

Chapter 3

Projected Gradient Descent

Contents

3.1	The Algorithm	39
3.2	Constrained minimization: $\mathcal{O}(1/\varepsilon^2)$ steps	39
3.3	Smooth constrained minimization: $\mathcal{O}(1/\varepsilon)$ steps	40
3.4	Strongly convex constrained minimization: $\mathcal{O}(\log(1/\varepsilon))$ steps	43
3.5	Projecting onto ℓ_1 -balls	44
3.6	Exercises	48

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}^* , 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

$$\mathbf{y}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t), \quad (3.1)$$

$$\mathbf{x}_{t+1} := \Pi_X(\mathbf{y}_{t+1}) := \operatorname{argmin}\{\|\mathbf{x} - \mathbf{y}_{t+1}\|^2 : \mathbf{x} \in X\}. \quad (3.2)$$

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*

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

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

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 $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$. \square

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and differentiable, $X \subseteq \mathbb{R}^d$ closed and convex, \mathbf{x}^* a minimizer of f over X ; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq R$, 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

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

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

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

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

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

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

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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

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

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and differentiable. Let $X \subseteq \mathbb{R}^d$ be a closed convex set, and assume that there is a minimizer \mathbf{x}^* 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)

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

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 $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$:

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

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

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

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)

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

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 $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 \leq \|\mathbf{y}_{t+1} - \mathbf{x}^*\|^2$ and not just $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \leq \|\mathbf{y}_{t+1} - \mathbf{x}^*\|^2$. Using this, we further get a critical improvement on the bound for $f(\mathbf{x}_t) - f(\mathbf{x}^*)$ which is now

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

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

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

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. \square

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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 μ) over X if

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

Theorem 3.6. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and differentiable. Let $X \subseteq \mathbb{R}^d$ be a 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}^* 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}^* are geometrically decreasing:

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

(ii)

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

Proof. In the strongly convex case, the “constrained” vanilla bound (3.6)

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

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

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

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

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

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

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

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

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

□

3.5 Projecting onto ℓ_1 -balls

Let $X = B_1(r, \mathbf{0}) = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i| \leq R\}$ be the ℓ_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.

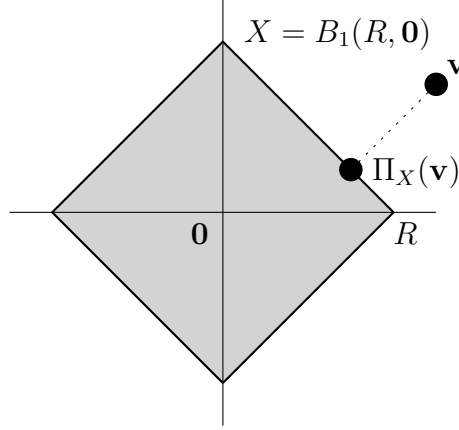


Figure 3.1: Projecting onto an ℓ_1 -ball

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 $B_1(0, 1)$, 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). \square

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 λ , with $\|\mathbf{x}' - \mathbf{v}\| = (1 - \lambda)\|\mathbf{x} - \mathbf{v}\|$, again contradicting the optimality of \mathbf{x} . \square

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

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

where

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

is the standard simplex.

This means, we have reduced the projection onto an ℓ_1 -ball to the projection onto the standard simplex; see Figure 3.2

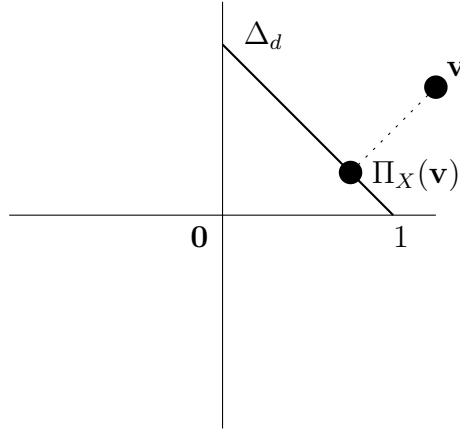


Figure 3.2: Projecting onto the standard simplex

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 Δ_d in the same way).

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

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

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

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

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

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

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

contradicting (3.9). \square

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

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

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

where

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

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

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

again contradicting (3.9). The expression for Θ_p is then obtained from

$$1 = \sum_{i=1}^p x_i^* = \sum_{i=1}^p v_i - p\Theta_p.$$

\square

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

$$\mathbf{x}^*(p) = (v_1 - \Theta_p, \dots, v_p - \Theta_p, 0, \dots, 0), \quad p \in \{1, \dots, n\}, \quad (3.10)$$

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

$$v_p - \Theta_p > 0, \quad (3.11)$$

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

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

it holds that

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

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

Theorem 3.14. *Let $\mathbf{v} \in \mathbb{R}^d$, $R \in \mathbb{R}_+$, $X = B_1(\mathbf{0}, R)$ the ℓ_1 -ball around $\mathbf{0}$ of radius R . The projection*

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

of \mathbf{v} onto $B_1(\mathbf{0}, R)$ can be computed in time $\mathcal{O}(n \log n)$.

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

3.6 Exercises

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

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

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

Exercise 21. Prove Lemma 3.13!

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

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