The real lesson from NAG (Nesterov): Nesterov correction

deego

November 10, 2018

Initial date: <year>

Abstract

I am trying to understand Nesterov. It seems to me that the net effect of Nesterov is to replace, in the update rule, some of the mometum by gradient. If that is so, then (a) that doesn't seem like much innovation to me. (b) I am surprised it produces better results (does it?) (c) We can then Nestorov that and replace even more momentum by gradient, and repeat that ad nauseam till we are eventually left with classical momentum.

[since the Q then, so ignore:] So, Nesterov seems illogical yet it produces better results. We explain that, and identify the real lesson of Nesterov. Then, we (a) argue that it is miniscule and almost not worth it. (b) Identify how to apply it to arbitrary function, like CM, adam, etc.

1 Introduction:

We shall cast our equations slightly differently, in order to bring home the point we are making.

2 Vanilla gradient descent (SGD):

A gradient term is computed. And, η time the gradient is added to Θ. Momentum is same as the current gradient (in other words, no time-averaging is done).

$$\boxed{g_t = D\left[\Theta_{t-1}\right]} \tag{1}$$

$$\boxed{((m_t = g_t))} \tag{2}$$

$$\boxed{\Theta_t = \Theta_{t-1} - \eta g_t} \tag{3}$$

3 Classical momentum (CM):

Momentum is now an average of the preceding momentum and the current gradient:

$$\boxed{g_t = D\left[\Theta_{t-1}\right]} \tag{4}$$

$$\boxed{m_t = \beta m_{t-1} + \alpha g_t} \tag{5}$$

$$\boxed{\Theta_t = \Theta_{t-1} - \eta m_t} \tag{6}$$

where $\alpha \equiv (1 - \beta)$, is not to be confused with the α that Adam uses. m_t simply averages the

new gradient with a previous momentum.

Compare SGD to CM. One updates Θ by η times g, the other updates Θ by η times m. Imagine a hydrid where instead of updating by either g_t or by m_t, you update by an average of the two. We argue that that's exactly what NAG does. Furthermore, it seems to use that if you re-apply the same logic, you will replace even more m by g, till at the end you are left with 100% SGD. If this argument holds, we are a little confused as to the logic of NAG, and surprised that it's touted to perform better than CM.

4 NAG:

This is identical to CM except that the gradient is computed at a modified Theta. Before computing the gradient, we don't know m_t because we don't yet know $g_{t.}$. However, we know a portion of m_t. Wouldn't it be nice to use that information to look forward where the gradient will lie? What portion of m_t do we know? We know βm_{t-1}. Therefore, we compute the gradient at a Θ that is modified by this amount. Doesn't that seem logical? Absolutely! Let's proceed!

$$\boxed{g_t = D\left[\Theta_{t-1} - \eta\beta m_{t-1}\right]} \tag{7}$$

$$\boxed{m_t = \beta m_{t-1} + \alpha g_t} \tag{8}$$

$$\boxed{\Theta_t = \Theta_{t-1} - \eta m_t} \tag{9}$$

That is NAG.

4.1 Re-cast NAG using ϕ :

Now, we shall re-cast it in terms of a quantity

$$\phi_t \equiv \Theta_t - \eta\beta m_t \tag{10}$$

We note that (1) re-casting makes absolutely no difference to the long-term convergence or the the evolution equations, (2) We can always get back Θ from ϕ at any time. (3) In any case, after a long time, the difference between the two vanishes. What are the evolution equations for ϕ_t? The first two are straightforward. It is Eq. 9 that needs to be re-cast carefully. We note that

$$\Theta_t = \phi_t + \eta\beta m_t \tag{11}$$

and substitute:

$$\phi_t + \eta\beta m_t = \phi_{t-1} + \eta\beta m_{t-1} - \eta m_t \tag{12}$$

which yields:

$$\phi_t = \phi_{t-1} - \eta\beta m_t - \eta\left(m_t - \beta m_{t-1}\right) \tag{13}$$

which yields, using Eq. 8,

$$\phi_t = \phi_{t-1} - \eta\beta m_t - \eta\left(\alpha g_t\right) \tag{14}$$

Let's re-write NAG in terms of ϕ :

$$\boxed{g_t = D\left[\phi_{t-1}\right]} \tag{15}$$

$$\boxed{m_t = \beta m_{t-1} + \alpha g_t} \tag{16}$$

and finally,

$$\boxed{\phi_t = \phi_{t-1} - \eta\left(\beta m_t + \alpha g_t\right)} \tag{17}$$

When seen in terms of ϕ, NAG looks very much like the hybrid of SGD and CM that we talked about. We argued above that replacing Θ by ϕ doesn't change the theory. All that we have done is replace some of the momentum term by the gradient term.

If this logic really makes sense, nothing stops us from applying it yet again, and replacing even more momentum by gradient, till we find ourselves all the way to vanilla SGD.

It is, then, very surprising that NAG works better than CM.

<<**Freenode Q ends here**>>

5 Lessons of Nesterov: Nesterov correction: (freenode: ignore!).

If Nesterov really does work better than CM, then the real lesson is that tiny difference between Θ and ϕ. Nesterov, in essence, says, Use ϕ but report Θ. What does Θ do? It says: "Let's consider the very final correction you made to ϕ. You added both a previous momentum term, as well as a current gradient. The latter has info, but the former was merely added for long-term navigation of saddle points. It is otherwise free of information in comparison to g_t. Therefore, won't you please

remove that before reporting the final Θ? Thus, please add back the entire term $\eta\beta m_t$ before reporting the Θ." This was the logic as applied to Nesterov. The report Theta, $R_t \equiv \phi_t + \eta\beta m_t$. If we take this to its logical next step, we don't need to correct for the portion of m_t that comes from g_t. We ONLY need to correct for past momenta. That is, we first notice that

$$\phi_t = \phi_{t-1} - \eta\left(\beta^2 m_{t-1} + \alpha(1+\beta)g_t\right) \tag{18}$$

Of this, the entire g_t term is acceptable, and we only need to correct for the m_{t-1} term. Thus, if we were to apply Nesterov's own logic "properly" to Nesterov, we should report the following **Nesterov correction** for NAG:

$$R_t = \phi_t + \eta\beta^2 m_{t-1} \tag{19}$$

which, notably, is slightly different from Θ_t

Let us apply the very same logic to CM and see what we get for **Nesterov correction** for CM:

$$\boxed{R_t = \Theta_t + \eta\beta m_{t-1}} \tag{20}$$

Note that if, like NAG, we were to remove m_t instead of m_{t-1}, we would end up with $R_t = \Theta_t$. (Also, by now, we are into unspecified territory, really, seeing as m_t has g_t and m_{t-1} in it, and we could remove arbitrary protions of m_t and m_{t-1} and keep an arbitrary portion of g_t around.)

That, in our opinion, should be the Nesterov correction for classical momentum. With this logic and ansatz, one can construct Nesterov corrections for Adam and any other methods.

Most importantly, however, we note that this is not a long-term effect. It may improve very short-term convergence, but it does not change the equations at all. In the long term, the difference

between Θ_t and R_t vanishes completely, and long-term convergence should not really show any improvement due to Nesterove correction. Note that while R_t is now your best guess for performance, you still need to keep Θ around. It is Θ that contains the evolution information.

We also note that in order to apply Nesterov correction, you need to store the current momentum, the previous momentum, and the η, β, etc, and probably more terms when using Adam.

We next argue a bit more formally that the Nesterov correction, saves, at most, one time step. We note that for arbitrary momentum method:

$$R_t = \Theta_{t-1} + O(\eta\alpha g_t) \tag{21}$$

And, that

$$\Theta_{t-1} = R_{t-1} + O(\eta\beta m_{t-2}) \tag{22}$$

Thus, R_t is a very small correction to Θ_{t-1} based on the latest gradient. If you did not use the latest gradient, all you are doing is removing this very tiny final contribution from the very final gradient. So, in effect, you are sticking with the previous step. Or, to put it in other words, the current Θ_t is different from R_{t+1} by a very small number: $\alpha\eta g_{t+1}$. By using Θ_t instead of R_t (or R_{t+1}), all we are doing is skipping that one very tiny (weighted by $\alpha\eta$) correction from the next gradient. In some sense, Θ_t is a partial step in between R_t and R_{t+1}. It adds noisy terms to R_t , but on the other hand, its difference from R_{t+1} is at most like a one-step difference.

In conclusion, we argue that the NAG logic is ill-specified, and taken to its logical conclusion, will return to vanilla SGD. We figure out the real reason NAG appears to work better than CM. With this logic, we find out that NAG itself doesn't apply the same logic fully. We identified a more proper Nesterov correction that can be applied to any arbitrary method that uses momentum. We

finally argue that the effect should be miniscule, should probably save at most one step, and should have zero consequence for long-term convergence. (In reality, however, NAG is known to work far better? I'm confused. Some links say that NAG doesn't really do any better than SGD).