1-5: Least-squares I

- $A : k \times n$. Usually
$$k > n$$
otherwise easily the minimum is zero.
- Analytical solution:
$$\begin{aligned} f(x) =& (Ax - b)^T (Ax - b) \\ =& x^T A^T Ax - 2b^T Ax + b^T b \end{aligned}$$
$$\nabla f(x) = 2A^T Ax - 2A^T b = 0$$

1-5: Least-squares II

- Regularization, weights:

$$\frac{1}{2}\lambda x^T x + w_1(Ax-b)_1^2 + \cdots + w_k(Ax-b)_k^2$$

2-4: Convex Combination and Convex Hull I

- Convex hull is convex

$$
\begin{aligned}
x &= \theta_1 x_1 + \cdots + \theta_k x_k \\
\bar{x} &= \bar{\theta}_1 \bar{x}_1 + \cdots + \bar{\theta}_{\bar{k}} \bar{x}_{\bar{k}}
\end{aligned}
$$

Then

$$
\begin{aligned}
& \alpha x + (1-\alpha)\bar{x} \\
=& \alpha\theta_1 x_1 + \cdots + \alpha\theta_k x_k + \\
& (1-\alpha)\bar{\theta}_1 \bar{x}_1 + \cdots + (1-\alpha)\bar{\theta}_{\bar{k}} \bar{x}_{\bar{k}}
\end{aligned}
$$

2-4: Convex Combination and Convex Hull II

Each coefficient is nonnegative and

$$\begin{aligned} & \alpha\theta_1 + \cdots + \alpha\theta_k + (1-\alpha)\bar{\theta}_1 + \cdots + (1-\alpha)\bar{\theta}_{\bar{k}} \\ = \; & \alpha + (1-\alpha) = 1 \end{aligned}$$

2-7: Euclidean Balls and Ellipsoid I

We prove that any

$$x = x_c + Au \text{ with } \|u\|_2 \le 1$$

satisfies

$$(x - x_c)^T P^{-1}(x - x_c) \le 1$$

Let

$$A = P^{1/2}$$

because P is symmetric positive definite.
Then

$$u^T A^T P^{-1} Au = u^T P^{1/2} P^{-1} P^{1/2} u \le 1.$$

2-10: Positive Semidefinite Cone I

- $\mathbf{S}_+^n$ is a convex cone. Let

$$X_1, X_2 \in S_+^n$$

For any $\theta_1 \geq 0, \theta_2 \geq 0$,

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

2-10: Positive Semidefinite Cone II

- 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

2-12: Interaction I

- 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

2-13: Affine Function I

- $f(S)$ is convex:

  Let

$$f(x_1) \in f(S), f(x_2) \in f(S)$$

$$\begin{aligned}
&\alpha f(x_1) + (1-\alpha) f(x_2) \\
=&\alpha(Ax_1 + b) + (1-\alpha)(Ax_2 + b) \\
=&A(\alpha x_1 + (1-\alpha)x_2) + b \\
\in& f(S)
\end{aligned}$$

2-13: Affine Function II

- $f^{-1}(C)$ 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,

$$\begin{aligned} &\alpha(Ax_1 + b) + (1-\alpha)(Ax_2 + b) \\ =&A(\alpha x_1 + (1-\alpha)x_2) + b \in C \end{aligned}$$

Thus

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

2-13: Affine Function III

- 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

2-13: Affine Function IV

For example, for projection

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

I: identity matrix

- Solution set of linear matrix inequality

$$C = \{S \mid S \preceq 0\} \text{ is convex}$$
$$f(x) = x_1 A_1 + \cdots + x_m A_m - B = Ax + b$$
$$f^{-1}(C) = \{x \mid f(x) \preceq 0\} \text{ is convex}$$

2-13: Affine Function V

But this isn't rigourous because of some problems in arguing

$$f(x) = Ax + b$$

A more formal explanation:

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

is convex

$$\begin{aligned} f(x) &= x_1 \text{vec}(A_1) + \cdots + x_m \text{vec}(A_m) - \text{vec}(B) \\ &= \begin{bmatrix} \text{vec}(A_1) & \cdots & \text{vec}(A_m) \end{bmatrix} x + (-\text{vec}(B)) \end{aligned}$$

2-13: Affine Function VI

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

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)

2-13: Affine Function VII

- We have that

$$f(x) = \begin{bmatrix} P^{1/2}x \\ c^T x \end{bmatrix} = \begin{bmatrix} P^{1/2} \\ c^T \end{bmatrix} x$$

is affine. Then

$$\begin{aligned} f^{-1}(C) &= \{x \mid f(x) \in C\} \\ &= \{x \mid x^T P x \le (c^T x)^2, c^T x \ge 0\} \end{aligned}$$

is convex

Perspective and linear-fractional function I

- 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

Perspective and linear-fractional function II

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

Perspective and linear-fractional function III

- We have

$$\begin{aligned}
\alpha\frac{x_1}{t_1} + (1-\alpha)\frac{x_2}{t_2} &= \frac{\alpha t_2 x_1 + (1-\alpha) t_1 x_2}{t_1 t_2} \\
&= \frac{\alpha t_2 x_1 + (1-\alpha) t_1 x_2}{\alpha t_1 t_2 + (1-\alpha) t_1 t_2} \\
&= \frac{\frac{\alpha t_2}{\alpha t_2 + (1-\alpha) t_1} x_1 + \frac{(1-\alpha) t_1}{\alpha t_2 + (1-\alpha) t_1} x_2}{\frac{\alpha t_2}{\alpha t_2 + (1-\alpha) t_1} t_1 + \frac{(1-\alpha) t_1}{\alpha t_2 + (1-\alpha) t_1} t_2}
\end{aligned}$$

Perspective and linear-fractional function IV

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

$$S \text{ is convex}$$

- 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,

$$\frac{\theta x_1 + (1-\theta)x_2}{\theta t_1 + (1-\theta)t_2} \in C?$$

Perspective and linear-fractional function VII

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 VIII

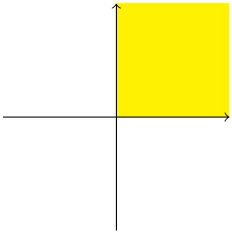
$$t_1\theta = \alpha t_2 - \alpha t_2\theta + \alpha t_1\theta$$

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

- K contains no line:

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

- Nonnegative orthant



Clearly all properties are satisfied

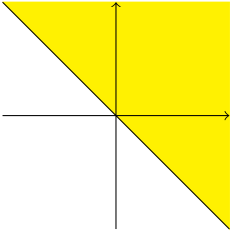
2-16: Generalized inequalities II

- Positive semidefinite cone:
  PD matrices are interior
- Nonnegative polynomial on $[0, 1]$
- When $n = 2$

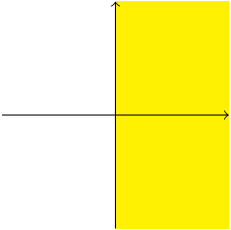
$$x_1 \geq -tx_2, \forall t \in [0, 1]$$

- $t = 1$

2-16: Generalized inequalities III

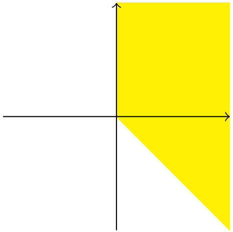


- $t = 0$



2-16: Generalized inequalities IV

- $\forall t \in [0, 1]$



- It really becomes a proper cone

2-17: I

- Properties:

$$x \preceq_K y, u \preceq_K v$$

  implies that

$$y - x \in K$$
$$v - u \in K$$

- From the definition of a convex cone,

$$(y - x) + (v - u) \in K$$

  Then

$$x + u \preceq_K y + v$$

2-18: Minimum and minimal elements I

- The minimum element

$$S \subseteq x_1 + K$$

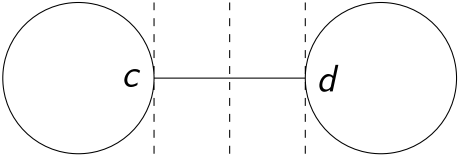
- A minimal element

$$(x_2 - K) \cap S = \{x_2\}$$

2-19: Separating hyperplane theorem I

- 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
$$a \equiv d - c, \quad b \equiv \frac{\|d\|^2 - \|c\|^2}{2}$$
forms a separating hyperplane $a^T x = b$

2-19: Separating hyperplane theorem II



$$a^T x = \Delta_2 \qquad b \qquad \Delta_1$$

$$a = d - c$$

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

$$b = \frac{\Delta_1 + \Delta_2}{2} = \frac{d^T d - c^T c}{2}$$

2-19: Separating hyperplane theorem III

Assume the result is wrong so there is $u \in D$ such that

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

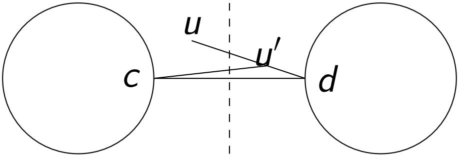
We will derive a point u' in D but it is closer to c than d. That is,

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

Then we have a contradiction

- The concept

2-19: Separating hyperplane theorem IV



$$a^T x = b$$

$$a^T u - b = (d-c)^T u - \frac{d^T d - c^T c}{2} < 0$$

implies that

$$(d-c)^T (u-d) + \frac{1}{2}\|d-c\|^2 < 0$$

2-19: Separating hyperplane theorem V

$$\begin{aligned}
&\frac{d}{dt}\|d + t(u - d) - c\|^2\Big|_{t=0} \\
=&2(d + t(u - d) - c)^T(u - d)\Big|_{t=0} \\
=&2(d - c)^T(u - d) < 0
\end{aligned}$$

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

$$\|d + t(u - d) - c\| < \|d - c\|$$

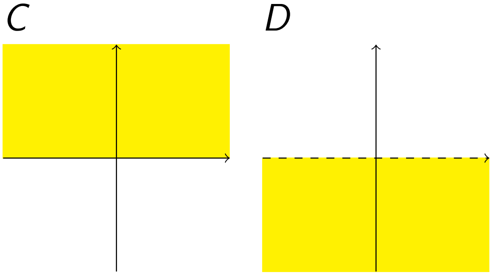
However,

$$d + t(u - d) \in D,$$

so there is a contradiction

2-19: Separating hyperplane theorem VI

- Strict separation



They are disjoint convex sets. However, no a, b such that

$$a^T x < b, \forall x \in C \text{ and } a^T x > b, \forall x \in D$$

2-20: Supporting hyperplane theorem I

Case 1: C has an interior region

- Consider 2 sets:

$$\text{interior of } C \text{ versus } \{x_0\},$$

  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

$$a^T x \leq a^T x_0, \forall x \in \text{interior of } C$$

2-20: Supporting hyperplane theorem II

- Then for all boundary point x we also have

$$a^T x \leq a^T x_0$$

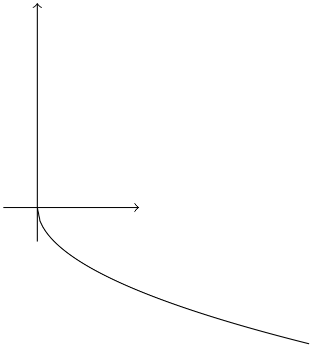
  because any boundary point is the limit of interior points

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
- We don't do a rigourous proof here

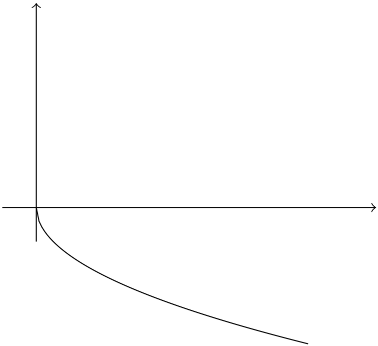
3-3: Examples on R I

- Example: $x^3, x \geq 0$



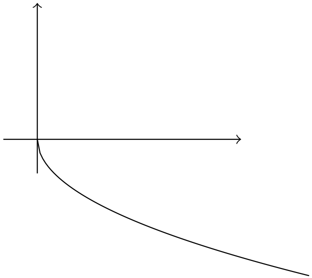
- Example: $x^{-1}, x > 0$

3-3: Examples on R II



- Example: $x^{1/2}, x \geq 0$

3-3: Examples on R III



3-4: Examples on R^n and $R^{m\times n}$ I

$$A, X \in R^{m\times n}$$

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

$$= \sum_j \sum_i A^T_{ji} X_{ij} = \sum_j \sum_i A_{ij} X_{ij}$$

3-7: First-order Condition I

- An open set: for any x, there is a ball covering x such that this ball is in the set
- Global underestimator:

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

$$z = f(x) + f'(x)(y - x)$$

3-7: First-order Condition II

- $\Rightarrow$: Because domain f is convex,

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

$$f(x + t(y - x)) \leq (1 - t)f(x) + tf(y)$$

$$f(y) \geq f(x) + \frac{f(x + t(y - x)) - f(x)}{t}$$

when $t \rightarrow 0$,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

- $\Leftarrow$:

3-7: First-order Condition III

For any $0 \le \theta \le 1$,

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

$$\begin{aligned} f(x) \ge & f(z) + \nabla f(z)^T (x - z) \\ = & f(z) + \nabla f(z)^T (1-\theta)(x-y) \end{aligned}$$

$$\begin{aligned} f(y) \ge & f(z) + \nabla f(z)^T (y - z) \\ = & f(z) + \nabla f(z)^T \theta (y-x) \end{aligned}$$

$$\theta f(x) + (1-\theta) f(y) \ge f(z)$$

3-7: First-order Condition IV

- First-order condition for strictly convex function:

$$f \text{ is strictly convex if and only if}$$
$$f(y) > f(x) + \nabla f(x)^T (y - x)$$

- $\Leftarrow$: it's easy by directly modifying $\geq$ to $>$

$$\begin{aligned} f(x) >& f(z) + \nabla f(z)^T (x - z) \\ =& f(z) + \nabla f(z)^T (1 - \theta)(x - y) \end{aligned}$$

$$f(y) > f(z) + \nabla f(z)^T (y - z) = f(z) + \nabla f(z)^T \theta (y - x)$$

3-7: First-order Condition V

- $\Rightarrow$: 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

$$\begin{aligned} f(x + t(y - x)) - f(x) &< tf(y) - tf(x) \\ &= \nabla f(x)^T t(y - x), \forall t \in (0, 1) \end{aligned}$$

3-7: First-order Condition VI

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

3-8: Second-order condition I

- Proof of the 2nd-order condition:
  We consider only the simpler condition of $n = 1$
- $\Rightarrow$

$$f(x+t) \geq f(x) + f'(x)t$$

$$\lim_{t \to 0} 2\frac{f(x+t) - f(x) - f'(x)t}{t^2}$$

$$= \lim_{t \to 0} \frac{2(f'(x+t) - f'(x))}{2t} = f''(x) \geq 0$$

3-8: Second-order condition II

- "$\Leftarrow$"

$$\begin{aligned} f(x+t) &= f(x) + f'(x)t + \frac{1}{2}f''(\bar{x})t^2 \\ &\geq f(x) + f'(x)t \end{aligned}$$

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

3-8: Second-order condition III

Using 1st-order condition for strictly convex function:

$$
\begin{aligned}
& f(y) \\
=& f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(\bar{x})(y-x) \\
>& f(x) + \nabla f(x)^T (y-x)
\end{aligned}
$$

- It's possible that f is strictly convex but

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

3-8: Second-order condition IV

- Example:

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

  Details omitted

3-9: Examples I

- Quadratic-over-linear

$$\frac{\partial f}{\partial x} = \frac{2x}{y}, \frac{\partial f}{\partial y} = -\frac{x^2}{y^2}$$

$$\frac{\partial^2 f}{\partial x \partial x} = \frac{2}{y}, \frac{\partial^2 f}{\partial x \partial y} = -\frac{2x}{y^2}, \frac{\partial^2 f}{\partial y \partial y} = \frac{2x^2}{y^3},$$

$$\begin{aligned} &\frac{2}{y^3} \begin{bmatrix} y \\ -x \end{bmatrix} \begin{bmatrix} y & -x \end{bmatrix} \\ =&\frac{2}{y^3} \begin{bmatrix} y^2 & -xy \\ -xy & x^2 \end{bmatrix} = 2 \begin{bmatrix} \frac{1}{y} & -\frac{x}{y^2} \\ -\frac{x}{y^2} & \frac{x^2}{y^3} \end{bmatrix} \end{aligned}$$

$$f(x) = \log \sum_k \exp x_k$$

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

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

Note that if

$$z_k = \exp x_k$$

3-10 II

then

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

Cauchy-Schwarz inequality

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

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

Note that

$$z_k > 0$$

3-12: Jensen's inequality I

- General form

$$f\left(\int p(z) z dz\right) \leq \int p(z) f(z) dz$$

- Discrete situation

$$f\left(\sum p_i z_i\right) \leq \sum p_i f(z_i), \sum p_i = 1$$

3-12: Jensen's inequality II

- Proof:

$$
\begin{aligned}
& f(p_1 z_1 + p_2 z_2 + p_3 z_3) \\
\leq & (1-p_3)\left(f\left(\frac{p_1 z_1 + p_2 z_2}{1-p_3}\right)\right) + p_3 f(z_3) \\
\leq & (1-p_3)\left(\frac{p_1}{1-p_3} f(z_1) + \frac{p_2}{1-p_3} f(z_2)\right) + p_3 f(z_3) \\
= & p_1 f(z_1) + p_2 f(z_2) + p_3 f(z_3)
\end{aligned}
$$

- Note that

$$
\frac{p_1}{1-p_3} + \frac{p_2}{1-p_3} = \frac{1-p_3}{1-p_3} = 1
$$

3-14: Positive weighted sum & composition with affine function I

- Composition with affine function:

  We know

  $$f(x) \text{ is convex}$$

  Is

  $$g(x) = f(Ax + b)$$

3-14: Positive weighted sum & composition with affine function II

convex?

$$
\begin{aligned}
& g((1-\alpha)x_1 + \alpha x_2) \\
=& f(A((1-\alpha)x_1 + \alpha x_2) + b) \\
=& f((1-\alpha)(Ax_1 + b) + \alpha(Ax_2 + b)) \\
\leq& (1-\alpha)f(Ax_1 + b) + \alpha f(Ax_2 + b) \\
=& (1-\alpha)g(x_1) + \alpha g(x_2)
\end{aligned}
$$

3-15: Pointwise maximum I

- Proof of the convexity

$$
\begin{aligned}
& f((1-\alpha)x_1+\alpha x_2) \\
=& \max(f_1((1-\alpha)x_1+\alpha x_2),\ldots,f_m((1-\alpha)x_1+\alpha x_2)) \\
\leq& \max((1-\alpha)f_1(x_1)+\alpha f_1(x_2),\ldots, \\
& \quad (1-\alpha)f_m(x_1)+\alpha f_m(x_2)) \\
\leq& (1-\alpha)\max(f_1(x_1),\ldots,f_m(x_1))+ \\
& \quad \alpha\max(f_1(x_2),\ldots,f_m(x_2)) \\
\leq& (1-\alpha)f(x_1)+\alpha f(x_2)
\end{aligned}
$$

3-15: Pointwise maximum II

- For
$$f(x) = x_{[1]} + \cdots + x_{[r]}$$
consider all
$$\binom{n}{r}$$
combinations

3-16: Pointwise supremum I

- 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 eigenvales 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

$$f(x_1, y_1) \leq g(x_1) + \epsilon$$

$$f(x_2, y_2) \leq g(x_2) + \epsilon$$

$$\begin{aligned} g(\theta x_1 + (1-\theta)x_2) &= \inf_{y \in C} f(\theta x_1 + (1-\theta)x_2, y) \\ &\leq f(\theta x_1 + (1-\theta)x_2, \theta y_1 + (1-\theta)y_2) \\ &\leq \theta f(x_1, y_1) + (1-\theta) f(x_2, y_2) \\ &\leq \theta g(x_1) + (1-\theta) g(x_2) + \epsilon \end{aligned}$$

3-19: Minimization II

- Note that the first inequality use the peroperty 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:

3-19: Minimization III

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

$$\min_y f(x, y)$$

3-19: Minimization IV

Because

$$C \succ 0,$$

the minimum occurs at

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

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

Then

$$\begin{aligned} g(x) &= x^T Ax - 2x^T BC^{-1}Bx + x^T BC^{-1}CC^{-1}B^T x \\ &= x^T(A - BC^{-1}B^T)x \end{aligned}$$

3-19: Minimization V

is convex. The second-order condition implies that

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

3-21: the conjugate function I

- This function is useful later
- When y is fixed, maximum happens at

$$y = f'(x) \tag{2}$$

by taking the derivative on x

3-21: the conjugate function II

- 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,

$$yx \text{ and } f(x)$$

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 $x = 0$,

$$z = -x_0 f'(x_0) + f(x_0) = -x_0 y + f(x_0) = -f^*(y)$$

3-21: the conjugate function IV

- f^* is convex: Given x,

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

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

3-22: examples I

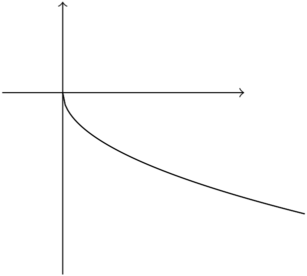
- negative logarithm

$$f(x) = -\log x$$

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

If $y < 0$, pictures of xy and $\log x$ are

3-22: examples II



Thus

$$xy + \log x$$

has maximum. Then

$$xy + \log x = -1 - \log(-y)$$

3-22: examples III

- strictly convex quadratic

$$Qx = y, x = Q^{-1}y$$

$$\begin{aligned}
& y^T x - \frac{1}{2} x^T Q x \\
= & y^T Q^{-1} y - \frac{1}{2} y^T Q^{-1} Q Q^{-1} y \\
= & \frac{1}{2} y^T Q^{-1} y
\end{aligned}$$

3-23: quasiconvex functions I

- Figure on slide:

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

  Both are convex
- The figure is an example showing that quasi convex may not be convex

3-26: properties of quasiconvex functions I

- Modified Jensen inequality:
  f quasiconvex if and only if

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

- $\Rightarrow$ Let

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

  S_Δ is convex

  $$x \in S_\Delta, y \in S_\Delta$$

  $$\theta x + (1-\theta)y \in S_\Delta$$

3-26: properties of quasiconvex functions II

$$f(\theta x + (1-\theta)y) \leq \Delta$$

and the result is obtained

- $\Leftarrow$ If results are wrong, there exists α such that S_α is not convex.

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

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

3-26: properties of quasiconvex functions III

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

"$\Rightarrow$"

$$f((1-t)x+ty) \le \max(f(x), f(y)) = f(x)$$

$$\frac{f(x+t(y-x))-f(x)}{t} \le 0$$

$$\lim_{t\to 0} \frac{f(x+t(y-x))-f(x)}{t} = \nabla f(x)^T(y-x) \le 0$$

3-26: properties of quasiconvex functions V

- $\Leftarrow$: If results are wrong, there exists α such that S_α 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

$$f(\theta x + (1 - \theta) y) > \alpha \geq \max\{f(x), f(y)\} \qquad (3)$$

3-26: properties of quasiconvex functions VI

Because f is differentiable, it is continuous. Without loss of generality, we have

$$f(z) \geq f(x), f(y), \forall z \text{ between } x \text{ and } y \tag{4}$$

Let's give a 1-D interpretation. From (3), we can find a ball surrounding

$$\theta x + (1-\theta)y$$

3-26: properties of quasiconvex functions VII

and two points x', y' such that

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

With (4),

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

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

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

3-26: properties of quasiconvex functions VIII

Then

$$\nabla f(z)^T(y-x) = 0, \forall \theta \in (0,1)$$

$$\begin{aligned} & f(x+\theta(y-x)) \\ =& f(x) + \nabla f(t)^T \theta(y-x) \\ =& f(x), \forall \theta \in [0,1) \end{aligned}$$

This contradicts (3).

3-27: Log-concave and log-convex functions I

- Powers:

$$\log(x^a) = a \log x$$

$$\log x \text{ is concave}$$

- Probability densities:

$$\log f(x) = -\frac{1}{2}(x - \bar{x})^T \Sigma^{-1}(x - \bar{x}) + \text{ constant}$$

Σ^{-1} is positive definite. Thus $\log f(x)$ is concave

3-27: Log-concave and log-convex functions II

- Cumulative Gaussian distribution

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

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

$$\frac{d^2}{d^2x} \log \Phi(x)$$

$$= \frac{(\int_{-\infty}^{x} e^{-u^2/2} du) e^{-x^2/2}(-x) - e^{-x^2/2} e^{-x^2/2}}{(\int_{-\infty}^{x} e^{-u^2/2} du)^2}$$

3-27: Log-concave and log-convex functions III

Need to prove that

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

- Because

$$x \geq u \text{ for all } u \in (-\infty, x],$$

3-27: Log-concave and log-convex functions IV

we have

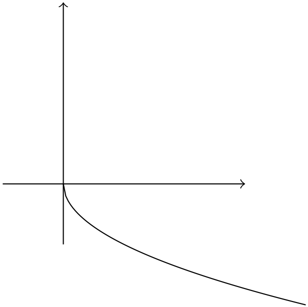
$$
\begin{aligned}
&\left(\int_{-\infty}^{x} e^{-u^2/2} du\right) x + e^{-x^2/2} \\
= &\int_{-\infty}^{x} x e^{-u^2/2} du + e^{-x^2/2} \\
\geq &\int_{-\infty}^{x} u e^{-u^2/2} du + e^{-x^2/2} \\
= &- e^{-u^2/2}\Big|_{-\infty}^{x} + e^{-x^2/2} \\
= &- e^{-x^2/2} + e^{-x^2/2} = 0
\end{aligned}
$$

3-27: Log-concave and log-convex functions V

This proof was given by a student (and polished by another student) of this course before

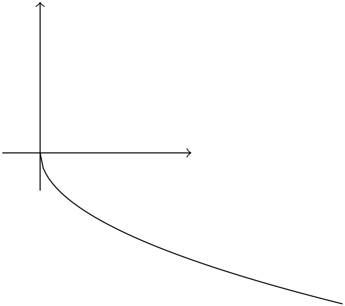
4-3: Optimal and locally optimal points I

- $f_0(x) = 1/x$



- $f_0(x) = x \log x$

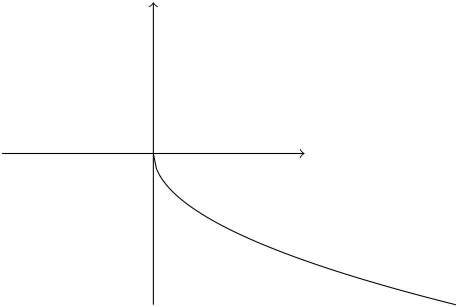
4-3: Optimal and locally optimal points II



$$f_0'(x) = 1 + \log x = 0$$
$$x = e^{-1} = 1/e$$

- $f_0(x) = x^3 - 3x$

4-3: Optimal and locally optimal points III



$$f_0'(x) = 3x^2 - 3 = 0$$

$$x = \pm 1$$

4-7: example I

$$
\begin{aligned}
& x_1/(1+x_2^2) \le 0 \\
\Leftrightarrow \quad & x_1 \le 0
\end{aligned}
$$

$$
\begin{aligned}
& (x_1+x_2)^2 = 0 \\
\Leftrightarrow \quad & x_1 + x_2 = 0
\end{aligned}
$$

4-9: Optimality criterion for differentiable f_0 I

- $\Leftarrow$: easy
  From first-order condition
  $$f_0(y) \geq f_0(x) + \nabla f_0(x)^T (y - x)$$
  Together with
  $$\nabla f_0(x)^T (y - x) \geq 0$$
  we have
  $$f_0(y) \geq f_0(x), \text{ for all feasible } y$$

4-9: Optimality criterion for differentiable f_0 II

- $\Rightarrow$ Assume the result is wrong. Then

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

Let

$$z(t) = ty + (1-t)x$$

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

$$\left.\frac{d}{dt} f_0(z(t))\right|_{t=0} = \nabla f_0(x)^T(y-x) < 0$$

4-9: Optimality criterion for differentiable f_0 III

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

4-10 I

- 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

4-10 II

$\Leftarrow$ Easy. For any feasible y,

$$Ay = b$$

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

So x is optimal

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

4-10 III

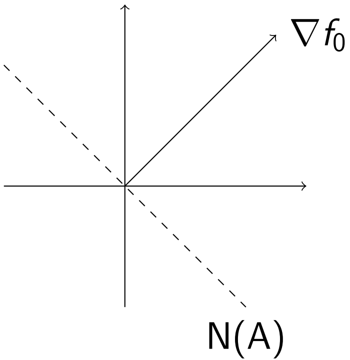
From optimality condition

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

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

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

4-10 IV



4-10 V

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

  $\Leftarrow$ Easy

4-10 VI

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

$$x \text{ is optimal}$$

4-10 VII

$\Rightarrow$ If $x_i = 0$, we claim

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

Otherwise,

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

Let

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

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

This violates the optimality condition

4-10 VIII

If $x_i > 0$, we claim

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

Otherwise, assume

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

Consider

$$y = x \text{ except } y_i = x_i/2 > 0$$

4-10 IX

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

4-23: examples I

- least-squares

$$\min x^T(A^TA)x - 2b^TAx + b^Tb$$

A^TA may not be invertibe $\Rightarrow$ pseudo inverse

- linear program with random cost

$$\bar{c} \equiv E(C)$$

$$\Sigma \equiv E_C((C-\bar{c})(C-\bar{c})^T)$$

4-23: examples II

$$
\begin{aligned}
\mathrm{Var}(C^T x) &= E_C((C^T x - \bar{c}^T x)(C^T x - \bar{c}^T x)) \\
&= E_C(x^T (C - \bar{c})(C - \bar{c})^T x) \\
&= x^T \Sigma x
\end{aligned}
$$

4-25: second-order cone programming I

- Cone was defined on slide 2-8

$$\{(x, t) \mid \|x\| \leq t\}$$

4-35: generalized inequality constraint I

- $f_i \in R^n \to R^{k_i}$ 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

4-37: LP and SOCP as SDP I

- LP and equivalent SDP

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

$$x_1 \begin{bmatrix} a_{11} & & \\ & \ddots & \\ & & a_{m1} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} & & \\ & \ddots & \\ & & a_{mn} \end{bmatrix}$$

$$- \begin{bmatrix} b_1 & & \\ & \ddots & \\ & & b_m \end{bmatrix} \preceq 0$$

4-37: LP and SOCP as SDP II

- For SOCP and SDP we will use results in 4-39:

$$\begin{bmatrix} tI_{p\times p} & A_{p\times q} \\ A^T & tI_{q\times q} \end{bmatrix} \succeq 0 \Leftrightarrow A^TA \preceq t^2 I_{q\times q}, t \geq 0$$

- Now

$$p = m, q = 1$$
$$A = A_i x + b_i, t = c_i^T x + d_i$$
$$\|A_i x + b_i\|^2 \leq (c_i^T x + d_i)^2,$$
$$c_i^T x + d_i \geq 0 \text{ from } t \geq 0$$

- Thus

$$\|A_i x + b_i\| \leq c_i^T x + d_i$$

4-39: matrix norm minimization I

- Following 4-38, we have the following equivalent problem

$$\begin{array}{ll} \min & t \\ \text{subject to} & \|A\|_2 \le t \end{array}$$

- We then use

$$\begin{aligned} \|A\|_2 \le t &\Leftrightarrow A^T A \preceq t^2 I, t \ge 0 \\ &\Leftrightarrow \begin{bmatrix} tI & A \\ A^T & tI \end{bmatrix} \succeq 0 \end{aligned}$$

4-39: matrix norm minimization II

to have the SDP

$$\begin{array}{rl} \min & t \\ \text{subject to} & \begin{bmatrix} tI & A(x) \\ A(x)^T & tI \end{bmatrix} \succeq 0 \end{array}$$

- Next we prove

$$\begin{bmatrix} tI_{p\times p} & A_{p\times q} \\ A^T & tI_{q\times q} \end{bmatrix} \succeq 0 \Leftrightarrow A^TA \preceq t^2 I_{q\times q}, t \geq 0$$

4-39: matrix norm minimization III

- $\Rightarrow$ we immediately have

$$t \geq 0$$

If $t > 0$,

$$\begin{aligned}
&\begin{bmatrix} -v^T A^T & tv^T \end{bmatrix} \begin{bmatrix} tI_{p\times p} & A_{p\times q} \\ A^T & tI_{q\times q} \end{bmatrix} \begin{bmatrix} -Av \\ tv \end{bmatrix} \\
=&\begin{bmatrix} -v^T A^T & tv^T \end{bmatrix} \begin{bmatrix} -tAv + tAv \\ -A^T Av + t^2 v \end{bmatrix} \\
=&t(t^2 v^T v - v^T A^T Av) \geq 0
\end{aligned}$$

$$v^T(t^2 I - A^T A)v \geq 0, \forall v$$

4-39: matrix norm minimization IV

and hence

$$t^2 I - A^T A \succeq 0$$

If $t = 0$

$$\begin{aligned}
&\begin{bmatrix} -v^T A^T & v^T \end{bmatrix} \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} -Av \\ v \end{bmatrix} \\
= &\begin{bmatrix} -v^T A^T & v^T \end{bmatrix} \begin{bmatrix} Av \\ -A^T Av \end{bmatrix} \\
= &-2v^T A^T Av \geq 0, \forall v
\end{aligned}$$

Therefore

$$A^T A \preceq 0$$

4-39: matrix norm minimization V

- $\Leftarrow$ Consider

$$\begin{aligned} &\begin{bmatrix} u^T & v^T \end{bmatrix} \begin{bmatrix} tI & A \\ A^T & tI \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \\ = &\begin{bmatrix} u^T & v^T \end{bmatrix} \begin{bmatrix} tu + Av \\ A^T u + tv \end{bmatrix} \\ = &tu^T u + 2v^T A^T u + tv^T v \end{aligned}$$

We hope to have

$$tu^T u + 2v^T A^T u + tv^T v \geq 0, \forall (u, v)$$

4-39: matrix norm minimization VI

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

$$\begin{aligned} & tu^T u + 2v^T A^T u + tv^T v \\ =& tv^T v - \frac{v^T A^T Av}{t} \\ =& \frac{1}{t} v^T (t^2 I - A^T A) v \geq 0. \end{aligned}$$

4-39: matrix norm minimization VII

Hence

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

If $t = 0$

$$A^T A \preceq 0$$

$$v^T A^T A v \leq 0, v^T A^T A v = \|Av\|^2 = 0$$

Thus

$$Av = 0, \forall v$$

$$\begin{bmatrix} u^T & v^T \end{bmatrix} \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

$$= \begin{bmatrix} u^T & v^T \end{bmatrix} \begin{bmatrix} 0 \\ A^T u \end{bmatrix} = 0 \geq 0$$

4-39: matrix norm minimization VIII

Thus

$$\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \succeq 0$$

4-40: Vector optimization I

- 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
$$f_0(\theta x + (1-\theta)y) \preceq_K \theta f_0(x) + (1-\theta) f_0(y)$$

4-41: optimal and pareto optimal points I

- See definition in slide 2-38
- Optimal

$$O \subseteq \{x\} + K$$

- Pareto optimal

$$(x - K) \cap O = \{x\}$$

5-3: Lagrange dual function I

- 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 λ, ν 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

$$\begin{aligned} & -\sup(-\cdots) = -\text{convex} \\ = & \inf(\cdots) = \text{concave} \end{aligned}$$

5-8: Lagrange dual and conjugate function I

$$
\begin{aligned}
& f_0^*(-A^T\lambda - c^T\nu) \\
= & \sup_x ((-A^T\lambda - c^T\nu)^T x - f_0(x)) \\
= & -\inf_x (f_0(x) + (A^T\lambda + c^T\nu)^T x)
\end{aligned}
$$

5-9: The dual problem I

- From 5-5, the dual problem is

$$\begin{aligned} \max \quad & g(\lambda, \nu) \\ \text{subject to} \quad & \lambda \succeq 0 \end{aligned}$$

- It can be simplified to

$$\begin{aligned} \max \quad & -b^T \nu \\ \text{subject to} \quad & A^T \nu - \lambda + c = 0 \\ & \lambda \succeq 0 \end{aligned}$$

5-9: The dual problem II

- Further,

$$\begin{array}{rl} \max & -b^T \nu \\ \text{subject to} & A^T \nu + c \succeq 0 \end{array}$$

5-10: weak and strong duality I

- We don't discuss the SDP problem on this slide because we omitted 5-7 on the two-way partitioning problem

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

5-12: inequality from LP I

- If we have only linear constraints, then constraint qualification holds

5-15: geometric interpretation I

- Explanation of $g(\lambda)$: when λ is fixed

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

  is a line. We lower Δ until it touches the boundary of G

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

- When

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

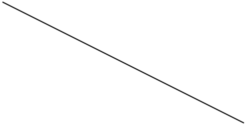
  so we see the point marked as $g(\lambda)$ on t-axis

5-15: geometric interpretation II

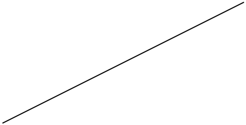
- We have $\lambda \geq 0$, so

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

must be like



rather than



5-15: geometric interpretation III

- Explanation of p^*:

  In G, only points satisfying

  $$u \leq 0$$

  are feasible

5-16 I

- We do not discuss a formal proof of

$$\text{Slater condition} \Rightarrow \text{strong duality}$$

  Instead, we explain this result by figures
- Reason of using A: G may not be convex
- Example:

$$\begin{aligned} \min \quad & x^2 \\ \text{subject to} \quad & x + 2 \leq 0 \end{aligned}$$

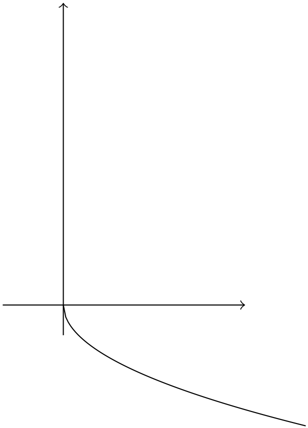
5-16 II

This is a convex optimization problem

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

is only a quadratic curve

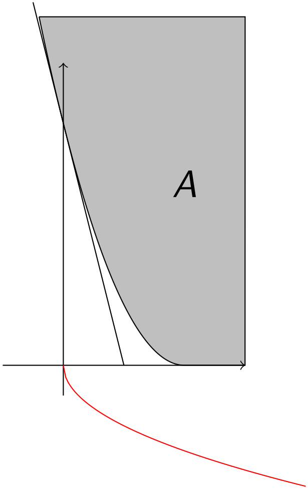
5-16 III



The curve is not convex

- However, A is convex

5-16 IV



5-16 V

- Primal problem:

$$x = -2$$

optimal objective value $= 4$

- Dual problem:

$$g(\lambda) = \min_x x^2 + \lambda(x+2)$$

$$x = -\lambda/2$$

$$\max_{\lambda \geq 0} -\frac{\lambda^2}{4} + 2\lambda$$

$$\text{optimal } \lambda = 4$$

5-16 VI

$$\text{optimal objective value } = -\frac{16}{4} + 8 = 4$$

- Proving that A is convex

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

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

$$f_1(x_1) \le u_1, f_0(x_1) \le t_1$$

$$f_1(x_2) \le u_2, f_0(x_2) \le t_2$$

Consider

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

5-16 VII

We have

$$f_1(x) \le \theta u_1 + (1-\theta)u_2$$

$$f_0(x) \le \theta t_1 + (1-\theta)t_2$$

So

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

- Why "non-vertical supporting hyperplane"?

  Then $g(\lambda)$ is well defined.
- Note that we have

$$\text{Slater condition} \Rightarrow \text{strong duality}$$

5-16 VIII

However, it's possible that Slater condition doesn't hold but strong duality holds

Example from exercise 5.22:

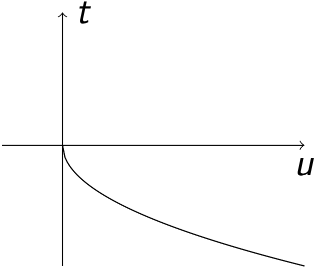
$$\begin{array}{rl} \min & x \\ \text{subject to} & x^2 \leq 0 \end{array}$$

Slater condition doesn't hold because no x satisfies

$$x^2 < 0$$

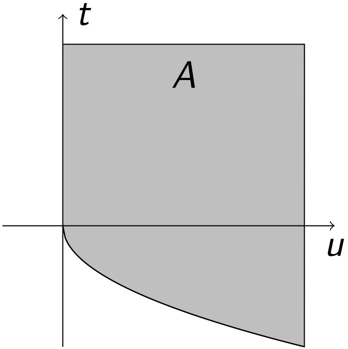
$$G = \{(x^2, x) \mid x \in R\}$$

5-16 IX



There is only one feasible point $(0, 0)$

5-16 X



5-16 XI

$$g(\lambda) = \min_x x + x^2\lambda$$

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

Dual problem

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

$$\lambda \to \infty, \text{ objective value } \to 0$$

$$d^* = 0, p^* = 0$$

Strong duality holds

5-17: complementary slackness I

- 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

5-17: complementary slackness II

- 4-10: x is optimal of

$$\begin{array}{rl} \min & f_0(x) \\ \text{subject to} & x_i \geq 0, \forall i \end{array}$$

if and only if

$$x_i \geq 0, \begin{cases} \nabla_i f_0(x) \geq 0 & x_i = 0 \\ \nabla_i f_0(x) = 0 & x_i > 0 \end{cases}$$

5-17: complementary slackness III

- From KKT condition

$$\nabla_i f_0(x) = \lambda_i$$
$$\lambda_i x_i = 0$$
$$\lambda_i \geq 0, x_i \geq 0$$

If

$$x_i > 0,$$

then

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

5-19: KKT conditions for convex problem I

- 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

5-19: KKT conditions for convex problem II

- Consider the situation of inequality constraints only:

$$\begin{array}{rl} \min & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, i = 1, \ldots, m \end{array}$$

- Consider an optimial solution x. We would like to see if x satisfies KKT condition
- We claim that

$$\nabla f_0(x) = \sum_{i:f_i(x)=0} -\lambda_i \nabla f_i(x) \tag{5}$$

5-19: KKT conditions for convex problem III

- Assume the result is wrong. Then,

$$\begin{aligned}\nabla f_0(x) = & \text{linear combination of } \{\nabla f_i(x) \mid f_i(x) = 0\} \\ & + \Delta,\end{aligned}$$

where

$$\Delta \neq 0 \text{ and } \Delta^T \nabla f_i(x) = 0, \forall i : f_i(x) = 0$$

5-19: KKT conditions for convex problem IV

- Then there exists $\alpha < 0$ such that

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

  satisfies

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

  and

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

- We claim that δx is feasible. That is,

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

5-19: KKT conditions for convex problem V

- We have

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

- However,

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

This contradicts the optimality condition that from slide 4-9, for any feasible direction δx,

$$\nabla f_0(x)^T \delta x \geq 0.$$

5-19: KKT conditions for convex problem VI

- We do not continue to prove

$$\nabla f_0(x) = \sum_{\lambda_i \geq 0, f_i(x)=0} -\lambda_i \nabla f_i(x) \tag{6}$$

  because the proof is not trivial
- However, what we want to say is that in proving (5), the proof is not rigourous because of $\approx$
- For linear the proof becomes rigourous
- This roughly give you a feeling that linear is different from non-linear

5-25 I

- Explanation of $f_0^*(\nu)$

$$
\begin{aligned}
&\inf_y (f_0(y) - \nu^T y) \\
= &- \sup_y (\nu^T y - f_0(y)) = -f_0^*(\nu)
\end{aligned}
$$

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

5-26 I

- 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 $\|\nu\|_* > 1$,

$$\nu^T y^* > 1, \|y^*\| \leq 1$$

5-26 II

$$\begin{aligned} &\inf \|y\| + \nu^T y \\ \leq &\|-y^*\| - \nu^T y^* < 0 \end{aligned}$$

$$\|-ty^*\| - \nu^T(ty^*) \to -\infty \text{ as } t \to \infty$$

Hence

$$\inf_y \|y\| + \nu^T y = -\infty$$

If $\|\nu\|_* \leq 1$, we claim that

$$\inf_y \|y\| + \nu^T y = 0$$

$$y = 0 \Rightarrow \|y\| + \nu^T y = 0$$

5-26 III

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

5-27: implicit constraint I

- The dual function

$$
\begin{aligned}
& c^T x + \nu^T (Ax - b) \\
= & -b^T \nu + x^T (A^T \nu + c)
\end{aligned}
$$

$$
\inf_{-1 \le x_i \le 1} x_i (A^T \nu + c)_i = -|(A^T \nu + c)_i|
$$

5-30: semidefinite program 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 produt between

  $$Z \text{ and } x_1F_1 + \cdots + x_nF_n - G$$

5-30: semidefinite program II

Trace is the component-wise product

$$
\begin{aligned}
& \operatorname{tr}(AB) \\
= & \sum_i (AB)_{ii} \\
= & \sum_i \sum_j A_{ij} B_{ji} = \sum_i \sum_j A_{ij} B_{ij}
\end{aligned}
$$

Note that we take the property that B is symmetric

7-4 I

- 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:

$$
\begin{aligned}
&\frac{\|a\|}{2} - \sum_i \lambda_i (a^T x_i + b - 1) + \sum_i \mu_i (a^T y_i + b + 1) \\
=&\frac{\|a\|}{2} + a^T\left(-\sum_i \lambda_i x_i + \sum_i \mu_i y_i\right) \\
&\quad + b\left(-\sum_i \lambda_i + \sum_i \mu_i\right) + \sum_i \lambda_i + \sum_i \mu_i
\end{aligned}
$$

Because of

$$
b\left(-\sum_i \lambda_i + \sum_i \mu_i\right)
$$

8-10: Dual of maximum margin problem II

we have

$$\inf_{a,b} L = \sum_i \lambda_i + \sum_i \mu_i +$$

$$\begin{cases} \inf_a \frac{\|a\|}{2} + a^T(-\sum_i \lambda_i x_i + \sum_i \mu_i y_i) & \text{if } \sum_i \lambda_i = \sum_i \mu_i \\ -\infty & \text{if } \sum_i \lambda_i \neq \sum_i \mu_i \end{cases}$$

For

$$\inf_a \frac{\|a\|}{2} + a^T(-\sum_i \lambda_i x_i + \sum_i \mu_i y_i)$$

8-10: Dual of maximum margin problem III

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:

8-10: Dual of maximum margin problem IV

- 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,

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

8-10: Dual of maximum margin problem V

Therefore

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

- If $\|v\| > 1/2$, let

$$a = \frac{-tv}{\|v\|}$$

$$\begin{aligned}
&\frac{\|a\|}{2} + v^T a \\
=&\frac{t}{2} - t\|v\| \\
=&t(\frac{1}{2} - \|v\|) \to -\infty \text{ if } t \to \infty
\end{aligned}$$

8-10: Dual of maximum margin problem VI

Thus

$$\inf_{a} \frac{\|a\|}{2} + v^T a = -\infty$$

- Finally,

$$\inf_{a,b} L = \sum_i \lambda_i + \sum_i \mu_i +$$

$$\begin{cases} 0 & \text{if } \sum_i \lambda_i = \sum_i \mu_i \text{ and} \\ & \|\sum_i \lambda_i x_i - \sum_i \mu_i y_i\| \leq 1/2 \\ -\infty & \text{otherwise} \end{cases}$$

8-14

$$\theta = \begin{bmatrix} \text{vec}(P) \\ q \\ r \end{bmatrix}, F(z) = \begin{bmatrix} \vdots \\ z_i z_j \\ \vdots \\ z_i \\ \vdots \\ 1 \end{bmatrix}$$

10-3: initial point and sublevel set I

- The condition that S is closed if

$$\text{domain of } f = R^n$$

Proof: Be definition S is closed if for every convergent sequence

$$\{x_i\} \text{ with } x_i \in S \text{ and } \lim_{i\to\infty} x_i = x^*,$$

then

$$x^* \in S.$$

10-3: initial point and sublevel set II

Because

$$\text{domain of } f = R^n$$

we have

$$x^* \in \text{domain of } f$$

Thus by the continuity of f,

$$\lim_{i \to \infty} f(x_i) = f(x^*) \le f(x_0)$$

and

$$x^* \in S$$

10-3: initial point and sublevel set III

- The condition that S is closed if

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

Proof: if not, from the definition of the closeness of S, there exists

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

such that

$$x_i \to x^* \notin \text{ domain } f$$

Thus

$$x^* \text{ is on the boundary}$$

10-3: initial point and sublevel set IV

Then

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

violates

$$f(x_i) \leq f(x_0), \forall i$$

Thus the assumption is wrong and S is closed

- Example

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

$$\text{domain } = R^n$$

10-3: initial point and sublevel set V

- Example

$$f(x) = -\sum_i \log(b_i - a_i^T x)$$

$$\text{domain } \neq R^n$$

We use the condition that

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

10-4: strong convexity and implications I

- 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

$$f(y) \leq f(x^0)$$

10-4: strong convexity and implications II

- 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

$$\nabla f(x) + m(y - x) = 0$$

10-4: strong convexity and implications III

$$\tilde{y} = x - \frac{\nabla f(x)}{m}$$

$$f(y) \geq f(x) + \nabla f(x)^T(\tilde{y} - x) + \frac{m}{2}\|\tilde{y} - x\|^2$$
$$= f(x) - \frac{1}{2m}\|\nabla f(x)\|^2, \forall y$$

Then

$$p^* \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2 > -\infty$$

and

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

10-5: descent methods I

- 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

$$t\nabla f(x)^T \Delta x \leq f(x + t\Delta x) - f(x) < 0$$

10-6: line search types I

- 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
$$f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^T \Delta x, \forall t \in (0, t^*)$$
Otherwise,
$$\exists \{t_k\} \to 0$$

10-6: line search types II

such that

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

$$\begin{aligned} &\lim_{t_k \to 0} \frac{f(x + t_k \Delta x) - f(x)}{t_k} \\ =&\nabla f(x)^T \Delta x \geq \alpha \nabla f(x)^T \Delta x \end{aligned}$$

However,

$$\nabla f(x)^T \Delta x < 0 \text{ and } \alpha \in (0, 1)$$

cause a contradiction

10-6: line search types III

- Graphical interpretation: the tangent line passes through $(0, f(x))$, so the equation is

$$\frac{y - f(x)}{t - 0} = \nabla f(x)^T \Delta x$$

  Because

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

  we see that the line of

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

10-6: line search types IV

is above that of

$$f(x) + t\nabla f(x)^T \Delta x$$

10-7 I

- Linear convergence. We consider exact line search; proof for backtracking line search is more complicated
- S closed and bounded

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

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

Solve

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

10-7 II

$$t = \frac{1}{M}$$

$$f(x_{\text{next}}) \le f(x - \frac{1}{M}\nabla f(x)) \le f(x) - \frac{1}{2M}\nabla f(x)^T \nabla f(x)$$

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

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

From slide 10-4,

$$-\|\nabla f(x)\|^2 \le -2m(f(x) - p^*)$$

10-7 III

Hence

$$f(x_{\text{next}}) - p^* \le \left(1 - \frac{m}{M}\right)(f(x) - p^*)$$

- Assume

$$x_1^k = \gamma(\frac{\gamma - 1}{\gamma + 1})^k, x_2^k = (-\frac{\gamma - 1}{\gamma + 1})^k,$$

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

$$\min_t \frac{1}{2}((x_1 - t x_1)^2 + \gamma(x_2 - t\gamma x_2)^2)$$

$$\min_t \frac{1}{2}(x_1^2(1 - t)^2 + \gamma x_2^2(1 - t\gamma)^2)$$

10-8 II

$$-x_1^2(1-t)+\gamma x_2^2(1-t\gamma)(-\gamma)=0$$

$$-x_1^2+tx_1^2-\gamma^2 x_2^2+\gamma^3 t x_2^2=0$$

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

$$t=\frac{x_1^2+\gamma^2 x_2^2}{x_1^2+\gamma^3 x_2^2}=\frac{\gamma^2(\frac{\gamma-1}{\gamma+1})^{2k}+\gamma^2(\frac{\gamma-1}{\gamma+1})^{2k}}{\gamma^2(\frac{\gamma-1}{\gamma+1})^{2k}+\gamma^3(\frac{\gamma-1}{\gamma+1})^{2k}}$$

$$=\frac{2\gamma^2}{\gamma^2+\gamma^3}=\frac{2}{1+\gamma}$$

$$x^{k+1}=x^k-t\nabla f(x^k)=\begin{bmatrix} x_1^k(1-t) \\ x_2^k(1-\gamma t) \end{bmatrix}$$

10-8 III

$$x_1^{k+1} = \gamma(\frac{\gamma-1}{\gamma+1})^k(\frac{\gamma-1}{1+\gamma}) = \gamma(\frac{\gamma-1}{\gamma+1})^{k+1}$$

$$\begin{aligned} x_2^{k+1} &= (-\frac{\gamma-1}{\gamma+1})^k(1-\frac{2\gamma}{1+\gamma}) \\ &= (-\frac{\gamma-1}{\gamma+1})^k(\frac{1-\gamma}{1+\gamma}) = (-\frac{\gamma-1}{\gamma+1})^{k+1} \end{aligned}$$

- Why gradient is orghogonal to the tangent line of the contour curve?

10-8 IV

Assume $f(g(t))$ is the countour with

$$g(0) = x$$

Then

$$\begin{aligned}
0 &= f(g(t)) - f(g(0)) \\
0 &= \lim_{t \to 0} \frac{f(g(t)) - f(g(0))}{t} \\
&= \lim_{t \to 0} \nabla f(g(t))^T \nabla g(t) \\
&= \nabla f(x)^T \nabla g(0)
\end{aligned}$$

10-8 V

where

$$x + t\nabla g(0)$$

is the tangent line

10-10 I

- linear convergence: from slide 10-7

$$f(x^k) - p^* \le c^k(f(x^0) - p^*)$$

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

is a straight line. Note that now k is the x-axis

10-11: steepest descent method I

- (unnormalized) steepest descent direction:

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

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

- We didn't discuss much about dual norm, but we can still explain some examples on 10-12

10-12 I

- Euclidean: Δx_{nsd} is by solving

$$\begin{array}{ll} \min & \nabla f^T v \\ \text{subject to} & \|v\| = 1 \end{array}$$

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

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

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

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

10-12 II

- Quadratic norm: Δx_{nsd} is by solving

$$\begin{array}{ll} \min & \nabla f^T v \\ \text{subject to} & v^T P v = 1 \end{array}$$

Now

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

where P is symmetric positive definite

10-12 III

- Let

$$w = P^{1/2} v$$

The optimization problem becomes

$$\begin{array}{ll} \min\limits_{w} & \nabla f^T P^{-1/2} w \\ \text{subject to} & \|w\| = 1 \end{array}$$

$$\begin{aligned} \text{optimal } w &= \frac{-P^{-1/2}\nabla f}{\|P^{-1/2}\nabla f\|} \\ &= \frac{-P^{-1/2}\nabla f}{\sqrt{\nabla f^T P^{-1} \nabla f}} \end{aligned}$$

10-12 IV

$$\text{optimal } v = \frac{-P^{-1}\nabla f}{\sqrt{\nabla f^T P^{-1} \nabla f}} = \Delta x_{\text{nsd}}$$

- Dual norm

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

Therefore

$$\begin{aligned} \Delta x_{\text{sd}} &= \sqrt{\nabla f^T P^{-1} \nabla f} \frac{-P^{-1}\nabla f}{\sqrt{\nabla f^T P^{-1} \nabla f}} \\ &= -P^{-1}\nabla f \end{aligned}$$

10-12 V

- Explanation of the figure:

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

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

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

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

- We omit the discusssion of l_1-norm

10-13 I

- The two figures are by using two P matrices
- The left one has faster convergence
- Gradient descent after change of variables

$$\bar{x} = P^{1/2}x, x = P^{-1/2}\bar{x}$$

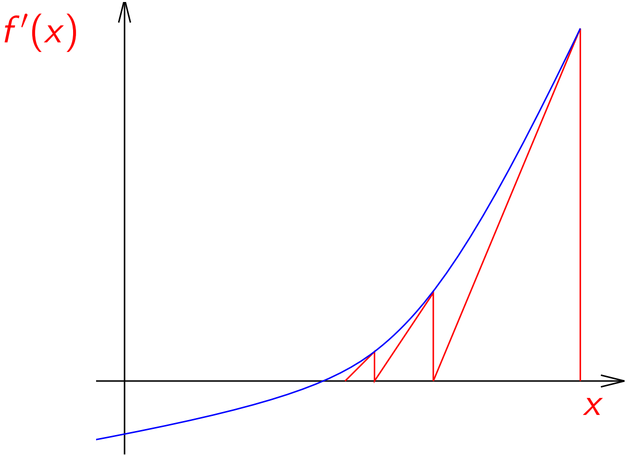
$$\min_x f(x) \Rightarrow \min_{\bar{x}} f(P^{-1/2}\bar{x})$$

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

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

$$x \leftarrow x - \alpha P^{-1}\nabla_x f(x)$$

10-14 I



10-14 II

- 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 $y = 0$

$$\boxed{x_{k+1} = x_k - f'(x_k)/f''(x_k)}$$

10-16 I

$$\hat{f}(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x)$$

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

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

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

$$f(x) - \inf_y \hat{f}(y) = \frac{1}{2}\lambda(x)^2$$

10-16 II

Norm of the Newton step in the quadratic Hessian norm

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

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

Directional derivative in the Newton direction

$$\begin{aligned}
&\lim_{t \to 0} \frac{f(x + t\Delta x_{\text{nt}}) - f(x)}{t} \\
=&\nabla f(x)^T \Delta x_{\text{nt}} \\
=&- \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) = -\lambda(x)^2
\end{aligned}$$

10-16 III

Affine invariant

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

Assume T is an invertable square matrix. Then

$$\bar{\lambda}(y) = \lambda(Ty)$$

Proof:

$$\nabla \bar{f}(y) = T^T \nabla f(Ty)$$
$$\nabla^2 \bar{f}(y) = T^T \nabla^2 f(Ty) T$$

10-16 IV

$$
\begin{aligned}
\bar{\lambda}(y)^2 &= \nabla \bar{f}(y)^T \nabla^2 \bar{f}(y)^{-1} \nabla \bar{f}(y) \\
&= \nabla f(Ty)^T T T^{-1} \nabla^2 f(Ty)^{-1} T^{-T} T^T \nabla f(Ty) \\
&= \nabla f(Ty)^T \nabla^2 f(Ty)^{-1} \nabla f(Ty) \\
&= \lambda(Ty)^2
\end{aligned}
$$

10-17 I

Affine invariant

$$\begin{aligned}\Delta y_{\text{nt}} &= -\nabla^2 \bar{f}(y)^{-1} \nabla \bar{f}(y) \\ &= -T^{-1} \nabla^2 f(Ty)^{-1} \nabla f(Ty) \\ &= T^{-1} \Delta x_{\text{nt}}\end{aligned}$$

Note that

$$y_k = T^{-1} x_k$$

so

$$y_{k+1} = T^{-1} x_{k+1}$$

10-17 II

But how about line search

$$
\begin{aligned}
&\nabla \bar{f}(y)^T \Delta y_{\text{nt}} \\
=&\nabla f(Ty)^T T T^{-1} \Delta x_{\text{nt}} \\
=&\nabla f(x)^T \Delta x_{\text{nt}}
\end{aligned}
$$

10-19 I

$$\eta \in (0, \frac{m^2}{L})$$

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

$$\frac{L}{2m^2}\|\nabla f(x_k)\| \leq \frac{1}{2}$$

10-20 I

$$
\begin{aligned}
& f(x_l) - f(x^*) \\
\leq & \frac{1}{2m} \|\nabla f(x_l)\|^2 \qquad \text{(from p10-4)} \\
\leq & \frac{1}{2m} \frac{4m^4}{L^2} (\frac{1}{2})^{2^{l-k} \cdot 2} \\
= & \frac{2m^3}{L^2} (\frac{1}{2})^{2^{l-k+1}} \leq \epsilon
\end{aligned}
$$

Let

$$
\epsilon_0 = \frac{2m^3}{L^2}
$$

10-20 II

$$\log_2 \epsilon_0 - 2^{l-k+1} \leq \log_2 \epsilon$$

$$2^{l-k+1} \geq \log_2(\epsilon_0/\epsilon)$$

$$l \geq k - 1 + \log_2 \log_2(\epsilon_0/\epsilon)$$

$$k \leq \frac{f(x_0) - p^*}{r}$$

In at most

$$\frac{f(x_0) - p^*}{r} + \log_2 \log_2(\epsilon_0/\epsilon)$$

10-20 III

iterations, we have

$$f(x_l) - f(x^*) \leq \epsilon$$

The second term is almost a constant. For example, if

$$\epsilon \approx 5 \cdot 10^{-20} \epsilon_0,$$

10-20 IV

then

$$
\begin{aligned}
& \log_2 \log_2 \frac{1}{5} 10^{20} \\
\approx \; & \log_2(1 + 19 \log_2 10) \\
\approx \; & \log_2(1 + 19 \cdot 3.322) \\
\approx \; & \log_2(64) = 6
\end{aligned}
$$

10-22 I

- On page 10-10, to reach

$$f(x^k) - p^* \approx 10^{-4},$$

  150 iterations are needed
- However, the cost per Newton iteration may be much higher
- Also for some applications we may not need a very accurate solution

10-29: implementation I

- If H is positive definite, then there exists unique L such that

$$H = LL^T$$

$$\begin{aligned}\lambda(x) &= (\nabla f(x) \nabla^2 f(x)^{-1} \nabla f(x))^{1/2} \\ &= (g^T L^{-T} L^{-1} g)^{1/2} = \|L^{-1} g\|_2\end{aligned}$$

10-30: example of dense Newton systems with structure I

$$\psi_i(x_i) : R \to R$$

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

10-30: example of dense Newton systems with structure II

$$\begin{aligned}
\nabla^2 f(x) &= \begin{bmatrix} \psi_1''(x_1) & & \\ & \ddots & \\ & & \psi_n''(x_n) \end{bmatrix} + A^T \nabla^2 \psi_0^2(Ax+b)A \\
&= D + A^T H_0 A
\end{aligned}$$

$$H_0 : p \times p$$

method 2:

$$\Delta x = D^{-1}(-g - A^T L_o w)$$

$$L_0^T A D^{-1}(-g - A^T L_0 w) = w$$

10-30: example of dense Newton systems with structure III

$$(I + L_0^T A D^{-1} A^T L_0) w = -L_0^T A D^{-1} g$$

Cost

$$\begin{aligned} & L_0 : p \times p, A : p \times n \\ & A^T L_0 : n \times p, \text{cost} : O(np^2) \\ & (L_0^T A) D^{-1} (A^T L_0) : O(p^2 n) \end{aligned}$$

Note that Cholesky factorization of H_0 costs

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

10-30: example of dense Newton systems with structure IV

as

$$p \ll n$$

Any problem fits into this framework? Logistic regression

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l} \log\left(1 + e^{-y_i\mathbf{w}^T\mathbf{x}_i}\right).$$

$$A = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_l^T \end{bmatrix}$$

10-30: example of dense Newton systems with structure V

$$\psi_0(\mathbf{t}) = C \sum_{i=1}^{l} \log(1 + e^{-y_i t_i})$$

$$\psi_0 : R^l \to R^1$$

This technique is useful if

$$\#\text{instances} \ll \#\text{features}$$

11-2 I

- For the constraint

$$Ax = b, A : p \times n$$

we assume

$$p < n$$

That is,

$$\#\text{constraints} < \#\text{variables}$$

This is reasonable. Otherwise in general the problem has a unique solution or is infeasible.

11-2 II

- With $p < n$ we can assume

$$\text{rank}(A) = p$$

11-3 I

- KKT matrix non-singular if and only if

$$Ax = 0, x \neq 0 \Rightarrow x^T P x > 0$$

$\Leftarrow$

If the result is wrong, then KKT matrix is singular

$$\exists \begin{bmatrix} x \\ v \end{bmatrix} \neq 0 \text{such that}$$

$$Px + A^T v = 0 \tag{7}$$

$$Ax = 0$$

11-3 II

Case 1: $x \neq 0$

$$x^T P x + x^T A^T v = x^T P x > 0 \text{ violates (7)}$$

Case 2: $x = 0$

$$A^T v = 0 \text{ and } v \neq 0 \text{ violates that } \text{rank}(A^T) = p,$$

where

$$A \in R^{n \times p}$$

That is, p columns of A^T are linear independent and hence $\text{rank}(A^T) < p$

11-3 III

$\Rightarrow$

If the result is wrong, $\exists x$ such that

$$Ax = 0, x \neq 0$$
$$x^T P x = 0$$

Since P is PSD, $P^{1/2}$ exists

$$(P^{1/2}x)^T(P^{1/2}x) = 0 \Rightarrow P^{1/2}x = 0 \Rightarrow Px = 0$$

$$\exists \begin{bmatrix} x \\ 0 \end{bmatrix} \neq 0 \text{ such that } \begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} = 0$$

contradicts the non-singularity

11-3 IV

- KKT matrix non-singular if and only if

$$P + A^T A \succ 0$$

$\Leftarrow$

If the result is wrong, the matrix is singular. That is, it does not have full rank. Thus, $\exists \begin{bmatrix} x \\ v \end{bmatrix} \neq 0$ such that

$$Px + A^T v = 0, Ax = 0$$

We claim that

$$x \neq 0$$

11-3 V

Otherwise,

$$x = 0, v \neq 0$$

imply that

$$A^T v = 0,$$

a contradiction to

$$\text{rank}(A^T) = p$$

That is, columns of A^T's p columns become linlinear dependent. Then

$$x^T(P + A^T A)x = x^T(-A^T v) = -(Ax)^T v = 0$$

leads to a contradiction

$\Rightarrow$

If

$$P + A^T A \not\succeq 0$$

$$\exists x \neq 0 \text{ such that } x^T P x + x^T A^T A x \leq 0$$

Because

$$P + A^T A \text{ is symmetric positive semi-definite,}$$

we have

$$\exists x \neq 0 \text{ such that } x^T P x + x^T A^T A x = 0$$

11-3 VII

Because P and A^TA are both PSD,

$$Ax = 0, x^TPx = 0$$

Then

$$\begin{bmatrix} x \\ 0 \end{bmatrix} \neq 0$$

is a solution of

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} = 0,$$

a contradiction

11-7: Newton decrement I

$$\nabla^2 f(x)\Delta x_{\text{nt}} + A^T w = -\nabla f(x)$$

$$\Delta x_{\text{nt}}^T \nabla^2 f(x)\Delta x_{\text{nt}} + 0 = -\Delta x_{\text{nt}}^T \nabla f(x) \tag{8}$$

$$\begin{aligned} & \frac{d}{dt} f(x + t\Delta x_{\text{nt}})\Big|_{t=0} \\ = \ & \nabla f(x)^T \Delta x_{\text{nt}} = -\lambda(x)^2 \end{aligned}$$

Note that

$$\nabla f(x)^T \Delta x_{\text{nt}} \leq 0$$

is from (8)

11-8 I

Original

$$\begin{array}{rl} \min & f(x) \\ \text{subject to} & Ax = b \end{array}$$

Let

$$x = Ty$$

New

$$\begin{array}{rl} \min\limits_{y} & f(Ty) = \bar{f}(y) \\ \text{subject to} & ATy = b = \bar{A}y, \end{array}$$

11-8 II

where

$$\bar{A} = AT$$

KKT system of the original one

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$$

New system

$$\begin{bmatrix} \nabla^2 \bar{f}(y) & \bar{A}^T \\ \bar{A} & 0 \end{bmatrix} \begin{bmatrix} \bar{v} \\ \bar{w} \end{bmatrix} = \begin{bmatrix} -\nabla \bar{f}(y) \\ 0 \end{bmatrix}$$

11-8 III

$$\begin{bmatrix} T^T\nabla^2 f(Ty)T & T^TA^T \\ AT & 0 \end{bmatrix} \begin{bmatrix} \bar{v} \\ \bar{w} \end{bmatrix} = \begin{bmatrix} -T^T\nabla f(Ty) \\ 0 \end{bmatrix}$$

If

$$x = Ty \Rightarrow \bar{v} = T^{-1}v, \bar{w} = w \text{ is a solution}$$

Let's omit the step size

$$\begin{aligned} y &\leftarrow y + \bar{v} \\ Ty &\leftarrow Ty + T\bar{v} \\ v &= T\bar{v} \\ x &\leftarrow x + v \end{aligned}$$

Thus invariant