Adam: A Method for Stochastic Optimization

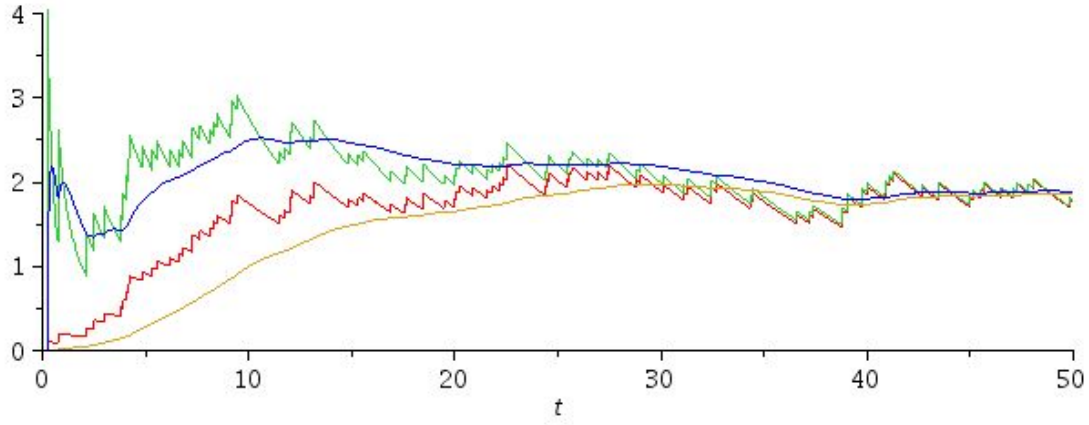
Authors: Diederik P. Kingma, Jimmy Lei Ba

CONS: Akshay Kamath

MAIN CONTRIBUTIONS:

1) Performs better than RMSProp and AdaGrad
2) Experiments indicate that Adam performs well
3) Initialization Bias is not really important
4) Proofs are badly written
5) Convergence Proofs are wrong

When does initialization bias matter?



$$m_t = \beta m_{t-1} + (1 - \beta)\theta_t$$

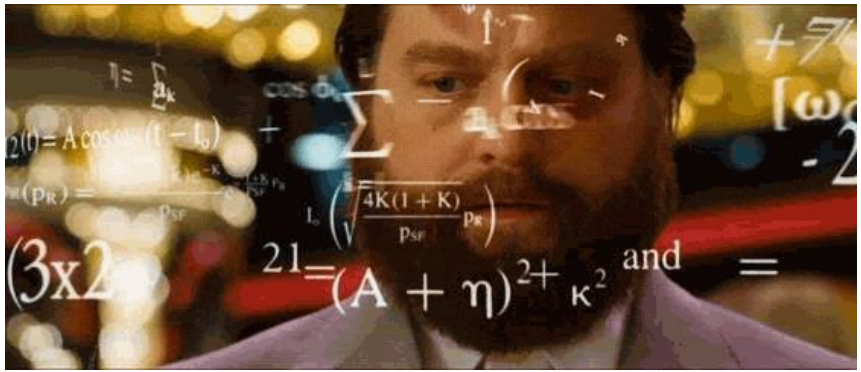
- Biased when β is large
- Only for the first few steps
- Sparse gradients occur much later during optimization

When does initialization bias matter?

```
for p, g, m, v, vhat in zip(params, grads, ms, vs, vhats):
    m_t = (self.beta_1 * m) + (1. - self.beta_1) * g
    v_t = (self.beta_2 * v) + (1. - self.beta_2) * K.square(g)
    if self.amsgrad:
        vhat_t = K.maximum(vhat, v_t)
        p_t = p - lr_t * m_t / (K.sqrt(vhat_t) + self.epsilon)
        self.updates.append(K.update(vhat, vhat_t))
    else:
        p_t = p - lr_t * m_t / (K.sqrt(v_t) + self.epsilon)

    self.updates.append(K.update(m, m_t))
    self.updates.append(K.update(v, v_t))
    new_p = p_t

    # Apply constraints.
    if getattr(p, 'constraint', None) is not None:
        new_p = p.constraint(new_p)

    self.updates.append(K.update(p, new_p))
return self.updates
```

KERAS IMPLEMENTATION DOES NOT CORRECT BIAS!

Primer on how to write a bad proof



1) Make the first arxiv post without a proof.
2) Pray that no one asks for it. Someone will.
3) Release a new version with a proof in the appendix.
4) Provide very little intuition about the proof technique.
5) Pray more.

A few errors(in case you missed them)

BUG IN LEMMA 10.4

For $\gamma < 1$, using the upper bound on the arithmetic-geometric series, $\sum_t t\gamma^t < \frac{1}{(1-\gamma)^2}$:

$$\sum_{t=1}^{T} \frac{\|g_{t,i}\|_2}{\sqrt{t(1-\beta_2)}} \sum_{j=0}^{T} t\gamma^j \leq \frac{1}{(1-\gamma)^2\sqrt{1-\beta_2}} \sum_{t=1}^{T} \frac{\|g_{t,i}\|_2}{\sqrt{t}}$$

UP NEXT: BUG IN THEOREM 10.5

3 THE NON-CONVERGENCE OF ADAM

With the problem setup in the previous section, we discuss fundamental flaw in the current exponential moving average methods like ADAM. We show that ADAM can fail to converge to an optimal solution even in simple one-dimensional convex settings. These examples of non-convergence contradict the claim of convergence in (Kingma & Ba, 2015), and the main issue lies in the following quantity of interest:

$$\Gamma_{t+1} = \left(\frac{\sqrt{V_{t+1}}}{\alpha_{t+1}} - \frac{\sqrt{V_t}}{\alpha_t} \right). \tag{2}$$

Thanks