# Unified analysis of gradient/subgradient descent APPM 4490/5490 Theory of Machine Learning

**Instructor**: Prof. Becker **Revision date**: 1/29/24

We'll solve $\min_{\mathbf{w}} f(\mathbf{w})$ via the following generic algorithm, with $B = \|\mathbf{w}_1 - \mathbf{w}^\star\|$,

---

**Require:** $\mathbf{w}_1$

1: **for** $t = 1, 2, \ldots, T$ **do**

2: $\quad\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{v}_t$

3: **end for**

---

where $\mathbf{v}$ is "gradient-like" (e.g., a gradient, subgradient, or a gradient in expectation, like $\mathbb{E}\,\mathbf{v}_t = \nabla f(\mathbf{w}_t)$).

**Lemma 1** (Lemma 14.1 in Shalev-Shwartz and Ben-David)**.** *Let $\{\mathbf{v}_t\}_{t=1}^T$ be arbitrary. No assumptions on $f$ (need not be convex or smooth). The generic algorithm sequence satisfies*

$$\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{w}^\star, \mathbf{v}_t \rangle \le \frac{\|\mathbf{w}_1 - \mathbf{w}^\star\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2 \tag{1}$$

*Proof.* (Sketch: just the good parts)

$$\begin{aligned}
\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{w}^\star, \mathbf{v}_t \rangle &= \frac{1}{2\eta} \sum_{t=1}^{T} \left( -\|\mathbf{w}_{t+1} - \mathbf{w}^\star\|^2 + \|\mathbf{w}_t - \mathbf{w}^\star\|^2 + \eta^2 \|\mathbf{v}_t\|^2 \right) \quad \text{complete-the-square and algebra} \\
&= \frac{1}{2\eta} \left( \|\mathbf{w}_1 - \mathbf{w}^\star\|^2 - \|\mathbf{w}_{T+1} - \mathbf{w}^\star\|^2 \right) + \frac{1}{2\eta} \sum_{t=1}^{T} \eta^2 \|\mathbf{v}_t\|^2 \quad \text{via telescoping sum} \\
&\le \frac{1}{2\eta} \|\mathbf{w}_1 - \mathbf{w}^\star\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2.
\end{aligned}$$

□

A variant of the above result, using a possibly non-constant stepsize $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{v}_t$, is known as Shor's Hyperplane Distance Convergence from his 1985 book; see Theorem 1 in Convergence Rates for Deterministic and Stochastic Subgradient Methods Without Lipschitz Continuity by Benjamin Grimmer (2019 SIOPT) for an elementary proof.

**Lemma 2** (Shor's Hyperplane Distance lemma)**.**

$$\langle \mathbf{w}_t - \mathbf{w}^\star, \mathbf{v}_t / \|\mathbf{v}_t\| \rangle \le \frac{\|\mathbf{w}_1 - \mathbf{w}^\star\|^2 + \sum_{t=1}^{T} \eta_t^2 \|\mathbf{v}_t\|^2}{2 \sum_{t=1}^{T} \eta_t \|\mathbf{v}_t\|}$$

Going back to the simpler Lemma 1, a basic corollary is the following:

**Corollary 3** (2nd part of Lemma 14.1)**.** *If $\|\mathbf{v}_t\| \le \rho$ (e.g., if $f$ is $\rho$-Lipschitz) and $\eta = \frac{B}{\rho\sqrt{T}}$ then*

$$\frac{1}{T} \sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{w}^\star, \mathbf{v}_t \rangle \le \rho \frac{B}{\sqrt{T}}$$

*Proof.* Plugging in $\|\mathbf{v}_t\|^2 \le \rho^2$ and $B^2 = \|\mathbf{w}_1 - \mathbf{w}^\star\|^2$ into the RHS of Eq. (1) gives $\frac{1}{2}\left(B^2/\eta + \eta T \rho^2\right)$ which is minimized at the given value of $\eta$ leading to $\rho B \sqrt{T}$. Dividing the LHS and RHS of Eq. (1) by $T$ gives the result. □

Now we'll see how to use these results

# 1 $f$ is convex but not smooth

Assume $f$ is $\rho$-Lipschitz so the corollary applies. If $f$ is convex, then we have a well-defined subdifferential, so we'll choose $\mathbf{v}_t \in \partial f(\mathbf{w}_t)$ to give us **subgradient descent**. By convexity and definition of subgradients,

$$f(\mathbf{w}_t) - f^\star \leq \langle \mathbf{w}_t - \mathbf{w}^\star, \mathbf{v}_t \rangle \tag{2}$$

so combining this with Corollary 3 immediately yields

**Corollary 4** (sub-gradient descent, Cor. 14.2)**.** *If $f$ is convex and $\rho$-Lipschitz, then subgradient descent (with $\eta = \frac{B}{\rho\sqrt{T}}$) yields*

$$\frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{w}_t) - f^\star) \leq \rho\frac{B}{\sqrt{T}}$$

*hence*

$$f(\mathbf{w}_{best}) - f^\star \leq \rho\frac{B}{\sqrt{T}} \tag{3}$$

*and*

$$f(\bar{\mathbf{w}}) - f^\star \leq \rho\frac{B}{\sqrt{T}} \tag{4}$$

where $\mathbf{w}_{\text{best}} \in \operatorname{argmin}_{\mathbf{w}\in\{\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_T\}} f(\mathbf{w})$ and $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$. If possible, we should use $\mathbf{w}_{\text{best}}$, but in some situations this is not easy. Subgradient descent is not a descent method, so it's not necessarily true that $\mathbf{w}_{\text{best}} = \mathbf{w}_T$. Couldn't we just evaluate $f(\mathbf{w}_t)$ and record the best iterate seen so far? Often we can do this, but sometimes $f$ is very expensive to evaluate (as will especially be the case when we do *stochastic* gradients which sample, and the true loss function $f$ is a population expectation that we can never calculate). In these case, we can do iterate averaging to get $\bar{\mathbf{w}}$, and this result follows because $f(\bar{\mathbf{w}}) \leq \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}_t)$ via Jensen's inequality.

**Commentary** Unlike gradient descent in the smooth case, here we have slower convergence $1/\sqrt{T}$ vs $1/T$ in the smooth case (or $1/T^2$ for Nesterov acceleration). Furthermore, we need to know the maximum number of iterations $T$ in advance in order to set the stepsize. In practice, like stochastic gradient methods, one might use a constant stepsize for a while, then reduce it: a stepsize "schedule."

# 2 f is smooth ($\nabla f$ is $\beta$-Lipschitz continuous)

We use the descent Lemma, which applies whenever $\nabla f$ is $\beta$-Lipschitz continuous, regardless of convexity:

$$f(\mathbf{y}) \leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{y} - \mathbf{w}\rangle + \frac{\beta}{2}\|\mathbf{y} - \mathbf{w}\|_2^2$$

and when applied to $\mathbf{w} = \mathbf{w}_t$, $\mathbf{y} = \mathbf{w}_t - \eta\mathbf{v}_t$ with $\mathbf{v}_t = \nabla f(\mathbf{w}_t)$ and $\eta = \beta^{-1}$ (this is **gradient descent**) after a bit of algebra gives

$$f(\mathbf{w}_{t+1}) \overset{\text{descent lem.}}{\leq} f(\mathbf{w}_t) + \left(\frac{\beta}{2}\eta^2 - \eta\right)\|\nabla f(\mathbf{w}_t)\|^2 \overset{\eta=\beta^{-1}}{\leq} f(\mathbf{w}_t) - \frac{1}{2\beta}\|\underbrace{\nabla f(\mathbf{w}_t)}_{\mathbf{v}_t}\|^2. \tag{5}$$

Also using that $f(\mathbf{w}^\star) \leq f(\mathbf{w}_{t+1})$ gives us another useful result:

$$f(\mathbf{w}) - f^\star \geq \frac{1}{2\beta}\|\nabla f(\mathbf{w})\|^2. \tag{6}$$

If we don't assume $f$ is convex, we can't expect to converge to the global minimizer, so there isn't a result about $f(\mathbf{w}_t) - f^\star \to 0$. Instead, we show convergence to a stationary point, meaning $\|\nabla f(\mathbf{w}_t)\| \to 0$.

**Corollary 5** (gradient descent, non-convex)**.** *If $\nabla f$ $\beta$-Lipschitz, then gradient descent with $\eta = \beta^{-1}$ yields*

$$\min_{t=1,\ldots,T} \|\nabla f(\mathbf{w}_t)\|^2 \le \frac{2\beta}{T}(f(\mathbf{w}_1) - f^\star)$$

*Proof.* Sum Eq. (5) from $t = 1, \ldots, T$ after re-arranging to get

$$\frac{1}{2\beta}\sum_{t=1}^{T} \|\nabla f(\mathbf{w}_t)\|^2 \le \sum_{t=1}^{T} f(\mathbf{w}_t) - f(\mathbf{w}_{t+1}) = f(\mathbf{w}_1) - f(\mathbf{w}_{T+1}) \le f(\mathbf{w}_1) - f^\star$$

since we had a telescoping series, and use $\min_{t=1,\ldots,T} \|\nabla f(\mathbf{w}_t)\|^2 \le \frac{1}{T}\sum_{t=1}^{T} \|\nabla f(\mathbf{w}_t)\|^2$ since the min is less than the average. □

In the convex case, we expect to converge to the global minimizer:

**Corollary 6** (gradient descent, convex)**.** *If $\nabla f$ $\beta$-Lipschitz, and $f$ is convex, then gradient descent with $\eta = \beta^{-1}$ yields*

$$f(\mathbf{w}_{T+1}) - f^\star \le \frac{\beta}{2T}\|\mathbf{w}_1 - \mathbf{w}^\star\|^2.$$

*Proof.* Using the main Lemma (Eq. 1) and replacing $\langle \mathbf{w}_t - \mathbf{w}^\star, \mathbf{v}_t\rangle$ with the bound in Eq. (2) (since gradients are subgradients) gives

$$\sum_{t=1}^{T} f(\mathbf{w}_t) - f^\star \le \frac{1}{2\eta}\|\mathbf{w}_1 - \mathbf{w}^\star\|^2 + \frac{\eta}{2}\sum_{t=1}^{T} \|\underbrace{\nabla f(\mathbf{w}_t)}_{\mathbf{v}_t}\|^2 \tag{7}$$

and the descent lemma Eq. (5) gives $f(\mathbf{w}_{t+1}) + \frac{1}{2\beta}\|\nabla f(\mathbf{w}_t)\|^2 \le f(\mathbf{w}_t)$, so combining with the above equation gives

$$\begin{aligned}
\sum_{t=1}^{T}\left( f(\mathbf{w}_{t+1}) + \frac{1}{2\beta}\|\nabla f(\mathbf{w}_t)\|^2 - f^\star\right) &\le \sum_{t=1}^{T} f(\mathbf{w}_t) - f^\star \quad \text{via descent lemma} \\
&\le \frac{\beta}{2}\|\mathbf{w}_1 - \mathbf{w}^\star\|^2 + \frac{1}{2\beta}\sum_{t=1}^{T}\|\nabla f(\mathbf{w}_t)\|^2 \quad \text{via Eq. (7)}
\end{aligned}$$

where we used $\eta = 1/\beta$. Now canceling the $\frac{1}{2\beta}\sum_{t=1}^{T}\|\nabla f(\mathbf{w}_t)\|^2$ from both sides gives

$$\sum_{t=1}^{T} f(\mathbf{w}_{t+1}) - f^\star \le \frac{\beta}{2}\|\mathbf{w}_1 - \mathbf{w}^\star\|^2$$

hence

$$f(\mathbf{w}_{T+1}) = f(\mathbf{w}_{\text{best}}) \le \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}_{t+1}) - f^\star \le \frac{\beta}{2T}\|\mathbf{w}_1 - \mathbf{w}^\star\|^2$$

where $\mathbf{w}_{T+1} = \mathbf{w}_{\text{best}}$ follows because the descent lemma implies that this is a descent method. □

A variant of the above fixed-stepsize case is to use the "Polyak" adaptive stepsize with $\eta_t = \frac{f(\mathbf{w}_t) - f^\star}{\|\nabla f(\mathbf{w}_t)\|^2}$ or similar (so $\eta_t \ge 1/(2\beta)$ if $\nabla f$ is $\beta$-Lipschitz). For proof techniques using that stepsize, see Revisiting the Polyak Step Size by Elad Hazan and Sham M. Kakade (2019).

Our last case to consider is if we're **strongly convex**, in which case we expect faster convergence, and $\mathbf{w}^\star$ is unique, and we expect a bound on $\|\mathbf{w}_t - \mathbf{w}^\star\|$. Note that if $f$ is $\mu$ strongly convex, then $f$ satisfies the $\mu$ Polyak-Lojasiewicz (PL) inequality

$$\frac{1}{2}\|\nabla f(\mathbf{w})\|^2 \ge \mu(f(\mathbf{w}) - f^\star) \tag{8}$$

(see Nesterov's 2018 book, Thm 2.1.5 and Eq 2.1.10 for a proof). Our result is

**Corollary 7** (gradient descent, strongly convex)**.** *If* $\nabla f$ *$\beta$-Lipschitz, and $f$ is $\mu$ strongly convex, then gradient descent with $\eta = \beta^{-1}$ yields*

$$f(\mathbf{w}_{T+1}) - f^\star \leq \underbrace{\left(1 - \frac{\mu}{\beta}\right)}_{c}^{T-1} (f(\mathbf{w}_1) - f^\star)\,.$$

This is linear convergence, which is asymptotically better than sublinear convergence. We think of $\kappa = \frac{\beta}{\mu}$ as the condition number, so $c = 1 - \kappa^{-1}$. We won't show it here, but Nesterov acceleration can improve $c$ to $c \approx 1 - \kappa^{-1/2}$ when $\kappa \gg 1$.

*Proof.*

$$f(\mathbf{w}_{t+1}) - f(\mathbf{w}_t) \leq \frac{-1}{2\beta}\|\nabla f(\mathbf{w}_t)\|^2 \leq \frac{-\mu}{\beta}(f(\mathbf{w}_t) - f^\star)$$

using the descent lemma for the first inequality and the PL inequality for the second inequality. Re-arranging and recursing gives

$$f(\mathbf{w}_{t+1}) - f^\star \leq \left(1 - \frac{\mu}{\beta}\right)(f(\mathbf{w}_t) - f^\star) \leq \left(1 - \frac{\mu}{\beta}\right)^{t-1}(f(\mathbf{w}_1) - f^\star)\,.$$

□