CS726 Optional Lecture: Fixed-Point Methods

1 Introduction to fixed-point methods

We use $\mathrm{Id} : \mathbb{R}^n \to \mathbb{R}^n$ to denote the identity operator $\mathrm{Id}(x) = x$. We will often (for this operator and others) drop the parenthesis when denoting the evaluation of an operator at a point x.

Definition 1. Let $T : \mathbb{R}^n \to \mathbb{R}^n$. Let $\|\cdot\|$ be any norm. We say T is nonexpansive if

$$\|Tx - Ty\| \leq \|x - y\| .$$

We say T is (γ-)contractive if there exists some $\gamma < 1$ such that

$$\|Tx - Ty\| \leq \gamma \|x - y\| .$$

We say T is (θ-)averaged if there exists some $\theta \in (0, 1)$ and some nonexpansive H such that

$$T = (1 - \theta)\mathrm{Id} + \theta H.$$

Definition 2. Let $T : \mathbb{R}^n \to \mathbb{R}^n$. $x \in \mathbb{R}^n$ is a fixed point of T if

$$Tx = x.$$

We let $\mathrm{Fix}\ T$ denote the set of all fixed points of T.

Definition 3. Fix some $x_0 \in \mathbb{R}^n$ and some $T : \mathbb{R}^n \to \mathbb{R}^n$. The Picard iteration is the sequence $(x_k)_{k=0}^{\infty}$ generated by

$$x_{k+1} = Tx_k. \tag{1}$$

With these three definitions out of the way, now we can describe a very general algorithmic blueprint, the analysis of which will be the topic of this lecture:

1. Start with a problem that we want to solve, which entails finding some point $x^\star \in \mathbb{R}^n$ with some special properties.
2. Rewrite the problem as a fixed-point problem: find an operator T such that $x^\star$ possesses the desired properties if $x^\star \in \mathrm{Fix}\ T$.
3. Run the Picard iteration (1), and obtain a convergence guarantee based upon the properties of T.

2 The contractive case

Example 1 (Strongly convex gradient descent)**.** *Suppose that f is L-smooth and m-strongly convex. Gradient descent with stepsize $\frac{1}{L}$ generates iterates*

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) = F(x_k)$$

where

$$F(x) = x - \frac{1}{L}\nabla f(x).$$

Note that if we find a fixed-point $x^\star$ of F, then we have

$$F(x^\star) = x^\star \iff x^\star = x^\star - \frac{1}{L}\nabla f(x^\star) \iff \nabla f(x^\star) = 0.$$

We can show that F is contractive. We compute

$$\begin{aligned}
\left\| \left(x - \frac{1}{L}\nabla f(x)\right) - \left(y - \frac{1}{L}\nabla f(y)\right) \right\|_2^2 &= \|x-y\|_2^2 + \frac{1}{L^2}\|\nabla f(x) - \nabla f(y)\|_2^2 - \frac{2}{L}\langle x-y, \nabla f(x) - \nabla f(y)\rangle \\
&\le \|x-y\|_2^2 + \frac{1}{L}\langle x-y, \nabla f(x) - \nabla f(y)\rangle - \frac{2}{L}\langle x-y, \nabla f(x) - \nabla f(y)\rangle \\
&= \|x-y\|_2^2 - \frac{1}{L}\langle x-y, \nabla f(x) - \nabla f(y)\rangle \\
&\le \|x-y\|_2^2 - \frac{m}{L}\|x-y\|_2^2
\end{aligned}$$

where in the first inequality we used the cocoercivity property

$$\langle x-y, \nabla f(x) - \nabla f(y)\rangle \ge \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2.$$

and in the second inequality we used the coercivity/strong monotonicity property

$$\langle x-y, \nabla f(x) - \nabla f(y)\rangle \ge m\|x-y\|_2^2.$$

Therefore F is a $\sqrt{1-\frac{m}{L}} \approx 1 - \frac{1}{2}\frac{m}{L}$ contraction.

Example 2 (Strongly convex projected gradient descent)**.** *In addition to the assumptions of the previous example, suppose also that $\mathcal{C}$ is a closed convex set, and let $P_{\mathcal{C}}(x) := \operatorname{argmin}_{y\in\mathcal{C}} \|y-x\|_2^2$ be the projection operator onto the set $\mathcal{C}$. Projected gradient descent generates iterates via*

$$x_{k+1} = P_{\mathcal{C}}\left(x_k - \frac{1}{L}\nabla f(x_k)\right) = P_{\mathcal{C}}\left(F(x_k)\right)$$

where the operator F was defined in the previous example. Since the composition of a nonexpansion with a contraction is still a contraction (with the same contraction factor), we immediately have that $P_{\mathcal{C}}F$ is a $\sqrt{1-\frac{m}{L}}$ contraction. Note that now a fixed point $x^\star$ of $P_{\mathcal{C}}F$ satisfies

$$x^\star = P_{\mathcal{C}}\left(x^\star - \frac{1}{L}\nabla f(x^\star)\right)$$

which was shown in the lecture on projected gradient descent to be equivalent to the first order optimality condition $-\nabla f(x^\star) \in N_{\mathcal{C}}(x^\star)$.

Theorem 1 (Banach Fixed-Point Theorem)**.** *Suppose T is a γ-contraction with respect to some norm $\|\cdot\|$. Then*

1. *T has a unique fixed point $x^\star$ (equivalently* Fix T *is a singleton)*
2. *The Picard iteration (1) starting from x_0 satisfies*

$$\|Tx_k - x_k\| \le \gamma^k \|Tx_0 - x_0\|.$$

3. *We have*

$$\|x_k - x^\star\| \le \frac{1}{1-\gamma}\|Tx_k - x_k\|$$

Remark 2. The last point implies that not only can we reduce the fixed-point error $\|Tx_k - x_k\|$ to 0 at a geometric rate, but we can also approach the (unique) fixed point $x^\star$ at a geometric rate.

Proof. If we run the Picard iteration (1) starting from some arbitrary x_0, then we have

$$\|Tx_k - x_k\| = \|TTx_{k-1} - Tx_{k-1}\| \leq \gamma \|Tx_{k-1} - x_{k-1}\|$$

so repeating this argument k times we obtain that

$$\|Tx_k - x_k\| \leq \gamma \|Tx_{k-1} - x_{k-1}\| \leq \cdots \leq \gamma^k \|Tx_0 - x_0\|.$$

Note that in particular this shows that the fixed-point error $\|Tx_k - x_k\|$ is strictly decreasing. Noting that $Tx_k = x_{k+1}$, this means that we have

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_k\| \leq \sum_{k=0}^{\infty} \gamma^k \|Tx_0 - x_0\| = \frac{\|Tx_0 - x_0\|}{1-\gamma}$$

using the geometric series formula $\sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$ for $\gamma \in [0, 1)$. Therefore the sequence $(x_k)_{k=0}^{\infty}$ is Cauchy, since for any $\varepsilon > 0$, there exists some integer K such that $\sum_{k=K}^{\infty} \|x_{k+1} - x_k\| < \varepsilon$ (since this infinite sum is finite), and this implies by the triangle inequality that for any $k_1, k_2 \geq K$ we have

$$\|x_{k_1} - x_{k_2}\| \leq \|x_{k_1} - x_K\| + \|x_K - x_{k_2}\| \leq \left(\sum_{k=K}^{k_1-1} \|x_{k+1} - x_k\|\right) + \left(\sum_{k=K}^{k_2-1} \|x_{k+1} - x_k\|\right) \leq 2\varepsilon.$$

Therefore the sequence $(x_k)_{k=0}^{\infty}$ converges to some limit $x^\star$. Now we show that $x^\star \in \operatorname{Fix} T$. By triangle inequality, for any k we have

$$\begin{aligned}\|Tx^\star - x^\star\| &= \|Tx^\star - Tx_k + Tx_k - x_k + x_k - x^\star\| \leq \|Tx^\star - Tx_k\| + \|Tx_k - x_k\| + \|x_k - x^\star\| \\ &\leq \gamma \|x_k - x^\star\| + \|Tx_k - x_k\| + \|x_k - x^\star\|\end{aligned}$$

(using γ-contractivity in the last step). But since all three of these terms can be made arbitrarily small by taking k sufficiently large, this implies that $\|Tx^\star - x^\star\|$ is arbitarily small, so we must have $\|Tx^\star - x^\star\| = 0$, that is, that $Tx^\star = x^\star$.

By adding and subtracting Tx_k and using that $Tx^\star = x^\star$, we have

$$\begin{aligned}&\|x_k - x^\star\| = \|x_k - Tx_k + Tx_k - x^\star\| \leq \|x_k - Tx_k\| + \|Tx_k - Tx^\star\| \leq \|x_k - Tx_k\| + \gamma \|x_k - x^\star\| \\ \implies &(1-\gamma) \|x_k - x^\star\| \leq \|x_k - Tx_k\| \\ \implies &\|x_k - x^\star\| \leq \frac{\|x_k - Tx_k\|}{1-\gamma}.\end{aligned}$$

Note that this last argument holds for an arbitrary $x^\star \in \operatorname{Fix} T$. Now letting $x_2^\star \in \operatorname{Fix} T$ be arbitrary as well, we can imagine initializing the Picard iteration with $x_0 = x_2^\star$, and then the previous argument implies that

$$\|x_2^\star - x^\star\| = \|x_0 - x^\star\| \leq \frac{\|Tx_0 - x_0\|}{1-\gamma} = \frac{\|Tx_2^\star - x_2^\star\|}{1-\gamma} = 0$$

so we must have $x_2^\star = x^\star$, that is, the set $\operatorname{Fix} T$ contains only one element (equivalently, T has a unique fixed point). □

3 The nonexpansive case

Now how can we find a fixed point of a nonexpansive operator? Generally nonexpansive operators are not guaranteed to have any fixed points, but assuming it has some, the Picard iteration (1) still may not converge to one:

Example 3 (Non-convergence of Picard iteration for nonexpansive operators)**.**

1. *(Rotation within the line) For $x \in \mathbb{R}$, let $T(x) = -x$. Then* Fix $T = \{0\}$*, but Picard iteration with T will oscillate between x_0 and $-x_0$ infinitely.*

2. *In higher dimensions, the previous example T is a special case of both reflection (about an axis) and rotation (around the origin), both of which are again nonexpansive and have fixed points which will not generally be approached by the Picard iteration.*

A sufficient condition for the Picard iteration to converge is for the operator T to be averaged. Before showing this, we first establish some general facts about nonexpansive and averaged operators and their fixed points.

Theorem 3.

$$\{\textit{contractive operators}\} \subset \{\textit{averaged operators}\} \subset \{\textit{nonexpansive operators}\}.$$

Proof. First we show that if T is averaged then it is nonexpansive. Since $T = (1-\theta)\mathrm{Id} + \theta H$ for some nonexpansive H, we have

$$\|Tx - Ty\| = \|(1-\theta)(x-y) + \theta(Hx - Hy)\| \leq (1-\theta)\|x-y\| + \theta\|Hx - Hy\| \leq (1-\theta+\theta)\|x-y\| = \|x-y\|.$$

Next we show that if T is γ-contractive then it is averaged. To find an α such that T is α-averaged, we write (for arbitrary $\alpha \in (0,1)$)

$$T(x) = (1-\alpha)x + \alpha\frac{T(x) - (1-\alpha)x}{\alpha}$$

and try to show that $x \mapsto \frac{T(x)-(1-\alpha)x}{\alpha}$ is nonexpansive. Then

$$\left\|\frac{Tx-(1-\alpha)x}{\alpha} - \frac{Ty-(1-\alpha)y}{\alpha}\right\| \leq \frac{1}{\alpha}\|Tx-Ty\| + \frac{1-\alpha}{\alpha}\|x-y\| \leq \frac{\gamma+1-\alpha}{\alpha}\|x-y\|.$$

Now to choose α so that $\frac{\gamma+1-\alpha}{\alpha} \leq 1$, by rearranging it suffices to set $\alpha \geq \frac{\gamma+1}{2}$. (It is possible to do this with an $\alpha < 1$ since $\gamma < 1$.) □

As we saw already, all contractive operators have exactly one fixed point, making the study of their fixed points not so interesting. For averaged and nonexpansive operators, the set of all fixed points may be empty, and it also may contain many points.

Example 4. *1.* Fix $\mathrm{Id} = \mathbb{R}^n$.

2. *Let $\mathcal{C}$ be a closed convex set, and let $P_{\mathcal{C}}(x) := \mathrm{argmin}_{y\in\mathcal{C}}\|y-x\|_2^2$ be the projection operator onto the set $\mathcal{C}$. As we have already seen, $P_{\mathcal{C}}$ is nonexpansive and* Fix $P_{\mathcal{C}} = \mathcal{C}$.

3. *Let $T(x) = x+1$ define $T : \mathbb{R} \to \mathbb{R}$. Then T is nonexpansive but* Fix $T = \emptyset$*. Also T is averaged, since we can write $T(x) = \frac{1}{2}x + \frac{1}{2}(x+2)$.*

The set of all fixed points of a nonexpansive operator (and thus also an averaged operator) is always closed and convex.

Theorem 4. *If $T : \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive, then* Fix T *is closed and convex.*

Proof. First we check that Fix T is closed. Note that T is continuous (since it is nonexpansive, which is equivalent to being 1-Lipschitz-continuous). Therefore $T-I$ is also continuous. Since the preimage of a closed set under a continuous function must also be closed, and the set $\{0\}$ is closed, we have that $(I-T)^{-1}(\{0\})$ is a closed set. Finally notice that the set $(I-T)^{-1}(\{0\})$ is equal to Fix T.

Now we show that Fix T is convex. Fix $x, y \in$ Fix T and let $\lambda \in (0,1)$. Our goal is to show that $\lambda x + (1-\lambda)y \in$ Fix T. For convenience let $w = \lambda x + (1-\lambda)y$. We have that

$$\|Tw - x\| = \|Tw - Tx\| \leq \|w - x\|$$

and similarly

$$\|Tw - y\| = \|Tw - Ty\| \leq \|w - y\|.$$

Therefore Tw is closer (nonstrictly) to x than w, and also Tw is closer to y than w. But w is already on the straight line between x and y, so this is only possible if $w = Tw$. (We can make this final step more rigorously, but the proof is better via picture anyways.) □

Example 5. *If $\lambda \in (0, 1)$, then*

$$\text{Fix } T = \text{Fix } ((1-\lambda)\text{Id} + \lambda T).$$

Proof. If $x \in \text{Fix } T$ then

$$((1-\lambda)\text{Id} + \lambda T)\, x = (1-\lambda)x + \lambda Tx = (1-\lambda)x + \lambda x = x$$

so $x \in \text{Fix } ((1-\lambda)\text{Id} + \lambda T)$. If $x \in \text{Fix } ((1-\lambda)\text{Id} + \lambda T)$, then we have

$$x = (1-\lambda)x + \lambda Tx \implies \lambda x = \lambda Tx \implies x \in \text{Fix } T.$$

□

This simple fact is very useful. Note that if some operator H is nonexpansive but not averaged, we can simply define $T = \frac{1}{2}\text{Id} + \frac{1}{2}H$, and then T will be averaged but also have the same fixed points as H by Example 5.

Example 6. *Revisiting our example of the nonexpansive operator $T(x) = -x$, after applying the transformation $T'(x) := (1-\theta)x + \theta T(x) = (1-2\theta)x$, since $\theta \in (0, 1)$ we will now shrink towards the origin (T' is actually a $(1-2\theta)$-contractive operator, but we don't generally expect to get a contractive operator from this recipe).*

The class of averaged operators contains countless operators of interest.

Example 7 (Non-strongly-convex gradient descent)**.** *Suppose that f is an L-smooth function which is convex but not strongly convex. Gradient descent with stepsize $\frac{1}{L}$ generates iterates*

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) = F(x_k)$$

where

$$F(x) = x - \frac{1}{L}\nabla f(x).$$

Now we will show that F is $\frac{1}{2}$-averaged. We can write

$$F(x) = \frac{1}{2}x + \frac{1}{2}\left(x - \frac{2}{L}\nabla f(x)\right)$$

so it suffices to show that the operator $x \mapsto x - \frac{2}{L}\nabla f(x)$ is nonexpansive. The key fact we need is that ∇f is $\frac{1}{L}$-cocoercive:

$$\langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2.$$

Using this, we can compute that

$$\begin{aligned}\left\|\left(x - \frac{2}{L}\nabla f(x)\right) - \left(y - \frac{2}{L}\nabla f(y)\right)\right\|_2^2 &= \|x-y\|_2^2 + \frac{4}{L^2}\|\nabla f(x) - \nabla f(y)\|_2^2 - \frac{4}{L}\langle x - y, \nabla f(x) - \nabla f(y)\rangle \\ &\leq \|x-y\|_2^2.\end{aligned}$$

Example 8 (Projections are $\frac{1}{2}$-averaged)**.** *Let $\mathcal{C}$ be a closed convex and nonempty set. Let*

$$P_{\mathcal{C}}(x) := \mathrm{argmin}_{y\in\mathcal{C}} \|x-y\|_2^2$$

be the projection operator onto the set $\mathcal{C}$. We already know that $P_{\mathcal{C}}$ is nonexpansive. In fact $P_{\mathcal{C}}$ is $\frac{1}{2}$-averaged. To show this, we need to show a cocoercivity property of $P_{\mathcal{C}}$:

$$\langle x-y, P_{\mathcal{C}}(x)-P_{\mathcal{C}}(y)\rangle \geq \|P_{\mathcal{C}}(x)-P_{\mathcal{C}}(y)\|_2^2 . \tag{2}$$

Assuming (2), we can similarly write

$$P_{\mathcal{C}}(x) = \frac{1}{2}x + \frac{1}{2}(2P_{\mathcal{C}}(x)-x)$$

and then we need to show that $2P_{\mathcal{C}}-\mathrm{Id}$ is nonexpansive. ($2P_{\mathcal{C}}-\mathrm{Id}$ is known as the overprojection operator.) Then using (2) we have

$$\begin{aligned}\|(2P_{\mathcal{C}}(x)-x)-(2P_{\mathcal{C}}(y)-y)\|_2^2 &= 4\|P_{\mathcal{C}}(x)-P_{\mathcal{C}}(y)\|_2^2 + \|x-y\|_2^2 - 4\langle x-y, P_{\mathcal{C}}(x)-P_{\mathcal{C}}(y)\rangle \\ &\leq \|x-y\|_2^2 .\end{aligned}$$

Remark 5. The property of being $\frac{1}{2}$-averaged is equivalent to the property of being firmly nonexpansive (mentioned in the lecture notes).

4 The KM Theorem

The fact that Picard iteration converges for averaged operators is still true with general norms, but we will just focus on the case of $\|\cdot\|_2$. First we need one useful identity.

Lemma 6. *Fix $x, y \in \mathbb{R}^n$ and suppose $\theta \in [0,1]$. Then*

$$\|\theta x + (1-\theta)y\|_2^2 = \theta\|x\|_2^2 + (1-\theta)\|y\|_2^2 - \theta(1-\theta)\|x-y\|_2^2 .$$

Proof. Expanding the LHS, we get

$$\|\theta x + (1-\theta)y\|_2^2 = \theta^2\|x\|_2^2 + (1-\theta)^2\|y\|_2^2 + 2\theta(1-\theta)\langle x, y\rangle. \tag{3}$$

Now we want to replace the inner product with an expression involving the norms of x, y, and $x-y$, which we can do using the Law of Cosines identity

$$2\langle x, y\rangle = \|x\|_2^2 + \|y\|_2^2 - \|x-y\|_2^2 . \tag{4}$$

Multiplying the equality (4) by $\theta(1-\theta)$ and subtracting it from (3) we obtain the desired equality. □

Theorem 7 (Krasnoselskii-Mann (KM) Theorem)**.** *Suppose $T = (1-\theta)\mathrm{Id} + \theta H$ for some nonexpansive H (that is, T is θ-averaged) and we run the Picard iteration (1) with initial point x_0. Also suppose that* Fix T *is nonempty. Then*

1. *(Fejer-monotonicity) For any $x^\star \in$ Fix T, the sequence $\|x_k - x^\star\|_2$ is monotonically non-increasing.*
2. *The fixed-point residual $\|Tx_k - x_k\|_2$ is monotonically non-increasing.*
3. *For any $k \geq 0$,*

$$\|Tx_k - x_k\|_2^2 \leq \frac{\theta}{1-\theta}\frac{\|x_0 - x^\star\|_2^2}{k+1}.$$

Remark 8. We might care about finding x with a small fixed-point residual $\|Hx - x\|_2$ with respect to H rather than T. However, we have

$$Tx - x = (1-\theta)x + \theta Hx - x = \theta\,(Hx - x)$$

so the fixed-point residuals $\|Hx - x\|_2$ and $\|Tx - x\|_2$ are equivalent up to a constant depending on θ (a fact that we will use in the proof). In particular the fixed-point error with respect to H is also monotone non-increasing.

Proof. Using Lemma 6, we have (for any $x^\star \in \operatorname{Fix} T$) that

$$\begin{aligned}
\|x_{k+1} - x^\star\|_2^2 &= \|Tx_k - Tx^\star\|_2 = \|((1-\theta)\mathrm{Id} + \theta H)\,x_k - ((1-\theta)\mathrm{Id} + \theta H)\,x^\star\|_2^2 \\
&= \|(1-\theta)(x_k - x^\star) + \theta(Hx_k - Hx^\star)\|_2^2 \\
&= (1-\theta)\,\|x_k - x^\star\|_2^2 + \theta\,\|Hx_k - Hx^\star\|_2^2 - \theta(1-\theta)\,\|x_k - Hx_k - x^\star + Hx^\star\|_2^2 \\
&= (1-\theta)\,\|x_k - x^\star\|_2^2 + \theta\,\|Hx_k - Hx^\star\|_2^2 - \theta(1-\theta)\,\|x_k - Hx_k\|_2^2 \\
&\le (1-\theta)\,\|x_k - x^\star\|_2^2 + \theta\,\|x_k - x^\star\|_2^2 - \theta(1-\theta)\,\|x_k - Hx_k\|_2^2 \\
&= \|x_k - x^\star\|_2^2 - \frac{1-\theta}{\theta}\,\|x_k - Tx_k\|_2^2
\end{aligned}$$

where we used that $Hx^\star = x^\star$, by Example 5, then nonexpansiveness of H, and then finally the relationship between the fixed-point residuals $\|Hx - x\|_2$ and $\|Tx - x\|_2$. Note that this already gives the Fejer monotonicity property (by using $\theta(1-\theta)\,\|x_k - Hx_k\|_2^2 \ge 0$).

Now by rearranging and using a standard telescoping argument, we have that

$$\frac{1-\theta}{\theta}\sum_{k=0}^{K}\|Tx_k - x_k\|_2^2 \le \sum_{k=0}^{K}\left(\|x_k - x^\star\|_2^2 - \|x_{k+1} - x^\star\|_2^2\right) = \|x_0 - x^\star\|_2^2 - \|x_{K+1} - x^\star\|_2^2 \le \|x_0 - x^\star\|_2^2\,.$$

Furthermore,

$$\|Tx_k - x_k\|_2 = \|Tx_k - Tx_{k-1}\|_2 \le \|x_k - x_{k-1}\|_2 = \|Tx_{k-1} - x_{k-1}\|_2$$

so the fixed-point error with respect to T is monotonically non-increasing. This implies that we can lower bound

$$\frac{1-\theta}{\theta}\sum_{k=0}^{K}\|Tx_k - x_k\|_2^2 \ge \frac{1-\theta}{\theta}(K+1)\,\|Tx_K - x_K\|_2^2\,.$$

Combining with our earlier bound and rearranging we can conclude that

$$\|Tx_K - x_K\|_2^2 \le \frac{\theta}{1-\theta}\frac{\|x_0 - x^\star\|_2^2}{K+1}.$$

□

Remark 9. It is also possible, without much more work, to show that the iterates $(x_k)_k$ converge to some fixed point of T, but (unlike for the fixed-point residual norm) this convergence can be arbitrarily slow.

Example 9 (Non-strongly-convex gradient descent, continued)**.** *Applying the KM Theorem to the example where f is an L-smooth, convex function, we have that the fixed-point residual*

$$\|Fx - x\|_2 = \left\|x - \frac{1}{L}\nabla f(x) - x\right\|_2 = \frac{1}{L}\,\|\nabla f(x)\|_2$$

is a multiple of the gradient norm. Therefore the theorem gives that

$$\|\nabla f(x_k)\|_2^2 \le \frac{L^2}{k+1}\,\|x_0 - x^\star\|_2^2\,.$$

It was shown in a previous lecture (using only the descent lemma, and therefore not assuming convexity) that

$$\min_{0\le t\le k} \|\nabla f(x_t)\|_2^2 \le \frac{2L}{k+1}\left(f(x_0) - f(x^\star)\right).$$

Using $\nabla f(x^\star) = 0$ and L-smoothness, we can show that

$$f(x_0) - f(x^\star) \le \langle \nabla f(x^\star), x_0 - x^\star\rangle + \frac{L}{2}\|x_0 - x^\star\|_2^2 = \frac{L}{2}\|x_0 - x^\star\|_2^2.$$

It was also shown on the homework that when f is convex and L-smooth we have $\min_{0\le t\le k}\|\nabla f(x_t)\|_2^2 = \|\nabla f(x_k)\|_2^2$ (using the cocoercivity property of ∇f).

We can easily generalize this result to projected gradient descent with the help of the following lemma. (We leave the details of this generalization as an exercise.)

Lemma 10. *If T_1 is θ_1-averaged and T_2 is θ_2-averaged, then T_1T_2 is $\theta_1 + \theta_2 - \theta_1\theta_2$ averaged.*

Proof. We have $T_1 = (1-\theta_1)\mathrm{Id} + \theta_1 H_1$ and $T_2 = (1-\theta_2)\mathrm{Id} + \theta_2 H_2$, where H_1, H_2 are nonexpansive. Then

$$\begin{aligned} T_1T_2 &= (1-\theta_1)T_2 + \theta_1 H_1 T_2 = (1-\theta_1)(1-\theta_2)\mathrm{Id} + (1-\theta_1)\theta_2 H_2 + \theta_1 H_1 T_2 \\ &= (1-(\theta_1+\theta_2-\theta_1\theta_2))\mathrm{Id} + (\theta_1+\theta_2-\theta_1\theta_2)\frac{(1-\theta_1)\theta_2 H_2 + \theta_1 H_1 T_2}{(\theta_1+\theta_2-\theta_1\theta_2)} \end{aligned}$$

so it remains to show that the second term is nonexpansive, which follows if we can show that $(1-\theta_1)\theta_2 H_2 + \theta_1 H_1 T_2$ is $(\theta_1+\theta_2-\theta_1\theta_2)$-Lipschitz. We then can compute that

$$\begin{aligned} \|(1-\theta_1)\theta_2 H_2 x + \theta_1 H_1 T_2 x - (1-\theta_1)\theta_2 H_2 y - \theta_1 H_1 T_2 y\| &\le (1-\theta_1)\theta_2\|H_2 x - H_2 y\| + \theta_1\|H_1T_2x - H_1T_2y\| \\ &\le ((1-\theta_1)\theta_2 + \theta_1)\|x-y\| \end{aligned}$$

as desired, using the facts that H_1, H_2, and T_2 are nonexpansive. □

Example 10 (Method of alternating projections)**.** *Suppose we are given two closed convex sets $\mathcal{C}_1$ and $\mathcal{C}_2$ such that $\mathcal{C}_1 \cap \mathcal{C}_2$ is nonempty. Given access to the two projection oracles $P_{\mathcal{C}_1}$ and $P_{\mathcal{C}_2}$, can we find a point $x \in \mathcal{C}_1 \cap \mathcal{C}_2$?*

The method of alternating projections generates the sequence

$$x_{k+1} = P_{\mathcal{C}_1} P_{\mathcal{C}_2} x_k.$$

Since each of $P_{\mathcal{C}_1}$ and $P_{\mathcal{C}_2}$ are averaged, by Lemma 10 so is their composition. Then we can immediately apply the KM Theorem.

5 More references

These notes barely scratch the surface of what algorithmic approaches are possible within the fixed-point methods framework. For more, some references are:

1. A Primer on Monotone Operator Methods by Ernest Ryu and Stephen Boyd
2. Large-Scale Convex Optimization: Algorithms & Analyses via Monotone Operators by Ernest Ryu and Wotao Yin.