Mathematical Toolkit **Autumn 2015**

Lecture 10: November 2, 2015

Lecturer: Madhur Tulsiani

1 Lanczos Method

We now describe a method, which uses the Krylov subspace, discussed in the context of conjugate gradient method, to solve the following problem: Given a symmetric matrix $A \in R^{n \times n}$ such that $A \succ 0$, a vector $v \in \mathbb{R}^n$ and a function $f : \mathbb{R} \to \mathbb{R}$, (approximately) compute $f(A) \cdot v$. Here, f is a function such that $|f(\lambda)| < \infty$ for each eigenvalue λ of A. If $v_1, \ldots, v_n$ are an orthonormal eigenbasis for A with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$, then

$$A \;=\; \sum_{i=1}^{n} \lambda_i \cdot v_i v_i^T \,.$$

We define $f(A)$ as

$$f(A) \;=\; \sum_{i=1}^{n} f(\lambda_i) \cdot v_i v_i^T$$

Thus, $f(A)$ is a matrix with eigenvector $v_1, \ldots, v_n$ and the corresponding eigenvalues as $f(\lambda_1), \ldots, f(\lambda_n)$. Thus, the definition generalizes $p(A)$ for a polynomial p, which can be defined directly iin terms of matrix powers, and which has eigenvectors $v_1, \ldots, v_n$ with eigenvalues $p(\lambda_1), \ldots, p(\lambda_n)$. A particularly useful application of this method is to compute $\exp(A) \cdot v$ for a vector v.

Note that we can of course compute $f(A)$ as listed above by computing all the eigenvalues and eigenvectors of A. However, the utility of the Lanczos method is that it allows us to compute a good approximation to $f(A) \cdot v$ using only a few matrix-vector multiplications. Solving systems of linear equations can be thought of as a special case with the function $f(x) = 1/x$.

Let $\Lambda(A)$ denote the interval $[\lambda_{\min}(A), \lambda_{\max}(A)]$. Then, using the order-$t$ Krylov subspace $\mathcal{K}_t(A, v)$, the Lanczos method computes a vector v' such that

$$\left\| v' - f(A) \cdot v \right\| \;\leq\; 2 \cdot \left(\min_{\deg(\mathrm{p}) \leq t} \; \max_{\lambda \in \Lambda(A)} \; |f(\lambda) - p(\lambda)| \right) \cdot \|v\| \,.$$

Thus, if f can be well-approximated in the interval $\Lambda(A)$ by a degree-t polynomial, the Lanczos method computes a good approximation to $f(A) \cdot v$. Also, like in the case of the conjugate gradient method, the Lanczos method works *without any knowledge of the best approximating polynomial.*

We now describe the method. Let $u_0, \ldots, u_t$ be an orthonormal basis for $\mathcal{K}_t(A, v)$ under the usual inner product on $\mathbb{R}^n$. We saw in the homewotk that such an orthonormal basis can be computed using only $O(t)$ matrix-vector multiplications. Let $U \in \mathbb{R}^{n \times (t+1)}$ be the matrix with columns $u_0, \ldots u_t$. Check the following:

Exercise 1.1 *Let $U \in \mathbb{R}^{n\times(t+1)}$ be as defined above. Show that*

1. *$U^TU = I_{t+1}$, where I_{t+1} is the $(t+1)\times(t+1)$ identity matrix.*
2. *UU^T is the matrix of the orthogonal projection onto the space $\mathcal{K}_t(A,v)$ i.e.,*

$$UU^Tw = \begin{cases} w & \text{if } w \in \mathcal{K}_t \\ 0 & \text{if } w \in \mathcal{K}_t^{\perp} \end{cases}$$

Define the matrix $B \in \mathbb{R}^{(t+1)\times(t+1)}$ as

$$B = U^TAU\,.$$

Exercise 1.2 *Let p be any polynomial with $deg(p) \le t$. Prove that*

$$p(A)\cdot v = Up(B)U^T\cdot v\,.$$

Note that the vector v here is the vector used to define the Krylov subspace $\mathcal{K}_t(A,v)$.

Since B is a $(t+1)\times(t+1)$ matrix, we can compute it's SVD in time poly(t). Thus, we can compute the matrix $f(B)$ in time poly(t). We define our approximation v' as

$$v' = Uf(B)U^T\cdot v\,.$$

From the exercise above, we know that v' is exact if $f(\cdot)$ is a polynomial of degree at most t. We show that also when f can be approximated by a polynomial, the error is small. Let $p(x)$ be any polynomial with $\deg(p) \le t$. Let $r(x) = f(x) - p(x)$. Thus, we have

$$\begin{aligned}
\left\|f(A)\cdot v - v'\right\| &= \left\|f(A)\cdot v - Uf(B)U^T\cdot v\right\| \\
&= \left\|p(A)\cdot v + r(A)\cdot v - Up(B)U^T\cdot v - Ur(B)U^T\cdot v\right\| \\
&= \left\|r(A)\cdot v - Ur(B)U^T\cdot v\right\| \\
&= \|r(A)\cdot v\| + \left\|Ur(B)U^T\cdot v\right\| \\
&\le \left(\|r(A)\|_2 + \left\|Ur(B)U^T\right\|_2\right)\cdot\|v\| \\
&\le \left(\max_{\lambda\in\Lambda(A)}|r(\lambda)| + \max_{\lambda'\in\Lambda(B)}\left|r(\lambda')\right|\right)\cdot\|v\|
\end{aligned}$$

Above, we use $\Lambda(B)$ to denote the interval $[\lambda_{\min}(B), \lambda_{\max}(B)]$. We can simplify the above bound using the folowing.

Exercise 1.3 *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{(t+1)\times(t+1)}$ be as above. Show that $\Lambda(B) \subseteq \Lambda(A)$.*

Thus, we get that for any polynomial p with $\deg(p) \le t$,

$$\left\|v' - f(A)\cdot v\right\| \le 2\cdot\left(\max_{\lambda\in\Lambda(A)}|f(\lambda) - p(\lambda)|\right)\cdot\|v\|\,.$$

Since, this is true for all polynomials p, we get that

$$\left\|v' - f(A)\cdot v\right\| \le 2\cdot\left(\min_{\deg(p)\le t}\ \max_{\lambda\in\Lambda(A)}\ |f(\lambda) - p(\lambda)|\right)\cdot\|v\|\,.$$

2 Basics of probability

We recall very briefly the basics of probability and random variables. For a much better and detailed introduction, please see the lecture notes by Terry Tao, linked from the course homepage.

2.1 The finite case

Let Ω be a finite set. Let $\mu : \Omega \to [0,1]$ be a function such that

$$\sum_{\omega \in \Omega} \mu(\omega) \;=\; 1.$$

We often refer to Ω as a sample space and the function μ as a probability distribution on this space. An event can be thought of as a subset of outcomes i.e., any $E \subseteq \Omega$ defines an event, and we define its probability as

$$\mathbb{P}[E] \;=\; \sum_{\omega \in E} \mu(\omega)\,.$$

A real-valued random variable over Ω is any function $X : \Omega \to \mathbb{R}$. We define

$$\mathbb{E}[X] \;=\; \sum_{\omega \in \Omega} \mu(\omega) \cdot X(\omega)\,.$$

Everything defined above can also be extended to countable spaces but we need to be careful about the convergence of the above summations.

2.2 The case of (uncountably) infinite probability spaces

Extending the idea of defining a probability *for each outcome* becomes problematic, when we try to extend it to uncountably infinite spaces. For example, let $\Omega = [0.1]$. Let $\mu : [0,1] \to [0,1]$ be a function, which we want to think of as a probability distribution. Define the set

$$S_n \;=\; \left\{ x \in [0,1] \mid \mu(x) \geq \frac{1}{n} \right\}.$$

Since we want the total probability to add up to 1, we must have $|S_n| \leq n$. Also,

$$\mathrm{Supp}(\mu) \;=\; \{x \in [0,1] \mid \mu(x) > 0\} \;\subseteq\; \cup_{n=1}^{\infty} S_n\,.$$

Since $\cup_{n=1}^{\infty} S_n$ is a countable set, $\mu(x) > 0$ only for countably many points x. Hence, it is problematic to think of the probability of the outcome x, for each $x \in [0,1]$. This can be resolved by only talking of probabilities of *events* for an allowed set of events obeying some nice properties. Such a set is known as a σ-algebra or a σ-field.

Definition 2.1 *Let 2^{Ω} denote the set of all subsets of Ω. A set $\mathcal{F} \subseteq 2^{\Omega}$ is called a* σ-field *(or σ-algebra) if*

1. $\emptyset \in \mathcal{F}$.

2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ *(where* $A^c = \Omega \setminus A$*).*

3. *For a (countable) sequence* $A_1, A_2, \ldots$ *such that each* $A_i \in \mathcal{F}$*, we have* $\cup_i A_i \in \mathcal{F}$*.*

We then think of the sets in $\mathcal{F}$ as the allowed events. We can now define probabilities as follows.

Definition 2.2 *Given a* σ*-field* $\mathcal{F} \subseteq 2^\Omega$*, a function* $\mu : \mathcal{F} \to [0,1]$ *is known as a* probability measure *if*

1. $\mu(\emptyset) = 0$*.*

2. $\mu(E^c) = 1 - \mu(E)$ *for all* $E \in \mathcal{F}$*.*

3. *For a (countable) sequence of disjoint sets* $E_1, E_2, \ldots$ *such that all* $E_i \in \mathcal{F}$*, we have*

$$\mu\left(\cup_i E_i\right) \;=\; \sum_i \mu(E_i)\,.$$

Note that the above definition do not say anything about unions of an uncountably infinite collection of sets. We can of course define probability measures on $\mathcal{F} = 2^\Omega$ and hence define $\mu(x)$ for all $x \in \Omega$. However, as we saw above, such measures will only have $\mu(x) > 0$ countably many x. Considet the following example.

Example 2.3 *Let* $\Omega = [0,1]$ *and* $\mathcal{F} = 2^\Omega$*. Let* $T = \left\{0, \frac{1}{3}, \frac{2}{3}, 1\right\}$*. For each* $S \in \mathcal{F}$*, define*

$$\mu(S) \;=\; \frac{|S \cap T|}{4}\,.$$

In the above example, $\mu(x) > 0$ only for the points in T. To think of the "uniform distribution" on the the space $[0,1]$, we would like that for $a, b \in [0,1]$ with $a < b$, the probability measure satisfies $\mu[a,b] = b - a$. It is a non-trivial result that such a probability measure indeed exists. This probability measure is known as the *Lebesgue* measure and is defined over a σ-algebra known as the *Borel* σ-algebra. The Borel σ-algebra does not contain all subsets of $[0,1]$ but does contain all intervals $[a,b]$. In fact, one can use the axiom of choice to show that a measure satisfying the requirement that $\mu([a,b]) = b - a$ cannot be defined over the σ-algebra 2^Ω.