Lecture Notes 8.5: Adaptive gradient descent and smooth convex losses

1 Adaptive rates in the convex setting

We previously considered the learning rates:

$$\eta_t = \frac{c}{\sqrt{\epsilon^2 + \sum_{i=1}^t \|\mathbf{g}_i\|^2}}$$

and showed that in the non-convex setting these rates obtain some adaptivity to the variance of the gradients. Let's show a similar

Let's consider again the case of convex losses. Previously, we studied losses $\mathcal{L}$ that are convex and Lipschitz. Now, we will consider $\mathcal{L}$ that are convex, Lipschitz, and *smooth*. We'll see that significantly better results are sometimes possible in this setting. The key idea is smoothness provides an *upper bound* on the function value via our standard smoothness lemma:

$$\mathcal{L}(\mathbf{w} + \delta) \leq \mathcal{L}(\mathbf{w}) + \langle \nabla \mathcal{L}(\mathbf{w}), \delta \rangle + \frac{H}{2}\|\delta\|^2$$

On the other hand, convexity provides an *lower bound* on the function value:

$$\mathcal{L}(\mathbf{w} + \delta) \geq \mathcal{L}(\mathbf{w}) + \langle \nabla \mathcal{L}(\mathbf{w}), \delta \rangle$$

By carefully combining these operations, we can provide improved convergence rates on convex losses. Let's start by revisiting our analysis of (non-stochastic) gradient descent for convex losses. Suppose that we start at the origin, so that $\mathbf{w}_1 = 0$. Then, with a constant learning rate η so that $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \mathcal{L}(\mathbf{w}_t)$, we previously proved:

Theorem 1. *Suppose $\mathcal{L}$ is G-Lipschitz and convex. Suppose $\|\mathbf{w}_\star - \mathbf{w}_1\| \leq D$. Set $\eta = \frac{D}{G\sqrt{T}}$. Then*

$$\sum_{t=1}^T \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) \leq GD\sqrt{T}$$

In particular, if $\hat{\mathbf{w}}$ is selected uniformly at random from $\mathbf{w}_1, \dots, \mathbf{w}_T$,

$$\mathbb{E}[\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_t)] \leq \frac{GD}{\sqrt{T}}$$

In fact, the proof of this theorem involved a more general inequality, which we generalized even further on the homework. Specifically, we have:

Theorem 2. *Suppose $\mathcal{L}$ is* any *differentiable function. Then gradient descent with learning rate η guarantees:*

$$\sum_{t=1}^T \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_\star \rangle \leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$$

If $\mathcal{L}$ is also convex:

$$\sum_{t=1}^T \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$$

Note that the first part of this theorem does not actually rely on convexity!

Proof. Let's again bound the change in $\|\mathbf{w}_t - \mathbf{w}_\star\|^2$:

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &= \|\mathbf{w}_t - \eta\nabla\mathcal{L}(\mathbf{w}_t) - \mathbf{w}_\star\|^2 \\ &= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\eta\langle\nabla\mathcal{L}(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_\star\rangle + \eta^2\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\end{aligned}$$

rearranging:

$$\langle\nabla\mathcal{L}(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_\star\rangle \leq \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2}{2\eta} - \frac{\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2$$

sum over t, and telescope the RHS:

$$\begin{aligned}\sum_{t=1}^{T}\langle\nabla\mathcal{L}(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_\star\rangle &\leq \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{2\eta} - \frac{\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 \\ &\leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\end{aligned}$$

And now the final result follows by convexity, since $\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) \leq \langle\nabla\mathcal{L}(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_\star\rangle$ □

Now, previously, we assumed that $\mathcal{L}$ was G-Lipschitz, and used this to bound the $\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2$ terms in the RHS of Theorem 2. This time, we'll use smoothness to make a more refined statement. Smoothness implies:

Lemma 3. *Suppose $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ is an H-smooth function. Then for any $\mathbf{w}$:*

$$\frac{\|\nabla\mathcal{L}(\mathbf{w})\|^2}{2H} \leq \mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star)$$

Warning: *the proof of this theorem may not work if $\mathcal{L}$ is only defined on a bounded subset of $\mathbb{R}^d$ - see if you can spot what would go wrong.*

Proof. From smoothness, we have for any δ:

$$\mathcal{L}(\mathbf{w} + \delta) \leq \mathcal{L}(\mathbf{w}) + \langle\nabla\mathcal{L}(\mathbf{w}), \delta\rangle + \frac{H}{2}\|\delta\|^2$$

Set $\delta = -\frac{\nabla\mathcal{L}(\mathbf{w})}{H}$ to obtain:

$$\begin{aligned}\mathcal{L}(\mathbf{w} - \nabla\mathcal{L}(\mathbf{w})/H) &\leq \mathcal{L}(\mathbf{w}) - \frac{\|\nabla\mathcal{L}(\mathbf{w})\|^2}{2H} \\ \mathcal{L}(\mathbf{w}_\star) &\leq \mathcal{L}(\mathbf{w}) - \frac{\|\nabla\mathcal{L}(\mathbf{w})\|^2}{2H}\end{aligned}$$

□

In words, this tells us that when $\mathcal{L}$ is smooth, then small function values imply small gradient values. This will help us prove faster convergence rates for convex functions because in Theorem 2, we notice that the error becomes smaller when $\nabla\mathcal{L}(\mathbf{w}_t)$ gets smaller. This will set up a nice feedback cycle: as GD converges, $\nabla\mathcal{L}(\mathbf{w}_t)$ will get smaller, which will cause GD to converge even faster, which will let $\nabla\mathcal{L}(\mathbf{w}_t)$ get even smaller, which again causes faster convergence and so on.

Let's see how we can use this Lemma 3 in concert with Theorem 2:

Theorem 4. *Suppose $\mathcal{L}$ is an H-smooth convex function. Set $\eta = \frac{1}{2H}$. Then gradient descent guarantees:*

$$\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) \leq 2H\|\mathbf{w}_\star - \mathbf{w}_1\|^2$$

In particular, if $\hat{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$, then

$$\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{2H\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{T}$$

Proof. From Theorem 2, we have

$$\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$$

now, from Lemma 3:

$$\leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} 2H(\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star))$$

Using our definition of η:

$$= H\|\mathbf{w}_\star - \mathbf{w}_1\|^2 + \frac{1}{2}\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)$$

rearranging:

$$\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) \leq 2H\|\mathbf{w}_\star - \mathbf{w}_1\|^2$$

Now, we need to show the last part of the Theorem. For this, note that if X is a random variable that takes on each values $\mathbf{w}_t$ with probability $1/T$, then $\mathbb{E}[X] = \hat{\mathbf{w}}$. Therefore, by Jensen:

$$\mathcal{L}(\hat{\mathbf{w}}) = \mathcal{L}(\mathbb{E}[X]) \leq \mathbb{E}[\mathcal{L}(X)] = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t)$$

Thus,

$$\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)$$

and so comparing the RHS to the bound we have already obtained, the result follows. □

This result is *significantly* better than the previous results we had for convex optimization: here we obtained $O(1/T)$ suboptimality, while previously it was only $O(1/\sqrt{T})$. The difference in assumptions is that this time we assumed H-smoothness, while last time we assumed G-Lipschitzness. It turns out that the $O(1/\sqrt{T})$ rate is *tight* for G-Lipschitz losses. That is, given any algorithm, there exists a G-Lipschitz convex function such that after T iterations, the algorithm cannot do better than $O(1/\sqrt{T})$ (see [1]).

However, in the smooth case we can actually do even better than $O(1/T)$, it is possible to obtain $O(1/T^2)$! Algorithms that achieve this rate (which is optimal) are called *accelerated* optimization algorithms. This was famously first achieved in 1983 by Yuri Nesterov in a much-studied algorithm. Unfortunately, the original paper is in Russian, but there are texts that cover it in various levels of detail [1, 2]. We will also show how to obtain this rate using a streamline presentation of the method of *linear coupling* [3], which is a more recent technique for designing accelerated algorithms.

2 Acceleration

To get some intuition behind linear coupling, observe that our ordinary analysis of gradient descent without using convexity shows that as long as $\eta \leq \frac{1}{H}$:

$$\mathcal{L}(\mathbf{w}_{t+1}) \leq \mathcal{L}(\mathbf{w}_t) - \frac{\eta}{2}\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$$

That is, *we make faster progress when* $\|\nabla \mathcal{L}(\mathbf{w}_t)\|$ *is large*. In contrast, Theorem 2 showed:

$$\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$$

so that if $\hat{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$,

$$\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta t} + \frac{\eta}{2T}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$$

That is, *we make faster progress when* $\|\nabla \mathcal{L}(\mathbf{w}_t)\|$ *is small*. Further, the analysis of gradient descent relating $\mathcal{L}(\mathbf{w}_{t+1})$ directly to $\mathcal{L}(\mathbf{w}_t)$ does not involve any averaging, but the analysis using convexity in Theorem 2 does involve averaging.

Intuitively, we might then expect to obtain a kind of "best of both worlds" result by combining these analysis styles: if the gradients are big, then the first style of analysis will give us some fast progress, but if the gradients are small, then the second style might be more helpful. In particular, we'll be able to show that the "function progress" proportional to $\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$ achieved by the first style of analysis can be used with clever algebra to "cancel out" the growing sum proportional to $\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$ in the second analysis.

In order to put this together properly, it will be helpful to note an even more general version of Theorem 2. This next result is the starting point of the field of *online learning*, although we will not need to go any further in this direction here.

Theorem 5. *Suppose* $\mathbf{g}_1, \dots, \mathbf{g}_T$ *is a* completely arbitrary *sequence of vectors. Define* $\mathbf{w}_t$ *by* $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_t$. *Then for any* $\mathbf{w}_\star$*:*

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

This result is remarkable general: it allows $\mathbf{g}_t$ to really be anything at all (even something generated by some evil adversary intent on messing you up in some way), and also allows $\mathbf{w}_\star$ to be anything at all. Note that while we are kind of doing gradient descent here, it's not really clear that the $\mathbf{g}_t$s actually represent gradients of anything. The proof of this result is actually identical to the proof of Theorem 2. Let's just spell it out below to make sure:

Proof. Again, we consider $\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2$. Note that this is only done "in analysis", at no point does the algorithm actually compute this quantity. That's a good thing, because $\mathbf{w}_\star$ could be anything!

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &= \|\mathbf{w}_t - \eta \mathbf{g}_t - \mathbf{w}_\star\|^2 \\ &= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\eta \langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_\star \rangle + \eta^2 \|\mathbf{g}_t\|^2\end{aligned}$$

rearrange:

$$\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_\star \rangle \leq \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\|\mathbf{g}_t\|^2$$

sum over t, and telescope:

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

and finally drop the negative term to conclude the desired result. □

This theorem will in some sense "free" us to do a lot of useful algebraic manipulations without worrying about the relationship between various vectors we produce and actual gradients of $\nabla\mathcal{L}(\mathbf{w}_t)$. Without further ado, let us describe our accelerated gradient descent algorithm:

Algorithm 1 Accelerated Gradient Descent

Input: Initial Point $\mathbf{w}_1$, smoothness constant H, time horizon T, learning rates $\eta_{\mathbf{w}}$ and $\eta_{\mathbf{y}}$:
Set $\eta_{\mathbf{w}} = \frac{1}{2H}$.
Set $\eta_{\mathbf{y}} = \frac{1}{H}$
Set $\mathbf{y}_0 = \mathbf{w}_1$
for $t = 1 \dots T$ **do**
  Set $\tau_t = \frac{t}{\sum_{i=1}^t i}$
  Set $\mathbf{x}_t = (1-\tau_t)\mathbf{y}_{t-1} + \tau_t \mathbf{w}_t$
  Set $\mathbf{g}_t = t\nabla\mathcal{L}(\mathbf{x}_t)$.
  Set $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_{\mathbf{w}}\mathbf{g}_t$.
  Set $\mathbf{y}_t = \mathbf{x}_t - \eta_{\mathbf{y}}\nabla\mathcal{L}(\mathbf{x}_t)$
end for

To gain some appreciation for what's happening here, first consider the case $\eta_{\mathbf{y}} = 0$. Then the formula for $\mathbf{x}_t$ reduces to:

$$\mathbf{x}_t = (1-\tau_t)\mathbf{x}_{t-1} + \tau_t\mathbf{w}_t$$

With a little bit of algebra, you can show that this is the same as:

$$\mathbf{x}_t = \frac{\sum_{i=1}^t i\mathbf{w}_i}{\sum_{i=1}^t i}$$

That is, $\mathbf{x}_t$ is a *weighted average* of the $\mathbf{w}_t$. Thus, we might expect some result like Theorem 4 to hold here, since that result involved averages.

However, the average is "perturbed" a bit by instead averaging with $\mathbf{y}_{t-1}$ instead of $\mathbf{x}_{t-1}$. Now, $\mathbf{y}_{t-1}$ is obtained from $\mathbf{x}_{t-1}$ simply by taking a gradient descent step with learning rate $\eta_{\mathbf{y}}$. Thus, we expect $\mathbf{y}_{t-1}$ to be a *better* point than $\mathbf{x}_{t-1}$ in terms of function value.

Theorem 6. *Suppose $\mathcal{L}$ is an H-smooth convex function. Let $\eta_{\mathbf{y}} \leq \frac{1}{2H}$ and $\eta_{\mathbf{w}} \leq \frac{\eta_{\mathbf{y}}}{2}$. Then if $\mathbf{w}_\star = \operatorname{argmin}\mathcal{L}(\mathbf{w})$, Algorithm 1 guarantees:*

$$\mathcal{L}(\mathbf{x}_T) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{2\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{T(T+1)\eta_{\mathbf{w}}}$$

In particular, with $\eta_{\mathbf{y}} = \frac{1}{2H}$ and $\eta_{\mathbf{w}} = \frac{1}{4H}$, we have:

$$\mathcal{L}(\mathbf{x}_T) - \mathcal{L}(\mathbf{w}_\star) \leq O\left(\frac{H\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{T^2}\right)$$

Proof. Let's start by examining the quantity $\sum_{t=1}^T t(\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_\star))$. By convexity, we have $\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_\star) \leq \langle\nabla\mathcal{L}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{w}_\star\rangle$, so

$$\begin{aligned}
\sum_{t=1}^T t(\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_\star)) &\leq \sum_{t=1}^T t\langle\nabla\mathcal{L}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{w}_\star\rangle \\
&= \sum_{t=1}^T t\langle\nabla\mathcal{L}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{w}_t\rangle + \sum_{t=1}^T t\langle\nabla\mathcal{L}(\mathbf{x}_t), \mathbf{w}_t - \mathbf{w}_\star\rangle \\
&= \sum_{t=1}^T \langle\nabla\mathcal{L}(\mathbf{x}_t), t(\mathbf{x}_t - \mathbf{w}_t)\rangle + \sum_{t=1}^T \langle\mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_\star\rangle
\end{aligned}$$

Now, notice that the second sum $\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_\star\rangle$ can be bounded by Theorem 5:

$$\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_\star\rangle \leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta_{\mathbf{w}}} + \frac{\eta_{\mathbf{w}}}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|^2$$

Note that even though the relationship between $\mathbf{g}_t$ and $\mathbf{w}_t$ is somewhat complicated, *this is totally irrelevant* because Theorem 5 works for *any* sequence of $\mathbf{g}_t$.

Next, look at the term $t(\mathbf{x}_t - \mathbf{w}_t)$. Let's do some tricky algebra using the definition of $\mathbf{x}_t$:

$$\mathbf{x}_t = (1-\tau_t)\mathbf{y}_{t-1} + \tau_t\mathbf{w}_t = \left(1 - \frac{t}{\sum_{i=1}^{t} i}\right)\mathbf{y}_{t-1} + \frac{t}{\sum_{i=1}^{t} i}\mathbf{w}_t$$
$$\left(\sum_{i=1}^{t} i\right)\mathbf{x}_t = \left(\left(\sum_{i=1}^{t} i\right) - t\right)\mathbf{y}_{t-1} + t\mathbf{w}_t$$

subtract $\left(\sum_{i=1}^{t-1} i\right)\mathbf{x}_t$ and $t\mathbf{w}_t$ from both sides:

$$\begin{aligned} t\mathbf{x}_t - t\mathbf{w}_t &= \left(\left(\sum_{i=1}^{t} i\right) - t\right)\mathbf{y}_{t-1} - \left(\sum_{i=1}^{t-1} i\right)\mathbf{x}_t \\ &= \left(\sum_{i=1}^{t-1} i\right)(\mathbf{y}_{t-1} - \mathbf{x}_t) \end{aligned}$$

Therefore, we have:

$$\langle \nabla\mathcal{L}(\mathbf{x}_t), t(\mathbf{x}_t - \mathbf{w}_t)\rangle = \left(\sum_{i=1}^{t-1} i\right)\langle \nabla\mathcal{L}(\mathbf{x}_t), \mathbf{y}_{t-1} - \mathbf{x}_t\rangle$$

Now, let's use convexity again: we have $\mathcal{L}(\mathbf{y}_{t-1}) \geq \mathcal{L}(\mathbf{x}_t) + \langle \nabla\mathcal{L}(\mathbf{x}_t), \mathbf{y}_{t-1} - \mathbf{x}_t\rangle$, so:

$$\langle \nabla\mathcal{L}(\mathbf{x}_t), t(\mathbf{x}_t - \mathbf{w}_t)\rangle \leq \left(\sum_{i=1}^{t-1} i\right)(\mathcal{L}(\mathbf{y}_{t-1}) - \mathcal{L}(\mathbf{x}_t))$$

Finally, let's relate $\mathcal{L}(\mathbf{y}_{t-1})$ to $\mathcal{L}(\mathbf{x}_{t-1})$: using our standard smoothness arguments, so long as $\eta_{\mathbf{y}} \leq \frac{1}{H}$, we have:

$$\mathcal{L}(\mathbf{y}_{t-1}) \leq \mathcal{L}(\mathbf{x}_{t-1}) - \frac{\eta_{\mathbf{y}}}{2}\|\nabla\mathcal{L}(\mathbf{x}_{t-1})\|^2$$

Thus:

$$\langle \nabla\mathcal{L}(\mathbf{x}_t), t(\mathbf{x}_t - \mathbf{w}_t)\rangle \leq \left(\sum_{i=1}^{t-1} i\right)(\mathcal{L}(\mathbf{x}_{t-1}) - \mathcal{L}(\mathbf{x}_t)) - \left(\sum_{i=1}^{t-1} i\right)\frac{\eta_{\mathbf{y}}}{2}\|\nabla\mathcal{L}(\mathbf{x}_{t-1})\|^2$$

Going back and putting this all together, we have shown:

$$\sum_{t=1}^{T} t(\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_\star)) \leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta_{\mathbf{w}}} + \frac{\eta_{\mathbf{w}}}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|^2 + \sum_{t=1}^{T}\left(\sum_{i=1}^{t-1} i\right)\left(\mathcal{L}(\mathbf{x}_{t-1}) - \mathcal{L}(\mathbf{x}_t) - \frac{\eta_{\mathbf{y}}}{2}\|\nabla\mathcal{L}(\mathbf{x}_{t-1})\|^2\right)$$

Now, notice that the sum

$$\sum_{t=1}^{T}\left(\sum_{i=1}^{t-1} i\right)(\mathcal{L}(\mathbf{x}_{t-1}) - \mathcal{L}(\mathbf{x}_t))$$

almost telescopes. If you want the sum to actually telescope, you should consider:

$$\sum_{t=1}^{T}\left(\sum_{i=1}^{t-1} i\right)\mathcal{L}(\mathbf{x}_{t-1}) - \left(\sum_{i=1}^{t} i\right)\mathcal{L}(\mathbf{x}_t) = -\left(\sum_{i=1}^{T} i\right)\mathcal{L}(\mathbf{x}_T)$$

Now, the difference between this telescoping sum and the sum we actually have is just:

$$\sum_{t=1}^{T} t\mathcal{L}(\mathbf{x}_t)$$

Thus, we can re-write our expression as:

$$\sum_{t=1}^{T} t(\mathcal{L}(\mathbf{x}_t) - \mathcal{L}(\mathbf{w}_\star)) \leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta_{\mathbf{w}}} + \frac{\eta_{\mathbf{w}}}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|^2 - \left(\sum_{i=1}^{T} i\right)\mathcal{L}(\mathbf{x}_T) + \sum_{t=1}^{T} t\mathcal{L}(\mathbf{x}_t) - \sum_{t=1}^{T}\left(\sum_{i=1}^{t-1} i\right)\frac{\eta_{\mathbf{y}}}{2}\|\nabla\mathcal{L}(\mathbf{x}_{t-1})\|^2$$

now, subtract $\sum_{t=1}^{T} t\mathcal{L}(\mathbf{x}_t)$ and add $\left(\sum_{i=1}^{T} i\right)\mathcal{L}(\mathbf{x}_T)$ to both sides:

$$\left(\sum_{i=1}^{T} i\right)(\mathcal{L}(\mathbf{x}_T) - \mathcal{L}(\mathbf{w}_\star)) \leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta_{\mathbf{w}}} + \frac{\eta_{\mathbf{w}}}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|^2 - \sum_{t=1}^{T}\left(\sum_{i=1}^{t-1} i\right)\frac{\eta_{\mathbf{y}}}{2}\|\nabla\mathcal{L}(\mathbf{x}_{t-1})\|^2$$

Now, let's finally start to use the identity $\sum_{i=1}^{t} i = \frac{t(t+1)}{2}$. We have:

$$\left(\frac{T(T+1)}{2}\right)(\mathcal{L}(\mathbf{x}_T) - \mathcal{L}(\mathbf{w}_\star)) \leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta_{\mathbf{w}}} + \frac{\eta_{\mathbf{w}}}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|^2 - \sum_{t=1}^{T}\frac{t(t-1)}{2}\frac{\eta_{\mathbf{y}}}{2}\|\nabla\mathcal{L}(\mathbf{x}_{t-1})\|^2$$

insert the definition of $\mathbf{g}_t$:

$$= \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta_{\mathbf{w}}} + \frac{\eta_{\mathbf{w}}}{2}\sum_{t=1}^{T} t^2\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2 - \sum_{t=1}^{T}\frac{t(t-1)}{2}\frac{\eta_{\mathbf{y}}}{2}\|\nabla\mathcal{L}(\mathbf{x}_{t-1})\|^2$$

reindex the last sum:

$$\begin{aligned}
&= \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta_{\mathbf{w}}} + \frac{\eta_{\mathbf{w}}}{2}\sum_{t=1}^{T} t^2\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2 - \sum_{t=1}^{T-1}\frac{t(t+1)}{2}\frac{\eta_{\mathbf{y}}}{2}\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2 \\
&\leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta_{\mathbf{w}}} + \frac{\eta_{\mathbf{w}}}{2}\sum_{t=1}^{T} t^2\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2 - \frac{\eta_{\mathbf{y}}}{4}\sum_{t=1}^{T-1} t^2\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2
\end{aligned}$$

Now, let us use our assumption that $\eta_{\mathbf{w}} \leq \eta_{\mathbf{y}}/2$:

$$\begin{aligned}
\left(\frac{T(T+1)}{2}\right)(\mathcal{L}(\mathbf{x}_T) - \mathcal{L}(\mathbf{w}_\star)) &\leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta_{\mathbf{w}}} + \frac{\eta_{\mathbf{y}}}{4}\sum_{t=1}^{T} t^2\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2 - \frac{\eta_{\mathbf{y}}}{4}\sum_{t=1}^{T-1} t^2\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2 \\
&= \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta_{\mathbf{w}}} + \frac{T^2\eta_{\mathbf{y}}}{4}\|\nabla\mathcal{L}(\mathbf{w}_T)\|^2
\end{aligned}$$

From Lemma 3:

$$\leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta_{\mathbf{w}}} + \frac{HT^2\eta_{\mathbf{y}}}{2}(\mathcal{L}(\mathbf{x}_T) - \mathcal{L}(\mathbf{w}_\star))$$

using $\eta_{\mathbf{y}} \leq \frac{1}{2H}$:

$$\leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta_{\mathbf{w}}} + \frac{T(T+1)}{4}(\mathcal{L}(\mathbf{x}_T) - \mathcal{L}(\mathbf{w}_\star))$$

rearranging:

$$\left(\frac{T(T+1)}{2}\right)(\mathcal{L}(\mathbf{x}_T) - \mathcal{L}(\mathbf{w}_\star)) \leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{\eta_{\mathbf{w}}}$$

from which the result follows. □

References

[1] Sébastien Bubeck et al. "Convex Optimization: Algorithms and Complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.

[2] Yurii Nesterov. "Introductory Lectures on Convex Optimization A Basic Course". In: ().

[3] Zeyuan Allen Zhu and Lorenzo Orecchia. "Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent". In: *Innovations in Theoretical Computer Science Conference, ITCS*. 2017.