Proximal Gradient Method Sequence Convergence

Pontus Giselsson

1 Introduction

We consider the problem of solving

$$\underset{x}{\text{minimize}}\, f(x) + g(x) \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable and $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is proper closed and convex. We also assume that a solution exists. A necessary and sufficient condition for $\bar{x}$ to be a solution is that

$$0 \in \partial(g + f)(x) = \partial g(x) + \nabla f(x).$$

This follows from Fermat's rule, the subdifferential sum rule (f has full domain) and that $\partial f(x) = \{\nabla f(x)\}$. Let us introduce the proximal gradient mapping

$$\text{prox}_{\gamma g}(I - \gamma \nabla f)$$

with $\gamma > 0$ and its fixed-point set F, defined as

$$F := \{x : x = \text{prox}_{\gamma}(x - \gamma \nabla f(x))\}.$$

The fixed-point set F is exactly the solution set to (1).

Proposition 1. *The set F satisfies*

$$F = \{x : 0 \in \partial g(x) + \nabla f(x)\}.$$

Proof. We have by Fermat's rule on the prox-grad step:

$$\begin{aligned} \bar{x} \in F \quad &\Leftrightarrow \quad \bar{x} = \text{prox}_{\gamma g}(\bar{x} - \gamma \nabla f(\bar{x})) \\ &\Leftrightarrow \quad 0 \in \partial g(\bar{x}) + \gamma(\bar{x} - (\bar{x} - \gamma \nabla f(\bar{x}))) \\ &\Leftrightarrow \quad 0 \in \partial g(\bar{x}) + \nabla f(\bar{x}) \end{aligned}$$

□

This implies that F is the same for all $\gamma > 0$. It also implies that solving (1) is equivalent to finding a fixed-point to the proximal gradient method. (This holds in the convex setting, in the nonconvex setting with nonconvex f, proximal gradient fixed-points are instead critical points.)

A fixed-point to the proximal gradient mapping can be found in many different ways. One option is to iterate the proximal gradient map gives, which gives the proximal gradient method. However, many other methods also exist.

We will base the sequence convergence results in this note on the following theorem.

Theorem 1. *Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable and that $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is proper closed convex. A sequence $(x_k)_{k\in\mathbb{N}}$ (not necessarily coming from the proximal gradient method) converges to a point in F (i.e., to a solution to (1)) if (and only if):*

(i) $\|\text{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k)) - x_k\| \to 0$ *as* $k \to \infty$

(ii) $(\|x_k - z\|)_{k\in\mathbb{N}}$ *converges for all* $z \in F$

We defer the proof of this result to Section 3. In the next section, we will show that the proximal gradient method satisfies these conditions in the convex setting under some step-size restrictions. Hence it generates a sequence that converges to a solution.

2 Convergence of proximal gradient method

We will provide two different results that hold in different settings. One is based on the descent lemma, and the other is based on operator theoretic arguments. The descent lemma approach is more general as it can be used to show convergence under more general assumptions (it can even tell us something in the nonconvex case), while the operator theoretic approach allows for larger step-size parameters. We provide two alternative proofs in the operator theoretic approach. Both results are based on Theorem 1.

The algorithm under consideration is the proximal gradient method

$$x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$$

where $\gamma_k > 0$ is a sequence of nonnegative scalar step-length variables. We assume that the following assumptions hold on the problem and algorithm parameters.

Assumption 1. *It holds that:*

(i) $f : \mathbb{R}^n \to \mathbb{R}$ *is convex and continuously differentiable*

(ii) For every x_k *and* x_{k+1} *there exists* $\beta_k \in [\eta, \eta^{-1}]$ *with* $\eta \in (0, 1)$*:*

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \tfrac{\beta_k}{2} \|x_k - x_{k+1}\|_2^2$$

where β_k *is a sort of local Lipschitz constant*

(iii) $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *is proper closed convex*

(iv) A minimizer exists (and $p^\star = \min_x (f(x) + g(x))$ *is optimal value)*

(v) Algorithm parameters $\gamma_k \in [\epsilon, \frac{2}{\beta_k} - \epsilon]$*, where* $\epsilon > 0$

We will show convergence of the proximal gradient method by proving that the generated sequence satisfies both conditions in Theorem 1 in the convex setting. The first condition holds even in the nonconvex setting. This stated below, and the proof does not rely on convexity of f.

Proposition 2. *Suppose that Assumption 1 holds. Then*

$$\|\text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k)) - x_k\| \to 0$$

as $k \to \infty$*.*

Proof. The proof uses:

(i) Assumption 1*(iii)*

(ii) Prox optimality condition: There exists $s_{k+1} \in \partial g(x_{k+1})$

$$0 = s_{k+1} + \gamma_k^{-1}(x_{k+1} - (x_k - \gamma_k \nabla f(x_k)))$$

(iii) Subgradient definition: $g(x_k) \geq g(x_{k+1}) + s_{k+1}^T (x_k - x_{k+1})$

The key inequality is

$$\begin{aligned}
f(x_{k+1}) + g(x_{k+1}) &\overset{(i)}{\leq} f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \tfrac{\beta_k}{2} \|x_{k+1} - x_k\|_2^2 + g(x_{k+1}) \\
&\overset{(iii)}{\leq} f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \tfrac{\beta_k}{2} \|x_{k+1} - x_k\|_2^2 + g(x_k) \\
&\quad - s_{k+1}^T (x_k - x_{k+1}) \\
&\overset{(ii)}{=} f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \tfrac{\beta_k}{2} \|x_{k+1} - x_k\|_2^2 + g(x_k) \\
&\quad + \gamma_k^{-1}(x_{k+1} - (x_k - \gamma_k \nabla f(x_k)))^T (x_k - x_{k+1}) \\
&= f(x_k) + g(x_k) - (\gamma_k^{-1} - \tfrac{\beta_k}{2}) \|x_{k+1} - x_k\|_2^2
\end{aligned}$$

Since $\gamma_k \in [\epsilon, \frac{2}{\beta_k} - \epsilon]$ (where $\epsilon > 0$ can be chosen such that the set is nonempty since $\beta_k \in [\eta, \eta^{-1}]$), there exists $\delta > 0$ such that

$$\gamma_k^{-1} \in [\tfrac{\beta_k}{2} + \delta, \delta^{-1}] \qquad \Rightarrow \qquad \gamma_k^{-1} - \tfrac{\beta_k}{2} > \delta > 0$$

Rearranging the inequality and using this gives

$$\delta \|x_{k+1} - x_k\|_2^2 \leq f(x_k) + g(x_k) - (f(x_{k+1}) + g(x_{k+1}))$$

Telescope summation gives for all $n \in \mathbb{N}$:

$$\begin{aligned}\delta \sum_{k=1}^{n} \|x_{k+1} - x_k\|_2^2 &\leq \sum_{k=1}^{n} (f(x_k) + g(x_k) - (f(x_{k+1}) + g(x_{k+1}))) \\ &= f(x_1) + g(x_1) - (f(x_{n+1}) + g(x_{n+1})) \\ &\leq f(x_1) + g(x_1) - p^\star < \infty\end{aligned}$$

where $p^\star = \min_x(f(x) + g(x))$ and $< \infty$ since $x_1 \in \mathrm{dom} g$. Since $\delta > 0$, this implies:

$$\|\mathrm{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k)) - x_k\| = \|x_{k+1} - x_k\| \to 0.$$

□

2.1 Convergence proof based on descent lemma

To show that the second condition in Theorem 1 holds based on the descent lemma, we need to half the allowed step-length range. The following result shows convergence. The proof is similar to the proof of Proposition 2 but uses in addition convexity of f.

Theorem 2. *Suppose that Assumption 1 holds and that in addition $\gamma_k \in [\epsilon, \frac{1}{\beta_k}]$. Then $(\|x_k - z\|)_{k\in\mathbb{N}}$ converges for all z in the fixed-point set F.*

Proof. The proof uses:

(i) Assumption 1*(iii)*

(ii) Prox optimality condition: There exists $s_{k+1} \in \partial g(x_{k+1})$

$$0 = s_{k+1} + \gamma_k^{-1}(x_{k+1} - (x_k - \gamma_k \nabla f(x_k)))$$

(iii) Subgradient definition: $g(z) \geq g(x_{k+1}) + s_{k+1}^T(z - x_{k+1})$ for all $z \in F$

(iv) Convexity of differentiable f: $f(z) \geq f(x_k) + \nabla f(x_k)^T(z - x_k)$ for all $z \in F$

(v) That: $(x_{k+1} - x_k)^T(z - x_{k+1}) = \frac{1}{2}(\|x_k - z\|_2^2 - \|x_{k+1} - z\|_2^2 - \|x_k - x_{k+1}\|_2^2)$

The key inequality is

$$\begin{aligned}f(x_{k+1}) + g(x_{k+1}) &\overset{(i)}{\leq} f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(x_{k+1}) \\ &\overset{(iii)}{\leq} f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(z) \\ &\quad - s_{k+1}^T(z - x_{k+1}) \\ &\overset{(ii)}{=} f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(z) \\ &\quad + \gamma_k^{-1}(x_{k+1} - (x_k - \gamma_k \nabla f(x_k)))^T(z - x_{k+1}) \\ &\overset{(iv)}{\leq} f(z) - \nabla f(x_k)^T(z - x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) \\ &\quad + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(z) + \gamma_k^{-1}(x_{k+1} - (x_k - \gamma_k \nabla f(x_k)))^T(z - x_{k+1}) \\ &= f(z) + g(z) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + \gamma_k^{-1}(x_{k+1} - x_k)^T(z - x_{k+1}) \\ &\overset{(v)}{=} f(z) + g(z) + \tfrac{\gamma_k^{-1}}{2}\|x_k - z\|_2^2 - \tfrac{1}{2}(\gamma_k^{-1} - \beta_k)\|x_{k+1} - x_k\|_2^2 \\ &\quad - \tfrac{\gamma_k^{-1}}{2}\|x_{k+1} - z\|_2^2\end{aligned}$$

Rearranging the inequality and multiplying by $2\gamma_k > 0$ gives

$$\begin{aligned}\|x_{k+1} - z\|_2^2 &\leq \|x_k - z\|_2^2 - (1 - \beta_k\gamma_k)\|x_{k+1} - x_k\|_2^2 \\ &\quad + 2\gamma_k((f(z) + g(z)) - (f(x_{k+1}) + g(x_{k+1}))) \\ &\leq \|x_k - z\|_2^2 - (1 - \beta_k\gamma_k)\|x_{k+1} - x_k\|_2^2\end{aligned}$$

since $\gamma_k \in [\epsilon, \frac{1}{\beta_k}]$ (where $\epsilon > 0$ can be chosen such that the set is nonempty since $\beta_k \in [\eta, \eta^{-1}]$). This concludes the proof. □

Proposition 2 and Theorem 2 together guarantee that the conditions of Theorem 1 are satisfied for sequences generated by the proximal gradient method under the assumption stated in Theorem 2.

2.2 Convergence proof based on operator properties

In this section, we keep the step-length requirement from Assumption 1 but instead restrict the assumption on f. We will assume that f is β-smooth, which is stronger than Assumption 1 *(iii)*. For β-smooth f, the gradient is β^{-1}-cocoercive (see below). First we provide a characterization of cocoercive operators.

Proposition 3. *Let $C : \mathbb{R}^n \to \mathbb{R}^n$ be β-cocoercive, i.e.,*

$$(Cx - Cy)^T(x - y) \geq \beta\|Cx - Cy\|_2^2$$

holds for all $x, y \in \mathbb{R}^n$. Then $C = \frac{1}{2\beta}(I + N)$ for some nonexpansive (1-Lipschitz) operator $N : \mathbb{R}^n \to \mathbb{R}^n$, i.e., an N that satisfies

$$\|Nx - Ny\|_2 \leq \|x - y\|_2$$

for all $x, y \in \mathbb{R}^n$.

Proof. That $N = 2\beta C - I$ is equivalent to that $C = \frac{1}{2\beta}(I + N)$. We therefore need to show that $2\beta C - I$ is nonexpansive. The definition

$$\begin{aligned}0 &\geq \beta\|Cx - Cy\|_2^2 - (Cx - Cy)^T(x - y) \\ &= \beta\|Cx - Cy\|_2^2 + \tfrac{1}{2}(-\|\sqrt{2\beta}(Cx - Cy)\|_2^2 - \|\sqrt{\tfrac{1}{2\beta}}(x - y)\|_2^2 + \|\sqrt{2\beta}(Cx - Cy) - \sqrt{\tfrac{1}{2\beta}}(x - y)\|_2^2) \\ &= \beta\|Cx - Cy\|_2^2 + \tfrac{1}{2}(-2\beta\|(Cx - Cy)\|_2^2 - \tfrac{1}{2\beta}\|(x - y)\|_2^2 + \tfrac{1}{2\beta}\|2\beta(Cx - Cy) - (x - y)\|_2^2) \\ &= -\tfrac{1}{4\beta}\|(x - y)\|_2^2 + \tfrac{1}{4\beta}\|(2\beta C - I)x - (2\beta C - Iy)\|_2^2,\end{aligned}$$

which is equivalent to that

$$\|(x - y)\|_2 \leq \|(2\beta C - I)x - (2\beta C - Iy)\|_2,$$

i.e. $2\beta C - I$ is nonexpansive. □

The result relies on the following properties of gradients of smooth convex functions and of proximal operators.

Proposition 4. *The gradient ∇f of a β-smooth and convex function $f : \mathbb{R}^n \to \mathbb{R}$ is β^{-1}-cocoercive, i.e., it satisfies*

$$(\nabla f(x) - \nabla f(y))^T(x - y) \geq \beta^{-1}\|\nabla f(x) - \nabla f(y)\|_2^2$$

for all $x, y \in \mathbb{R}^n$. Further $(I - \gamma\nabla f)$ is nonexpansive (1-Lipschitz) if $\gamma \in (0, \frac{2}{\beta})$.

Proof. The first result on cocoercivity can be found in the note on duality correspondence. Next, we prove the Lipschitz constant of $(I-\gamma\nabla f)$. We invoke Proposition 3 and let N be the nonexpansive such that $\nabla f = \frac{\beta}{2}(I+N)$. Then $I-\gamma\nabla f = I-\frac{\gamma\beta}{2}(I+N) = (1-\frac{\gamma\beta}{2})I-\frac{\gamma\beta}{2}N$. Now, let $\gamma \in (0, \frac{2}{\beta})$, then

$$\begin{aligned}\|(I-\gamma\nabla f)(x)-(I-\gamma\nabla f)(y)\|_2 &= \|(1-\tfrac{\gamma\beta}{2})x-\tfrac{\gamma\beta}{2}Ny-(1-\tfrac{\gamma\beta}{2})y-\tfrac{\gamma\beta}{2}Ny\|_2\\ &\leq \|(1-\tfrac{\gamma\beta}{2})(x-y)\|_2+\|\tfrac{\gamma\beta}{2}(Nx-Ny)\|_2\\ &\leq (1-\tfrac{\gamma\beta}{2})\|(x-y)\|_2+\tfrac{\gamma\beta}{2}\|(Nx-Ny)\|_2\\ &\leq \|x-y\|_2,\end{aligned}$$

where nonexpansiveness of N has been used in the last inequality. □

Cocercivity of ∇f does not hold if f is nonconvex. We will provide two proofs of the sequence convergences result in this section. The first uses the inequality that defines the cocoercive operator ∇f. The second uses that the gradient mapping is nonexpansive. This second proof becomes simpler. Both proofs also use the following result.

Proposition 5. *The proximal operator of a closed convex functions $g : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is 1-cocoercive, i.e., it satisfies*

$$(\text{prox}_{\gamma g}(x)-\text{prox}_{\gamma g}(y))^T(x-y) \geq \|\text{prox}_{\gamma g}(x)-\text{prox}_{\gamma g}(y)\|_2^2$$

for all $x, y \in \mathbb{R}^n$. Consequently it is 1-Lipschitz continuous (nonexpansive), i.e.,

$$\|\text{prox}_{\gamma g}(x)-\text{prox}_{\gamma g}(y)\|_2 \leq \|x-y\|_2$$

for all $x, y \in \mathbb{R}^n$.

Proof. We claim that $\text{prox}_{\gamma g} = \nabla r^*$, where $r(x) = \gamma g(x) + \frac{1}{2}\|x\|_2^2$. First, we note that r is 1-strongly convex, hence r^* is differentiable, ∇r^* is 1-cocoercive, hence 1-Lipschitz. The claim follows since:

$$\begin{aligned}\text{prox}_{\gamma g}(z) &= \text{argmin}(g(x)+\tfrac{1}{2\gamma}\|x-z\|_2^2)\\ &= \text{argmax}(-\gamma g(x)-\tfrac{1}{2}\|x-z\|_2^2)\\ &= \text{argmax}(z^Tx-(\tfrac{1}{2}\|x\|_2^2+\gamma g(x)))\\ &= \text{argmax}(z^Tx-r(x))\\ &= \nabla r^*(z)\end{aligned}$$

where we have used the subgradient formula for r^* and that $\partial r^*(z) = \{\nabla r^*(z)\}$ due to convexity of the conjugate. □

Theorem 3. *Suppose that Assumption 1 holds and that in addition f is β-smooth, i.e., $\gamma_k \in [\epsilon, \frac{2}{\beta}-\epsilon]$. Then $(\|x_k - z\|)_{k\in\mathbb{N}}$ converges for all z in the fixed-point set F.*

Proof. For all $z \in F$, it holds that

$$\begin{aligned}\|x_{k+1}-z\|_2^2 &= \|\text{prox}_{\gamma_k g}(x_k-\gamma_k\nabla f(x_k))-\text{prox}_{\gamma_k g}(z-\gamma_k\nabla f(z))\|_2^2\\ &\leq \|(x_k-\gamma_k\nabla f(x_k))-(z-\gamma_k\nabla f(z))\|_2^2\\ &= \|x_k-z\|_2^2-2\gamma_k(\nabla f(x_k)-\nabla f(z))^T(x_k-z)+\gamma_k^2\|\nabla f(x_k)-\nabla f(z))\|_2^2\\ &\leq \|x_k-z\|_2^2-\tfrac{2\gamma_k}{\beta}\|\nabla f(x_k)-\nabla f(z))\|_2^2+\gamma_k^2\|\nabla f(x_k)-\nabla f(z))\|_2^2\\ &\leq \|x_k-z\|_2^2-\gamma_k(\tfrac{2}{\beta}-\gamma_k)\|\nabla f(x_k)-\nabla f(z))\|_2^2\end{aligned}$$

where we have used Propositions 4 and 5 in the inequalities. This proves the claim. □

We also provide an alternative proof based on operator properties.

Proof. Since $\text{prox}_{\gamma_k g}$ and $(I-\gamma_k \nabla f)$ for $\gamma_k \in [\epsilon, \frac{2}{\beta}-\epsilon]$ are nonexpansive (Propositions 5 and 4), it holds for all $x, y \in \mathbb{R}^n$ that

$$\|\text{prox}_{\gamma_k g}(I-\gamma_k \nabla f)x - \text{prox}_{\gamma_k g}(I-\gamma_k \nabla f)y\|_2 \leq \|(I-\gamma_k \nabla f)x - (I-\gamma_k \nabla f)y\|_2 \leq \|x-y\|_2.$$

Let in particular $x = x_k$ (which implies $x_{k+1} = \text{prox}_{\gamma_k g}(I-\gamma_k \nabla f)x$) and $y = z$ to get the result. □

This second proof show the stronger property that the proximal gradient operator is nonexpansive.

Proposition 2 and Theorem 3 together guarantee that the conditions of Theorem 1 are satisfied for sequences generated by the proximal gradient method under the assumption stated in Theorem 3.

2.3 Strong convexity

In this section, we repeat the above results in the strongly convex setting. The computations are very similar to in the convex case. There are two properties that make the analysis simpler. The solution set to (1) is a singleton, i.e., the solution is unique. The method sequence is a *contraction* towards the solution. This is also referred to as that the algorithm converges linearly (or geometrically). In such cases, sequence convergence does not rely on Theorem 1.

The two convergence results are respectively based on the descent lemma and an operator theoretic approach. The operator theoretic approach shows linear convergence for a larger set of algorithm parameters and it provides a better (smaller) linear convergence rate. Again, we provide two different proofs in the operator theoretic approach.

2.3.1 Convergence proof based on descent lemma

We assume in addition to what is assumed in Theorem 2 that f is σ-strongly convex for some $\sigma > 0$. The proof is very similar to the proof of Theorem 2 but uses strong convexity of f.

Theorem 4. *Suppose that Assumption 1 holds and that in addition $\gamma_k \in [\epsilon, \frac{1}{\beta_k}]$ and that f is σ-strongly convex with $\sigma > 0$. Then the sequence $(x_k)_{k\in\mathbb{N}}$ generated by the proximal gradient method converges to the unique solution $z \in F$ with linear rate as follows:*

$$\|x_{k+1}-z\|_2^2 \leq \tfrac{1-\gamma_k\sigma}{1+\gamma_k\sigma}\|x_k-z\|_2^2 = \left(1-\tfrac{2\gamma_k\sigma}{1+\gamma_k\sigma}\right)\|x_k-z\|_2^2.$$

Proof. The proof uses:

(i) Assumption 1*(iii)*

(ii) Prox optimality condition: There exists $s_{k+1} \in \partial g(x_{k+1})$

$$0 = s_{k+1} + \gamma_k^{-1}(x_{k+1} - (x_k - \gamma_k \nabla f(x_k)))$$

(iii) Subgradient definition: $g(z) \geq g(x_{k+1}) + s_{k+1}^T(z - x_{k+1})$ for all $z \in F$

(iv) Strong convexity of differentiable f: $f(z) \geq f(x_k) + \nabla f(x_k)^T(z-x_k) + \frac{\sigma}{2}\|x_k - z\|_2^2$ for all $z \in F$

(v) That: $(x_{k+1}-x_k)^T(z-x_{k+1}) = \frac{1}{2}(\|x_k-z\|_2^2 - \|x_{k+1}-z\|_2^2 - \|x_k-x_{k+1}\|_2^2)$

The key inequality is

$$\begin{aligned}
f(x_{k+1}) + g(x_{k+1}) &\overset{(i)}{\leq} f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(x_{k+1}) \\
&\overset{(iii)}{\leq} f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(z) \\
&\quad - s_{k+1}^T(z - x_{k+1}) \\
&\overset{(ii)}{=} f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(z) \\
&\quad + \gamma_k^{-1}(x_{k+1} - (x_k - \gamma_k \nabla f(x_k)))^T(z - x_{k+1}) \\
&\overset{(iv)}{\leq} f(z) - \nabla f(x_k)^T(z - x_k) - \tfrac{\sigma}{2}\|x_k - z\|_2^2 + \nabla f(x_k)^T(x_{k+1} - x_k) \\
&\quad + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 + g(z) + \gamma_k^{-1}(x_{k+1} - (x_k - \gamma_k \nabla f(x_k)))^T(z - x_{k+1}) \\
&= f(z) + g(z) + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|_2^2 - \tfrac{\sigma}{2}\|x_k - z\|_2^2 + \gamma_k^{-1}(x_{k+1} - x_k)^T(z - x_{k+1}) \\
&\overset{(v)}{=} f(z) + g(z) + \tfrac{1}{2}(\gamma_k^{-1} - \sigma)\|x_k - z\|_2^2 - \tfrac{1}{2}(\gamma_k^{-1} - \beta_k)\|x_{k+1} - x_k\|_2^2 \\
&\quad - \tfrac{\gamma_k^{-1}}{2}\|x_{k+1} - z\|_2^2
\end{aligned}$$

Rearranging the inequality and multiplying by $2\gamma_k > 0$ gives

$$\begin{aligned}
\|x_{k+1} - z\|_2^2 &\leq (1 - \gamma_k\sigma)\|x_k - z\|_2^2 - (1 - \beta_k\gamma_k)\|x_{k+1} - x_k\|_2^2 \\
&\quad + 2\gamma_k((f(z) + g(z)) - (f(x_{k+1}) + g(x_{k+1}))) \\
&\leq (1 - \gamma_k\sigma)\|x_k - z\|_2^2 + 2\gamma_k((f(z) + g(z)) - (f(x_{k+1}) + g(x_{k+1}))) \qquad (2)
\end{aligned}$$

where we have used $\gamma_k \in [\epsilon, \frac{1}{\beta_k}]$ (where $\epsilon > 0$ can be chosen such that the set is nonempty since $\beta_k \in [\eta, \eta^{-1}]$). Next use that $s \in \partial g(z)$ exists, i.e.,

$$g(x_{k+1}) \geq g(z) + s^T(x_{k+1} - z)$$

and that f is σ-strongly convex

$$f(x_{k+1}) \geq f(z) + \nabla f(z)^T(x_{k+1} - z) + \tfrac{\sigma}{2}\|z - x_{k+1}\|_2^2.$$

By subdifferential sum rule we have $0 \in \partial(f+g)(z) = \partial g(z) + \nabla f(z)$. Let $s \in \partial g(z)$ be such that $0 = s + \nabla f(z)$. Add the two previous inequalities to get

$$\begin{aligned}
f(x_{k+1}) + g(x_{k+1}) &\geq f(z) + g(z) + (\nabla f(z) + s)^T(x_{k+1} - z) + \tfrac{\sigma}{2}\|z - x_{k+1}\|_2^2 \\
&\geq f(z) + g(z) + \tfrac{\sigma}{2}\|z - x_{k+1}\|_2^2
\end{aligned}$$

Therefore, the inequality in (2) can be continued as

$$\begin{aligned}
\|x_{k+1} - z\|_2^2 &\leq (1 - \gamma_k\sigma)\|x_k - z\|_2^2 + 2\gamma_k((f(z) + g(z)) - (f(x_{k+1}) + g(x_{k+1}))) \\
&\leq (1 - \gamma_k\sigma)\|x_k - z\|_2^2 - \gamma_k\sigma\|x_{k+1} - z\|_2^2.
\end{aligned}$$

Rearranging gives

$$\|x_{k+1} - z\|_2^2 \leq \tfrac{1-\gamma_k\sigma}{1+\gamma_k\sigma}\|x_k - z\|_2^2 = \left(1 - \tfrac{2\gamma_k\sigma}{1+\gamma_k\sigma}\right)\|x_k - z\|_2^2$$

This concludes the proof. □

Theorem 4 directly guarantees that the sequence converges (linearly) to the unique fixed-point. Therefore, we do not need to rely on Theorem 1 to show sequence convergence.

2.3.2 Convergence proof based on operator properties

We assume in addition to what is assumed in Theorem 3 that f is σ-strongly convex for some $\sigma > 0$. The proof is very similar to the proof of Theorem 3 but uses strong convexity of f.

The result relies on the following properties of gradients of smooth strongly convex functions.

Proposition 6. *The gradient ∇f of a β-smooth and σ-strongly convex function $f : \mathbb{R}^n \to \mathbb{R}$ with $\beta \geq \sigma > 0$ satisfies*

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \tfrac{1}{\beta+\sigma} \|\nabla f(x) - \nabla f(y)\|_2^2 + \tfrac{\sigma\beta}{\beta+\sigma} \|x - y\|_2^2$$

for all $x, y \in \mathbb{R}^n$. Further $(I - \gamma \nabla f)$ is $\max(1 - \gamma\sigma, \gamma\beta - 1)$-contractive (Lipschitz with parameter less than 1) if $\gamma \in (0, \frac{2}{\beta})$.

Proof. Assume first that $\beta > \sigma$. We have f is β-smooth and σ-strongly convex if and only if $f - \frac{\sigma}{2}\|\cdot\|_2^2$ is convex and $(\beta - \sigma)$-smooth. This is equivalent to that $\nabla(f - \frac{\sigma}{2}\|\cdot\|_2^2) = \nabla f - \sigma I$ is $\frac{1}{\beta-\sigma}$ cocoercive (see note on duality correspondence), i.e.

$$\begin{aligned}(\nabla f(x) - \sigma x - (\nabla f(y) - \sigma y))^T (x - y) &\geq \tfrac{1}{\beta-\sigma} \|\nabla f(x) - \sigma x - (\nabla f(y) - \sigma y)\|_2^2 \\ &= \tfrac{1}{\beta-\sigma} \left(\|\nabla f(x) - \nabla f(y)\|_2^2 + \sigma^2 \|x - y\|_2^2 - 2\sigma (\nabla f(x) - \nabla f(y))^T (x - y) \right).\end{aligned}$$

Rearrangement and using $(1 + \frac{2\sigma}{\beta-\sigma}) = \frac{\beta+\sigma}{\beta-\sigma}$ gives

$$\tfrac{\beta+\sigma}{\beta-\sigma} (\nabla f(x) - \nabla f(y))^T (x - y) \geq \tfrac{1}{\beta-\sigma} \left(\|\nabla f(x) - \nabla f(y)\|_2^2 + \sigma^2 \|x - y\|_2^2 \right) + \sigma \|x - y\|_2^2$$

which is equivalent to

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \tfrac{1}{\beta+\sigma} \left(\|\nabla f(x) - \nabla f(y)\|_2^2 + \sigma^2 \|x - y\|_2^2 \right) + \sigma \tfrac{\beta-\sigma}{\beta+\sigma} \|x - y\|_2^2$$

which is equivalent to

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \tfrac{1}{\beta+\sigma} \left(\|\nabla f(x) - \nabla f(y)\|_2^2 + \beta\sigma \|x - y\|_2^2 \right).$$

To prove that $I - \gamma \nabla f$ is contractive, we start again from that $\nabla f - \sigma I$ is $\frac{1}{\beta-\sigma}$-cocoercive. We invoke Proposition 3 and let N be the nonexpansive such that $\nabla f - \sigma I = \frac{\beta-\sigma}{2}(I + N)$. Then $I - \gamma \nabla f = I - \frac{\gamma(\beta-\sigma)}{2}(I + N) - \sigma\gamma I = (1 - \frac{\gamma(\beta+\sigma)}{2})I - \frac{\gamma(\beta-\sigma)}{2} N$. Now, let $\gamma \in (0, \frac{2}{\beta})$, then

$$\begin{aligned}\|(I - \gamma \nabla f)(x) - (I - \gamma \nabla f)(y)\|_2 &= \|(1 - \tfrac{\gamma(\beta+\sigma)}{2})x - \tfrac{\gamma(\beta-\sigma)}{2} Nx - (1 - \tfrac{\gamma(\beta+\sigma)}{2})y - \tfrac{\gamma(\beta-\sigma)}{2} Ny\|_2 \\ &\leq \|(1 - \tfrac{\gamma(\beta+\sigma)}{2})(x - y)\|_2 + \|\tfrac{\gamma(\beta-\sigma)}{2}(Nx - Ny)\|_2 \\ &\leq |(1 - \tfrac{\gamma(\beta+\sigma)}{2})| \|(x - y)\|_2 + \tfrac{\gamma(\beta-\sigma)}{2} \|Nx - Ny\|_2 \\ &\leq (|(1 - \tfrac{\gamma(\beta+\sigma)}{2})| + \tfrac{\gamma(\beta-\sigma)}{2}) \|x - y\|_2 \\ &\leq \max(1 - \gamma\sigma, \gamma\beta - 1) \|x - y\|_2.\end{aligned}$$

To prove the case $\beta = \sigma$, some extra care is needed to avoid division by 0. □

Also here, we will provide two proofs of the linear convergence result. The first uses the inequality describing ∇f, while the other uses that it is contractive. The second proof is simpler, and the first proof also needs following result, which we state without a proof.

Proposition 7. *The gradient ∇f of a differentiable and σ-strongly convex and β-smooth function $f : \mathbb{R}^n \to \mathbb{R}$ with $\beta > \sigma > 0$ satisfies*

$$\sigma \|x - y\|_2 \leq \|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2$$

for all $x, y \in \mathbb{R}^n$.

Theorem 5. *Suppose that Assumption 1 holds and that in addition f is β-smooth and σ-strongly convex and that $\gamma_k \in [\epsilon, \frac{2}{\beta} - \epsilon]$. Then the sequence $(x_k)_{k\in\mathbb{N}}$ generated by the proximal gradient method converges to the unique solution $z \in F$ with linear rate as follows:*

$$\|x_{k+1} - z\|_2 \leq \max(1 - \sigma\gamma_k, \beta\gamma_k - 1)\|x_k - z\|_2.$$

Proof. For all $z \in F$, it holds that

$$\begin{aligned}
\|x_{k+1} - z\|_2^2 &= \|\mathrm{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k)) - \mathrm{prox}_{\gamma_k g}(z - \gamma_k \nabla f(z))\|_2^2 \\
&\leq \|(x_k - \gamma_k \nabla f(x_k)) - (z - \gamma_k \nabla f(z))\|_2^2 \\
&= \|x_k - z\|_2^2 - 2\gamma_k(\nabla f(x_k) - \nabla f(z))^T(x_k - z) + \gamma_k^2\|\nabla f(x_k) - \nabla f(z))\|_2^2 \\
&\leq \|x_k - z\|_2^2 - \tfrac{2\gamma_k}{\beta+\sigma}(\|\nabla f(x_k) - \nabla f(z)\|_2^2 + \sigma\beta\|x_k - z\|_2^2) + \gamma_k^2\|\nabla f(x_k) - \nabla f(z))\|_2^2 \\
&\leq (1 - \tfrac{2\gamma_k\sigma\beta}{\beta+\sigma})\|x_k - z\|_2^2 - \gamma_k(\tfrac{2}{\beta+\sigma} - \gamma_k)\|\nabla f(x_k) - \nabla f(z))\|_2^2
\end{aligned}$$

where we have used Propositions 6 and 5 in the inequalities. Now, let us apply Proposition 7 to conclude that

$$-\gamma_k(\tfrac{2}{\beta+\sigma} - \gamma_k)\|\nabla f(x_k) - \nabla f(z)\|_2^2 \leq -\gamma_k(\tfrac{2}{\beta+\sigma} - \gamma_k)\|x_k - z\|_2^2 \begin{cases} \sigma^2 & \text{if } \gamma_k \in [\epsilon, \frac{2}{\beta+\sigma}] \\ \beta^2 & \text{if } \gamma_k \in (\frac{2}{\beta+\sigma}, \frac{2}{\beta} - \epsilon] \end{cases}$$

Therefore,

$$\begin{aligned}
\|x_{k+1} - z\|_2^2 &\leq \|x_k - z\|_2^2 \begin{cases} \left(1 - \frac{2\gamma_k\sigma\beta}{\beta+\sigma} - \gamma_k\sigma^2(\frac{2}{\beta+\sigma} - \gamma_k)\right) & \text{if } \gamma_k \in [\epsilon, \frac{2}{\beta+\sigma}] \\ \left(1 - \frac{2\gamma_k\sigma\beta}{\beta+\sigma} - \gamma_k\beta^2(\frac{2}{\beta+\sigma} - \gamma_k)\right) & \text{if } \gamma_k \in (\frac{2}{\beta+\sigma}, \frac{2}{\beta} - \epsilon] \end{cases} \\
&= \|x_k - z\|_2^2 \begin{cases} \left(1 - \frac{2\gamma_k\sigma(\beta+\sigma)}{\beta+\sigma} + \gamma_k^2\sigma^2\right) & \text{if } \gamma_k \in [\epsilon, \frac{2}{\beta+\sigma}] \\ \left(1 - \frac{2\gamma_k\beta(\beta+\sigma)}{\beta+\sigma} + \gamma_k^2\beta^2\right) & \text{if } \gamma_k \in (\frac{2}{\beta+\sigma}, \frac{2}{\beta} - \epsilon] \end{cases} \\
&= \|x_k - z\|_2^2 \begin{cases} \left(1 - 2\gamma_k\sigma + \gamma_k^2\sigma^2)\right) & \text{if } \gamma_k \in [\epsilon, \frac{2}{\beta+\sigma}] \\ \left(1 - 2\gamma_k\beta + \gamma_k^2\beta^2)\right) & \text{if } \gamma_k \in (\frac{2}{\beta+\sigma}, \frac{2}{\beta} - \epsilon] \end{cases} \\
&= \|x_k - z\|_2^2 \begin{cases} (1 - \gamma_k\sigma)^2 & \text{if } \gamma_k \in [\epsilon, \frac{2}{\beta+\sigma}] \\ (1 - \gamma_k\beta)^2 & \text{if } \gamma_k \in (\frac{2}{\beta+\sigma}, \frac{2}{\beta} - \epsilon] \end{cases}
\end{aligned}$$

Taking the square root we get

$$\|x_{k+1} - z\|_2 \leq \max(1 - \sigma\gamma_k, \beta\gamma_k - 1)\|x_k - z\|_2.$$

This proves the claim. □

We also provide a very short proof using operator properties of $\mathrm{prox}_{\gamma_k g}$ and $I - \gamma_k \nabla f$.

Proof. Since $\mathrm{prox}_{\gamma_k g}$ is nonexpansive (Proposition 5) and $(I - \gamma_k \nabla f)$ for $\gamma_k \in [\epsilon, \frac{2}{\beta} - \epsilon]$ is $\max(1 - \gamma_k\sigma, \gamma_k\beta - 1)$-Lipschitz (Proposition 6), it holds for all $x, y \in \mathbb{R}^n$ that

$$\begin{aligned}
\|\mathrm{prox}_{\gamma_k g}(I - \gamma_k \nabla f)x - \mathrm{prox}_{\gamma_k g}(I - \gamma_k \nabla f)y\|_2 &\leq \|(I - \gamma_k \nabla f)x - (I - \gamma_k \nabla f)y\|_2 \\
&\leq \max(1 - \gamma_k\sigma, \gamma_k\beta - 1)\|x - y\|_2.
\end{aligned}$$

Let in particular $x = x_k$ (which implies $x_{k+1} = \mathrm{prox}_{\gamma_k g}(I - \gamma_k \nabla f)x$) and $y = z$ to get the result. □

Theorem 5 directly guarantees that the sequence converges (linearly) to the unique fixed-point. Therefore, we do not need to rely on Theorem 1 to show sequence convergence.

3 Proof of Theorem 1

The proof of Theorem 1 relies on convergent subsequences. Let us first define a subsequence, and state a well known result on when a sequence possesses a convergence subsequence.

Let $(x_k)_{k\in\mathbb{N}}$ be an infinite sequence. A subsequence takes an infinite subset of elements in (x_k), e.g., all odd elements or all even elements. We denote a subsequence to (x_k) by $(x_{n_k})_{k\in\mathbb{N}}$. A subsequence $(x_{n_k})_{k\in\mathbb{N}}$ is convergent if it converges to some limit point, i.e, $x_{n_k} \to \bar{x}$. The full sequence does not necessarily converge if its convergent subsequences converge:

$$(x_k)_{k\in\mathbb{N}} = (-1^k)_{k\in\mathbb{N}} \text{ does not converge, while } (x_{2k})_{k\in\mathbb{N}} = (1)_{k\in\mathbb{N}} \text{ does}$$

The Bolzano-Weierstrass theorem implies the following. Assume that $(x_k)_{k\in\mathbb{N}}$ with $x_k \in \mathbb{R}^n$ is a bounded sequence. Then at least one convergent subsequence exists. We denote by $\mathcal{B}$ the set of limit points of convergent subsequences to $(x_k)_{k\in\mathbb{N}}$, i.e.,

$$\hat{x} \in \mathcal{B} \quad \Leftrightarrow \quad \text{there exists a subsequence } (x_{n_k})_{k\in\mathbb{N}} \text{ with } x_{n_k} \to \hat{x}$$

The Bolzano-Weierstrass theorem implies that $\mathcal{B} \neq \emptyset$ if the sequence $(x_k)_{k\in\mathbb{N}}$ is bounded.

The roadmap to proving Theorem 1 is that we first note that a convergent subsequence exits since the second condition implies that the sequence is bounded, i.e., $\mathcal{B} \neq \emptyset$. The next step is to show that any convergent (sub)sequence that satisfies the first condition in the theorem has a limit point in the set F, i.e. $\mathcal{B} \subseteq F$. The last part applies Browder's Lemma.

Lemma 1. *Assume that S is a nonempty set and that $(x_k)_{k\in\mathbb{N}}$ is a sequence such that $(\|x_k - z\|)_{k\in\mathbb{N}}$ converges for all $z \in S$. Then*

$$x_k \to \bar{x} \in S \qquad \Leftrightarrow \qquad \mathcal{B} \subseteq S$$

After we have proven this lemma, it is enough to show that for any convergent subsequence, its limit point is in F. We first prove Browder's lemma.

Proof. A sequence converges if and only if it is bounded and $\mathcal{B}$ is a singleton. By assumption, $(x_k)_{k\in\mathbb{N}}$ is bounded and $\mathcal{B}$ is nonempty. It is therefore enough to show that $\mathcal{B}$ cannot contain two distinct points. Let $\bar{x}, \bar{y} \in \mathcal{B}$ and assume they are different, i.e., $\bar{x} \neq \bar{y}$. Denote the corresponding subsequences $x_{l_k} \to \bar{x}$ and $x_{n_k} \to \bar{y}$. Since $\bar{x}$ and $\bar{y}$ belong to S, the sequences $(\|x_k - \bar{x}\|)_{k\in\mathbb{N}}$ and $(\|x_k - \bar{y}\|)_{k\in\mathbb{N}}$ converge. Now, since for all $k \in \mathbb{N}$,

$$2x_k^T(\bar{x} - \bar{y}) = \|x_k - \bar{y}\|_2^2 - \|x_k - \bar{x}\|_2^2 + \|\bar{x}\|_2^2 - \|\bar{y}\|_2^2,$$

the sequence $(2x_k^T(\bar{x} - \bar{y}))_{k\in\mathbb{N}}$ converges too. We let $2x_k^T(\bar{x} - \bar{y}) \to \theta$. Taking the limit along x_{l_k} and x_{n_k} gives

$$0 = \theta - \theta = \lim_{k\to\infty} 2x_{l_k}^T(\bar{x} - \bar{y}) - \lim_{k\to\infty} 2x_{n_k}^T(\bar{x} - \bar{y}) = \bar{x}^T(\bar{x} - \bar{y}) - \bar{y}^T(\bar{x} - \bar{y}) = \|\bar{x} - \bar{y}\|_2.$$

Hence $\bar{x} = \bar{y}$ and $\mathcal{B}$ is a singleton. This concludes the proof. □

We next finish the proof by showing that the limit points of all convergent subsequences belong to F. To prove this, we need the following results.

Proposition 8. *Suppose that $S \subset \mathbb{R}^n$ is a nonempty closed set and that $(x_k)_{k\in\mathbb{N}}$ is a convergent sequence $x_k \to \bar{x}$ and all $x_k \in S$. Then $\bar{x} \in S$.*

Proposition 9. *The graph of the subdifferential ∂g of a closed convex function g, i.e., the set*

$$\text{gph}\partial g = \{(x, u) : u \in \partial g(x)\},$$

is a closed set.

We use these results to show that any convergent sequence in the graph of ∂g converges to a point in the graph.

Corollary 1. *Suppose that g is closed convex and that $(x_k, u_k)_{k\in\mathbb{N}}$ converges $(x_k u_k) \to (\bar{x}, \bar{u})$ and that all $(x_k, u_k) \in \text{gph}\partial g$. Then $(\bar{x}, \bar{u}) \in \text{gph}\partial g$, i.e., $\bar{u} \in \partial g(\bar{x})$.*

Proof. Apply Propositions 8 and 9. □

Let $(x_{n_k})_{k\in\mathbb{N}}$ be any convergent (sub)sequence with limit point $\bar{x}$ and define $v_{n_k} = \text{prox}_{\gamma_{n_k} g}(x_{n_k} - \gamma_{n_k}\nabla f(x_{n_k}))$, which by assumption satisfies $\|v_{n_k} - x_{n_k}\| \to 0$ as $k \to \infty$. The proximal gradient optimality condition is

$$\partial g(v_{n_k}) + \nabla f(x_{n_k}) \ni \gamma_k^{-1}(x_{n_k} - v_{n_k})$$

Let $w_{n_k} = \gamma_k^{-1}(x_{n_k} - v_{n_k}) - \nabla f(x_{n_k})$ and

$$w_{n_k} \in \partial g(v_{n_k}) \qquad \Leftrightarrow \qquad (v_{n_k}, w_{n_k}) \in \text{gph}\partial g.$$

Now, since $\|v_{n_k} - x_{n_k}\| \to 0$ and $x_{n_k} \to \bar{x}$, we conclude

$$\begin{aligned} v_{n_k} &= v_{n_k} - x_{n_k} + x_{n_k} \to 0 + \bar{x} = \bar{x} \\ w_{n_k} &= \gamma_k^{-1}(x_{n_k} - v_{n_k}) - \nabla f(x_{n_k}) \to 0 - \nabla f(\bar{x}) = -\nabla f(\bar{x}) \end{aligned}$$

where the second conclusion follows from uniform upper and lower boundedness of γ_k and continuity of ∇f. Therefore $(v_{n_k}, w_{n_k}) \in \text{gph}\partial g$ satisfies $(v_{n_k}, w_{n_k}) \to (\bar{x}, -\nabla f(\bar{x}))$. An appeal to Corollary 1 implies that $(\bar{x}, -\nabla f(\bar{x})) \in \text{gph}\partial g$, or equivalently $-\nabla f(\bar{x}) \in \partial g(\bar{x})$ which is equivalent to that $\bar{x} \in F$.

This concludes the proof.