymbol{A} \in \mathbb{R}^{m\times n} \setminus \{\boldsymbol{0}\}$ with $m \ge n$. Consider

$$\min \|\boldsymbol{X}\|_* \quad \text{s.t.} \quad \langle \boldsymbol{A}, \boldsymbol{X}\rangle = 1.$$

Then an optimal solution is

$$\hat{\boldsymbol{X}} = \boldsymbol{U} \begin{bmatrix} \frac{1}{\sigma_1(\boldsymbol{A})} & & \\ & 0 & \\ & & \ddots \\ & & & 0 \end{bmatrix} \boldsymbol{V}^\top,$$

where $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is an SVD of $\boldsymbol{A}$.

Proof

First, check that $\hat{\boldsymbol{X}}$ is feasible:

$$\begin{aligned}
\langle \boldsymbol{A}, \hat{\boldsymbol{X}} \rangle &= \left\langle \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \boldsymbol{U} \begin{bmatrix} \frac{1}{\sigma_1(\boldsymbol{A})} & & \\ & 0 & \\ & & \ddots \\ & & & 0 \end{bmatrix} \boldsymbol{V}^\top \right\rangle \\
&= \mathrm{Tr}\Big(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{V} \begin{bmatrix} \frac{1}{\sigma_1(\boldsymbol{A})} & & \\ & 0 & \\ & & \ddots \\ & & & 0 \end{bmatrix}^\top \boldsymbol{U}^\top \Big) \\
&= \mathrm{Tr}\Big(\boldsymbol{\Sigma} \begin{bmatrix} \frac{1}{\sigma_1(\boldsymbol{A})} & & \\ & 0 & \\ & & \ddots \\ & & & 0 \end{bmatrix}^\top \Big) \quad \text{(cyclic invariance of trace, and } \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}) \\
&= \mathrm{Tr} \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} = 1.
\end{aligned}$$

Next, note that $\|\hat{\boldsymbol{X}}\|_2 = \dfrac{1}{\sigma_1(\boldsymbol{A})}$.

For any other feasible $\boldsymbol{X}$ (i.e. $\langle \boldsymbol{A}, \boldsymbol{X} \rangle = 1$), the nuclear norm can be written using the dual norm relationship:

$$\|\boldsymbol{X}\|_* = \sup\{\langle \boldsymbol{X}, \boldsymbol{Y} \rangle : \|\boldsymbol{Y}\|_2 \le 1\}.$$

Take

$$\bar{\boldsymbol{Y}} = \frac{\boldsymbol{A}}{\sigma_1(\boldsymbol{A})} = \frac{\boldsymbol{A}}{\|\boldsymbol{A}\|_2},$$

so $\|\bar{\boldsymbol{Y}}\|_2 = 1$. Then

$$\|\boldsymbol{X}\|_* \ge \langle \boldsymbol{X}, \bar{\boldsymbol{Y}} \rangle = \left\langle \boldsymbol{X}, \frac{\boldsymbol{A}}{\sigma_1(\boldsymbol{A})} \right\rangle = \frac{1}{\sigma_1(\boldsymbol{A})} \langle \boldsymbol{X}, \boldsymbol{A} \rangle = \frac{1}{\sigma_1(\boldsymbol{A})}.$$

Thus any feasible $\boldsymbol{X}$ has nuclear norm at least $1/\sigma_1(\boldsymbol{A})$, which is exactly $\|\hat{\boldsymbol{X}}\|_*$. Therefore $\hat{\boldsymbol{X}}$ is optimal.

Lec 21

21.1 Matrix Completion Problem

Given partially specified $\boldsymbol{M} \in (\mathbb{R} \cup \{?\})^{m \times n}$, the goal is to fill in the missing entries to obtain a completion with as low rank as possible.

Theorem

Candès–Recht Relaxation (CR Relaxation).

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_* \quad \text{subject to} \quad \boldsymbol{X}(\Omega) = \boldsymbol{M}(\Omega),$$

where

$$\Omega = \{(i,j) \in \{1,\dots,m\} \times \{1,\dots,n\} : \boldsymbol{M}(i,j) \neq ?\}.$$

Extension of CR Relaxation to noisy $\boldsymbol{M}$:

$$\min_{\boldsymbol{X}} \frac{1}{2} \sum_{(i,j)\in\Omega} \left(\boldsymbol{X}(i,j) - \boldsymbol{M}(i,j)\right)^2 + \gamma \|\boldsymbol{X}\|_*.$$

Norms are proper, closed, convex functions and are typically nonsmooth, so we are in a good setting for proximal / first-order methods.

We will use *APGD* and therefore need prox for the nuclear norm $\|\cdot\|_*$.

Theorem

Let $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ with $m \geq n$, and let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be its SVD with singular values $\sigma_1, \dots, \sigma_n$ on the diagonal of $\boldsymbol{\Sigma}$. Then

$$\operatorname{prox}_{t\|\cdot\|_*}(\boldsymbol{X}) = \boldsymbol{U} \operatorname{diag}(\max(\sigma_1 - t, 0), \dots, \max(\sigma_n - t, 0)) \boldsymbol{V}^\top.$$

That is, the proximal operator for $\|\cdot\|_*$ performs soft–thresholding on the singular values.

Proof

For matrices, the proximal operator is

$$\operatorname{prox}_{t\|\cdot\|_*}(\boldsymbol{X}) = \arg\min_{\boldsymbol{Z}} \left\{ t\|\boldsymbol{Z}\|_* + \frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{X}\|_{\mathrm{F}}^2 \right\}.$$

Use unitary invariance of both $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_*$. Write $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ with $\boldsymbol{U}, \boldsymbol{V}$ orthogonal, and set

$$\boldsymbol{W} := \boldsymbol{U}^\top \boldsymbol{Z} \boldsymbol{V}.$$

Then

$$\|\boldsymbol{Z} - \boldsymbol{X}\|_{\mathrm{F}} = \|\boldsymbol{U}^\top \boldsymbol{Z} \boldsymbol{V} - \boldsymbol{\Sigma}\|_{\mathrm{F}} = \|\boldsymbol{W} - \boldsymbol{\Sigma}\|_{\mathrm{F}}, \qquad \|\boldsymbol{Z}\|_* = \|\boldsymbol{U}^\top \boldsymbol{Z} \boldsymbol{V}\|_* = \|\boldsymbol{W}\|_*.$$

Hence

$$\arg\min_{\boldsymbol{Z}} \left\{ t\|\boldsymbol{Z}\|_* + \frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{X}\|_{\mathrm{F}}^2 \right\} = \boldsymbol{U} \left(\arg\min_{\boldsymbol{W}} \left\{ t\|\boldsymbol{W}\|_* + \frac{1}{2}\|\boldsymbol{W} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 \right\} \right) \boldsymbol{V}^\top.$$

So it suffices to solve

$$\min_{\boldsymbol{W}} \; t\|\boldsymbol{W}\|_* + \frac{1}{2}\|\boldsymbol{W} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2.$$

Lemma

For any $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ with $m \geq n$,

$$\|\boldsymbol{W}\|_* \;\geq\; |\boldsymbol{W}(1,1)| + \cdots + |\boldsymbol{W}(n,n)|.$$

Proof

Proof of Lemma.
By the dual characterization of the nuclear norm,

$$\|\boldsymbol{W}\|_* = \sup\{\langle \boldsymbol{W}, \boldsymbol{Y}\rangle : \|\boldsymbol{Y}\|_2 \leq 1\}.$$

Choose

$$\hat{\boldsymbol{Y}} = \mathrm{diag}(\pm 1, \ldots, \pm 1),$$

with signs chosen so that the (i,i) entry of $\hat{\boldsymbol{Y}}$ has the same sign as $\boldsymbol{W}(i,i)$. Then $\|\hat{\boldsymbol{Y}}\|_2 = 1$ and

$$\|\boldsymbol{W}\|_* \;\geq\; \langle \boldsymbol{W}, \hat{\boldsymbol{Y}}\rangle = \sum_{i=1}^{n} \boldsymbol{W}(i,i)\,\hat{\boldsymbol{Y}}(i,i) = \sum_{i=1}^{n} |\boldsymbol{W}(i,i)|.$$

Now consider the objective

$$\Phi(\boldsymbol{W}) := t\|\boldsymbol{W}\|_* + \frac{1}{2}\|\boldsymbol{W} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2.$$

Write $\boldsymbol{W}_0$ for the diagonal matrix formed from the diagonal of $\boldsymbol{W}$. From the lemma, $\|\boldsymbol{W}\|_* \geq \|\boldsymbol{W}_0\|_*$, and since $\boldsymbol{\Sigma}$ is diagonal,

$$\|\boldsymbol{W} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 = \|\boldsymbol{W}_0 - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 + (\text{nonnegative off-diagonal terms}),$$

so $\|\boldsymbol{W} - \boldsymbol{\Sigma}\|_{\mathrm{F}} \geq \|\boldsymbol{W}_0 - \boldsymbol{\Sigma}\|_{\mathrm{F}}$ with equality only when $\boldsymbol{W}$ is diagonal. Thus $\Phi(\boldsymbol{W}) \geq \Phi(\boldsymbol{W}_0)$, and any minimizer of Φ must be diagonal.
Let $\boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_n)$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$. Then

$$\|\boldsymbol{W}\|_* = \sum_{i=1}^{n} |w_i|, \qquad \|\boldsymbol{W} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 = \sum_{i=1}^{n} (w_i - \sigma_i)^2.$$

So the problem decouples across i:

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} \left\{ t\|\boldsymbol{w}\|_1 + \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{\sigma}\|_2^2 \right\}, \quad \boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)^\top.$$

This is exactly the proximal operator of $t\|\cdot\|_1$ at $\boldsymbol{\sigma}$, whose solution is componentwise soft–thresholding:

$$w_i = \begin{cases} \sigma_i - t, & \sigma_i \geq t, \\ 0, & \sigma_i < t. \end{cases}$$

Hence

$$\operatorname{prox}_{t\|\cdot\|_*}(\boldsymbol{X}) = \boldsymbol{U}\ \operatorname{diag}\big(\max(\sigma_i - t, 0)\big)_{i=1}^{n}\ \boldsymbol{V}^\top,$$

which completes the proof.

21.2 Economy-sized SVD

Given $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ with $m \geq n$,

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{U}(:, 1\!:\!n)\,\boldsymbol{\Sigma}(1\!:\!n, 1\!:\!n)\,\boldsymbol{V}(:, 1\!:\!n)^\top.$$

This is the **economy-sized SVD**.

In practice, for the nuclear-norm proximal operator we only need the nonzero singular values and corresponding singular vectors, so we use the economy-sized SVD, especially when $m \gg n$.

21.3 Rank–Sparsity Decomposition

Given $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, seek $\boldsymbol{L}, \boldsymbol{M} \in \mathbb{R}^{m\times n}$ such that

$$\boldsymbol{A} = \boldsymbol{L} + \boldsymbol{M},$$

with

- $\boldsymbol{L}$ sparse (most entries zero),
- $\boldsymbol{M}$ low rank ($\operatorname{rank}(\boldsymbol{M}) \ll \min(m, n)$).

Application: background identification / removal from video frames (Problem Set 6). Think of columns of $\boldsymbol{A}$ as vectorized frames: background is nearly low rank, moving objects are sparse.

If $\boldsymbol{A}$ itself were simultaneously very sparse and very low rank, the decomposition would be ill-posed. So we assume $\boldsymbol{L}$ is sparse and high rank, while $\boldsymbol{M}$ is dense and low rank.

Theorem

Convex relaxation (Candès et al. 2011; Chandrasekaran et al. 2011).

$$\min_{\boldsymbol{L},\boldsymbol{M}} \|\boldsymbol{L}\|_{\ell_1} + \gamma\|\boldsymbol{M}\|_* \quad \text{subject to} \quad \boldsymbol{L} + \boldsymbol{M} = \boldsymbol{A},$$

where

$$\|\boldsymbol{L}\|_{\ell_1} = \sum_{i=1}^{m}\sum_{j=1}^{n} |\boldsymbol{L}(i,j)|, \quad \gamma > 0.$$

We solve this with **ADMM** (Alternating Direction Method of Multipliers).

21.3.1 ADMM for two blocks

General form:

$$\min_{\boldsymbol{x},\boldsymbol{y}} f(\boldsymbol{x}) + g(\boldsymbol{y}) \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} = \boldsymbol{c},$$

where $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{A} \in \mathbb{R}^{p\times n}$, $\boldsymbol{B} \in \mathbb{R}^{p\times m}$, $\boldsymbol{c} \in \mathbb{R}^p$.

Augmented Lagrangian:

$$\mathcal{L}_A(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\mu}) = f(\boldsymbol{x}) + g(\boldsymbol{y}) + \boldsymbol{\mu}^\top(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} - \boldsymbol{c}) + \frac{\rho}{2}\|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} - \boldsymbol{c}\|_2^2,$$

where $\rho > 0$.

ADMM iteration:

$$\begin{aligned}
&\boldsymbol{y}^0 \text{ arbitrary}, \quad \boldsymbol{\mu}^0 \text{ arbitrary},\\
&\text{for } k = 0, 1, 2, \ldots\\
&\quad \boldsymbol{x}^{k+1} := \arg\min_{\boldsymbol{x}} \ \mathcal{L}_A(\boldsymbol{x}, \boldsymbol{y}^k; \boldsymbol{\mu}^k),\\
&\quad \boldsymbol{y}^{k+1} := \arg\min_{\boldsymbol{y}} \ \mathcal{L}_A(\boldsymbol{x}^{k+1}, \boldsymbol{y}; \boldsymbol{\mu}^k),\\
&\quad \boldsymbol{\mu}^{k+1} := \boldsymbol{\mu}^k + \rho\,(\boldsymbol{A}\boldsymbol{x}^{k+1} + \boldsymbol{B}\boldsymbol{y}^{k+1} - \boldsymbol{c}).
\end{aligned}$$

21.3.2 Specialization to Rank–Sparsity

Here the constraint is $\boldsymbol{L} + \boldsymbol{M} = \boldsymbol{A}$; the augmented Lagrangian is

$$\mathcal{L}_A(\boldsymbol{L}, \boldsymbol{M}; \boldsymbol{\Lambda}) = \|\boldsymbol{L}\|_{\ell_1} + \gamma\|\boldsymbol{M}\|_* + \langle \boldsymbol{\Lambda}, \boldsymbol{L} + \boldsymbol{M} - \boldsymbol{A}\rangle + \frac{\rho}{2}\|\boldsymbol{L} + \boldsymbol{M} - \boldsymbol{A}\|_{\mathrm{F}}^2,$$

where $\boldsymbol{\Lambda}$ is the matrix of Lagrange multipliers and $\langle\cdot,\cdot\rangle$ is the Frobenius inner product.

Step 1: $\boldsymbol{L}$-update.

$$\begin{aligned}
\boldsymbol{L}^{k+1} &:= \arg\min_{\boldsymbol{L}} \ \mathcal{L}_A(\boldsymbol{L}, \boldsymbol{M}^k; \boldsymbol{\Lambda}^k)\\
&= \arg\min_{\boldsymbol{L}} \left\{ \|\boldsymbol{L}\|_{\ell_1} + \langle \boldsymbol{\Lambda}^k, \boldsymbol{L}\rangle + \frac{\rho}{2}\|\boldsymbol{L} + \boldsymbol{M}^k - \boldsymbol{A}\|_{\mathrm{F}}^2 + (\text{constants}) \right\}.
\end{aligned}$$

Complete the square:

$$\boldsymbol{L}^{k+1} = \arg\min_{\boldsymbol{L}} \left\{ \|\boldsymbol{L}\|_{\ell_1} + \frac{\rho}{2}\left\|\boldsymbol{L} + \boldsymbol{M}^k - \boldsymbol{A} + \frac{1}{\rho}\boldsymbol{\Lambda}^k\right\|_{\mathrm{F}}^2 + (\text{constants}) \right\}.$$

Thus

$$\boldsymbol{L}^{k+1} = \operatorname{prox}_{\frac{1}{\rho}\|\cdot\|_{\ell_1}}\left(-\boldsymbol{M}^k + \boldsymbol{A} - \frac{1}{\rho}\boldsymbol{\Lambda}^k\right),$$

i.e. entrywise soft–thresholding with parameter $1/\rho$.

Step 2: $\boldsymbol{M}$-update.

$$\begin{aligned}
\boldsymbol{M}^{k+1} &:= \arg\min_{\boldsymbol{M}} \ \mathcal{L}_A(\boldsymbol{L}^{k+1}, \boldsymbol{M}; \boldsymbol{\Lambda}^k)\\
&= \arg\min_{\boldsymbol{M}} \left\{ \gamma\|\boldsymbol{M}\|_* + \langle \boldsymbol{\Lambda}^k, \boldsymbol{M}\rangle + \frac{\rho}{2}\|\boldsymbol{L}^{k+1} + \boldsymbol{M} - \boldsymbol{A}\|_{\mathrm{F}}^2 \right\}\\
&= \arg\min_{\boldsymbol{M}} \left\{ \gamma\|\boldsymbol{M}\|_* + \frac{\rho}{2}\left\|\boldsymbol{M} + \boldsymbol{L}^{k+1} - \boldsymbol{A} + \frac{1}{\rho}\boldsymbol{\Lambda}^k\right\|_{\mathrm{F}}^2 \right\}.
\end{aligned}$$

So

$$\boldsymbol{M}^{k+1} = \operatorname{prox}_{\frac{\gamma}{\rho}\|\cdot\|_*}\left(-\boldsymbol{L}^{k+1} + \boldsymbol{A} - \frac{1}{\rho}\boldsymbol{\Lambda}^k\right),$$

i.e. singular value soft–thresholding (using economy-sized SVD).

Step 3: Multiplier update.

$$\boldsymbol{\Lambda}^{k+1} := \boldsymbol{\Lambda}^k + \rho(\boldsymbol{L}^{k+1} + \boldsymbol{M}^{k+1} - \boldsymbol{A}).$$

21.3.3 ADMM Motivation: Method of Multipliers

Consider

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0},$$

with $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^p$.

The augmented Lagrangian is

$$\mathcal{L}_A(\boldsymbol{x}; \boldsymbol{\mu}) = f(\boldsymbol{x}) + \boldsymbol{\mu}^\top \boldsymbol{h}(\boldsymbol{x}) + \frac{1}{2}\rho\|\boldsymbol{h}(\boldsymbol{x})\|_2^2.$$

Given $\hat{\boldsymbol{\mu}}$, define $\hat{\boldsymbol{x}}$ by

$$\nabla_{\boldsymbol{x}} \mathcal{L}_A(\hat{\boldsymbol{x}}; \hat{\boldsymbol{\mu}}) = \boldsymbol{0},$$

that is,

$$\nabla f(\hat{\boldsymbol{x}}) + \boldsymbol{J}(\hat{\boldsymbol{x}})^\top \hat{\boldsymbol{\mu}} + \rho\, \boldsymbol{J}(\hat{\boldsymbol{x}})^\top \boldsymbol{h}(\hat{\boldsymbol{x}}) = \boldsymbol{0},$$

where

$$\boldsymbol{J}(\hat{\boldsymbol{x}}) = \begin{pmatrix} \nabla h_1(\hat{\boldsymbol{x}})^\top \\ \vdots \\ \nabla h_p(\hat{\boldsymbol{x}})^\top \end{pmatrix} \in \mathbb{R}^{p \times n}$$

is the Jacobian.

Compare with the KKT conditions:

$$\begin{aligned} &\boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0}, \\ &\nabla f(\boldsymbol{x}^*) + \boldsymbol{J}(\boldsymbol{x}^*)^\top \boldsymbol{\mu}^* = \boldsymbol{0}. \end{aligned}$$

If we define

$$\boldsymbol{\mu}^{\text{new}} := \hat{\boldsymbol{\mu}} + \rho\, \boldsymbol{h}(\hat{\boldsymbol{x}}),$$

the stationarity equation above resembles the KKT stationarity equation with updated multiplier.

Thus the **method of multipliers** (augmented Lagrangian method) updates

$$\begin{aligned} \boldsymbol{x}^{k+1} &:= \arg\min_{\boldsymbol{x}} \ \mathcal{L}_A(\boldsymbol{x}; \boldsymbol{\mu}^k), \\ \boldsymbol{\mu}^{k+1} &:= \boldsymbol{\mu}^k + \rho\, \boldsymbol{h}(\boldsymbol{x}^{k+1}). \end{aligned}$$

ADMM can be viewed as applying this idea but approximating the primal minimization by alternating minimization over blocks of variables ($\boldsymbol{x}$-block, $\boldsymbol{y}$-block), which leads to the simple, proximal-based updates used in rank–sparsity decomposition.

Lec 22

22.1 CVX

22.1.1 Semidefinite Programming

CVX is a modeling language for convex optimization.

It accepts a high-level description of an optimization problem, and translates it to an input suitable for a solver; invokes solver; returns optimal values to user's code.

The solvers in CVX are for **semidefinite programming (SDP)**.

SDP

$$\min\langle \boldsymbol{C}, \boldsymbol{X}\rangle \quad \text{S.T.} \quad \langle \boldsymbol{A}_1, \boldsymbol{X}\rangle = b_1, \ldots, \langle \boldsymbol{A}_m, \boldsymbol{X}\rangle = b_m,\ \boldsymbol{X} \succeq 0$$

where $\boldsymbol{X} \in \mathbb{S}^n$, given $\boldsymbol{C}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_m \in \mathbb{S}^n$, $b_1, \ldots, b_m \in \mathbb{R}$.

Fact

Fact. All of the convex programs considered this semester (ℓ_1 LS, SVM, Huber, Candès-Recht, Rank-Sparsity) are special cases of SDP.

CVX is equipped with rules for translating many convex problems to SDP.

22.1.2 Examples

Example

(Setting up Candès-Recht relaxation of matrix completion in CVX.)
Assume we have $\boldsymbol{M} \in \mathbb{R}^{m\times n}$, say $\boldsymbol{M}(i,j) = \text{NaN}$ for missing entries. So $\boldsymbol{M}, m, n$ are already program variables.

```
cvx_begin
variable X(m, n)
minimize norm_nuc(X)
subject to
for i = 1 :  m
  for j = 1 :  n
    if  ~  isnan(M(i, j))
      X(i, j) == M(i, j)
    end
  end
end
cvx_end
```

This code segment is inserted inside a MATLAB function. The variables $\boldsymbol{M}$, m, and n are defined

before the segment. The variable $\boldsymbol{X}$ will be defined upon completion.

Example

(Noisy C-R relaxation.)
$\boldsymbol{M}, m, n$ as above, `gamma` also defined.

```
[is, js] = find( ~ isnan(M))
nk = length(is)
cvx_begin
variable X(m, n)
variable z(nk)
minimize (quad_form(z, eye(nk)) / 2 + gamma * norm_nuc(X))
subject to
for k = 1 :  nk
  z(k) == X(is(k), js(k)) - M(is(k), js(k))
end
cvx_end
```

CVX formulation of $\min \frac{1}{2}\|\boldsymbol{z}\|^2 + \gamma\|\boldsymbol{X}\|_*$ subject to $z[s(i,j)] = \boldsymbol{X}(i,j) - \boldsymbol{M}(i,j),\ \forall (i,j) \in \Omega$, where $s(i,j)$ is sequential numbering $(i,j) \in \Omega$.

CVX implements **Disciplined Convex Programming (DCP)**. CVX determines convexity using syntactic rules.

Example

Consider

$$\min \langle \text{some function} \rangle + \gamma \sum_{i=1}^{n} t_i \quad \text{S.T.} \quad t_i \geq x_i,\ t_i \geq -x_i,\ \forall i = 1:n.$$

OK in CVX: write as above.
NOT OK in CVX: $t_i == \text{abs}(x_i)$.

22.1.3 Second-order Methods

CVX standard solvers are **Interior-point Methods** (CO 663 / CO 666).

Interior-point methods are second-order methods. On each iteration, solve large system of linear equations. Obtain high accuracy, fast convergence, but each iteration is costly in terms of space and time.

The methods in this class are all first-order methods, that is, **gradient-based**, no matrices (except matrices native to problem like C-R relaxation). First-order methods have slower convergence, lower accuracy, but much faster iterations.

22.2 Nonconvex Optimization

22.2.1 Solvers

1. Convex relaxation

(i) C-R relaxation of matrix completion

(ii) Relaxation of rank-sparsity

(iii) Compressive sensing

Nonconvex problem:

$$\min \ \ \mathrm{nnz}(\boldsymbol{x}) \quad \text{S.T.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

where $\mathrm{nnz}(\cdot)$ is the function of the number of nonzero entries. This is NP-hard.

Convex relaxation:

$$\min \ \ \|\boldsymbol{x}\|_1 \quad \text{S.T.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}.$$

2. Alternating algorithms Regression: given $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{y} \in \mathbb{R}^m$, want

$$\boldsymbol{x} := \arg\min\{\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|\}.$$

Robust Regression: ℓ entries of $\boldsymbol{y}$ are arbitrarily corrupted. Still want $\boldsymbol{x}$ fits the noncorrupted entries.

AMRR-Alternating Minimization for robust regression.

Initialize: guess ℓ corrupted entries randomly, for example. Let S^0 be the guess of uncorrupted entries, so

$$S^0 \subseteq \{1, \ldots, m\}, \ |S^0| = m - \ell.$$

Now solve:

$$\boldsymbol{x}^0 := \arg\min \|\boldsymbol{A}(S^0, :)\boldsymbol{x} - \boldsymbol{y}(S^0)\|.$$

Let $S^1 :=$ indices of the $m - \ell$ smallest entries of $|r_1^0|, \ldots, |r_m^0|$ where $\boldsymbol{r}^0 = \boldsymbol{A}\boldsymbol{x}^0 - \boldsymbol{y}$.

Continue alternating: find $\boldsymbol{x}^0, S^1, \boldsymbol{x}^1, S^2, \boldsymbol{x}^2, \ldots$ until convergence.

Theorem

(Bhatia, Jain, Kar) AMRR yields an $\boldsymbol{x}^k$ such that

$$\|\boldsymbol{x}^k - \boldsymbol{x}^*\| \le \varepsilon$$

in $\mathcal{O}(\log \|\boldsymbol{q}^*\|/\varepsilon)$ iterations under some fairly strong assumptions on $\boldsymbol{A}$. Here $\boldsymbol{q}^*$ is the residual vector of $\boldsymbol{x}^*$.

From experience, AMRR performs well for a problem arising in an image application, where assumptions on $\boldsymbol{A}$ may not hold.

Nonnegative Matrix Factorization (NMF):

(Gillis Book) Given $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{A} \geq 0$ (entrywise), given $r \leq \min(m, n)$. Seek $\boldsymbol{W}, \boldsymbol{H}$ such that $\boldsymbol{W} \in \mathbb{R}^{m \times r}, \boldsymbol{H} \in \mathbb{R}^{r \times n}$, $\boldsymbol{W} \geq 0, \boldsymbol{H} \geq 0, \boldsymbol{A} \approx \boldsymbol{W}\boldsymbol{H}$.

Example

(A term-doc matrix.) Interpretation: each document is a nonnegative sum of topics, so more interpretable than SVD.
Other applications of NMF: **Blind Sound Source Separation, Hyperspectral Imaging, Microarray Experiments**.

3. *GD* / *SGD* / Conjugate Gradient (Gradient Descent / Stochastic Gradient Descent / Conjugate Gradient.)

Lec 23

23.1 Alternating algorithms

23.1.1 Nonnegative Matrix Factorization

Theorem

Given $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, given $r \leq \min(m,n)$ (some variants: NMF figures out r), want $\boldsymbol{W} \in \mathbb{R}^{m\times r}, \boldsymbol{H} \in \mathbb{R}^{r\times n}, \boldsymbol{W} \geq \boldsymbol{0}, \boldsymbol{H} \geq \boldsymbol{0}, \boldsymbol{A} \approx \boldsymbol{W}\boldsymbol{H}$.

Claim

Geometry of NMF:
Say each column of $\boldsymbol{A}$ normalized in 1-norm (sum of entries of each column equals 1)
In this case, we can assume that columns of $\boldsymbol{W}, \boldsymbol{H}$ are similarly normalized.
Why?
Say $\boldsymbol{A} \approx \boldsymbol{W}\boldsymbol{H}$.
Rescale columns of $\boldsymbol{W}$ to have 1-norm to 1, that is, define $\bar{\boldsymbol{W}} = \boldsymbol{W}\boldsymbol{D}$, $\boldsymbol{D}$ is diagonal, $\boldsymbol{D} \succ \boldsymbol{0}$, entries chosen so that $\underbrace{\boldsymbol{e}^\top}_{\mathbb{R}^m}\bar{\boldsymbol{W}} = \underbrace{\boldsymbol{e}^\top}_{\mathbb{R}^r}$.
Let $\bar{\boldsymbol{H}} = \boldsymbol{D}^{-1}\boldsymbol{H}$ so that $\boldsymbol{e}^\top\bar{\boldsymbol{W}}\bar{\boldsymbol{H}} = \boldsymbol{e}^\top\bar{\boldsymbol{H}}$.
We know $\bar{\boldsymbol{W}}\bar{\boldsymbol{H}} = \boldsymbol{W}\boldsymbol{H} \approx \boldsymbol{A}$, and $\underbrace{\boldsymbol{e}^\top}_{\mathbb{R}^m}\boldsymbol{A} = \underbrace{\boldsymbol{e}^\top}_{\mathbb{R}^n}$ by assumption.
So may as well take columns of $\bar{\boldsymbol{H}}$ to have 1-norm equal to 1. ($\boldsymbol{e}^\top\bar{\boldsymbol{H}} \approx \boldsymbol{e}^\top\boldsymbol{A} = \boldsymbol{e}^\top$)
Under assumption that columns of $\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{A}$ are all normalized, NMF looks as follows:

$$\Delta^m = m\text{-dimensional standard simplex} = \{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{e}^\top\boldsymbol{x} = 1, \boldsymbol{x} \geq \boldsymbol{0}\}.$$

Columns of $\boldsymbol{A}$ lie in Δ^m as do columns of $\boldsymbol{W}$.
Columns of $\boldsymbol{H}$ are coefficients of convex combinations.
$\boldsymbol{A} = \boldsymbol{W}\boldsymbol{H}$ means each column of $\boldsymbol{A}$ is a convex combination of columns of $\boldsymbol{W}$.

When posed as optimization, for example

$$\min \ \|\boldsymbol{A} - \boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}} \quad \text{S.T.} \quad \boldsymbol{W} \in \mathbb{R}^{m\times r}, \boldsymbol{H} \in \mathbb{R}^{r\times n}, \boldsymbol{W} \geq \boldsymbol{0}, \boldsymbol{H} \geq \boldsymbol{0}$$

is NP-hard.

Special case: columns of $\boldsymbol{W}$ appear as columns of $\boldsymbol{A}$.
In context of term-doc matrices, for each topic there is a *pure document* only about that topic, and it is comprehensive.
This is called *Separable Case* in literature.
Gillis and Vavasis proposed algorithm **SPA** to solve separable with noise.

For nonseparable case: **Alternating Nonnegative Least Squares (ANLS)**.
If $\boldsymbol{W}$ is given, optimal $\boldsymbol{H}$ solves:

$$\arg\min_{\boldsymbol{H}} \{\|\boldsymbol{A}-\boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}} : \boldsymbol{H}\in\mathbb{R}^{r\times n},\ \boldsymbol{H}\geq\boldsymbol{0}\}$$

is convex in $\boldsymbol{H}$.
In fact, this is quadratic programming.
Even better: decouples into n problems of size r, for $k=1:n$

$$\arg\min_{\boldsymbol{H}(:,k)} \left\{\|\boldsymbol{A}(:,k)-\boldsymbol{W}\boldsymbol{H}(:,k)\|^2 : \boldsymbol{H}(:,k)\geq\boldsymbol{0}\right\}.$$

ANLS: alternately optimize for $\boldsymbol{H}$ as above, then fix $\boldsymbol{H}$ and optimize for $\boldsymbol{W}$.
Optimizing for $\boldsymbol{W}$ yields m quadratic programming problems each of r variables.
This could be slow.

Another faster algorithm in practice: **Hierarchical Alternating Least Squares (HALS)**.
Consider row ℓ of $\boldsymbol{H}$.
Fix all other entries of $\boldsymbol{H}$.
Optimal choice for row:

$$\arg\min_{\boldsymbol{H}(\ell,:)\geq 0} \left\|\boldsymbol{A}-\sum_{k\neq\ell}\boldsymbol{W}(:,k)\boldsymbol{H}(k,:)-\boldsymbol{W}(:,\ell)\boldsymbol{H}(\ell,:)\right\|_{\mathrm{F}}^2.$$

Simpler notation: $\boldsymbol{h}:=\boldsymbol{H}(\ell,:)^\top$, $\boldsymbol{w}:=\boldsymbol{W}(:,\ell)$, $\boldsymbol{B}^\ell=\boldsymbol{A}-\sum_{k\neq\ell}\boldsymbol{W}(:,k)\boldsymbol{H}(k,:)$,

$$\arg\min_{\boldsymbol{h}\geq\boldsymbol{0}} \left\{\|\boldsymbol{B}^\ell-\underbrace{\boldsymbol{w}\boldsymbol{h}^\top}_{\mathbb{R}^{m\times n}}\|_{\mathrm{F}}^2\right\}.$$

The problem separates over j:

$$\arg\min_{h_j\geq 0}\left\|\boldsymbol{B}^\ell(:,j)-\boldsymbol{w}h_j\right\|^2 = \arg\min_{h_j\geq 0}\left\{\boldsymbol{w}^\top\boldsymbol{w}\,h_j^2-2\boldsymbol{w}^\top\boldsymbol{B}^\ell(:,j)h_j+\|\boldsymbol{B}^\ell(:,j)\|^2\right\}.$$

This is a parabola:

$$h_j=\max\left\{0,\frac{\boldsymbol{w}^\top\boldsymbol{B}^\ell(:,j)}{\boldsymbol{w}^\top\boldsymbol{w}}\right\}.$$

So HALS applies this formula to update every entry of $\boldsymbol{H}$, then every entry of $\boldsymbol{W}$, then back to $\boldsymbol{H}$, etc.

Both HALS and ANLS have a descent property: $\|\boldsymbol{A}-\boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}}$ decreases on every iteration.
To make HALS efficient, must compute $\boldsymbol{B}^\ell$ via updating previous values.

23.2 Gradient Descent and Related Methods

23.2.1 *GD*

Suppose *GD* applied to $f : \mathbb{R}^n \to \mathbb{R}$, L-smooth, not necessarily convex.
Assume f bounded below by $f_{\min}$.
Then after k iterations of *GD* with stepsize $\frac{1}{L}$,

$$\min_{0 \le j \le k-1} \|\nabla f(\boldsymbol{x}^j)\| \le \sqrt{\frac{2L(f(\boldsymbol{x}^0) - f_{\min})}{k}}$$

also,

$$\lim_{k \to \infty} \nabla f(\boldsymbol{x}^k) = \boldsymbol{0}.$$

What if L not known?
Can use **Backtrack Line-Search** (Problem Set 3):

$$\boldsymbol{x}^{k+1} := \boldsymbol{x}^k - 2^{\ell_k} \nabla f(\boldsymbol{x}^k)$$

where $\ell_k \in \mathbb{Z}$ determined adaptively (Armijo's Condition).
Assuming f is L-smooth, previous theorem holds with worse constants for Backtrack Line-Search.
Flaw with Backtrack Line-Search: each iteration is more expensive than using step-size $\frac{1}{L}$, because of the need to evaluate $f(\boldsymbol{x}^k - 2^{\ell} \nabla f(\boldsymbol{x}^k))$ for multiple values of ℓ for each k.
Published Line-Search Algorithms are more efficient than *Backtrack Line-Search*.
In convex case, can improve on *GD* via *AGD*. But *AGD* is not usable in nonconvex cases.

23.2.2 *CGD*

Nonlinear Conjugate Gradient Descent (*CGD*) often outperforms *GD*.
(Linear) Conjugate Gradient (Hestenes and Stiefel 1952) solves

$$\min \underbrace{\frac{1}{2} \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}}_{f(\boldsymbol{x})}, \quad \boldsymbol{A} \in \mathbb{S}^n, \ \boldsymbol{A} \succ 0$$

which is same as solving $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

Given $\boldsymbol{x}_0 \in \mathbb{R}^n$ arbitrarily,
(1) $\boldsymbol{g}_0 := \boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b}$ $\quad (\ = \nabla f(\boldsymbol{x}_0)\ )$
(2) $\boldsymbol{p}_0 := -\boldsymbol{g}_0$

for $k = 0, 1, 2, \ldots$

(3) $\alpha_k := \dfrac{-\boldsymbol{g}_k^\top \boldsymbol{p}_k}{\boldsymbol{p}_k^\top \boldsymbol{A} \boldsymbol{p}_k}$
(4) $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$
(5) $\boldsymbol{g}_{k+1} := \boldsymbol{g}_k + \alpha_k \boldsymbol{A} \boldsymbol{p}_k$ $\quad (\ = \boldsymbol{A}\boldsymbol{x}_{k+1} - \boldsymbol{b} = \nabla f(\boldsymbol{x}_{k+1})\ )$

(6) $\beta_{k+1} := \dfrac{\boldsymbol{g}_{k+1}^\top \boldsymbol{g}_{k+1}}{\boldsymbol{g}_k^\top \boldsymbol{g}_k}$
(7) $\boldsymbol{p}_{k+1} := -\boldsymbol{g}_{k+1} + \beta_{k+1}\boldsymbol{p}_k$.

Step (3) is called **Exact Line-Search**

$$\alpha_k := \arg\min_{\alpha} \ f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k).$$

Exact Line-Search is generally not used in optimization because of little benefit for expensive computation (use Backtrack Line-Search instead), however in current settings it's OK.
Fact: in exact arithmetic, *CG* returns the exact minimizer of f after $\leq n$ iterations.
Thanks to success of *CG*, there are many attempts to extend to an arbitrary f.

(Fletcher-Reeves 1960s.)

- replace $\boldsymbol{g}_k$ with $\nabla f(\boldsymbol{x}_k)$
- use inexact line-search for α_k
- many formulas for β_k
- occasional restart: set $\beta_{k+1} := 0$

Say μ is the modulus of strong convexity, L is the modulus of smoothness.

(Nonlinear) *CGD*	(Convex Functions) *AGD*
Quadratics	Convex functions
$\left(1 - \sqrt{\frac{\mu}{L}}\right)^k$ (better constant)	$\left(1 - \sqrt{\frac{\mu}{L}}\right)^k$
No guarantees for nonquadratics	Guaranteed for all convex functions
No hyperparameters	Needs μ, L

Lec 24

24.1 Motivation

A software package to embed high-dimensional data in a lower dimension.

Example usage: if embed into 2 dimensions, then obtain a visual representation of data.

Premise: realistic high dimensional data actually lies on a lower dimensional manifold (a priori unknown). Goal is to identify the manifold.

If low dimensional manifold is a linear subspace, then can be found via SVD. If it is an affine subspace, then translate data by its mean and apply SVD. Both are called **Principal Components Analysis (PCA)**.

If manifold is curved, need something like UMAP.

24.2 UMAP Algorithm

Given n vectors in $\mathbb{R}^d$,

(i) Compute k-nearest-neighbor graph on data. For each data item, find its closest neighbor, second-closest neighbor, etc up to k. This yields a directed graph with distances on arcs.

(ii) Assign weights to edges based on distances.

(iii) Symmetrize the graph (yielding an undirected graph).

(iv) Use **Spectral Method** to produce an embedding.

(v) *SGD* to improve spectral embedding.

Step 4 Spectral Embedding

Let $\boldsymbol{p}_i \in \mathbb{R}^d, \forall i = 1 : n$. The **Euclidean Distance Matrix (EDM)** is an $n \times n$ symmetric matrix whose (i, j) entry is $\|\boldsymbol{p}_i - \boldsymbol{p}_j\|^2$.

The **Gram Matrix** is also $n \times n$ symmetric, (i, j) entry is $\boldsymbol{p}_i^\top \boldsymbol{p}_j$. Given Gram Matrix, it's straightforward to obtain EDM:

$$\|\boldsymbol{p}_i - \boldsymbol{p}_j\|^2 = \boldsymbol{p}_i^\top \boldsymbol{p}_i - 2\boldsymbol{p}_i^\top \boldsymbol{p}_j + \boldsymbol{p}_j^\top \boldsymbol{p}_j.$$

To convert an EDM to a Gram Matrix is more complicated. Must assume: $\sum_{i=1}^n \boldsymbol{p}_i = \boldsymbol{0}$.

Add up one row of EDM, say row i:

$$\sum_{j=1}^n \|\boldsymbol{p}_i - \boldsymbol{p}_j\|^2 = n\boldsymbol{p}_i^\top \boldsymbol{p}_i + \sum_{j=1}^n \boldsymbol{p}_j^\top \boldsymbol{p}_j.$$

Sum this over i, letting $\rho = \sum_{j=1}^n \boldsymbol{p}_j^\top \boldsymbol{p}_j$, obtain $2n\rho$.

So from sum of EDM entries, we obtain ρ. Once ρ is known, we obtain $\boldsymbol{p}_i^\top \boldsymbol{p}_i$ for all i, then obtain $\boldsymbol{p}_i^\top \boldsymbol{p}_j$ for all i, j by

$$\|\boldsymbol{p}_i - \boldsymbol{p}_j\|^2 = \boldsymbol{p}_i^\top \boldsymbol{p}_i - 2\boldsymbol{p}_i^\top \boldsymbol{p}_j + \boldsymbol{p}_j^\top \boldsymbol{p}_j.$$

Thus all entries of Gram Matrix are recovered.

From Gram Matrix, we can almost recover $\boldsymbol{p}_i$'s.

Let

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{p}_1^\top \\ \vdots \\ \boldsymbol{p}_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \qquad \boldsymbol{G} = \boldsymbol{P}\boldsymbol{P}^\top.$$

Observe: $\operatorname{rank}(\boldsymbol{G}) \leq d$.

Let $\boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top$ be eigendecomposition of $\boldsymbol{G}$. At most d entries of $\boldsymbol{D}$ are nonzero, say entries $1:d$.

$$\boldsymbol{G} = \boldsymbol{Q}(:, 1\!:\!d)\, \boldsymbol{D}(1\!:\!d, 1\!:\!d)\, \boldsymbol{Q}(:, 1\!:\!d)^\top = \bar{\boldsymbol{P}}\, \bar{\boldsymbol{P}}^\top,$$

where $\bar{\boldsymbol{P}} := \boldsymbol{Q}(:, 1\!:\!d)\, \boldsymbol{D}(1\!:\!d, 1\!:\!d)^{1/2}$.

Fact: if $\boldsymbol{G} = \boldsymbol{P}\boldsymbol{P}^\top = \bar{\boldsymbol{P}}\, \bar{\boldsymbol{P}}^\top$ for $\boldsymbol{P}, \bar{\boldsymbol{P}} \in \mathbb{R}^{n \times d}$, then exists orthogonal $\boldsymbol{Q} \in \mathbb{R}^{d \times d}$ such that $\bar{\boldsymbol{P}} = \boldsymbol{P}\boldsymbol{Q}$.

So given EDM $\boldsymbol{E}$, recover $\boldsymbol{P}$ via

$$\boldsymbol{E} \xrightarrow{\sum \boldsymbol{p}_i = \boldsymbol{0}} \boldsymbol{G} \xrightarrow{\text{Eigendecomposition}} \boldsymbol{P} \qquad \text{(up to unknown rotation)}.$$

Given a data set with knowledge of pairwise distances:

(i) Square distances.

(ii) Form *Fake EDM*.

(iii) Form Gram Matrix.

(iv) Form Eigendecomposition.

(v) Keep top d eigenvalues (d is the desired embedding dimension).

Eigenvectors yield coordinates in $\mathbb{R}^d$.

Step 5 *SGD*

Say p, q are probability distributions over same set of events. **Cross-Entropy** of q with respect to p is $-\mathbb{E}_p[\log q]$.

Example: flipping a coin with true heads probability p but guessed value q.

$$\operatorname{CE}(p, q) = -p \log q - (1 - p) \log(1 - q).$$

This is convex in q, minimized at $p = q$.

UMAP idea: assume a hidden undirected graph among data points with edge probabilities from Step 3. For embedded points $\boldsymbol{x}_i \in \mathbb{R}^d$,

$$\Pr[(i,j) \text{ edge}] = \frac{1}{1 + a\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^{2b}},$$

with hyperparameters $a, b > 0$.

UMAP minimizes Cross-Entropy:

$$\min_{\boldsymbol{x}} \; - \sum_{e=(i,j)} \mathbb{E}_{p_e}[\log q_e].$$

UMAP uses a variant of *SGD*: on each iteration pick edge (i,j) with probability proportional to true weight and update $\boldsymbol{x}_i, \boldsymbol{x}_j$ using gradient of

$$-p_{ij} \log\left(\frac{1}{1 + a\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^{2b}}\right) - (1 - p_{ij}) \log\left(1 - \frac{1}{1 + a\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^{2b}}\right).$$

Then perform **Negative Sampling**. Choose random i, j with weight 0 and update similarly.

Negative sampling originates from **Word2Vec** by Mikolov et al.

Does this really minimize Cross-Entropy?

Unsure..(similar but not same)

A pattern in *SGD* for modern machine learning: start with a principled approach, then insert many heuristics (layer normalizations, etc), then keep trying until it works well.