3 Projected Gradient

Goal

White-box attack, projection, convergence of projected gradient, convergence rate of function value under convexity.

Alert 3.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.

This note is likely to be updated again soon.

We remind that $\langle \cdot, \cdot \rangle$ is the inner product defined in Lecture 0, and $\|\mathbf{w}\|_2 := \sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}$.

Definition 3.2: Problem

In this lecture we consider the constrained smooth minimization problem

$$f_\star = \inf_{\mathbf{w} \in C} f(\mathbf{w}), \tag{3.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$, as before, is continuously differentiable and now $\mathbf{w}$ is constrained in a closed set C. We will consider both convex and nonconvex f, and convex and nonconvex C.

Example 3.3: White-box attack

Let $\mathbf{f}(\mathbf{x}; \mathbf{w}) \in \mathbb{R}^c$ be an image classifier, where $\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}$ is an input image, $\mathbf{w}$ are the weights of the classifier, and c is the number of classes. Here $f_k(\mathbf{x}; \mathbf{w}) \in \mathbb{R}$ represents the relative confidence our classifier has for class $k \in \{1, \ldots, c\}$. In a white-box attack, we assume complete access to the classifier and aim to construct a perturbed image $\mathbf{x} + \boldsymbol{\delta}$ such that

$$\operatorname*{argmax}_k f_k(\mathbf{x} + \boldsymbol{\delta}; \mathbf{w}) \neq y(\mathbf{x}) =: y.$$

Of course, we are interested in small perturbations $\|\boldsymbol{\delta}\| \leq \epsilon$ that are within the perturbation budget $\epsilon > 0$, so that $\mathbf{x} + \boldsymbol{\delta}$ should have been predicted with the same label as $\mathbf{x}$ according to say a human observer. Such (minimally) perturbed images are dubbed adversarial examples by Szegedy et al. (2014).

We can formulate the above white-box attack as the constrained minimization problem:

$$\min_{\boldsymbol{\delta} \in C} \ f_y(\mathbf{x} + \boldsymbol{\delta}; \mathbf{w}), \quad \text{where} \quad C = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \leq \epsilon, \mathbf{0} \leq \boldsymbol{\delta} + \mathbf{x} \leq \mathbf{1}\}. \tag{3.2}$$

We remind that both $(\mathbf{x}, y)$ and $\mathbf{w}$ are given. Assuming $\mathbf{f}$ is smooth, (3.2) falls into our main problem (3.1) of this lecture.

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus (2014). "Intriguing properties of neural networks". In: *International Conference on Learning Representations (ICLR)*.

Theorem 3.4: An algorithm for univariate convex functions

For any univariate convex function f and convex interval $C = [a, b]$, we have

$$\mathrm{P}_C\left(\operatorname*{argmin}_{w \in \mathbb{R}} f(w)\right) \subseteq \operatorname*{argmin}_{w \in C} f(w),$$

where $\mathrm{P}_C(t) = \mathrm{P}_{[a,b]}(t) := \min\{b, \max\{t, a\}\}$ is the projection of t onto the convex interval $C = [a, b]$.

Proof: Exercise. ■

Alert 3.5: Does it work on high dimensions?

The immediate question we have is to what extent can we generalize Theorem 3.4?

- Does it still hold if C is not an interval (i.e. convex)?
- Does it still hold if f is not convex (or more generally unimodal, i.e. decreasing up to some point and then increasing)?
- Does it still hold on high dimensions, even when both f and C are convex?

The answer is no for all of the above! To repeat, in general:

$$\mathrm{P}_C\left(\operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})\right) \not\subseteq \operatorname*{argmin}_{\mathbf{w}\in C} f(\mathbf{w}),$$

with one *notable exception*, namely when

$$\operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) \subseteq C.$$

Alert 3.6: What if I never leave my cubical?

Consider the following natural alternative where we only move the iterates *within* the constraint set C:

$$\tilde{\mathbf{w}} \in \operatorname*{argmin}_{\mathbf{w}_\eta\in C}\ f(\mathbf{w}_\eta), \qquad \text{where} \quad \mathbf{w}_\eta := \mathbf{w} - \eta\cdot\nabla f(\mathbf{w}),$$

which can even be iterated ($\mathbf{w} \leftarrow \tilde{\mathbf{w}}$). Computing $\tilde{\mathbf{w}}$ already poses some challenge since we still need to solve a univariate constrained minimization problem, especially when the constraint set C is complicated.

Somewhat surprisingly, a more disturbing issue of this seemingly natural approach is that it can easily get trapped at points with no apparent meaning. Consider the following counterexample:

$$\min_{w_1+w_2=1,\mathbf{w}\geq\mathbf{0}}\ \frac{1}{2}(w_1^2+w_2^2).$$

Clearly, $\mathbf{w}_\star = (\frac{1}{2},\frac{1}{2})$ is the only sensible (global) minimizer. However, with $\mathbf{w}=(1,0)$ we have $\mathbf{w}_\eta = (1-\eta, 0)$. Thus, $\tilde{\mathbf{w}} = \mathbf{w} = (1,0)$, i.e. we are stuck at a point that has no apparent optimality.

It turns out that it is important to leave the constraint set (e.g. one's comfort zone ☺) first and then get projected back in order to make progress, at least for gradient descent. This idea, unnatural at first glance, turns out to be quite intuitive, as we will see momentarily. Barzilai and Borwein (1988) took a similar idea to an even more surprising extent.

Barzilai, J. and J. M. Borwein (1988). "Two-Point Step Size Gradient Methods". *IMA Journal of Numerical Analysis*, vol. 8, no. 1, pp. 141–148.

Definition 3.7: (Euclidean) projection to a closed set

Let $C \subseteq \mathbb{R}^d$ be a closed set. We define the (Euclidean) projection of a point $\mathbf{w}\in\mathbb{R}^d$ to C as:

$$\mathrm{P}_C(\mathbf{w}) := \operatorname*{argmin}_{\mathbf{z}\in C}\ \|\mathbf{z}-\mathbf{w}\|_2,$$

i.e., points in C that are closest to the given point $\mathbf{w}$. Projection is obviously non-empty (assuming of course

$C \neq \emptyset$): Choose any $\mathbf{z}_0 \in C$. Then,

$$\operatorname*{argmin}_{\mathbf{z}\in C} \|\mathbf{z}-\mathbf{w}\|_2 \quad = \quad \operatorname*{argmin}_{\mathbf{z}\in C\cap \mathsf{B}(\mathbf{w};\|\mathbf{w}-\mathbf{z}_0\|_2)} \|\mathbf{z}-\mathbf{w}\|_2,$$

where $\mathsf{B}(\mathbf{w};r)$ is the ball with center $\mathbf{w}$ and radius r. Existence of a minimizer of the latter problem, hence also the former, now immediately follows from Theorem 0.33.

The following two cases are geometrically obvious:

- $\mathrm{P}_C(\mathbf{w}) = \mathbf{w}$ iff $\mathbf{w}$ lies in C.
- $\mathrm{P}_C(\mathbf{w}) \in \partial C$ if $\mathbf{w} \notin C$.

The projection is always unique (i.e. singleton) iff C is convex (Bunt 1934; Motzkin 1935). To this day, the only if part remains a long open problem when the space is infinite dimensional.

Bunt, L. N. H. (1934). "Bijdrage tot de theorie de convexe puntverzamelingen". PhD thesis. University of Groningen.
Motzkin, T. S. (1935). "Sur quelques propriétés caractéristiques des ensembles convexes". *Atti della Reale Accademia Nazionale dei Lincei*, vol. 21, no. 6, pp. 562–567.

Example 3.8: Projection onto the cube

Let us consider the following projection:

$$\operatorname*{argmin}_{\mathbf{a}\le\boldsymbol{\delta}\le\mathbf{b}} \|\boldsymbol{\delta}-\boldsymbol{\gamma}\|_2 \quad = \quad \operatorname*{argmin}_{\mathbf{a}\le\boldsymbol{\delta}\le\mathbf{b}} \|\boldsymbol{\delta}-\boldsymbol{\gamma}\|_2^2.$$

The key observation we make here is that the above problem is separable, namely that we can solve each entry in $\boldsymbol{\delta}$ separately. This allows us to reduce to the univariate problem:

$$\operatorname*{argmin}_{a\le\delta\le b} |\delta-\gamma|^2.$$

Using Theorem 3.4, we then have

$$\boldsymbol{\delta}_\star = (\boldsymbol{\gamma}\vee\mathbf{a})\wedge\mathbf{b}, \quad \text{where} \quad a\vee b := \max\{a,b\} \text{ and } a\wedge b := \min\{a,b\}.$$

If we choose the norm $\|\cdot\| = \|\cdot\|_\infty$ in Example 3.3 (the so-called ℓ_∞ attack), projecting onto the constraint set C there can be reduced to our projection here.

Example 3.9: Projection onto the Euclidean ball

Consider projecting a point $\mathbf{w}$ onto the unit Euclidean ball:

$$\operatorname*{argmin}_{\|\mathbf{z}\|_2\le 1} \|\mathbf{w}-\mathbf{z}\|_2^2.$$

This problem is not separable because of the constraint. Nevertheless, we can solve this problem using a change of variables:

$$\operatorname*{argmin}_{0\le r\le 1} \min_{\|\bar{\mathbf{z}}\|_2=1} \|\mathbf{w}-r\cdot\bar{\mathbf{z}}\|_2^2 = -2r\langle\mathbf{w},\bar{\mathbf{z}}\rangle + r^2 + \|\mathbf{w}\|_2^2.$$

The problem is separable now since $\bar{\mathbf{z}}$ does not depend on r. Using the Cauchy-Schwarz inequality (see Definition 0.10) we know $\bar{\mathbf{z}} = \mathbf{w}/\|\mathbf{w}\|_2$. Plugging in we then solve for r:

$$\operatorname*{argmin}_{r\in[0,1]} -2r\|\mathbf{w}\|_2 + r^2.$$

Using Theorem 3.4 we know $r_\star = \|\mathbf{w}\|_2 \wedge 1$, yielding finally

$$\mathbf{w}_\star = (\tfrac{1}{\|\mathbf{w}\|_2}\wedge 1)\cdot\mathbf{w}.$$

It can be proved that the Euclidean ball is the only convex set whose projection is simply a scaling of the input (Yu 2013).

Yu, Y.-L. (2013). "On Decomposing the Proximal Map". In: *Advances in Neural Information Processing Systems 27 (NIPS)*.

Exercise 3.10: Projection onto the nonnegative orthant

Let $C = \mathbb{R}^d_+$ be the nonnegative orthant. Find the formula for $\mathrm{P}_C(\mathbf{w})$. (Exercise 0.35 may be handy.)
When is $\mathrm{P}_C(\mathbf{w})$ unique (i.e. a singleton)?

Exercise 3.11: Projection onto the *discrete* cube

Let $C = \{\pm 1\}^d$ be the *discrete* cube. Find the formula for $\mathrm{P}_C(\mathbf{w})$. (Exercise 0.35 may be handy.)
When is $\mathrm{P}_C(\mathbf{w})$ unique (i.e. a singleton)?

Theorem 3.12: Optimality condition of Euclidean projection (Cheney and Goldstein 1959)

Let $C \subseteq \mathbb{R}^d$ be closed. Then, $C \ni \bar{\mathbf{w}} \in \mathrm{P}_C(\mathbf{w})$ iff

$$\forall \mathbf{z} \in C, \quad \tfrac{1}{2}\|\mathbf{w} - \mathbf{z}\|_2^2 \geq \tfrac{1}{2}\|\mathbf{w} - \bar{\mathbf{w}}\|_2^2. \tag{3.3}$$

If C is also convex, then $C \ni \bar{\mathbf{w}} = \mathrm{P}_C(\mathbf{w})$ iff

$$\forall \mathbf{z} \in C, \quad \langle \mathbf{z} - \bar{\mathbf{w}}, \mathbf{w} - \bar{\mathbf{w}} \rangle \leq 0. \tag{3.4}$$

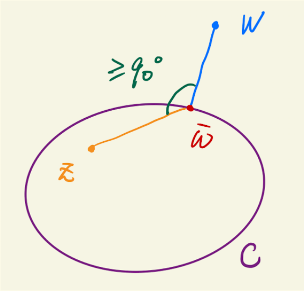
Proof: The first claim (3.3) is simply the definition. To prove the second claim, take any $\mathbf{z} \in C$ we know $\lambda\mathbf{z} + (1-\lambda)\bar{\mathbf{w}} \in C$, hence by (3.3):

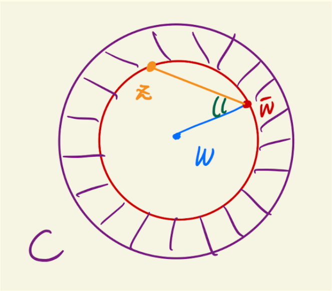
$$\begin{aligned} \tfrac{1}{2}\|\mathbf{w} - \bar{\mathbf{w}}\|_2^2 \leq \tfrac{1}{2}\|\mathbf{w} - \lambda\mathbf{z} - (1-\lambda)\bar{\mathbf{w}}\|_2^2 &\iff 0 \leq \tfrac{1}{2}\lambda^2\|\mathbf{z} - \bar{\mathbf{w}}\|_2^2 - \lambda\langle \mathbf{w} - \bar{\mathbf{w}}, \mathbf{z} - \bar{\mathbf{w}} \rangle \\ &\iff \langle \mathbf{w} - \bar{\mathbf{w}}, \mathbf{z} - \bar{\mathbf{w}} \rangle \leq \tfrac{1}{2}\lambda\|\mathbf{z} - \bar{\mathbf{w}}\|_2^2. \end{aligned}$$

Since $\lambda \in [0, 1]$ is arbitrary, letting $\lambda \to 0$ completes our proof. ■

Geometrically, (3.4) means we have an obtuse angle (at $\bar{\mathbf{w}}$), or equivalently

$$\tfrac{1}{2}\|\mathbf{z} - \mathbf{w}\|_2^2 \geq \tfrac{1}{2}\|\bar{\mathbf{w}} - \mathbf{w}\|_2^2 + \tfrac{1}{2}\|\mathbf{z} - \bar{\mathbf{w}}\|_2^2. \tag{3.5}$$





Cheney, E. W. and A. A. Goldstein (1959). "Newton's Method for Convex Programming and Tchebycheff Approximation". *Numerische Mathematik*, vol. 1, pp. 253–268.

Remark 3.13: Projected gradient as minimizing quadratic upper bound, again

We are now ready to naturally extend gradient descent to the constrained setting, using projections. Recall from (2.19) the quadratic upper bound:

$$\inf_{\mathbf{w}\in C} f(\mathbf{w}) \le \min_{\mathbf{w}\in C} f(\mathbf{w}_t) + \langle \mathbf{w}-\mathbf{w}_t, \nabla f(\mathbf{w}_t)\rangle + \frac{1}{2\eta_t}\|\mathbf{w}-\mathbf{w}_t\|_2^2,$$

where we have added the constraint set C on both sides. A moment's thought (by completing the square Euclidean norm) confirms that the solution to the right-hand side is exactly

$$\mathrm{P}_C(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)) =: \mathbf{w}_{t+1}. \tag{3.6}$$

By definition $\mathbf{w}_t \in C$ for all t and $f(\mathbf{w}_t)$ monotonically decreases if $\eta_t \le \frac{1}{\mathsf{L}_2^{[1]}}$ (so that the quadratic upper bound holds in the first place).

Definition 3.14: Stationary condition for constrained minimization

The iteration in (3.6) motivates us to define the following stationary condition for our constrained minimization problem (3.1):

$$\forall \eta \in [0, \tfrac{1}{\mathsf{L}_2^{[1]}}], \quad \mathbf{w}_* = \mathrm{P}_C(\mathbf{w}_* - \eta\nabla f(\mathbf{w}_*)). \tag{3.7}$$

Clearly, condition (3.7) is necessary (otherwise the iteration (3.6) would strictly decrease f further).

Using (3.3), we can rewrite the stationary condition on $\mathbf{w}_*$ as:

$$\begin{aligned}\forall \mathbf{w}\in C,\quad \tfrac{1}{2}\|\mathbf{w}-\mathbf{w}_*+\eta\nabla f(\mathbf{w}_*)\|_2^2 \ge \tfrac{1}{2}\|\mathbf{w}_*-\mathbf{w}_*+\eta\nabla f(\mathbf{w}_*)\|_2^2 &\iff \tfrac{1}{2}\|\mathbf{w}-\mathbf{w}_*\|_2^2 + \eta\,\langle \mathbf{w}-\mathbf{w}_*, \nabla f(\mathbf{w}_*)\rangle \ge 0\\ \text{if } \eta>0 \text{ is arbitrary or } C \text{ is convex} &\iff \langle \mathbf{w}-\mathbf{w}_*, \nabla f(\mathbf{w}_*)\rangle \ge 0.\end{aligned}$$

We have in fact proved the following equivalence when $\mathbf{w}_*$ is in a convex set C:

$$\begin{aligned}\forall \mathbf{w}\in C,\ \langle \mathbf{w}-\mathbf{w}_*, \nabla f(\mathbf{w}_*)\rangle \ge 0 &\iff \exists \eta>0,\ \mathbf{w}_* = \mathrm{P}_C(\mathbf{w}_*-\eta\nabla f(\mathbf{w}_*))\\ &\iff \forall \eta \ge 0,\ \mathbf{w}_* = \mathrm{P}_C(\mathbf{w}_*-\eta\nabla f(\mathbf{w}_*)).\end{aligned}$$

When f is also convex, using (0.4) we know this necessary condition is also sufficient.

Needless to say, when $C = \mathbb{R}^d$, we recover Fermat's condition (see Theorem 0.38) for unconstrained minimization.

Algorithm 3.15: Projected gradient descent (Goldstein 1964; Levitin and Polyak 1966)

```
Algorithm: Projected gradient descent for constrained smooth minimization
Input: $\mathbf{w}_0 \in C$
for $t = 0, 1, \ldots$ do
    compute gradient $\nabla f(\mathbf{w}_t)$
    choose step size $\eta_t > 0$
    $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \cdot \nabla f(\mathbf{w}_t)$                          // update
    $\mathbf{w}_{t+1} \leftarrow \mathrm{P}_C(\mathbf{w}_{t+1})$          // projecting back to the constraint
```

Compare to the gradient descent Algorithm 2.4, we merely added a projection step in each iteration. It is clear that projected gradient is a strict generalization, and reduces to gradient descent when $C = \mathbb{R}^d$.

Goldstein, A. A. (1964). "Convex programming in Hilbert space". *Bulletin of the American Mathematical Society*, vol. 70, no. 5, pp. 709–710.

Levitin, E. S. and B. T. Polyak (1966). "Constrained Minimization Methods". *USSR Computational Mathematics and Mathematical Physics*, vol. 6, no. 5, pp. 1–50. [English translation in *Zh. Vȳchisl. Mat. mat. Fiz.* vol. 6, no. 5, pp. 787–823, 1965].

Theorem 3.16: Convergence of projected gradient for L-smooth functions

Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathsf{L} = \mathsf{L}_2^{[1]}$-smooth (w.r.t. $\|\cdot\|_2$) and bounded from below (i.e. $f_\star > -\infty$). Let C be convex. If the step size $\eta_t \in [\alpha, \frac{2}{\mathsf{L}} - \beta]$ for some $\alpha, \beta > 0$, then the sequence $\{\mathbf{w}_t\} \subseteq C$ generated by Algorithm 2.4 satisfies $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \to \mathbf{0}$. Moreover,

$$\min_{0\le t\le T-1} \left\| \frac{\mathbf{w}_{t+1} - \mathbf{w}_t}{\eta_t} \right\|_2 \le \sqrt{\frac{f(\mathbf{w}_0) - f_\star}{\alpha\beta\mathsf{L}T/2}}. \tag{3.8}$$

Proof: As before, using L-smoothness:

$$f(\mathbf{w}_{t+1}) \le f(\mathbf{w}_t) + \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla f(\mathbf{w}_t)\rangle + \tfrac{\mathsf{L}}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \le f(\mathbf{w}_t) + (\tfrac{\mathsf{L}}{2} - \tfrac{1}{\eta_t})\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2, \tag{3.9}$$

where the inequality

$$\langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla f(\mathbf{w}_t)\rangle + \tfrac{1}{\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \le 0 \iff \langle \mathbf{w}_t - \mathbf{w}_{t+1}, \mathbf{w}_t - \eta_t\nabla f(\mathbf{w}_t) - \mathbf{w}_{t+1}\rangle \le 0$$

follows from (3.4), since $\mathbf{w}_{t+1}$ is the projection of $\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)$.

Therefore, if $\eta_t \in [\alpha, \frac{2}{\mathsf{L}} - \beta]$ and $\mathbf{w}_{t+1} - \mathbf{w}_t \ne \mathbf{0}$, we *strictly* decrease the function value. Rearranging (3.9) we have

$$\left\| \frac{\mathbf{w}_{t+1} - \mathbf{w}_t}{\eta_t} \right\|_2^2 \le \frac{f(\mathbf{w}_t) - f(\mathbf{w}_{t+1})}{\eta_t(1 - \frac{\eta_t \mathsf{L}}{2})} \le \frac{f(\mathbf{w}_t) - f(\mathbf{w}_{t+1})}{\alpha\beta\mathsf{L}/2}.$$

Summing from $t = 0$ to $t = T - 1$:

$$\sum_{t=0}^{T-1} \left\| \frac{\mathbf{w}_{t+1} - \mathbf{w}_t}{\eta_t} \right\|_2^2 \le \frac{f(\mathbf{w}_0) - f(\mathbf{w}_T)}{\alpha\beta\mathsf{L}/2} \le \frac{f(\mathbf{w}_0) - f_\star}{\alpha\beta\mathsf{L}/2}.$$

Therefore, the sequence $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2$ is square summable hence $\mathbf{w}_{t+1} - \mathbf{w}_t \to \mathbf{0}$ and the bound (3.8) holds (recall that the sum of T numbers is at least T times the smallest number). ■

As before, choosing $\alpha = \beta = \frac{1}{\mathsf{L}}$ optimizes the bound:

$$\min_{0\le t\le T-1} \mathsf{L} \cdot \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \le \sqrt{\frac{2\mathsf{L}[f(\mathbf{w}_0) - f_\star]}{T}}.$$

The observations in Remark 2.20 continue to apply. In particular, we can use Armijo's backtracking line search to avoid estimating L directly: we find the smallest $k \in \mathbb{N}$ such that

$$f\left(\mathbf{w}^{(k)}\right) \le f(\mathbf{w}) + \left\langle \mathbf{w}^{(k)} - \mathbf{w}, \nabla f(\mathbf{w}) \right\rangle + \tfrac{1}{2\eta^{(k)}}\|\mathbf{w}^{(k)} - \mathbf{w}\|_2^2,$$

where $\mathbf{w}^{(k)} := \mathrm{P}_C\left(\mathbf{w} - \eta^{(k)}\nabla f(\mathbf{w})\right)$ and $\eta^{(k)} := \frac{\eta}{2^k}$.

Theorem 3.17: Convergence rate of projected gradient in terms of function value

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and $\mathsf{L} = \mathsf{L}_2^{[1]}$-smooth, $C \subseteq \mathbb{R}^d$ be closed convex, and η_t is chosen so that (3.10) below holds, then for all $\mathbf{w} \in C$ and $t \geq 1$, the sequence $\{\mathbf{w}_t\} \subseteq C$ generated by Algorithm 3.15 satisfy:

$$f(\mathbf{w}_t) \leq f(\mathbf{w}) + \frac{\|\mathbf{w} - \mathbf{w}_0\|_2^2}{2t\bar{\eta}_t}, \quad \text{where} \quad \bar{\eta}_t := \frac{1}{t}\sum_{s=0}^{t-1} \eta_s,$$

Proof: As before, using L-smoothness we have for all $\mathbf{w} \in C$:

$$\begin{aligned} f(\mathbf{w}_{t+1}) &\leq f(\mathbf{w}_t) + \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla f(\mathbf{w}_t)\rangle + \tfrac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \qquad (3.10)\\ &\leq f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, \nabla f(\mathbf{w}_t)\rangle + \tfrac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2 - \tfrac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 \\ &\leq f(\mathbf{w}) + \tfrac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2 - \tfrac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2, \end{aligned}$$

where the second inequality follows from $\mathbf{w}_{t+1}$ being the projection to the convex set C (see (3.5)) and the last inequality is due to the convexity of f (see Theorem 0.29). Take $\mathbf{w} = \mathbf{w}_t$ we see that

$$f(\mathbf{w}_{t+1}) \leq f(\mathbf{w}_t),$$

i.e., the algorithm is descending. Summing from $t = 0$ to $t = T - 1$:

$$T\bar{\eta}_T \cdot [f(\mathbf{w}_T) - f(\mathbf{w})] \leq \sum_{t=0}^{T-1} \eta_t[f(\mathbf{w}_{t+1}) - f(\mathbf{w})] \leq \frac{1}{2}\|\mathbf{w} - \mathbf{w}_0\|_2^2, \quad \text{where} \quad \bar{\eta}_T := \frac{1}{T}\sum_{t=0}^{T-1} \eta_t.$$

(To derive the left inequality, apply $f(\mathbf{w}_{t+1} \geq f(\mathbf{w}_T)$.) Dividing both sides by $T\bar{\eta}_T$ completes the proof. ■

If there exists a minimizer $\mathbf{w}_\star$, then we have

$$f(\mathbf{w}_t) - f_\star \leq \frac{\mathsf{L}\|\mathbf{w}_\star - \mathbf{w}_0\|_2^2}{2t},$$

where we have chosen $\eta_t \equiv 1/\mathsf{L}$ to minimize the bound. So the function value converges to the global minimum (thanks to convexity) at the rate of $O(1/t)$. As before, the dependence on L and $\mathbf{w}_0$ makes intuitive sense. Again, the rate of convergence does not depend on d, the dimension!

It can be proved using fixed point theorems that the iterate $\mathbf{w}_t$ also converges (provided that a minimizer exists). Moreover, $\mathbf{w}_t$ converges (to a stationary point) even when f is nonconvex, provided that it is "definable" (whatever that means ☺).

Remark 3.18: Open-loop step size

Our proof in Theorem 3.17 actually revealed that the projected gradient Algorithm 3.15 converges to the minimum function value if the step size η_t satisfies the following conditions:

$$\eta_t \to 0 \qquad \text{and} \qquad \sum_t \eta_t = \infty,$$

where the first condition is needed so that $\eta_t \leq \frac{1}{\mathsf{L}}$ eventually (no matter what L actually is, as long as it is finite), i.e. (3.10) holds. Such step sizes are called open-loop, meaning that it does not depend on $\mathbf{w}_t$ and hence can be decided beforehand. We will meet this step size again shortly, and then repeatedly.

Needless to say, step sizes that do depend on the iterates $\mathbf{w}_t$, such as Amijo's rule, are called closed-loop, i.e. forming a closed loop

$$\begin{tikzcd} \mathbf{w}_t \arrow[r, bend left] & \eta_t \arrow[l, bend left, "t\leftarrow t+1"] \end{tikzcd}.$$

Alert 3.19: Devil is in the details

The projected gradient Algorithm 3.15 is an elegant way to deal with constraints. However, we remind that projection, even onto a convex set, is itself a constrained minimization problem! For complicated constraints, computing its projection is already very challenging, let alone that we need to do it in every iteration of the projected gradient algorithm. Thus, projected gradient is applicable only when projection to the constraint set is relatively cheap or even in closed-form, as shown in some of the examples.