CS 726: Constrained, Projection-Based, Optimization

Jelena Diakonikolas

Fall 2022

In this lecture note, we discuss how what we have learned so far can be generalized to constrained settings, where we assume that computing projections onto the feasible set $\mathcal{X}$ can be done efficiently (and we provide some examples for which this is true).

1 Optimality Conditions, Revisited

We begin by describing the setup we will be working with. Recall that our basic optimization problem is

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}). \tag{P}$$

For this lecture, we will additionally be assuming that:

- The norm of the space is Euclidean, $\|\cdot\| = \|\cdot\|_2$;
- f is L-smooth w.r.t. $\|\cdot\|_2$;
- $\mathcal{X}$ is closed and convex. For the algorithms that we describe, we will additionally assume that Euclidean projection onto $\mathcal{X}$ (defined in this lecture note) is efficiently computable.

Important: because here we are working with constrained optimization settings, the gradient at an optimal solution $\mathbf{x}^*$ need not be zero. For example, consider minimizing a univariate linear function $f(x) = 3x$ on the interval $[1, 2]$. This function is minimized at $x^* = 1$ and the gradient is equal to 3 at any point, including the solution $x^* = 1$.

To talk about optimality conditions and understand what is geometrically happening around (local) solutions, it is useful to define normal cones.

Definition 1.1 (Normal Cone)**.** Let $\mathcal{X}$ be a closed convex set. At any point $\mathbf{x} \in \mathcal{X}$, the normal cone $N_{\mathcal{X}}(\mathbf{x})$ is defined by

$$N_{\mathcal{X}}(\mathbf{x}) = \Big\{\mathbf{z} \in \mathbb{R}^d : \langle \mathbf{z}, \mathbf{y} - \mathbf{x}\rangle \le 0,\ \forall \mathbf{y} \in \mathcal{X}\Big\}. \tag{1}$$

Observe that, by definition, a normal cone is indeed a cone: if $\mathbf{z} \in N_{\mathcal{X}}(\mathbf{x})$ then also $t\mathbf{z} \in N_{\mathcal{X}}(\mathbf{x})$, for all $t > 0$.

Normal cone characterizes local solutions. In particular, if for some $\mathbf{x}^* \in \mathcal{X}$ we have that $-\nabla f(\mathbf{x}^*) \in N_{\mathcal{X}}(\mathbf{x}^*)$, then, equivalently (by the definition of $N_{\mathcal{X}}(\mathbf{x}^*)$), for all $\mathbf{y} \in \mathcal{X}$,

$$\langle -\nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^*\rangle \le 0 \quad \Leftrightarrow \quad \langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^*\rangle \ge 0. \tag{2}$$

But (2) is precisely our definition of stationarity from the first lecture note! In particular, we can summarize what we have already proved about points $\mathbf{x}^*$ that satisfy (2) (equivalent to $-\nabla f(\mathbf{x}^*) \in N_{\mathcal{X}}(\mathbf{x}^*)$).

Theorem 1.2. *Given (P), if $\mathbf{x}^* \in \mathcal{X}$ is a local solution to (P), then $-\nabla f(\mathbf{x}^*) \in N_{\mathcal{X}}(\mathbf{x}^*)$. If f is convex, then $-\nabla f(\mathbf{x}^*) \in N_{\mathcal{X}}(\mathbf{x}^*)$ if and only if $\mathbf{x}^*$ is a* global *solution to (P). Additionally, if f is strongly convex, then $\mathbf{x}^*$ such that $-\nabla f(\mathbf{x}^*) \in N_{\mathcal{X}}(\mathbf{x}^*)$ exists and is the unique solution to (P).*

2 Euclidean Projection

Euclidean projection of a point $\mathbf{x}$ onto a set $\mathcal{X}$ is also known as the orthogonal projection and it corresponds to choosing the point form $\mathcal{X}$ that is closest to $\mathbf{x}$. Formally, given a point $\mathbf{x} \in \mathbb{R}^d$, the Euclidean projection is defined by

$$P_{\mathcal{X}}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{y} \in \mathcal{X}} \|\mathbf{y} - \mathbf{x}\|_2. \tag{3}$$

We can further characterize Euclidean projections using the following lemma.

Lemma 2.1. *Given a closed convex set $\mathcal{X}$, for any $\mathbf{x} \in \mathbb{R}^d$, $P_{\mathcal{X}}(\mathbf{x})$ exists and is unique. It is further characterized by $-(P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}) \in N_{\mathcal{X}}(\mathbf{x})$, and, as a consequence, by the following inequality:*

$$(\forall \mathbf{y} \in \mathcal{X}) : \quad \langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}, \mathbf{y} - P_{\mathcal{X}}(\mathbf{x}) \rangle \geq 0. \tag{4}$$

Proof. Observe that, given $\mathbf{x} \in \mathbb{R}^d$, the optimization problems $\min_{\mathbf{y} \in \mathcal{X}} \|\mathbf{y} - \mathbf{x}\|_2$ and $\min_{\mathbf{y} \in \mathcal{X}} \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$ have the same solutions, thus both define $P_{\mathcal{X}}(\mathbf{x})$. The latter problem has a strongly convex objective, which (by results proved in the first lecture note) implies that $P_{\mathcal{X}}(\mathbf{x})$ is attained and unique.

For the remaining claims, the gradient of $\frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$ at $P_{\mathcal{X}}(\mathbf{x})$ equals $P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}$. As $P_{\mathcal{X}}(\mathbf{x})$ minimizes $\frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$ over $\mathbf{y} - \mathcal{X}$, we have $-(P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}) \in N_{\mathcal{X}}(\mathbf{x})$, by Theorem 1.2. The remaining inequality is equivalent to $-(P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}) \in N_{\mathcal{X}}(\mathbf{x})$, by the definition of a normal cone (see (2)). □

We now provide some examples of sets $\mathcal{X}$ for which it is easy to compute Euclidean projections.

For the first three examples, proving that the stated expressions are indeed Euclidean projections for the sets in question can be done by verifying that the stationarity condition (4) holds, in light of Lemma 2.1 and Theorem 1.2.

Example 2.2. For the non-negative orthant $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} \geq \mathbf{0}\}$, where '$\geq'$ is applied element-wise, the Euclidean projection for any point $\mathbf{x} \in \mathbb{R}^d$ is given by

$$P_{\mathcal{X}}(\mathbf{x}) = \max\{\mathbf{x}, \, \mathbf{0}\},$$

where the max operation is applied element-wise.

Example 2.3. Given vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ such that $\mathbf{a} \leq \mathbf{b}$ element-wise, the hyperrectangle corresponding to these vectors is $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}\}$, where, as before, '$\leq'$ is applied element-wise. Euclidean projection on $\mathcal{X}$ in this case is given by

$$P_{\mathcal{X}}(\mathbf{x}) = \max\{\mathbf{a}, \, \min\{\mathbf{x}, \, \mathbf{b}\}\}.$$

Example 2.4. For a unit centered Euclidean ball, $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$, the projection operator is given by

$$P_{\mathcal{X}}(\mathbf{x}) = \begin{cases} \mathbf{x}, & \text{if } \mathbf{x} \in \mathcal{X}, \\ \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, & \text{if } \mathbf{x} \notin \mathcal{X}. \end{cases}$$

How would you compute a Euclidean projection for a ball that is not centered (at $\mathbf{0}$)? How would you compute it if the radius were $R > 0$, $R \neq 1$?

Finally, the Euclidean projection can be computed efficiently for the following sets (details omitted, but code implementing projections can be found online).

Example 2.5. For the ℓ_1 ball $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_1 \leq 1\}$ and the probability simplex $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} \geq \mathbf{0}, \langle \mathbf{1}, \mathbf{x} \rangle = 1\}$, the projection operator $P_{\mathcal{X}}(\mathbf{x})$ can be computed efficiently, in $O(d \log(d))$ time.

We now provide some useful results that further characterize Euclidean projections and will be useful when characterizing gradient mapping (in the next section).

Lemma 2.6. *Let $\mathcal{X}$ be a closed convex set. Then $\langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle \geq 0$, $\forall \mathbf{y} \in \mathcal{X}$, with equality if and only if $\mathbf{y} = P_{\mathcal{X}}(\mathbf{x})$.*

Proof. To prove the lemma, we add and subtract $P_{\mathcal{X}}(\mathbf{x})$ within the inner product term of interest, so that we can make use of Inequality (4) from Lemma 2.1. In particular,

$$\begin{aligned}\langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{y}, \mathbf{x} - \mathbf{y}\rangle = \; & \langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{y}, \mathbf{x} - P_{\mathcal{X}}(\mathbf{x})\rangle + \langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{y}, P_{\mathcal{X}}(\mathbf{x}) - \mathbf{y}\rangle \\ & \langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{y}, \mathbf{x} - P_{\mathcal{X}}(\mathbf{x})\rangle + \|P_{\mathcal{X}}(\mathbf{x}) - \mathbf{y}\|_2^2.\end{aligned}$$

In the last expression, $\langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{y}, \mathbf{x} - P_{\mathcal{X}}(\mathbf{x})\rangle \geq 0$ by Lemma 2.1, and if $P_{\mathcal{X}}(\mathbf{x}) = \mathbf{y}$, then $\langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{y}, \mathbf{x} - P_{\mathcal{X}}(\mathbf{x})\rangle = 0$. The last term, $\|P_{\mathcal{X}}(\mathbf{x}) - \mathbf{y}\|_2^2$, is always non-negative, and equal to zero if and only if $P_{\mathcal{X}}(\mathbf{x}) = \mathbf{y}$. Hence, $\langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{y}, \mathbf{x} - \mathbf{y}\rangle \geq 0$, with equality if and only if $P_{\mathcal{X}}(\mathbf{x}) = \mathbf{y}$. □

Lemma 2.7. *Let $\mathcal{X} \in \mathbb{R}^d$ be a closed convex set. Then $P_{\mathcal{X}}$ is a non-expansive operator, that is,*

$$(\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d): \quad \|P_{\mathcal{X}}(\mathbf{x}) - P_{\mathcal{X}}(\mathbf{y})\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2.$$

Proof. There are two main approaches that come to mind when trying to prove expressions such as $\|P_{\mathcal{X}}(\mathbf{x}) - P_{\mathcal{X}}(\mathbf{y})\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2$. The first is to use triangle inequality. Unfortunately, it is unclear how to make this work, as all the terms you would end up with would be non-negative. Another approach would be to instead look at $\|P_{\mathcal{X}}(\mathbf{x}) - P_{\mathcal{X}}(\mathbf{y})\|_2^2$, add and subtract $\mathbf{x} - \mathbf{y}$ inside the norm, and then expand the square. This approach seems more promising, as it does not require using any inequalities right at the beginning, so we do not "lose" anything in the process. This is the approach we take here. In particular,

$$\begin{aligned}\|P_{\mathcal{X}}(\mathbf{x}) - P_{\mathcal{X}}(\mathbf{y})\|_2^2 &= \|P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x} - (P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}) + \mathbf{x} - \mathbf{y}\|_2^2 \\ &= \|\mathbf{x} - \mathbf{y}\|^2 + \|P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}\|_2^2 + \|P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}\|_2^2 \\ &\quad + 2\langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}, \mathbf{x} - \mathbf{y}\rangle - 2\langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}, P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}\rangle - 2\langle P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}, \mathbf{x} - \mathbf{y}\rangle \\ &= \|\mathbf{x} - \mathbf{y}\|^2 + \|P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}\|_2^2 + \|P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}\|_2^2 \\ &\quad + 2\langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}, \mathbf{x} - P_{\mathcal{X}}(\mathbf{y})\rangle - 2\langle P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}, \mathbf{x} - \mathbf{y}\rangle.\end{aligned}$$

Now, observe that the inner product terms in the last expression are "easy" to complete to negative quadratic terms by adding and subtracting $P_{\mathcal{X}}(\mathbf{x})$ in the first term and $P_{\mathcal{X}}(\mathbf{y})$ in the second one. Doing so leads to

$$\begin{aligned}\|P_{\mathcal{X}}(\mathbf{x}) - P_{\mathcal{X}}(\mathbf{y})\|_2^2 = \; & \|\mathbf{x} - \mathbf{y}\|_2^2 - \|P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}\|_2^2 - \|P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}\|_2^2 \\ & + 2\langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}, P_{\mathcal{X}}(\mathbf{x}) - P_{\mathcal{X}}(\mathbf{y})\rangle - 2\langle P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}, \mathbf{x} - P_{\mathcal{X}}(\mathbf{y})\rangle \\ & \|\mathbf{x} - \mathbf{y}\|_2^2 - \|P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}\|_2^2 - \|P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}\|_2^2 \\ & + 2\langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}, P_{\mathcal{X}}(\mathbf{x}) - P_{\mathcal{X}}(\mathbf{y})\rangle - 2\langle P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}, P_{\mathcal{X}}(\mathbf{x}) - P_{\mathcal{X}}(\mathbf{y})\rangle - 2\langle P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}, \mathbf{x} - P_{\mathcal{X}}(\mathbf{x})\rangle.\end{aligned}$$

We can now complete the term by observing that $-\|P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}\|_2^2 - \|P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}\|_2^2 - 2\langle P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}, \mathbf{x} - P_{\mathcal{X}}(\mathbf{x})\rangle \leq 0$ and both

$$\langle P_{\mathcal{X}}(\mathbf{x}) - \mathbf{x}, P_{\mathcal{X}}(\mathbf{x}) - P_{\mathcal{X}}(\mathbf{y})\rangle \leq 0 \;\text{ and }\; \langle P_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}, P_{\mathcal{X}}(\mathbf{x}) - P_{\mathcal{X}}(\mathbf{y})\rangle \geq 0$$

due to Lemma 2.1. □

3 Projected Gradient Descent and Gradient Mapping

To motivate gradient mapping (and projected gradient descent – PGD), we go back to one of the two ways we used to arrive at gradient descent algorithm (in unconstrained settings). In particular, assuming that our function were smooth and starting from some point $\mathbf{x}_k$, smoothness tells us that

$$(\forall \mathbf{y} \in \mathbb{R}^d): \quad f(\mathbf{y}) \leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{y} - \mathbf{x}_k\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}_k\|_2^2. \tag{5}$$

We have seen that we can arrive at gradient descent with step size $\frac{1}{L}$ by simply minimizing the right-hand side of (5) (GD with other step sizes can be obtained by observing that the same inequality holds if L is replaced by any number larger than L):

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{y} \in \mathbb{R}^d} \left\{ f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{y} - \mathbf{x}_k\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}_k\|_2^2 \right\} = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k). \tag{GD}$$

Of course, if our optimization problem is constrained, in general we get no guarantees from GD, as it may step outside the feasible set $\mathcal{X}$. Thus a reasonable question to ask is whether we would get a method that behaves "similar to GD" (i.e., for which we can get similar convergence guarantees) by minimizing the right-hand side of (5) over the feasible set $\mathcal{X}$. Clearly, by doing so, we could guarantee that $\mathbf{x}_{k+1} \in \mathcal{X}$ (so no need to worry about feasibility) and that (assuming $\mathbf{x}_k \in \mathcal{X}$) $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$, so we never increase the function value. Now let us look more closely at what the resulting step would be:

$$\begin{aligned}
\mathbf{x}_{k+1} &= \operatorname*{argmin}_{\mathbf{y}\in\mathcal{X}} \Big\{ f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{y} - \mathbf{x}_k\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}_k\|_2^2 \Big\} \\
&= \operatorname*{argmin}_{\mathbf{y}\in\mathcal{X}} \Big\{ \frac{L}{2}\|\mathbf{y} - \mathbf{x}_k + \frac{1}{L}\nabla f(\mathbf{x}_k)\|_2^2 \Big\} \qquad \text{(PGD}_\text{L}) \\
&= P_{\mathcal{X}}\Big(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\Big).
\end{aligned}$$

Since this method is projecting gradient descent step $\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$ back to the feasible set $\mathcal{X}$, it is known as the projected gradient descent (PGD). Of course, similar to GD, it makes sense to also take step sizes different than $\frac{1}{L}$; we would however usually considered step sizes that are at most $\frac{1}{L}$.

Further, we can write this method in a form that reveals what the "descent" direction that replaces the negative gradient is in this case. In particular, we have that

$$\begin{aligned}
\mathbf{x}_{k+1} &= \mathbf{x}_k - \mathbf{x}_k + P_{\mathcal{X}}\Big(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\Big) \\
&= \mathbf{x}_k - \frac{1}{L}L\Big(\mathbf{x}_k - P_{\mathcal{X}}\Big(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\Big)\Big).
\end{aligned}$$

The quantity $L\Big(\mathbf{x}_k - P_{\mathcal{X}}\Big(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\Big)$ that replaces the gradient when going from gradient descent to projected gradient descent has a name—it is called gradient mapping—and it is useful when trying to extend the analysis from unconstrained to constrained settings. Observe that this quantity is precisely the gradient at $\mathbf{x}_k$ in unconstrained settings. We formally define gradient mapping below and then prove that it leads to similar properties we had from using gradient instead; that is, we can view it as a proxy for the gradient: (i) when its norm is small, we get a near-stationary point (if zero, then the point is stationary) and (ii) taking a step in the direction of the gradient mapping leads to a sufficient descent condition, similar to the one we had for GD.

Definition 3.1. Given a closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$, a point $\mathbf{x} \in \mathbb{R}^d$, a differentiable function f, and a constant $\eta > 0$, gradient mapping (w.r.t. f and $\mathcal{X}$, at point $\mathbf{x}$, and with constant η) is defined by

$$G_\eta(\mathbf{x}) = \eta\Big(\mathbf{x} - P_{\mathcal{X}}\Big(\mathbf{x} - \frac{1}{\eta}\nabla f(\mathbf{x})\Big).$$

Observe that this definition allows us to write PGD with more general step sizes as

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{\eta}G_\eta(\mathbf{x}_k). \qquad \text{(PGD)}$$

Lemma 3.2. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a closed convex set and consider minimizing a function $f : \mathbb{R}^d \to \mathbb{R}$ that is L-smooth on $\mathcal{X}$ over $\mathcal{X}$. Denote $\bar{\mathbf{x}} = P_{\mathcal{X}}(\mathbf{x} - \frac{1}{\eta}\nabla f(\mathbf{x}))$ (so that $G_\eta(\mathbf{x}) = \eta(\mathbf{x} - \bar{\mathbf{x}})$). If, for some $\epsilon > 0$, $\|G_\eta(\mathbf{x})\|_2 \leq \epsilon$, then*

$$\begin{aligned}
&-\nabla f(\bar{\mathbf{x}}) \in N_{\mathcal{X}}(\bar{\mathbf{x}}) + \mathcal{B}(\boldsymbol{0}, \epsilon(L/\eta + 1)) \\
\Leftrightarrow\ &(\forall \mathbf{y} \in \mathcal{X}) : \quad \langle \nabla f(\bar{\mathbf{x}}), \mathbf{y} - \bar{\mathbf{x}}\rangle \geq -\epsilon\Big(\frac{L}{\eta} + 1\Big)\|\mathbf{y} - \bar{\mathbf{x}}\|_2,
\end{aligned}$$

where $\mathcal{B}(\mathbf{a}, r)$ denotes the Euclidean ball of radius $r > 0$, centered at $\mathbf{a} \in \mathbb{R}^d$. As a consequence,

$$(\forall \mathbf{y} \in \mathcal{X} \cap \mathcal{B}(\bar{\mathbf{x}}, 1)) : \quad \langle \nabla f(\bar{\mathbf{x}}), \mathbf{y} - \bar{\mathbf{x}}\rangle \geq -\epsilon\Big(\frac{L}{\eta} + 1\Big). \qquad (6)$$

In particular, if $G_\eta(\mathbf{x}) = 0$, then $\mathbf{x} = \bar{\mathbf{x}}$ and

$$-\nabla f(\mathbf{x}) \in N_{\mathcal{X}}.$$

Proof. Suppose that $\|G_\eta(\mathbf{x})\|_2 \leq \epsilon$. From the definition of $\bar{\mathbf{x}}$, we have that

$$\bar{\mathbf{x}} = P_{\mathcal{X}}\Big(\mathbf{x} - \frac{1}{\eta}\nabla f(\mathbf{x})\Big) = \operatorname*{argmin}_{\mathbf{y}\in\mathcal{X}} \Big\{\frac{1}{2}\Big\|\mathbf{u} - \mathbf{x} + \frac{1}{\eta}\nabla f(\mathbf{x})\Big\|_2^2\Big\}.$$

Hence, by Theorem 1.2, we have

$$-\Big(\bar{\mathbf{x}} - \mathbf{x} + \frac{1}{\eta}\nabla f(\mathbf{x})\Big) \in N_{\mathcal{X}}(\bar{\mathbf{x}}). \tag{7}$$

From the definition of $G_\eta(\mathbf{x})$, if $G_\eta(\mathbf{x}) = 0$, then $\bar{\mathbf{x}} = \mathbf{x}$ and (7) reduces to $-\nabla f(\mathbf{x}) \in N_{\mathcal{X}}(\mathbf{x})$, which proves the last claim from the statement of the lemma. Otherwise, it is useful to express (7) as $-\nabla f(\bar{\mathbf{x}}) - \mathbf{e} \in N_{\mathcal{X}}(\bar{\mathbf{x}})$ where $\mathbf{e}$ can be seen as the "error" vector violating the stationarity condition $-\nabla f(\bar{\mathbf{x}}) \in N_{\mathcal{X}}(\bar{\mathbf{x}})$. Adding and subtracting $-\frac{1}{n}\nabla f(\bar{\mathbf{x}})$ from the left-hand side of Eq. (7), we now have

$$\begin{aligned} & -\frac{1}{\eta}\nabla f(\bar{\mathbf{x}}) + \frac{1}{\eta}\nabla f(\bar{\mathbf{x}}) - \Big(\bar{\mathbf{x}} - \mathbf{x} + \frac{1}{\eta}\nabla f(\mathbf{x})\Big) \in N_{\mathcal{X}}(\bar{\mathbf{x}}) \\ \Leftrightarrow\ & -\nabla f(\bar{\mathbf{x}}) - \Big(\nabla f(\mathbf{x}) - \nabla f(\bar{\mathbf{x}}) + \eta(\bar{\mathbf{x}} - \mathbf{x})\Big) \in N_{\mathcal{X}}(\bar{\mathbf{x}}) \end{aligned}$$

Thus, the "error" vector is

$$\begin{aligned} \mathbf{e} &= \nabla f(\mathbf{x}) - \nabla f(\bar{\mathbf{x}}) + \eta(\bar{\mathbf{x}} - \mathbf{x}) \\ &= \nabla f(\mathbf{x}) - \nabla f(\bar{\mathbf{x}}) + \eta G_\eta(\mathbf{x}). \end{aligned}$$

Using smoothness of f and the assumption that $\|G_\eta\|_2 \leq \epsilon$, we further have

$$\|\mathbf{e}\|_2 \leq L\|\mathbf{x} - \mathbf{x}_b\|_2 + \|G_\eta(\mathbf{x})\|_2 = \Big(\frac{L}{\eta} + 1\Big)\|G_\eta(\mathbf{x})\|_2 \leq \Big(\frac{L}{\eta} + 1\Big)\epsilon.$$

Hence, we can conclude that

$$-\nabla f(\bar{\mathbf{x}}) \in N_{\mathcal{X}}(\bar{\mathbf{x}}) + \mathbf{e} \subset N_{\mathcal{X}}(\bar{\mathbf{x}}) + \mathcal{B}\Big(\mathbf{0}, \Big(\frac{L}{\eta} + 1\Big)\epsilon\Big),$$

as $\mathbf{e} \in \mathcal{B}\Big(\mathbf{0}, \Big(\frac{L}{\eta} + 1\Big)\epsilon\Big)$. The remaining claims follow directly from this result, using the definition of a normal cone. □

The next lemma can be interpreted as a "descent lemma" for PGD.

Lemma 3.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth function and let $\mathcal{X}$ be a closed convex set. Let*

$$\bar{\mathbf{x}} = P_{\mathcal{X}}\Big(\mathbf{x} - \frac{1}{\eta}\nabla f(\mathbf{x})\Big),$$

where $\mathbf{x} \in \mathcal{X}$, and observe that $G_\eta(\mathbf{x}) = \eta(\mathbf{x} - \bar{\mathbf{x}})$. If $\eta \geq L$, then

$$f(\bar{\mathbf{x}}) \leq f(\mathbf{x}) - \frac{1}{2\eta}\|G_\eta(\mathbf{x})\|_2^2.$$

Proof. As f is L-smooth and $\eta \geq L > 0$, we have

$$\begin{aligned} f(\bar{\mathbf{x}}) &\leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \bar{\mathbf{x}} - \mathbf{x}\rangle + \frac{L}{2}\|\bar{\mathbf{x}} - \mathbf{x}\|_2^2 \\ &\leq -\frac{1}{\eta}\langle \nabla f(\mathbf{x}), G_\eta(\mathbf{x})\rangle + \frac{1}{2\eta}\|G_\eta(\mathbf{x})\|_2^2. \end{aligned}$$

Thus, to prove the lemma, it suffices to show that $-\frac{1}{\eta}\langle \nabla f(\mathbf{x}), G_\eta(\mathbf{x})\rangle \leq -\frac{1}{\eta}\|G_\eta(\mathbf{x})\|_2^2$. This statement is equivalent to:

$$\begin{aligned} & \langle \nabla f(\mathbf{x}), G_\eta(\mathbf{x})\rangle \geq \|G_\eta(\mathbf{x})\|_2^2 \\ \Leftrightarrow\quad & \langle \nabla f(\mathbf{x}) - G_\eta(\mathbf{x}), G_\eta(\mathbf{x})\rangle \geq 0 \\ \Leftrightarrow\quad & \langle \nabla f(\mathbf{x}) - \eta(\mathbf{x} - \bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}}\rangle \geq 0 \\ \Leftrightarrow\quad & \Big\langle P_{\mathcal{X}}\Big(\mathbf{x} - \frac{1}{\eta}\nabla f(\mathbf{x})\Big) - \Big(\mathbf{x} - \frac{1}{\eta}\nabla f(\mathbf{x})\Big), \mathbf{x} - P_{\mathcal{X}}\Big(\mathbf{x} - \frac{1}{\eta}\nabla f(\mathbf{x})\Big)\Big\rangle \geq 0. \end{aligned}$$

The last inequality holds by Lemma 2.1, hence the proof is complete. □

3.1 Convergence of PGD for Possibly Nonconvex Objectives

Based on Lemmas 3.2 and 3.3, we can now immediately obtain the following convergence result for PGD.

Theorem 3.4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth function and let $\mathcal{X}$ be a closed convex set. Starting with an arbitrary initial point $\mathbf{x}_0 \in \mathcal{X}$, consider the sequence of PGD iterates defined by (PGD) for $k \geq 0$. If $\eta \geq 0$ and f is bounded below by $f_* > -\infty$ on $\mathcal{X}$, then for all $k \geq 0$,*

$$\min_{0 \leq i \leq k} \|G_\eta(\mathbf{x}_i)\|_2^2 \leq \frac{2\eta(f(\mathbf{x}_0) - f_*)}{k+1}.$$

As a consequence, given any $\epsilon > 0$, for any $k \geq \lceil \frac{2\eta(f(\mathbf{x}_0)-f_*)(L/\eta+1)^2}{\epsilon^2} \rceil$, there exists $\hat{\mathbf{x}} \in \{\mathbf{x}_0, \dots, \mathbf{x}_k\}$ such that

$$-\nabla f(\hat{\mathbf{x}}) \in N_{\mathcal{X}}(\hat{\mathbf{x}}) + \mathcal{B}(\boldsymbol{0}, \epsilon).$$

Proof. Observe that by the definition of PGD, it is necessarily the case that $\mathbf{x}_i \in \mathcal{X}$, for all $i \geq 0$. By Lemma 3.3, for any $i \geq 0$,

$$f(\mathbf{x}_{i+1}) - f(\mathbf{x}_i) \leq -\frac{1}{2\eta}\|G_\eta(\mathbf{x}_i)\|_2^2.$$

Summing the last inequality over $0 \leq i \leq k$ and using that $f(\mathbf{x}_{k+1}) \geq f_*$, the first claim of the theorem follows after a simple rearrangement.

For the second claim, observe first that, based on the first claim, for any $\bar{\epsilon} > 0$,

$$\min_{0 \leq i \leq k} \|G_\eta(\mathbf{x}_i)\|_2^2 \leq \bar{\epsilon}^2$$

for $k \geq \frac{2\eta(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{\bar{\epsilon}^2} - 1$. Hence, for any $k \geq \lceil \frac{2\eta(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{\bar{\epsilon}^2} \rceil - 1$, there exists $\mathbf{y} \in \{\mathbf{x}_0, \dots, \mathbf{x}_k\}$ such that $\|G_\eta(\mathbf{y})\|_2 \leq \bar{\epsilon}$. By Lemma 3.2, for $\hat{\mathbf{x}} = P_{\mathcal{X}}(\mathbf{y} - \frac{1}{\eta}\nabla f(\mathbf{y})) \in \{\mathbf{x}_0, \dots, \mathbf{x}_{k+1}\}$, it holds

$$-\nabla f(\hat{\mathbf{x}}) \in N_{\mathcal{X}}(\hat{\mathbf{x}}) + \mathcal{B}(\mathbf{0}, \bar{\epsilon}(L/\eta + 1)).$$

To complete the proof, it remains to choose $\bar{\epsilon}$ so that $\bar{\epsilon}(L/\eta + 1)) = \epsilon$. □

3.2 Convergence of PGD for Convex Objectives

It is possible to further follow the analogy between GD and PGD to analyze the convergence of PGD in the case where the objective function f is additionally convex (we still assume that $\eta \geq L$). In this subsection, we further assume that the minimum of f on $\mathcal{X}$ is attained at some $\mathbf{x}^* \in \mathcal{X}$.

As before, convexity allows us to estimate the minimum value of f on $\mathcal{X}$, equal to $f(\mathbf{x}^*)$. In particular,

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle. \tag{8}$$

Hence, the optimality gap for point $\mathbf{x}_{k+1}$ is bounded by

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) - \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle. \tag{9}$$

To further bound $f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)$, observe that we can write the iterates of PGD as solutions to the following optimization problem:

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{X}} \{\langle \nabla f(\mathbf{x}_k), \mathbf{x} \rangle + \frac{\eta}{2}\|\mathbf{x} - \mathbf{x}_k\|_2^2\}.$$

Let $h(\mathbf{x}) = \langle \nabla f(\mathbf{x}_k), \mathbf{x} \rangle + \frac{\eta}{2}\|\mathbf{x} - \mathbf{x}_k\|_2^2$, so that $\mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} h(\mathbf{x})$. Since h is a quadratic function with its Hessian equal to $\eta \mathbf{I}$,, we have that for any $\mathbf{x}, \mathbf{y}$,

$$h(\mathbf{x}) = h(\mathbf{y}) + \langle \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\eta}{2}\|\mathbf{x} - \mathbf{y}\|_2^2.$$

In particular, for $\mathbf{x} = \mathbf{x}^*$ and $\mathbf{y} = \mathbf{x}_{k+1}$, we have

$$\begin{aligned} h(\mathbf{x}^*) &= h(\mathbf{x}_{k+1}) + \langle \nabla h(\mathbf{x}_{k+1}), \mathbf{x}^* - \mathbf{x}_{k+1} \rangle + \frac{\eta}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2 \\ &\geq h(\mathbf{x}_{k+1}) + \frac{\eta}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2, \end{aligned}$$

where the inequalities holds because $\langle \nabla h(\mathbf{x}_{k+1}), \mathbf{x}^* - \mathbf{x}_{k+1}\rangle$, as $\mathbf{x}_{k+1}$ minimizes h over $\mathcal{X}$ and $\mathbf{x}^* \in \mathcal{X}$. Plugging in the definition of h into the last inequality and simplifying, we have that

$$\langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_{k+1}\rangle \geq \frac{\eta}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 - \frac{\eta}{2}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 + \frac{\eta}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2. \tag{10}$$

Writing $\langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k\rangle = \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_{k+1}\rangle + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle$ and plugging (10) into (9), we obtain the following bound on the optimality gap:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) - \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle - \frac{\eta}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 + \frac{\eta}{2}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 - \frac{\eta}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2.$$

For $\eta \geq L$, we have that $f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) - \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle - \frac{\eta}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 \leq 0$, as f is L-smooth. Hence, obtain the same bound as we had for gradient descent:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \frac{\eta}{2}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 - \frac{\eta}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2. \tag{11}$$

Same as for gradient descent, telescoping (11) and using that $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) \leq \cdots \leq f(\mathbf{x}_0)$, we now obtain:

$$(k+1)(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)) \leq \frac{\eta}{2}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 - \frac{\eta}{2}\|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2 \leq \frac{\eta}{2}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2.$$

Thus,

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \frac{\eta\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{2(k+1)}.$$

We can further conclude from the last inequality that for any $\epsilon > 0$, $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon$ after at most $k = \lceil \frac{\eta\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2\epsilon} \rceil$ iterations.

Similar to GD, it is possible to extend the analysis of the smooth convex case to the smooth strongly convex case by using the quadratic lower bound we get from strong convexity in place of (8) and recover the same convergence bound that we had for GD. The proof is left as an exercise.

Exercises

1. Prove that the expressions for projections provided in Examples 2.2–2.4 are correct.

2. Given an L-smooth function f, a closed convex set $\mathcal{X}$, and their gradient mapping G_η with $\eta \geq L$, prove that G_η is 2η-Lipschitz continuous.

3. Let $f : \mathbb{R}^d$ be a convex L-smooth function and let $\mathcal{X} \subseteq \mathbb{R}^d$ be convex and closed. Let $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ (assume at least one such $\mathbf{x}^*$ exists). Consider the following algorithm, for $k \geq 0$:

$$\mathbf{x}_{k+1} = \alpha_k\mathbf{x}_0 + (1-\alpha_k)P_{\mathcal{X}}\Big(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\Big), \tag{12}$$

where $\mathbf{x}_0 \in \mathcal{X}$ is an arbitrary initial point and $\alpha_k \in (0,1)$, $\forall k \geq 0$.

(i) Argue that $\mathbf{x}^* = P_{\mathcal{X}}\Big(\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*)\Big)$.

(ii) Prove that for all $k \geq 0$, $\|\mathbf{x}_k - \mathbf{x}^*\|_2 \leq \|\mathbf{x}_0 - \mathbf{x}^*\|_2$.

(iii) It is possible to show (and you are not asked to show this) that when $\alpha_k = \frac{1}{k+1}$, we have

$$\Big\|\mathbf{x}_k - P_{\mathcal{X}}\Big(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\Big)\Big\|_2 = O\Big(\frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2}{k+1}\Big). \tag{13}$$

Let $\bar{\mathbf{x}}_k = P_{\mathcal{X}}\Big(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\Big)$. Argue that Part (ii) combined with Eq. (13) implies that

$$f(\bar{\mathbf{x}}_k) - f(\mathbf{x}^*) = O\Big(\frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{k+1}\Big).$$

Hint: Recall the properties of projections and gradient mapping that we proved in class and in homework assignments.

4. Consider the following convex composite optimization problem $\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$, where $F(\mathbf{x}) = f(\mathbf{x}) + \psi(\mathbf{x})$, $f : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth and $\psi : \mathbb{R}^d \to \mathbb{R}$ is convex and lower semicontinuous. In particular, when ψ is the indicator function of a convex set, the considered minimization problem corresponds to minimizing a smooth convex function over a closed convex set. Your goal in this question is to show that we can design an accelerated algorithm (provided in this question) that for any $\epsilon > 0$ can construct a point $\mathbf{y}_k$ with $F(\mathbf{y}_k) - F(\mathbf{x}^*) \le \epsilon$ within $k = O\Big(\sqrt{\frac{L}{\epsilon}}\|\mathbf{x}^* - \mathbf{x}_0\|_2\Big)$ iterations, where $\mathbf{x}_0 \in \mathrm{dom}(\psi)$ is the initial point of the algorithm and $\mathbf{x}^* \in \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$.

(i) Is F smooth in general? Is its minimizer unique?

Consider the following algorithm:

$$\begin{aligned} \mathbf{x}_k &= \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_{k-1}, \\ \mathbf{v}_k &= \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} M_k(\mathbf{x}), \\ \mathbf{y}_k &= \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_k, \end{aligned} \tag{14}$$

where

$$M_k(\mathbf{x}) = \sum_{i=0}^{k} a_i \langle \nabla f(\mathbf{x}_i), \mathbf{x} - \mathbf{x}_i\rangle + A_k\psi(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2, \tag{15}$$

$\mathbf{y}_0 = \mathbf{v}_0$, $\mathbf{x}_0 \in \mathrm{dom}(\psi)$ is an arbitrary initial point, $a_i > 0$, $\forall i$, and $A_k = \sum_{i=0}^k a_i$.

(ii) Let $U_k = f(\mathbf{y}_k) + \frac{1}{A_k}\sum_{i=0}^k a_i\psi(\mathbf{v}_i)$. Prove that U_k is a valid upper bound on $F(\mathbf{y}_k)$, i.e., $U_k \ge F(\mathbf{y}_k)$.

(iii) Prove that L_k defined by

$$L_k = \frac{1}{A_k}\sum_{i=0}^{k} a_i f(\mathbf{x}_i) + \frac{1}{A_k}M_k(\mathbf{v}_k) - \frac{1}{2A_k}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 \tag{16}$$

is a valid lower bound for $F(\mathbf{x}^*)$, i.e., $L_k \le F(\mathbf{x}^*)$.

(iv) Let $G_k = U_k - L_k$. Prove that:

$$\begin{aligned} A_kU_k - A_{k-1}U_{k-1} \le \frac{A_kL}{2}\|\mathbf{y}_k - \mathbf{x}_k\|_2^2 + \langle \nabla f(\mathbf{x}_k), A_k\mathbf{y}_k - A_{k-1}\mathbf{y}_{k-1} - a_k\mathbf{x}_k\rangle \\ + a_k(f(\mathbf{x}_k) + \psi(\mathbf{v}_k)). \end{aligned} \tag{17}$$

Prove that:

$$M_k(\mathbf{v}_k) - M_{k-1}(\mathbf{v}_{k-1}) \ge a_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\rangle + a_k\psi(\mathbf{v}_k) + \frac{1}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2,$$

and, thus,

$$A_kL_k - A_{k-1}L_{k-1} \ge a_k(f(\mathbf{x}_k) + \psi(\mathbf{v}_k)) + a_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\rangle + \frac{1}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2. \tag{18}$$

Combine Eq. (17) and Eq. (18) to conclude that

$$A_kG_k - A_{k-1}G_{k-1} \le \Big(\frac{{a_k}^2}{A_k}L - 1\Big)\frac{\|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2}{2}. \tag{19}$$

In particular, $A_kG_k - A_{k-1}G_{k-1} \le 0$ whenever $\frac{{a_k}^2}{A_k} \le \frac{1}{L}$. Argue that $a_k = \frac{k+1}{2L}$ suffices here.

(v) Show that under a suitable choice of $A_0 = a_0$, we have $A_0G_0 \le \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2$.

(vi) Combine Parts (iv) and (v) to argue that, under the choices of a_i, A_i, $i \ge 0$, from Parts (iv) and (v), we have that, for any $\epsilon > 0$, (14) produces a point $\mathbf{y}_k \in \mathbb{R}^d$ such that $F(\mathbf{y}_k) - F(\mathbf{x}^*) \le \epsilon$ within at most $k = O\Big(\sqrt{\frac{L}{\epsilon}}\|\mathbf{x}^* - \mathbf{x}_0\|_2\Big)$ iterations.