Exercise: Convergence rate of Adagrad

For any three vectors $a, a', b \in \mathbb{R}^d$, we define

$$a \odot b = \begin{pmatrix} a_1 b_1 \\ \vdots \\ a_d b_d \end{pmatrix}, \qquad a \oslash b \text{ or } \frac{a}{b} = \begin{pmatrix} a_1/b_1 \\ \vdots \\ a_d/b_d \end{pmatrix},$$

$$\|a\|_b^2 = \sum_{i=1}^n a_i^2 b_i, \qquad \langle a, a' \rangle_b = \sum_{i=1}^d a_i a_i' b_i$$

$$a^{\odot 2} = a \odot a,$$

Algorithm 1: Adagrad

1. Pick $x_0 \in \mathbb{R}^d$.
2. Until termination condition, iterate:

   Generate ξ_{k+1} independent of the past

   $$g_{k+1} = \begin{pmatrix} g_{k+1}(1) \\ \vdots \\ g_{k+1}(d) \end{pmatrix} = \nabla f_x \left(x_k, \xi_{k+1}\right)$$

   $$\gamma_{k+1}(j) = \frac{\alpha}{\sqrt{\sum_{s=0}^k g_{s+1}^2(j)}}, \ \forall j = 1, \ldots, d$$

   $$x_{k+1} = x_k - \gamma_{k+1} \odot g_{k+1}$$

The convergence guarantees of Adagrad are given in the theorem below.

Theorem 1. *Suppose that*

1. *$f(\cdot, \xi)$ is convex for all ξ*
2. *There exists $x^* \in \arg\min F$, where $F(x) = \mathbb{E}[f(x, \xi)]$*
3. *There exists $D > 0$ s.t. for all $k \geq 0$, for all $i \in \{1, \ldots, d\}$: $\left|x_{k,i} - x_i^*\right| \leq D$*
4. *For all $x, \xi : \|\nabla f(x, \xi)\| \leq G$.*

Then the iterates of Adagrad satisfy

$$\mathbb{E}\left[F\left(\bar{x}_K\right) - F\left(x^*\right)\right] \leq \frac{dG}{\sqrt{K}} \left(\frac{D^2}{\alpha} + 2\alpha\right)$$

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

1. Justify that $f\left(x_k, \xi_{k+1}\right) - f\left(x^*, \xi_{k+1}\right) \leq \langle g_{k+1}, x_k - x^* \rangle$. One can use the convexity of $f(\,\cdot\,, \xi_{k+1})$ and the fact that, for any convex function φ, we have $\varphi(y) - \varphi(x) \geq \langle \nabla \varphi(x), y - x \rangle$.

2. Using the relation $x_{k+1} = x_k - \gamma_{k+1} \odot g_{k+1}$, show that

$$\frac{1}{2}\left\|x_k - x^*\right\|^2_{\gamma_{k+1}^{-1}} - \frac{1}{2}\left\|x_{k+1} - x^*\right\|^2_{\gamma_{k+1}^{-1}} + \frac{1}{2}\left\|g_{k+1}\right\|^2_{\gamma_{k+1}} = \left\langle g_{k+1}, x_k - x^* \right\rangle.$$

3. Deduce that

$$\sum_{k=0}^{K-1} f\left(x_k, \xi_{k+1}\right) - f\left(x^*, \xi_{k+1}\right) \leq \sum_{k=0}^{K-1}\left(\frac{1}{2}\left\|x_k - x^*\right\|^2_{\gamma_{k+1}^{-1}} - \frac{1}{2}\left\|x_{k+1} - x^*\right\|^2_{\gamma_{k+1}^{-1}}\right) + \sum_{k=0}^{K-1}\frac{1}{2}\left\|g_{k+1}\right\|^2_{\gamma_{k+1}}. \quad (1)$$

From now on, we will control the two terms in equation (1) separately.

Control of the first term.

4. Prove that for any $z, a, b \in \mathbb{R}^d$ we have $\|z\|_a^2 - \|z\|_b^2 = \|z\|^2_{a-b}$.

5. Deduce that the first term in (1) can be rewritten as

$$\sum_{k=0}^{K-1}\left(\left\|x_k - x^*\right\|^2_{\gamma_{k+1}^{-1}} - \left\|x_{k+1} - x^*\right\|^2_{\gamma_{k+1}^{-1}}\right) = \left\|x_0 - x^*\right\|^2_{\gamma_1^{-1}} - \left\|x_K - x^*\right\|^2_{\gamma_K^{-1}} + \sum_{k=1}^{K-1}\left\|x_k - x^*\right\|^2_{\left(\gamma_{k+1}^{-1} - \gamma_k^{-1}\right)}$$

6. Show that the Adagrad learning rates γ_k's have coordinates that are non-increasing over time:

$$\forall j \in \{1, \ldots, d\}: \quad \gamma_{k+1}(j) \leq \gamma_k(j).$$

7. Using Assumption 3 in the theorem, deduce that the first term in (1) can be controlled as

$$\sum_{k=0}^{K-1}\left(\left\|x_k - x^*\right\|^2_{\gamma_{k+1}^{-1}} - \left\|x_{k+1} - x^*\right\|^2_{\gamma_{k+1}^{-1}}\right) \leq \sum_{j=1}^{d}\left[\frac{D^2}{\gamma_1(j)} + \sum_{k=1}^{K-1} D^2\left(\frac{1}{\gamma_{k+1}(j)} - \frac{1}{\gamma_k(j)}\right)\right].$$

8. Using Assumption 4 from the theorem, prove that $\gamma_K(j) \geq \frac{\alpha}{G\sqrt{K}}$.

9. Deduce the final upper bound on the first term

$$\sum_{k=0}^{K-1}\left(\left\|x_k - x^*\right\|^2_{\gamma_{k+1}^{-1}} - \left\|x_{k+1} - x^*\right\|^2_{\gamma_{k+1}^{-1}}\right) \leq \frac{D^2 d G\sqrt{K}}{\alpha}.$$

Control of the second part of (1): $\sum_{k=0}^{K-1}\left\|g_{k+1}\right\|^2_{\gamma_{k+1}}$.

10. For any $k \in \{0, K-1\}$, define $a_k^{(i)} = g_{k+1,i}^2 \geq 0$. Justify that

$$\sum_{k=0}^{K-1}\left\|g_{k+1}\right\|^2_{\gamma_{k+1}} = \alpha \sum_{k=0}^{K-1}\sum_{i=1}^{d}\frac{a_k^{(i)}}{\sqrt{\sum_{s=0}^{k} a_s^{(i)}}}.$$

We will use the following Lemma

Lemma 1. *Let (a_k) be a sequence of nonnegative numbers. Then*

$$\sum_{k=0}^{K-1}\frac{a_k}{\sqrt{\sum_{s=0}^{k} a_s}} \leq 2\sqrt{\sum_{s=0}^{K-1} a_s}$$

The proof of this lemma is a bonus question at the end of the exercise. For now, we will simply use this result without proof.

11. Using the lemma above, show that

$$\sum_{k=0}^{K-1} \|g_{k+1}\|^2_{\gamma_{k+1}} \leq 2\alpha dG\sqrt{K}.$$

12. Deduce that

$$\sum_{k=0}^{K-1} f\left(x_k, \xi_{k+1}\right) - f\left(x^*, \xi_{k+1}\right) \leq \frac{dD^2G\sqrt{K}}{\alpha} + 2\alpha dG\sqrt{K}.$$

13. Applying the expectation on both sides, and deduce that

$$\frac{dD^2G}{\alpha\sqrt{K}} + 2\frac{\alpha dG}{\sqrt{K}} \;\geq\; \frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1} F(x_k) - F(x^*)\right].$$

14. Using the convexity of F, conclude the proof of the theorem.

Bonus question: Prove the lemma.