ECS 231
Gradient descent methods for solving large scale eigenvalue problems

Generalized symmetric definite eigenvalue problem

- Generalized symmetric definite eigenvalue problem

$$Au = \lambda Bu$$

  where A and B are $n \times n$ symmetric, and B positive definite,
- All eigenvalues and eigenvectors are real
- Denote the eigenvalues by $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and their associated eigenvectors by u_1, u_2, $\ldots$, u_n and u_i are normalized, i.e., $\|u_i\|_B = 1$ for $i = 1, 2, \ldots, n$
- $\|\cdot\|_B$ is defined through the B-inner product

$$\langle x, y \rangle_B = \langle Bx, y \rangle \equiv y^T Bx.$$

Rayleigh quotient and minimization principles

- **Rayleigh Quotient** is defined by

$$\rho(x) = \frac{x^T A x}{x^T B x}$$

- Minimization principle:

$$\lambda_1 = \lambda_{\min} = \min_x \rho(x)$$
$$u_1 = \mathrm{argmin}_x \rho(x)$$

- In general for $i > 1$

$$\lambda_i = \min_{x \perp_B u_j,\, 1 \le j < i} \rho(x)$$
$$u_i = \mathrm{argmin}_{x \perp_B u_j,\, 1 \le j < i} \rho(x)$$

where by $x \perp_B y$ we mean that $\langle x, y \rangle_B = 0$.

Optimization methods for eigen-computation

- By the minimization principles, various optimization techniques can be *naturally* employed to compute λ_1 or the first few λ_i and their associated eigenvectors.
- Gradient (steepest) Decent (GD) and Conjugate Gradient (CG) methods are based on minimizing the Rayleigh Quotient $\rho(x)$.
- Two useful quantities:
  - The gradient of $\rho(x)$:

$$\nabla\rho(x) = \frac{2}{x^T Bx}[Ax - \rho(x)\, Bx] = \frac{2}{x^T Bx} r(x)$$

  - The Hessian of $\rho(x)$:

$$\nabla^2\rho(x) = \frac{2}{x^T Bx}[A - \rho(x)\, B - \nabla\rho(x)(Bx)^T - (Bx)\nabla\rho(x)^T]$$

Line search

- Minimizing the Rayleigh quotient along the direction of the gradient $r = r(x)$ is equivalently to solve the single-variable optimization problem

$$\min_t \rho(x + tr) = \rho(x + t_{\rm opt} r)$$

Steepest decent (SD) method

One step of the SD:

- Given an approximation $\mathbf{x}$ to u_1 and $\|x\|_B = 1$
- Compute the search direction:
  the (opposite) direction of the gradient $r = \nabla\rho(x)$,
- solve the problem

$$\inf_t \rho(x + tr) = \rho(x + t_{\text{opt}} r)$$

- update

$$x := x + t_{\text{opt}} r.$$

Steepest decent (SD) method

Given an initial approximation x_0 to u_1, and a relative tolerance `rtol`, the algorithm attempts to compute an approximate pair to (λ_1, u_1) with the prescribed `rtol`.

$x_0 = x_0/\|x_0\|_B$, $\rho_0 = x_0^T A x_0$, $r_0 = Ax_0 - \rho_0 B x_0$
for $i = 0, 1, \ldots,$ do
    if $\|r_i\|/(\|Ax_i\|_2 + |\rho_i|\,\|Bx_i\|_2) \leq$ `rtol`, break
    solve $\inf_t \rho(x_i + tr_i)$ for t_{opt}
    $\widehat{x} = x_i + t_{\text{opt}}\, r_i$, $x_{i+1} = \widehat{x}/\|\widehat{x}\|_B$
    $\rho_{i+1} = x_{i+1}^T A x_{i+1}$, $r_{i+1} = Ax_{i+1} - \rho_{i+1} B x_{i+1}$
end
return (ρ_i, x_i) as an approximate eigenpair to (λ_1, u_1)

Steepest decent (SD) method

Alternatively, the SD method can also be reformulated under Rayleigh-Ritz subspace projection framework:

```
select initial vector $x_0$, and compute $\rho_0 = \rho(x_0)$
for $i = 0, 1, \ldots$ until convergence do
    compute $r_i = Ax_i - \rho_i Bx_i$
    compute $H = Z^T AZ$ and $S = Z^T BZ$, where $Z = [x_i, r_i]$
    compute the smallest eigenpair $(\gamma_1, w_1)$ of $(H, S)$
    update $\rho_{i+1} = \gamma_1$, $x_{i+1} = Zw_1$
end
return $(\rho_i, x_i)$ as an approximate eigenpair to $(\lambda_1, u_1)$
```

Steepest decent (SD) method

Remarks:

1. Convergence analysis
   - The case $B = I$, *locally*, the convergence rate is

$$\frac{\rho_{i+1} - \lambda_1}{\rho_i - \lambda_1} \sim \left(\frac{1-\xi}{1+\xi}\right)^2, \quad \xi = \frac{\lambda_2 - \lambda_1}{\lambda_n - \lambda_1}.$$

     [Faddeev and Faddeeva'63] and [Knyazev and Skorokhodov'91]:
   - For the case $B \neq I$, [Yang'93].
2. Entending the search space

$$\rho_{\text{new}} = \min_{z \in \text{span}\{x_i, r_i\}} \rho(z)$$

$$\Downarrow$$

$$\rho_{\text{new}} = \min_{z \in \mathcal{K}_m(A - \rho_i B, x_i)} \rho(z)$$

Steepest decent (SD) method

3. Preconditioning the search direction

$$r_i \qquad \Rightarrow \qquad K_i r_i$$

   where K_i is a preconditioner (more later)

4. Introducing block implementation

$$R(X) = [r(x_1^{(i)}), r(x_2^{(i)}), \ldots, r(x_k^{(i)})]$$

Further reading: R.-C. Li, Rayleigh quotient based optimization methods for eigenvalue problems, in "Matrix Functions and Matrix Equations", Z. Bai et al ed., World Scientific, 2015.

Conjugate gradient method

- The Conjugate Gradient (CG) method was originally proposed in 1950s by Hestenes and Stiefel for solving linear system

$$Hx = b$$

  where H is symmetric positive definite.
- In the 1960s, it was extended by Fletcher and Reeves as an iterative method for **nonlinear** optimization problems. The extension is almost verbatim.
- Because of the optimality properties of Rayleigh quotients, it is natural to apply the CG method to compute a few eigenpairs of $A - \lambda B$.

CG for linear systems: review

- Define
$$\phi(x) = \frac{1}{2}x^T Hx - x^T b. \tag{1}$$
- the gradient $\nabla\phi(x) = Hx - b = r(x)$ (the residual vector)
  the Hessian matrix $H(x) = H$.
- $\phi(x)$ is a quadratic functional in x. It is convex and has a unique local and global minimum at $x = H^{-1}b$.
- Given an initial guess x_0, the CG method iteratively produces a sequence of approximations x_i and p_i, such that
$$\phi(x_{i+1}) = \min_{\alpha} \phi(x_i + \alpha p_i).$$
  where p_i are conjugate searching directions, i.e., $p_i^T Hp_j = 0$ for $i \neq j$, and $p_0 = r(x_0)$.

CG for linear systems: review

CG algorithm:

1. Give an initial guess x_0, compute $r_0 = Ax_0 - b$, and set $p_0 = r_0$;
2. For $i = 0, 1, \ldots,$ do

$$\begin{aligned}
\alpha_i &= \mathrm{argmin}_\alpha \phi(x_i + \alpha p_i) = \frac{r_i^T A r_i}{p_i^T A p_i} \\
x_{i+1} &= x_i + \alpha_i p_i \\
r_{i+1} &= r_i - \alpha_i H p_i \\
\beta_i &= \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i} \\
p_{i+1} &= r_{i+1} + \beta_i p_i
\end{aligned}$$

Note that β_i is chosen so that $p_{i+1}^T H p_i = 0$.

CG for linear systems: review

- In the absence of roundoff errors, it can be proved that
  1. $r_i^T r_j = 0, \quad$ for $0 \leq i, j \leq \ell$ and $i \neq j$
  2. $\text{subspan}\{r_0, r_1, \ldots, r_\ell\} = \text{subspan}\{p_0, p_1, \ldots, p_\ell\}$
     $= \text{subspan}\{r_0, Hr_0, \ldots, H^\ell r_0\}$
  3. $p_0, p_1, \ldots, p_\ell$ are linearly independent and
     $p_i^T H p_j = 0, \quad$ for $0 \leq i, j \leq \ell$ and $i \neq j$
  4. $\phi(x_\ell) = \min_{t_0,\ldots,t_\ell} \phi(x_0 + t_0 p_0 + t_1 p_1 + \cdots + t_\ell p_\ell)$.
- The CG method converges in at most n steps, *a direct method*, is a consequence of these properties.

CG method for eigenvalue computation

- In extending the CG method, the key is to recognize that the residual $r(x)$ in the linear system case plays the role of the gradient direction for $\phi(x)$.
- For the eigenproblem of $A - \lambda B$, the objective function is the Rayleigh quotient

  $$\rho(x) = \frac{x^T Ax}{x^T Bx}$$

  whose gradient $\nabla \rho(x)$ is collinear to

  $$r(x) = Ax - \rho(x)\, Bx.$$

CG method for eigenvalue computation

Given an initial approximation x_0 to u_1, and a relative tolerance `rtol`, the algorithm attempts to compute an approximate pair to (λ_1, u_1) with the prescribed `rtol`.

$x_0 = x_0/\|x_0\|_B$, $\rho_0 = x_0^T A x_0$, $r_0 = Ax_0 - \rho_0 B x_0$, $p_0 = r_0$;
for $i = 0, 1, \ldots,$ do
    if $\|r_i\|/(\|Ax_i\|_2 + |\rho_i| \, \|Bx_i\|_2) \le$ `rtol`, break;
    solve $\inf_t \rho(x_i + tp_i) = \rho(x_i + t_{\text{opt}} p_i)$.
    $\alpha_i = t_{\text{opt}}$
    $\hat{x} = x_i + \alpha_i \, p_i$, $x_{i+1} = \hat{x}/\|\hat{x}\|_B$;
    $\rho_{i+1} = x_{i+1}^T A x_{i+1}$, $r_{i+1} = Ax_{i+1} - \rho_{i+1} B x_{i+1}$,
    choose β_i and update $p_{i+1} = r_{i+1} + \beta_i p_i$,
endfor
Return (ρ_i, x_i) as an approximate eigenpair to (λ_1, u_1).

Conjugate gradient method

- Different choice of β_i leads to the different version of the CG method. Common choices:

$$\beta_i = \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i} \quad \text{or} \quad \beta_i = \frac{r_{i+1}^T (r_{i+1} - r_i)}{r_i^T r_i}$$

- Choose β_i, together with α_i, to minimize the Rayleigh quotient on the projection subspace

$$\mathrm{subspan}\{x_{i+1}, r_{i+1}, p_i\} = \mathrm{subspan}\{x_{i+1}, r_{i+1}, x_i\}$$

leads to the **locally optimal** method. – *the state of the art?*

Ref.: A. Knyazev, Toward the optimal preconditioned eigensolver: locally optimal block preconditioned conjugate gradient method.

SIAM J. Sci. Comput. 23(2):517-541, 2001

- Open problem: no quantitative estimate on the convergence rate of the CG method is available yet.