Mathematical Programming Computation (2024) 16:185–228
https://doi.org/10.1007/s12532-024-00252-0

FULL LENGTH PAPER

A new computational framework for log-concave density estimation

Wenyu Chen[1] · **Rahul Mazumder**[2] · **Richard J. Samworth**[3]

Received: 1 May 2022 / Accepted: 26 June 2023 / Published online: 30 April 2024
© The Author(s) 2024

Abstract

In statistics, log-concave density estimation is a central problem within the field of nonparametric inference under shape constraints. Despite great progress in recent years on the statistical theory of the canonical estimator, namely the log-concave maximum likelihood estimator, adoption of this method has been hampered by the complexities of the non-smooth convex optimization problem that underpins its computation. We provide enhanced understanding of the structural properties of this optimization problem, which motivates the proposal of new algorithms, based on both randomized and Nesterov smoothing, combined with an appropriate integral discretization of increasing accuracy. We prove that these methods enjoy, both with high probability and in expectation, a convergence rate of order $1/T$ up to logarithmic factors on the objective function scale, where T denotes the number of iterations. The benefits of our new computational framework are demonstrated on both synthetic and real data, and our implementation is available in a github repository `LogConcComp` (Log-Concave Computation).

Research supported in part by grants from the Office of Naval Research: N000141812298, N000142112841, the National Science Foundation and IBM to Rahul Mazumder. The research of Richard J. Samworth was supported by EPSRC Grants EP/N031938/1 and EP/P031447/1, as well as ERC Advanced Grant 101,019,498.

✉ Rahul Mazumder
rahulmaz@mit.edu

Wenyu Chen
wenyu.main@gmail.com

Richard J. Samworth
r.samworth@statslab.cam.ac.uk

1 MIT Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA

2 MIT Sloan School of Management, Operations Research Center, MIT Center for Statistics, Massachusetts Institute of Technology, Cambridge, MA, USA

3 Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge, UK

Keywords Dual averaging · Large-scale computation · Log-concavity · Nesterov smoothing · Randomized smoothing · Shape-constrained density estimation

Mathematics Subject Classification 90C06 · 90C25 · 62G05 · 62G07

1 Introduction

In Statistics, the field of nonparametric inference under shape constraints dates back at least to [33], who studied the nonparametric maximum likelihood estimator of a decreasing density on the non-negative half line. But it is really over the last decade or so that researchers have begun to realize its full potential for addressing key contemporary data challenges such as (multivariate) density estimation and regression. The initial allure is the flexibility of a nonparametric model, combined with estimation methods that can often avoid the need for tuning parameter selection, which can often be troublesome for other nonparametric techniques such as those based on smoothing. Intensive research efforts over recent years have revealed further great attractions: for instance, these procedures frequently attain optimal rates of convergence over relevant function classes. Moreover, it is now known that shape-constrained procedures can possess intriguing adaptation properties, in the sense that they can estimate particular subclasses of functions at faster rates, even (nearly) as well as the best one could do if one were told in advance that the function belonged to this subclass.

Typically, however, the implementation of shape-constrained estimation techniques requires the solution of an optimization problem, and, despite some progress, there are several cases where computation remains a bottleneck and hampers the adoption of these methods by practitioners. In this work, we focus on the problem of log-concave density estimation, which has become arguably the central challenge in the field because the class of log-concave densities enjoys stability properties under marginalization, conditioning, convolution and linear transformations that make it a very natural infinite-dimensional generalization of the class of Gaussian densities [60].

The univariate log-concave density estimation problem was first studied in [68], and fast algorithms for the computation of the log-concave maximum likelihood estimator (MLE) in one dimension are now available through the `R` packages `logcondens` [27] and `cnmlcd` [49]. [20] introduced and studied the multivariate log-concave maximum likelihood estimator, but their algorithm, which is described below and implemented in the `R` package `LogConcDEAD` [18], is slow; for instance, [20] report a running time of 50 s for computing the bivariate log-concave MLE with 500 observations, and 224 min for computing the log-concave MLE in four dimensions with 2,000 observations. An alternative, interior point method for a suitable approximation was proposed by [46]. Recent progress on theoretical aspects of the computational problem in the computer science community includes [2], who proved that there exists a polynomial time algorithm for computing the log-concave maximum likelihood estimator. We are unaware of any attempt to implement this algorithm. [57] compute an approximation to the log-concave MLE by considering $-\log p$ as a piecewise affine maximum function, using the log-sum-exp operator to approximate the non-smooth operator, a Riemann

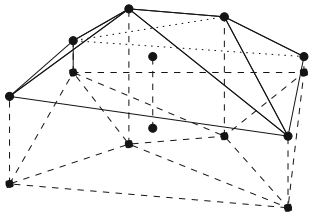

Fig. 1 An illustration of a tent function, taken from [20]

sum to compute the integral and its gradient, and obtain a solution via L-BFGS. This reformulation means that the problem is no longer convex.

To describe the problem more formally, let $\mathcal{C}_d$ denote the class of proper, convex lower-semicontinuous functions $\varphi : \mathbb{R}^d \to (-\infty, \infty]$ that are coercive in the sense that $\varphi(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$. The class of upper semi-continuous log-concave densities on $\mathbb{R}^d$ is denoted as

$$\mathcal{P}_d := \left\{ p : \mathbb{R}^d \to [0, \infty) : p = e^{-\varphi} \text{ for some } \varphi \in \mathcal{C}_d, \int_{\mathbb{R}^d} p = 1 \right\}.$$

Given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, [20, Theorem 1] proved that whenever the convex hull C_n of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is d-dimensional, there exists a unique

$$\hat{p}_n \in \operatorname*{argmax}_{p \in \mathcal{P}_d} \frac{1}{n} \sum_{i=1}^{n} \log p(\boldsymbol{x}_i). \tag{1}$$

If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are regarded as realizations of independent and identically distributed random vectors on $\mathbb{R}^d$, then the objective function in (1) is a scaled version of the log-likelihood function, so $\hat{p}_n$ is called the *log-concave MLE*. The existence and uniqueness of this estimator is not obvious, because the infinite-dimensional class $\mathcal{P}_d$ is non-convex, and even the class of negative log-densities $\left\{\varphi \in \mathcal{C}_d : \int_{\mathbb{R}^d} e^{-\varphi} = 1\right\}$ is non-convex. In fact, the estimator belongs to a finite-dimensional subclass; more precisely, for a vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n) \in \mathbb{R}^n$, define $\operatorname{cef}[\boldsymbol{\phi}] \in \mathcal{C}_d$ to be the (pointwise) largest function with

$$\operatorname{cef}[\boldsymbol{\phi}](\boldsymbol{x}_i) \le \phi_i$$

for $i = 1, \ldots, n$. [20] proved that $\hat{p}_n = e^{-\operatorname{cef}[\boldsymbol{\phi}^*]}$ for some $\boldsymbol{\phi}^* \in \mathbb{R}^n$, and refer to the function $-\operatorname{cef}[\boldsymbol{\phi}^*]$ as a 'tent function'; see the illustration in Fig. 1. [20] further defined the non-smooth, convex objective function $f : \mathbb{R}^n \to \mathbb{R}$ by

$$f(\boldsymbol{\phi}) \equiv f(\phi_1, \ldots, \phi_n) := \frac{1}{n} \sum_{i=1}^{n} \phi_i + \int_{C_n} \exp\left\{-\operatorname{cef}[\boldsymbol{\phi}](x)\right\} dx, \tag{2}$$

and proved that $\boldsymbol{\phi}^* = \operatorname{argmin}_{\boldsymbol{\phi} \in \mathbb{R}^n} f(\boldsymbol{\phi})$.

The two main challenges in optimizing the objective function f in (2) are that the value and subgradient of the integral term are hard to evaluate, and that it is

non-smooth, so vanilla subgradient methods lead to a slow rate of convergence. To address the first issue, [20] computed the exact integral and its subgradient using the `qhull` algorithm [4] to obtain a triangulation of the convex hull of the data, evaluating the function value and subgradient over each simplex in the triangulation. However, in the worst case, the triangulation can have $O(n^{d/2})$ simplices [50]. The non-smoothness is handled via Shor's r-algorithm [66, Chapter 3], as implemented by [42]. In Sect. 2, we characterize the subdifferential of the objective function in terms of the solution of a linear program (LP), and show that the solution lies in a known, compact subset of $\mathbb{R}^n$. This understanding allows us to introduce our new computational framework for log-concave density estimation in Sect. 3, based on an accelerated version of a dual averaging approach [53]. This relies on smoothing the objective function, and encompasses two popular strategies, namely Nesterov smoothing [52] and randomized smoothing [25, 48, 73], as special cases. A further feature of our algorithm is the construction of approximations to gradients of our smoothed objective, and this in turn requires an approximation to the integral in (2). While a direct application of the theory of [25] would yield a rate of convergence for the objective function of order $n^{1/4}/T + 1/\sqrt{T}$ after T iterations, we show in Sect. 4 that by introducing finer approximations of both the integral and its gradient as the iteration number increases, we can obtain an improved rate of order $1/T$, up to logarithmic factors. Moreover, we translate the optimization error in the objective into a bound on the error in the log-density, which is uncommon in the literature in the absence of strong convexity. A further advantage of our approach is that we are able to extend it in Sect. 5 to the more general problem of quasi-concave density estimation [46, 65], thereby providing a computationally tractable alternative to the discrete Hessian approach of [46]. Section 6 illustrates the practical benefits of our methodology in terms of improved computational timings on simulated data. Additional experimental details and applications on real data sets are provided in Appendix A. Proofs of all main results can be found in Appendix B, and background on the field of nonparametric inference under shape constraints can be found in Appendix C.

Notation: We write $[n] := \{1, 2, \ldots, n\}$, let $\mathbf{1} \in \mathbb{R}^n$ denote the all-ones vector, and denote the cardinality of a set S by $|S|$. For a Borel measurable set $C \subseteq \mathbb{R}^d$, we use $\mathrm{vol}(C)$ to denote its volume (i.e. d-dimensional Lebesgue measure). We write $\|\cdot\|$ for the Euclidean norm of a vector. For $\mu > 0$, a convex function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be μ-strongly convex if $\boldsymbol{\phi} \mapsto f(\boldsymbol{\phi}) - \frac{\mu}{2}\|\boldsymbol{\phi}\|^2$ is convex. The notation $\partial f(\boldsymbol{\phi})$ denotes the subdifferential (set of subgradients) of f at $\boldsymbol{\phi}$. Given a real-valued sequence (a_n) and a positive sequence (b_n), we write $a_n = \tilde{O}(b_n)$ if there exist $C, \gamma > 0$ such that $a_n \leq C b_n \log^{\gamma}(1 + n)$ for all $n \in \mathbb{N}$.

2 Understanding the structure of the optimization problem

Throughout this paper, we assume that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ are distinct and that their convex hull $C_n := \mathrm{conv}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ has nonempty interior, so that $n \geq d + 1$ and $\Delta := \mathrm{vol}(C_n) > 0$. This latter assumption ensures the existence and uniqueness of a minimizer of the objective function in (2) [28, Theorem 2.2]. Recall that we define

the *lower convex envelope* function [59] $\mathrm{cef} : \mathbb{R}^n \to \mathcal{C}_d$ by

$$\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x}) \equiv \mathrm{cef}[(\phi_1, \ldots, \phi_n)](\boldsymbol{x}) := \sup\big\{g(\boldsymbol{x}) : g \in \mathcal{C}_d, g(\boldsymbol{x}_i) \le \phi_i \ \forall i \in [n]\big\}. \tag{3}$$

As mentioned in the introduction, in computing the MLE, we seek

$$\boldsymbol{\phi}^* := \operatorname*{argmin}_{\boldsymbol{\phi}\in\mathbb{R}^n} f(\boldsymbol{\phi}), \tag{4}$$

where

$$f(\boldsymbol{\phi}) := \frac{1}{n}\mathbf{1}^\top\boldsymbol{\phi} + \int_{C_n} \exp\{-\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x})\}\, d\boldsymbol{x} =: \frac{1}{n}\mathbf{1}^\top\boldsymbol{\phi} + I(\boldsymbol{\phi}). \tag{5}$$

Note that (4) can be viewed as a stochastic optimization problem by writing

$$f(\boldsymbol{\phi}) = \mathbb{E}F(\boldsymbol{\phi}, \boldsymbol{\xi}), \tag{6}$$

where $\boldsymbol{\xi}$ is uniformly distributed on C_n and where, for $\boldsymbol{x} \in C_n$,

$$F(\boldsymbol{\phi}, \boldsymbol{x}) := \frac{1}{n}\mathbf{1}^\top\boldsymbol{\phi} + \Delta e^{-\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x})}. \tag{7}$$

Let $\boldsymbol{X} := [\boldsymbol{x}_1 \cdots \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$, and for $\boldsymbol{x} \in \mathbb{R}^d$, let $E(\boldsymbol{x}) := \big\{\boldsymbol{\alpha} \in \mathbb{R}^n : \boldsymbol{X}^\top\boldsymbol{\alpha} = \boldsymbol{x}, \mathbf{1}_n^\top\boldsymbol{\alpha} = 1, \boldsymbol{\alpha} \ge 0\big\}$ denote the set of all weight vectors for which $\boldsymbol{x}$ can be written as a weighted convex combination of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Thus $E(\boldsymbol{x})$ is a compact, convex subset of $\mathbb{R}^n$. The cef function is given by a linear program (LP) [2, 46]:

$$\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x}) = \inf_{\boldsymbol{\alpha}\in E(\boldsymbol{x})} \boldsymbol{\alpha}^\top\boldsymbol{\phi}. \tag{Q_0}$$

If $\boldsymbol{x} \notin C_n$, then $E(\boldsymbol{x}) = \emptyset$, and, with the standard convention that $\inf\emptyset := \infty$, we see that (Q_0) agrees with (3). From the LP formulation, it follows that $\boldsymbol{\phi} \mapsto \mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x})$ is concave, for every $\boldsymbol{x} \in \mathbb{R}^d$.

Given a pair $\boldsymbol{\phi} \in \mathbb{R}^n$ and $\boldsymbol{x} \in C_n$, an optimal solution to (Q_0) may not be unique, in which case the map $\boldsymbol{\phi} \mapsto \mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x})$ is not differentiable [7, Proposition B.25(b)]. Noting that the infimum in (Q_0) is attained whenever $\boldsymbol{x} \in C_n$, let

$$\begin{aligned} A[\boldsymbol{\phi}](\boldsymbol{x}) &:= \mathrm{conv}\big(\big\{\boldsymbol{\alpha} \in E(\boldsymbol{x}) : \boldsymbol{\alpha}^\top\boldsymbol{\phi} = \mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x})\big\}\big) \\ &= \big\{\boldsymbol{\alpha} \in E(\boldsymbol{x}) : \boldsymbol{\alpha}^\top\boldsymbol{\phi} = \mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x})\big\}. \end{aligned}$$

Danskin's theorem [7, Proposition B.25(b)] applied to $-\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x})$ then yields that for each $\boldsymbol{x} \in C_n$, the subdifferential of $F(\boldsymbol{\phi}, \boldsymbol{x})$ with respect to $\boldsymbol{\phi}$ is given by

$$\partial F(\boldsymbol{\phi}, \boldsymbol{x}) := \left\{\frac{1}{n}\mathbf{1} - \Delta e^{-\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x})}\boldsymbol{\alpha} : \boldsymbol{\alpha} \in A[\boldsymbol{\phi}](\boldsymbol{x})\right\}. \tag{8}$$

Since both f and $F(\cdot, \boldsymbol{x})$ are finite convex functions on $\mathbb{R}^n$ (for each fixed $\boldsymbol{x} \in C_n$ in the latter case), by [17, Proposition 2.3.6(b) and Theorem 2.7.2], the subdifferential of f at $\boldsymbol{\phi} \in \mathbb{R}^n$ is given by

$$\partial f(\boldsymbol{\phi}) := \big\{\mathbb{E}\boldsymbol{G}(\boldsymbol{\phi}, \boldsymbol{\xi}) : \boldsymbol{G}(\boldsymbol{\phi}, \boldsymbol{x}) \in \partial F(\boldsymbol{\phi}, \boldsymbol{x}) \text{ for each } \boldsymbol{x} \in C_n\big\}. \tag{9}$$

Observe that given any $\boldsymbol{\phi} \in \mathbb{R}^n$, the function $\boldsymbol{x} \mapsto -\mathrm{cef}\big[\boldsymbol{\phi} + \log I(\boldsymbol{\phi})\mathbf{1}\big](\boldsymbol{x})$ (where $I(\boldsymbol{\phi})$ is the integral defined in (5)) is a log-density. It is also convenient to let $\bar{\boldsymbol{\phi}} \in \mathbb{R}^n$ be such that $\exp\{-\mathrm{cef}[\bar{\boldsymbol{\phi}}]\}$ is the uniform density on C_n, so that $f(\bar{\boldsymbol{\phi}}) = \log \Delta + 1$. Proposition 1 below (an extension of [2, Lemma 2]) provides uniform upper and lower bounds on this log-density, whenever the objective function f evaluated at $\boldsymbol{\phi}$ is at least as good as that at $\bar{\boldsymbol{\phi}}$. In more statistical language, these bounds hold whenever the log-likelihood of the density $\exp\big\{-\mathrm{cef}\big[\boldsymbol{\phi} + \log I(\boldsymbol{\phi})\mathbf{1}\big](\cdot)\big\}$ is at least as large as that of the uniform density on the convex hull of the data, so in particular, they must hold for the log-concave MLE (i.e. when $\boldsymbol{\phi} = \boldsymbol{\phi}^*$). Let $\phi^0 := (n-1) + d(n-1)\log\big(2n + 2nd\log(2nd)\big) + \log \Delta$ and $\phi_0 := -1 - d\log\big(2n + 2nd\log(2nd)\big) + \log \Delta$.

Proposition 1 *For any $\boldsymbol{\phi} \in \mathbb{R}^n$ such that $f(\boldsymbol{\phi}) \le \log \Delta + 1$, we have $\phi_0 \le \phi_i + \log I(\boldsymbol{\phi}) \le \phi^0$ for all $i \in [n]$.*

The following corollary is an immediate consequence of Proposition 1.

Corollary 1 *Suppose that $\boldsymbol{\phi} \in \mathbb{R}^n$ satisfies $I(\boldsymbol{\phi}) = 1$ and $f(\boldsymbol{\phi}) \le f(\bar{\boldsymbol{\phi}}) = \log \Delta + 1$. Then $\boldsymbol{\phi}^* \in \mathbb{R}^n$ defined in (4) satisfies*

$$\|\boldsymbol{\phi} - \boldsymbol{\phi}^*\| \le \sqrt{n}(\phi^0 - \phi_0).$$

Corollary 1 gives a sense in which any $\boldsymbol{\phi} \in \mathbb{R}^n$ for which the objective function is 'good' cannot be too far from the optimizer $\boldsymbol{\phi}^*$; here, 'good' means that the objective should be no larger than that of the uniform density on the convex hull of the data. Moreover, an upper bound on the integral $I(\boldsymbol{\phi})$ provides an upper bound on the norm of any subgradient $\boldsymbol{g}(\boldsymbol{\phi})$ of f at $\boldsymbol{\phi}$.

Proposition 2 *Any subgradient $\boldsymbol{g}(\boldsymbol{\phi}) \in \mathbb{R}^n$ of f at $\boldsymbol{\phi} \in \mathbb{R}^n$ satisfies $\|\boldsymbol{g}(\boldsymbol{\phi})\|^2 \le \max\big\{1/n + 1/4, I(\boldsymbol{\phi})^2\big\}$.*

3 Computing the log-concave MLE

As mentioned in the introduction, subgradient methods [56, 66] tend to be slow for minimizing the objective function f defined in (5) [20]. Our alternative approach involves the minimizing the representation of f given in (6) via smoothing techniques, which offer superior computational guarantees and practical performance in our numerical experiments.

3.1 Smoothing techniques

We present two smoothing techniques to find the minimizer $\boldsymbol{\phi}^* \in \mathbb{R}^n$ of the nonsmooth convex optimization problem (4). By Proposition 1, we have that $\boldsymbol{\phi}^* \in \boldsymbol{\Phi}$, where

$$\boldsymbol{\Phi} := \{\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n) \in \mathbb{R}^n : \phi_0 \leq \phi_i \leq \phi^0 \text{ for } i \in [n]\}, \tag{10}$$

with $\phi_0, \phi^0 \in \mathbb{R}$. In what follows we present two smoothing techniques: one based on Nesterov smoothing [52] and the second on randomized smoothing [25].

3.1.1 Nesterov smoothing

Recall that the non-differentiability in f in (5) is due to the LP (Q_0) potentially having multiple optimal solutions. Therefore, following [52], we consider replacing this LP with the following quadratic program (QP):

$$q_u[\boldsymbol{\phi}](\boldsymbol{x}) := \inf_{\boldsymbol{\alpha} \in E(\boldsymbol{x})} \left(\boldsymbol{\alpha}^\top \boldsymbol{\phi} + \frac{u}{2} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|^2 - \frac{u}{2} \right), \tag{Q_u}$$

where $\boldsymbol{\alpha}_0 := (1/n)\mathbf{1} \in \mathbb{R}^n$ is the center of $E(\boldsymbol{x})$, and where $u \geq 0$ is a regularization parameter that controls the extent of the quadratic regularization of the objective. With this definition, we have $q_0[\boldsymbol{\phi}](\boldsymbol{x}) = \text{cef}[\boldsymbol{\phi}](\boldsymbol{x})$. For $u > 0$, due to the strong convexity of the function $\boldsymbol{\alpha} \mapsto \boldsymbol{\alpha}^\top \boldsymbol{\phi} + (u/2)\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|^2$ on the convex polytope $E(\boldsymbol{x})$, (Q_u) admits a unique solution that we denote by $\boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x})$. It follows again from Danskin's theorem that $\boldsymbol{\phi} \mapsto q_u[\boldsymbol{\phi}](\boldsymbol{x})$ is differentiable for such u, with gradient $\nabla_{\boldsymbol{\phi}} q_u[\boldsymbol{\phi}](\boldsymbol{x}) = \boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x})$.

Using $q_u[\boldsymbol{\phi}](\boldsymbol{x})$ instead of $q_0[\boldsymbol{\phi}](\boldsymbol{x})$ in (5), we obtain a smooth objective $\boldsymbol{\phi} \mapsto \tilde{f}_u(\boldsymbol{\phi})$, given by

$$\tilde{f}_u(\boldsymbol{\phi}) := \frac{1}{n}\mathbf{1}^\top \boldsymbol{\phi} + \int_{C_n} \exp\{-q_u[\boldsymbol{\phi}](\boldsymbol{x})\}\, d\boldsymbol{x} = \mathbb{E}\tilde{F}_u(\boldsymbol{\phi}, \boldsymbol{\xi}), \tag{11}$$

where $\tilde{F}_u(\boldsymbol{\phi}, \boldsymbol{x}) := (1/n)\mathbf{1}^\top \boldsymbol{\phi} + \Delta \exp\{-q_u[\boldsymbol{\phi}](\boldsymbol{x})\}$, and where $\boldsymbol{\xi}$ is again uniformly distributed on C_n. We may differentiate under the integral (e.g. [45, Theorem 6.28]) to see that the partial derivatives of $\tilde{f}_u$ with respect to each component of $\boldsymbol{\phi}$ exist, and moreover they are continuous (because $\boldsymbol{\phi} \mapsto \boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x})$ is continuous by Proposition 5), so $\nabla_{\boldsymbol{\phi}} \tilde{f}_u(\boldsymbol{\phi}) = \mathbb{E}[\tilde{\boldsymbol{G}}_u(\boldsymbol{\phi}, \boldsymbol{\xi})]$, where

$$\tilde{\boldsymbol{G}}_u(\boldsymbol{\phi}, \boldsymbol{x}) := \nabla_{\boldsymbol{\phi}} \tilde{F}_u(\boldsymbol{\phi}, \boldsymbol{x}) = \frac{1}{n}\mathbf{1} - \Delta e^{-q_u[\boldsymbol{\phi}](\boldsymbol{x})} \boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x}). \tag{12}$$

Proposition 3 below presents some properties of the smooth objective $\tilde{f}_u$.

Proposition 3 *For any $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, we have (a) $0 \leq \tilde{f}_u(\boldsymbol{\phi}) - \tilde{f}_{u'}(\boldsymbol{\phi}) \leq \frac{u-u'}{2} e^{u'/2} I(\boldsymbol{\phi})$ for $u' \in [0, u]$;*
(b) For every $u \geq 0$, the function $\boldsymbol{\phi} \mapsto \tilde{f}_u(\boldsymbol{\phi})$ is convex and $\Delta e^{-\phi_0 + u/2}$-Lipschitz;

(c) For every $u \geq 0$, the function $\boldsymbol{\phi} \mapsto \tilde{f}_u(\boldsymbol{\phi})$ has $\Delta e^{-\phi_0+u/2}(1+u^{-1})$-Lipschitz gradient;
(d) $\mathbb{E}\big(\|\tilde{\boldsymbol{G}}_u(\boldsymbol{\phi},\boldsymbol{\xi}) - \nabla_{\boldsymbol{\phi}} \tilde{f}_u(\boldsymbol{\phi})\|^2\big) \leq (\Delta e^{-\phi_0+u/2})^2$ for every $u \geq 0$.

3.1.2 Randomized smoothing

Our second smoothing technique is randomized smoothing [25, 48, 73]: we take the expectation of a random perturbation of the argument of f. Specifically, for $u \geq 0$, let

$$\bar{f}_u(\boldsymbol{\phi}) := \mathbb{E} f(\boldsymbol{\phi} + u\boldsymbol{z}), \tag{13}$$

where $\boldsymbol{z}$ is uniformly distributed on the unit ℓ_2-ball in $\mathbb{R}^n$. Thus, similar to Nesterov smoothing, $\bar{f}_0 = f$, and the amount of smoothing increases with u. From a stochastic optimization viewpoint, we can write

$$\bar{f}_u(\boldsymbol{\phi}) = \mathbb{E} F(\boldsymbol{\phi} + u\boldsymbol{z}, \boldsymbol{\xi}) \quad \text{and} \quad \nabla_{\boldsymbol{\phi}} \bar{f}_u(\boldsymbol{\phi}) = \mathbb{E}\boldsymbol{G}(\boldsymbol{\phi} + u\boldsymbol{z}, \boldsymbol{\xi})$$

where $\boldsymbol{G}(\boldsymbol{\phi} + u\boldsymbol{v}, \boldsymbol{x}) \in \partial F(\boldsymbol{\phi} + u\boldsymbol{v}, \boldsymbol{x})$, and where the expectations are taken over independent random vectors $\boldsymbol{z}$, distributed uniformly on the unit Euclidean ball in $\mathbb{R}^n$, and $\boldsymbol{\xi}$, distributed uniformly on C_n. Here the gradient expression follows from, e.g., [48, Lemma 3.3(a)], [73, Lemma 7]; since $F(\boldsymbol{\phi} + u\boldsymbol{v}, \boldsymbol{x})$ is differentiable almost everywhere with respect to $\boldsymbol{\phi}$, the expression for $\bar{f}_u(\boldsymbol{\phi})$ does not depend on the choice of subgradient.

Proposition 4 below lists some properties of $\bar{f}_u$ and its gradient. It extends [73, Lemmas 7 and 8] by exploiting special properties of the objective function to sharpen the dependence of the bounds on n.

Proposition 4 *For any $u \geq 0$ and $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, we have*
(a) $0 \leq \bar{f}_u(\boldsymbol{\phi}) - f(\boldsymbol{\phi}) \leq I(\boldsymbol{\phi}) u e^u \sqrt{\frac{2\log n}{n+1}}$;
(b) $\bar{f}_{u'}(\boldsymbol{\phi}) \leq \bar{f}_u(\boldsymbol{\phi})$ for any $u' \in [0, u]$;
(c) $\boldsymbol{\phi} \mapsto \bar{f}_u(\boldsymbol{\phi})$ is convex and $\Delta e^{-\phi_0+u}$-Lipschitz;
(d) $\boldsymbol{\phi} \mapsto \bar{f}_u(\boldsymbol{\phi})$ has $\Delta e^{-\phi_0+u} n^{1/2}/u$-Lipschitz gradient;
(e) $\mathbb{E}\big(\big\|\boldsymbol{G}(\boldsymbol{\phi}+u\boldsymbol{z},\boldsymbol{\xi}) - \nabla \bar{f}_u(\boldsymbol{\phi})\big\|^2\big) \leq (\Delta e^{-\phi_0+u})^2$ whenever $\boldsymbol{G}(\boldsymbol{\phi}+u\boldsymbol{v},\boldsymbol{x}) \in \partial F(\boldsymbol{\phi}+u\boldsymbol{v},\boldsymbol{x})$ for every $\boldsymbol{v} \in \mathbb{R}^n$ with $\|\boldsymbol{v}\| \leq 1$ and $\boldsymbol{x} \in C_n$.

3.2 Stochastic first-order methods for smoothing sequences

Our proposed algorithm for computing the log-concave MLE is given in Algorithm 1. It relies on the choice of a smoothing sequence of f, which may be constructed using Nesterov or randomized smoothing, for instance. For a non-negative sequence $(u_t)_{t\in\mathbb{N}_0}$, this smoothing sequence is denoted by $(\ell_{u_t})_{t\in\mathbb{N}_0}$, where $\ell_{u_t} := \tilde{f}_{u_t}$ is given by (11) or $\ell_{u_t} := \bar{f}_{u_t}$ is given by (13). In Algorithm 1, $P_{\boldsymbol{\Phi}} : \mathbb{R}^n \to \boldsymbol{\Phi}$ denotes the projection operator onto the closed convex set $\boldsymbol{\Phi}$, which is essentially a threshold clipping operator. In fact, Algorithm 1 is a modification of an algorithm due to [25], and

can be regarded as an accelerated version of the dual averaging scheme [53] applied to (ℓ_{u_t}).

Algorithm 1 Accelerated stochastic dual averaging on a smoothing sequence with increasing grids

Input: Smoothing sequence (ℓ_{u_t}) whose gradients have Lipschitz constants $(L_t)_{t\in\mathbb{N}_0}$; initialization $\boldsymbol{\phi}_0 \in \mathbb{R}^n$; learning rate sequence $(\eta_t)_{t\in\mathbb{N}}$ of positive real numbers; number of iterations $T \in \mathbb{N}$

1: $\boldsymbol{\phi}_0^{(x)} = \boldsymbol{\phi}_0^{(y)} = \boldsymbol{\phi}_0^{(z)} = \boldsymbol{\phi}_0$, $\boldsymbol{s}_t = \mathbf{0} \in \mathbb{R}^n$, $\theta_0 = 1$
2: **for** $t = 0, \ldots, T-1$ **do**
3: Compute an approximation $\boldsymbol{g}_t$ of $\nabla_{\boldsymbol{\phi}} \ell_{u_t}\big(\boldsymbol{\phi}_t^{(y)}\big)$; see Sect. 3.2.1
4: $\boldsymbol{s}_{t+1} = \boldsymbol{s}_t + \boldsymbol{g}_t/\theta_t$
5: $\theta_{t+1} = \frac{2}{1+\sqrt{1+4/\theta_t^2}}$
6: $\boldsymbol{\phi}_{t+1}^{(z)} = P_{\boldsymbol{\Phi}}\big(\boldsymbol{\phi}_0 - \frac{\boldsymbol{s}_t}{L_{t+1}+\eta_{t+1}/\theta_{t+1}}\big)$
7: $\boldsymbol{\phi}_{t+1}^{(x)} = (1-\theta_t)\boldsymbol{\phi}_t^{(x)} + \theta_t \boldsymbol{\phi}_{t+1}^{(z)}$
8: $\boldsymbol{\phi}_{t+1}^{(y)} = (1-\theta_{t+1})\boldsymbol{\phi}_{t+1}^{(x)} + \theta_{t+1}\boldsymbol{\phi}_{t+1}^{(z)}$
9: **end for**
10: **return** $\boldsymbol{\phi}_{t+1}^{(x)}$

3.2.1 Approximating the gradient of the smoothing sequence

In Line 3 of Algorithm 1, we need to compute an approximation of the gradient $\nabla_{\boldsymbol{\phi}} \ell_u$, for a general $u \geq 0$. A key step in this process is to approximate the integral $I(\cdot)$, as well as a subgradient of I, at an arbitrary $\boldsymbol{\phi} \in \mathbb{R}^n$. [20] provide explicit formulae for these quantities, based on a triangulation of C_n, using tools from computational geometry. For practical purposes, [21] apply a Taylor expansion to approximate the analytic expression. The R package `LogConcDEAD` [18] uses this method to evaluate the exact integral at each iteration, but since this is time-consuming, we will only use this method at the final stage of our proposed algorithm as a polishing step[1].

An alternative approach is to use numerical integration.[2] Among deterministic schemes, [57] observed empirically that the simple Riemann sum with uniform weights appears to perform the best among several multi-dimensional integration techniques. Random (Monte Carlo) approaches to approximate the integral are also possible: given a collection of grid points $\mathcal{S} = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m\}$, we approximate the integral as $I_{\mathcal{S}}(\boldsymbol{\phi}) := (\Delta/m)\sum_{\ell=1}^m \exp\{-\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{\xi}_\ell)\}$. This leads to an approximation of the objective f given by

$$f(\boldsymbol{\phi}) \approx \frac{1}{n}\mathbf{1}^\top \boldsymbol{\phi} + I_{\mathcal{S}}(\boldsymbol{\phi}) =: f_{\mathcal{S}}(\boldsymbol{\phi}). \tag{14}$$

[1] Once our algorithm terminates at $\tilde{\boldsymbol{\phi}}_T$, say, we evaluate the integral $I(\tilde{\boldsymbol{\phi}}_T)$ in the same way as [20]. Our final output, then is $\boldsymbol{\phi}_T := \tilde{\boldsymbol{\phi}}_T + \log I(\tilde{\boldsymbol{\phi}}_T)\mathbf{1}$; this final step not only improves the objective function, but also guarantees that $\exp[-\mathrm{cef}[\boldsymbol{\phi}_T](\cdot)]$ is a log-concave density.

[2] Yet another option involves sampling from a log-concave density [20, 22]. [2] discuss interesting polynomial-time sampling methods to approximate $I(\cdot)$, but as noted by [57], these methods may not be practically efficient, and we do not pursue them here.

Since $f_{\mathcal{S}}$ is a finite, convex function on $\mathbb{R}^n$, it has a subgradient at each $\boldsymbol{\phi} \in \mathbb{R}^n$, given by

$$\boldsymbol{g}_{\mathcal{S}}(\boldsymbol{\phi}) := \frac{1}{m} \sum_{\ell=1}^{m} \boldsymbol{G}(\boldsymbol{\phi}, \boldsymbol{\xi}_\ell).$$

As the effective domain of $\mathrm{cef}[\boldsymbol{\phi}](\cdot)$ is C_n, we consider grid points $\mathcal{S} \subseteq C_n$.

We now illustrate how these ideas allow us to approximate the gradient of the smoothing sequence, and initially consider Nesterov smoothing, with $\ell_u = \tilde{f}_u$. If $\mathcal{S} = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m\} \subseteq C_n$ denotes a collection of grid points (either deterministic or Monte Carlo based), then $\nabla_{\boldsymbol{\phi}} \ell_u$ can be approximated by $\tilde{\boldsymbol{g}}_{u,\mathcal{S}}$, where

$$\tilde{\boldsymbol{g}}_{u,\mathcal{S}}(\boldsymbol{\phi}) := \frac{1}{n}\mathbf{1} - \frac{\Delta}{m} \sum_{j=1}^{m} e^{-q_u[\boldsymbol{\phi}](\boldsymbol{\xi}_j)} \alpha_u^*[\boldsymbol{\phi}](\boldsymbol{\xi}_j). \tag{15}$$

In fact, we distinguish the cases of deterministic and random $\mathcal{S}$ by writing this approximation as $\tilde{\boldsymbol{g}}^{\mathrm{D}}_{u,\mathcal{S}}$ and $\tilde{\boldsymbol{g}}^{\mathrm{R}}_{u,\mathcal{S}}$ respectively.

For the randomized smoothing method with $\ell_u = \bar{f}_u$, the approximation is slightly more involved. Given m grid points $\mathcal{S} = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m\} \subseteq C_n$ (again either deterministic or random), and independent random vectors $z_1, \ldots, z_m$, each uniformly distributed on the unit Euclidean ball in $\mathbb{R}^n$, we can approximate $\nabla_{\boldsymbol{\phi}} \ell_u(\boldsymbol{\phi})$ by

$$\bar{\boldsymbol{g}}^{\circ}_{u,\mathcal{S}}(\boldsymbol{\phi}) = \frac{1}{n}\mathbf{1} - \frac{\Delta}{m} \sum_{j=1}^{m} e^{-\mathrm{cef}[\boldsymbol{\phi}+uz_j](\boldsymbol{\xi}_j)} \alpha^*[\boldsymbol{\phi} + uz_j](\boldsymbol{\xi}_j), \tag{16}$$

with $\circ \in \{\mathrm{D}, \mathrm{R}\}$ again distinguishing the cases of deterministic and random $\mathcal{S}$.

3.2.2 Related literature

As mentioned above, Algorithm 1 is an accelerated version of the dual averaging method of [53], which to the best of our knowledge has not been studied in the context of log-concave density estimation previously. Nevertheless, related ideas have been considered for other optimization problems (e.g. [25, 70]). Relative to previous work, our approach is quite general, in that it applies to both of the smoothing techniques discussed in Sect. 3.1, and allows the use of both deterministic and random grids to approximate the gradients of the smoothing sequence. Another key difference with earlier work is that we allow the grid $\mathcal{S}$ to depend on t, so we write it as $\mathcal{S}_t$, with $m_t := |\mathcal{S}_t|$; in particular, inspired by both our theoretical results and numerical experience, we take (m_t) to be a suitable increasing sequence.

Table 1 Summary of options for smoothing and gradient approximation methods

Options	Smoothing	Approximation	Options	Smoothing	Approximation
1	$\tilde{f}_u$	$\tilde{\boldsymbol{g}}^{\mathrm{D}}_{u,\mathcal{S}}$	3	$\bar{f}_u$	$\bar{\boldsymbol{g}}^{\mathrm{D}}_{u,\mathcal{S}}$
2	$\tilde{f}_u$	$\tilde{\boldsymbol{g}}^{\mathrm{R}}_{u,\mathcal{S}}$	4	$\bar{f}_u$	$\bar{\boldsymbol{g}}^{\mathrm{R}}_{u,\mathcal{S}}$

4 Theoretical analysis of optimization error of Algorithm 1

We have seen in Propositions 3 and 4 that the two smooth functions $\tilde{f}_u$ and $\bar{f}_u$ enjoy similar properties — according to Proposition 3(a) to (c) and Proposition 4(a) to (d), both $\tilde{f}_u$ and $\bar{f}_u$ satisfy the following assumption:

Assumption 1 [Assumptions on smoothing sequence] There exists $r \geq 0$ such that for any $\boldsymbol{\phi} \in \boldsymbol{\Phi}$,
(a) we can find $B_0 > 0$ with $f(\boldsymbol{\phi}) \leq \ell_u(\boldsymbol{\phi}) \leq f(\boldsymbol{\phi}) + B_0 I(\boldsymbol{\phi})u$ for all $u \in [0, r]$;
(b) $\ell_{u'}(\boldsymbol{\phi}) \leq \ell_u(\boldsymbol{\phi})$ for all $u' \in [0, u]$;
(c) for each $u \in [0, r]$, the function $\boldsymbol{\phi} \mapsto \ell_u(\boldsymbol{\phi})$ is convex and has B_1/u-Lipschitz gradient, for some $B_1 > 0$.

Recall from Sect. 3 that we have four possible choices corresponding to a combination of the smoothing and integral approximation methods, as summarized in Table 1.

Once we select an option, in line 3 of Algorithm 1, we can take

$$\boldsymbol{g}_t = \check{\boldsymbol{g}}^{\circ}_{u_t,\mathcal{S}_t}(\boldsymbol{\phi}_t^{(y)}),$$

where $\check{} \in \{\tilde{}, \bar{}\}$ and $\circ \in \{\mathrm{D}, \mathrm{R}\}$. To encompass all four approximation choices in Line 3 of Algorithm 1, we make the following assumption on the gradient approximation error $\boldsymbol{e}_t := \boldsymbol{g}_t - \nabla_{\boldsymbol{\phi}} \ell_{u_t}(\boldsymbol{\phi}_t^{(y)})$:

Assumption 2 [Gradient approximation error] There exists $\sigma > 0$ such that

$$\mathbb{E}\big(\|\boldsymbol{e}_t\|^2 | \mathcal{F}_{t-1}\big) \leq \sigma^2/m_t \quad \text{for all } t \in \mathbb{N}_0, \tag{17}$$

where $\mathcal{F}_{t-1}$ denotes the σ-algebra generated by all random sources up to iteration $t-1$ (with $\mathcal{F}_{-1}$ denoting the trivial σ-algebra).

When $\mathcal{S}$ is a Monte Carlo random grid (options 2 and 4), the approximate gradient $\boldsymbol{g}_t$ is an average of m_t independent and identically distributed random vectors, each being an unbiased estimator of $\nabla \ell_{u_t}(\boldsymbol{\phi}_t^{(y)})$. Hence, (17) holds true with σ^2 determined by the bounds in Proposition 3(d) (option 2) and Proposition 4(e) (option 4). For a deterministic Riemann grid $\mathcal{S}$ and Nesterov's smoothing technique (option 1), $\boldsymbol{e}_t$ is deterministic, and arises from using $\tilde{\boldsymbol{g}}_{u,\mathcal{S}}(\boldsymbol{\phi})$ in (15) to approximate $\nabla_{\boldsymbol{\phi}} \tilde{f}_u(\boldsymbol{\phi}) = \mathbb{E}[\tilde{\boldsymbol{G}}_u(\boldsymbol{\phi}, \boldsymbol{\xi})]$. For the deterministic Riemann grid and randomized smoothing (option 3), the error $\boldsymbol{e}_t$ can be decomposed into a random estimation error term (induced by

$z_1, \ldots, z_{m_t}$) and a deterministic approximation error term (induced by $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{m_t}$) as follows:

$$\boldsymbol{e}_t = \frac{1}{m_t}\sum_{j=1}^{m_t}\big(\boldsymbol{G}(\boldsymbol{\phi}_t^{(y)} + u_t z_j, \boldsymbol{\xi}_j) - \mathbb{E}[\boldsymbol{G}(\boldsymbol{\phi}_t^{(y)} + u_t z, \boldsymbol{\xi}_j)|\mathcal{F}_{t-1}]\big) + \left(\frac{1}{m_t}\sum_{j=1}^{m_t}\mathbb{E}[\boldsymbol{G}(\boldsymbol{\phi}_t^{(y)} + u_t z, \boldsymbol{\xi}_j)|\mathcal{F}_{t-1}] - \mathbb{E}[\boldsymbol{G}(\boldsymbol{\phi}_t^{(y)} + u_t z, \boldsymbol{\xi})|\mathcal{F}_{t-1}]\right).$$

It can be shown using this decomposition that $\mathbb{E}(\|\boldsymbol{e}_t\|^2|\mathcal{F}_{t-1}) = O(1/m_t)$ under regularity conditions.

Theorem 1 below establishes our desired computational guarantees for Algorithm 1. We write $D := \sup_{\boldsymbol{\phi},\tilde{\boldsymbol{\phi}}\in\boldsymbol{\Phi}} \|\boldsymbol{\phi} - \tilde{\boldsymbol{\phi}}\|$ for the diameter of $\boldsymbol{\Phi}$.

Theorem 1 *Suppose that Assumptions 1 and 2 hold, and define the sequence* $(\theta_t)_{t\in\mathbb{N}_0}$ *by* $\theta_0 := 1$ *and* $\theta_{t+1} := 2\big(1 + \sqrt{1 + 4/\theta_t^2}\big)^{-1}$ *for* $t \in \mathbb{N}_0$. *Let* $u > 0$, *let* $u_t := \theta_t u$ *and take* $L_t = B_1/u_t$ *and* $\eta_t = \eta$ *for all* $t \in \mathbb{N}_0$ *as input parameters to Algorithm 1. Writing* $M_T^{(1)} := \sqrt{\sum_{t=0}^{T-1} m_t^{-1}}$ *and* $M_T^{(1/2)} := \sum_{t=0}^{T-1} m_t^{-1/2}$, *we have for any* $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ *that*

$$\mathbb{E}[f(\boldsymbol{\phi}_T^{(x)})] - f(\boldsymbol{\phi}) \le \frac{B_1 D^2}{Tu} + \frac{4B_0 I(\boldsymbol{\phi})u}{T} + \frac{\eta D^2}{T} + \frac{\sigma^2 (M_T^{(1)})^2}{T\eta} + \frac{2D\sigma M_T^{(1/2)}}{T}. \tag{18}$$

In particular, taking $\boldsymbol{\phi} = \boldsymbol{\phi}^*$, *and choosing* $u = (D/2)\sqrt{B_1/B_0}$ *and* $\eta = (\sigma M_T^{(1)})/D$, *we obtain*

$$\varepsilon_T := \mathbb{E}[f(\boldsymbol{\phi}_T^{(x)})] - f(\boldsymbol{\phi}^*) \le \frac{4\sqrt{B_0 B_1}D}{T} + \frac{2\sigma D M_T^{(1)}}{T} + \frac{2D\sigma M_T^{(1/2)}}{T}. \tag{19}$$

Moreover, if we further assume that $\mathbb{E}(\boldsymbol{e}_t|\mathcal{F}_{t-1}) = \boldsymbol{0}$ *(e.g. by using options 2 and 4), then we can remove the last term of both inequalities above.*

For related general results that control the expected optimization error for smoothed objective functions, see, e.g., [52], [67], [70], [25]. With deterministic grids (corresponding to options 1 and 3), if we take $|\mathcal{S}_t| = m$ for all t, then $M_T^{(1/2)} = T/\sqrt{m}$, and the upper bound in (19) does not converge to zero as $T \to \infty$. On the other hand, if we take $|\mathcal{S}_t| = t^2$, for example, then $\sup_{T\in\mathbb{N}} M_T^{(1)} < \infty$ and $M_T^{(1/2)} = \tilde{O}(1)$, and we find that $\varepsilon_T = \tilde{O}(1/T)$. For random grids (options 2 and 4), if we take $|\mathcal{S}_t| = m$ for all t, then $M_T^{(1)} = \sqrt{T/m}$ and we recover the $\varepsilon_T = O(1/\sqrt{T})$ rate for stochastic subgradient methods [56]. This can be improved to $\varepsilon_T = \tilde{O}(1/T)$ with $m_t = t$, or even $\varepsilon_T = O(1/T)$ if we choose $(m_t)_t$ such that $\sum_{t=0}^{\infty} m_t^{-1} < \infty$.

A direct application of the theory of [25] would yield an error rate of $\varepsilon_T = O(n^{1/4}/T + 1/\sqrt{T})$. On the other hand, Theorem 1 shows that, owing to the increasing sequence of grid sizes used to approximate the gradients in Step 3 of Algorithm 1,

Table 2 Comparion of constants in Assumption 1 for different smoothing schemes with $u \in [0, r]$

	B_0	B_1	σ ($\eta \propto \sigma$)	u ($\propto \sqrt{B_1/B_0}$)	$\sqrt{B_0 B_1}$
Nesterov	$1/2$	$\Delta e^{-\phi_0+r/2}(r+1)$	$\Delta e^{-\phi_0+r/2}$	$C_1 e^{r/4}\sqrt{(r+1)}O(1)$	
Randomized	$\sqrt{2n^{-1}\log n}e^r$	$\Delta e^{-\phi_0+r}\sqrt{n}$	$\Delta e^{-\phi_0+r}$	$C_1\tilde{O}(\sqrt{n})$	$C_1 e^r \tilde{O}(1)$

Here, σ corresponds to random grid points (options 2 and 4), the optimal η is taken to be proportional to σ, the optimal u is proportional to $\sqrt{B_1/B_0}$, we take $C_1 = \sqrt{\Delta e^{-\phi_0}}$; $\sqrt{B_0 B_1}$ determines the first term in the error rate

we can improve this rate to $\tilde{O}(1/T)$. Note however, that this improvement is in terms of the number of iterations T, and not the total number of stochastic oracle queries (equivalently, the total number of LPs (Q_0)), which is given by $T_{\text{query}} := \sum_{t=0}^{T-1} m_t$. [1] and [51] have shown that the optimal expected number of stochastic oracle queries is of order $1/\sqrt{T_{\text{query}}}$, which is attained by the algorithm of [25]. For our framework, by taking $m_t = t$, we have $T_{\text{query}} = \sum_{t=0}^{T-1} m_t = \tilde{O}(T^2)$, so after T_{query} stochastic oracle queries, our algorithm also attains the optimal error on the objective function scale, up to a logarithmic factor. Other advantages of our algorithm and the theoretical guarantees provided by Theorem 1 relative to the earlier contributions of [25] are that we do not require an upper bound on $I(\boldsymbol{\phi})$ and are able to provide a unified framework that includes Nesterov smoothing and an alternative gradient approximation approach by numerical integration in addition to randomized smoothing scheme with stochastic gradients. Moreover, we can exploit the specific structure of the log-concave density estimation problem to provide much better Lipschitz constants for the randomized smoothing sequence than would be obtained using the generic constants of [25]. For example, our upper bound in Proposition 4(a) is of order $O(n^{-1/2}\log^{1/2} n)$, whereas a naive application of the general theory of [25] would only yield a bound of $O(1)$. A further improvement in our bound comes from the fact that it now involves $I(\boldsymbol{\phi})$ directly, as opposed to an upper bound on this quantity.

In Theorem 1, the computational guarantee depends upon B_0, B_1, σ in Assumptions 1 and 2. In light of Propositions 3 and 4, Table 2 illustrates how these quantities, and hence the corresponding guarantees, differ according to whether we use Nesterov smoothing or randomized smoothing.

The randomized smoothing procedure requires solving LPs, whereas Nesterov's smoothing technique requires solving QPs. While both of these problems are highly structured and can be solved efficiently by off-the-shelf solvers (e.g., [36]), we found the LP solution times to be faster than those for the QP. Additional computational details are discussed in Sect. 6.

Note that Theorem 1 presents error bounds in expectation, though for option 1, since we use Nesterov's smoothing technique and the Riemann sum approximation of the integral, the guarantee in Theorem 1 holds without the need to take an expectation. Theorem 2 below presents corresponding high-probability guarantees. For simplicity, we present results for options 2 and 4, which rely on the following assumption:

Assumption 3 Assume that $\mathbb{E}(\boldsymbol{e}_t|\mathcal{F}_{t-1}) = \mathbf{0}$ and that $\mathbb{E}\big(e^{\|\boldsymbol{e}_t\|^2/\sigma_t^2} \mid \mathcal{F}_{t-1}\big) \le e$, where $\sigma_t = \sigma/\sqrt{m_t}$.

Theorem 2 *Suppose that Assumptions 1 and 3 hold, and define the sequence* $(\theta_t)_{t\in\mathbb{N}_0}$ *by* $\theta_0 := 1$ *and* $\theta_{t+1} := 2\big(1+\sqrt{1+4/\theta_t^2}\big)^{-1}$ *for* $t \in \mathbb{N}_0$. *Let* $u > 0$, *let* $u_t := \theta_t u$ *and take* $L_t = B_1/u_t$ *and* $\eta_t = \eta$ *for all* $t \in \mathbb{N}_0$ *as input parameters to Algorithm 1. Writing* $M_T^{(2)} := \sqrt{\sum_{t=0}^{T-1} m_t^{-2}}$ *and* $M_T^{(1)} := \sqrt{\sum_{t=0}^{T-1} m_t^{-1}}$, *and choosing* $u = (D/2)\sqrt{B_1/B_0}$ *and* $\eta = (\sigma M_T^{(1)})/D$ *as in Theorem 1, for any* $\delta \in (0,1)$, *we have with probability at least* $1-\delta$ *that*

$$f(\boldsymbol{\phi}_T^{(x)}) - f(\boldsymbol{\phi}^*) \le \frac{2\sqrt{B_0 B_1}D}{T} + \frac{\sigma D M_T^{(1)}}{T} + \frac{4\sigma D M_T^{(1)}\sqrt{\log\frac{2}{\delta}}}{T} + \frac{4\sigma D \max\big\{M_T^{(2)}\sqrt{2e\log\frac{2}{\delta}}, m_0^{-1}\log\frac{2}{\delta}\big\}}{M_T^{(1)} T}.$$

For option 3, we would need to consider the approximation error from the Riemann sum, and the final error rate would include additional $O(1/T)$ terms. We omit the details for brevity.

Finally in this section, we relate the error of the objective to the error in terms of $\boldsymbol{\phi}$, as measured through the squared L_2 distance between the corresponding lower convex envelope functions.

Theorem 3 *For any* $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, *we have*

$$\int_{C_n} \big\{\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x}) - \mathrm{cef}[\boldsymbol{\phi}^*](\boldsymbol{x})\big\}^2 d\boldsymbol{x} \le 2e^{\phi^0}\big\{f(\boldsymbol{\phi}) - f(\boldsymbol{\phi}^*)\big\}. \tag{20}$$

5 Beyond log-concave density estimation

In this section, we extend our computational framework beyond the log-concave density family, through the notion of s-concave densities. For $s \in \mathbb{R}$, define domains $\mathcal{D}_s$ and $\psi_s : \mathcal{D}_s \to \mathbb{R}$ by

$$\mathcal{D}_s := \begin{cases} [0,\infty) & \text{if } s < 0 \\ (-\infty,\infty) & \text{if } s = 0 \\ (-\infty, 0] & \text{if } s > 0. \end{cases} \quad \text{and} \quad \psi_s(y) := \begin{cases} y^{1/s} & \text{if } s < 0 \\ e^{-y} & \text{if } s = 0 \\ (-y)^{1/s} & \text{if } s > 0. \end{cases}$$

Definition 1 [s-concave density, [65]] For $s \in \mathbb{R}$, the class $\mathcal{P}_s(\mathbb{R}^d)$ of s-concave density functions on $\mathbb{R}^d$ is given by

$$\mathcal{P}_s(\mathbb{R}^d) := \Big\{p(\cdot) : p = \psi_s \circ \varphi \text{ for some } \varphi \in \mathcal{C}_d \text{ with } \mathrm{Im}(\varphi) \subseteq \mathcal{D}_s \cup \{\infty\}, \int_{\mathbb{R}^d} p = 1\Big\}.$$

For $s = 0$, the family of s-concave densities reduces to the family of log-concave densities. Moreover, for $s_1 < s_2$, we have $\mathcal{P}_{s_2}(\mathbb{R}^d) \subseteq \mathcal{P}_{s_1}(\mathbb{R}^d)$ [23, p. 86]. The s-concave

density family introduces additional modelling flexibility, in particular allowing much heavier tails when $s < 0$ than the log-concave family, but we note that there is no guidance available in the literature on how to choose s.

For the problem of s-concave density estimation, we discuss two estimation methods, both of which have been previously considered in the literature, but for which there has been limited algorithmic development. The first is based on the maximum likelihood principle (Sect. 5.1), while the other is based on minimizing a Rényi divergence (Sect. 5.2).

5.1 Computation of the s-concave maximum likelihood estimator

[65] proved that a maximum likelihood estimator over $\mathcal{P}_s(\mathbb{R}^d)$ exists with probability one for $s \in (-1/d, \infty)$ and $n > \max\left(\frac{dr}{r-d}, d\right)$, where $r := -1/s$, and does not exist if $s < -1/d$. [24] provide some statistical properties of this estimator when $d = 1$. The maximum likelihood estimation problem is to compute

$$\hat{p}_n := \operatorname*{argmax}_{p \in \mathcal{P}_s(\mathbb{R}^d)} \sum_{i=1}^{n} \log p(\boldsymbol{x}_i), \tag{21}$$

or equivalently,

$$\operatorname*{argmax}_{\varphi \in \mathcal{C}_d : \mathrm{Im}(\varphi) \subseteq \mathcal{D}_s \cup \{\infty\}} \frac{1}{n} \sum_{i=1}^{n} \log \psi_s \circ \varphi(\boldsymbol{x}_i) \quad \text{subject to} \quad \int_{\mathbb{R}^d} \psi_s \circ \varphi(\boldsymbol{x})\, d\boldsymbol{x} = 1. \tag{22}$$

We establish the following theorem:

Theorem 4 *Let $s \in [0, 1]$ and suppose that the convex hull C_n of the data is d-dimensional (so that the s-concave MLE $\hat{p}_n$ exists and is unique). Then computing $\hat{p}_n$ in (21) is equivalent to the convex minimization problem of computing*

$$\boldsymbol{\phi}^* := \operatorname*{argmin}_{\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n) \in \mathcal{D}_s^n} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log \psi_s(\phi_i) + \int_{C_n} \psi_s\big(\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x})\big)\, d\boldsymbol{x} \right\}, \tag{23}$$

in the sense that $\hat{p}_n = \psi_s \circ \mathrm{cef}[\boldsymbol{\phi}^*]$.

Remark 1 The equivalence result in Theorem 4 holds for any s (outside $[0, 1]$) as long as the s-concave MLE exists. However, when $s \in [0, 1]$, (23) is a convex optimization problem. The family of s-concave densities with $s < 0$ appears to be more useful from a statistical viewpoint as it allows for heavier tails than log-concave densities, but the MLE cannot be then computed via convex optimization. Nevertheless, the entropy minimization methods discussed in Sect. 5.2 can be used to obtain s-concave density estimates for $s > -1$.

5.2 Quasi-concave density estimation

Another route to estimate an s-concave density (or even a more general class) is via the following problem:

$$\check{\varphi} := \operatorname*{argmin}_{\varphi \in \mathcal{C}_d : \mathrm{dom}(\varphi) = C_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \varphi(\boldsymbol{x}_i) + \int_{C_n} \Psi\big(\varphi(\boldsymbol{x})\big)\, d\boldsymbol{x} \right\}, \tag{24}$$

where $\Psi : \mathbb{R} \to (-\infty, \infty]$ is a decreasing, proper convex function. When $\Psi(y) = e^{-y}$, (24) is equivalent to the MLE for log-concave density estimation (1), by [20, Theorem 1]. This problem, proposed by [46], is called *quasi-concave* density estimation. [46, Theorem 4.1] show that under some assumptions on Ψ, there exists a solution to (24), and if Ψ is strictly convex, then the solution is unique. Furthermore, if Ψ is differentiable on the interior of its domain, then the optimal solution to the dual of (24) is a probability density p such that $p = -\Psi'(\varphi)$, and the dual problem can be regarded as minimizing different distances or entropies (depending on Ψ) between the empirical distribution of the data and p. In particular, when $\beta \geq 1$ and $\Psi(y) = \mathbb{1}_{\{y \leq 0\}}(-y)^{\beta}/\beta$, and when $\beta < 0$ and $\Psi(y) = -y^{\beta}/\beta$ for $y \geq 0$ (with $\Psi(y) = \infty$ otherwise), the dual problem of (24) is essentially minimizing the Rényi divergence and we have the primal-dual relationship $p = |\varphi|^{\beta - 1}$. In fact, this amounts to estimating an s-concave density via Rényi divergence minimization with $\beta = 1 + 1/s$ and $s \in (-1, \infty) \backslash \{0\}$. We therefore consider the problem

$$\min_{\substack{\varphi \in \mathcal{C}_d : \mathrm{dom}(\varphi) = C_n \\ \mathrm{Im}(\varphi) \subseteq \mathcal{D}_s}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \varphi(\boldsymbol{x}_i) + \frac{1}{|1 + 1/s|} \int_{C_n} |\varphi(\boldsymbol{x})|^{1 + 1/s}\, d\boldsymbol{x} \right\}. \tag{25}$$

The proof of Theorem 5 is similar to that of Theorem 4, and is omitted for brevity.

Theorem 5 *Given a decreasing proper convex function Ψ, the quasi-concave density estimation problem (24) is equivalent to the following convex problem:*

$$\boldsymbol{\phi}^* := \operatorname*{argmin}_{\boldsymbol{\phi} \in \mathcal{D}_s^n} \left\{ \frac{1}{n} \mathbf{1}^{\top} \boldsymbol{\phi} + \int_{C_n} \Psi\big(\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x})\big)\, d\boldsymbol{x} \right\}, \tag{26}$$

in the sense that $\check{\varphi} = \mathrm{cef}[\boldsymbol{\phi}^*]$, with corresponding density estimator $\tilde{p}_n = -\Psi' \circ \mathrm{cef}[\boldsymbol{\phi}^*]$.

The objective in (26) is convex, so our computational framework can be applied to solve this problem.

6 Computational experiments on simulated data

In this section, we present numerical experiments to study the different variants of our algorithm and compare them with existing methods based on convex optimization

for the log-concave MLE. Our results are based on large-scale synthetic datasets with $n \in \{5{,}000, 10{,}000\}$ observations generated from standard d-dimensional normal and Laplace distributions with $d = 4$. Code for our experiments is available from the github repository `LogConcComp` available at:

https://github.com/wenyuC94/LogConcComp.

All computations were carried out on the MIT Supercloud Cluster [58] on an Intel Xeon Platinum 8260 machine, with 24 CPUs and 24GB of RAM. Our algorithms were written in Python; we used Gurobi [36] to solve the LPs and QPs.

Our first comparison method is that of [20], implemented in the `R` package `LogConcDEAD` [18], and denoted by CSS. The CSS algorithm terminates when $\|\boldsymbol{\phi}^{(t)} - \boldsymbol{\phi}^{(t-1)}\|_\infty \leq \tau$, and we consider $\tau \in \{10^{-2}, 10^{-3}, 10^{-4}\}$. Our other competing approach is the randomized smoothing method of [25], with random grids of a fixed grid size, which we denote here by RS-RF-m, with m being the grid size. To the best of our knowledge, this method has not been used to compute the log-concave MLE previously.

We denote the different variants of our algorithm as Alg- V, where Alg $\in$ {RS,NS} represents Algorithm 1 with Randomized smoothing and Nesterov smoothing, and $V \in$ {DI, RI} represents whether we use deterministic or random grids of increasing grid sizes to approximate the gradient. Further details of our input parameters are given in Appendix A.3.

Figure 2 presents the relative objective error, defined for an algorithm with iterates $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_t$ as

$$\texttt{relobj}(t) := \left| \frac{\min_{s \in [t]} f(\boldsymbol{\phi}_s) - f(\boldsymbol{\phi}^*)}{f(\boldsymbol{\phi}^*)} \right|, \tag{27}$$

against time (in minutes) and number of iterations. In the definition of the relative objective error in (27) above, $\boldsymbol{\phi}^*$ is taken as the CSS solution with tolerance $\tau = 10^{-4}$. The figure shows that randomized smoothing appears to outperform Nesterov smoothing in terms of the time taken to reach a desired relative objective error, since the former solves an LP (Q_0), whereas the latter has to solve a QP (Q_u); the number of iterations taken by the different methods is, however, similar. There is no clear winner between randomized and deterministic grids, and both appear to perform well.

Table 3 compares our proposed methods against the CSS solutions with different tolerances, in terms of running time, final objective function, and distances of the algorithm outputs to the optimal solution $\boldsymbol{\phi}^*$ and the truth $\boldsymbol{\phi}^{\text{truth}}$. We find that all of our proposals yield marked improvements in running time compared with the CSS solution: with $n = 10{,}000$, $d = 4$ and $\tau = 10^{-4}$, CSS takes more than 20 h for all of the data sets we considered, whereas the RS-DI variant is around 50 times faster. The CSS solution may have a slightly improved objective function value on termination, but as shown in Table 3, all of our algorithms achieve an optimization error that is small by comparison with the statistical error, and from a statistical perspective, there is no point in seeking to reduce the optimization error further than this. Table 4 shows that the distances $\|\boldsymbol{\phi}^* - \boldsymbol{\phi}^{\text{truth}}\|/n^{1/2}$ are well concentrated around their means (i.e. do not vary greatly over different random samples drawn from the underlying distribution), which

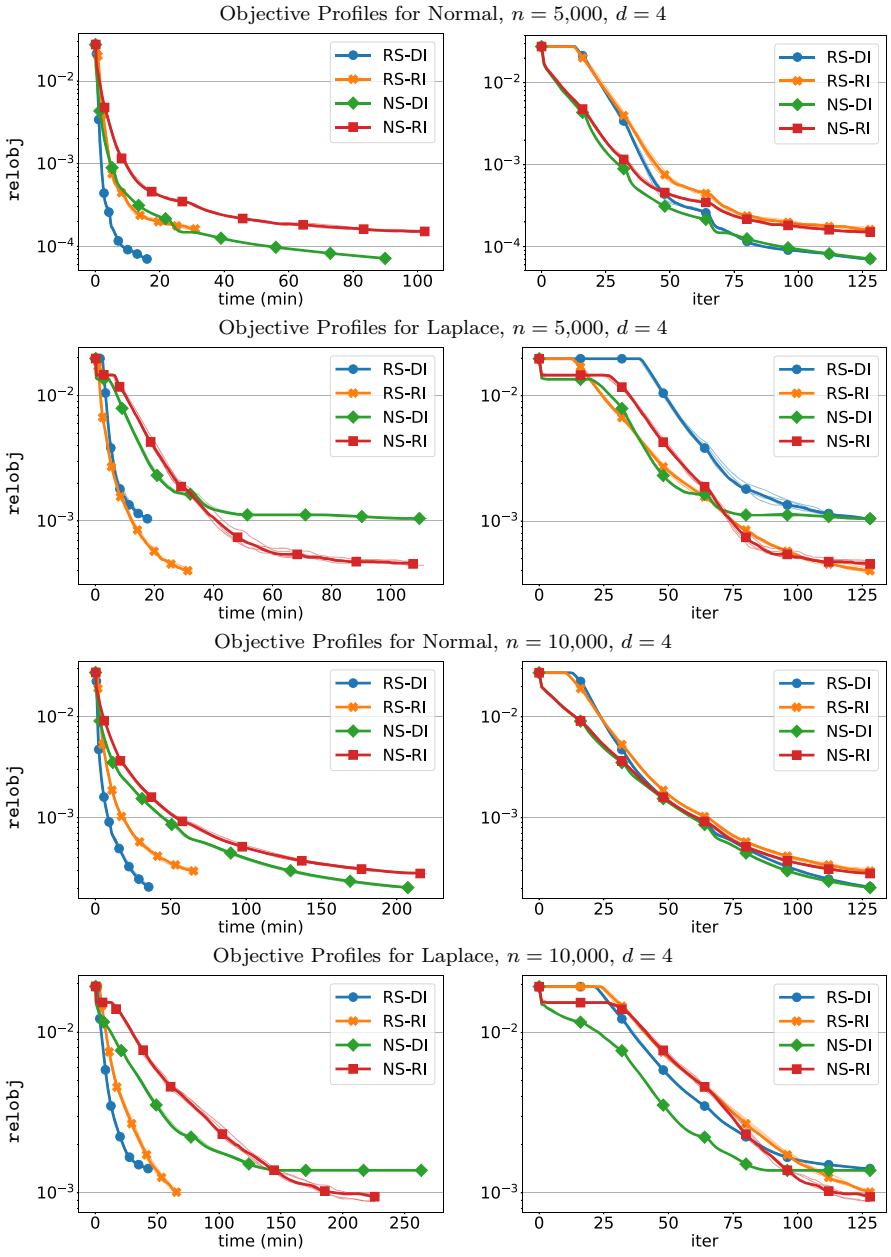

Fig. 2 Plots on a log-scale of relative objective versus time (mins) [left panel] and number of iterations [right panel]. For each of our four synthetic data sets, we ran five repetitions of each algorithm, so each bold line corresponds to the median of the profiles of the corresponding algorithm, and each thin line corresponds to the profile of one repetition. For the right panel, we show the profiles up to 128 iterations

Table 3 Comparison of our proposed methods with the CSS solution [20] and RS-RF [25]

Normal, $n = 5{,}000$, $d = 4$										
algo	param	obj	relobj	runtime	dopt	dtruth	iter	tO	aO	hO
CSS	1e-2	6.5209	1.1e-03	10.15	0.1955	0.2788				
	1e-3	6.5146	9.8e-05	110.04	0.0612	0.2465				
	1e-4	6.5140	0.0e-00	829.55	0.0000	0.2454				
RS-DI	None	6.5144	7.0e-05	16.04	0.0227	0.2499	128	6.18M	48.31K	20.23K
RS-RI	None	6.5150	1.6e-04	31.05	0.0289	0.2502	128	6.88M	53.75K	32.00K
NS-DI	None	6.5144	7.1e-05	89.94	0.0259	0.2497	128	6.18M	48.31K	20.23K
NS-RI	None	6.5149	1.5e-04	102.23	0.0312	0.2502	128	6.88M	53.75K	32.00K
RS-RF	5000	6.5174	5.2e-04	30.22	0.0575	0.2552	1024	5.12M	5.00K	5.00K
	10000	6.5168	4.4e-04	25.48	0.0429	0.2508	512	5.12M	10.00K	10.00K
	20000	6.5164	3.7e-04	23.05	0.0552	0.2567	256	5.12M	20.00K	20.00K
	40000	6.5158	2.9e-04	21.31	0.0344	0.2496	128	5.12M	40.00K	40.00K
	80000	6.5150	1.6e-04	42.25	0.0288	0.2499	128	10.24M	80.00K	80.00K
Laplace, $n = 5{,}000$, $d = 4$										
algo	param	obj	relobj	runtime	dopt	dtruth	iter	tO	aO	hO
CSS	1e-2	7.9183	3.0e-02	30.77	2.5100	2.5449				
	1e-3	7.6994	1.1e-03	387.34	0.5985	0.6514				
	1e-4	7.6908	0.0e-00	1007.80	0.0000	0.2552				
RS-DI	None	7.6988	1.0e-03	17.64	0.0592	0.2304	128	7.24M	56.54K	34.84K
RS-RI	None	7.6939	4.0e-04	31.32	0.0424	0.2632	128	6.88M	53.75K	32.00K
NS-DI	None	7.6989	1.0e-03	109.68	0.0640	0.2259	128	7.24M	56.54K	34.84K
NS-RI	None	7.6943	4.5e-04	107.57	0.0362	0.2601	128	6.88M	53.75K	32.00K
RS-RF	5000	7.7048	1.8e-03	31.43	0.0628	0.2732	1024	5.12M	5.00K	5.00K
	10000	7.7059	2.0e-03	27.33	0.0621	0.2745	512	5.12M	10.00K	10.00K
	20000	7.6986	1.0e-03	24.53	0.0527	0.2696	256	5.12M	20.00K	20.00K
	40000	7.6979	9.2e-04	22.68	0.0852	0.2776	128	5.12M	40.00K	40.00K
	80000	7.6949	5.4e-04	43.27	0.0349	0.2614	128	10.24M	80.00K	80.00K
Normal, $n = 10{,}000$, $d = 4$										
algo	param	obj	relobj	runtime	dopt	dtruth	iter	tO	aO	hO
CSS	1e-2	6.5634	4.1e-04	24.72	0.1018	0.1911				
	1e-3	6.5612	7.3e-05	181.01	0.0462	0.1854				
	1e-4	6.5607	0.0e-00	-	0.0000	0.1859				
RS-DI	None	6.5621	2.1e-04	35.40	0.0411	0.1939	128	6.67M	52.14K	21.85K
RS-RI	None	6.5626	3.0e-04	65.18	0.0443	0.1939	128	6.88M	53.75K	32.00K
NS-DI	None	6.5620	2.0e-04	207.32	0.0429	0.1953	128	6.67M	52.14K	21.85K
NS-RI	None	6.5625	2.8e-04	215.51	0.0452	0.1959	128	6.88M	53.75K	32.00K
RS-RF	5000	6.5690	1.3e-03	64.80	0.1097	0.2205	1024	5.12M	5.00K	5.00K

Table 3 continued

Normal, $n = 5{,}000$, $d = 4$										
algo	param	obj	relobj	runtime	dopt	dtruth	iter	tO	aO	hO
	10000	6.5704	1.5e-03	56.26	0.0470	0.1854	512	5.12M	10.00K	10.00K
	20000	6.5656	7.5e-04	50.17	0.0412	0.1890	256	5.12M	20.00K	20.00K
	40000	6.5638	4.7e-04	46.24	0.0478	0.1948	128	5.12M	40.00K	40.00K
	80000	6.5627	3.0e-04	89.52	0.0446	0.1948	128	10.24M	80.00K	80.00K
Laplace, $n = 10{,}000$, $d = 4$										
algo	param	obj	relobj	runtime	dopt	dtruth	iter	tO	aO	hO
CSS	1e-2	8.1796	5.8e-02	57.46	3.9044	3.9328				
	1e-3	7.7327	4.6e-04	–	0.3470	0.4081				
	1e-4	7.7292	0.0e-00	–	0.0000	0.2025				
RS-DI	None	7.7401	1.4e-03	42.70	0.0886	0.1825	128	8.14M	63.60K	39.20K
RS-RI	None	7.7370	1.0e-03	65.67	0.0753	0.2295	128	6.88M	53.75K	32.00K
NS-DI	None	7.7399	1.4e-03	263.40	0.0801	0.1724	128	8.14M	63.60K	39.20K
NS-RI	None	7.7365	9.4e-04	225.80	0.0480	0.2159	128	6.88M	53.75K	32.00K
RS-RF	5000	7.7541	3.2e-03	64.93	0.1051	0.2499	1024	5.12M	5.00K	5.00K
	10000	7.7543	3.2e-03	57.87	0.0918	0.2435	512	5.12M	10.00K	10.00K
	20000	7.7468	2.3e-03	51.25	0.1157	0.2529	256	5.12M	20.00K	20.00K
	40000	7.7511	2.8e-03	46.31	0.0907	0.2423	128	5.12M	40.00K	40.00K
	80000	7.7378	1.1e-03	89.13	0.0529	0.2230	128	10.24M	80.00K	80.00K

On a single dataset, we ran 5 repetitions of each algorithm with different random seeds and report the median statistics. Here, `obj` and `relobj` denote the objective and relative objective error, respectively, `runtime` denotes the running time (in minutes), `dopt` and `dtruth` denote the (Euclidean) distances between the algorithm outputs and the optimal solution and the truth, respectively, `iter` denotes the number of iterations, `tO` denotes the total number of oracles (grid points), `aO` denotes the average number of oracles (grid points) per iteration, and `hO` denotes the harmonic average of grid sizes (which equals $T/(M_T^{(1)})^2$). For CSS, `param` is the tolerance τ; for RS-RF, `param` is the (fixed) grid size m. Here '-' means the running time of the corresponding algorithm exceeded 20 h

Table 4 Statistics of the distance between the optimal solution and truth

	mean	std.error	min	25			
normal ($n = 5{,}000$)	0.2565	0.0093	0.2415	0.2480	0.2574	0.2629	0.2745
Laplace ($n = 5{,}000$)	0.2590	0.0114	0.2366	0.2508	0.2578	0.2676	0.2825

For each type of data set, we drew 40 random samples of the sizes given, and computed the log-concave MLE by CSS with tolerance 10^{-4}

provides further reassurance that our solutions are sufficiently accurate for practical purposes. On the other hand, the CSS solution with tolerance 10^{-3} is not always sufficiently reliable in terms of its statistical accuracy, e.g. for a Laplace distribution with $n = 5{,}000$. Our further experiments on real data sets reported in Appendix A.4 provide qualitatively similar conclusions.

Finally, Fig. 3 compares our proposed multistage increasing grid sizes (RS-DI/RS-RI) (see Tables 5 and 6) with the fixed grid size (RS-RF) proposed by [25], under the

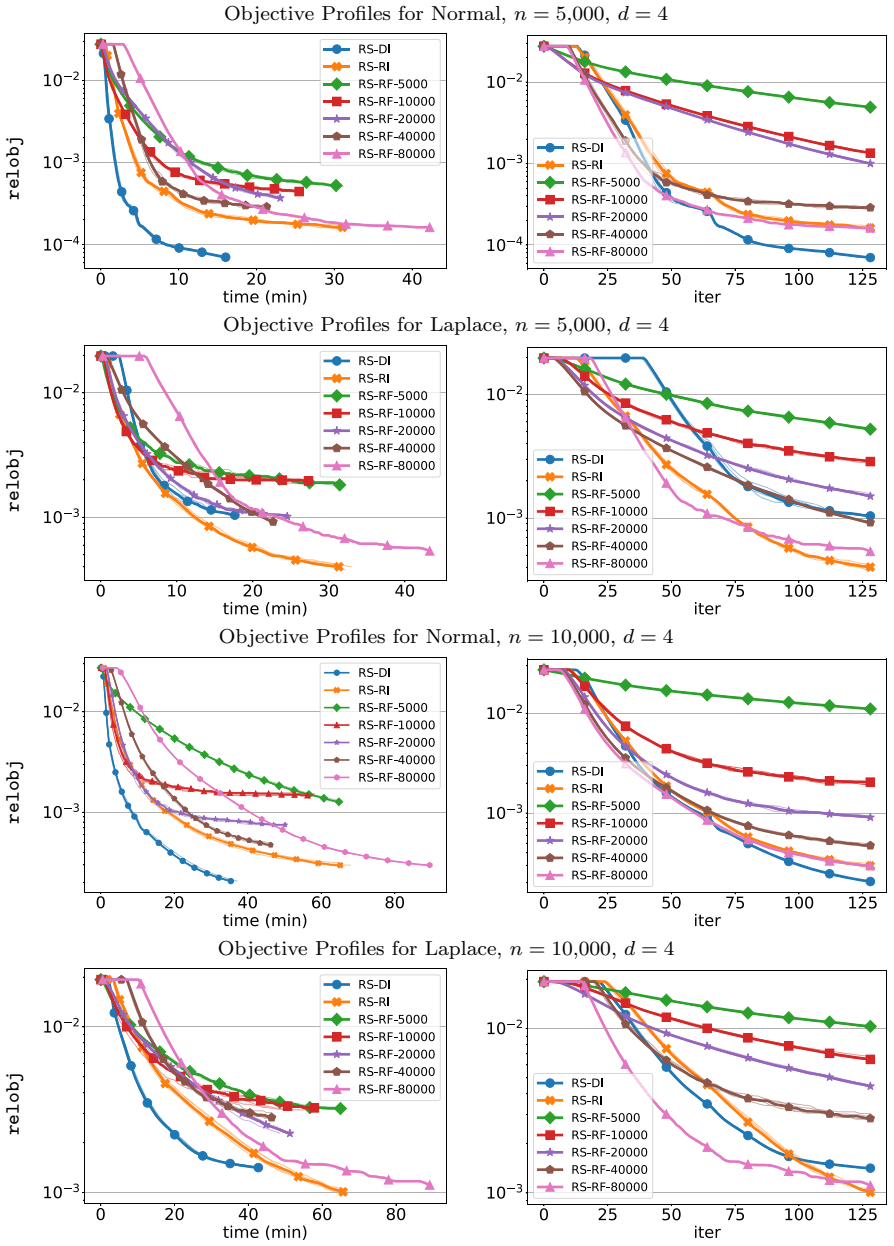

Fig. 3 Plots on a log-scale of relative objective versus time (mins) [left panel] and number of iterations [right panel]

randomized smoothing setting. We see that the benefits of using the increasing grid sizes as described by our theory carry over to improved practical performance, both in terms of running time and number of iterations.

Acknowledgements The authors acknowledge MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this paper. We also thank the anonymous reviewers for constructive comments on an earlier version that helped to improve this paper.

Funding 'Open Access funding provided by the MIT Libraries' Research supported in part by grants from the Office of Naval Research: N000141812298, N000142112841, N000142212665, the National Science Foundation and IBM to Rahul Mazumder. The research of Richard J. Samworth was supported by EPSRC grants EP/N031938/1 and EP/P031447/1, as well as ERC Advanced Grant 101019498.

Data availability All real data analyzed during this study are publicly available. URLs are included in this published article. The code to generate synthetic data is available in the `LogConcComp` repository: https://github.com/wenyuC94/LogConcComp.

Code availability The full code for computation in the current study is available in the `LogConcComp` repository: https://github.com/wenyuC94/LogConcComp.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

A Additional implementational and experimental details

A.1 Initialization: non-convex method

[20] show that the negative log-density $-\log \hat{p}_n(\cdot)$ of the log-concave MLE is a piecewise-affine convex function over its domain C_n. This allows us to parametrize these functions as $\varphi(\boldsymbol{x}) := \max_{j\in[m]}\{\boldsymbol{a}_j^\top \boldsymbol{x} + b_j\}$ for $x \in C_n$, where $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in \mathbb{R}^d$ and $b_1, \ldots, b_m \in \mathbb{R}$. We can then reformulate the problem as $\min_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_m\in\mathbb{R}^d, b_1,\ldots,b_m\in\mathbb{R}} f_0(\boldsymbol{a}, \boldsymbol{b})$, with non-convex objective

$$f_0(\boldsymbol{a}, \boldsymbol{b}) := \frac{1}{n}\sum_{i=1}^{n} \max_{j\in[m]}\{\boldsymbol{a}_j^\top \boldsymbol{x}_i + b_j\} + \int_{C_n} e^{-\max_{j\in[m]}\{\boldsymbol{a}_j^\top \boldsymbol{x}+b_j\}}\, d\boldsymbol{x}. \tag{28}$$

To approximate the integral, we use the same simple Riemann grid points mentioned in Sect. 3.2.1. Subgradients of the objective (28) are straightforward to compute via the chain rule and the subgradient of the maximum function (see, e.g., [8]). After standardizing each coordinate of our data to have mean zero and unit variance, which does not affect the final outcome due to the affine equivariance of the log-concave MLE [28, Remark 2.4], we generate $m = 10$ initializing hyperplanes from a standard

$(d+1)$-dimensional Gaussian distribution. We then obtain the initializer for our main algorithm by applying a vanilla subgradient method to the objective (28) [56, 66] with stepsize $t^{-1/2}$ at the tth iteration, terminating when the difference in the objective function at successive iterations drops below 10^{-4}, or after 100 iterations, whichever is the sooner. This technique is related to the non-convex method for log-concave density estimation proposed by [57], who considered a smoothed version of (28). Our goal here is only to seek a good initializer rather than the global optimum, and we found that the approach described above was effective in this respect, as well as being faster to compute than the method of [57].

A.2 Final polishing step

As mentioned in Sect. 3.2.1, once our algorithm terminates at $\tilde{\boldsymbol{\phi}}_T$, say, we evaluate the integral $I(\tilde{\boldsymbol{\phi}}_T)$ in the same way as [20]. Our final output, then is $\boldsymbol{\phi}_T := \tilde{\boldsymbol{\phi}}_T + \log I(\tilde{\boldsymbol{\phi}}_T)\mathbf{1}$; this final step not only improves the objective function, but also guarantees that $\exp[-\mathrm{cef}[\boldsymbol{\phi}_T](\cdot)]$ is a log-concave density. This can be shown by following the same arguments as in Steps 2-3 in the proof of Theorem 4.

A.3 Input parameter settings

According to Theorem 1, we should take $u = \frac{D}{2}\sqrt{\frac{B_1}{B_0}}$. By Table 2, for randomized smoothing, this is approximately $\frac{D}{2}C_1\sqrt{n}$, where $C_1 = \sqrt{\Delta e^{-\phi_0}}$. In our experiments for randomized smoothing, we chose $u = Dn^{1/4}/2$, while for Nesterov smoothing, we chose $u = D/2$. According to Theorem 1, $\eta = \sigma M_T^{(1)}/D$, where we took $\sigma = 10^{-4}$ for RS-RI and RS-DI, and $\sigma = 10^{-3}$ for NS-RI and NS-DI. For the competing RS-RF-m method, we present the better of the results from $\sigma \in \{10^{-3}, 10^{-4}\}$.

To illustrate the increasing grid size strategy we take in the experiments, we first present in Table 5 some potential schemes to achieve the $\tilde{O}(1/T)$ error rate on the objective function scale. In our experiments, we used the multi-stage increasing grid size scheme with $C_1 = 8$ and $\rho_1 = 2$. For the random grid (RI), we take $C = 5{,}000$

Table 5 Examples of increasing grid size ($|\mathcal{S}_t| = m_t$) schemes to achieve $\tilde{O}(1/T)$ rate (i.e. $M_T^{(1/2)} = \tilde{O}(1)$ for deterministic $\mathcal{S}_t$ and $M_T^{(1)} = \tilde{O}(1)$ for random $\mathcal{S}_t$)

Schemes	Grid Sizes	$M_T^{(1/2)} = \tilde{O}(1)$	$M_T^{(1)} = \tilde{O}(1)$
Exponential	$m_t = C\rho^t$	$\rho > 1$	$\rho > 1$
Polynomial	$m_t = Ct^\beta$	$\beta \geq 2$	$\beta \geq 1$
Multi-stage	$m_t = C\rho^a$ for $t \in \left[\mathbb{1}_{\{a\geq 1\}} + C_1\rho_1^{a-1}\mathbb{1}_{\{a\geq 2\}}, C_1\rho_1^a\right]$	$\rho_1^2 \leq \rho$	$\rho_1 \leq \rho$

Here, C and C_1 are positive constants. For the multi-stage scheme, $a \geq 1$ denotes the current stage number

Table 6 Summary of increasing grid size strategy (illustrated with $n = 5000$ observations from a Laplace distribution in four dimensions)

Stage number a		1	2	3	4
Stage length		16	16	32	64
DI	$m_{0,t}$	18	22	26	30
DI	m_t	10,656	23,582	45,969	81,558
RI	m_t	10,000	20,000	40,000	80,000

We take a four stage grid strategy and 128 iterations in total, with stage lengths shown in second line. For deterministic grids (denoted by DI), we use $m_{0,t}$ to determine the grid size (third line in the table), and the fourth line of the table is the corresponding grid size. For random grids (denoted by RI), the fifth line is the grid size of random sample

and $\rho = 2$. For the deterministic grid (DI), we first choose an axis-aligned grid with $m_{0,t}$ points in each dimension that encloses the convex hull C_n of the data. Then m_t is the number of these grid points that fall within C_n. Table 6 provides an illustration of this multi-stage strategy used in the numerical experiments for a Laplace distribution with $n = 5{,}000$ and $d = 4$. Code for the other settings is available in the github repository `LogConcComp`.

A.4 Experimental results on real data sets

We provide additional simulation results on three real data sets:

- **Stock returns:** The Stock returns real data consist of daily returns of four stocks[3] over $n = 10{,}000$ randomly sampled days between 1970 and 2010, normalized so that each dimension has unit variance. The real data are available at https://stooq.com/db/h/.
- **Census:** The Census real data consist of percentages of the population of different age groups (18-24, 25-44, 45-64 and 65+) for $n = 10{,}000$ randomly sampled Census tracts based on the 2015–2019 5-year ACS (American Community Survery),[4] and the data are normalized so that each dimension has unit variance. The data and description are available at https://www.census.gov/topics/research/guidance/planning-databases/2021.html.
- **Gas turbine:** The Gas turbine real data consist of 4 sensor measures[5] aggregated over one hour from a gas turbine for $n = 10{,}000$ hours between 2011 and 2015, normalized so that each dimension has unit variance. The data are available at https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set.

Table 7, Figs. 4 and 5 provide simulation results that correspond to those in Table 3, Figs. 2 and 3 respectively, but for three three real data sets. The table and figures reveal

[3] International Business Machines Corporation (IBM.US), JPMorgan Chase & Co. (JPM.US), Caterpillar Inc. (CAT.US), 3 M Company (MMM.US)

[4] pct_Pop_18_24_ACS_15_19, pct_Pop_25_44_ACS_15_19, pct_Pop_45_64_ACS_15_19, pct_Pop_65plus_ACS_15_19.

[5] Ambient temperature (AT), Ambient pressure (AP), Carbon monoxide (CO), Nitrogen oxides (NOx).

Table 7 Comparison of our proposed methods with the CSS solution [20] and RS-RF [25], but on 3 real datasets

Stock returns, $n = 10{,}000$, $d = 4$									
algo	param	obj	relobj	runtime	dopt	iter	tO	aO	hO
CSS	1e-2	6.3395	6.7e-02	315.19	5.4659				
	1e-3	5.9458	8.7e-04	–	0.5130				
	1e-4	5.9406	0.0e-00	–	0.0000				
RS-DI	None	5.9589	3.1e-03	38.47	0.1428	128	7.79M	60.86K	30.22K
RS-RI	None	5.9778	6.3e-03	59.65	0.2032	128	6.88M	53.75K	32.00K
NS-DI	None	5.9506	1.7e-03	254.91	0.0792	128	7.79M	60.86K	30.22K
NS-RI	None	5.9672	4.5e-03	228.82	0.1362	128	6.88M	53.75K	32.00K
RS-RF	5000	6.0003	1.0e-02	61.98	0.2015	1024	5.12M	5.00K	5.00K
	10000	5.9886	8.1e-03	54.20	0.2354	512	5.12M	10.00K	10.00K
	20000	5.9852	7.5e-03	49.16	0.2141	256	5.12M	20.00K	20.00K
	40000	5.9940	9.0e-03	46.25	0.3194	128	5.12M	40.00K	40.00K
	80000	5.9665	4.4e-03	84.59	0.1055	128	10.24M	80.00K	80.00K
Census, $n = 10{,}000$, $d = 4$									
algo	param	obj	relobj	runtime	dopt	iter	tO	aO	hO
CSS	1e-2	5.4458	9.4e-03	71.36	0.9222				
	1e-3	5.3953	1.2e-05	812.49	0.0098				
	1e-4	5.3952	0.0e-00	–	0.0000				
RS-DI	None	5.3995	8.0e-04	31.97	0.0478	128	6.19M	48.33K	25.79K
RS-RI	None	5.4003	9.4e-04	63.32	0.0506	128	6.88M	53.75K	32.00K
NS-DI	None	5.3992	7.3e-04	199.74	0.0453	128	6.19M	48.33K	25.79K
NS-RI	None	5.3992	7.4e-04	223.34	0.0475	128	6.88M	53.75K	32.00K
RS-RF	5000	5.4100	2.7e-03	69.86	0.1047	1024	5.12M	5.00K	5.00K
	10000	5.4093	2.6e-03	60.79	0.0796	512	5.12M	10.00K	10.00K
	20000	5.4074	2.3e-03	55.89	0.1236	256	5.12M	20.00K	20.00K
	40000	5.4059	2.0e-03	49.04	0.0587	128	5.12M	40.00K	40.00K
	80000	5.3998	8.6e-04	94.51	0.0477	128	10.24M	80.00K	80.00K
Gas turbine, $n = 10{,}000$, $d = 4$									
algo	param	obj	relobj	runtime	dopt	iter	tO	aO	hO
CSS	1e-2	5.5920	5.7e-03	95.59	0.5908				
	1e-3	5.5617	2.7e-04	–	0.0994				
	1e-4	5.5602	0.0e-00	–	0.0000				
RS-DI	None	5.5693	1.6e-03	34.14	0.0897	128	7.08M	55.28K	25.53K
RS-RI	None	5.5633	5.7e-04	61.90	0.0493	128	6.88M	53.75K	32.00K
NS-DI	None	5.5689	1.6e-03	230.47	0.0914	128	7.08M	55.28K	25.53K
NS-RI	None	5.5622	3.7e-04	224.88	0.0499	128	6.88M	53.75K	32.00K

Table 7 continued

Stock returns, $n = 10{,}000$, $d = 4$									
algo	param	obj	relobj	runtime	dopt	iter	tO	aO	hO
RS-RF	5000	5.5694	1.7e-03	67.28	0.1111	1024	5.12M	5.00K	5.00K
	10000	5.5670	1.2e-03	61.22	0.0794	512	5.12M	10.00K	10.00K
	20000	5.5673	1.3e-03	53.08	0.0455	256	5.12M	20.00K	20.00K
	40000	5.5657	9.9e-04	47.80	0.0547	128	5.12M	40.00K	40.00K
	80000	5.5632	5.5e-04	92.44	0.0501	128	10.24M	80.00K	80.00K

Details are given in the caption of Table 3

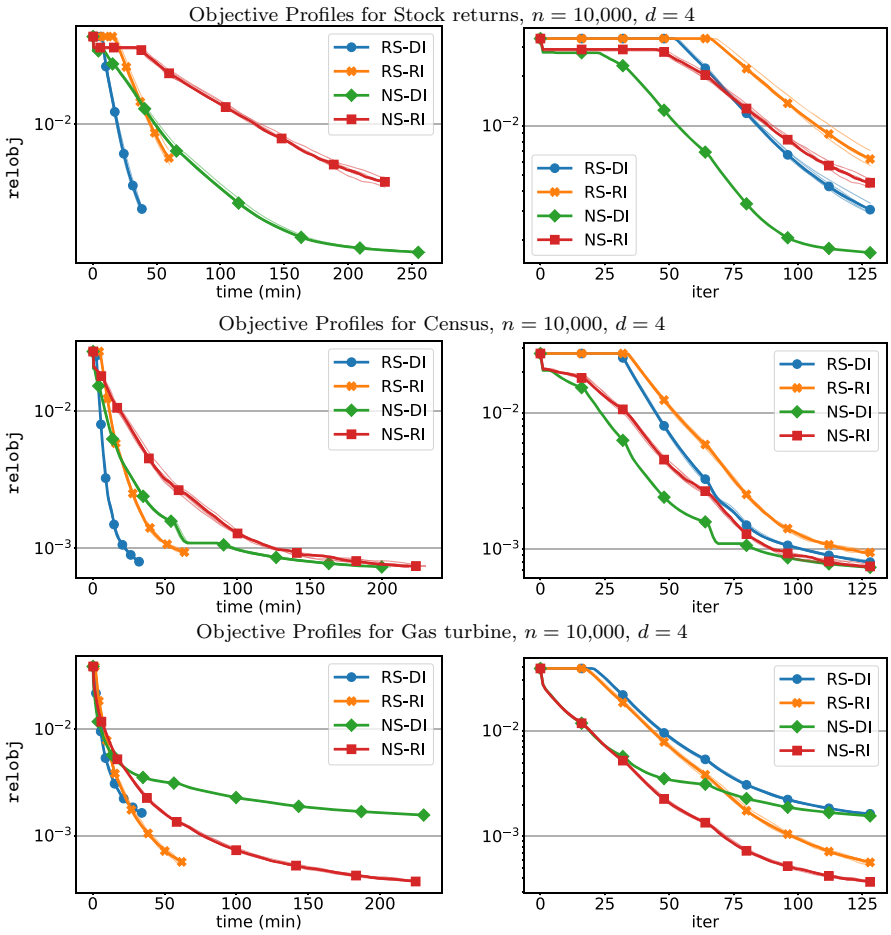

Fig. 4 Additional plots on a log-scale of Relative objective versus time (mins) [left panel] and number of iterations [right panel]. Details are given in the caption of Fig. 2

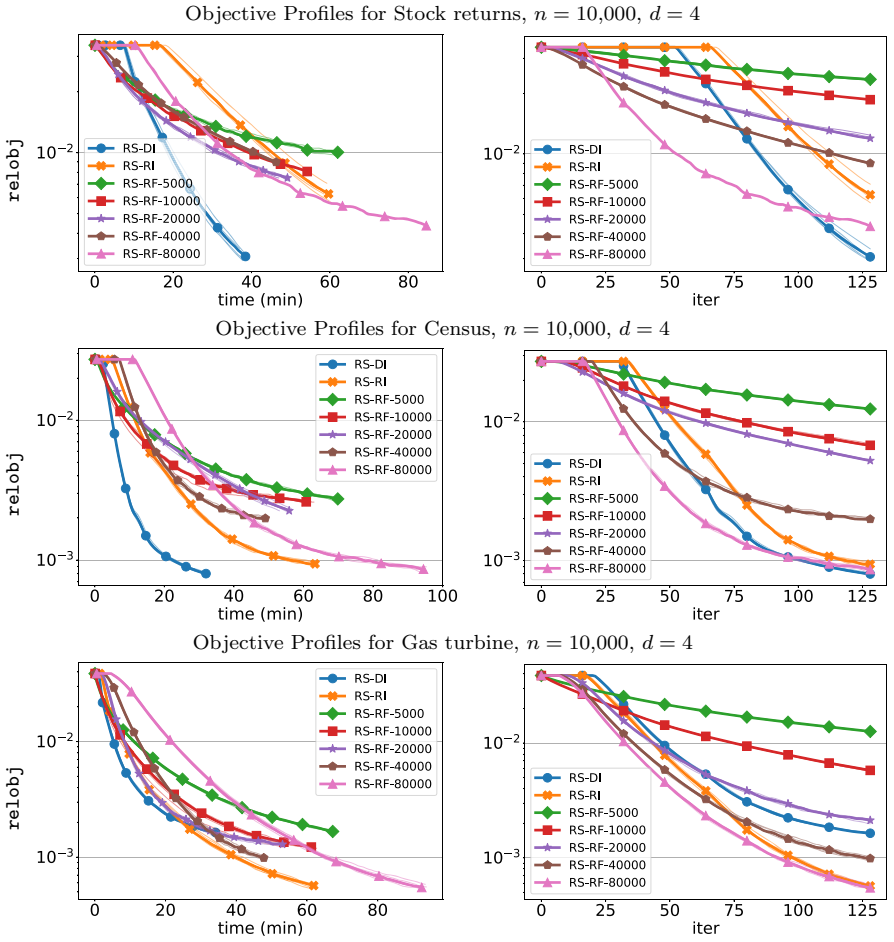

Fig. 5 Plots on a log-scale of relative objective versus time (mins) [left panel] and number of iterations [right panel]

a qualitatively very similar story to that presented for the simulated data in Sect. 6: the main conclusion is that our randomized smoothing approaches are significantly more computationally efficient than both the Nesterov smoothing and CSS methods.

B Proofs

B.1 Proofs of Propositions 1 and 2

The proof of Proposition 1 is adapted from the proof of [2, Lemma 2], which in turn is based on [12, Lemma 8].

Proof (Proof of Proposition 1) The proof has three parts.

Part 1. We first prove that $\phi_i^* \in [\phi_0, \phi^0]$ for all $i \in [n]$; or equivalently

$$\begin{aligned} \log \hat{p}_n(\boldsymbol{x}_i) &\geq -(n-1) - d(n-1)\log\big(2n + 2nd\log(2nd)\big) - \log \Delta \\ \log \hat{p}_n(\boldsymbol{x}_i) &\leq 1 + d\log\big(2n + 2nd\log(2nd)\big) - \log \Delta \end{aligned} \tag{29}$$

for all $i \in [n]$. Define

$$M := \sup_{\boldsymbol{x}\in\mathbb{R}^d} \hat{p}_n(\boldsymbol{x}) = \max_{i\in[n]} \hat{p}_n(\boldsymbol{x}_i) \quad \text{and} \quad R := \frac{\max_{i\in[n]} \hat{p}_n(\boldsymbol{x}_i)}{\min_{i\in[n]} \hat{p}_n(\boldsymbol{x}_i)}.$$

We proceed to obtain an upper bound on R. To this end, let $\bar{p}_n$ denote the uniform density over C_n. If $\hat{p}_n(\boldsymbol{x}_i) = \bar{p}_n(\boldsymbol{x}_i) = 1/\Delta$ for all i, then (29) holds. So we may assume that $\hat{p}_n \neq \bar{p}_n$, so that $R > 1$ and $M > 1/\Delta$. For a density p on $\mathbb{R}^d$ and for $t \in \mathbb{R}$, let $L_p(t) := \{\boldsymbol{x} \in \mathbb{R}^d : p(\boldsymbol{x}) \geq t\}$ denote the super-level set of p at height t. Since $\hat{p}_n$ is supported on C_n, and since $\hat{p}_n(\boldsymbol{x}) \geq \min_{i\in[n]} \hat{p}_n(\boldsymbol{x}_i) = M/R$ for $\boldsymbol{x} \in C_n$, it follows by [12, Lemma 8],[6] that when $R \geq e$,

$$\Delta = \mathrm{vol}(C_n) \leq \mathrm{vol}\big(L_{\hat{p}_n}(M/R)\big) \leq \frac{e\log^d R}{M}. \tag{30}$$

On the other hand, since $\inf_{\boldsymbol{x}\in C_n} \hat{p}_n(\boldsymbol{x}) = M/R$, we have $(M/R)\cdot\Delta \leq 1$, so for $R < e$, we have $M \leq R/\Delta < e/\Delta$. We deduce that

$$M \leq \frac{e\log_+^d R}{\Delta}, \tag{31}$$

for all $R > 1$, where $\log_+(x) := 1 \vee \log x$. Now, by the optimality of $\hat{p}_n$, we have

$$\begin{aligned} n\log(1/\Delta) &= \sum_{i=1}^n \log \bar{p}_n(\boldsymbol{x}_i) \leq \sum_{i=1}^n \log \hat{p}_n(\boldsymbol{x}_i) \\ &\leq \min_{i\in[n]} \log \hat{p}_n(\boldsymbol{x}_i) + (n-1)\max_{i\in[n]} \log \hat{p}_n(\boldsymbol{x}_i) \\ &= \log(M/R) + (n-1)\log M, \end{aligned} \tag{32}$$

so that $R \leq (M\Delta)^n$. It follows that when $R \geq e$, we have from (31) and the fact that $\log y \leq y$ for $y > 0$ that

$$R = \frac{R^2}{R} \leq \frac{e^{2n}\log^{2nd} R}{R} \leq e^{2n}(2nd)^{2nd}. \tag{33}$$

Since (33) holds trivially when $R < e$, we may combine (33) with (31) to obtain

$$\log M \leq 1 + d\log\big(2n + 2nd\log(2nd)\big) - \log \Delta \tag{34}$$

[6] In fact, the factor of e is omitted in the statement of [12, Lemma 8] but one can see from the authors' inequalities (27) and (28) that it should be present.

Moreover, from (32) and (34), we also have

$$\begin{aligned}\log(M/R) &\geq -n\log\Delta - (n-1)\log M \\ &\geq -(n-1) - d(n-1)\log\big(2n + 2nd\log(2nd)\big) - \log\Delta,\end{aligned}$$

as required.

Part 2. Now we extend the above result to all $\boldsymbol{\phi} \in \mathbb{R}^n$ such that $I(\boldsymbol{\phi}) = 1$ and $f(\boldsymbol{\phi}) \leq f(\bar{\boldsymbol{\phi}})$, where $\bar{\boldsymbol{\phi}}$ is defined just after Proposition 1. The key observation here is that the proof of Part 1 applies to any density with log-likelihood at least that of the uniform distribution over C_n. In particular, for any $\boldsymbol{\phi}$ satisfying these conditions, the density $p \in \mathcal{F}_d$ given by $p(\boldsymbol{x}) = \exp\{-\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x})\}$ has log-likelihood at least that of the uniform distribution over C_n, so

$$-\phi^0 \leq \min_{i\in[n]} \log p(\boldsymbol{x}_i) \leq \sup_{\boldsymbol{x}\in\mathbb{R}^d} \log p(\boldsymbol{x}) \leq -\phi_0,$$

as required.

Part 3. We now consider the case for a general $\boldsymbol{\phi} \in \mathbb{R}^n$ with $f(\boldsymbol{\phi}) \leq f(\bar{\boldsymbol{\phi}})$. Let $\tilde{\boldsymbol{\phi}} := \boldsymbol{\phi} + \log I(\boldsymbol{\phi})\mathbf{1}$, so that $I(\tilde{\boldsymbol{\phi}}) = 1$ and $\mathrm{cef}[\tilde{\boldsymbol{\phi}}](\cdot) = \mathrm{cef}[\boldsymbol{\phi}](\cdot) + \log(I(\boldsymbol{\phi}))$. Furthermore,

$$f(\tilde{\boldsymbol{\phi}}) = \frac{1}{n}\mathbf{1}^\top\boldsymbol{\phi} + \log I(\boldsymbol{\phi}) + 1 \leq \frac{1}{n}\mathbf{1}^\top\boldsymbol{\phi} + I(\boldsymbol{\phi}) = f(\boldsymbol{\phi}) \leq f(\bar{\boldsymbol{\phi}}).$$

The result therefore follows by Part 2. □

Proof (Proof of Proposition 2) Recall our notation from Sect. 2 that $\boldsymbol{\alpha}^*[\boldsymbol{\phi}](\boldsymbol{x}) \in A[\boldsymbol{\phi}](\boldsymbol{x})$ denotes a solution to (Q_0) at $\boldsymbol{x} \in C_n$. Recall further from (9) that any subgradient $\boldsymbol{g}(\boldsymbol{\phi})$ of f at $\boldsymbol{\phi}$ is of the form

$$\boldsymbol{g}(\boldsymbol{\phi}) = \frac{1}{n}\mathbf{1} - \boldsymbol{\gamma},$$

where $\boldsymbol{\gamma} := \Delta\mathbb{E}\big[\boldsymbol{\alpha}^*[\boldsymbol{\phi}](\boldsymbol{\xi})\exp\{-\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{\xi})\}\big]$ and $\boldsymbol{\xi}$ is uniformly distributed on C_n. Since $\boldsymbol{\alpha}^*$ lies in the simplex $\{\boldsymbol{\alpha} \in \mathbb{R}^n : \boldsymbol{\alpha} \geq 0, \mathbf{1}^\top\boldsymbol{\alpha} = 1\}$, we have $\boldsymbol{\gamma} \geq 0$ and

$$\mathbf{1}^\top\boldsymbol{\gamma} = \Delta\mathbb{E}[\mathbf{1}^\top\alpha^*[\boldsymbol{\phi}](\boldsymbol{\xi})\exp\{-\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{\xi})\}] = \Delta\mathbb{E}[\exp\{-\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{\xi})\}] = I(\boldsymbol{\phi}).$$

In particular, $\|\boldsymbol{\gamma}\|_1 = I(\boldsymbol{\phi})$, so

$$\|\boldsymbol{g}(\boldsymbol{\phi})\|^2 = \frac{1}{n} - \frac{2}{n}\mathbf{1}^\top\boldsymbol{\gamma} + \|\boldsymbol{\gamma}\|^2 \leq \frac{1}{n} - \frac{2}{n}I(\boldsymbol{\phi}) + I(\boldsymbol{\phi})^2.$$

If $I(\boldsymbol{\phi}) \leq 1/2$, then $\|\boldsymbol{g}(\boldsymbol{\phi})\|^2 \leq 1/4 + 1/n$; if $I(\phi) > 1/2$, then $\|\boldsymbol{g}(\boldsymbol{\phi})\|^2 \leq I(\boldsymbol{\phi})^2$. Therefore, $\|\boldsymbol{g}(\boldsymbol{\phi})\|^2 \leq \max\big\{1/4 + 1/n, I(\boldsymbol{\phi})^2\big\}$. □

B.2 Proofs of Proposition 3 and Proposition 4

The proof of Proposition 3 is based on the following properties of the quadratic program $q_u[\boldsymbol{\phi}](\boldsymbol{x})$ defined in (Q_u), as well as its unique optimizer $\boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x})$:

Proposition 5 *For $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ and $\boldsymbol{x} \in C_n$, we have*
(a) $\|\boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x})\| \le 1$ for any $\boldsymbol{x} \in C_n$ and $\boldsymbol{\phi} \in \boldsymbol{\Phi}$;
(b) $q_{u'}[\boldsymbol{\phi}](\boldsymbol{x}) + (u'-u)/2 \le q_u[\boldsymbol{\phi}](\boldsymbol{x}) \le q_{u'}[\boldsymbol{\phi}](\boldsymbol{x})$ for $u' \in [0,u]$;
(c) $\phi_0 - \frac{u}{2} \le q_u[\boldsymbol{\phi}](\boldsymbol{x}) \le \phi^0$ for all $u \ge 0$, $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ and $\boldsymbol{x} \in C_n$;
(d) $\|\boldsymbol{\alpha}_u^*[\tilde{\boldsymbol{\phi}}](\boldsymbol{x}) - \boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x})\| \le (1/u)\|\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}\|$ for any $u > 0$, $\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}} \in \boldsymbol{\Phi}$, and any $\boldsymbol{x} \in C_n$.

Proof The proof exploits ideas from [52]. For (a), observe that $\boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x}) \in E(\boldsymbol{x}) \subseteq \{\boldsymbol{\alpha} \in \mathbb{R}^n : \mathbf{1}_n^\top \boldsymbol{\alpha} = 1, \boldsymbol{\alpha} \ge 0\}$, and this simplex is the convex hull of $n+1$ points that all lie in the closed unit Euclidean ball in $\mathbb{R}^n$.

The lower bound in (b) follows immediately from the definition of the quadratic program in (Q_u). For the upper bound, for $u' \in [0,u]$, we have

$$\begin{aligned} & q_u[\boldsymbol{\phi}](\boldsymbol{x}) - q_{u'}[\boldsymbol{\phi}](\boldsymbol{x}) \\ &= \boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x})^\top \boldsymbol{\phi} + \frac{u}{2}\left\|\boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x}) - \boldsymbol{\alpha}_0\right\|^2 - \boldsymbol{\alpha}_{u'}^*[\boldsymbol{\phi}](\boldsymbol{x})^\top \boldsymbol{\phi} \\ &\quad - \frac{u'}{2}\left\|\boldsymbol{\alpha}_{u'}^*[\boldsymbol{\phi}](\boldsymbol{x}) - \boldsymbol{\alpha}_0\right\|^2 + \frac{u'-u}{2} \\ &\le \frac{u-u'}{2}\left(\left\|\boldsymbol{\alpha}_{u'}^*[\boldsymbol{\phi}](\boldsymbol{x}) - \boldsymbol{\alpha}_0\right\|^2 - 1\right) \le \frac{u-u'}{2}\left(1 - \frac{2}{n} + \frac{1}{n} - 1\right) \le 0. \end{aligned}$$

(c) For all $u \ge 0$, $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ and $\boldsymbol{x} \in C_n$, we have

$$\begin{aligned} q_u[\boldsymbol{\phi}](\boldsymbol{x}) = \inf_{\boldsymbol{\alpha} \in E(\boldsymbol{x})}\left(\boldsymbol{\alpha}^\top \boldsymbol{\phi} + \frac{u}{2}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|^2 - \frac{u}{2}\right) &\ge \inf_{\boldsymbol{\alpha} \in E(\boldsymbol{x})} \boldsymbol{\alpha}^\top \boldsymbol{\phi} - \frac{u}{2} \\ &\ge \phi_0 \inf_{\boldsymbol{\alpha} \in E(\boldsymbol{x})} \boldsymbol{\alpha}^\top \mathbf{1}_n - \frac{u}{2} = \phi_0 - \frac{u}{2}. \end{aligned}$$

Similarly,

$$q_u[\boldsymbol{\phi}](\boldsymbol{x}) \le \phi^0 \sup_{\boldsymbol{\alpha} \in E(\boldsymbol{x})} \boldsymbol{\alpha}^\top \mathbf{1}_n + \frac{u}{2} \sup_{\boldsymbol{\alpha} \in E(\boldsymbol{x})} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|^2 - \frac{u}{2} \le \phi^0.$$

(d) Observe that

$$\boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{\alpha} \in E(\boldsymbol{x})}\left(\boldsymbol{\alpha}^\top \boldsymbol{\phi} + \frac{u}{2}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|^2\right) = \operatorname*{argmin}_{\boldsymbol{\alpha} \in E(\boldsymbol{x})}\left\|\boldsymbol{\alpha} - \left(\boldsymbol{\alpha}_0 - \frac{\boldsymbol{\phi}}{u}\right)\right\|^2,$$

so $\boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x})$ is the Euclidean projection of $\boldsymbol{\alpha}_0 - (\boldsymbol{\phi}/u)$ onto $E(\boldsymbol{x})$. Since this projection is an ℓ_2-contraction, we deduce that

$$\left\|\boldsymbol{\alpha}_u^*[\tilde{\boldsymbol{\phi}}](\boldsymbol{x}) - \boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x})\right\| \le \left\|\boldsymbol{\alpha}_0 - \frac{\tilde{\boldsymbol{\phi}}}{u} - \left(\boldsymbol{\alpha}_0 - \frac{\boldsymbol{\phi}}{u}\right)\right\| = \frac{1}{u}\|\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}\|,$$

as required. □

Proof (of Proposition 3) (a) For $u \ge 0$ and $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, let

$$\tilde{I}_u(\boldsymbol{\phi}) := \int_{C_n} e^{-q_u[\boldsymbol{\phi}](\boldsymbol{x})}\, d\boldsymbol{x} = \Delta\mathbb{E}(e^{-q_u[\boldsymbol{\phi}](\boldsymbol{\xi})}),$$

where $\boldsymbol{\xi}$ is uniformly distributed on C_n, so that $\tilde{I}_0(\boldsymbol{\phi}) = I(\boldsymbol{\phi})$. By definition of $\tilde{f}_u$, we have for $u' \in [0, u]$ that

$$\tilde{f}_u(\boldsymbol{\phi}) - \tilde{f}_{u'}(\boldsymbol{\phi}) = \tilde{I}_u(\boldsymbol{\phi}) - \tilde{I}_{u'}(\boldsymbol{\phi}) = \Delta\mathbb{E}\big(e^{-q_u[\boldsymbol{\phi}](\boldsymbol{\xi})} - e^{-q_{u'}[\boldsymbol{\phi}](\boldsymbol{\xi})}\big) \ge 0, \tag{35}$$

where the inequality follows from Proposition 5(b). Hence, for every $u \ge 0$ and $\boldsymbol{\phi} \in \boldsymbol{\Phi}$,

$$\tilde{I}_u(\boldsymbol{\phi}) \le e^{u/2}\tilde{I}_0(\boldsymbol{\phi}) = e^{u/2} I(\boldsymbol{\phi}). \tag{36}$$

Now, from (35), Proposition 5(b) and (36), we deduce that

$$\tilde{f}_u(\boldsymbol{\phi}) - \tilde{f}_{u'}(\boldsymbol{\phi}) \le \big(e^{(u-u')/2} - 1\big)\tilde{I}_{u'}(\boldsymbol{\phi}) \le \frac{u-u'}{2} e^{u'/2} I(\boldsymbol{\phi}),$$

as required.

(b) For each $\boldsymbol{x} \in C_n$, the function $\boldsymbol{\phi} \mapsto q_u[\boldsymbol{\phi}](\boldsymbol{x})$ is the infimum of a set of affine functions of $\boldsymbol{\phi}$, so it is concave. Moreover, $y \mapsto e^{-y}$ is a decreasing convex function, so $\boldsymbol{\phi} \mapsto e^{-q_u[\boldsymbol{\phi}](\boldsymbol{x})}$ is convex, and it follows that $\boldsymbol{\phi} \mapsto (1/n)\mathbf{1}^\top\boldsymbol{\phi} + \Delta\mathbb{E}(e^{-q_u[\boldsymbol{\phi}](\boldsymbol{\xi})}) = \tilde{f}_u(\boldsymbol{\phi})$ is convex. Similarly to the proof of Proposition 2, any subgradient $\tilde{\boldsymbol{g}}_u(\phi)$ of $\tilde{f}_u$ at $\boldsymbol{\phi}$ satisfies

$$\|\tilde{\boldsymbol{g}}_u(\boldsymbol{\phi})\|^2 \le \max\left\{\frac{1}{4} + \frac{1}{n}, \tilde{I}_u(\boldsymbol{\phi})^2\right\} \le \max\left\{\frac{1}{4} + \frac{1}{n}, \Delta^2 e^{-2\phi_0 + u}\right\}. \tag{37}$$

But $\boldsymbol{\phi}^* \in \boldsymbol{\Phi}$, so $\Delta^2 e^{-2\phi_0+u} \ge (\Delta e^{-\phi_0})^2 \ge I(\boldsymbol{\phi}^*)^2 = 1 \ge 1/4 + 1/n$. Hence, $\tilde{f}_u$ is $e^{-\phi_0+u/2}$-Lipschitz.

(c) To establish the Lipschitz property of $\nabla_{\boldsymbol{\phi}}\tilde{f}_u$, for any $\boldsymbol{x} \in C_n$, any $\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}} \in \boldsymbol{\Phi}$, and $t \in [0, 1]$, we define

$$\eta(t) := e^{-q_u[\boldsymbol{\phi} + t(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi})](\boldsymbol{x})}.$$

Then

$$\eta'(t) = -e^{-q_u[\boldsymbol{\phi}+t(\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi})](\boldsymbol{x})}(\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top \boldsymbol{\alpha}_u^*[\boldsymbol{\phi}+t(\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi})](\boldsymbol{x}).$$

By the mean value theorem there exists $t_0 \in [0, 1]$ such that

$$\begin{aligned}
&\left|e^{-q_u[\tilde{\boldsymbol{\phi}}](\boldsymbol{x})} - e^{-q_u[\boldsymbol{\phi}](\boldsymbol{x})}\right| = |\eta(1)-\eta(0)| = |\eta'(t_0)| \\
&\le e^{-q_u[\boldsymbol{\phi}+t_0(\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi})](\boldsymbol{x})}\|\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi}\|\left\|\boldsymbol{\alpha}_u^*[\boldsymbol{\phi}+t_0(\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi})](\boldsymbol{x})\right\| \\
&\le e^{-\phi_0+u/2}\|\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi}\|, \qquad (38)
\end{aligned}$$

where the final bound follows from Proposition 5(a) and (c). Now, for any $\boldsymbol{x} \in C_n$, we have by (38) as well as Proposition 5(a), (c) and (d) that

$$\begin{aligned}
&\left\|e^{-q_u[\tilde{\boldsymbol{\phi}}](\boldsymbol{x})}\boldsymbol{\alpha}_u^*[\tilde{\boldsymbol{\phi}}](\boldsymbol{x}) - e^{-q_u[\boldsymbol{\phi}](\boldsymbol{x})}\boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x})\right\| \\
&= \left\|\left(e^{-q_u[\tilde{\boldsymbol{\phi}}](\boldsymbol{x})} - e^{-q_u[\boldsymbol{\phi}](\boldsymbol{x})}\right)\boldsymbol{\alpha}_u^*[\tilde{\boldsymbol{\phi}}](\boldsymbol{x}) + e^{-q_u[\boldsymbol{\phi}](\boldsymbol{x})}\left(\boldsymbol{\alpha}_u^*[\tilde{\boldsymbol{\phi}}](\boldsymbol{x}) - \boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{x})\right)\right\| \\
&\le e^{-\phi_0+u/2}\|\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi}\| + e^{+}u/2-\phi_0\frac{1}{u}\|\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi}\| = e^{-\phi_0+u/2}(1+u^{-1})\|\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi}\|.
\end{aligned}$$

It follows that for any $\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}} \in \boldsymbol{\Phi}$, we have

$$\begin{aligned}
\left\|\nabla_{\tilde{\boldsymbol{\phi}}}\tilde{f}_u(\tilde{\boldsymbol{\phi}}) - \nabla_{\boldsymbol{\phi}}\tilde{f}_u(\boldsymbol{\phi})\right\| &\le \Delta\mathbb{E}\left\|e^{-q_u[\tilde{\boldsymbol{\phi}}](\boldsymbol{\xi})}\boldsymbol{\alpha}_u^*[\tilde{\boldsymbol{\phi}}](\boldsymbol{\xi}) - e^{-q_u[\boldsymbol{\phi}](\boldsymbol{\xi})}\boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{\xi})\right\| \\
&\le \Delta e^{-\phi_0+u/2}(1+u^{-1})\|\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi}\|,
\end{aligned}$$

as required.

(d) For $u \ge 0$ and $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, it follows from Proposition 5(a) and (c) that

$$\begin{aligned}
\mathbb{E}\left(\|\tilde{\boldsymbol{G}}_u(\boldsymbol{\phi},\boldsymbol{\xi}) - \nabla\tilde{f}_u(\boldsymbol{\phi})\|^2\right) &= \mathbb{E}\left\|\Delta e^{-q_u[\boldsymbol{\phi}](\boldsymbol{\xi})}\boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{\xi}) - \Delta\mathbb{E}\left(e^{-q_u[\boldsymbol{\phi}](\boldsymbol{\xi})}\boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{\xi})\right)\right\|^2 \\
&\le \Delta^2\mathbb{E}\left\|e^{-q_u[\boldsymbol{\phi}](\boldsymbol{\xi})}\boldsymbol{\alpha}_u^*[\boldsymbol{\phi}](\boldsymbol{\xi})\right\|^2 \le (\Delta e^{-\phi_0+u/2})^2,
\end{aligned}$$

as required. □

Proposition 6 *If $\boldsymbol{z}$ is uniformly distributed on the unit ℓ_2-ball in $\mathbb{R}^n$, then $\mathbb{E}(\|\boldsymbol{z}\|_\infty) \le \sqrt{\frac{2\log n}{n+1}}$.*

Proof By [71, Proposition 3], we have that $\boldsymbol{z} \overset{d}{=} U^{1/n}\boldsymbol{z}'$, where $U \sim \mathcal{U}[0,1]$, where $\boldsymbol{z}'$ is uniformly distributed on the unit sphere in $\mathbb{R}^n$, and where U and $\boldsymbol{z}'$ are independent. Thus,

$$\mathbb{E}(\|\boldsymbol{z}\|_\infty) = \mathbb{E}(U^{1/n})\mathbb{E}(\|\boldsymbol{z}'\|_\infty) = \frac{n}{n+1}\mathbb{E}(\|\boldsymbol{z}'\|_\infty). \qquad (39)$$

Moreover, if $\boldsymbol{\zeta} \sim \mathcal{N}(0, I_n)$, then $\|\boldsymbol{\zeta}\|$ and $\boldsymbol{\zeta}/\|\boldsymbol{\zeta}\|$ are independent, and $z' \stackrel{d}{=} \boldsymbol{\zeta}/\|\boldsymbol{\zeta}\|$. It follows that

$$\mathbb{E}(\|z'\|_\infty) = \mathbb{E}\left(\frac{\|\boldsymbol{\zeta}\|_\infty}{\|\boldsymbol{\zeta}\|}\right) \cdot \frac{\mathbb{E}(\|\boldsymbol{\zeta}\|)}{\mathbb{E}(\|\boldsymbol{\zeta}\|)}$$
$$= \frac{\mathbb{E}(\|\boldsymbol{\zeta}\|_\infty)}{\mathbb{E}(\|\boldsymbol{\zeta}\|)} = \frac{\sqrt{2\log n}\cdot\Gamma(n/2)}{2^{1/2}\Gamma\left((n+1)/2\right)} \le \frac{1}{n}\sqrt{2(n+1)\log n}, \tag{40}$$

where the final bound follows from bounds on the gamma function, e.g. [29, Lemma 12]. The result follows from (39) and (40). □

Proof (of Proposition 4) (a) By Jensen's inequality,

$$\bar{f}_u(\boldsymbol{\phi}) = \mathbb{E} f(\boldsymbol{\phi} + uz) \ge f(\boldsymbol{\phi}). \tag{41}$$

For the upper bound, let $\boldsymbol{v} \in \mathbb{R}^n$ have $\|\boldsymbol{v}\| \le 1$ and, for some $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, let $\bar{\boldsymbol{\phi}} := \boldsymbol{\phi} + u\boldsymbol{v}$. For any $\boldsymbol{\alpha} \in E(\boldsymbol{x})$, we have

$$|\boldsymbol{\alpha}^\top \bar{\boldsymbol{\phi}} - \boldsymbol{\alpha}^\top \boldsymbol{\phi}| = u|\boldsymbol{\alpha}^\top \boldsymbol{v}| \le u\|\boldsymbol{\alpha}\|_1 \|\boldsymbol{v}\|_\infty \le u.$$

Therefore, for any $\boldsymbol{x} \in C_n$, we have $\mathrm{cef}[\bar{\boldsymbol{\phi}}](\boldsymbol{x}) \ge \mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x}) - u$. Hence

$$I(\bar{\boldsymbol{\phi}}) = \Delta\mathbb{E}[\exp\{-\mathrm{cef}[\bar{\boldsymbol{\phi}}](\boldsymbol{\xi})\}] \le e^u \Delta\mathbb{E}[\exp\{-\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{\xi})\}] = e^u I(\boldsymbol{\phi}). \tag{42}$$

Recall that all subgradients of I at $\tilde{\boldsymbol{\phi}} \in \mathbb{R}^n$ are of the form $-\boldsymbol{\gamma}(\tilde{\boldsymbol{\phi}})$, where

$$\boldsymbol{\gamma}(\tilde{\boldsymbol{\phi}}) = \Delta\mathbb{E}\big(\boldsymbol{\alpha}\exp\{-\mathrm{cef}[\tilde{\boldsymbol{\phi}}](\boldsymbol{\xi})\}\big)$$

for some $\boldsymbol{\alpha} \in A[\tilde{\boldsymbol{\phi}}](\boldsymbol{x})$. Moreover, as we saw in the proof of Proposition 2, $\|\boldsymbol{\gamma}(\tilde{\boldsymbol{\phi}})\|_1 = I(\tilde{\boldsymbol{\phi}})$. We deduce from [59, Theorem 24.7] that

$$\bar{f}_u(\boldsymbol{\phi}) - f(\boldsymbol{\phi}) = \mathbb{E}\big(I(\boldsymbol{\phi} + uz) - I(\boldsymbol{\phi})\big) \le u \sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}, \|\boldsymbol{v}\|\le 1} I(\boldsymbol{\phi} + u\boldsymbol{v})\mathbb{E}\big(\|z\|_\infty\big)$$
$$\le I(\boldsymbol{\phi})ue^u\mathbb{E}\big(\|z\|_\infty\big) \le I(\boldsymbol{\phi})ue^u\sqrt{\frac{2\log n}{n+1}},$$

where the final inequality uses Proposition 6.

(b) By the convexity of f, we have

$$\bar{f}_{u'}(\boldsymbol{\phi}) = \mathbb{E}\big(f(\boldsymbol{\phi} + u'z)\big) \le \frac{u'}{u}\bar{f}_u(\boldsymbol{\phi}) + \left(1 - \frac{u'}{u}\right) f(\boldsymbol{\phi}) \le \bar{f}_u(\boldsymbol{\phi}),$$

where the last inequality uses property (a).

(c) For each $\boldsymbol{v} \in \mathbb{R}^n$ with $\|\boldsymbol{v}\| = 1$, the map $\boldsymbol{\phi} \mapsto f(\boldsymbol{\phi} + u\boldsymbol{v})$ is convex, so $\boldsymbol{\phi} \mapsto \mathbb{E}\big(f(\boldsymbol{\phi} + uz)\big) = \bar{f}_u(\boldsymbol{\phi})$ is convex. The proof of the Lipschitz property is very similar to that of Proposition 3(b) and is omitted for brevity.

(d) As in the proof of (a), for any $\boldsymbol{v} \in \mathbb{R}^n$ with $\|\boldsymbol{v}\| \leq 1, \boldsymbol{x} \in C_n$ and $\boldsymbol{\alpha} \in A[\boldsymbol{\phi}+u\boldsymbol{v}](\boldsymbol{x})$, we have

$$\left\|\boldsymbol{\alpha} e^{-\operatorname{cef}[\boldsymbol{\phi}+u\boldsymbol{v}](\boldsymbol{x})}\right\| \leq e^{-\phi_0+u}.$$

Since $\nabla_{\boldsymbol{\phi}} \bar{f}_u(\boldsymbol{\phi}) = n^{-1}\mathbf{1} - \Delta\mathbb{E}\big(\boldsymbol{\alpha}^*[\boldsymbol{\phi}+u\boldsymbol{z}](\boldsymbol{\xi}) e^{-\operatorname{cef}[\boldsymbol{\phi}+u\boldsymbol{z}](\boldsymbol{\xi})}\big)$, where $\boldsymbol{\alpha}^*[\boldsymbol{\phi}+u\boldsymbol{v}](\boldsymbol{x}) \in A[\boldsymbol{\phi}+u\boldsymbol{v}](\boldsymbol{x})$, we have by [73, Lemma 8] that $\nabla_{\boldsymbol{\phi}} \bar{f}_u$ is $\Delta e^{-\phi_0+u} n^{1/2}/u$-Lipschitz.

(e) The proof is very similar to the proof of Proposition 3(d) and is omitted for brevity. □

B.3 Proofs of Theorem 1 and Theorem 2

We will make use of the following lemma:

Lemma 1 *[Lemma 4.2 of [25]] Let $\big(\ell_{u_t}(\boldsymbol{\phi})\big)_t$ be a smoothing sequence such that $\boldsymbol{\phi} \mapsto \ell_{u_t}(\boldsymbol{\phi})$ has L_t-Lipschitz gradient. Assume that $\ell_{u_t}(\boldsymbol{\phi}) \leq \ell_{u_{t-1}}(\boldsymbol{\phi})$ for $\boldsymbol{\phi} \in \boldsymbol{\Phi}$. Let $(\boldsymbol{\phi}_t^{(x)})_{t=0}^T, (\boldsymbol{\phi}_t^{(y)})_{t=0}^T, (\boldsymbol{\phi}_t^{(z)})_{t=0}^T$ be the sequences generated by Algorithm 1. Let $\boldsymbol{g}_t$ denote an approximation of $\nabla \ell_{u_t}(\boldsymbol{\phi}_t^{(y)})$ with error $\boldsymbol{e}_t = \boldsymbol{g}_t - \nabla \ell_{u_t}(\boldsymbol{\phi}_t^{(y)})$. Then for any $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ and $t \in \mathbb{N}$, we have*

$$\frac{1}{\theta_t^2}\ell_{u_t}(\boldsymbol{\phi}_{t+1}^{(x)}) \leq \sum_{\tau=0}^{t} \frac{1}{\theta_\tau}\ell_{u_\tau}(\boldsymbol{\phi}) + \frac{1}{2}\left(L_{t+1} + \frac{\eta_{t+1}}{\theta_{t+1}}\right)\|\boldsymbol{\phi} - \boldsymbol{\phi}_0\|^2 + \sum_{\tau=0}^{t} \frac{\|\boldsymbol{e}_\tau\|^2}{2\theta_\tau \eta_\tau} + \sum_{\tau=0}^{t} \frac{1}{\theta_\tau}\langle \boldsymbol{e}_\tau, \boldsymbol{\phi} - \boldsymbol{\phi}_\tau^{(z)}\rangle.$$

Recall the definition of the diameter D of $\boldsymbol{\Phi}$ given just before Theorem 1.

Corollary 2 *Fix $u, \eta > 0$, and assume that Assumption 1 holds with $r \geq u$. Suppose in Algorithm 1 that $u_t = \theta_t u$, $L_t = B_1/u_t$ and $\eta_t = \eta$. Let $(\boldsymbol{\phi}_t^{(x)})_{t=0}^T, (\boldsymbol{\phi}_t^{(y)})_{t=0}^T, (\boldsymbol{\phi}_t^{(z)})_{t=0}^T$ be the sequences generated by Algorithm 1 and let $\boldsymbol{e}_t = \boldsymbol{g}_t - \nabla \ell_{u_t}(\boldsymbol{\phi}_t^{(y)})$. Then for any $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, we have*

$$f(\boldsymbol{\phi}_T^{(x)}) - f(\boldsymbol{\phi}) \leq \frac{B_1 D^2}{Tu} + \frac{\eta D^2}{T} + \frac{1}{T\eta}\sum_{t=0}^{T-1}\|\boldsymbol{e}_t\|^2 + \theta_{T-1}^2 \sum_{t=0}^{T-1} \frac{1}{\theta_t}\langle \boldsymbol{e}_t, \boldsymbol{\phi} - \boldsymbol{\phi}_t^{(z)}\rangle + \frac{4B_0 I(\boldsymbol{\phi})u}{T}.$$

Proof By induction, we have that $\theta_t \leq 2/(t+2)$ and $\sum_{\tau=0}^t 1/\theta_\tau = 1/\theta_t^2$ for all $t = 0, 1, \ldots, T$ [25, 67]. Using Assumption 1, we have

$$\frac{1}{\theta_{T-1}^2}\big(f(\boldsymbol{\phi}_T^{(x)}) - f(\boldsymbol{\phi})\big) \leq \frac{1}{\theta_{T-1}^2}\ell_{u_{T-1}}(\boldsymbol{\phi}_T^{(x)}) - \sum_{t=0}^{T-1}\frac{1}{\theta_t}\big(\ell_{u_t}(\boldsymbol{\phi}) - B_0 u_t\big)$$

$$= \frac{1}{\theta_{T-1}^2}\ell_{u_{T-1}}(\boldsymbol{\phi}_T^{(x)}) - \sum_{t=0}^{T-1}\frac{1}{\theta_t}\ell_{u_t}(\boldsymbol{\phi}) + TB_0 I(\boldsymbol{\phi})u.$$

Hence, by Lemma 1,

$$\begin{aligned}
&f(\boldsymbol{\phi}_T^{(x)}) - f(\boldsymbol{\phi}) \\
&\quad\le \frac{\theta_{T-1}^2}{2}\|\boldsymbol{\phi}-\boldsymbol{\phi}_0\|^2\left(L_T + \frac{\eta_T}{\theta_T}\right) + \sum_{t=0}^{T-1}\frac{\theta_{T-1}^2}{2\theta_t\eta_t}\|\boldsymbol{e}_t\|^2 \\
&\qquad + \theta_{T-1}^2\sum_{t=0}^{T-1}\frac{1}{\theta_t}\langle \boldsymbol{e}_t, \boldsymbol{\phi}-\boldsymbol{\phi}_t^{(z)}\rangle + \theta_{T-1}^2 TB_0 I(\boldsymbol{\phi})u \\
&\quad\le \frac{B_1 D^2\theta_{T-1}^2}{2u\theta_T} + \frac{\eta_T\theta_{T-1}^2 D^2}{2\theta_T} + \sum_{t=0}^{T-1}\frac{\theta_{T-1}^2}{2\theta_t\eta_t}\|\boldsymbol{e}_t\|^2 \\
&\qquad + \theta_{T-1}^2\sum_{t=0}^{T-1}\frac{1}{\theta_t}\langle \boldsymbol{e}_t, \boldsymbol{\phi}-\boldsymbol{\phi}_t^{(z)}\rangle + \theta_{T-1}^2 TB_0 I(\boldsymbol{\phi})u \\
&\quad\le \frac{B_1 D^2}{Tu} + \frac{\eta_T D^2}{T} + \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_t}\|\boldsymbol{e}_t\|^2 + \theta_{T-1}^2\sum_{t=0}^{T-1}\frac{1}{\theta_t}\langle \boldsymbol{e}_t, \boldsymbol{\phi}-\boldsymbol{\phi}_t^{(z)}\rangle \\
&\qquad + \frac{4B_0 I(\boldsymbol{\phi})u}{T},
\end{aligned}$$

where we have used the facts that $\theta_{T-1}^2/\theta_T = \theta_T/(1-\theta_T) \le 2/T$ and $\theta_{T-1}^2/\theta_t \le \theta_{T-1} \le 2/T$ for $t \in \{0, 1, \ldots, T-1\}$ and $\theta_{T-1}^2 \le 4/T^2$. □

Proof (of Theorem 1) According to Corollary 2, it suffices to bound $\theta_{T-1}^2\sum_{t=0}^{T-1}\frac{1}{\theta_t}\mathbb{E}\big(\langle \boldsymbol{e}_t, \boldsymbol{\phi}-\boldsymbol{\phi}_t^{(z)}\rangle\big)$ and $\sum_{t=0}^{T-1}\mathbb{E}\big(\|\boldsymbol{e}_t\|^2\big)$. To this end, we have by Assumption 2 that

$$\begin{aligned}
\mathbb{E}\big(\langle \boldsymbol{e}_t, \boldsymbol{\phi}-\boldsymbol{\phi}_t^{(z)}\rangle\big) &= \mathbb{E}\Big(\mathbb{E}\big(\langle \boldsymbol{e}_t, \boldsymbol{\phi}-\boldsymbol{\phi}_t^{(z)}\rangle \,\big|\, \mathcal{F}_{t-1}\big)\Big) \le \mathbb{E}\Big(\mathbb{E}\big(\|\boldsymbol{e}_t\|\|\boldsymbol{\phi}-\boldsymbol{\phi}_t^{(z)}\| \mid \mathcal{F}_{t-1}\big)\Big) \\
&\le D\cdot\mathbb{E}\Big(\mathbb{E}\big(\|\boldsymbol{e}_t\| \mid \mathcal{F}_{t-1}\big)\Big) \le D\cdot\mathbb{E}\Big(\sqrt{\mathbb{E}\big(\|\boldsymbol{e}_t\|^2 \,\big|\, \mathcal{F}_{t-1}\big)}\Big) \le \frac{D\sigma}{\sqrt{m_t}}.
\end{aligned} \tag{43}$$

We deduce that

$$\theta_{T-1}^2\sum_{t=0}^{T-1}\frac{1}{\theta_t}\mathbb{E}\big(\langle \boldsymbol{e}_t, \boldsymbol{\phi}-\boldsymbol{\phi}_t^{(z)}\rangle\big) \le \frac{2D\sigma M_T^{(1/2)}}{T}. \tag{44}$$

Moreover, by Assumption 2 again,

$$\sum_{t=0}^{T-1} \mathbb{E}\big(\|\boldsymbol{e}_t\|^2\big) \le \sigma^2 \sum_{t=0}^{T-1} \frac{1}{m_t} = \sigma^2 (M_T^{(1)})^2. \tag{45}$$

The bound (18) follows from Corollary 2, together with (44) and (45), and the bound (19) then follows directly from the parameter choice of u and η and the fact that $I(\boldsymbol{\phi}^*) = 1$.

Finally, if $\mathbb{E}(\boldsymbol{e}_t \mid \mathcal{F}_{t-1}) = \mathbf{0}$, then

$$\mathbb{E}\big(\langle \boldsymbol{e}_t, \boldsymbol{\phi} - \boldsymbol{\phi}_t^{(z)} \rangle\big) = \mathbb{E}\big(\mathbb{E}(\langle \boldsymbol{e}_t, \boldsymbol{\phi} - \boldsymbol{\phi}_t^{(z)} \rangle \mid \mathcal{F}_{t-1})\big) = \mathbb{E}\big(\langle \mathbb{E}(\boldsymbol{e}_t | \mathcal{F}_{t-1}), \boldsymbol{\phi} - \boldsymbol{\phi}_t^{(z)} \rangle\big) = 0$$

where the second equality uses the fact that $\boldsymbol{\phi} - \boldsymbol{\phi}_t^{(z)}$ is $\mathcal{F}_{t-1}$-measurable. This allows us to remove the last term of the two inequalities in the theorem. □

Proof (of Theorem 2) According to Corollary 2, it suffices to obtain a high-probability bound for $\theta_{T-1}^2 \sum_{t=0}^{T-1} \frac{1}{\theta_t} \langle \boldsymbol{e}_t, \boldsymbol{\phi} - \boldsymbol{\phi}_t^{(z)} \rangle$ and $\sum_{t=0}^{T-1} \|\boldsymbol{e}_t\|^2$. Writing $\boldsymbol{\zeta}_t := \boldsymbol{\phi} - \boldsymbol{\phi}_t^{(z)}$, we have from the proof of Theorem 1 that $(1/\theta_t)\langle \boldsymbol{e}_t, \boldsymbol{\zeta}_t \rangle$ is a martingale difference sequence under Assumption 3. Note that $\boldsymbol{\zeta}_t$ is $\mathcal{F}_{t-1}$-measurable, and we will now show that $\langle \boldsymbol{e}_t, \boldsymbol{\zeta}_t \rangle$ is $\sqrt{2}\sigma_t D$ sub-Gaussian, conditional on $\mathcal{F}_{t-1}$.

For any $x \in \mathbb{R}$, we have $e^x \le x + e^{x^2}$. Hence, for $\lambda \in \mathbb{R}$ such that $\lambda^2 \sigma_t^2 D^2 \le 1$, we have by the conditional version of Jensen's inequality that

$$\begin{aligned}\mathbb{E}\big(e^{\lambda \langle \boldsymbol{e}_t, \boldsymbol{\zeta}_t \rangle} \mid \mathcal{F}_{t-1}\big) &\le \mathbb{E}\big(\lambda \langle \boldsymbol{e}_t, \boldsymbol{\zeta}_t \rangle \mid \mathcal{F}_{t-1}\big) + \mathbb{E}\big(e^{\lambda^2 \langle \boldsymbol{e}_t, \boldsymbol{\zeta}_t \rangle^2} \mid \mathcal{F}_{t-1}\big) \\ &\le \mathbb{E}\big(e^{\lambda^2 \|\boldsymbol{e}_t\|^2 D^2} | \mathcal{F}_{t-1}\big) \le e^{\lambda^2 \sigma_t^2 D^2}.\end{aligned}$$

On the other hand, if $\lambda^2 \sigma_t^2 D^2 > 1$, then since $2ab \le a^2 + b^2$ for all $a, b \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{E}\big(e^{\lambda \langle \boldsymbol{e}_t, \boldsymbol{\zeta}_t \rangle} \mid \mathcal{F}_{t-1}\big) &\le e^{\lambda^2 \sigma_t^2 D^2/2} \mathbb{E}\big(e^{\langle \boldsymbol{e}_t, \boldsymbol{\zeta}_t \rangle^2/(2\sigma_t^2 D^2)} \mid \mathcal{F}_{t-1}\big) \\ &\le e^{\lambda^2 \sigma_t^2 D^2/2} \mathbb{E}\big(e^{\|\boldsymbol{e}_t\|^2/(2\sigma_t^2)} \mid \mathcal{F}_{t-1}\big) \le e^{\lambda^2 \sigma_t^2 D^2/2} e^{1/2} \le e^{\lambda^2 \sigma_t^2 D^2}.\end{aligned}$$

We deduce that $\langle \boldsymbol{e}_t, \boldsymbol{\zeta}_t \rangle / \theta_t$ is $(\sqrt{2}\sigma_t D)/\theta_t$ sub-Gaussian, conditional on $\mathcal{F}_{t-1}$. Applying the Azuma–Hoeffding inequality (e.g. [3]) therefore yields that for every $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\bigg(\theta_{T-1}^2 \sum_{t=0}^{T-1} \frac{1}{\theta_t} \langle \boldsymbol{e}_t, \boldsymbol{\zeta}_t \rangle \ge \epsilon\bigg) &\le \exp\bigg(-\frac{\epsilon^2}{4D^2 \theta_{T-1}^2 \sum_{t=0}^{T-1} \sigma_t^2 \theta_{T-1}^2/\theta_t^2}\bigg) \\ &\le \exp\bigg(-\frac{T^2 \epsilon^2}{16 D^2 \sigma^2 (M_T^{(1)})^2}\bigg),\end{aligned}$$

where the last inequality uses the facts that $\theta_{T-1} \le \theta_t$ and $\theta_{T-1} \le 2/T$. Therefore, for every $\delta \in (0, 1)$, we have with probability at least $1 - \delta/2$ that

$$\theta_{T-1}^2 \sum_{t=0}^{T-1} \frac{1}{\theta_t} \langle \boldsymbol{e}_t, \boldsymbol{\zeta}_t \rangle \le \frac{4\sigma D M_T^{(1)} \sqrt{\log(2/\delta)}}{T}. \tag{46}$$

Next we will turn to finding a tail bound for $\sum_{t=0}^{T-1} \|\boldsymbol{e}_t\|^2$. By Assumption 3 and Jensen's inequality, we have

$$\sum_{t=0}^{T-1} \mathbb{E}\big(\|\boldsymbol{e}_t\|^2 \,\big|\, \mathcal{F}_{t-1}\big) \le \sum_{t=0}^{T-1} \sigma_t^2 \log\big(\mathbb{E}(e^{\|\boldsymbol{e}_t\|^2/\sigma_t^2} | \mathcal{F}_{t-1})\big) \le \sigma^2 (M_T^{(1)})^2. \tag{47}$$

Now define the random variables $\Xi_t := \|\boldsymbol{e}_t\|^2 - \mathbb{E}\big(\|\boldsymbol{e}_t\|^2 \,\big|\, \mathcal{F}_{t-1}\big)$. Then by Markov's inequality, for every $\epsilon > 0$,

$$\mathbb{P}\big(\Xi_t > \epsilon \,\big|\, \mathcal{F}_{t-1}\big) \le \mathbb{P}\big(\|\boldsymbol{e}_t\|^2/\sigma_t^2 > \epsilon/\sigma_t^2 \,\big|\, \mathcal{F}_{t-1}\big) \le e^{-\epsilon/\sigma_t^2} \mathbb{E}\big(e^{\|\boldsymbol{e}_t\|^2/\sigma_t^2} \,\big|\, \mathcal{F}_{t-1}\big) \le e^{1-\epsilon/\sigma_t^2}.$$

Moreover, by Markov's inequality again, and then Jensen's inequality, we have for every $\epsilon > 0$ that

$$\begin{aligned}\mathbb{P}\big(\Xi_t < -\epsilon \,\big|\, \mathcal{F}_{t-1}\big) &\le e^{-\epsilon/\sigma_t^2} \mathbb{E}\big(e^{\mathbb{E}(\|\boldsymbol{e}_t\|^2/\sigma_t^2 | \mathcal{F}_{t-1}) - \|\boldsymbol{e}_t\|^2/\sigma_t^2} \,\big|\, \mathcal{F}_{t-1}\big) \\ &\le e^{-\epsilon/\sigma_t^2} e^{\mathbb{E}(\|\boldsymbol{e}_t\|^2/\sigma_t^2 | \mathcal{F}_{t-1})} \le e^{1-\epsilon/\sigma_t^2}.\end{aligned}$$

It follows by, e.g., [25, Lemma F.7] that Ξ_t is sub-exponential with parameters $\lambda_t := 1/(2\sigma_t^2) = m_t/(2\sigma^2)$ and $\tau_t^2 := 16e\sigma_t^4 = 16e\sigma^4/m_t^2$, in the sense that

$$\mathbb{E}\big(e^{\lambda \Xi_t} \,\big|\, \mathcal{F}_{t-1}\big) \le e^{8e\lambda^2\sigma_t^4}, \tag{48}$$

for $|\lambda| \le 1/(2\sigma_t^2)$.

Now define $\Lambda_T := \min_{t=0,\dots,T-1} \lambda_t = m_0/(2\sigma^2)$ (as we assume (m_t) is increasing) and $C_T := \big(\sum_{t=0}^{T-1} \tau_t^2\big)^{1/2} = 4e^{1/2}\sigma^2 M_T^{(2)}$. We claim that $\sum_{t=0}^{T-1} \Xi_t$ is sub-exponential with parameters Λ_T and C_T, and prove this by induction on T. The base case $T = 1$ holds by (48), so suppose it holds for a given $T \in \mathbb{N}$. Then for $\lambda \in \mathbb{R}$ with $|\lambda| \le \min(\Lambda_T, \lambda_T) = \Lambda_{T+1}$, we have

$$\begin{aligned}\mathbb{E}\left\{\exp\left(\lambda \sum_{t=0}^{T} \Xi_t\right)\right\} &= \mathbb{E}\left[\exp\left(\lambda \sum_{t=0}^{T-1} \Xi_t\right) \mathbb{E}\big\{\exp\big(\lambda \Xi_T \,\big|\, \mathcal{F}_{T-1}\big)\big\}\right] \\ &\le e^{(\lambda^2 C_T^2 + 16e\lambda^2\sigma_T^4)/2} = e^{\lambda^2 C_{T+1}^2/2},\end{aligned}$$

which proves the claim by induction. We deduce by, e.g. [10, Lemma 1.4.1], that for every $\epsilon > 0$ and $T \in \mathbb{N}$,

$$\mathbb{P}\Big(\sum_{t=0}^{T-1} \Xi_t \geq \epsilon\Big) \leq \exp\Big(-\min\Big\{\frac{\epsilon^2}{2C_T^2}, \frac{\Lambda_T \epsilon}{2}\Big\}\Big).$$

In other words, with probability at least $1 - \delta/2$,

$$\sum_{t=0}^{T-1} \Xi_t \leq 4\sigma^2 \max\Big\{M_T^{(2)}\sqrt{2e\log\frac{2}{\delta}}, \frac{1}{m_0}\log\frac{2}{\delta}\Big\}. \tag{49}$$

Applying (46), (47) and (49) in Corollary 2, together with a union bound, yields that with probability at least $1 - \delta$,

$$\begin{aligned} &f(\boldsymbol{\phi}_T^{(x)}) - f(\boldsymbol{\phi}) \\ &\leq \frac{B_1 D^2}{Tu} + \frac{4B_0 I(\boldsymbol{\phi})u}{T} + \frac{\eta D^2}{T} + \frac{\sigma^2 (M_T^{(1)})^2}{T\eta} + \frac{4\sigma D M_T^{(1)}\sqrt{\log(2/\delta)}}{T} \\ &\quad + \frac{4\sigma^2 \max\big\{M_T^{(2)}\sqrt{2e\log(2/\delta)}, m_0^{-1}\log(2/\delta)\big\}}{T\eta}. \end{aligned}$$

Taking the same choices of $\boldsymbol{\phi}$, u and η as in Theorem 1, we obtain the final result. □

B.4 Proof of Theorem 3

To prove Theorem 3, we first introduce the following lemma. Recall that $\mathcal{C}_d$ denotes the class of proper, convex lower-semicontinuous functions $\varphi : \mathbb{R}^d \to (-\infty, \infty]$ that are coercive in the sense that $\varphi(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$. Recall further from [28, Theorem 2.2] that if P is a distribution on $\mathbb{R}^d$ with $\int_{\mathbb{R}^d} \|\boldsymbol{x}\|\, dP(\boldsymbol{x}) < \infty$ and $P(H) < 1$ for all hyperplanes H, then the strictly convex function $\Gamma : \mathcal{C}_d \to (-\infty, \infty]$ given by

$$\Gamma(\varphi) := \int_{\mathbb{R}^d} \varphi(\boldsymbol{x})\, dP(\boldsymbol{x}) + \int_{\mathbb{R}^d} e^{-\varphi(\boldsymbol{x})}\, d\boldsymbol{x} \tag{50}$$

has a unique minimizer $\varphi^* \in \mathcal{C}_d$ satisfying $\Gamma(\varphi^*) \in \mathbb{R}$.

Lemma 2 *Let P be a distribution on $\mathbb{R}^d$ with $\int_{\mathbb{R}^d} \|x\|\, dP(x) < \infty$ and $P(H) < 1$ for all hyperplanes H, and let $\varphi^* := \operatorname{argmin}_{\varphi \in \mathcal{C}_d} \Gamma(\varphi)$. Then*

(1) *For any $\lambda \in [0, 1]$, and $\varphi, \tilde{\varphi} \in \mathcal{C}_d$, we have*

$$\begin{aligned} &\Gamma\big(\lambda\varphi + (1-\lambda)\tilde{\varphi}\big) \\ &\leq \lambda\Gamma(\varphi) + (1-\lambda)\Gamma(\tilde{\varphi}) - \frac{\lambda(1-\lambda)}{2}\int_{\mathbb{R}^d} e^{-\max\{\varphi(\boldsymbol{x}), \tilde{\varphi}(\boldsymbol{x})\}}\big\{\varphi(\boldsymbol{x}) - \tilde{\varphi}(\boldsymbol{x})\big\}^2\, d\boldsymbol{x}. \end{aligned}$$

(51)

Here, when $\max\{\varphi(\boldsymbol{x}), \tilde{\varphi}(\boldsymbol{x})\} = \infty$, *we define the integrand to be zero.*

(2) *Furthermore, if* $\varphi \in \mathcal{C}_d$ *is such that* $\max\{\varphi(\boldsymbol{x}), \varphi^*(\boldsymbol{x})\} \leq \phi^0$ *for all* $\boldsymbol{x} \in \mathrm{dom}(\varphi) \cap \mathrm{dom}(\varphi^*)$, *then*

$$\Gamma(\varphi) - \Gamma(\varphi^*) \geq \frac{1}{2} e^{-\phi^0} \int_{\mathrm{dom}(\varphi)\cap\mathrm{dom}(\varphi^*)} \{\varphi(\boldsymbol{x}) - \varphi^*(\boldsymbol{x})\}^2 \, d\boldsymbol{x}. \tag{52}$$

Proof **(1)** Fix $\varphi, \tilde{\varphi} \in \mathcal{C}_d$ with $\max\{\Gamma(\varphi), \Gamma(\tilde{\varphi})\} < \infty$ (because otherwise the result is clear). For any $M \in \mathbb{R}$, the function $y \mapsto e^{-y}$ is e^{-M}-strongly convex on $y \leq M$. Therefore, for any $\lambda \in [0, 1]$, we have $\Gamma\big(\lambda\varphi + (1-\lambda)\tilde{\varphi}\big) \geq \Gamma(\varphi^*) > -\infty$, so

$$\begin{aligned}
&\Gamma\big(\lambda\varphi + (1-\lambda)\tilde{\varphi}\big) - \lambda\Gamma(\varphi) - (1-\lambda)\Gamma(\tilde{\varphi}) \\
&= \int_{\mathbb{R}^d} \big[e^{-\{\lambda\varphi(\boldsymbol{x}) + (1-\lambda)\tilde{\varphi}(\boldsymbol{x})\}} - \lambda e^{-\varphi(\boldsymbol{x})} - (1-\lambda) e^{-\tilde{\varphi}(\boldsymbol{x})}\big] \, d\boldsymbol{x} \\
&\leq -\frac{\lambda(1-\lambda)}{2} \int_{\mathbb{R}^d} e^{-\max\{\varphi(\boldsymbol{x}), \tilde{\varphi}(\boldsymbol{x})\}} \{\varphi(\boldsymbol{x}) - \varphi^*(\boldsymbol{x})\}^2 \, d\boldsymbol{x},
\end{aligned} \tag{53}$$

as required.

(2) By (53), we have for any $\lambda \in (0, 1)$ that

$$\begin{aligned}
\frac{\lambda(1-\lambda)}{2} e^{-\phi^0} \int_{\mathrm{dom}(\varphi)\cap\mathrm{dom}(\varphi^*)} &\{\varphi(\boldsymbol{x}) - \varphi^*(\boldsymbol{x})\}^2 \, d\boldsymbol{x} \\
&\leq \lambda\Gamma(\varphi) + (1-\lambda)\Gamma(\varphi^*) - \Gamma\big(\lambda\varphi + (1-\lambda)\varphi^*\big) \\
&\leq \lambda\{\Gamma(\varphi) - \Gamma(\varphi^*)\},
\end{aligned}$$

where the last inequality follows by definition of φ^*. We deduce that

$$\frac{1}{2} e^{-\phi^0} \int_{\mathrm{dom}(\varphi)\cap\mathrm{dom}(\varphi^*)} \{\varphi(\boldsymbol{x}) - \varphi^*(\boldsymbol{x})\}^2 \, d\boldsymbol{x} \leq \frac{\Gamma(\varphi) - \Gamma(\varphi^*)}{1-\lambda}.$$

The result follows on taking $\lambda \searrow 0$. □

Proof (of Theorem 3) In Lemma 2, let P be the empirical distribution of $\{\boldsymbol{x}_i\}_{i=1}^n$. From the proof of Theorem 2 of [20] (see also the proof of Theorem 4), Problem (50) is equivalent to (5) in the sense that $\Gamma(\mathrm{cef}[\boldsymbol{\phi}]) = f(\boldsymbol{\phi})$ for all $\boldsymbol{\phi} \in \mathbb{R}^n$, and $\varphi^* = \mathrm{cef}[\boldsymbol{\phi}^*]$. Now fix $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ and let $\varphi = \mathrm{cef}[\boldsymbol{\phi}]$, so that $\varphi(\boldsymbol{x}) \leq \phi^0$ for all $\boldsymbol{x} \in C_n$. From (52), we obtain

$$\frac{1}{2} e^{-\phi^0} \int_{C_n} \{\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x}) - \mathrm{cef}[\boldsymbol{\phi}^*](\boldsymbol{x})\}^2 \, d\boldsymbol{x} \leq \Gamma(\mathrm{cef}[\boldsymbol{\phi}]) - \Gamma(\mathrm{cef}[\boldsymbol{\phi}^*]) = f(\boldsymbol{\phi}) - f(\boldsymbol{\phi}^*),$$

as required. □

B.5 Proofs of Theorem 4 and Theorem 5

The proof of Theorem 4 is based on following proposition:

Proposition 7 *Given $s \in \mathbb{R}$, $\varphi \in \mathcal{C}_d$ and $c > 0$, there exists $\tilde{\varphi} \in \mathcal{C}_d$, such that $\psi_s \circ \varphi(\cdot) = c \cdot \psi_s \circ \tilde{\varphi}(\cdot)$.*

Proof (of Proposition 7) Given s, φ and c, define

$$\tilde{\varphi}(\cdot) := \begin{cases} \varphi(\cdot) + \log c & \text{if } s = 0 \\ c^{-s}\varphi(\cdot) & \text{if } s \neq 0. \end{cases}$$

Then $\tilde{\varphi} \in \mathcal{C}_d$, and for $\boldsymbol{x} \in \mathcal{D}_s$,

$$\psi_s \circ \tilde{\varphi}(\boldsymbol{x}) = \begin{cases} \big(c^{-s}\varphi(\boldsymbol{x})\big)^{1/s} & \text{if } s < 0 \\ \exp\big(-\varphi(\boldsymbol{x}) - \log c\big) & \text{if } s = 0 \\ \big(-c^{-s}\varphi(\boldsymbol{x})\big)^{1/s} & \text{if } s > 0 \end{cases} = c^{-1} \cdot \psi_s \circ \varphi(\boldsymbol{x}),$$

as required. □

Proof (of Theorem 4) The proof is split into four steps. The first three steps hold for any $s \in \mathbb{R}$, while in Step 4, we show the convexity of the objective when $s \in [0, 1]$.

Step 1: We claim that any solution $\hat{p}_n$ to (21) is supported on C_n, so that $\varphi^*(\boldsymbol{x}) = \infty$ when $\boldsymbol{x} \notin C_n$, where φ^* is the solution to (22). Indeed, suppose for a contradiction that $p = \psi_s \circ \varphi \in \mathcal{P}_s(\mathbb{R}^d)$ is such that $\sum_{i=1}^n \log p(\boldsymbol{x}_i) > -\infty$, and that $\int_{C_n} p(\boldsymbol{x})\, d\boldsymbol{x} = c < 1 = \int_{\mathbb{R}^d} p(\boldsymbol{x})\, d\boldsymbol{x}$. We may assume that $c > 0$, because otherwise $p(\boldsymbol{x}) = 0$ for almost all $\boldsymbol{x} \in C_n$, which would mean that $\sum_{i=1}^n \log p(\boldsymbol{x}_i) = -\infty$ since $p \in \mathcal{P}_s(\mathbb{R}^d)$. Define $\bar{\varphi} \in \mathcal{C}_d$ by

$$\bar{\varphi}(\boldsymbol{x}) := \begin{cases} \varphi(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in C_n \\ \infty & \text{otherwise,} \end{cases}$$

so that $\int_{\mathbb{R}^d} \psi_s \circ \bar{\varphi}(\boldsymbol{x})\, d\boldsymbol{x} = \int_{\mathbb{R}^d} \psi_s \circ \varphi(\boldsymbol{x})\, d\boldsymbol{x} = c < 1$. Applying Proposition 7 to $\bar{\varphi} \in \mathcal{C}_d$ and $c > 0$, we can find $\tilde{\varphi} \in \mathcal{C}_d$ with $\int_{\mathbb{R}^d} \psi_s \circ \tilde{\varphi}(\boldsymbol{x})\, d\boldsymbol{x} = \int_{\mathbb{R}^d} c^{-1} \cdot \psi_s \circ \bar{\varphi}(\boldsymbol{x})\, d\boldsymbol{x} = 1$, and

$$\sum_{i=1}^n \log \psi_s \circ \tilde{\varphi}(\boldsymbol{x}_i) = \sum_{i=1}^n \log \psi_s \circ \varphi(\boldsymbol{x}_i) - n \log c > \sum_{i=1}^n \log \psi_s \circ \varphi(\boldsymbol{x}_i).$$

This establishes our desired contradiction.

Step 2: We claim that any solution φ^* to (22) satisfies

$$\varphi^* = \operatorname*{argmin}_{\substack{\varphi \in \mathcal{C}_d : \operatorname{Im}(\varphi) \subseteq \mathcal{D}_s \cup \{\infty\}, \\ \operatorname{dom}(\varphi) = C_n}} \left\{ -\frac{1}{n} \sum_{i=1}^n \log \psi_s \circ \varphi(\boldsymbol{x}_i) + \int_{C_n} \psi_s \circ \varphi(\boldsymbol{x})\, d\boldsymbol{x} \right\}. \quad (54)$$

Indeed, for any $\varphi \in \mathcal{C}_d$ such that $\mathrm{dom}(\varphi) = C_n$ and $\int_{C_n} \psi_s \circ \varphi(\boldsymbol{x})\, d\boldsymbol{x} = c \neq 1$, we can again apply Proposition 7 to φ and c to obtain $\tilde{\varphi}$. Then

$$-\frac{1}{n}\sum_{i=1}^{n} \log \psi_s \circ \varphi(\boldsymbol{x}_i) + \int_{C_n} \psi_s \circ \varphi(\boldsymbol{x})\, d\boldsymbol{x} = -\frac{1}{n}\sum_{i=1}^{n} \log \psi_s \circ \varphi(\boldsymbol{x}_i) + c$$
$$> -\frac{1}{n}\sum_{i=1}^{n} \log \psi_s \circ \varphi(\boldsymbol{x}_i) + \log c + 1 = -\frac{1}{n}\sum_{i=1}^{n} \log \psi_s \circ \tilde{\varphi}(\boldsymbol{x}_i) + \int_{C_n} \psi_s \circ \tilde{\varphi}(\boldsymbol{x})\, d\boldsymbol{x},$$

so $\int_{C_n} \psi_s \circ \varphi^*(\boldsymbol{x})\, d\boldsymbol{x} = 1$, which establishes our claim.

Step 3: Letting $\boldsymbol{\phi}^* = (\phi_1^*, \ldots, \phi_n^*)$ denote an optimal solution to (23), we claim that $\mathrm{cef}[\boldsymbol{\phi}^*](\boldsymbol{x}_i) = \phi_i^*$ holds for all $i \in [n]$. Indeed, for any $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n) \in \mathcal{D}_s^n$, if there exists $i^* \in [n]$ such that $\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x}_{i^*}) \neq \phi_{i^*}$, then we can define $\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_1, \ldots, \tilde{\phi}_n) \in \mathcal{D}_s^n$ such that $\tilde{\phi}_i = \mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x}_i)$ for all $i \in [n]$. We now claim that $\mathrm{cef}[\tilde{\boldsymbol{\phi}}] = \mathrm{cef}[\boldsymbol{\phi}]$. On the one hand, by (3), $\tilde{\phi}_i = \mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x}_i) \leq \phi_i$ for any $i \in [n]$. From the LP expression (Q_0), $\mathrm{cef}[\tilde{\boldsymbol{\phi}}] \leq \mathrm{cef}[\boldsymbol{\phi}]$. On the other hand, since $\mathrm{cef}[\boldsymbol{\phi}](\cdot)$ is a convex function with $\mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x}_i) = \tilde{\phi}_i \leq \tilde{\phi}_i$ for any i, we have $\mathrm{cef}[\tilde{\boldsymbol{\phi}}](\cdot) \geq \mathrm{cef}[\boldsymbol{\phi}](\cdot)$. It follows that $\mathrm{cef}[\tilde{\boldsymbol{\phi}}] = \mathrm{cef}[\boldsymbol{\phi}]$, and $\tilde{\boldsymbol{\phi}}$ with a smaller objective than $\boldsymbol{\phi}$. This establishes our claim, and shows that (23) is equivalent to (54) in the sense that $\hat{p}_n$ and $\boldsymbol{\phi}^*$ satisfy $\hat{p}_n = \psi_s(\mathrm{cef}[\boldsymbol{\phi}^*])$.

Step 4: When $s = 0$, the function $\boldsymbol{\phi} \mapsto \frac{1}{n}\mathbf{1}^\top \boldsymbol{\phi}$ is convex on $\mathbb{R}^n$; when $s > 0$, the function $\boldsymbol{\phi} \mapsto -\frac{1}{ns}\sum_{i=1}^n \log(-\phi_i)$ is convex on $(-\infty, 0]^n$. Moreover, when $s \leq 1$, the function ψ_s is decreasing and convex, and since $\boldsymbol{\phi} \mapsto \mathrm{cef}[\boldsymbol{\phi}](\boldsymbol{x})$ is concave for every $\boldsymbol{x} \in \mathbb{R}^n$, the result follows. □

C Background on shape-constrained inference

Entry points to the field of nonparametric inference under shape constraints include the book by [34], as well as the 2018 special issue of the journal *Statistical Science* [61]. Other canonical problems in shape constraints that involve non-trivial computational issues include isotonic regression ([6, 9, 13, 30, 38, 55, 72, 74]) and convex regression ([11, 14, 31, 35, 40, 41, 64]), or combinations and variants of these ([16]).

Beyond papers already discussed, early theoretical work on log-concave density estimation includes [54], [26], [69], [19], [28], [63], [62] and [15]. Sometimes, the class $\mathcal{P}_d$ is considered as a special case of the class of s-concave densities ([24, 37, 39, 46, 65]); see also Sect. 5. Much recent work has focused on rates of convergence, which are best understood in the Hellinger distance d_{H}, given by

$$d_{\mathrm{H}}^2(p, q) := \int_{\mathbb{R}^d} (p^{1/2} - q^{1/2})^2.$$

For the case of correct model specification, i.e. where $\hat{p}_n$ is computed from an independent and identically distributed sample of size n from $p_0 \in \mathcal{P}_d$, it is now known

[44, 47] that

$$\sup_{p_0 \in \mathcal{P}_d} \mathbb{E} d_{\mathrm{H}}^2(\hat{p}_n, p_0) \leq K_d \cdot \begin{cases} n^{-4/5} & \text{when } d = 1 \\ n^{-2/(d+1)} \log n & \text{when } d \geq 2, \end{cases}$$

where $K_d > 0$ depends only on d, and that this risk bound is minimax optimal (up to the logarithmic factor when $d \geq 2$). See also [12] for an earlier result in the case $d \geq 4$, and [71] for an alternative approach to high-dimensional log-concave density estimation that seeks to evade the curse of dimensionality in the additional presence of symmetry constraints. It is further known that when $d \leq 3$, the log-concave maximum likelihood estimator can adapt to certain subclasses of log-concave densities, including log-concave densities whose logarithms are piecewise affine [32, 43]. See also [5] for recent work on extensions to the misspecified setting (where the true distribution from which the data are drawn does not have a log-concave density).

References

1. Agarwal, A., Bartlett, P.L., Ravikumar, P., Wainwright, M.J.: Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. IEEE Trans. Inf. Theory **5**(58), 3235–3249 (2012)
2. Axelrod, B., Diakonikolas, I., Stewart, A., Sidiropoulos, A., Valiant, G.: A polynomial time algorithm for log-concave maximum likelihood via locally exponential families. In: Advances in Neural Information Processing Systems, pp. 7723–7735 (2019)
3. Azuma, K.: Weighted sums of certain dependent random variables. Tohoku Math. J. Second Ser. **19**(3), 357–367 (1967)
4. Barber, C.B., Dobkin, D.P., Huhdanpaa, H.: The quickhull algorithm for convex hull. Tech. rep., Technical Report GCG53, Geometry Center, University of Minnesota (1993)
5. Barber, R.F., Samworth, R.J.: Local continuity of log-concave projection, with applications to estimation under model misspecification. Bernoulli, to appear (2021)
6. Bellec, P.C.: Sharp oracle inequalities for least squares estimators in shape restricted regression. Ann. Stat. **46**(2), 745–780 (2018)
7. Bertsekas, D.: Nonlinear Programming. Athena Scientific Optimization and Computation Series. Athena Scientific (2016). https://books.google.com/books?id=TwOujgEACAAJ
8. Bertsekas, D.P.: Convex Optimization Theory. Athena Scientific Belmont (2009)
9. Brunk, H., Barlow, R.E., Bartholomew, D.J., Bremner, J.M.: Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression. John Wiley & Sons (1972)
10. Buldygin, V.V., Kozachenko, Y.V.: Metric Characterization of Random Variables and Random Processes, vol. 188. American Mathematical Society (2000)
11. Cai, T.T., Low, M.G.: A framework for estimation of convex functions. Stat. Sin. **25**, 423–456 (2015)
12. Carpenter, T., Diakonikolas, I., Sidiropoulos, A., Stewart, A.: Near-optimal sample complexity bounds for maximum likelihood estimation of multivariate log-concave densities. In: Conference On Learning Theory, pp. 1234–1262 (2018)
13. Chatterjee, S., Guntuboyina, A., Sen, B.: On risk bounds in isotonic and other shape restricted regression problems. Ann. Stat. **43**(4), 1774–1800 (2015)
14. Chen, W., Mazumder, R.: Multivariate convex regression at scale. arXiv preprint arXiv:2005.11588 (2020)
15. Chen, Y., Samworth, R.J.: Smoothed log-concave maximum likelihood estimation with applications. Stat. Sin. **23**, 1373–1398 (2013)
16. Chen, Y., Samworth, R.J.: Generalized additive and index models with shape constraints. J. R. Stat. Soc. B **78**, 729–754 (2016)
17. Clarke, F.H.: Optimization and Nonsmooth Analysis, vol. 5. SIAM, Philadelphia (1990)

18. Cule, M., Gramacy, R.B., Samworth, R.: LogConcDEAD: an R package for maximum likelihood estimation of a multivariate log-concave density. J. Stat. Softw. **29**(2), 1–20 (2009). https://doi.org/10.18637/jss.v029.i02. (https://www.jstatsoft.org/v029/i02)
19. Cule, M., Samworth, R.: Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. Electron. J. Stat. **4**, 254–270 (2010)
20. Cule, M., Samworth, R., Stewart, M.: Maximum likelihood estimation of a multi-dimensional log-concave density. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **72**(5), 545–607 (2010)
21. Cule, M.L., Dümbgen, L.: On an auxiliary function for log-density estimation. arXiv preprint arXiv:0807.4719 (2008)
22. Dalalyan, A.S.: Theoretical guarantees for approximate sampling from smooth and log-concave densities. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **79**(3), 651–676 (2017)
23. Dharmadhikari, S., Joag-Dev, K.: Unimodality, Convexity, and Applications. Elsevier (1988)
24. Doss, C.R., Wellner, J.A.: Global rates of convergence of the MLEs of log-concave and s-concave densities. Ann. Stat. **44**(3), 954 (2016)
25. Duchi, J.C., Bartlett, P.L., Wainwright, M.J.: Randomized smoothing for stochastic optimization. SIAM J. Optim. **22**(2), 674–701 (2012)
26. Dümbgen, L., Rufibach, K.: Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. Bernoulli **15**(1), 40–68 (2009)
27. Dümbgen, L., Rufibach, K.: logcondens: computations related to univariate log-concave density estimation. J. Stat. Softw. **39**, 1–28 (2011)
28. Dümbgen, L., Samworth, R., Schuhmacher, D.: Approximation by log-concave distributions, with applications to regression. Ann. Stat. **39**(2), 702–730 (2011)
29. Dümbgen, L., Samworth, R.J., Wellner, J.A.: Bounding distributional errors via density ratios. Bernoulli **27**, 818–852 (2021)
30. Durot, C., Lopuhaä, H.P.: Limit theory in monotone function estimation. Stat. Sci. **33**(4), 547–567 (2018)
31. Fang, B., Guntuboyina, A.: On the risk of convex-constrained least squares estimators under misspecification. Bernoulli **25**(3), 2206–2244 (2019)
32. Feng, O.Y., Guntuboyina, A., Kim, A.K., Samworth, R.J.: Adaptation in multivariate log-concave density estimation. Ann. Stat. **49**, 129–153 (2021)
33. Grenander, U.: On the theory of mortality measurement: part ii. Scand. Actuar. J. **1956**(2), 125–153 (1956)
34. Groeneboom, P., Jongbloed, G.: Nonparametric Estimation under Shape Constraints, vol. 38. Cambridge University Press, Cambridge (2014)
35. Guntuboyina, A., Sen, B.: Global risk bounds and adaptation in univariate convex regression. Probab. Theory Relat. Fields **163**(1–2), 379–411 (2015)
36. Gurobi Optimization, LLC: Gurobi optimizer reference manual (2021). http://www.gurobi.com
37. Han, Q.: Global empirical risk minimizers with "shape constraints" are rate optimal in general dimensions. The Annals of Statistics, to appear (2021)
38. Han, Q., Wang, T., Chatterjee, S., Samworth, R.J.: Isotonic regression in general dimensions. Ann. Stat. **47**(5), 2440–2471 (2019)
39. Han, Q., Wellner, J.A.: Approximation and estimation of s-concave densities via rényi divergences. Ann. Stat. **44**(3), 1332 (2016)
40. Han, Q., Wellner, J.A.: Multivariate convex regression: global risk bounds and adaptation. arXiv preprint arXiv:1601.06844 (2016)
41. Hildreth, C.: Point estimates of ordinates of concave functions. J. Am. Stat. Assoc. **49**(267), 598–619 (1954)
42. Kappel, F., Kuntsevich, A.V.: An implementation of Shor's r-algorithm. Comput. Optim. Appl. **15**(2), 193–205 (2000)
43. Kim, A.K., Guntuboyina, A., Samworth, R.J.: Adaptation in log-concave density estimation. Ann. Stat. **46**(5), 2279–2306 (2018)
44. Kim, A.K., Samworth, R.J.: Global rates of convergence in log-concave density estimation. Ann. Stat. **44**(6), 2756–2779 (2016)
45. Klenke, A.: Probability Theory: A Comprehensive Course. Springer Science & Business Media (2014)
46. Koenker, R., Mizera, I.: Quasi-concave density estimation. Ann. Stat. **38**(5), 2998–3027 (2010)
47. Kur, G., Dagan, Y., Rakhlin, A.: The log-concave maximum likelihood estimator is optimal in high dimensions. arXiv preprint arXiv:1903.05315v3 (2019)

48. Lakshmanan, H., De Farias, D.P.: Decentralized resource allocation in dynamic networks of agents. SIAM J. Optim. **19**(2), 911–940 (2008)
49. Liu, Y., Wang, Y.: A fast algorithm for univariate log-concave density estimation. Austr. N. Z. J. Stat. **60**(2), 258–275 (2018)
50. McMullen, P.: The maximum numbers of faces of a convex polytope. Mathematika **17**(2), 179–184 (1970)
51. Nemirovsky, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley (1983)
52. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
53. Nesterov, Y.: Primal-dual subgradient methods for convex problems. Math. Program. **120**(1), 221–259 (2009)
54. Pal, J.K., Woodroofe, M., Meyer, M.: Estimating a Polya frequency function. Lecture Notes-Monograph Series pp. 239–249 (2007)
55. Pananjady, A., Samworth, R.J.: Isotonic regression with unknown permutations: Statistics, computation, and adaptation. arXiv preprint arXiv:2009.02609 (2020)
56. Polyak, B.T.: Introduction to Optimization. Optimization Software Inc., Publications Division, New York (1987)
57. Rathke, F., Schnörr, C.: Fast multivariate log-concave density estimation. Comput. Stat. Data Anal. **140**, 41–58 (2019)
58. Reuther, A., Kepner, J., Byun, C., Samsi, S., Arcand, W., Bestor, D., Bergeron, B., Gadepally, V., Houle, M., Hubbell, M., et al.: Interactive supercomputing on 40,000 cores for machine learning and data analysis. In: 2018 IEEE High Performance extreme Computing Conference (HPEC), pp. 1–6. IEEE (2018)
59. Rockafellar, R.T.: Convex Analysis. Princeton University Press (1997)
60. Samworth, R.J.: Recent progress in log-concave density estimation. Stat. Sci. **33**(4), 493–509 (2018)
61. Samworth, R.J., Sen, B.: Editorial: special issue on nonparametric inference under shape constraints. Stat. Sci. **33**, 469–472 (2018)
62. Samworth, R.J., Yuan, M.: Independent component analysis via nonparametric maximum likelihood estimation. Ann. Stat. **40**(6), 2973–3002 (2012)
63. Schuhmacher, D., Hüsler, A., Dümbgen, L.: Multivariate log-concave distributions as a nearly parametric model. Stat. Risk Model. Appl. Finance Insur. **28**(3), 277–295 (2011)
64. Seijo, E., Sen, B.: Nonparametric least squares estimation of a multivariate convex regression function. Ann. Stat. **39**(3), 1633–1657 (2011)
65. Seregin, A., Wellner, J.A.: Nonparametric estimation of multivariate convex-transformed densities. Ann. Stat. **38**(6), 3751–3781 (2010)
66. Shor, N.Z.: Minimization Methods for Non-Differentiable Functions. Springer-Verlag (1985)
67. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. submitted to SIAM Journal on Optimization **1** (2008)
68. Walther, G.: Detecting the presence of mixing with multiscale maximum likelihood. J. Am. Stat. Assoc. **97**(458), 508–513 (2002)
69. Walther, G.: Inference and modeling with log-concave distributions. Stat. Sci. **24**(3), 319–327 (2009)
70. Xiao, L.: Dual averaging methods for regularized stochastic learning and online optimization. J. Mach. Learn. Res. **11**, 2543–2596 (2010)
71. Xu, M., Samworth, R.J.: High-dimensional nonparametric density estimation via symmetry and shape constraints. Ann. Stat. **49**, 650–672 (2021)
72. Yang, F., Barber, R.F.: Contraction and uniform convergence of isotonic regression. Electron. J. Stat. **13**(1), 646–677 (2019)
73. Yousefian, F., Nedić, A., Shanbhag, U.V.: On stochastic gradient and subgradient methods with adaptive steplength sequences. Automatica **48**(1), 56–67 (2012)
74. Zhang, C.H.: Risk bounds in isotonic regression. Ann. Stat. **30**(2), 528–555 (2002)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.