AM 221: Advanced Optimization Spring 2016

Prof. Yaron Singer *Section 5 — Wednesday, Feb. 24th*

1 Convex functions

a. Let $f_1, \ldots, f_k$ be k convex functions from $\mathbb{R}^n$ to $\mathbb{R}$, show that the function g defined by

$$g(\mathbf{x}) = \max_{1 \leq i \leq k} f_i(\mathbf{x}),\ x \in \mathbb{R}^n$$

is convex.

b. Let $f_1, \ldots, f_k$ be k convex functions from $\mathbb{R}^n$ to $\mathbb{R}$ and let $\lambda_1, \ldots, \lambda_k$ be k non-negative reals, show that the function g defined by:

$$g(\mathbf{x}) = \sum_{i=1}^{i} \lambda_k f_i(\mathbf{x})$$

is convex.

c. Let f be a convex function from $\mathbb{R}^n$ to $\mathbb{R}$. Show that for any $(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^n \times \mathbb{R}^n$ the function $f_{\mathbf{x},\mathbf{d}} : \mathbb{R} \to \mathbb{R}$ defined by:

$$f_{\mathbf{x},\mathbf{d}}(\lambda) = f(\mathbf{x} + \lambda \mathbf{d}),\ \lambda \in \mathbb{R}$$

is convex.

Remark. The last property is important, because it implies that the exact line search problem in the gradient descent algorithm is a single-dimensional convex optimization problem.

2 Spectral theory

Throughout this section A is a matrix in $\mathbb{R}^{n \times n}$.

Definition 1. *$\lambda \in \mathbb{R}$ is an* eigenvalue *of A iff there exists $\mathbf{x} \in \mathbb{R}^n \setminus \{0\}$ such that $A\mathbf{x} = \lambda \mathbf{x}$. $\mathbf{x}$ is then called an* eigenvector *associated with λ. The set of all eigenvectors for λ, $E_\lambda \stackrel{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \lambda \mathbf{x}\}$ (note that we also include 0) is called the* eigenspace *associated with λ.*

Remark. Note that 0 is an eigenvalue of A is equivalent to saying that E_0 is not the trivial space $\{0\}$. Which is exactly saying that the kernel (or nullspace) of A is non trivial, *i.e* A is a singular matrix.

Remark. How to compute eigenvalues? λ is an eigenvalue of A iff the matrix $A - \lambda I_n$ (where I_n is the identity matrix of size n) is singular. In other words, the eigenvalues of A are the solutions to the equation $\det(A - \lambda I_n) = 0$. This amounts to finding the roots of a polynomial of degree n in λ.

The following theorem is a central theorem in spectral theory.

Theorem 2 (Spectral theorem)**.** *If A is symmetric, then there exists an orthogonal matrix P such that $A = P^\intercal DP$, with $D = diag(\lambda_1, \dots, \lambda_n)$. $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A (they are not necessarily all distinct) and they are all real.*

Remark. Remember that a matrix P in $\mathbb{R}^{n\times n}$ is orthogonal iff $P^\intercal P = I_n$. In particular, P is invertible and its inverse is $P^\intercal$. Furthermore, P preserves the norm, that is:

$$\|P\mathbf{x}\| = \|x\|,\ \mathbf{x} \in \mathbb{R}^n$$

To verify that fact, note that $\|P\mathbf{x}\|^2 = (P\mathbf{x})^\intercal(P\mathbf{x}) = \mathbf{x}^\intercal P^\intercal P\mathbf{x} = \mathbf{x}^\intercal\mathbf{x} = \|\mathbf{x}\|^2$.

Corollary 3. *Let A be a symmetric matrix and let us denote by λ_{max} its largest eigenvalue and by λ_{min} its smallest eigenvalue, then we have:*

$$\max_{\|\mathbf{x}\|=1} \mathbf{x}^\intercal A\mathbf{x} = \lambda_{max}$$
$$\min_{\|\mathbf{x}\|=1} \mathbf{x}^\intercal A\mathbf{x} = \lambda_{min}$$

Remark. We deduce from this corollary that A is semi-definite positive iff $\lambda_{min} \geq 0$, *i.e* all the eigenvalues of A are non-negative, and that A is definite positive iff $\lambda_{min} > 0$, *i.e* all its eigenvalues are positive.

Proof. Using the spectral theorem we can write:

$$\mathbf{x}^\intercal A\mathbf{x} = \mathbf{x}^\intercal P^\intercal DP\mathbf{x} = (P\mathbf{x})^\intercal D(Px),\ \mathbf{x} \in \mathbb{R}^n,\ \|x\| = 1$$

since P is invertible an preserves the norm, for any $\mathbf{y}$ such that $\|\mathbf{y}\| = 1$, we can find a unique $\mathbf{x}$ with $\|\mathbf{x}\| = 1$ such that $P\mathbf{x} = \mathbf{y}$. Hence:

$$\max_{\|\mathbf{y}\|=1} \mathbf{y}^\intercal D\mathbf{y} = \max_{\|\mathbf{x}\|=1} \mathbf{x}^\intercal A\mathbf{x}$$

Let us now look at the quantity $\mathbf{y}^\intercal D\mathbf{y}$. By expanding the matrix product, we have:

$$\mathbf{y}^\intercal D\mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2 \leq \lambda_{max} \sum_{i=1}^{n} y_i^2 = \lambda_{max}\|\mathbf{y}\|^2 = \lambda_{max} \tag{1}$$

where the inequality uses that λ_{max} is the largest eigenvalue and where we used that $\|\mathbf{y}\| = 1$. Furthermore, when $\mathbf{y}$ is such that $y_i = 1$ for some i such that $\lambda_i = 1$ and $y_i = 0$ otherwise, we see that the inequality (1) is in fact an equality. This proves:

$$\max_{\|\mathbf{x}\|=1} \mathbf{x}^\intercal A\mathbf{x} = \lambda_{max}$$

The proof for the minimum can be done in a similar way. □

Corollary 4. *Let A be a symmetric matrix and let us denote by λ_{max} (resp. λ_{min}) its largest (resp. smallest) eigenvalue, then:*

$$\lambda_{min}\mathbf{x}^\intercal\mathbf{x} \leq \mathbf{x}^\intercal A\mathbf{x} \leq \lambda_{max}\mathbf{x}^\intercal\mathbf{x},\ \mathbf{x} \in \mathbb{R}^n$$

Proof. Simply apply Corollary 3 to $\frac{\mathbf{x}}{\|\mathbf{x}\|}$. □

Definition 5. *We define a partial order $\preceq$ on matrices by:*

$$A \preceq B \Leftrightarrow B - A \text{ is positive semi-definite}$$

Using the definition of semi-definite positives matrices, this can be reformulated as:

$$A \preceq B \Leftrightarrow \mathbf{x}^\intercal A\mathbf{x} \leq \mathbf{x}^\intercal B\mathbf{x},\ \mathbf{x} \in \mathbb{R}^n$$

3 Additional remarks

3.1 Computing gradients and hessians

Last week, we saw how computing a Taylor expansion is a convenient way to compute gradients and Hessians. A precise formulation of this technique is captured by the following proposition (credits to Weiwei Pan for making this correct)

Proposition 6. *Consider $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$ such that f is twice-differentiable at $\mathbf{x}$. Assume that for all $\mathbf{h} \in \mathbb{R}^n$:*

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \mathbf{h}^\intercal\mathbf{a} + \mathbf{h}^\intercal A\mathbf{h} + o(\|\mathbf{h}\|^2)$$

then $\mathbf{a} = \nabla f(\mathbf{h})$ and $H_f(\mathbf{x}) = \frac{1}{2}(A + A^\intercal)$. In particular, if A is symmetric, $H_f(\mathbf{x}) = A$.

3.2 Strong convexity and convergence rate of gradient descent

We saw in class that a twice-differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex iff there exist $m > 0$ and $M > 0$ such that:

$$mI_n \preceq H_f(\mathbf{x}) \preceq MI_n,\ \mathbf{x} \in \mathbb{R}^n$$

using the definition of the partial order $\preceq$, this is equivalent to:

$$m\mathbf{y}^\intercal\mathbf{y} \leq \mathbf{y}^\intercal H_f(\mathbf{x})\mathbf{y} \leq M\mathbf{y}^\intercal\mathbf{y},\ \mathbf{x} \in \mathbb{R}^n,\ \mathbf{y} \in \mathbb{R}^n$$

Using Corollary 3 this is equivalent to saying that m (resp. M) is a lower (resp. upper) bound on the smallest (resp. largest) eigenvalue of $H_f(\mathbf{x})$ for all $\mathbf{x}$. In particular, $H_f(\mathbf{x})$ needs to be definite positive for all $\mathbf{x}$.

Example. Let us consider again the example function from class:

$$f(\mathbf{x}) = 4x_1^2 - 4x_1x_2 + 2x_2^2$$

We can rewrite:

$$f(\mathbf{x}) = \mathbf{x}^\intercal A\mathbf{x} \quad \text{with} \quad A = \begin{pmatrix} 4 & -2 \\ -2 & 2 \end{pmatrix}$$

Using Proposition 6 it is then easy to see that $H_f(\mathbf{x})$ is constant equal to A. The eigenvalues of A can be computed by solving the equation $\det(A - \lambda I_2) = 0$, which is equivalent to:

$$(4-\lambda)(2-\lambda) - 4 = 0$$

This equation has two solutions $3-\sqrt{5}$, $3+\sqrt{5}$. This implies that f is strongly convex for $m = 3-\sqrt{5}$ and $M = 3+\sqrt{5}$.

Once we know the strong convexity constants m, M of a given function we can use that to compute the convergence rate of the gradient descent algorithm. Remember that we showed in class that at the kth iteration:

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \leq \left(1 - \frac{m}{M}\right)^k \left(f(\mathbf{x}^0) - f(\mathbf{x}^*)\right)$$

For the example above, the term $\left(1 - \frac{m}{M}\right)$ is approximately equal to 0.85. This is the amount by which the error of current solution shrinks at every iteration of the gradient descent algorithm.

When $\frac{m}{M}$ is very close to zero, the convergence will be slow. This is commonly referred to as an *ill-conditioned* problem. Newton's method that we will cover next week uses the Hessian of f to define the step size and circumvent ill-conditioned problems.

3.3 Stopping criterion

In the formulation of gradient descent seen in class, the stopping criterion was written as:

$$\textbf{while } \|\nabla f(\mathbf{x}^{(k)})\| > \varepsilon$$

In practice we would like to know how to set ε such that when the while loop terminates, the error of the solution $f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)$ is smaller than some $\delta > 0$. Fortunately, we have the following bound (seen in class) induced by strong convexity:

Proposition 7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a m-strongly convex function, then:*

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \frac{1}{2m}\|\nabla f(\mathbf{x})\|^2, \ \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n$$

In particular, for $\mathbf{y} = \mathbf{x}^*$, this implies that setting $\varepsilon = \sqrt{2m\delta}$ in the stopping criterion of gradient descent will guarantee that the error of the solution is smaller than δ when the algorithm terminates.