JMLR: Workshop and Conference Proceedings vol 40:1–13, 2015

# Bandit Convex Optimization: $\sqrt{T}$ Regret in One Dimension

**Sébastien Bubeck** SEBUBECK@MICROSOFT.COM
**Ofer Dekel** OFERD@MICROSOFT.COM
*Microsoft Research, 1 Microsoft Way, Redmond, WA 98052, USA*

**Tomer Koren** TOMERK@TECHNION.AC.IL
*Technion—Israel Institute of Technology, Haifa 32000, Israel*

**Yuval Peres** PERES@MICROSOFT.COM
*Microsoft Research, 1 Microsoft Way, Redmond, WA 98052, USA*

## Abstract

We analyze the minimax regret of the adversarial bandit convex optimization problem. Focusing on the one-dimensional case, we prove that the minimax regret is $\widetilde{\Theta}(\sqrt{T})$ and partially resolve a decade-old open problem. Our analysis is non-constructive, as we do not present a concrete algorithm that attains this regret rate. Instead, we use minimax duality to reduce the problem to a Bayesian setting, where the convex loss functions are drawn from a worst-case distribution, and then we solve the Bayesian version of the problem with a variant of Thompson Sampling. Our analysis features a novel use of convexity, formalized as a "local-to-global" property of convex functions, that may be of independent interest.

## 1. Introduction

Online convex optimization with bandit feedback, commonly known as bandit convex optimization, can be described as a $T$-round game, played by a randomized player in an adversarial environment. Before the game begins, the adversarial environment chooses an arbitrary sequence of $T$ bounded convex functions $f_1, \ldots, f_T$, where each $f_t : \mathcal{K} \mapsto [0,1]$ and $\mathcal{K}$ is a fixed convex and compact set in $\mathbb{R}^n$. On round $t$ of the game, the player chooses a point $X_t \in \mathcal{K}$ and incurs a loss of $f_t(X_t)$. The player observes the value of $f_t(X_t)$ and nothing else, and she uses this information to improve her choices going forward. The player's performance is measured in terms of her $T$-round *regret*, defined as $\sum_{t=1}^T f_t(X_t) - \min_{x\in\mathcal{K}} \sum_{t=1}^T f_t(x)$. In words, the regret compares the player's cumulative loss to that of the best fixed point in hindsight.

While regret measures the performance of a specific player against a specific loss sequence, the inherent difficulty of the game is measured using the notion of *minimax regret*. Informally, the game's minimax regret is the regret of an optimal player when she faces the worst-case loss sequence. Characterizing the minimax regret of bandit convex optimization is one of the most elusive open problems in the field of online learning. For general bounded convex loss functions, Flaxman et al. (2005) presents an algorithm that guarantees a regret of $\widetilde{O}(T^{5/6})$—and this is the best known upper bound on the minimax regret of the game. Better regret rates can be guaranteed if additional assumptions are made: for Lipschitz functions the regret is $\widetilde{O}(T^{3/4})$ (Flaxman et al., 2005), for Lipschitz and strongly convex losses the regret is $\widetilde{O}(T^{2/3})$ (Agarwal et al., 2010), and for smooth functions the regret is $\widetilde{O}(T^{2/3})$ (Saha and Tewari, 2011). In all of the aforementioned settings, the best known lower bound on minimax regret is $\Omega(\sqrt{T})$ (Dani et al., 2008), and the challenge is to

bridge the gap between the upper and lower bounds. In a few special cases, the gap is resolved and we know that the minimax regret is exactly $\widetilde{\Theta}(\sqrt{T})$; specifically, when the loss functions are both smooth and strongly-convex (Hazan and Levy, 2014), when they are Lipschitz and linear (Dani et al., 2008; Abernethy et al., 2008), or when they are Lipschitz and drawn i.i.d. from a fixed and unknown distribution (Agarwal et al., 2011).

In this paper, we resolve the open problem in the one-dimensional case, where $\mathcal{K} = [0, 1]$, by proving that the minimax regret with arbitrary bounded convex loss functions is $\widetilde{\Theta}(\sqrt{T})$. Formally, we prove the following theorem.

**Theorem 1** (main result)**.** *There exists a randomized player strategy that relies on bandit feedback and guarantees an expected regret of $O(\sqrt{T}\log T)$ against any sequence of convex loss functions $f_1, \ldots, f_T : [0, 1] \mapsto [0, 1]$.*

The one-dimensional case has received very little special attention, and the best published result is the $\widetilde{O}(T^{5/6})$ bound mentioned above, which holds in any dimension. However, by discretizing the domain $[0, 1]$ appropriately and applying a standard multi-armed bandit algorithm, one can prove a tighter bound of $\widetilde{O}(T^{2/3})$; see Bubeck et al. (2015) for details. It is worth noting that replacing the convexity assumption with a Lipschitz assumption also gives an upper bound of $\widetilde{O}(T^{2/3})$ (Kleinberg, 2004). However, obtaining the tight upper bound of $\widetilde{O}(\sqrt{T})$ requires a more delicate analysis, which is the main focus of this paper.

Our tight upper bound is non-constructive, in the sense that we do not describe an algorithm that guarantees a $\widetilde{O}(\sqrt{T})$ regret for any loss sequence. Instead, we use minimax duality to reduce the problem of bounding the adversarial minimax regret to the problem of upper bounding the analogous *maximin regret* in a *Bayesian* setting. Unlike our original setting, where the sequence of convex loss functions is chosen adversarially, the loss functions in the Bayesian setting are drawn from a probability distribution, called the *prior*, which is known to the player. The idea of using minimax duality to study minimax regret is not new (see, e.g., Abernethy et al., 2009; Gravin et al., 2014); however, to the best of our knowledge, we are the first to apply this technique to prove upper bounds in a bandit feedback scenario.

After reducing our original problem to the Bayesian setting, we design a novel algorithm for Bayesian bandit convex optimization (in one dimension) that guarantees $\widetilde{O}(\sqrt{T})$ regret for any prior distribution. Since our main result is non-constructive to begin with, we are not at all concerned with the computational efficiency of this algorithm. We first discretize the domain $[0, 1]$ and treat each discrete point as an arm in a multi-armed bandit problem. We then apply a variant of the classic Thompson Sampling strategy (Thompson, 1933) that is designed to exploit the fact that the loss functions are all convex. We adapt the analysis of Thompson Sampling in Russo and van Roy (2014) to our algorithm and extend it to arbitrary joint prior distributions over sequences of loss functions (not necessarily i.i.d. sequences).

The significance of the convexity assumption is that it enables us to obtain regret bounds that scale *logarithmically* with the number of arms, which turns out to be the key property that leads to the desired $\widetilde{O}(\sqrt{T})$ upper bound. Intuitively, convexity ensures that a change to the loss value of one arm influences the loss values in many of the adjacent arms. Therefore, even the worst case prior distribution cannot hide a small loss in one arm without globally influencing the loss of many other arms. Technically, this aspect of our analysis boils down to a basic question about convex functions: given two convex functions $f : \mathcal{K} \mapsto [0, 1]$ and $g : \mathcal{K} \mapsto [0, 1]$ such that $f(x) < \min_y g(y)$ at some point $x \in \mathcal{K}$, how small can $\|f - g\|$ be (where $\|\cdot\|$ is an appropriate norm over the function space)?

In other words, if two convex functions differ locally, how similar can they be globally? We give an answer to this question in the one-dimensional case.

The paper is organized as follows. We begin in Section 2 where we define the setting of Bayesian online optimization, establish basic techniques for the analysis of Bayesian online algorithms, and demonstrate how to readily recover some of the known minimax regret bounds for the full information case by bounding the Bayesian regret. Then, in Section 3, we prove the key structural lemma by which we exploit the convexity of the loss functions. Section 4 is the main part of the paper, where we give our algorithm for Bayesian bandit convex optimization (in one dimension) and analyze its regret. We conclude the paper in Section 5 with a few remarks and open problems.

## 2. From Adversarial to Bayesian Regret

In this section, we show how regret bounds for an adversarial online optimization setting can be obtained via a Bayesian analysis. Before explaining this technique in detail, we first formalize two variants of the online optimization problem: the adversarial setting and the Bayesian setting.

We begin with the standard, adversarial online optimization setup. As described above, in this setting the player plays a $T$-round game, during which he chooses a sequence of points $X_{1:T}$,[1] where $X_t \in \mathcal{K}$ for all $t$. The player's randomized policy for choosing $X_{1:T}$ is defined by a sequence of deterministic functions $\rho_{1:T}$, where each $\rho_t : [0,1]^{2(t-1)} \mapsto \Delta(\mathcal{K})$ (here $\Delta(\mathcal{K})$ is the set of probability distributions over $\mathcal{K}$). On round $t$, the player uses $\rho_t$ and her past observations to define the probability distribution

$$\pi_t \;=\; \rho_t\big(X_1, f_1(X_1), \ldots, X_{t-1}, f_{t-1}(X_{t-1})\big)\;,$$

and then draws a concrete point $X_t \sim \pi_t$. Even though $\rho_t$ is a deterministic function, the probability distribution $\pi_t$ is itself a random variable, because it depends on the player's past random actions $X_1, \ldots, X_{t-1}$.

The player's cumulative loss at the end of the game is the random quantity $\sum_{t=1}^{T} f_t(X_t)$ and her expected regret against the sequence $f_{1:T}$ is

$$R(\rho_{1:T}; f_{1:T}) \;=\; \mathbb{E}\left[\sum_{t=1}^{T} f_t(X_t)\right] \;-\; \min_{x\in\mathcal{K}} \sum_{t=1}^{T} f_t(x)\;.$$

The difficulty of the game is measured by its minimax regret, defined as

$$\min_{\rho_{1:T}} \sup_{f_{1:T}} \; R(\rho_{1:T}; f_{1:T})\;.$$

We now turn to introduce the Bayesian online optimization setting. In the Bayesian setting, we assume that the sequence of loss functions $F_{1:T}$, where each $F_t : \mathcal{K} \mapsto [0,1]$ is convex, is drawn from a probability distribution $\mathcal{F}$ called the *prior distribution*. Note that $\mathcal{F}$ is a distribution over the entire sequence of losses, and not over individual functions in the sequence. Therefore, it can encode arbitrary dependencies between the loss functions on different rounds. However, we assume that this distribution is known to the player, and can be used to design her policy. The player's

1. Throughout the paper, we use the notation $a_{s:t}$ as shorthand for the sequence $a_s, \ldots, a_t$.

*Bayesian regret* is defined as

$$R(\rho_{1:T}; \mathcal{F}) \;=\; \mathbb{E}\left[\sum_{t=1}^{T} F_t(X_t) - \sum_{t=1}^{T} F_t(X^\star)\right] \;,$$

where $X^\star$ is the point in $\mathcal{K}$ with the smallest cumulative loss at the end of the game, namely the random variable

$$X^\star \;=\; \operatorname*{arg\,min}_{x \in \mathcal{K}} \sum_{t=1}^{T} F_t(x) \;. \tag{1}$$

The difficulty of online optimization in a Bayesian environment is measured using the maximin Bayesian regret, defined as

$$\sup_{\mathcal{F}} \min_{\rho_{1:T}} \; R(\rho_{1:T}; \mathcal{F}) \;.$$

In words, the maximin Bayesian regret is the regret of an optimal Bayesian strategy over the worst possible prior $\mathcal{F}$.

It turns out that the two online optimization settings we described above are closely related. The following theorem, which is a consequence of a generalization of the von Neumann minimax theorem, shows that the minimax adversarial regret and maximin Bayesian regret are equal.

**Theorem 2.** *It holds that*

$$\min_{\rho_{1:T}} \sup_{f_{1:T}} \; R(\rho_{1:T}; f_{1:T}) \;=\; \sup_{\mathcal{F}} \min_{\rho_{1:T}} \; R(\rho_{1:T}; \mathcal{F}) \;.$$

For completeness, we include a proof of this fact in Bubeck et al. (2015). As a result, instead of analyzing the minimax regret directly, we can analyze the maximin Bayesian regret. That is, our new goal is to design a prior-dependent player policy that guarantees a small regret against *any* prior distribution $\mathcal{F}$.

### 2.1. Bayesian Analysis with Full Feedback

As a warm-up, we first consider the Bayesian setting where the player receives full-feedback. Namely, on round $t$, after the player draws a point $X_t \sim \pi_t$ and incurs a loss of $F_t(X_t)$, we assume that she observes the entire loss function $F_t$ as feedback. We show how minimax duality can be used to recover the known $O(\sqrt{T})$ regret bounds for this setting. For simplicity, we focus on the concrete setting where $\mathcal{K} = \Delta_n$ (the $n$-dimensional simplex), and where the convex loss functions $F_{1:T}$ are also 1-Lipschitz with respect to the $L_1$-norm (with probability one).

The evolution of the game is specified by a filtration $\mathcal{H}_{1:T}$, where each $\mathcal{H}_t$ denotes the history observed by the player up to and including round $t$ of the game; formally, $\mathcal{H}_t$ is the sigma-field generated by the random variables $X_{1:t}$ and $F_{1:t}$. To simplify notations, we use the shorthand $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{H}_{t-1}]$ to denote expectation conditioned on the history before round $t$. The analogous shorthands $\mathbb{P}_t(\cdot)$ and $\mathrm{Var}_t(\cdot)$ are defined similarly.

Recall that the player's policy can rely on the prior $\mathcal{F}$. A natural deterministic policy is to choose, based on the random variable $X^\star$ defined in Eq. (1), actions $X_{1:T}$ according to

$$\forall\, t \in [T] \;, \qquad X_t \;=\; \mathbb{E}_t[X^\star] \;. \tag{2}$$

In other words, the player uses her knowledge of the prior and her observations so far to calculate a posterior distribution over loss functions, and then chooses the expected best-point-in-hindsight. Notice that the sequence $X_{1:T}$ is a martingale (in fact, a Doob martingale), whose elements are vectors in the simplex.

The following lemma shows that the expected instantaneous (Bayesian) regret of the strategy on each round $t$ can be upper bounded in terms of the variation of the sequence $X_{1:T}$ on that round.

**Lemma 3.** *Assume that with probability one, the loss functions $F_{1:T}$ are convex and 1-Lipschitz with respect to some norm $\|\cdot\|$. Then the strategy defined in Eq. (2) guarantees $\mathbb{E}[F_t(X_t) - F_t(X^\star)] \le \mathbb{E}[\|X_t - X_{t+1}\|]$ for all $t$.*

*Proof.* By the subgradient inequality, we have $F_t(X_t) - F_t(X^\star) \le \nabla F_t(X_t) \cdot (X_t - X^\star)$ for all $t$. The Lipschitz assumption implies that $\|\nabla F_t(X_t)\|_* \le 1$, where $\|\cdot\|_*$ is the norm dual of $\|\cdot\|$. Using Eq. (2), noting that $X_t, F_t \in \mathcal{H}_t$, and taking the conditional expectation, we get

$$\mathbb{E}_{t+1}\left[F_t(X_t) - F_t(X^\star)\right] \;\le\; \nabla F_t(X_t) \cdot (X_t - \mathbb{E}_{t+1}[X^\star]) \;=\; \nabla F_t(X_t) \cdot (X_t - X_{t+1}) \;.$$

Finally, applying Holder's inequality on the right-hand side and taking expectations proves the lemma. □

To bound the total variation of $X_{1:T}$, we use a bound of Neyman (2013) on the total variation of martingales in the simplex.

**Lemma 4** (Neyman, 2013)**.** *For any martingale $Z_1, \ldots, Z_{T+1}$ in the $n$-dimensional simplex, one has*

$$\mathbb{E}\left[\sum_{t=1}^{T} \|Z_t - Z_{t+1}\|_1\right] \;\le\; \sqrt{\tfrac{1}{2} T \log n}\;.$$

Lemma 4 and Lemma 3 together yield a $O(\sqrt{T \log n})$ bound on the maximin Bayesian regret of online convex optimization on the simplex with full-feedback. Theorem 2 then implies the same bound over the minimax regret in the corresponding adversarial setting, recovering the well-known bounds in this case (e.g., Kivinen and Warmuth, 1997). We remark that essentially the same technique can be used to retrieve known dimension-free regret bounds in the Euclidean setting, e.g., when $\mathcal{K}$ is an Euclidean ball and the losses are Lipschitz with respect to the $L_2$ norm; in this case, the $L_2$ total variation of the martingale $X_{1:T}$ can be shown to be bounded by $O(\sqrt{T})$ with no dependence on $n$.[2]

### 2.2. Regret Analysis of Bayesian Bandits

The analysis in this section builds on the technique introduced by Russo and van Roy (2014). While their analysis is stated for prior distributions that are i.i.d. (namely, $\mathcal{F}$ is a product distribution),[3] we show that it extends to arbitrary prior distributions with essentially no modifications.

2. This follows from the fact that a martingale in $\mathbb{R}^n$ can always be projected to a martingale in $\mathbb{R}^2$ with the same magnitude of increments; namely, given a martingale $Z_1, Z_2, \ldots$ in $\mathbb{R}^n$ one can show that there exists a martingale sequence $\widetilde{Z}_1, \widetilde{Z}_2, \ldots$ in $\mathbb{R}^2$ such that $\|Z_t - Z_{t+1}\|_2 = \|\widetilde{Z}_t - \widetilde{Z}_{t+1}\|_2$ for all $t$.
3. More precisely, the distributions Russo and van Roy (2014) consider are i.i.d. only when conditioned on the true outcome distribution, which is itself chosen at random.

We begin by restricting our attention to finite decision sets $\mathcal{K}$, and denote $K = |\mathcal{K}|$. (When we get to the analysis of Bayesian bandit convex optimization, $\mathcal{K}$ will be an appropriately chosen grid of points in $[0,1]$.) In the bandit case, the history $\mathcal{H}_t$ is the sigma-field generated by the random variables $X_{1:t}$ and $F_1(X_1), \ldots, F_t(X_t)$. Following Russo and van Roy (2014), we consider the following quantities related to the filtration $\mathcal{H}_{1:T}$:

$$\forall\, x \in \mathcal{K}\,, \qquad \begin{aligned} r_t(x) &= \mathbb{E}_t\big[F_t(x) - F_t(X^\star)\big]\,, \\ v_t(x) &= \mathrm{Var}_t\big(\mathbb{E}_t[F_t(x) \mid X^\star]\big)\,. \end{aligned} \tag{3}$$

The random quantity $r_t(x)$ is the expected regret incurred by playing the point $x$ on round $t$, conditioned on the history. Hence, the cumulative expected regret of the player equals $\mathbb{E}[\sum_{t=1}^T r_t(X_t)]$. The random variable $v_t(x)$ is a proxy for the information revealed about $X^\star$ by choosing the point $x$ on round $t$. Intuitively, if the value of $F_t(x)$ varies significantly as a function of the random variable $X^\star$, then observing the value of $F_t(x)$ should reveal much information on the identity of $X^\star$. (More precisely, $v_t(x)$ is the amount of variance in $F_t(x)$ explained by the random variable $X^\star$.)

The following lemma can be viewed as an analogue of Lemma 4 in the bandit setting.

**Lemma 5.** *For any player strategy and any prior distribution $\mathcal{F}$, it holds that*

$$\mathbb{E}\left[\sum_{t=1}^T \sqrt{\mathbb{E}_t[v_t(X_t)]}\right] \;\leq\; \sqrt{\tfrac{1}{2} T \log K}\,.$$

The proof uses tools from information theory to relate the quantity $v_t(X_t)$ to the *decrease in entropy* of the random variable $X^\star$ due to the observation on round $t$; the total decrease in entropy is necessarily bounded, which gives the bound in the lemma. For completeness, we give a proof in Bubeck et al. (2015).

Lemma 5 suggests a generic way of obtaining regret bounds for Bayesian algorithms: first bound the instantaneous regret $\mathbb{E}_t[r_t(X_t)]$ of the algorithm in terms of $\sqrt{\mathbb{E}_t[v_t(X_t)]}$ for all $t$, then sum the bounds and apply the lemma. Russo and van Roy (2014) refer to the ratio $\mathbb{E}_t[r_t(X_t)]/\sqrt{\mathbb{E}_t[v_t(X_t)]}$ as the *information ratio*, and show that for Thompson Sampling over a set of $K$ points (under an i.i.d. prior $\mathcal{F}$) this ratio is always bounded by $\sqrt{K}$, with no assumptions on the structure of the functions $F_{1:T}$. In the sequel, we show that this $\sqrt{K}$ factor can be improved to a polylogarithmic term in $K$ (albeit using a different algorithm) when $F_{1:T}$ are univariate convex functions.

## 3. Leveraging Convexity: The Local-to-Global Lemma

To obtain the desired regret bound, our analysis must somehow take advantage of some special property of convex functions. In this section, we specify which property of convex functions is leveraged in our proof.

To gain some intuition, consider the following prior distribution, which is not restricted to convex functions: draw a point $X^\star$ uniformly in $[0,1]$ and set all of the loss functions to be the same function, $F_t(x) = \mathbb{1}_{x \neq X^\star}$ (the indicator of $x \neq X^\star$). Regardless of the player's policy, she will almost surely miss the point $X^\star$, observe the loss sequence $1, \ldots, 1$, and incur a regret of $T$. The reason for this high regret is that the prior was able to hide the good point $X^\star$ in each of the loss functions without modifying them globally. However, if the loss functions are required to be convex, it is impossible to design a similar example. Specifically, any local modification to a convex

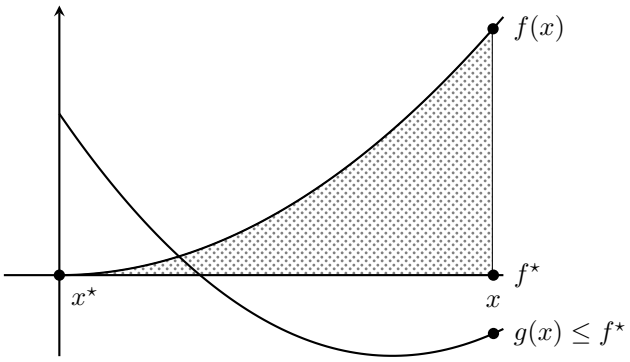


Figure 1: An illustration of the local-to-global lemma. The $L_2$ distance between the reference convex function $f$ to a convex function $g$ in the interval $[x^\star, x]$, where $x^\star$ is the minimizer of $f$ and $x$ is a point such that $g(x) \le f(x^\star)$, can be lower bounded in terms of the shaded area that depicts the energy of the function $f$ in the same interval.

function necessarily changes the function globally (namely, at many different points). This intuitive argument is formalized in the following lemma; here we denote by $\|g\|_\nu^2 = \int g^2 d\nu$ the $L_2$-norm of a function $g : [0,1] \mapsto \mathbb{R}$ with respect to a probability measure $\nu$.

**Lemma 6** (Local-to-global lemma)**.** *Let $f, g : [0,1] \mapsto \mathbb{R}$ be strictly convex functions. Denote $x^\star = \arg\min_{x\in[0,1]} f(x)$ and $f^\star = f(x^\star)$, and let $x \in [0,1]$ such that $g(x) \le f^\star < f(x)$. Then for any probability measure $\nu$ supported on $[x^\star, x]$, we have*

$$\frac{\|f-g\|_\nu^2}{(f(x)-g(x))^2} \;\ge\; \nu(x^\star) \cdot \frac{\|f-f^\star\|_\nu^2}{(f(x)-f^\star)^2}\,.$$

To understand the statement of the lemma, it is convenient to think of $f$ as a reference convex function, to which we compare another convex function $g$; see Fig. 1. If $g$ substantially differs from $f$ at one point $x$ (in the sense that $g(x) \le f^\star$), then the lemma asserts that $g$ must also differ from $f$ globally (in the sense that $\|f-g\|_\nu^2$ is large).

*Proof.* Let $X$ be a random variable distributed according to $\nu$. To prove the lemma, we must show that

$$\frac{\mathbb{E}(f(X)-g(X))^2}{(f(x)-g(x))^2} \;\ge\; \mathbb{P}(X = x^\star) \cdot \frac{\mathbb{E}(f(X)-f^\star)^2}{(f(x)-f^\star)^2}\,.$$

Without loss of generality, we can assume that $x > x^\star$. Let $x_0$ be the unique point such that $f(x_0) = g(x_0)$, and if such a point does not exist let $x_0 = x^\star$. Note that $x_0 < x$, and observe that $g$ is below (resp. above) $f$ on $[x_0, x]$ (resp. $[x^\star, x_0]$).

**Step 1:** We first prove that, without loss of generality, one can assume that $g$ is linear with $g(x_0) = f(x_0)$. Indeed, consider $\tilde{g}$ to be the linear extension of the chord of $g$ between $x$ and $x_0$. Then, we claim that:

$$\frac{\mathbb{E}(f(X)-g(X))^2}{(f(x)-g(x))^2} \;\ge\; \frac{\mathbb{E}(f(X)-\tilde{g}(X))^2}{(f(x)-\tilde{g}(x))^2}. \tag{4}$$

Indeed, the denominator is the same on both side of the inequality, and by convexity $\tilde{g}$ is always closer to $f$ than $g$. Thus, in what follows we assume that $g$ is linear and $g(x_0) = f(x_0)$.

**Step 2:** We show now that one can assume $g(x) = f^\star$. Let $\tilde{g}$ be the linear function such that $\tilde{g}(x) = f^\star$ and $\tilde{g}(x_0) = f(x_0)$. Similarly to the previous step, we have to show that Eq. (4) holds true. We will show that $h(y) = \big(f(y) - g(y)\big)/\big(f(y) - \tilde{g}(y)\big)$ is non-increasing on $[x^\star, x]$, which would imply Eq. (4). A simple approximation argument shows that without of generality one may assume that $f$ is differentiable, in which case $h$ is also differentiable. Observe that $h'(y)$ has the same sign as $u(y) = f'(y)(g(y) - \tilde{g}(y)) - g'(y)(f(y) - \tilde{g}(y)) + \tilde{g}'(y)(f(y) - g(y))$ . Moreover, $u'(y) = f''(y)(g(y) - \tilde{g}(y))$ since $g'' = \tilde{g}'' = 0$, and thus $u$ is decreasing on $[x_0, x]$ and increasing on $[x^\star, x_0]$ (recall that by convexity $f''(y) \geq 0$). Since $u(x_0) \leq 0$ (in fact $u(x_0) = 0$ in the case $x_0 \neq x^\star$), this implies that $u$ is non-positive, and thus $h$ is non-increasing, which concludes this step.

**Step 3:** It remains to show that when $g$ is linear with $g(x) = f^\star$, then

$$\mathbb{E}(f(X) - g(X))^2 \;\geq\; \mathbb{P}(X = x^\star) \cdot \mathbb{E}(f(X) - f^\star)^2 \,. \tag{5}$$

For notational convenience, we assume $f^\star = 0$. By monotonicity of $f$ and $g$ on $[x^\star, x]$, one has $|f(y) - g(y)| \;\geq\; |f(y) - f(x_0)|$ for all $y \in [x^\star, x]$. Therefore, it holds that

$$\mathbb{E}(f(X) - g(X))^2 \;\geq\; \mathbb{E}(f(X) - f(x_0))^2 \;\geq\; \mathrm{Var}(f(X)) \;=\; \mathbb{E}f^2(X) - (\mathbb{E}f(X))^2. \tag{6}$$

Now using Cauchy-Schwarz one has $\mathbb{E}f(X) = \mathbb{E}f(X)\mathbb{1}\{X \neq x^\star\} \leq \sqrt{\mathbb{P}(X \neq x^\star) \cdot \mathbb{E}f^2(X)}$, which together with Eq. (6) yields Eq. (5). □

## 4. Algorithm for Bayesian Convex Bandits

In this section we present and analyze our algorithm for one-dimensional bandit convex optimization in the Bayesian setting, over $\mathcal{K} = [0, 1]$. Recall that in Bayesian setting, there is a prior distribution $\mathcal{F}$ over a sequence $F_{1:T}$ of loss functions over $\mathcal{K}$, such that each function $F_t$ is convex (but not necessarily Lipschitz) and take values in $[0, 1]$ with probability one.

Before presenting the algorithm, we make the following simplification: given $\epsilon > 0$, we discretize the interval $[0, 1]$ to a grid $\mathcal{X}_\epsilon = \{x_1, \ldots, x_K\}$ of $K = 1/\epsilon^2$ equally-spaced points and treat $\mathcal{X}_\epsilon$ as the de facto decision set, restricting all computations as well as the player's decisions to this finite set. We may do so without loss of generality: it can be shown (see Bubeck et al., 2015) that for any sequence of convex loss functions $F_1, \ldots, F_T : \mathcal{K} \mapsto [0, 1]$, the $T$-round regret (of any algorithm) with respect to $\mathcal{X}_\epsilon$ is at most $2\epsilon T$ larger than its regret with respect to $\mathcal{K}$, and we will choose $\epsilon$ to be small enough so that this difference is negligible.

After fixing a grid $\mathcal{X}_\epsilon$, we introduce the following definitions. We define the random variable $X^\star = \arg\min_{x \in \mathcal{X}_\epsilon} \sum_{t=1}^T F_t(x)$, and for all $t$ and $i, j \in [K]$ let

$$\begin{aligned} \alpha_{i,t} \;&=\; \mathbb{P}_t(X^\star = x_i) \,, \\ f_t(x_i) \;&=\; \mathbb{E}_t[F_t(x_i)] \,, \\ f_{j,t}(x_i) \;&=\; \mathbb{E}_t[F_t(x_i) \mid X^\star = x_j] \,. \end{aligned} \tag{7}$$

In words, $X^\star$ is the optimal action in hindsight, and $\alpha_t = (\alpha_{1,t}, \ldots, \alpha_{K,t})$ is the posterior distribution of $X^\star$ on round $t$. The function $f_t : \mathcal{X}_\epsilon \mapsto [0, 1]$ is the expected loss function on round $t$ given

**Inputs**: prior distribution $\mathcal{F}$, tolerance parameter $\epsilon > 0$

Let $K = 1/\epsilon^2$ and $\mathcal{X}_\epsilon = \{x_1, \ldots, x_K\}$ with $x_i = i/K$ for all $i \in [K]$ ;
**For** round $t = 1$ to $T$:
For all $i \in [K]$, compute $\alpha_{i,t}$, $f_t(x_i)$ and $f_{i,t}(x_i)$ defined in Eq. (7) ;
Find $i_t^\star = \arg\min_i f_t(x_i)$ and let $x_t^\star = x_{i_t^\star}$ ;
Define the set

$$S_t \;=\; \left\{ i \in [K] \;:\; f_{i,t}(x_i) \le f_t(x_t^\star) \;\;\text{and}\;\; \alpha_{i,t} \ge \tfrac{\epsilon}{K} \right\} \;; \tag{8}$$

Sample $X_t$ from the distribution $\pi_t = (\pi_{1,t}, \ldots, \pi_{K,t})$ over $\mathcal{X}_\epsilon$, given by

$$\forall\, i \in [K]\,, \qquad \pi_{i,t} \;=\; \tfrac{1}{2}\alpha_{i,t} \cdot \mathbb{1}\{i \in S_t\} + (1 - \tfrac{1}{2}\alpha_t(S_t)) \cdot \mathbb{1}\{i = i_t^\star\} \;, \tag{9}$$

where we denote $\alpha_t(S) = \sum_{i \in S} \alpha_{i,t}$ ;
Play $X_t$ and observe feedback $F_t(X_t)$ ;

Figure 2: A modified Thompson Sampling strategy that guarantees $\widetilde{O}(\sqrt{T})$ expected Bayesian regret for any prior distribution $\mathcal{F}$ over convex functions $F_1, \ldots, F_T : [0,1] \mapsto [0,1]$.

the feedbacks observed in previous rounds, and for each $j \in [K]$, the function $f_{j,t} : \mathcal{X}_\epsilon \mapsto [0,1]$ is the expected loss function on round $t$ conditioned on $X^\star = x_j$ and on the history.

Using the above definitions, we can present our algorithm, shown in Fig. 2. On each round $t$ the algorithm computes, using the knowledge of the prior $\mathcal{F}$ and the feedback observed in previous rounds, the posterior $\alpha_t$ and the values $f_t(x_i)$ and $f_{i,t}(x_i)$ for all $i \in [K]$. Also, it computes the minimizer $x_t^\star$ of the expected loss $f_t$ over the set $\mathcal{X}_\epsilon$, which is the point that has the smallest expected loss on the current round. Instead of directly sampling the decision from the posterior $\alpha_t$ (as Thompson Sampling would do), we make the following two simple modifications. First, we add a forced exploitation on the optimizer $x_t^\star$ of the expected loss to ensure that the player chooses this point with probability at least $\frac{1}{2}$. Second, we transfer the probability mass assigned by the posterior to points not represented in the set $S_t$, towards $x_t^\star$. The idea is that playing a point $x_i$ with $i \notin S_t$ is useless for the player, either because it has a very low probability mass, or because playing $x_i$ would not be (much) more profitable to the player than simply playing $x_t^\star$ on round $t$, *even if she is told that $x_i$ is the optimal point at the end of the game*.

The main result of this section is the following regret bound attained by our algorithm.

**Theorem 7.** *Let $F_1, \ldots, F_T : [0,1] \mapsto [0,1]$ be a sequence of convex loss functions drawn from an arbitrary prior distribution $\mathcal{F}$. For any $\epsilon > 0$, the Bayesian regret of the algorithm described in Fig. 2 over $\mathcal{X}_\epsilon$ is upper-bounded by*

$$10\sqrt{T}\log\frac{2K}{\epsilon} + 10\epsilon T\sqrt{\log\frac{2K}{\epsilon}}\;.$$

*In particular, for $\epsilon = 1/\sqrt{T}$ we obtain an upper bound of $O(\sqrt{T}\log T)$ over the regret.*

*Proof.* We bound the Bayesian regret of the algorithm (with respect to $\mathcal{X}_\epsilon$) on a per-round basis, via the technique described in Section 2.2. Namely, we fix a round $t$ and bound $\mathbb{E}_t[r_t(X_t)]$ in terms of

$\mathbb{E}_t[v_t(X_t)]$ (see Eq. (3)). Since the round is fixed throughout, we omit the round subscripts from our notation, and it is understood that all variables are fixed to their state on round $t$.

First, we bound the expected regret incurred by the algorithm on round $t$ in terms of the posterior $\alpha$ and the expected loss functions $f, f_1, \ldots, f_K$.

**Lemma 8.** *With probability one, it holds that*

$$\mathbb{E}_t[r_t(X_t)] \;\leq\; \sum_{i \in S} \alpha_i (f(x_i) - f_i(x_i)) + \epsilon \,. \tag{10}$$

The proofs of all of our intermediate lemmas are deferred to Bubeck et al. (2015). Next, we turn to lower bound the information gain of the algorithm (as defined in Eq. (3)). Recall our notation $\|g\|_\nu^2$ that stands for the $L_2$-norm of a function $g : \mathcal{K} \mapsto \mathbb{R}$ with respect to a probability measure $\nu$ over $\mathcal{K}$; specifically, for a measure $\nu$ supported on the finite set $\mathcal{X}_\epsilon$ we have $\|g\|_\nu^2 = \sum_{i=1}^K \nu_i g^2(x_i)$.

**Lemma 9.** *With probability one, we have*

$$\mathbb{E}_t[v_t(X_t)] \;\geq\; \sum_{i \in S} \alpha_i \|f - f_i\|_\pi^2 \,. \tag{11}$$

We now set to relate between the right-hand sides of Eqs. (10) and (11), in a way that would allow us to use Lemma 5 to bound the expected cumulative regret of the algorithm. In order to accomplish that, we first relate each regret term $f(x_i) - f_i(x_i)$ to the corresponding information term $\|f - f_i\|_\pi^2$. Since $f$ and the $f_i$'s are all convex functions, this is given by the local-to-global lemma (Lemma 6) which lower-bounds the global quantity $\|f - f_i\|_\pi^2$ in terms of the local quantity $f(x_i) - f_i(x_i)$.

To apply the lemma, we establish some necessary definitions. For all $i \in S$, define $\epsilon_i = \epsilon\,|x_i - x^\star|$, and let $S_i = S \cap [x_i, x^\star]$ be the neighborhood of $x_i$ that consists of all points in $S$ lying between (and including) $x_i$ and the optimizer $x^\star$ of $f$. Now, define weights $w_i$ for all $i \in S$ as follows:

$$\forall\, i^\star \neq i \in S \,, \qquad w_i \;=\; \sum_{j \in S_i} \pi_j \left( \frac{f(x_j) - f(x^\star) + \epsilon_j}{f(x_i) - f(x^\star) + \epsilon_i} \right)^2 , \qquad \text{and} \qquad w_{i^\star} \;=\; \pi_{i^\star} \,. \tag{12}$$

With these definitions, Lemma 6 can be used to prove the following.

**Lemma 10.** *For all $i \in S$ it holds that $\|f - f_i\|_\pi^2 \;\geq\; \frac{1}{4} w_i (f(x_i) - f_i(x_i))^2 - \epsilon^2$.*

Now, averaging the inequality of the lemma with respect to $\alpha$ over all $i \in S$ and using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$, we obtain

$$\sqrt{\sum_{i \in S} \alpha_i w_i (f(x_i) - f_i(x_i))^2} \;\leq\; 2 \sqrt{\sum_{i \in S} \alpha_i \|f - f_i\|_\pi^2} + 2\epsilon \,.$$

On the other hand, the Cauchy-Schwarz inequality gives

$$\sum_{i \in S} \alpha_i (f(x_i) - f_i(x_i)) \;\leq\; \sqrt{\sum_{i \in S} \frac{\alpha_i}{w_i}} \cdot \sqrt{\sum_{i \in S} \alpha_i w_i (f(x_i) - f_i(x_i))^2} \,.$$

Combining the two inequalities and recalling Lemmas 8 and 9, we get

$$\mathbb{E}_t[r_t(X_t)] \;\leq\; 2\sqrt{\sum_{i\in S}\frac{\alpha_i}{w_i}}\cdot\left(\sqrt{\mathbb{E}_t[v_t(X_t)]}+\epsilon\right)+\epsilon\,. \tag{13}$$

It remains to upper bound the sum $\sum_{i\in S}\alpha_i/w_i$. This is accomplished in the following lemma.

**Lemma 11.** *We have*

$$\sum_{i\in S}\frac{\alpha_i}{w_i} \;\leq\; 20\log\frac{2K}{\epsilon}\,.$$

Finally, plugging the bound of the lemma into Eq. (13) and using Lemma 5, yields the stated regret bound. □

## 5. Discussion and Open Problems

We proved that the minimax regret of adversarial one-dimensional bandit convex optimization is $\widetilde{O}(\sqrt{T})$ by designing an algorithm for the analogous Bayesian setting and then using minimax duality to upper-bound the regret in the adversarial setting. Our work raises interesting open problems. The main open problem is whether one can generalize our analysis from the one-dimensional case to higher dimensions (say, even $n = 2$). While much of our analysis generalizes to higher dimensions, the key ingredient of our proof, namely the local-to-global lemma (Lemma 6) is inherently one-dimensional. We hope that the components of our analysis, and especially the local-to-global lemma, will inspire the design of efficient algorithms for adversarial bandit convex optimization, even though our end result is a non-constructive bound.

The Bayesian algorithm used in our analysis is a modified version of the classic Thompson Sampling strategy. A second open question is whether or not the same regret guarantee can be obtained by vanilla Thompson Sampling, without any modification. However, if it turns out that unmodified Thompson Sampling is sufficient, the proof is likely to be more complex: our analysis is greatly simplified by the observation that the instantaneous regret of our algorithm is controlled by its instantaneous information gain on each and every round—a claim that does not hold for Thompson Sampling.

Finally, we note that our reasoning together with Proposition 5 of Russo and van Roy (2014) allows to recover effortlessly Theorem 4 of Bubeck et al. (2012), which gives the worst-case minimax regret for online linear optimization with bandit feedback on a discrete set in $\mathbb{R}^n$. It would be interesting to see if this proof strategy also allows to exploit geometric structure of the point set. For instance, could the techniques described here give an alternative proof of Theorem 6 of Bubeck et al. (2012)?

### Acknowledgements

We thank Jian Ding and Ronen Eldan for helpful discussions during the early stages of this work. Parts of this work were done while TK was visiting at Microsoft Research, Redmond; partial support is gratefully acknowledged.

## References

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

J. Abernethy, A. Agarwal, P. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, 2010.

A. Agarwal, D. Foster, D. Hsu, S. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

S. Bubeck, N. Cesa-Bianchi, and S. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012.

S. Bubeck, O. Dekel, T. Koren, and Y. Peres. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. *arXiv preprint arXiv:1502.06398*, 2015.

V. Dani, T. Hayes, and S. Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

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

N. Gravin, Y. Peres, and B. Sivan. Towards optimal algorithms for prediction with expert advice. *Arxiv preprint arXiv:1409.3040*, 2014.

E. Hazan and K. Levy. Bandit convex optimization: Towards tight bounds. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

A. Neyman. The maximal variation of martingales of probabilities and repeated games with incomplete information. *Journal of Theoretical Probability*, 26(2):557–567, 2013.

D. Russo and B. van Roy. An information-theoretic analysis of thompson sampling. *arXiv preprint arXiv:1403.5341*, 2014.

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

W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Bulletin of the American Mathematics Society*, 25:285–294, 1933.