A Deterministic Convergence Analysis

Here we analyze the AdaRem method in full batch setting and show that when the objective function is L-lipschitz convex, AdaRem converges with rate $O(1/k)$.

Notation Given two vectors $u, v \in \mathbb{R}^d$, we use $\langle u, v\rangle$ for inner product, $u \odot v$ for element-wise product, u/v to denote element-wise division. Given a vector $x \in \mathbb{R}^d$ we denote its i-th coordinate by x_i and its ℓ_2-norm by $\|x\|_2$. For a vector x_t in the t-th iteration, the i-th coordinate of x_t is denoted as $x_{t,i}$ by adding a subscript i. We use g_i to denote $\nabla f(x)_i$.

Claim A.1. *If $f : \mathbb{R}^d \to \mathbb{R}$ is L-lipschitz convex, then for all x, y,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2$$

Theorem A.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a L-lipschitz convex function and $x^* = \arg\min_x f(x)$. For AdaRem, we set adjustment coefficientat a_t bounded: $c \leq a_t \leq 2$, and weight decay factor $\lambda = 0$. Then, AdaRem with learning rate $\eta \leq \frac{1}{2L}$ satisfies the following:*

$$f(x_k) \leq f(x^*) + \frac{\|x_0 - x^*\|_2^2}{2c\eta k}$$

Proof. For $i = 0, ...,$ define

$$x_{i+1} = x_i - t_i \odot g_i$$

$$t_i = \eta a_i$$

First, by convexity of f, we have:

$$f(x_i) \leq f(x^*) + \langle g_i, x_i - x^*\rangle \,. \tag{1}$$

Further, as f is L-lipschitz, by the previous lemma,

$$\begin{aligned} f(x_{i+1}) &\leq f(x_i) + \langle g_i, x_{i+1} - x_i\rangle + \frac{L}{2}\|x_{i+1} - x_i\|_2^2 \\ &= f(x_i) + \langle g_i, -t_i \odot g_i\rangle + \frac{L}{2}\|t_i \odot g_i\|_2^2 \\ &= f(x_i) - \sum_{j=1}^{d} g_{i,j}^2 t_{i,j} + \frac{L}{2}\sum_{j=1}^{d} g_{i,j}^2 t_{i,j}^2 \\ &= f(x_i) - \sum_{j=1}^{d} t_{i,j}(1 - \frac{t_{i,j}L}{2}) g_{i,j}^2 \\ &\leq f(x_i) - \sum_{j=1}^{d} \frac{t_{i,j}}{2} g_{i,j}^2 \end{aligned} \tag{2}$$

where the last inequality follows as $t_{i,j}L = \eta a_{i,j} L \leq 1$. In particular, the above shows that AdaRem is monotonic: the objective value is non-decreasing. Combining the above two equations we get,

$$
\begin{aligned}
f(x_{i+1}) &\leq f(x^*) + \langle g_i, x_i - x^* \rangle - \sum_{j=1}^{d} \frac{t_{i,j}}{2} g_{i,j}^2 \\
&= f(x^*) + \sum_{j=1}^{d} \frac{(x_i - x^*)_j^2}{2t_{i,j}} - \left[\sum_{j=1}^{d} \frac{(x_i - x^*)_j^2}{2t_{i,j}} - \langle g_i, x_i - x^* \rangle + \sum_{j=1}^{d} \frac{t_{i,j}}{2} g_{i,j}^2 \right] \\
&= f(x^*) + \sum_{j=1}^{d} \frac{(x_i - x^*)_j^2}{2t_{i,j}} - \sum_{j=1}^{d} \frac{1}{2t_{i,j}} \left[(x_i - x^*)_j^2 - 2t_{i,j} g_{i,j} (x_i - x^*)_j + {t_{i,j}}^2 g_{i,j}^2 \right] \\
&= f(x^*) + \sum_{j=1}^{d} \frac{(x_i - x^*)_j^2}{2t_{i,j}} - \sum_{j=1}^{d} \frac{\left((x_i - x^*)_j - t_{i,j} g_{i,j} \right)^2}{2t_{i,j}} \\
&= f(x^*) + \sum_{j=1}^{d} \frac{(x_i - x^*)_j^2 - (x_{i+1} - x^*)_j^2}{2t_{i,j}} \\
&\leq f(x^*) + \frac{1}{2c\eta} \left(\|x_i - x^*\|_2^2 - \|x_{i+1} - x^*\|_2^2 \right).
\end{aligned} \tag{3}
$$

where the last inequality follows as $t_{i,j} \geq c\eta$. Summing the above equations for $i = 0, ..., k-1$, we get

$$
\sum_{i=0}^{k-1} \left(f\left(x_{i+1}\right) - f\left(x^*\right) \right) \leq \frac{1}{2c\eta} \left(\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2 \right) \leq \frac{\|x_0 - x^*\|_2^2}{2c\eta}
$$

Finally, by Equation (2), $f(x_0), ..., f(x_k)$ is non-increasing. Therefore, $f(x_k) - f(x^*) \leq f(x_i) - f(x^*)$ for all $i < k$. Thus

$$
k \cdot \left(f\left(x_k\right) - f\left(x^*\right) \right) \leq \frac{\|x_0 - x^*\|_2^2}{2c\eta}.
$$

The theorem now follows. $\square$