计算方法

刘景铖

计算机软件新技术国家重点实验室
南京大学

回顾

- 矩阵多项式的特征空间：$\boldsymbol{p(A)\,v = p(\lambda)v}$
  - $A \to p(A)$
  - 特征空间$\{\lambda_1, \lambda_2, \cdots, \lambda_n\} \to \{p(\lambda_1), p(\lambda_2), \cdots, p(\lambda_n)\}$
- **(Cayley-Hamilton)存在n次多项式$\boldsymbol{p(A)}$ 使得 $\boldsymbol{p(A)A = I}$**
  - 记$\boldsymbol{q(x) = 1 - xp(x)}$
  - 则$\boldsymbol{n}$次多项式$q_A(x) = \det(\boldsymbol{A - xI})$满足$\boldsymbol{q(A) = 0}$
  - 非平凡的：注意$\det(\boldsymbol{A - xI})$仅仅为一个数

Cayley-Hamilton

考虑关于$\boldsymbol{x}$的$\boldsymbol{n}$次多项式$q_A(x) = \det(\boldsymbol{A} - \boldsymbol{xI})$

定理**(Cayley-Hamilton):** $\boldsymbol{q_A(A) = 0}$为全零矩阵

注意$q_A(0) = \det(\boldsymbol{A}) \neq \boldsymbol{0}$

$$\boldsymbol{A^{-1}} = \frac{(-\boldsymbol{1})^{\boldsymbol{n-1}}}{\det(\boldsymbol{A})}(\boldsymbol{A^{n-1}} + \boldsymbol{c_{n-1}A^{n-2}} + \cdots + \boldsymbol{c_1 I})$$

$$\boldsymbol{A^{-1}} = \boldsymbol{p(A)}$$

如何计算 $\boldsymbol{q(x)}$? 高斯消元?

Richardson iteration

一个特殊情形: 给定实数$0 < a < 1$, 如何用 a 表示出$a^{-1}b$?

$$b + (1-a)b + (1-a)^2 b + \cdots = \frac{b}{1-(1-a)} = a^{-1}b$$

等价于$\boldsymbol{x_0 = 0, x_{k+1} = (1-a)x_k + b}$

Richardson iteration（对于对称正定的矩阵$\boldsymbol{A}$）:

$$\boldsymbol{x_0 = 0}$$

$$\boldsymbol{\overrightarrow{x_{k+1}} = (I - \alpha A)\overrightarrow{x_k} + \alpha \vec{b} = \overrightarrow{x_k} - \alpha(A\overrightarrow{x_k} - \vec{b})}$$

Richardson iteration

一个特殊情形: 给定实数$0 < a < 1$, 如何用 a 表示出$a^{-1}b$?

$$b + (1-a)b + (1-a)^2 b + \cdots = \frac{b}{1-(1-a)} = a^{-1}b$$

等价于$\boldsymbol{x_0 = 0}$, $\boldsymbol{x_{k+1} = (1-a)x_k + b}$

Richardson iteration:

$$\boldsymbol{x_0 = 0}$$

$$\boldsymbol{\overrightarrow{x_{k+1}} = (I - \alpha A)\overrightarrow{x_k} + \alpha \vec{b} = \overrightarrow{x_k} - \alpha(A\overrightarrow{x_k} - \vec{b})}$$

矩阵多项式的视角：特征空间：$\boldsymbol{p(A)\, v = p(\lambda) v}$

- $A \to p(A)$
- 特征空间$\{\lambda_1, \lambda_2, \cdots, \lambda_n\} \to \{p(\lambda_1), p(\lambda_2), \cdots, p(\lambda_n)\}$

另一个视角: 对于对称的 A, 这正好是目标函数$\frac{1}{2}\boldsymbol{x^T A x - b^T x}$的梯度下降方法!

$$\boldsymbol{\nabla\left(\frac{1}{2}x^T A x - b^T x\right) = \frac{1}{2}\left(A + A^T\right)x - b}$$

An Optimization Perspective

给定连续可导f: $\mathbb{R}^n \to \mathbb{R}$, 最小化$f$

例子:

最小二乘法: $f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$

向量投影: $f(\boldsymbol{c}) = \left\|\boldsymbol{c}\,\vec{\boldsymbol{a}} - \vec{\boldsymbol{b}}\right\|_2^2$

求伪逆(Pseudoinverse): $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2$, subject to $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$

对称矩阵的特征值: $f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\,\boldsymbol{x}$, subject to $x^T x = 1$

线性规划: $f(\boldsymbol{x}) = \boldsymbol{c}^T\boldsymbol{x}$, subject to $\boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}$

主成分分析 (Principal component analysis): $f(\boldsymbol{C}) = \left\|\boldsymbol{X} - \boldsymbol{C}\boldsymbol{C}^T\boldsymbol{X}\right\|_{\boldsymbol{F}}$, subject to $C\,C^T = I_{d\times d}$

离散组合优化: 最小生成树，最小割、最大流（网络流），最短路径，最大匹配；最大独立集、团、支配集，最小覆盖，背包问题，最大割，max-SAT，旅行商问题，最长路径（哈密顿路径）……

Optimality notion

给定连续可导f: $\mathbb{R}^n \to \mathbb{R}$, 最小化$f$

x_***是一个全局最优解(Global minimum)**，如果$f(x) \geq f(x_*), \forall x$

x_***是一个局部最优解(Local minimum)**，如果$\exists \delta > 0, \forall x: |x - x_*| < \delta$，均有$f(x) \geq f(x_*)$

Optimality condition from calculus

给定连续可导$f\colon \mathbb{R}^n \to \mathbb{R}$, 和一个点 x_0

$$\nabla f(x) \coloneqq \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n}\right)$$

$$f(x) \approx f(x_0) + \nabla f(x_0) \cdot (x - x_0), \qquad \forall x$$

特别地，可以选取$x - x_0 = \alpha\left(\nabla f(x_0)\right)^T$

$$f\left(x_0 + \alpha\left(\nabla f(x_0)\right)^T\right) - f(x_0) \approx \alpha \|\nabla f(x_0)\|_2^2$$

当α足够小的时候，α的符号决定f局部的增减性

当 $\nabla f(x_0) = 0$，则点 x_0为驻点(stationary point)

此时需要看二阶导数

Optimality condition from calculus

给定连续可导f: $\mathbb{R}^n \to \mathbb{R}$

$$\mathbf{H}_f = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

通常也把Hessian矩阵H_f记作 $\nabla^2 f$

Optimality condition from calculus

给定连续可导f: $\mathbb{R}^n \to \mathbb{R}, \forall x$

$$f(x) \approx f(x_0) + \nabla f(x_0) \cdot (x - x_0) + \frac{1}{2}(x - x_0)^{\mathrm{T}} \; H_f(x_0) \, (x - x_0)$$

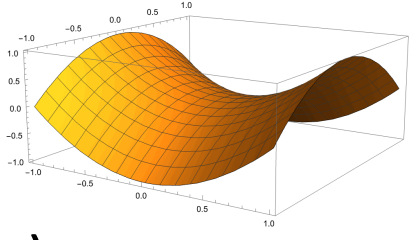
如果 $\nabla f(x_0) = 0$

$H_f(x_0) \succ 0$: 局部最小

$H_f(x_0) \prec 0$: 局部最大

$H_f(x_0)$同时有正的和负的特征值：saddle point

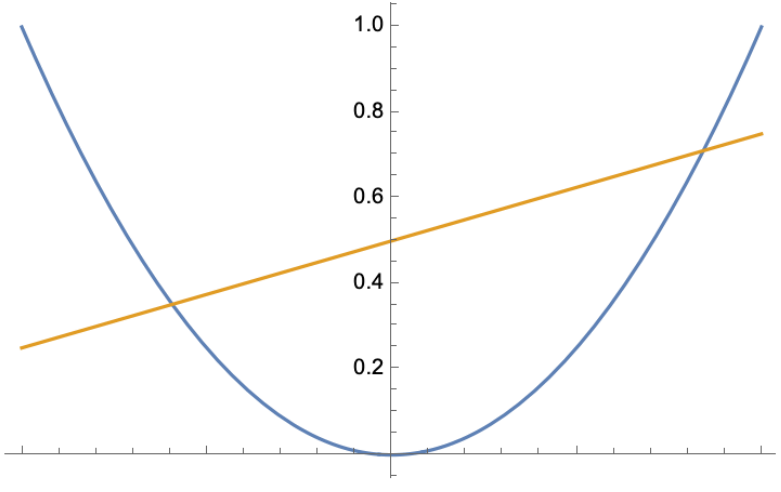
$H_f(x_0)$不可逆：不一定是局部极值 (Morse theory)

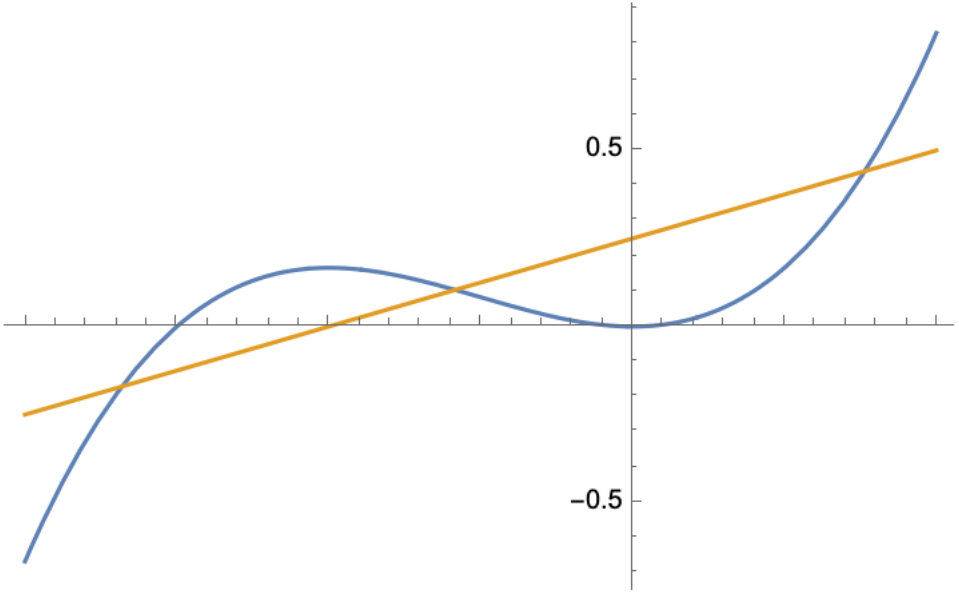


Convexity

如果给定的不仅仅是一个点上的导数，而是一个邻域上导数的信息，那么判断标准可以有所简化。

例子：凸函数(convex function)





Convexity

如果给定的不仅仅是一个点上的导数，而是一个邻域上导数的信息，那么判断标准可以有所简化。

例子：凸函数(convex function)的等价刻画

- Jensen's inequality:

$$\forall t \in [0,1], \forall x, y \in \mathbb{R}^n, f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

- 一阶条件（对可导的f）：

$$\forall x, y \in \mathbb{R}^n, f(x) \geq f(y) + \nabla f(y) \cdot (x - y)$$

- 二阶条件（对二阶可导的f）：

$$\forall x \in \mathbb{R}^n, H_f(x) \succcurlyeq 0$$

Steepest descent

给定连续可导f: $\mathbb{R}^n \to \mathbb{R}$, 最小化$f$

使用函数求根的方法，找极值点

- ∇f: $\mathbb{R}^n \to \mathbb{R}^n$，需要找到$\nabla f = 0$

梯度下降方法:

- 在当前的点，找到“最速下降”的方向 (steepest descent)
- $f(x) \approx f(x_0) + \nabla f(x_0) \cdot (x - x_0)$
- 假设$f(x) = f(x_0) + \nabla f(x_0) \cdot (x - x_0)$
- “最速下降”的方向是什么？
  - 找到单位长度的向量y，使得 $\nabla f(x_0) \cdot y$ 长度最大化

Steepest descent

给定向量$\vec{a}$，求单位长度的向量$\vec{b}$，使得$|\langle \vec{a}, \vec{b} \rangle|$最大化。

Cauchy-Schwarz不等式：

$$|\langle \vec{a}, \vec{b} \rangle| \leq \|\vec{a}\|_2 \|\vec{b}\|_2,$$

当且仅当存在乘数(scalar)使得$\vec{a} = c\,\vec{b}$时，等号可以取到

- 假设$f(x) = f(x_0) + \nabla f(x_0) \cdot (x - x_0)$
- “最速下降”的方向是什么？
  - 找到单位长度的向量y，使得$\nabla f(x_0) \cdot y$长度最大化
  - $y = \alpha \nabla f(x_0)$

回到Richardson iteration

一个特殊情形: 给定实数$0 < a < 1$, 如何用 a 表示出$a^{-1}b$?

$$b + (1-a)b + (1-a)^2 b + \cdots = \frac{b}{1-(1-a)} = a^{-1}b$$

等价于$\boldsymbol{x_0 = 0, x_{k+1} = (1-a)x_k + b}$

Richardson iteration:

$$\boldsymbol{x_0 = 0}$$

$$\boldsymbol{\overrightarrow{x_{k+1}} = (I - \alpha A)\overrightarrow{x_k} + \alpha \vec{b} = \overrightarrow{x_k} - \alpha(A\overrightarrow{x_k} - \vec{b})}$$

矩阵多项式的视角：特征空间：$\boldsymbol{p(A)\, v = p(\lambda) v}$

- $A \to p(A)$
- 特征空间$\{\lambda_1, \lambda_2, \cdots, \lambda_n\} \to \{p(\lambda_1), p(\lambda_2), \cdots, p(\lambda_n)\}$

另一个视角: 对于对称的 A, 这正好是目标函数$\frac{1}{2}\boldsymbol{x^T A x - b^T x}$的梯度下降方法!

$$\boldsymbol{\nabla\left(\frac{1}{2}x^T A x - b^T x\right) = \frac{1}{2}\left(A + A^T\right)x - b = A\overrightarrow{x_k} - \vec{b}}$$

Richardson iteration

$$A^{-1} = p(A)$$

可否近似$p(A)$?
等价地，记$q(x) = 1 - xp(x)$；需要寻找$q(0) = 1$，$q(x) \approx 0, \forall x > 0$

要解$Ax = b$，即找出$x_* = p(A)b \in span\{b, Ab, A^2b, ..\}$

Richardson iteration: 设$x_0 = 0$,

$$x_{k+1} = (I - \alpha A)x_k + \alpha\, b$$

误差$e_k = x_k - x_*$满足$e_k = (I - \alpha A)e_{k-1} = (I - \alpha A)^k e_0$
收敛性：当且仅当$\rho(I - \alpha A) = \max\{|1 - \alpha\lambda_1|, |1 - \alpha\lambda_n|\} < 1$
可以令$1 - \alpha\lambda_1 = -(1 - \alpha\lambda_n)$即$\alpha = \frac{2}{\lambda_1 + \lambda_n}$
此时$\rho(I - \alpha A) = \frac{\lambda_n - \lambda_1}{\lambda_1 + \lambda_n}$，收敛需要$O\left(\left(1 + \frac{\lambda_n}{\lambda_1}\right) \cdot \log\frac{1}{\epsilon}\right)$步
回顾：$\frac{\lambda_n}{\lambda_1}$正是条件数！

Richardson iteration

$$A^{-1} = p(A)$$

可否近似$p(A)$?
等价地，记$q(x) = 1 - xp(x)$；需要寻找$q(0) = 1$，$q(x) \approx 0, \forall x > 0$

要解$Ax = b$，即找出$x_* = p(A)b \in span\{b, Ab, A^2b, ..\}$

Richardson iteration: 设$x_0 = 0$,

$$x_{k+1} = (I - \alpha A)x_k + \alpha\, b$$

注意，这里的$x_k = p_k(A)b \in span\{b, Ab, A^2b, \dots, A^{k-1}b\}$；**Krylov子空间**

可以证明: 相当于考虑$q(x) = 1 - xp(x) = (1 - \alpha x)^k$

$$-Ae_k = b - Ax_k = \big(I - Ap_k(A)\big)b = q_k(A)b$$
$$-e_k = x_* - x_k = (I - p_k(A)A)x_* = q_k(A)x_*$$

Chebyshev iteration（选讲）

$$A^{-1} = p(A)$$

目标：近似$p(A)$

等价地，记$q(x) = 1 - xp(x)$；需要寻找$q(0) = 1$，$q(x) \approx 0, \forall x > 0$

要解$Ax = b$，即找出$x_* = p(A)b \in span\{b, Ab, A^2b, ..\}$

Chebyshev iteration:

$$x_{k+1} = (I - \alpha_k A)x_k + \alpha_k b$$

误差$e_k = x_k - x_*$满足$e_k = \prod_i (I - \alpha_i A)\, e_0$

$$\|e_k\| \le \left\| \prod_i (I - \alpha_i A) \right\| \|e_0\|$$

要最小化$\|\prod_i (I - \alpha_i A)\|$: **Chebyshev**多项式！

注意，这里的$x_k \in span\{b, Ab, A^2b, ..\}$

回到Richardson iteration: Steepest descent

Richardson iteration:

$$x_0 = 0$$

$$\overrightarrow{x_{k+1}} = (I - \alpha A)\overrightarrow{x_k} + \alpha \vec{b} = \overrightarrow{x_k} - \alpha(A\overrightarrow{x_k} - \vec{b})$$

对于对称的 A, 这正好是目标函数$\frac{1}{2}x^T Ax - b^T x$的梯度下降方法!

$$\nabla\left(\frac{1}{2}x^T Ax - b^T x\right) = \frac{1}{2}\left(A + A^T\right)x - b = A\overrightarrow{x_k} - \vec{b}$$

目标函数$\frac{1}{2}x^T Ax - b^T x$是convex的：$A$是正定的

利用strong convexity，也可以证明梯度下降的收敛时间正比于条件数$\frac{\lambda_n}{\lambda_1}$

参考： Convex Optimization by Stephen Boyd and Lieven Vandenberghe

Conjugate gradients

给定正定矩阵$\boldsymbol{A}$，可定义内积：

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{A}} := \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{A} \boldsymbol{y}$$

- 线性
- 对称性
- 正定性$\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{A}} > \boldsymbol{0}\ \forall \boldsymbol{x} \neq \boldsymbol{0}$

如果$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{A}} = \boldsymbol{0}$则称$\boldsymbol{x}, \boldsymbol{y}$关于$\boldsymbol{A}$共轭

内积自带的范数：$\|\boldsymbol{x}\|_{\boldsymbol{A}}^{\boldsymbol{2}} := \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{A} \boldsymbol{x}$

令$\boldsymbol{x}_*$满足$\boldsymbol{A}\boldsymbol{x}_* = \boldsymbol{b}$，考虑关于$\boldsymbol{x}$的函数：

$$\frac{\boldsymbol{1}}{\boldsymbol{2}} \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{A} \boldsymbol{x} - \boldsymbol{b}^{\boldsymbol{T}} \boldsymbol{x} = \frac{\boldsymbol{1}}{\boldsymbol{2}} \|\boldsymbol{x} - \boldsymbol{x}_*\|_{\boldsymbol{A}}^{\boldsymbol{2}} + \boldsymbol{constant}$$

Conjugate gradients

给定正定矩阵A，可定义内积：

$$\langle x, y\rangle_A := x^T A y$$

性质：共轭向量一定是线性无关的。因此可以组成一组基。

假设$p_1, p_2, \ldots, p_k$互相共轭，并且是线性相关的：存在$\alpha_1, \ldots, \alpha_k$使得：

$$\alpha_1 p_1 + \cdots + \alpha_k p_k = 0$$

另一方面

$$(\alpha_1 p_1 + \cdots + \alpha_k p_k)^T A(\alpha_1 p_1 + \cdots + \alpha_k p_k)$$
$$= \alpha_1^2 \|p_1\|_A^2 + \alpha_2^2 \|p_2\|_A^2 + \cdots + \alpha_k^2 \|p_k\|_A^2 > 0$$

与上式矛盾

Conjugate gradients

$$A^{-1} = p(A)$$

目标：近似$p(A)$

等价地，记$q(x) = 1 - xp(x)$；需要寻找$q(0) = 1$，$q(x) \approx 0, \forall x > 0$

要解$Ax = b$，即找出$x_* = p(A)b \in span\{b, Ab, A^2b, ..\}$

考虑所有这样的迭代算法：

记Krylov子空间$K_0 = \{0\}, K_i = span\{b, Ab, ... A^{i-1}b\}$

设$x_i = arg\ min_{x \in K_i} \|x - x_*\|_A^2$，其中$x_*$满足$Ax_* = b$

引理：记$v_i := x_i - x_{i-1}$，则$\{v_i\}$两两共轭：$v_i^T A v_j = 0, \forall i \neq j$

Conjugate gradients

记Krylov子空间$K_0 = \{0\}, K_i = span\{b, Ab, \dots A^{i-1}b\}$

设$x_i = arg\ min_{x \in K_i} \|x - x_*\|_A^2$，其中$x_*$满足$Ax_* = b$

引理：记$v_i := x_i - x_{i-1}$，则$\{v_i\}$两两共轭：$v_i^{\mathrm{T}} A v_j = 0, \forall i \neq j$

证明：假设$i < j$. 注意到x_j的定义是K_j中最小化$\|x - x_*\|_A^2$的，所以$\nabla \frac{1}{2}\|x_j - x_*\|_A^2 = Ax_j - b$必须与$K_j$正交，因此与$K_{j-1}$正交。类似地$Ax_{j-1} - b$必须与$K_{j-1}$正交。因此，

$$Av_j = Ax_j - Ax_{j-1} \in K_{j-1}^{\perp}$$

而$v_i \in K_i \subset K_{j-1}$, 因此$v_i^T A v_j = 0$

一个推论：$K_i = span\{v_1, v_2, \dots, v_i\}$

Conjugate gradients

记Krylov子空间$K_0 = \{0\}, K_i = span\{b, Ab, \dots A^{i-1}b\}$

设$x_i = arg\ min_{x \in K_i} \|x - x_*\|_A^2$，其中$x_*$满足$Ax_* = b$

引理：记$v_i := x_i - x_{i-1}, r_i := b - Ax_i$，则

$$v_i = \frac{v_i^T r_{i-1}}{\|r_{i-1}\|^2}(r_{i-1} - \frac{r_{i-1}^T A v_{i-1}}{v_{i-1}^T A v_{i-1}} v_{i-1})$$

证明(sketch): 首先注意到$Ax_{i-1} - b \in K_{i-1}^{\perp}$, 但是$Ax_{i-1} - b \in K_i$，因此$K_i = span\{v_1, v_2, \dots, v_{i-1}, r_{i-1}\}$

换言之，可以写出$v_i = c_0 r_{i-1} + \sum_{j=1}^{i-1} c_j v_j$

要确定c_0，考虑$v_i^T r_{i-1} = c_0 \|r_{i-1}\|^2$

要确定c_{i-1}，注意到$v_j^T A v_i = 0, \forall j < i-1$，令$v_{i-1}^T A v_i = 0$即可

类似地，令$v_j^T A v_i = 0$可得到其它的$c_j = 0$

Conjugate gradients

记Krylov子空间$K_0 = \{0\}, K_i = span\{b, Ab, \dots A^{i-1}b\}$
设$x_i = arg\ min_{x \in K_i} \|x - x_*\|_A^2$，其中$x_*$满足$Ax_* = b$

引理：记$v_i := x_i - x_{i-1}, r_i := b - Ax_i$，则

$$v_i = \frac{v_i^T r_{i-1}}{\|r_{i-1}\|^2}(r_{i-1} - \frac{r_{i-1}^T A v_{i-1}}{v_{i-1}^T A v_{i-1}} v_{i-1})$$

简化：令$d_i = \frac{\|r_{i-1}\|^2}{v_i^T r_{i-1}} v_i$,
即有$x_i = x_{i-1} + \frac{\|r_{i-1}\|^2}{d_i^T A d_i} d_i, d_i = r_{i-1} + \frac{\|r_{i-1}\|^2}{\|r_{i-2}\|^2} d_{i-1}$

可以证明：conjugate gradients在运行k步之后，误差最多是
正比于从所有$\deg(q) < k$且$q(0) = 1$的多项式中，最小的$\|q\|_\infty$

事实上：
$\|x_k - x_*\|_A^2 \le inf_{q(0)=1, deg(q) \le k}\ max_i q(\lambda_i)^2 \cdot \|b\|_{A^{-1}}^2$

与高斯消元相比：对于稀疏矩阵，恰当运用迭代法可以收敛更快
最多迭代n+1次

Conjugate Gradient Method

$x_0 =$ initial guess
$d_0 = r_0 = b - Ax_0$
for $k = 0, 1, 2, \dots, n-1$
  if $r_k = 0$, **stop**, **end**
  $\alpha_k = \frac{r_k^T r_k}{d_k^T A d_k}$
  $x_{k+1} = x_k + \alpha_k d_k$
  $r_{k+1} = r_k - \alpha_k A d_k$
  $\beta_k = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$
  $d_{k+1} = r_{k+1} + \beta_k d_k$
end