Convergence rate of ADAM: solution

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

1. Show that
$$f\left(x_k, \xi_{k+1}\right) - f\left(x^*, \xi_{k+1}\right) \leq \left\langle \nabla f\left(x_k, \xi_{k+1}\right), x_k - x^* \right\rangle$$
A convex and differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$ always satisfies
$$\phi(b) \geq \phi(a) + \langle \nabla\phi(a), b - a\rangle, \quad \forall a, b \in \mathbb{R}^d.$$
Applying this inequality to the function $\phi = f(\cdot, \xi_{k+1})$ and $a = x_k^*, b = x^*$ yields the result.

2. Using the relation $m_{k+1} = \beta_1 m_k + (1-\beta_1)\nabla f\left(x_k, \xi_{k+1}\right)$, show that $\left\langle \nabla f\left(x_k, \xi_{k+1}\right), x_k - x^* \right\rangle = \langle m_{k+1}, x_k - x^*\rangle + \frac{\beta_1}{1-\beta_1}\left(\langle m_{k+1}, x_{k+1} - x^*\rangle - \langle m_k, x_k - x^*\rangle\right) + \frac{\beta_1}{1-\beta_1}\|m_{k+1}\|^2_{\hat{\gamma}_{k+1}}$.
We recall that $x_{k+1} = x_k - \hat{\gamma}_{k+1} m_{k+1}$. Therefore,
$$\begin{aligned}
&\langle m_{k+1}, x_k - x^*\rangle + \frac{\beta_1}{1-\beta_1}\Big(\Big\langle m_{k+1}, \underbrace{x_{k+1} - x^*}_{x_k - x^* - \hat{\gamma}_{k+1} m_{k+1}}\Big\rangle - \langle m_k, x_k - x^*\rangle\Big) + \frac{\beta_1}{1-\beta_1}\|m_{k+1}\|^2_{\hat{\gamma}_{k+1}} \\
&= \langle m_{k+1}, x_k - x^*\rangle + \frac{\beta_1}{1-\beta_1}\Big(\Big\langle m_{k+1} - m_k, x_k - x^*\Big\rangle - \cancel{\langle m_{k+1}, \hat{\gamma}_{k+1} m_{k+1}\rangle}\Big) + \cancel{\frac{\beta_1}{1-\beta_1}\|m_{k+1}\|^2_{\hat{\gamma}_{k+1}}} \\
&= \Big\langle \frac{1}{1-\beta_1}(m_{k+1} - \beta_1 m_k), x_k - x^*\Big\rangle \\
&= \left\langle \nabla f\left(x_k, \xi_{k+1}\right), x_k - x^*\right\rangle \quad \text{by definition of } m_{k+1}.
\end{aligned}$$

3. Show that
$$\langle m_{k+1}, x_k - x^*\rangle = \frac{1}{2}\|x_k - x^*\|^2_{\hat{\gamma}_{k+1}^{-1}} - \frac{1}{2}\|x_{k+1} - x^*\|^2_{\hat{\gamma}_{k+1}^{-1}} + \frac{1}{2}\|m_{k+1}\|^2_{\hat{\gamma}_{k+1}}.$$
We can follow the same steps as in the proof of Adagrad's convergence rate.
$$\begin{aligned}
&\frac{1}{2}\|x_k - x^*\|^2_{\hat{\gamma}_{k+1}^{-1}} - \frac{1}{2}\Big\| \underbrace{x_{k+1} - x_k}_{= -\hat{\gamma}_{k+1} m_{k+1}} + x_k - x^*\Big\|^2_{\hat{\gamma}_{k+1}^{-1}} + \frac{1}{2}\|m_{k+1}\|^2_{\hat{\gamma}_{k+1}} \\
&= \cancel{\frac{1}{2}\|x_k - x^*\|^2_{\hat{\gamma}_{k+1}^{-1}}} - \frac{1}{2}\|\hat{\gamma}_{k+1} m_{k+1}\|^2_{\hat{\gamma}_{k+1}^{-1}} - \cancel{\frac{1}{2}\|x_k - x^*\|^2_{\hat{\gamma}_{k+1}^{-1}}} - \left\langle -\hat{\gamma}_{k+1} m_{k+1}, x_k - x^*\right\rangle_{\hat{\gamma}_{k+1}^{-1}} + \frac{1}{2}\|m_{k+1}\|^2_{\hat{\gamma}_{k+1}} \\
&= -\frac{1}{2}\cancel{\left(\sum_{j=1}^d \left(\hat{\gamma}_{k+1}(j) m_{k+1}(j)\right)^2 \hat{\gamma}_{k+1}^{-1}(j)\right)} - \left(\sum_{j=1}^d \Big(-\hat{\gamma}_{k+1}(j) m_{k+1}(j)\Big)\Big(x_k(j) - x^*(j)\Big)\hat{\gamma}_{k+1}^{-1}(j)\right) \\
&\quad + \frac{1}{2}\cancel{\left(\sum_{j=1}^d \left(m_{k+1}(j)\right)^2 \hat{\gamma}_{k+1}(j)\right)} \\
&= \left\langle m_{k+1}, x_k - x^*\right\rangle.
\end{aligned}$$

4. Show that

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

We can combine Questions 1, 2, and 3 to get:

$$\begin{aligned}
\sum_{k=0}^{K-1} f\left(x_k, \xi_{k+1}\right) - f\left(x^*, \xi_{k+1}\right) &\leq \sum_{k=0}^{K-1} \left\langle \nabla f\left(x_k, \xi_{k+1}\right), x_k - x^* \right\rangle \qquad \text{by Question 1.} \\
&= \sum_{k=0}^{K-1} \Big(\langle m_{k+1}, x_k - x^* \rangle + \frac{\beta_1}{1-\beta_1} \left(\langle m_{k+1}, x_{k+1} - x^* \rangle - \langle m_k, x_k - x^* \rangle \right) \\
&\qquad + \frac{\beta_1}{1-\beta_1} \left\|m_{k+1}\right\|^2_{\hat{\gamma}_{k+1}} \Big) \qquad \text{by Question 2.} \\
&= \sum_{k=0}^{K-1} \frac{1}{2}\left\|x_k - x^*\right\|^2_{\hat{\gamma}^{-1}_{k+1}} - \frac{1}{2}\left\|x_{k+1} - x^*\right\|^2_{\hat{\gamma}^{-1}_{k+1}} + \frac{1}{2}\left\|m_{k+1}\right\|^2_{\hat{\gamma}_{k+1}} \\
&\quad + \sum_{k=0}^{K-1} \frac{\beta_1}{1-\beta_1} \left(\langle m_{k+1}, x_{k+1} - x^* \rangle - \langle m_k, x_k - x^* \rangle \right) \\
&\quad + \sum_{k=0}^{K-1} \frac{\beta_1}{1-\beta_1} \left\|m_{k+1}\right\|^2_{\hat{\gamma}_{k+1}} \qquad \text{by Question 3.} \\
&= \sum_{k=0}^{K-1} \frac{1}{2}\left\|x_k - x^*\right\|^2_{\hat{\gamma}^{-1}_{k+1}} - \frac{1}{2}\left\|x_{k+1} - x^*\right\|^2_{\hat{\gamma}^{-1}_{k+1}} \\
&\quad + \frac{\beta_1}{1-\beta_1} \left(\langle m_K, x_K - x^* \rangle - \langle m_0, x_0 - x^* \rangle \right) \\
&\quad + \left(\frac{\beta_1}{1-\beta_1} + \frac{1}{2} \right) \sum_{k=0}^{K-1} \left\|m_{k+1}\right\|^2_{\hat{\gamma}_{k+1}}.
\end{aligned}$$

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

We have $\alpha_k < \alpha_{k-1}$. Moreover, $1 - \beta_1^{k+1} \geq 1 - \beta_1^k$, and $\hat{v}_{k+1} \geq \hat{v}_k$ by definition of $\hat{v}_{k+1}$, so we obtain that $(\hat{\gamma}_k)_k$ is a decreasing sequence:

$$\hat{\gamma}_{k+1} = \frac{\alpha_k}{(1-\beta_1^{k+1})\sqrt{\hat{v}_{k+1}}} < \frac{\alpha_{k-1}}{(1-\beta_1^k)\sqrt{\hat{v}_k}} = \hat{\gamma}_k.$$

To prove the relation $\hat{\gamma}_k \geq \frac{\alpha_0\sqrt{1-\beta_2}}{\sqrt{k}G}$, we first prove an intermediary result by induction:

$$\forall k \in \mathbb{N}: \quad \hat{v}_k \leq \frac{G^2}{1-\beta_2} \quad \text{and} \quad |v_k| \leq G^2.$$

Indeed,

- We have $|v_1| = (1-\beta_2)\left|\nabla f(x_0,\xi_1)\right|^2 \leq (1-\beta_2)G^2$. Moreover,

$$\begin{aligned}\hat{v}_1 &= \max\left(\hat{v}_0, \frac{v_1}{1-\beta_2}\right) = \max\left(0, \frac{(1-\beta_2)G^2}{1-\beta_2}\right)\\ &= G^2 \leq \frac{G^2}{1-\beta_2}.\end{aligned}$$

- Suppose we have $\hat{v}_k \leq \frac{G^2}{1-\beta_2}$ and $v_k \leq G^2$ for some fixed $k \in \mathbb{N}$, and let us show that $\hat{v}_{k+1} \leq \frac{G^2}{1-\beta_2}$ and $v_{k+1} \leq G^2$. We have

$$|v_{k+1}| = \left|\beta_2 v_k + (1-\beta_2)\nabla f(x_k,\xi_{k+1})^2\right| \leq \beta_2 G^2 + (1-\beta_2)G^2 = G^2.$$

  Therefore,

$$\begin{aligned}\hat{v}_{k+1} &= \max\left(\hat{v}_k, \frac{v_{k+1}}{1-\beta_2^{k+1}}\right)\\ &\leq \max\left(\frac{G^2}{1-\beta_2}, \frac{G^2}{1-\beta_2^{k+1}}\right)\\ &= \frac{G^2}{1-\beta_2}.\end{aligned}$$

Therefore, we have proved that for any $k \in \mathbb{N}$, the following property holds:

$$\hat{v}_k \leq \frac{G^2}{1-\beta_2} \quad \text{and} \quad |v_k| \leq G^2.$$

To conclude, it now suffices to recall the definition of $\hat{\gamma}_k$:

$$\begin{aligned}\hat{\gamma}_k &= \frac{\alpha_{k-1}}{(1-\beta_1^k)\sqrt{\hat{v}_k}} \leq \frac{\alpha_0}{\sqrt{k}\sqrt{\hat{v}_k}} \qquad \text{since } 1-\beta_1^k \leq 1\\ &\leq \frac{\alpha_0\sqrt{1-\beta_2}}{\sqrt{k}\,G}, \qquad \text{by the intermediary property.}\end{aligned}$$

6. Show that

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

We have

$$\begin{aligned}&\sum_{k=0}^{K-1} \frac{1}{2}\left\|x_k - x^*\right\|^2_{\hat{\gamma}^{-1}_{k+1}} - \frac{1}{2}\left\|x_{k+1} - x^*\right\|^2_{\hat{\gamma}^{-1}_{k+1}} = \sum_{k=0}^{K-1} \frac{1}{2}\left\|x_k - x^*\right\|^2_{\hat{\gamma}^{-1}_{k+1}} - \sum_{k=1}^{K} \frac{1}{2}\left\|x_k - x^*\right\|^2_{\hat{\gamma}^{-1}_{k}}\\
&= \frac{1}{2}\left\|x_0 - x^*\right\|^2_{\hat{\gamma}_1^{-1}} \underbrace{-\frac{1}{2}\left\|x_K - x^*\right\|^2_{\hat{\gamma}_K^{-1}}}_{\leq 0} + \sum_{k=1}^{K-1}\frac{1}{2}\left\|x_k - x^*\right\|^2_{\hat{\gamma}^{-1}_{k+1} - \hat{\gamma}^{-1}_{k}}\\
&\leq \sum_{j=1}^{d}\frac{1}{2}\frac{(x_{0,j} - x^*_j)^2}{\hat{\gamma}_{1,j}} + 0 + \frac{1}{2}\sum_{k=1}^{K-1}\sum_{j=1}^{d}(x_{k,j} - x^*_j)^2\left(\frac{1}{\hat{\gamma}_{k+1,j}} - \frac{1}{\hat{\gamma}_{k,j}}\right)\\
&\leq \sum_{j=1}^{d}\frac{1}{2}\frac{D^2}{\hat{\gamma}_{1,j}} + \frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{K-1} D^2\left(\frac{1}{\hat{\gamma}_{k+1,j}} - \frac{1}{\hat{\gamma}_{k,j}}\right) \qquad \text{since } \frac{1}{\hat{\gamma}_{k+1,j}} - \frac{1}{\hat{\gamma}_{k,j}} \geq 0\\
&= \frac{D^2}{2}\sum_{i=1}^{d}\frac{1}{\hat{\gamma}_{K,i}}.\end{aligned}$$

7. Show that

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

We use the following inequality, true for any $a, b \in \mathbb{R}$:

$$ab \leq \frac{a^2}{2} + \frac{b^2}{2}.$$

We obtain:

$$\begin{aligned}
\langle m_K, x_k - x^* \rangle &= \sum_{j=1}^{d} m_{K,j}(x_{K,j} - x^*_j) = \sum_{j=1}^{d} m_{K,j} \sqrt{\hat{\gamma}_{K,j}} \, \frac{x_{K,j} - x^*_j}{\sqrt{\hat{\gamma}_{K,j}}} \\
&\leq \frac{1}{2} \sum_{j=1}^{d} m^2_{K,j} \hat{\gamma}_{K,j} \; + \frac{1}{2} \sum_{j=1}^{d} \frac{(x_{K,j} - x^*_j)^2}{\hat{\gamma}_{K,j}} \\
&\leq \frac{1}{2} \|m_K\|^2_{\hat{\gamma}_K} + \frac{1}{2} \sum_{j=1}^{d} \frac{D^2}{\hat{\gamma}_{K,j}} \\
&\leq \frac{1}{2} \sum_{k=0}^{K-1} \|m_{k+1}\|^2_{\hat{\gamma}_{k+1}} + \frac{D^2}{2} \sum_{i=1}^{d} \frac{1}{\hat{\gamma}_{K,i}}.
\end{aligned}$$

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

We have $1 - \beta_1^{k+1} \geq 1 - \beta_1$ and $\hat{v}_{k+1} \geq v_{k+1}$, hence

$$\gamma_{k+1} = \frac{\alpha_k}{(1-\beta_1)\sqrt{v_{k+1}}} \geq \frac{\alpha_k}{\left(1 - \beta_1^{k+1}\right)\sqrt{\hat{v}_{k+1}}} = \hat{\gamma}_{k+1}.$$

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.

Before justifying these inequalities, it will be useful to prove the following relations by induction

$$m_{k,i} = (1 - \beta_1) \sum_{j=1}^{k} \beta_1^{k-j} g_j \tag{1}$$

$$v_{k,i} = (1 - \beta_2) \sum_{j=1}^{k} \beta_2^{k-j} g_j^2. \tag{2}$$

- We have $m_{1,i} = \beta_1 m_0 + (1 - \beta_1) g_1 = (1 - \beta_1) g_1 = (1 - \beta_1) \sum_{j=1}^{1} \beta_1^{1-j} g_j$.
  Similarly, $v_{1,i} = \beta_2 v_0 + (1 - \beta_2) g_1^2 = (1 - \beta_2) g_1^2 = (1 - \beta_2) \sum_{j=1}^{1} \beta_2^{1-j} g_j^2$.
- Suppose the relations (1) and (2) hold for some $k \geq 1$. Let us prove that they hold for $k + 1$ as well. We have

$$\begin{aligned}
m_{k+1,i} &= \beta_1 m_{k,i} + (1 - \beta_1) g_{k+1} \\
&= \beta_1 \cdot (1 - \beta_1) \sum_{j=1}^{k} \beta_1^{k-j} g_j + (1 - \beta_1) g_{k+1}
\end{aligned}$$

$$= (1-\beta_1)\sum_{j=1}^{k}\beta_1^{k+1-j}g_j + (1-\beta_1)g_{k+1}$$
$$= (1-\beta_1)\sum_{j=1}^{k+1}\beta_1^{k+1-j}g_j.$$

We proceed similarly for v_k.

$$\left(m_{k,i}\right)^2\hat{\gamma}_{k,i} \leq \left(m_{k,i}\right)^2\gamma_{k,i} \qquad \text{true by Question 8}$$

$$= \left(m_{k,i}\right)^2\frac{\alpha_{k-1}}{(1-\beta_1)\sqrt{v_k}}$$
$$= \frac{\alpha_{k-1}}{(1-\beta_1)}\frac{\left((1-\beta_1)\sum_{j=1}^{k}\beta_1^{k-j}g_j\right)^2}{\sqrt{(1-\beta_2)\sum_{j=1}^{k}\beta_2^{k-j}g_j^2}} \qquad \text{by equations (1) and (2)}$$
$$\leq \frac{\alpha_{k-1}(1-\beta_1)}{\sqrt{1-\beta_2}}\frac{\left(\sum_{j=1}^{k}\left(\beta_2^{\frac{k-j}{4}}\left|g_j\right|^{\frac{1}{2}}\right)\left(\beta_1\beta_2^{-1/2}\right)^{\frac{k-j}{2}}\left(\beta_1^{k-j}\left|g_j\right|\right)^{\frac{1}{2}}\right)^2}{\sqrt{\sum_{j=1}^{k}\beta_2^{k-j}g_j^2}}$$

since $\left(\beta_2^{\frac{k-j}{4}}\left|g_j\right|^{\frac{1}{2}}\right)\left(\beta_1\beta_2^{-1/2}\right)^{\frac{k-j}{2}}\left(\beta_1^{k-j}\left|g_j\right|\right)^{\frac{1}{2}} = \beta_1^{k-j}g_j$

$$\leq \frac{\alpha_{k-1}(1-\beta_1)}{\sqrt{1-\beta_2}}\left(\sum_{j=1}^{k}\left(\frac{\beta_1^2}{\beta_2}\right)^{k-j}\right)^{\frac{1}{2}}\left(\sum_{j=1}^{k}\beta_1^{k-j}\left|g_j\right|\right)$$

by Hölder's inequality with $p=q=4$, and $r=2$

$$\leq \frac{\alpha_{k-1}(1-\beta_1)}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}}\sum_{j=1}^{k}\beta_1^{k-j}\left|g_j\right|$$

since $\sum_{j=1}^{k}\left(\frac{\beta_1^2}{\beta_2}\right)^{k-j} = \frac{1-\left(\frac{\beta_1^2}{\beta_2}\right)^k}{1-\frac{\beta_1^2}{\beta_2}} \leq \frac{1}{1-\frac{\beta_1^2}{\beta_2}}$.

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

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

First, since $(\alpha_k)_k$ is decreasing, we have

$$\sum_{k=j}^{K-1}\alpha_k\beta_1^{k-j} \leq \alpha_j\sum_{\ell=0}^{K-1-j}\beta_1^{\ell} \leq \frac{\alpha_j}{1-\beta_1}.$$

We let $C = \frac{(1-\beta_1)}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}}$. By the previous question, we get

$$\begin{aligned}\sum_{k=0}^{K-1}\left(m_{k+1,i}\right)^2\hat{\gamma}_{k+1,i} &\leq C\sum_{k=0}^{K-1}\alpha_k\sum_{j=1}^{k+1}\beta_1^{k+1-j}\left|g_j\right| = C\sum_{j=1}^{K}\left|g_j\right|\sum_{\substack{k\leq K-1\\ k+1\geq j}}\alpha_k\beta_1^{k+1-j}\\ &= C\sum_{j'=0}^{K-1}\left|g_{j'+1}\right|\sum_{\substack{k\leq K-1\\ k\geq j'}}\alpha_k\beta_1^{k+j'} \leq \sum_{j'=0}^{K-1}\frac{C\alpha_{j'}}{1-\beta_1}\\ &= \frac{1}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}}\sum_{k=0}^{K-1}\alpha_k\left|\nabla_i f\left(x_k,\xi_{k+1}\right)\right|.\end{aligned}$$

11. Show that

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

We apply Cauchy-Schwarz's inequality to the result of Question 11

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

It suffices to prove that $\sum_{k=0}^{K-1}\alpha_k^2 \leq 1+\log(K)$ to obtain the desired result. We have

$$\begin{aligned}\sum_{k=0}^{K-1}\alpha_k^2 &= \alpha_0^2\sum_{k=0}^{K-1}\frac{1}{k+1} = \alpha_0^2\sum_{k=1}^{K}\frac{1}{k}\\ &= \alpha_0^2 + \alpha_0^2\sum_{k=2}^{K}\frac{1}{k} \leq \alpha_0^2\left(1+\sum_{k=1}^{K}\int_{k-1}^{k}\frac{1}{x}dx\right)\\ &= \alpha_0^2\left(1+\int_1^K\frac{1}{x}dx\right) = \alpha_0^2\left(1+\log(K)\right).\end{aligned}$$

12. Conclude

Before solving the question, we first note that the result of Question 11 ensures that:

$$\left\|m_{k+1}\right\|_{\hat{\gamma}_{k+1}}^2 \leq d\cdot\frac{\alpha_0\sqrt{1+\ln(K)}}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}}\sqrt{\sum_{k=0}^{K-1}\left(\nabla_i f\left(x_k,\xi_{k+1}\right)\right)^2} \leq \frac{\alpha_0\, d\sqrt{1+\ln(K)}}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}}\sqrt{KG^2}. \tag{3}$$

Moreover, we have also proved $\hat{\gamma}_k \geq \frac{\alpha_0\sqrt{1-\beta_2}}{\sqrt{k}G}$ in Question 5, so that

$$\sum_{i=1}^{d}\frac{1}{\hat{\gamma}_{K,i}} \leq \frac{dG\sqrt{K}}{\alpha_0\sqrt{1-\beta_2}} \tag{4}$$

We now bring everything together. We have

$$\begin{aligned}
\sum_{k=0}^{K-1} f\left(x_k, \xi_{k+1}\right) - f\left(x^*, \xi_{k+1}\right) &\leq \frac{\beta_1}{1-\beta_1}\left(\langle m_K, x_K - x^*\rangle - \left\langle \cancelto{=0}{m_0}, x_0 - x^*\right\rangle\right) \\
&+ \sum_{k=0}^{K-1}\left(\frac{1}{2}\left\|x_k - x^*\right\|^2_{\hat{\gamma}_{k+1}^{-1}} - \frac{1}{2}\left\|x_{k+1} - x^*\right\|^2_{\hat{\gamma}_{k+1}^{-1}}\right) \\
&+ \left(\frac{\beta_1}{1-\beta_1} + \frac{1}{2}\right)\sum_{k=0}^{K-1}\left\|m_{k+1}\right\|^2_{\hat{\gamma}_{k+1}} \qquad \text{by Question 4} \\
&\leq \frac{\beta_1}{1-\beta_1}\left(\frac{1}{2}\sum_{k=0}^{K-1}\left\|m_{k+1}\right\|^2_{\hat{\gamma}_{k+1}} + \frac{D^2}{2}\sum_{i=1}^{d}\frac{1}{\hat{\gamma}_{K,i}}\right) \qquad \text{by Question 7} \\
&+ \frac{D^2}{2}\sum_{i=1}^{d}\frac{1}{\hat{\gamma}_{K,i}} \qquad \text{by Question 6} \\
&+ \left(\frac{\beta_1}{1-\beta_1} + \frac{1}{2}\right)\sum_{k=0}^{K-1}\left\|m_{k+1}\right\|^2_{\hat{\gamma}_{k+1}} \\
&= \frac{D^2}{2(1-\beta_1)}\sum_{i=1}^{d}\frac{1}{\hat{\gamma}_{K,i}} + \left(\frac{1+2\beta_1}{2(1-\beta_1)}\right)\sum_{k=0}^{K-1}\left\|m_{k+1}\right\|^2_{\hat{\gamma}_{k+1}} \\
&\leq \frac{D^2}{2(1-\beta_1)}\frac{dG\sqrt{K}}{\alpha_0\sqrt{1-\beta_2}} \qquad \text{by equation (4)} \\
&+ \left(\frac{1+2\beta_1}{2(1-\beta_1)}\right)\frac{\alpha_0\, d\sqrt{1+\ln(K)}}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}}\sqrt{KG^2} \qquad \text{by equation (3)}
\end{aligned}$$

We can now conclude as in the proof of Adagrad's convergence rate. Taking the expectation on both sides and dividing by K, we can write:

$$\begin{aligned}
&\frac{D^2}{2(1-\beta_1)}\frac{dG\sqrt{K}}{\alpha_0\sqrt{1-\beta_2}} + \left(\frac{1+2\beta_1}{2(1-\beta_1)}\right)\frac{\alpha_0\, d\sqrt{1+\ln(K)}}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}}\sqrt{KG^2} \\
&\geq \frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}\mathbb{E}_k\left[f\left(x_k, \xi_{k+1}\right) - f\left(x^*, \xi_{k+1}\right)\right]\right] \qquad \text{using } \mathbb{E}X = \mathbb{E}\big[\mathbb{E}(X|Y)\big]\ \forall X, Y \\
&= \frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1} F(x_k) - F(x^*)\right] \qquad \text{by definition of } F \\
&\geq \mathbb{E}\left[F(\bar{x}_K) - F(x^*)\right] \qquad \text{by convexity.}
\end{aligned}$$