A : \( k \times n \). Usually \( k > n \)

otherwise easily the minimum is zero.

Analytical solution:

\[
\begin{align*}
f(x) &= (Ax - b)^T (Ax - b) \\
&= x^T A^T Ax - 2b^T Ax + b^T b
\end{align*}
\]

\[
\nabla f(x) = 2A^T Ax - 2A^T b = 0
\]
Regularization, weights:

\[
\frac{1}{2} \lambda x^T x + w_1 (Ax - b)^2_1 + \cdots + w_k (Ax - b)^2_k
\]
Convex hull is convex

\[ x = \theta_1 x_1 + \cdots + \theta_k x_k \]
\[ \bar{x} = \bar{\theta}_1 \bar{x}_1 + \cdots + \bar{\theta}_k \bar{x}_k \]

Then

\[ \alpha x + (1 - \alpha)\bar{x} = \alpha \theta_1 x_1 + \cdots + \alpha \theta_k x_k + \]
\[ (1 - \alpha)\bar{\theta}_1 \bar{x}_1 + \cdots + (1 - \alpha)\bar{\theta}_k \bar{x}_k \]
Each coefficient is nonnegative and

\[ \alpha \theta_1 + \cdots + \alpha \theta_k + (1 - \alpha) \bar{\theta}_1 + \cdots + (1 - \alpha) \bar{\theta}_k = \alpha + (1 - \alpha) = 1 \]
We prove that any

\[ x = x_c + Au \text{ with } \|u\|_2 \leq 1 \]

satisfies

\[ (x - x_c)^T P^{-1} (x - x_c) \leq 1 \]

Let

\[ A = P^{1/2} \]

because \( P \) is symmetric positive definite. Then

\[ u^T A^T P^{-1} Au = u^T P^{1/2} P^{-1} P^{1/2} u \leq 1. \]
$S^+_n$ is a convex cone. Let $X_1, X_2 \in S^+_n$. For any $\theta_1 \geq 0, \theta_2 \geq 0$,

$$z^T(\theta_1 X_1 + \theta_2 X_2)z = \theta_1 z^T X_1 z + \theta_2 z^T X_2 z \geq 0$$
Example:

\[
\begin{bmatrix}
    x & y \\
    y & z
\end{bmatrix} \in \mathbf{S}_+^2
\]

is equivalent to

\[x \geq 0, z \geq 0, xz - y^2 \geq 0\]

If \(x > 0\) or \((z > 0)\) is fixed, we can see that

\[z \geq \frac{y^2}{x}\]

has a parabolic shape
When $t$ is fixed,

$$\{(x_1, x_2) \mid -1 \leq x_1 \cos t + x_2 \cos 2t \leq 1\}$$

gives a region between two parallel lines.

This region is convex.
**f(\(S\)) is convex:**

Let

\[
\begin{align*}
    f(x_1) &\in f(S), f(x_2) \in f(S) \\
    \alpha f(x_1) + (1 - \alpha)f(x_2) &\in f(S) \\
    \alpha(Ax_1 + b) + (1 - \alpha)(Ax_2 + b) &\in f(S)
\end{align*}
\]
$f^{-1}(C)$ convex:

$x_1, x_2 \in f^{-1}(C)$

means that

$Ax_1 + b \in C, Ax_2 + b \in C$

Because $C$ is convex,

$\alpha(Ax_1 + b) + (1 - \alpha)(Ax_2 + b)$

$= A(\alpha x_1 + (1 - \alpha) x_2) + b \in C$

Thus

$\alpha x_1 + (1 - \alpha) x_2 \in f^{-1}(C)$
Scaling:
\[ \alpha S = \{ \alpha x \mid x \in S \} \]

Translation
\[ S + a = \{ x + a \mid x \in S \} \]

Projection
\[ T = \{ x_1 \in R^m \mid (x_1, x_2) \in S, x_2 \in R^n \}, S \subseteq R^m \times R^n \]

Scaling, translation, and projection are all affine functions
For example, for projection

\[ f(x) = \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 0 \]

- Solution set of linear matrix inequality

\[ C = \{ S \mid S \preceq 0 \} \text{ is convex} \]

\[ f(x) = x_1 A_1 + \cdots + x_m A_m - B = Ax + b \]

\[ f^{-1}(C) = \{ x \mid f(x) \preceq 0 \} \text{ is convex} \]
But this isn’t rigorous because of some problems in arguing

\[ f(x) = Ax + b \]

A more formal explanation:

\[ C = \{ s \in \mathbb{R}^{p^2} \mid \text{mat}(s) \in S^p \ \text{and} \ \text{mat}(s) \preceq 0 \} \]

is convex

\[
\begin{align*}
f(x) &= x_1 \text{vec}(A_1) + \cdots + x_m \text{vec}(A_m) - \text{vec}(B) \\
&= \left[ \begin{array}{c} \text{vec}(A_1) \\ \vdots \\ \text{vec}(A_m) \end{array} \right] x + (-\text{vec}(B))
\end{align*}
\]

\[ f^{-1}(C) = \{ x \mid \text{mat}(f(x)) \in S^p \ \text{and} \ \text{mat}(f(x)) \preceq 0 \} \]

is convex
Hyperbolic cone:

\[ C = \{(z, t) \mid z^T z \leq t^2, \, t \geq 0\} \]

is convex (by drawing a figure in 2 or 3 dimensional space)

We have that

\[ f(x) = \begin{bmatrix} P^{1/2}x \\ c^T x \end{bmatrix} \]
is affine. Then

\[ f^{-1}(C) = \{ x \mid f(x) \in C \} \]
\[ = \{ x \mid x^T P x \leq (c^T x)^2, c^T x \geq 0 \} \]

is convex
Image convex: if $S$ is convex, check if

\[ \{ P(x, t) \mid (x, t) \in S \} \]

convex or not

Note that $S$ is in the domain of $P$
Assume \((x_1, t_1), (x_2, t_2) \in S\)

We hope

\[\alpha \frac{x_1}{t_1} + (1 - \alpha) \frac{x_2}{t_2} = P(A, B),\]

where

\((A, B) \in S\)
We have

\[ \alpha \frac{x_1}{t_1} + (1 - \alpha) \frac{x_2}{t_2} = \frac{\alpha t_2 x_1 + (1 - \alpha) t_1 x_2}{t_1 t_2} = \frac{\alpha t_2 x_1 + (1 - \alpha) t_1 x_2}{\alpha t_1 t_2 + (1 - \alpha) t_1 t_2} \]

\[ = \frac{\alpha t_2}{\alpha t_2 + (1 - \alpha) t_1} x_1 + \frac{(1 - \alpha) t_1}{\alpha t_2 + (1 - \alpha) t_1} x_2 \]

\[ = \frac{\alpha t_2}{\alpha t_2 + (1 - \alpha) t_1} t_1 + \frac{(1 - \alpha) t_1}{\alpha t_2 + (1 - \alpha) t_1} t_2 \]
Let 

\[ \theta = \frac{\alpha t_2}{\alpha t_2 + (1 - \alpha) t_1} \]

We have 

\[ \frac{\theta x_1 + (1 - \theta) x_2}{\theta t_1 + (1 - \theta) t_2} = \frac{A}{B} \]

Further 

\((A, B) \in S\)

because 

\((x_1, t_1), (x_2, t_2) \in S\)
Perspective and linear-fractional function $V$

and

$S$ is convex

- Inverse image is convex
- Given $C$ a convex set

$$P^{-1}(C) = \{(x, t) \mid P(x, t) = x/t \in C\}$$

is convex
Perspective and linear-fractional function

VI

Let

\[(x_1, t_1) : x_1/t_1 \in C\]
\[(x_2, t_2) : x_2/t_2 \in C\]

Do we have

\[\theta(x_1, t_1) + (1 - \theta)(x_2, t_2) \in P^{-1}(C)\]?

That is,

\[\frac{\theta x_1 + (1 - \theta)x_2}{\theta t_1 + (1 - \theta)t_2} \in C?\]
Let
\[
\frac{\theta x_1 + (1 - \theta) x_2}{\theta t_1 + (1 - \theta) t_2} = \alpha \frac{x_1}{t_1} + (1 - \alpha) \frac{x_2}{t_2},
\]

Earlier we had
\[
\theta = \frac{\alpha t_2}{\alpha t_2 + (1 - \alpha) t_1}
\]

Then
\[
(\alpha (t_2 - t_1) + t_1) \theta = \alpha t_2
\]
Perspective and linear-fractional function

\[ t_1 \theta = \alpha t_2 - \alpha t_2 \theta + \alpha t_1 \theta \]

\[ \alpha = \frac{t_1 \theta}{t_1 \theta + (1 - \theta) t_2} \]
2-16: Generalized inequalities 1

- $K$ contains no line:

\[ \forall x \text{ with } x \in K \text{ and } -x \in K \Rightarrow x = 0 \]

- Nonnegative polynomial on $[0,1]$

- When $n = 2$

\[ x_1 \geq -tx_2, \forall t \in [0,1] \]

- $t = 1$
$t = 0$
\( \forall t \in [0,1] \)

- It really becomes a proper cone
Properties:

\[ x \preceq_K y, \ u \preceq_K v \]

implies that

\[ x - y \in K \]
\[ u - v \in K \]

From the definition of a convex cone,

\[ (x - y) + (u - v) \in K \]

Then

\[ x + u \preceq_K y + v \]
The minimum element

\[ S \subseteq x_1 + K \]

A minimal element

\[ (x_2 - K) \cap S = \{x_2\} \]
We consider a simplified situation and omit part of the proof.

Assume

$$\inf_{u \in C, v \in D} \|u - v\| > 0$$

and minimum attained at $c, d$

We will show that

$$a = d - c, \quad b = \frac{||d||^2 - ||c||^2}{2}$$

forms a separating hyperplane $a^T x = b$
2-19: Separating hyperplane theorem II

\[ a^T x = \Delta_2 \quad b \quad \Delta_1 \]

\[ a = d - c \]

\[ \Delta_1 = (d - c)^T d, \quad \Delta_2 = (d - c)^T c \]

\[ b = \frac{\Delta_1 + \Delta_2}{2} = \frac{d^T d - c^T c}{2} \]
Assume the result is wrong so there is $u \in D$ such that

$$a^T u - b < 0$$

We will derive a point $u'$ in $D$ but closer to $c$ than $d$. That is,

$$\|u' - c\| < \|d - c\|$$

Then we have a contradiction

- The concept
The separating hyperplane theorem IV states that if \( a^T x = b \), then \( a^T u - b = (d - c)^T u - \frac{d^T d - c^T c}{2} < 0 \) implies that

\[
(d - c)^T (u - d) + \frac{1}{2} \| d - c \|^2 < 0
\]
\[
\frac{d}{dt} \| d + t(u - d) - c \|^2 \bigg|_{t=0} = 2(d + t(u - d) - c)^T (u - d) \bigg|_{t=0} = 2(d - c)^T (u - d) < 0
\]

There exists a small \( t \in (0, 1) \) such that

\[
\| d + t(u - d) - c \| < \| d - c \|
\]

However,

\[
d + t(u - d) \in D,
\]

so there is a contradiction
Case 1: $C$ has an interior region

- Consider 2 sets:
  
  interior of $C$ versus $\{x_0\}$,

  where $x_0$ is any boundary point

- If $C$ is convex, then interior of $C$ is also convex

- Then both sets are convex

- We can apply results in slide 2-19 so that there exists $a$ such that

  \[ a^T x \leq a^T x_0, \forall x \in \text{interior of } C \]
Then for all boundary point \( x \) we also have

\[
a^T x \leq a^T x_0
\]

because any boundary point is the limit of interior points

Case 2: \( C \) has no interior region

In this situation, \( C \) is like a line in \( R^3 \) (so no interior). Then of course it has a supporting hyperplane
$A, X \in R^{m \times n}$

$$\text{tr}(A^T X) = \sum_j (A^T X)_{jj}$$

$$= \sum_j \sum_i A_{ji}^T X_{ij} = \sum_j \sum_i A_{ij} X_{ij}$$
An open set: for any $x$, there is a ball covering $x$ such that this ball is in the set.

Global underestimator:

$$
\frac{z - f(x)}{y - x} = f'(x)
$$

$$
z = f(x) + f'(x)(y - x)
$$
⇒: Because domain $f$ is convex,

for all $0 < t \leq 1$, $x + t(y - x) \in \text{domain } f$

\[
\begin{align*}
  f(x + t(y - x)) &\leq (1 - t)f(x) + tf(y) \\
  f(y) &\geq f(x) + \frac{f(x + t(y - x)) - f(x)}{t}
\end{align*}
\]

when $t \to 0$,

\[
f(y) \geq f(x) + \nabla f(x)^T (y - x)
\]
For any $0 \leq \theta \leq 1$,

\[ z = \theta x + (1 - \theta)y \]

\[ f(x) \geq f(z) + \nabla f(z)^T (x - z) \]
\[ = f(z) + \nabla f(z)^T (1 - \theta)(x - y) \]

\[ f(y) \geq f(z) + \nabla f(z)^T (y - z) \]
\[ = f(z) + \nabla f(z)^T \theta(y - x) \]

\[ \theta f(x) + (1 - \theta)f(y) \geq f(z) \]
First-order condition for strictly convex function:

\[ f \] is strictly convex if and only if

\[ f(y) > f(x) + \nabla f(x)^T(y - x) \]

\[ \iff \] it’s easy by directly modifying \( \geq \) to \( > \)

\[ f(x) > f(z) + \nabla f(z)^T(x - z) = f(z) + \nabla f(z)^T(1 - \theta)(x - y) \]

\[ f(y) > f(z) + \nabla f(z)^T(y - z) = f(z) + \nabla f(z)^T \theta(y - x) \]
3-7: First-order Condition V

⇒: Assume the result is wrong. From the 1st-order condition of a convex function, \( \exists x, y \) such that \( x \neq y \) and

\[
\nabla f(x)^T (y - x) = f(y) - f(x) \tag{1}
\]

For this \((x, y)\), from the strict convexity

\[
f(x + t(y - x)) - f(x) < tf(y) - tf(x) = \nabla f(x)^T t(y - x), \quad \forall t \in (0, 1)
\]
Therefore,

\[ f(x + t(y - x)) < f(x) + \nabla f(x)^T t(y - x), \forall t \in (0, 1) \]

However, this contradicts the first-order condition:

\[ f(x + t(y - x)) \geq f(x) + \nabla f(x)^T t(y - x), \forall t \in (0, 1) \]

This proof was given by a student of this course before.
Proof of the 2nd-order condition: We consider only the simpler condition of \( n = 1 \)

\[
\begin{align*}
\lim_{{t \to 0}} 2 \frac{f(x + t) - f(x) - f'(x)t}{t^2} &= \lim_{{t \to 0}} \frac{2(f'(x + t) - f'(x))}{2t} \\
&= f''(x) \geq 0
\end{align*}
\]
3-8: Second-order condition II

“⇐”

\[ f(x + t) = f(x) + f'(x)t + \frac{1}{2}f''(\bar{x})t^2 \geq f(x) + f'(x)t \]

by 1st-order condition

- The extension to general \( n \) is straightforward
- If \( \nabla^2 f(x) \succ 0 \), then \( f \) is strictly convex
Using 1st-order condition for strictly convex function:

\[ f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(\bar{x})(y - x) \]

Note that in our proof of 1st-order condition for strictly convex functions we used 2nd-order condition of convex function, so no contradiction here.
It’s possible that $f$ is strictly convex but

$$\nabla^2 f(x) \not\succ 0$$

Example:

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

Details omitted
Quadratic-over-linear

\[
\frac{\partial f}{\partial x} = \frac{2x}{y}, \quad \frac{\partial f}{\partial y} = -\frac{x^2}{y^2}
\]

\[
\frac{\partial^2 f}{\partial x \partial x} = \frac{2}{y}, \quad \frac{\partial^2 f}{\partial x \partial y} = -\frac{2x}{y^2}, \quad \frac{\partial^2 f}{\partial y \partial y} = \frac{2x^3}{y^3},
\]

\[
\frac{2}{y^3} \begin{bmatrix} y & -x \\ -x & 0 \end{bmatrix} \begin{bmatrix} y & -x \\ -x & x^2 \end{bmatrix} = 2 \begin{bmatrix} \frac{1}{y} & -\frac{x}{y^2} \\ -\frac{x}{y^2} & \frac{x^2}{y^3} \end{bmatrix}
\]
\[ f(x) = \log \sum_k \exp x_k \]

\[ \nabla f(x) = \begin{bmatrix} e^{x_1} \\ \sum_k e^{x_k} \\ \vdots \\ e^{x_n} \\ \sum_k e^{x_k} \end{bmatrix} \]

\[ \nabla^2_{ii} f = \frac{(\sum_k e^{x_k}) e^{x_i} - e^{x_i} e^{x_i}}{(\sum_k e^{x_k})^2}, \quad \nabla^2_{ij} f = \frac{-e^{x_i} e^{x_j}}{(\sum_k e^{x_k})^2}, \quad i \neq j \]

Note that if

\[ z_k = \exp x_k \]
then

$$(zz^T)_{ij} = z_i(z^T)_j = z_i z_j$$

Cauchy-Schwarz inequality

$$(a_1 b_1 + \cdots + a_n b_n)^2 \leq (a_1^2 + \cdots + a_n^2)(b_1^2 + \cdots + b_n^2)$$

$$a_k = v_k \sqrt{z_k}, \quad b_k = \sqrt{z_k}$$

Note that

$$z_k > 0$$
3-12: Jensen’s inequality I

- General form

\[
f(\int p(z)zdz) \leq \int p(z)f(z)dz
\]

- Discrete situation

\[
f(\sum p_i z_i) \leq \sum p_i f(z_i), \quad \sum p_i = 1
\]
3-12: Jensen’s inequality II

- **Proof:**

\[
\begin{align*}
    f(p_1z_1 + p_2z_2 + p_3z_3) &
    
    \leq (1 - p_3) \left( f \left( \frac{p_1z_1 + p_2z_2}{1 - p_3} \right) \right) + p_3 f(z_3) \\
    &
    
    \leq (1 - p_3) \left( \frac{p_1}{1 - p_3} f(z_1) + \frac{p_2}{1 - p_3} f(z_2) \right) + p_3 f(z_3) \\
    &
    
    = p_1 f(z_1) + p_2 f(z_2) + p_3 f(z_3)
\end{align*}
\]

- **Note that**

\[
\frac{p_1}{1 - p_3} + \frac{p_2}{1 - p_3} = \frac{1 - p_3}{1 - p_3} = 1
\]
Composition with affine function:

We know

\[ f(x) \text{ is convex} \]

Is

\[ g(x) = f(Ax + b) \]
Positive weighted sum & composition with affine function II

convex?

\[ g((1 - \alpha)x_1 + \alpha x_2) \]
\[ = f(A((1 - \alpha)x_1 + \alpha x_2) + b) \]
\[ = f((1 - \alpha)(Ax_1 + b) + \alpha(Ax_2 + b)) \]
\[ \leq (1 - \alpha)f(Ax_1 + b) + \alpha f(Ax_2 + b) \]
\[ = (1 - \alpha)g(x_1) + \alpha g(x_2) \]
Proof of the convexity

\begin{align*}
  f((1 - \alpha)x_1 + \alpha x_2) \\
  &= \max(f_1((1 - \alpha)x_1 + \alpha x_2), \ldots, f_m((1 - \alpha)x_1 + \alpha x_2)) \\
  &\leq \max((1 - \alpha)f_1(x_1) + \alpha f_1(x_2), \ldots, \\
  &\hspace{1cm} (1 - \alpha)f_m(x_1) + \alpha f_m(x_2)) \\
  &\leq (1 - \alpha) \max(f_1(x_1), \ldots, f_m(x_1)) + \\
  &\hspace{1cm} \alpha \max(f_1(x_2), \ldots, f_m(x_2)) \\
  &\leq (1 - \alpha)f(x_1) + \alpha f(x_2)
\end{align*}
For

\[ f(x) = x[1] + \cdots + x[r] \]

consider all

\[ \binom{n}{r} \]

combinations
The proof is similar to pointwise maximum.

Support function of a set $C$:

When $y$ is fixed,

$$f(x, y) = y^T x$$

is linear (convex) in $x$.

Maximum eigenvalues of symmetric matrix

$$f(X, y) = y^T X y$$

is a linear function of $X$ when $y$ is fixed.
Proof:

Let $\epsilon > 0$. \(\exists y_1, y_2 \in C\) such that

\[
\begin{align*}
  f(x_1, y_1) &\leq g(x_1) + \epsilon \\
  f(x_2, y_2) &\leq g(x_2) + \epsilon \\
  g(\theta x_1 + (1 - \theta) x_2) &= \inf_{y \in C} f(\theta x_1 + (1 - \theta) x_2, y) \\
  &\leq f(\theta x_1 + (1 - \theta) x_2, \theta y_1 + (1 - \theta) y_2) \\
  &\leq \theta f(x_1, y_1) + (1 - \theta) f(x_2, y_2) \\
  &\leq \theta g(x_1) + (1 - \theta) g(x_2) + \epsilon
\end{align*}
\]
3-19: Minimization II

- Note that the first inequality use the property that $C$ is convex to have

\[
\theta y_1 + (1 - \theta) y_2 \in C
\]

- Because the above inequality holds for all $\epsilon > 0$,

\[
g(\theta x_1 + (1 - \theta) x_2) \leq \theta g(x_1) + (1 - \theta) g(x_2)
\]
first example:
The goal is to prove

\[ A - BC^{-1}B^T \succeq 0 \]

Instead of a direct proof, here we use the property in this slide. First we have that \( f(x, y) \) is convex in \((x, y)\) because

\[
\begin{bmatrix}
A & B \\
B^T & C
\end{bmatrix} \succeq 0
\]

Consider

\[
\min_y f(x, y)
\]
Because $C \succ 0$,

the minimum occurs at

$$2Cy + 2B^T x = 0$$

$$y = -C^{-1}B^T x$$

Then

$$g(x) = x^T Ax - 2x^T BC^{-1}Bx + x^T BC^{-1}CC^{-1}B^T x$$

$$= x^T (A - BC^{-1}B^T)x$$
is convex. The second-order condition implies that

\[ A - BC^{-1}B^T \succeq 0 \]
This function is useful later

When \( y \) is fixed, maximum happens at

\[
y = f'(x)
\]  

(2)

by taking the derivative on \( x \)
Explanation of the figure: when $y$ is fixed

$$z = xy$$

is a straight line passing through the origin, where $y$ is the slope of the line. Check under which $x$, $yx$ and $f(x)$ have the largest distance

From the figure, the largest distance happens when (2) holds
3-21: the conjugate function III

About the point

\((0, -f^*(y))\)

The tangent line is

\[
\frac{z - f(x_0)}{x - x_0} = f'(x_0)
\]

where \(x_0\) is the point satisfying

\[
y = f'(x_0)
\]

When \(x = 0\),

\[
z = -x_0f'(x_0) + f(x_0) = -x_0y + f(x_0) = -f^*(y)
\]
$f^\ast$ is convex: Given $x$, 

$$y^T x - f(x)$$

is linear (convex) in $y$. Then we apply the property of pointwise supremum

3-21: the conjugate function IV
negative logarithm

\[ f(x) = -\log x \]

\[ \frac{\partial}{\partial x}(xy + \log x) = y + \frac{1}{x} = 0 \]

If \( y < 0 \), the picture of

\[ xy + \log x \]
Then

\[ xy + \log x = -1 - \log(-y) \]
strictly convex quadratic

\[ Qx = y, \quad x = Q^{-1}y \]

\[
y^T x - \frac{1}{2} x^T Q x = y^T Q^{-1} y - \frac{1}{2} y^T Q^{-1} Q Q^{-1} y = \frac{1}{2} y^T Q^{-1} y
\]
3-23: quasiconvex functions I

- Figure on slide:

\[ S_\alpha = [a, b], \quad S_\beta = (-\infty, c] \]

Both are convex

- The figure is an example showing that quasi convex may not be convex
Modified Jensen inequality:

\[ f(\theta x + (1 - \theta)y) \leq \max\{f(x), f(y)\}, \forall x, y, \theta \in [0, 1]. \]

⇒ Let

\[ \Delta = \max\{f(x), f(y)\} \]

\( S_\Delta \) is convex

\[ x \in S_\Delta, y \in S_\Delta \]

\[ \theta x + (1 - \theta)y \in S_\Delta \]
f(\theta x + (1 - \theta)y) \leq \Delta

and the result is obtained

\iff \text{results are wrong, there exists } \alpha \text{ such that } S_\alpha \text{ is not convex.}

\exists x, y, \theta \text{ with } x, y \in S_\alpha, \theta \in [0, 1] \text{ such that}

\theta x + (1 - \theta)y \notin S_\alpha
Then

\[ f(\theta x + (1 - \theta)y) > \alpha \geq \max\{f(x), f(y)\} \]

This violates the assumption

- First-order condition (this is exercise 3.43):
3-26: properties of quasiconvex functions

IV

\[ f((1-t)x + ty) \leq \max(f(x), f(y)) = f(x) \]

\[ \frac{f(x + t(y - x)) - f(x)}{t} \leq 0 \]

\[ \lim_{t \to 0} \frac{f(x + t(y - x)) - f(x)}{t} = \nabla f(x)^T (y - x) \leq 0 \]
If results are wrong, there exists $\alpha$ such that $S_\alpha$ is not convex.

There exist $x, y, \theta$ with $x, y \in S_\alpha, \theta \in [0, 1]$ such that

$$\theta x + (1 - \theta)y \notin S_\alpha$$

Then

$$f(\theta x + (1 - \theta)y) > \alpha \geq \max\{f(x), f(y)\} \quad (3)$$
Because $f$ is differentiable, it is continuous. Without loss of generality, we have

$$f(z) \geq f(x), f(y), \forall z \text{ between } x \text{ and } y$$

$$z = x + \theta(y - x), \theta \in (0, 1)$$

$$\nabla f(z)^T(-\theta(y - x)) \leq 0$$

$$\nabla f(z)^T(y - x - \theta(y - x)) \leq 0$$

Then

$$\nabla f(z)^T(y - x) = 0, \forall \theta \in (0, 1)$$
3-26: properties of quasiconvex functions

\[ f(x + \theta(y - x)) = f(x) + \nabla f(t)^T \theta(y - x) = f(x), \forall \theta \in [0, 1) \]

This contradicts (3).
3-27: Log-concave and log-convex functions I

- **Powers:**
  \[
  \log(x^a) = a \log x
  \]
  \(\log x\) is concave

- **Probability densities:**
  \[
  \log f(x) = -\frac{1}{2}(x - \bar{x})^T \Sigma^{-1}(x - \bar{x}) + \text{constant}
  \]
  \(\Sigma^{-1}\) is positive definite. Thus \(\log f(x)\) is concave
3-27: Log-concave and log-convex functions II

- Cumulative Gaussian distribution

\[ \log \Phi(x) = \log \int_{-\infty}^{x} e^{-u^2/2} du \]

\[ \frac{d}{dx} \log \Phi(x) = \frac{e^{-x^2/2}}{\int_{-\infty}^{x} e^{-u^2/2} du} \]

\[ \frac{d^2}{d^2x} \log \Phi(x) = \frac{(\int_{-\infty}^{x} e^{-u^2/2} du) e^{-x^2/2} (-x) - e^{-x^2/2} e^{-x^2/2}}{(\int_{-\infty}^{x} e^{-u^2/2} du)^2} \]
Need to prove that

\[
\left( \int_{-\infty}^{x} e^{-u^2/2} \, du \right) x + e^{-x^2/2} > 0
\]

Because

\[x \geq u \text{ for all } u \in (-\infty, x],\]
we have

\[
\left( \int_{-\infty}^{x} e^{-u^2/2} \, du \right) x + e^{-x^2/2} \\
= \int_{-\infty}^{x} xe^{-u^2/2} \, du + e^{-x^2/2} \\
\geq \int_{-\infty}^{x} ue^{-u^2/2} \, du + e^{-x^2/2} \\
= -e^{-u^2/2} \bigg|_{-\infty}^{x} + e^{-x^2/2} \\
= -e^{-x^2/2} + e^{-x^2/2} = 0
\]
This proof was given by a student (and polished by another student) of this course before
4-3: Optimal and locally optimal points I

- $f_0(x) = 1/x$

- $f_0(x) = x \log x$
4-3: Optimal and locally optimal points II

\[ f_0'(x) = 1 + \log x = 0 \]

\[ x = e^{-1} = \frac{1}{e} \]

- \( f_0(x) = x^3 - 3x \)
4-3: Optimal and locally optimal points III

\[ f_0'(x) = 3x^2 - 3 = 0 \]

\[ x = \pm 1 \]
Optimality criterion for differentiable $f_0$

\[ f_0(y) \geq f_0(x) + \nabla f_0(x)^T (y - x) \]

Together with

\[ \nabla f_0(x)^T (y - x) \geq 0 \]

we have

\[ f_0(y) \geq f_0(x), \text{ for all feasible } y \]
⇒ Assume the result is wrong. Then

\[ \nabla f_0(x)^T(y - x) < 0 \]

Let

\[ z(t) = ty + (1 - t)x \]

\[ \frac{d}{dt} f_0(z(t)) = \nabla f_0(z(t))^T(y - x) \]

\[ \left. \frac{d}{dt} f_0(z(t)) \right|_{t=0} = \nabla f_0(x)^T(y - x) < 0 \]
There exists \( t \) such that

\[
f_0(z(t)) < f_0(x)
\]

Note that

\[
z(t)
\]

is feasible because

\[
f_i(z(t)) \leq tf_i(x) + (1 - t)f_i(y) \leq 0
\]

and

\[
A(tx + (1-t)y) = tAx + (1-t)Ay = tb + (1-b)b = b
\]
4-9: Optimality criterion for differentiable $f_0$ IV
Unconstrained problem:

Let

\[ y = x - t\nabla f_0(x) \]

It is feasible (unconstrained problem). Optimality condition implies

\[ \nabla f_0(x)^T (y - x) = -t\|\nabla f_0(x)\|^2 \geq 0 \]

Thus

\[ \nabla f_0(x) = 0 \]

Equality constrained problem
Easy. For any feasible $y$,

$$Ay = b$$

$$\nabla f_0(x)(y-x) = -\nu^T A(y-x) = -\nu^T (b-b) = 0 \geq 0$$

So $x$ is optimal

$\Rightarrow$: more complicated. We only do a rough explanation
From optimality condition

$$\nabla f_0(x)^T \nu = \nabla f_0(x)^T ((x + \nu) - x) \geq 0, \forall \nu \in N(A)$$

$N(A)$ is a subspace in 2-D. Thus

$$\nu \in N(A) \Rightarrow -\nu \in N(A)$$
We have

\[ \nabla f_0(x)^T \nu = 0, \forall \nu \in N(A) \]

\[ \nabla f_0(x) \perp N(A), \nabla f_0(x) \in R(A^T) \]

\[ \Rightarrow \exists \nu \text{ such that } \nabla f_0(x) + A^T \nu = 0 \]

Minimization over nonnegative orthant

\[ \Leftarrow \text{Easy} \]
For any $y \succeq 0$,

$$\nabla_i f_0(x)(y_i - x_i) = \begin{cases} 
\nabla_i f_0(x)y_i \geq 0 & \text{if } x_i = 0 \\
0 & \text{if } x_i > 0.
\end{cases}$$

Therefore,

$$\nabla f_0(x)^T (y - x) \geq 0$$

and

$x$ is optimal
If $x_i = 0$, we claim

$$\nabla_i f_0(x) \geq 0$$

Otherwise,

$$\nabla_i f_0(x) < 0$$

Let

$$y = x \text{ except } y_i \to \infty$$

$$\nabla f_0(x)^T(y - x) = \nabla_i f_0(x)(y_i - x_i) \to -\infty$$

This violates the optimality condition
If $x_i > 0$, we claim

$$\nabla_i f_0(x) = 0$$

Otherwise, assume

$$\nabla_i f_0(x) > 0$$

Consider

$$y = x \text{ except } y_i = x_i / 2 > 0$$
It is feasible. Then

$$\nabla f_0(x)^T(y-x) = \nabla_i f_0(x)(y_i-x_i) = -\nabla_i f_0(x)x_i/2 < 0$$

violates the optimality condition. The situation for

$$\nabla_i f_0(x) < 0$$

is similar
\[ \bar{c} \equiv E(C) \]

\[ \Sigma \equiv E_C((C - \bar{c})(C - \bar{c})) \]

\[ \text{Var}(C^T x) = E_C((C^T x - \bar{c}^T x)(C^T x - \bar{c}^T x)) \]
\[ = E_C(x^T (C - \bar{c})(C - \bar{c})^T x) \]
\[ = x^T \Sigma x \]
Cone was defined on slide 2-8

\[ \{(x, t) \mid \|x\| \leq t\} \]
4-35: generalized inequality constraint I

- \( f_i \in \mathbb{R}^n \rightarrow \mathbb{R}^{k_i} \) \( K_i \)-convex:

\[
f_i(\theta x + (1 - \theta)y) \preceq_{K_i} \theta f_i(x) + (1 - \theta)f_i(y)
\]

- See page 3-31
LP and equivalent SDP

\[ Ax = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \]

\[ x_1 \begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{bmatrix} \leq 0 \]
For SOCP and SDP we will use results in 4-39:
\[
\begin{bmatrix}
  tl_{p \times p} & A_{p \times q} \\
  A^T & tl_{q \times q}
\end{bmatrix} \succeq 0 \iff A^T A \leq t^2 l_{q \times q}, \; t \geq 0
\]

Now
\[
p = m, \; q = 1
\]
\[
A = A_i x + b_i, \; t = c_i^T x + d_i
\]
\[
\|A_i x + b_i\|^2 \leq (c_i^T x + d_i)^2,
\]
\[
c_i^T x + d_i \geq 0
\]

Thus
\[
\|A_i x + b_i\| \leq c_i^T x + d_i
\]
Following 4-38, we have the following equivalent problem

\[
\begin{align*}
\text{min} & \quad t \\
\text{subject to} & \quad \|A\|_2 \leq t
\end{align*}
\]

We then use

\[
\|A\|_2 \leq t \iff A^T A \preceq t^2 I, \quad t \geq 0
\]

\[
\iff \begin{bmatrix} tl & A \\ A^T & tl \end{bmatrix} \succeq 0
\]
to have the SDP

$$\begin{align*}
\min \quad & t \\
\text{subject to} \quad & \begin{bmatrix}
tl & A(x) \\
A(x)^T & tl
\end{bmatrix} \succeq 0
\end{align*}$$

Next we prove

$$\begin{bmatrix}
tl_{p \times p} & A_{p \times q} \\
A^T & tl_{q \times q}
\end{bmatrix} \succeq 0 \iff A^T A \preceq t^2 I_{q \times q}, \quad t \geq 0$$
we immediately have

\[ t \geq 0 \]

If \( t > 0 \),

\[
\begin{bmatrix}
-\mathbf{v}^T \mathbf{A}^T & t\mathbf{v}^T
\end{bmatrix}
\begin{bmatrix}
tI_{p \times p} & \mathbf{A}_{p \times q} \\
\mathbf{A}^T & tI_{q \times q}
\end{bmatrix}
\begin{bmatrix}
-\mathbf{A}\mathbf{v} \\
t\mathbf{v}
\end{bmatrix}
= \begin{bmatrix}
-\mathbf{v}^T \mathbf{A}^T & t\mathbf{v}^T
\end{bmatrix}
\begin{bmatrix}
-t\mathbf{A}\mathbf{v} + t\mathbf{A}\mathbf{v} \\
-\mathbf{A}^T \mathbf{A}\mathbf{v} + t^2 \mathbf{v}
\end{bmatrix}
= t(t^2 \mathbf{v}^T \mathbf{v} - \mathbf{v}^T \mathbf{A}^T \mathbf{A} \mathbf{v}) \geq 0
\]

\[ \mathbf{v}^T (t^2 \mathbf{I} - \mathbf{A}^T \mathbf{A}) \mathbf{v} \geq 0, \forall \mathbf{v} \]
and hence

\[ t^2 I - A^T A \succeq 0 \]

If \( t = 0 \)

\[
\begin{bmatrix}
-v^T A^T & v^T
\end{bmatrix}
\begin{bmatrix}
0 & A
\end{bmatrix}
\begin{bmatrix}
-Av \\
\nu
\end{bmatrix}
\]

\[ = \begin{bmatrix}
-v^T A^T & v^T
\end{bmatrix}
\begin{bmatrix}
Av \\
-A^T A \nu
\end{bmatrix}
\]

\[ = -2v^T A^T A \nu \geq 0, \forall \nu \]

Therefore

\[ A^T A \preceq 0 \]
Consider

\[
\begin{bmatrix}
u^T & v^T
\end{bmatrix}
\begin{bmatrix}
tI & A \\
A^T & tI
\end{bmatrix}
\begin{bmatrix}
u \\
v
\end{bmatrix}
\]

\[
= \begin{bmatrix}
u^T & v^T
\end{bmatrix}
\begin{bmatrix}
tu + Av \\
A^T u + tv
\end{bmatrix}
\]

\[
= tu^T u + 2v^T A^T u + tv^T v
\]

We hope to have

\[
tu^T u + 2v^T A^T u + tv^T v \geq 0, \forall (u, v)
\]
If \( t > 0 \)

\[
\min_u tu^T u + 2v^T A^T u + tv^T v
\]

has optimum at

\[
u = \frac{-Av}{t}
\]

We have

\[
tu^T u + 2v^T A^T u + tv^T v
\]

\[
= tv^T v - \frac{v^T A^T A v}{t}
\]

\[
= \frac{1}{t} v^T (t^2 I - A^T A)v \geq 0.
\]
Hence

\[
\begin{bmatrix}
tI & A \\
A^T & tI
\end{bmatrix} \succeq 0
\]

If \( t = 0 \)

\[
A^T A \preceq 0
\]

\[
\nu^T A^T A \nu \leq 0, \quad \nu^T A^T A \nu = \| A \nu \|^2 = 0
\]

Thus

\[
A \nu = 0, \quad \forall \nu
\]

\[
\begin{bmatrix}
u^T & u^T
\end{bmatrix}
\begin{bmatrix}
0 & A \\
A^T & 0
\end{bmatrix}
\begin{bmatrix}
u \\
u
\end{bmatrix}
\]

\[
= \begin{bmatrix}
u^T & u^T
\end{bmatrix}
\begin{bmatrix}
0 \\
A^T u
\end{bmatrix}
= 0 \geq 0
\]
Thus

\[
\begin{bmatrix}
0 & A \\
A^T & 0
\end{bmatrix} \succeq 0
\]
Though

\[ f_0(x) \text{ is a vector} \]

note that

\[ f_i(x) \text{ is still } R^n \to R^1 \]

\( K \)-convex

See 3-31 though we didn’t discuss it earlier
**Optimal**

\[ O \subseteq \{x\} + K \]

**Pareto optimal**

\[ (x - K) \cap O = \{x\} \]
Note that \( g \) is concave no matter if the original problem is convex or not

\[
f_0(x) + \sum \lambda_i f_i(x) + \sum \nu_i h_i(x)
\]

is convex (linear) in \( \lambda, \nu \) for each \( x \)
5-3: Lagrange dual function II

Use pointwise supremum on 3-16

$$\sup_{x \in D} (-f_0(x) - \sum \lambda_i f_i(x) - \sum \nu_i h_i(x))$$

is convex. Hence

$$\inf (f_0(x) + \sum \lambda_i f_i(x) + \sum \nu_i h_i(x))$$

is concave. Note that

$$- \sup (\cdots) = -\text{convex}$$

$$= \inf (\cdots) = \text{concave}$$
5-8: Lagrange dual and conjugate function

\[ f_0^*(\mathbf{y}) = \sup_{\mathbf{x}} \left( \mathbf{y}^T \mathbf{x} - f_0(\mathbf{x}) \right) = -\inf_{\mathbf{x}} \left( f_0(\mathbf{x}) + \mathbf{y}^T (\mathbf{A}^T \lambda + \mathbf{c}^T \nu) \right) \]
We don’t discuss the SDP problem on this slide because we omitted 5-7
5-11: Slater’s constraint qualification I

- We omit the proof because of no time
- “linear inequality do not need to hold with strict inequality”: for linear inequalities we DO NOT need constraint qualification
- We will see some explanation later
If we have only linear constraints, then constraint qualification holds.
Explanation of $g(\lambda)$: when $\lambda$ is fixed

$$\lambda u + t = \Delta$$

is a line. We lower $\Delta$ until it touches the boundary of $G$

The $\Delta$ value then becomes $g(\lambda)$

When

$$u = 0 \Rightarrow t = \Delta$$

so we see the point marked as $g(\lambda)$ on $t$-axis
We have $\lambda \geq 0$, so

\[ \lambda u + t = \Delta \]

must be like

rather than
Explanation of $p^*$: 
In $G$, only points satisfying

$$u \leq 0$$

are feasible
We do not discuss a formal proof of Slater condition $\Rightarrow$ strong duality.

Instead, we explain this result by figures.

Reason of using $A$: $G$ may not be convex.

Example:

$$\min x^2$$

subject to $x + 2 \leq 0$
This is a convex optimization problem

\[ G = \{(x + 2, x^2) \mid x \in R\} \]

is only a quadratic curve
The curve is not convex

However, $A$ is convex
5-16 IV

- Primal problem:
  \[ x = -2 \]
  optimal objective value = 4

- Dual problem:
  \[ g(\lambda) = \min_x x^2 + \lambda(x + 2) \]
  \[ x = -\lambda/2 \]
  \[ \max_{\lambda \geq 0} -\frac{\lambda^2}{4} + 2\lambda \]
  optimal \( \lambda = 4 \)
optimal objective value $= -\frac{16}{4} + 8 = 4$

- Proving that $A$ is convex

$$(u_1, t_1) \in A, (u_2, t_2) \in A$$

$\exists x_1, x_2 \text{ such that}$

$$f_1(x_1) \leq u_1, f_0(x_1) \leq t_1$$
$$f_1(x_2) \leq u_2, f_0(x_2) \leq t_2$$

Consider

$$x = \theta x_1 + (1 - \theta)x_2$$
We have

\[ f_1(x) \leq \theta u_1 + (1 - \theta)u_2 \]

\[ f_0(x) \leq \theta t_1 + (1 - \theta)t_2 \]

So

\[
\begin{bmatrix}
  u \\
  t
\end{bmatrix} = \theta \begin{bmatrix}
  u_1 \\
  t_1
\end{bmatrix} + (1 - \theta) \begin{bmatrix}
  u_2 \\
  t_2
\end{bmatrix} \in A
\]

Note that we have

Slater condition \(\Rightarrow\) strong duality

However, it’s possible that Slater condition doesn’t hold but strong duality holds
Example from exercise 5.22:

$$ \min x $$

subject to $x^2 \leq 0$

Slater condition doesn’t hold because no $x$ satisfies $x^2 < 0$

$$ G = \{(x^2, x) \mid x \in \mathbb{R}\} $$
There is only one feasible point \((0, 0)\)
\[ g(\lambda) = \min_x x + x^2 \lambda \]

\[ x = \begin{cases} 
-1/(2\lambda) & \text{if } \lambda > 0 \\
-\infty & \text{if } \lambda = 0 
\end{cases} \]

**Dual problem**

\[ \max_{\lambda \geq 0} -1/(4\lambda) \]

\[ \lambda \to \infty, \text{ objective value } \to 0 \]

\[ d^* = 0, p^* = 0 \]

**Strong duality holds**
In deriving the inequality we use

\[ h_i(x^*) = 0 \text{ and } f_i(x^*) \leq 0 \]

Complementary slackness
compare the earlier results in 4-10
For the problem on p5-16, neither slater condition nor KKT condition holds

\[ 1 \neq \lambda 0 \]

Therefore, for convex problems,

\[ \text{KKT} \Rightarrow \text{optimality} \]

but not vice versa.

Next we explain why for linear constraints we don’t need constraint qualification
Consider the situation of inequality constraints only:

\[
\begin{align*}
\min & \quad f_0(x) \\
\text{subject to} & \quad f_i(x) \leq 0, \ i = 1, \ldots, m
\end{align*}
\]

Consider an optimal solution \( x \). We would like to prove that \( x \) satisfies KKT condition.

Because \( x \) is optimal, from the optimality condition on slide 4-9, for any feasible direction \( \delta x \),

\[
\nabla f_0(x)^T \delta x \geq 0.
\]
5-19: KKT conditions for convex problem

- A feasible $\delta x$ means

$$f_i(x + \delta x) \leq 0, \forall i$$

- Because

$$f_i(x + \delta x) \approx f_i(x) + \nabla f_i(x)^T \delta x$$

from

$$f_i(x) \leq 0, \forall i$$

we have

$$\nabla f_i(x)^T \delta x \leq 0 \text{ if } f_i(x) = 0.$$
We claim that
\[ \nabla f_0(x) = \sum_{\lambda_i \geq 0, f_i(x) = 0} -\lambda_i \nabla f_i(x) \] (4)

Assume the result is wrong. First let’s consider
\[ \nabla f_0(x) = \text{linear combination of } \{ \nabla f_i(x) \mid f_i(x) = 0 \} + \Delta, \]
where
\[ \Delta \neq 0 \text{ and } \Delta^T \nabla f_i(x) = 0, \forall i : f_i(x) = 0. \]
Then there exists $\alpha < 0$ such that

$$\delta x \equiv \alpha \Delta$$

satisfies

$$\nabla f_i(x)^T \delta x = 0 \text{ if } f_i(x) = 0$$

$$f_i(x + \delta x) \leq 0 \text{ if } f_i(x) < 0$$

and

$$\nabla f_0(x)^T \delta x = \alpha \Delta^T \Delta < 0$$

This contradicts the optimality condition.
By a similar setting we can further prove (4).

Assume

$$\nabla f_0(x) = \sum_{i:f_i(x)=0} -\lambda_i \nabla f_i(x)$$

and there exists $i'$ such that

$$\lambda_{i'} < 0, \nabla f_{i'}(x) \neq 0, \text{ and } f_{i'}(x) = 0$$
Let

$$\bar{\lambda} = \arg \min_{\lambda} \| \nabla f_{i'}(x) - \sum_{i: i \neq i', f_i(x) = 0} \nabla f_i(x) \lambda_i \|$$

$$\Delta = \| \nabla f_{i'}(x) - \sum_{i: i \neq i', f_i(x) = 0} \nabla f_i(x) \bar{\lambda}_i \|$$

Then

$$\nabla f_i(x)^T \Delta = 0, \forall i \neq i', f_i(x) = 0 \quad (5)$$
We have

$$\nabla f_i'(x)^T \Delta \geq 0$$

If

$$\nabla f_i'(x)^T \Delta = 0$$

then

$$-\lambda_i \nabla f_i'(x)$$

can be rearranged to use linear combination of

$$\{\nabla f_i(x) \mid i \neq i', f_i(x) = 0\}$$
Otherwise

\[ \nabla f_{i'}(x)^T \Delta > 0 \]

- Let

\[ \delta x = \alpha \Delta, \alpha < 0. \]

- From (5),

\[ \nabla f_i(x)^T \delta x = 0, \forall i \neq i', f_i(x) = 0 \]

\[ \nabla f_{i'}(x)^T \delta x = \alpha \nabla f_{i'}(x)^T \Delta < 0. \]

Hence \( \delta x \) is a feasible direction.
However, 

\[ \nabla f_0(x)^T \delta x = -\alpha \lambda_i \nabla f_i(x)^T \Delta < 0 \]

contradicts the optimality condition.

This proof is not rigorous because of \( \approx \).

For linear the proof becomes rigorous.
Explanation of $f_0^*(\nu)$

\[
\inf_y (f_0(y) - \nu^T y) = -\sup_y (\nu^T y - f_0(y)) = -f_0^*(\nu)
\]

where $f_0^*(\nu)$ is the conjugate function
The original problem

\[ g(\lambda, \nu) = \inf_x \| Ax - b \| = \text{constant} \]

Dual norm:

\[ \| \nu \|_* \equiv \sup \{ \nu^T y \mid \| y \| \leq 1 \} \]

If \( \| \nu \|_* > 1 \),

\[ \nu^T y^* > 1, \| y^* \| \leq 1 \]
\[
\begin{align*}
\inf \| y \| + \nu^T y & \\
\leq \| - y^* \| - \nu^T y^* & < 0 \\
\| - ty^* \| - \nu^T (ty^*) & \to -\infty \text{ as } t \to \infty
\end{align*}
\]

Hence
\[
\inf_y \| y \| + \nu^T y = -\infty
\]

If \( \| \nu \|_* \leq 1 \), we claim that
\[
\inf_y \| y \| + \nu^T y = 0
\]

\[ y = 0 \Rightarrow \| y \| + \nu^T y = 0 \]
If \( \exists y \) such that
\[
\|y\| + \nu^T y < 0
\]
then
\[
\| - y\| < -\nu^T y
\]
We can scale \( y \) so that
\[
\sup\{\nu^T y \mid \|y\| \leq 1\} > 1
\]
but this causes a contradiction.
The dual function

\[ c^T x + \nu^T (A x - b) = - b^T \nu + x^T (A^T \nu + c) \]

\[ \inf_{-1 \leq x_i \leq 1} x_i (A^T \nu + c)_i = - |(A^T \nu + c)_i| \]
From 5-29 we need that $Z$ is non-negative in the dual cone of $S_k^+$

Dual cone of $S_k^+$ is $S_k^+$ (we didn’t discuss dual cone so we assume this result)

Why

$$\text{tr}(Z(\cdots))$$?

We are supposed to do component-wise product between

$$Z \text{ and } x_1 F_1 + \cdots + x_n F_n - G$$
Trace is the component-wise product

$$\text{tr}(AB) = \sum_i (AB)_{ii} = \sum_i \sum_j A_{ij} B_{ji} = \sum_i \sum_j A_{ij} B_{ij}$$

Note that we take the property that $B$ is symmetric
Uniform noise

\[ p(z) = \begin{cases} 
\frac{1}{2a} & \text{if } |z| \leq a \\
0 & \text{otherwise}
\end{cases} \]
8-10: Dual of maximum margin problem I

Largangian:

\[
\frac{\|a\|}{2} - \sum_i \lambda_i (a^T x_i + b - 1) + \sum_i \mu_i (a^T y_i + b + 1)
\]

\[
= \frac{\|a\|}{2} + a^T (-\sum_i \lambda_i x_i + \sum_i \mu_i y_i)
\]

\[
+ b (-\sum_i \lambda_i + \sum_i \mu_i) + \sum_i \lambda_i + \sum_i \mu_i
\]

Because of

\[
b (-\sum_i \lambda_i + \sum_i \mu_i)
\]
we have

\[
\inf_{a,b} L = \begin{cases} 
\inf_a \frac{\|a\|}{2} - \sum_i \lambda_i a^T x_i + \sum_i \mu_i a^T y_i & \text{if } \sum_i \lambda_i = \sum_i \mu_i \\
-\infty & \text{if } \sum_i \lambda_i \neq \sum_i \mu_i 
\end{cases}
\]

For

\[
\inf_a \frac{\|a\|}{2} - \sum_i \lambda_i a^T x_i + \sum_i \mu_i a^T y_i
\]
we can denote it as

\[ \inf_a \frac{\|a\|}{2} + v^T a \]

where \( v \) is a vector. We cannot do derivative because \( \|a\| \) is not differentiable. Formal solution:
Case 1: If $\|v\| \leq 1/2$:

$$a^T v \geq -\|a\| \|v\| \geq -\frac{\|a\|}{2}$$

so

$$\inf_a \frac{\|a\|}{2} + v^T a \geq 0.$$ 

However,

$$a = 0 \rightarrow \frac{\|a\|}{2} + v^T a = 0.$$
Therefore

\[ \inf_a \frac{\|a\|}{2} + v^T a = 0. \]

- If \( \|v\| > 1/2 \), let

\[
a = \frac{-tv}{\|v\|}
\]

\[
\frac{\|a\|}{2} + v^T a
\]

\[
= \frac{t}{2} - t\|v\|
\]

\[
=t\left(\frac{1}{2} - \|v\|\right) \to \infty \text{ if } t \to -\infty
\]
8-10: Dual of maximum margin problem

Thus

\[
\inf_a \frac{\|a\|}{2} + v^T a = -\infty
\]
\[
\theta = \begin{bmatrix}
\text{vec}(P) \\
q \\
r
\end{bmatrix}, \quad F(z) = \begin{bmatrix}
\vdots \\
Z_i Z_j \\
\vdots \\
Z_i \\
\vdots \\
1
\end{bmatrix}
\]
The condition that $S$ is closed if

$$f(x) \to \infty \text{ as } x \to \text{ boundary of domain } f$$

Proof: if not, consider

$$\{x_i\} \subset S$$

such that

$$x_i \to \text{ boundary}$$

Then

$$f(x_i) \to \infty > f(x^0)$$
and

S is not closed

Example

$$f(x) = \log\left(\sum_{i=1}^{m} \exp(a_i^T x + b_i)\right)$$

domain = $\mathbb{R}^n$
Example

\[ f(x) = - \sum_{i} \log(b_i - a_i^T x) \]

domain \( \neq \mathbb{R}^n \)

We use the condition that

\[ f(x) \to \infty \text{ as } x \to \text{ boundary of domain } f \]
S is bounded. Otherwise, there exists a set

\[ \{ y_i \mid y_i = x + \Delta_i \} \subset S \]

satisfying

\[ \lim_{i \to \infty} |\Delta_i| = \infty \]

Then

\[ f(y_i) \geq f(x) + \nabla f(x)^T \Delta_i + \frac{m}{2} \| \Delta_i \|^2 \to \infty \]

This contradicts

\[ f(y) \leq f(x^0) \]
Proof of \( p^* > -\infty \)

and

\[
f(x) - p^* \leq \frac{1}{2m} \|\nabla f(x)\|^2
\]

From

\[
f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|x - y\|^2
\]

Minimize the right-hand side with respect to \( y \)

\[
\nabla f(x) + m(y - x) = 0
\]
\[ \tilde{y} = x - \frac{\nabla f(x)}{m} \]

\[ f(y) \geq f(x) + \nabla f(x)^T(\tilde{y} - x) + \frac{m}{2}\|\tilde{y} - x\|^2 \]

\[ = f(x) - \frac{1}{2m}\|\nabla f(x)\|^2, \forall y \]

Then

\[ p^* \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2 \]

and

\[ f(x) - p^* \leq \frac{1}{2m}\|\nabla f(x)\|^2 \]
10-5: descent methods

- If
  \[ f(x + t\Delta x) < f(x) \]
  then
  \[ \nabla f(x)^T \Delta x < 0 \]

Proof: From the first-order condition of a convex function

\[ f(x + t\Delta x) \geq f(x) + t\nabla f(x)^T \Delta x \]

Then

\[ t\nabla f(x)^T \Delta x \leq f(x + t\Delta x) - f(x) < 0 \]
Why \( \alpha \in (0, \frac{1}{2})? \)

The use of \(1/2\) is for convergence though we won’t discuss details.

Finite termination of backtracking line search. We argue that \( \exists t^* > 0 \) such that

\[
f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^T \Delta x, \quad \forall t \in (0, t^*)
\]

Otherwise,

\[
\exists \{t_k\} \to 0
\]
such that

\[ f(x + t_k \Delta x) \geq f(x) + \alpha t_k \nabla f(x)^T \Delta x, \forall k \]

\[
\lim_{t_k \to 0} \frac{f(x + t_k \Delta x) - f(x)}{t_k} = \nabla f(x)^T \Delta x \geq \alpha \nabla f(x)^T \Delta x
\]

However,

\[ \nabla f(x)^T \Delta x < 0 \text{ and } \alpha > 0 \]

cause a contradiction
Geometric interpretation: the tangent line passes through \((0, f(x))\), so the equation is

\[
y - f(x) \over t - 0 = \nabla f(x)^T \Delta x
\]

Because

\[
\nabla f(x)^T \Delta x < 0,
\]

we see that the line of

\[
f(x) + \alpha t \nabla f(x)^T \Delta x
\]

is above that of

\[ f(x) + t \nabla f(x)^T \Delta x \]
Linear convergence. We consider exact line search; proof for backtracking line search is more complicated

$S$ closed and bounded

$$\nabla^2 f(x) \preceq M I, \forall x \in S$$

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|^2$$

Solve

$$\min_t f(x) - t \nabla f(x)^T \nabla f(x) + \frac{t^2 M}{2} \nabla f(x)^T \nabla f(x)$$
\[ t = \frac{1}{M} \]

\[
f(x_{\text{next}}) \leq f(x - \frac{1}{M} \nabla f(x)) \leq f(x) - \frac{1}{2M} \nabla f(x)^T \nabla f(x) \]

The first inequality is from the fact that we use exact line search

\[
f(x_{\text{next}}) - p^* \leq f(x) - p^* - \frac{1}{2M} \nabla f(x)^T \nabla f(x) \]

From slide 10-4,

\[-\|\nabla f(x)\|^2 \leq -2m(f(x) - p^*)\]
Hence

\[ f(x_{\text{next}}) - p^* \leq (1 - \frac{m}{M})(f(x) - p^*) \]
Assume

\[ x_1^k = \gamma \left( \frac{\gamma - 1}{\gamma + 1} \right)^k, \quad x_2^k = \left( -\frac{\gamma - 1}{\gamma + 1} \right)^k, \]

\[ \nabla f(x_1, x_2) = \begin{bmatrix} x_1 \\ \gamma x_2 \end{bmatrix} \]

\[ \min_t \frac{1}{2} \left( (x_1 - tx_1)^2 + \gamma(x_2 - t\gamma x_2)^2 \right) \]

\[ \min_t \frac{1}{2} \left( x_1^2(1 - t)^2 + \gamma x_2^2(1 - t\gamma)^2 \right) \]
\[-x_1^2(1 - t) + \gamma x_2^2(1 - t\gamma)(-\gamma) = 0\]

\[-x_1^2 + tx_1^2 - \gamma^2 x_2^2 + \gamma^3 tx_2^2 = 0\]

\[t(x_1^2 + \gamma^3 x_2^2) = x_1^2 + \gamma^2 x_2^2\]

\[t = \frac{x_1^2 + \gamma^2 x_2^2}{x_1^2 + \gamma^3 x_2^2} = \frac{\gamma^2 \left(\frac{\gamma - 1}{\gamma + 1}\right)^{2k} + \gamma^2 \left(\frac{\gamma - 1}{\gamma + 1}\right)^{2k}}{\gamma^2 \left(\frac{\gamma - 1}{\gamma + 1}\right)^{2k} + \gamma^3 \left(\frac{\gamma - 1}{\gamma + 1}\right)^{2k}}\]

\[= \frac{2\gamma^2}{\gamma^2 + \gamma^3} = \frac{2}{1 + \gamma}\]

\[x^{k+1} = x^k - t\nabla f(x^k) = \begin{bmatrix} x_1^k(1 - t) \\ x_2^k(1 - \gamma t) \end{bmatrix}\]
\[ x_1^{k+1} = \gamma \left( \frac{\gamma - 1}{\gamma + 1} \right)^k \left( \frac{\gamma - 1}{1 + \gamma} \right) = \gamma \left( \frac{\gamma - 1}{\gamma + 1} \right)^{k+1} \]

\[ x_2^{k+1} = \left( -\frac{\gamma - 1}{\gamma + 1} \right)^k \left( 1 - \frac{2\gamma}{1 + \gamma} \right) \]

\[ = \left( -\frac{\gamma - 1}{\gamma + 1} \right)^k \left( \frac{1 - \gamma}{1 + \gamma} \right) = \left( -\frac{\gamma - 1}{\gamma + 1} \right)^{k+1} \]

Why gradient is orthogonal to the tangent line of the contour curve?
Assume \( f(g(t)) \) is the contour with

\[
g(0) = x
\]

Then

\[
0 = f(g(t)) - f(g(0))
\]

\[
0 = \lim_{t \to 0} \frac{f(g(t)) - f(g(0))}{t}
\]

\[
= \lim_{t \to 0} \nabla f(g(t))^T \nabla g(t)
\]

\[
= \nabla f(x)^T \nabla g(0)
\]
where

\[ x + t \nabla g(0) \]

is the tangent line
linear convergence: from slide 10-7

\[ f(x^k) - p^* \leq c^k (f(x^0) - p^*) \]

\[ \log(c^k (f(x^0) - p^*)) = k \log c + \log(f(x^0) - p^*) \]

is a straight line. Note that now \( k \) is the \( x \)-axis.
(unnormalized) steepest descent direction:

$$\Delta x_{sd} = \|\nabla f(x)\|_* \Delta x_{nsd}$$

Here $\| \cdot \|_*$ is the dual norm

We didn’t discuss much about dual norm, but we can still explain some examples on 10-12
Euclidean: $\Delta x_{\text{nsd}}$ is by solving

$$\min \quad \nabla f^T v$$

subject to  \quad \|v\| = 1

$$\nabla f^T v = \|\nabla f\| \|v\| \cos \theta = -\|\nabla f\|$$ \quad \text{when} \quad \cos \theta = \pi

$$\Delta x_{\text{nsd}} = \frac{-\nabla f(x)}{\|\nabla f(x)\|}$$

$$\|\nabla f(x)\|_* = \|\nabla f(x)\|$$

$$\|\nabla f(x)\|_* \Delta x_{\text{nsd}} = \|\nabla f(x)\|_* \frac{-\nabla f(x)}{\|\nabla f(x)\|} = -\nabla f(x)$$
Quadratic norm: $\Delta x_{nsd}$ is by solving

$$
\min \nabla f^T v
\text{subject to } v^T P v = 1
$$

Now

$$
\|v\|_P = \sqrt{v^T P v},
$$

where $P$ is symmetric positive definite.
Let \( w = P^{1/2}v \)

The optimization problem becomes

\[
\min_w \nabla f^T P^{-1/2}w
\]

subject to \( \|w\| = 1 \)

optimal \( w = \frac{-P^{-1/2}\nabla f}{\|P^{-1/2}\nabla f\|} = \frac{-P^{-1/2}\nabla f}{\sqrt{\nabla f^T P^{-1}\nabla f}} \)
optimal $v = \frac{-P^{-1}\nabla f}{\sqrt{\nabla f^T P^{-1} \nabla f}}$

- Dual norm

$$\|z\|_* = \|P^{-1/2}z\|$$

Therefore

$$\Delta x_{sd} = \sqrt{\nabla f^T P^{-1} \nabla f} \frac{-P^{-1}\nabla f}{\sqrt{\nabla f^T P^{-1} \nabla f}} = -P^{-1}\nabla f$$
Explanation of the figure:

\[-\nabla f(x)^T \Delta x_{\text{nsd}} = \| -\nabla f(x) \| \| \Delta x_{\text{nsd}} \| \cos \theta\]

\| -\nabla f(x) \| \text{ is a constant. From a point } \Delta x_{\text{nsd}} \text{ on the boundary, the projected point on } -\nabla f(x) \text{ indicates}

\| \Delta x_{\text{nsd}} \| \cos \theta\]

In the figure, we see that the chosen \( \Delta x_{\text{nsd}} \) has the largest \( \| \Delta x_{\text{nsd}} \| \cos \theta \)

We omit the discussion of \( l_1 \)-norm
The two figures are by using two P matrices
The left one has faster convergence
Gradient descent after change of variables

\[ \bar{x} = P^{1/2}x, \quad x = P^{-1/2}\bar{x} \]

\[ \min_x f(x) \Rightarrow \min_{\bar{x}} f(P^{-1/2}\bar{x}) \]

\[ \bar{x} \leftarrow \bar{x} - \alpha P^{-1/2}\nabla_x f(P^{-1/2}\bar{x}) \]

\[ P^{1/2}x \leftarrow P^{1/2}x - \alpha P^{-1/2}\nabla_x f(x) \]

\[ x \leftarrow x - \alpha P^{-1} \nabla_x f(x) \]
$f'(x)$
Solve

\[ f'(x) = 0 \]

Fining the tangent line at \( x_k \):

\[
\frac{y - f'(x_k)}{x - x_k} = f''(x_k)
\]

\( x_k \): the current iterate

Let \( y = 0 \)

\[
x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}
\]
\[ \hat{f}(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(x)(y - x) \]

\[ \nabla \hat{f}(y) = 0 = \nabla f(x) + \nabla^2 f(x)(y - x) \]

\[ y - x = - \nabla^2 f(x)^{-1} \nabla f(x) \]

\[ \inf_y \hat{f}(y) = f(x) - \frac{1}{2} \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \]

\[ f(x) - \inf_y \hat{f}(y) = \frac{1}{2} \lambda(x)^2 \]
Norm of the Newton step in the quadratic Hessian norm

\[ \Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x) \]

\[ \Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt} = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) = \lambda(x)^2 \]

Directional derivative in the Newton direction

\[ \lim_{t \to 0} \frac{f(x + t\Delta_{nt}) - f(x)}{t} \]

\[ = \nabla f(x)^T \Delta x_{nt} \]

\[ = - \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) = -\lambda(x)^2 \]
Affine invariant

\[ \bar{f}(y) \equiv f(Ty) = f(x) \]

Assume \( T \) is an invertable square matrix. Then

\[ \bar{\lambda}(y) = \lambda(Ty) \]

Proof:

\[ \nabla \bar{f}(y) = T^T \nabla f(Ty) \]

\[ \nabla^2 \bar{f}(y) = T^T \nabla^2 f(Ty) T \]
\[ \tilde{\lambda}(y)^2 = \nabla \bar{f}(y)^T \nabla^2 \bar{f}(y)^{-1} \nabla \bar{f}(y) \]
\[ = \nabla f(Ty)^T TTT^{-1} \nabla^2 f(Ty)^{-1} T^{-T} T^T \nabla f(Ty) \]
\[ = \nabla f(Ty)^T \nabla^2 f(Ty)^{-1} \nabla f(Ty) \]
\[ = \lambda(Ty)^2 \]
Affine invariant

\[
\Delta y_{nt} = \nabla^2 \bar{f}(y)^{-1} \nabla \bar{f}(y)
\]

\[
= T^{-1} \nabla^2 f(Ty)^{-1} \nabla f(Ty)
\]

\[
= T^{-1} \Delta x_{nt}
\]

Note that

\[
y_k = T^{-1} x_k
\]

so

\[
y_{k+1} = T^{-1} x_{k+1}
\]
But how about line search

\[ \nabla \bar{f}(y)^T \Delta y_{nt} = \nabla f(Ty)^T T T^{-1} \Delta x_{nt} = \nabla f(x)^T \Delta x_{nt} \]
\[ \eta \in (0, \frac{m^2}{L}) \]

\[ \| \nabla f(x_k) \| \leq \eta \leq \frac{m^2}{L} \]

\[ \frac{L}{2m^2} \| \nabla f(x_k) \| \leq \frac{1}{2} \]
\begin{align*}
f(x_l) - f(x^*) &
\leq \frac{1}{2m} \left\| \nabla f(x_l) \right\|^2 
\leq \frac{1}{2m} \frac{4m^4}{L^2} \left( \frac{1}{2} \right)^{2^{l-k+1}} 
= \frac{2m^3}{L^2} \left( \frac{1}{2} \right)^{2^{l-k+1}} 
\leq \epsilon
\end{align*}

Let

\[ \epsilon_0 = \frac{2m^3}{L^2} \]
\[ \log_2 \epsilon_0 - 2^{l-k+1} \leq \log_2 \epsilon \]
\[ 2^{l-k+1} \geq \log_2 (\epsilon_0/\epsilon) \]
\[ l \geq k - 1 + \log_2 \log_2 (\epsilon_0/\epsilon) \]
\[ k \leq \frac{f(x_0) - p^*}{r} \]

In at most
\[ \frac{f(x_0) - p^*}{r} + \log_2 \log_2 (\epsilon_0/\epsilon) \]
iterations, we have
\[ f(x_l) - f(x^*) \leq \epsilon \]
\[ \lambda(x) = (\nabla f(x) \nabla^2 f(x)^{-1} \nabla f(x))^{1/2} = (g^T L^{-T} L^{-1} g)^{1/2} = \| L^{-1} g \|_2 \]
10-30: example of dense Newton systems with structure $I$

$$\nabla f(x) = \begin{bmatrix} \psi_1'(x_1) \\ \vdots \\ \psi_n'(x_n) \end{bmatrix} + A^T \nabla \psi_0(Ax + b)$$

$$\nabla^2 f(x) = \begin{bmatrix} \psi_1''(x_1) \\ \vdots \\ \psi_n''(x_n) \end{bmatrix} + A^T \nabla^2 \psi_0^2(Ax + b)A$$

method 2:

$$\Delta x = D^{-1}(-g - A^T L_0 w)$$
10-30: example of dense Newton systems with structure II

\[
L_0^T AD^{-1}(-g - A^T L_0 w) = w
\]
\[
(I + L_0^T AD^{-1} A^T L_0)w = -L_0^T AD^{-1} g
\]

Cost

\[
L_0 : p \times p
\]
\[
A^T L_0 : n \times p, \text{ cost } : O(np)
\]
\[
(L_0^T A)D^{-1}(A^T L_0) : O(p^2 n)
\]
10-30: example of dense Newton systems with structure III

Note that Cholesky factorization of $H_0$ costs

$$\frac{1}{3}p^3 \leq p^2 n$$

as

$$p \ll n$$