262B-Lecture 10

Date created: 2021.02.18
N. of Pages: 12

Conjugate gradient method:

$$\min f(x) \longrightarrow \frac{1}{2} x^T Q x - b^T x$$

Algorithm: $x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}$ $\quad k = 0, 1, \ldots$

Instead of $d^{(k)} = -\nabla f(x^{(k)})$, we generate

$d^{(0)}, d^{(1)}, \ldots, d^{(n-1)}$ based on $-\nabla f(x^{(0)}), -\nabla f(x^{(1)})$, $\ldots, -\nabla f(x^{(n-1)})$ in such a way that the spanning rule on subspaces is preserved.

Formula:

$$\begin{cases} d^{(k)} = -\nabla f(x^{(k)}) + \sum_{i=0}^{k-1} \dfrac{(d^{(i)})^T Q \nabla f(x^{(k)})}{(d^{(i)})^T Q (d^{(i)})} d^{(i)} \\ d^{(0)} = -\nabla f(x^{(0)}) \end{cases}$$

Also, since $\alpha^{(k)}$ is optimal: $\boxed{(d^{(i)})^T \nabla f(x^{(k+1)}) = 0 \quad i = 0, \ldots, k}$

Also, span of $d^{(0)}, \ldots, d^{(k)}$

$=$ span of $\nabla f(x^{(0)}), \ldots, \nabla f(x^{(k)})$

$\Rightarrow$ $\nabla f(x^{(k+1)})$ is orthogonal to $\nabla f(x^{(0)}), \ldots, \nabla f(x^{(k)})$

Simplify:

$$\begin{cases} x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)} \\ d^{(k)} = -\nabla f(x^{(k)}) + \dfrac{\nabla f(x^{(k)})^T \nabla f(x^{(k)})}{\nabla f(x^{(k-1)})^T \nabla f(x^{(k-1)})} d^{(k-1)} \\ d^{(0)} = -\nabla f(x^{(0)}) \end{cases}$$

And $\alpha^{(k)}$: $\min_{\alpha} f(x^{(k)} + \alpha d^{(k)})$

Thm: If $f(x)$ is quadratic, then conjugate gradient stops after at most $\underline{n}$ steps.

→ Finite-time convergence

Note: Algorithm can be used for arbitrary $f(x)$,

but convergence is not finite.

$\min \frac{1}{2} x^T Q x + b^T x$

Conjugate gradient = Quasi-newton with exact line search & $\underline{D^{(0)} = I}$

(true for arbitrary $f(x)$)

Convergence rate: $\min \frac{1}{2} x^T \underbrace{Q}_{\succ 0} x$ (no b)

$$\Rightarrow \begin{cases} x^{(1)} \in x^{(0)} + \text{span}\{\nabla f(x^{(0)})\} \\ x^{(2)} \in x^{(0)} + \text{span}\{\underbrace{\nabla f(x^{(0)})}_{Qx^{(0)}}, \underbrace{\nabla f(x^{(1)})}_{Qx^{(1)}}\} \end{cases}$$

$$\Rightarrow x^{(2)} \in x^{(0)} + \text{span}\{Qx^{(0)}, Q^2x^{(0)}\}$$

$$\vdots$$

$$x^{(k+1)} \in x^{(0)} + \underbrace{\text{span}\{Qx^{(0)}, Q^2x^{(0)}, \dots, Q^{k+1}x^{(0)}\}}$$

$$x^{(0)} + Q\left(\sum_{j=0}^{k} \beta_j Q^j\right) x^{(0)}$$

$P^k(Q)$: polynomial of degree $\leq k$

$$\Rightarrow \boxed{x^{(k+1)} = \left(I + Q P^k(Q)\right) x^{(0)}}$$

previously: coefficients are optimal

$$\left(x^{(k+1)} = \underset{x \in \mathcal{M}^{(k)}}{\arg\min} f(x) \right)$$

$$\Rightarrow \left(f(x^{(k+1)}) = \min_{P^k(\cdot)} \frac{1}{2} (x^{(0)})^T Q \left(I + Q P^k(Q) \right)^2 \right) (x^{(0)})$$

(*) $\quad - \frac{1}{2} x^{(k+1)T} Q x^{(k+1)}$

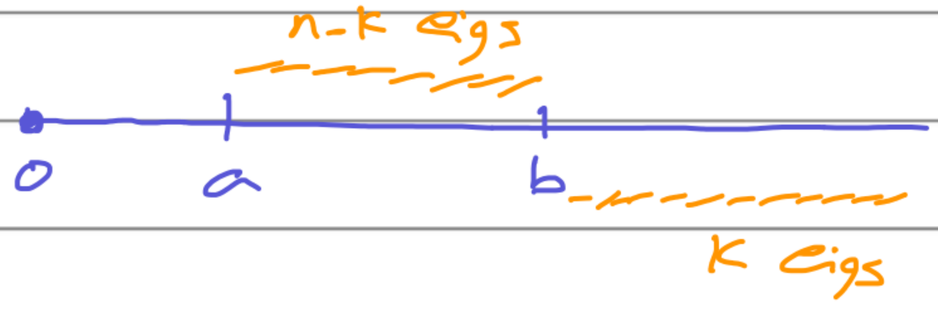
If we denote eigenvalues of Q with $\lambda_1, \dots, \lambda_n$ and eigenvectors as $v_1, \dots, v_n$, then

$$x^{(0)} = \underbrace{\mu_1}_{\in \mathbb{R}} v_1 + \underbrace{\mu_2}_{\in \mathbb{R}} v_2 + \dots + \underbrace{\mu_n}_{\in \mathbb{R}} v_n$$

$$(*) \Rightarrow f(x^{(k+1)}) \leq \max_{1 \leq i \leq n} \left(1 + \lambda_i P^k(\lambda_i) \right)^2 f(x^{(0)})$$

$$\forall\, k, \quad \forall\, P^k(\cdot)$$

Thm: Assume Q has $n-k$ eigenvalues in the interval $[a, b]$ where $a > 0$ and k eigenvalues in the interval (b, ∞).



After $k+1$ steps of conjugate gradients:

$$f(x^{(k+1)}) \leq \left(\frac{b-a}{b+a}\right)^2 f(x^{(0)})$$

Compare it with gradient method:
(with optimal stepsize)

$$f(x^{(k+1)}) \leq \underbrace{\left(\frac{c.d.(Q)-1}{c.d.(Q)+1}\right)^2}_{\frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}}} f(x^{(k)})$$

So, after $k+1$ iterations of conjugate gradient the k largest eigs drop and condition number changes from $\frac{\lambda_{max}}{\lambda_{min}}$ to $\frac{b}{a}$.

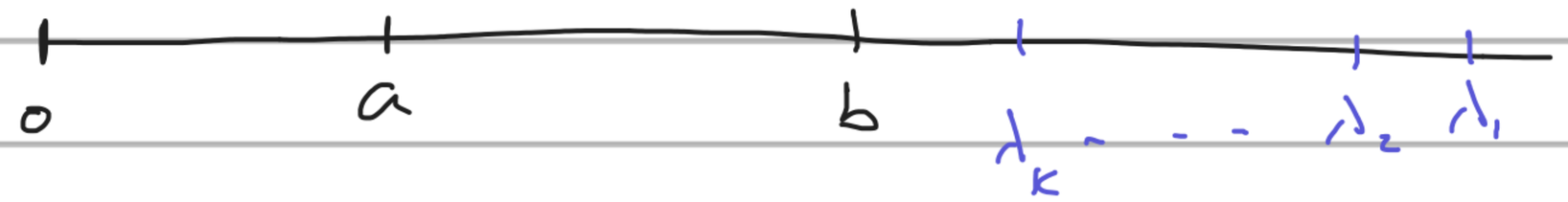
Ex: $\underbrace{\times\times\times\times}_{\text{around } 1}$ $\underbrace{\times}_{10^4}$ $\Rightarrow$ c.d. = huge

$\downarrow$

gradient alg = slow

If you use conjugate gradient $\Rightarrow$ c.d. after on iteration $\simeq 1$

proof:



pick $p^k(\cdot)$:

$$1 + \lambda p^k(\lambda) = \frac{2}{(a+b)\lambda_1 \cdots \lambda_k} \left(\frac{a+b}{2} - \lambda\right)(\lambda_1 - \lambda) \cdots (\lambda_k - \lambda)$$

degree $= k+1$

at $\lambda = 0$: value $= 1$

satisfies both properties

① : $1 + \lambda_i p^k(\lambda_i) = 0 \qquad i = 1, \ldots, k$

② : $\lambda \le b \Rightarrow \frac{\lambda_i - \lambda}{\lambda_i} \le 1 \qquad i = 1, \ldots, 1$

$$\Rightarrow f(x^{(k+1)}) \le \max_{1 \le i \le n} \left(1 + \lambda_i p^k(\lambda_i)\right)^2 f(x^{(0)})$$

① , ② $\Rightarrow$ $$f(x^{(k+1)}) \le \max_{a \le \lambda \le b} \left(\frac{\lambda - \frac{a+b}{2}}{\frac{a+b}{2}}\right)^2 f(x^{(0)})$$

$$= \left(\frac{b-a}{b+a}\right)^2 f(x^{(0)})$$

Homework : If Q is low rank ($\text{rank} = m$)

$\Rightarrow$ Conjugate gradient solves the problem

in at most $\underline{m}$ steps.

If Q has clustered eigs , $Q \succ 0$

$\lambda_1, \dots, \lambda_n \longrightarrow$ ● (repeated t_1 times) ● (repeated t_2 times) ⋯ ● (repeated t_m times)

$\longrightarrow$ m : number of clusters

Then, Conjugate gradient solves the problem

in at most $m+1$ steps.

Newton's method but use Conjugate gradient

to find Newton's step :

$x^{(k+1)} = x^{(k)} + \alpha^{(k)} \Delta x^{(k)}$: $\nabla f(x^{(k)})^2 \Delta x^{(k)} = -\nabla f(x^{(k)})$

Then instead of complexity of $O(n^3)$,
it could be much lower if $\nabla^2 f(x_*)$
is structured and $x^{(0)}$ is close to x_*.

Ex: $\nabla^2 f(x_*)$: ① ill-conditioned → conjugate gradient gets rid of bad eigs

② almost low-rank

③ eigs are almost clustered.

⋮

Coordinate descent :

Goal : Low-complexity algorithm

Here, instead of iterating over n entries,
what if we iterate over a single entry ?

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)})$$

↓ instead, pick entry $i \in \{1, \dots, n\}$

① $x^{(k+1)} = x^{(k)} - \alpha^{(k)} \dfrac{\partial f(x^{(k)})}{\partial x_i} e_i$

$e_i = [0 \ \dots \ 0 \ 1 \ 0 \ \dots \ 0]$ ← i^{th} coordinate

or

② $x^{(k+1)} = \underset{x_i \in \mathbb{R}}{\arg\min} \ f(\underbrace{x_1^{(k)}, \dots, x_{i-1}^{(k)}}_{\text{Known}}, \underbrace{x_i}_{\text{variable}}, \underbrace{x_{i+1}^{(k)}, \dots, x_n^{(k)}}_{\text{Known}})$

strategy ①: one improvement step for entry i

strategy ②: univariate opt $\Rightarrow$ low-complexity

later on, we will study convergence in a general setting subject to constraints.

Machine learning application:

$\min \ g(x) + \underbrace{|x|_1}_{\text{regularizer}}$ ⟶ $\min_{x_i} \ g(x_i) + |x_i|$

non-smooth

1-d non-smooth ⟶ closed-form

Convergence rate for constant stepsize:

Assumptions:

- $\left| \frac{\partial f(x+te_i)}{\partial x_i} - \frac{\partial f(x)}{\partial x_i} \right| \leq L|t|$

  $\forall x \in \mathbb{R}^n, \forall t \in \mathbb{R}$

  $\forall i \in \{1, \dots n\}$

Lipschitz continuity of ~~gradient~~ partial derivatives

- $\nabla^2 f(x) \succeq \underbrace{m}_{>0} I \quad \forall x$

  similar result holds for non-convex case

- pick stepsize $\alpha^{(k)} = \alpha < \frac{1}{L}$

Alg: $x^{(k+1)} = x^{(k)} - \alpha \frac{\partial f(x^{(k)})}{\partial x_i} e_i \rightarrow y$

$f(x^{(k+1)}) \leq f(x^{(k)})$

$+ \nabla f(x^{(k)})^T \left(-\alpha \frac{\partial f(x^{(k)})}{\partial x_i} \right) e_i$

$+ \frac{L}{2} \left\| -\alpha \frac{\partial f(x^{(k)})}{\partial x_i} \right\|^2$ (*)

Thm:

$f(x+y) \leq f(x) + \nabla f(x)^T y + \frac{L}{2}\|y\|^2 \rightarrow$ redo for partial derivatives

$$(*) \Rightarrow f(x^{(k+1)}) \leq f(x^{(k)}) - \left(\frac{\partial f(x^{(k)})}{\partial x_i}\right)^2 \times \alpha\left(1 - \frac{\alpha L}{2}\right) \quad (**)$$

How to pick $\underline{i}$:

1 - cyclic : $1, 2, \ldots, n, 1, 2, \ldots, n, 1, 2, \ldots, n, \ldots$

2 - random : pick a coordinate in a uniform way

call them $i_1, i_2, \ldots, i_k, \ldots$

Let's analyze case $\underline{2}$:

Take expected value with respect to i_k from $(**)$:

$$\Rightarrow \mathop{\mathbb{E}}_{i_k}\left(f(x^{(k+1)}) - f_*\right) \leq \left(f(x^{(k)}) - f_*\right)$$

$$- \alpha\left(1 - \frac{\alpha L}{2}\right) \times \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial f(x^{(k)})}{\partial x_i}\right)^2$$

$$2m\left(f(x^{(k)}) - f_*\right) \leq \left(\|\nabla f(x^{(k)})\|^2\right)$$

$$\Rightarrow \underset{i_k}{\mathbb{E}} \left(f(x^{(k+1)}) - f_* \right) \leq$$

$$\left(f(x^{(k)}) - f_* \right) \times \left(1 - \frac{m}{n} \alpha (2 - \alpha L) \right)$$

Take expected value with respect to $i_0, \ldots, i_{k-1}$

$$\Rightarrow \left(\phi_{k+1} - f_* \right) \leq \left(\phi_k - f_* \right) \left(1 - \frac{m}{n} \alpha (2 - \alpha L) \right)$$

where $\phi_k = \underset{\{i_0, i_1, \ldots, i_{k-1}\}}{\mathbb{E}} \left(f(x^{(k)}) \right)$

if α : small then this term is in $(0, 1)$

$\Rightarrow$ Linear convergence on $\phi_k - f_*$

$\Rightarrow$ Linear convergence on expected value of $f(x)$.