Mathematical Toolkit **Autumn 2025**

Lecture 10: October 30, 2025

Lecturer: Madhur Tulsiani

1 The conjugate gradient method

In the previous lecture, we saw the steepest descent or gradient descent method for finding a solution to the linear system $Ax = b$ for $A \succ 0$. The method guarantees $\|x_t - u\| \leq \varepsilon \cdot \|x_0 - u\|$ after $t = O(\kappa \cdot \log(1/\varepsilon))$ iterations, where κ is the condition number of the matrix A, and u is the unique (unknown) solution to the system. We will see that the conjugate gradient can obtain a similar guarantee in $O(\sqrt{\kappa} \cdot \log(1/\varepsilon))$ iterations.

For the steepest descent method, if we start from $x_0 = 0$, we get

$$x_t - u \;=\; (I - \eta A)(-u)\,,$$

which gives $x_t = p(A) \cdot b$ for some polynomial p of degree at most t. The conjugate gradient method just takes this idea of finding an x of the form $p(A) \cdot b$ and runs with it. The method finds an $x_t = p_t(A) \cdot b$ where p_t is the *best* polynomial of degree at most t i.e., the polynomial which minimizes the function $\frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$. However, the method does not explicitly work with polynomials. Instead we use the simple observation that any vector of the form $p_t(A) \cdot b$ lies in the subspace $\text{Span}\left(\{b, Ab, \ldots, A^t b\}\right)$ and the method finds the best vector in the subspace at every time t.

Definition 1.1 *Let $\varphi : V \to V$ be a linear operator on a vector space V and let $v \in V$ be a vector. The* Krylov subspace *of order t defined by φ and v is defined as*

$$\mathcal{K}_t(\varphi, v) \;:=\; \text{Span}\left(\left\{v, \varphi(v), \ldots, \varphi^{t-1}(v)\right\}\right)\,.$$

Thus, at step t of the conjugate gradient method, we find the best vector in the space $\mathcal{K}_t(A, b)$ (we will just write the subspace as $\mathcal{K}_t$ since A and b are fixed for the entire argument). The trick of course is to be able to do this in an iterative fashion so that we can quickly update the minimizer in the space $\mathcal{K}_{t-1}$ to the minimizer in the space $\mathcal{K}_t$. This can be done by expressing the minimizer in $\mathcal{K}_{t-1}$ in terms of a convenient orthonormal basis $\{w_0, \ldots, w_{t-1}\}$ for $\mathcal{K}_{t-1}$. It turns out that if we work with a basis which is orthonormal with respect to the inner product $\langle \cdot, \cdot \rangle_A$, at step t we only need to update the component of the minimizer along the new vector w_t we get to obtain a basis for $\mathcal{K}_t$.

1.1 The algorithm

Recall that we defined the inner product $\langle x, y \rangle_A := \langle Ax, y \rangle$ where $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^n$. As before we consider the function $\frac{1}{2}\langle Ax, x \rangle - \langle b, x \rangle$ and pick

$$x_t \;:=\; \operatorname*{arg\,min}_{x \in \mathcal{K}_t} f(x)\,.$$

This can also be thought of as finding the closest point to u in the space $\mathcal{K}_t$ (under the distance $\|\cdot\|_A$) since

$$\begin{aligned} f(x) \;=\; \frac{1}{2}\langle Ax, x \rangle - \langle b, x \rangle \;&=\; \frac{1}{2}\langle Ax, x \rangle - \langle Au, x \rangle \;=\; \frac{1}{2}\langle x, x \rangle_A - \langle u, x \rangle_A \\ &=\; \frac{1}{2} \cdot \left(\|x - u\|_A^2 - \|u\|_A^2 \right)\,, \end{aligned}$$

which gives

$$x_t \;=\; \operatorname*{arg\,min}_{x \in \mathcal{K}_t} f(x) \;=\; \operatorname*{arg\,min}_{x \in \mathcal{K}_t} \|x - u\|_A\,.$$

We have already seen how to compute find the characterize the closest point in a subspace, to a given point. Let $\{w_0, \ldots, w_{t-1}\}$ be an orthonormal basis for $\mathcal{K}_t$ under the inner product $\langle \cdot, \cdot \rangle_A$. Completing this to an orthonormal basis $\{w_0, \ldots, w_{n-1}\}$ for $\mathbb{R}^n$, let u be expressible as

$$u \;=\; \sum_{i=0}^{n-1} c_i \cdot w_i \;=\; \sum_{i=0}^{n-1} \langle u, w_i \rangle_A \cdot w_i\,.$$

Then we know that the closest point x_t in $\mathcal{K}_t$, under the distance $\|\cdot\|_A$ is given by

$$x_t \;=\; \sum_{i=0}^{t-1} \langle u, w_i \rangle_A \cdot w_i \;=\; \sum_{i=0}^{t-1} \langle Au, w_i \rangle \cdot w_i \;=\; \sum_{i=0}^{t-1} \langle b, w_i \rangle \cdot w_i$$

Note that even thhough we do not know u, we can find x_t given an orthonormal basis $\{w_0, \ldots, w_{t-1}\}$, since we can compute $\langle b, w_i \rangle$ for all w_i. This gives the following algorithm:

- Start with $w_0 = b / \|b\|_A$ as an orthonormal basis for $\mathcal{K}_1$.
- Let $x_t \;=\; \sum_{i=0}^{t-1} \langle b, w_i \rangle \cdot w_i$ for a basis $\{w_0, \ldots, w_{t-1}\}$ orthonormal under the inner product $\langle \cdot, \cdot \rangle_A$.
- Extend $\{w_0, \ldots, w_{t-1}\}$ to a basis of $\mathcal{K}_{t+1}$ by defining

$$v_t \;=\; A^t b - \sum_{i=0}^{t-1} \langle A^t b, w_i \rangle_A \cdot w_i \quad \text{and} \quad w_t \;=\; \frac{v_t}{\sqrt{\langle v_t, v_t \rangle_A}}\,.$$

- Update $x_{t+1} = x_t + \langle b, w_t \rangle \cdot w_t$.

Notice that the basis extension step here seems to require $O(t)$ matrix-vector multiplications in the t^{th} iteration and thus we will need $O(t^2)$ matrix-vector multiplications in total for t iterations. This would negate the quadratic advantage we are trying to gain over steepest descent. However, in the homework you will see a way of extending the basis using only $O(1)$ matrix-vector multiplications in each step.

1.2 Bounding the number of iterations

Since x_t lies in the subspace $\mathcal{K}_t$, we have $x_t = p(A) \cdot b$ for some polynomial p of degree at most $t-1$. Thus,

$$x_t - u = p(A) \cdot b - u = p(A) \cdot A \cdot u - u = (I - p(A) \cdot A) \cdot (x_0 - u),$$

since $x_0 = 0$. We can think of $I - p(A)A$ as a polynomial $q(A)$, where $\deg(q) \leq t$ and $q(0) = 1$. Recall from last lecture that the minimizer of $f(x)$ is the same as the minimizer of $\langle x-u, x-u \rangle_A = \|x-u\|_A^2$. Since $p(A)b$ is the minimizer of $f(x)$ in $\mathcal{K}_t$, we have

$$\|x_t - u\|_A^2 = \min_{q \in Q_t} \|q(A)(x_o - u)\|_A^2 ,$$

where Q_t is the set of polynomials defined as

$$Q_t := \{q \in \mathbb{R}[z] \mid \deg(q) \leq t, q(0) = 1\} .$$

Use the fact that if λ is an eigenvalue of a matrix M, then λ^t is an eigenvalue of M^t (with the same eigenvector) to prove that the following.

Exercise 1.2 *Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A. Then for any polynomial q and any $v \in \mathbb{R}^n$,*

$$\|q(A)v\|_A \leq \left(\max_i |q(\lambda_i)| \right) \cdot \|v\|_A .$$

Using the above, we get that

$$\|x_t - u\|_A \leq \left(\min_{q \in Q_t} \max_i |q(\lambda_i)| \right) \cdot \|x_0 - u\|_A .$$

Thus, the problem of bounding the norm of $x_t - u$ is reduced to finding a polynomial q of degree at most t such that $q(0) = 1$ and $q(\lambda_i)$ is small for all i.

Exercise 1.3 *Verify that using $q(z) = \left(1 - \frac{2z}{\lambda_1 + \lambda_n}\right)^t$ recovers the guarantee of the steepest descent method.*

Note that the conjugate gradient method itself does not need to know anything about the optimal polynomials in the above bound. The polynomials are only used in the analysis of the bound. The following claim, which can be proved by using slightly modified Chebyshev polynomials, suffices to obtain the desired bound on the number of iterations.

Claim 1.4 *For each $t \in \mathbb{N}$, there exists a polynomial $q_t \in Q_t$ such that*

$$|q_t(z)| \;\leq\; 2 \cdot \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^t \quad \forall z \in [\lambda_1, \lambda_n]\,.$$

We will prove the claim later using Chebyshev polynomials. However, using the claim we have that

$$\|x_t - u\|_A \;\leq\; \left(\min_{q \in Q_t} \max_i |q(\lambda_i)|\right) \cdot \|x_0 - u\|_A \;\leq\; 2 \cdot \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^t \cdot \|x_0 - u\|_A\,.$$

Thus, $O(\sqrt{\kappa}\log(1/\varepsilon))$ iterations suffice to ensure that $\|x_t - u\|_A \leq \varepsilon \cdot \|x_0 - u\|_A$.

1.3 Chebyshev polynomials

The Chebyshev polynomial of degree t is given by the expression

$$P_t(z) \;=\; \frac{1}{2} \cdot \left[\left(z + \sqrt{z^2-1}\right)^t + \left(z - \sqrt{z^2-1}\right)^t\right]\,.$$

Note that this is a polynomial since the odd powers of $\sqrt{z^2-1}$ will cancel from the two expansions. For $z \in [-1,1]$ this can also be written as

$$P_t(z) \;=\; \cos\left(t \cos^{-1}(z)\right)\,,$$

which shows that $P_t(z) \in [-1,1]$ for all $z \in [-1,1]$.

Using these polynomials, we can define the required polynomials q_t as

$$q_t(z) \;=\; \frac{P_t\left(\frac{\lambda_1+\lambda_n-2z}{\lambda_n-\lambda_1}\right)}{P_t\left(\frac{\lambda_1+\lambda_n}{\lambda_n-\lambda_1}\right)}\,.$$

The denominator is a constant which does not depend on z and the numerator is a polynomial of degree t in z. Hence $\deg(q_t) = t$. Also, the denominator ensures that $q_t(0) = 1$. Finally, for $z \in [\lambda_1, \lambda_n]$, we have $\left|\frac{\lambda_1+\lambda_n-2z}{\lambda_n-\lambda_1}\right| \leq 1$. Hence, the numerator is in the range $[-1,1]$ for all $z \in [\lambda_1, \lambda_n]$. This gives

$$|q_t(z)| \;\leq\; \frac{1}{P_t\left(\frac{\lambda_1+\lambda_n}{\lambda_n-\lambda_1}\right)} \;\leq\; 2 \cdot \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^t \;=\; 2 \cdot \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^t\,.$$

The last bound above can be computed directly from the first definition of the Chebyshev polynomials.

An detailed treatment of the conjugate gradient method, and a related method called the Lanczos Method, which also uses the Krylov subspace, can be found in the excellent monograph by Vishnoi [Vis13].

2 Basics of probability: the finite case

Probability theory is a mathematical framework used to model uncertainty and variability in nature. It is by no means the only contender for this role, but has weathered many trials through time. A good deal of probability theory was developed long before being formalized in the way that we're familiar with now, which is due to Kolmogorov. One could cite the works of Laplace, Poisson, Gauss, to name a few. So in some sense the formalization we present here is not strictly necessary, at least for most simple problems. But it does place the whole field on a very stable foundation, which is also helpful whenever something challenges our grasp of this otherwise intuitive discipline.

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 Probability spaces

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

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

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

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

2.2 Random Variables and Expectation

In a finite probability space, a real-valued random variable over Ω is any function $X : \Omega \to \mathbb{R}$. So a random variable is technically neither random (it's quite deterministic) nor a variable (it's a function), but it's a terminology that has stuck.

In a finite probability space, we define the expectation of a random variable X as:

$$\mathbb{E}[X] := \sum_{\omega \in \Omega} \nu(\omega) \cdot X(\omega).$$

An extremely useful fact about expectation is that is a linear transformation from the space of random variables to $\mathbb{R}$. In particular, if X and Y are random variables, then $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, and $\mathbb{E}[c \cdot X] = c \cdot \mathbb{E}[X]$.

Proposition 2.1 (Linearity of Expectation) *For any two random variables X and Y, $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, and $\mathbb{E}[c \cdot X] = c \cdot \mathbb{E}[X]$.*

Proof: This follows directly from the definition.

$$\mathbb{E}[X+Y] = \sum_{\omega \in \Omega} \nu(\omega) \cdot (X(\omega) + Y(\omega)) = \sum_{\omega \in \Omega} \nu(\omega) \cdot X(\omega) + \sum_{\omega \in \Omega} \nu(\omega) \cdot Y(\omega) = \mathbb{E}[X] + \mathbb{E}[Y].$$

The proof for $\mathbb{E}[c \cdot X] = c \cdot \mathbb{E}[X]$ is similar. ■

Example: Card shuffling Suppose we unwrap a fresh deck of cards and shuffle it until the cards are completely random. How many cards do we expect to be in the same position as they were at the start? To solve this, let's think formally about what we are asking. We are looking for the expected value of a random variable X denoting the number of cards that end in the same position as they started. We can write X as a sum of indicator random variables X_i, one for each card, where $X_i = 1$ if the ith card ends in position i and $X_i = 0$ otherwise. These X_i are easy to analyze: $\mathbb{P}(X_i = 1) = 1/n$ where n is the number of cards. $\mathbb{P}(X_i = 1)$ is also $\mathbb{E}[X_i]$. Now we use linearity of expectation:

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \ldots + X_n] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n] = 1.$$

So, this is interesting: no matter how large a deck we are considering, the expected number of cards that end in the same position as they started is 1.

References

[Vis13] Nisheeth K. Vishnoi, $Lx = b$, Foundations and Trends® in Theoretical Computer Science **8** (2013), no. 1–2, 1–141. 5