# 4 Subgradient Algorithms

**Goal**

Subgradient, optimality condition, subdifferential calculus, necessity of non-descending.

**Alert 4.1: Convention**

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

This note is likely to be updated again soon.

**Definition 4.2: Problem**

In this lecture we consider the generic minimization problem:

$$\inf_{\mathbf{w}\in C} \ f(\mathbf{w})$$

where $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a convex function and $C \subseteq \mathbb{R}^d$ is a closed convex set. We do not pose any smoothness or structural assumption on $f$.

**Example 4.3: (Soft-margin) Support Vector Machines (SVM)**

Given a binary dataset $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$, SVM aims at finding a hyperplane that minimizes the following objective function:

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \ \frac{1}{n}\sum_{i=1}^{n}(1 - y_i\hat{y}_i)_+ + C\|\mathbf{w}\|_2^2, \quad \text{where} \quad \hat{y}_i := \langle \mathbf{w}, \mathbf{x}_i\rangle + b,$$

each term in the first summation is called the hinge loss (for the $i$-th training example) and the second Euclidean norm is called the (inverse) margin (of the hyperplane parameterized by normal vector $\mathbf{w}$ and offset $b$). The hinge loss equals 0 if $y_i\hat{y}_i \geq 1$ (in which case our prediction $\text{sign}(\hat{y}_i)$ would coincide with the groundtruth $y_i$) while we pay a linear penalty if $y_i\hat{y}_i \leq 1$. Notably, we still pay a small penalty when $0 < y_i\hat{y}_i < 1$, i.e., even when our prediction is correct: $\text{sign}(\hat{y}_i) = y_i$.

The hinge loss is not differentiable, due to the kink at origin. While many algorithms have been developed for optimizing SVM, it eventually became clear that the classic subgradient algorithm, when applied directly to the above formulation, is as competitive (Shalev-Shwartz et al., 2011).

Shalev-Shwartz, S., Y. Singer, N. Srebro, and A. Cotter (2011). "Pegasos: primal estimated sub-gradient solver for SVM". *Mathematical Programming*, vol. 127, pp. 3–30.

**Definition 4.4: Subgradient and subdifferential**

We define the subdifferential of a *convex* function $f$ at some point $\mathbf{w}$ as the set:

$$\partial f(\mathbf{w}) := \{\mathbf{g} \in \mathbb{R}^d : \forall \mathbf{z}, \ f(\mathbf{z}) \geq f(\mathbf{w}) + \langle \mathbf{z} - \mathbf{w}; \mathbf{g}\rangle\}$$

Any $\mathbf{g} \in \partial f(\mathbf{w})$ is called a subgradient of $f$ at $\mathbf{w}$. It is clear from the definition that the subdifferential is always closed and convex.

### Theorem 4.5: Optimality condition for nonsmooth minimization

$\mathbf{w}_\star \in \operatorname{argmin} f$ *iff* $\mathbf{0} \in \partial f(\mathbf{w}_\star)$.

*Proof:* Clear from the definition. ■

The subdifferential and the above theorem extend naturally to nonconvex functions. However, the subtlety is that for an arbitrary nonconvex function, even though the subdifferential is nonempty at the minimizers, it can be empty at other points. We need convexity (or more generally local Lipschitz continuity) to have nonempty subdifferential everywhere.

### Remark 4.6: Subdifferential as limits

One remarkable property of convex functions (and more generally locally Lipschitz continuous functions) is that they are differentiable almost everywhere. This allows us to define the subdifferential as the following limit (Clarke, 1990):

$$\partial f(\mathbf{w}) = \operatorname{conv}\{\mathbf{g} : \exists \mathbf{z} \to \mathbf{w}, \nabla f(\mathbf{z}) \to \mathbf{g}\}.$$

It then follows that when $f$ is continuously differentiable, its subdifferential reduces to the singleton $\partial f = \{\nabla f\}$.

Clarke, F. H. (1990). “Optimization and Nonsmooth Analysis”. reprinted from the 1983 edition. SIAM.

### Alert 4.7: Subdifferential calculus: Successes and failures

The usual calculus rules for derivatives no longer hold for subdifferentials. For instance:

- $\partial(\alpha f) = \alpha \cdot \partial f$ only when $\alpha > 0$ while the equality typically fails when $\alpha < 0$.
- $\partial(f + g) \supseteq \partial f + \partial g$, where equality is attained when one of the function is continuous at the point (and the other function is finite).
- $\partial(f \circ g) = (\nabla g) \cdot (\partial f)$ when $f : \mathbb{R}^p \to \mathbb{R}$ is convex and $g : \mathbb{R}^d \to \mathbb{R}^p$ is continuously differentiable. Note that $\nabla g \in \mathbb{R}^{d\times p}$ whose $j$-th column is the gradient of $g_j$ and $\partial f \subseteq \mathbb{R}^p$ so the multiplication makes sense. We may derive this result using Remark 4.6. Note that even though we still use the same notation $\partial$, the subdifferential here is not necessarily the same one as in Definition 4.4 (so Theorem 4.5 may not apply), unless of course $f \circ g$ is convex.

### Example 4.8: Subdifferential calculation

Consider the positive part function $\ell(t) = (t)_+$ that appeared in SVM. Using Remark 4.6 we obtain its subdifferential as

$$\partial \ell(t) = \begin{cases} 1, & t > 0 \\ 0, & t < 0 \\ [0,1], & t = 0 \end{cases}.$$

We may also verify the above formula from the definition:

$$\forall s, \quad (s)_+ \ge (t)_+ + g(s - t).$$

Indeed, for $t > 0$, choosing $s > t$ we obtain $g \le 1$ while choosing $0 < s < t$ we obtain $g \ge 1$ hence $g = 1$. For $t = 0$, choosing $s > 0$ we obtain $g \le 1$ while choosing $s < 0$ we obtain $g \ge 0$, i.e. $g \in [0,1]$. Lastly, for $t < 0$, choosing $0 > s > t$ we obtain $g \le 0$ while choosing $s < t$ we obtain $g \ge 0$.

### Example 4.9: Envelope function

Let $f(\mathbf{w}) = \max_{i \in I} f_i(\mathbf{w})$ be the upper envelope of the continuously differentiable functions $f_i$. To apply the last rule in Alert 4.7, we need to compute the subdifferential of the max function $h(\mathbf{t}) = \max_{i \in I} t_i$, i.e.

$$\text{find all } \mathbf{g} \text{ s.t. } \forall \mathbf{s},\ h(\mathbf{s}) \geq h(\mathbf{t}) + \langle \mathbf{s} - \mathbf{t}, \mathbf{g} \rangle . \tag{4.1}$$

Let $I^* = I^*(\mathbf{t}) = \{i \in I : h(\mathbf{t}) = t_i\}$ be the *active* indices. Then, we claim

$$\partial h(\mathbf{t}) = \operatorname{conv}\{\mathbf{e}_i : i \in I^*(\mathbf{t})\},$$

where $\mathbf{e}_i$ is 0 except 1 in the $i$-th entry. Indeed, choose any $i \in I^*$ and let $\mathbf{g} = \mathbf{e}_i$, we have

$$h(\mathbf{s}) \geq h(\mathbf{t}) + \langle \mathbf{s} - \mathbf{t}, \mathbf{g} \rangle \iff h(\mathbf{s}) \geq t_i + s_i - t_i \iff h(\mathbf{s}) \geq s_i.$$

Thus, $\mathbf{e}_i \in \partial h(\mathbf{t})$.

To prove the converse,

- we first note that $\mathbf{g}$ must be nonnegative, for otherwise we may just let some component of $\mathbf{s}$ go to $-\infty$, pushing the right-hand side of (4.1) to $\infty$ while capping the left-hand side.
- choose any $j \notin I^*$ and set $\mathbf{s} = \mathbf{t} + \delta \mathbf{e}_j$ for some small $\delta \geq 0$ such that $h(\mathbf{s}) = h(\mathbf{t})$. It then follows from (4.1) that $g_j \leq 0$ and hence $g_j = 0$.
- let $\mathbf{s} = \mathbf{t} + \delta \mathbf{1}$, then from (4.1) we obtain for any subgradient $\mathbf{g}$:

  $$h(\mathbf{t}) + \delta = h(\mathbf{s}) \geq h(\mathbf{t}) + \delta \langle \mathbf{1}; \mathbf{g} \rangle ,$$

  whence follows $\langle \mathbf{1}; \mathbf{g} \rangle = 1$ since $\delta$ is arbitrary. This completes our proof as the set $\{\mathbf{g} \geq \mathbf{0} : \langle \mathbf{1}; \mathbf{g} \rangle = 1, g_j = 0\ \forall j \notin I^*\}$ was exactly our claim for the subdifferential.

### Exercise 4.10: More subdifferentials

- Compute the subdifferential of the absolute function $f(t) = |t|$.
- We mentioned before that any norm is not differentiable at the origin. Prove that

  $$\partial \|\mathbf{0}\| = \{\mathbf{g} : \|\mathbf{g}\|_\circ \leq 1\}.$$

### Alert 4.11: It is differentiable almost everywhere, so what?!

Let us consider the following function

$$f(x, y) = |x| + \tfrac{1}{2} y^2.$$

Obviously, we have a unique minimizer at $(x_\star, y_\star) = (0, 0)$. Assuming we are at $(x, y) = (0, 1)$ and we choose the subgradient $\mathbf{g} = (1, 1)$. Let us try to find an optimal step size:

$$\min_{\eta \geq 0} |\eta| + \tfrac{1}{2}(1 - \eta)^2,$$

leading to $\eta = 0$. Thus, we are stuck at an obviously suboptimal point $(x, y) = (0, 1)$ had we chosen a wrong subgradient and used Cauchy's rule to find an optimal step size!

### Alert 4.12: The subtlety in nonsmooth optimization

Given a direction vector $\mathbf{d} \in \mathbb{R}^d$, we define the directional derivative

$$f'(\mathbf{w}; \mathbf{d}) := \lim_{t \downarrow 0} \frac{f(\mathbf{w} + t\mathbf{d}) - f(\mathbf{w})}{t}, \tag{4.2}$$

which always exists for a convex function (when $\mathbf{w} \in \operatorname{int} \operatorname{dom} f$). Under mild regularity conditions, it can be shown that

$$\operatorname*{argmin}_{\|\mathbf{d}\|_2 \le 1} f'(\mathbf{w}; \mathbf{d}) = - \operatorname*{argmin}_{\mathbf{d} \in \partial f(\mathbf{w})} \|\mathbf{d}\|_2, \quad \text{since } f'(\mathbf{w}; \mathbf{d}) = \sigma_{\partial f(\mathbf{w})}(\mathbf{d}), \tag{4.3}$$

which can be interpreted as the *steepest* descent direction at $\mathbf{w}$. Therefore, a natural algorithm for minimizing nonsmooth convex function is:

**Algorithm:** The minimum-point subgradient algorithm, may NOT converge

```
Input: $\mathbf{w}_0 \in \operatorname{dom} f$
for $t = 0, 1, \ldots$ do
    $\mathbf{d}_t \leftarrow \operatorname*{argmin}_{\mathbf{d} \in \partial f(\mathbf{w}_t)} \|\mathbf{d}\|_2$        // choose the minimum subgradient
    choose step size $\eta_t$        // e.g. Cauchy's rule: $\eta_t = \operatorname*{argmin}_{\eta \ge 0} f(\mathbf{w}_t - \eta_t \mathbf{d}_t)$
    $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \mathbf{d}_t$
```

This algorithm is appealing in two aspects:

- It reduces to gradient descent when $f$ is smooth.
- It is a descent algorithm, i.e. if $\mathbf{w}_t$ is not optimal, then $\mathbf{d}_t \neq \mathbf{0}$ according to Theorem 4.5. Choosing the step size suitably then guarantees $f(\mathbf{w}_{t+1}) < f(\mathbf{w}_t)$ (see (4.3) and (4.2)).

Surprisingly, as the following example shows, the descending property may prevent the algorithm from converging to a global minimum!

### Example 4.13: Cauchy's rule no longer works

Consider the following nonsmooth function

$$f(x, y) = \begin{cases} 5\sqrt{9x^2 + 16y^2}, & x > |y| \\ 9x + 16|y|, & x \le |y| \end{cases},$$

which is not differentiable on the ray $x \le 0, y = 0$, but $f(x, y) \to -\infty$ as $x \to -\infty$. We verify the following subdifferential:

- For $x < 0$, $\partial f(x, 0) = \{(9, v) : |v| \le 16\}$.
- If we approach $(0, 0)$ from the upper case, we have

  $$\nabla f(x, y) = (u, v) = (\tfrac{45x}{r}, \tfrac{80y}{r}), \quad \text{where} \quad r = \sqrt{9x^2 + 16y^2}.$$

  Thus, considering $x > |y|$, we have $(u/15)^2 + (v/20)^2 = 1, u > 9$. If we approach $(0, 0)$ from the lower case, we have $\partial f(x, 0)$. Thus,

  $$\partial f(0, 0) = \{(u, v) : (u/15)^2 + (v/20)^2 \le 1, u \ge 9\}.$$

Wolfe (1975) showed that starting with $x > |y| > (9/16)^2|x|$, Cauchy's rule leads to a polygonal path of successively orthogonal segments that converges to the origin! Intuitively, the algorithm only "sees" the upper case and converges myopically to its minimizer.

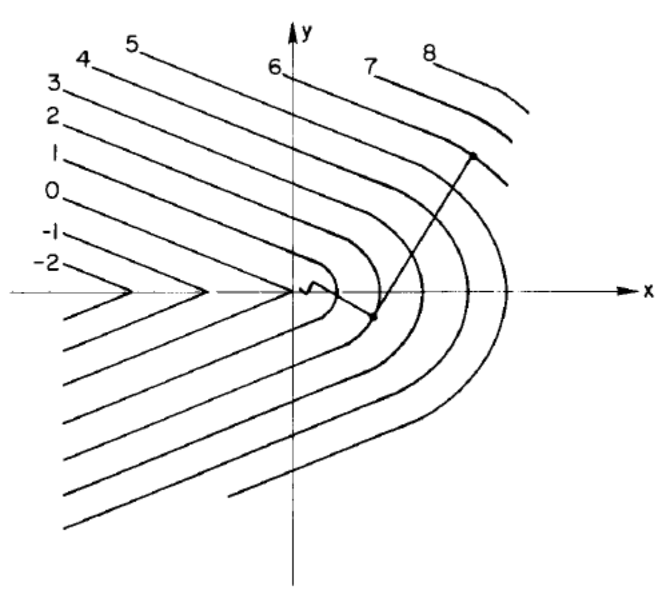

Fig. 1. Contours of $f$ and steepest descent path.

Wolfe, P. (1975). “A method of conjugate subgradients for minimizing nondifferentiable functions”. *Mathematical Programming Study*, vol. 3, pp. 145–173.

## Algorithm 4.14: The subgradient algorithm

```
Algorithm: The subgradient algorithm
Input: $\mathbf{w}_0 \in C$
for $t = 0, 1, \ldots$ do
    choose $\mathbf{d}_t \in \partial f(\mathbf{w}_t)$
    optional: $\mathbf{d}_t \leftarrow \mathbf{d}_t / \|\mathbf{d}_t\|_2$                // normalize
    choose step size $\eta_t$                                   // e.g.   $\eta_t = O(1/t)$
    $\mathbf{w}_{t+1} \leftarrow \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathbf{d}_t)$
```

It turns out that asking the algorithm to always descend is too much when the function is nonsmooth. Instead, we resort to open loop rules:

- $\eta_t \to 0, \sum_t \eta_t = \infty$, e.g. $\eta_t = O(1/\sqrt{t})$
- $\sum_t \eta_t = \infty, \sum_t \eta_t^2 < \infty$, e.g. $\eta_t = O(1/t)$
- $\eta_t \equiv \eta$
- $\eta_t = \eta^t$

When the minimum value $f_\star$ is known in advance, we may also use the following rule (Polyak, 1969):

$$\eta_t = \frac{f(\mathbf{w}_t) - f_\star}{\|\mathbf{d}_t\|}.$$

Polyak, B. (1969). “Minimization of unsmooth functionals”. *USSR Computational Mathematics and Mathematical Physics*, vol. 9, no. 3, pp. 14–29.

### Example 4.15: To normalize or not?

Consider minimizing the convex function $f(w) = w^4$. We obtain the iterates with or without normalizing the (sub)gradient:

$$\begin{aligned} w_{t+1} &= w_t - 4\eta_t w_t^3 = (1 - 4\eta_t w_t^2) w_t \\ \bar{w}_{t+1} &= \bar{w}_t - \eta_t \operatorname{sign}(\bar{w}_t) = \operatorname{sign}(\bar{w}_t)(|\bar{w}_t| - \eta_t). \end{aligned}$$

It is clear that the latter $\bar{w}_t \to 0$ as long as $\eta_t \to 0$ and $\sum_t \eta_t = \infty$. However, for the former, if we start with $w_1 = 1$ and $\eta_t = 1/t$, then

$$w_t^2 \geq 1/\eta_t \implies w_{t+1}^2 = (4\eta_t w_t^2 - 1)^2 w_t^2 \geq (4w_t - 1)^2 w_t^2 \geq 9w_t^2 \geq 9t \geq t+1 = 1/\eta_{t+1},$$

i.e. $|w_t| \to \infty$.

### Proposition 4.16: Euclidean projection to convex set is nonexpansion

*Let $C \subseteq \mathbb{R}^d$ be a closed convex set. Then its (Euclidean) projection $\mathrm{P}_C$ is nonexpansive:*

$$\forall \mathbf{w}, \mathbf{z} \in \mathbb{R}^d, \quad \|\mathrm{P}_C(\mathbf{w}) - \mathrm{P}_C(\mathbf{z})\|_2 \leq \|\mathbf{w} - \mathbf{z}\|_2.$$

*Proof:* Applying the optimality condition in Theorem 19.12 we obtain:

$$\begin{aligned} \langle \mathrm{P}_C(\mathbf{z}) - \mathrm{P}_C(\mathbf{w}), \mathbf{w} - \mathrm{P}_C(\mathbf{w}) \rangle &\leq 0 \\ \langle \mathrm{P}_C(\mathbf{w}) - \mathrm{P}_C(\mathbf{z}), \mathbf{z} - \mathrm{P}_C(\mathbf{z}) \rangle &\leq 0. \end{aligned}$$

Adding the two inequalities:

$$\begin{aligned} \langle \mathrm{P}_C(\mathbf{w}) - \mathrm{P}_C(\mathbf{z}), \mathbf{z} - \mathrm{P}_C(\mathbf{z}) - \mathbf{w} + \mathrm{P}_C(\mathbf{w}) \rangle \leq 0 &\iff \|\mathrm{P}_C(\mathbf{w}) - \mathrm{P}_C(\mathbf{z})\|_2^2 \leq \langle \mathbf{w} - \mathbf{z}, \mathrm{P}_C(\mathbf{w}) - \mathrm{P}_C(\mathbf{z}) \rangle \\ &\implies \|\mathrm{P}_C(\mathbf{w}) - \mathrm{P}_C(\mathbf{z})\|_2^2 \leq \|\mathbf{w} - \mathbf{z}\|_2 \cdot \|\mathrm{P}_C(\mathbf{w}) - \mathrm{P}_C(\mathbf{z})\|_2. \end{aligned}$$

Canceling the common factor completes our proof. ■

It is crucial that the set $C$ is convex, for otherwise the projection may not even be single-valued.

### Theorem 4.17: Convergence of the subgradient algorithm

*Let $C \subseteq \mathbb{R}^d$ be a closed convex set and $f : C \to \mathbb{R}$ be an $\mathsf{L} = \mathsf{L}^{[0]}$-Lipschitz continuous convex function (w.r.t. $\|\cdot\|_2$). Start with $\mathbf{w}_0 \in C$, for any $\mathbf{w} \in C$, the sequence generated by Algorithm 4.14 (without normalizing the subgradient) satisfies:*

$$\min_{0 \leq t \leq T-1} f(\mathbf{w}_t) - f(\mathbf{w}) \leq \sum_{t=0}^{T-1} \frac{\eta_t}{\sum_{s=0}^{T-1} \eta_s} (f(\mathbf{w}_t) - f(\mathbf{w})) \leq \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2 + \mathsf{L}^2 \sum_{t=0}^{T-1} \eta_t^2}{2 \sum_{s=0}^{T-1} \eta_s}.$$

*Proof:* According to the update rule in line 5 of Algorithm 4.14:

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 &= \|\mathrm{P}_C(\mathbf{w}_t - \eta_t \mathbf{d}_t) - \mathbf{w}\|_2^2 \\ [\mathbf{w} \in C] \quad &= \|\mathrm{P}_C(\mathbf{w}_t - \eta_t \mathbf{d}_t) - \mathrm{P}_C(\mathbf{w})\|_2^2 \\ [\text{projections are nonexpansive}] \quad &\leq \|\mathbf{w}_t - \eta_t \mathbf{d}_t - \mathbf{w}\|_2^2 \\ &= \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2 \|\mathbf{d}_t\|_2^2 - 2\eta_t \langle \mathbf{w}_t - \mathbf{w}, \mathbf{d}_t \rangle \\ [\mathbf{d}_t \text{ is a subgradient}, \eta_t \geq 0] \quad &\leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2 \|\mathbf{d}_t\|_2^2 + 2\eta_t (f(\mathbf{w}) - f(\mathbf{w}_t)) \end{aligned}$$

$$[\partial f \text{ is bounded by } \mathsf{L}] \quad \le \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2 \mathsf{L}^2 + 2\eta_t(f(\mathbf{w}) - f(\mathbf{w}_t)).$$

Telescoping we obtain

$$\|\mathbf{w}_T - \mathbf{w}\|_2^2 \le \|\mathbf{w}_0 - \mathbf{w}\|_2^2 + \mathsf{L}^2 \sum_{t=0}^{T-1} \eta_t^2 + 2\sum_{t=0}^{T-1} \frac{\eta_t}{\sum_{s=0}^{T-1}\eta_s}(f(\mathbf{w}) - f(\mathbf{w}_t)) \cdot \sum_{s=0}^{T-1}\eta_s.$$

Thus,

$$\min_{0\le t\le T-1} f(\mathbf{w}_t) - f(\mathbf{w}) \le \sum_{t=0}^{T-1} \frac{\eta_t}{\sum_{s=0}^{T-1}\eta_s}(f(\mathbf{w}_t) - f(\mathbf{w})) \le \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2 + \mathsf{L}^2 \sum_{t=0}^{T-1}\eta_t^2}{2\sum_{s=0}^{T-1}\eta_s},$$

as claimed. ■

The bound on the right-hand side vanishes iff $\sum_t \eta_t \to \infty$ and $\eta_t \to 0$.

If we fix a tolerance $\epsilon > 0$ beforehand, then setting $\eta_t = c/\mathsf{L}^2 \cdot \epsilon$ for some constant $c \in ]0, 2[$ leads to:

$$\min_{0\le t\le T-1} f(\mathbf{w}_t) - f(\mathbf{w}) \le \epsilon,$$

as long as $T \ge \frac{\mathsf{L}^2\|\mathbf{w}_0 - \mathbf{w}\|_2^2}{c(2-c)} \cdot \frac{1}{\epsilon^2}$. The same claim holds for $\bar{\mathbf{w}}_T := \sum_{t=0}^{T-1} \frac{\eta_t}{\sum_{s=0}^{T-1}\eta_s}\mathbf{w}_t$.

The choices $\min_{0\le t\le T-1} f(\mathbf{w}_t)$, $f(\bar{\mathbf{w}}_T)$, or $f(\mathbf{w}_T)$ are all used in practice.

**Theorem 4.18: Convergence of the subgradient algorithm: strongly convex**

*Under the same setting as in Theorem 4.17, if $f$ is additionally $\sigma$-strongly convex (w.r.t. the norm $\|\cdot\|_2$), then with $\eta_t = \frac{1}{\sigma(t+1)}$ we have*

$$\min_{0\le t\le T-1} f(\mathbf{w}_t) - f(\mathbf{w}) \le \sum_{t=0}^{T-1} \frac{1}{T}(f(\mathbf{w}_t) - f(\mathbf{w})) \le \frac{\mathsf{L}^2 \sum_{t=0}^{T-1}\frac{1}{t+1}}{2\sigma T}.$$

*Proof:* From the proof of Theorem 4.17:

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 &\le \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2\|\mathbf{d}_t\|_2^2 - 2\eta_t\langle \mathbf{w}_t - \mathbf{w}, \mathbf{d}_t\rangle \\ [\sigma\text{-strong convexity}] \quad &\le (1 - \sigma\eta_t)\|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2\|\mathbf{d}_t\|_2^2 + 2\eta_t(f(\mathbf{w}) - f(\mathbf{w}_t)) \\ [\partial f \text{ is bounded by } \mathsf{L}] \quad &\le \tfrac{t}{t+1}\|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2\mathsf{L}^2 + 2\eta_t(f(\mathbf{w}) - f(\mathbf{w}_t)). \end{aligned}$$

Telescoping we obtain

$$T\|\mathbf{w}_T - \mathbf{w}\|_2^2 \le \frac{\mathsf{L}^2}{\sigma^2}\sum_{t=0}^{T-1}\frac{1}{t+1} + \frac{2}{\sigma}\sum_{t=0}^{T-1} f(\mathbf{w}) - f(\mathbf{w}_t).$$

Thus,

$$\min_{0\le t\le T-1} f(\mathbf{w}_t) - f(\mathbf{w}) \le \sum_{t=0}^{T-1} \frac{1}{T}(f(\mathbf{w}_t) - f(\mathbf{w})) \le \frac{\mathsf{L}^2 \sum_{t=0}^{T-1}\frac{1}{t+1}}{2\sigma T},$$

as claimed. ■

If we define $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{w}_t$, then obviously

$$f(\bar{\mathbf{w}}_T) - f(\mathbf{w}) \le \sum_{t=0}^{T-1} \frac{1}{T}(f(\mathbf{w}_t) - f(\mathbf{w})) \le \frac{\mathsf{L}^2 \sum_{t=0}^{T-1}\frac{1}{t+1}}{2\sigma T}.$$

We note that $\sum_{t=0}^{T-1} \frac{1}{t+1} = \Theta(\log T)$ hence the right-hand side above converges to 0 at rate $O(\frac{\log T}{T})$.