The entropic barrier: a simple and optimal universal self-concordant barrier

Sébastien Bubeck * Ronen Eldan †

December 8, 2014

Abstract

We prove that the Fenchel dual of the log-Laplace transform of the uniform measure on a convex body in $\mathbb{R}^n$ is a $(1 + o(1))n$-self-concordant barrier. This gives the first construction of a universal barrier for convex bodies with optimal self-concordance parameter. The proof is based on basic geometry of log-concave distributions, and elementary duality in exponential families.

1 Introduction

Let $\mathcal{K} \subset \mathbb{R}^n$ be a convex body, namely a compact convex set with a non-empty interior. Our main result is:

**Theorem 1** Let $f : \mathbb{R}^n \to \mathbb{R}$ be defined for $\theta \in \mathbb{R}^n$ by

$$f(\theta) = \log \left( \int_{x \in \mathcal{K}} \exp(\langle \theta, x \rangle) dx \right).$$

Then the Fenchel dual $f^* : \text{int}(\mathcal{K}) \to \mathbb{R}$, defined for $x \in \text{int}(\mathcal{K})$ by $f^*(x) = \sup_{\theta \in \mathbb{R}^n} \langle \theta, x \rangle - f(\theta)$, is a $(1 + \varepsilon_n)n$-self-concordant barrier on $\mathcal{K}$, with $\varepsilon_n \leq 100 \sqrt{\log(n)/n}$, for any $n \geq 80$.

In Section 2 we recall the definition of a $\nu$-self-concordant barrier and its importance in mathematical optimization. We give another point of view on $f^*$ in Section 3, where we show that it corresponds to the negative entropy of a specific element in a canonical exponential family for $\mathcal{K}$. For this reason we refer to $f^*$ as the **entropic barrier** for $\mathcal{K}$. Finally, we prove Theorem 1 in Section 4. Technical lemmas on log-concave distributions are gathered in Section 5, where in particular we derive the sharp bound $\mathbb{E}X^3 \leq 2$ for a real isotropic log-concave random variable $X$.

*Microsoft Research and Princeton University; sebubeck@microsoft.com.
†University of Washington; roneneldan@gmail.com.
2 Context and related work

For a $C^3$-smooth function $g : \mathbb{R}^n \to \mathbb{R}$, denote by $\nabla^2 g[\cdot, \cdot]$ its Hessian which we understand as a bilinear form over $\mathbb{R}^n$. Likewise, by $\nabla^3 g[\cdot, \cdot, \cdot, \cdot]$ we denote its third derivative tensor. We first recall the definition, introduced in Nesterov and Nemirovski [1994], of a self-concordant barrier.

**Definition 1** A function $g : \text{int}(\mathcal{K}) \to \mathbb{R}$ is a barrier for $\mathcal{K}$ if

$$g(x) \xrightarrow{x \to \partial \mathcal{K}} +\infty.$$  

A $C^3$-smooth convex function $g : \text{int}(\mathcal{K}) \to \mathbb{R}$ is self-concordant if for all $x \in \text{int}(\mathcal{K}), h \in \mathbb{R}^n$,

$$\nabla^3 g(x)[h, h, h] \leq 2(\nabla^2 g(x)[h, h])^{3/2}.$$  

Furthermore it is $\nu$-self-concordant if in addition for all $x \in \text{int}(\mathcal{K}), h \in \mathbb{R}^n$,

$$\nabla g(x)[h] \leq \sqrt{\nu} \cdot \nabla^2 g(x)[h, h].$$  

Self-concordant barriers are central objects in the theory of Interior Point Methods (IPMs). The latter class of algorithms has revolutionized mathematical optimization, starting with Karmarkar [1984]. Roughly speaking, an IPM minimizes the linear function $x \in \mathcal{K} \mapsto \langle c, x \rangle$ (for some given $c \in \mathbb{R}^n$) by tracing the central path $(x(t))_{t \in (0, +\infty)}$ of a self-concordant barrier $g$ for $\mathcal{K}$, where $x(t) \in \text{argmin}_x \langle c, x \rangle + \frac{1}{t} g(x)$. The key property of $\nu$-self-concordant barriers is that a step of Newton’s method on the function $x \mapsto \langle c, x \rangle + \frac{1}{t} g(x)$ allows to move from $x((1 - 1/\sqrt{\nu} t))$ to (approximately) $x(t)$, see e.g. Nesterov [2004] for more details. In other words in $O(\sqrt{\nu})$ steps of Newton’s method on $g$ one can approximately minimize a linear function on $\mathcal{K}$.

From a theoretical point of view, one of the most important results in the theory of IPM is Nesterov and Nemirovski’s construction of the universal barrier, which is a $\nu$-self-concordant barrier that always satisfies $\nu \leq Cn$, for some universal constant $C > 0$. To the best of our knowledge, Theorem 1 is the first improvement (for convex bodies) over this seminal result: we show that in fact there always exists a barrier with self-concordance parameter $\nu = (1 + o(1))n$. Up to the second-order term, this improved self-concordance parameter is also optimal, as one must have $\nu \geq n$ for some convex sets (such as a simplex or a hypercube, see [Proposition 2.3.6., Nesterov and Nemirovski [1994]]).

Interestingly, in the case of homogeneous convex cones, an (immediate) generalization of our construction turns out to be identical to Nesterov and Nemirovski’s universal barrier, as proved in Güler [1996]. This connection is nontrivial, and somewhat mysterious to us. In this case our analysis provides a new perspective on the universal barrier, and it allows to improve the bound on its self-concordance parameter to $\nu \leq n$. We note that the recent paper Hildebrand [2014] introduces a new construction which obtains the same bound for general convex cones. Hildebrand’s proof is however much more abstract than ours. In particular his self-concordance barrier, which he calls the canonical barrier, is only defined implicitly as the (convex) potential for the Cheng-Yau metric. We also observe that, while for convex optimization one can assume without loss of generality that $\mathcal{K}$ is a convex cone, there are other applications of the theory of self-concordant barriers where it is important to have a barrier for convex bodies too. We briefly describe such an application in the next section.
It is important to note that the universal, canonical, and entropic barriers are not (immediately at least) relevant in practice. Indeed, the computational effort to implement an IPM depends on the complexity of calculating gradients and Hessians for the barrier. The key to the practical success of IPM is that for important classes of convex sets there exist self-concordant barriers with efficiently computable gradients and Hessians. While this is certainly not immediately the case for the entropic barrier, there is some hope: for instance, its inverse Hessian corresponds to the covariance matrix of a simply described log-concave distribution (a similar statement is true for the universal barrier, but the distribution is more complicated to describe). Furthermore, it can be seen that given a membership oracle to the universal barrier, but the distribution is more complicated to describe. Therefore, it is an elementary calculation to recover a basic duality result for exponential families (see e.g., Wainwright and Jordan [2008]).

Finally we note that even in the simplest situation where $\mathcal{K}$ is a polytope, it remained open until very recently (Lee and Sidford [2014]) to find an efficiently computable barrier with self-concordance parameter nearly matching the one of the universal barrier. We hope that our new barrier will help making progress in finding efficient and optimal barriers, as it is fundamentally different from all previously considered barriers.

## 3 A canonical exponential family

In this section we introduce and briefly study the canonical exponential family $\{p_\theta, \theta \in \mathbb{R}^n\}$ associated with $\mathcal{K}$. For $\theta \in \mathbb{R}^n$, let $p_\theta$ be the probability measure on $\mathbb{R}^n$ whose density with respect to the Lebesgue measure at $x \in \mathbb{R}^n$ is

$$\exp(\langle \theta, x \rangle - f(\theta))1\{x \in \mathcal{K}\},$$

where $f$ is as in (1). In other words $f$ is the log-partition function for this exponential family. We denote $x(\theta) := \mathbb{E}_{X \sim p_\theta} X$. It is well-known (see e.g., [Section 3, Klartag [2006]]) that $\theta \mapsto x(\theta)$ is a bijection between $\mathbb{R}^n$ and $\text{int}(\mathcal{K})$ (we denote $x \in \text{int}(\mathcal{K}) \mapsto \theta(x)$ for the inverse mapping, which is onto $\mathbb{R}^n$), and that $f$ is strictly convex, $C^\infty$-smooth, and $\nabla f(\theta) = x(\theta)$. With these observations it is an elementary calculation to recover a basic duality result for exponential families (see e.g. [Theorem 3.4., Wainwright and Jordan [2008]]), namely that $f^*(x) = -H(p_{\theta(x)})$, where $H(p)$ is the differential entropy of $p$, defined by

$$H(p) := -\int_{\mathbb{R}^n} p(x) \log p(x) dx.$$

Hence the name entropic barrier for $f^*$. Recall also that $\nabla f^*(x) = x(\theta)$.

We will also need higher moments of $f$ and $f^*$. Let $\Sigma(\theta) := \mathbb{E}_{X \sim p_\theta} (X - x(\theta))(X - x(\theta))^\top$, and $T(\theta) := \mathbb{E}_{X \sim p_\theta} (X - x(\theta)) \otimes (X - x(\theta)) \otimes (X - x(\theta))$. It is again an easy exercise (partly done in Klartag [2006]) to show that $\nabla^2 f(\theta) = \Sigma(\theta)$, $\nabla^3 f(\theta) = T(\theta)$ and $\nabla^2 f^*(x) = \Sigma(\theta(x))^{-1}$ (see for example [(2.15), Nemirovski [2004]] for the latter equality).

We summarize the above in a lemma.

**Lemma 1** The functions $f, f^*$ satisfy the following.
(i) The function $f$ is strictly convex on $\mathbb{R}^n$ and the function $f^*$ is strictly convex in the interior of $\mathcal{K}$.

(ii) The function $\theta(\cdot) = \nabla f^*(\cdot)$ is a bijection between the interior of $\mathcal{K}$ and $\mathbb{R}^n$.

(iii) One has for all $\theta \in \mathbb{R}^n$,
$$
\nabla^2 f(\theta) = \mathbb{E}_{X \sim \theta}(X - x(\theta))(X - x(\theta))^\top = \Sigma(\theta).
$$
and
$$
\nabla^3 f(\theta) = \mathbb{E}_{X \sim \theta}(X - x(\theta)) \otimes (X - x(\theta)) \otimes (X - x(\theta)) = T(\theta).
$$

(iv) One has for all $x \in \text{int}(\mathcal{K})$,
$$
\nabla^2 f^*(x) = \left(\nabla^2 f(\theta(x))\right)^{-1} = \left(\mathbb{E}_{X \sim \theta(z)}(X - x)(X - x)^\top\right)^{-1} = \Sigma(\theta(x))^{-1}.
$$

Next we describe an application where the connection between the entropic barrier and the canonical exponential family $\{p_\theta\}$ is crucial.

An application to the linear bandit problem

We consider a sequential extension of linear optimization, known as online linear optimization. It can be described as the following sequential game: at each time step $t = 1, \ldots, T$, a player selects an action $x_t \in \mathcal{K}$, and simultaneously an adversary selects a cost vector $c_t \in \mathcal{K}^\circ$ (where $\mathcal{K}^\circ$ is the polar of $\mathcal{K}$). Both the action and the cost are selected as a function of the history $(x_s, c_s)_{s < t}$, and possibly external randomness (independent for the player and the adversary). The player’s performance at the end of the game is measured through the regret:
$$
R_T = \sum_{t=1}^T \langle c_t, x_t \rangle - \min_{x \in \mathcal{K}} \sum_{t=1}^T \langle c_t, x \rangle,
$$
which compares her cumulative cost to the best cumulative cost she could have obtained in hindsight with a fixed action, if she had known the sequence of costs played by the adversary. This problem has a long history, and a wealth of applications, see, e.g., Cesa-Bianchi and Lugosi [2006]. A far more challenging scenario is when the player only receives a limited feedback on the cost function. Of particular interest is the bandit feedback, where the player only observes her incurred cost $\langle c_t, x_t \rangle \in \mathbb{R}$, rather than the full cost vector $c_t \in \mathbb{R}^n$. See Bubeck and Cesa-Bianchi [2012] for a recent survey on bandit problems. In the following we show how Theorem 1 gives a new point of view on some known results for online linear optimization with bandit feedback.

Since the seminal work of Abernethy et al. [2008] it is known that self-concordant barriers play an important role in the design of good player’s strategies. More precisely the latter paper proposed to run Mirror Descent (which was originally introduced in Nemirovski and Yudin [1983]) with a self-concordant barrier as the mirror map. In addition to the choice of a barrier, one also needs to choose a sampling scheme, that is a mapping from actions to distributions over actions. A key insight of Abernethy et al. [2008] is that the barrier and the sampling scheme should “match” each other, in the sense that the Hessian of the barrier should be approximately proportional to
the inverse covariance of the sampling scheme. In Abernethy et al. [2008] this is achieved with a sampling scheme supported on the Dikin’s ellipsoid, and they prove that with the universal barrier this yields \( \mathbb{E}R_T = O(n^{3/2}/\sqrt{T \log T}) \). By using the entropic barrier together with the sampling scheme \( x \mapsto p_\theta(x) \), it is easy to see that one can improve the bound to \( \mathbb{E}R_T = O(n\sqrt{T \log T}) \), thus matching the state of the art bound of Bubeck et al. [2012] (which, up to the logarithmic factor, is the best possible universal bound). The improvement over Abernethy et al. [2008] is due to the fact that the sampling scheme \( p_\theta(x) \) makes a much better use of the available “space” around \( x \) than the Dikin’s ellipsoid sampling. We also note that Bubeck et al. [2012] obtained their bound via exponential weights on a discretization of \( K \), while it is easy to see that Mirror Descent with the entropic barrier and its associated sampling scheme exactly corresponds to continuous exponential weights, a strategy introduced in Cover [1991] for the full information case. In both cases one has to use the John’s exploration described in Bubeck et al. [2012] to obtain the bound mentioned above, though one can envisage more efficient alternatives such as those described in Hazan et al. [2014].

4 Proof of Theorem 1

Since \( \nabla f(\mathbb{R}^n) = \text{int}(K) \), a basic property of the Fenchel transform is that \( f^* \) is a barrier for \( K \). Next we show that \( f^* \) is self-concordant on \( K \) by proving that \( f \) is self-concordant on \( \mathbb{R}^n \) (the implication then follows from [Section 2.2., Nemirovski [2004]])). By definition, and using equation (3) and (4), \( f \) is self-concordant if for any \( \theta, h \in \mathbb{R}^n \),

\[
\mathbb{E}_{X \sim p_\theta} \langle X - x(\theta), h \rangle^3 \leq 2 \left( \mathbb{E}_{X \sim p_\theta} \langle X - x(\theta), h \rangle^2 \right)^{3/2}.
\]

Noting that \( p_\theta \) is a log-concave measure one immediately obtains the above equation with a worse numerical constant from [(2.21), Ledoux [2001]]. The numerical constant 2 can be obtained via the following lemma, whose proof can be found in Section 5.

**Lemma 2** Let \( X \) be a real log-concave and centered random variable. Then

\[
\mathbb{E}X^3 \leq 2 \left( \mathbb{E}X^2 \right)^{3/2}.
\]

We now move to the main part of the proof, which is to bound the self-concordance parameter \( \nu \) of \( f^* \) by \((1 + \varepsilon_n)n\). By setting \( h = (\nabla^2 f^*(x))^{-1/2} w \), equation (2) becomes

\[
\left\langle (\nabla^2 f^*(x))^{-1/2} \nabla f^*(x), w \right\rangle \leq \sqrt{\nu \langle w, w \rangle}.
\]

and therefore, (2) is equivalent to

\[
\left\langle (\nabla^2 f^*(x))^{-1} \nabla f^*(x), \nabla f^*(x) \right\rangle \leq \nu.
\]

Thus, according to equation (5), we have to show that for any \( \theta \in \mathbb{R}^n \),

\[
\langle \Sigma(\theta) \theta, \theta \rangle \leq (1 + \varepsilon_n)n.
\]
In other words, considering the random variable \( Y = \left\langle \frac{\theta}{\|\theta\|}, X \right\rangle \), with \( X \sim p_\theta \), the proof will be concluded by showing that

\[
\text{Var}(Y) \leq \frac{n}{\|\theta\|^2} (1 + \varepsilon_n). \tag{6}
\]

We denote by \( \rho \) the density of \( Y \), which is proportional to

\[
\text{Vol}_{n-1}(K \cap \{ y\theta/\|\theta\| + \theta^\perp \}) \exp(y\|\theta\|). \tag{7}
\]

At this point we observe that, without loss of generality, we can assume that \( \rho \) is a \( C^\infty \)-smooth function in the interior of its support. Indeed, consider a sequence \( K_1 \subset K_2 \subset \ldots \) of convex bodies with a \( C^\infty \)-smooth boundary which satisfy \( \bigcup_k K_k = K \), and let \( p^{(k)}_\theta \) be the canonical exponential family associated with \( K_k \). Then for all \( \theta \) we have that \( p^{(k)}_\theta \) converges weakly to \( p_\theta \), which implies that the covariance matrix of \( p^{(k)}_\theta \) converges (in operator norm) to \( \Sigma(\theta) \) as \( k \to \infty \). Therefore, it is enough to verify equation (6) for \( p^{(k)}_\theta \). By the smoothness and compactness of \( K_k \), the marginal \( \rho \) will be a smooth function on its support.

The most technical step of the proof is the following lemma, which relies on the log-concavity properties of the 1-dimensional marginals of the uniform measure on \( K \), and which states that \( Y \) is "locally" sub-Gaussian. We give a proof of this lemma at the end of this section.

**Lemma 3** Let \( y_0 \in \text{argmax}_{y \in \mathbb{R}} \rho(y) \), \( M = \frac{\sqrt{7n \log(n)}}{\|\theta\|} \), and \( \sigma^2 = \frac{n}{\|\theta\|^2} \left( 1 - \frac{1}{\sqrt{7n \log(n)/n}} \right) \). There exists \( \zeta : [-M, M] \to [0, 1] \), increasing on \([-M, 0]\), decreasing on \([0, M] \), and with \( \zeta(0) = 1 \), such that for any \( y \in [-M, M] \),

\[
\rho(y + y_0) = \rho(y_0) \zeta(y) \exp\left( -\frac{y^2}{2\sigma^2} \right).
\]

The above lemma implies that, conditionally on \( |Y - y_0| \leq M \), the random variable \( |Y - y_0| \) is stochastically dominated by \( |N(0, \sigma^2)| \). Indeed, the density of \( |Y| \), conditioned on \( |Y - y_0| \leq M \), with respect to the law of \( |N(0, \sigma^2)| \) is equal to \( q(y) := Z(\zeta(y) + \zeta(-y)) 1_{|y| \leq M} \) for a normalization constant \( Z \). Since \( (\zeta(y) + \zeta(-y)) \) is non-increasing, we learn that there exists \( t > 0 \) such that \( q(y) \geq 1 \) for \( y \in [0, t] \) and \( q(y) \leq 1 \) for \( y > t \) which confirms the assertion.

This implies in particular

\[
\mathbb{E}(|Y - y_0|^2 \mid |Y - y_0| \leq M) \leq \sigma^2. \tag{8}
\]

It remains to show that the above conditional variance bound implies (6). For this we use another technical result, whose proof can be found in Section 5:

**Lemma 4** Let \( \varepsilon > 0 \), and \( X \) a real log-concave random variable with density \( \lambda \). Let \( x_1 < x_0 < x_2 \) be three points satisfying \( \lambda(x_1) < \varepsilon \lambda(x_0) \) and \( \lambda(x_2) < \varepsilon \lambda(x_0) \). Then, with \( c(\varepsilon) = \left( 1 + \frac{2}{\log(1/\varepsilon)} \right)^3 \left( 1 + \frac{2}{\log(1/\varepsilon)} + \frac{2}{\log^2(1/\varepsilon)} \right) \), one has

\[
(1 - 2c(\varepsilon)\varepsilon \log^2(1/\varepsilon)) \text{Var}(X) \leq \int_{x_1}^{x_2} (x - x_0)^2 \lambda(x) dx \leq \mathbb{E}(|X - x_0|^2 \mid X \in [x_1, x_2]).
\]
Thanks to Lemma 3, we know that
\[
\max(\rho(y_0 - M), \rho(y_0 + M)) \leq \left(\frac{1}{n}\right)^\frac{3}{2} \left(1 - \sqrt{\frac{\log(n)}{n}}\right) \rho(y_0),
\]
and thus Lemma 4 together with (8) imply that, with \(\varepsilon = \left(\frac{1}{n}\right)^\frac{3}{2} \left(1 - \sqrt{\frac{\log(n)}{n}}\right)\),
\[
\left(1 - \frac{70}{n}\right) \text{Var}(Y) \leq \sigma^2,
\]
which proves (6).

We now conclude the proof of Theorem 1 with the proof of Lemma 3.

**Proof** Let \(\lambda\) be the 1-dimensional marginal of the uniform measure on \(K\) in the direction \(\theta/\|\theta\|\), that is for \(y \in \mathbb{R}\),
\[
\lambda(y) = \frac{\text{Vol}_{n-1}(K \cap \{y\theta/\|\theta\| + \theta^\perp\})}{\text{Vol}(K)}.
\]
We already observed in (7) that
\[
\rho(y) = \frac{\lambda(y) \exp(y\|\theta\|)}{\int_{\mathbb{R}} \lambda(s) \exp(s\|\theta\|)ds}.
\]
It will be useful to consider the functions \(u(y) = \log \lambda(y)\) and
\[
v(y) = \log \rho(y) = u(y) + y\|\theta\| - \log \left(\int_{\mathbb{R}} \lambda(s) \exp(s\|\theta\|)ds\right).
\]
Since we assumed that \(K\) is a \(C^\infty\)-smooth domain, it is clear that \(\lambda\) and \(\rho\) are also \(C^\infty\) on the interior of their support \([a,b]\), where
\[
a = \inf\{s \in \mathbb{R} : \lambda(s) > 0\}, \quad b = \sup\{s \in \mathbb{R} : \lambda(s) > 0\}.
\]
The key observation is that, thanks to the Brunn-Minkowski inequality, \(\lambda\) is \(n\)-concave on its support (see, e.g., Borell [1975]), or in other words \(\lambda^{1/n}\) is a concave function on the interval \((a,b)\). We now obtain a simple differential inequality by using the following lemma, whose proof can be found in Section 5:

**Lemma 5** Let \(\varphi \in C^2((a,b))\), and \(\zeta(x) = \log \varphi(x)\). Then
\[
\varphi \text{ is } n\text{-concave in } (a,b) \iff \zeta'' \leq -\frac{1}{n} (\zeta')^2 \text{ in } (a,b).
\]
The result of the lemma directly yields
\[
v''(y) = u''(y) \leq -\frac{1}{n} (u'(y))^2 = -\frac{1}{n} (v'(y) - \|\theta\|)^2.
\]
It is useful to rewrite the above inequality in terms of \(w(y) := v'(y + y_0) + \frac{\|\theta\|^2 v}{n - \|\theta\|y}, \quad y \in (a',b')\), with \(a' = a - y_0, b' = \min(b - y_0, n/\|\theta\|)\), in which case one easily obtains
\[
-w'(y) \geq \frac{1}{n} \left( w(y)^2 - 2w(y) \frac{\|\theta\| n}{n - \|\theta\| y} \right).
\]
Observe that \( w(0) = 0 \), as \( y_0 \) is a local maximum of the smooth function \( v \). A simple application of Gronwall’s inequality teaches us that \( w \) is non-positive in the interval \([0, b')\) and non-negative in the interval \((a', 0)\). We conclude that, for \( y \in (a', b')\),

\[
v(y + y_0) - v(y_0) = \int_0^y v'(s + y_0) ds \\
= -\int_0^y \frac{\|\theta\|^2 s}{n - \|\theta\| s} ds + \int_0^y w(s) ds \\
= \|\theta\| y + n \log(1 - \|\theta\| y/n) + \int_0^y w(s) ds \\
= -\left(1 - \sqrt{\frac{7 \log(n)}{n}}\right) \frac{y^2}{2n/\|\theta\|^2} + n h \left(\frac{\|\theta\| y}{\sqrt{\frac{7 \log(n)}{n}}}\right) + \int_0^y w(s) ds,
\]

where \( h(x, \varepsilon) = x + (1 - \varepsilon)x^2 + \log(1 - x) \). For any \( \varepsilon \in (0, 1) \), \( x \mapsto h(x, \varepsilon) \) is increasing on \([-\varepsilon/(1 - \varepsilon), 0]\) and decreasing on \([0, 1]\). In particular, denoting \( a'' = \min(a', \sqrt{7n \log(n)/\|\theta\|}) \) and \( \sigma^2 \) as in the statement of the lemma, we showed that there exists a function \( \xi : (a'', b') \to \mathbb{R}_- \), increasing on \((a'', 0)\), decreasing on \([0, b')\), with \( \xi(0) = 0 \), and such that

\[
v(y + y_0) - v(y_0) = -\frac{y^2}{2\sigma^2} + \xi(y).
\]

This easily concludes the proof.

\[\square\]

5 Technical lemmas

In this section we prove Lemma 2, 4, 5 that were used in the preceding section to prove Theorem 1. In each case we first restate the lemma before going into the proof.

**Lemma 2** Let \( X \) be a real log-concave and centered random variable. Then

\[
\mathbb{E}X^3 \leq 2 \left(\mathbb{E}X^2\right)^{3/2}.
\]

**Proof** We may assume that \( X \) is supported in a compact interval \([-M, M]\). Indeed, if that is not the case, we can define \( X_k \) to have the law of \( X \) conditioned on the interval \([-k, k]\); if the result for the lemma holds for every \( X_k \), we may take limits and deduce its correctness for \( X \).

Define \( g(x) = 1 - x^2 \). Denote by \( P_g \) the family of log-concave probability measures \( \mu \) on \( \mathbb{R} \), supported on \([-M, M]\) which satisfy \( \int g d\mu \geq 0 \). Let \( \Psi : P_g \to \mathbb{R} \) be a convex functional. According to [Fradelizi and Guédon, 2004, Theorem 1, Theorem 2], which characterizes the extremal points of the set \( P_g \), the supremum \( \sup_{\mu \in P_g} \Psi(\mu) \) is attained either on a Dirac measure \( \delta_x \) for some \( x \in \mathbb{R} \) or on a measure \( \mu \), which satisfies the following:

(i) The density of \( \mu \) is log-affine on its support, hence there exist \( a, b \in [-M, M] \), \( a < b \) and \( c \in \mathbb{R} \) such that

\[
\frac{d\mu}{dx} = Z^{-1} 1_{[a,b]} e^{cx},
\]

where

\[
Z = \frac{1}{\phi(b) - \phi(a)},
\]

\( \phi(x) = \int_{-\infty}^x \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt \)

and

\[
\phi(a) = 1 - \frac{1}{\sqrt{2\pi}c/s} \int_{a/s}^{-1/s} \frac{e^{-t^2/2}}{\sqrt{2\pi}c/s} dt,
\]

\( s = \sqrt{7 \log(n)/\|\theta\|} \)

or

(ii) \( \mu \) is a Dirac measure \( \delta_x \) for some \( x \in \mathbb{R} \).

In either case

\[
\int g d\mu \geq \int g d\delta_x = 1 - x^2 \rightarrow \infty \quad \text{as} \quad x \rightarrow \pm \infty
\]

and

\[
\int \Psi(\mu) \geq \int \Psi(\delta_x) = \Psi(\delta_x) = \Psi(\delta_x) = \Psi(\delta_x).
\]

This proves the statement.
where $Z$ is chosen such that $\mu$ is a probability measure.

(ii) One has $\int g \, d\mu = 0$.

(iii) There is $\xi \in \{-1, 1\}$ such that for all $x \in [a, b]$ one has $\xi \int_a^x g \, d\mu \geq 0$.

Consider the linear functional $\Psi(\mu) = \int (x^3 - 3x) \, d\mu$. It is clear that the statement of the lemma will follow if we establish that

$$\sup_{\mu \in P_g} \Psi(\mu) \leq 2. \quad (10)$$

(note that for centered measures $\mu$, $\Psi(\mu)$ is exactly $\int x^3 \, d\mu(x)$). Now, it is clear that if $\mu$ is a Dirac measure satisfying $\int g \, d\mu \geq 0$, we have that

$$\Psi(\mu) \leq \sup_{x \in [-1, 1]} (x^3 - 3x) = 2.$$  

Therefore, according to the above, it is enough to verify that the bound of equation (10) holds for measures of the form (9), which satisfy the constraint

$$\int x^2 \, d\mu(x) = 1, \quad (11)$$

and such that either $-1 \leq a \leq 1$ or $-1 \leq b \leq 1$ (this is a direct consequence of (iii) above).

In other words, the lemma will be established if we show that

$$H(a, b, c) := Z(a, b, c)^{-1} \int_a^b (x^3 - 3x) e^{cx} \, dx \leq 2$$

for all $a, b, c$ such that either $|a| \leq 1$ or $|b| \leq 1$ and such that

$$G(a, b, c) := Z(a, b, c)^{-1} \int_a^b (x^2 - 1) e^{cx} \, dx = 0. \quad (13)$$

We will continue the proof under the assumption that $|a| \leq 1$. The proof under the assumption $|b| \leq 1$ is similar.

A calculation gives

$$Z(a, b, c) = \int_a^b e^{cx} \, dx = \frac{e^{cb} - e^{ca}}{c}, \quad (14)$$

and by integrating by parts, we have that

$$H(a, b, c) = \frac{1}{cZ(a, b, c)} (x^3 - 3x) e^{cx} \bigg|_a^b - \frac{1}{cZ(a, b, c)} \int_0^1 3(x^2 - 1) e^{cx} \, dx \quad (15)$$

$$\quad \overset{(14), (13)}{=} \frac{(b^3 - 3b)e^{cb} - (a^3 - 3a)e^{ca}}{e^{cb} - e^{ca}}.$$  

It will be useful to set $r = e^{c(b-a)}$ and, respectively, define $c(a, b, r) = \log(r) / (b - a)$.

Next, note that for any $a \in [-1, 1]$ and any $r > 0$, one has $G(a, 1, c(a, 1, r)) \leq 0$. Moreover, observe that $G(a, b, c(a, b, r))$ is continuously increasing with respect to $b$ when $b \geq 1$. Thus, by
the intermediate value theorem, there exists a unique point $b > a$ such that $G(a, b, c(a, b, r)) = 0$. Let us try to obtain a formula for this point. A straightforward calculation yields

$$G(a, b, c) = \frac{(bc(bc - 2) + 2 - c^2)e^{bc} - (ac(ac - 2) + 2 - c^2)e^{ac}}{c^2(e^{bc} - e^{ac})},$$

which means that the constraint $G(a, b, c) = 0$ is equivalent to

$$r = \frac{ac(ac - 2) + 2 - c^2}{bc(bc - 2) + 2 - c^2} = \frac{a \log(r)\left(a \log(r) - 2\right) + 2 - \left(\log(r)\right)^2}{b \log(r)\left(b \log(r) - 2\right) + 2 - \left(\log(r)\right)^2}.$$

Or in other words

$$r \left(b \log(r) (b \log(r) - 2(b - a)) + 2(b - a)^2 - \log^2(r)\right) = a \log(r) (a \log(r) - 2(b - a)) + 2(b - a)^2 - \log^2(r).$$

At this point, we see that $b$ can be expressed as a function of $(a, r)$ as a root of a quadratic equation. Out of the two possible roots of this equation, it is easily checked that exactly one is greater than $a$. We get the explicit formula

$$b = b(a, r) := \frac{\log(r) (\text{sgn}(r - 1)g(a, r) - a(r + 1)) + 2a(r - 1)}{q(a, r)},$$

where

$$g(a, r) = \sqrt{-(a^2 - 2)(r - 1)^2 + r (a^2 + r - 1) \log^2(r) - 2r(r - 1) \log(r)},$$

and

$$q(a, r) = (2r + r \log^2(r) - 2r \log(r) - 2).$$

Using this formula, proving equation (12) under the constraint (13) amounts to showing that

$$H(a, b(a, r), c(a, b(a, r), r)) \leq 2, \ \forall (a, r) \in [-1, 1] \times (0, \infty).$$

(16)

Plugging the definition of $b(a, r)$ and $c(a, r)$ into (15) gives

$$H(a, r) := H(a, b(a, r), c(a, b(a, r), r)) = \frac{r (b(a, r)^3 - 3b(a, r)) - (a^3 - 3a)}{r - 1}.$$

It turns out that $H(a, r)$ is monotone decreasing in both $a$ and $r$ in the domain $[-1, 1] \times (0, \infty)$ (up to a removable discontinuity on $[-1, 1] \times \{1\}$). Moreover, it is straightforward to check that $\lim_{r \to 0^+} H(-1, r) = 2$. These two facts complete the proof of inequality (16). We omit further details of this proof.
**Lemma 4** Let $\varepsilon > 0$, and $X$ a real log-concave random variable with density $\lambda$. Let $x_1 < x_0 < x_2$ be three points satisfying $\lambda(x_1) < \varepsilon \lambda(x_0)$ and $\lambda(x_2) < \varepsilon \lambda(x_0)$. Then, with $c(\varepsilon) = \left(1 + \frac{2}{\log(1/\varepsilon)}\right)^3 \left(1 + \frac{2}{\log(1/\varepsilon)} + \frac{2}{\log^2(1/\varepsilon)}\right)$, one has

$$
(1 - 2c(\varepsilon)\varepsilon \log^2(1/\varepsilon)) \Var(X) \leq \int_{x_1}^{x_2} (x - x_0)^2 \lambda(x) dx \\
\leq \mathbb{E}(|X - x_0|^2 | X \in [x_1, x_2]).
$$

**Proof** We only have to prove the first inequality, as the second one is obviously true. By rescaling and translating, we can assume without loss of generality that $x_0 = 0$ and $\mathbb{E}(X^2) = 1$. Under these conditions we will prove the slightly stronger following bound:

$$
\int_{x_1}^{x_2} x^2 \lambda(x) dx \geq 1 - 2 \left(1 + \frac{2}{\log(1/\varepsilon)}\right)^3 \left(1 + \frac{2}{\log(1/\varepsilon)} + \frac{2}{\log^2(1/\varepsilon)}\right) \varepsilon \log^2(1/\varepsilon).
$$

(17)

We prove that in fact

$$
\int_{-\infty}^{+\infty} x^2 \lambda(x) dx \leq \left(1 + \frac{2}{\log(1/\varepsilon)}\right)^3 \left(1 + \frac{2}{\log(1/\varepsilon)} + \frac{2}{\log^2(1/\varepsilon)}\right) \varepsilon \log^2(1/\varepsilon),
$$

(18)

and an identical computation yields the same upper bound for the integral on $(-\infty, x_1]$, which then concludes the proof of (17).

First note that, by log-concavity of $\lambda$, one has for any $x > x_2$,

$$
\frac{\lambda(x)}{\lambda(x_2)} \leq \left(\frac{\lambda(x_2)}{\lambda(x_0)}\right)^{\frac{x-x_2}{x_2}} \leq \varepsilon \frac{x-x_2}{x_2}.
$$

Using [Lemma 5.5 (a), Lovász and Vempala [2007]] one has $\lambda(0) \leq 1$, and thus $\lambda(x_2) \leq \varepsilon$, which together with the above display yields

$$
\lambda(x) \leq \varepsilon \exp\left(-\frac{x-x_2}{x_2} \log(1/\varepsilon)\right).
$$

This directly implies that

$$
\int_{x_2}^{+\infty} x^2 \lambda(x) dx \leq \frac{x_2}{\log(1/\varepsilon)} \left(1 + \frac{2}{\log(1/\varepsilon)} + \frac{2}{\log^2(1/\varepsilon)}\right) \varepsilon \log^2(1/\varepsilon)^2.
$$

Finally, using [Lemma 5.7, Lovász and Vempala [2007]] it is easy to see that without loss of generality one can assume that $x_2 \leq \log(1/\varepsilon) + 2$, and thus the above display directly yields (18). ▮

**Lemma 5** Let $\varphi \in C^2((a, b))$, and $\zeta(x) = \log \varphi(x)$. Then

$$
\varphi \text{ is } n\text{-concave in } (a, b) \iff \zeta'' \leq -\frac{1}{n} (\zeta')^2 \text{ in } (a, b).
$$

**Proof** Denote $\psi(x) = \varphi^{1/n}(x)$ and $\xi(x) = \log \psi(x) = \frac{1}{n} \zeta(x)$. Then

$$
\xi''(x) = \frac{\psi''(x)}{\psi(x)} - \frac{\psi'(x)^2}{\psi(x)^2} = \frac{\varphi^{1/n}(x)''}{\varphi^{1/n}(x)} - \xi'(x)^2.
$$

So $\varphi^{1/n}$ is concave if and only if $\xi'' \leq - (\xi')^2$ which proves the fact. ▮
Acknowledgements

The authors would like to thank Jacob Abernethy and Elad Hazan and Bo’az Klartag for interesting discussions. We are grateful to François Glineur for pointing out the connection with the universal barrier in the case of homogeneous convex cones.

References

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

C. Borell. Convex set functions in d-space. *Periodica Mathematica Hungarica*, 6(2):111–136, 1975.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

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

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

T. M. Cover. Universal portfolios. *Mathematical finance*, 1(1):1–29, 1991.

M. Fradelizi and O. Guédon. The extreme points of subsets of s-concave probabilities and a geometric localization theorem. *Discrete and Computational Geometry*, 31(2):327–335, 2004.

O. Güler. Barrier functions in interior point methods. *Mathematics of Operations Research*, 21(4): 860–885, 1996.

E. Hazan, Z. Karnin, and R. Meka. Volumetric spanners: an efficient exploration basis for learning. In *Proceedings of The 27th Conference on Learning Theory*, 2014.

R. Hildebrand. Canonical barriers on convex cones. *Mathematics of Operations Research*, 39: 841–850, 2014.

N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4: 373–395, 1984.

B. Klartag. On convex perturbations with a bounded isotropic constant. *Geometric and Functional Analysis*, 16(6):1274–1290, 2006.

M. Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, 2001.
Y. T. Lee and A. Sidford. Path-finding methods for linear programming: Solving linear programs in $\tilde{O}(\sqrt{\text{rank}})$ iterations and faster algorithms for maximum flow. In *55th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2014.

L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

A. Nemirovski. Interior point polynomial time methods in convex programming. *Lecture Notes*, 2004.

A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004.

Y. Nesterov and A. Nemirovski. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.