Improved Regret for Zeroth-Order Adversarial Bandit Convex Optimisation

Tor Lattimore

DeepMind, London
lattimore@google.com

Abstract. We prove that the information-theoretic upper bound on the minimax regret for zeroth-order adversarial bandit convex optimisation is at most $O(d^{2.5}\sqrt{n}\log(n))$, where d is the dimension and n is the number of interactions. This improves on $O(d^{9.5}\sqrt{n}\log(n)^{7.5})$ by Bubeck et al. (2017). The proof is based on identifying an improved exploratory distribution for convex functions.

Mathematics Subject Classification (2010). 52A20; 62C05.
Keywords. Bandit convex optimisation, online learning.

1 Introduction

Let $\mathcal{K} \subset \mathbb{R}^d$ be a convex body (convex, compact with non-empty interior). At the start of the game, an adversary secretly chooses a sequence $(f_t)_{t=1}^n$ with $f_t \in \mathcal{G}$ and $\mathcal{G}$ is a subset of all convex functions from $\mathcal{K}$ to $[0,1]$. Then, in each round t, the learner chooses an action $x_t \in \mathcal{K}$, possibly with randomisation, and observes only the loss $f_t(x_t)$. The minimax regret over n rounds is

$$\mathfrak{R}_n^*(\mathcal{G}) = \inf_{\text{policies}} \sup_{(f_t)_{t=1}^n \in \mathcal{G}^n} \max_{x \in \mathcal{K}} \mathbb{E}\left[\sum_{t=1}^n f_t(x_t) - f_t(x)\right], \tag{1}$$

where the inf is over all policies of the learner that determine the actions $(x_t)_{t=1}^n$. The expectation integrates over the randomness of the actions $(x_t)_{t=1}^n$. Our contribution is a proof of the following theorem.

Theorem 1. *Suppose that $\mathcal{K}$ contains a unit-radius Euclidean ball and $\mathcal{G}$ is the set of all convex functions from $\mathcal{K}$ to $[0,1]$. Then there exists a universal constant c such that*

$$\mathfrak{R}_n^*(\mathcal{G}) \leq c d^{2.5}\sqrt{n}\log\left(n\,\mathrm{diam}(\mathcal{K})\right),$$

where $\mathrm{diam}(\mathcal{K}) = \max_{x,y\in\mathcal{K}} |x-y|$ *is the diameter of $\mathcal{K}$ and $|\cdot|$ is the standard Euclidean norm.*

As in previous work, we make use of a simple reduction that allows us to restrict slightly the class of functions available to the adversary [2, 4]. Define a constant $m = 1/((n+1)\,\mathrm{diam}(\mathcal{K})^2)$ and let $\mathscr{F}$ be the space of all convex functions $\mathcal{K}$ to $[0,1]$ that are

(a) n-Lipschitz: $f(x) - f(y) \leq n|x - y|$ for all $x, y \in \mathcal{K}$; and

(b) m-strongly convex: for all $x, y \in \mathcal{K}$ and $\lambda \in [0,1]$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) - \frac{1}{2} m \lambda (1-\lambda) |x-y|^2 .$$

Proposition 16 in Section 5 shows that it suffices to prove Theorem 1 with $\mathscr{F}$ rather than $\mathcal{G}$. Briefly, the reduction works by showing that bounded convex functions from $\mathcal{K}$ to $[0,1]$ cannot have large directional derivatives except close to the boundary $\mathcal{K}$ and must have near-minimisers that are not too close to the boundary. This means the learner can play on subset of $\mathcal{K}$ on which the loss functions are n-Lipschitz without sacrificing much in terms of the regret. Strong convexity is introduced by adding a small quadratic to all loss functions, which has only a negligible impact because m is small.

The next (known) theorem serves as our starting point. It follows from the machinery developed by Bubeck et al. [3], Bubeck and Eldan [2] and Russo and Van Roy [13].

Theorem 2. *Let $\alpha, \beta \in \mathbb{R}$ be non-negative and for $f \in \mathscr{F}$, let $f_\star = \min_{x \in \mathcal{K}} f(x)$. Suppose that for any $\bar{f} \in \mathscr{F}$ and finitely supported distribution μ on $\mathscr{F}$ with the discrete σ-algebra there exists a probability measure ρ on $\mathcal{K}$ such that*

$$\int_{\mathcal{K}} \bar{f}(x)\,\mathrm{d}\rho(x) - \int_{\mathscr{F}} f_\star \,\mathrm{d}\mu(f) \leq \alpha + \sqrt{\beta \int_{\mathscr{F}} \int_{\mathcal{K}} (\bar{f}(x) - f(x))^2 \,\mathrm{d}\rho(x)\,\mathrm{d}\mu(f)} . \quad (2)$$

Then there exists a universal constant c such that

$$\mathfrak{R}_n^*(\mathscr{F}) \leq n\alpha + c\sqrt{\beta n d \log\left(n\,\mathrm{diam}(\mathcal{K})\right)} .$$

A hand-wavy intuition for Theorem 2 is as follows. Standard tools from minimax theory show that the inf and sup can be exchanged in Eq. (1), provided that the adversary is allowed to randomise their choices. In other words, the adversary does not lose power if at the start of the game they must announce to the learner from which distribution the sequence of loss functions will be sampled. Based on this, it suffices to bound the Bayesian regret for any prior. When $\bar{f} = \int_{\mathscr{F}} f \,\mathrm{d}\mu(f)$, the left-hand side of Eq. (2) is the instantaneous Bayesian regret for a learner sampling x from ρ and when the posterior is μ. Meanwhile, the right-hand side is the variance of the observation when sampling x from ρ and f from μ, which can be upper bounded by an information gain. When this is large, the learner gains information. Since there is only so much information, bounding the instantaneous regret in terms of the information gain leads to a bound on the cumulative regret [13]. Some subtlety is hidden, and curious readers should investigate the original source [2, §4]. The distribution ρ in Theorem 2 is called an exploratory distribution. Bubeck and Eldan [2] established the conditions of Theorem 2 with $\alpha = O(1/n)$ and $\beta = O(d^{21} \operatorname{polylog}(n))$. The next theorem improves on this result.

Theorem 3. *For any $\bar{f} \in \mathscr{F}$ and distribution μ over $\mathscr{F}$, there exists a probability measure ρ on $\mathcal{K}$ such that Eq.* (2) *holds with $\alpha = 1/n$ and $\beta = cd^4 \log(nd/m)$, where c is a universal constant.*

Briefly, the measure realising Theorem 3 is a mixture over probability measures on the boundaries of level sets of $\bar{f}$. Combining Theorems 2 and 3 with the reduction in Proposition 16 proves Theorem 1.

Related work Online convex optimisation is usually studied under the assumption that the learner has access to the (sub-)gradient $\nabla f_t(X_t)$, or even the whole function f_t. A number of perspectives on this vast literature can be found in recent books and notes by Cesa-Bianchi and Lugosi [5], Hazan [8] and Orabona [12]. There is far less work when the learner only has access to point evaluations. A natural idea is to use importance-weighting to estimate the gradients. At least with current tools for estimating the gradient, however, the resulting bias/variance tradeoff leads to suboptimal regret [11].

The function class $\mathscr{F}$ is omitted from the following regret bounds, to emphasise that the assumptions vary in minor ways. Information-theoretic means were used by Bubeck et al. [3] to show that the minimax regret is $\mathfrak{R}_n^* \leq c \log(n)\sqrt{n}$ when $d = 1$ and c is a universal constant. The multi-dimensional problem was considered by Hazan and Li [10], who showed that the minimax regret is $O(n^{1/2} \operatorname{polylog}(n))$, but with an exponential dependence on the dimension. Shortly after, Bubeck and Eldan [2] generalised the information-theoretic machinery to prove that $\mathfrak{R}_n^* \leq cd^{11} \log(n)^4 \sqrt{n}$, breaking the exponential dependence on the dimension while retaining square root dependence on the horizon. None of these works provide an efficient algorithm. More recently, Bubeck et al. [4] used kernel-based estimators and tools from online convex optimisation to show that $\mathfrak{R}_n^* \leq cd^{9.5} \log(n)^{7.5} \sqrt{n}$. Furthermore, their algorithm can be implemented in polynomial time (with reasonable assumptions) with the price that the dimension-dependence increases to $d^{10.5}$. They conjectured that the optimal regret is $\mathfrak{R}_n^* \leq cd^{1.5}\sqrt{n} \operatorname{polylog}(n)$. The best known lower bound is $\mathfrak{R}_n^* \geq cd\sqrt{n}$, which holds even when the function class is restricted to linear functions [6]. There is also a line of work that exploits strong convexity to obtain better bounds. In particular, if one is prepared to make an additional assumption on smoothness, then Hazan and Levy [9] proved that $\mathfrak{R}_n^* = O(d^{1.5}\sqrt{(n/m)\log(n)})$. The polynomial dependence on m means that enforcing strong convexity by adding a small quadratic to the losses blows up the bound and leads to suboptimal rates.

The new results reduce the dependence on the dimension in the information-theoretic upper bound, but in a way that does not lead naturally to an algorithm.

A positive aspect of the approach is that the proof is geometric and makes use of standard tools in asymptotic convex geometry.

Preliminaries The standard Euclidean norm is $|\cdot|$. The d-dimensional sphere is $S^d = \{x \in \mathbb{R}^{d+1} : |x| = 1\}$. The Minkowski sum of sets A and B is denoted $A + B$. When x is a vector, $A + x = A + \{x\}$. Let vol_p denote the p-dimensional Hausdorff measure on $\mathbb{R}^d$, normalised to coincide with the Lebesgue measure. Given $x, y \in \mathbb{R}^d$, $[x, y] = \{tx + (1-t)y : t \in [0, 1]\}$ is the chord connecting x and y. The sets $[x, y)$ and $(x, y]$ and (x, y) are defined similarly but with appropriate end-points removed. When $x \neq \mathbf{0}$, the hyperplane with normal x is $x^\perp = \{y \in \mathbb{R}^d : \langle x, y\rangle = 0\}$ and $P_x(y) = \arg\min_{z \in x^\perp} |y - z|$ is the Euclidean projection of y onto x. For convex body $K \subset \mathbb{R}^d$, the set $P_x(K)$ is called the shadow of K in direction x. The infimum of a convex function $f : K \to \mathbb{R}$ is $f_\star = \inf_{x \in K} f(x)$. The boundary of K is ∂K.

Outline The proof of Theorem 1 has four main parts. The first provides a key tool for combining exploratory distributions (Section 2). The next step develops the main idea, which is that sampling from the level sets of $\bar f$ is a good exploratory distribution for large subsets of $\mathscr{F}$. The proof of Theorem 3 is found here (Section 3). In Section 4 we prove some technical lemmas, while the argument that Theorems 2 and 3 imply Theorem 1 is given in Section 5.

Acknowledgements Many thanks to my colleagues András György and Marcus Hutter for their useful suggestions.

2 Combining exploratory distributions

The plan is to establish the conditions of Theorem 2 for suitable values of α and β, which means that for any $\bar f \in \mathscr{F}$ and distribution μ on $\mathscr{F}$ we need to find a probability measure ρ on $\mathcal{K}$ satisfying Eq. (2). To make the problem more manageable, we first prove that exploratory distributions can be combined.

Lemma 4. *Let $\bar f \in \mathscr{F}$ and $\mathscr{F} = \cup_{i=1}^k \mathscr{F}_i$. Assume there exist probability measures $(\rho_i)_{i=1}^k$ on $\mathcal{K}$ such that for all $i \in \{1, \ldots, k\}$,*

$$\int_{\mathcal{K}} \bar f(x)\,\mathrm{d}\rho_i(x) - f_\star \leq \alpha + \sqrt{\beta \int_{\mathcal{K}} (\bar f(x) - f(x))^2\,\mathrm{d}\rho_i(x)} \quad \text{for all } f \in \mathscr{F}_i\,. \tag{3}$$

Then there exists a probability measure ρ on $\mathcal{K}$ such that

$$\int_{\mathcal{K}} \bar f(x)\,\mathrm{d}\rho(x) - \int_{\mathscr{F}} f_\star\,\mathrm{d}\mu(f) \leq \alpha + \sqrt{\beta k \int_{\mathscr{F}} \int_{\mathcal{K}} \left(\bar f(x) - f(x)\right)^2 \mathrm{d}\rho(x)\,\mathrm{d}\mu(f)}\,.$$

Proof. The argument is algebraically identical to that used by Russo and Van Roy [14] to bound the information ratio for Thompson sampling. Assume without loss of generality that $(\mathscr{F})_{i=1}^k$ are disjoint and $\mu(\mathscr{F}_i) > 0$ for all $i \in \{1, \ldots, k\}$. Let μ_i be the probability measure obtained by conditioning μ on $\mathscr{F}_i$: $\mu_i(A) = \mu(A \cap \mathscr{F}_i)/\mu(\mathscr{F}_i)$. Then, letting $q_i = \mu(\mathscr{F}_i)$ and $\rho = \sum_{i=1}^k q_i \rho_i$,

$$\begin{aligned}
\int_{\mathcal{K}} \bar{f}(x)\, \mathrm{d}\rho(x) - \int_{\mathscr{F}} f_\star \, \mathrm{d}\mu(f) &= \sum_{i=1}^k q_i \left(\int_{\mathcal{K}} \bar{f}(x)\, \mathrm{d}\rho_i(x) - \int_{\mathscr{F}} f_\star \, \mathrm{d}\mu_i(f) \right) \\
&\leq \alpha + \sum_{i=1}^k q_i \sqrt{\beta \int_{\mathscr{F}} \int_{\mathcal{K}} \left(\bar{f}(x) - f(x)\right)^2 \mathrm{d}\rho_i(x)\, \mathrm{d}\mu_i(f)} \\
&\leq \alpha + \sqrt{\beta k \sum_{i=1}^k q_i^2 \int_{\mathscr{F}} \int_{\mathcal{K}} \left(\bar{f}(x) - f(x)\right)^2 \mathrm{d}\rho_i(x)\, \mathrm{d}\mu_i(f)} \\
&\leq \alpha + \sqrt{\beta k \sum_{i,j=1}^k q_i q_j \int_{\mathscr{F}} \int_{\mathcal{K}} \left(\bar{f}(x) - f(x)\right)^2 \mathrm{d}\rho_i(x)\, \mathrm{d}\mu_j(f)} \\
&= \alpha + \sqrt{\beta k \int_{\mathscr{F}} \int_{\mathcal{K}} \left(\bar{f}(x) - f(x)\right)^2 \mathrm{d}\rho(x)\, \mathrm{d}\mu(f)}\,,
\end{aligned}$$

where the first inequality follows from the assumption in Eq. (3) and Jensen's inequality and the second follows from Cauchy-Schwarz. □

3 Proof of Theorem 3

The first lemma shows that convex functions with different minimisers must differ along suitably chosen rays. The situation is illustrated in Fig. 1.

Lemma 5. *Let $\mathcal{D} \subset \mathbb{R}^d$ be a convex body and $f, g : \mathcal{D} \to \mathbb{R}$ be convex functions with f minimised at $x_\star \in \mathcal{D}$ and $f_\star \leq g_\star$. Suppose that $\epsilon > 0$ and $x_\star \notin K = \{x \in \mathcal{D} : g(x) \leq g_\star + \epsilon\}$. Assume that $x, y \in \partial K$ and $x = \psi x_\star + (1-\psi) y$ with $\psi \in (0, 1)$. Then,*

$$(f(x) - g(x))^2 + (f(y) - g(y))^2 \geq \frac{1}{2} \psi^2 \left(g_\star + \epsilon - f_\star\right)^2 .$$

Proof. Since $x \in (x_\star, y)$, the definition of the level set means that $g(x) = g_\star + \epsilon$. Furthermore, $g(y) \leq g_\star + \epsilon$, with a strict inequality only possible if y is on the boundary of $\mathcal{D}$. Suppose that $\psi f_\star + (1-\psi) f(y) \geq g_\star + \epsilon$. Then

$$(f(y) - g(y))^2 \geq \left(\frac{\psi}{1-\psi}\right)^2 (g_\star + \epsilon - f_\star)^2 \geq \frac{1}{2} \psi^2 (g_\star + \epsilon - f_\star)^2 .$$

Otherwise, the linear case is optimal and the result follows by minimising over the possible values of $f(y)$.

$$\begin{aligned}(g(y)-f(y))^2+(g(x)-f(x))^2 &= (g(y)-f(y))^2+(g_\star+\epsilon-f(x))^2 \\ &\geq \min_{a\in\mathbb{R}}\left[(g(y)-a)^2+(g_\star+\epsilon-\psi f_\star-(1-\psi)a)^2\right] \\ &= \frac{\left(\psi(g_\star+\epsilon-f_\star)+(1-\psi)(g_\star+\epsilon-g(y))\right)^2}{2-2\psi+\psi^2} \\ &\geq \frac{1}{2}\psi^2(g_\star+\epsilon-f_\star)^2\,,\end{aligned}$$

where the second equality follows by solving the quadratic and the last inequality is naive bounding. □

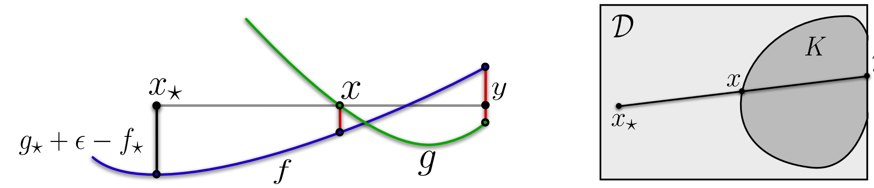

Figure 1: In the figure on the left, the horizontal grey line marks $g_\star + \epsilon = g(x)$. Note that $g(y) \neq g_\star + \epsilon$ is only possible when y is on the boundary of $\mathcal{D}$. The proof of Lemma 5 shows that at least one of the red vertical lines is roughly ψ times the length of the black line. The figure on the right shows the top down view.

Lemma 5 shows that sampling uniformly from entry/exit points of a ray emanating from minimisers of f through a level set of $\bar{f}$ is a good exploratory distribution for f. The depth of the cut made by the ray relative to the length of the ray determines the constant β, with deeper cuts being more informative. This idea cannot be applied directly because the resulting exploratory distribution depends on f. The next lemma connects surface integrals over level sets to rays emanating from a point outside the level set, effectively showing that sampling from a suitable probability measure on a level set of $\bar{f}$ is a good exploratory distribution for many functions $f \in \mathscr{F}$.

Before the statement of the next lemma, we define a quantity that measures a kind of average depth/distance ratio for rays emanating from a point and passing through a convex body. The concepts are illustrated in Fig. 2. Let $K \subset \mathbb{R}^d$ be a convex body with $\mathbf{0} \in K$ and $x \notin K$. Let $\pi_K(x, \cdot) : P_x(K) \to \partial K$ be an inverse of the projection P_x defined by

$$\pi_K(x,z) = \arg\min_{y\in K\cap\{z+tx:t\in\mathbb{R}\}} \langle y, x\rangle\,.$$

For $x \in \mathbb{R}^d$ and $z \in P_x(K)$, define the depth/distance ratio by

$$\Psi_K(x,z) = \frac{\mathrm{vol}_1(K\cap[x,\pi_K(x,z)])}{\mathrm{vol}_1([x,\pi_K(x,z)])}\,.$$

The average depth/distance ratio with respect to the uniform probability measure on $P_x(K)$ is

$$\Psi_K(x) = \frac{1}{\mathrm{vol}_{d-1}(P_x(K))} \int_{P_x(K)} \Psi_K(x,z)\, \mathrm{dvol}_{d-1}(z)\,.$$

The maximum depth/distance ratio is $\Psi_K^\infty(x) = \max_{z \in P_x(K)} \Psi_K(x,z)$. A number of properties of Ψ are collected in Section 4. The most important is that $z \mapsto \Psi_K(x,z)$ is concave on its domain $P_x(K)$.

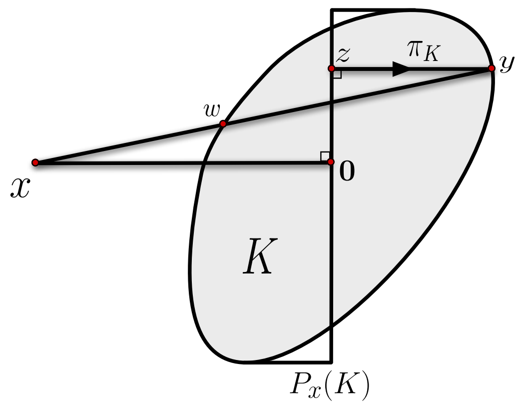

Figure 2: For the configuration in the figure, $\pi_K(x,z) = y$ and $\Psi_K(x,z) = |y-w|/|y-x|$, which is the ratio of the depth of K along the chord $[x,y]$ and the distance of the chord. The quantity $\Psi_K(x)$ averages $\Psi_K(x,z)$ with respect to the uniform probability measure over all z in the shadow $P_x(K)$.

Lemma 6. *Let $\mathcal{D} \subset \mathbb{R}^d$ be a convex body and $f, g : \mathcal{D} \to \mathbb{R}$ be convex, $\epsilon > 0$ and $K = \{x \in \mathcal{D} : g(x) \le g_\star + \epsilon\}$. Let $x_\star \in \mathcal{D}$ be the minimiser of f and assume that $\mathbf{0} \in K$ and $\Psi_K(x_\star) \in [1/(128d), 1/(32d)]$. Then,*

$$\int_{\partial K} g \,\mathrm{d}\rho - f_\star \le 2^{14} d \sqrt{\max_{\theta \in S^{d-1}} \left(\frac{\mathrm{vol}_{d-1}(\partial K)}{\mathrm{vol}_{d-1}(P_\theta(K))}\right) \int_{\partial K} (f-g)^2 \,\mathrm{d}\rho}\,.$$

where $\rho = \mathrm{vol}_{d-1} / \mathrm{vol}_{d-1}(\partial K)$ is the surface area probability measure on ∂K.

Proof of Lemma 6. Abbreviate $\Psi(z) = \Psi_K(x_\star, z)$ and $\pi(z) = \pi_K(x_\star, z)$ and $P = P_{x_\star}$. Define $\lambda : P(K) \to \mathbb{R}$ and $\Lambda : P(K) \to P(K)$ by $\lambda(z) = 1 - \Psi(z)$ and $\Lambda(z) = \lambda(z) z$. The rest of the proof is divided into three steps. The first uses Lemma 15 to show that a large fraction of rays through K from $x_\star$ are reasonably deep. The second step connects surface integrals over ∂K to rays cutting K from $x_\star$ and the third puts together the pieces using Lemma 5.

Step 1: Preponderance of deep cuts Let $B \subset P(K)$ be the set given by

$$B = \left\{ z \in P(K) : \frac{\Psi(z)}{\frac{1}{\mathrm{vol}_{d-1}(P(K))} \int_{P(K)} \Psi \,\mathrm{dvol}_{d-1}} \in [1/4, 16] \right\}\,,$$

which is the set of points in $P(K)$ where $\Psi(z)$ is close to its mean. By Lemma 15 and concavity of Ψ (Lemma 8),

$$\frac{\mathrm{vol}_{d-1}(B)}{\mathrm{vol}_{d-1}(P(K))} \geq \frac{1}{32}\,. \tag{4}$$

Furthermore, by the assumptions in the lemma,

$$\frac{1}{\mathrm{vol}_{d-1}(P(K))} \int_{P(K)} \Psi \,\mathrm{dvol}_{d-1} \in \left[\frac{1}{128d}, \frac{1}{32d}\right],$$

which implies that for $z \in B$,

$$1 - \lambda(z) = \Psi(z) \in \left[\frac{2^{-9}}{d}, \frac{1}{2d}\right]. \tag{5}$$

Step 2: Surface area to rays Let $C = \pi(B) \subset \partial K$. Define a function $\kappa : C \to \partial K$ so that $[x_\star, y] \cap K = [\kappa(y), y]$, which is the point w in Fig. 2 and is chosen so that for any $z \in P(K)$,

$$\kappa(\pi(z)) = \Psi(z)x_\star + (1 - \Psi(z))\pi(z)\,. \tag{6}$$

Let $D = \kappa(C)$. The goal in this step is to establish Eq. (7) below, which makes the connection between rays and surface area. Let $\varphi : \mathcal{D} \to [0, \infty)$ be measurable. Then the following holds:

$$\begin{aligned} \int_C \varphi \,\mathrm{dvol}_{d-1} &\geq \int_B \varphi \circ \pi \,\mathrm{dvol}_{d-1} \\ \int_D \varphi \,\mathrm{dvol}_{d-1} &\geq \frac{1}{2e} \int_B \varphi \circ \kappa \circ \pi \,\mathrm{dvol}_{d-1}\,. \end{aligned} \tag{7}$$

The first inequality is true because the projection of ∂K onto $P(K)$ only decreases surface area. For the second inequality, differentiating Λ at $z \in B$ yields $\mathrm{D}\,\Lambda(z) = \lambda(z)\,\mathrm{Id} + z\nabla\lambda(z)^\top$, which exists vol_{d-1}-a.e. by convexity of λ. Hence,

$$\det(\mathrm{D}\,\Lambda(z)) = \lambda(z)^{d-1}\left(1 + \frac{\langle \nabla\lambda(z), z\rangle}{\lambda(z)}\right) \geq \lambda(z)^{d-1}\left(2 - \frac{\lambda(\mathbf{0})}{\lambda(z)}\right) \geq \frac{1}{2e}\,. \tag{8}$$

where the equality is Sylvester's determinant theorem, the first inequality from convexity of λ, so that $\lambda(\mathbf{0}) \geq \lambda(z) - \langle \nabla\lambda(z), z\rangle$. The last inequality follows because for $z \in B$, $\lambda(z) \geq 1 - 1/(2d)$ and $\lambda(\mathbf{0}) \leq 1$. The claim follows because, by the definition of Λ and Eq. (6), $P \circ \kappa \circ \pi = \Lambda$, which implies that

$$\int_D \varphi \,\mathrm{dvol}_{d-1} \geq \int_{\Lambda B} \varphi \circ \kappa \circ \pi \circ \Lambda^{-1} \,\mathrm{dvol}_{d-1} \geq \frac{1}{2e} \int_B \varphi \circ \kappa \circ \pi \,\mathrm{dvol}_{d-1}\,,$$

where in the first inequality we again used the fact that projections decrease surface area and in the second we used Eq. (8).

Step 3: Combining Let $z \in B$ and $y = \pi(z) \in \partial K$ and $x = \kappa(y) = \Psi(z)x_\star + (1 - \Psi(z))y \in \partial K$. Hence, by Eq. (5) and Lemma 5, the following holds on B,

$$(f-g)^2 \circ \pi + (f-g)^2 \circ \kappa \circ \pi \geq \frac{1}{2}(g_\star + \epsilon - f_\star)^2 \Psi^2 \geq \frac{(g_\star + \epsilon - f_\star)^2}{2^{19} d^2} . \tag{9}$$

Therefore, by Eq. (4), Eq. (7) and Eq. (9), and recalling the definition of ρ in the lemma statement,

$$\begin{aligned}
\int_{\partial K} (f-g)^2 \,\mathrm{d}\rho &\geq \frac{1}{\mathrm{vol}_{d-1}(\partial K)} \int_{C \cup D} (f-g)^2 \,\mathrm{dvol}_{d-1} \\
&\geq \frac{1}{2e\,\mathrm{vol}_{d-1}(\partial K)} \int_B \left((f-g)^2 \circ \pi + (f-g)^2 \circ \kappa \circ \pi\right) \mathrm{dvol}_{d-1} \\
&\geq \frac{(g_\star + \epsilon - f_\star)^2}{2^{20} e d^2} \left(\frac{\mathrm{vol}_{d-1}(B)}{\mathrm{vol}_{d-1}(\partial K)}\right) \\
&\geq \frac{(g_\star + \epsilon - f_\star)^2}{2^{25} e d^2} \min_{\theta \in S^{d-1}} \left(\frac{\mathrm{vol}_{d-1}(P_\theta(K))}{\mathrm{vol}_{d-1}(\partial K)}\right) .
\end{aligned}$$

Hence, since $f_\star \leq g_\star$ by assumption and $g \leq g_\star + \epsilon$ on ∂K,

$$\int_{\partial K} g \,\mathrm{d}\rho - f_\star \leq 2^{14} d \sqrt{\max_{\theta \in S^{d-1}} \left(\frac{\mathrm{vol}_{d-1}(\partial K)}{\mathrm{vol}_{d-1}(P_\theta(K))}\right) \int_{\partial K} (f-g)^2 \,\mathrm{d}\rho} . \qquad \square$$

Remark 7. A convex body $K \subset \mathbb{R}^d$ is in minimal surface area position if its surface area measure (as a measure on the sphere) is isotropic [1, §2.3]. Giannopoulos and Papadimitrakis [7] show that for convex bodies K in minimal surface area position,

$$\max_{\theta \in S^{d-1}} \left(\frac{\mathrm{vol}_{d-1}(\partial K)}{\mathrm{vol}_{d-1}(P_\theta(K))}\right) \leq 2d ,$$

which is sharp when K is a cube. Furthermore, for any convex body K, there exists a linear bijection $T : \mathbb{R}^d \to \mathbb{R}^d$ such that TK is in minimal surface area position.

Proof of Theorem 3. The proof is broken into three parts. First we construct the basic exploratory distribution using Lemma 6. In the second step we define the partitions. The final step puts together the pieces. Without loss of generality, choose coordinates on $\mathcal{K}$ such that $\mathbf{0}$ is the minimiser of $\bar{f}$.

Step 1: Constructing exploratory distributions Let $\epsilon > 0$ and define level set $K_\epsilon = \{x : \bar{f}(x) \leq \bar{f}_\star + \epsilon\}$. Let $\mathscr{F}_\epsilon$ be the set of all $f \in \mathscr{F}$ with minimisers $x_\star$ for which

$$\Psi_{K_\epsilon}(x_\star) \in \left[\frac{1}{128d}, \frac{1}{32d}\right] .$$

Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a linear bijection such that TK_ϵ is in minimal surface area position and define probability measure ρ_ϵ on ∂K by

$$\rho_\epsilon = \frac{\mathrm{vol}_{d-1} \circ T}{\mathrm{vol}_{d-1}(\partial(TK_\epsilon))}\,,$$

which is the pullback of the normalised surface area measure on $\partial(TK_\epsilon)$. That is, for any measurable $\varphi : \mathcal{K} \to \mathbb{R}$,

$$\int_{\partial K_\epsilon} \varphi \, \mathrm{d}\rho_\epsilon = \frac{1}{\mathrm{vol}_{d-1}(\partial(TK_\epsilon))} \int_{\partial(TK_\epsilon)} \varphi \circ T^{-1} \, \mathrm{dvol}_{d-1}\,.$$

Let $f \in \mathscr{F}_\epsilon$ with minimiser $x_\star \in \mathcal{K}$. Since T is a linear bijection, both $f \circ T^{-1}$ and $\bar{f} \circ T^{-1}$ are convex functions from $\mathcal{D} = T\mathcal{K} \to [0,1]$. By Lemma 9,

$$\Psi_{TK_\epsilon}(Tx_\star) = \Psi_{K_\epsilon}(x_\star) \in \left[\frac{1}{128d}, \frac{1}{32d}\right].$$

Hence, by Lemma 6 and Remark 7, for any $f \in \mathscr{F}_\epsilon$,

$$\begin{aligned}
\int_{\partial K_\epsilon} \bar{f} \, \mathrm{d}\rho_\epsilon - f_\star &= \frac{1}{\mathrm{vol}_{d-1}(\partial(TK_\epsilon))} \int_{\partial(TK_\epsilon)} \bar{f} \circ T^{-1} \, \mathrm{dvol}_{d-1} - f_\star \\
&\leq 2^{14} d \sqrt{\frac{2d}{\mathrm{vol}_{d-1}(\partial(TK_\epsilon))} \int_{\partial(TK_\epsilon)} (\bar{f} - f)^2 \circ T^{-1} \, \mathrm{dvol}_{d-1}} \\
&= 2^{14} \sqrt{2d^3 \int_{\partial K_\epsilon} (\bar{f} - f)^2 \, \mathrm{d}\rho_\epsilon}\,,
\end{aligned}$$

which shows that ρ_ϵ satisfies Eq. (3) for all $f \in \mathscr{F}_\epsilon$.

Step 2: Constructing a partition Let $\gamma = 1 + 1/(9d)$ and $\epsilon_0 > 0$ a constant to be tuned later and $\mathcal{E} = \epsilon_0\{1, \gamma, \gamma^2, \ldots\} \cap [0,1]$. Let $f \in \mathscr{F}$ and $\mathbf{0} \neq x_\star \in \mathcal{K}$ be its minimiser. By a continuity argument (Corollary 13), there exists an $\epsilon > 0$ such that $\Psi_{K_\epsilon}(x_\star) = 1/(64d)$. The definition of $\mathcal{E}$ shows that if $\epsilon \geq \epsilon_0$, then there exists a $\delta \in \mathcal{E}$ such that $\delta \in [\epsilon/\gamma, \epsilon]$. By convexity of $\bar{f}$ and monotonicity of level sets, $\frac{1}{\gamma} K_\epsilon \subset K_{\epsilon/\gamma} \subset K_\delta \subset K_\epsilon$. By Corollary 12,

$$\Psi_{K_\delta}(x_\star) \in \left[\frac{1}{128d}, \frac{1}{32d}\right].$$

Hence, by the definition of $\mathscr{F}_\epsilon$,

$$\bigcup_{\epsilon \geq \epsilon_0} \mathscr{F}_\epsilon \subset \bigcup_{\epsilon \in \mathcal{E}} \mathscr{F}_\epsilon\,.$$

The previous step demonstrated the existence of an exploratory distribution for each $\mathscr{F}_\epsilon$. It remains to tune ϵ_0 and handle the functions not in $\bigcup_{\epsilon \in \mathcal{E}} \mathscr{F}_\epsilon$. Define

$$\mathscr{F}_0 = \left\{ f \in \mathscr{F} : \bar{f}_\star - f_\star \leq 2(\bar{f}_\star - f(\mathbf{0})) \right\} \cup \left\{ f \in \mathscr{F} : f_\star \geq \bar{f}_\star - 1/n \right\}.$$

Shortly we show that $\mathscr{F} = \mathscr{F}_0 \cup \bigcup_{\epsilon \in \mathcal{E}} \mathscr{F}_\epsilon$ for a suitable choice of ϵ_0. Before that, let us check that the Dirac at $\mathbf{0}$ is a good exploratory distribution for $f \in \mathscr{F}_0$. Let ρ_0 be a Dirac at $\mathbf{0}$. Then,

$$\int_{\mathcal{K}} \bar{f} \, \mathrm{d}\rho_0 - f_\star = \bar{f}_\star - f_\star \leq \frac{1}{n} + 2|\bar{f}_\star - f(\mathbf{0})| = \frac{1}{n} + \sqrt{4 \int_{\mathcal{K}} \left(\bar{f} - f\right)^2 \mathrm{d}\rho_0} \,.$$

On the other hand, if $f \notin \mathscr{F}_0$ with minimiser $x_\star$, then by the definition of $\mathscr{F}_0$ and the assumption that f is n-Lipschitz,

$$\frac{1}{2n} \leq f(\mathbf{0}) - f(x_\star) \leq n|x_\star| \,. \tag{10}$$

Let $\epsilon > 0$ be such that $\Psi_{K_\epsilon}(x_\star) = 1/(64d)$. Using the assumption that $\bar{f}$ is m-strongly convex means that

$$K_\epsilon \subset \left\{x \in \mathbb{R}^d : |x| \leq \sqrt{2\epsilon/m}\right\} \,.$$

Therefore,

$$\frac{1}{64d} = \Psi_{K_\epsilon}(x_\star) \leq \frac{2\sqrt{2\epsilon/m}}{|x_\star|} \overset{\text{Eq. (10)}}{\leq} 4n^2\sqrt{2\epsilon/m} \,.$$

Rearranging shows that $\epsilon \geq m/(2^{11}d^2n^4) \triangleq \epsilon_0$. Noticing the Eq. (10) also shows that $x_\star \neq \mathbf{0}$, it follows that $f \in \bigcup_{\epsilon \in \mathcal{E}} \mathscr{F}_\epsilon$. Altogether we have shown that $\mathscr{F}$

$$\mathscr{F} = \mathscr{F}_0 \cup \bigcup_{\epsilon \in \mathcal{E}} \mathscr{F}_\epsilon$$

and that for each subset in the union there exists a good exploratory distribution.

Step 3: Combining By definition, $|\mathcal{E}| \leq \lceil \log_\gamma(1/\epsilon_0) \rceil$ with $\gamma = 1 + 1/(9d)$. Combining Lemma 4 with the exploratory distributions and partitions in steps 1 and 2 completes the proof. □

4 Technical lemmas

Here we collect the necessary lemmas. The first four concern concavity, invariance and approximate monotonicity of Ψ.

Lemma 8. *Let $K \subset \mathbb{R}^d$ be a compact convex set with $\mathbf{0} \in K$. Then, for any $x \notin K$ the function $z \mapsto \Psi_K(x, z)$ is concave on $P_x(K)$.*

Proof. Abbreviate $\Psi(z) = \Psi_K(x, z)$. Parameterise K by

$$K = \{z - \alpha x : f(z) \leq \alpha \leq g(z), z \in P_x(K), \alpha \in \mathbb{R}\} \,,$$

where $f : P_x(K) \to \mathbb{R}$ is convex and $g : P_x(K) \to \mathbb{R}$ is concave (Fig. 3). Parameterise the chord connecting x and $z - g(z)x$ by $y(t) = (1-t)x + tz - tg(z)x$. Then,

$$\begin{aligned} 1 - \Psi(z) &= \sup\{t \in (0,1) : y(t) \notin K\} \\ &= \max\{t \in (0,1) : f(tz) \geq t + tg(z) - 1\} \\ &= \max\{1/s : sf(z/s) - g(z) + s - 1 \geq 0, s \in (0,\infty)\}\,. \end{aligned}$$

But $(z, s) \mapsto sf(z/s) - g(z) + s - 1$ is the perspective of f minus a concave function and hence is convex on $P_x(K) \times (0,\infty)$. Hence, since $u \mapsto 1/u$ is convex for $u > 0$, $1 - \Psi(z)$ is convex and therefore Ψ is concave. □

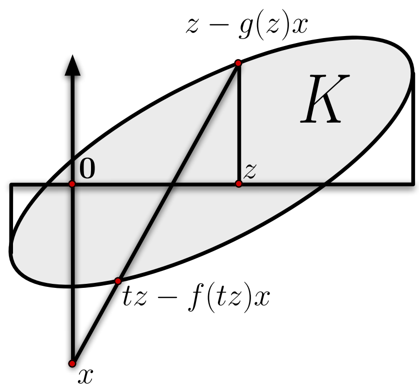

Figure 3: The construction used in the proof of Lemma 8.

Lemma 9. *Let $K \subset \mathbb{R}^d$ be a convex body with $\mathbf{0} \in K$ and let $x \notin K$. Then $\Psi_K(x) = \Psi_{TK}(Tx)$ for all linear bijections $T : \mathbb{R}^d \to \mathbb{R}^d$.*

Proof. Let $\Lambda = P_{Tx} \circ T \circ \pi_K(x, \cdot)$, which is a linear bijection between $P_x(K)$ and $P_{Tx}(TK)$. Furthermore, when $z \in P_x(K)$ and $y = \pi_K(x, z)$, then $Ty = \pi_{TK}(Tx, \Lambda(x))$, which means that

$$\Psi_{TK}(Tx, \Lambda(z)) = \frac{\mathrm{vol}_{d-1}([Tx, Ty] \cap TK)}{\mathrm{vol}_{d-1}([Tx, Ty])} = \frac{\mathrm{vol}_{d-1}([x, y] \cap K)}{\mathrm{vol}_{d-1}([x, y])} = \Psi_K(x, z)\,.$$

Then,

$$\begin{aligned} \Psi_K(x) &= \frac{1}{\mathrm{vol}_{d-1}(P_x(K))} \int_{P_x(K)} \Psi_K(x, z)\, \mathrm{dvol}_{d-1}(z) \\ &= \frac{1}{\mathrm{vol}_{d-1}(P_x(K))} \int_{P_x(K)} \Psi_{TK}(Tx, \Lambda(z))\, \mathrm{dvol}_{d-1}(z) \\ &= \frac{1}{\mathrm{vol}_{d-1}(P_{Tx}(TK))} \int_{\Lambda(P_x(K))} \Psi_{TK}(Tx, y)\, \mathrm{dvol}_{d-1}(y) \\ &= \Psi_{TK}(Tx)\,. \end{aligned}$$

where the first equality is the definition of $\Psi_K(x)$, the second inequality follows because $\Psi_K(x,z) = \Psi_{TK}(TK, \Lambda(z))$ and the third by a change of measure. $\square$

Lemma 10. *Let A and B be convex bodies such that $\mathbf{0} \in A \subset B \subset \gamma A$ for some $\gamma > 1$. Assume that $\Psi^{\infty}_{\gamma A}(x) \leq 1/2$. Then, for any $x \notin \gamma A$,*

$$\Psi_{\gamma A}(x) \leq (2\gamma - 1)^{d+1}\, \Psi_B(x)\,.$$

Proof. Let $\Lambda(z) = z/(2\gamma - 1)$. We claim that

$$\Psi_{\gamma A}(x, z) \leq (2\gamma - 1)^2 \Psi_B(x, \Lambda(z))\,. \tag{11}$$

Setting the proof of this aside for a moment, the consequence is that

$$\begin{aligned}\int_{P_x(\gamma A)} \Psi_{\gamma A}(x,z)\, \mathrm{dvol}_{d-1}(z) &\leq (2\gamma-1)^2 \int_{P_x(\gamma A)} \Psi_B(x, \Lambda(z))\, \mathrm{dvol}_{d-1}(z) \\ &= (2\gamma-1)^{d+1} \int_{\Lambda(P_x(\gamma A))} \Psi_B(x,y)\, \mathrm{dvol}_{d-1}(y) \\ &\leq (2\gamma-1)^{d+1} \int_{P_x(B)} \Psi_B(x,y)\, \mathrm{dvol}_{d-1}\,.\end{aligned}$$

The lemma follows since $\mathrm{vol}_{d-1}(P_x(B)) \leq \mathrm{vol}_{d-1}(P_x(\gamma A))$. It remains to establish Eq. (11), which is a high-school exercise in length chasing. The quantities that follow are defined in the caption of Fig. 4.

$$\frac{|p-u|}{|p-x|} = \frac{|y-w|}{|y-x|} = \frac{|y-v||u-q|}{|y-x||v-q|} = \frac{|u-q|}{|v-q|}\Psi_{\gamma A}(x,z) = \frac{1}{\gamma}\Psi_{\gamma A}(x,z)\,.$$

Hence, there exists a $t \in [r, z]$ such that $\Psi_B(x,t) \geq \Psi_{\gamma A}(x,z)/\gamma$. Furthermore,

$$\frac{|u-w|}{|y-p|} = 1 - \frac{|p-u|}{|p-x|} = 1 - \frac{|u-q|}{|v-q|}\Psi_{\gamma A}(x,z) \qquad \frac{|u-w|}{|y-q|} = \frac{|v-u|}{|v-q|} = \frac{\gamma-1}{\gamma}\,.$$

Dividing one by the other yields

$$\begin{aligned}\frac{|r-q|}{|z-q|} &= \frac{|p-q|}{|y-q|} \\ &= 1 - \frac{|y-p|}{|y-q|} \\ &= 1 - \left(\frac{\gamma-1}{\gamma}\right)\frac{1}{1-\Psi_{\gamma A}(x,z)/\gamma} \in \left[\frac{1}{2\gamma-1}, \frac{1}{\gamma}\right]\,,\end{aligned}$$

where the final relation holds because $\Psi_{\gamma A}(x,z) \in [0, 1/2]$ by assumption. Therefore $r \in [\Lambda(z), z]$ and hence, by the concavity of Ψ_B,

$$\Psi_B(x, \Lambda(z)) \geq \frac{1}{2\gamma-1}\max_{t\in[\Lambda(z),z]} \Psi_B(x,t) \geq \frac{\Psi_{\gamma A}(x,z)}{\gamma(2\gamma-1)} \geq \frac{\Psi_{\gamma A}(x,z)}{(2\gamma-1)^2}\,. \qquad \square$$

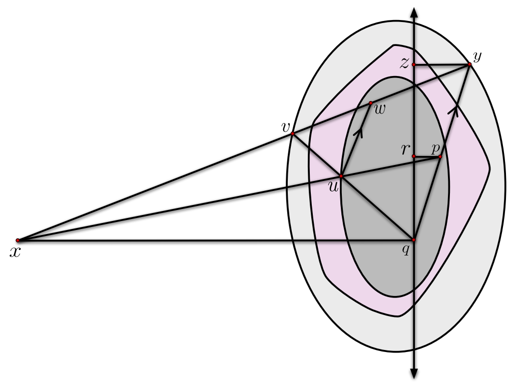

Figure 4: The construction used in the proof of Lemma 10, which has $q = \mathbf{0}$, $y = \pi_{\gamma A}(x, z)$. The point v is chosen so that $[x, y] \cap \gamma A = [v, y]$. The point u is such that $[v, q] \cap A = [u, q]$ and p is the intersection of $[q, y]$ and the affine hull $\mathrm{aff}(\{x, u\})$. Lastly, w is the point in $\mathrm{aff}(\{x, y\})$ such that $[u, w]$ is parallel to $[q, y]$ and $r = P_x(p)$.

Lemma 11. *If $A \subset \mathbb{R}^d$ is a convex body with $\mathbf{0} \in A$ and $\gamma > 1$. Then $\Psi_{\gamma A}(x) \geq \Psi_A(x)$ for all $x \notin \gamma A$.*

Proof. The first step is to argue that $\Psi_{\gamma A}(x, \gamma z) \geq \Psi_A(x, z)$, which follows using the notation in Fig. 5 because

$$\Psi_{\gamma A}(x, \gamma z) = \frac{|\gamma y - w|}{|\gamma y - x|} = \frac{|\gamma y - u|}{|\gamma y - \gamma x|} \geq \Psi_{\gamma A}(\gamma x, \gamma z) = \Psi_A(x, z)\,.$$

Hence,

$$\begin{aligned}
\Psi_{\gamma A}(x) &= \frac{1}{\mathrm{vol}_{d-1}(P_x(\gamma A))} \int_{P_x(\gamma A)} \Psi_{\gamma A}(x, z) \,\mathrm{dvol}_{d-1}(z) \\
&\geq \frac{1}{\mathrm{vol}_{d-1}(P_x(\gamma A)} \int_{P_x(\gamma A)} \Psi_A(x, z/\gamma) \,\mathrm{dvol}_{d-1}(z) \\
&= \frac{\gamma^{d-1}}{\mathrm{vol}_{d-1}(P_x(\gamma A))} \int_{P_x(A)} \Psi_A(x, y) \,\mathrm{dvol}_{d-1}(y) \\
&= \frac{1}{\mathrm{vol}_{d-1}(P_x(A))} \int_{P_x(A)} \Psi_A(x, y) \,\mathrm{dvol}_{d-1}(y) \\
&= \Psi_A(x)\,.
\end{aligned}$$

□

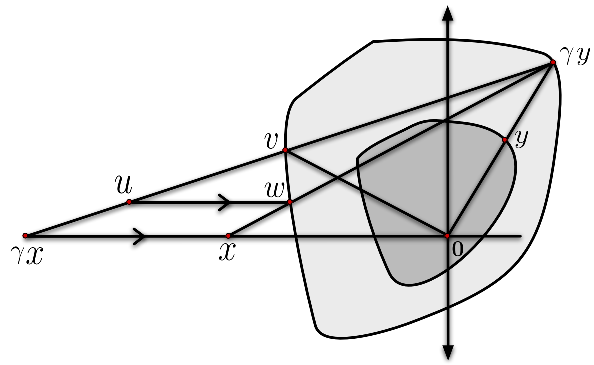

Figure 5: The construction used in the proof of Lemma 11. The point y is $\pi_A(z)$ for some z and $\gamma y = \pi_{\gamma A}(\gamma z)$. The points v and w are chosen so that $[\gamma y, \gamma x] \cap \gamma A = [v, \gamma y]$ and $[\gamma y, x] \cap A = [\gamma y, w]$. Finally, u is chosen on the chord $[\gamma x, \gamma y]$ so that $u - w$ is parallel to x.

Corollary 12. *Suppose that A and B are convex bodies and $\mathbf{0} \in \frac{1}{\gamma} A \subset B \subset A$ for $\gamma = 1 + 1/(9d)$. Then, for any $x \notin A$ with $\Psi_A(x) \leq 1/(2d)$,*

$$\frac{\Psi_B(x)}{\Psi_A(x)} \in [1/2, 2] \,.$$

Proof. The second part of Lemma 14 and concavity of $z \mapsto \Psi_A(x, z)$ shows that $\Psi_A^\infty(x) \leq 1/2$. By assumption, $\frac{1}{\gamma} A \subset B \subset A$, which implies that $\frac{1}{\gamma} B \subset \frac{1}{\gamma} A \subset B$ and so by Lemmas 10 and 11,

$$\begin{aligned} \Psi_B(x) &\leq (2\gamma - 1)^{d+1} \, \Psi_A(x) && \text{(Lemma 10)} \\ &\leq (2\gamma - 1)^{d+1} \, \Psi_{A/\gamma}(x) && \text{(Lemma 11)} \\ &\leq (2\gamma - 1)^{2d+2} \, \Psi_B(x) \,, && \text{(Lemma 10)} \end{aligned}$$

where the last inequality also uses the fact that $\Psi_B^\infty(x) \leq \Psi_A^\infty(x)$. The result follows from the choice of γ and naive bounding. □

Corollary 13. *Suppose that $f \in \mathscr{F}$ is minimised at $\mathbf{0}$ and let $K_\epsilon = \{y \in \mathcal{K} : f(y) \leq f_\star + \epsilon\}$. Then, for any $x \neq \mathbf{0}$, there exists an $\epsilon > 0$ such that*

$$\Psi_{K_\epsilon}(x) = \frac{1}{64d} \,.$$

Proof. That $\epsilon \mapsto \Psi_{K_\epsilon}^\infty(x)$ is continuous and non-decreasing is straightforward. Let $\epsilon_{\max}$ be the smallest value such that

$$\Psi_{K_{\epsilon_{\max}}}^\infty(x) = 1/64 \,,$$

which by the second part of Lemma 14 means that $\Psi_{K_{\epsilon_{\max}}}(x) \geq 1/(64d)$. Strong convexity of f ensures level sets contract to a point as ϵ tends to zero and hence,

$\lim_{\epsilon\to 0}\Psi_{K_\epsilon}(x)=0$. Hence, by the intermediate value theorem it suffices to show that $\epsilon\mapsto\Psi_{K_\epsilon}(x)$ is continuous for $\epsilon\in(0,\epsilon_{\max}]$. Let $\epsilon\in(0,\epsilon_{\max}]$ and $\gamma>1$. By convexity of f, $\frac{1}{\gamma}K_{\gamma\epsilon}\subset K_\epsilon\subset K_{\gamma\epsilon}$. Repeating the argument in the proof of Corollary 12 shows that $\Psi_{K_\epsilon}(x)$ tends to $\Psi_{K_{\gamma\epsilon}}(x)$ as γ tends to 1. □

The next two lemmas are probably known. They concern the law of a concave random variable under the uniform probability measure on the domain, which is shown to have constant mass about its expectation.

Lemma 14. *Let $A\subset\mathbb{R}^{d-1}$ be convex and $\varphi:A\to[0,\infty)$ be concave. Then,*

$$\frac{1}{\mathrm{vol}_{d-1}(A)}\int_A\varphi^2\,\mathrm{dvol}_{d-1}\le 2^{5/2}\left(\frac{\int_A\varphi\,\mathrm{dvol}_{d-1}}{\mathrm{vol}_{d-1}(A)}\right)^2 .$$

Furthermore, $|\varphi|_\infty\le\frac{d}{\mathrm{vol}_{d-1}(A)}\int_A\varphi\,\mathrm{dvol}_{d-1}$.

Proof. Let $B=\{(x,y):x\in A,\,|y|\le\varphi(x)\}\subset\mathbb{R}^d$, which is convex. Define $\theta=(0,\dots,0,1)$ and $h(t)=\mathrm{vol}_{d-1}(B\cap(\theta^\perp+t\theta))$, which is $\frac{1}{d-1}$-concave by Brunn's concavity principle and hence log-concave by the arithmetic-geometric mean inequality. Then,

$$\begin{aligned}
\frac{1}{\mathrm{vol}_{d-1}(A)}\int_A\varphi^2\,\mathrm{dvol}_{d-1}&=\frac{1}{\mathrm{vol}_{d-1}(A)}\int_B|\langle x,\theta\rangle|\,\mathrm{dvol}_d(x)\\
&\le\frac{1}{\mathrm{vol}_{d-1}(A)}\left(\mathrm{vol}_d(B)\int_B\langle x,\theta\rangle^2\,\mathrm{dvol}_d(x)\right)^{1/2}\\
&=\frac{1}{\mathrm{vol}_{d-1}(A)}\left(\mathrm{vol}_d(B)\int_{-\infty}^{\infty}t^2h(t)\,\mathrm{d}t\right)^{1/2}\\
&\le\frac{1}{\mathrm{vol}_{d-1}(A)}\left(\frac{2\,\mathrm{vol}_d(B)}{h(0)^2}\left(\int_{-\infty}^{\infty}h(t)\,\mathrm{d}t\right)^3\right)^{1/2}\\
&=2^{5/2}\left(\frac{\int_A\varphi\,\mathrm{dvol}_{d-1}}{\mathrm{vol}_{d-1}(A)}\right)^2 ,
\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz and the second from Brunn's concavity principle and corollary 2.24 in the notes by Tkocz [15]. The last equality follows since $\int_{-\infty}^{\infty}h(t)\,\mathrm{d}t=\mathrm{vol}_d(B)=2\int_A\varphi\,\mathrm{dvol}_{d-1}$. For the second part, let $x\in A$ be a point with $|\varphi|_\infty=\varphi(x)$ and let $C\subset B$ be the convex hull of $(x,\varphi(x))$ and $A\times\{0\}$. Then,

$$\int_A\varphi\,\mathrm{dvol}_{d-1}=\mathrm{vol}_d(B)\ge\mathrm{vol}_d(C)=\frac{\varphi(x)\,\mathrm{vol}_{d-1}(A)}{d} ,$$

where the inequality is due to convexity of B and the second equality by the formula for the volume of a cone. □

Lemma 15. *Let $A \subset \mathbb{R}^{d-1}$ be convex and $\varphi : A \to [0, \infty)$ be concave and $\nu = \mathrm{dvol}_{d-1} / \mathrm{vol}_{d-1}(A)$ be the uniform probability measure on A. Then,*

$$\nu\left(\left\{x \in A : \frac{\varphi(x)}{\int_A \varphi \, \mathrm{d}\nu} \in [1/4, 16]\right\}\right) \geq 1/32 \,.$$

Proof. By Markov's inequality,

$$\nu\left(\left\{x \in A : \varphi(x) \leq 16 \int_A \varphi \, \mathrm{d}\nu\right\}\right) \geq 1 - \frac{1}{16} \,.$$

On the other hand, by the Paley–Zygmund inequality and Lemma 14,

$$\nu\left(\left\{x \in A : \varphi(x) \geq \frac{1}{4} \int_A \varphi \, \mathrm{d}\nu\right\}\right) \geq \left(1 - \frac{1}{4}\right)^2 \frac{\left(\int_A \varphi \, \mathrm{d}\nu\right)^2}{\int_A \varphi^2 \, \mathrm{d}\nu} \geq \frac{9}{2^{13/2}} \,.$$

Combining the previous two displays and naive simplification yields the result. □

5 Lipschitz and strong convexity relaxation

The last ingredient of the proof is to show that the Lipschitz and strong convexity assumptions are indeed mild. More or less the same argument has been used elsewhere [4, 2].

Proposition 16. *Theorem 1 follows from Theorems 2 and 3.*

Proof. Let $\mathscr{F}$ be the set of n-Lipschitz and m-strongly convex functions from convex body $\mathcal{J} \subset \mathbb{R}^d$ to $[0, 1]$. Theorems 2 and 3 show that

$$\mathfrak{R}_n^*(\mathscr{F}) \leq c d^{2.5} \sqrt{n} \log(n \,\mathrm{diam}(\mathcal{J})/m) \,. \tag{12}$$

Suppose that $\mathcal{K} \subset \mathbb{R}^d$ is a convex body with $B_1^n \subset \mathcal{K}$ and let $(f_t)_{t=1}^n$ be an arbitrary sequence of convex functions from $\mathcal{K}$ to $[0, 1]$, possibly non-Lipschitz and non-strongly convex. Define $\mathrm{dist}(x, A) = \min_{y \in A} |x - y|$ and $\mathcal{J} = \{x \in \mathcal{K} : \mathrm{dist}(x, \partial\mathcal{K}) \geq 1/n\}$, which is a convex subset of $\mathcal{K}$. Next, let $(f'_t)_{t=1}^n$ be the sequence of convex functions from $\mathcal{J} \to [0, 1]$ given by

$$f'_t(x) = \left(\frac{n}{n+1}\right)\left(f_t(x) + \frac{1}{n}\left(\frac{|x|}{\mathrm{diam}(\mathcal{K})}\right)^2\right) \in [0, 1] \,.$$

Boundedness of f_t ensures that f'_t is n-Lipschitz and m-strongly convex with $m = 1/((n + 1)\,\mathrm{diam}(\mathcal{K})^2)$. Running the policy for witnessing Eq. (12),

$$\begin{aligned} c d^{2.5} \sqrt{n} \log(n \,\mathrm{diam}(\mathcal{J})/m) &\geq \max_{y \in \mathcal{J}} \mathbb{E}\left[\sum_{t=1}^n f'_t(x_t) - f'_t(x)\right] \\ &\geq \frac{n}{n+1}\left[\max_{y \in \mathcal{J}} \mathbb{E}\left[\sum_{t=1}^n f_t(x_t) - f_t(y)\right] - 1\right] \\ &\geq \frac{n}{n+1}\left[\max_{x \in \mathcal{K}} \mathbb{E}\left[\sum_{t=1}^n f_t(x_t) - f_t(x)\right] - 2\right] \qquad (13) \end{aligned}$$

Only the last inequality presents any challenge. To see why this is true, let $x \in \mathcal{K}$ be the minimiser of $\sum_{t=1}^n f_t$ and let $y = (1 - 1/n)x$. Since $B_1^n \subset \mathcal{K}$, it follows that $y \in K$. Furthermore, since f_t is convex and bounded in $[0, 1]$,

$$f(y) \leq (1 - 1/n) f(x) + \frac{1}{n} f(\mathbf{0}) \leq f(x) + \frac{1}{n} \,. \qquad \square$$

The result follows by rearranging Eq. (13), noting that $\operatorname{diam}(\mathcal{J}) \leq \operatorname{diam}(\mathcal{K})$ and by substituting the value of m.

References

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

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

[3] 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.

[4] 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, 2017.

[5] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[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.

[7] A. Giannopoulos and M. Papadimitrakis. Isotropic surface area measures. *Mathematika*, 46(1):1–13, 1999.

[8] E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

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

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

[11] X. Hu, LA. Prashanth, A. György, and Cs. Szepesvári. (bandit) convex optimization with biased noisy gradient oracles. In *Artificial Intelligence and Statistics*, pages 819–828, 2016.

[12] F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

[13] 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.

[14] D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(1):2442–2471, 2016. ISSN 1532-4435.

[15] T. Tkocz. *Asymptotic Convex Geometry Lecture Notes*. 2018.