Opti Mid3

tyang27

November 2019

1 Properties of matrices

Positive definite (PD), and Positive semidefinite (PSD).

1.1 Definite

Let $\mathrm{A} \in \mathbb{R}^{n \times n}$ be a symmetric matrix. A is

- PSD iff $\forall$ d $\in \mathbb{R}^n$ nonzero, $\mathrm{d}^T \mathrm{Ad} \geq 0$.
- PD iff $\forall$ d $\in \mathbb{R}^n$ nonzero, $\mathrm{d}^T \mathrm{Ad} > 0$.
- PSD iff all eigenvalues of A nonnegative (≥ 0).
- PD iff all eigenvalues of A positive (> 0). (Invertible!).

1.2 PSD and invertible implies PD

If $\mathrm{A} \in \mathbb{R}^{n \times n}$ is a symmetric, invertible, PSD matrix, then A is PD, and A^{-1} is PD as well.

- Since invertible, all eigenvalues are nonzero.
- Since PSD, eigenvalues are nonnegative.
- Since eigenvalues are nonzero and nonnegative, must be positive and PD.
- Since eigenvalues of inverse are just 1 over eigenvalue, which is still positive, so inverse is also PD.

$$eigenvalues(c\mathrm{A}) = c\lambda_i$$
$$eigenvalues(c\mathrm{I} + \mathrm{A}) = c + \lambda_i$$

2 Higher dimension Taylor

Note that hessian is symmetric.
Given function $f : \mathbb{R}^n \to \mathbb{R}$, sufficiently smooth, $\bar{\mathrm{x}} \in \mathbb{R}^n, \mathrm{d} \in \mathbb{R}^n$. Consider a function around the point $\bar{\mathrm{x}}$ in direction d,

$$h(\alpha) = f(\bar{\mathrm{x}} + \alpha \mathrm{d})$$

1st order:

$$h(\alpha) = f(\bar{\mathrm{x}} + \alpha \mathrm{d}) = f(\bar{\mathrm{x}}) + (\nabla f^T(\mathrm{z}_1)\mathrm{d})\alpha$$

2nd order:

$$h(\alpha) = f(\bar{\mathrm{x}} + \alpha \mathrm{d}) = f(\bar{\mathrm{x}}) + (\nabla f^T(\bar{\mathrm{x}})\mathrm{d})\alpha + \frac{1}{2}(\mathrm{d}^T \nabla^2 f(\mathrm{z}_2)\mathrm{d})\alpha^2$$

Where z is on the line segment between $\bar{\mathrm{x}},\ \bar{\mathrm{x}} + \mathrm{d}$.

2.1 Descent direction

Let $f : \mathbb{R}^n \to \mathbb{R}$ c1, $\mathrm{x} \in \mathbb{R}^n$. d is a descent direction iff

$$\nabla f^T(\bar{\mathrm{x}})\mathrm{d} < 0$$

Note: steepest direction when $\nabla f(\bar{\mathrm{x}}) \neq 0$ is

$$\mathrm{d} = -\nabla f(\bar{\mathrm{x}})$$

Proof:

$$f(\bar{\mathrm{x}} + \alpha \mathrm{d}) = f(\bar{\mathrm{x}}) + [\nabla f^T(\mathrm{z}_1)\mathrm{d}]_{-}[\alpha]_{+} < f(\bar{\mathrm{x}}))$$

3 Extrema

Suppose $S \subseteq \mathbb{R}^n$ open, $\bar{\mathrm{x}} \in S$, $f : S \to \mathbb{R}$ ...

3.1 First order necessary condition

...and f is c1.

$$\bar{\mathrm{x}} \text{ is a local min} \implies \nabla f(\bar{\mathrm{x}}) = 0$$

Proof: By contradiction, if gradient is not zero, then there would be a descent direction.

3.2 Second order necessary condition

...and f is c2.

$$\bar{\mathrm{x}} \text{ is a local min} \implies \nabla^2 f(\bar{\mathrm{x}}) \text{ is PSD}$$

Proof: If hessian is not PSD, then $\exists$ $\mathrm{d} \in \mathbb{R}^n$ nonzero s.t. $\mathrm{d}^T \nabla^2 f(\bar{\mathrm{x}})\mathrm{d} < 0$.

$$f(x) = f(\bar{x}) + [\nabla f^T(\bar{\mathrm{x}})\mathrm{d}]_0 \alpha + \frac{1}{2}[\mathrm{d}^T \nabla^2 f(\bar{\mathrm{x}})\mathrm{d}]_- [\alpha^2]_+ \leq f(\bar{\mathrm{x}})$$

However, $\bar{\mathrm{x}}$ is a local min, so we have reached a contradiction.

3.3 Second order sufficient condition

...and f is c2.

$$\nabla f(\bar{\mathrm{x}}) = 0,\ \nabla^2 f(\bar{\mathrm{x}}) \text{ is PD} \implies \bar{\mathrm{x}} \text{ is a strict local min}$$

Proof:

$$f(x) = f(\bar{x}) + [\nabla f^T(\bar{\mathrm{x}})\mathrm{d}]_0 \alpha + \frac{1}{2}[\mathrm{d}^T \nabla^2 f(\mathrm{z})\mathrm{d}]_+ [\alpha^2]_+ > f(\bar{\mathrm{x}})$$

Note: Must be PD and not PSD, since we could have inflection/saddle points.

4 Convexity

Note: we don't really care about concavity, because f is convex iff $-f$ is concave, so we only need to think about one.

4.1 Convex set

$S \subseteq \mathbb{R}^n$ is convex set if

$$\forall \mathrm{x}, \mathrm{y} \in S,\ \lambda \in [0, 1],\ \lambda \mathrm{x} + (1 - \lambda)\mathrm{y} \in S$$

4.2 Convex function

$S \subseteq \mathbb{R}^n$ nonempty convex set, $f : S \to \mathbb{R}$.

- f is a convex function if
$$\forall \mathrm{x}, \mathrm{y} \in S,\ \lambda \in [0, 1],\ f(\lambda \mathrm{x} + (1 - \lambda)\mathrm{y}) \leq \lambda f(\mathrm{x}) + (1 - \lambda) f(\mathrm{y})$$
- f is a strictly convex function if
$$\forall \mathrm{x}, \mathrm{y} \in S,\ \mathrm{x} \neq \mathrm{y},\ \lambda \in [0, 1],\ f(\lambda \mathrm{x} + (1 - \lambda)\mathrm{y}) < \lambda f(\mathrm{x}) + (1 - \lambda) f(\mathrm{y})$$

4.2.1 Using first degree taylor

$S \subset \mathbb{R}^n$ open convex set, $f : S \to \mathbb{R}$ c1 function. f is...

- convex iff
$$\forall\ \mathrm{x}, \hat{\mathrm{x}} \in S,\ f(\mathrm{x}) \geq f(\hat{\mathrm{x}}) + \nabla f(\hat{\mathrm{x}})^T (\mathrm{x} - \hat{\mathrm{x}})$$
- stritly convex iff
$$\forall\ \mathrm{x}, \hat{\mathrm{x}} \in S,\ \mathrm{x} \neq \hat{\mathrm{x}},\ f(\mathrm{x}) > f(\hat{\mathrm{x}}) + \nabla f(\hat{\mathrm{x}})^T (\mathrm{x} - \hat{\mathrm{x}})$$

4.2.2 Using hessian

$S \subset \mathbb{R}^n$ open convex set, $f : S \to \mathbb{R}$ c2 function.

- f is convex iff $\nabla^2 f(\mathrm{x})$ is PSD $\forall$ x $\in S$.
  Proof:
  $\Longleftarrow$ $\forall$ $\mathrm{x}, \hat{\mathrm{x}} \in S$

$$f(\mathrm{x}) = f(\hat{\mathrm{x}}) + \nabla f(\hat{\mathrm{x}})^T(\mathrm{x} - \hat{\mathrm{x}}) + \left[\frac{1}{2}(\mathrm{x} - \hat{\mathrm{x}})^T \nabla^2 f(\mathrm{z})_{\geq 0} (\mathrm{x} - \hat{\mathrm{x}})_{\geq 0}\right]_{\geq 0}$$

$$f(\mathrm{x}) \geq f(\hat{\mathrm{x}}) + \nabla f(\hat{\mathrm{x}})^T(\mathrm{x} - \hat{\mathrm{x}})$$

  $\Longrightarrow$ converse

$$\text{not PSD} \implies \exists \ \mathrm{d} \in \mathbb{R}^n \text{ s.t. } \mathrm{d}^T \nabla^2 f(\mathrm{x}) \mathrm{d} < 0$$

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

$$f(\mathrm{x}) < f(\hat{\mathrm{x}}) + \nabla f(\hat{\mathrm{x}})^T(x - \hat{\mathrm{x}})$$

- if $\nabla^2 f$ is PD, then strictly convex (converse not true e.g. x^4 is strictly convex but has $\nabla^2 f(x)$ at $x = 0$).
  Proof:
  $\Longrightarrow$ $\forall$ $\mathrm{x}, \hat{\mathrm{x}} \in S$, $\mathrm{x} \neq \hat{\mathrm{x}}$

$$f(\mathrm{x}) = f(\hat{\mathrm{x}}) + \nabla^T f(\hat{\mathrm{x}})(\mathrm{x} - \hat{\mathrm{x}}) + \left[\frac{1}{2}(\mathrm{x} - \hat{\mathrm{x}})^T \nabla^2 f(\mathrm{z})_{>0} (\mathrm{x} - \hat{\mathrm{x}})_{>0}\right]_{>0}$$

4.2.3 Using epigraph

$S \subseteq \mathbb{R}^n$, $f : S \to \mathbb{R}$. Suppose S nonempty convex set and $f : S \to \mathbb{R}$.

$$f \text{ is convex iff } \mathit{epigraph}(f) = \left\{ \begin{bmatrix} \mathrm{x} \\ y \end{bmatrix} \mid \mathrm{x} \in S, y \in \mathbb{R}, y \geq f(\mathrm{x}) \right\} \text{ is a convex set}$$

4.3 Support theorem

$S \subseteq \mathbb{R}^n$ convex set, x is boundary point.

$$\exists \text{ hyperplane containing x s.t. } S \text{ is completely in an associated half space}$$

4.4 Global is local minimizer for convex function

Suppose $S \subseteq \mathbb{R}^n$ open convex set, $f : S \to \mathbb{R}$ convex, c1 function.

$$\hat{\mathrm{x}} \text{ is global minimizer} \iff \hat{\mathrm{x}} \text{ is local minimizer} \iff \nabla f(\hat{\mathrm{x}}) = 0$$

Proof:
Global min is a local min. Local min have gradient of 0, by first order necessary condition. If gradient is 0, then first order taylor is $f(\mathrm{x}) \geq f(\hat{\mathrm{x}}) + 0$ so it is a global minimizer.

Furthermore, if function is strictly convex, the global min is unique and strict.

5 Newton's method

Assume f is c2, $\hat{\mathrm{x}}$ is initial guess. Take second degree Taylor.

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

Derive wrt x, set to 0, then solve for x

$$\begin{aligned} g'(\mathrm{x}) = 0 &= \nabla f(\hat{\mathrm{x}}) + \nabla^2 f(\hat{\mathrm{x}})(\mathrm{x} - \hat{\mathrm{x}}) \\ \nabla^2 f(\hat{\mathrm{x}})(\mathrm{x} - \hat{\mathrm{x}}) &= -\nabla f(\hat{\mathrm{x}}) \\ \mathrm{x} - \hat{\mathrm{x}} &= -[\nabla^2 f(\hat{\mathrm{x}})]^{-1} \nabla f(\hat{\mathrm{x}}) \\ \mathrm{x} &= \hat{\mathrm{x}} - [\nabla^2 f(\hat{\mathrm{x}})]^{-1} \nabla f(\hat{\mathrm{x}}) \\ \mathrm{x} &= \hat{\mathrm{x}} - \mathrm{d} \end{aligned}$$

Assume f is convex, $\nabla^2 f(\hat{\mathrm{x}})$ is invertible, $\nabla f(\hat{\mathrm{x}}) \neq 0$. d is a descent direction.
Recall, a vector d is a descent direction if $\nabla f(\hat{\mathrm{x}})^T \mathrm{d} < 0$.

$$\nabla f(\hat{\mathrm{x}})^T \mathrm{d} = \left[-\nabla f(\hat{\mathrm{x}})^T [\nabla^2 f(\hat{\mathrm{x}})]^{-1}_{>0} \nabla f(\hat{\mathrm{x}})_{>0}\right]_{<0}$$

Thus, d is a descent direction.

5.1 Algorithm

$x_0 \in \mathbb{R}^n$
for $k = 1 \ldots$ until $\nabla f(x_k) = 0$
$\quad \mathrm{x}_{k+1} = \mathrm{x}_k - [\nabla^2 f(\hat{\mathrm{x}}_k)]^{-1} \nabla f(\mathrm{x}_k)$

6 Steepest descent

6.1 Algorithm

$\mathrm{x}_0 \in \mathbb{R}^n$
for $k = 1 \ldots$ until $\nabla f(x_k) = 0$
$\quad \alpha_* = \min_{\alpha > 0} f(\mathrm{x}_k - \alpha \nabla f(\mathrm{x}_k))$
$\quad \mathrm{x}_{k+1} = \mathrm{x}_k - \alpha_* \nabla f(\mathrm{x}_k)$

6.2 Note

This algorithm moves orthogonally. That is, $\nabla f(\mathrm{x}_k) \perp \nabla f(\mathrm{x}_{k+1})$.

$$g'(\alpha_*) = 0 = \nabla f(\mathrm{x}_k - \alpha \nabla f(\mathrm{x}_k))[-\nabla f(\mathrm{x}_k)] = \nabla f(\mathrm{x}_{k+1})^T \nabla f(\mathrm{x}_k)$$

7 Ferkas theorem

$\mathrm{A} \in \mathbb{R}^{m \times n},\ \mathrm{b} \in \mathbb{R}^n$. Either:

1. $\exists\, \mathrm{x} \in \mathbb{R}^n$ s.t. $\mathrm{Ax} = \mathrm{b},\ \mathrm{x} \geq 0$
2. $\exists\, \mathrm{y} \in \mathbb{R}^m$ s.t. $\mathrm{b}^T \mathrm{y} > 0,\ \mathrm{A}^T \mathrm{y} \leq 0$

Proof: Consider (LP0), (DP0), where c = 0

- If 1 is true,

$$\begin{aligned} \exists\, \mathrm{x} \text{ s.t. } \mathrm{Ax} = \mathrm{b},\ \mathrm{x} \geq 0 & \\ \text{(LP0) is feasible} & \quad \text{From construction} \\ \mathrm{b}^T \mathrm{y} \leq \mathrm{c}^T \mathrm{x} = 0 & \quad \text{By weak duality} \\ \mathrm{b}^T \mathrm{y} \not> 0 & \end{aligned}$$

  So 2 is false.

- If 2 is false,

$$\begin{gathered} \nexists\, \mathrm{y} \text{ s.t.} \mathrm{b}^T \mathrm{y} > 0,\ \mathrm{A}^T \mathrm{y} \leq 0 \\ \forall\, \mathrm{y} \text{ s.t. } \mathrm{b}^T \mathrm{y} \leq 0 \text{ or } \mathrm{A}^T \mathrm{y} > 0 \\ \mathrm{y} = 0 \text{ maximizes } \mathrm{b}^T \mathrm{y} \text{ Optimality} \\ \mathrm{c}^T \mathrm{x} = \mathrm{b}^T \mathrm{y} = 0 \text{ By strong duality} \\ \text{(LP0) is feasible} \end{gathered}$$

  So 1 is true.

8 Gordons theorem

$\mathrm{A} \in \mathbb{R}^{m \times n}$. Either:

1. $\exists$ x $\in \mathbb{R}^n$ s.t. $\mathrm{Ax} = 0, \mathrm{x} \geq 0, \mathrm{x}$ nonzero.
2. $\exists$ y $\in \mathbb{R}^m$ s.t. $\mathrm{A}^T\mathrm{y} < 0$

Proof:

- 2 is true iff

$$\exists\ \mathrm{y}, \varepsilon,\ \varepsilon > 0 \text{ s.t. } A^T\mathrm{y} + \varepsilon 1 \leq 0$$

iff (use linalg)

$$\exists \begin{bmatrix} \mathrm{y} \\ \varepsilon \end{bmatrix},\ \begin{bmatrix} 0 \\ 1 \end{bmatrix}^T \begin{bmatrix} \mathrm{y} \\ \varepsilon \end{bmatrix} > 0\ (\mathrm{b}'^T\mathrm{y}' > 0)$$
$$\begin{bmatrix} \mathrm{A}^T & 1 \end{bmatrix} \begin{bmatrix} \mathrm{y} \\ \varepsilon \end{bmatrix} \leq 0\ (\mathrm{A}'^T\mathrm{y}' \leq 0)$$

iff (use Farkas 2→1)

$$\nexists\ \mathrm{x} \in \mathbb{R}^n \text{ s.t. } \begin{bmatrix} \mathrm{A} \\ 1^T \end{bmatrix} \mathrm{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}\ (\mathrm{A}'^T\mathrm{y}' = \mathrm{b}')$$

iff

$$\mathrm{Ax} = 0$$
$$\sum x = 1$$
$$\mathrm{x} \geq 0$$

We know that $\sum x = 1$ iff x is nonzero. In the forward direction, obvious. In the reverse direction, we can scale x without affecting $\mathrm{Ax} = 0$. So 1 is true.

8.1 Active constraint

$$\mathcal{A}_{\mathrm{x}} = \{i \mid g_i(\mathrm{x}) = 0\}$$

Let $m = |\mathcal{A}_x|$

- **Obvious:** If x feasible in (P) is a local min of (P), then $\nexists$ d that is simultaneously a descent direction for f and for all active $g_i,\ i \in \mathcal{A}_{\mathrm{x}}$. (By contradiction, if there was a descent direction, would not be a local min).
- **LinAlg:** If x feasible in (P) is a local min of (P), then

$$\nexists \text{ d s.t. } J^T\mathrm{d} < 0$$
$$J^T = \begin{bmatrix} \nabla^T f(\mathrm{x}) \\ \dots \\ \nabla^T g_i(\mathrm{x}) \end{bmatrix},\ i \in \mathcal{A}_{\mathrm{x}}$$

- **Gordon's 2→1:** If x feasible in (P) is a local min of (P), then

$$\exists\ \lambda \geq 0 \text{ nonzero s.t. } J^T\lambda = 0$$

- **Inactive/CompSlack:** If x feasible in (P) and is a local min of (P) then $\exists \begin{bmatrix} \beta \\ \lambda \end{bmatrix} \geq 0$ s.t.

$$[\nabla f(\mathrm{x}), \dots \nabla g_i(\mathrm{x})] \begin{bmatrix} \beta \\ \lambda \end{bmatrix} = 0$$

$$\lambda_i * g_i(\mathrm{x}) = 0 \implies \sum \lambda_i g_i(\mathrm{x}) = 0 \implies \lambda \cdot g(\mathrm{x}) = 0$$

- **Fritz John**: If x feasible in (P) and is a local min of (P), then

$$\exists\ \lambda, \beta \text{ s.t.}$$
$$\beta\nabla f(\mathrm{x}) + \nabla g(\mathrm{x})\lambda = 0$$
$$\lambda^T g(\mathrm{x}) = 0$$
$$\lambda, \beta \geq 0$$

8.2 KKT

$$(P)\ \min f(\mathrm{x})$$

$$\text{s.t. } g(\mathrm{x}) \leq 0$$

If x feasible and local min of (P) and x satisfies a constraint qualification that $\{\nabla g_i(\mathrm{x}) \mid i \in \mathcal{A}_{\mathrm{x}}\}$ are linearly indepedent, then $\exists\ \lambda \in \mathbb{R}^m$ s.t.

$$\nabla f(\mathrm{x}) + \nabla g(\mathrm{x})\lambda = 0$$

$$\lambda \cdot g(\mathrm{x}) = 0$$

$$\lambda \geq 0$$

Find all KKT points

- Check when no active constraints (interior), aka all lambdas 0, aka $\nabla f(\mathrm{x}) = 0$.
- Check when one constraint active.
- Check when multiple constraints active.

Find and verify KKT point

- Guess point.
- Check which constraints are active, $g(x) = 0$ and say that the $\lambda = 0$ for inactive, so that $\lambda \cdot g(\mathrm{x}) = 0$
- Calculate ∇f and ∇g for active constraints. Check nonzero and linear independent, and write that it satisfies constraint qualification.
- Set $\nabla f(\mathrm{x}) + \lambda_i \nabla g_i(\mathrm{x}) = 0$, solve for λ_i.
- Verify that $\lambda_i > 0$.