CPSC 532D — 14. (STOCHASTIC) GRADIENT DESCENT

Danica J. Sutherland

University of British Columbia, Vancouver

Fall 2023

We haven't yet really talked in this course about any optimization algorithms to actually *implement* our learning algorithms ERM, RLM, or SRM.

By far the most common optimization algorithm used in machine learning is (stochastic) gradient descent and its variants. We're going to do one particular analysis of SGD, which can both be used for learning bounds in and of itself – SGD *automatically* generalizes, in some settings – and is useful for knowing how long it takes to optimize a function.

This presentation pretty much follows Chapter 14 of Shalev-Shwartz and Ben-David [SSBD]. For much much more about optimization, some good resources are graduate courses by Michael Friedlander (536M) and Mark Schmidt ("5XX"), the books of Boyd and Vandenbreghe [BV04], Nocedal and Wright [NW06], and Bubeck [Bub15], and the recent survey of Garrigos and Gower [GG23].

1 STOCHASTIC GRADIENT DESCENT

Gradient descent tries to find $\min_w f(w)$ for some function f, such as $L_S(f_w)$. Here w should be some finite-dimensional parameter; in kernel methods, we'd typically use the representer theorem, though there's also something called "kernel gradient descent."

We start at some initial point w_1, often either 0 or a sample from, say, $\mathcal{N}(0, \sigma^2 I)$. We then update according to the rule

$$w_{t+1} = w_t - \eta_t \nabla f(w_t);$$

$\eta_t > 0$ is known as either the "learning rate" or the "step size," although note that it's not actually the size of the step since $\|w_{t+1} - w_t\| = \eta_t \|\nabla f(w_t)\|$.

One way to motivate this is to say that we should only "trust" the gradient direction locally, and then should re-check it regularly. Another way is to notice that this update actually minimizes the local quadratic approximation given by

$$g(w) = f(w_t) + \langle \nabla f(w_t), w - w_t \rangle + \frac{1}{2\eta} \|w - w_t\|^2;$$

If instead of $\frac{1}{2\eta}\|w - w_t\|^2$ we use $\frac{1}{2}(w-w_t)\nabla^2 f(w_t)(w-w_t)$, i.e. the second-order Taylor expansion, this is called Newton's method. Each step of Newton's method often improves your loss much more than gradient descent, but each step is also much more computationally expensive.

if f is $\frac{1}{\eta}$-strongly convex, then g will be a global lower bound for f. Even if not, though, it'll be an okay approximation locally.

We repeat this until we decide to stop, after T steps, and then return a result: this might be w_T (the "last iterate"), $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$ (the "average iterate"), $w_{\hat{t}}$ for $\hat{t} \in \arg\min_{t \in [T]} f(w_t)$ (the "best iterate"), the best iterate according to a validation set, or some other scheme.

For more, visit https://cs.ubc.ca/~dsuth/532D/23w1/.

We're going to assume that η_t follows a fixed scheme independent of the data (probably constant), and that we optimize for a fixed number of steps T, also chosen independently of the data. In practice, other schemes might be better; for instance, it's often better to use a *backtracking* scheme to adaptively choose η_t.

1.1 *Projected gradient descent*

Now, we often have some constraint on the possible parameter: for instance, we might require that $\|w\| \leq \mathrm{B}$. One way to adapt gradient descent to this setting is to do what's called *projected gradient descent*: if we require $w \in \mathcal{W}$, define $\mathrm{proj}_{\mathcal{W}}(w) \in \arg\min_{w' \in \mathcal{W}} \|w - w'\|$.

For instance, $\mathcal{W} = \{w : \|w\| \leq \mathrm{B}\}$ gives the projection operator

$$\mathrm{proj}_{\{w:\|w\|\leq \mathrm{B}\}}(w) = \begin{cases} w & \text{if } \|w\| \leq \mathrm{B} \\ \frac{\mathrm{B}}{\|w\|} w & \text{if } \|w\| > \mathrm{B}. \end{cases}$$

proposition 1. *Let $\mathcal{W}$ be a closed convex set, and define $\mathrm{proj}_{\mathcal{W}}(w) = \arg\min_{w' \in \mathcal{W}} \|w' - w\|$. This projection is unique, and for all $v \in \mathcal{W}$,*

$$\left\|\mathrm{proj}_{\mathcal{W}}(w) - v\right\| \leq \|w - v\|.$$

Proof. First, a minimizer must exist since the objective is continuous and the domain is closed. For uniqueness: first, if $w \in \mathcal{W}$ then clearly w is the unique minimizer. Otherwise, suppose that $v, v' \in \mathcal{W}$ both minimize $\|w - \cdot\|$. Since $\mathcal{W}$ is convex, we must have $\frac{1}{2}v + \frac{1}{2}v' \in \mathcal{W}$. But

$$\begin{aligned} \left\|w - \frac{v+v'}{2}\right\|^2 &= \left\|\frac{1}{2}(w-v) + \frac{1}{2}(w-v')\right\|^2 \\ &= \frac{1}{4}\|w-v\|^2 + \frac{1}{4}\left\|w-v'\right\|^2 + \frac{1}{2}\langle w-v, w-v'\rangle \\ &\leq \frac{1}{4}\|w-v\|^2 + \frac{1}{4}\left\|w-v'\right\|^2 + \frac{1}{2}\|w-v\|\left\|w-v'\right\| \\ &= \|w-v\|^2, \end{aligned}$$

since $\|w - v\| = \|w - v'\|$. Since Cauchy-Schwartz is an equality only when the two vectors are parallel or antiparallel, and they have the same norm, this inequality is an equality only if $w - v = \pm(w - v')$. Thus either $w - v = w - v'$ so that $v = v'$, or else $w - v = v' - w$, in which case $w = \frac{1}{2}(v + v')$, and hence $w \in \mathcal{W}$.

For the second part, let $\hat{w} = \mathrm{proj}_{\mathcal{W}}(w)$. Since $\mathcal{W}$ is convex, $\hat{w} + \alpha(v - \hat{w}) \in \mathcal{W}$ for all $\alpha \in [0, 1]$. By definition of $\mathrm{proj}_{\mathcal{W}}$, we then have

$$\begin{aligned} \|\hat{w} - w\|^2 &\leq \|\hat{w} + \alpha(v - \hat{w}) - w\|^2 \\ &= \|\hat{w} - w\|^2 + 2\alpha\langle \hat{w} - w, v - \hat{w}\rangle + \alpha^2\|v - \hat{w}\|^2, \end{aligned}$$

and so

$$\langle \hat{w} - w, v - \hat{w}\rangle \geq -\tfrac{1}{2}\alpha^2\|v - \hat{w}\| 2.$$

Since this holds for all $\alpha \geq 0$, we necessarily have $\langle \hat{w} - w, v - \hat{w} \rangle \geq 0$. Thus

$$\begin{aligned} \|w - v\|^2 &= \|w - \hat{w} + \hat{w} - v\|^2 \\ &= \|w - \hat{w}\|^2 + \|\hat{w} - v\|^2 + 2\langle w - \hat{w}, \hat{w} - v \rangle \\ &\geq \|\hat{w} - v\|^2 . \end{aligned}$$

□

Projected gradient descent updates according to

$$w_{t+1} = \mathrm{proj}_{\mathcal{W}}\Big(w_t - \eta_t \nabla f(w_t)\Big).$$

1.2 *(Projected) subgradient descent*

Sometimes the function f isn't differentiable. It turns out that if f is convex, the algorithm doesn't actually need f to be differentiable; it's enough to get a *subgradient* of f.

Recall that differentiable convex f always lie above their tangent planes:

$$\forall w', \qquad f(w') \geq f(w) + \langle \nabla f(w), w' - w \rangle.$$

DEFINITION 2. A *subgradient* of a function $f : \mathbb{R}^d \to \mathbb{R}$ at w is a vector g such that

$$\forall w', \qquad f(w') \geq f(w) + \langle g, w' - w \rangle.$$

The *subdifferential* of f at w, written $\partial f(w)$, is the set of all subgradients of f at w.

[SSBD] calls this the "differential set."

Notice that the subdifferential is always a closed convex set.

PROPOSITION 3. *If a convex function f is differentiable at w, then $\partial f(w) = \{\nabla f(w)\}$.*

The fact that $\nabla f(w) \in \partial f(w)$ is immediate from the first-order characterization of convex functions. That it is unique is a little harder to show, but if there were more than one then f wouldn't be differentiable at w [Roc70, Theorem 25.1].

It's possible for the subdifferential to have more than one element, though: consider $f(x) = |x|$ at $x = 0$. We have $|x'| \geq |0| + g(x' - 0) = gx'$ for any g with $|g| \leq 1$, so $\partial f(0) = [-1, 1]$.

PROPOSITION 4. *f is convex iff $\partial f(w)$ is nonempty for all w in the interior of its domain.*

The following is probably the most commonly useful way to find subgradients:

PROPOSITION 5. *Suppose that $f(w) = \max_{i \in [k]} f_i(w)$, where each function f_i is convex. For any point w, if $j \in \arg\max_{i \in [k]} f_i(w)$, then $\partial f_j(w) \subseteq \partial f(w)$.*

Proof. By definition, $f(w') \geq f_j(w')$ for all w', and $f(w) = f_j(w)$. Thus, if $g \in \partial f_j(w)$,

$$f(w') \geq f_j(w') \geq f_j(w) + \langle g, w' - w \rangle = f(w) + \langle g, w' - w \rangle.$$

□

In subgradient descent, rather than following the gradient, we follow any subgradient:

$$w_{t+1} = \text{proj}_{\mathcal{W}}(w_t - \eta g_t) \quad \text{for } g_t \in \partial f(w_t).$$

The converse also holds: for functions f defined on an open set, if all of their subgradients have norm at most ρ, then f is ρ-Lipschitz. To see this, bound $f(w) - f(w')$ and $f(w') - f(w)$ by subgradients, and use Cauchy-Schwartz.

proposition 6. *If f is ρ-Lipschitz, then for all points w in the interior of its domain, for all $g \in \partial f(w)$, it holds that $\|g\| \le \rho$.*

Proof. For any w in the interior of the domain of f, let $g \in \partial f(w)$. Then we have for some $\varepsilon > 0$ that $w + \varepsilon g/\|g\|$ is in the domain of f, and so since g is a subgradient

$$f\left(w + \varepsilon \frac{g}{\|g\|}\right) - f(w) \ge \left\langle g, w + \frac{\varepsilon}{\|g\|} g - w \right\rangle = \varepsilon \|g\|.$$

Since f is ρ-Lipschitz, however, we know that

$$\varepsilon \|g\| \le f\left(w + \varepsilon \frac{g}{\|g\|}\right) - f(w) \le \rho \left\| w + \varepsilon \frac{g}{\|g\|} - w \right\| = \rho\varepsilon. \qquad \square$$

1.3 *Analysis for convex, Lipschitz functions*

The most common case in the literature is to analyze β-smooth functions, functions whose gradient is β-Lipschitz. We're going to instead assume that f is Lipschitz.

Because the proof is simpler, we'll analyze the *average iterate* $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$. The average iterate works better if you're using a constant learning rate than

We're going to upper bound $f(\bar{w}) - f(w^*)$, where as we often do we let w^* be any arbitrary weight vector with $\|w^*\| \le B$ (though probably $w^* \in \arg\min_w f(w)$ is nicest, the proof won't use that).

By Jensen's inequality,

$$f(\bar{w}) - f(w^*) = f\left(\frac{1}{T}\sum_{t=1}^{T} w_t\right) - f(w^*) \le \frac{1}{T}\sum_{i=1}^{T} f(w_t) - f(w^*). \tag{1}$$

Now, since f is convex, we know that

$$f(w^*) - f(w_t) \ge \langle g_t, w^* - w_t \rangle \qquad \text{for } g_t \in \partial f(w_t),$$

or equivalently

$$f(w_t) - f(w^*) \le \langle g_t, w_t - w^* \rangle \qquad \text{for } g_t \in \partial f(w_t).$$

Thus

$$f(\bar{w}) - f(w^*) \le \frac{1}{T}\sum_{t=1}^{T} \langle g_t, w^* - w_t \rangle.$$

Applying Lemma 7 below will yield that for projected subgradient descent,

$$f(\bar{w}) - f(w^*) \le \frac{1}{2\eta T}\|w_1 - w^*\|^2 + \frac{\eta}{2T}\sum_{t=1}^{T} \|g_t\|^2.$$

If f is ρ-Lipschitz, Proposition 6 tells us that $\|g_t\| \le \rho$, and so for any w^* with

$\|w_1 - w^*\| \le \mathrm{B}$,

$$f(\bar{w}) - f(w^*) \le \frac{\mathrm{B}^2}{2\eta\mathrm{T}} + \frac{1}{2}\eta\rho.$$

Recalling that $ax + b/x$ is minimized at $x = \sqrt{b/a}$ with value $2\sqrt{ab}$, the bound is optimized when $\eta = \sqrt{\rho\mathrm{T}}/\mathrm{B}$, giving

$$f(\bar{w}) \le \inf_{w^*:\|w^*\|\le\mathrm{B}} f(w^*) + \mathrm{B}\sqrt{\frac{\rho}{\mathrm{T}}}.$$

Alternatively, we can phrase the bound for any η as

$$f(\bar{w}) \le \inf_{w^*:\|w^*\|\le\frac{\sqrt{\rho\mathrm{T}}}{\eta}} f(w^*) + \frac{\rho}{\eta}.$$

Thus, to compete with any possible w^* for a sequence of optimizations with longer and longer T, we want $\sqrt{\mathrm{T}}/\eta \to \infty$ and $1/\eta \to 0$, i.e. $\eta = \omega(1)$, $\eta = o(\sqrt{\mathrm{T}})$. As with stability bounds, if we don't establish an upper bound on $\|w^*\|$ the rate might be bad, depending on how fast $\inf_{w:\|w\|\le\mathrm{B}} f(w)$ shrinks as $\mathrm{B} \to \infty$.

LEMMA 7. *Let $v_1, \dots, v_\mathrm{T}$ be an arbitrary sequence of vectors. If*

$$w_{t+1} = \mathrm{proj}_{\mathcal{W}}(w_t - \eta v_t),$$

where $\mathcal{W}$ is a closed convex set, we have for any $w^* \in \mathcal{W}$ that

$$\sum_{t=1}^{\mathrm{T}} \langle w_t - w^*, v_t\rangle \le \frac{1}{2\eta}\|w_1 - w^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{\mathrm{T}}\|v_t\|^2.$$

$\mathbb{R}^d$ is closed and convex, if you don't actually want to project.

Proof. First notice that, without using any particular properties yet, it holds that

$$\begin{aligned}\langle w_t - w^*, v_t\rangle &= \frac{1}{\eta}\langle w_t - w^*, \eta v_t\rangle \\ &= \frac{1}{2\eta}\left(-\left\|w_t - w^* - \eta v_t\right\|^2 + \|w_t - w^*\|^2 + \eta^2\|v_t\|^2\right).\end{aligned}$$

As $w_{t+1} = \mathrm{proj}_{\mathcal{W}}(w_t - \eta v_t)$, Proposition 1 implies $\|w_{t+1} - w^*\| \le \left\|w_t - \eta v_t - w^*\right\|$, so

$$\langle w_t - w^*, v_t\rangle \le \frac{1}{2\eta}\left(-\|w_{t+1} - w^*\|^2 + \|w_t - w^*\|^2\right) + \frac{\eta}{2}\|v_t\|^2.$$

Let $d_t = \|w_t - w^*\|^2$. Summing this inequality over t yields

$$\sum_{t=1}^{\mathrm{T}}\langle w_t - w^*, v_t\rangle \le \frac{1}{2\eta}\sum_{t=1}^{\mathrm{T}}(-d_{t+1}) + \frac{1}{2\eta}\sum_{t=1}^{\mathrm{T}} d_t + \frac{\eta}{2}\sum_{t=1}^{\mathrm{T}}\|v_t\|^2. \tag{2}$$

The first two sums mostly cancel in a telescoping sum, leaving us with

$$\sum_{t=1}^{\mathrm{T}}\langle w_t - w^*, v_t\rangle \le \frac{1}{2\eta}\|w_1 - w^*\|^2 - \frac{1}{2\eta}\|w_{\mathrm{T}+1} - w^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{\mathrm{T}}\|v_t\|^2.$$

The result follows from $\|w_{\mathrm{T}+1} - w^*\| \ge 0$. □

2 Stochastic (projected) (sub)gradient descent

If we're trying to minimize $L_S(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(w, z_i)$ for very large m, it can be wasteful to use the whole dataset in calculating a (sub)gradient. Instead, stochastic gradient descent goes in a random direction, which is *on average* the direction of the (sub)gradient:

$$w_{t+1} = \text{proj}_{\mathcal{W}}(w_t - \eta \hat{g}_t) \quad \text{for } \mathbb{E}[\hat{g}_t \mid w_t] \in \partial f(w_t).$$

Our analysis will assume that the $\hat{g}_t \mid w_t$ are independent, although the $\hat{g}_t$ will probably be marginally dependent since which $\hat{g}_1$ we take will affect where w_2 is.

For instance, when $f(w) = \frac{1}{m}\sum_{i=1}^{m} f_i(w)$, we could pick $\hat{g}_t \in \partial f_t(w_t)$. More generally, we could use $\hat{g}_t \in \partial\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} f_i\right)(w_t)$ for some random *minibatch* $\mathcal{I} \subseteq [m]$; larger $|\mathcal{I}|$ will have lower variance, but higher computational cost.

Even more interestingly, we could also consider minimizing the function $f(w) = \mathbb{E}_{z\sim\mathcal{D}}\,\ell(w, z)$ by sampling a fresh $z_t \sim \mathcal{D}$, defining $f_t(w) = \ell(w, z_t)$, and taking $\hat{g}_t \in \partial f_t(w_t)$. This is sometimes called *pure SGD*, but the analysis only works if we do only one pass over our dataset: if we repeat samples, then the $\hat{g}_t \mid w_t$ won't be independent anymore.

theorem 8. *Let $\mathcal{W}$ be a closed convex set and $f : \mathcal{W} \to \mathbb{R}$ convex. For any $w_1, w^* \in \mathcal{W}$, let $w_{t+1} = \text{proj}_{\mathcal{W}}(w_t - \eta\hat{g}_t)$ with independent $\hat{g}_t \mid w_t$ such that $\mathbb{E}[\hat{g}_t \mid w_t] \in \partial f(w_t)$. Letting $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$,*

$$\mathop{\mathbb{E}}_{\hat{g}_{1:T}}[f(\bar{w})] \le f(w^*) + \frac{1}{2\eta}\|w_1 - w^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\hat{g}_{1:t}}\|\hat{g}_t\|^2.$$

Thus, if $\|w_1 - w^*\| \le B$, $\mathbb{E}\|\hat{g}_t\|^2 \le \rho^2$, and $\eta = \frac{B}{\rho\sqrt{T}}$,

$$\mathop{\mathbb{E}}_{\hat{g}_{1:T}}[f(\bar{w})] \le f(w^*) + \frac{B\rho}{\sqrt{T}}.$$

We'll have $\mathbb{E}\|\hat{g}_t\|^2 \le \rho^2$ if $\hat{g}_t$ is a stochastic subgradient for a ρ-Lipschitz loss.

We could also take a $\rho/2$-Lipschitz loss and add $\mathcal{N}(0, \rho/(2d))$ noise to its gradients, if we felt like it.

Notice that if we want to minimize f up to ε accuracy, it requires $T = \left(\frac{B\rho}{\varepsilon}\right)^2$ steps. A $1/\varepsilon^2$ rate isn't particularly great in optimization: for "nice" functions (e.g. smooth, strongly convex functions) gradient descent actually gets a $\log\frac{1}{\varepsilon}$ rate.

Optimizers call a $\log\frac{1}{\varepsilon}$ rate "linear," because it looks linear on a log-scale plot.

Proof. Use $\hat{g}_{t:\tau}$ to denote $(\hat{g}_t, \hat{g}_{t+1}, \dots, \hat{g}_\tau)$. As in (1), Jensen's inequality tells us that

$$\mathop{\mathbb{E}}_{\hat{g}_{1:T}}[f(\bar{w}) - f(w^*)] \le \mathop{\mathbb{E}}_{\hat{g}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^*)\right]. \tag{3}$$

Using linearity of expectation, let's consider each term of the form

$$\mathop{\mathbb{E}}_{\hat{g}_{1:T}}[f(w_t) - f(w^*)] = \mathop{\mathbb{E}}_{\hat{g}_{1:t-1}}[f(w_t) - f(w^*)],$$

since $w_t = \text{proj}_{\mathcal{W}}(w_{t-1} - \eta\hat{g}_{t-1})$ doesn't depend on $\hat{g}_{t:T}$. Now, recall that the *mean* of

the (sub)gradient estimator, $g_t = \mathbb{E}[\hat{g}_t \mid w_t]$, is a subgradient of f: $g_t \in \partial f(w_t)$. Thus we know that

$$f(w_t) - f(w^*) \le \langle w_t - w^*, \mathop{\mathbb{E}}_{\hat{g}_t}[\hat{g}_t \mid \hat{g}_{1:t-1}]\rangle = \mathop{\mathbb{E}}_{\hat{g}_t}\Big[\langle w_t - w^*, \hat{g}_t\rangle \;\Big|\; \hat{g}_{1:t-1}\Big].$$

Taking the expectation with respect to $\hat{g}_{1:t-1}$ of both sides, we get

$$\mathop{\mathbb{E}}_{\hat{g}_{1:t-1}}[f(w_t) - f(w^*)] \le \mathop{\mathbb{E}}_{\hat{g}_{1:t-1}}\mathop{\mathbb{E}}_{\hat{g}_t}\Big[\langle w_t - w^*, \hat{g}_t\rangle \;\Big|\; \hat{g}_{1:t-1}\Big] = \mathop{\mathbb{E}}_{\hat{g}_{1:t}}\langle w_t - w^*, \hat{g}_t\rangle,$$

by the law of total expectation. Summing over t and using (3), we get that

$$\mathop{\mathbb{E}}_{\hat{g}_{1:T}}[f(\bar{w}) - f(w^*)] \le \mathop{\mathbb{E}}_{\hat{g}_{1:T}}\left[\frac{1}{T}\sum_{i=1}^{T}\langle w_t - w^*, \hat{g}_t\rangle\right]. \tag{4}$$

Now, *inside* the expectation, the sequence of realized $\hat{g}_{1:t}$ is just some set of vectors, and we can apply Lemma 7, getting the first result. The second follows by using that $ax + b/x = 2\sqrt{ab}$ for $x = \sqrt{b/a}$. □

Aside (not covered in class)

This choice of a constant learning rate that depends on how long we'll optimize for is a little weird. Often, it makes sense to instead use a learning rate that decays over time (so we get in the general vicinity of a good solution first, then hone in as we get closer).

LEMMA 9. *Let $v_1, \dots, v_T$ be an arbitrary sequence of vectors, and $\eta_t = \alpha/\sqrt{t}$ for some $\alpha > 0$. If*

$$w_{t+1} = \mathrm{proj}_{\mathcal{W}}(w_t - \eta_t v_t),$$

where $\mathcal{W}$ is a closed convex set with diameter at most $2B$. We have for any $w^* \in \mathcal{W}$ that

The diameter *of $\mathcal{W}$ is $\max_{w,w' \in \mathcal{W}} \|w - w'\|$.*

$$\sum_{t=1}^{T}\langle w_t - w^*, v_t\rangle \le \left(\frac{2B^2}{\alpha} + \alpha \max_{t\in[T]}\|v_t\|^2\right)\sqrt{T}.$$

Proof. The start of the proof is the same as Lemma 7, but instead of a telescoping sum in (2), we get (recalling the notation $d_t = \|w_t - w^*\|^2$) that

$$\begin{aligned}\sum_{t=1}^{T}\langle w_t - w^*, v_t\rangle &\le \sum_{t=1}^{T}\frac{d_t}{2\eta_t} - \sum_{t=1}^{T}\frac{d_{t+1}}{2\eta_t} + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|v_t\|^2 \\ &= \frac{d_1}{2\eta_1} + \frac{1}{2}\sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)d_t - \frac{d_{T+1}}{2\eta_T} + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|v_t\|^2.\end{aligned}$$

We'll use that if $a_t, b_t \ge 0$ we have $\sum_t a_t b_t \le \left(\sum_t a_t\right)\max_t b_t$, then

$$\sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) = \frac{1}{\eta_T} - \frac{1}{\eta_1} = \frac{\sqrt{T} - 1}{\alpha}.$$

Also, since $\frac{1}{\sqrt{t}}$ is decreasing,

$$\sum_{t=1}^{T} \eta_t = \alpha \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \le \alpha \sum_{t=1}^{T} \int_{t-1}^{t} \frac{1}{\sqrt{x}} dx = \alpha \int_{0}^{T} \frac{1}{\sqrt{x}} dx = 2\alpha\sqrt{T}.$$

Thus

$$\sum_{t=1}^{T} \langle w_t - w^*, v_t \rangle \le \frac{1}{2\alpha} d_1 + \frac{1}{2\alpha}(\sqrt{T} - 1) \max_{t \in [T]} d_t - \frac{\sqrt{T}}{2\alpha} d_{T+1} + \alpha\sqrt{T} \max_t \|v_t\|^2 .$$

We know that $\|w_t - w^*\| \le 2B$ for all t, so each $d_t \le 4B^2$. Then

$$\sum_{t=1}^{T} \langle w_t - w^*, v_t \rangle \le \frac{2B^2}{\alpha} + \frac{2B^2}{\alpha}(\sqrt{T}-1) + \alpha\sqrt{T} \max_t \|v_t\|^2 = \left(\frac{2B^2}{\alpha} + \alpha \max_t \|v_t\|^2 \right) \sqrt{T}. \quad \square$$

If we have $\|v_t\| \le \rho$ and choose $\alpha = \frac{B}{\rho}\sqrt{2}$, so that $\eta_t = \frac{B}{\rho}\sqrt{\frac{2}{T}}$, plugging into (4) gives

$$\mathop{\mathbb{E}}_{\hat{g}_{1:T}} f(\bar{w}) \le f(w^*) + \frac{2\sqrt{2}B\rho}{\sqrt{T}}.$$

This choice now lets us optimize without pre-committing to a given length, and the result is only slightly worse: there's a constant $2\sqrt{2}$ factor, and we needed slightly stronger versions of B ($\sup_w \|w - w^*\|$ instead of just $\|w_1 - w^*\|$) and ρ (worst-case instead of root-mean-square bound on $\|\hat{g}_t\|$).

Strongly convex functions (not in class)

We can get better rates if we assume that f is λ-strongly convex, not just convex. Theorem 14.11 of Shalev-Shwartz and Ben-David [SSBD] shows suboptimality of

$$\frac{\rho^2}{2\lambda T}(1 + \log T);$$

the log T factor can be removed if we instead output a *tail average* $\frac{2}{T} \sum_{t=T/2+1}^{T} w_t$, and the version with a log still holds for the last iterate.

Also see Garrigos and Gower [GG23] for more related results (especially Section 9).

References

[Bub15] Sébastien Bubeck. "Convex Optimization: Algorithms and Complexity." *Foundations and Trends in Machine Learning* 8.3-4 (2015). arXiv: 1405.4980.

[BV04] Stephen Boyd and Lieven Vandenbreghe. *Convex Optimization*. Cambridge University Press, 2004.

[GG23] Guillaume Garrigos and Robert M. Gower. *Handbook of Convergence Theorems for (Stochastic) Gradient Methods*. 2023. arXiv: 2301.11235.

[NW06] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2006.

[Roc70] R. Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[SSBD] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.