ADFOCS '21 Summer School: Adaptive Gradient Descent Algorithms

Alina Ene*

1 Lecture I: Adaptive Gradient Descent

We consider the convex minimization problem $\min_{x \in K} f(x)$ where $f \colon \mathbb{R}^d \to \mathbb{R}$ and $K \subseteq \mathbb{R}^d$ are convex. We assume for simplicity that f is differentiable. Throughout, we use the ℓ_2-norm to measure distances, and we denote it by $\|\cdot\|$:

$$\|x\| = \sqrt{\sum_{i=1}^{d} x_i^2}$$

The projection of a point $x \in \mathbb{R}^d$ onto K is the unique point in K that is closest to x, i.e.,

$$\Pi_K(x) = \arg\min_{u \in K} \|u - x\|^2$$

We assume the following oracle access to f and K:

- Gradient oracle for f: given a vector x, it returns the gradient $\nabla f(x)$.
- Projection oracle for K: given $x \in \mathbb{R}^d$, it returns $\Pi_K(x)$.

Throughout, we let $x^* \in \arg\min f(x)$ be an optimal solution, which we assume it exists and it has finite value.

1.1 AdaGrad algorithm for (essentially) unconstrained optimization

We first consider the more basic setting where the problem is essentially unconstrained. For reasons that will become clear later, we need to work with a feasible domain that has bounded diameter, which prevents us from directly working with $K = \mathbb{R}^d$. Thus we will consider the following setting, which is essentially unconstrained: we assume we are given a finite value R such that $\max_{x,y \in K} \|x - y\| \leq R$ and K contains a global minimum x^* (i.e., we have $x^* \in K$ and $\nabla f(x^*) = 0$). For example, K could be a ball of radius R that contains x^*.

AdaGrad algorithm We now present a version of gradient descent that adaptively sets the step size based on the information observed (the gradients). The algorithm is a scalar version of the celebraded AdaGrad algorithm [5, 10].

Algorithm 1 AdaGrad algorithm (scalar version) [5, 10].

Let $x_1 \in K$, $R \geq \max_{x,y \in K} \|x - y\|$
For $t = 1, \ldots, T$:

$$\eta_t = \frac{R}{\sqrt{\sum_{i=1}^{t} \|\nabla f(x_i)\|^2}}$$

$$x_{t+1} = \arg\min_{u \in K} \left\{ \langle \nabla f(x_t), u - x_t \rangle + \frac{1}{2\eta_t} \|u - x_t\|^2 \right\}$$

Return $\overline{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$

*Department of Computer Science, Boston University. aene@bu.edu

Next, we present an analysis of the algorithm in both the non-smooth settings, and show that it achieves convergence guarantees that are analogous to those of gradient descent. Crucially, the algorithm does so without knowing whether the function is smooth or not or the smoothness parameter. These properties are very useful in practice, and AdaGrad and other algorithms based on it are are some of the most popular optimization algorithms in deep learning and beyond. We also note that practical implementations of AdaGrad such as those in Tensorflow and Pytorch set $R = 1$, as $K = \mathbb{R}^d$ and finding a suitable upper bound R on the distance to the optimum requires additional tuning.

The analysis we present here is based on the work [7].

Analysis for non-smooth functions We first consider the setting where f is non-smooth. As in the standard gradient descent setting, we will need to assume that the gradients are bounded, i.e., we have $\|\nabla f(x)\| \leq G$ for all $x \in K$. This is satisfied, for example, when f is G-Lipschitz, i.e., we have $|f(x) - f(y)| \leq G \|x - y\|$ for all x, y.

We use convexity twice[1] and obtain:

$$f(\overline{x}_T) - f(x^*) \leq \frac{1}{T} \sum_{t=1}^{T} (f(x_t) - f(x^*)) \leq \frac{1}{T} \sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle \tag{1}$$

Next, we upper bound the inner product terms $\langle \nabla f(x_t), x_t - x^* \rangle$. To this end, we apply the first-order optimality condition[2] for x_{t+1}:

$$\left\langle \nabla f(x_t) + \frac{1}{\eta_t} (x_{t+1} - x_t), x_{t+1} - x^* \right\rangle \leq 0$$

Rearranging and using the identity $ab = \frac{1}{2}(a+b)^2 - \frac{1}{2}a^2 - \frac{1}{2}b^2$, we obtain

$$\begin{aligned} \langle \nabla f(x_t), x_{t+1} - x^* \rangle &\leq \frac{1}{\eta_t} \langle x_t - x_{t+1}, x_{t+1} - x^* \rangle \\ &= \frac{1}{2\eta_t} \left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 - \|x_t - x_{t+1}\|^2 \right) \end{aligned} \tag{2}$$

Note that the above bounds $\langle \nabla f(x_t), x_{t+1} - x^* \rangle$ whereas what we are interested in is actually $\langle \nabla f(x_t), x_t - x^* \rangle$. To address this discrepancy, we write

$$\begin{aligned} \langle \nabla f(x_t), x_t - x^* \rangle &= \langle \nabla f(x_t), x_{t+1} - x^* \rangle + \langle \nabla f(x_t), x_{t+1} - x_t \rangle \\ &\overset{(2)}{\leq} \frac{1}{2\eta_t} \|x_t - x^*\|^2 - \frac{1}{2\eta_t} \|x_{t+1} - x^*\|^2 + \langle \nabla f(x_t), x_{t+1} - x_t \rangle - \frac{1}{2\eta_t} \|x_t - x_{t+1}\|^2 \end{aligned} \tag{3}$$

We furter bound the last two terms above using the Cauchy-Schwartz inequality and the inequality $ab \leq \frac{\lambda}{2}a^2 + \frac{1}{2\lambda}b^2$

[1] The first inequality is via the definition of convexity:

$$f(\overline{x}_T) = f\left(\frac{1}{T} \sum_{t=1}^{T} x_t \right) \leq \frac{1}{T} \sum_{t=1}^{T} f(x_t)$$

The second inequality is via the first-order characterization of convexity:

$$f(x^*) \geq f(x_t) + \langle \nabla f(x_t), x^* - x_t \rangle \Rightarrow f(x_t) - f(x^*) \leq \langle \nabla f(x_t), x^* - x_t \rangle$$

[2] Recall that the first-order optimality condition for $x^* \in \arg\min_{x \in K} \phi(x)$ states that we have

$$\langle \nabla \phi(x^*), x - x^* \rangle \geq 0 \quad \forall x \in K$$

In the unconstrained setting $K = \mathbb{R}^d$, we recover the usual optimality condition $\nabla \phi(x^*) = 0$. In the constrained setting, we may have $\nabla \phi(x^*) \neq 0$ but none of the directions $x - x^*$ with $x \in K$ is a descent direction. Indeed, if we have $\langle \nabla \phi(x^*), x - x^* \rangle < 0$ for some $x \in K$, then we can make a small step $\eta \in (0, 1]$ in the direction $x - x^*$ and improve the objective value:

$$\phi(x^* + \eta(x - x^*)) \approx \phi(x^*) + \eta \langle \nabla \phi(x^*), x - x^* \rangle < \phi(x^*)$$

which holds for any $\lambda > 0$:[3]

$$\begin{aligned}\langle \nabla f(x_t), x_{t+1} - x_t \rangle - \frac{1}{2\eta_t} \|x_t - x_{t+1}\|^2 &\leq \|\nabla f(x_t)\| \|x_{t+1} - x_t\| - \frac{1}{2\eta_t} \|x_t - x_{t+1}\|^2 \\ &\leq \frac{\eta_t}{2} \|\nabla f(x_t)\|^2 + \frac{1}{2\eta_t} \|x_{t+1} - x_t\|^2 - \frac{1}{2\eta_t} \|x_t - x_{t+1}\|^2 \\ &= \frac{\eta_t}{2} \|\nabla f(x_t)\|^2 \end{aligned} \tag{4}$$

We plug in (4) into (3):

$$\langle \nabla f(x_t), x_t - x^* \rangle \leq \frac{1}{2\eta_t} \|x_t - x^*\|^2 - \frac{1}{2\eta_t} \|x_{t+1} - x^*\|^2 + \frac{\eta_t}{2} \|\nabla f(x_t)\|^2 \tag{5}$$

Summing up and collecting terms:

$$\begin{aligned}\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle &\leq \sum_{t=2}^{T} \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \underbrace{\|x_t - x^*\|^2}_{\leq R^2} + \frac{1}{2\eta_1} \underbrace{\|x_2 - x^*\|^2}_{\leq R^2} + \sum_{t=1}^{T} \frac{\eta_t}{2} \|\nabla f(x_t)\|^2 \\ &\leq \frac{R^2}{2\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2} \|\nabla f(x_t)\|^2 \end{aligned} \tag{6}$$

Above, we crucially relied on our assumption that K has bounded diameter in order to telescope the sums.

Finally, we analyze the two terms above. By the definition of the step sizes, we have

$$\frac{R^2}{\eta_T} = R \sqrt{\sum_{t=1}^{T} \|\nabla f(x_t)\|^2} \tag{7}$$

$$\sum_{t=1}^{T} \eta_t \|\nabla f(x_t)\|^2 = R \sum_{t=1}^{T} \frac{\|\nabla f(x_t)\|^2}{\sqrt{\sum_{i=1}^{t} \|\nabla f(x_i)\|^2}} \tag{8}$$

We can bound the last sum using the following result (we defer the proof of these inequalities to the exercises[4]). For any positive scalars $a_1, a_2, \ldots, a_n > 0$, we have

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

Thus we obtain

$$\sum_{t=1}^{T} \frac{\|\nabla f(x_t)\|^2}{\sqrt{\sum_{i=1}^{t} \|\nabla f(x_i)\|^2}} \leq 2 \sqrt{\sum_{t=1}^{T} \|\nabla f(x_t)\|^2} \tag{9}$$

Combining (6), (7), (8), (9):

$$\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle \leq \frac{3}{2} R \sqrt{\sum_{t=1}^{T} \|\nabla f(x_t)\|^2} \tag{10}$$

Note that the above result holds for any function f (either non-smooth or smooth). To complete the analysis for non-smooth functions, we use our assumption that the gradients are bounded ($\|\nabla f(x)\| \leq G$ for all x), and obtain

$$\sqrt{\sum_{t=1}^{T} \|\nabla f(x_t)\|^2} \leq G \sqrt{T} \tag{11}$$

[3] We can show the last inequality as follows. For $\lambda > 0$, we can write

$$ab = \left(\sqrt{\lambda} a \right) \left(\frac{1}{\sqrt{\lambda}} b \right) \leq \frac{1}{2} \left(\sqrt{\lambda} a \right)^2 + \frac{1}{2} \left(\frac{1}{\sqrt{\lambda}} b \right)^2 = \frac{1}{2} \lambda a^2 + \frac{1}{2\lambda} b^2$$

[4] The idea is to replace the sum by an integral. Think of $\frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}}$ as $\frac{dx}{\sqrt{x}}$ and recall that $\int \frac{dx}{\sqrt{x}} = \sqrt{x}$.

Plugging in (11) into (1) gives our final convergence guarantee:

$$f(\overline{x}_T) - f(x^*) \leq O\left(\frac{RG}{\sqrt{T}}\right)$$

Analysis for smooth functions If f is smooth, we can strengthen the above analysis and obtain an $\frac{1}{T}$ convergence. We will use the following result which we will prove in the exercises. Let f be a convex function that is β-smooth, i.e., we have $\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$ for all x, y. We have

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

Note that the above can be viewed as a stronger version of the inequality we obtain from convexity:

$$\underbrace{f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle}_{\text{convexity inequality}} + \underbrace{\frac{1}{2\beta} \|\nabla f(y) - \nabla f(x)\|^2}_{\text{extra term (from smoothness)}}$$

We can use the above result to strengthen (1) and gain an extra term proportional to the gradient. Setting $y = x_t$ and $x = x^*$ in the above inequality and using that $\nabla f(x^*) = 0$, we obtain

$$f(x_t) - f(x^*) \leq \langle \nabla f(x_t), x_t - x^* \rangle - \frac{1}{2\beta} \|\nabla f(x_t)\|^2$$

Thus

$$f(\overline{x}_T) - f(x^*) \leq \frac{1}{T} \sum_{t=1}^{T} (f(x_t) - f(x^*)) \leq \frac{1}{T} \left(\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle - \underbrace{\frac{1}{2\beta} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2}_{\text{gain}} \right) \tag{12}$$

As we have shown in (10), we have

$$\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle \leq \frac{3}{2} R \underbrace{\sqrt{\sum_{t=1}^{T} \|\nabla f(x_t)\|^2}}_{\text{loss}}$$

Crucially, we can use the gain term to cancel most of the loss. Indeed, letting $z = \sqrt{\sum_{t=1}^{T} \|\nabla f(x_t)\|^2}$, note that the gain is proportional to z^2 whereas the loss is proportional to z. Thus, once z grows large enough, the gain will overpower the loss. More precisely,

$$\begin{aligned}
\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle - \frac{1}{2\beta} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 &\leq \frac{3}{2} R \sqrt{\sum_{t=1}^{T} \|\nabla f(x_t)\|^2} - \frac{1}{2\beta} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \\
&\leq \max_{z \geq 0} \left\{ \frac{3}{2} R z - \frac{1}{2\beta} z^2 \right\} \\
&= O\left(\beta R^2\right)
\end{aligned} \tag{13}$$

On the last line, we used the fact that $\phi(z) = az - bz^2$ with $a, b > 0$ is concave and it is maximized at $z^* = \frac{a^2}{4b}$.

Plugging in (13) into (12) gives our final convergence guarantee:

$$f(\overline{x}_T) - f(x^*) \leq O\left(\frac{\beta R^2}{T}\right)$$

1.2 AdaGrad+ algorithm for constrained optimization

The algorithm and the analysis above strongly relied on the fact that we had $\nabla f(x^*) = 0$ at a point x^* in the feasible domain, i.e., the domain contained a global minimum. For general constraints, we have $\nabla f(x^*) \neq 0$, and we need a different choice of step sizes. The reason for this is that the norm of the gradients $\|\nabla f(x_t)\|$ does not necessarily go down as we approach the optimum. Thus the step sizes $\eta_t \propto \frac{1}{\sqrt{\sum_{i=1}^t \|\nabla f(x_t)\|^2}}$ may become too small, which is problematic in the smooth setting where we want the steps to be constant.

One approach we can take here is to set the step sizes based on the iterate movement $\|x_t - x_{t-1}\|^2$ instead; whereas the gradient may not decrease, we intuitively expect that the movement does decrease as we approach the optimum. This brings us to the following extension of AdaGrad to the constrained setting:

Algorithm 2 AdaGrad+ algorithm (scalar version) [7].

Let $x_1 \in K$, $\eta_1 > 0$, $R \geq \max_{x,y \in K} \|x - y\|$
For $t = 1, \ldots, T$:

$$x_{t+1} = \arg\min_{u \in K} \left\{ \langle \nabla f(x_t), u - x_t \rangle + \frac{1}{2\eta_t} \|u - x_t\|^2 \right\}$$
$$\frac{1}{\eta_{t+1}^2} = \frac{1}{\eta_t^2} \left(1 + \frac{\|x_{t+1} - x_t\|^2}{R^2} \right)$$

Return $\overline{x}_T = \frac{1}{T} \sum_{t=1}^T x_t$

The main update scales the movement $\|x_{t+1} - x_t\|^2$ by R^2 to ensure that the step sizes do not decrease too fast, i.e., we have $\eta_{t+1} = \Theta(\eta_t)$.

By unrolling the recurrence in the step size update, we can see that the step sizes have the following closed-form:

$$\eta_t = \frac{R}{\sqrt{\frac{R^2}{\eta_1^2} + \sum_{i=1}^{t-1} \frac{\|x_{i+1} - x_i\|^2}{\eta_i^2}}}$$

In the unconstrained setting $K = \mathbb{R}^d$, we have $x_{i+1} = x_i - \eta_i \nabla f(x_i)$ and thus the steps become

$$\eta_t = \frac{R}{\sqrt{\frac{R^2}{\eta_1^2} + \sum_{i=1}^{t-1} \|\nabla f(x_i)\|^2}}$$

which closely mirror the AdaGrad step sizes, with the following two differences: the additional term $\frac{R^2}{\eta_1^2}$ and the fact that the step is "off-by-one," i.e., it does not include the latest gradient $\|\nabla f(x_t)\|^2$. The additional term can be made as small as we would like and it is beneficial to have for numerical stability. Practical implementations of AdaGrad such as those in Tensorflow and Pytorch add a small additional term, typically set to 10^{-10}. The off-by-one iterate is unavoidable in the constrained setting, since knowing the latest iterate movement $\|x_{t+1} - x_t\|^2$ requires first computing x_{t+1}. The off-by-one iterate introduces additional complications in the analysis, and the convergence bounds shown in [7] have additional logarithmic factors compared to the analysis above (the non-smooth bound is sub-optimal by a $\sqrt{\ln T}$ factor and the smooth bound by a $\ln \beta$ factor). We refer the interested reader to [7] for the analysis.

In Lecture III, we will see a different approach for setting the step sizes that removes these difficulties and achieves optimal convergence.

2 Lecture II: Adaptive Accelerated Gradient Descent

In the previous lecture, we introduced the AdaGrad algorithm that achieves the optimal $\frac{1}{\sqrt{T}}$ convergence in the non-smooth setting and the $\frac{1}{T}$ convergence in the smooth setting, analogously to standard gradient descent. The $\frac{1}{T}$ convergence is not optimal for smooth functions, and several algorithms have been developed that achieve a faster $\frac{1}{T^2}$ convergence, which is optimal. The first such algorithm was developed by Nesterov, called accelerated gradient descent (AGD). Many different variants of AGD have been developed since (we refer the reader to, e.g., [11, 2]). In

this lecture, we will see one such variant (AGD+) due to [8, 3] and an adaptive version of AGD+ (AdaAGD+) due to [7]. The analysis we present here is based on the works [3, 7].

As before, we consider the problem $\min_{x \in K} f(x)$. Here we focus on the case where f is β-smooth. We note that, analogously to AdaGrad, AdaAGD+ automatically adapts to the problem structure. The interested reader can find an analysis of AdaAGD+ for non-smooth functions in [7].

Acceleration via better lower bounds In the primer lectures, we saw that the main driving forces behind the standard gradient descent algorithm and its analysis are suitable upper and lower bounds on the function. Indeed, if x_t is the current iterate, smoothness gives us a quadratic upper bound:

$$f(x) \leq \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\beta} \|x - x_t\|^2}_{\text{quadratic function of } x} \quad \forall x$$

The upper bound allows us to make large steps without overshooting: if we let x_{t+1} be the minimizer of the upper bound, we are guaranteed that $f(x_{t+1}) \leq f(x_t)$.

Convexity gives us an affine lower bound on the function:

$$f(x) \geq \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle}_{\text{affine function of } x} \quad \forall x$$

This affine lower bound is exploited both in the algorithm and the analysis. In the algorithm, we use it as part of the update:

$$x_{t+1} = \arg\min_{x \in K} \left\{ \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle}_{\text{affine lower bound}} + \frac{1}{2\eta_t} \underbrace{\|x - x_t\|^2}_{\text{proximity term}} \right\}$$

In the analysis, we apply the lower bound with $x = x^*$ and obtain an upper bound on the sub-optimality of the current iterate.

In the basic gradient descent algorithm, we always use the current iterate to construct our upper and lower bounds. In this lecture, we will depart from this approach and aim to construct much better lower bounds on the function. Indeed, suppose we have queried the gradient at the points $x_1, \ldots, x_t$. Thus we have learned the following lower bounds on the optimum via convexity:

$$f(x^*) \geq f(x_i) + \langle \nabla f(x_i), x^* - x_i \rangle \quad \forall i \in [t] \tag{14}$$

The maximum of these lower bounds gives us the maximum amount of information about $f(x^*)$. However, this is a complicated piece-wise affine function that is difficult to work with. Instead, we can take a convex combination of these affine lower bounds, which is simpler than the maximum since it is affine. More precisely, let us choose any non-negative weights $a_1, \ldots, a_t \geq 0$. Let $A_t = \sum_{i=1}^t a_i$. By combining the inequalities (14) using weights $a_1, \ldots, a_t \geq 0$, we obtain:

$$f(x^*) \geq \frac{1}{A_t} \sum_{i=1}^t \left(a_i f(x_i) + a_i \langle \nabla f(x_i), x^* - x_i \rangle \right)$$

Following the gradient descent paradigm, our main update will minimize the lower bound plus a proximity term. A particularly simple choice of proximity term is given by the distance to the *initial* solution, which we will adopt here. Finally, we make one more key modification to the GD approach: we will separate the sequence $\{x_t\}$ at which we query gradients from the sequence that we will use to obtain our solutions, which we will denote by $\{z_t\}$. Thus our approach so far can be summarized as follows. We will design a sequence $\{x_t\}$ at which to query the gradients. Using the queried gradients, we update a sequence $\{z_t\}$ which minimizes the lower bound plus the distance to the initial solution:

$$\begin{aligned} z_t &= \arg\min_{x \in K} \left\{ \frac{1}{A_t} \sum_{i=1}^t \left(a_i f(x_i) + a_i \langle \nabla f(x_i), x - x_i \rangle + \frac{1}{2\eta_t} \|x - z_0\|^2 \right) \right\} \\ &= \arg\min_{x \in K} \left\{ \sum_{i=1}^t a_i \langle \nabla f(x_i), x \rangle + \frac{1}{2\eta_t} \|x - z_0\|^2 \right\} \end{aligned}$$

Following standard GD, we will output a convex combination of $\{z_t\}$. A natural choice is to use the convex combination given by the weights $\{a_t\}$, i.e., our output is

$$\overline{z}_T := \sum_{t=1}^{T} \frac{a_t}{A_T} z_t$$

Let us now discuss how to choose $\{x_t\}$. Naturally, x_t should be set based on the z iterates that we have observed so far, namely $z_1, \ldots, z_{t-1}$. Once again, we use a convex combination based on the weights $\{a_t\}$, but with a twist:

$$x_t = \frac{\sum_{i=1}^{t-1} a_i z_i + a_t z_{t-1}}{A_t}$$

Ideally, we would want x_t to be a convex combination of $z_1, \ldots, z_t$, but of course we have a chicken-and-egg problem since we need x_t to compute z_t. The workaround is to use z_{t-1} as a proxy for z_t. The intuition for this is that the proximity term encourages the iterates not to move too much, and thus we hope that z_t is close to z_{t-1}. The choice of x_t above will also follow organically from the analysis, as we will see.

The right choice of weights $\{a_t\}$ will also follow organically from the analysis. Thus it only remains to discuss how to choose the step sizes $\{\eta_t\}$. In the non-adaptive setting, we follow the gradient descent approach and set $\eta_t = \frac{1}{\beta}$.

Putting everything together, we have the following algorithm:

Algorithm 3 AGD+ [8, 4].

Let $z_0 \in K$, $\eta_1 > 0$, $a_t = \frac{t}{2}$, $A_t = \sum_{i=1}^{t} a_i = \frac{t(t+1)}{4}$, $R \geq \max_{x,y \in K} \|x - y\|$.
For $t = 1, \ldots, T$:

$$\begin{aligned}
\eta_t &= \frac{1}{\beta} \\
x_t &= \frac{\sum_{i=1}^{t-1} a_i z_i + a_t z_{t-1}}{A_t} \\
z_t &= \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \langle \nabla f(x_i), x - x_i \rangle + \frac{1}{2\eta_t} \|x - z_0\|^2 \right\}
\end{aligned}$$

Return $\overline{z}_T := \sum_{t=1}^{T} \frac{a_t}{A_T} z_t$

Analysis Let

$$\overline{z}_t := \sum_{i=1}^{t} \frac{a_i}{A_t} z_i$$

$\{\overline{z}_t\}$ is our main solution sequence, and thus we are interested in upper bounding the optimality gap $f(\overline{z}_t) - f(x^*)$. To this end, we need to lower bound $f(x^*)$. We follow the approach described above and lower bound $f(x^*)$ as follows. Let $x_1, \ldots, x_t$ be the query points we have chosen so far. By convexity, we have

$$f(x^*) \geq f(x_i) + \langle \nabla f(x_i), x^* - x_i \rangle \quad \forall i \in [t]$$

We combine these lower bounds with weights $a_i \geq 0$ and obtain

$$A_t f(x^*) \geq \sum_{i=1}^{t} a_i f(x_i) + \sum_{i=1}^{t} a_i \langle \nabla f(x_i), x^* - x_i \rangle$$

Now, a crucial step is to connect the RHS above to the main update:

$$\begin{aligned}
z_t &= \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \langle \nabla f(x_i), x \rangle + \frac{1}{2\eta_t} \|x - z_0\|^2 \right\} \\
&= \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \langle \nabla f(x_i), x - x_i \rangle + \frac{1}{2\eta_t} \|x - z_0\|^2 \right\}
\end{aligned}$$

We can do so as follows:

$$\begin{aligned}
A_t f(x^*) &\geq \sum_{i=1}^{t} a_i f(x_i) + \sum_{i=1}^{t} a_i \langle \nabla f(x_i), x^* - x_i \rangle \\
&= \sum_{i=1}^{t} a_i f(x_i) - \frac{1}{2\eta_t} \|x^* - z_0\|^2 + \sum_{i=1}^{t} a_i \langle \nabla f(x_i), x^* - x_i \rangle + \frac{1}{2\eta_t} \|x^* - z_0\|^2 \\
&\geq \sum_{i=1}^{t} a_i f(x_i) - \frac{1}{2\eta_t} \|x^* - z_0\|^2 + \min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \langle \nabla f(x_i), x - x_i \rangle + \frac{1}{2\eta_t} \|x - z_0\|^2 \right\} \\
&= \sum_{i=1}^{t} a_i f(x_i) - \frac{1}{2\eta_t} \|x^* - z_0\|^2 + \sum_{i=1}^{t} a_i \langle \nabla f(x_i), z_t - x_i \rangle + \frac{1}{2\eta_t} \|z_t - z_0\|^2
\end{aligned}$$

Thus we have obtained the following lower bound on $f(x^*)$:

$$f(x^*) \geq \underbrace{\frac{1}{A_t} \left(\sum_{i=1}^{t} a_i f(x_i) - \frac{1}{2\eta_t} \|x^* - z_0\|^2 + \sum_{i=1}^{t} a_i \langle \nabla f(x_i), z_t - x_i \rangle + \frac{1}{2\eta_t} \|z_t - z_0\|^2 \right)}_{:= L_t}$$

and thus

$$f(\overline{z}_t) - f(x^*) \leq f(\overline{z}_t) - L_t$$

Thus we can focus on upper bounding $f(\overline{z}_t) - L_t$. To this end, we analyze how the lower bounds are evolving. We have

$$\begin{aligned}
A_{t-1} L_{t-1} - A_t L_t &= -a_t f(x_t) - a_t \langle \nabla f(x_t), z_t - x_t \rangle \\
&\quad + \sum_{i=1}^{t-1} a_i \langle \nabla f(x_i), z_{t-1} - z_t \rangle \\
&\quad + \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|x^* - z_0\|^2 \\
&\quad + \frac{1}{2\eta_{t-1}} \|z_{t-1} - z_0\|^2 - \frac{1}{2\eta_t} \|z_t - z_0\|^2
\end{aligned} \tag{15}$$

By looking at the term $\sum_{i=1}^{t-1} a_i \langle \nabla f(x_i), z_{t-1} - z_t \rangle$, we see an opportunity to leverage the optimality condition:

$$z_{t-1} = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t-1} a_i \langle \nabla f(x_i), x \rangle + \frac{1}{2\eta_{t-1}} \|x - z_0\|^2 \right\}$$

The optimality condition for z_{t-1} (applied with $x = z_t$) gives us

$$\left\langle \sum_{i=1}^{t-1} a_i \nabla f(x_i) + \frac{1}{\eta_{t-1}} (z_{t-1} - z_0), z_{t-1} - z_t \right\rangle \leq 0$$

By rearranging and using the identity $ab = \frac{1}{2}(a+b)^2 - \frac{1}{2}a^2 - \frac{1}{2}b^2$, we obtain

$$\begin{aligned}
\sum_{i=1}^{t-1} a_i \langle \nabla f(x_i), z_{t-1} - z_t \rangle &\leq \frac{1}{\eta_{t-1}} \langle z_0 - z_{t-1}, z_{t-1} - z_t \rangle \\
&= \frac{1}{2\eta_{t-1}} \left(\|z_0 - z_t\|^2 - \|z_0 - z_{t-1}\|^2 - \|z_{t-1} - z_t\|^2 \right)
\end{aligned} \tag{16}$$

Plugging in (16) into (15), we obtain

$$\begin{aligned}
A_{t-1}L_{t-1} - A_tL_t &\leq -a_t f(x_t) - a_t \langle \nabla f(x_t), z_t - x_t \rangle \\
&+ \frac{1}{2\eta_{t-1}} \|z_0 - z_t\|^2 - \frac{1}{2\eta_{t-1}} \|z_0 - z_{t-1}\|^2 - \frac{1}{2\eta_{t-1}} \|z_{t-1} - z_t\|^2 \\
&+ \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|x^* - z_0\|^2 \\
&+ \frac{1}{2\eta_{t-1}} \|z_{t-1} - z_0\|^2 - \frac{1}{2\eta_t} \|z_t - z_0\|^2 \\
&= -a_t f(x_t) - a_t \langle \nabla f(x_t), z_t - x_t \rangle \\
&+ \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|x^* - z_0\|^2 - \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|z_0 - z_t\|^2 - \frac{1}{2\eta_{t-1}} \|z_{t-1} - z_t\|^2 \qquad (17)
\end{aligned}$$

Using the above inequality, we can now analyze how much $A_t \left(f(\overline{z}_t) - L_t \right)$ changed in the current iteration. We have

$$\begin{aligned}
&A_t \left(f(\overline{z}_t) - L_t \right) - A_{t-1} \left(f(\overline{z}_{t-1}) - L_{t-1} \right) \\
&= A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) + A_{t-1}L_{t-1} - A_tL_t \\
&\overset{(17)}{\leq} A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) - a_t f(x_t) - a_t \langle \nabla f(x_t), z_t - x_t \rangle \\
&+ \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|x^* - z_0\|^2 - \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|z_0 - z_t\|^2 - \frac{1}{2\eta_{t-1}} \|z_{t-1} - z_t\|^2 \qquad (18)
\end{aligned}$$

We further bound the first terms above as follows:

$$\begin{aligned}
&A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) - a_t f(x_t) \\
&= A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) - (A_t - A_{t-1}) f(x_t) \\
&= A_t \underbrace{\left(f(\overline{z}_t) - f(x_t) \right)}_{\text{smoothness}} - A_{t-1} \underbrace{\left(f(\overline{z}_{t-1}) - f(x_t) \right)}_{\text{convexity}} \\
&\leq A_t \left(\langle \nabla f(x_t), \overline{z}_t - x_t \rangle + \frac{\beta}{2} \|\overline{z}_t - x_t\|^2 \right) - A_{t-1} \langle \nabla f(x_t), \overline{z}_{t-1} - x_t \rangle \qquad (19)
\end{aligned}$$

We plug in (19) into (18) and collect terms:

$$\begin{aligned}
&A_t \left(f(\overline{z}_t) - L_t \right) - A_{t-1} \left(f(\overline{z}_{t-1}) - L_{t-1} \right) \\
&\leq \langle \nabla f(x_t), A_t \left(\overline{z}_t - x_t \right) - A_{t-1} \left(\overline{z}_{t-1} - x_t \right) - a_t \left(z_t - x_t \right) \rangle + A_t \frac{\beta}{2} \|\overline{z}_t - x_t\|^2 \\
&+ \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|x^* - z_0\|^2 - \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|z_0 - z_t\|^2 - \frac{1}{2\eta_{t-1}} \|z_{t-1} - z_t\|^2 \\
&= \left\langle \nabla f(x_t), \underbrace{A_t \overline{z}_t - A_{t-1} \overline{z}_{t-1} - a_t z_t}_{=0} \right\rangle + A_t \frac{\beta}{2} \|y_t - x_t\|^2 \\
&+ \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|x^* - z_0\|^2 - \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|z_0 - z_t\|^2 - \frac{1}{2\eta_{t-1}} \|z_{t-1} - z_t\|^2 \\
&= A_t \frac{\beta}{2} \|\overline{z}_t - x_t\|^2 + \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|x^* - z_0\|^2 - \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|z_0 - z_t\|^2 - \frac{1}{2\eta_{t-1}} \|z_{t-1} - z_t\|^2 \qquad (20)
\end{aligned}$$

Let us now inspect the distance terms above:

$$\underbrace{A_t \frac{\beta}{2} \|\overline{z}_t - x_t\|^2 - \frac{1}{2\eta_{t-1}} \|z_{t-1} - z_t\|^2}_{\text{want it to be small}} + \underbrace{\left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|x^* - z_0\|^2}_{\text{telescopes}} \underbrace{- \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|z_0 - z_t\|^2}_{\leq 0}$$

The ideal scenario is when the second distance term $\frac{1}{2\eta_{t-1}} \|z_{t-1} - z_t\|^2$ is at least as big as the first distance term $A_t \frac{\beta}{2} \|\overline{z}_t - x_t\|^2$. A moment's thought reveales that we can choose x_t to try to make this happen. Consider setting

x_t so that $\overline{z}_t - x_t$ is a scalar multiple of $z_t - z_{t-1}$:

$$\begin{aligned}
\overline{z}_t - x_t &= \lambda_t \left(z_t - z_{t-1}\right) \\
\Rightarrow x_t &= \overline{z}_t + \lambda_t \left(z_{t-1} - z_t\right) \\
&= \frac{A_{t-1}\overline{z}_{t-1} + a_t z_t}{A_t} + \lambda_t \left(z_{t-1} - z_t\right) \\
&= \frac{A_{t-1}\overline{z}_{t-1} + A_t\lambda_t z_{t-1} + \left(a_t - A_t\lambda_t\right) z_t}{A_t}
\end{aligned}$$

We need to compute x_t without access to z_t, so we need to set

$$a_t - A_t\lambda_t = 0 \Rightarrow \lambda_t = \frac{a_t}{A_t}$$

Thus we have arrived at the following choice for x_t:

$$x_t = \frac{A_{t-1}\overline{z}_{t-1} + a_t z_{t-1}}{A_t} = \frac{\sum_{i=1}^{t-1} a_i z_i + a_t z_{t-1}}{A_t}$$

Therefore we obtain

$$\overline{z}_t - x_t = \frac{a_t}{A_t}\left(z_t - z_{t-1}\right) \tag{21}$$

Plugging (21) into (20), we obtain

$$\begin{aligned}
&A_t\left(f(\overline{z}_t) - L_t\right) - A_{t-1}\left(f(\overline{z}_{t-1}) - L_{t-1}\right) \\
&\leq \left(\frac{\beta}{2}\frac{a_t^2}{A_t} - \frac{1}{2\eta_{t-1}}\right)\left\|z_t - z_{t-1}\right\|^2 + \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}}\right)\left\|x^* - z_0\right\|^2 - \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}}\right)\left\|z_0 - z_t\right\|^2
\end{aligned} \tag{22}$$

Plugging in $\eta_t = \frac{1}{\beta}$ into (22), we obtain

$$A_t\left(f(\overline{z}_t) - L_t\right) - A_{t-1}\left(f(\overline{z}_{t-1}) - L_{t-1}\right) \leq \frac{\beta}{2}\left(\frac{a_t^2}{A_t} - 1\right)\left\|z_t - z_{t-1}\right\|^2 \tag{23}$$

Naturally, the ideal scenario is when the RHS is non-positive. A moment's thought reveals that we can choose the weights $\{a_t\}$ to make this happen. Indeed, we need

$$\frac{a_t^2}{A_t} - 1 \leq 0 \Leftrightarrow a_t^2 \leq \sum_{i=1}^{t} a_i$$

Since $\sum_{i=1}^{t} i = \Theta(t^2)$, a simple choice is to set $a_t = ct$ for some constant c. We have $A_t = c\sum_{i=1}^{t} i = c\frac{t(t+1)}{2}$. Thus the inequality $c^2t^2 \leq c\frac{t(t+1)}{2}$ holds if we set $c = \frac{1}{2}$. This gives the choice $a_t = \frac{t}{2}$ that we stated in the algorithm.[5]

Plugging in our choice of $a_t = \frac{t}{2}$ into (23), we obtain

$$A_t\left(f(\overline{z}_t) - L_t\right) - A_{t-1}\left(f(\overline{z}_{t-1}) - L_{t-1}\right) \leq 0 \tag{24}$$

Summing up over all iterations and using that $a_1 = A_1 = \frac{1}{2}$ and $x_1 = z_0$, we obtain

$$\begin{aligned}
A_T\left(f(\overline{z}_T) - L_T\right) &\leq A_1\left(f(z_1) - L_1\right) \\
&= \frac{1}{2}\left(\beta\left\|x^* - z_0\right\|^2 + \underbrace{f(z_1) - f(x_1) - \left\langle\nabla f(x_1), z_1 - x_1\right\rangle}_{\leq \frac{\beta}{2}\|z_1 - x_1\|^2 \text{ by smoothness}} - \beta\left\|z_1 - x_1\right\|^2\right) \\
&\leq \frac{\beta\left\|x^* - z_0\right\|^2}{2}
\end{aligned}$$

Finally, since $A_T = \Theta(T^2)$, we have obtained the following convergence guarantee:

$$f(\overline{z}_T) - f(x^*) \leq f(\overline{z}_T) - L_T \leq \Theta\left(\frac{\beta\left\|x^* - z_0\right\|^2}{T^2}\right)$$

[5]Another choice is the following. Instead of making the coefficient non-positive, we can try to make it equal to 0. Thus we want that $a_t^2 = \sum_{i=1}^{t} a_i$ for all $t \geq 1$. Rearranging, we want $a_t^2 - a_t - a_{t-1}^2 = 0$, which is a quadratic equation in a_t. Solving the quadratic equation and picking the positive solution, we obtain the recurrence $a_t = \frac{1}{2}\left(1 + \sqrt{4a_{t-1}^2 + 1}\right)$ for all $t \geq 1$ and $a_0 = 0$.

Adaptive step sizes Similarly to Adagrad, we can make the algorithm adaptive by setting the step sizes proportional to the movement of the main iterate sequence $\{z_t\}$. We can show that the algorithm convergences at the rate $O\left(\frac{R^2\beta\ln\beta}{T^2}\right)$ [7].

Algorithm 4 AdaAGD+ (scalar version) [7] algorithms.

Let $z_0 \in K$, $\eta_1 > 0$, $a_t = \frac{t}{2}$, $A_t = \sum_{i=1}^t a_i = \frac{t(t+1)}{4}$, $R \geq \max_{x,y\in K} \|x-y\|$.
For $t = 1, \ldots, T$:

$$\frac{1}{\eta_t^2} = \frac{1}{\eta_{t-1}^2}\left(1 + \frac{\|z_{t-1} - z_{t-2}\|^2}{R^2}\right) \quad \forall t \geq 2$$
$$x_t = \frac{\sum_{i=1}^{t-1} a_i z_i + a_t z_{t-1}}{A_t}$$
$$z_t = \arg\min_{x\in K}\left\{\sum_{i=1}^t a_i \langle \nabla f(x_i), x - x_i\rangle + \frac{1}{2\eta_t}\|x - z_0\|^2\right\}$$

Return $\overline{z}_T := \sum_{t=1}^T \frac{a_t}{A_T} z_t$

3 Lecture III: Adaptive Extra-Gradient for Variational Inequalities

Variational (VI) are a general framework that capture several optimization of interest. Two important examples of optimization problems that fit into this framework are the problem of minimizing a function on which we have focused so far and the problem of finding a Nash equilibrium in a 2-player game.

Here we have a feasible domain $K \subseteq \mathbb{R}^d$ and a vector-valued function $F\colon K \to \mathbb{R}^d$; F is often referred to as an operator. The variational inequality problem asks for a strong solution, which is a point $x^* \in K$ satisfying

$$\langle F(x^*), x^* - x\rangle \leq 0 \quad \forall x \in K \tag{25}$$

The above condition may remind the reader of the optimality condition we used extensively in our study of convex minimization. Indeed, this is not a coincidence. If we let $F(x) = \nabla f(x)$, we can see that the variational inequality problem captures the problem $\min_{x\in K} f(x)$.

The operator analogue of convexity is monotonicity. An operator F is *monotone* if it satisfies

$$\langle F(x) - F(y), x - y\rangle \geq 0 \quad \forall x, y \in K \tag{26}$$

In the exercises, we will show that f is convex if and only if ∇f is a monotone operator.

The operator analogue of smoothness is Lipschitzness. An operator F is β-Lipschitz if it satisfies

$$\|F(x) - F(y)\| \leq \beta\|x - y\| \quad \forall x, y \in K \tag{27}$$

In this lecture, we consider the problem of finding a strong solution to a VI problem. We will assume throughout that F is a monotone operator and K is a convex set with finite diameter.

2-player games and min-max optimization We have seen above that convex minimization is a special case of monotone VI. We now discuss another important class of optimization problems that are special cases of monotone VI: finding Nash equilibria of 2-player zero-sum games and more generally, min-max optimization. In 2-player zero-sum games, we have two players, Alice and Bob, that are playing a game such as rock-paper-scissors. In this game, each player simultaneously selects a a strategy to play: either rock, paper, or scissors. Rock beats scissors, scissors beats paper, and paper beats rock. We can give this game an optimization perspective by writing down a payoff matrix $\mathbb{A} \in \mathbb{R}^{3\times 3}$ that encodes whether Alice wins or not. The rows of $\mathbb{A}$ correspond to Alice's strategies (rock, paper, or scissors) and the columns of $\mathbb{A}$ correspond to Bob's strategies (rock, paper, or scissors). The entries of $\mathbb{A}$ associate a payoff (e.g., the loser pays the winner \$1):

$$\mathbb{A} = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}$$

Bob's payoff matrix is $-\mathbb{A}$, and the game is called zero-sum since the players' payoffs sum up to 0. We will allow the players to use randomness to choose their strategies, i.e., each player chooses a probability distribution $p = (p_1, p_2, p_3) \in \Delta_3$ over the strategies (these are called mixed strategies, and the deterministic strategies such as picking rock ($p = (1, 0, 0)$) are called pure strategies). If Alice chooses $p \in \Delta_3$ and Bob chooses $q \in \Delta_3$, the expected payoff to Alice is $f(p, q) := p^\top \mathbb{A} q$. Alice wants to maximize this expected payoff, whereas Bob wants to minimize it since it is a zero-sum game. If the players play optimally, the players can even play sequentially and it does not matter which player goes first. In other words, we have the following fundamental result (the minimax theorem, due to Von Neumann, which follows from LP duality):

$$\max_{p \in \Delta_3} \min_{q \in \Delta_3} f(p, q) = \min_{q \in \Delta_3} \max_{p \in \Delta_3} f(p, q)$$

where $\max_{p \in \Delta_3} \min_{q \in \Delta_3} f(p, q)$ means that Bob goes first and Alice goes second.

An important solution concept for zero-sum games is a Nash equilibrium. A pair of mixed strategies (p^*, q^*) is a mixed Nash equilibrium if neither player can strictly improve their expected payoff by switching to a different strategy given that the other player's strategy remains the same., i.e., we have

$$f(p, q^*) \leq f(p^*, q^*) \leq f(p^*, q) \quad \forall p, q \in \Delta_3$$

Another fundamental result, due to Nash, states that a mixed Nash equilibrium always exists (the interested reader can derive a proof of this result using the minimax theorem above).

The problem of finding a mixed Nash equilibrium is a special case of the variational inequality problem. The corresponding operator $F \colon \mathbb{R}^6 \to \mathbb{R}^6$ is given by

$$F((p, q)) = (-\nabla_p f(p, q), \nabla_q f(p, q)) = (-\mathbb{A} q, \mathbb{A} p)$$

The reader can readily verify that a strong solution (p^*, q^*) to the resulting variational inequality is a mixed Nash equilibrium.

More generally, we can consider the min-max optimization problem $\min_{u \in U} \max_{v \in V} f(u, v)$ where $f(u, v)$ is convex in u and concave in v. Analogously to the games setting, the corresponding monotone operator is $F((u, v)) = (\nabla_u f(u, v), -\nabla_v f(u, v))$.

We now turn our attention to designing algorithms for solving monotone VIs.

Gradient descent does not converge As we have seen, for convex minimization, the operator corresponds to the gradient. Thus it is very natural to consider the gradient descent approach for variational inequalities as well:

$$x_{t+1} = \arg\min_{x \in K} \left\{ \langle F(x_t), x - x_t \rangle + \frac{1}{2\eta} \|x - x_t\|^2 \right\}$$

A moment's thought reveals that, unfortunately, the above algorithm fails to converge even for very simple min-max optimization problems. Indeed, consider the problem $\min_{u \in \mathbb{R}^d} \max_{v \in \mathbb{R}^d} f(u, v)$ where $f(u, v) = u^\top v$, i.e., a 2-player game objective with identity payoff matrix. The Nash equilibrium is $(0, 0)$. The operator is $F((u, v)) = (v, -u)$ and it is 1-Lipschitz. The gradient descent iterates $x_t = (u_t, v_t)$ evolve as follows:

$$\begin{pmatrix} u_{t+1} \\ v_{t+1} \end{pmatrix} = \begin{pmatrix} u_t - \eta v_t \\ v_t + \eta u_t \end{pmatrix}$$

As we can observe from Figure 3, the iterates do not converge to the Nash equilibrium $(0, 0)$.

Extra-gradient algorithm We now introduce a different approach for solving variational inequalities, shown in the following algorithm. In the non-adaptive setting, we choose the step sizes similarly to gradient descent.

Algorithm 5 Extra-gradient algorithm [9].

Let $z_0 \in K$.
For $t = 1, \ldots, T$, update:

$$x_t = \arg\min_{u \in K} \left\{ \langle F(z_{t-1}), u \rangle + \frac{1}{2\eta_t} \|u - z_{t-1}\|^2 \right\}$$

$$z_t = \arg\min_{u \in K} \left\{ \langle F(x_t), u \rangle + \frac{1}{2\eta_t} \|u - z_{t-1}\|^2 \right\}$$

Return $\overline{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$.

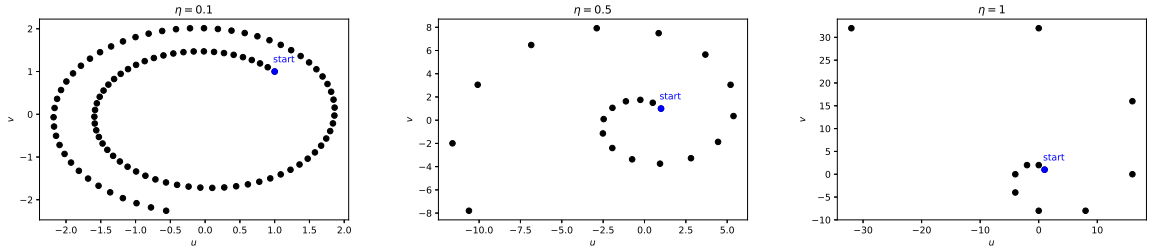

Figure 1: The gradient descent iterates for $\min_{u\in\mathbb{R}} \max_{v\in\mathbb{R}} uv$. The initial solution is $(u_0, v_0) = (1, 1)$.

To gain a bit more insight into the algorithm, let us consider our more familiar unconstrained convex minimization problem: $K = \mathbb{R}^d$ and $F = \nabla f$. We have

$$\begin{aligned} x_{t+1} &= z_t - \eta_t \nabla f(z_t) \\ z_{t+1} &= z_t - \eta_t \nabla f(x_{t+1}) \end{aligned}$$

Thus the algorithm is looking ahead to see what the gradient looks like after performing the GD update from z_t, and uses this extrapolated gradient $\nabla f(z_t - \eta_t \nabla f(z_t))$ instead of the current gradient $\nabla f(z_t)$.

Analysis of extra-gradient We will analyze the convergence of the algorithm via the error function, which is defined as follows.

$$\mathrm{Err}(x) := \sup_{y\in K} \langle F(y), x - y\rangle \tag{28}$$

The error function measures convergence to a weak solution, i.e., a point $x \in K$ satisfying

$$\langle F(y), x - y\rangle \le 0 \quad \forall y \in K \tag{29}$$

If F is monotone and continuous, a weak solution is a strong solution and vice-versa.

Using the definition of the error function (28), the definition of $\overline{x}_T = \frac{1}{T}\sum_{t=1}^T x_t$, and the monotonicity of F (26), we obtain

$$\begin{aligned} \mathrm{Err}(\overline{x}_T) &= \sup_{y\in K} \langle F(y), \overline{x}_T - y\rangle \\ &= \sup_{y\in K} \left(\frac{1}{T} \sum_{t=1}^T \langle F(y), x_t - y\rangle \right) \\ &\le \sup_{y\in K} \left(\frac{1}{T} \sum_{t=1}^T \langle F(x_t), x_t - y\rangle \right) \end{aligned} \tag{30}$$

We fix an arbitrary point $y \in K$, and we analyze $\sum_{t=1}^T \langle F(x_t), x_t - y\rangle$. A key step is to split each inner product as follows:

$$\langle F(x_t), x_t - y\rangle = \langle F(x_t), z_t - y\rangle + \langle F(z_{t-1}), x_t - z_t\rangle + \langle F(x_t) - F(z_{t-1}), x_t - z_t\rangle \tag{31}$$

The intuition behind the split is that we can upper bound the first two terms using the optimality conditions for z_t and x_t respectively. The third term is a loss term that will be offset by the gains that we will obtain from the first two terms.

By the optimality condition for z_t, we have

$$\left\langle F(x_t) + \frac{1}{\eta_{t-1}} (z_t - z_{t-1}), z_t - y \right\rangle \le 0$$

Rearranging and using the identity $ab = \frac{1}{2}(a+b)^2 - \frac{1}{2}a^2 - \frac{1}{2}b^2$, we obtain

$$\begin{aligned} \langle F(x_t), z_t - y\rangle &\le \frac{1}{\eta_{t-1}} \langle z_{t-1} - z_t, z_t - y\rangle \\ &= \frac{1}{2\eta_{t-1}} \left(\|z_{t-1} - y\|^2 - \|z_t - y\|^2 - \|z_{t-1} - z_t\|^2 \right) \end{aligned}$$

Similarly, by the optimality condition for x_t, we have

$$\begin{aligned}\langle F(z_{t-1}), x_t - z_t\rangle &\leq \frac{1}{\eta_{t-1}} \langle z_{t-1} - x_t, x_t - z_t\rangle \\ &= \frac{1}{2\eta_{t-1}} \left(\|z_{t-1} - z_t\|^2 - \|x_t - z_{t-1}\|^2 - \|x_t - z_t\|^2 \right)\end{aligned}$$

Plugging into (31), we obtain

$$\begin{aligned}\langle F(x_t), x_t - y\rangle \leq \frac{1}{2\eta_{t-1}} &\left(\|z_{t-1} - y\|^2 - \|z_t - y\|^2 \right) \\ &+ \langle F(x_t) - F(z_{t-1}), x_t - z_t\rangle - \frac{1}{2\eta_{t-1}} \left(\|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2 \right)\end{aligned}$$

The first two terms telescope for uniform step sizes $\eta_t = \eta$. When the step sizes are non-uniform, as it is the case for the adaptive algorithms described below, the assumption that K has bounded diameter allows us to telescope the sums as follows:

$$\begin{aligned}\frac{1}{2\eta_{t-1}} \left(\|z_{t-1} - y\|^2 - \|z_t - y\|^2 \right) &= \frac{1}{2\eta_{t-1}} \|z_{t-1} - y\|^2 - \frac{1}{2\eta_t} \|z_t - y\|^2 + \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \underbrace{\|z_t - y\|^2}_{\leq R^2} \\ &\leq \underbrace{\frac{1}{2\eta_{t-1}} \|z_{t-1} - y\|^2 - \frac{1}{2\eta_t} \|z_t - y\|^2}_{\text{telescopes}} + \underbrace{\left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right)}_{\text{telescopes}} R^2\end{aligned}$$

Thus we have obtained

$$\begin{aligned}\langle F(x_t), x_t - y\rangle \leq &\underbrace{\frac{1}{2\eta_{t-1}} \|z_{t-1} - y\|^2 - \frac{1}{2\eta_t} \|z_t - y\|^2}_{\text{telescopes}} + \underbrace{\left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right)}_{\text{telescopes}} R^2 \\ &+ \underbrace{\langle F(x_t) - F(z_{t-1}), x_t - z_t\rangle}_{\text{loss}} - \underbrace{\frac{1}{2\eta_{t-1}} \left(\|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2 \right)}_{\text{gain}}\end{aligned} \tag{32}$$

The final step is to upper bound the net loss. We do so separately for "non-smooth" (i.e., non-Lipschitz) and "smooth" (i.e., Lipschitz) operators.

Non-smooth setting As in the convex minimization setting, we will need to assume that the operator norms are bounded, i.e., we have $\|F(x)\| \leq G$ for all $x \in K$. We proceed similarly to the gradient descent analysis. As in the primer lecture, we will use a uniform step size $\eta_t = \eta$ where η will follow from the analysis.

Using Cauchy-Schwartz, the triangle inequality, the bounded operator assumption, and the inequality $ab \leq \frac{\lambda}{2}a^2 + \frac{1}{2\lambda}b^2$, we obtain

$$\begin{aligned}\langle F(x_t) - F(z_{t-1}), x_t - z_t\rangle &\leq \|F(x_t) - F(z_{t-1})\| \|x_t - z_t\| \\ &\leq (\|F(x_t)\| + \|F(z_{t-1})\|) \|x_t - z_t\| \\ &\leq 2G \|x_t - z_t\| \\ &\leq 2\eta G^2 + \frac{1}{2\eta} \|x_t - z_t\|^2\end{aligned}$$

Plugging into (32) and summing over all iterations, we obtain

$$\sum_{t=1}^{T} \langle F(x_t), x_t - y\rangle \leq \frac{1}{2\eta} \|z_0 - y\|^2 + 2\eta G^2 T$$

We set η to balance the two terms:

$$\eta = \frac{\|z_0 - y\|}{2G\sqrt{T}}$$

and obtain

$$\sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \leq 2G \|z_0 - y\| \sqrt{T}$$
$$\leq 2GR\sqrt{T}$$

where we have let $R \geq \max_{x,y \in K} \|x - y\|$.

Plugging into (30) gives our convergence guarantee:

$$\mathrm{Err}(\overline{x}_T) \leq O\left(\frac{GR}{\sqrt{T}}\right)$$

Smooth setting We now consider the setting where F is β-Lipschitz. As in the primer lecture, we will use a uniform step size $\eta_t = \eta$ where η will follow from the analysis. Using Cauchy-Schwartz and the Lipschitz property, we obtain

$$\begin{aligned}\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle &\leq \|F(x_t) - F(z_{t-1})\| \|x_t - z_t\| \\ &\leq \beta \|x_t - z_{t-1}\| \|x_t - z_t\| \\ &\leq \frac{\beta}{2} \|x_t - z_{t-1}\|^2 + \frac{\beta}{2} \|x_t - z_t\|^2\end{aligned}$$

Plugging into (32), we obtain

$$\begin{aligned}\langle F(x_t), x_t - y \rangle \leq \frac{1}{2\eta} \left(\|z_{t-1} - y\|^2 - \|z_t - y\|^2 \right) \\ + \frac{\beta}{2} \|x_t - z_{t-1}\|^2 + \frac{\beta}{2} \|x_t - z_t\|^2 - \frac{1}{2\eta} \left(\|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2 \right)\end{aligned}$$

We set $\eta = \frac{1}{\beta}$ so that the terms cancel, and obtain

$$\sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \leq \frac{\beta}{2} \|z_0 - y\|^2 \leq \frac{\beta R^2}{2}$$

Plugging into (30) gives our convergence guarantee:

$$\mathrm{Err}(\overline{x}_T) \leq O\left(\frac{\beta R^2}{T}\right)$$

In contrast to smooth convex minimization, the above convergence rate is optimal for variational inequalities with Lipschitz operators.

Adaptive algorithm I Similarly to the AdaGrad+ algorithm, we can make the algorithm adaptive by setting the step sizes based on the iterate movement. Specifically, we can use the distances appearing in the gain term in (32), since those are precisely the terms that will offset the loss. This leads us to the following algorithm, where once again we have normalized the update so that the step sizes change by at most a constant factor.

Algorithm 6 Adaptive extra-gradient algorithm with steps based on the iterate movement [1].

Let $z_0 \in K$, $\eta_0 > 0$.
For $t = 1, \ldots, T$, update:

$$x_t = \arg\min_{u \in K} \left\{ \langle F(z_{t-1}), u \rangle + \frac{1}{2\eta_{t-1}} \|u - z_{t-1}\|^2 \right\}$$
$$z_t = \arg\min_{u \in K} \left\{ \langle F(x_t), u \rangle + \frac{1}{2\eta_{t-1}} \|u - z_{t-1}\|^2 \right\}$$
$$\frac{1}{\eta_t^2} = \frac{1}{\eta_{t-1}^2} \left(1 + \frac{\|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2}{2R^2} \right)$$

Return $\overline{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$.

We can show that the above algorithm converges at a rate of $O\left(GR\sqrt{\frac{\ln(GT/R)}{T}}\right)$ for non-Lipschitz operators and $O\left(\frac{\beta \ln \beta R^2}{T}\right)$. We refer the reader to [1, 7] for more details.

Adaptive algorithm II We now introduce a different approach for setting the step sizes that uses the norm of the operator differences $\|F(x_t) - F(z_{t-1})\|^2$. We also added an extra term to the update for z_t; this term helps improve the constants, and it can be safely omitted without affecting the asymptotic convergence.

Algorithm 7 Adaptive extra-gradient algorithm with steps based on the operator differences [6].

Let $z_0 \in K$, $\eta_0 > 0$.
For $t = 1, \ldots, T$, update:

$$x_t = \arg\min_{u \in K} \left\{ \langle F(z_{t-1}), u \rangle + \frac{1}{2\eta_{t-1}} \|u - z_{t-1}\|^2 \right\}$$
$$\eta_t = \frac{R}{\sqrt{\sum_{s=1}^{t} \|F(x_s) - F(z_{s-1})\|^2}}$$
$$z_t = \arg\min_{u \in K} \left\{ \langle F(x_t), u \rangle + \frac{1}{2\eta_{t-1}} \|u - z_{t-1}\|^2 \right\}$$

Return $\overline{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$.

We can show that the algorithm converges at the rate $O\left(\frac{RG}{\sqrt{T}}\right)$ for non-Lipschitz operators and $O\left(\frac{\beta R^2}{T}\right)$ for Lipschitz operators [6]. Both rates are optimal for variational inequalities.

References

[1] Francis Bach and Kfir Y. Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pages 164–194. PMLR, 2019.

[2] Nikhil Bansal and Anupam Gupta. Potential-function proofs for first-order methods. *CoRR*, abs/1712.04581, 2017.

[3] Michael Cohen, Jelena Diakonikolas, and Lorenzo Orecchia. On acceleration with noise-corrupted gradients. In *International Conference of Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 1018–1027. PMLR, 2018.

[4] Michael B Cohen, Jelena Diakonikolas, and Lorenzo Orecchia. On acceleration with noise-corrupted gradients. *arXiv preprint arXiv:1805.12591*, 2018.

[5] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[6] Alina Ene and Huy L. Nguyen. Adaptive and universal algorithms for variational inequalities with optimal convergence. *CoRR*, abs/2010.07799, 2020.

[7] Alina Ene, Huy L Nguyen, and Adrian Vladu. Adaptive gradient methods for constrained convex optimization. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[8] Alexander Vladimirovich Gasnikov and Yu E Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58(1):48–64, 2018.

[9] G.M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12:747–756, 1976.

[10] H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory (COLT)*, pages 244–256. Omnipress, 2010.

[11] Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5, 1998.