Improved Regret for Zeroth-Order Stochastic Convex Bandits

Tor Lattimore LATTIMORE@DEEPMIND.COM
DeepMind, London

András György AGYORGY@DEEPMIND.COM
DeepMind, London

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We present an efficient algorithm for stochastic bandit convex optimisation with no assumptions on smoothness or strong convexity and for which the regret is bounded by $O(d^{4.5}\sqrt{n}\,\mathrm{polylog}(n))$, where n is the number of interactions and d is the dimension.

Keywords: Bandits, zeroth order convex optimisation, ellipsoid method.

1. Introduction

Let $\mathcal{K} \subset \mathbb{R}^d$ be a convex body (compact and convex with non-empty interior) and $\mathcal{L} : \mathcal{K} \to [0,1]$ be a convex function. A learner and environment interact sequentially over n rounds. In each round t the learner chooses $X_t \in \mathcal{K}$ and observes $\mathcal{L}(X_t) + \xi_t$, where $(\xi_t)_{t=1}^n$ is a sequence of centered conditionally subgaussian random variables: $\mathbb{E}[\xi_t|\mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[\exp(\xi_t^2)|\mathcal{F}_{t-1}] \leq 2$ with $\mathcal{F}_{t-1} = \sigma(X_1, \xi_1, \ldots, X_{t-1}, \xi_{t-1}, X_t)$ the σ-algebra generated by the history. The learner's aim is to minimise the regret, which is $\mathfrak{R}_n = \mathbb{E}\left[\sum_{t=1}^n (\mathcal{L}(X_t) - \mathcal{L}_\star)\right]$, where $\mathcal{L}_\star = \min_{x\in\mathcal{K}} \mathcal{L}(x)$ and the expectation is over the noise and any randomness in the learner's actions. Our contribution is an efficient and conceptually simple algorithm for the above problem, known as stochastic zeroth-order bandit convex optimisation, and a proof of the following theorem bounding its regret.

Theorem 1 *Assume $\mathcal{K}$ contains a unit radius Euclidean ball. Then there exists an algorithm for which*

$$\mathfrak{R}_n \leq \mathrm{const} \cdot d^{4.5}\sqrt{n}\log(n\,\mathrm{diam}(\mathcal{K}))^{3/2} \max\left(1, \frac{\log(n\,\mathrm{diam}(\mathcal{K}))^{1/2}}{d}\right),$$

where const *is a universal constant and* $\mathrm{diam}(\mathcal{K}) = \max_{x,y\in\mathcal{K}} \|x-y\|$ *with* $\|\cdot\|$ *being the standard Euclidean norm.*

Because the dependence on the diameter of $\mathcal{K}$ in Theorem 1 is logarithmic, it is enough to prove the statement for Lipschitz losses, as given in the next theorem. Theorem 1 then follows by running Algorithm 2 for Lipschitz losses on a slightly restricted and scaled version of $\mathcal{K}$, which is a procedure adopted by Bubeck et al. (2017a) and Lattimore (2020), and summarised in Appendix F.

Theorem 2 *Suppose that $\mathcal{L}(x) - \mathcal{L}(y) \leq \|x - y\|$ for all $x, y \in \mathcal{K}$ and that $\mathcal{K}$ contains a unit radius Euclidean ball. Then the regret of Algorithm 2 is bounded as in Theorem 1 with another universal constant* const[1].

1. Throughout we use const to denote universal constants, but their exact value may differ for every appearance.

The algorithm uses the ellipsoid method, so needs to approximately compute minimum volume ellipsoids of $\mathcal{K}$ intersected with a finite number of half-spaces. The computational complexity (and accuracy) of this procedure depends on the representation of $\mathcal{K}$. A number of situations are discussed in Section 8.

Organisation After the related work and preliminaries, we introduce a surrogate loss function and explain its role in defining a separation oracle to be used by the ellipsoid method (Section 4). There follows the construction and analysis of the separation oracle (Section 5), the main algorithm (Section 6) and then the proof of Theorem 2 (Section 7). The computational complexity under different assumptions is explained in Section 8, which is followed by a discussion of practical considerations (Appendix A).

2. Related work

Bandit convex optimisation is studied under a wide range of assumptions. The most notable distinction is between the adversarial and stochastic settings. In the adversarial setting the loss function changes from round to round and the learner is evaluated relative to the best single point in hindsight. Assumptions on the regularity of the loss function also play a role. The major considerations are: *(a)* smoothness, *(b)* strong convexity, *(c)* Lipschitzness and *(d)* the diameter of the constraint set.

Adversarial setting Algorithms designed for the adversarial setting can be used in the stochastic setting. The only caveat is that standard proofs in the adversarial setting implicitly assume that the noise is homogeneous (independent of the action). For example, they cannot handle classical Bernoulli noise where $\xi_t = 1$ with probability $\mathcal{L}(X_t)$ and $\xi_t = 0$ otherwise. We expect, however, that most analysis can be generalised to handle dependent noise as well.

Research on zeroth-order bandit convex optimisation was initialised by Kleinberg (2005) and Flaxman et al. (2005). Both use importance-weighting to approximate the gradient of the loss, which is then used in gradient descent. In our setting this leads to a regret of $O(n^{5/6}d^{2/3})$ or $O(n^{3/4}\sqrt{d\operatorname{diam}(\mathcal{K}) + d^2})$. While the dependence on n was later shown to be suboptimal, these algorithms are simple to implement and the regret depends only quite weakly on the dimension.

Making assumptions on smoothness and/or strong convexity leads to improved bounds for gradient-based algorithms. For example, Saha and Tewari (2011) refined the arguments of Flaxman et al. (2005) to prove that for smooth loss functions, a gradient-based algorithm can achieve a regret of $O(dn^{2/3}\log(n)^{1/3})$ with a polynomial dependence on the diameter of $\mathcal{K}$. Hazan and Levy (2014) assume strong convexity and smoothness and design a gradient-based algorithm for which the regret is $O(d^{1.5}\sqrt{n\log(n)})$. Ito (2020) showed that when the optimum is not too close to the boundary, then this can be improved to $O(d\sqrt{n\log(n)})$, which matches the $\Omega(d\sqrt{n})$ lower bound of Dani et al. (2008) up to a logarithmic factor. There has been much effort – and a few failed attempts – to show that $O(\sqrt{n}\operatorname{polylog}(n))$ regret is possible for gradient-based methods assuming only that the losses are Lipschitz, rather than smooth and/or strongly convex. It seems, however, that there may be fundamental limitations, at least with current estimation techniques (Hu et al., 2016). Another kind of assumption is that the learner can evaluate the loss function at two (or more) points simultaneously, which reduces the variance of gradient estimation and leads to $O(\sqrt{dn})$ regret (Shamir, 2017). This argument depends on the evaluations having the identical noise (or no noise), and hence the idea cannot be used in the stochastic setting studied here.

The breakthrough by Bubeck et al. (2015) showed that $O(\sqrt{n}\operatorname{polylog}(n))$ regret is possible in the adversarial setting when $d = 1$ without any regularity conditions. Hazan and Li (2016) showed that $\tilde{O}(\sqrt{n})$ is also possible in higher dimensions, but with an exponential dependence on the dimension and computation. Like our work, these authors also use the ellipsoid method, though this is about where the similarity ends. Shortly after, Bubeck et al. (2017a) used the mirror-descent framework and a kernel-based loss estimator to construct a polynomial-time (but not terribly practical) algorithm for the adversarial setting with regret at most $O(d^{10.5}\sqrt{n}\log(n)^{7.5})$. A key ingredient of this work is a kernel-based method for estimating a convex function based on bandit feedback. The surrogate loss used in the present work is closely related, as we explain in detail at the end of Section 4.

The best known bound without strong convexity or smoothness is by Lattimore (2020), who showed that the minimax regret is at most $O(d^{2.5}\sqrt{n}\log(n))$. Their analysis, however, relies on the information-theoretic machinery developed by Russo and Van Roy (2014) and Bubeck and Eldan (2018), and does not yield an algorithm.

Stochastic setting Perhaps surprisingly, stochastic bandit convex optimisation has received far less attention. Agarwal et al. (2013) developed and analysed an algorithm for which the regret is at most $O(d^{16}\sqrt{n}\operatorname{polylog}(n))$, showing for the first time that the optimal dependence on the horizon is at most $n^{1/2}\operatorname{polylog}(n)$. Much like ours, their algorithm is based on the ellipsoid method to incrementally focus on the minimiser of $\mathcal{L}$. Where the methods diverge is the separation oracle. While Agarwal et al. (2013) use the tetrahedron construction of Nemirovsky and Yudin (1983), we use a new gadget that proves to be statistically more efficient.

There is an enormous literature on zeroth-order stochastic convex optimisation, where the learner aims to approximately minimise a convex function using noisy function evaluations. Here the best known bound, at least without any kind of smoothness assumption, is by Belloni et al. (2015). They show that a version of simulated annealing finds an ϵ-optimal minimiser with $O(d^6/\epsilon^2)$ evaluations. This algorithm makes $O(1/\epsilon^2)$ noisy evaluations of the loss function at every point along its trajectory, which generally leads to an $\Omega(n^{2/3})$ cumulative regret. Adapting this algorithm to achieve a $\tilde{O}(\sqrt{n})$ regret in the present setting seems to require new ideas.

3. Preliminaries

Before the analysis, we introduce the necessary notation and remind the reader about elementary results on the ellipsoid method and Orlicz norms.

Notation The Euclidean norm on $\mathbb{R}^d$ is $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$ and for a positive definite matrix $P \in \mathbb{R}^{d \times d}$, $\|x\|_P = \sqrt{x^\top P x}$. The centered ball of radius r is $B_r = \{x \in \mathbb{R}^d : \|x\| \leq r\}$ and its boundary is the $(d-1)$-dimensional sphere $S_r = \{x \in \mathbb{R}^d : \|x\| = r\}$. The uniform (rotational invariant) probability measure on the latter is $\mathcal{H}(S_r)$. The zero vector is $\mathbf{0}$ and the identity matrix is I. The context will always inform the reader about the dimension of these quantities. The Gaussian distribution with mean μ and covariance matrix Σ is denoted by $\mathcal{N}(\mu, \Sigma)$. The density of the standard d-dimensional Gaussian $\mathcal{N}(\mathbf{0}, I)$ is $\rho(z) = (1/(2\pi))^{d/2} \exp(-\|x\|^2/2)$. The (Lebesgue) volume of a measurable set $A \subset \mathbb{R}^d$ is denoted by $\operatorname{vol}(A) = \int_A \mathrm{d}x$. The convex hull of A is $\operatorname{conv}(A)$ and its (Euclidean) diameter is $\operatorname{diam}(A) = \sup_{x,y \in A} \|x - y\|$. The image of A under a function $f : \mathbb{R}^d \to \mathbb{R}$ is denoted by $f(A) = \{f(x) : x \in A\}$. Let Lip_ℓ be the space of functions

$f : \mathbb{R}^d \to \mathbb{R}$ for which $|f(x) - f(y)| \leq \ell \|x - y\|$ for all $x, y \in \mathbb{R}^d$. The space of affine bijections from $\mathbb{R}^d \to \mathbb{R}^d$ is $\mathrm{Aff}(\mathbb{R}^d)$. For a positive integer s, $[s] = \{1, \ldots, s\}$.

Orlicz norms Given a real-valued random variable X, let $\|X\|_{\psi_1} = \inf\{t : \mathbb{E}\left[\exp(|X|/t)\right] \leq 2\}$ and $\|X\|_{\psi_2} = \inf\{t : \mathbb{E}\left[\exp(|X|^2/t^2)\right] \leq 2\}$. Recall that $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}$ (Vershynin, 2018, Lemma 2.7.7). Furthermore, if $(X_t)_{t=1}^n$ are a sequence of independent and identically distributed random variables with mean μ, then there exists a universal constant $C \in (0, 1)$ such that for all $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{t=1}^n X_t - \mu \geq \epsilon\right) \leq \exp\left(-\frac{Cn\epsilon^2}{\|X\|_{\psi_2}^2}\right) \qquad \text{(Vershynin, 2018, §2.6.2, §2.8.1)}$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{t=1}^n X_t - \mu \geq \epsilon\right) \leq \exp\left(-Cn \min\left(\frac{\epsilon^2}{\|X\|_{\psi_1}^2}, \frac{\epsilon}{\|X\|_{\psi_1}}\right)\right) . \tag{1}$$

Our assumption on the noise (ξ_t) is that the conditional law of the noise given the history has $\|\cdot\|_{\psi_2}$-norm of at most 1. For readers unfamiliar with Orlicz norms, a random variable X has $\|X\|_{\psi_2} < \infty$ if and only if it is subgaussian:

$$\left(\|X\|_{\psi_2} < \infty\right) \Leftrightarrow \left(\exists \sigma^2 > 0 \ \forall \alpha \in \mathbb{R}, \ \mathbb{E}[\exp(\alpha(X - \mathbb{E}[X]))] \leq \exp(\alpha^2\sigma^2/2)\right) .$$

Similarly, random variables with $\|X\|_{\psi_1} < \infty$ are sub-exponential, though conventions on the definition of sub-exponential random variables vary. These connections are explained in Chapter 2 of Vershynin (2018). On the positive side, the behaviour of Orlicz norms under composition is easier algebraically than moment generating functions, which justifies our choice. More negatively, Orlicz norms and convolutions are not so well behaved, which sometimes introduces additional constants. The universal constants are not especially large but are left unspecified for simplicity. For readers bold enough to try and implement the algorithm (which uses these confidence intervals), we provide recommendations in Appendix A.

Extending the loss function Many quantities in our analysis are simplified by assuming that $\mathcal{L}$ is defined on all of $\mathbb{R}^d$. For $x \notin \mathcal{K}$ we define $\mathcal{L}(x) = \sup_{y \in \mathcal{K}} \sup_{g \in \partial \mathcal{L}(y)} \left(\mathcal{L}(y) + \langle g, x - y\rangle\right)$, where $\partial \mathcal{L}(y) \subset \mathbb{R}^d$ is the set of subgradients of $\mathcal{L}$ at y. The extended function is the supremum of linear 1-Lipschitz functions and hence is convex and 1-Lipschitz. Note that the extended loss function is only used in the analysis, as we will prove that our algorithm only evaluates $\mathcal{L}$ outside $\mathcal{K}$ with negligible probability.

Affine transformations and noisy functions Our learner interacts with the environment by choosing actions $X \in \mathcal{K}$, possibly with randomisation, and observing $Y = \mathcal{L}(X) + \xi$ where $\mathbb{E}[\exp(\xi^2)|X] \leq 2$. When the learner interacts with the environment this way, we abuse notation by writing $Y \sim \mathcal{L}(X)$. Very often these interactions are made via an affine reparameterisation. Given any $A \in \mathrm{Aff}(\mathbb{R}^d)$ and $f = \mathcal{L} \circ A$, we write $Y \sim f(X)$ to mean that $Y = f(X) + \xi = \mathcal{L}(A(X)) + \xi$ where $\mathbb{E}[\exp(\xi^2)|X] \leq 2$ and $\mathbb{E}[\xi|X] = 0$. We refer to f and $\mathcal{L}$ in this context as noisy convex functions.

Ellipsoid method The ellipsoid method is a classical tool in convex optimisation, which is explained in detail by Grötschel et al. (2012, Chapter 3). Given a convex body $\mathcal{K}$, let $\mathcal{E}(\mathcal{K})$ denote the ellipsoid of smallest volume containing $\mathcal{K}$, which exists and is unique. Our algorithm uses the shallow cut ellipsoid method as a component. The key inequality is below.

Lemma 3 (Grötschel et al. 2012, §3.3.21) *Suppose that $E = \mathcal{E}(\mathcal{K})$ is the minimum volume ellipsoid of a convex body $\mathcal{K} \subset \mathbb{R}^d$ and $A \in \mathrm{Aff}(\mathbb{R}^d)$ is such that $A(B_r) = E$ with $r > 0$. Let $\eta \in \mathbb{R}^d$ be a unit vector. Then,*

$$\mathrm{vol}(\mathcal{E}(\mathcal{K} \cap A(\{x : \langle x, \eta\rangle \le r/(2d)\}))) \le \exp\left(-\frac{1}{20d}\right)\mathrm{vol}(E) \triangleq \gamma\,\mathrm{vol}(E)\,.$$

We will also make use of the fact that if $E = A(B_r)$ is the minimum volume ellipsoid of $\mathcal{K}$ with $A \in \mathrm{Aff}(\mathbb{R}^d)$, then $A(B_{r/d}) \subset \mathcal{K}$ (Artstein-Avidan et al., 2015, Remark 2.1.17).

Constants Let us state upfront a number of constants. The meaning and intuition for these choices will be described later. We assume the following equalities:

$$r = 10^5 d^2\sqrt{2\log(1/\delta)} \qquad \lambda = \frac{d}{2r\sqrt{2\log(1/\delta)}} \qquad \ell = \frac{d}{r}\,\mathrm{diam}(\mathcal{K}) \qquad \gamma = \exp\left(-\frac{1}{20d}\right),$$

where δ is the smallest positive root of

$$\delta\log(1/\delta)^7 = \left(2^{11}\cdot 10^{50} d^{21} n^2\,\mathrm{diam}(\mathcal{K})^2\right)^{-1}\,. \tag{2}$$

Note that δ satisfies $\delta = \lambda^4/(10^{26}\,\mathrm{diam}(\mathcal{K})^2 n^2 d^5 r^6)$ and is upper bounded by the right-hand side of Eq. (2). Furthermore, $\lambda = \Theta(1/(d\log(1/\delta))) < 1$ and $\ell = \Theta(\mathrm{diam}(\mathcal{K})/(d\sqrt{\log(1/\delta)}))$.

4. Outline and surrogate losses

Our algorithm combines the ellipsoid method with a surrogate loss function that can be estimated efficiently. To set the stage, we remind the reader how the ellipsoid method can be used for optimisation.

Ellipsoid method Suppose for a moment that the learner has access to the actual gradients of $\mathcal{L}$. A simple method for minimising $\mathcal{L}$ is to compute a sequence of convex sets $(\mathcal{K}_k)_{k=1}^{\infty}$ inductively, starting with $\mathcal{K}_1 = \mathcal{K}$. The inductive update operates as follows. Given $\mathcal{K}_k$, the learner computes the minimum volume ellipsoid $E_k = \mathcal{E}(\mathcal{K}_k)$ and $A_k \in \mathrm{Aff}(\mathbb{R}^d)$ such that $A_k(B_r) = E_k$. Next, let $f_k = \mathcal{L} \circ A_k$ and $z \in \mathbb{R}^d$ be any point with $\|z\| \le r/(2d)$ and $\nabla f_k(z) \neq \mathbf{0}$ and

$$\begin{aligned}\mathcal{K}_{k+1} &= \mathcal{K}_k \cap \{x \in \mathbb{R}^d : \langle x - A_k(z), \nabla\mathcal{L}(A_k(z))\rangle \le 0\} \\ &= \mathcal{K}_k \cap A_k(\{x \in \mathbb{R}^d : \langle x - z, \nabla f_k(z)\rangle \le 0\})\,.\end{aligned}$$

By convexity, $\mathcal{K}_{k+1}$ contains the minimiser of $\mathcal{L}$, and by Lemma 3,

$$\mathrm{vol}(\mathcal{K}_{k+1}) \le \mathrm{vol}(E_{k+1}) \le \gamma\,\mathrm{vol}(E_k) \le \gamma^{k-1}\,\mathrm{vol}(E_1) \le \gamma^{k-1}\,\mathrm{diam}(\mathcal{K})^d\,\mathrm{vol}(B_1)\,.$$

With a judicious choice of a stopping rule, it is easy to show that this procedure finds an approximate minimiser of $\mathcal{L}$ in just logarithmically many iterations (Bubeck, 2015). Note the unorthodox choice of z: in the standard method, it would be chosen at the center of the current ellipsoid ($z = \mathbf{0}$).

In the terminology of the ellipsoid method, the gradient $\nabla f_k(z)$ is providing a shallow cut separation oracle. Adapting this idea to the noisy setting leads to a serious complication. Namely, the gradient of f_k is hard to estimate without many samples from $\mathcal{L}$. The novelty employed here is to use a surrogate loss function for which the gradient *can* be estimated efficiently while still providing a shallow separation oracle. This replaces the tetrahedron construction used by Agarwal et al. (2013).

A surrogate loss Let $f \in \mathrm{Lip}_\ell$ be convex and X be a random variable with law $\mathcal{N}(\mathbf{0}, I)$. Define $g : \mathbb{R}^d \to \mathbb{R}$ by

$$g(z) = \mathbb{E}\left[f(X) + \frac{f((1-\lambda)X + \lambda z) - f(X)}{\lambda}\right]. \tag{3}$$

Convexity of f implies that g is convex and $g \leq f$ pointwise. We will show that

(a) g and its gradients can be estimated statistically efficiently using noisy evaluations of f; and
(b) ∇g can be used as a separation oracle in place of ∇f in the argument above, provided the point z is chosen carefully.

To see why (b) might be true, let $x_\star = \arg\min_{x \in B_r} f(x)$. Convexity of g and the fact that $g \leq f$ pointwise means that if we can find a $z \in \mathbb{R}^d$ with $\|z\| = r/(2d)$ and $g(z) \geq f(x_\star)$, then

$$0 \geq f(x_\star) - g(z) \geq g(x_\star) - g(z) \geq \langle \nabla g(z), x_\star - z \rangle . \tag{4}$$

Hence, using $\nabla g(z)$ as a separation oracle will not eliminate the minimiser of f. An example is illustrated in Fig. 1. The main challenge is to show that points z for which $g(z) \geq f(x_\star)$ exist and can be identified in a statistically and computationally efficient manner.

That g and its gradients can be estimated efficiently using zeroth-order information is straightforward. Let $Y = f(X) + \xi$ where $\mathbb{E}[\exp(\xi^2)|X] \leq 2$ and $\mathbb{E}[\xi|X] = 0$ and define

$$\mathcal{R}_z(x) = \frac{\rho\left(\frac{x - \lambda z}{1-\lambda}\right)}{(1-\lambda)^d \rho(x)} \qquad \mathfrak{I}_z(x) = \mathbf{1}\left(\mathcal{R}_z(x) \leq \frac{e}{(1-\lambda)^d} \text{ and } \|x\| \leq r/d\right).$$

The former quantity is the Radon-Nikodym derivative between the laws of $(1-\lambda)X + \lambda z$ and X. By a change of measure,

$$g(z) = \mathbb{E}\left[f(X) + \frac{f((1-\lambda)X + \lambda z) - f(X)}{\lambda}\right] = \mathbb{E}\left[Y\left(1 - \frac{1}{\lambda} + \frac{\mathcal{R}_z(X)}{\lambda}\right)\right].$$

The expectation can now be estimated by sampling. A minor annoyance is that $\mathcal{R}_z(X)$ is large with low probability, which compels us to use the truncated version. Nevertheless, an easy calculation shows that $\mathfrak{I}_z(X) = 1$ with overwhelming probability, and so

$$\hat{g}(z) = Y\left(1 - \frac{1}{\lambda} + \frac{\mathcal{R}_z(X)}{\lambda}\right)\mathfrak{I}_z(X) \tag{5}$$

is a nearly unbiased estimate of $g(z)$ that simultaneously has well-behaved moments. Similarly, integrating by parts shows that the gradient of g satisfies $\nabla g(z) = \frac{1}{(1-\lambda)^2}\mathbb{E}\left[Y(X - \lambda z)\mathcal{R}_z(X)\right]$. Note, g is differentiable even if f is not. Like Eq. (5), this means that a nearly unbiased estimator of $\nabla g(z)$ is

$$\widehat{\nabla} g(z) = \frac{Y(X - \lambda z)}{(1-\lambda)^2}\mathcal{R}_z(X)\mathfrak{I}_z(X). \tag{6}$$

The next two lemmas establish the key properties of the surrogate loss. The first lemma shows that the above estimators are nearly unbiased and have small Orlicz norms. Note, the former property would hold immediately if the estimators had been defined without the truncation $\mathfrak{I}_z(X)$. The proof is routine and is provided in Appendix D.

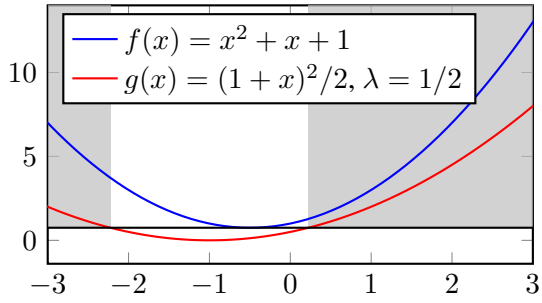

Figure 1: Using the gradient of g as a separation oracle for minimising f is not safe in general. For example, when $x = -3/4$, then $\langle \nabla g(z), x_\star - z \rangle \geq 0$. Convexity of g and the fact that $g \leq f$ ensures that whenever z is such that $g(z) \geq f(x_\star)$ (shaded region), then $\langle \nabla g(z), x_\star - z \rangle \leq 0$.

Lemma 4 *Suppose that $f \in \mathrm{Lip}_\ell$ is convex and $f(B_{r/d}) \subset [0,1]$ and g is defined as in Eq. (3). Then $g(z) \leq f(z)$ for all $z \in \mathbb{R}^d$. Furthermore, when $\|z\| = r/(2d)$, then there exists a universal constant $c = 135$ such that*

(a) $|\mathbb{E}[\hat{g}(z)] - g(z)| \leq 1/(8rn)$.
(b) $\mathbb{E}[\langle \widehat{\nabla g}(z) - \nabla g(z), \eta \rangle] \leq 1/(8rn)$ *for all* $\eta \in B_1$.
(c) $\|\hat{g}(z)\|_{\psi_2} \leq \dfrac{c}{\lambda}$.
(d) $\|\langle \widehat{\nabla g}(z), \eta \rangle\|_{\psi_1} \leq c$ *for all* $\eta \in B_1$.

The next lemma shows there exist z for which $g(z) \geq f(\mathbf{0}) \geq f(x_\star)$, and they can be found with constant probability by sampling. Furthermore, the gap between $g(z)$ and $f(\mathbf{0})$ can be controlled in terms of the regret. The proof is given in Appendix E.

Lemma 5 *Suppose that $Z \sim \mathcal{H}(S_{r/(2d)})$ and $\Delta = \mathbb{E}[f(X)] - \min_{x \in B_r} f(x)$. Then,*

$$\mathbb{P}\left(g(Z) - f(\mathbf{0}) \geq \frac{\lambda r(\Delta - 1/n)}{8 \cdot 10^8 d^{5/2}} \right) \geq 1/48 \,.$$

Relation to the kernel-based method of Bubeck et al. (2017b) The surrogate loss g is closely related to the kernel-based estimator used by Bubeck et al. (2017b). There are some minor differences in how the truncation is done, but setting these aside, their algorithm samples $X \in \mathbb{R}^d$ from some time-dependent distribution on $\mathcal{K}$ and observes $\mathcal{L}(X)$. The algorithm then uses importance-weighting as above to estimate

$$h(z) = \mathbb{E}\left[\mathcal{L}((1-\lambda)X + \lambda z)\right] .$$

Except for shifting and scaling this is the same as the estimator in Eq. (3). The law of X is a smoothed version of an exponential weights distribution constructed from previous loss estimates. Both our analyses exploit the fact that g (or h) can be estimated efficiently with samples. This is where the similarity ends. Bubeck et al. (2017b) focus on properties needed to use an estimator in conjunction with exponential weights, while Lemma 5 is designed for using g as part of a separation oracle. Whether or not our analysis can be used to improve the bounds in the adversarial setting by Bubeck et al. (2017b) is an interesting question. Curiously, Bubeck (2016) conjectured that it may be possible to prove an $\tilde{O}(d^{4.5}\sqrt{n})$ bound on the regret for the kernel-based method in the stochastic setting. It is not at all obvious whether or not this is a coincidence.

5. Separation oracle

Lemmas 4 and 5 provide the tools to define an efficient separation oracle. Lemma 5 and Eq. (4) show that when z is sampled uniformly from $S_{r/(2d)}$, then with constant probability, the pair z and $\nabla g(z)$ provide a separation oracle. By sampling sufficiently many values of z randomly, at least one will be suitable with high probability, and Lemma 4 shows that the relevant values can be estimated efficiently. This idea is implemented in Algorithm 1. In its initialisation phase, the algorithm randomly selects several candidate z-values, $Z_1, \dots, Z_u$. The algorithm then repeatedly queries the loss function until it can certify that a specific point is suitable and returns a half-space that includes the optimal point with high probability. The separation oracle is randomised and the number of times the loss function is queried before the algorithm terminates is a random variable that depends on both the initialisation and the random observations.

input: function f
set $u = \lceil 48 \log(3n^2) \rceil$ and sample $(Z_s)_{s=1}^u \sim \mathcal{H}(S_{r/(2d)})$.
for $m = 1, 2, \dots$**:**
sample $X_m \sim \mathcal{N}(\mathbf{0}, I)$ and $Y_m \sim f(X)$ and $Y'_m \sim f(\mathbf{0})$
for $s = 1, \dots, u$**:**

$$\hat{\theta}_s = \frac{1}{m} \sum_{i=1}^m \left(Y_i \left(1 - \frac{1}{\lambda} + \frac{\mathcal{R}_{Z_s}(X_i)}{\lambda} \right) \mathfrak{I}_{Z_s}(X_i) - Y'_i \right) \qquad \text{estimator of } g(Z_s) - f(\mathbf{0})$$

$$\widehat{\nabla g}_s = \frac{1}{m} \sum_{i=1}^m \frac{Y_i (X_i - \lambda Z_s)}{(1-\lambda)^2} \mathcal{R}_{Z_s}(X_i) \mathfrak{I}_{Z_s}(X_i) \qquad \text{estimator of } \nabla g(Z_s)$$

$$\tilde{\theta}_s = \max \left(0, \hat{\theta}_s - \left(\frac{c}{\lambda} + 1 \right) \sqrt{\frac{\log(1/\epsilon)}{Cm}} - \frac{1}{8rn} \right) \qquad \text{lower confidence bound on } g(Z_s) - f(\mathbf{0})$$

if $m \geq \left(\frac{4cr}{\tilde{\theta}_s} \right)^2 \frac{\log(1/\epsilon)}{C}$ **then**
set $s_\star = s$, $m_\star = m$, and **return** $\left\{ x : \langle \widehat{\nabla g}_s, x - Z \rangle \leq 0 \right\}$

Algorithm 1: Separation oracle

Lemma 6 *Suppose that Algorithm 1 is initialised with a noisy convex function $f \in \mathrm{Lip}_\ell$ and let $\Delta = \mathbb{E}[f(X)] - \min_{x \in B_r} f(x)$, where X has law $\mathcal{N}(\mathbf{0}, I)$. Assume that $x \in B_r$ is a point for which $f(x) \leq f(\mathbf{0}) + \delta$. Let $m_\star$ be the number of iterations before Algorithm 1 returns a half-space H. Provided that $m_\star$ is finite, $H \cap S_{r/(2d)} \neq \emptyset$. Furthermore, with probability at least $1 - 1/n^2$ the following hold:*

(1) Correctness: $x \in H$ or $m_\star > n$.

(2) Running time: if $\Delta \geq 2/n$, then $m_\star \leq \mathrm{const} \left(\frac{d^5}{\Delta^2 \lambda^2} \max \left(1, \frac{1}{\lambda^2 r^2} \right) \log(n) \right)$.

where const *is a universal constant.*

Proof Let $\theta_s = g(Z_s) - f(\mathbf{0})$ and $\hat{\theta}_{s,m}$ and $\widehat{\nabla g}_{s,m}$ and $\tilde{\theta}_{s,m}$ be the values of $\hat{\theta}_s$, $\widehat{\nabla g}_s$ and $\tilde{\theta}_s$ as defined in Algorithm 1 in the mth iteration. Note the definition of $s_\star \in [u]$ and $m_\star$ in Algorithm 1.

To begin, provided the algorithm returns, the half-space returned has a non-empty intersection with $S_{r/(2d)}$ because $\{Z_s\}_{s=1}^u$ are all in $S_{r/(2d)}$ by definition.

Concentration of measure Let $\epsilon = 1/(3un^3)$, where $u = \lceil 48 \log(3n^2) \rceil$ is defined in Algorithm 1. A union bound in combination with Lemma 4 and Eq. (1) shows that with probability at least $1 - 2/(3n^2)$ the following hold for all $m \in [n]$ and $s \in [u]$:

$$|\hat{\theta}_{s,m} - \theta_s| \le \frac{1}{8rn} + \left(\frac{c}{\lambda} + 1\right)\sqrt{\frac{\log(1/\epsilon)}{Cm}}$$
$$|\langle \widehat{\nabla} g_{s,m} - \nabla g(Z_s), x - Z_s \rangle| \le \frac{\|x - Z_s\|}{8rn} + \|x - Z_s\| \max\left(\frac{c\log(1/\epsilon)}{Cm}, c\sqrt{\frac{\log(1/\epsilon)}{Cm}}\right). \quad (7)$$

By Lemma 5 and the assumption in the statement of the present lemma that $\Delta \ge 2/n$, for any $s \in [u]$, with probability at least $1/48$,

$$\theta_s \ge \frac{\lambda r \Delta}{16 \cdot 10^8 d^{5/2}}. \quad (8)$$

Therefore, with probability at least $1 - (47/48)^u \ge 1 - 1/(3n^2)$, there exists an $s \in [u]$ such that Eq. (8) holds. In the remaining two parts we prove the correctness and bound the number of queries under the event where all of the above concentration properties hold, which by a union bound has probability at least $1 - 1/n^2$.

Proof of correctness By the assumed concentration properties, $\tilde{\theta}_{s,m} \le \max(0, \theta_s)$ for all $m \le n$. Furthermore, whenever the algorithm halts with $m_\star \le n$, combining the stopping condition with Eq. (7) gives

$$\begin{aligned}\langle \widehat{\nabla} g_{s_\star, m_\star}, x - Z_{s_\star} \rangle &\le \langle \nabla g(Z_{s_\star}), x - Z_{s_\star} \rangle + \frac{\|x - Z_{s_\star}\|}{8rn} + \frac{\|x - Z_{s_\star}\|\theta_{s_\star}}{4r} \\ &\le \langle \nabla g(Z_{s_\star}), x - Z_{s_\star} \rangle + \frac{1}{4n} + \frac{\theta_{s_\star}}{2},\end{aligned}$$

where we used the fact that $x \in B_r$ and $Z_{s_\star} \in S_{r/(2d)} \subset B_r$, and that $c, r > 1$. Since g is convex and $g(x) \le f(x)$ by the first part of Lemma 4, and by the assumption that $f(x) \le f(\mathbf{0}) + \delta$, it follows that $f(\mathbf{0}) + \delta \ge f(x) \ge g(x) \ge g(Z_{s_\star}) + \langle \nabla g(Z_{s_\star}), x - Z_{s_\star} \rangle$, which shows that $\langle \nabla g(Z_{s_\star}), x - Z_{s_\star} \rangle \le \delta - \theta_{s_\star}$. Therefore, whenever $m_\star \le n$,

$$\langle \widehat{\nabla} g(Z_{s_\star}), x - Z_{s_\star} \rangle \le \langle \nabla g(Z_{s_\star}), x - Z_{s_\star} \rangle + \frac{\theta_{s_\star}}{2} + \frac{1}{4n} \le \delta + \frac{1}{4n} - \frac{\theta_{s_\star}}{2} \le \frac{1}{2n} - \frac{\theta_{s_\star}}{2} \le 0,$$

where the last inequality follows since $n \ge m_\star \ge 1/\tilde{\theta}_{s_\star, m_\star} \ge 1/\theta_{s_\star}$. The result follows from the definition of the half-space returned by Algorithm 1.

Bound on running time Let $s \in [u]$ be such that Eq. (8) holds and

$$M = \left\lceil \frac{\log(1/\epsilon)}{C} \max\left(\left(\frac{8c/\lambda + 4}{\theta_s}\right)^2, \left(\frac{8cr}{\theta_s}\right)^2\right) \right\rceil \le \text{const} \left(\frac{d^5}{\lambda^2 \Delta^2} \max\left(1, \frac{1}{\lambda^2 r^2}\right) \log(n)\right),$$

where $\text{const} \approx 10^{22}$ is a suitably large universal constant. By construction, the algorithm has returned after $m_\star \leq M$ iterations provided that $\tilde{\theta}_{s,M} \geq \frac{1}{2}\theta_s$, which holds because

$$\begin{aligned}\tilde{\theta}_{s,M} &= \hat{\theta}_{s,M} - \left(\frac{c}{\lambda} + 1\right)\sqrt{\frac{\log(1/\epsilon)}{CM}} - \frac{1}{8rn} \geq \theta_s - \frac{2}{8rn} - \left(\frac{2c}{\lambda} + 1\right)\sqrt{\frac{\log(1/\epsilon)}{CM}} \\ &\geq \frac{3\theta_s}{4} - \left(\frac{2c}{\lambda} + 1\right)\sqrt{\frac{\log(1/\epsilon)}{CM}} \geq \frac{\theta_s}{2}\,,\end{aligned}$$

where in the first inequality we used the assumed concentration property and the second that for $M \leq n$, $\theta_s \geq 1/n$, $r > 1$ and that $C < 1$. ■

6. Algorithm

The algorithm combines the ellipsoid method described at the start of Section 4 with the shallow cut separation oracle defined in Section 5.

let $\mathcal{K}_1 = \mathcal{K}$

for $k = 1, 2, \ldots$**:**

let $E_k \leftarrow \mathcal{E}(\mathcal{K}_k)$ and $A_k \in \text{Aff}(\mathbb{R}^d)$ be such that $A_k(B_r) = E_k$

run Algorithm 1 with $f_k = \mathcal{L} \circ A_k$ until oracle returns half-space H_k

update the constraint set: $\mathcal{K}_{k+1} \leftarrow \mathcal{K}_k \cap A_k(H_k)$

Algorithm 2

7. Proof of Theorem 2

The proof is decomposed into three steps: (1) defining and bounding the probability of a good event, (2) bounding the total number of iterations, and (3) combining the previous parts. Before these, let us introduce a little common notation. By translating coordinates, assume without loss of generality that $B_1 \subset \mathcal{K}$. Let $x_\star = \arg\min_{x \in \mathcal{K}} \mathcal{L}(x)$ and, using Lemma 7, let $\mathcal{T} \subset \mathcal{K}$ be a collection of $d+1$ points such that *(a)* $x_\star \in \mathcal{T}$, *(b)* $\max_{y \in \mathcal{T}} \|x - y\| \leq \delta$, and *(c)*,

$$\text{vol}(\text{conv}(\mathcal{T})) \geq \frac{1}{2d!}\left(\frac{3}{4}\right)^{(d-1)/2}\left(\frac{\delta}{\text{diam}(\mathcal{K})}\right)^d .$$

Since $\mathcal{L} \in \text{Lip}_1$, it follows that $\mathcal{L}(x) \leq \mathcal{L}(x_\star) + \delta$ for all $x \in \mathcal{T}$.

Step 1: Good event Recall from the definition of Algorithm 2 that $f_k = \mathcal{L} \circ A_k$. Since $A_k(B_r)$ is the minimum volume ellipsoid of $\mathcal{K}_k$, by the remark after Lemma 3, $A_k(B_{r/d}) \subset \mathcal{K}_k \subset \mathcal{K} \subset \text{diam}(\mathcal{K})B_1$. Therefore, for any $x, y \in \mathbb{R}^d$,

$$f_k(x) - f_k(y) \leq \|A_k(x - y)\| \leq \frac{d\,\text{diam}(\mathcal{K})}{r}\|x - y\| = \ell\|x - y\|\,,$$

which means that $f_k \in \mathrm{Lip}_\ell$. Let $\Delta_k = \int f_k(x)\,\mathrm{d}\rho(x) - f(A_k^{-1}(x_\star))$, which is the instantaneous expected regret of querying $A_k(X)$; note that $2\Delta_k$ is an upper bound on the regret per round in iteration k, since $f(\mathbf{0}) \leq \int f_k(x)\,\mathrm{d}\rho(x)$ by convexity. Let G_0 be the trivial event that always holds and G_k be the event that the following both hold:

(a) $\mathcal{T} \subset \mathcal{K}_{k+1}$ if there is an iteration $k+1$.

(b) The number of rounds in iteration k is at most

$$N_k \leq \mathrm{const}\left(\frac{d^5}{\Delta_k^2\lambda^2}\max\left(1, \frac{1}{\lambda^2 r^2}\right)\log(n)\right), \tag{9}$$

where const is the universal constant of Lemma 6.

Let k be fixed and suppose that G_k holds so that $\mathcal{T} \subset \mathcal{K}_k$, which implies that $A_k^{-1}(\mathcal{T}) \subset B_r$. Therefore, the minimiser of f_k is in B_r and furthermore that for any $x \in A_k^{-1}(\mathcal{T})$ we have $f_k(x) \leq f_k(\mathbf{0}) + \delta$. Hence, by Lemma 6 and a union bound over $x \in \mathcal{T}$, $\mathbb{P}(G_{k+1} \mid G_k) \geq 1 - (d+1)/n^2$, where we used the naive bound that $N_k \leq n$ in case $\Delta_k \leq 1/(2n)$.

As promised, next we show that the algorithm only queries $\mathcal{L}$ outside of $\mathcal{K}$ with low probability. Suppose that $X \sim \mathcal{N}(\mathbf{0}, I)$, then by Lemma 8, $\mathbb{P}(\|X\| \geq 6\sqrt{d\log(1/\delta)}) \leq \delta$. Algorithm 1 is initialised with $f_k = \mathcal{L} \circ A_k$ and hence queries $\mathcal{L}$ either at $A_k(\mathbf{0})$ or $A_k(X)$ where X is a standard Gaussian. Since $A_k(B_{r/d}) \subset \mathcal{K}$ and $6\sqrt{d\log(1/\delta)} \leq r/d$, it follows by a union bound that the algorithm never queries $\mathcal{L}$ outside $\mathcal{K}$ with probability at least $1 - n\delta \geq 1 - 1/n$.

Finally, define the good event $G = \cap_k G_k \cap \{\text{algorithm never queries outside } \mathcal{K}\}$. Since there are at most n iterations, by induction and a union bound it follows that $\mathbb{P}(G) \geq 1 - (d+2)/n$.

Step 2: Bounding the number of iterations Let $k_\star$ be the number of iterations. On the event G, by Lemma 3 and the fact that $H_k \cap B_{r/(2d)} \neq \emptyset$,

$$\frac{1}{2d!}\left(\frac{3}{4}\right)^{(d-1)/2}\left(\frac{\delta}{\mathrm{diam}(\mathcal{K})}\right)^d \leq \mathrm{vol}(\mathcal{T}) \leq \mathrm{vol}(\mathcal{K}_k) \leq \gamma^{k-1}\,\mathrm{vol}(E_1) \leq \gamma^{k-1}\,\mathrm{diam}(\mathcal{K})^d\,\mathrm{vol}(B_1)\,.$$

Hence, by the definition of $\gamma = \exp(-1/(20d))$, the number of iterations is bounded by $k_\star \leq \mathrm{const}\, d^2 \log\left(\mathrm{diam}(\mathcal{K})/\delta\right)$, where const is a suitably large universal constant.

Step 3: Combining Decomposing the regret,

$$\begin{aligned}
\mathfrak{R}_n &\leq n\mathbb{E}\left[\mathbf{1}_{G^c}\right] + 2\mathbb{E}\left[\mathbf{1}_G \sum_{k=1}^{k_\star} N_k\Delta_k\right] \\
&\leq (d+2) + \mathrm{const}\left(\frac{d^{5/2}}{\lambda}\max\left(1, \frac{1}{\lambda r}\right)\mathbb{E}\left[\mathbf{1}_G \sum_{k=1}^{k_\star}\sqrt{N_k\log(n)}\right]\right) \\
&\leq (d+2) + \mathrm{const}\left(\frac{d^{5/2}}{\lambda}\max\left(1, \frac{1}{\lambda r}\right)\sqrt{\mathbb{E}\left[\mathbf{1}_G k_\star \sum_{k=1}^{k_\star} N_k\right]\log(n)}\right) \\
&= (d+2) + \mathrm{const}\left(\frac{d^{5/2}}{\lambda}\max\left(1, \frac{1}{\lambda r}\right)\sqrt{\mathbb{E}\left[\mathbf{1}_G k_\star\right] n\log(n)}\right) \\
&\leq \mathrm{const}\left(d^{9/2}\sqrt{n}\left(\log(n\,\mathrm{diam}(\mathcal{K}))\right)^{3/2}\max\left(1, \frac{\sqrt{\log(n\,\mathrm{diam}(\mathcal{K}))}}{d}\right)\right),
\end{aligned}$$

where in the first inequality we used the fact that the losses are bounded in $[0,1]$ and $\mathbb{P}(G) \geq 1 - (d+2)/n$. The second inequality is true by the definition of the good event and the fact that Eq. (9) implies that

$$N_k \Delta_k \leq \text{const} \left(\frac{d^{5/2}}{\lambda} \max\left(1, \frac{1}{\lambda r}\right) \right) \sqrt{N_k \log(n)} \,.$$

The final inequalities follow from the bound on the number of iterations, the fact that $\sum_{k=1}^{k_\star} N_k = n$ and naive simplification.

8. Computation

The only computationally heavy part of Algorithm 2 is the update of the minimum volume ellipsoid. Recall the algorithm is initialised with $\mathcal{K}_1 = \mathcal{K}$. Then, in each round the following steps are needed: *(a)* Find ellipsoid $E_k = \mathcal{E}(\mathcal{K}_k)$; *(b)* use Algorithm 1 to find a shallow cut half-space H_k,;and *(c)* update $\mathcal{K}_{k+1} = \mathcal{K}_k \cap H_k$. Let $A_k \in \text{Aff}(\mathbb{R}^d)$ be such that $A_k(B_r) = E_k$. The two essential properties of these calculations are: (1) that $A_k(B_{\alpha r}) \subset \mathcal{K}$ where α is large enough so that $X \in B_{\alpha r}$ with high probability when $X \sim \mathcal{N}(\mathbf{0}, I)$; and (2) that $\text{vol}(E_{k+1}) \leq \gamma \, \text{vol}(E_k)$ for some $\gamma < 1$. The complexity of (approximately) computing E_k depends on the representation of $\mathcal{K}_k$. Below we outline three special cases.

Unconstrained version Consider the case where $\mathcal{L}$ is defined on all of $\mathbb{R}^d$ and the learner knows the minimum lies in B_R for some given radius R. Then a simple modification of Algorithm 2 yields an efficient algorithm that only needs elementary matrix operations. Specifically, let the algorithm be initialised with $\mathcal{K}_1 = B_R$. Subsequently,

(a) Set $E_k = \mathcal{K}_k$;
(b) Use Algorithm 1 to find a shallow cut half-space H_k;
(c) Update $\mathcal{K}_{k+1} = \mathcal{E}(\mathcal{E}_k \cap H_k)$.

The difference is that $\mathcal{K}_{k+1}$ is taken to be the minimum volume ellipsoid of $E_k \cap H_k$, rather than the intersection $\mathcal{K}_k \cap H_k$. The point is that $\mathcal{E}(E_k \cap H_k)$ has a closed form solution (Grötschel et al., 2012, §3.1). Nothing changes in the analysis. The volume of $(E_k)_k$ still shrinks exponentially and because $\mathcal{L}$ is defined everywhere, the condition that the shrunken ellipsoid is contained in the domain of $\mathcal{L}$ is satisfied automatically.

Polytopes When $\mathcal{K}$ is a polytope, then $\mathcal{K} \cap H_k$ is also a polytope. Sadly, however, finding the minimum volume ellipsoid for a polytope represented as an intersection of half-spaces is computationally intractable. Like in the unconstrained case, a simple modification of the algorithm is efficient and retains the same guarantees. Given a polytope $\mathcal{P}$ and an ellipsoid E' with $\mathcal{P} \subset E'$, there exists an efficient algorithm for finding another ellipsoid $E = A(B_r) \subset E'$ such that $A(B_{\alpha r}) \subset \mathcal{P} \subset A(B_r)$ where $\alpha = (d(d+1))^{-1/2}$. The algorithm for computing E is itself based on the ellipsoid method and is explained in the proof of Theorem 4.6.3 of Grötschel et al. (2012). The modified algorithm is initialised with $\mathcal{K}_1 = \mathcal{K}$ and an ellipsoid $E_1 = A_1(B_r)$ for which $A_1(B_{\alpha r}) \subset \mathcal{K} \subset A_1(B_r)$. Subsequently:

(a) Use Algorithm 1 to find a shallow cut half-space H_k.
(b) Compute $E'_{k+1} = \mathcal{E}(E_k \cap H_k)$ and $\mathcal{K}_{k+1} = \mathcal{K}_k \cap H_k$.
(c) Compute ellipsoid $A_{k+1}(B_r) = E_{k+1} \subset E'_{k+1}$ such that $A_{k+1}(B_{\alpha r}) \subset \mathcal{K}_{k+1} \subset A_{k+1}(B_r)$.

Separation oracle Like for polytopes, when $\mathcal{K}$ is represented by a (weak) separation oracle, there exists an efficient algorithm for finding an ellipsoid E and associated affine map A such that $A(B_{r\alpha}) \subset \mathcal{K} \subset A(B_r)$, where $\alpha = 1/(d^{1/2}(d+1))$. For details, see Theorem 4.6.1 (Grötschel et al., 2012). Notice that α is a factor of $d^{-1/2}$ worse than what was achievable for a polytope, but still large enough that $X \in B_{\alpha r}$ with high probability given our choice of r.

9. Discussion

Other surrogate losses What is interesting about our choice of surrogate loss is that it is a lower bound on the true loss (it is optimistic), but is also a poor approximation. Meanwhile, the surrogate losses used in many previous works are good approximations, but suffer from high variance that leads to suboptimal rates (Flaxman et al., 2005, and many followups). Interestingly, the standard approach based on smoothing leads to a surrogate loss that is an upper bound on the true loss.

Another observation is g (as defined in Eq. (3)) and its gradients can be estimated from single-point feedback, so some of the theory developed here may potentially have applications in the adversarial model.

Other approaches There are many ideas to explore to improve the bound. One can try alternative loss functions or replace the ellipsoid method with something else like Vaidya's cutting plane method. As mentioned, Agarwal et al. (2013) borrowed the tetrahedron construction of Nemirovsky and Yudin (1983), which was used for noise-free zeroth order optimisation. But this latter method already has a poor dependence on the dimension, which has been improved (for example) by Protasov (1996). There is no immediate reduction from noise free convex optimisation to the bandit case, however. One must check that the algorithm is *(a)* robust to noise and *(b)* the regret can be controlled.

References

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964.

A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013.

S. Artstein-Avidan, A. Giannopoulos, and V. D. Milman. *Asymptotic geometric analysis, Part I*, volume 202. American Mathematical Soc., 2015.

A. Belloni, T. Liang, H. Narayanan, and A. Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *Conference on Learning Theory*, pages 240–265, 2015.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

S. Bubeck. Kernel-based methods for convex bandits, part 3. https://blogs.princeton.edu/imabandit/2016/08/10/kernel-based-methods-for-convex-bandits-part-3/, 2016. Accessed: June 2021.

S. Bubeck and R. Eldan. Exploratory distributions for convex functions. *Mathematical Statistics and Learning*, 1(1):73–100, 2018.

S. Bubeck, O. Dekel, T. Koren, and Y. Peres. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In *Proceedings of the 28th Conference on Learning Theory*, pages 266–278, Paris, France, 2015. JMLR.org.

S. Bubeck, Y-T. Lee, and R. Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85, 2017a.

S. Bubeck, Y.T. Lee, and R. Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 72–85, New York, NY, USA, 2017b. ACM. ISBN 978-1-4503-4528-6.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Conference on Learning Theory*, pages 355–366, 2008.

A Flaxman, A Kalai, and HB McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *SODA'05: Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

M. Grötschel, L. Lovász, and A. Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.

E. Hazan and K. Levy. Bandit convex optimization: Towards tight bounds. In *Advances in Neural Information Processing Systems*, pages 784–792, 2014.

E. Hazan and Y. Li. An optimal algorithm for bandit convex optimization. *arXiv preprint arXiv:1603.04350*, 2016.

X. Hu, Prashanth L.A., A. György, and Cs. Szepesvári. (Bandit) convex optimization with biased noisy gradient oracles. In *AISTATS*, pages 819–828, 2016.

S. Ito. An optimal algorithm for bandit convex optimization with strongly-convex and smooth loss. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2229–2239. PMLR, 26–28 Aug 2020.

Bernd Kind and Peter Kleinschmidt. On the maximal volume of convex bodies with few vertices. *Journal of Combinatorial Theory, Series A*, 21(1):124 – 128, 1976.

R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704. MIT Press, 2005.

T. Lattimore. Improved regret for zeroth-order adversarial bandit convex optimisation. *Mathematical Statistics and Learning*, 2(3/4):311–334, 2020.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4, 2011.

A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

V Yu Protasov. Algorithms for approximate calculation of the minimum of a convex function from its values. *Mathematical Notes*, 59(1):69–74, 1996.

D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591. Curran Associates, Inc., 2014.

A. Saha and A. Tewari. 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, 2011.

O. Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *The Journal of Machine Learning Research*, 18(1):1703–1713, 2017.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Appendix A. Practical considerations

The only polynomial-time algorithms with $O(\sqrt{n}\,\mathrm{polylog}(n))$ regret without smoothness or strong convexity are by Agarwal et al. (2013) and Bubeck et al. (2017a), with the latter designed for the adversarial setting. Below we discuss the computation requirements of these algorithms.

Bubeck et al. (2017a): The algorithm is based on continuous exponential weights and needs to solve two computationally heavy sub-problems in every round: *(a)* repeatedly approximately sampling from an approximately log-concave probability measure and *(b)* approximately optimising a convex function. Implementing this algorithm would be a daunting task, except in one dimension for which there is a simplified version.

Agarwal et al. (2013): Like our algorithm, an elimination argument is used and approximate minimum volume ellipsoids need to be computed periodically. The algorithm is otherwise elementary, though care would be required to carefully handle the many cases in the tetrahedron construction borrowed from Nemirovsky and Yudin (1983). As mentioned, the regret of this algorithm is $O(d^{16}\sqrt{n}\,\mathrm{polylog}(n))$, which hints towards worse performance. We are not aware of an existing implementation.

Algorithm 2 can be implemented in an afternoon and in low dimensional problems is numerically robust and extremely fast. There are, however, a number of modifications that are essential before the algorithm is able to get off the ground.

1. The theoretically recommended value of δ is too conservative. This is largely a consequence of the analysis and is due both to naive rounding and the decision to use a single δ for all small error terms.
2. The theoretically recommended value of r is also much too conservative. We recommend $r = d^2\sqrt{\log(1/\delta)}$ as a heuristically well-motivated choice for which we did not observe any problems in experiments.

3. Replace the finite-time confidence bounds in Algorithm 1 with (asymptotic) statistical hypothesis tests. Possibly variance-aware concentration inequalities would suffice, but in practice the number of samples is large enough that the Gaussian approximation is reasonable.

With these changes we were able to produce the regret plot and accompanying piece of abstract art in Fig. 2.

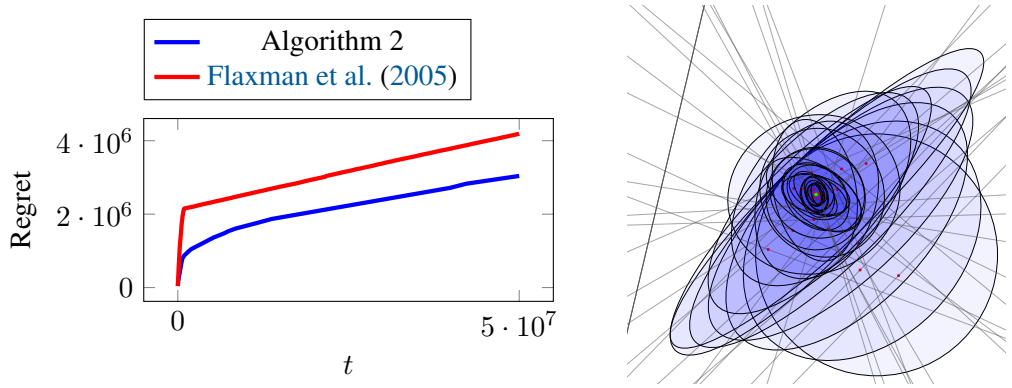

Figure 2: The regret of a modified version of Algorithm 2 and the algorithm proposed by Flaxman et al. (2005) with $\mathcal{L}(x) = \|x\|_1$ and $d = 2$ and $\mathcal{K}$ is a ball of radius 4 and center $(2, 2)$. The parameters of the algorithm by Flaxman et al. (2005) were tuned to the values recommended by theory, so a tuned version of this algorithm may have better performance. The piecewise linear regret of this algorithm is caused by the fixed learning rate. The shaded circles in the right-hand figure are the sequence of ellipsoids generated by Algorithm 2. The lines mark the half-spaces and the green dot is the minimiser of $\mathcal{L}$.

Appendix B. Technical inequalities

Lemma 7 *Let $\mathcal{K}$ be a convex body with $B_1 \subset \mathcal{K}$ and let $x \in \mathcal{K}$. Then there exists a set $\mathcal{T}$ of $d+1$ points such that*

(a) $x \in \mathcal{T}$;
(b) $\max_{y \in \mathcal{T}} \|x - y\| \leq \delta$;
(c) $\mathrm{vol}(\mathrm{conv}(\mathcal{T})) \geq \frac{1}{2d!} \left(\frac{3}{4}\right)^{(d-1)/2} \left(\frac{\delta}{\mathrm{diam}(\mathcal{K})}\right)^d$.

Proof By way of a rotation, assume without loss of generality that $x = (\alpha, 0, \ldots, 0)$ with $\alpha \in [0, \mathrm{diam}(\mathcal{K})]$. Let $U = \{y \in B_1 : y_1 = -1/2\}$, which is a $(d-1)$-dimensional ball with radius $\sqrt{3}/2$ and take $y_1, \ldots, y_d \in U$ to be the corners of a $(d-1)$-dimensional simplex in U of maximum volume.[2] Then, let $\epsilon = \delta / \mathrm{diam}(\mathcal{K})$ and $x_i = (1-\epsilon)x + \epsilon y_i$, which is chosen so that $\|x - x_i\| \leq \delta$.

2. Note that the construction and the following calculations are valid for $d = 1$, in which case $U = \{y_1\}$ is a single point with 0-dimensional volume 1.

Finally, let $\mathcal{T} = \{x, x_1, \ldots, x_d\}$. Then, denoting by vol_d the volume in $\mathbb{R}^d$ to avoid confusion,

$$\begin{aligned}\mathrm{vol}_d(\mathrm{conv}(x, x_1, \ldots, x_d)) &= \epsilon^d \,\mathrm{vol}_d(\mathrm{conv}(x, y_1, \ldots, y_d)) \geq \frac{\epsilon^d}{2d}\,\mathrm{vol}_{d-1}(\mathrm{conv}(y_1, \ldots, y_d)) \\ &= \frac{\epsilon^d}{2d}\left(\frac{3}{4}\right)^{(d-1)/2} \frac{d^{d/2}}{(d-1)^{(d-1)/2}(d-1)!} \geq \frac{\epsilon^d}{2d!}\left(\frac{3}{4}\right)^{(d-1)/2},\end{aligned}$$

where the first inequality follows because the height of the cone with base $(y_1, \ldots, y_d)$ and point x is at least $1/2$ and the second equality follows because the volume of the largest $(d-1)$-simplex inside B_1^{d-1} is $d^{d/2}/((d-1)^{(d-1)/2}(d-1)!)$. The latter can be easily verified by using that the largest $(d-1)$-simplex with diameter 1 is a regular simplex with volume $\sqrt{d/2^{d-1}}/(d-1)!$ (see, e.g., Kind and Kleinschmidt, 1976) and that the $(d-1)$-regular simplex spanned by the standard unit vectors in $\mathbb{R}^d$ has a diameter of $\sqrt{2}$ and the radius of its circumscribed sphere is $\sqrt{1-1/d}$. ■

Appendix C. Concentration

This whole section is a rather monotonous application of standard tail bounds.

Lemma 8 *Let $X \sim \mathcal{N}(\mathbf{0}, I)$ and $\delta \leq 1/e$. Then $\mathbb{P}\left(\|X\| \geq 6\sqrt{d\log(1/\delta)}\right) \leq \delta$.*

Proof This is a naive simplification of a tail bound for the χ-squared distribution (Laurent and Massart, 2000, Lemma 1). ■

Lemma 9 *Suppose that $\eta \sim \mathcal{H}(S_1)$ and $x \in S_1$ is arbitrary. Then,*

$$\mathbb{P}\left(|\langle \eta, x\rangle| \geq \epsilon\sqrt{\frac{\pi}{2d}}\right) \geq 1 - \epsilon \qquad \text{for all } \epsilon \in [0, \sqrt{2d/\pi}]\,.$$

Proof $\mathbb{P}(|\langle \eta, x\rangle| \geq a)$ is the surface area of the hyperspherical cap of height $1-a$ divided by the surface area of a hemisphere, which can be written in terms of the regularised incomplete beta function (Li, 2011):

$$\mathbb{P}\left(|\langle \eta, x\rangle| \geq a\right) = I_{1-a^2}\left(\frac{d-1}{2}, \frac{1}{2}\right).$$

Let $a = \epsilon\sqrt{\pi/(2d)}$. Differentiating shows that $a \mapsto I_{1-a^2}((d-1)/2, 1/2)$ is convex on $[0,1]$ whenever $d \geq 3$ and concave otherwise (famously linear when $d = 3$). When $d = 1$, then $|\langle \eta, x\rangle| = 1$ and when $d = 2$, then by concavity and the fact that $I_0(1/2, 1/2) = 1$ and $I_1(1/2, 1/2) = 0$,

$$I_{1-a^2}\left(\frac{d-1}{2}, \frac{1}{2}\right) \geq 1 - a = 1 - \epsilon\sqrt{\frac{\pi}{4}} \geq 1 - \epsilon\,.$$

Finally, when $d \geq 3$, using convexity and letting $\mathrm{B}(\alpha, \beta)$ be the beta function,

$$I_{1-a^2}\left(\frac{d-1}{2}, \frac{1}{2}\right) \geq 1 - a\left(\frac{\partial}{\partial x} I_{1-x^2}\left(\frac{d-1}{2}, \frac{1}{2}\right)\Big|_{x=0}\right) = 1 - \frac{\epsilon\sqrt{2\pi/d}}{\mathrm{B}\left(\frac{d-1}{2}, \frac{1}{2}\right)} \geq 1 - \epsilon\,. \qquad ■$$

Lemma 10 *Let $\eta \sim \mathcal{H}(S_1)$. Then, $\mathbb{E}[\eta_1^2] = 1/d$ and $\mathbb{E}[\eta_1^4] = 3/(d^2+2d)$.*

Proof The first is trivial by symmetry. For the second, recall that the cumulative density function of $|\eta_1|$ is $\mathbb{P}(|\eta_1| \geq a) = I_{1-a^2}((d-1)/2, 1/2)$ with I the regularised incomplete beta function. Differentiating yields the density which is integrated to obtain the result. ■

Lemma 11 *Let $a \in \mathbb{R}^d$ be a vector with non-negative entries and $\eta \sim \mathcal{H}(S_1)$. Then,*

$$\mathbb{P}\left(\langle a, \eta^2\rangle \geq \frac{\|a\|_1}{2d}\right) \geq \frac{1}{12}\,,$$

where η^2 is interpreted coordinatewise.

Proof By linearity of expectation and Lemma 10, $\mathbb{E}[\langle a, \eta^2\rangle] = \|a\|_1/d$. On the other hand, by the second part of Lemma 10,

$$\mathbb{E}[\langle a, \eta^2\rangle^2] = \sum_{i,j} a_i a_j \mathbb{E}[\eta_i^2 \eta_j^2] \leq \sum_{i,j} a_i a_j \sqrt{\mathbb{E}[\eta_i^4]\mathbb{E}[\eta_j^4]} = \sum_{i,j} \frac{3a_i a_j}{d^2+2d} = \frac{3\|a\|_1^2}{d^2+2d}\,.$$

Hence, by the Payley-Zygmund inequality, for any $\theta \in [0,1]$,

$$\mathbb{P}(\langle a, \eta^2\rangle \geq \theta\|a\|_1/d) \geq (1-\theta)^2 \frac{\mathbb{E}[\langle a, \eta^2\rangle]^2}{\mathbb{E}[\langle a, \eta^2\rangle^2]} \geq (1-\theta)^2 \frac{\|a\|_1^2/d^2}{3\|a\|_1^2/(d^2+2d)} \geq \frac{1}{3}(1-\theta)^2\,.$$

The result is completed by choosing $\theta = 1/2$. ■

Lemma 12 *Suppose $X \sim \mathcal{N}(\mathbf{0}, I)$ and $\|z\| = r/(2d)$. Then,*

$$\mathbb{P}\left(\mathcal{R}_z((1-\lambda)X + \lambda z) \geq \frac{e}{(1-\lambda)^d}\right) \leq \delta \quad \textit{and} \quad \mathbb{P}\left(\mathcal{R}_z(X) \leq \frac{\exp(-2/(1-\lambda)^2)}{(1-\lambda)^d}\right) \leq 2\delta\,.$$

Proof Using the definitions and elementary tail bounds for the standard Gaussian,

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{R}_z((1-\lambda)X + \lambda z) \geq \frac{e}{(1-\lambda)^d}\right) \\
&= \mathbb{P}\left(\exp\left(-\frac{1}{2}\|X\|^2 + \frac{1}{2}\|(1-\lambda)X + \lambda z\|^2\right) \geq e\right) \\
&\leq \mathbb{P}\left(\exp\left(\lambda(1-\lambda)\langle X, z\rangle + \frac{\lambda^2\|z\|^2}{2}\right) \geq e\right) \\
&= \mathbb{P}\left(\lambda(1-\lambda)\langle X, z\rangle + \frac{\lambda^2\|z\|^2}{2} \geq 1\right) \\
&\leq \mathbb{P}\left(\lambda(1-\lambda)\langle X, z\rangle \geq \lambda(1-\lambda)\|z\|\sqrt{2\log(1/\delta)}\right) \\
&\leq \delta\,,
\end{aligned}$$

where in the second to last inequality we used the assumption in the lemma statement that $\|z\| = r/(2d)$ and in the last inequality we used naive simplification in combination with the fact that $\langle X, z\rangle/\|z\|$ is a standard Gaussian and the well known tail bound (Abramowitz and Stegun, 1964, §26),

$$\int_\epsilon^\infty \rho(x)\,\mathrm{d}x \leq \frac{1}{\sqrt{2\pi}}\frac{1}{\epsilon}\exp(-\epsilon^2/2) \qquad \text{for all } \epsilon > 0\,.$$

For the second part,

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{R}_z(X) \leq \frac{\exp(-2/(1-\lambda)^2)}{(1-\lambda)^d}\right) \\
&= \mathbb{P}\left(\exp\left(\frac{1}{2}\|X\|^2 - \frac{1}{2(1-\lambda)^2}\|X - \lambda z\|^2\right) \leq \exp\left(-\frac{2}{(1-\lambda)^2}\right)\right) \\
&= \mathbb{P}\left(\frac{1}{2}\|X\|^2\left(2\lambda - \lambda^2\right) + \frac{\lambda^2}{2}\|z\|^2 - \lambda\langle X, z\rangle \geq 2\right) \\
&\leq \mathbb{P}\left(\lambda\|X\|^2 + \frac{\lambda^2}{2}\|z\|^2 - \lambda\langle X, z\rangle \geq 2\right) \\
&\leq \mathbb{P}\left(\lambda\|X\|^2 \geq 1\right) + \mathbb{P}\left(\frac{\lambda^2}{2}\|z\|^2 - \lambda\langle X, z\rangle \geq 1\right) \\
&\leq \mathbb{P}\left(\lambda\|X\|^2 \geq 1\right) + \mathbb{P}\left(\lambda\langle X, z\rangle \geq \frac{1}{2}\right) \\
&\leq 2\delta\,.
\end{aligned}$$

where in the second last inequality we used the fact that $\lambda\|z\| \leq 1$ and in the final inequality we used Lemma 8, the assumption that $\|z\| = r/(2d)$, the definitions of r and λ, and the above tail bound for the Gaussian. ■

Lemma 13 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a non-negative function and let $\|z\| \leq r/(2d)$ and $X \sim \mathcal{N}(\mathbf{0}, I)$ and $Y = (1-\lambda)X + \lambda z$. Then,*

$$\begin{aligned}
\mathbb{E}[f(Y)] &\leq \exp(2)\mathbb{E}[f(X)] + \sqrt{\delta\mathbb{E}[f(Y)^2]} \\
\mathbb{E}[f(Y)] &\geq \exp(-8)\mathbb{E}[f(X)] - \sqrt{\delta\mathbb{E}[f(X)^2]}\,.
\end{aligned}$$

Proof Let $A = \{x : \mathcal{R}_z(x) \leq e/(1-\lambda)^d\}$. Then, by Lemma 12,

$$\begin{aligned}
\mathbb{E}[f(Y)] &= \mathbb{E}[f(Y)\mathbf{1}_A(Y)] + \mathbb{E}[f(Y)\mathbf{1}_{A^c}(Y)] \\
&\leq \mathbb{E}[f(Y)\mathbf{1}_A(Y)] + \sqrt{\mathbb{E}[f(Y)^2]\mathbb{E}[\mathbf{1}_{A^c}(Y)]} \\
&\leq \mathbb{E}[f(Y)\mathbf{1}_A(Y)] + \sqrt{\delta\mathbb{E}[f(Y)^2]} \\
&= \mathbb{E}[f(X)\mathbf{1}_A(X)\mathcal{R}_z(X)] + \sqrt{\delta\mathbb{E}[f(Y)^2]} \\
&\leq \frac{e}{(1-\lambda)^d}\mathbb{E}[f(X)] + \sqrt{\delta\mathbb{E}[f(Y)^2]} \\
&\leq \exp(2)\mathbb{E}[f(X)] + \sqrt{\delta\mathbb{E}[f(Y)^2]}\,.
\end{aligned}$$

For the other direction, let $B = \{x : \mathcal{R}_z(x) \geq \exp(-2/(1-\lambda)^2)/(1-\lambda)^d\}$. Then, by Lemma 12,

$$\begin{aligned}
\mathbb{E}[f(Y)] &= \mathbb{E}[f(X)\mathcal{R}_z(X)] \\
&= \mathbb{E}[f(X)\mathcal{R}_z(X)\mathbf{1}_B(X)] + \mathbb{E}[f(X)\mathcal{R}_z(X)\mathbf{1}_{B^c}(X)] \\
&\geq \frac{\exp(-2/(1-\lambda)^2)}{(1-\lambda)^d}\mathbb{E}[f(X)\mathbf{1}_B(X)] \\
&\geq \frac{\exp(-2/(1-\lambda)^2)}{(1-\lambda)^d}\mathbb{E}[f(X)] - \frac{\exp(-2/(1-\lambda)^2)}{(1-\lambda)^d}\mathbb{E}[f(X)\mathbf{1}_{B^c}(X)] \\
&\geq \frac{\exp(-2/(1-\lambda)^2)}{(1-\lambda)^d}\mathbb{E}[f(X)] - \frac{\exp(-2/(1-\lambda)^2)}{(1-\lambda)^d}\sqrt{2\delta\mathbb{E}[f(X)^2]} \\
&\geq \exp(-8)\mathbb{E}[f(X)] - \sqrt{\delta\mathbb{E}[f(X)^2]}\,. \qquad \blacksquare
\end{aligned}$$

Lemma 14 *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a non-negative convex function with $h(\mathbf{0}) = 0$ and $X \sim \mathcal{N}(\mathbf{0}, I)$. Then $\mathbb{E}\left[\|X\|^2 h(X)\right] \geq \frac{d}{32}\mathbb{E}[h(X)]$.*

Proof Note that $\|X\|^2$ has a χ-squared distribution with mean d and $\mathbb{E}[\|X\|^4] = d^2 + 2d$. Therefore, by the Payley-Zygmund inequality,

$$\mathbb{P}\left(\|X\|^2 \geq \frac{d}{2}\right) \geq \frac{1}{4}\frac{\mathbb{E}[\|X\|^2]^2}{\mathbb{E}[\|X\|^4]} = \frac{d^2}{4(d^2+2d)} \geq \frac{1}{8}\,.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}[\|X\|^2 h(X)] \geq \mathbb{E}\left[\frac{d}{2}h(X)\mathbf{1}(\|X\|^2 \geq d/2)\right] \\
&\geq \frac{d}{4}\mathbb{E}\left[h(X)\mathbf{1}(\|X\|^2 \geq d/2\right] + \frac{d}{4}\mathbb{E}\left[h(X)\mathbf{1}(\|X\|^2 \geq d/2)\right] \\
&\geq \frac{d}{4}\mathbb{E}\left[h(X)\mathbf{1}(\|X\|^2 \geq d/2)\right] + \frac{d}{4}\frac{\mathbb{P}(\|X\|^2 \geq d/2)}{\mathbb{P}(\|X\|^2 \leq d/2)}\mathbb{E}\left[h(X)\mathbf{1}(\|X\|^2 < d/2)\right] \\
&\geq \frac{d}{4}\mathbb{E}\left[h(X)\mathbf{1}(\|X\|^2 \geq d/2)\right] + \frac{d}{32}\mathbb{E}\left[h(X)\mathbf{1}(\|X\|^2 < d/2)\right] \\
&\geq \frac{d}{32}\mathbb{E}\left[h(X)\right]\,. \qquad \blacksquare
\end{aligned}$$

Appendix D. Proof of Lemma 4

That $g(z) \leq f(z)$ follows immediately from convexity. Moving now to (a) and (b). Let $U = (1-\lambda)X + \lambda z$. Then, by Lemma 8 and the definition of r,

$$\mathbb{P}(\|X\| \geq r/d) \leq \mathbb{P}(\|X\| \geq 6\sqrt{d\log(1/\delta)}) \leq \delta\,.$$

Similarly,

$$\mathbb{P}(\|U\| \geq r/d) \leq \mathbb{P}(\|X\| + \lambda\|z\| \geq r/d) \leq \mathbb{P}(\|X\| \geq 6\sqrt{d\log(1/\delta)}) \leq \delta\,.$$

Combining the above with Lemma 12, Lemma 13 and a union bound,

$$\begin{aligned}\mathbb{P}(\mathfrak{I}_z(U) \neq 1) &\leq 2\delta \leq \zeta \triangleq \exp(9)\sqrt{\delta} \\ \mathbb{P}(\mathfrak{I}_z(X) \neq 1) &\leq \delta + \exp(8)(2\delta + \sqrt{\delta}) \leq \zeta\,.\end{aligned}$$

Let us now prove (a):

$$\begin{aligned}\mathbb{E}[\hat{g}(z)] &= \mathbb{E}\left[f(X)\left(1 - \frac{1}{\lambda} + \frac{\mathcal{R}_z(X)}{\lambda}\right)\mathfrak{I}_z(X)\right] \\ &= \mathbb{E}\left[f(X)\left(1 - \frac{1}{\lambda}\right)\mathfrak{I}_z(X) + \frac{f(U)}{\lambda}\mathfrak{I}_z(U)\right] \\ &= g(z) + \mathbb{E}\left[f(X)\left(1 - \frac{1}{\lambda}\right)(\mathfrak{I}_z(X) - 1) + \frac{f(U)}{\lambda}(1 - \mathfrak{I}_z(U))\right] \\ &\leq g(z) + \frac{1}{\lambda}\sqrt{\mathbb{E}\left[f(X)^2\right]\mathbb{P}(\mathfrak{I}_z(X) \neq 1)} + \frac{1}{\lambda}\sqrt{\mathbb{E}\left[f(U)^2\right]\mathbb{P}(\mathfrak{I}_z(U) \neq 1)} \\ &\leq g(z) + \frac{1}{\lambda}\sqrt{\zeta(1 + \ell^2\mathbb{E}[\|X\|^2)} + \frac{1}{\lambda}\sqrt{\zeta(1 + \ell^2\mathbb{E}[\|(1-\lambda)X + \lambda z\|^2])} \\ &\leq g(z) + \frac{1}{\lambda}\sqrt{\zeta(1 + d\ell^2)} + \frac{1}{\lambda}\sqrt{\zeta(1 + \ell^2\mathbb{E}[\|(1-\lambda)X + \lambda z\|^2])} \\ &\leq g(z) + \frac{2}{\lambda}\sqrt{\zeta(1 + (d+1)\ell^2)} \\ &\leq g(z) + \frac{1}{8rn}\,.\end{aligned}$$

A symmetric argument yields (a). In a similar manner, for any $\eta \in B_1$,

$$\begin{aligned}\mathbb{E}\left[\langle\widehat{\nabla g}(z) - \nabla g(z), \eta\rangle\right] &= \mathbb{E}\left[\langle X, \eta\rangle f(U)(\mathfrak{I}_z(U) - 1)\right] \\ &\leq \sqrt{\mathbb{E}[\|X\|^2 f(U)^2]\mathbb{P}(\mathfrak{I}_z(U) \neq 1)} \\ &\leq \sqrt{\zeta\mathbb{E}[\|X\|^2(1 + \ell^2 + \ell^2\|X\|^2)]} \\ &\leq \sqrt{\zeta(d + (d^2 + 3d)\ell^2)} \\ &\leq \frac{1}{8rn}\,.\end{aligned}$$

where we used the fact that $\mathbb{E}[\|X\|^4] = d^2 + 2d$. We move now to parts (c) and (d). For (c), notice that

$$\begin{aligned}\|\hat{g}(z)\|_{\psi_2} &= \left\|Y\left(1 - \frac{1}{\lambda} + \frac{\mathcal{R}_z(X)}{\lambda}\right)\mathfrak{I}_z(X)\right\|_{\psi_2} \leq \frac{2\exp(1)}{\lambda}\|Y\mathfrak{I}_z(X)\|_{\psi_2} \\ &= \frac{2\exp(1)}{\lambda}\|(f(X) + \xi)\mathfrak{I}_z(X)\|_{\psi_2} \leq \frac{2\exp(1)}{\lambda}\left(1 + \sqrt{1/\log(2)}\right)\,,\end{aligned}$$

For (d),

$$\begin{aligned}
\|\langle\widehat{\nabla g}(z),\eta\rangle\|_{\psi_1} &= \left\|\frac{Y\langle X-\lambda z,\eta\rangle}{(1-\lambda)^2}\mathcal{R}_z(X)\mathfrak{I}_z(X)\right\|_{\psi_1} \\
&\le \|Y\mathfrak{I}_z(X)\|_{\psi_2}\left\|\frac{\langle X-\lambda z,\eta\rangle}{(1-\lambda)^2}\mathcal{R}_z(X)\mathfrak{I}_z(X)\right\|_{\psi_2} \\
&\le \exp(3)\|Y\mathfrak{I}_z(X)\|_{\psi_2}\|\langle X-\lambda z,\eta\rangle\|_{\psi_2} \\
&\le \exp(3)\|Y\mathfrak{I}_z(X)\|_{\psi_2}\left(\|\langle X,\eta\rangle\|_{\psi_2}+\lambda\|\langle z,\eta\rangle\|_{\psi_2}\right) \\
&\le \exp(3)\|Y\mathfrak{I}_z(X)\|_{\psi_2}\left(\|\langle X,\eta\rangle\|_{\psi_2}+\lambda\|\langle z,\eta\rangle\|_{\psi_2}\right) \\
&\le \exp(3)\|Y\mathfrak{I}_z(X)\|_{\psi_2}\left(2\sqrt{\frac{2}{3}}+\lambda\|z\|\right) \\
&\le \exp(3)\|Y\mathfrak{I}_z(X)\|_{\psi_2}\left(2\sqrt{\frac{2}{3}}+\sqrt{2}\right) \\
&\le \exp(3)(1+\sqrt{1/\log(2)})\left(2\sqrt{\frac{2}{3}}+\sqrt{2}\right).
\end{aligned}$$

Appendix E. Proof of Lemma 5

Let $x_\star = \arg\min_{x\in B_r} f(x)$ and $\nabla f(\mathbf{0})$ be any subgradient of f at $\mathbf{0}$. Then, since f is convex,

$$f(x_\star) \ge f(\mathbf{0}) + \langle\nabla f(\mathbf{0}), x_\star\rangle \ge f(\mathbf{0}) - \|x_\star\|\|\nabla f(\mathbf{0})\| \ge f(\mathbf{0}) - r\|\nabla f(\mathbf{0})\|\,.$$

Rearranging shows that $\|\nabla f(\mathbf{0})\| \ge (f(\mathbf{0}) - f(x_\star))/r$. Let $h(z) = f(z) - f(\mathbf{0}) - \langle\nabla f(\mathbf{0}), z\rangle$. Clearly h is non-negative and convex. Furthermore, since $f \in \mathrm{Lip}_\ell$, it follows also that $h \in \mathrm{Lip}_{2\ell}$. Let

$$\Psi(z) = \frac{1}{\lambda}\mathbb{E}\left[h((1-\lambda)X + \lambda z) - h(X)\right].$$

Plugging in the definitions shows that

$$\begin{aligned}
g(z) &= \mathbb{E}[f(X)] + \langle\nabla f(\mathbf{0}), z\rangle + \frac{1}{\lambda}\mathbb{E}\left[h((1-\lambda)X+\lambda z) - h(X)\right] \\
&= \mathbb{E}[f(X)] + \langle\nabla f(\mathbf{0}), z\rangle + \Psi(z)\,.
\end{aligned}$$

By Lemma 15 below, when $Z \sim \mathcal{H}(S_{r/(2d)})$, then with probability at least $1/12$,

$$\begin{aligned}
&\max\left(\Psi(Z) + \langle\nabla f(\mathbf{0}), Z\rangle,\ \Psi(-Z) - \langle\nabla f(\mathbf{0}), Z\rangle\right) \\
&\qquad \ge \frac{\lambda\|Z\|\langle Z, \nabla f(\mathbf{0})\rangle|}{10^7} - \frac{10^4 r^4\,\mathrm{diam}(\mathcal{K})\sqrt{\delta}}{\lambda}\,.
\end{aligned}$$

Furthermore, by Lemma 9, with probability at least $1 - 1/24$,

$$|\langle\nabla f(\mathbf{0}), Z\rangle| \ge \frac{1}{24}\sqrt{\frac{\pi}{2}}\|\nabla f(\mathbf{0})\|\|Z\|d^{-1/2}\,.$$

Therefore, with probability at least $1/24$,

$$\begin{aligned}
\max(g(Z), g(-Z)) - f(\mathbf{0}) &\geq \mathbb{E}[f(X)] - f(\mathbf{0}) + \frac{\lambda \|Z\|^2 \|\nabla f(\mathbf{0})\|}{2 \cdot 10^8 \sqrt{d}} - \frac{10^4 r^4 \operatorname{diam}(\mathcal{K}) \sqrt{\delta}}{\lambda} \\
&\geq \mathbb{E}[f(X)] - f(\mathbf{0}) + \frac{\lambda \|Z\|^2 (f(\mathbf{0}) - f(x_\star))}{2 \cdot 10^8 r \sqrt{d}} - \frac{10^4 r^4 \operatorname{diam}(\mathcal{K}) \sqrt{\delta}}{\lambda} \\
&\geq \frac{\lambda \|Z\|^2 (\mathbb{E}[f(X)] - f(x_\star))}{2 \cdot 10^8 r \sqrt{d}} - \frac{10^4 r^4 \operatorname{diam}(\mathcal{K}) \sqrt{\delta}}{\lambda} \\
&= \frac{\lambda \Delta r}{8 \cdot 10^8 d^{5/2}} - \frac{10^4 r^4 \operatorname{diam}(\mathcal{K}) \sqrt{\delta}}{\lambda} \\
&\geq \frac{\lambda (\Delta - 1/n) r}{8 \cdot 10^8 d^{5/2}} \,,
\end{aligned}$$

where in the second inequality used the convexity of f and that $x_\star \in B_r$, the third holds since $\mathbb{E}[f(X)] \geq f(\mathbf{0})$ by Jensen's inequality, while the last inequality holds by the definition of δ (in fact, δ is defined to make this inequality true). The result now follows since $\mathcal{H}(S_{r/(2d)})$ is rotationally invariant, the laws of Z and $-Z$ are the same, so that with probability at least $1/48$,

$$g(Z) - f(\mathbf{0}) \geq \frac{\lambda (\Delta - 1/n) r}{8 \cdot 10^8 d^{5/2}} \,. \qquad \blacksquare$$

Lemma 15 *Let $u \in \mathbb{R}^d$ and let Ψ and h be defined as above. Then, with probability at least $1/12$,*

$$\max\left(\Psi(Z) + \langle Z, u\rangle,\, \Psi(-Z) - \langle Z, u\rangle\right) \geq \frac{\lambda \|Z\| |\langle Z, u\rangle|}{10^7} - \frac{10^4 r^4 \operatorname{diam}(\mathcal{K}) \sqrt{\delta}}{\lambda} \,.$$

Proof Recall that $X \sim \mathcal{N}(\mathbf{0}, I)$ and $h \in \operatorname{Lip}_{2\ell}$ is convex, non-negative and with $h(\mathbf{0}) = 0$. The first call of business is to resolve an annoying edge case. Since h is non-negative, by definition $\Phi(z) \geq -\mathbb{E}[h(X)]/\lambda$. Suppose that $\mathbb{E}[h(X)] \leq 10^4 r^4 \operatorname{diam}(\mathcal{K}) \sqrt{\delta}$, then for any z,

$$\max(\Psi(z) + \langle z, u\rangle, \Psi(-z) - \langle z, u\rangle) \geq |\langle z, u\rangle| - \frac{10^4 r^4 \operatorname{diam}(\mathcal{K}) \sqrt{\delta}}{\lambda} \,.$$

Since $\lambda \|z\| \leq 1$, this implies the result. For the remainder, assume that $\mathbb{E}[h(X)] > 10^4 r^4 \operatorname{diam}(\mathcal{K}) \sqrt{\delta}$. Integrating by parts shows that

$$\nabla \Psi(z) = \frac{1}{1-\lambda} \mathbb{E}\left[X h((1-\lambda)X + \lambda z)\right] \text{ and } \nabla^2 \Psi(z) = \frac{\lambda}{(1-\lambda)^2} \mathbb{E}\left[X X^\top h((1-\lambda)X + \lambda z)\right] \,.$$

Let $H = \mathbb{E}[X X^\top h(X)]$ and $z \in S_{r/(2d)}$ and suppose that

$$\|z\|_H^2 \geq \frac{\|z\|^2 \operatorname{tr}(H)}{2d} \,. \tag{10}$$

We will prove that this implies the lemma's conclusion. The proof is concluded by Lemma 11 showing that when $Z \sim \mathcal{H}(S_{r/(2d)})$, then

$$\mathbb{P}\left(\|Z\|_H^2 \geq \frac{\|Z\|^2 \operatorname{tr}(H)}{2d}\right) \geq \frac{1}{12} \,. \tag{11}$$

To see this, notice that H is positive semidefinite and symmetric by definition, hence $H = U^\top A U$ where U is a rotation matrix and $A = \langle a_1, \ldots, a_d \rangle$ is a diagonal matrix with the eigenvalues of H, which are non-negative. Letting $Z' = UZ/\|Z\|$, it is easy to see that Z' is uniformly distributed over S_1 and $\|Z\|_H^2/\|Z\|^2 = Z'^\top A Z = \langle a, Z'^2 \rangle$, where Z'^2 is interpreted coordinatewise and $a = (a_1, \ldots, a_d)$. Therefore, since $\operatorname{tr}(H) = \|a\|_1$, Lemma 11 implies Eq. (11).

For the remainder, we prove the lemma under the assumption that Eq. (10) holds and also, without loss of generality, that $\langle z, u \rangle \geq 0$. The core idea is to argue that $\Psi(\mathbf{0})$ is relatively small and that either $\Psi(\pm z)$ is negligible relative to $|\langle z, u \rangle|$ or that Ψ has a lot of curvature so that one of $\Psi(\pm z)$ dominates $|\langle z, u \rangle|$. We get started with three facts, all of which follow arduously from Lemma 13:

$$\text{for all } s \in [0,1] \qquad \mathbb{E}[\langle z, X \rangle^2 h((1-\lambda)X + \lambda s z)] \leq 500 \|z\|_H^2 \tag{12}$$

$$\mathbb{E}[\langle z, X \rangle^2 h((1-\lambda)X] \geq \frac{\|z\|_H^2}{1500} \tag{13}$$

$$\mathbb{E}[h((1-\lambda)X + \lambda z)] \leq 10 \mathbb{E}\left[h(X)\right] . \tag{14}$$

The proof is delayed until the end of the section. Moving on, by Lemma 14 and Eq. (10),

$$\begin{aligned} \|z\|_H^2 &\geq \frac{\|z\|^2 \operatorname{tr}(H)}{2d} = \frac{\|z\|^2}{2d} \operatorname{tr}(\mathbb{E}[X X^\top h(X)]) \\ &= \frac{\|z\|^2}{2d} \mathbb{E}[\|X\|^2 h(X)] \geq \frac{\|z\|^2}{64} \mathbb{E}[h(X)] \geq \mathbb{E}[h(X)] \,, \end{aligned} \tag{15}$$

where in the last step we used that $\|z\|^2 = r^2/(4d^2) > 64$. Second, by convexity of h and integrating by parts,

$$\begin{aligned} \Psi(\mathbf{0}) &= \frac{1}{\lambda} \mathbb{E}[h((1-\lambda)X) - h(X)] \geq -\mathbb{E}[\langle X, \nabla h(X) \rangle] \\ &= \mathbb{E}[h(X)] - \mathbb{E}[\|X\|^2 h(X)] \geq -\mathbb{E}[\|X\|^2 h(X)] \,. \end{aligned}$$

Therefore, repeating at first the argument in Eq. (15), and using that $\Psi(\mathbf{0}) \leq 0$ because h is non-negative and convex with $h(\mathbf{0}) = 0$,

$$\lambda \|z\|_H^2 \geq \frac{\lambda \|z\|^2}{2d} \mathbb{E}[\|X\|^2 h(X)] \geq \frac{\lambda \|z\|^2}{2d} |\Psi(\mathbf{0})| \geq 6000 |\Psi(\mathbf{0})| \,, \tag{16}$$

where the final inequality follows from the the fact that $z \in S_{r/(2d)}$ and the definition of r and λ. We move now to study the curvature of Ψ. Let $\eta = z/\|z\|$. Using the fact that for differentiable f,

$f(t) = f(0) + \int_0^t \int_0^s f''(r)\,\mathrm{d}r\,\mathrm{d}s$ and also convexity of h shows that

$$\begin{aligned}
\frac{1}{2}\left(\Psi(z) + \Psi(-z)\right) &= \Psi(\mathbf{0}) + \frac{1}{2}\int_0^{\|z\|}\int_0^t \left(\|\eta\|^2_{\nabla^2\Psi(s\eta)} + \|\eta\|^2_{\nabla^2\Psi(-s\eta)}\right)\mathrm{d}s\,\mathrm{d}t \\
&\geq \Psi(\mathbf{0}) + \int_0^{\|z\|}\int_0^t \|\eta\|^2_{\nabla^2\Psi(\mathbf{0})}\,\mathrm{d}s\,\mathrm{d}t \\
&= \Psi(\mathbf{0}) + \frac{1}{2}\|z\|^2_{\nabla^2\Psi(\mathbf{0})} \\
&= \Psi(\mathbf{0}) + \frac{\lambda}{2(1-\lambda)^2}\mathbb{E}[\langle z, X\rangle^2 h((1-\lambda)X)] \\
&\geq \Psi(\mathbf{0}) + \frac{\lambda\|z\|^2_H}{3000} \\
&\geq \frac{\lambda\|z\|^2_H}{6000}\,, \qquad (17)
\end{aligned}$$

where the first inequality follows from the convexity of h, the second from Eq. (13) and the third from Eq. (16). On the other hand,

$$\begin{aligned}
\Psi(z) &= \Psi(\mathbf{0}) + \int_0^1 \langle z, \nabla\Psi(sz)\rangle\,\mathrm{d}s \\
&= \Psi(\mathbf{0}) + \frac{1}{1-\lambda}\int_0^1 \mathbb{E}[\langle z, X\rangle h((1-\lambda)X + \lambda sz)]\,\mathrm{d}s \\
&\geq \Psi(\mathbf{0}) - \frac{1}{1-\lambda}\int_0^1 \sqrt{\mathbb{E}[\langle z, X\rangle^2 h((1-\lambda)X + \lambda sz)]\mathbb{E}[h((1-\lambda)X + \lambda sz)]}\,\mathrm{d}s \\
&\geq \Psi(\mathbf{0}) - \frac{1}{1-\lambda}\sqrt{5000\|z\|^2_H\mathbb{E}[h(X)]} \\
&\geq -\frac{\lambda\|z\|^2_H}{6000} - \frac{1}{1-\lambda}\sqrt{5000\|z\|^2_H\mathbb{E}[h(X)]} \\
&\geq -\lambda\|z\|^2_H\left(\frac{1}{6000} + \frac{1}{1-\lambda}\frac{\sqrt{64\cdot 5000}}{\lambda\|z\|}\right) \\
&\geq -\frac{600\lambda\|z\|^2_H}{\lambda\|z\|} = -\frac{600\|z\|^2_H}{\|z\|}\,. \qquad (18)
\end{aligned}$$

where in the first inequality we used Cauchy-Schwarz and in the second we used Eqs. (12) and (14). The third inequality follows from Eq. (16) and the fourth from Eq. (15). The last step is true since $\lambda\|z\| = 1/(4\sqrt{\log(1/\delta)}) < 1$ and $\lambda < 4\cdot 10^{-5}$. The result is finally completed by considering two cases. Suppose that

$$\frac{600\|z\|^2_H}{\|z\|} \leq \frac{1}{2}\langle z, u\rangle\,. \qquad (19)$$

Then $\Psi(z) + \langle z, u\rangle \geq \langle z, u\rangle/2$ is immediate from Eq. (18). On the other hand, if Eq. (19) does not hold, then

$$\max\left(\Psi(z) + \langle z, u\rangle, \Psi(-z) - \langle z, u\rangle\right) \geq \frac{1}{2}\left(\Psi(z) + \Psi(-z)\right) \geq \frac{\lambda\|z\|^2_H}{6000} \geq \frac{\lambda\|z\|\langle z, u\rangle}{10^7}\,.$$

where in the first inequality we used that the max is larger than the average. The second inequality follows from Eq. (17) and the third from the assumption that Eq. (19) does not hold.

Change of measure The proof is completed by proving Eqs. (12) to (14). All results follow from the change of measure result in Lemma 13. Since $\mathbb{E}[h(X)] \geq 10^4 r^4 \operatorname{diam}(\mathcal{K})\sqrt{\delta}$, by Eq. (15),

$$\|z\|_H^2 \geq \mathbb{E}[h(X)] \geq 10^4 r^4 \operatorname{diam}(\mathcal{K})\sqrt{\delta}\,.$$

In order to apply Lemma 13 we crudely bound moments of h. Let $Y = (1-\lambda)X + \lambda s z$ for some $s \in [0,1]$. Using the fact that $\|X\|^2$ is χ-squared distributed,

$$\begin{aligned}
&\mathbb{E}[\|X\|^6] = (d+4)^3 \\
&\mathbb{E}[\|Y\|^6] \leq (1-\lambda)\|X\|^6 + \lambda\|z\|^6 \leq (d+4)^3 + \frac{\lambda r^6}{(2d)^6}\,.
\end{aligned}$$

Now we prove Eqs. (12) to (14). The argument in all cases starts with algebraic manipulation, followed by an application of Lemma 13 and then collecting constants. We start with Eq. (12):

$$\begin{aligned}
\mathbb{E}\left[\langle z, X\rangle^2 h(Y)\right] &= \frac{1}{(1-\lambda)^2}\mathbb{E}\left[\langle z, Y - \lambda s z\rangle^2 h(Y)\right] \\
&\leq 4\left(\mathbb{E}\left[\langle z, Y\rangle^2 h(Y)\right] + \lambda^2\|z\|^4 s^2 \mathbb{E}[h(Y)]\right) \\
&\leq 4\left(\mathbb{E}\left[\langle z, Y\rangle^2 h(Y)\right] + \|z\|^2 \mathbb{E}[h(Y)]\right) \\
&\leq 4\left(\mathbb{E}\left[\langle z, Y\rangle^2 h(Y)\right] + 64\|z\|_H^2\right) \\
&\leq 4\left(72\mathbb{E}\left[\langle z, X\rangle^2 h(X)\right] + \sqrt{\delta\mathbb{E}[\langle z, Y\rangle^4 h(Y)^2]}\right) \\
&= 4\left(72\|z\|_H^2 + \sqrt{\delta\mathbb{E}[\langle z, Y\rangle^4 h(Y)^2]}\right) \\
&\leq 4\left(72\|z\|_H^2 + 2\ell\|z\|^2\sqrt{\delta\mathbb{E}[\|Y\|^6]}\right) \\
&\leq 4\left(72\|z\|_H^2 + \ell r\left((d+4)^3 + \frac{\lambda r^6}{(2d)^6}\right)^{1/2}\sqrt{\delta}\right) \\
&\leq 4\left(72\|z\|_H^2 + r^4 \operatorname{diam}(\mathcal{K})\sqrt{\delta}\right) \\
&\leq 4 \cdot 73\|z\|_H^2 \\
&\leq 500\|z\|_H^2\,.
\end{aligned}$$

where in the first inequality we used that $1/(1-\lambda)^2 \leq 2$ and $(x+y)^2 \leq 2x^2 + 2y^2$. In the second that $\lambda\|z\| \leq 1$ and $s \in [0,1]$. The third inequality follows from Eq. (15) and the fourth from Lemma 13. The fourth inequality is true because $h \in \operatorname{Lip}_{2\ell}$ and by Cauchy-Schwarz. The last

inequality holds by our assumption on δ. Moving on to Eq. (13),

$$\begin{aligned}
\mathbb{E}\left[\langle z, X\rangle^2 h((1-\lambda)X)\right] &\geq \mathbb{E}\left[\langle z, (1-\lambda)X\rangle^2 h((1-\lambda)X)\right] \\
&\geq \exp(-8)\mathbb{E}\left[\langle z, X\rangle^2 h(X)\right] - \sqrt{\delta \mathbb{E}\left[\langle z, X\rangle^4 h(X)^2\right]} \\
&= \exp(-8)\|z\|_H^2 - \sqrt{\delta \mathbb{E}\left[\langle z, X\rangle^4 h(X)^2\right]} \\
&\geq \exp(-8)\|z\|_H^2 - 2\ell\|z\|^2\sqrt{\delta \mathbb{E}\left[\|X\|^6\right]} \\
&\geq \exp(-8)\|z\|_H^2 - \frac{\ell r^2\sqrt{\delta(d+4)^3}}{d^2} \\
&\geq \frac{\exp(-8)}{2}\|z\|_H^2 \\
&\geq \frac{\|z\|_H^2}{1500}\,,
\end{aligned}$$

where in the first inequality we used that $\lambda \in (0,1)$. The second follows from Lemma 13 and the third from the fact that $h \in \mathrm{Lip}_{2\ell}$ and Cauchy-Schwarz. The fourth is from our moment bound and the fifth from the assumption on δ. Finally, for Eq. (14),

$$\mathbb{E}[h(Y)] \leq 8\mathbb{E}[h(X)] + \sqrt{\delta\mathbb{E}[h(Y)^2]} \leq 8\mathbb{E}[h(X)] + (2\ell)\sqrt{\delta\mathbb{E}[\|Y\|^2]} \leq 10\mathbb{E}[h(X)]\,,$$

where in the first inequality we used Lemma 13, the second we used Cauchy-Schwarz and then the assumption on δ. ■

Appendix F. Non-Lipschitz case: the proof of Theorem 1

The following reduction shows how to use Algorithm 2 for non-Lipschitz zeroth-order stochastic bandit convex optimisation. A proof of all claims that follow are given by Lattimore (2020). Let

$$\mathcal{K}' = \left\{nx \in \mathcal{K} : \min_{y \in \partial\mathcal{K}} \|x - y\| \geq 1/n\right\} \qquad\qquad \mathcal{L}'(x) = \mathcal{L}(x/n)\,.$$

Then, using the fact that $\mathcal{L}(\mathcal{K}) \subset [0,1]$, it follows that $\mathcal{L}' : \mathcal{K}' \to [0,1]$ is 1-Lipschitz and there exists an $x \in \mathcal{K}'$ such that $\mathcal{L}'(x) \leq \min_{x\in\mathcal{K}} \mathcal{L}(x) + 1/n$. Furthermore, $\mathrm{diam}(\mathcal{K}') \leq n\,\mathrm{diam}(\mathcal{K})$. Then, simulating $\mathcal{L}'$ in the obvious manner using $\mathcal{L}$, run Algorithm 2 on $\mathcal{K}'$ and $\mathcal{L}'$. Letting $(X'_t)_{t=1}^n$ be the sequence of decisions of the algorithm and $(X_t)_{t=1}^n$ be given by $X_t = X'_t/n$, we have

$$\mathfrak{R}_n = \mathbb{E}\left[\sum_{t=1}^n \mathcal{L}(X_t) - \mathcal{L}_\star\right] \leq 1 + \mathbb{E}\left[\sum_{t=1}^n \mathcal{L}'(X_t) - \mathcal{L}'_\star\right]\,,$$

where $\mathcal{L}'_\star = \min_{x\in\mathcal{K}'} \mathcal{L}'(x)$. Now Theorem 2 provides a bound on the right-hand side, giving

$$\mathfrak{R}_n \leq 1 + \mathrm{const}\cdot d^{4.5}\sqrt{n}\log(n\,\mathrm{diam}(\mathcal{K}))^{3/2}\max\left(1, \frac{\log(n\,\mathrm{diam}(\mathcal{K}))^{1/2}}{d}\right)\,.$$

This proves Theorem 1