Notes on numerical methods for high-dimensional models
Young Ahn (@chiyahn)

---

1 Basics

1.1 LU Decomposition

Suppose we want to solve a system $A\mathbf{x} = \mathbf{b}$ with n rows. Gaussian eliminiation returns an augmented upper traiangular matrix of A, U, by backward substitution. Let $M^{(1)}, ..., M^{(n)}$ be the row operation matrices used for Gaussian elimination. Then we have

$$M^{(n)}M^{(n-1)}...M^{(2)}M^{(1)}A = U \tag{1}$$

Note that the row operation matrices are all lower triangular matrices and inverses of triangular matrices are all lower triangular matrices. Thus, defining $L = \left(M^{(n)}M^{(n-1)}...M^{(2)}M^{(1)}\right)^{-1}$, we have

$$A = LU \tag{2}$$

1.2 Cholesky Decomposition

Suppose that A is a symmetric positive definite (spd) matrix. Given an LU decomposition $A = LU$ (2), factor out the diagonal elements of U to decopmose them into

$$\begin{aligned} U &= \begin{pmatrix} u_{11} & & & & \\ & u_{22} & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & u_{nn} \end{pmatrix} \begin{pmatrix} 1 & u_{12}/u_{11} & \dots & \dots & u_{1n}/u_{11} \\ & 1 & u_{23}/u_{22} & \dots & u_{2n}/u_{22} \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ & & & & 1 \end{pmatrix} && (3) \\ &= D\overline{U} && (4) \end{aligned}$$

so $A = LD\overline{U}$, but since A is symmetric we have $A = LDL^T$. As A is positive definite, elements of D are all positive, so we have $A = GG^T$ where $G = LD^{1/2}$.

1.3 QR Decomposition

Let A be an $m \times n$ matrix of rank n and $\mathbf{a}_1, ..., \mathbf{a}_n$ denote its columns. Gram-Schmidt process finds coefficients $r_{12}, ..., r_{n-1,n}$ satisfying

$$\begin{aligned}
\mathbf{q}_1 &= \mathbf{a}_1 \\
\mathbf{q}_2 &= \mathbf{a}_2 - r_{12}\mathbf{a}_1 \\
&\vdots \\
\mathbf{q}_n &= \mathbf{a}_n - r_{n-1,n}\mathbf{a}_{n-1} - \cdots - r_{1n}\mathbf{a}_n
\end{aligned}$$

where $\mathbf{q}_1, ..., \mathbf{q}_n$ are orthogonal to each other. This procedure yields $A = \mathbf{Q}_1\mathbf{R}_1$. Defining $\mathbf{Q} = \mathbf{Q}_1\mathbf{\Lambda}^{-1/2}$ and $\mathbf{Q} = \mathbf{\Lambda}^{1/2}\mathbf{Q}_1$ where $\mathbf{\Lambda} = \text{diag}(\mathbf{q}_1^T\mathbf{q}_1, ..., \mathbf{q}_n^T\mathbf{q}_n)$ yields the decomposition $A = \mathbf{QR}$ where $\mathbf{Q}$ is an orthonomal matrix.

2 Stationary iteration methods

2.1 Splitting rule and convergence analysis

2.1.1 Splitting rule

Suppose we want to solve a large system $A\mathbf{x} = \mathbf{b}$ Let M be a matrix with a generalized inverse M^{-1} that is *informative* about A so that A can be decomposed into $A = M - N$ where N denotes a relatively small residual matrix. This yields the system $M\mathbf{x} = N\mathbf{x} + b$ which leads to the fixed point iteration of the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + M^{-1}(\mathbf{b} - A\mathbf{x}_k) \tag{5}$$

Note that $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$ is a residual term coming from $\mathbf{x}_k$.

2.1.2 Convergence analysis

Define the error of the estimate $\mathbf{x}_k$ as $\mathbf{e}_k = \mathbf{x} - \mathbf{x}_k = A^{-1}\mathbf{r}_k$. Note that (5) yields the following system

$$\mathbf{x}_k = M^{-1}\mathbf{b} + (I - M^{-1}A)\mathbf{x}_{k-1} \tag{6}$$

$$\mathbf{x} = M^{-1}\mathbf{b} + (I - M^{-1}A)\mathbf{x} \tag{7}$$

Taking the difference, we have

$$\begin{aligned}
\mathbf{e}_k &= \mathbf{x} - \mathbf{x}_k && (8) \\
&= (I - M^{-1}A)(\mathbf{x} - \mathbf{x}_{k-1}) && (9) \\
&= (I - M^{-1}A)\mathbf{e}_{k-1} && (10)
\end{aligned}$$

Hence, taking $T = I - M^{-1}A$, we have the dynamics $\mathbf{e}_k = T\mathbf{e}_{k-1}$, which has a convergence point iff $\rho(T) < 1$ where $\rho(T)$ is the spectral radius of T.

Examples 2.1 (Jacobi method)**.** Take $M = D$, where D is a square matrix of diagonal elements of A.

2.2 Gradient descent methods

Note that the fixed point iteration of (5) does not update M^{-1} matrix. Instead, we can adopt a gradient descent method to adopt

$$\mathbf{x}_{K+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k \tag{11}$$

where $\mathbf{p}_k$ is the search direction and α_k is the step size. The gradient descent method adopts $\mathbf{p}_k = \mathbf{r}_k$. To find the step size α_k, note that the objective function of $A\mathbf{x} = \mathbf{b}$ is

$$\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x} \tag{12}$$

Hence, replace $\mathbf{x}$ with the updated $\mathbf{x}_{k+1}$ gives

$$\phi(\mathbf{x}_{k+1}) = \frac{1}{2}(\mathbf{x} + \alpha_k\mathbf{r}_k)^T A(\mathbf{x} + \alpha_k\mathbf{r}_k) - \mathbf{b}^T(\mathbf{x} + \alpha_k\mathbf{r}_k) \tag{13}$$

Taking the derivatve with respect to α_k gives

$$\alpha_k = \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_k^T A\mathbf{r}_k} \tag{14}$$

2.3 Conjugate gradient

Definition 2.1 (Energy norm)**.** Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix. Then the energy norm of $\mathbf{x} \in \mathbb{R}^n$ with respect to A is defined as

$$\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^T A\mathbf{x}} \tag{15}$$

Definition 2.2 (Krylov subspace)**.** Let $A \in \mathbb{R}^{n\times n}$ and $\mathbf{x} \in \mathbb{R}^n$. The kth Krylov subspace of A with respect to $\mathbf{x}$ is given by

$$\mathcal{K}_k(A, \mathbf{x}) = \text{span}\{\mathbf{x}, A\mathbf{x}, ..., A^{k-1}\mathbf{x}\} \tag{16}$$

Lemma 2.1. [Mutual conjugacy with respect to a spd matrix implies linear independence] Let $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ be a series of vectors that are mutually conjugate to a spd matrix A, i.e.,

$$\langle \mathbf{v}_i, A\mathbf{v}_j\rangle = 0 \qquad \forall i \neq j \tag{17}$$

Then, $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ are linearly independent.

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix. Unlike the gradient descent method that finds the search direction and step by solving the associated least square of the residual, the CG iteration proceeds by finding solutions that minimizes the energy norm of the residual. The search directions and step sizes satisfy

$$\mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \frac{\langle \mathbf{r}_k, \mathbf{r}_k \rangle}{\langle \mathbf{p}_k, A\mathbf{p}_k \rangle} \mathbf{p}_k \tag{18}$$

$$\alpha_k = \frac{\langle \mathbf{r}_k, \mathbf{r}_k \rangle}{\langle \mathbf{p}_k, A\mathbf{p}_k \rangle} \tag{19}$$

It turns out that this procedure makes the residual vectors orthogonal to each other and search directions conjugate with respect to A. Hence, with Lemma 2.1, it can be shown that $\{\mathbf{r}_1, ..., \mathbf{r}_k\}$ forms an orthogonal basis of $\mathcal{K}_k(A; \mathbf{r}_0)$ and $\mathbf{x}_k - \mathbf{x}_0 \in \mathcal{K}_k(A; \mathbf{r}_0)$. Therefore, $\mathbf{x}_k$ can be regarded as the projection of $\mathbf{x}$ on the Krylov subspace $\mathcal{K}_k(A; \mathbf{r}_0)$ and as k increases the solution is more closely approximated to the actual one as the search space increases.

2.4 Preconditioning

Note that the CG requires A to have a small condition number to work. This is done by multiplying a matrix to solve

$$P^{-1} A \left(P^{-1}\right)^T P^T \mathbf{x} = P^{-1} b \tag{20}$$

Note that $P^{-1} A \left(P^{-1}\right)^T$ is psd if A is psd.

2.5 Krylov subspace methods

As seen in CG, Krylov subspace solvers is based on finding a solution by projecting $\mathbf{x}$ on a Krylov subspace. To achieve this, computing an appropriate orthogonal basis for the Krylov subspace is crucial.

2.5.1 Arnoldi process

Suppose we want to find the orthonormal basis for the Krylov subspace.

First begin with $\mathbf{r}_0$ to compute the first basis vector $\mathbf{q}_1$ by normalization. To find $\mathbf{q}_2$ that is orthogonal to $\mathbf{q}_1$, note first that $A\mathbf{q}_1$ is in the same direction as $A\mathbf{r}_0$, and we require

$$A\mathbf{q}_1 = h_{11}\mathbf{q}_1 + h_{21}\mathbf{q}_2 \tag{21}$$

Multiply both sides by $\mathbf{q}_1^T$, we have $h_{11} = \mathbf{q}_1^T A\mathbf{q}_1$. Using the fact that $\mathbf{q}_2$ is a unit vector, we have $h_{21} = \|A\mathbf{q}_1 - h_{11}\mathbf{q}_1$, which yields $\mathbf{q}_2$. This can be generalized to compute up to $\mathbf{q}_{k+1}$ vectors, and we have that

$$AQ_k = Q_{k+1}H_{k+1,k} \tag{22}$$

where Q_{k+1} is the matrix containing of the $(k+1)$ vectors of the orthogonal basis for Krylov subspace and $H_{k+1,k}$ is an $(k+1) \times k$ matrix whose jth column is $(h_{1j}, ..., h_{j+1,j}, 0, ..., 0)^T$. Multiplying both sides by Q_k^T gives

$$Q_k^T AQ_k = H_{k,k} \tag{23}$$

2.5.2 GMRES

Let $\mathbf{x}_k$ be a projection of $\mathbf{x}_0$ onto the Krylov subspace. Suppose we have Q_k, Q_{k+1}, and $H_{k+1,k}$ ready from the Arnoldi process. Then we can write

$$\mathbf{x}_k = \mathbf{x}_0 + Q_k\mathbf{z} \tag{24}$$

for some $\mathbf{z} \in \mathbb{R}^k$. Then we can rewrite the residual as

$$\begin{aligned} \mathbf{b} - A\mathbf{x}_k &= \mathbf{b} - A\mathbf{x}_0 - AQ_k\mathbf{z} & (25)\\ &= \mathbf{r}_0 - Q_{k+1}H_{k+1,k}\mathbf{z} & (26) \end{aligned}$$

whose norm is identical as the norm of

$$Q_{k+1}^T\mathbf{r}_0 - H_{k+1,k}\mathbf{z} \tag{27}$$

so that the minimization problem is now the one for $\mathbb{R}^k$ rather than $\mathbb{R}^n$.

2.5.3 MINRES

MINRES is a special case of GMRES when A is symmetric. If A is symmetric , then $H_{k,k}$ must be symmetric. Hence, $H_{k,k}$ is a tridiagonal matrix so the rest of GMRES algorithm can be simplified further.