Mila Optim Crash Course - Adam and RMSProp

Charles Guille-Escuret

March 2024

1 History

$$\texttt{GD:}\ w_{t+1} = w_t - \eta \nabla f(w_t)$$

Problem: difference of curvature in badly conditioned quadratics.

$$\texttt{GD momentum:}\quad w_{t+1} = w_t - \eta \nabla f(w_t) + \beta(w_t - w_{t-1})$$

$$\texttt{equivalently:}\quad w_{t+1} = w_t - \eta(\sum_{i=0}^{t+1} \beta^i w_{t-i})$$

Idea: using a different learning rate per coordinate, adjusted by the scale of the gradients in that coordinate.

$$\texttt{AdaGrad (Duchi et al., 2011):}\quad w_{t+1} = w_t - \eta G_t^{-1/2} \nabla f(w_t)$$

$$G_t = \sum_{i=0}^{t} g_i g_i^\top$$

Too expensive, use diagonal instead

$$w_{t+1} = w_t - \eta \times diag(G_t + \epsilon Id)^{-1/2} \nabla f(w_t)$$

if we note $v_t := diag(G_t)$, with elementwise division:

$$w_{t+1} = w_t - \eta \frac{\nabla f(w_t)}{\sqrt{v_t + \epsilon}}$$

Problem: G_t goes to infinity with time, learning rate shrink.

•**RMSProp, 2012** (Root Mean Square Propagation) introduced by Geoff Hinton as a Coursera lecture (lecture 6 on neural networks).

•Never published, but over 7500 citations !

•Lecture was given by Geoff Hinton, but RMSProp was invented by his student, Tijmen Tieleman. History did not give the right credits.

`RMSProp (T. Tieleman, G. Hinton, 2012):` $w_{t+1} = w_t - \eta \frac{\nabla f(w_t)}{\sqrt{v_t + \epsilon}}$

with

$$v_{t+1} = \beta_2 v_t + (1 - \beta_2)[\nabla f(w_t)]^2$$

•Denominator identical to Adam !

Momentum works, RMSProp works, why not combine them ?

•Hinton writes: "Momentum does not help as much as it normally does. Needs more investigation.", and "It works best if the RMS of the recent gradients is used to divide the correction rather than the jump in the direction of accumulated corrections".

•Hinton's momentum was by adding $\beta(w_t - w_{t-1})$. This is not equivalent anymore to $\sum_{i=0}^{t+1} \beta^i w_{t-i}$! Adam adds momentum similarly to the later, and introduces bias correction:

(Adaptive Moment Estimation)

`Adam (D.P Kingma, J. Ba, 2014):` $w_{t+1} = w_t - \eta \frac{\bar{m}_t}{\sqrt{\bar{v}_t} + \epsilon}$

$$m_{t+1} = \beta_1 m_t + (1 - \beta_1)\nabla f(w_{t+1})$$
$$v_{t+1} = \beta_2 v_t + (1 - \beta_2)[\nabla f(w_{t+1})]^2$$
$$\bar{m}_t = \frac{m_t}{1 - \beta_1^t}$$
$$\bar{v}_t = \frac{v_t}{1 - \beta_2^t}$$

•momentum of non-rescaled direction, instead of rescaled updates.

2 Bias correction and momentum expression

Contrary to what many have said, no difference between doing

$$m_{t+1} = \beta_1 m_t + (1 - \beta_1)\nabla f$$

$$m_{t+1} = \beta_1 m_t + \nabla f$$

besides a rescaling of η by $1 - \beta_1$.

•m_t and v_t are biased towards 0. Bias correction fixes that and means we don't have to "pick up speed".

•The direction does not change ! Only magnitude. With $\beta_1 = 0.9$ and $\beta_2 = 0.999$, at step 1, we multiply by $\frac{0.032}{0.1} = 0.32$ the step size at init. Then reaches a minimum of 0.152 at the 12th step, and increases back to 0.3 at 100 steps and 0.8 at 1000 steps.

3 Adam's theoretical guarantees

Not going to go over proof !

Was found to use unreasonable assumptions AND to be wrong.

Credit to David Martínez Rubio MSc thesis (2017).

•Assumes distance between all iterates is bounded.

•Implicitly assumes $\frac{\sqrt{v_{t,i}}}{\alpha_t}$ increases with t.

•Implicitly assumes $|\nabla f_t(x_t)_i| \leq 1$

•After a telescoping sum, obtain $\sum_{j=0}^{T-t} t\gamma^j$, should be $\sum_{j=0}^{T-t} \sqrt{t+j}\gamma^j$. Then they bound $\sum_{j=0}^{T-t} j\gamma^j$ by mistake, and their resulting last step is proven to not be true in general.

•Multiple papers claiming to prove convergence without changing the update step.

4 Intuition of why second moments works

Questions:

•is the update in the span of gradients ?

→ Nesterov's lower bound doesnt apply !

•is the optimizer rotation invariant ?

→ Adam only works in the standard base, which is not very well captured today.

4.1 Link to Newton

Assume quadratic with diagonal hessian Q, $f(x) = x^\top Qx$. Sample x_t from a noise ϵ with anisotropic variance σ^2. We have $\nabla f(x_t)_i = Q_{ii}\epsilon_i$. Thus, $v_t = Q_{ii}^2 \times EMA(\epsilon_i^2)$. v_t ends up approximating $Q_{ii}^2 \times \sigma^2$.

The update of RMSProp becomes

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{Q_{ii}^2\sigma^2}} m_t = w_t - \frac{\eta}{\sigma}(\nabla^2 f(w_t))^{-1}\nabla f(w_t)$$

This is Newton's method !

4.2 Link to noise in the gradient

In the above expression, if each coordinate has its own std σ_i, then the learning rate is rescaled by σ_i^{-1}. Variance reduction !

•simple example with two functions, one coordinate identical, the other coordinate with one flat and one highly curved function.

4.3 Link to feature imbalance

•Need to scale the update based on how frequent a feature is, thus why it works well on NLP and not on images.

4.4 Link to Transformer architecture

•Different blocks means different behavior, means different hyperparameters. Having LR adaptive by weight is useful.

5 Connection to other methods

SignGD : with $\beta_1 = 1$ and $\beta_2 = 1$, recover SignGD. Explains its stabilization properties.

NAdam : Compute m_t and v_t identically, but update in the average of m_t and $\nabla f(w_t)$, similarly to Nesterov's trick

Amsgrad : $\hat{v}_t = \max_s(v_s)$. Doesn't work so well. Based on a mistake in Adam's original paper, better theoretical guarantees.