Preconditioned SD method

1. To compute the smallest eigenpair \((\lambda_1, u_1)\) of the matrix pair \((A, B)\), where \(A^T = A\) and \(B^T = B > 0\), the SD method, while always makes progress in driving the Rayleigh quotient towards a minimum, can be extremely slow in practice. The pre-conditioned SD (PSD) method is designed to overcome its slow convergence by modifying (preconditioning) its search (gradient) direction. The PSD method can be simply viewed as an application of the vanilla SD method after a linear transformation to the Rayleigh quotient.

2. Recall the kernel of the SD method is

\[
t_{\text{opt}} = \arg\min_t \rho(x + t r(x))
\]

\[
y = x + t_{\text{opt}} r(x),
\]

where \(r(x) = Ax - \rho(x)Bx\).

3. Consider transformation

\[
\tilde{x} = Lx,
\]

where \(L\) is \(n \times n\) and nonsingular. Then the Rayleigh Quotient

\[
\rho(x) = \frac{x^T Ax}{x^T Bx} = \frac{\tilde{x}^T (L^{-T} A L^{-1}) \tilde{x}}{\tilde{x}^T (L^{-T} B L^{-1}) \tilde{x}}
\]

(2)

corresponds to the eigenproblem of the matrix pencil \(L^{-T} AL^{-1} - \lambda L^{-T} BL^{-1}\).

4. Adopt the notational convention that the same symbol with and without a tilde is for \(A - \lambda B\) and for \(L^{-T} AL^{-1} - \lambda L^{-T} BL^{-1}\), respectively. For example,

\[
\tilde{\rho}(\tilde{x}) = \frac{\tilde{x}^T L^{-T} A L^{-1} \tilde{x}}{\tilde{x}^T L^{-T} B L^{-1} \tilde{x}} \equiv \rho(x),
\]

\[
\tilde{r}(\tilde{x}) = L^{-T} A L^{-1} \tilde{x} - \tilde{\rho}(\tilde{x}) L^{-T} B L^{-1} \tilde{x} \equiv L^{-T} r(x).\]

5. The kernel of the SD method for \(L^{-T} AL^{-1} - \lambda L^{-T} BL^{-1}\) is

\[
\tilde{t}_{\text{opt}} = \arg\min_{\tilde{t}} \tilde{\rho}(\tilde{x} + \tilde{t} \tilde{r}(\tilde{x})),
\]

\[
\tilde{y} = \tilde{x} + \tilde{t}_{\text{opt}} \tilde{r}(\tilde{x}).
\]

6. Eliminating the tilde variables to get back to the original variables, we have

\[
y = x + \tilde{t}_{\text{opt}} (L^T L)^{-1} r(x) \equiv x + \tilde{t}_{\text{opt}} K r(x),
\]

where \(K = (L^T L)^{-1}\) is the so-called pre-conditioner.

Furthermore, notice that

\[
\tilde{\rho}(\tilde{x} + \tilde{t} \tilde{r}(\tilde{x})) = \tilde{\rho}(L(x + \tilde{t} K r(x))) = \rho(x + \tilde{t} K r(x)).
\]

Therefore in terms of variable \(x\) and \(y\), the SD method for the transformed eigenproblem can be stated as, after dropping the tildes on the \(t\)-parameters,

\[
t_{\text{opt}} = \arg\min_{t} \rho(x + t K r(x)),
\]

\[
y = x + t_{\text{opt}} K r(x).\]

(4)
7. Comparing (1) and (4), we see the difference is the modification of the search direction:
\[ r(x) \quad \rightarrow \quad Kr(x) \]
by the selected pre-conditioner \( K \).

8. PSD algorithm: Given an initial approximation \( x_0 \) to \( u_1 \), and a relative tolerance \( rtol \), with proper selection of preconditioners \( K_i \), the following algorithm attempts to compute an approximate pair to \((\lambda_1, u_1)\) with the prescribed \( rtol \):

\[
\begin{align*}
1 & \quad x_0 = x_0/\|x_0\|_B, \rho_0 = x_0^T A x_0, r_0 = A x_0 - \rho_0 B x_0; \\
2 & \quad \text{for } i = 0, 1, \ldots, \text{do} \\
3 & \quad \quad \text{if } \|r_i\|/(\|Ax_i\|_2 + |\rho_i| \|Bx_i\|_2) \leq rtol, \text{ break; } \\
4 & \quad \quad \text{solve } \inf_t \rho(x_i + tK_i r_i) \text{ for } t_{opt}; \\
5 & \quad \quad \hat{x} = x_i + t_{opt} K_i r_i, x_{i+1} = \hat{x}/\|\hat{x}\|_B; \\
6 & \quad \rho_{i+1} = x_{i+1}^T A x_{i+1}, r_{i+1} = A x_{i+1} - \rho_{i+1} B x_{i+1}; \\
7 & \quad \text{end} \\
8 & \quad \text{Return } (\rho_i, x_i) \text{ as an approximate eigenpair to } (\lambda_1, u_1).
\end{align*}
\]

9. Convergence analysis does show that the rates could be dramatically improved with suitable pre-conditioners.

Samokish’s theorem: Let \( t_{opt} = \arg\min_t \rho(x + tK r(x)) \) and \( y = x + tK r(x) \), and denote the smallest positive and largest eigenvalue of \( K(A - \lambda_1 B) \) by \( \gamma \) and \( \Gamma \). If
\[
\tau \left( \sqrt{\Gamma} + \epsilon \right) \epsilon < 1,
\]
where
\[
\epsilon = \sqrt{\|B^{1/2} KB^{1/2}\|_2 [\rho(x) - \lambda_1]}, \quad \tau = \frac{2}{\gamma + \Gamma},
\]
then
\[
\rho(y) - \lambda_1 \leq \left[ \frac{\Delta + \tau \sqrt{\Gamma} \epsilon}{1 - \tau (\sqrt{\Gamma} + \epsilon) \epsilon} \right]^2 [\rho(x) - \lambda_1], \tag{5}
\]
where \( \kappa = \Gamma/\gamma, \Delta = (\kappa - 1)/(\kappa + 1) \).

10. Discussions on selecting good pre-conditioners

- By Samokish’s theorem, asymptotically, we have
\[
\frac{\rho(y) - \lambda_1}{\rho(x) - \lambda_1} \sim \left( \frac{\kappa - 1}{\kappa + 1} \right)^2, \quad \kappa = \frac{\Gamma}{\gamma}.
\]
\( \kappa \) is the conditioning of \( K(A - \lambda_1 B) \), after its zero eigenvalues discarded, but not the eigenvalues of \( A - \lambda_1 B \).
- One important aspect of the theorem lies as to what constitutes a good pre-conditioner, namely those making \( \Gamma/\gamma \) as close to 1 as possible.
- Since \( \Gamma/\gamma = 1 \) for \( K = (A - \lambda_1 B)^\dagger \), \( K \approx (A - \lambda_1 B)^\dagger \) would be a good pre-conditioner, the best one could hope for although albeit impractical.
- Naturally this suggests that, for example, to let \( K = (L^T L)^{-1} \approx A - \lambda_1 B \), an incomplete Cholesky decomposition.
- In practice, however, \( \lambda_1 \) is not available to begin with. A remedy would be to estimate a lower bound \( \mu \) of \( \lambda_1 \) and compute \( L^T L \approx A - \mu B \), instead.
- In some practical situations, \( A \) is also positive definite. In such cases, often simply \( \mu = 0 \) is chosen.
Pre-conditioned CG method

1. The preconditioned version of the CG method can be similarly viewed as the application of the *vanilla* CG method after a linear transformation

\[ \tilde{x} = Lx \]

done on the Rayleigh quotient \( \rho(x) \), as we did for the SD method.

2. PCG algorithm: Given an initial approximation \( x_0 \) to \( u_1 \), and a relative tolerance \( \text{rtol} \), with proper selection of preconditioners \( K_i \), the following PCG algorithm attempts to compute an approximate pair to \((\lambda_1, u_1)\) with the prescribed \( \text{rtol} \):

1. \[ x_0 = x_0/\|x_0\|_B, \rho_0 = x_0^T A x_0, r_0 = A x_0 - \rho_0 B x_0, p_0 = K r_0; \]
2. \( \text{for } i = 0, 1, \ldots, \text{do} \)
3. \( \text{if } \|r_i\|/(\|A x_i\|_2 + |\rho_i| \|B x_i\|_2) \leq \text{rtol}, \text{break;} \)
4. \( \text{solve } \inf_t \rho(x_i + t p_i) \text{ for } t_{\text{opt}}; \)
5. \( \hat{x} = x_i + t_{\text{opt}} p_i, x_{i+1} = \hat{x}/\|\hat{x}\|_B; \)
6. \( \rho_{i+1} = x_{i+1}^T A x_{i+1}, r_{i+1} = A x_{i+1} - \rho_{i+1} B x_{i+1}; \)
7. \( \text{choose } \beta_i \text{ and update } p_{i+1} = K r_{i+1} + \beta_i p_i; \)
8. \( \text{end} \)
9. \( \text{return } (\rho_i, x_i) \text{ as an approximate eigenpair to } (\lambda_1, u_1). \)

3. Comparing the CG method and its pre-conditioned version PCG, we see the difference is the modification of the residual from \( r_i \) to \( K r_i \) by the selected pre-conditioner \( K \).

4. Our discussions on selecting a good pre-conditioner for the SD method are often followed for the pre-conditioned CG method and its many variations. Numerical tests support this practice.

5. Over years, there are various heuristics on the convergence rates of the pre-conditioned CG method, but none is rigorously proven.