ECE 8803: Online Decision Making in Machine Learning **Fall 2022**

Lecture 12: October 6

Lecturer: Guanghui Wang

12.1. Recap: Online Gradient Descent and Adaptive OCO

Last lecture, we introduced the online gradient descent (OGD) algorithm, which is another important algorithm for solving OCO (apart from FTRL). In each round t, OGD updates the decision by

$$\begin{cases} \boldsymbol{z}_t = \boldsymbol{w}_t - \eta \nabla \ell_t(\boldsymbol{w}_t) \\ \boldsymbol{w}_{t+1} = \Pi_{\mathcal{B}}(\boldsymbol{z}_t), \end{cases} \tag{12.1}$$

where $\Pi_{\mathcal{B}}(\boldsymbol{z}) = \arg\min_{\boldsymbol{z}' \in \mathcal{B}} \|\boldsymbol{z} - \boldsymbol{z}'\|_2$ is the projection operator. We showed that, by setting the step size $\eta = \frac{D}{G\sqrt{T}}$, OGD can ensure a regret bound of order $O(DG\sqrt{T})$.

In the second part of Tuesday's lecture, we discussed several adaptive OCO algorithms. One of the main features of adaptive OCO algorithms is that, they enjoy the $O(\sqrt{T})$-type of regret bound in general, and can *automatically* achieve tighter bounds when the problems at hand are structured/easy. Last lecture, we learned two kinds of adaptive algorithms: the first class of algorithms can adaptive to the *norm of gradients*, and enjoy the following regret bound:

$$R_T = \mathcal{O}\left(D \sqrt{\sum_{t=1}^{T} ||\boldsymbol{\ell}_t||_2^2} \right). \tag{12.2}$$

The bound above reduces to $O(DG\sqrt{T})$ in the worst case (that is, based on Assumption 2 stated in the last lecture, the norm of all gradients are bounded by a constant, so the above regret bound will always upper bounded by $O(DG\sqrt{T})$). On the other hand, it can automatically become much more tighter when the cumulative sum of the norm of the gradients are small. The main idea of this kind of algorithms is to apply a *gradeint-dependant time-variant* step size:

$$\eta_t := \frac{D}{\sqrt{\sum_{s=1}^{t} ||\nabla f_t(\boldsymbol{w}_t)||_2^2}}. \tag{12.3}$$

The second class of algorithms are Adagrad and its variants, which can adapt to *the sparsity of the data* (to be more persise, the sparsity of gradients). The main idea of Adagrad to set

an *individual learning rate*

$$\eta_{t,i} = \frac{1}{\sqrt{\sum_{s=1}^{t} g_{s,i}^2}}$$

corresponding to each coordinate i. We showed that the regret bound of Adagrad is on the order of

$$R_T = \mathcal{O}\left(D' \cdot \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} g_{t,i}^2}\right), \tag{12.4}$$

which automatically become tighter when some dimensions of the gradients are sparse. Finally, we also briefly reviewed other popular Adagrad-type of algorithms, including RMSprop and Adam.

12.2. Adapting to Changing Environments

Today, we are going to discuss a new type of adaptive OCO algorithms, which can adapt to *changing environments*. This line of research started by questioning whether regret is the best metric for OCO. Specifically, in all of the previous lectures, we use *regret* as the performance measure for our OCO algorithms, which is defined as

$$R_T = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{w}^*).$$

Here, $\boldsymbol{w}^*$ is the so-called "best decision in hindsight", given by $\boldsymbol{w}^* = \operatorname{argmin}_{\boldsymbol{w}\in\mathcal{B}} \sum_{t=1}^{T} \ell_t(\boldsymbol{w})$. The rationale (or basic assumption) behind this performance metric is that *there at least exists one fixed decision $\boldsymbol{w}$ in $\mathcal{B}$ that is reasonably good during T iterations.* However, this is too optimistic and may not hold in *changing environments*, where data are evolving and the optimal decision is drifting over time. For example, in the stock market, different stocks can be the best one in different days; In ads recommendation, the flavor of the customers may also change over time. In these cases, there does not exist one single comparator in $\mathcal{B}$ that is good enough for T rounds, and we want to compare with *different comparators in different time periods.*

Motivating Example To make the argument more solid, let's have a look at the following simple example. Consider a 1-dimensional OCO problem. Let $\mathcal{B} = [-1, 1]$. For the first $\frac{T}{2}$ rounds, the loss function is $\ell_t(w) = (w-1)^2$. From round $\frac{T}{2}+1$ to round T, the loss function becomes $\ell_t(w) = (w+1)^2$. In this example, the "best decision in hindsight" is

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathcal{B}} \sum_{t=1}^{T} \ell_t(\boldsymbol{w}) = 0.$$

However, it's easy to see that $\boldsymbol{w}^*$ performs terribly, since it always suffer a constant loss at each round, and the cumulative loss is T! Clearly, a more reasonable choice is to compare with the policy with zero cumulative loss, that is, $w = 1$ in the first half of the learning period, and $\boldsymbol{w} = -1$ in the second half.

Dynamic regret To address the limitation of regret, researchers have proposed a more stringent metric, called *dynamic regret*, in which the learner competes with a sequence of changing comparators $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathcal{B}$:

$$\text{D-}R_T(\mathbf{u}_1, \ldots, \mathbf{u}_T) = \sum_{t=1}^{T} f_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t).$$

For the motivating example above, a reasonable sequence of comparators would be $\mathbf{u}_1 = \cdots = \mathbf{u}_{T/2} = 1$, $\mathbf{u}_{T/2+1} = \cdots = \mathbf{u}_T = -1$. However, this is true since we already know all loss functions before the learning process. This kind of prior knowledge is not accessible in the general OCO. For this reason, when defining the dynamic regret, we assume the sequence of comparators $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathcal{B}$ is chosen by the adversary (or the environment), and it *is not known* to the learner (even after the learning process). Our goal is to be *competitive with any sequence of comparators*.

Algorithms for minimizing dynamic regret Next, let's see how to minimzie the dynamic regret. It turns out that, we can use the exact OGD algorithm defined in (12.1) and obtain the following conclusion:

$$\text{D-}R_T(\mathbf{u}_1, \ldots, \mathbf{u}_T) = O((V_T + 1)\sqrt{T}).$$

In class, we went over the proof step-by-step, and here we provide the proof in the Appendix. It is actully very similar to that of OGD for regret minimization, and the main difference is about how to deal with the telescope sum, and how to draw connections to V_T.

The theorem above shows that, OGD with an appropriately chosen step-size can achieve an $O((V_T + 1)\sqrt{T})$ regret bound. Thus, as long as $V_T = o(\sqrt{T})$, the regret bound is still sub-linear with respect to T. For instance, in our motivating example, one reasonable sequence of comparators is $\mathbf{u}_1 = \cdots = \mathbf{u}_{T/2} = 1$, $\mathbf{u}_{T/2+1} = \cdots = \mathbf{u}_T = -1$. For this sequence, $V_T = 2$ which is a constant, so in this case OGD can still achieve an $O(\sqrt{T})$ dynamic regret bound.

Finally, a follow-up question is whether we can improve this result. The answer is yes, and it is recently improved to $O(\sqrt{(V_T + 1)T})$ by the Ader algorithm. This algorithm is an elegant combination of OCO with learning with expert advice. In class, we discussed this algorithm. In the following, we only discuss the basic idea. Basically, in the proof of the dynamic regret for OGD, we can upper bound the dynamic regret in the form of $O(\frac{V_T}{\eta} + \eta T)$, where η is the step size for OGD. To minimize this upper bound, the best choice for η is on the order of $\frac{\sqrt{V_T}}{\sqrt{T}}$. However, we cannot set η like this, because we do not know V_T even after the learning process. To address this issue, Ader borrows the idea of learning with expert advice. Since we do not know the optimal step size, we can instead run multiple OGD simultaneously with different step sizes, and use a meta algorithm to learn the best OGD on the fly. The step sizes for these OGD algorithms are carefully designed, such that at least one OGD is reasonably good. We refer to the Ader paper (Zhang et al., 2018) for more details.

12.3. From regret to optimization convergence

We end this lecture with a cool implication of the regret bounds that we have derived: we can use these to show that *stochastic* optimization methods, like stochastic gradient descent (SGD) converge to the true minimum of an objective function $f(\boldsymbol{w})$. Due to time limitation, we only briefly mentioned this in class, but it is indeed an important topic.

Concretely, consider a convex optimization problem

$$\min_{\boldsymbol{w}\in\mathcal{B}} \ell(\boldsymbol{w})$$

where the decision set $\mathcal{B}$ is convex, and the function $\ell(\cdot)$ is convex in its argument. Here, we write the stochastic gradient update as

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \eta(\nabla\ell(\boldsymbol{w}_t) + \boldsymbol{n}_t) \tag{12.5}$$

where, as before, $\boldsymbol{w}_t$ is the projection of $\boldsymbol{z}_t$ into the constrained decision set $\mathcal{B}$. Here, $\nabla\ell(\boldsymbol{w}_t) + \boldsymbol{n}_t$ is a *noisy* version of the gradient at time step t, where $\boldsymbol{n}_t$ is some random bounded iid noise. These noisy gradients are commonly obtained in the application of training of large-scale ML models. Importantly, this means that the updates $\{\boldsymbol{w}_t\}_{t\geq 1}$ will be *random. In what follows, all expectations will be taken only over the randomness in the noise* $\{\boldsymbol{n}_t\}_{t\geq 1}$ *used in the SGD updates.*

We want to show that some *averaged* version of the model, converges to the true minimum, given by $\boldsymbol{w}^*$, at a particular rate. In particular, can we show that $\mathbb{E}[\overline{\boldsymbol{w}}_T]$ where $\overline{\boldsymbol{w}}_T := \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t$, becomes arbitrarily close to $\boldsymbol{w}^*$ as T increases? There turn out to be many ways of showing this, but we will now show one particularly elegant way using the OCO paradigm. We notice that Equation (12.5) can be written exactly as a step of online gradient descent on the *OLO* problem, with loss vectors given by $\boldsymbol{\ell}_t := \nabla\ell(\boldsymbol{w}_t) + \boldsymbol{n}_t$. If the original objective function $\ell(\cdot)$ and the noise are bounded, these loss vectors are clearly bounded. Therefore, we obtain the regret bound

$$R_{T,\mathsf{lin}} := \sum_{t=1}^{T}\langle\boldsymbol{\ell}_t, \boldsymbol{w}_t - \boldsymbol{w}^*\rangle = \mathcal{O}(DG\sqrt{T}), \tag{12.6}$$

where D represents the "size" of the decision set $\boldsymbol{w}$ and G represents the smoothness level of the original function $\ell(\cdot)$. Crucially, this regret bound holds for all realizations of the random noise used in the SGD updates[1].

We will now show how Equation (12.6) can be used to show that

$$\mathbb{E}\left[\ell(\overline{\boldsymbol{w}}_T)\right] \to \ell(\boldsymbol{w}^*)$$

for the choice of learning rate $\eta = \frac{D}{G\sqrt{T}}$. We split this proof up into three parts: the first two parts use important properties of convexity, and the third part uses the *unbiasedness* of the SGD updates, i.e. that they are equal to the true gradient in expectation. The intermediate steps are marked in blue and red for ease of reading.

1. Note that it can be verified from the proof below that a bound that only holds on the *expected* regret would suffice; however, this is typically not the type of bound that we obtain in OCO problems.

First, we crucially use convexity of the function $\ell(\cdot)$ to get:

$$\begin{aligned}\mathbb{E}\left[\ell(\overline{\boldsymbol{w}}_T) - \ell(\boldsymbol{w}^*)\right] &= \mathbb{E}\left[\ell\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t\right) - \ell(\boldsymbol{w}^*)\right] \\ &\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\ell(\boldsymbol{w}_t) - \ell(\boldsymbol{w}^*)\right] = \mathbb{E}\left[\frac{1}{T}(\sum_{t=1}^{T}\ell(\boldsymbol{w}_t) - \ell(\boldsymbol{w}^*))\right]\end{aligned}$$

This inequality uses the 0-th order definition of the convex function (and is commonly called *Jensen's inequality*). In other words, recall that for any $\theta \in (0,1)$, we had $\ell(\theta\boldsymbol{w}_1+(1-\theta)\boldsymbol{w}_2) \leq \theta\ell(\boldsymbol{w}_1) + (1-\theta)\ell(\boldsymbol{w}_2)$. We can apply this inequality with $\theta = 1/2$ to get:

$$\ell\left(\frac{1}{2}(\boldsymbol{w}_1 + \boldsymbol{w}_2)\right) \leq \frac{1}{2}\left(\ell(\boldsymbol{w}_1) + \ell(\boldsymbol{w}_2)\right),$$

and extending this inequality to T arguments $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T$ gives us the desired inequality.

Second, we use the first-order definition of convexity to get

$$\begin{aligned}&\ell(\boldsymbol{w}^*) - \ell(\boldsymbol{w}_t) \geq \langle\nabla\ell(\boldsymbol{w}_t), \boldsymbol{w}^* - \boldsymbol{w}_t\rangle \\ \implies\ &\ell(\boldsymbol{w}_t) - \ell(\boldsymbol{w}^*) \leq \langle\nabla\ell(\boldsymbol{w}_t), \boldsymbol{w}_t - \boldsymbol{w}^*\rangle.\end{aligned}$$

Substituting this below gives us

$$\mathbb{E}\left[\frac{1}{T}(\sum_{t=1}^{T}\ell(\boldsymbol{w}_t) - \ell(\boldsymbol{w}^*))\right] \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle\nabla\ell(\boldsymbol{w}_t), \boldsymbol{w}_t - \boldsymbol{w}^*\rangle\right].$$

Finally, we use the fact that the stochastic gradient updates are equal to the gradient in expectation, i.e. $\mathbb{E}[\boldsymbol{\ell}_t] = \nabla\ell(\boldsymbol{w}_t)$. This allows us to connect directly to the regret bound in Equation (12.6)! In fact, we get:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle\nabla\ell(\boldsymbol{w}_t), \boldsymbol{w}_t - \boldsymbol{w}^*\rangle\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle\boldsymbol{\ell}_t, \boldsymbol{w}_t - \boldsymbol{w}^*\rangle\right] = \mathbb{E}\left[\frac{R_{T,\mathsf{lin}}}{T}\right] = \mathcal{O}\left(\frac{DG}{\sqrt{T}}\right).$$

Putting these steps together completes the proof of optimization convergence for the choice of learning rate $\eta = \frac{D}{G\sqrt{T}}$. In practice, SGD uses a time-varying, decaying step size given by $\eta_t = \frac{D}{G\sqrt{t}}$; this can be also shown to converge through OCO-regret (in a manner similar to the "doubling trick" that you studied in HW 1, Problem 3). We do not discuss this in detail.

In summary, we have shown that *regret* bounds for OCO imply *convergence* to the optimum of stochastic gradient methods, which are heavily used in practice.

References

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 1323–1333, 2018.

Appendix A. Proof of the dynamic regret for OGD

Like the proof we did in Lecture 11, we first do the following one-step analysis.

$$\begin{aligned}
\ell_t(\boldsymbol{w}_t) - \ell_t(\mathbf{u}_t) &\le \langle \nabla \ell_t(\boldsymbol{w}_t), \boldsymbol{w}_t - \mathbf{u}_t \rangle \\
&= \frac{1}{\eta} \langle \boldsymbol{z}_t - \boldsymbol{w}_t, \boldsymbol{w}_t - \mathbf{u}_t \rangle \\
&\le \frac{1}{2\eta} \left[\|\boldsymbol{z}_t - \boldsymbol{w}_t\|_2^2 + \|\boldsymbol{w}_t - \mathbf{u}_t\|_2^2 - \|\boldsymbol{z}_t - \mathbf{u}_t\|_2^2 \right] \\
&\le \frac{1}{2\eta} \left[\|\boldsymbol{z}_t - \boldsymbol{w}_t\|_2^2 + \|\boldsymbol{w}_t - \mathbf{u}_t\|_2^2 - \|\boldsymbol{w}_{t+1} - \mathbf{u}_t\|_2^2 \right] \\
&= \frac{1}{2\eta} \left[\eta^2 \|\nabla \ell_t(\boldsymbol{w}_t)\|_2^2 + \|\boldsymbol{w}_t - \mathbf{u}_t\|_2^2 - \|\boldsymbol{w}_{t+1} - \mathbf{u}_t\|_2^2 \right] \\
&= \frac{\eta \|\nabla \ell_t(\boldsymbol{w}_t)\|_2^2}{2} + \frac{\|\boldsymbol{w}_t - \boldsymbol{w}^*\|_2^2 - \|\boldsymbol{w}_{t+1} - \mathbf{u}_t\|_2^2}{2\eta} \\
&= \frac{\eta G^2}{2} + \frac{\|\boldsymbol{w}_t - \mathbf{u}_t\|_2^2 - \|\boldsymbol{w}_{t+1} - \mathbf{u}_t\|_2^2}{2\eta}
\end{aligned} \tag{12.7}$$

Please compare this proof with the proof in the appendix of Lecture 11. You can see that the are *exactly* the same. The only difference is that we replace $\boldsymbol{w}^*$ with $\mathbf{u}_t$.

Sum it over from 1 to T, we have

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{w}^*) \le \frac{\eta G^2 T}{2} + \sum_{t=1}^{T} \frac{\|\boldsymbol{w}_t - \mathbf{u}_t\|_2^2 - \|\boldsymbol{w}_{t+1} - \mathbf{u}_t\|_2^2}{2\eta}.$$

Note that the second term is no longer a telescope sum, so the proof in Lecture 11 will not go through. To fix this issue, we choose a different path to handle this term. We first expend the square norms:

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{\|\boldsymbol{w}_t - \mathbf{u}_t\|_2^2 - \|\boldsymbol{w}_{t+1} - \mathbf{u}_t\|_2^2}{2\eta} \\
&= \frac{1}{2\eta} \sum_{t=1}^{T} \left(\|\boldsymbol{w}_t - \mathbf{u}_t\|_2^2 - \|\boldsymbol{w}_{t+1} - \mathbf{u}_t\|_2^2 \right) \\
&= \frac{1}{2\eta} \sum_{t=1}^{T} \left(\|\boldsymbol{w}_t\|_2^2 + \|\mathbf{u}_t\|_2^2 - 2\boldsymbol{w}_t^\top \mathbf{u}_t - \|\boldsymbol{w}_{t+1}\|_2^2 - \|\mathbf{u}_t\|_2^2 + 2\boldsymbol{w}_{t+1}^\top \mathbf{u}_t \right) \\
&= \sum_{t=1}^{T} \frac{1}{2\eta} \left(\|\boldsymbol{w}_t\|_2^2 - \|\boldsymbol{w}_{t+1}\|_2^2 \right) + \sum_{t=1}^{T} \frac{1}{2\eta} \left(2\mathbf{u}_t^\top (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) \right),
\end{aligned} \tag{12.8}$$

For the first term in the right hind side of the inequality, note that it is a telescope sum, so

$$\frac{1}{2\eta} \sum_{t=1}^{T} (\|\boldsymbol{w}_t\|_2^2 - \|\boldsymbol{w}_{t+1}\|_2^2) = \frac{1}{2\eta} (\|\boldsymbol{w}_1\|_2^2 - \|\boldsymbol{w}_{T+1}\|_2^2) \le \frac{D^2}{2\eta}.$$

For the second term of (12.8), we do some rearranging and hope to draw relation to V_T:

$$\begin{aligned}
&\frac{1}{\eta}\sum_{t=1}^{T}\left(\mathbf{u}_t^\top(\boldsymbol{w}_{t+1}-\boldsymbol{w}_t)\right)\\
&=\frac{1}{\eta}\sum_{t=2}^{T}\boldsymbol{w}_t^\top(\mathbf{u}_{t-1}-\mathbf{u}_t)+\frac{1}{\eta}\mathbf{u}_T^\top\boldsymbol{w}_{T+1}-\frac{1}{\eta}\mathbf{u}_1^\top\boldsymbol{w}_1\\
&\le\frac{1}{\eta}\sum_{t=2}^{T}\boldsymbol{w}_t^\top(\mathbf{u}_{t-1}-\mathbf{u}_t)+\frac{2D^2}{\eta}\\
&\le\frac{1}{\eta}\sum_{t=2}^{T}\|\boldsymbol{w}_t^\top\|_2\|\mathbf{u}_{t-1}-\mathbf{u}_t\|_2+\frac{2D^2}{\eta}\\
&\le\frac{D^2}{\eta}\sum_{t=2}^{T}\|\mathbf{u}_{t-1}-\mathbf{u}_t\|_2+\frac{2D^2}{\eta}\\
&=\frac{D^2V_T}{\eta}+\frac{2D^2}{\eta},
\end{aligned}\tag{12.9}$$

where the inequalities are based on Assumptions 1 and 2 in Lecture 11 and Cauchy–Schwarz. To summarize, we have

$$\text{D-}R_T\le\frac{D^2V_T}{\eta}+\frac{2D^2}{\eta}+\frac{D^2}{2\eta}+\frac{\eta G^2T}{2}=\frac{1}{\eta}(D^2V_T+2.5D^2)+\eta\frac{G^2T}{2},$$

and we can finish the proof by plugging in the value of η.