Mathematics for Machine Learning
Problem Set 4
COMS 4995-002 : Summer B 2024

Instructor: Samuel Deng
samdeng@cs.columbia.edu
Due: 08/06/2024 11:59 PM ET.

Problem 1

Proof of the spectral theorem through optimization [30 points].

In this problem, you will prove one of the most important theorems of linear algebra: the *spectral theorem*. Recall that the spectral theorem we stated in class states that, if $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a *symmetric* matrix, then it has an eigendecomposition

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top,$$

where $\mathbf{V}$ is composed of an orthonormal basis of eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$ and $\mathbf{\Lambda}$ is a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_d$. Although we've been using this throughout the course, we haven't proven this result yet, and many introductory linear algebra courses neglect showing a proof. In this problem, you will see that performing successive equality constrained optimization gives you the spectral theorem.

Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a fixed symmetric matrix. For this problem, we will attempt to prove the spectral theorem for $\mathbf{A}$. This problem will heavily rely on the optimization of the associated quadratic form, which we denote $Q_{\mathbf{A}} : \mathbb{R}^d \to \mathbb{R}$,

$$Q_{\mathbf{A}}(\mathbf{x}) := \mathbf{x}^\top \mathbf{A}\mathbf{x}.$$

The overall strategy of this proof is to find the vector that maximizes the quadratic form out of all vectors with norm 1. Then, we will find the next vector that is orthogonal to the first, also constrained to have norm 1. We find a third vector that is orthogonal to the first two with norm 1, repeating this process. The vectors we find will happen to be eigenvectors.

First, consider the constrained optimization problem:

$$\begin{aligned} \text{maximize} \quad & \mathbf{x}^\top \mathbf{A}\mathbf{x} \\ \text{subject to} \quad & \|\mathbf{x}\| = 1. \end{aligned}$$

Maximizing an objective is equivalent to minimizing the negative of the objective, so consider the equivalent problem:

$$\begin{aligned} \text{minimize} \quad & -\mathbf{x}^\top \mathbf{A}\mathbf{x} \\ \text{subject to} \quad & \|\mathbf{x}\|^2 = 1. \end{aligned} \tag{1}$$

Note that, above, requiring $\|\mathbf{x}\|^2 = 1$ is the same as requiring that $\|\mathbf{x}\| = 1$, but the square is more mathematically convenient for our purposes. Suppose that a minimum exists for the above optimization problem presented in (1).[1]

[1] Technically, we should prove that a minimum indeed exists first, by showing that the constraint set is

Problem 1(a) [3 points] The problem in (1) is an constrained optimization problem with equality constraints. Write the Lagrangian $L : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ for this problem, and state clearly what the constraint function $h : \mathbb{R}^d \to \mathbb{R}$ is.

You should have the Lagrangian from Problem 1(a), which now unconstrains the optimization problem. From our techniques in lecture, we now know that finding a minimum for this Lagrangian amounts to first applying the first-order necessary conditions to the unconstrained Lagrangian.

Problem 1(b) [3 points] Using the Lagrangian you found in Problem 1(a), find the gradient $\nabla_{\mathbf{x}} L(\mathbf{x}, \lambda)$ with respect to $\mathbf{x}$. Then, find the gradient $\nabla_{\lambda} L(\mathbf{x}, \lambda)$ with respect to λ. State both these gradients clearly.

Problem 1(c) [3 points] Conclude that the optimal $\mathbf{x}^* \in \mathbb{R}^d$ of the Lagrangian satisfies

$$\mathbf{A}\mathbf{x}^* = \lambda \mathbf{x}^*,$$

where $\|\mathbf{x}^*\| = 1$. Prove that the only non-regular point is $\mathbf{x} = \mathbf{0}$, which implies that $\mathbf{x}^*$ is a global minimum to the optimization problem in (1).

By solving Problem 1(c), you effectively showed the existence of an eigenvector $\mathbf{x}^*$ corresponding to the eigenvalue λ, the unique Lagrange multiplier from solving the constrained optimization problem in (1). Let us now denote this solution $\mathbf{v}_1$ (formerly known as $\mathbf{x}^*$) and the corresponding eigenvalue as λ_1 (formerly known as λ, the Lagrange multiplier).

Now, to find the second eigenvector, consider the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & \mathbf{x}^\top \mathbf{A}\mathbf{x} \\ \text{subject to} \quad & \|\mathbf{x}\| = 1 \\ & \mathbf{x}^\top \mathbf{v}_1 = 0. \end{aligned}$$

This is, again, equivalent to a corresponding minimization problem:

$$\begin{aligned} \text{minimize} \quad & -\mathbf{x}^\top \mathbf{A}\mathbf{x} \\ \text{subject to} \quad & \|\mathbf{x}\|^2 = 1 \\ & \mathbf{x}^\top \mathbf{v}_1 = 0. \end{aligned} \tag{2}$$

That is, our goal is now to find a second eigenvector that is unit length and is orthogonal to the first eigenvector we found, $\mathbf{v}_1$. In the optimization problem above, $\mathbf{v}_1$ is fixed.

compact. We will not deal with compactness because that will require a bit of basic topology, outside the scope of this course.

Problem 1(d) [3 points] The problem in (2) is a constrained optimization problem with 2 equality constraints. Write the Lagrangian $L : \mathbb{R}^d \times \mathbb{R}^2 \to \mathbb{R}$ for this problem, and state clearly what the constraint functions $h_1 : \mathbb{R}^d \to \mathbb{R}$ and $h_2 : \mathbb{R}^d \to \mathbb{R}$ are.

Use variable λ_2 to refer to the *first* Lagrange multiplier (the multiplier corresponding to h_1, the first constraint) and use the variable $\mu_{2,1}$ to refer to the *second* Lagrange multiplier (the multiplier corresponding to h_2, the second constraint).

Now, again use the first-order necessary conditions to solve the Lagrangian in Problem 1(d).

Problem 1(e) [3 points] Using the first order necessary conditions, prove that the optimal minimizer $\mathbf{x}^* \in \mathbb{R}^d$ satisfies:

$$\mathbf{A}\mathbf{x}^* = \lambda_2 \mathbf{x}^* + \frac{\mu_{2,1}}{2}\mathbf{v}_1, \tag{3}$$

while satisfying $\|\mathbf{x}^*\| = 1$ and $\mathbf{v}_1^\top \mathbf{x}^* = 0$.

The equation you found in Problem 1(e) *almost* says that $\mathbf{x}^*$ is an eigenvector with eigenvalue λ_2, but not quite. In order to show this, it would suffice to show that $\mu_{2,1} = 0$.

Problem 1(f) [3 points] Prove that $\mu_{2,1} = 0$ in Equation (3). Conclude that the optimal $\mathbf{x}*$ for the optimization problem in (2) satisfies

$$\mathbf{A}\mathbf{x}^* = \lambda_2 \mathbf{x}^*.$$

Prove that, by how the constraint functions h_1 and h_2 are set up, there are no non-regular points.

Hint: It may be helpful to take the inner product of both sides of Equation (3) by $\mathbf{v}_1$. Then, use the constraints on $\mathbf{x}^*$, particularly $\mathbf{v}_1^\top \mathbf{x}^* = 0$. The assumption that $\mathbf{A}$ is symmetric may also help.

In all, Problem 1(f) gives us some $\mathbf{x}^*$ that satisfies

$$\mathbf{A}\mathbf{x}^* = \lambda_2 \mathbf{x}^*$$

where $\mathbf{v}_1^\top \mathbf{x}^* = 0$. Because $\mathbf{x}^*$ is orthogonal to $\mathbf{v}_1$ and is an eigenvector of $\mathbf{A}$ with eigenvalue λ_2, let us denote it as $\mathbf{v}_2$. Now, we have orthonormal eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ with eigenvalues λ_1 and λ_2, by design.

It should now be clear how to proceed. To find the third eigenvector, consider the following

optimization problem:

$$\begin{aligned} \text{maximize} \quad & \mathbf{x}^\top \mathbf{A}\mathbf{x} \\ \text{subject to} \quad & \|\mathbf{x}\| = 1 \\ & \mathbf{x}^\top \mathbf{v}_1 = 0 \\ & \mathbf{x}^\top \mathbf{v}_2 = 0. \end{aligned} \tag{4}$$

Using the techniques we've established in this problem already, it shouldn't be too hard to find the third eigenvector $\mathbf{v}_3$ and eigenvalue λ_3, in a similar fashion.

Problem 1(g) [8 points] Prove that solving the above optimization problem in (4) yields a third eigenvector $\mathbf{v}_3$ with corresponding eigenvalue λ_3 that is orthogonal to $\mathbf{v}_1$ and $\mathbf{v}_2$ and has unit length, $\|\mathbf{v}_3\| = 1$.

Hint: Set up a Lagrangian and solve it, just as in the previous problems. If we denote the Lagrangian's multipliers $\lambda_1, \mu_{3,1}$, and $\mu_{3,2}$, you should reach a point where you obtain:

$$\mathbf{A}\mathbf{x}^* = \mu_{3,1}\mathbf{v}_1 + \mu_{3,2}\mathbf{v}_2 + \lambda_3 \mathbf{x}^*,$$

and showing that $\mu_{3,1} = \mu_{3,2} = 0$ using the technique from Problem 1(f) should give you your third eigenvector.

Continuing in this way, it should be clear that we can obtain eigenvalues $\lambda_1, \ldots, \lambda_d$ and an orthonormal basis of eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$, proving the spectral theorem.

Finally, we will use the spectral theorem, which we just showed, to prove an important property of quadratic forms. Namely, we will show that the eigenvector corresponding to the largest eigenvalue maximizes quadratic forms over all unit vectors, and, correspondingly, the eigenvector corresponding to the smallest eigenvalue minimizes quadratic forms.

Problem 1(h) [4 points] Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be symmetric. Prove that the optimization problem

$$\begin{aligned} \text{maximize} \quad & \mathbf{x}^\top \mathbf{A}\mathbf{x} \\ \text{subject to} \quad & \|\mathbf{x}\| = 1. \end{aligned}$$

is optimized at $\mathbf{v}_1$, the eigenvector corresponding to the largest eigenvalue λ_1 of $\mathbf{A}$, and the optimal value is λ_1. Also prove that the optimization problem

$$\begin{aligned} \text{minimize} \quad & \mathbf{x}^\top \mathbf{A}\mathbf{x} \\ \text{subject to} \quad & \|\mathbf{x}\| = 1. \end{aligned}$$

is optimized at $\mathbf{v}_d$, the eigenvector corresponding to the smallest eigenvalue λ_d of $\mathbf{A}$, and the optimal value is λ_d. Conclude that:

$$\lambda_d \leq \mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_1 \quad \text{for all } \|\mathbf{x}\| = 1.$$

Hints: This problem does not require you to use the Lagrangian method. Instead, you should use the spectral theorem on $\mathbf{A}$ to obtain $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$. The property of orthogonal matrices that $\|\mathbf{V}^\top\mathbf{x}\| = \|\mathbf{x}\|$ should also help.

This last problem, Problem 1(h), shows us the close connection between quadratic forms, these objects from calculus we've studied extensively, and the eigenvectors and eigenvalues of the matrix $\mathbf{A}$ that defines the quadratic form.

Problem 2

Verifying convexity [21 points total].

In this problem, you will verify a couple of commonly used properties that are helpful in identifying convex functions. Recall that, a function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be *convex* if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) \quad \text{for all } \alpha \in [0,1]. \tag{5}$$

This definition of convexity can be simply stated as: "the line segment between any two points lies above the function." One commonly occurring convex function is the *affine function*, which is simply the class of linear functions plus an offset.

Problem 2(a) [3 points] Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$. Consider any function $f : \mathbb{R}^d \to \mathbb{R}^n$ of the form

$$f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}.$$

Prove that all such functions are convex using the definition above.

Hint: It may be helpful to break up $\mathbf{b}$ into $\alpha\mathbf{b}$ and $(1-\alpha)\mathbf{b}$.

Another commonly considered function is the one that outputs the *norm* of a given vector.

Problem 2(b) [3 points] Consider the function $f : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f(\mathbf{x}) := \|\mathbf{x}\|.$$

Prove that f is convex.

Hint: The triangle inequality for norms you proved in PS1 may help you.

It is also often very useful to combine convex functions together. One common way to combine convex functions together is by summing them with nonnegative coefficients.

Problem 2(c) [3 points] Let $c_1, \ldots, c_n \geq 0$ be positive scalars. Suppose that $f_1, \ldots, f_n$ are convex functions, with $f_i : \mathbb{R}^d \to \mathbb{R}$. Prove that the function $g : \mathbb{R}^d \to \mathbb{R}$ defined by their pointwise sum,

$$g(\mathbf{x}) := \sum_{i=1}^{n} c_i f_i(\mathbf{x})$$

is convex.

Hint: Start with considering $\sum_{i=1}^{n} c_i f_i(\alpha\mathbf{x}+(1-\alpha)\mathbf{y})$ and use the definition of convexity on each f_i.

Another way is to compose convex functions with other convex functions. Two useful convex function composition properties that we won't prove are the following:

Lemma 1. Let $h : \mathbb{R}^d \to \mathbb{R}$ be a convex function and let $g : \mathbb{R} \to \mathbb{R}$ be convex and nondecreasing on the image of h, i.e. $\{y \in \mathbb{R} : h(\mathbf{x}) = y, \text{ for some } \mathbf{x} \in \mathbb{R}^d\}$. Then, the composition function $f(\mathbf{x}) = g(h(\mathbf{x}))$ is convex.

Lemma 2. Let $g : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then, the pre-composition with an affine function $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x} + \mathbf{b})$ is convex.

Problem 2(d) [3 points] Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Prove, using the properties you've already seen in this problem, that the least squares objective function

$$f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

is convex. Also, for $\gamma > 0$, prove that the ridge regression objective function

$$f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \gamma\|\mathbf{w}\|^2$$

is convex. You may use any statements in the previous problems, Lemma 1, and Lemma 2 to prove these claims.

Typically, recognizing that a function is a combination of other simpler convex pieces is enough to verify convexity. However, sometimes it is easier to verify it from the equivalent *first-order definition of convexity* if the function is differentiable. Recall that a function can also be said to be *convex* if, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}). \tag{6}$$

This definition of convexity can be simply stated as: "the function's tangents/linearizations all lie below the function." We stated without proof that this definition of convexity is equivalent to the first definition we stated above. We will prove this claim here.

Problem 2(e) [3 points] First, prove the $\implies$ direction, that (5) implies (6). That is, let $f : \mathbb{R}^d \to \mathbb{R}$ satisfy, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) \quad \text{for all } \alpha \in [0, 1].$$

Prove that:

$$f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}).$$

Hints: Note that:

$$f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) = f(\mathbf{x} + \alpha(\mathbf{y} - \mathbf{x})).$$

Does this look familiar, perhaps in the form of the directional derivative? Rearrange

the right-hand side so we get something that looks like a directional derivative. Take the limit as $\alpha \to 0$ to obtain the result.

Problem 2(f) [3 points] Now, prove the $\Longleftarrow$ direction, that (6) implies (5). That is, let $f : \mathbb{R}^d \to \mathbb{R}$ satisfy, for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$,

$$f(\mathbf{v}) + \nabla f(\mathbf{v})^\top (\mathbf{u} - \mathbf{v}) \leq f(\mathbf{u}).$$

Prove that, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) \quad \text{for all } \alpha \in [0,1].$$

Hints: The first-order definition applies for any choices of $\mathbf{u}$ and $\mathbf{v}$. In the statement of Definition (5), let $\alpha \in [0,1]$ be fixed and let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Consider

$$\mathbf{z} = (1-\alpha)\mathbf{x} + \alpha\mathbf{y},$$

a point on the line segment in between $\mathbf{x}$ and $\mathbf{y}$. Apply the first-order definition (6) first to the pair $\mathbf{v} = \mathbf{z}$ and $\mathbf{u} = \mathbf{x}$. Then apply (6) to the pair $\mathbf{v} = \mathbf{z}$ and $\mathbf{u} = \mathbf{y}$. This gives you two inequalities; try to combine them to obtain (5).

Problem 2(e) and Problem 2(f) now gives you another way to verify whether a function is convex, as long as it's differentiable. We won't prove this here, but, in lecture, we saw a third characterization of convexity for twice-differentiable functions. A twice-differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be *convex* if its Hessian $\nabla^2 f(\mathbf{x})$ at all points $\mathbf{x} \in \mathbb{R}^d$ is positive semidefinite. We call this the *second-order definition of convexity.*

Problem 2(g) [3 points] Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$ be fixed. Using the first-order definition of convexity, prove that the least squares objective

$$f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

is convex. Also use the second-order definition of convexity to prove that it is convex.

Problem 3

OLS-specific gradient descent [24 points total].

In this problem, we will analyze the properties of gradient descent applied specifically to the least squares optimization problem. Throughout this problem, let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$ be fixed. Also, suppose that $n \geq d$ and $\text{rank}(\mathbf{X}) = d$ for simplicity; that is, assume that $\mathbf{X}$ is full-rank. In this case, a minimizer surely exists, and we already know from the first lecture that it takes the form:

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Recall from lecture that, for general *convex* and β-smooth functions, gradient descent with learning rate $\eta = 1/\beta$ guaranteed:

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{\beta}{2T} \left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_T - \mathbf{x}^*\|^2 \right).$$

We saw that the least squares objective

$$f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

is convex and β-smooth, where $\lambda_{\max}(2\mathbf{X}^\top \mathbf{X}) = \beta$, so as a corollary, we are have the guarantee:

$$\|\mathbf{X}\mathbf{w}_T - \mathbf{y}\|^2 - \|\mathbf{X}\mathbf{w}^* - \mathbf{y}\|^2 \leq \frac{\beta}{2T} (\|\mathbf{w}_0 - \mathbf{w}^*\|^2 - \|\mathbf{w}_T - \mathbf{w}^*\|^2).$$

However, this result was a corollary for the general case of convex and β-smooth functions, which is a much broader class of functions than the least squares objective. If we already know that we're specifically performing gradient descent on the least squares objective, can we analyze the behavior of gradient descent directly?

It turns out that we can. We will be proving a guarantee that is a slightly different flavor from the one in lecture for this problem. Whereas the guarantee on gradient descent in lecture proved convergence in the function values $f(\mathbf{w}_T)$ and $f(\mathbf{w}^*)$, we will be proving convergence in distance to the minimizer in the input space. That is, we will bound $\|\mathbf{w}_T - \mathbf{w}^*\|$.

We've seen time and again that the gradient and Hessian of the least squares objective are:

$$\nabla f(\mathbf{w}) = 2(\mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y}) \tag{7}$$

$$\nabla^2 f(\mathbf{w}) = 2\mathbf{X}^\top \mathbf{X}. \tag{8}$$

Notice that the Hessian in Equation (8) does not depend on $\mathbf{w}$, so for notational convenience, we will denote the Hessian $[\nabla^2 f]$ throughout this problem. For some step size $\eta > 0$, we know that the gradient descent update step looks like:

$$\mathbf{w}_t = \mathbf{w}_{t-1} - 2\eta(\mathbf{X}^\top \mathbf{X}\mathbf{w}_{t-1} - \mathbf{X}^\top \mathbf{y}). \tag{9}$$

Problem 3(a) [4 points] Denote the global minimizer $\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$. Prove that we can rewrite the gradient update step in Equation (9) as:

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta[\nabla^2 f](\mathbf{w}_{t-1} - \mathbf{w}^*).$$

Hint: Notice that we can rewrite $\mathbf{X}^\top \mathbf{y}$ as $(\mathbf{X}^\top \mathbf{X})(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$.

Problem 3(a) is quite helpful because it introduces $\mathbf{w}^*$, the minimum value. This is helpful because it'll help us track how far we are from the minimizer, in terms of $\mathbf{w}_t$. Let us denote the integer T as the total number of steps we run gradient descent for.

Problem 3(b) [4 points] Prove the following expression for any step $1 \leq t \leq T$,

$$\mathbf{w}_t - \mathbf{w}^* = (\mathbf{I} - \eta[\nabla^2 f])(\mathbf{w}_{t-1} - \mathbf{w}^*),$$

where $\mathbf{I}$ is the $d \times d$ identity matrix. Use this expression to conclude that, for any integer $T \geq 1$,

$$\mathbf{w}_T - \mathbf{w}^* = (\mathbf{I} - \eta[\nabla^2 f])^T(\mathbf{w}_0 - \mathbf{w}^*).$$

Hint: Subtract $\mathbf{w}^*$ from both sides of the expression in Problem 3(a).

Problem 3(b) almost has the quantity we want to track, but we are interested in the distance, $\|\mathbf{w}_T - \mathbf{w}^*\|$, not the vector $\mathbf{w}_T - \mathbf{w}^*$ itself.

Problem 3(c) [4 points] Prove the following expression for any integer $T \geq 1$:

$$\|\mathbf{w}_T - \mathbf{w}^*\|^2 \leq (\mathbf{w}_0 - \mathbf{w}^*)^\top (\mathbf{I} - \eta[\nabla^2 f])^{2T}(\mathbf{w}_0 - \mathbf{w}^*). \tag{10}$$

Also, prove that for any symmetric matrix $\mathbf{A}$, the matrix $(\mathbf{I} - \mathbf{A})^k$ is symmetric for any $k \geq 1$. Conclude that the function

$$f(\mathbf{v}) = \mathbf{v}^\top (\mathbf{I} - \eta[\nabla^2 f])^{2T}\mathbf{v}$$

is a quadratic form.

We are now again dealing with a quadratic form. In order to analyze quadratic forms, we know from the end of Problem 1 (specifically, Problem 1(h)), that it may help to analyze the eigenvalues of the underlying symmetric matrix.

Problem 3(d) [4 points] Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a symmetric matrix and consider the matrix $(\mathbf{I} - \mathbf{A})^k$ for integer $k \geq 1$. Prove that if λ is an eigenvalue of $\mathbf{A}$, then $(1-\lambda)^k$ is an eigenvalue of $(\mathbf{I} - \mathbf{A})^k$. Conclude that if λ is an eigenvalue of $[\nabla^2 f]$, then $(1 - \eta\lambda)^{2T}$

is an eigenvalue of the matrix $(\mathbf{I} - \eta[\nabla^2 f])^{2T}$.

From Problem 3(d), we have obtained an expression for the eigenvalues of the somewhat messy-looking matrix $(\mathbf{I} - \eta[\nabla^2 f])^{2T}$. Recall from Problem 1(h) that, for any symmetric matrix $\mathbf{A}$, the quadratic form is bounded by

$$\lambda_{\min}(\mathbf{A}) \leq \mathbf{v}^\top \mathbf{A}\mathbf{v} \leq \lambda_{\max}(\mathbf{A}) \quad \text{for all } \|\mathbf{v}\| = 1,$$

where $\lambda_{\max}(\mathbf{A})$ is the largest eigenvalue of $\mathbf{A}$ and $\lambda_{\min}(\mathbf{A})$ is the smallest eigenvalue. In our theorem on gradient descent for β-smooth functions from lecture, we let

$$\beta = \lambda_{\max}(2\mathbf{X}^\top \mathbf{X}) = \lambda_{\max}([\nabla^2 f])$$

and chose the learning rate $\eta = 1/\beta$. We will use this same learning rate, $\eta = 1/\beta$. For simplicity of notation, we refer to $\lambda_{\min} := \lambda_{\min}([\nabla^2 f])$ and $\lambda_{\max} := \lambda_{\max}([\nabla^2 f])$. Using this notation, the learning rate is $\eta = 1/\beta = 1/\lambda_{\max}$.

Problem 3(e) [4 points] Let λ be any eigenvalue of $[\nabla^2 f]$. Show that, setting $\eta = 1/\beta$, we can bound

$$(1 - \eta\lambda)^{2T} \leq \left(1 - \frac{\lambda_{\min}}{\lambda_{\max}}\right)^{2T}.$$

Finally, we will use these facts we've accumulated to prove our result.

Problem 3(f) [4 points] Using Problems 3(c), 3(d), and 3(e), show that Equation (10) yields the inequality:

$$\|\mathbf{w}_T - \mathbf{w}^*\|^2 \leq \left(1 - \frac{\lambda_{\min}}{\lambda_{\max}}\right)^{2T} \|\mathbf{w}_0 - \mathbf{w}^*\|^2.$$

Hint: Divide the vectors $\mathbf{w}_0 - \mathbf{w}^*$ in Equation (10) by their norm, $\|\mathbf{w}_0 - \mathbf{w}^*\|$ and put the norm back by multiplying by $\|\mathbf{w}_0 - \mathbf{w}^*\|^2$ again. You may find Problem 1(h) helpful (but possibly not necessary).

Sometimes, the ratio $\kappa := \frac{\lambda_{\max}}{\lambda_{\min}}$ is referred to as the "condition number" of a symmetric matrix. Then, using the famous inequality $1 + x \leq e^x$ for all $x \in \mathbb{R}$ (optional exercise: you can prove this using the first-order definition of convex functions!), we can conclude that:

$$\|\mathbf{w}_T - \mathbf{w}^*\|^2 \leq e^{-2T/\kappa} \|\mathbf{w}_0 - \mathbf{w}^*\|^2, \tag{11}$$

which says that our distance to the minimizer $\mathbf{w}^*$ actually decreases *exponentially* fast in T, all else held constant. This gives a slightly different flavor of guarantee than the one in class.

Programming Part

Gradient descent and OLS (25 points total). In this problem, you will apply gradient descent to OLS and verify that it corresponds to the analytical solution we've seen all semester.

In order to start this programming part, download the file `ps4.ipynb` from Course Content on the course webpage. Your submission for this part will be the same `ps4.ipynb` file modified with your code; see HW Submission on the course webpage for additional instructions.