11/17/16

① #4 posted, due Fri Dec 2.

② Vote on final exam.

Last class:

n-dim Gaussian dist For $x \in \mathbb{R}^n$

$$x \sim N_n(\mu, \Sigma)$$

① mean vector $\mu = (E[x_1], \dots, E[x_n])$

② $n \times n$ covariance matrix $\Sigma(i,j) = E[(x_i - \mu_i)(x_j - \mu_j)]$
↳ symmetric, PSD

③ For any $a \in \mathbb{R}^n$, R.V. $y = a^T x$ has univariate normal distribution.

Claim: For any n-dim dist over $\mathbb{R}^n$, $\exists$ n-dim Gaussian dist $\bar{y}$ same quadratic moments;
ie $\mathbb{E}[x_i x_j] = \mathbb{E}[y_i y_j]$.

Idea: Let $\Sigma = \sum_i \lambda_i v_i v_i^T$ be spectral decomp.
Set $y = \sum_k \sqrt{\lambda_k}\, w_k\, v_k$ for $w_k$ iid Gaussian (standard)
$w_k$.

$$y = \sum_k \sqrt{\lambda_k}\, w_k\, v_k$$

Claim: $\mathbb{E}[y_i y_j] = \mathbb{E}[x_i x_j]$ (a) and $y$ is $n$-dim Gaussian (b)

(a): $\mathbb{E}[y_i y_j] = \mathbb{E}\left[\left(\sum_k \sqrt{\lambda_k}\, w_k\, v_k(i)\right)\left(\sum_{k'} \lambda_{k'}\, w_{k'}\, v_{k'}(j)\right)\right]$

$$= \sum_k \sum_{k'} \sqrt{\lambda_k \lambda_{k'}}\, v_k(i)\, v_k(j)\, \mathbb{E}[w_k w_{k'}]$$

If $k = k' \Rightarrow \mathbb{E}[w_k^2] = 1$ (variance of std Gaussian)

If $k \neq k' \Rightarrow \mathbb{E}[w_k w_{k'}] = 0$ (indep. std Gaussian, mean 0 each)

$$\therefore = \sum_k \lambda_k\, v_k(i)\, v_k(j)$$

$$= \sum_k \lambda_k\, v_k v_k^T(i,j)$$

$$= \Sigma(i,j) \leftarrow \text{covariance matrix of } \{x\}$$

$$= \mathbb{E}[x_i x_j].$$

Pf of (b): Since $\Sigma \succeq 0$, evecs $\{v_i\}$ form o.n. basis for $\mathbb{R}^n$.

$\therefore$ For any $a \in \mathbb{R}^n$, $a = \sum_i \alpha_i v_i$ for $\{\alpha_i\} \subseteq \mathbb{R}$.

Recall $y = \sum_k \sqrt{\lambda_k}\, w_k\, v_k$.

$$\therefore \quad a^T y = \left(\sum_i \alpha_i v_i^T\right)\left(\sum_j \sqrt{\lambda_j} w_j v_j\right)$$

$$= \sum_i \alpha_i \sqrt{\lambda_i}\, w_i \qquad (v_k \text{ are o.n. basis})$$

($\alpha_i \sqrt{\lambda_i}$: $\mathbb{R}$; $w_i$: std Gaussian)

which is normal since sum of Gaussians is Gaussian.

Conclusion: Any dist over $\mathbb{R}^n$ is mapped to n-variate Gaussian dist w same 2nd moments.

Claim: Same trick applies to deg-2 pseudodistributions since "pseudo-covariance matrix" is also PSD!

# Algorithms for convex problems

## Chapter 9 (Boyd/Vandenberghe): Unconstrained minimization.

**Goal:** compute $p^* = \min f(x) \Leftarrow f(x): \mathbb{R}^n \longmapsto \mathbb{R}$ convex, twice continuously differentiable

Recall: If $f$ is differentiable, & convex, then $x^* \in \mathbb{R}^n$ is optimal iff $\nabla f(x^*) = 0$. (9.2)

**High-level idea to achieve goal:**

1. Choose "suitable" $x^{(0)}$ (initial pt) in $\mathbb{R}^n$.
2. Iteratively "improve" current pt until "close" to $p^*$.

Q: (1*) How to choose $x^{(0)}$?

(2*) How to design each iterative step?

(3*) How to bound # iterations needed to get "$\epsilon$-optimal" soln?

(1*) **Initial point: $x^{(0)}$**

- For our purposes, if $f$ is continuous & $\text{dom} f = \mathbb{R}^n$, any $x^{(0)} \in \mathbb{R}^n$ works.
- More generally, need:
  1. $x^{(0)} \in \text{dom} f$
  2. sublevel set $S = \{x \in \text{dom} f \mid f(x) \le f(x^{(0)})\}$ is closed.

(3*) Recall $f$ convex iff $\nabla^2 f(x) \succeq 0$.

Strongly convex: $\exists\, m > 0$ s.t. $\forall x \in S$ $\nabla^2 f(x) \succeq mI$. (9.7)

With this, we can strengthen first-order conditions for convexity:

Fact: For any $x, y \in S$,

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \underbrace{\frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)}_{*}$$

for some $z \in [x, y]$. (Taylor's thm)

If $f$ strongly convex, then $* \ge \frac{m}{2}\|y-x\|_2^2$.

$\therefore$ $f(y) \ge \underbrace{f(x) + \nabla f(x)^T (y-x)}_{\text{first-order conds.}} + \frac{m}{2}\|y-x\|_2^2$. (9.8)

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}\|y-x\|_2^2 \quad (9.8)$$

These strengthened 1st-order conditions allow us to bound $f(x) - p^*$:

Fix $x$. RHS of (9.8) is convex quadratic (unconstrained) fcn of $y$.

s.t. $\tilde{y} = x - \frac{1}{m}\nabla f(x)$ minimizes RHS.

Plug into (9.8): $f(y) \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2$. (9.9)

This works $\forall\, y \in S$, s.t. choose $y = y^* \leftarrow$ optimal pt:

$$p^* = f(y^*) \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2$$

$$\Rightarrow \text{if } \|\nabla f(x)\|_2^2 \leq \sqrt{2m\epsilon}, \text{ then } f(x) - p^* \leq \epsilon.$$

Q: We had $\nabla^2 f(x) \succeq mI$. We can also get $\nabla^2 f(x) \preceq MI$ for some $M \in \mathbb{R}$.

↳ (9.9) $\leq f(y) \leq f(x^{(0)})$. Since convex fcn unbounded unless $\nabla f(x) = 0\ \forall\, x$, this implies $S$ is bounded

$\Rightarrow \lambda_{max}(\nabla^2 f(x))$, which is continuous fcn of $x$ on $S$, ie bounded,

$\Rightarrow \exists\, M > 0$ s.t. $\nabla^2 f(x) \preceq MI\ \forall\, x \in S$. ∎

A similar argument via Taylor's thm yields:

$\forall x, y \in S$: $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|_2^2$

$\Rightarrow p^* \leq f(x) - \frac{1}{2M} \|\nabla f(x)\|_2^2$.

$\therefore \exists\ m, M$ s.t. $mI \leq \nabla^2(f(x)) \leq MI$.

↳ will be important for complexity analysis

↳ typically, $m$ & $M$ not known in practice

↳ geometric interpretation (roughly):

$\frac{M}{m}$ is upper bound on "anisotropy" of $S$, ie

small "anisotropy" $\Rightarrow$ $S$ roughly same "width" in all directions

large "anisotropy" $\Rightarrow$ $S$ has larger width in some directions.

(2*) How to design iteration step

Technique 1:

Alg 9.3 Gradient-descent

Input: starting pt $x \in \text{dom} f$.

Repeat

1. (Pick dir) set $\Delta x := -\nabla f(x)$.
2. (Pick step size) Line search:
   Choose step size $t$ via exact or backtracking search.
3. Update. set $x = x + t\Delta x$.

until "stopping criterion" satisfied

↳ typically $\|\nabla f(x)\|_2 \leq \eta$ for small $\eta > 0$.

Comments: ① How to choose $t$?

↳ we'll use: $t = \underset{s \geq 0}{\text{argmin}}\, f(x + s\Delta x)$ (exact line search)

② In order to guarantee $f(x^{(k+1)}) < f(x^{(k)})$, must have

$$\nabla f(x^{(k)})^T \Delta x^{(k)} < 0 \qquad \text{("descent method")}$$

$\uparrow$ (use first-order convexity conds.)

Note: $\Delta x^{(k)} = -\nabla f(x^{(k)})$ satisfies this.