Optimization for Data Science

Bernd Gärtner, Martin Jaggi

1 Gradient Descent

A few quick notes:1

We assume $f : \mathbb{R}^d \to \mathbb{R}$ is a differentiable convex function and that it has a global minimum $\mathbf{x}^* \in \mathbb{R}^d$, and then our goal is finding a $\mathbf{x} \in \mathbb{R}^d$ with $\varepsilon > 0$ such that:

$$f(\mathbf{x}) - f(\mathbf{x}^*) < \varepsilon$$

So We're not necessarily chasing the global minimum, just any point in the domain that is minimal.

2 The Algorithm

The old familliar:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)$$

Nothing shocking here. We are just doing our best at every step to reduce the function. $\mathbf{x}_0 \in \mathbb{R}^d$ is simply random.

γ there is the stepsize. Larger stepsize means larger steps, which means We may "overshoot" our target. Furthermore We had

$$f(\mathbf{x}_t + \mathbf{v}_t) = f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{v}_t) + r(\mathbf{x}_t - \mathbf{v}_t) \approx f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{v}_t)$$

So We're losing that r term, the direction is according to the gradient and for the time being γ is fixed (though it may make sense to vary it in accordance to stuff).

3 Vanilla Analysis

Okay so up first We let $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$ which gives us

$$\begin{aligned}
\mathbf{x}_{t+1} &= \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t) \\
\mathbf{x}_{t+1} &= \mathbf{x}_t - \gamma \mathbf{g}_t \\
\mathbf{x}_{t+1} - \mathbf{x}_t &= -\gamma \mathbf{g}_t \\
\frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1}) &= \mathbf{g}_t
\end{aligned}$$

And now We do

$$\mathbf{g}^\top (\mathbf{x}_t - \mathbf{x}^*) = \frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top (\mathbf{x}_t - \mathbf{x}^*)$$

What are We up to here? We are project the well, this is weird. The problem I'm having here is that $\mathbf{g}_t$ seems to be pointing away from the minimum, while oh they are both away from the minimum. Weird way of phrasing it. So We taking the gradient at time t, and project on to that step We take *from* the global minimum *to* the point $\mathbf{x}_t$. So this is just getting the overlap of those two directions (plus some scaling by lengths, but whatever).

At this point see appendix for cosine law thingie.

So, by the cosine law We've got

$$\begin{aligned}
\|\mathbf{a} - \mathbf{b}\|^2 &= \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\mathbf{a}^\top \mathbf{b} \\
2\mathbf{a}^\top \mathbf{b} &= \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2 \\
\mathbf{a}^\top \mathbf{b} &= \frac{1}{2}\big(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2\big)
\end{aligned}$$

Slotting in our vectors there We get

$$\begin{aligned}
&\frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top (\mathbf{x}_t - \mathbf{x}^*) = \\
&\qquad \frac{1}{\gamma}\frac{1}{2}\big(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|(\mathbf{x}_t - \mathbf{x}_{t+1} - (\mathbf{x}_t - \mathbf{x}^*)\|^2\big) \\
&\qquad\qquad = \frac{1}{2\gamma}\big(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_{t+1}\|^2\big)
\end{aligned}$$

All that happens then is We stare at it for a little while and observe that $\mathbf{x}_t - \mathbf{x}_{t+1} = \gamma \mathbf{g}_t$, so

$$\frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top(\mathbf{x}_t - \mathbf{x}^*) = \frac{1}{2\gamma}\big(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_{t+1}\|^2\big)$$
$$\frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top(\mathbf{x}_t - \mathbf{x}^*) = \frac{1}{2\gamma}\big(\|\gamma\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_{t+1}\|^2\big)$$
$$\frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top(\mathbf{x}_t - \mathbf{x}^*) = \frac{1}{2\gamma}\big(\gamma^2\|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_{t+1}\|^2\big)$$
$$\frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top(\mathbf{x}_t - \mathbf{x}^*) = \frac{\gamma}{2}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_{t+1}\|^2\big)$$

Now, if We look at that second term, it's a bit silly. Take $t = 0$:

$$\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_1\|^2$$

What is that? The first term is the vector from the start of the sequence to the end, and the second term is the vector from the first step in the sequence to the end. We're left then with the (squared) change in magnitude due to the first step. With $t = 1$, We'll be left with the change in magnitude due to the second step and so forth. Ultimately, We're just splitting the distance from the start to the end lots of little chunks! This becomes useful if We think about summing over our series, like so:

$$\frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top(\mathbf{x}_t - \mathbf{x}^*) = \frac{\gamma}{2}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_{t+1}\|^2\big)$$
$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^*) = \frac{\gamma}{2}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_{t+1}\|^2\big)$$
$$\sum_{t=0}^{T-1}\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^*) = \sum_{t=0}^{T-1}\left(\frac{\gamma}{2}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_{t+1}\|^2\big)\right)$$
$$\sum_{t=0}^{T-1}\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^*) = \sum_{t=0}^{T-1}\left(\frac{\gamma}{2}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_{t+1}\|^2\big)\right)$$
$$\sum_{t=0}^{T-1}\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^*) = \sum_{t=0}^{T-1}\left(\frac{\gamma}{2}\|\mathbf{g}_t\|^2\right) + \frac{1}{2\gamma}\big(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_T\|^2\big)$$

Up to this point We are precise. Now We are going to drop the distance between our last point and the global minimum. I suppose an argument for doing is is that there is not much point in keeping track of the last bit - T steps is what We are given, so that is what We work with. In theory You can get a tighter bound by keeping this term around I suppose.

$$\sum_{t=0}^{T-1} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*) = \sum_{t=0}^{T-1} \left(\frac{\gamma}{2} \|\mathbf{g}_t\|^2 \right) + \frac{1}{2\gamma} \left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_T\|^2 \right)$$
$$\sum_{t=0}^{T-1} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*) \leq \sum_{t=0}^{T-1} \left(\frac{\gamma}{2} \|\mathbf{g}_t\|^2 \right) + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Ah I bet We'd drop the other term there too, but We can't - dropping the negative term makes the bound looser and We're okay with that, but We can't make it tighter without solid argumentation so We can't drop a positive norm.

Anyway, We haven't used anything about our function f yet, just the differentiability and the gradient descent structure. Using first order characterization of convexity We can say

$$\begin{aligned}
f(\mathbf{x}^*) &\geq f(\mathbf{x}_t) + \nabla f(\mathbf{x})^\top (\mathbf{x}^* - \mathbf{x}_t) \\
f(\mathbf{x}^*) - f(\mathbf{x}_t) &\geq \nabla f(\mathbf{x})^\top (\mathbf{x}^* - \mathbf{x}_t) \\
-(f(\mathbf{x}^*) - f(\mathbf{x}_t)) &\leq -\nabla f(\mathbf{x})^\top (\mathbf{x}^* - \mathbf{x}_t) \\
-(f(\mathbf{x}^*) - f(\mathbf{x}_t)) &\leq -\nabla f(\mathbf{x})^\top (\mathbf{x}^* - \mathbf{x}_t) \\
f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq \nabla f(\mathbf{x})^\top (\mathbf{x}_t - \mathbf{x^*}_t) \\
f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq \mathbf{g}^\top (\mathbf{x}_t - \mathbf{x^*}_t)
\end{aligned}$$

Wee. That last line is particularly leading, You probably see where We're going with it:

$$\sum_{t=0}^{T-1} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*) \leq \sum_{t=0}^{T-1} \left(\frac{\gamma}{2} \|\mathbf{g}_t\|^2 \right) + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$
$$\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \sum_{t=0}^{T-1} \left(\frac{\gamma}{2} \|\mathbf{g}_t\|^2 \right) + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

With this We get an average error (by simply dividing by T) upper bound, and so the smallest error attained along our series of $0..T$ is certainly less than that.

4 Lipschitz Convex Functions

Simple - We bound the gradient:

Theorem 4.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function, $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq R$, $\|\mathbf{g}_t\| \leq B$, $\forall \mathbf{x}$. Choosing stepsize*

$$\gamma = \frac{R}{B\sqrt{T}}$$

gradient descent yields

$$\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{RB}{\sqrt{T}}$$

Proof. As before We have

$$\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \sum_{t=0}^{T-1} \left(\frac{\gamma}{2} \|\mathbf{g}_t\|^2 \right) + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Using our bound for the gradients We get

$$\begin{aligned}
\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq \sum_{t=0}^{T-1} \left(\frac{\gamma}{2} \|\mathbf{g}_t\|^2 \right) + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 \\
\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq \sum_{t=0}^{T-1} \frac{\gamma B^2}{2} + \frac{R^2}{2\gamma} \\
\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq \frac{T\gamma B^2}{2} + \frac{R^2}{2\gamma}
\end{aligned}$$

Now We differentiate the right side w.r.t. γ to find an extremum:

$$\begin{aligned}
\frac{d}{d\gamma} \left(\frac{T\gamma B^2}{2} + \frac{R^2}{2\gamma} \right) &= \frac{TB^2}{2} + \frac{d}{d\gamma} \left(\frac{R^2}{2\gamma} \right) \\
&= \frac{TB^2}{2} + \frac{d}{d\gamma} \left(\frac{R^2}{2} \gamma^{-1} \right) \\
&= \frac{TB^2}{2} - \frac{R^2}{2} \gamma^{-2}
\end{aligned}$$

Now We set it to zero

$$
\begin{aligned}
0 &= \frac{TB^2}{2} - \frac{R^2}{2}\gamma^{-2} \\
-\frac{TB^2}{2} &= -\frac{R^2}{2}\gamma^{-2} \\
TB^2 &= R^2\gamma^{-2} \\
\frac{TB^2}{R^2} &= \gamma^{-2} \\
\frac{R^2}{TB^2} &= \gamma^2 \\
\frac{R}{B\sqrt{T}} &= \gamma
\end{aligned}
$$

We know this is a local minimum since the second derivative is positive. Using the above We get

$$
\begin{aligned}
\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq \frac{T\gamma B^2}{2} + \frac{R^2}{2\gamma} \\
\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq \frac{R}{B\sqrt{T}}\frac{TB^2}{2} + \frac{B\sqrt{T}}{R}\frac{R^2}{2} \\
\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq \frac{\sqrt{T}RB}{2} + \frac{B\sqrt{T}R}{2} \\
\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq RB\sqrt{T} \\
\frac{\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*)}{T} &\leq \frac{RB\sqrt{T}}{T} \\
\frac{\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*)}{T} &\leq \frac{RB}{\sqrt{T}}
\end{aligned}
$$

□

So say We want

$$
\min_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \varepsilon
$$

Well, the minimum is less than the average so We can safely say

$$
\min_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{RB}{\sqrt{T}} \leq \varepsilon
$$

And then it's just messing about:

$$\begin{aligned}\varepsilon &\geq \frac{RB}{\sqrt{T}}\\ \sqrt{T}\varepsilon &\geq RB\\ \sqrt{T} &\geq \frac{RB}{\varepsilon}\\ T &\geq \frac{R^2B^2}{\varepsilon^2}\end{aligned}$$

So We get $\mathcal{O}(\frac{1}{\varepsilon^2})$.

5 Smooth convex functions

Alright, so, smooth functions (not necessarily convex for now) are functions bounded from above by some parabaloid, like so:

Definition 5.1. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be a differentiable function, and let $X \subseteq \mathbf{dom}(f)$ be a convex set, if for some $L \geq 0$ it is true that*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2,\ \forall \mathbf{x},\mathbf{y} \in X$$

then f is smooth over X with parameter L. If $X = \mathbf{dom}(f)$, then f is simply smooth.

So convexity bounds from below, smoothness bounds from above. For polynomials We can say

Lemma 5.2. *Let $f(\mathbf{x}) = \mathbf{x}^\top Q\mathbf{x} + \mathbf{b}^\top\mathbf{x} + c$ where Q is symmetric and everything is adequately dimensioned, then f is smooth with parameter $2\|Q\|$.*

Basically smooth with the second derivative.

Lemma 5.3. *Let $f : \mathbf{R}^d \to \mathbb{R}$ be convex and differentiable, then the following two statements are equivalent:*
i) f is smooth with parameter L
ii) $\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \leq L\|\mathbf{y}-\mathbf{x}\|$

Lemma 5.4. *Operations that preserve function smoothness are scaled positive sums of said functions and composition with a linear function, where if g is a linear function then $f(g(x))$ leaves things smooth and $g(f(x))$ scales the smoothness parameter by $\|A\|^2$, where A is the spectral norm of the linear operator.*

Lemma 5.5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable and smooth function with parameter L, then with*

$$\gamma = \frac{1}{L}$$

gradient descent step satisfies

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|f(\mathbf{x}_t)\|^2$$

This is called sufficient decrease - since We're getting a decrease

Proof. Is not too bad actually. By smoothness We have

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top\left(-\frac{1}{L}\mathbf{g}_t\right) + \frac{L}{2}\left\|-\frac{1}{L}\mathbf{g}_t\right\|^2 \\
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) - \frac{1}{L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \\
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2
\end{aligned}$$

□

Theorem 5.6. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function, and furthermore let f be smooth with parameter L, then with stepsize*

$$\gamma = \frac{1}{L}$$

gradient descent satisfies

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^*\|^2,\ T > 0.$$

Proof. Well, first We use sufficient decrease to get some handle on the gradients here. This makes sense, after all, if the graph is bounded from above, this has a lot to say about the gradient. From sufficient decrease We have

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \\
f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) &\leq -\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \\
f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) &\geq \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2
\end{aligned}$$

From vanilla analysis We had:

$$\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \sum_{t=0}^{T-1} \left(\frac{\gamma}{2} \|\mathbf{g}_t\|^2 \right) + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Focusing on the gradient term there, with sufficient decrease We get We have

$$\begin{aligned} \sum_{t=0}^{T-1} \left(\frac{\gamma}{2} \|\mathbf{g}_t\|^2 \right) = \frac{1}{2L} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 &\leq \sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) \\ &\leq f(\mathbf{x}_0) - f(\mathbf{x}_T) \end{aligned}$$

Plugging this back into vanilla analysis leads to

$$\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \sum_{t=0}^{T-1} \left(\frac{\gamma}{2} \|\mathbf{g}_t\|^2 \right) + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

$$\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

We can bring over those two f terms to get

$$\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

$$\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) - (f(\mathbf{x}_0) - f(\mathbf{x}_T)) \leq \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

$$\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) - f(\mathbf{x}_0) + f(\mathbf{x}_T) \leq \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

$$\sum_{t=1}^{T} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Alright so then by sufficient decrease We know that $f(\mathbf{x}_{t+1} \leq f(\mathbf{x}_t$, so the last iterate in the sequence is the smallest, and the smallest is certainly smaller than the average yielding

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{L}{2T} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

If We wish the error to be lower than ε (letting $R = \|\mathbf{x}_0 - \mathbf{x}^*\|$) We get

$$\frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 \leq \varepsilon$$
$$\frac{LR^2}{2T} \leq \varepsilon$$
$$\frac{LR^2}{2\varepsilon} \leq T$$

□

6 Accelerated Gradient Descent

6.1 Hinton's video

Can be found here.

So We are at some point $\mathbf{x}$. Let's say We've already iterated a couple of times.

One of the terms We'll have at this point is momentum, and We go in accordance to the momentum, which is to say We'll just make a step in the direction of the momentum.

Having done so, We now have some $\mathbf{x}'$ at which We arrived by going in the direction of the momentum (I don't think We scale the momentum vector before taking the first step, by the by, so no γ yet.). Now at this new point $\mathbf{x}'$, We take the derivative and take a standard gradient descent step. Now We have a point $\mathbf{x}''$, and this will be the starting point for the next iteration.

The last thing We do now before starting all over again is We is We update the momentum. We add the gradient We just computed at $\mathbf{x}'$ to the momentum, so now the momentum vector points from $\mathbf{x}$ to $\mathbf{x}''$, and We attenuate it a bit by multiplying the momentum by some constant close to 1, like 0.95.

6.2 Their version

They have

$$\begin{aligned}
\mathbf{y}_{t+1} &= \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t), \\
\mathbf{z}_{t+1} &= \mathbf{z}_t - \frac{t+1}{2L}\nabla f(\mathbf{x}_t), \\
\mathbf{x}_{t+1} &= \frac{t+1}{t+3}\mathbf{y}_{t+1} + \frac{2}{t+3}\mathbf{z}_{t+1}.
\end{aligned}$$

Which is equivalent to Hinton's video, of course, just a bit messed about. L there is the smoothness parameter of f.

Alright so first We have $\mathbf{y}_{t+1}$ which is just the standard smooth step - the best step We could take, and it's in accordance with the result We got for

gradient descent with smooth functions. We're just noting it down, but this is not our new $\mathbf{x}_{t+1}$.

$\mathbf{z}_{t+1}$ is the momentum term. We have the previous momentum, and We are adding the gradient at the current point to it. Weird thing is is that We are not really attenuating the momentum here? It's not being multiplied by .99 or whatever.

Finally We have our new $\mathbf{x}_{t+1}$. It's a combination of $\mathbf{y}_{t+1}$ and $\mathbf{z}_{t+1}$. As time increases, We basically leave the $\mathbf{y}_{t+1}$ untouched. I guess this is where We attenuate $\mathbf{z}_{t+1}$, since We are scaling it by the reciprocal of t. So instead of keeping an attenuated version of momentum, We keep an unattenuated one and just scale it before adding it to previous coordinates.

6.3 Inconsistencies

Alright so in Hinton's version (that We like lol) We take a large step based on momentum, and then correct course with a gradient from that gambled point and update the momentum with the gradient.

In our version, We seemingly take the gradient at the very beginning, which I do not like.

Let's think of the gradient as the focal point, since We are only allowed to compute one at a time.

Alright so the algorithms are equivalent (surprise), I think. Our $\mathbf{x}_t$ is Hinton's $\mathbf{x}'$, the point We arrive at after taking the momentum step. We then take the gradient at this point, correct our momentum with it (yielding $\mathbf{z}_{t+1}$). $\mathbf{y}_{t+1}$ has to be our "corrected" step $\mathbf{x}''$ in Hinton's version.

Let's try the first step. We are at $\mathbf{x}_0$, from here We calculate a gradient and calculate what it would be like to take that step, which is our $\mathbf{y}_1$. Then We compute the slightly corrected momentum $\mathbf{z}_1$. Then We compute our next landing point by basically going to $\mathbf{y}_1$ and adding on to that our new momentum, which is the sum of old momentum plus new gradient.

My problem is that to me, the point of accelerated momentum is that We take a big momentum step, and then adjust our landing position with some actual information from the gradient. This is Hinton's video. In the notes' version however, it feels as if though We are taking the gradient and adding it to the momentum and taking a great big leap, which would not compensate for the risk We took in the leap to begin with.

The resolution is that the point We take the gradient at *is* the result of a momentum leap. We take a leap, calculate an adjustment, and take another leap. The adjustment calculated at time $t+1$ adjusts for the risk We took in leaping at time t.

This is hard to follow. Hinton's version is clear, so let's just keep that fixed in our minds. In the notes' version, We arrive at a point via a correction and leap from time t. We calculate a gradient correction $t+1$. This gradient correction $t+1$ is in fact adjusting the leap We took at time t.

7 Appendix

7.1 Cosine Rule

Cosine rule is just a generalization of $(a-b)^2 = a^2 + b^2 - 2ab$ in the case where a, b do not lie in the same plane. Letting them be vectors We get

$$\|\mathbf{a} - \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\mathbf{a}^\top \mathbf{b}$$

So just their individual lengths minus an interaction term twice. We can expand on that interaction term since recall $\mathbf{a}^\top \mathbf{b} = \|\mathbf{a}\|\|\mathbf{b}\|\cos(\theta)$ giving us

$$\|\mathbf{a} - \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\|\mathbf{a}\|\|\mathbf{b}\|\cos(\theta)$$

Makes this a bit more apparent - the interaction term is just their lengths scaled by their overlap. Basically this generalization is account for the case where there is an angle between the things We are multiplying - in the original case, everything lines up perfectly so We get $-2ab$, but in this new general case We scale the interaction down since the components are no longer parallel (and think about the perpendicular case - $\mathbf{a}, \mathbf{b}$ point in completely different directions so when You look at the length of $\mathbf{a} - \mathbf{b}$ You just get their individual lengths added. Ain't it neat?