· Alternating Convex optimization

given nonconvex problem with variable $(x_1, x_2 \ldots x_n) \in \mathbb{R}^n$

$I_1, \ldots I_k \subset \{1, \ldots, n\}$ are index subset with $\bigcup_j I_j = \{1, \ldots n\}$

Suppose problem is convex in subset of variables $x_i, i \in I_j$

alternating convex optimization method: cycle through j, in each step optimizing over variable $x_i, i \in I_j$

eg. $f(x, A, b) = \|Ax - b\|$

f is convex in A when x, b fixed

x when A, b fixed

b when A, x fixed

f is jointly convex in X and B when A fixed $\quad I_1 = \{A, B\}$

A and B when X fixed $\quad I_2 = \{B, X\}$

f is not convex in A and x when b fixed

$$\min_{A,b,x} \|Ax - b\| \quad \text{s.t.} \quad A \in \mathcal{A},\ b \in \mathcal{B},\ x \in X \implies \text{Repeat } \left\{ \begin{array}{l} \min_x \|Ax - b\| \\ \text{s.t.} \quad x \in X \\ \\ \min_{A,b} \|Ax - b\| \\ \text{s.t.} \quad A \in \mathcal{A} \quad b \in \mathcal{B} \end{array} \right\}$$

special case: bi-convex problem

$X = (u, v)$ problem is convex in u/v with v/u fixed

alternate optimizing over u and v

eg. Non-negative matrix factorization

$\min. \|A - XY\|_F \qquad X \in \mathbb{R}^{m \times k} \quad Y \in \mathbb{R}^{k \times n} \quad$ data $A \in \mathbb{R}^{m \times n}$

s.t. $X_{ij}, Y_{ij} \geq 0$

when k is small: express A as product of low-rank matrices

if there's no constraint, this problem is solved by truncating the SVD of A

when $k=1$:

Take the positive part of A, and take the 1st singular vectors of SVD of A

perron-frobenius theorem:

for a pos matrix, the singular vectors associated with 1st singular value can be chosen to be positive alternating convex optimization:

solve QPs to optimize over X, then Y, then X ...

there are a handful of non-convex problems that can be solved exactly.

for example: can minimize any quadratic function, convex or not,

subject to a single quadratic inequality, convex not

eg. $\min_x \ x^T P x$ s.t. $x^T x \le 1$ $\Rightarrow$ mini eigenvec

· Conjugate gradient method

· Three classes of methods for linear equations

methods to solve linear system $Ax=b$ $A \in \mathbb{R}^{n\times n}$

1. dense direct (factor-solve method)
   - runtime depends only on size, (almost) independent of data
   - work well for n up to a few thousands

2. Sparse direct (factor solve method)
   - runtime depends on size, sparsity pattern, (almost) independent of data
   - work well for n up to 10^4 or 10^5 or more
   - require good heuristic for ordering

3. indirect (iterative methods)
   - runtime depends on data, size, sparsity, required accuracy
   - require tuning, preconditioning
   - good choices in many cases, only choice for $n=10^6$ or larger

· Symmetric positive definite linear system

SPD system of equations: $Ax=b$ $A\in\mathbb{R}^{n\times n}$ $A=A^T \succ 0$

eg. newton step $\nabla^2 f(x)\, dx = -\nabla f(x)$

(regularized) least square $(A^TA+uI)x = A^Tb$

minimize convex quadratic function $\frac{1}{2}x^TAx - b^Tx$

- Conjugate gradient overview

solve SPD system $Ax=b$ $(A \in \mathbb{R}^{n\times n} \succ 0)$ $\begin{cases} \text{in theory in } n \text{ iterations} \\ \text{each iteration requires a few inner products in } \mathbb{R}^n, \\ \quad \text{and one matrix-vector multiplication } z \to Az \end{cases}$

With round-off error, CG can work poorly

for some A, b, can get good approximation in $\ll n$ iterations

- Solution and Error

$x^* = A^{-1}b$ is the solution

x^* minimizes $f(x) = \frac{1}{2}x^TAx - b^Tx$

$\nabla f(x) = Ax - b$ (the residual)

$$\begin{aligned} f(x) - f^* &= \tfrac{1}{2}x^TAx - b^Tx - \tfrac{1}{2}x^{*T}Ax^* + b^Tx^* \\ &= \tfrac{1}{2}x^TAx - b^Tx - \tfrac{1}{2}b^TA^{-1}AA^{-1}b + b^TA^{-1}b \\ &= \tfrac{1}{2}x^TAx + \tfrac{1}{2}bA^{-1}b - b^Tx \\ &= \tfrac{1}{2}x^TAx + \tfrac{1}{2}x^{*T}Ax^* - x^{*T}Ax \\ &= \tfrac{1}{2}(x-x^*)^TA(x-x^*) = \tfrac{1}{2}\|x-x^*\|_A^2 \end{aligned}$$

a relative measure (comparing to 0)

$$\tau = \frac{f(x)-f^*}{f(0)-f^*} = \frac{\|x-x^*\|_A^2}{\|x^*\|_A^2}$$

- Residual

$r = b - Ax$ is called the residual at x

$\nabla f = Ax - b = A(x-x^*) = -r$

$$f(x) - f^* = \tfrac{1}{2}(x-x^*)^TA(x-x^*) = \tfrac{1}{2}r^TA^{-1}r = \tfrac{1}{2}\|r\|_{A^{-1}}^2$$

a commonly used measure of relative accuracy $\eta = \|r\|/\|b\|$

b is the residual norm of $x^*=0$

if $\|r\|/\|b\| \geq 1$, the solution is outperformed by $x=0$

· Krylov Subspace (controllability subspace)

$K_k = \text{span}\{b, Ab, \cdots, A^{k-1}b\}$
$= \{p(A)b \mid p \text{ polynomial}, \text{degree}(p) < k\}$

define kylov sequence $x^{(1)}, x^{(2)} \cdots$ as

$x^{(k)} = \underset{x \in K_k}{\text{argmin}}\, f(x) = \underset{x \in K_k}{\text{argmin}}\, \|x - x^*\|_A^2$

the CG algorithm (among others) generates the krylov sequence

· Properties of Krylov Sequence

$f(x^{(k+1)}) \le f(x^{(k)})$ (optimizing over a bigger and bigger subspace), but $\|r\|$ can increase
$x^{(n)} = x^*$ even if $K_n \ne \mathbb{R}^n$

cayley-hamilton theorem:

$\chi(s) = \det(sI - A) = s^n + a_1 s^{n-1} + \cdots + a_n$

by cayley-hamilton theorem:

$\chi(A) = A^n + a_1 A^{n-1} + \cdots a_n I = 0$

$\therefore\ A^{-1} = -\frac{1}{a_n} A^{n-1} - \frac{a_1}{a_n} A^{n-2} \cdots \frac{a_{n-1}}{a_n} I$

the solution of $Ax=b$ is in the span of krylov subspace $K_n = \{b\ Ab \cdots A^{n-1}b\}$

eg. $b = 0$

b is an eigenvector of A

$x^{(k)} = P_k(A)\, b$, where P_k is a polynomial with degree $(P_k) < k$

$x^{(k+1)} = x^{(k)} + \alpha_k r^{(k)} + \beta_k (x^{(k)} - x^{(k-1)})$ for some α_k β_k (basis of CG algorithm)

· Spectral analysis of krylov Sequence

$A = Q \Lambda Q^T$ $\quad Q$ orthogonal, $\Lambda = \mathrm{diag}(\lambda_1 \cdots \lambda_n)$

define $y = Q^T x \quad \bar{b} = Q^T b \quad y^* = Q^T x^*$

$$f(x) = \tfrac{1}{2} x^T A x + b^T x$$
$$= \tfrac{1}{2} x^T Q \Lambda Q^T x + b^T Q Q^T x$$
$$= \tfrac{1}{2} y^T \Lambda y - \bar{b}^T y \quad = \bar{f}(y)$$
$$= \sum_i^n \tfrac{1}{2} \lambda_i y_i^2 - \bar{b}_i y_i$$

$$y_i^* = \frac{\bar{b}_i}{\lambda_i} \qquad f^* = -\tfrac{1}{2} \sum_i^n \left(\bar{b}_i^2 / \lambda_i\right)$$

krylov sequence in terms of y

$$y^{(k)} = \underset{y \in \bar{K}_k}{\arg\min}\ \bar{f}(y) = \underset{y \in \bar{K}_k}{\arg\min}\ \tfrac{1}{2} y^T \Lambda y - \bar{b}^T y \qquad \bar{K}_k = \{ \bar{b}, \Lambda \bar{b} \cdots \Lambda^{k-1} \bar{b} \}$$

$$y_i^{(k)} = P_k(\lambda_i)\, \bar{b}_i \qquad \mathrm{degree}(p_k) < k$$

$$p_k = \underset{\mathrm{degree}(p) < k}{\arg\min} \sum_i^n \tfrac{1}{2} \bar{b}_i^2 \cdot \lambda_i \cdot P(\lambda_i)^2 - \bar{b}_i^2 P(\lambda_i)$$
$$= \underset{\mathrm{degree}(p) < k}{\arg\min} \sum_i^n \bar{b}_i^2 \left(\tfrac{1}{2} \lambda_i P(\lambda_i)^2 - P(\lambda_i) \right)$$

$$f(x^{(k)}) - f^* = \bar{f}(y^{(k)}) - f^*$$
$$= \min_{\deg(p) < k} \sum_i \bar{b}_i^2 \left(\tfrac{1}{2} \lambda_i P(\lambda_i)^2 - P(\lambda_i) \right) - -\tfrac{1}{2} \sum_i^n \left(\bar{b}_i^2 / \lambda_i \right)$$
$$= \min_{\deg(p) < k} \tfrac{1}{2} \sum_i \bar{b}_i^2 \left(\frac{(\lambda_i P(\lambda_i) - 1)^2}{\lambda_i} \right)$$
$$= \min_{\deg(p) < k} \tfrac{1}{2} \sum_i y_i^{*2} \lambda_i \underbrace{\left(\lambda_i P(\lambda_i) - 1 \right)}_{\text{new polynomial } q}{}^2$$
$$= \min_{\deg(q) \le k \quad q(0) = 1} \tfrac{1}{2} \sum_i y_i^{*2} \lambda_i \quad q(\lambda_i)^2$$
$$= \min_{\deg(q) \le k \quad q(0) = 1} \tfrac{1}{2} \sum_i \bar{b}_i^2 \frac{q(\lambda_i)^2}{\lambda_i}$$

$$\tau^{(k)} = \frac{f(x^{(k)}) - f^*}{f(0) - f^*} = \frac{f(x^{(k)}) - f^*}{\|x^*\|_A} = \frac{f(x^{(k)}) - f^*}{\|y^*\|_\Lambda}$$

$$= \frac{\min_{\deg(q) \le k,\ q(0)=1} \sum_i^n y_i^* \lambda_i q(\lambda_i)^2}{\sum_i^n y_i^* \lambda_i}$$ a weighted sum of $q(\lambda_i)^2$

$$\le \min_{\deg(q) \le k,\ q(0)=1} \max_{i=1\ldots n} q(\lambda_i)^2$$

if $\exists$ q of degree k, $q(0)=1$, that is small on spectrum (eigenvalue) of A,
then $f(x^{(k)}) - f^*$ is small

if eigenvalue are clustered in k groups, then $y^{(k)}$ is a good approximate solution