# Chapter 3

# Projected and Proximal Gradient Descent

## Contents

## 3.1 The Algorithm

Another way to control gradients in (2.5) is to minimize $f$ over a closed convex subset $X \subseteq \mathbb{R}^d$. For example, we may have a constrained optimization problem to begin with (for example the LASSO in Section 1.6.2), or we happen to know some region $X$ containing a global minimum $\mathbf{x}^\star$, so that we can restrict our search to that region. In this case, gradient descent also works, but we need an additional *projection step*. After all, it can happen that some iteration of (2.1) takes us "into the wild" (out of $X$) where we have no business to do. *Projected* gradient descent is the following modification. We choose $\mathbf{x}_0 \in X$ arbitrary and for $t \geq 0$ define

$$\begin{aligned} \mathbf{y}_{t+1} &:= \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t), && (3.1) \\ \mathbf{x}_{t+1} &:= \Pi_X(\mathbf{y}_{t+1}) := \operatorname*{argmin}_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}_{t+1}\|^2. && (3.2) \end{aligned}$$

This means, after each iteration, we project the obtained iterate $\mathbf{y}_{t+1}$ back to $X$. This may be very easy (think of $X$ as the unit ball in which case we just have to scale $\mathbf{y}_{t+1}$ down to length 1 if it is longer). But it may also be very difficult. In general, computing $\Pi_X(\mathbf{y}_{t+1})$ means to solve an auxiliary convex constrained minimization problem in each step! Here, we're just assuming that we can do this. The projection is well-defined since $d_{\mathbf{y}} := \|\mathbf{x} - \mathbf{y}\|^2$ has bounded sublevel sets. Moreover, $d_{\mathbf{y}}(\mathbf{x})$ is strictly convex, so the minimum over $X$ (that exists by continuity of $d_{\mathbf{y}}$ and compactness of $X$ intersected with any nonempty sublevel set) is unique by Lemma 1.15.

## 3.2 Constrained minimization: $\mathcal{O}(1/\varepsilon^2)$ steps

To show that the vanilla analysis still goes through, we need the following

**Fact 3.1.** *Let $X \subseteq \mathbb{R}^d$ convex, $\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$. Then*

*(i)* $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top (\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$.

*(ii)* $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.

*Proof.* $\Pi_X(\mathbf{y})$ is by definition a minimizer of the (differentiable) convex function $d_{\mathbf{y}}(\mathbf{x}) = \|\mathbf{x} - \mathbf{y}\|^2$ over $X$, and (i) is just the equivalent optimality condition of Lemma 1.17. Part (ii) follows from (i) via the (by now well-known) equation $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$. □

If we minimize $f$ over a compact convex set $X$, we get the existence of a minimizer and a bound $R$ for the initial distance to it for free; assuming that $f$ is *continuously* differentiable, we also have a bound $L$ for the gradient norms over $X$. In this case, our vanilla analysis yields a much more useful result than the one in Theorem 2.1, with the same stepsize and the same number of steps.

**Theorem 3.2.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, $X \subseteq \mathbb{R}^d$ closed and convex, $\mathbf{x}^\star$ a minimizer of $f$ over $X$; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$, and that $\|\nabla f(\mathbf{x})\| \leq L$ for all $\mathbf{x} \in X$. Choosing the constant stepsize*

$$\gamma := \frac{R}{L\sqrt{T}},$$

*projected gradient descent (3.1) with $\mathbf{x}_0 \in X$ yields*

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{RL}{\sqrt{T}}.$$

*Proof.* The only required changes to the vanilla analysis are that in steps (2.3) and (2.4), $\mathbf{x}_{t+1}$ needs to be replaced by $\mathbf{y}_{t+1}$ as this is the real next (non-projected) gradient descent iterate after these steps, we therefore get

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2\right). \tag{3.3}$$

From Fact 3.1 (ii) (with $\mathbf{x} = \mathbf{x}^\star, \mathbf{y} = \mathbf{y}_{t+1}$), we obtain $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2$, hence we get

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right)$$

and return to the previous vanilla analysis for the remainder of the proof. □

## 3.3 Smooth constrained minimization: $\mathcal{O}(1/\varepsilon)$ steps

First, we define smoothness relative to a closed and convex subset $X \subseteq \mathbb{R}^d$.

**Definition 3.3.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, $L \in \mathbb{R}_+$. Furthermore, let $X \subseteq \mathbb{R}^d$ be a closed convex set. $f$ is called* smooth over $X$ *(with parameter $L$) if*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X. \tag{3.4}$$

For example, the globally non-smooth function $f(x) = x^4$ is smooth over any closed interval $X$, but with $L$ depending on $X$, see Exercise 16.

To minimize a smooth $f$ over $X$, we use projected gradient descent again. The runtime turns out to be the same as in the unconstrained case.

**Theorem 3.4.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Let $X \subseteq \mathbb{R}^d$ be a closed convex set, and assume that there is a minimizer $\mathbf{x}^\star$ of $f$ over $X$; furthermore, suppose that $f$ is smooth over $X$ with parameter $L$ according to (3.4). When choosing the stepsize*

$$\gamma := \frac{1}{L},$$

*projected gradient descent (3.1) with $\mathbf{x}_0 \in X$ satisfies the following two properties.*

*(i) Function values are monotone decreasing (this is not obvious from the following inequality, but you are asked to prove this in Exercise 19):*

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2, \quad t \geq 0.$$

*(ii)*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

*Proof.* For (i), we proceed similar to the proof of the "unconstrained" Theorem 2.6, except that we now need to deal with projected gradient descent. We again start from smoothness (3.4) but then use $\mathbf{y}_{t+1} = \mathbf{x}_t - \nabla f(\mathbf{x}_t)/L$,

followed by the usual equation $2\mathbf{v}^\top\mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2$:

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t+1}-\mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\ &= f(\mathbf{x}_t) - L(\mathbf{y}_{t+1}-\mathbf{x}_t)^\top(\mathbf{x}_{t+1}-\mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\ &= f(\mathbf{x}_t) - \frac{L}{2}\left(\|\mathbf{y}_{t+1}-\mathbf{x}_t\|^2 + \|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 - \|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2\right) \\ &\quad + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\ &= f(\mathbf{x}_t) - \frac{L}{2}\|\mathbf{y}_{t+1}-\mathbf{x}_t\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2. \end{aligned}$$

This proves (i). The plan is as in the proof of Theorem 2.6 to use the resulting inequality

$$\frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{L}{2}\sum_{t=0}^{T-1}\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2 \tag{3.5}$$

to control the sum of squared gradients in the bound (2.5) of the vanilla analysis. We have shown in the proof of Theorem 3.2 that (2.5) also holds in the constrained case, so this is a good start. Unfortunately, (3.5) now has an extra term compared to the bound (2.8) that we derived in the unconstrained case. To take care of this term, we observe that we can actually improve (2.5) in such a way that the extra term is absorbed. Let us go back to the "constrained" inequality (3.3)

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1}-\mathbf{x}^\star\|^2\right),$$

to which we then applied Fact 3.1(ii) to get back on the vanilla track. In doing so, we dropped a term that now becomes significant. In (3.3), Fact 3.1(ii) actually yields $\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2 + \|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2 \leq \|\mathbf{y}_{t+1}-\mathbf{x}^\star\|^2$ and not just $\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2 \leq \|\mathbf{y}_{t+1}-\mathbf{x}^\star\|^2$. Using this, we further get a critical improvement on the bound for $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ which is now

$$\frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2\right). \tag{3.6}$$

Summing this up for $t = 0, \dots, T-1$ (and using $\gamma = 1/L$), we get

$$\sum_{t=0}^{T-1}(f(\mathbf{x}_t)-f(\mathbf{x}^\star)) \leq \frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2+\frac{L}{2}\|\mathbf{x}_0-\mathbf{x}^\star\|^2-\frac{L}{2}\sum_{t=0}^{T-1}\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2.$$

Plugging in the bound (3.5) for the sum of squared gradients, the extra term of $\frac{L}{2}\sum_{t=0}^{T-1}\|\mathbf{y}_{t+1}-\mathbf{x}_{t+1}\|^2$ is exactly absorbed and we arrive at the bound (2.9) from which statement (ii) of the theorem follows as before. □

## 3.4 Strongly convex constrained minimization: $\mathcal{O}(\log(1/\varepsilon))$ steps

Assuming that $f$ is smooth *and* strongly convex over a set $X$, we can also prove fast convergence of projected gradient descent. This does not require any new ideas, we have seen all the ingredients before.

We first need to define strong convexity with respect to a set $X$ which reads as expected.

**Definition 3.5.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, $\mu \in \mathbb{R}_+, \mu > 0$. Furthermore, let $X \subseteq \mathbb{R}^d$ be a closed convex set. $f$ is called* strongly convex *(with parameter $\mu$) over $X$ if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X. \tag{3.7}$$

**Theorem 3.6.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Let $X \subseteq \mathbb{R}^d$ be a nonempty closed and convex set and suppose that $f$ is smooth over $X$ with parameter $L$ according to (3.4) and strongly convex over $X$ with parameter $\mu > 0$ according to (3.7). Exercise 20 asks you to prove that there is a unique minimizer $\mathbf{x}^\star$ of $f$ over $X$. Choosing*

$$\gamma := \frac{1}{L},$$

*projected gradient descent (3.1) with arbitrary $\mathbf{x}_0$ satisfies the following two properties.*

*(i) Squared distances to $\mathbf{x}^\star$ are geometrically decreasing:*

$$\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2 \leq \left(1-\frac{\mu}{L}\right)\|\mathbf{x}_t-\mathbf{x}^\star\|^2, \quad t \geq 0.$$

*(ii)*

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^t \|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

*Proof.* In the strongly convex case, the "constrained" vanilla bound (3.6)

$$\frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2\right)$$

on $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ can be strengthened to

$$\frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2\right) - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2 \tag{3.8}$$

Now we proceed as in the proof of Theorem 2.11. Rewriting the latter bound into a bound on $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2$ yields

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq 2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + (1 - \mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2.$$

Again, we show that the noise in this bound disappears. From Theorem 3.4 (i), we know that

$$f(\mathbf{x}^\star) - f(\mathbf{x}_t) \leq f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2,$$

and hence the noise can be bounded as follows, using $\gamma = 1/L$:

$$\begin{aligned} & 2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 \\ = \; & \frac{2}{L}(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 \\ \leq \; & -\frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 + \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 = 0. \end{aligned}$$

□

## 3.5 Projecting onto $\ell_1$-balls

$\ell_1$-regularized problems appear among the most commonly used models in machine learning and signal processing, and we have already discussed the Lasso as an important example of that class. We will now address how to perform projected gradient as an efficient optimization for

$\ell_1$-constrained problems. Let

$$X = B_1(R) := \left\{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_1 = \sum_{i=1}^{d} |x_i| \leq R \right\}$$

be the $\ell_1$-ball of radius $R > 0$ around $\mathbf{0}$, i.e. the set of all points with 1-norm at most $R$. Our goal is to compute $\Pi_X(\mathbf{v})$ for a given vector $\mathbf{v}$, i.e. the projection of $\mathbf{v}$ onto $X$; see Figure 3.1.

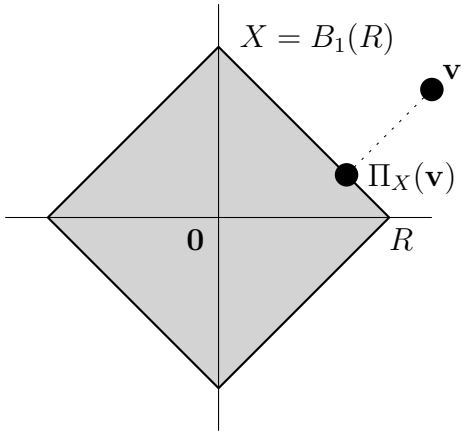


Figure 3.1: Projecting onto an $\ell_1$-ball

At first sight, this may look like a rather complicated task. Geometrically, $X$ is a *cross polytope* (square for $d = 2$, octahedron for $d = 3$), and as such it has $2^d$ many facets. But we can start with some basic simplifying observations.

**Fact 3.7.** *We may assume w.l.o.g. that (i) $R = 1$, (ii) $v_i \geq 0$ for all $i$, and (iii) $\sum_{i=1}^{d} v_i > 1$.*

*Proof.* If we project $\mathbf{v}/R$ onto $B_1(1)$, we obtain $\Pi_X(\mathbf{v})/R$ (just scale Figure 3.1), so we can restrict to the case $R = 1$. For (ii), we observe that simultaneously flipping the signs of a fixed subset of coordinates in both $\mathbf{v}$ and $\mathbf{x} \in X$ yields vectors $\mathbf{v}'$ and $\mathbf{x}' \in X$ such that $\|\mathbf{x} - \mathbf{v}\| = \|\mathbf{x}' - \mathbf{v}'\|$; thus, $\mathbf{x}$ minimizes the distance to $\mathbf{v}$ if and only if $\mathbf{x}'$ minimizes the distance to $\mathbf{v}'$. Hence, it suffices to compute $\Pi_X(\mathbf{v})$ for vectors with nonnegative entries. If $\sum_{i=1}^{d} v_i \leq 1$, we have $\Pi_X(\mathbf{v}) = \mathbf{v}$ and are done, so the interesting case is (iii). □

**Fact 3.8.** *Under the assumptions of Fact 3.7,* $\mathbf{x} = \Pi_X(\mathbf{v})$ *satisfies* $x_i \geq 0$ *for all* $i$ *and* $\sum_{i=1}^d x_i = 1$.

*Proof.* If $x_i < 0$ for some $i$, then $(-x_i - v_i)^2 \leq (x_i - v_i)^2$ (since $v_i \geq 0$), so flipping the $i$-th sign in $\mathbf{x}$ would yield another vector in $X$ at least as close to $\mathbf{v}$ as $\mathbf{x}$, but such a vector can't exist by strict convexity of the squared distance. And if $\sum_{i=1}^d x_i < 1$, then $\mathbf{x}' = \mathbf{x} + \lambda(\mathbf{v} - \mathbf{x}) \in X$ for some small positive $\lambda$, with $\|\mathbf{x}' - \mathbf{v}\| = (1-\lambda)\|\mathbf{x} - \mathbf{v}\|$, again contradicting the optimality of $\mathbf{x}$. □

**Corollary 3.9.** *Under the assumptions of Fact 3.7,*

$$\Pi_X(\mathbf{v}) = \operatorname*{argmin}_{\mathbf{x} \in \Delta_d} \|\mathbf{x} - \mathbf{v}\|^2,$$

*where*

$$\Delta_d := \left\{ \mathbf{x} \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x_i \geq 0 \ \forall i \right\}$$

*is the* standard simplex.

This means, we have reduced the projection onto an $\ell_1$-ball to the projection onto the standard simplex; see Figure 3.2.

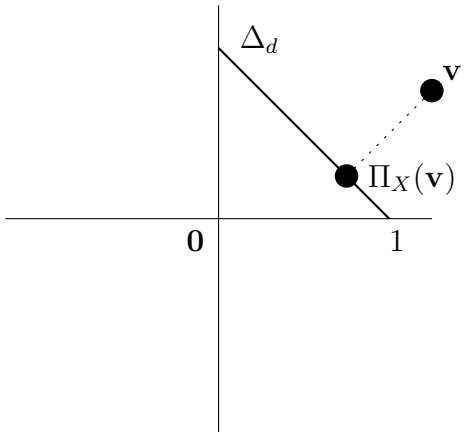


Figure 3.2: Projecting onto the standard simplex

To address the latter task, we make another assumption that can be established by suitably permuting the entries of $\mathbf{v}$ (which just permutes the entries of its projection onto $\Delta_d$ in the same way).

**Fact 3.10.** *We may w.l.o.g. assume that* $v_1 \geq v_2 \geq \cdots \geq v_d$.

**Lemma 3.11.** *Let* $\mathbf{x}^\star := \operatorname{argmin}_{\mathbf{x}\in\Delta_d} \|\mathbf{x}-\mathbf{v}\|^2$. *Under the assumption of Fact 3.10, there exists (a unique)* $p \in \{1,\ldots,d\}$ *such that*

$$\begin{aligned} x_i^\star &> 0, \quad i \leq p, \\ x_i^\star &= 0, \quad i > p. \end{aligned}$$

*Proof.* We are using the optimality criterion of Lemma 1.17:

$$\nabla d_{\mathbf{v}}(\mathbf{x}^\star)^\top(\mathbf{x}-\mathbf{x}^\star) = 2(\mathbf{x}^\star-\mathbf{v})^\top(\mathbf{x}-\mathbf{x}^\star) \geq 0, \quad \mathbf{x}\in\Delta_d. \tag{3.9}$$

By $\sum_{i=1}^d x_i^\star = 1$, there is at least one positive entry in $\mathbf{x}^\star$. It remains to show that we can't have $x_i^\star = 0$ and $x_{i+1}^\star > 0$. Indeed, in this situation, we could decrease $x_{i+1}^\star$ by some small positive $\varepsilon$ and simultaneously increase $x_i^\star$ to $\varepsilon$ to obtain a vector $\mathbf{x}\in\Delta_d$ such that

$$(\mathbf{x}^\star-\mathbf{v})^\top(\mathbf{x}-\mathbf{x}^\star) = (0-v_i)\varepsilon - (x_{i+1}^\star - v_{i+1})\varepsilon = \varepsilon(\underbrace{v_{i+1}-v_i}_{\leq 0} - \underbrace{x_{i+1}^\star}_{>0}) < 0,$$

contradicting the optimality (3.9). □

But we can say even more about $\mathbf{x}^\star$.

**Lemma 3.12.** *Under the assumption of Fact 3.10, and with* $p$ *as in Lemma 3.11,*

$$x_i^\star = v_i - \Theta_p, \quad i \leq p,$$

*where*

$$\Theta_p = \frac{1}{p}\Big(\sum_{i=1}^p v_i - 1\Big).$$

*Proof.* Suppose $x_i^\star - v_i < x_j^\star - v_j$ for some $i,j \leq p$. As before, we could then decrease $x_j^\star > 0$ by some small positive $\varepsilon$ and simultaneously increase $x_i^\star$ by $\varepsilon$ to obtain $\mathbf{x}\in\Delta_d$ such that

$$(\mathbf{x}^\star-\mathbf{v})^\top(\mathbf{x}-\mathbf{x}^\star) = (x_i^\star-v_i)\varepsilon - (x_j^\star-v_j)\varepsilon = \varepsilon(\underbrace{(x_i^\star-v_i)-(x_j^\star-v_j)}_{<0}) < 0,$$

again contradicting (3.9). The expression for $\Theta_p$ is then obtained from

$$1 = \sum_{i=1}^p x_i^\star = \sum_{i=1}^p (v_i - \Theta_p) = \sum_{i=1}^p v_i - p\Theta_p.$$

□

Let us summarize the situation: we now have $d$ candidates for $\mathbf{x}^\star$, namely the vectors

$$\mathbf{x}^\star(p) := (v_1 - \Theta_p, \dots, v_p - \Theta_p, 0, \dots, 0), \quad p \in \{1, \dots, d\}, \tag{3.10}$$

and we just need to find the right one. In order for candidate $\mathbf{x}^\star(p)$ to comply with Lemma 3.11, we must have

$$v_p - \Theta_p > 0, \tag{3.11}$$

and this actually ensures $\mathbf{x}^\star(p)_i > 0$ for all $i \leq p$ by the assumption of Fact 3.10 and therefore $\mathbf{x}^\star(p) \in \Delta_d$. But there could still be several values of $p$ satisfying (3.11). Among them, we simply pick the one for which $\mathbf{x}^\star(p)$ minimizes the distance to $\mathbf{v}$. It is not hard to see that this can be done in time $\mathcal{O}(d \log d)$, by first sorting $v$ and then carefully updating the values $\Theta_p$ and $\|\mathbf{x}^\star(p) - \mathbf{v}\|^2$ as we vary $p$ to check all candidates.

But actually, there is an even simpler criterion that saves us from comparing distances.

**Lemma 3.13.** *Under the assumption of Fact 3.10, with $\mathbf{x}^\star(p)$ as in (3.10), and with*

$$p^\star := \max\Big\{p \in \{1, \dots, d\} : v_p - \frac{1}{p}\Big(\sum_{i=1}^{p} v_i - 1\Big) > 0\Big\},$$

*it holds that*

$$\operatorname*{argmin}_{\mathbf{x} \in \Delta_d} \|\mathbf{x} - \mathbf{v}\|^2 = \mathbf{x}^\star(p^\star).$$

The proof is Exercise 21. Together with our previous reductions, we obtain the following result.

**Theorem 3.14.** *Let $\mathbf{v} \in \mathbb{R}^d$, $R \in \mathbb{R}_+$, $X = B_1(R)$ the $\ell_1$-ball around $\mathbf{0}$ of radius $R$. The projection*

$$\Pi_X(\mathbf{v}) = \operatorname*{argmin}_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{v}\|^2$$

*of $\mathbf{v}$ onto $B_1(R)$ can be computed in time $\mathcal{O}(d \log d)$.*

This can be improved to time $\mathcal{O}(d)$, based on the observation that a given $p$ can be compared to the value $p^\star$ in Lemma 3.13 in linear time, without the need to presort $\mathbf{v}$ [DSSSC08].

## 3.6 Proximal gradient descent

Many optimization problems in applications come with additional structure. An important class of objective functions is composed as

$$f(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x}) \tag{3.12}$$

where $g$ is a "nice" function, where as $h$ is a "simple" additional term, which however doesn't satisfy the assumptions of niceness which we used in the convergence analysis so far. In particular, an important case is when $h$ is not differentiable.

The classical gradient step for unconstrained minimization of a function $g$ can be equivalently written as

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^d} \; g(\mathbf{x}_t) + \nabla g(\mathbf{x}_t)^\top(\mathbf{y} - \mathbf{x}_t) + \frac{1}{2\gamma}\|\mathbf{y} - \mathbf{x}_t\|^2 \tag{3.13}$$

$$= \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^d} \; \frac{1}{2\gamma}\|\mathbf{y} - (\mathbf{x}_t - \gamma\nabla g(\mathbf{x}_t))\|^2 \,. \tag{3.14}$$

To obtain the last equality, we have just completed the quadratic $\|\mathbf{v}\|^2 + 2\mathbf{v}^\top\mathbf{w} + \|\mathbf{w}\|^2 = \|\mathbf{v} + \mathbf{w}\|^2$ for $\mathbf{v} := \gamma\nabla g(\mathbf{x}_t)$ and $\mathbf{w} := \mathbf{y} - \mathbf{x}_t$. Here it is crucial that $\mathbf{v}$ is independent of the optimization variable $\mathbf{y}$, so therefore the term can be ignored when taking the $\operatorname{argmin}$. The scaling by $\frac{1}{2\gamma}$ is also irrelevant but we keep it for better illustrating the next step.

The interpretation of the above equivalent reformulation of the classic gradient step is important for us, and is what has enabled the previous convergence analysis in Section 2.4 for smooth unconstrained optimization: For the particular choice of stepsize $\gamma := \frac{1}{L}$ which we have used, the above formulation shows that the gradient descent step exactly minimizes the local quadratic model of $g$ at our current iterate $\mathbf{x}_t$, formed by the smoothness property with parameter $L$ as defined in (2.7).

Our goal in this section is to minimize $f = g + h$, instead of only the smooth part $g$ alone. The idea of the proximal gradient method is to modify the simple quadratic model (3.13) above, so as to make it a valid model for $f$, that is a model which upper bounds $f$ at all points. The simplest way to do this is to just treat the $h$ function separately by adding it unmodified.

We obtain the update equation for *proximal gradient descent*

$$\mathbf{x}_{t+1} := \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^d} \; g(\mathbf{x}_t) + \nabla g(\mathbf{x}_t)^\top(\mathbf{y}-\mathbf{x}_t) + \frac{1}{2\gamma}\|\mathbf{y}-\mathbf{x}_t\|^2 + h(\mathbf{y}) \tag{3.15}$$

$$= \operatorname*{argmin}_{\mathbf{y}} \; \frac{1}{2\gamma}\|\mathbf{y}-(\mathbf{x}_t-\gamma\nabla g(\mathbf{x}_t))\|^2 + h(\mathbf{y}) \; . \tag{3.16}$$

The last formulation makes clear that the resulting update tries to combine the two goals, staying close to the classic gradient update, as well as also to minimize $h$.

### 3.6.1 The proximal gradient algorithm

We define the *proximal mapping* for a given function $h$, and parameter $\gamma > 0$:

$$\operatorname{prox}_{h,\gamma}(\mathbf{z}) := \operatorname*{argmin}_{\mathbf{y}} \left\{ \frac{1}{2\gamma}\|\mathbf{y}-\mathbf{z}\|^2 + h(\mathbf{y}) \right\}$$

An iteration of *proximal gradient descent* is defined as

$$\mathbf{x}_{t+1} := \operatorname{prox}_{h,\gamma}(\mathbf{x}_t - \gamma\nabla g(\mathbf{x}_t)) \; . \tag{3.17}$$

This same update step can also be written in different form as

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma G_\gamma(\mathbf{x}_t) \tag{3.18}$$

for $G_{h,\gamma}(\mathbf{x}) := \frac{1}{\gamma}\Big(\mathbf{x} - \operatorname{prox}_{h,\gamma}(\mathbf{x}-\gamma\nabla g(\mathbf{x}))\Big)$ being the so called generalized gradient of $f$.

**A generalization of gradient descent.** The proximal gradient descent method (3.17) is also known as generalized gradient descent. In the special case $h \equiv 0$, we of course recover classic gradient descent.

More interestingly, it is also a generalization of projected gradient descent as we have discussed in the previous sections. Given a closed convex set $X$, the *indicator function* of the set $X$ is given as the convex function

$$\begin{aligned} \boldsymbol{\iota}_X : \mathbb{R}^d &\to \mathbb{R}\cup+\infty \\ \mathbf{x} \mapsto \boldsymbol{\iota}_X(\mathbf{x}) &:= \begin{cases} 0 & \text{if } \mathbf{x}\in X, \\ +\infty & \text{otherwise.} \end{cases} \end{aligned} \tag{3.19}$$

When using the indicator function of our constraint set $X$ as $h \equiv \boldsymbol{\iota}_X$, it is easy to see that the proximal mapping simply becomes

$$\begin{aligned}\operatorname{prox}_{h,\gamma}(\mathbf{z}) &:= \operatorname*{argmin}_{\mathbf{y}} \left\{ \frac{1}{2\gamma}\|\mathbf{y}-\mathbf{z}\|^2 + \boldsymbol{\iota}_X(\mathbf{y}) \right\} \\ &= \operatorname*{argmin}_{\mathbf{y}\in X} \ \|\mathbf{y}-\mathbf{z}\|^2 = \Pi_X(\mathbf{z}) \ ,\end{aligned}$$

which is the projection of $\mathbf{z}$ onto $X$.

As we will see, the convergence of proximal gradient will be as fast as classic gradient descent. However, this still comes not entirely for free. In every iteration, we now have to additionally compute the proximal mapping. This can be very expensive if $h$ is complex. Nevertheless, for some important examples of $h$ the proximal mapping is efficient to compute, such as for the $\ell_1$-norm.

### 3.6.2 Convergence in $\mathcal{O}(1/\varepsilon)$ steps

Interestingly, the vanilla convergence analysis for smooth functions as in Theorem 2.6 directly applies for the more general case of proximal gradient descent. Intuitively, this means that proximal method only "sees" the nice smooth part $g$ of the objective, and is not impacted by the additional $h$ which it treats separately in each step.

**Theorem 3.15.** *Let $g : \mathbb{R}^d \to \mathbb{R}$ be convex and smooth with parameter $L$, and also $h$ convex and $prox_{h,\gamma}(\mathbf{x}) := \operatorname{argmin}_{\mathbf{z}}\{\|\mathbf{x}-\mathbf{z}\|^2/(2\gamma) + h(\mathbf{z})\}$ can be computed. Choosing the fixed stepsize*

$$\gamma := \frac{1}{L},$$

*proximal gradient descent (3.17) with arbitrary $\mathbf{x}_0$ satisfies*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

*Proof.* The proof follows the vanilla analysis for the smooth case, applying it only to $g$, while always keeping $h$ separate, as in (3.15). We leave the details as Exercise 22 for the reader. □

## 3.7 Exercises

**Exercise 19.** *Prove that in Theorem 3.4 (i),*

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t).$$

**Solution:** By definition of projected gradient descent we have

$$\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\| \leq \|\mathbf{y}_{t+1} - \mathbf{x}_t\| = \gamma \|\nabla f(\mathbf{x}_t)\|.$$

The inequality holds because of (3.1) (by definition, $\mathbf{x}_{t+1}$ is the point closest to $\mathbf{y}_{t+1}$ in $X$). The equality holds because of (3.2) (by definition, $\mathbf{y}_{t+1} = \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)$). Combining the above inequality with the step size $\gamma = 1/L$ and squaring yields

$$\|\nabla f(\mathbf{x}_t)\|^2 \geq L^2 \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2.$$

The desired inequality now easily follows from Theorem 3.4 (i).

**Exercise 20.** *Prove that under the assumptions of Theorem 3.6, $f$ has a unique minimizer $\mathbf{x}^\star$ over any nonempty closed and convex set $X \subseteq \mathbb{R}^d$!*

**Solution:** This is just a slight generalization of Exercise 17. Let $\mathbf{x}, \mathbf{y} \in X, \mathbf{x} \neq \mathbf{y}, \lambda \in (0,1)$ and $\mathbf{z} := \lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in X$ (by convexity of $X$). As $\mathbf{z} \neq \mathbf{x}, \mathbf{y}$, strong convexity over $X$ (3.7) yields

$$\begin{aligned} f(\mathbf{x}) &> f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{z}) = f(\mathbf{z}) + (1-\lambda) \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{y}), \\ f(\mathbf{y}) &> f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{z}) = f(\mathbf{z}) + \lambda \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{x}). \end{aligned}$$

Adding up these two inequalities with multiples $\lambda$ and $1-\lambda$, respectively, the gradient terms cancel, and we get

$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) > f(\mathbf{z}).$$

This is strict convexity over $X$ and it implies that there can't be more than one minimizer of $f$ over $X$ (otherwise, any convex combination would have still smaller value). To prove that there is a minimizer, we show that every sublevel set intersected with $X$ is bounded (and closed, since the sublevel set as well as $X$ are closed). Hence, the function attains a minimum $\mathbf{x}^\star$ over this intersection, and this minimum must be a minimizer of $f$ over $X$.

To see boundedness, suppose $\mathbf{y} \in X$ such that $f(\mathbf{y}) \leq \alpha$. Furthermore, assume w.l.o.g. that $\mathbf{0} \in X$ (otherwise, we translate our coordinate system so that this holds). By strong convexity over $X$, we then have

$$\alpha \geq f(\mathbf{y}) \geq f(\mathbf{0}) - \nabla f(\mathbf{0})^\top(-\mathbf{y}) + \frac{\mu}{2}\|\mathbf{y}\|^2 \geq f(\mathbf{0}) - \|\nabla f(\mathbf{0})\|\|\mathbf{y}\| + \frac{\mu}{2}\|\mathbf{y}\|^2,$$

using the Cauchy-Schwarz inequality ($\mathbf{v}^\top\mathbf{w} \leq \|\mathbf{v}\|\|\mathbf{w}\|$). Hence,

$$\|\mathbf{y}\|\left(\frac{\mu}{2}\|\mathbf{y}\| - \|\nabla f(\mathbf{0})\|\right) \leq \alpha - f(\mathbf{0}),$$

which implies that $\|\mathbf{y}\|$ is bounded.

**Exercise 21.** *Prove Lemma 3.13!*

**Hint:** *It can be useful to prove that with $\mathbf{x}^\star(p)$ as in (3.10),*

$$\mathbf{x}^\star(p) = \operatorname{argmin}\{\|\mathbf{x} - \mathbf{v}\| : \sum_{i=1}^{d} x_i = 1, x_{p+1} = \cdots = x_d = 0\}.$$

**Solution:** Let's assume that

$$\mathbf{x}^\star(p) = \operatorname{argmin}\{\|\mathbf{x} - \mathbf{v}\| : \sum_{i=1}^{d} x_i = 1, x_{p+1} = \cdots = x_d = 0\}$$

is true. And also assume that $\Pi_X(\mathbf{v}) = \mathbf{x}^\star(p)$ which also means that $\mathbf{x}^\star(p) = \operatorname{argmin}_{\mathbf{x}\in X} \|\mathbf{x} - \mathbf{v}\|^2$ . Now suppose Lemma 3.13 is wrong, which means that we can find $p' > p$, ($p' \geq p + 1$) with $\mathbf{x}^\star(p')$ as in (3.10), which means that we also get

$$\mathbf{x}^\star(p') = \operatorname{argmin}\{\|\mathbf{x} - \mathbf{v}\| : \sum_{i=1}^{d} x_i = 1, x_{p'+1} = \cdots = x_d = 0\}.$$

Here we are minimizing $\|\mathbf{x} - \mathbf{v}\|$ with less constraint than in the previous case with $\mathbf{x}^\star(p)$ (components $p+1$ to $p'$ do not have to be equal to 0), which implies that $\|\mathbf{x}^\star(p') - \mathbf{v}\| \leq \|\mathbf{x}^\star(p) - \mathbf{v}\|$ . Combining this with the previous assumption of $\mathbf{x}^\star(p) = \Pi_X(\mathbf{v})$ we get $\|\mathbf{x}^\star(p') - \mathbf{v}\| = \|\mathbf{x}^\star(p) - \mathbf{v}\|$ . And since we are projecting on a convex set we know that the projection is unique, and thus $\mathbf{x}^\star(p') = \mathbf{x}^\star(p)$ . However, from the way $\mathbf{x}^\star(p)$ and $\mathbf{x}^\star(p')$ are defined using (3.10), we know that the $p + 1$ component of $\mathbf{x}^\star(p)$ is equal to 0, and that of $\mathbf{x}^\star(p')$ is strictly positive which leads to a contradiction.

**Exercise 22.** *Prove Theorem 3.15!*

# Bibliography


[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. https://web.stanford.edu/~boyd/cvxbook/.

[Dav59] William C. Davidon. Variable metric method for minimization. Technical Report ANL-5990, AEC Research and Development, 1959.

[Dav91] William C. Davidon. Variable metric method for minimization. *SIAM J. Optimization*, 1(1):1–17, 1991.

[DSSSC08] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the 1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279, 07 2008.

[Gol70] D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[Gre70] J. Greenstadt. Variations on variable-metric methods. *Mathematics of Computation*, 24(109):1–22, 1970.

[KNS16] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition. In *ECML PKDD 2016: Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer International Publishing, Cham, September 2016.

[Nes12] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[Noc80] J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980.

[NP06] Yurii Nesterov and B.T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, Aug 2006.

[NSL+15] Julie Nutini, Mark W Schmidt, Issam H Laradji, Michael P Friedlander, and Hoyt A Koepke. Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Selection. In *ICML*, pages 1632–1641, 2015.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *J. R. Statist. Soc. B*, 58(1):267–288, 1996.

[Vis14] Nisheeth Vishnoi. Lecture notes on fundamentals of convex optimization, 2014. https://tcs.epfl.ch/files/content/sites/tcs/files/Lec3-Fall14-Web.pdf.

[Zim16] Judith Zimmermann. *Information Processing for Effective and Stable Admission*. PhD thesis, ETH Zurich, 2016. .