Data-Dependent Algorithms for Bandit Convex Optimization

Mehryar Mohri
Courant Institute and Google
251 Mercer Street
New York, NY 10012
mohri@cims.nyu.edu

Scott Yang
Courant Institute
251 Mercer Street
New York, NY 10012
yangs@cims.nyu.edu

Abstract

We present algorithms with strong data-dependent regret guarantees for the problem of bandit convex optimization. In the process, we develop a general framework from which all previous main results in this setting can be recovered. The key method is the introduction of adaptive regularization. By appropriately adapting the exploration scheme, we show that one can derive regret guarantees which can be significantly more favorable than those previously known.

1 Introduction

Bandit convex optimization is a general framework for sequential decision making under uncertainty. It generalizes the well-known multi-armed bandit scenario, and captures the exploration-exploitation trade-off inherent to many online machine learning problems. At the same time, bandit convex optimization also remains an area of online convex optimization in which existing regret guarantees are not satisfactory.

We consider the setting of bandit convex optimization. Let $\mathcal{K} \subset \mathbb{R}^n$ be a compact convex set, and let $\{f_t\}_{t=1}^{\infty}$ be a sequence of convex functions. At each round $t = 1, 2, \ldots, T$, the learner selects a point $x_t \in \mathcal{K}$ and incurs loss $f_t(x_t)$. The learner's objective is to minimize his regret, defined by:

$$\mathrm{Reg}_T = \max_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x_t) - f_t(x),$$

that is the difference between his cumulative loss and that of the best fixed point in hindsight. In contrast to the standard online learning or online convex optimization scenario, in bandit convex optimization, the learner has access only to the value $f_t(x_t)$ and not any higher-order information.

This scenario was first studied by [9], and has been further investigated by [1, 2, 12, 10, 4, 14].

It still remains an open question whether one can efficiently obtain the desired $\mathcal{O}(\sqrt{T})$ regret in the purely strongly convex, purely strongly smooth, or purely Lipschitz settings. To make progress in this direction, we will build upon recent advances in other areas of the online convex optimization literature. Specifically, we will draw from the techniques in adaptive regularization presented in [3, 7, 11] as well as ideas from the "learning faster from easy data" paradigm studied in [8, 5, 13, 6] to derive a pair of efficient adaptive algorithms with minimal assumptions on the function's loss sequence.

Specifically, our algorithms will provide strong data-dependent guarantees, so that while their regret will never be worse than that of previous algorithms in the same setting, they can also be much better depending on how favorable and "easy" the actual data is. Moreover, the algorithms we present are any-time and automatically adjust to the data, so that they can run without any a priori tuning or unreasonable parameter specification.

Algorithm 1 AdaBCO-Lipschitz

Input: $\eta_0 = \frac{1}{2nC}$, ν-self concordant barrier $\mathcal{R}$, $C > 0$ constant
Initialize: $x_1 = \operatorname{argmin}_{x \in \mathcal{K}} \mathcal{R}(x)$.
for $t = 1, \dots, T$: **do**
  Choose a constant $L_t \geq 0$ such that $|f_t(x) - f_t(y)| \leq L_t |x - y|$.
  Choose matrix $Q_t \succcurlyeq 0$ such that $f_t(x) \geq f_t(x_t) + g_t^\top (x - x_t) + \frac{1}{2} \|Q_t(x - x_t)\|_2^2$.
  Define $\tilde{B}_{t,s} = \left(\nabla^2 R(x_s) + (\eta_s 1_{\{s<t\}}) Q_{1:s}\right)^{-1/2}$ and

$$\eta_t = \left(\sum_{s=1}^{t} 2 \left(2L_s \frac{1}{n} \sum_{j=1}^{n} \lambda_j(\tilde{B}_{t,s}) n^2 C^2 \right)^{1/3} \right)^{-3/4} \left(\frac{\nu \log(T)}{2} \right)^{3/4}$$

  Let $B_t = \left[\nabla^2 \mathcal{R}(x_t) + \eta_t Q_{1:t}\right]^{-1/2}$.
  Define $\delta_t = \left(2 \frac{n^3 C^2 \eta_t}{L_t \sum_{j=1}^{n} \lambda_j(B_t)} \right)^{1/3}$.
  Sample $u \sim \mathcal{S}^n$ uniformly, where $\mathcal{S}^n = \{x \in \mathbb{R}^n | \|x\|_2 = 1\}$ the unit sphere.
  Select $y_t = x_t + \delta_t B_t u \in W_1(x_t) \subset \mathcal{K}$, incurring loss $f_t(y_t)$.
  Construct the gradient estimate $\widehat{g}_t = n f_t(y_t)(\delta_t B_t)^{-1} u$.
  Update $x_{t+1} = \operatorname{argmin}_{x \in \mathcal{K}} g_{1:t}^\top x + \frac{1}{2} \sum_{s=1}^{t} \|(x - x_s)\|_{Q_s}^2 + \frac{1}{\eta_t} \mathcal{R}(x)$.
end for

2 Adaptive bandit convex optimization

We will now present Algorithm 1, AdaBCO-Lipschitz, an adaptive algorithm that requires minimal regularity assumptions on the loss functions.

Unlike previous algorithms in the literature, AdaBCO-Lipschitz does not need the learner to specify a priori a fixed level of convexity for the sequence of loss functions encountered during learning. This is often an unreasonable requirement, particularly in a truly online setting, and so instead, AdaBCO-Lipschitz allows the learner to specify the convexity of functions as it sees them, and the algorithm is constructed such that the regret bound will automatically adapt to this data. This is achieved via dynamic tuning of the sampling ellipsoids and learning rates.

Moreover, it is important to realize that computing parameters in real-time is never more difficult than computing bounds that hold uniformly over all rounds at the start in a truly online scenario, so AdaBCO-Lipschitz is never more difficult to implement than previous algorithms.

Theorem 1 (Adaptive BCO using dynamic Lipschitz bounds)**.** *Let $\mathcal{K}$ be a convex set and $\mathcal{R}$ a ν-self-concordant barrier over $\mathcal{K}$. Assume that $|f| \leq C$. Then Algorithm 2 provides the regret bound:*

$$\sum_{t=1}^{T} \mathbb{E}[f_t(y_t) - f_t(x)] \leq \mathbb{E}\left[5(\nu \log(T))^{\frac{1}{4}} \left(\sum_{t=1}^{T} \left(L_t n C^2 \sum_{j=1}^{n} \lambda_j(\tilde{B}_{t,s}) \right)^{\frac{1}{3}} \right)^{\frac{3}{4}} \right]$$

As mentioned before, Algorithm 1 differs from previous algorithms in the literature in that it does not require a priori assumptions on the convexity or smoothness of the loss functions. One can adjust these convexity matrices and smoothness constants dynamically over time, and the regret bound will adapt.

The construction of dynamic sampling ellipsoids and data-dependent learning rates here involves some subtlety due to their interdependence. The optimal a posteriori learning rate depends on the sampling ellipsoid, and the radius of the sampling ellipsoid depends on the learning rate. Building an on-line approximation to the optimal rate also involves an abstract result on normalized sums.

Moreover, AdaBCO differs from previous algorithms in that it treats convexity as a matrix parameter instead of a scalar parameter. This is based on the insight that, for minimizing regret, convexity of the loss function is closely tied to convexity of the self-concordant barrier's Hessian, and that one

Algorithm 2 AdaBCO-Smooth

Input: $\eta_0 = \frac{1}{2nC}$, ν-self concordant barrier $\mathcal{R}$, $C > 0$ constant
Initialize: $x_1 = \operatorname{argmin}_{x \in \mathcal{K}} \mathcal{R}(x)$.
for $t = 1, \dots, T$: **do**
  Choose a constant $\beta_t \geq 0$ such that $f_t(x) \leq f_t(y) + \nabla f_t(y)^\top (x - y) + \frac{\beta_t}{2}\|x - y\|_2^2$.
  Choose matrix $Q_t \succcurlyeq 0$ such that $f_t(x) \geq f_t(x_t) + g_t^\top (x - x_t) + \frac{1}{2}\|Q_t(x - x_t)\|_2^2$.
  Define $\tilde{B}_{t,s} = \left(\nabla^2 R(x_s) + \eta_s 1_{\{s<t\}} Q_{1:s}\right)^{-1/2}$ and

$$\eta_t = \left(\sum_{s=1}^{t} \sqrt{4\beta_s \frac{1}{n} \sum_{j=1}^{n} \lambda_j(\tilde{B}_{t,s}^2) n^2 C^2}\right)^{-2/3} (\nu \log(T))^{2/3}$$

  Let $B_t = \left[\nabla^2 \mathcal{R}(x_t) + \eta_t Q_{1:t}\right]^{-1/2}$.
  Define $\delta_t = \left(\frac{n^3 C^2 \eta_t}{\beta_t \sum_{j=1}^{n} \lambda_j(B_t^2)}\right)^{1/4}$.
  Sample $u \sim \mathcal{S}^n$ uniformly, where $\mathcal{S}^n = \{x \in \mathbb{R}^n | \|x\|_2 = 1\}$ the unit sphere.
  Select $y_t = x_t + \delta_t B_t u \in W_1(x_t) \subset \mathcal{K}$, incurring loss $f_t(y_t)$.
  Construct the gradient estimate $\widehat{g}_t = n f_t(y_t)(\delta_t B_t)^{-1} u$.
  Update $x_{t+1} = \operatorname{argmin}_{x \in \mathcal{K}} g_{1:t}^\top x + \frac{1}{2}\sum_{s=1}^{t} \|(x - x_s)\|_{Q_s}^2 + \frac{1}{\eta_t}\mathcal{R}(x)$.
end for

can bound regret in terms of the average eigenvalue of the sum of these matrices as opposed to to the minimal eigenvalue. Essentially, the algorithm can "borrow" convexity from the self-concordant barrier if the convexity of the loss function is not strong enough to achieve the desired regret. This becomes particularly useful when the learner is querying points near the decision set's boundary, and the Hessian of $\mathcal{R}$ has large eigenvalues in the direction of the boundary (often the case because $\mathcal{R} \nearrow \infty$ at $\partial\mathcal{K}$).

One can also construct a similar algorithm for functions that are dynamically $C^{1,1}$ with the following guarantee:

Theorem 2 (Adaptive BCO using dynamic smoothness bounds)**.** *Let $\mathcal{K}$ be a convex set and $\mathcal{R}$ a ν-self-concordant barrier over $\mathcal{K}$. Assume that $|f| \leq C$. Then the following regret bound holds for Algorithm 2:*

$$\sum_{t=1}^{T} \mathbb{E}[f_t(y_t) - f_t(x)] \leq \mathbb{E}\left[\frac{5}{2}(\nu \log(T))^{\frac{1}{3}} \left(\sum_{t=1}^{T} \sqrt{4\beta_t n C^2 \sum_{j=1}^{n} \lambda_j(\tilde{B}_{t,s}^2)}\right)^{\frac{2}{3}}\right]$$

3 Applications and comparisons with previous results

The data-dependent nature of the above algorithm provides two important implications.

The first is that they allow us to easily produce regret bounds in a variety of new situations, where the learner experiences loss functions with various levels of smoothness and convexity. In particular, we can identify new scenarios where the optimal $\tilde{\mathcal{O}}(\sqrt{T})$ regret is achievable by navigating the relationship between smoothnesss and convexity.

The second is that these algorithms also automatically adapt to the smoothness and convexity of these scenarios. These new cases do not require any a priori insight or tuning. The algorithms presented in this paper adaptively determine optimal sampling ellipsoids and learning rates, which lead to strong guarantees.

In particular, they allow the learner to recover existing regret bounds without modifying the algorithms. Properties such as strong convexity or smoothness are processed adaptively and online, so that if, e.g., a sequence of loss functions is found to be approximately strongly convex (which will become clear in the following results), then the strongly convex guarantee will apply. If the sequence

of loss functions is better than strongly convex, then the algorithm will give an even better guarantee. Thus, these algorithms are prime examples of algorithms that "learn faster from easy data".

We present first the results for Algorithm 1.

Corollary 1 (Power law asymptotics for the dynamically Lipschitz and strongly convex scenario)**.** *Assume that there exists $\alpha \in \mathbb{R}$ such that*

$$L_t n C^2 \sum_{j=1}^{n} \lambda_j((\nabla^2 \mathcal{R}(x_t) + \eta_t Q_{1:t})^{-1/2}) = \mathcal{O}(t^\alpha).$$

Then the inequality $\sum_{t=1}^{T} \mathbb{E}[f_t(y_t) - f_t(x)] \leq \tilde{\mathcal{O}}(T^{\frac{3+\alpha}{4}})$ holds.

In particular, $\tilde{\mathcal{O}}(\sqrt{T})$ regret is attainable for $\alpha \leq -1$.

Moreover, $\tilde{\mathcal{O}}(T^{2/3})$ regret is adaptively attained for strongly convex functions, and $\tilde{\mathcal{O}}(T^{3/4})$ regret is adaptively attained for Lipschitz functions.

We would like to stress that the above reductions are worst-case guarantees. The data-dependent nature of the regret bound above implies that it can do much better on easier data, and that we do not need to know about these optimistic settings in advance of running the algorithm. The algorithms above will automatically take advantage of them.

This is not the case with any of the previous algorithms in bandit convex optimization, as they do not provide any sort of data-dependent guarantees. In particular, they do not account for the convexity of the self-concordant barrier at all, so that they provide much weaker bounds when an algorithm plays points at which the Hessian of the barrier has large average eigenvalues. For self-concordant barriers such as the log-barrier function, having larger average eigenvalues corresponds to being closer to the boundary of the set.

We now present the accompanying results for smooth functions.

Corollary 2 (Power law asymptotics for the dynamically smooth and strongly convex scenario)**.** *Assume that there exists $\alpha \in \mathbb{R}$ such that*

$$\beta_t n C^2 \sum_{j=1}^{n} \lambda_j((\nabla^2 \mathcal{R}(x_t) + \eta_t Q_{1:t})^{-1}) = \mathcal{O}(t^\alpha).$$

Then the inequality $\sum_{t=1}^{T} \mathbb{E}[f_t(y_t) - f_t(x)] \leq \tilde{\mathcal{O}}(T^{\frac{2+\alpha}{3}})$ holds.

In particular, $\tilde{\mathcal{O}}(\sqrt{T})$ regret is attainable for $\alpha \leq \frac{-1}{2}$.

Moreover, $\tilde{\mathcal{O}}(T^{1/2})$ regret is adaptively attained for smooth and strongly convex functions, while $\tilde{\mathcal{O}}(T^{2/3})$ regret is adaptively attained for smooth functions.

4 Conclusion

We presented efficient algorithms with regret bounds that adapt to the bandit convex optimization problem. Unlike previous algorithms, the algorithms we give do not require a priori assumptions of strong convexity or strong smoothness. Instead, they can process these parameters in an online fashion, which is much more suitable for the settling on online convex optimization.

Moreover, our algorithms adapt to the sequence of loss functions seen, so that they adapt to the problem at hand, recovering, in the worst case, all previous known bounds in this scenario without external tuning of learning rates or prior knowledge of function regularity.

They also provide data-dependent guarantees, so that on "easier data", the algorithms learn faster and the bounds become tighter. This is a desirable attribute that previous algorithms do not possess, and is due to a careful online tuning of the sampling ellipsoid and learning rate as well as the exploiting the connection between the convexity of the loss function and the convexity of the self-concordant barrier. In particular, we are able to use these algorithms and their accompanying bounds to provide problem-dependent conditions under which one can obtain the $\tilde{\mathcal{O}}(\sqrt{T})$ regret, including in the purely Lipschitz or purely smooth settings.

References

[1] Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *COLT*, pages 263–274, 2008.

[2] Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pages 28–40, 2010.

[3] Peter L. Bartlett, Elad Hazan, and Alexander Rakhlin. Adaptive online gradient descent. In *NIPS*, pages 65–72, 2007.

[4] Sébastien Bubeck and Ronen Eldan. Multi-scale exploration of convex functions and bandit convex optimization. *CoRR*, abs/1507.06580, 2015.

[5] Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *COLT*, volume 23, pages 42.1–42.23, 2012.

[6] Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316, 2014.

[7] John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *COLT*, pages 257–269, 2010.

[8] Eyal Even-Dar, Michael Kearns, Yishay Mansour, and Jennifer Wortman. Regret to the best vs. regret to the average. In *COLT*, volume 4539, pages 233–247, 2007.

[9] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *SODA*, pages 385–394, 2005.

[10] Elad Hazan and Kfir Y. Levy. Bandit convex optimization: Towards tight bounds. In *NIPS*, pages 784–792, 2014.

[11] H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In *COLT*, pages 244–256, 2010.

[12] Ankan Saha and Ambuj Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *AISTATS*, pages 636–642, 2011.

[13] Amir Sani, Gergely Neu, and Alessandro Lazaric. Exploiting easy data in online optimization. In *NIPS*, pages 810–818, 2014.

[14] Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *COLT*, volume 30, pages 3–24, 2013.