On the convergence of the ADAM algorithm

DSAIDIS optimization and neural networks workshop
15 November 2022

Olivier Fercoq

TELECOM Paris
IP PARIS

Optimization problem

$$\min_{x \in \mathbb{R}^p} \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)]$$

- $F(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)]$
- We look for $x^* \in \arg\min_{x \in \mathbb{R}^p} F(x)$, or at least a local minimizer
- Distribution $\mathcal{D}$ is unknown but we can generate samples

Example: Multilayer perceptron on MNIST

- $\mathcal{D}$ is the uniform distribution over $N = 10,000$ images of digits, together with their label
- For sample $\xi = (I, y)$ and model parameters θ, $f(\theta, (I, y)) = \ell(\mathrm{MLP}(I, \theta), y)$
  - ▷ ℓ is the categorical cross entropy $\ell(\hat{y}, y) = -\sum_{d=0}^{9} \mathbb{1}_{y_d=1} \log(\hat{y}_d)$
  - ▷ MLP with 2 layers of 32 neurons with relu activation followed by a 10-neuron layer with soft-max activation $\rightarrow p = 26,506$ parameters
- Objective: $\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{N} f(\theta, \xi_i)$

Stochastic gradient I

Setup

$\min_{x \in \mathbb{R}^p} \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)]$

Idea

For $x \in \mathbb{R}^p$, $\xi \sim \mathcal{D}$, $\nabla f(x, \xi)$ is an unbiased estimator of $\nabla F(x)$

Algorithm

$x_0 \in \mathbb{R}^p$
For $k \geq 0$:
 $\xi_{k+1} \sim \mathcal{D}$
 $x_{k+1} = x_k - \alpha_k \nabla f(x_k, \xi_{k+1})$

Stochastic gradient II

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k, \xi_{k+1})$$

Convergence speed

If F is convex with bounded stochastic gradients & $\alpha_k = \frac{a}{\sqrt{k+b}}$:

$$\mathbb{E}\left[F(\bar{x}_k^\alpha) - F(x^*)\right] \leq \frac{\mathbb{E}[\|x_0 - x^*\|^2] + G\sum_{l=0}^k \alpha_l^2}{2\sum_{l=0}^k \alpha_l} \in O\left(\frac{\ln(k)}{\sqrt{k}}\right)$$

Advantages

- 1 sample per iteration
- Good result with only one pass over the data
- Speed of convergence independent from N

Drawbacks

- Limited precision for $k > N$
- The choice of the sequence α_k is problem-dependent

Presentation of ADAM

ADAM: stochastic gradient with adaptive moment estimation [Kingma and Ba, 2015] – 124,000 citations

$x_0^p \in \mathbb{R}^p, m_0 = 0 \in \mathbb{R}^p, v_0 = \hat{v}_0 = 0 \in \mathbb{R}_+^p$

$$m_{k+1} = \beta_1 m_k + (1 - \beta_1)\nabla f(x_k, \xi_{k+1})$$

$$\hat{m}_{k+1} = \frac{m_{k+1}}{1 - \beta_1^{k+1}}$$

$$v_{k+1} = \beta_2 v_k + (1 - \beta_2)\nabla f(x_k, \xi_{k+1})^2$$

$$\hat{v}_{k+1} = \max\left(\hat{v}_k, \frac{v_{k+1}}{1 - \beta_2^{k+1}}\right)$$

$$x_{k+1} = x_k - \underbrace{\frac{\alpha_k}{\epsilon + \sqrt{\hat{v}_{k+1}}}}_{\text{adaptive learning rate}} \hat{m}_{k+1}$$

Presentation of ADAM

ADAM: stochastic gradient with adaptive moment estimation [Kingma and Ba, 2015] – 124,000 citations

$x_0^p \in \mathbb{R}^p$, $m_0 = 0 \in \mathbb{R}^p$, $v_0 = \hat{v}_0 = 0 \in \mathbb{R}_+^p$

$$
\begin{aligned}
m_{k+1} &= \beta_1 m_k + (1-\beta_1)\nabla f(x_k, \xi_{k+1}) \longleftarrow \text{estimate of 1st moment} \\
\hat{m}_{k+1} &= \frac{m_{k+1}}{1-\beta_1^{k+1}} \\
v_{k+1} &= \beta_2 v_k + (1-\beta_2)\nabla f(x_k, \xi_{k+1})^2 \\
\hat{v}_{k+1} &= \max\left(\hat{v}_k, \frac{v_{k+1}}{1-\beta_2^{k+1}}\right) \\
x_{k+1} &= x_k - \underbrace{\frac{\alpha_k}{\epsilon + \sqrt{\hat{v}_{k+1}}}}_{\substack{\text{adaptive} \\ \text{learning rate}}} \hat{m}_{k+1}
\end{aligned}
$$

Presentation of ADAM

ADAM: stochastic gradient with adaptive moment estimation
[Kingma and Ba, 2015] – 124,000 citations
$x_0^p \in \mathbb{R}^p$, $m_0 = 0 \in \mathbb{R}^p$, $v_0 = \hat{v}_0 = 0 \in \mathbb{R}_+^p$

$$
\begin{aligned}
m_{k+1} &= \beta_1 m_k + (1-\beta_1)\nabla f(x_k, \xi_{k+1}) \longleftarrow \text{estimate of 1st moment} \\
\hat{m}_{k+1} &= \frac{m_{k+1}}{1-\beta_1^{k+1}} \\
v_{k+1} &= \beta_2 v_k + (1-\beta_2)\nabla f(x_k, \xi_{k+1})^2 \longleftarrow \text{estimate of 2nd moment} \\
\hat{v}_{k+1} &= \max\left(\hat{v}_k, \frac{v_{k+1}}{1-\beta_2^{k+1}}\right) \\
x_{k+1} &= x_k - \underbrace{\frac{\alpha_k}{\epsilon + \sqrt{\hat{v}_{k+1}}}}_{\substack{\text{adaptive} \\ \text{learning rate}}} \hat{m}_{k+1}
\end{aligned}
$$

Presentation of ADAM

ADAM: stochastic gradient with adaptive moment estimation
[Kingma and Ba, 2015] – 124,000 citations
$x_0^p \in \mathbb{R}^p, m_0 = 0 \in \mathbb{R}^p, v_0 = \hat{v}_0 = 0 \in \mathbb{R}^p_+$

$$m_{k+1} = \beta_1 m_k + (1-\beta_1)\nabla f(x_k, \xi_{k+1}) \longleftarrow \text{estimate of 1st moment}$$

$$\hat{m}_{k+1} = \frac{m_{k+1}}{1-\beta_1^{k+1}} \longleftarrow \text{bias correction step}$$

$$v_{k+1} = \beta_2 v_k + (1-\beta_2)\nabla f(x_k, \xi_{k+1})^2 \longleftarrow \text{estimate of 2nd moment}$$

$$\hat{v}_{k+1} = \max\left(\hat{v}_k, \frac{v_{k+1}}{1-\beta_2^{k+1}}\right)$$

$$x_{k+1} = x_k - \frac{\alpha_k}{\underbrace{\epsilon + \sqrt{\hat{v}_{k+1}}}_{\substack{\text{adaptive}\\ \text{learning rate}}}} \hat{m}_{k+1}$$

Presentation of ADAM

ADAM: stochastic gradient with adaptive moment estimation
[Kingma and Ba, 2015] – 124,000 citations
$x_0^p \in \mathbb{R}^p$, $m_0 = 0 \in \mathbb{R}^p$, $v_0 = \hat{v}_0 = 0 \in \mathbb{R}^p_+$

$$m_{k+1} = \beta_1 m_k + (1-\beta_1)\nabla f(x_k, \xi_{k+1}) \longleftarrow \text{estimate of 1st moment}$$

$$\hat{m}_{k+1} = \frac{m_{k+1}}{1-\beta_1^{k+1}} \longleftarrow \text{bias correction step}$$

$$v_{k+1} = \beta_2 v_k + (1-\beta_2)\nabla f(x_k, \xi_{k+1})^2 \longleftarrow \text{estimate of 2nd moment}$$

$$\hat{v}_{k+1} = \max\left(\hat{v}_k, \frac{v_{k+1}}{1-\beta_2^{k+1}}\right) \longleftarrow \text{technical condition}$$

$$x_{k+1} = x_k - \underbrace{\frac{\alpha_k}{\epsilon + \sqrt{\hat{v}_{k+1}}}}_{\substack{\text{adaptive}\\ \text{learning rate}}} \hat{m}_{k+1}$$

Presentation of ADAM

ADAM: stochastic gradient with adaptive moment estimation
[Kingma and Ba, 2015] – 124,000 citations
$x_0^p \in \mathbb{R}^p$, $m_0 = 0 \in \mathbb{R}^p$, $v_0 = \hat{v}_0 = 0 \in \mathbb{R}_+^p$

$$m_{k+1} = \beta_1 m_k + (1-\beta_1)\nabla f(x_k, \xi_{k+1}) \longleftarrow \text{estimate of 1st moment}$$

$$\hat{m}_{k+1} = \frac{m_{k+1}}{1-\beta_1^{k+1}} \longleftarrow \text{bias correction step}$$

$$v_{k+1} = \beta_2 v_k + (1-\beta_2)\nabla f(x_k, \xi_{k+1})^2 \longleftarrow \text{estimate of 2nd moment}$$

$$\hat{v}_{k+1} = \max\left(\hat{v}_k, \frac{v_{k+1}}{1-\beta_2^{k+1}}\right) \longleftarrow \text{technical condition}$$

$$x_{k+1} = x_k - \underbrace{\frac{\alpha_k}{\epsilon + \sqrt{\hat{v}_{k+1}}}}_{\text{adaptive learning rate}} \hat{m}_{k+1} \longleftarrow \frac{a}{b}\text{: elementwise division}$$

A few insights

- $\dfrac{\alpha_k}{\epsilon + \sqrt{\hat{v}_{k+1}}}$

  Adaptive step size that depends:

  - on the amplitude of the objective function
  - and on the noise in the stochastic gradients

- α_k has no unit

  Easy to tune independently on the problem

- m_{k+1} vs $\nabla f(x_k, \xi_{k+1})$

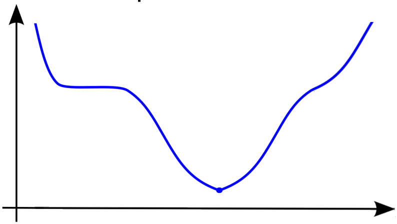
  Less variance but $\mathbb{E}[m_{k+1} \mid x_k] \neq \nabla F(x_k)$

- $\hat{v}_k$ is a vector

  Coordinate-dependent learning rate

What if there is no noise?

- We choose $\beta_1 = \beta_2 = 0$ and we have $\nabla f(x, \xi) = \nabla F(x)$
- $x_{k+1} = x_k - \alpha_k \frac{\nabla F(x_k)}{\epsilon + |\nabla F(x_k)|}$
- Far from the optimum, the norm of the gradient does not influence the algorithm
- Good for quasi-convex landscapes



Moment or momentum?

- Suppose $\beta_2 = 0$ and $\nabla f(x, \xi) = \nabla F(x)$ (no noise)
- $m_{k+1} = \beta_1 m_k + (1 - \beta_1)\nabla F(x_k)$
- $x_{k+1} = x_k - \alpha/\epsilon m_{k+1} = x_k - \alpha/\epsilon(1-\beta_1)\nabla F(x_k) - \alpha/\epsilon\beta_1 m_k$
  $= x_k - \alpha/\epsilon(1 - \beta_1)\nabla F(x_k) - \beta_1(x_{k-1} - x_k)$
- We recognize the heavy ball method

Convergence theorem

Suppose that

- $f(\cdot, \xi)$ is convex for all ξ (local behaviour)
- $\exists x^* \in \arg\min F$, $F(x) = \mathbb{E}[f(x, \xi)]$
- For all k, for all i, $|x_{k,i} - x^*_i| \leq D$
- For all x, ξ, for all i, $|\nabla_i f(x, \xi)| \leq G$
- $\alpha_k = \frac{\alpha_0}{\sqrt{k+1}}$
- $\beta_1^2 < \beta_2 < 1$

Then the iterates of Adam satisfy

$$\mathbb{E}[F(\bar{x}_K) - F(x^*)]$$
$$\leq \frac{dD^2}{2(1-\beta_1)} \frac{\sqrt{1-\beta_2}\,G}{\alpha_0(\sqrt{K}+K)} + \frac{1+2\beta_1}{2(1-\beta_1)} \frac{\alpha_0 \sqrt{1+\ln(K)}\,G}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}\sqrt{K}}$$

where $\bar{x}_K = \frac{1}{K}\sum_{k=0}^{K-1} x_k$

Sketch of proof 1

We will denote $\hat{\gamma}_{k+1} = \frac{\alpha_k}{(1-\beta_1^{k+1})(\epsilon+\sqrt{\hat{v}_{k+1}})}$ so that $x_{k+1} = x_k - \hat{\gamma}_{k+1} m_{k+1}$.

- $f(x_k, \xi_{k+1}) - f(x^*, \xi_{k+1}) \leq \langle \nabla f(x_k, \xi_{k+1}), x_k - x^* \rangle$
- Using the relation $m_{k+1} = \beta_1 m_k + (1-\beta_1)\nabla f(x_k, \xi_{k+1})$, we get

$$\langle \nabla f(x_k, \xi_{k+1}), x_k - x^* \rangle = \langle m_{k+1}, x_k - x^* \rangle + \frac{\beta_1}{1-\beta_1}\Big(\langle m_{k+1}, x_{k+1} - x^* \rangle - \langle m_k, x_k - x^* \rangle\Big) + \frac{\beta_1}{1-\beta_1}\|m_{k+1}\|^2_{\hat{\gamma}_{k+1}}$$

- We make appear nearly telescoping terms in the first term

$$\langle m_{k+1}, x_k - x^* \rangle = \frac{1}{2}\|x_k - x^*\|^2_{\hat{\gamma}_{k+1}^{-1}} - \frac{1}{2}\|x_{k+1} - x^*\|^2_{\hat{\gamma}_{k+1}^{-1}} + \frac{1}{2}\|m_{k+1}\|^2_{\hat{\gamma}_{k+1}}$$

Sketch of proof 2

- We can now sum

$$
\begin{aligned}
&\sum_{k=0}^{K-1} f(x_k, \xi_{k+1}) - f(x^*, \xi_{k+1}) \\
&\qquad \le \frac{\beta_1}{1-\beta_1}\Big(\langle m_K, x_K - x^* \rangle - \langle m_0, x_0 - x^* \rangle \Big) \\
&\qquad\qquad + \sum_{k=0}^{K-1} \left(\frac{1}{2}\|x_k - x^*\|^2_{\hat{\gamma}_{k+1}^{-1}} - \frac{1}{2}\|x_{k+1} - x^*\|^2_{\hat{\gamma}_{k+1}^{-1}} \right) \\
&\qquad\qquad + \left(\frac{\beta_1}{1-\beta_1} + \frac{1}{2} \right) \sum_{k=0}^{K-1} \|m_{k+1}\|^2_{\hat{\gamma}_{k+1}}
\end{aligned}
$$

- The main term is $\sum_{k=0}^{K-1} \|m_{k+1}\|^2_{\hat{\gamma}_{k+1}}$

Sketch of proof 3

The step size almost surely compensates the error term

- Denote $g_{k+1} = \nabla_i f(x_k, \xi_{k+1})$ and $\gamma_{k+1} = \frac{\alpha_k}{(1-\beta_1)\sqrt{v_{k+1}}} \geq \hat{\gamma}_{k+1}$

$$
\begin{aligned}
(m_{k,i})^2 \hat{\gamma}_{k,i} &\leq (m_{k,i})^2 \gamma_{k,i} \\
&= \frac{\alpha_{k-1}}{(1-\beta_1)} \frac{\left((1-\beta_1)\sum_{j=1}^k \beta_1^{k-j} g_j\right)^2}{\sqrt{(1-\beta_2)\sum_{j=1}^k \beta_2^{k-j} g_j^2}} \\
&= \frac{\alpha_{k-1}(1-\beta_1)}{\sqrt{1-\beta_2}} \frac{\left(\sum_{j=1}^k \left(\beta_2^{\frac{k-j}{4}} |g_j|^{\frac{1}{2}}\right)\left(\beta_1 \beta_2^{1/2}\right)^{\frac{k-j}{2}} \left(\beta_1^{k-j}|g_j|\right)^{\frac{1}{2}}\right)^2}{\sqrt{\sum_{j=1}^k \beta_2^{k-j} g_j^2}} \\
&\leq \frac{\alpha_{k-1}(1-\beta_1)}{\sqrt{1-\beta_2}} \left(\sum_{j=1}^k \left(\frac{\beta_1^2}{\beta_2}\right)^{k-j}\right)^{\frac{1}{2}} \left(\sum_{j=1}^k \beta_1^{k-j} |g_j|\right) \\
&\leq \frac{\alpha_{k-1}(1-\beta_1)}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}} \sum_{j=1}^k \beta_1^{k-j} |g_j|
\end{aligned}
$$

Sketch of proof 5

- By remarking that $\sum_{k=j}^{K-1} \alpha_k \beta_1^{k-j} \leq \frac{\alpha_j}{1-\beta_1}$, we get

$$\sum_{k=0}^{K-1} (m_{k+1,i})^2 \hat{\gamma}_{k,i} \leq \frac{1}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}} \sum_{k=0}^{K-1} \alpha_k |\nabla_i f(x_k, \xi_{k+1})|$$

$$\leq \frac{\alpha_0 \sqrt{1+\ln(K)}}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}} \sqrt{\sum_{k=0}^{K-1} (\nabla_i f(x_k, \xi_{k+1}))^2}$$

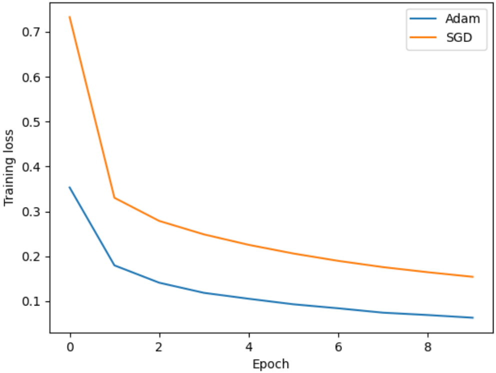
$$\leq \frac{\alpha_0 \sqrt{1+\ln(K)}}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}} G\sqrt{K}$$

- We only apply the expectation on ξ_k in the end:

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[f(x_k, \xi_{k+1}) - f(x^*, \xi_{k+1})] \in O(\frac{\ln K}{\sqrt{K}})$$

Numerical illustration on MNIST

Default Keras parameters
SGD: $\alpha_k = 0.01$
ADAM: $\alpha_k = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-7}$



1 epoch = 10,000 samples = 10 seconds

Conclusion

- Adam = several improvements over SGD, that combine well
- Tuning of learning rate is easier
- Behaviour on convex as well as non convex problems is good