Zeroth-Order Online Convex Optimization

Yan Pan
Carnegie Mellon University
ypan2@andrew.cmu.edu

1 Introduction

In online machine learning, the objective function is changing over time, and online convex optimization is needed to find the optimal solution for online learning problems (Shalev-Shwartz et al., 2012). While stochastic gradient methods, such as stochastic gradient descent, can be applied on such learning problems, in many applications such as online advertising, gradient-based algorithms are no longer feasible since gradient information is hard to obtain. We refer such setting to "zeroth-order" or "bandit" online convex optimization, where we optimize without gradient information. Then, optimization is particularly hard in the online setting, since only limited information can be queried for each function.

In this project, we review important algorithms in zeroth-order online convex optimization, where they perform gradient descent without any gradient information. The setting is closedly connected to the online learning problem in class, but is extended to the bandit setting. We provide an overview of the online convex optimization framework, where the setting will follow the framework proposed by Zinkevich (2003). Then, we prove how stochastic projected gradient descent can achieve $O(\sqrt{T})$ regret in the online setting, which is the foundation of most online optimization algorithms. Then, we analyze the algorithm of Flaxman et al. (2005), where Stoke's theorem is used to approximate the gradient with function value at one point. We prove its regret of $O(T^{3/4})$ using the stochastic projected gradient descent framework. Furthermore, we show how similar ideas can be applied to slightly different settings with modifications, leading to more optimal algorithms in the two-point feedback setting (Agarwal and Dekel, 2010) and smooth setting (Saha and Tewari, 2011). In the end, we also discuss an application of the algorithm in game theory described by Bravo et al. (2018).

2 Online Convex Optimization

2.1 Convex Optimization

We start with the definitions of convex sets and convex functions. A set $\mathcal{D}$ is *convex* if for every $x, y \in \mathcal{D}$, $0 \leq \lambda \leq 1$, we have $\lambda x + (1-\lambda)y \in \mathcal{D}$. A function $f : \mathcal{D} \to \mathbb{R}$ is *convex* if $\mathcal{D}$ is

convex and for every $x, y \in \mathcal{D}$, $0 \leq \lambda \leq 1$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

In general, we can efficiently optimize convex functions with a variety of first-order (gradient-based) algorithms such as gradient descent, Adagrad (Duchi et al., 2011), Adam (Kingma and Ba, 2015).

In online optimization, the main difference is that instead of a fixed function f, we now have a sequence of functions $f_1, \ldots, f_T$. At every timestep t, the player is asked to choose a point x_t before knowing any information about f_t, and then $f_t(x_t)$ and/or $\nabla f_t(x_t)$ is revealed to the player. The player may use the information to decide on x_{t+1}. This can be thought as an repeated game between the player who chooses x_t and an adversarial who chooses f_t. In general, there is no guarantee about the functions f_t, besides being convex, and Lipschitz or smooth. The goal of online optimization is to minimize the *regret* of the player, defined as

$$R_T := \sum_{t=1}^{T} f_t(x_t) - \min_{x^* \in \mathcal{D}} \sum_{t=1}^{T} f_t(x^*). \tag{1}$$

Although the game seems hard, since the functions can be arbitrary, most classical convex optimization algorithms, such as stochastic gradient descent, works well on the problem (Zinkevich, 2003). The key is that in Equation (1), the regret is defined as the difference between the sum of losses and the minimizer of the sum of the functions, instead of the sum of minimizer of each function. When the functions f_t are very arbitrary, such that they do not have a good common minimizer x^*, then the baseline is already very bad, so we would not be too worse compared to the baseline. When the functions are close to each other with a common minimizer x^*, then most gradients will be pointing to x^*, in which case gradient information will be useful. In the next section, we will see how online SGD achieves a regret of $O(\sqrt{T})$.

2.2 Convergence of Online SGD

We provide the algorithm for online projected stochastic gradient descent in Algorithm 1. We show the convergence rate of Algorithm 1 in Theorem 2.1.

Algorithm 1 Projected stochastic gradient descent algorithm for online convex optimization (Zinkevich, 2003).

Require: Initialize $x_1 \in \mathcal{D}$
  for $t \leftarrow 1, \ldots, T$ **do**
    $g_t \leftarrow$ random vector with $\mathbb{E}[g_t] = \nabla f_t(x_t)$
    $x_{t+1} \leftarrow P_{\mathcal{D}}(x_t - \eta g_t)$
  end for

Theorem 2.1. *Let $\mathcal{D} \subseteq \mathbb{R}^d$ be a convex set and let $f_1, \ldots, f_T : \mathcal{D} \to \mathbb{R}$ be convex functions. Suppose $\|g_t\| \leq G$ for some constant $G > 0$ in Algorithm 1. Then, if we run Algorithm 1 with $\eta = \frac{R}{G\sqrt{T}}$, the expected regret is upper bounded by*

$$R_T \leq RG\sqrt{T}.$$

Proof. By the lower linear bound (Boyd and Vandenberghe, 2004) of convex functions,

$$\begin{aligned} f_t(x_t) - f_t(x^*) &\leq \langle \nabla f_t(x_t), x_t - x^* \rangle \\ &= \mathbb{E}[\langle g_t, x_t - x^* \rangle]. \end{aligned}$$

Then, since $\mathcal{D}$ is convex, for any $x \in \mathbb{R}^d$, $y \in \mathcal{D}$, $\|P_{\mathcal{D}}(x) - y\| \leq \|x - y\|$. Then,

$$\begin{aligned} \|x_{t+1} - x^*\|^2 &= \|P_{\mathcal{D}}(x_t - \eta g_t) - x^*\|^2 \\ &\leq \|x_t - \eta g_t - x^*\|^2 \\ &= \|x_t - x^*\|^2 + \eta^2 \|g_t\|^2 - 2\eta \langle g_t, x_t - x^* \rangle \\ &\leq \|x_t - x^*\|^2 + \eta^2 G^2 + 2\eta (f_t(x^*) - f_t(x_t)). \end{aligned}$$

Rearranging terms we get

$$f_t(x_t) - f_t(x^*) \leq \frac{1}{2\eta} \big(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 + \eta^2 G^2 \big).$$

Then, summing up and taking expectation, we have

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}[f_t(x_t)] - \sum_{t=1}^{T} f_t(x^*) &\leq \frac{1}{2\eta} (\|x_0 - x^*\|^2 - \|x_{T+1} - x^*\|^2 + \eta^2 G^2 T) \\ &\leq \frac{\|x_0 - x^*\|^2}{2\eta} + \frac{\eta G^2 T}{2} \\ &\leq \frac{R^2}{2\eta} + \frac{\eta G^2 T}{2}. \end{aligned}$$

Let $\eta = \frac{R}{G\sqrt{T}}$, we have $R_T \leq RG\sqrt{T}$. □

3 Zeroth-Order Online Convex Optimization

In zeroth-order online convex optimization, access to the gradient is no longer feasible. At every timestep, only the function value $f_t(x_t)$ is observed, and we need to pick x_{t+1} immediately without any further information. In the offline setting, the problem is not particularly hard, since there are various ways to approximate the geometry of the function. For example, we can approximate the partial derivative of f at x by

$$\langle \nabla f(x), e_i \rangle = \lim_{\delta \to 0} \frac{f(x + \delta e_i) - f(x)}{\delta} \approx \frac{f(x + \delta e_i) - f(x)}{\delta}$$

for some sufficiently small delta. However, in $\mathbb{R}^d$, we would need d estimates to obtain an estimate of the full gradient, but this is impossible in online optimization, since once we observed $f_t(x_t)$, the function changes to f_{t+1}, and we can no longer obtain any information about f_t.

Flaxman et al. (2005) provides a randomized solution to approximate the stochastic gradient of the objective function with only one point feedback. The key is to use Stoke's theorem to formulate the relationship between the gradient and function value over distributions on the unit ball and sphere. Formally, let $\mathcal{S}$ be the unit sphere and $\mathcal{B}$ be the unit closed ball, by Stoke's theorem, we have

$$\nabla \int_{\delta\mathcal{B}} f(x+v)\ dv = \int_{\delta\mathcal{S}} f(x+u)\frac{u}{\|u\|}\ du$$

Then, if we define $\widehat{f}(x_t) = \mathbb{E}_{u\sim\mathcal{B}}[f(x+\delta u)]$, we have

$$\nabla \widehat{f}(x_t) = \mathop{\mathbb{E}}_{u\sim\mathcal{S}}\left[\frac{d}{\delta} f(x+\delta u)u\right]. \tag{2}$$

When f is Lipschitz, $\widehat{f}$ is close to f since it is the "average" of f in a small neighborhood of f, so if we optimize $\widehat{f}$, we also approximately optimize f. This leads to Algorithm 2.

Algorithm 2 Algorithm for zeroth-order online convex optimization with $O(T^{3/4})$ regret by Flaxman et al. (2005).

for $t \leftarrow 1, \ldots, T$ **do**
  Draw $u_t \sim \mathcal{S}$ uniformly at random
  $g_t \leftarrow \frac{d}{\delta} f_t(x_t + \delta u_t)u_t$
  $x_{t+1} \leftarrow P_{\mathcal{D}}(x_t - \eta g_t)$
end for

We make the following assumptions. In the online convex optimization setting, we have a convex and compact constraint set $\mathcal{D} \subseteq \mathbb{R}^d$, with $r\mathcal{B} \subseteq \mathcal{D} \subseteq R\mathcal{B}$. We are given a sequence of functions $f_1, \ldots, f_T : \mathcal{D} \to \mathbb{R}$, which are all convex and L-Lipschitz. In addition, each f_t is bounded by $\|f_t\|_\infty := \sup_{x\in\mathcal{D}}(x) \le M$. If the assumptions are satisfied, we have Theorem 3.1 that establishes the convergence rate for Algorithm 2.

Theorem 3.1. *If the assumptions hold, then the expected regret of Algorithm 2 is upper bounded by*

$$R_T \le 2T^{3/4}\sqrt{(2+1/r)RMdL} = O(T^{3/4}).$$

Lemma 3.2. *Let* $\widehat{f}_t(x_t) = \mathbb{E}_{u_t\sim\mathcal{B}}[f_t(x_t+\delta u_t)]$, *then* $\mathbb{E}[g_t] = \nabla\widehat{f}(x_t)$.

Proof. Using the fact that $\text{vol}(\delta\mathcal{B}) = \frac{\delta}{d}\text{vol}(\delta\mathcal{S})$, we have

$$\begin{aligned}\mathbb{E}[g_t] &= \frac{1}{\text{vol}(\delta\mathcal{S})}\frac{d}{\delta}\int_{\delta\mathcal{S}} f_t(x_t+u)\frac{u}{\|u\|}\,du \\ &= \frac{1}{\text{vol}(\delta\mathcal{B})}\nabla\int_{\delta\mathcal{B}} f_t(x_t+u)\,du \\ &= \nabla\widehat{f}_t(x_t).\end{aligned}$$

□

Lemma 3.3. *The optimum in $(1-\delta/r)\mathcal{D}$ satisfy*

$$\min_{x^*\in(1-\delta/r)\mathcal{D}}\sum_{t=1}^{T} f_t(x^*) \le \frac{\delta LT}{r} + \min_{x^*\in\mathcal{D}}\sum_{t=1}^{T} f_t(x^*).$$

Proof. Since f_t is L-Lipschitz, $\sum_{t=1}^{T} f_t$ is LT-Lipschitz, so let

$$x^* := \arg\min_{x\in\mathcal{D}}\sum_{t=1}^{T} f_t(x),$$

we have

$$\begin{aligned}\min_{x^*\in(1-\delta/r)\mathcal{D}}\sum_{t=1}^{T} f_t(x^*) &\le \sum_{t=1}^{T} f_t(P_{(1-\delta/r)\mathcal{D}}(x^*)) \\ &\le \sum_{t=1}^{T} f_t(x^*) + \sum_{t=1}^{T}|f_t(P_{(1-\delta/r)\mathcal{D}}(x^*)) - f_t(x^*)| \\ &\le \sum_{t=1}^{T} f_t(x^*) + TL\|P_{(1-\delta/r)\mathcal{D}}(x^*) - x^*\| \\ &\le \sum_{t=1}^{T} f_t(x^*) + \frac{\delta LT}{r}.\end{aligned}$$

□

Lemma 3.4. *For any point $x \in (1-\delta/r)\mathcal{D}$, $x + \delta\mathcal{B} \subseteq \mathcal{D}$.*

Proof. This follows from

$$(1-\delta/r)\mathcal{D} + \delta\mathcal{B} \subseteq (1-\delta/r)\mathcal{D} + (\delta/r)\mathcal{D} \subseteq \mathcal{D}.$$

□

Lemma 3.5. *For any $x \in \mathcal{D}$, $|\widehat{f}_t(x) - f_t(x)| \le \delta L$.*

Proof. Since f_t is L-Lipschitz, we have

$$\begin{aligned}
|\widehat{f}_t(x) - f_t(x)| &= \left| \frac{1}{\mathrm{vol}(\delta\mathcal{B})} \int_{\delta\mathcal{B}} f(x+u)\, du - f(x) \right| \\
&\le \frac{1}{\mathrm{vol}(\delta\mathcal{B})} \int_{\delta\mathcal{B}} |f(x+u) - f(x)|\, du \\
&\le \frac{1}{\mathrm{vol}(\delta\mathcal{B})} \int_{\delta\mathcal{B}} L\|u\|\, du \\
&\le \frac{1}{\mathrm{vol}(\delta\mathcal{B})} \int_{\delta\mathcal{B}} \delta L\, du \\
&\le \delta L.
\end{aligned}$$

□

Proof of Theorem 3.1. We follow the proof of the general online gradient descent. We first bound $\|g_t\|$ using the upper bound of $|f_t|$

$$\|g_t\| = \left\| \frac{d}{\delta} f_t(x_t + \delta u_t) u_t \right\| \le \frac{Md}{\delta}.$$

Let $G := \frac{Md}{\delta}$. By Theorem 2.1, we have the regret of $\{\widehat{f}_t\}$ is bounded by

$$\widehat{R}_T \le RG\sqrt{T} = \frac{RMd\sqrt{T}}{\delta}.$$

Then by Lemma 3.5, we have

$$\begin{aligned}
R_T &= \sum_{t=1}^{T} f_t(x_t) - \min_{x^* \in \mathcal{D}} \sum_{t=1}^{T} f_t(x^*) \\
&\le \sum_{t=1}^{T} f_t(x_t) - \min_{x^* \in \mathcal{D}} \sum_{t=1}^{T} \widehat{f}_t(x^*) + \delta LT \\
&\le \sum_{t=1}^{T} f_t(x_t) - \min_{x^* \in (1-\delta/r)\mathcal{D}} \sum_{t=1}^{T} \widehat{f}_t(x^*) + (1+1/r)\delta LT \\
&= \widehat{R}_T + \sum_{t=1}^{T} |\widehat{f}_t(x_t) - f_t(x_t)| + (1+1/r)\delta LT \\
&\le \widehat{R}_T + (2+1/r)\delta LT \\
&\le \frac{RMd\sqrt{T}}{\delta} + (2+1/r)\delta LT.
\end{aligned}$$

Choosing $\delta = T^{-1/4}\sqrt{\frac{RMd}{(2+1/r)L}}$, we have

$$R_T \le 2T^{3/4}\sqrt{(2+1/r)RMdL} = O(T^{3/4}).$$

□

3.1 Two-Point Estimates

One potential problem with the bound in Theorem 3.1 is that it is dependent on the upper bound for $\|f_t\|_\infty$. When the function values are too large, the variance of the stochastic gradient will be large, which affects the performance of the algorithm. In particular, an adversarial could simply add a constant term to each function to slow down the convergence of the algorithm, without changing the regret objective. Such assumption is not realistic and is typically unnecessary in general convex optimization problems.

Agarwal and Dekel (2010) shows that when we have two-point estimates of the gradient, we can improve the estimation of the stochastic gradient and remove the dependence on $\|f_t\|_\infty$. The key idea is that we can approximate the gradient by

$$\nabla f(x) \approx \mathop{\mathbb{E}}_{u\sim\mathcal{S}}\left[\frac{d}{2\delta}(f(x+\delta u) - f(x-\delta u))u\right].$$

Algorithm 3 Algorithm for zeroth-order online convex optimization with two point feedback, with $O(\sqrt{T})$ regret by Agarwal and Dekel (2010).

for $t \leftarrow 1, \ldots, T$ **do**
  Draw $u_t \sim \mathcal{S}$ uniformly at random
  $g_t \leftarrow \frac{d}{2\delta}(f_t(x_t + \delta u_t) - f_t(x_t - \delta u_t))u_t$
  $x_{t+1} \leftarrow P_{\mathcal{D}}(x_t - \eta g_t)$
end for

Theorem 3.6. *If the assumptions hold, with infinitesimal δ, the expected regret of Algorithm 3 is upper bounded by*

$$R_T \leq RLd\sqrt{T}.$$

In particular, with $\delta = O(\frac{1}{\sqrt{T}})$, we have $R_T = O(\sqrt{T})$.

Proof. We bound $\|g_t\|$ by

$$\begin{aligned}
\|g_t\| &= \left\|\frac{d}{2\delta}(f_t(x_t+\delta u) - f_t(x_t-\delta u))u\right\| \\
&\leq \frac{d}{2\delta}|f_t(x_t+\delta u) - f_t(x_t-\delta u)| \\
&\leq \frac{d}{2\delta}L\|2\delta u\| \\
&= Ld.
\end{aligned}$$

Then, let $G := Ld$, by Theorem 2.1, we have the regret of $\{\widehat{f}_t\}$ is bounded by

$$\widehat{R}_T \leq RG\sqrt{T} = RLd\sqrt{T}.$$

Then, similar to the proof of Theorem 3.1, we have

$$\begin{aligned} R_T &\leq \widehat{R}_T + (2 + 1/r)\delta LT \\ &\leq RLd\sqrt{T} + (2 + 1/r)\delta LT. \end{aligned}$$

Choosing $\delta \to 0^+$, we have

$$R_T \leq RLd\sqrt{T}.$$

In particular, it suffices to choose $\delta = O(\frac{1}{\sqrt{T}})$ to get $R_T = O(\sqrt{T})$. □

When we have two point feedback, we can see that we can get nearly as good as when we have the full gradient, in which case we achieve a regret of at most $RL\sqrt{T}$. There is only an additional multiplier of d and an extra term that is negligible if we choose sufficiently small δ.

3.2 Smooth Objective Function

The gradient estimation given by Flaxman et al. (2005) is a generic idea that can be used to turn first-order online convex optimization algorithms to zeroth-order algorithms. Similar ideas can be applied to other convex optimization algorithms. Saha and Tewari (2011) shows how the algorithm can be applied to an interior point method proposed by Abernethy et al. (2008); Abernethy and Rakhlin (2009). Specifically, in the original algorithm, a vector is sampled from the unit sphere, while in the algorithm by Saha and Tewari (2011), a ν-self-concordant barrier function R is used and the vector is sampled from an ellipsoid based on R. The algorithm is described in Algorithm 4.

Algorithm 4 Zeroth-order online convex optimization algorithm for smooth functions by Saha and Tewari (2011).

Require: R is a ν-self-concordant barrier function for $\mathcal{D}$
  for $t \leftarrow 1, \ldots, T$ **do**
    $A_t \leftarrow \sqrt{(\nabla^2 R(x_t))^{-1}}$
    Draw $u_t \sim \mathcal{S}$ uniformly at random
    $g_t \leftarrow \frac{d}{\delta} f_t(x_t + \delta A_t u_t) A_t^{-1} u_t$
    $x_{t+1} \leftarrow \arg\min_{x \in \mathcal{D}} \eta \sum_{s=1}^{t} \langle g_s, x \rangle + R(x)$
  end for

We need an additional assumption that f is L-smooth.

Definition 3.7. A convex function $f : \mathcal{D} \to \mathbb{R}$ is *L-smooth* if for all $x, y \in \mathcal{D}$,

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

Lemma 3.8. *Let* $\widehat{f}_t(x_t) = \mathbb{E}_{u \sim \mathcal{B}}[f_t(x_t + \delta A_t u)]$, *then* $\mathbb{E}[g_t] = \nabla \widehat{f}(x_t)$.

Proof. We define $F_t(x) := f_t(A_t x)$, which allows us to apply the chain rule.

$$\begin{aligned}
\mathbb{E}[g_t] &= \mathop{\mathbb{E}}_{u\sim\mathcal{S}}\left[\frac{d}{\delta} f_t(x_t + \delta A_t u) A_t^{-1} u\right] \\
&= A_t^{-1} \mathop{\mathbb{E}}_{u\sim\mathcal{S}}\left[\frac{d}{\delta} f_t(x_t + \delta A_t u) u\right] \\
&= A_t^{-1} \mathop{\mathbb{E}}_{u\sim\mathcal{S}}\left[\frac{d}{\delta} F_t(A_t^{-1} x_t + \delta u) u\right] \\
&= A_t^{-1} \widehat{F}_t(A_t^{-1} x_t) \\
&= A_t^{-1} A_t \widehat{f}_t(x_t) \\
&= \widehat{f}_t(x_t).
\end{aligned}$$

□

Theorem 3.9. *Assume that $f_1, \ldots, f_T : \mathcal{D} \to \mathbb{R}$ are L-smooth, with $\|f_t\|_\infty \le M$, then running Algorithm 4 with appropriate choices of η, δ, the expected regret is upper bounded by*

$$R_t \le 3(L\nu \log T)^{1/3} (MdR)^{2/3} T^{2/3} + \left(\frac{2M}{R} + RL\right)\sqrt{T} = \tilde{O}(T^{2/3}).$$

We will not show the proof of the theorem here, and will refer the reader to the original paper by Saha and Tewari (2011). But notice that $\tilde{O}(T^{2/3})$ is a significant improvement compared to the $O(T^{3/4})$ bound. It is a question that whether we can use this idea to turn other first-order online convex optimization algorithm into bandit algorithms and achieve better regret bound.

3.3 Application in Game Theory

Multi-person repeated concave games can be thought as a special case for online learning. We assume that each player has a fixed concave utility function $u_i : \mathcal{A} \to \mathbb{R}$ and the set of all possible actions $\mathcal{A} = \prod_{i=1}^N \mathcal{A}_i$ is convex. Then, notice that while the utility functions are fixed, the players can only control their own actions, but not the actions of other players. Then, for each player i, we can define

$$f_{t,i}(x_{t,i}) = u_i(x_{t,i}, x_{t,-i})$$

which is concave. Hence, the players can use the online learning framework as a strategy to find actions. When the game is unknown to the player, it is a zeroth-order online optimization problem, since gradient information is not available. Then, the algorithm by Flaxman et al. (2005) can be applied to solve for future actions. Bravo et al. (2018) shows that for concave N-person monotone games, if each player chooses his or her action according to the algorithm by Flaxman et al. (2005), the game would converge to a Nash equilibrium.

4 Discussion

The bandit optimization is a topic that I wanted to look into since I learned online optimization in the convex optimization course at CMU. I was surprised that there can be such a simple way of using first-order algorithms in the zeroth-order setting. The proof was also simple and elegant and can be summarized in a few pages. While I was looking for examples of online learning, I realized that many natural applications of online learning does not involve the definition of a differentiable model like in machine learning, and more of them are like "bandits." As a result, the bandit optimization framework can be applied to a variety of such problems. I was also excited to see that there is a connection between bandit optimization and game theory. I realized that online learning is a very generic and powerful setting, and it would always be a great research idea to try the online learning framework when studying sequential decision making problems.

5 Conclusion

In conclusion, this report provides a comprehensive review of important algorithms in zeroth-order online convex optimization. We started with an overview of the online convex optimization framework and demonstrated how stochastic projected gradient descent can achieve $O(\sqrt{T})$ regret in the online setting. We then analyzed the algorithm of Flaxman et al. (2005), which approximates the gradient with function value at one point using Stoke's theorem and proved its expected regret of $O(T^{3/4})$. We also explored how similar ideas can be applied to slightly different settings, leading to more optimal algorithms in the two-point feedback and smooth settings. Finally, we discussed an application of the algorithm in game theory described by Bravo et al. (2018). This report highlights the significance of online convex optimization in machine learning and provides insights into various algorithms that can be used to optimize online learning problems in the absence of gradient information.

References

Abernethy, J., Hazan, E., and Rakhlin, A. (2008). Competing in the dark: An efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory, COLT 2008*.

Abernethy, J. and Rakhlin, A. (2009). Beating the adaptive bandit with high probability. In *2009 Information Theory and Applications Workshop*, pages 280–289. IEEE.

Agarwal, A. and Dekel, O. (2010). Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pages 28–40. Citeseer.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Bravo, M., Leslie, D., and Mertikopoulos, P. (2018). Bandit learning in concave n-person games. *Advances in Neural Information Processing Systems*, 31.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7).

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2005). Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Saha, A. and Tewari, A. (2011). Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 636–642. JMLR Workshop and Conference Proceedings.

Shalev-Shwartz, S. et al. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936.