ADFOCS '21 Summer School: Adaptive Gradient Descent Algorithms

A tentative plan is to cover the first 4 exercises in the first session, and the remaining exercises in the second session.

In the following exercises, we give a proof of the results we used in our analysis of adaptive methods.

Exercise 1. Let $a_1, a_2, \ldots, a_n > 0$ be positive scalars. Show that

$$\sqrt{\sum_{i=1}^{n} a_i} \leq \sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \leq 2\sqrt{\sum_{i=1}^{n} a_i}$$

Hint: Approximate the sum by an integral.

Exercise 2. Let f be a differentiable function that is convex and β-smooth. Show that, for all x, y, we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\beta} \|\nabla f(y) - \nabla f(x)\|^2$$

Adagrad with per-coordinate step sizes In the next exercise, we consider a version of the Adagrad algorithm from Lecture I that uses per-coordinate step sizes. Let $\mathbf{D} \in \mathbb{R}^{d \times d}$ be a diagonal matrix with positive diagonal entries. We let $\mathbf{D}_i$ denote the i-th diagonal entry of $\mathbf{D}$. We also let

$$\|x\|_{\mathbf{D}} := \sqrt{x^\top \mathbf{D} x} = \sqrt{\sum_{i=1}^{d} \mathbf{D}_i x_i^2}$$

We let $\|x\|_\infty$ denote the ℓ_∞-norm of x: $\|x\|_\infty = \max_{i \in [d]} |x_i|$. We let $\nabla_i f(x)$ denote the i-th coordinate of $\nabla f(x)$.

Algorithm 1 Adagrad algorithm with per-coordinate step sizes.

```
Let $x_1 \in K$, $R_\infty \geq \max_{x,y \in K} \|x - y\|_\infty$
For $t = 1, \ldots, T$:
    $\mathbf{D}_{t,i} = \frac{1}{R_\infty} \sqrt{\sum_{s=1}^{t} (\nabla_i f(x_s))^2}$        $\forall i \in [d]$
    $x_{t+1} = \arg\min_{u \in K} \left\{ \langle \nabla f(x_t), u - x_t \rangle + \frac{1}{2} \|u - x_t\|_{\mathbf{D}_t}^2 \right\}$
Return $\overline{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$
```

In the unconstrained setting ($K = \mathbb{R}^d$), the update becomes

$$x_{t+1} = x_t - \mathbf{D}_t^{-1} \nabla f(x_t)$$

$$\Rightarrow x_{t+1,i} = x_{t,i} - \underbrace{\frac{R_\infty}{\sqrt{\sum_{s=1}^{t} (\nabla_i f(x_s))^2}}}_{:=\eta_{t,i}} \nabla_i f(x_t)$$

Thus the algorithm can be interpreted as building a diagonal preconditioner based on the gradient information seen so far. This approach is beneficial in practice since the coordinates may be very different (e.g., we can set high learning rates for informative features).

Exercise 3. Extend the analysis that we saw in Lecture I to Algorithm 1. You may find the following inequality helpful:

$$\|x - y\|_{\mathbf{D}}^2 \le \|x - y\|_\infty^2 \operatorname{tr}(\mathbf{D})$$

where $\operatorname{tr}(\mathbf{D}) = \sum_{i=1}^d \mathbf{D}_i$ is the trace of the diagonal matrix $\mathbf{D}$.

A different adaptive algorithm for constrained optimization In the following exercises, we will analyze a different algorithm for constrained optimization. The algorithm is based on a variant of gradient descent that uses extrapolation, which we will discuss in more detail in Lecture III. In contrast to the Adagrad+ algorithm that we saw in Lecture I that uses the iterate movement $\|x_{t+1} - x_t\|^2$ to set the step sizes, this algorithm uses the gradient differences $\|\nabla f(x_{t+1}) - \nabla f(x_t)\|^2$.

Algorithm 2 An adaptive version of the Past Extra-Gradient algorithm.

```
Let $x_0 = z_0 \in K$, $\eta_0 > 0$.
For $t = 1, \ldots, T$, update:
    $x_t = \arg\min_{u \in K} \left\{ \langle \nabla f(x_{t-1}), u \rangle + \frac{1}{2\eta_{t-1}} \|u - z_{t-1}\|^2 \right\}$
    $\eta_t = \frac{R}{\sqrt{\sum_{i=1}^t \|\nabla f(x_i) - \nabla f(x_{i-1})\|^2}}$
    $z_t = \arg\min_{u \in K} \left\{ \langle \nabla f(x_t), u \rangle + \frac{1}{2\eta_{t-1}} \|u - z_{t-1}\|^2 + \left( \frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|u - x_t\|^2 \right\}$
Return $\overline{x}_T = \frac{1}{T} \sum_{t=1}^T x_t$.
```

Exercise 4. As in lecture, start by upper bounding the function value difference $f(\overline{x}_T) - f(x^*)$ in terms of the inner products $\langle \nabla f(x_t), x_t - x^* \rangle$.

Next, we analyze the inner product $\langle \nabla f(x_t), x_t - x^* \rangle$. It is useful to split it as follows:

$$\begin{aligned} \langle \nabla f(x_t), x_t - x^* \rangle &= \langle \nabla f(x_t), z_t - x^* \rangle + \langle \nabla f(x_t), x_t - z_t \rangle \\ &= \langle \nabla f(x_t), z_t - x^* \rangle + \langle \nabla f(x_{t-1}), x_t - z_t \rangle + \langle \nabla f(x_t) - \nabla f(x_{t-1}), x_t - z_t \rangle \end{aligned}$$

The above split is a good one to consider, since it aligns well with the optimality conditions.

Exercise 5. Upper bound $\langle \nabla f(x_t), z_t - x^* \rangle$ and $\langle \nabla f(x_{t-1}), x_t - z_t \rangle$ using the optimality condition for z_t and x_t.

Next, we consider the third term $\langle \nabla f(x_t) - \nabla f(x_{t-1}), x_t - z_t \rangle$. The intuition is that the two gradients $\nabla f(x_t)$ and $\nabla f(x_{t-1})$ are close, and thus their difference $\|\nabla f(x_t) - \nabla f(x_{t-1})\|$ is small. Assuming this intuition is correct, we can try to show that the inner product is small by upper bounding it in terms of $\|\nabla f(x_t) - \nabla f(x_{t-1})\|$.

Exercise 6. Show that, if $K = \mathbb{R}^d$, we have

$$\langle \nabla f(x_t) - \nabla f(x_{t-1}), x_t - z_t \rangle \le \eta_t \|\nabla f(x_t) - \nabla f(x_{t-1})\|^2$$

Hint: Note that x_t is also the minimizer of $\|u - x_t\|^2$.

It turns out that we can prove the above result for general K as well. You may assume this fact without proof, and proceed with the analysis.

By combining the above inequality with the two inequalities that we obtained via the optimality conditions, we obtain an upper bound on $\langle \nabla f(x_t), x_t - x^* \rangle$. Next, as in lecture, we sum up these inequalities and telescope as much as possible.

Exercise 7. Show that we have

$$\sum_{t=1}^T \langle \nabla f(x_t), x_t - x^* \rangle \le \underbrace{O(R) \sqrt{\sum_{t=1}^T \|\nabla f(x_t) - \nabla f(x_{t-1})\|^2}}_{\text{loss}} - \underbrace{\sum_{t=1}^T \frac{1}{\eta_{t-1}} \left(\|x_t - z_{t-1}\|^2 + \|x_{t-1} - z_{t-1}\|^2 \right)}_{\text{gain}} + O(1)$$

Using the above exercise, derive the final convergence rate for non-smooth functions.

Exercise 8. Suppose that f is non-smooth and the gradients are bounded: $\|\nabla f(x_t)\| \le G$ for all $t \in [T]$. Derive the appropriate convergence guarantee for the algorithm.

Next, we consider the setting where f is smooth. Our starting point is the result from the previous exercise:

$$\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle \le \underbrace{O(R)\sqrt{\sum_{t=1}^{T} \|\nabla f(x_t) - \nabla f(x_{t-1})\|^2}}_{\text{loss}} - \underbrace{\sum_{t=1}^{T} \frac{1}{\eta_{t-1}} \left(\|x_t - z_{t-1}\|^2 + \|x_{t-1} - z_{t-1}\|^2 \right)}_{\text{gain}}$$

As in the lecture, the goal is to use the gain to offset the loss, and achieve faster convergence.

Exercise 9. Use the smoothness of f to lower bound the terms appearing in the gain in terms of the terms appearing in the loss.

Exercise 10. Show that there is a transition point where the gain offsets the loss, and upper bound the net loss. Derive the appropriate convergence guarantee for the algorithm.