Learning rate warmup for Adam

Jacob Hilton

September 30, 2019

Learning rate warmup for SGD and its variants has a long history, but is not that well-understood. Here we explore one possible motivation for learning rate warmup in the context of Adam, which was originally identified by Liu et al. [2019]. We simplify their analysis and reduce their "Rectified Adam" (RAdam) algorithm to the following scheme with almost identical behaviour: **multiply the learning rate at step t by**

$$\sqrt{\frac{2}{1+\beta_2^t}-1},$$

where β_2 is the Adam second moment hyperparameter.

1 Background

The adjustments made by Adam [Kingma and Ba, 2014] act on each parameter dimension independently, so let's consider a one-dimensional parameter θ. The Adam update rule may be written as

$$\theta_{t+1} \leftarrow \theta_t + \underbrace{\frac{\alpha}{\sqrt{v_t}+\epsilon}}_{\substack{\text{adaptive} \\ \text{learning rate}}} m_t,$$

where $m_t := \text{ewma}_{\beta_1}(g)_t$, $v_t := \text{ewma}_{\beta_2}\left(g^2\right)_t$,[1] and g_t is the tth stochastic gradient. Here we have introduced the notation for an exponentially-weighted moving average

$$\text{ewma}_\beta(X)_t := \frac{X_t + \beta X_{t-1} + \beta^2 X_{t-2} + \cdots + \beta^{t-1} X_1}{1 \quad + \beta \quad + \beta^2 \quad + \cdots + \beta^{t-1}}.$$

Adam can be thought of as momentum SGD with an "adaptive learning rate", which we have labeled in the update rule above. The adaptive learning rate is a function of v_t, which is an estimate of the expected squared gradient. We would like the adaptive learning rate not to vary too wildly, which is why we usually take β_2 to be large (0.999 is the recommended default). (It's fine for m_t to vary wildly, as long as our steps are sufficiently small, since the noise will cancel out over multiple steps, but we will not be data-efficient if our adaptive learning rate varies too much, and we may be unstable if it is too large.)

However, towards the start of training, taking β_2 to be large does not save us from the fact that we have simply not yet seen many gradients, which means that our expected squared gradient estimator v_t may have high variance. As a result, RAdam takes the cautious approach of taking smaller steps towards the start of training to account for this. More precisely, under the assumption of i.i.d. Gaussian gradients, it is possible to analytically compute the effect of having not seen many gradients on the variance of the adaptive learning rate, and RAdam downscales the adaptive learning rate to cancel this effect out.

[1] Kingma and Ba [2014] and Liu et al. [2019] denote these expressions by $\hat{m}_t$ and $\hat{v}_t$ and refer to them as bias-corrected moment estimates.

2 The variance of an EWMA

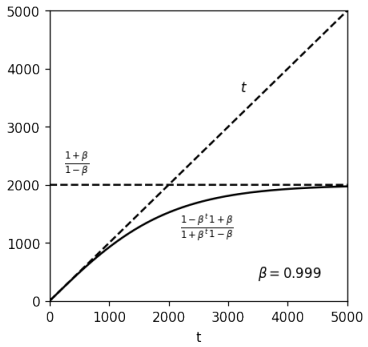
Thus learning rate warmup may be motivated by the effect of t on the variance of the adaptive learning rate, which is a function of $v_t = \text{ewma}_{\beta_2}\left(g^2\right)_t$. We would therefore like to analyze the variance of a generic EWMA. Assuming $X_1, \ldots, X_t$ are i.i.d. with variance σ^2, we may use the fact that $\text{Var}\left[aX\right] = a^2 \text{Var}\left[X\right]$ and that $\text{Var}\left[X + Y\right] = \text{Var}\left[X\right] + \text{Var}\left[Y\right]$ for independent X and Y to deduce that

$$\text{Var}\left[\text{ewma}_\beta\left(X\right)_t\right] = \frac{\sigma^2 + \beta^2\sigma^2 + \cdots + \beta^{2(t-1)}\sigma^2}{\left(1 + \beta + \cdots + \beta^{t-1}\right)^2} = \frac{\sigma^2}{\frac{1-\beta^t}{1+\beta^t}\frac{1+\beta}{1-\beta}}.$$

We refer to the denominator of the final expression here as the effective sample size[2] of the EWMA. This terminology is motivated by the fact that, in a similar fashion, the variance of the unweighted average of $X_1, \ldots, X_t$

$$\text{Var}\left[\frac{X_1 + \cdots + X_t}{t}\right] = \frac{\sigma^2 + \cdots + \sigma^2}{t^2} = \frac{\sigma^2}{t}.$$

Note that the effective sample size of the above EWMA is roughly t when t is small, and converges to $\frac{1+\beta}{1-\beta}$ as t tends to ∞:



So under the assumption of i.i.d. gradients,

$$\frac{\text{Var}\left[v_t\right]}{\text{Var}\left[v_\infty\right]} = \frac{1 + \beta_2^t}{1 - \beta_2^t},$$

where by an abuse of notation $\text{Var}\left[v_\infty\right] := \lim_{t\to\infty} \text{Var}\left[v_t\right]$. In other words, the effect of t being finite on the variance of v_t is to multiply it by $\frac{1+\beta_2^t}{1-\beta_2^t}$.

3 Learning rate warmup: RAdam

The RAdam update rule may be written as

$$\theta_{t+1} \leftarrow \theta_t + \underbrace{r_t \frac{\alpha}{\sqrt{v_t} + \epsilon}}_{\substack{\text{adaptive} \\ \text{learning rate}}} m_t,$$

where m_t and v_t are as above, and r_t is a function of t called the "variance rectification term". In other words, it is the Adam update rule with the adaptive learning rate multiplied by r_t, which has the effect of learning rate warmup. Occasionally r_t is undefined, in which case RAdam falls back to momentum SGD.

The variance rectification term r_t is chosen in such a way that the variance of the adaptive learning rate is constant, and $r_t \to 1$ as $t \to \infty$. Since the variance of the adaptive learning rate is $r_t^2 \alpha^2 \text{Var}\left[\frac{1}{\sqrt{v_t}+\epsilon}\right]$, this

[2] `https://en.wikipedia.org/wiki/Effective_sample_size`

is equivalent (approximating ϵ as 0) to taking

$$r_t = \sqrt{\frac{\operatorname{Var}\left[\frac{1}{\sqrt{v_\infty}}\right]}{\operatorname{Var}\left[\frac{1}{\sqrt{v_t}}\right]}}.$$

Under the assumption of i.i.d. Gaussian gradients, Liu et al. [2019] show that this is approximately

$$r_t = \begin{cases} \sqrt{\frac{(\rho_t-4)(\rho_t-2)\rho_\infty}{(\rho_\infty-4)(\rho_\infty-2)\rho_t}}, & \text{if } \rho_t > 4 \\ \text{undefined}, & \text{if } \rho_t \leq 4 \end{cases},$$

where

$$\rho_t = \rho_\infty - \frac{2t\beta_2^t}{1-\beta_2^t} \qquad \text{and} \qquad \rho_\infty = \frac{2}{1-\beta_2} - 1.$$

4 Simplifying RAdam

We may bypass much of the analysis of Liu et al. [2019] by simply using the approximation

$$\operatorname{Var}\left[\frac{1}{\sqrt{v_t}}\right] \approx c \operatorname{Var}\left[v_t\right],$$

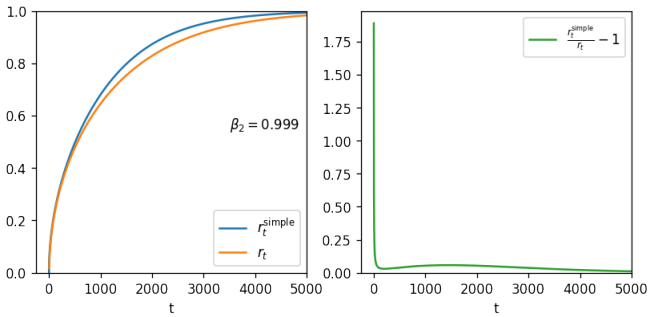
where c does not depend on t. This holds in the limit as the effective sample size of v_t tends to ∞ because of the delta method[3], and in practice it is a good approximation as long as t is not very small. Thus we may instead use the variance rectification term

$$r_t^{\text{simple}} = \sqrt{\frac{\operatorname{Var}\left[v_\infty\right]}{\operatorname{Var}\left[v_t\right]}} = \sqrt{\frac{1-\beta_2^t}{1+\beta_2^t}} = \sqrt{\frac{2}{1+\beta_2^t} - 1},$$

as originally claimed. Here we have assumed that the gradients are i.i.d., but did not need to assume that they are Gaussian (though the above approximation is less realistic than that of Gaussian gradients).

Note that r_t^{simple} is the square root of the ratio of the effective sample size of v_t to the effective sample size of v_∞ (i.e., the limit of the effective sample size of v_t as $t \to \infty$). In other words, it cancels out the effect of t being finite on the variance of v_t, in the sense that $\operatorname{Var}\left[r_t^{\text{simple}} v_t\right]$ is constant (under the assumption of i.i.d. gradients) and $r_t^{\text{simple}} \to 1$ as $t \to \infty$.

In practice, r_t and r_t^{simple} are very close:



Moreover, for typical values of β_2, r_t is only undefined (causing RAdam to fall back to SGD momentum) when $t \leq 4$. So the two versions of RAdam should have almost identical performance.

Between the two versions of RAdam, my personal recommendation is to use the simplified version. The original version uses more accurate approximations, but the simplified version uses a more understandable

[3] `https://en.wikipedia.org/wiki/Delta_method`

function, has a more natural interpretation in terms of effective sample sizes, and does not need to fall back to momentum SGD. Falling back to momentum SGD makes the algorithm more complex, and it would make more theoretical sense to me to instead take $r_t = 0$.

When I asked Liu about my simplification, they agreed with my analysis, but disagreed with my recommendation, saying (notation adapted to context):

> Comparing these two approximations, I feel our approximation is better as it is more accurate. The main difference is that we are trying to estimate $\text{Var}\left[\sqrt{\frac{1}{v_t}}\right]$ as a whole, instead of estimating $\text{Var}\left[\sqrt{v_t}\right]$. The adaptive learning rate is the inverse of $\sqrt{v_t}$, whose variance (i.e. $\text{Var}\left[\sqrt{\frac{1}{v_t}}\right]$) is hard to estimate from $\text{Var}\left[\sqrt{v_t}\right]$. At the same time, if $g_i \sim \mathcal{N}(0,1)$, then $\text{Var}\left[\frac{1}{\sqrt{g_0^2+g_1^2}}\right]$ is divergent, while $\text{Var}\left[\sqrt{g_0^2+g_1^2}\right]$ is not. Since the adaptive learning rate is designed as the inverse, I think it's better to use the current approximation. Meanwhile, I agree with you that your approximation is more simple, and in most cases, I feel these two approximations should result in a similar empirical performance.

5 Discussion

We have explored a potential problem with Adam, whereby the variance of the adaptive learning rate is too high towards the start of training due to a lack of gradient samples. We have seen that this can be resolved using learning rate warmup, by multiplying the learning rate at step t by

$$\sqrt{\frac{2}{1+\beta_2^t} - 1},$$

where β_2 is the Adam second moment hyperparameter.

OpenAI et al. [2019] used a learning rate of 0 for the first several hours of training after surgery, allowing the adaptive learning rate to settle down before actually changing the network. This provides some evidence that the variance of adaptive learning rate is genuinely a problem in practice.

None of this analysis precludes the possibility that there are other reasons to use learning rate warmup. Indeed, given that learning rate warmup is older than Adam, it seems highly likely that there are other reasons. So this scheme is probably best thought of simply as a way to warm up the learning rate gradually over the course of around the first $\frac{5}{-\ln(\beta_2)} \approx \frac{5}{1-\beta_2}$ steps (the point at which r_t^{simple} first exceeds around 0.99). This is justified by the analysis presented, but may not be justified by other considerations.

Note that this scheme couples the Adam hyperparameter β_2 to the speed of learning rate warmup, so extra care should be taken under this scheme when adjusting β_2.

References

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

OpenAI, C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dbiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. de Oliveira Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, and S. Zhang. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.