1 Lecture 8: Conjugate gradients and Krylov subspaces

1.1 Krylov subspaces

What's a standard non-convex problem? Finding the eigenvalues of a matrix. Below we list the standard methods for solving linear equations, and for solving eigenvalue equations.

	$Ax = b$	$Ax = \lambda x$ (non convex)
Basic	Gradient descent	Power methods
Accelerated	Chebyshev iteration	Chebyshev iteration
Accelerated and step size free	Conjugate gradient	Lanczos

Remark 1.1 (Chebyshev)**.** *The Chebyshev flow requires step sizes to be carefully chosen while the "accelerated and step size free" methods do not.*

Definition 1.2 (Krylov subspace)**.** For a matrix $A \in \mathbb{R}^{nxn}$ and a vector $b \in \mathbb{R}^n$, the Krylov sequence of order t is $b, Ab, A^2b, ..., A^tb$. Define the Krylov subspace as the span $\left[b, Ab, A^2b, ..., A^t \subset \mathbb{R}^n\right]$.

Fact 1 (Polynomial connection)**.** A useful fact, if A has eigenvectors $u_1, ...u_n$ and $t \geqslant rank(A)$ and $< b, u_i > \neq 0$, then $u_i \in K_t \; \forall i$

Suppose I have a vector $v \in K_t(A, b)$, then $\iff \exists \alpha_i : \; v = \alpha_0 b + \alpha_1 Ab + \cdots \alpha_t A^t b$. If we define $p(A) \sum_{i=1}^{t} \alpha_i A^i$ then $v = p(A)b$. Then $K_t(A, b) = \{p(A)b : \deg(p) \leqslant t\}$.

Suppose we have a symmetric matrix $A \in \mathbb{R}^{n\times n}$ that has orthonormal eigenvectors $u_1 \cdots u_n$ and ordered eigenvalues $\lambda_1 \geqslant \lambda_2 \ldots \geqslant \lambda_n$. Now suppose we write b in this basis, $b = \alpha_1 u_1 + ... + \alpha_n u_n$ with $\alpha_i =< u_i, b >$.

Remark 1.3 (Orthonormal eigenvectors)**.**

$$< u_i, u_j >= 0 \; i \neq j$$
$$< u_i, u_i >= 1$$

Remark 1.4. $p(A)u_i = p(\lambda_i)u_i$

Subsequently:

$$\begin{aligned} A &= \sum \lambda_i u_i \boldsymbol{u_i}^\top \\ p(A) &= \sum p(\lambda_i) u_i \boldsymbol{u_i}^\top \\ p(A)b &= \alpha_1 p(\lambda_1) u_1 + \alpha_2 p(\lambda_2) u_2 + \ldots + \alpha_n p(\lambda_n) u_n \end{aligned}$$

We want to find a polynomial such that $p(A)b \approx \alpha_1 u_1$. Ideally, we would have $p(\lambda_1) = 1$ and $p(\lambda_i) = 0$ for $i > 1$. One thing that'll do this is if we get $p(\lambda_1) = 1$

and $\max_{i>1} p(\lambda_i)$ as small as possible. This will give us a close approximation to the top eigenvalue.

What's an easy polynomial that'll get us pretty close to this? $p(\lambda) = \frac{\lambda^t}{\lambda_1^t}$. From this we get $p(\lambda_1) = 1$ and $p(\lambda_2) = (\frac{\lambda_2}{\lambda_1})^t$. We want $p(\lambda_2)$ to get small so we care about how close λ_2 is to λ_1.
$\lambda_1 = (1+\epsilon)\lambda_2$ then you need $p(\lambda_2) = \frac{1}{(1+\epsilon)^t}$ if you want $p(\lambda_2)$ to get small.

Remark 1.5 ($\angle$ notation)**.** $\tan\angle(a,b)$ *is the tangent of the angle between a and b*

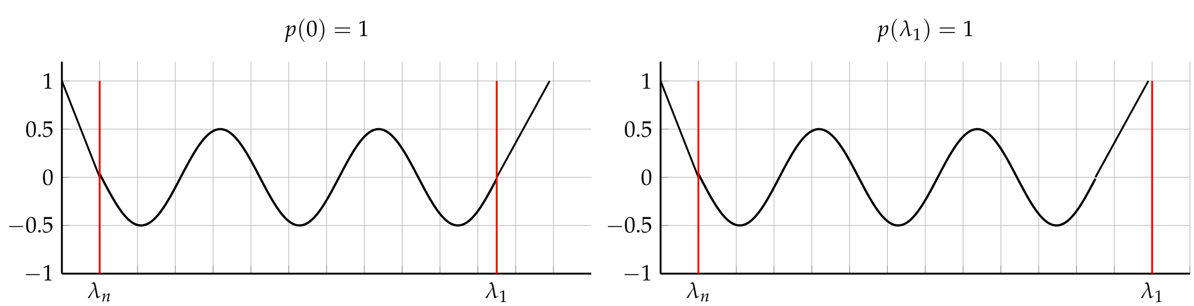
Theorem 1.6. $\tan\angle(p(A)b, u_1) \leqslant max_{j>1} \frac{|p(\lambda_j)|}{|p(\lambda_1)|} \tan\angle(b,u)$

Proof. Define $\theta = \angle(u_1, b)$. By this, we get $\sin^2\theta = \sum_{j>1}\alpha_j^2$ and $\cos^2\theta = |\alpha_1|^2$ and $\tan^2\theta = \sum_{j>1}\frac{|\alpha_j^2|}{|\alpha_1|^2}$. Now we can write $\tan^2\angle(p(A)b, u_1) = \sum_{j>1}\frac{|p(\lambda_j)\alpha_j|^2}{|p(\lambda_1)\alpha_1|^2} \leqslant \max_{j>1}\frac{|p(\lambda_j)|^2}{|p(\lambda_1)|^2}\sum_{j>1}\frac{\alpha_j|^2}{|\alpha_1|^2}$.
We note that this last sum $\sum_{j>1}\frac{\alpha_j|^2}{|\alpha_1|^2}$ is just $\tan\theta$ we have our desired result ■

Apply this to $p(\lambda) = \frac{\lambda^t}{\lambda_1^t}$ and $\lambda_1 = (1+\epsilon)\lambda_2$. This implies $\tan\angle(p(A)b, u_1) \leqslant \frac{1}{(1+\epsilon)t}\tan\angle(u_1, b)$. IF there is a big gap between λ_1 and λ_2 this converges quickly but it can be slow if $\lambda_1 \approx \lambda)2$.

Definition 1.7 (Power method)**.**

$$x_0 = \frac{b}{\|b\|}$$
$$x_t = \frac{Ax_{t-1}}{\|Ax_{t-1}\|}$$



1.2 Applying Chebyshev polynomials

So, as in prior lectures, we need to normalize our chebyshev polynomials. However, now we want to ensure that $p(\lambda_1) = 1$ so that we are picking out the first eigenvalue with the correct scaling.

Lemma 1.8. *A suitably rescaled degree t Chebyshev polynomial achieves*

$$\min_{p(\lambda_1)=1} \max_{\lambda \in [\lambda_2, \lambda_n]} p(\lambda) \leqslant \frac{2}{(1+\sqrt{\epsilon})^t} \tag{1}$$

where $\epsilon = \frac{\lambda_1}{\lambda_2} - 1$

	$Ax = b$	$Ax = \lambda x$ (non convex)
ϵ	$\frac{1}{\kappa} = \frac{\alpha}{\beta}$	$\frac{\lambda_1}{\lambda_2} - 1$

1.3 Conjugate gradient method

We want to solve $Ax = b,\ A \geqslant 0$.

$$\begin{aligned} x_0 = 0 &: \text{ "solution"} \\ r_0 = b &: \text{ "residual"} \\ p_0 = r_0 &: \text{ "search direction"} \end{aligned}$$

For t = 1,2, ...

$$\begin{aligned} \eta_t = \frac{\|r_t\|}{< p_{t-1}, Ap_{t-1} >} &: \text{ "step size"} \\ x_t = x_{t-1} + \eta_t p_{t-1} \\ r_t = r_{t-1} - \eta_t A r_{t-1} \\ p_t = r_t + \frac{\|r_t\|^2}{\|r_{t-1}^2\|} P_{t-1} \end{aligned}$$

Proof. Proof by induction. Show that 1-3 are true initially and stay true when the update rule is applied. ■

Lemma 1.9.

1. *span* $< r_0, ...r_{t-1} >= K_t(A, b)$
2. $j < t\ \ < r_t, r_j >= 0,\ r_t \perp K_t(a, b)$
3. $i \neq j :\ p_i \top A p_j = 0$*: "Conjugacy"*

Lemma 1.10. *Let* $\|u\|_A = \sqrt{\boldsymbol{u}^\top A u}$ *and* $< u, v >_A = \boldsymbol{u}^\top A v$ *and* $e_t = x^* - x_t$*. Then* e_t *minimizes* $\|x^* - x\|_A$ *over all vectors* $x \in K_{t-1}$*.*

Proof. We know that $x_t \in K_{t-1}$. Let $x \in K_{t-1}$. Define $x = x_t + \delta$. Then $e = x^* - x = x_t + \delta$. Lets compute the error in the A norm.

$$\begin{aligned}\|x^* - x\|_A^2 &= \|e_t + \delta\|^\top A(e_t + \delta) \\ e &= x^* - x \\ e &= x^* - x = e_t + \delta \\ \|x^* - x\|_A^2 &= e_t^\top A e_t + \delta^\top A\delta + 2e_t^\top A\delta \\ A\delta &\in K_{t-1}\end{aligned}$$

Want to argue that the last term $2e_t^\top A\delta = 0$ because e_t is orthogonal to the Krylov subspace. By definition $e_t^\top A = r_t$. ■