Exercise: Convergence rate of ADAM

We now introduce an algorithm often used to train neural network models: Adam, which stands for stochastic gradient with adaptive moment estimation. Its main ingredients are an adaptive estimation of the first and second moments of the stochastic gradient and coordinate-wise step sizes. The idea is to design an exponential moving average of previous gradients and square gradients to estimate its moments. Finally, instead of just using the estimate of $\nabla F(x)$ to set the step-size, ADAM uses it directly as a means of reducing the variance of the stochastic gradient. The algorithm uses parameters $\alpha > 0, \beta_1 \in [0,1]$, $\beta_2 \in [0,1]$ and $\epsilon > 0$. It is initialized with a fixed x_0 and $m_0 = v_0 = 0$.

Algorithm: Adam

We recall the following notation: For any two vectors $a, b \in \mathbb{R}^d$, we define

$$ab = \begin{pmatrix} a_1 b_1 \\ \vdots \\ a_d b_d \end{pmatrix}, \qquad \frac{a}{b} = \begin{pmatrix} a_1/b_1 \\ \vdots \\ a_d/b_d \end{pmatrix}, \qquad \|a\|_b^2 = \sum_{i=1}^{n} a_i^2 b_i, \qquad \langle a, a' \rangle_b = \sum_{i=1}^{d} a_i a_i' b_i,$$

if these quantities exist. The Adam algorithm can be written as follows.

Algorithm 1: Adam

Inputs: $\alpha > 0, \beta_1 \in [0,1]$, $\beta_2 \in [0,1]$, $\epsilon > 0$

Start from $x_0 \in \mathbb{R}^d$ and let $v_0 = \hat{v}_0 = m_0 = 0$.

Until termination condition, iterate

$$\begin{aligned}
&\xi_{k+1} \sim \mathcal{D} \\
&m_{k+1} = \beta_1 m_k + (1-\beta_1) \nabla f(x_k, \xi_{k+1}) \\
&\hat{m}_{k+1} = \frac{m_{k+1}}{1-\beta_1^{k+1}} \\
&v_{k+1} = \beta_2 v_k + (1-\beta_2) \nabla f(x_k, \xi_{k+1})^2 \\
&\hat{v}_{k+1} = \max\left(\hat{v}_k, \frac{v_{k+1}}{1-\beta_2^{k+1}}\right) \\
&x_{k+1} = x_k - \frac{\alpha_k}{\epsilon + \sqrt{\hat{v}_{k+1}}} \hat{m}_{k+1}
\end{aligned}$$

This algorithm is widely used in the training of neural networks. Its convergence properties for convex functions are given in the following theorem.

Theorem 1. *Suppose that*

1. *$f(\cdot, \xi)$ is convex for all ξ*

2. $\exists x^* \in \arg\min F, F(x) = \mathbb{E}[f(x,\xi)]$

3. *For all k, for all i,* $\left|x_{k,i} - x_i^*\right| \leq D$

4. *For all* x, ξ*, for all i,* $\left|\nabla_i f(x,\xi)\right| \leq G$

5. $\alpha_k = \frac{\alpha_0}{\sqrt{k+1}}$

6. $\beta_1^2 < \beta_2 < 1, \epsilon = 0$

Then the iterates of Adam satisfy

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

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

We will prove this theorem as an exercise.

Exercise

We define $\hat{\gamma}_{k+1} = \frac{\alpha_k}{\left(1-\beta_1^{k+1}\right)\sqrt{\hat{v}_{k+1}}}$ (since $\epsilon = 0$) so that $x_{k+1} = x_k - \hat{\gamma}_{k+1} m_{k+1}$.

1. Show that
$$f\left(x_k, \xi_{k+1}\right) - f\left(x^*, \xi_{k+1}\right) \leq \left\langle \nabla f\left(x_k, \xi_{k+1}\right), x_k - x^* \right\rangle$$

2. Using the relation $m_{k+1} = \beta_1 m_k + \left(1-\beta_1\right)\nabla f\left(x_k, \xi_{k+1}\right)$, show that
$\left\langle \nabla f\left(x_k, \xi_{k+1}\right), x_k - x^* \right\rangle = \left\langle m_{k+1}, x_k - x^* \right\rangle + \frac{\beta_1}{1-\beta_1}\left(\left\langle m_{k+1}, x_{k+1} - x^* \right\rangle - \left\langle m_k, x_k - x^* \right\rangle\right) + \frac{\beta_1}{1-\beta_1}\left\|m_{k+1}\right\|_{\hat{\gamma}_{k+1}}^2$.

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

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

5. Show that $\left(\hat{\gamma}_k\right)_k$ is a coordinate-wise decreasing sequence (i.e. $\hat{\gamma}_{k,i} > \hat{\gamma}_{k+1,i}$ for all i,k) and that $\hat{\gamma}_k \geq \frac{\alpha_0\sqrt{1-\beta_2}}{\sqrt{k}G}$.

6. Show that
$$\sum_{k=0}^{K-1} \frac{1}{2}\left\|x_k - x^*\right\|_{\hat{\gamma}_{k+1}^{-1}}^2 - \frac{1}{2}\left\|x_{k+1} - x^*\right\|_{\hat{\gamma}_{k+1}^{-1}}^2 \leq \frac{D^2}{2}\sum_{i=1}^{d}\frac{1}{\hat{\gamma}_{K,i}}.$$

7. Show that

$$\langle m_K, x_K - x^* \rangle \leq \frac{1}{2} \|m_K\|^2_{\hat{\gamma}_K} + \frac{D^2}{2} \sum_{i=1}^{d} \frac{1}{\hat{\gamma}_{K,i}} \leq \frac{1}{2} \sum_{k=0}^{K-1} \|m_{k+1}\|^2_{\hat{\gamma}_{k+1}} + \frac{D^2}{2} \sum_{i=1}^{d} \frac{1}{\hat{\gamma}_{K,i}}.$$

8. Define $\gamma_{k+1} = \frac{\alpha_k}{(1-\beta_1)\sqrt{v_{k+1}}}$. Show that $\gamma_{k+1} \geq \hat{\gamma}_{k+1}$.

9. Let $x, y, z \in \mathbb{R}^d_+$ be nonnegative vectors and let p, q, r be positive real numbers such that $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$. We recall that the Hölder inequality ensures that $\sum_{j=1}^{d} x_j y_j z_j \leq \|x\|_p \|y\|_q \|z\|_r$. Define $g_{k+1}(i) = \partial_i f(x_k, \xi_{k+1})$. In this question, we will slightly abuse notation and write g_j rather than $g_j(i)$ for short. Justify each one of the following equalities and inequalities.

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

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

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

11. Show that

$$\sum_{k=0}^{K-1} \left(m_{k+1,i}\right)^2 \hat{\gamma}_{k+1,i} \leq \frac{\alpha_0 \sqrt{1+\ln(K)}}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}} \sqrt{\sum_{k=0}^{K-1} \left(\partial_i f(x_k, \xi_{k+1})\right)^2}.$$

12. Conclude