Adam

A Preprint

Authors propose Adam, a method for efficient stochastic optimization that only requires first-order gradients with little memory requirement. The method computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients; the name Adam is derived from adaptive moment estimation. The main idea is combination of AdaGrad (which handles sparse gradients) and RMSProp which works well in on-line and non-stationary settings.

Algorithm 1 Adam: Stochastic Optimization

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for moment estimates
Require: $f(\theta)$: Stochastic objective function
Require: θ_0: Initial parameters

$m_0 \leftarrow 0$ ▷ Initialize 1st moment vector
$v_0 \leftarrow 0$ ▷ Initialize 2nd moment vector
$t \leftarrow 0$ ▷ Initialize timestep
while θ_t not converged do
    $t \leftarrow t + 1$
    $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ ▷ Compute gradient
    $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ ▷ Update biased 1st moment, it calculates an exponentially moving average of gradients
    $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ ▷ Update biased 2nd moment, exponentially moving average of element-wise squared gradients
    $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ ▷ Bias-corrected 1st moment
    $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ ▷ Bias-corrected 2nd moment
    $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \varepsilon)$ ▷ Update parameters
end while

Moving moments are the uncentered variance of the gradient. However, these moving averages are initialized as (vectors of) 0's, leading to moment estimates that are biased towards zero, especially during the initial timesteps, and especially when the decay rates are small thats why Bias correction is made.

In Stochastic Gradient Descent (SGD), we don't calculate the true gradient because it's too computationally expensive. Instead, we use a small batch of data (a minibatch) to get a stochastic estimate of the gradient. This estimate is noisy—it points in roughly the right direction, but it fluctuates and jitters with every batch.A moving average acts as a low-pass filter. It smooths out the high-frequency jitters (the noise from individual minibatches) to reveal the underlying, more stable trend (the "true"gradient direction). This leads to a smoother, more direct path towards the minimum. Briefly saying momentum is setting direction, controls velocity and currect gradient is setting corrections.

Effective Step Size:

$$\Delta_t = \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t}}$$

- Upper bound (assuming $\varepsilon = 0$):

$$|\Delta_t| \leq \begin{cases} \alpha \cdot \frac{1-\beta_1}{\sqrt{1-\beta_2}}, & \text{if } 1 - \beta_1 > \sqrt{1 - \beta_2} \\ \alpha, & \text{otherwise} \end{cases}$$

- In sparse gradients: effective steps may reach upper bound.
- In denser updates: step size reduces.

When $1-\beta_1 = \sqrt{1-\beta_2}$:

$$\left|\frac{\hat{m}_t}{\sqrt{\hat{v}_t}}\right| < 1 \quad \Rightarrow \quad |\Delta_t| < \alpha$$

Typical behavior:

$$\left|\frac{\mathbb{E}[g]}{\sqrt{\mathbb{E}[g^2]}}\right| \leq 1 \Rightarrow \left|\frac{\hat{m}_t}{\sqrt{\hat{v}_t}}\right| \approx 1$$

Interpretation:

[leftmargin=1.5em]Δ_t is roughly bounded by $\alpha \rightarrow$ defines a trust region. Facilitates prior choice of α. Adam scales steps relative to gradient confidence.

Signal-to-Noise Ratio (SNR):

$$\text{SNR}_t := \frac{\hat{m}_t}{\sqrt{\hat{v}_t}}$$

- Smaller SNR $\Rightarrow$ smaller steps.
- SNR typically $\rightarrow 0$ near optimum (automatic annealing).

Scale-Invariance:

$$g_t \mapsto cg_t \Rightarrow \frac{c\hat{m}_t}{\sqrt{c^2\hat{v}_t}} = \frac{\hat{m}_t}{\sqrt{\hat{v}_t}}$$

Bias Correction for Second Moment Estimate

Update rule:

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

Assuming $v_0 = 0$, this expands to:

$$v_t = (1-\beta_2)\sum_{i=1}^{t} \beta_2^{t-i} g_i^2 \tag{1}$$

Expected value:

$$\mathbb{E}[v_t] = \mathbb{E}[g_t^2] \cdot (1-\beta_2^t) + \zeta \tag{2}$$

where ζ is small for slowly changing gradient statistics.

Bias-corrected estimate:

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

Note: same correction is used for first moment m_t with β_1.

If a function is convex then $\forall x, y \in R^d$:

$$f(y) \geq f(x) + \nabla f(x)^T (y-x)$$

A function $f : R^d$ßR is convex if $\forall x, y \in R^d, \forall \lambda$

$$\lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda) y)$$