Bonus 3 - Adam with Nesterov Momentum

Murali Krishnan Rajasekharan Pillai

December 5, 2019

We aim to improve upon the Adam optimization algorithm with the Nesterov's accelerated gradient.[1] The standard Stochastic Gradient Algorithm is as follows:

Algorithm 1 Stochastic Gradient Descent

```
Require: $\alpha_0, \cdots, \alpha_T$: The learning rates for each time-step (presumably annealed)
Require: $f_i(\theta)$: Stochastic objective function parameterized by $\theta$ and indexed by time-step $i$
Require: $\theta_0$: The initial parameters
while $\theta_t$ not converged do
    $t \leftarrow t + 1$
    $\boldsymbol{g}_t \leftarrow \nabla_{\theta_{t-1}} f_t(\theta_{t-1})$
    $\theta_t \leftarrow \theta_{t-1} - \alpha_t \boldsymbol{g}_t$
end while
return $\theta_t$
```

Classical momentum accumulates a decaying sum (with a decay factor μ) of the previous updates into a momentum vector $\boldsymbol{m}$ and replaces the original gradient step in Algorithm 1 with that vector. That is we modify the algorithm to include the following at each time-step:

$$\boldsymbol{m}_t \leftarrow \mu \boldsymbol{m}_{t-1} + \alpha_t \boldsymbol{g}_t$$
$$\theta_t \leftarrow \theta_{t-1} - \boldsymbol{m}_t$$

This algorithm allows the algorithm to move faster along the dimensions of low curvature where the update is consistently small but in the same direction, and slower along turbulent dimensions where the direction of update is significantly oscillating.

Nesterov's accelerated gradient can be re-written as a kind of improved momentum. From above we can see that the update of the gradient with typical momentum update would be

$$\theta_t = \theta_{t-1} - (\mu \boldsymbol{m}_{t-1} + \alpha_t \boldsymbol{g}_t)$$

As we can observe above, the momentum step $\mu \boldsymbol{m}_{t-1}$ doesn't depend on the current gradient $\boldsymbol{g}_t$, so we can get a higher-quality gradient step direction by updating the parameters with the momentum step before computing the gradient. Hence the proposed changes would be reflected in the following way:

$$\boldsymbol{g}_t \leftarrow \nabla_{\theta_{t-1}} f_t(\theta_{t-1} - \mu \boldsymbol{m}_{t-1})$$
$$\boldsymbol{m}_t \leftarrow \mu \boldsymbol{m}_{t-1} + \alpha_t \boldsymbol{g}_t$$
$$\theta_t \leftarrow \theta_{t-1} - \boldsymbol{m}_t$$

While classical momentum and Nesterov's accelerated gradient define $\boldsymbol{m}$ as a decaying sum over the previous updates; however, Adaptive moment estimations (Adam) defines it instead as decaying mean over the previous gradients as follows:

$$\boldsymbol{m}_t \leftarrow \mu \boldsymbol{m}_{t-1} + (1 - \mu)\boldsymbol{g}_t$$
$$\theta_t \leftarrow \theta_{t-1} - \alpha_t \frac{\boldsymbol{m}_t}{1 - \mu_t}$$

Using the previous gradients instead of the previous updates allows the algorithm to continue changing the direction even when the learning rate has annealed significantly toward the end of training, resulting in more precise fine-grained convergence.

Modifying Adam's Momentum

Let as index μ by timestep as $\mu_1, \cdots, \mu_T$ in order to aid clarity. Before modifying Adam's update rule, let's re-write Nesterov's Accelerate Gradient (NAG) to be more straightforward to implement.

$$\begin{aligned} \boldsymbol{g}_t &\leftarrow \nabla_{\theta_{t-1}} f_t\left(\theta_{t-1}\right) \\ \boldsymbol{m}_t &\leftarrow \mu_t \boldsymbol{m_{t-1}} + \alpha_t \boldsymbol{g}_t \\ \theta_t &\leftarrow \theta_{t-1} - \left(\mu_{t+1} \boldsymbol{m}_t + \alpha_t \boldsymbol{g}_t\right) \end{aligned}$$

Using the same trick in Adam's momentum: let's rewrite Adam's update step in terms of $\boldsymbol{m}_{t-1}$ and $\boldsymbol{g}_t$ as:

$$\theta_t \leftarrow \theta_{t-1} \alpha_t \left(\frac{\mu_t \boldsymbol{m}_{t-1}}{1 - \prod_{i=1}^{t} \mu_i} + \frac{(1-\mu_t)\boldsymbol{g}_t}{1 - \prod_{i=1}^{t} \mu_i} \right)$$

The we substitute next momentum step for the current one, as dicsussed above:

$$\theta_t \leftarrow \theta_{t-1} \alpha_t \left(\frac{\mu_{t+1} \boldsymbol{m}_t}{1 - \prod_{i=1}^{t+1} \mu_i} + \frac{(1-\mu_t)\boldsymbol{g}_t}{1 - \prod_{i=1}^{t} \mu_i} \right)$$

When we change Adam this way, we get the Nesterov-Accelerated Adam, as follows:

Algorithm 2 Adam with Nesterov Momentum

```
Require: $\alpha_0, \cdots, \alpha_T; \mu_0, \cdots, \mu_T; \nu; \epsilon$: Hyper-parameters
$\boldsymbol{m}_0; \boldsymbol{n}_0 \leftarrow 0$ (first/second moment vectors)
while $\theta_t$ not converged do
    $\boldsymbol{g}_t \leftarrow \nabla_{\theta_{t-1}} f_t\left(\theta_{t-1}\right)$
    $\boldsymbol{m}_t \leftarrow \mu_t \boldsymbol{m}_{t-1} + (1-\mu_t)\boldsymbol{g}_t$
    $\boldsymbol{n}_t \leftarrow \nu \boldsymbol{m}_{t-1} + (1-\nu)\boldsymbol{g}_t^2$
    $\hat{\boldsymbol{m}} \leftarrow \left(\mu_{t+1} \frac{\boldsymbol{m}_t}{1-\prod_{i=1}^{t+1} \mu_i}\right) + \left(\frac{(1-\mu_t)\boldsymbol{g}_t}{1-\prod_{i=1}^{t} \mu_i}\right)$
    $\hat{\boldsymbol{n}} \leftarrow \frac{\nu \boldsymbol{n}_t}{1-\nu_t}$
    $\theta_t \leftarrow \theta_{t-1} - \frac{\alpha_t}{\sqrt{\hat{\boldsymbol{n}}_t} + \epsilon}$
end while
return $\theta_t$
```

References

[1] Timothy Dozat. Incorporating nesterov momentum into adam. 2016.