Lecture Notes 8: Smooth Convex losses and Acceleration

Instructor: Ashok Cutkosky

1 Smooth convex losses

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:

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

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

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

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 $\|\mathbf{w}_\star - \mathbf{w}_1\| \le D$. Set $\eta = \frac{D}{G\sqrt{T}}$. Then*

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

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

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

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:*

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

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

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

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:

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

sum over t, and telescope the RHS:

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

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

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 $\mathbf{w}$:*

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

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:

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

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

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

□

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:*

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

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

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

Proof. From Theorem 2, we have

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

now, from Lemma 3:

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

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:

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

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:

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

Thus,

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

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. The original paper is in Russian, but there are texts that cover it in various levels of detail [1, 2].

2 Momentum

Before getting into the technical details of acceleration, let's discuss a possibly more intuitive algorithm. This is the idea of *momentum*. Momentum is probably the single most ubiquitous modification added to SGD/GD in machine learning. The basic idea can be viewed via a kind of "physical intuition". If we consider the objective function $\mathcal{L}(\mathbf{w})$ as a kind of "contour map" in which $\mathcal{L}(\mathbf{w})$ indicates the height of the landscape at some "d-dimensional" GPS coordinates $\mathbf{w}$, then we might view gradient descent as an algorithm that finds a local minimum of $\mathcal{L}$ by picking repeatedly taking a step in the direction of steepest incline of the landscape.

However, if you were to drop a smooth bowling ball at some point on the landscape, it actually would not always take the direction of steepest incline: as it rolls it would pick up some *momentum* that would keep it going uphill occasionally. Although going uphill seems bad, overall we might feel this is a good thing because the momentum will allow the ball to roll down hill faster. An important point here is the idea of *friction* - without some friction to decrease the kinetic energy of the bowling ball, it will happily roll uphill exactly as far as it has rolled downhill and make zero progress. By adding an appropriate amount of friction, we could hope to limit the uphill motion while still getting the gains for the downhill motion. Intuitively, the regular gradient descent is like a bowling ball that has a near-infinite amount of friction associated with its motion so that it cannot pick up any momentum.

This idea is captured by defining an auxiliary variable $\mathbf{v}$, which represents the "velocity" of the parameter $\mathbf{w}$. $\mathbf{v}$ is updated using the gradient of $\mathcal{L}$, so that $\mathcal{L}$ is playing the role of a "potential energy":

$$\mathbf{v}_t = \beta \mathbf{v}_{t-1} - \eta \nabla \mathcal{L}(\mathbf{w}_t) \tag{1}$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \mathbf{v}_t \tag{2}$$

Here the coefficient $\beta \in [0, 1]$ plays the role of "friction". $\beta = 0$ corresponds to "infinite friction" - all of the kinetic energy is immediately removed and so no velocity gets to carry over to the next iterate. Inspecting the equations shows that with $\beta = 0$, the update is identical to regular gradient descent. $\beta = 1$ corresponds to zero friction. Typically β is set to some large value like 0.9 or 0.99.

2.1 Better integrators

Algorithm 1 Gradient Descent with Momentum

```
Input: Initial Point $\mathbf{w}_1$, learning rates $\eta_1, \dots, \eta_T$, momentum parameters $\beta_1, \dots, \beta_T$:
Set $\mathbf{v}_0 = 0$.
for $t = 1 \dots T$ do
    if Using two-step integrator then
        Set $\mathbf{v}_{t-\frac{1}{2}} = \beta_t \mathbf{v}_{t-1}$.
        Set $\mathbf{w}_{t+\frac{1}{2}} = \mathbf{w}_t - \mathbf{v}_{t-\frac{1}{2}}$
        Set $\mathbf{g}_t = \nabla \mathcal{L}(\mathbf{w}_{t+\frac{1}{2}})$.
    else
        Set $\mathbf{g}_t = \nabla \mathcal{L}(\mathbf{w}_t)$
    end if
    Set $\mathbf{v}_t = \beta_t \mathbf{v}_{t-1} - \eta_t \mathbf{g}_t$.
    Set $\mathbf{w}_{t+1} = \mathbf{w}_t - \mathbf{v}_t$.
end for
```

We can view the momentum update as numerically integrating the following physical "equations of motion":

$$\mathbf{v} = \frac{d\mathbf{w}}{dt}$$

$$\frac{d\mathbf{v}}{dt} = -(1-\beta)\mathbf{v} + \eta \nabla \mathcal{L}(\mathbf{w})$$

The momentum update described in equations (1) and (2) correspond to the most basic Euler integration step. However, there are more accurate ways to simulate this differential equation. Specifically, consider the following "two-step" integrator:

$$\mathbf{v}_{t-\frac{1}{2}} = \beta \mathbf{v}_{t-1}$$

$$\mathbf{w}_{t+\frac{1}{2}} = \mathbf{w}_t - \mathbf{v}_{t-\frac{1}{2}}$$

$$\mathbf{v}_t = \beta \mathbf{v}_{t-1} - \eta \nabla \mathcal{L}(\mathbf{w}_{t+\frac{1}{2}})$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \mathbf{v}_t$$

The intuition here is that the equation $\frac{d\mathbf{v}}{dt} = -(1-\beta)\mathbf{v} + \eta\nabla\mathcal{L}(\mathbf{w})$ containts two terms, a $-(1-\beta)\mathbf{v}$ and a $\eta\nabla\mathcal{L}(\mathbf{w})$. If we naively evaluate $\nabla\mathcal{L}(\mathbf{w})$ at $\mathbf{w}_t$, we might miss some behavior in which the ball starts rolling up the hill during this step. By instead evaluating at the "lookahead" point $\mathbf{w}_t + \beta\mathbf{v}_t$, we can help to counteract this by detecting if just the momentum on its own would cause the ball to start rolling uphill.

This leads to Algorithm 1 in which we generalize the setting to allow for time-varying β and η.

Unfortunately, in order to rigorously analyze this algorithm in the convex setting and show acceleration, we will completely abandon this intuition and make an argument that follows a mess of algebra. Further, we will need to set a very delicate schedule for β_t, and re-write the algorithm into very different looking (but totally equivalent) form. This is a big issue with accelerated algorithms: although there are reasonable intuitive explanations in terms of physics like the above, the analyses are typically very strange-looking.

3 Acceleration

Now, let us forget for a momentum about momentum and just think about some equations we've seen so far at a very high level and how we might be able to use them to get a faster convergence. Our approa ch is a streamlined presentation of *linear coupling* [3], which is a more recent technique for designing accelerated algorithms.

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

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

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

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

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

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

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* $\mathbf{w}_\star$*:*

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

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:

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

sum over t, and telescope:

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

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 2 Accelerated Gradient Descent via Momentum

```
Input: Initial Point $\mathbf{w}_1$, smoothness constant $H$, time horizon $T$, learning rate $\eta$
Set $\mathbf{y}_1 = \mathbf{w}_1$
Set $\alpha_0 = 0$, $\alpha_1 = 1$.
for $t = 1 \dots T$ do
    Set $\tau_t = \frac{\alpha_t}{\sum_{i=1}^t \alpha_t}$
    Set $\mathbf{x}_t = (1 - \tau_t)\mathbf{w}_t + \tau_t\mathbf{y}_t$
    Set $\mathbf{g}_t = \alpha_t\nabla\mathcal{L}(\mathbf{x}_t)$.
    Set $\mathbf{y}_{t+1} = \mathbf{y}_t - \eta\mathbf{g}_t$.
    Set $\mathbf{w}_{t+1} = \mathbf{x}_t - \eta\nabla\mathcal{L}(\mathbf{x}_t)$
    Set $\alpha_{t+1}$ to satisfy $\alpha_{t+1}^2 - \alpha_{t+1} = \sum_{i=1}^t \alpha_i$ (use quadratic formula).
end for
```

Now, this algorithm looks extremely weird, but with a little work we can massage it to look like the momentum algorithm. Specifically, set

$$\mathbf{v}_t = \frac{1}{\alpha_t}(\mathbf{y}_t - \mathbf{w}_t) - \eta\nabla\mathcal{L}(\mathbf{x}_t)$$

Observe that we have:

$$\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{x}_t - \eta\nabla\mathcal{L}(\mathbf{x}_t) \\
&= (1 - \tau_t)\mathbf{w}_t + \tau_t\mathbf{y}_t - \eta\nabla\mathcal{L}(\mathbf{x}_t) \\
&= \mathbf{w}_t + \tau_t(\mathbf{y}_t - \mathbf{w}_t) - \eta\nabla\mathcal{L}(\mathbf{x}_t)
\end{aligned}$$

using the fact that $\alpha_t = \sum_{i=1}^t \alpha_i$, so that $\tau_t = \frac{1}{\alpha_t}$:

$$\begin{aligned}
&= \mathbf{w}_t + \frac{1}{\alpha_t}(\mathbf{y}_t - \mathbf{w}_t) - \eta\nabla\mathcal{L}(\mathbf{x}_t) \\
&= \mathbf{w}_t + \mathbf{v}_t
\end{aligned}$$

Further, we have the important relationship:

$$\begin{aligned}\mathbf{y}_{t+1} - \mathbf{w}_{t+1} &= \mathbf{y}_t - \eta\alpha_t \nabla\mathcal{L}(\mathbf{x}_t) - \mathbf{w}_{t+1}\\ &= \mathbf{y}_t - \eta\alpha_t \nabla\mathcal{L}(\mathbf{x}_t) - \mathbf{x}_t + \eta\nabla\mathcal{L}(\mathbf{x}_t)\\ &= \mathbf{y}_t - (1-\tau_t)\mathbf{w}_t + \tau_t\mathbf{y}_t - \eta(\alpha_t - 1)\nabla\mathcal{L}(\mathbf{x}_t)\\ &= (1-\tau_t)(\mathbf{y}_t - \mathbf{w}_t) - \eta(\alpha_t - 1)\nabla\mathcal{L}(\mathbf{x}_t)\end{aligned}$$

Now, observe two more magical properties of our definition for α_t:

$$\begin{aligned}\alpha_t - 1 &= \frac{\alpha_t^2 - \alpha_t}{\alpha_t} = \frac{\sum_{i=1}^{t-1}\alpha_i}{\alpha_t} = \frac{\alpha_{t-1}^2}{\alpha_t}\\ 1-\tau_t &= \frac{\sum_{i=1}^{t-1}\alpha_i}{\sum_{i=1}^{t}\alpha_i} = \frac{\alpha_{t-1}^2}{\alpha_t^2}\end{aligned}$$

Combining this with the above implies:

$$\mathbf{y}_{t+1} - \mathbf{w}_{t+1} = \frac{\alpha_{t-1}^2}{\alpha_t}\mathbf{v}_t$$

or, with t instead of $t+1$ for all $t \geq 2$:

$$\mathbf{y}_t - \mathbf{w}_t = \frac{\alpha_{t-2}^2}{\alpha_{t-1}}\mathbf{v}_{t-1} \tag{3}$$

$$\mathbf{v}_t = \frac{\alpha_{t-2}^2}{\alpha_t\alpha_{t-1}}\mathbf{v}_{t-1} - \eta\nabla\mathcal{L}(\mathbf{x}_t)$$

Finally, observe that

$$\begin{aligned}\mathbf{x}_t &= \mathbf{w}_t + \tau_t(\mathbf{y}_t - \mathbf{w}_t)\\ &= \mathbf{w}_t + \frac{1}{\alpha_t}(\mathbf{y}_t - \mathbf{w}_t)\end{aligned}$$

using equation (3):

$$= \mathbf{w}_t + \frac{\alpha_{t-2}^2}{\alpha_{t-1}\alpha_t}\mathbf{v}_{t-1}$$

Thus, defining $\beta_1 = 0$ and $\beta_t = \frac{\alpha_{t-2}^2}{\alpha_t\alpha_{t-1}}$ for all $t \geq 2$ yields the updates for all $t \geq 1$:

$$\begin{aligned}\mathbf{x}_t &= \mathbf{w}_t + \beta_t\mathbf{v}_{t-1}\\ \mathbf{v}_t &= \beta_t\mathbf{v}_{t-1} - \eta\nabla\mathcal{L}(\mathbf{x}_t)\\ \mathbf{w}_{t+1} &= \mathbf{w}_t + \mathbf{v}_t\end{aligned}$$

which is exactly the form of the two-step momentum integrator used in Algorithm 1.

3.1 Acceleration analysis

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

$$\mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{9\|\mathbf{w}_\star - \mathbf{w}_0\|^2}{2T^2\eta}$$

In particular, with $\eta = \frac{1}{H}$, we have:

$$\mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_\star) \leq O\left(\frac{H\|\mathbf{w}_\star - \mathbf{w}_0\|^2}{T^2}\right)$$

Before proving the Theorem, let's show a small proposition to gain some intuition for how the α_t behave:

Proposition 7. *For all* $t \geq 1$*,*

$$\frac{t^2}{9} \leq \sum_{i=1}^{t} \alpha_i \leq t^2$$

Further, $\alpha_t = \sum_{i=1}^{t} \alpha_i$*.*

Proof. The second part of the Proposition is immediate from the definition of α_t. The tricky part is proving the first statement. For this, we proceed by induction. The base case for $t = 1$ is clear from definition of α_1. Let us assume the statement holds for some t. By definition of α_{t+1}, we have:

$$\begin{aligned} \alpha_{t+1} &= \frac{1 + \sqrt{1 + 4\sum_{i=1}^{t-1} \alpha_i}}{2} \\ &\leq 1 + \sqrt{\sum_{i=1}^{t-1} \alpha_i} \end{aligned}$$

So therefore:

$$\sum_{i=1}^{t+1} \alpha_i \leq \sum_{i=1}^{t} \alpha_i + 1 + \sqrt{\sum_{i=1}^{t-1} \alpha_i}$$

using the induction assumption:

$$\begin{aligned} &\leq t^2 + 1 + t \\ &\leq (t+1)^2 \end{aligned}$$

For the lower bound, we have

$$\alpha_{t+1} \geq \sqrt{\sum_{i=1}^{t-1} \alpha_i}$$

so that:

$$\begin{aligned} \sum_{i=1}^{t+1} \alpha_i &\geq \sum_{i=1}^{t} \alpha_i + \sqrt{\sum_{i=1}^{t-1} \alpha_i} \\ &\geq \frac{t^2}{9} + \frac{t}{3} \end{aligned}$$

using $t \geq 1$:

$$\begin{aligned} &\geq \frac{t^2}{9} + \frac{2t}{9} + \frac{1}{9} \\ &= \frac{(t+1)^2}{9} \end{aligned}$$

□

This shows that $\sum_{i=1}^{t} \alpha_i$ grows quadratically in t. Now, we are ready to prove Theorem 6:

Proof of Theorem 6. Let's start by examining the quantity $\sum_{t=1}^T \alpha_t(\mathcal{L}(\mathbf{x}_t)-\mathcal{L}(\mathbf{w}_\star))$. By convexity, we have $\mathcal{L}(\mathbf{x}_t)-\mathcal{L}(\mathbf{w}_\star) \le \langle \nabla\mathcal{L}(\mathbf{x}_t), \mathbf{x}_t-\mathbf{w}_\star\rangle$, so

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

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

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

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

Next, look at the term $\alpha_t(\mathbf{x}_t-\mathbf{y}_t)$. Let's do some tricky algebra using the definition of $\mathbf{x}_t$:

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

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

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

Therefore, we have:

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

Now, let's use convexity again: we have $\mathcal{L}(\mathbf{w}_t) \ge \mathcal{L}(\mathbf{x}_t) + \langle \nabla\mathcal{L}(\mathbf{x}_t), \mathbf{w}_t-\mathbf{x}_t\rangle$, so:

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

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

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

Now, eventually we are going to want the last sum to telescope. So far there are two obstacles. First, there is a $\mathbf{w}$ instead of a $\mathbf{x}$, and second the coefficients on the $\mathcal{L}(\mathbf{w}_t)$ and $\mathcal{L}(\mathbf{x}_t)$ are the same. Let's fix the second problem first: subtract $\sum_{t=1}^T \alpha_t\mathcal{L}(\mathbf{x}_t)$ from both sides to get,

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

Now, we can see that if it weren't for the $\mathbf{x}$ vs $\mathbf{w}$ mismatch, the last sum would indeed telescope. Let's use smoothness to relate $\mathcal{L}(\mathbf{x}_t)$ to $\mathcal{L}(\mathbf{w}_{t+1})$. So long as $\eta \le \frac{1}{H}$, we have:

$$\begin{aligned}\mathcal{L}(\mathbf{w}_{t+1}) &\le \mathcal{L}(\mathbf{x}_t) - \frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2\\ -\mathcal{L}(\mathbf{x}_t) &\le -\mathcal{L}(\mathbf{w}_{t+1}) - \frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{x}_t)\|^2\end{aligned}$$

Thus, we have:

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

Finally, the sum $\sum_{t=1}^T\left[\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathcal{L}(\mathbf{w}_t) - \left(\sum_{i=1}^{t}\alpha_i\right)\mathcal{L}(\mathbf{w}_{t+1})\right]$ telescopes, yielding:

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

rearrange:

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

Now, let's finally start to use the definition of α_t. Note that we have $\alpha_t^2 = \sum_{i=1}^t\alpha_i$. Thus:

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

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

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

using Proposition 7:

$$\frac{T^2}{9}(\mathcal{L}(\mathbf{x}_T) - \mathcal{L}(\mathbf{w}_\star)) \le \frac{\|\mathbf{w}_\star - \mathbf{y}_1\|^2}{2\eta}$$

from which the result follows by observing that $\mathbf{y}_1 = \mathbf{w}_1$. □

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.