Maximum Likelihood Estimation of a Multi-Dimensional Log-Concave Density Through SOS-Convexity

Wei Hu
real.huwei@gmail.com

Max Cytrynbaum
mcytrynbaum@gmail.com

Abstract

We study a tractable family of log-concave density estimators via maximum likelihood estimation. Specifically, we let $s(x)$ be a multivariate convex polynomial and consider densities of the form $\exp(-s(x))$. While checking if a polynomial is convex is NP-hard in general, sos-convexity can be enforced using semi-definite programming. We restrict $s(x)$ to be an sos-convex polynomial and formulate this problem as a convex optimization. An algorithm using projected stochastic gradient method is proposed, in which biased gradient estimates are obtained through an Markov chain Monte Carlo (MCMC) sampling procedure that is efficient for log-concave densities. We motion towards the theoretical properties of our estimator, including consistency and asymptotic equivalence with the max-likelihood log-concave estimator discussed in [Cule et al. (2010)].

1 Introduction

Consider the general maximum likelihood estimation problem. We are given independent and identically distributed random vectors $\{x_i\}_{i=1}^n$ with an unknown probability density f and asked to find a density that maximizes the likelihood in some restricted class of densities $\mathcal{F}$. Formally, we want to maximize

$$\ell(f) = \sum_{i=1}^{n} \log f(x_i) \tag{1.1}$$

under the constraint $f \in \mathcal{F}$.

In a nonparametric setting, the function class $\mathcal{F}$ is often defined by a shape constraint. Log-concavity is a common shape constraint with a variety of well-studied properties. A density $f(x)$ is said to be log-concave if $f(x) = \exp(-s(x))$, where $s(x)$ is a convex function. Note that this nests several well known distributions, including normal, exponential, logistic, gamma (with shape parameter at least 1), and many more. For example, the multivariate normal distribution $N(\mu, \Sigma)$ has a density proportional to $f(x) = \exp(-s(x))$, where $s(x) = \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$ is clearly convex.

In [Cule et al. (2010)] is studied the case where $\mathcal{F} = \{\text{log-concave densities}\}$. They show that, with probability 1, there exists a unique optimal solution $\hat{f}_n$ of (1.1). The paper also derives the form of the maximum likelihood estimator, showing that $\log \hat{f}_n$ is a piecewise-affine "tent function" on the convex hull C_n of the data and $-\infty$ outside C_n. Therefore $\hat{f}_n$ is in general not smooth, even on C_n. Computing $\hat{f}_n$ is formulated as a non-differential convex optimization problem which is solved through Shor's r-algorithm.

We consider a different problem, restricting our attention to the class of log-concave functions $f(x) = \exp(-s(x))$ where $s(x)$ is an sos-convex polynomial. While checking a polynomial's convexity is in general strongly NP-hard as shown in [Ahmadi et al. (2013)], an algebraic sum of squares (sos) based sufficient condition for polynomial convexity, termed sos-convexity, can be formulated as a semidefinite program and can thus be checked efficiently. Using this insight, we propose an algorithm based on projected stochastic gradient method, study its convergence rate, and demonstrate some of the properties of this estimator. In our algorithm, Markov chain Monte Carlo (MCMC) sampling is used for approximating the gradient. Importantly for our purposes, there exist efficient MCMC samplers for log-concave densities. For these procedures, it can be shown that a Markov chain with stationary distribution $f(x) = \exp(-s(x))$ is rapidly mixing. For more details and further references, see e.g. [Narayanan and Rakhlin (2013)] or [Lovász and Vempala (2006)].

The rest of the paper is organized as follows. In Section 2, we give a brief introduction to sos-convexity. In Section 3, we formulate our maximum likelihood estimation problem and discuss some of its properties. In Section 4, we propose an algorithm of calculating the estimator and show its convergence rate. In Section 5, we prove some theoretical properties of log-sos-concave densities as well as our estimator.

2 Convexity and SOS-Convexity

In this section, we briefly review the basic definitions and results regarding polynomial convexity and sos-convexity. Throughout this paper, we denote the number of variables and degree of polynomials by p and d respectively.

2.1 Nonnegativity and Sum of Squares

A polynomial g is said to be *nonnegative* or *positive semidefinite (psd)* if $g(x) \geq 0$ for all $x = (x_1, \cdots, x_p) \in \mathbb{R}^p$. It is NP-hard to decide whether or not a polynomial is psd. An obvious sufficient condition for a polynomial g to be psd is that it is *sum of squares (sos)*, i.e., it has the form

$$g(x) = \sum_{i=1}^{m} g_i(x)^2$$

for some polynomials $g_1, \cdots, g_m$.

From a computational point of view, sos is an appealing concept because of its relation to semidefinite programming. It can be easily verified that a polynomial g in p variables and of even degree d is sos if and only if there exists a positive semidefinite matrix Q such that

$$g(x) = z^T Q z, \tag{2.1}$$

where z is the vector of all monomials with degree at most $d/2$:

$$z = (1, x_1, \cdots, x_p, x_1^2, x_1 x_2, \cdots, x_p^{d/2})^T.$$

Note that (2.1) is equivalent to a set of linear equality constraints equating the coefficients of all monomials on both sides. Therefore the the condition that g is sos is equivalent to the semidefiniteness constraint $Q \succeq 0$ together with a set of linear equations, which is the feasible set of a semidefinite program. This can be handled efficiently.

2.2 Convexity and SOS-convexity

A polynomial g is said to be (globally) convex if for all $x, y \in \mathbb{R}^p$ and $\lambda \in [0,1]$, we have

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y).$$

In [Ahmadi et al. (2013)] is proved that it is NP-hard to decide whether or not a polynomial is convex. This motivates the introduction of a sum-of-squares certificate of polynomial convexity, termed sos-convexity, which is an analogue of sum of squares being a sufficient condition of nonnegativity.

Before we introduce the concept of sos-convexity, we give the definition of sos-matrix as follows.

Definition 2.1 (sos-matrix)**.** *A symmetric polynomial matrix $U(x) \in \mathbb{R}[x]^{m\times m}$ is an sos-matrix if there exists a factorization $U(x) = V(x)^T V(x)$, where $V(x) \in \mathbb{R}[x]^{s\times m}$ is also a polynomial matrix (for some $s \in \mathbb{N}$).*

We know that a polynomial g is convex if and only if its hessian $\nabla^2 g$ satisfies $\nabla^2 g(x) \succeq 0$ for any $x \in \mathbb{R}^p$. Clearly, an sos-matrix $U(x)$ is positive semidefinite for any x. Based on this, sos-convexity is defined as follows, which was formally introduced in [Helton and Nie (2010)].

Definition 2.2 (sos-convexity)**.** *Let g be a polynomial over $\mathbb{R}^p$. We say that g is sos-convex if its Hessian $\nabla^2 g$ is an sos-matrix.*

Any sos-convex polynomial is convex, because its hessian is an sos-matrix, and thus is positive semidefinite for any x.

Do there exist polynomials that are convex but not sos-convex? This question was answered in the affirmative in [Ahmadi and Parrilo (2011)]. In [Ahmadi and Parrilo (2013)] is proved the following result about the gap between convex ans sos-convex polynomials:

Theorem 2.3. *Consider polynomials of degree d over $\mathbb{R}^p$. Except in the cases (1) $p = 1$, (2) $d = 2$, and (3) $(p, d) = (2, 4)$, there exists a polynomial g such that g is convex but not sos-convex.*

Nevertheless, the tractability of checking sos-convexity through semidefinite programming makes it an attractive alternative of polynomial convexity. We have the following

Proposition 2.4. *For a polynomial $g(x)$ in p variables, define*

$$h(x, y) = y^T[\nabla^2 g(x)]y$$

as a polynomial in $2p$ variables. Then

(1) g is convex if and only if h is psd.

(2) g is sos-convex if and only if h is sos.

Hence, to determine whether g is sos-convex is equivalent to determine whether h is sos, which can be formulated as a semidefiniteness constraint together with a set of linear equality constraints, as discussed before.

3 The Log-SOS-Concave Maximum Likelihood Estimation Problem

In this section, we formalize the problem of maximum likelihood estimation of a log-concave density $f(x) = \exp(-s(x))$, where s is an sos-convex polynomial (we say that such a density f is "log-sos-concave"). We will also discuss some properties of this problem.

3.1 Notations

First, we introduce some notations.

Consider a polynomial $g(x)$ over $\mathbb{R}^p$, where $x = (x_1, \cdots, x_p)$. We use multi-indices $\alpha \in \mathbb{N}^p$ to represent monomials — for $\alpha = (\alpha_1, \cdots, \alpha_p) \in \mathbb{N}^p$, $x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_p^{\alpha_p}$. Then we can write any polynomial g as

$$g(x) = \sum_{\alpha \in \mathbb{N}^p} \theta_\alpha x^\alpha$$

where θ_α is the coefficient of x^α.

Let $v_d(x)$ denote the vector of all monomials in x with degree at most d, i.e.,

$$v_d(x) = (x_1, \cdots, x_p, x_1^2, x_1 x_2, \cdots, x_p^d).$$

Then any polynomial g with degree at most d can be written as

$$g(x) = \sum_{\alpha \in \mathbb{N}^p, |\alpha| \leq d} \theta_\alpha x^\alpha = \theta^T v_d(x) \tag{3.1}$$

where $|\alpha| = \sum_{i=1}^{p} |\alpha_i|$, and θ is the vector of all coefficients. The order of these coefficients corresponds to the order of monomials in $v_d(x)$.

The length of $v_d(x)$ is $\binom{p+d}{d}$. Therefore We parametrize polynomials of degree d as coefficient vectors of length $\binom{p+d}{d}$.

3.2 Problem Setup

We consider the family of log-sos-concave densities:

$$f(x) = \frac{e^{-s(x)}}{Z}$$

where $s(x)$ is an sos-convex polynomial and Z is a normalization constant. We assume that f is supported on a compact convex set $K \in \mathbb{R}^p$. Specifically, let $K \subseteq [-R, R]^p$ for some $R > 0$. This assumption is made mainly for computational purposes. Since log-concave densities have "light tails", this assumption will affect little when R is large enough. Then we have

$$Z = \int_K e^{-s(x)} dx.$$

We further assume that the degree of s is at most d (d is even), then from (3.1), s can be represented as

$$s(x) = \sum_{\alpha \in \mathbb{N}^p, |\alpha| \leq d} \theta_\alpha x^\alpha = \theta^T v_d(x).$$

With increasing degree d, we have a sequence of classes of densities $\{\mathcal{F}_d\}_{d=0,2,4,\cdots}$, where $\mathcal{F}_d = \big\{f(x) \propto e^{-s(x)} \text{ (on } K) : s \text{ is a sos-convex polynomial in } p \text{ variables with degree at most } d\big\}$. The set of all log-sos-concave densities (supported on K) is $\mathcal{F} = \bigcup_{d=0,2,4,\cdots} \mathcal{F}_d$. Therefore, with increasing d, the maximum likelihood within $\mathcal{F}_d$ will converge to that within $\mathcal{F}$.

Now we focus on the density class $\mathcal{F}_d$. This class is parametrized by a length-$\binom{p+d}{d}$ vector θ, and then we write the density $f(x)$ in $\mathcal{F}_d$ with parameter θ as $f(x;\theta)$. Given n i.i.d. samples $x_1, \cdots, x_n \in K$, the log-likelihood is

$$\ell(\theta) = \sum_{i=1}^n \log f(x_i) = -\sum_{i=1}^n s(x_i) - n \log Z = -\sum_{i=1}^n \theta^T v_d(x_i) - n \log \int_K e^{-\theta^T v_d(x)} dx.$$

Therefore the maximum likelihood estimation problem can be summarized by the following optimization:

$$\begin{array}{ll} \underset{\theta}{\text{minimize}} & \frac{1}{n}\theta^T \sum_{i=1}^n v_d(x_i) + \log \int_K e^{-\theta^T v_d(x)} dx \\ \text{s.t.} & s(x) = \theta^T v_d(x) \text{ is sos-convex.} \end{array} \tag{3.2}$$

3.3 Properties

Problem (3.2) turns out to be a convex optimization. Denote the objective function in (3.2) by $F(\theta)$. The gradient and Hessian of F are:

$$\nabla F(\theta) = \frac{1}{n}\sum_{i=1}^n v_d(x_i) + \frac{\int_K e^{-\theta^T v_d(x)}(-v_d(x))dx}{\int_K e^{-\theta^T v_d(x)}dx} = \frac{1}{n}\sum_{i=1}^n v_d(x_i) - \mathbb{E}_\theta[v_d(X)] \tag{3.3}$$

and

$$\begin{aligned} \nabla^2 F(\theta) =& \frac{(\int_K e^{-\theta^T v_d(x)} v_d(x) v_d(x)^T dx)(\int_K e^{-\theta^T v_d(x)} dx) - (-\int_K e^{-\theta^T v_d(x)} v_d(x) dx)(\int_K e^{-\theta^T v_d(x)}(-v_d(x)^T) dx)}{(\int_K e^{-\theta^T v_d(x)} dx)^2} \\ =& \mathbb{E}_\theta[v_d(X)v_d(X)^T] - \mathbb{E}_\theta[v_d(X)] \cdot \mathbb{E}_\theta[v_d(X)]^T = \mathbb{V}_\theta[v_d(X)] \end{aligned} \tag{3.4}$$

where X is a random variable with density $f(x;\theta)$, and $\mathbb{E}_\theta[g(X)]$ and $\mathbb{V}_\theta[g(X)]$ are respectively the expectation and covariance matrix of a (vector-valued) function g of a random vector X with density $f(x;\theta)$. In other words,

$$\mathbb{E}_\theta[g(X)] = \int_K g(x) f(x;\theta) dx,$$

$$\mathbb{V}_\theta[g(X)] = \mathbb{E}_\theta[g(X)g(X)^T] - \mathbb{E}_\theta[g(X)]\mathbb{E}_\theta[g(X)]^T.$$

Since the covariance matrix of any random vector is positive semidefinite, we have $\nabla^2 F(\theta) \succeq 0$ for all θ, which immediately indicates that F is a convex function. Additionally, the feasible set of (3.2) is clearly convex. Hence, (3.2) is a convex optimization problem.

We use $\|\cdot\|$ to denote the standard l_2-norm. We can give an upper-bound of $\|\nabla F\|$:

Lemma 3.1. *Let $M = 2R^d\sqrt{\binom{p+d}{d}}$. For any θ, we have $\|\nabla F(\theta)\| \le M$.*

Proof. Since $K \subseteq [-R, R]^p$, we have $v(x) \in [-R^d, R^d]^{\binom{p+d}{d}}$ for all $x \in K$. Hence $\frac{1}{n}\sum_{i=1}^{n} v_d(x_i) \in [-R^d, R^d]^{\binom{p+d}{d}}$ and $\mathbb{E}_\theta(v_d(X)) \in [-R^d, R^d]^{\binom{p+d}{d}}$, which implies $\nabla g(\theta) \in [-2R^d, 2R^d]^{\binom{p+d}{d}}$ according to (3.3). Therefore

$$\|\nabla g(\theta)\| \leq \sqrt{\binom{p+d}{d}(2R^d)^2} = 2R^d\sqrt{\binom{p+d}{d}} = M.$$

□

Another important property of F is that its gradient is *Lipschitz-continuous*, i.e., there exists a constant L such that

$$\|\nabla F(\theta) - \nabla F(\eta)\| \leq L\|\theta - \eta\|$$

for any $\theta, \eta \in \mathbb{R}^{\binom{p+d}{d}}$. This fact is an important ingredient for showing convergence of our algorithm (in Section 4) and is a useful assumption in convex optimization. We have the following

Lemma 3.2. *Let $L = 2R^{2d}\binom{p+d}{d}$, then ∇F is Lipschitz-continuous with constant L.*

Proof. It suffices to prove that for any θ, the maximum eigenvalue of $\nabla^2 g(\theta)$ is at most L.

Recall that we use multi-indices $\alpha \in \mathbb{N}^p$ for the monomials as well as the vector θ, and thus we can also use it to index the rows and columns in $\nabla^2 F(\theta)$. The entry at row α and column β of $\nabla^2 F(\theta)$ $(\alpha, \beta \in \mathbb{N}^p, |\alpha| \leq d, |\beta| \leq d)$ is

$$[\nabla^2 F(\theta)]_{\alpha,\beta} = \mathbb{E}_\theta[X^{\alpha+\beta}] - \mathbb{E}_\theta[X^\alpha]\mathbb{E}_\theta[X^\beta].$$

For any $x \in K \in [-R, R]^d$, we have $|x^\alpha| \leq R^{|\alpha|} \leq R^d$ and similarly $|x^\beta| \leq R^d, |x^{\alpha+\beta}| \leq R^{2d}$. Then we have

$$\big|\mathbb{E}_\theta[X^\alpha]\big| \leq R^d, \big|\mathbb{E}_\theta[X^\beta]\big| \leq R^d, \big|\mathbb{E}_\theta[X^{\alpha+\beta}]\big| \leq R^{2d}$$

and hence

$$\big|[\nabla^2 F(\theta)]_{\alpha,\beta}\big| = \big|\mathbb{E}_\theta[X^{\alpha+\beta}] - \mathbb{E}_\theta[X^\alpha]\mathbb{E}_\theta[X^\beta]\big| \leq 2R^{2d}. \tag{3.5}$$

Note that (3.5) holds for any row α and column β.

By Gershgorin circle theorem, for any eigenvalue λ of $\nabla^2 g(\theta)$, there exists some row α such that

$$\big|\lambda - [\nabla^2 F(\theta)]_{\alpha,\alpha}\big| \leq \sum_{\beta \neq \alpha} \big|[\nabla^2 F(\theta)]_{\alpha,\beta}\big|,$$

which implies

$$\lambda \leq [\nabla^2 F(\theta)]_{\alpha,\alpha} + \sum_{\beta \neq \alpha} \big|[\nabla^2 F(\theta)]_{\alpha,\beta}\big| \leq \sum_{\beta} \big|[\nabla^2 F(\theta)]_{\alpha,\beta}\big| \leq \binom{p+d}{d} \cdot 2R^{2d} = L,$$

completing the proof. □

4 Algorithm

We solve (3.2) using *projected stochastic gradient method*. This generates a sequence $\{\theta_k\}_{k\geq 0}$ through the recursion

$$\theta_{k+1} = P_{sos}\big(\theta_k - \alpha_k(\nabla F(\theta_k) + \xi_k)\big), k = 0, 1, 2, \cdots \tag{4.1}$$

where the initial point θ_0 is feasible for (3.2), $\{\alpha_k\}$ is a sequence of deterministic positive stepsizes, ξ_k is the (stochastic) error in the gradient evaluation, and $P_{sos}(\gamma)$ is the projection of $\gamma \in \mathbb{R}^{\binom{p+d}{d}}$ onto the feasible set of (3.2). The projection operator can be written as an optimization problem

$$\begin{aligned} P_{sos}(\gamma): \qquad & \underset{\theta}{\text{minimize}} \quad \|\theta - \gamma\|_2^2 \\ & \text{s.t.} \qquad \theta^T v_d(x) \text{ is sos-convex.} \end{aligned} \tag{4.2}$$

4.1 Approximating the Gradient

In the algorithm we need to evaluate the gradient $\nabla F(\theta)$ whose expression is given in (3.3). To do this it suffices to evaluate the expectation $\mathbb{E}_\theta[v_d(X)]$ for any given θ. Since we only know the distribution $f(x;\theta)$ up to a constant factor, i.e. $f(x;\theta) \propto e^{-\theta^T v_d(x)}$ on K and 0 outside K, we can use Markov chain Monte Carlo (MCMC) to approximate the expectation. MCMC is an extremely powerful approach of sampling from and calculating the expectations with respect to high-dimensional probability distributions. In particular, we adopt the method proposed in [Narayanan and Rakhlin (2013)] which studies how to sample from time-varying log-concave distributions using a single Markov chain.

To be more specific, suppose a sequence of log-concave distributions with measures $\mu_0, \mu_1, \cdots$ is tracked; they are supported on a compact convex set K and their densities are only known up to a constant — assume that we know $\frac{d\mu_k(x)}{dx} \propto e^{-s_k(x)}$ (on K) for a convex function s_k. (In our case, μ_k has density $f(x;\theta_k)$.) A single Markov chain based on *Dikin Walk* are run for sampling form $\mu_0, \mu_1, \cdots$ successively. Suppose that the chain is run for μ_k for τ_k steps and that at the end of these τ_k steps the current distribution on the chain is $\hat{\mu}_k$ $(k = 0, 1, 2, \cdots)$. It is proved in [Narayanan and Rakhlin (2013)] that the "distance" between $\hat{\mu}_k$ and μ_k can be arbitrarily small provided that τ_k is large enough, i.e.,

$$\sqrt{\int_K (\frac{d\hat{\mu}_k}{d\mu_k} - 1)^2 d\mu_k} \leq \varepsilon_k. \tag{4.3}$$

The required value of τ_k can be calculated "on the fly". It is related to ε_{k-1}, s_k and s_{k-1}, and is $O(\log(\frac{1}{\varepsilon_k}))$ when ε_{k-1}, s_k and s_{k-1} are fixed. Note that (4.3) implies a bound on the total variation distance between $\hat{\mu}_k$ and μ_k:

$$\int_K \big|d\hat{\mu}_k - d\mu_k\big| \leq \varepsilon_k. \tag{4.4}$$

On the other hand, following the work in [Honorio (2012)], we characterize the asymptotic behavior of a family of samplers that includes importance sampling and MCMC:

Definition 4.1. *A (B, V, S, D)-sampler takes S samplers and produces biased estimates of $\nabla F(\theta)$ as $\nabla F(\theta) + \xi$, where the error ξ satisfies*

(1) $\mathbb{E}[\|\xi\|] \leq \frac{B}{S} + O(\frac{1}{S^2})$

(2) $\mathbb{V}ar[\|\xi\|] \leq \frac{V}{S} + O(\frac{1}{S^2})$

for $B \geq 0, V \geq 0$ *and* $(\forall\theta)\|\theta\| \leq D$.

For simplicity, we can rewrite the above conditions for (B, V, S, D)-sampler as

(1) $\mathbb{E}[\|\xi\|] \leq \frac{B}{S}$

(2) $\mathbb{V}ar[\|\xi\|] \leq \frac{V}{S}$.

Here the constants B and V are not necessarily the same as in Definition 4.1.

4.2 Projection as an SDP

It has been shown in Section 2 that the sos-convexity constraint in (4.2) is equivalent to a semidefiniteness constraint together with a set of linear equalities. In addition, a quadratic objective function can be transformed into a linear objective with a linear matrix inequality, which is stated in e.g. [Boyd and Vandenberghe (2004)]. Therefore, the projection (4.2) can be formulated as an SDP.

4.3 Stepsizes and Convergence Rate

Let θ^* be an optimal solution of problem (3.2). Projected gradient method is a special case of *proximal-gradient* method where the common choice of stepsizes for objective with L-Lipschitz-continuous gradient is the constant $\alpha_k = \frac{1}{L}(k = 0, 1, \cdots)$. In the error-free situation ($\xi_k = 0$), this can achieve an $O(1/n)$ convergence rate for objective values, while an *accelerated* case of proximal-gradient method can achieve $O(1/n^2)$.

4.3.1 Convergence Rate for Deterministic Errors

In [Schmidt et al. (2011)] is considered basic and accelerated proximal-gradient methods where errors are present in the calculation of gradients or the proximity operator. Since we only care about projected gradient method with errors in gradients, we present here the special versions of Proposition 1 and Proposition 2 in [Schmidt et al. (2011)]:

Lemma 4.2 (Basic proximal-gradient method)**.** *Assume we iterate* (4.1) *with* $\alpha_k = \frac{1}{L}(k = 0, 1, \cdots)$. *Then, for any* $n \geq 1$*, we have*

$$F(\overline{\theta}_n) - F(\theta^*) \leq \frac{L}{2n}\big(\|\theta_0 - \theta^*\| + \frac{2}{L}\sum_{k=0}^{n-1}\|\xi_k\|\big)^2 \tag{4.5}$$

where $\overline{\theta}_n = \frac{\sum_{k=1}^{n}\theta_k}{n}$.

If $\{\|\xi_k\|\}$ is summable (e.g. $\|\xi_k\| = O(\frac{1}{k^{1+\varepsilon}}), \varepsilon > 0$, for $k \geq 1$), then the well-known $O(1/n)$ convergence rate for basic proximal gradient method still holds as if there were no errors.

Lemma 4.3 (Accelerated proximal-gradient method)**.** *Assume we change* (4.1) *to the following:*

$$\begin{cases} \theta_{k+1} = P_{sos}\big(\eta_k - \frac{1}{L}(\nabla F(\theta_k) + \xi_k)\big) \\ \eta_k = \theta_k + \frac{k-1}{k+2}(\theta_k - \theta_{k-1}), k = 1, 2, \cdots \end{cases} \tag{4.6}$$

Then, for any $n \geq 1$, we have

$$F(\theta_n) - F(\theta^*) \leq \frac{2L}{(n+1)^2}\big(\|\theta_0 - \theta^*\| + \frac{2}{L}\sum_{k=0}^{n-1}(k+1)\|\xi_k\|\big)^2. \tag{4.7}$$

In this case, we require the sequence $\{(k+1)\|\xi_k\|\}$ to be summable to achieve the $O(1/n^2)$ rate as in the error-free situation.

4.3.2 Bounding the Error Terms

We give high-probability bounds for the error terms $\sum_{k=0}^{n-1}\|\xi_k\|$ and $\sum_{k=0}^{n-1}(k+1)\|\xi_k\|$ in (4.5) and (4.7) respectively. We assume that all visited points $\theta_0, \theta_1, \cdots, \theta_{n-1}$ are bounded, i.e., $\|\theta_k\| \leq D$.

As stated in [Honorio (2012)], in the case that a new Markov chain is run in each iteration, the errors $\xi_0, \xi_1, \cdots, \xi_{n-1}$ are independent given $\theta_0, \cdots, \theta_{n-1}$. This is not true, however, if we adopt the sampling scheme in [Narayanan and Rakhlin (2013)], where a single long chain is run throughout all iterations. Nevertheless, a weaker condition suffices for our purpose. Let $\mathcal{F}_k$ be the σ-algebra generated by $\theta_0, \theta_1, \cdots, \theta_k$. Suppose S_k samples are taken in the calculation of $\nabla F(\theta_k)$, by the definition of (B, V, S, D)-sampler we have

$$\begin{aligned} \mathbb{E}[\|\xi_k\|\big|\mathcal{F}_k] &\leq \frac{B}{S_k} \\ \mathbb{V}ar[\|\xi_k\|\big|\mathcal{F}_k] &\leq \frac{V}{S_k}. \end{aligned} \tag{4.8}$$

With (4.8), we can prove the high-probability bounds of the error terms.

Theorem 4.4. *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the error term $\sum_{k=0}^{n-1}\|\xi_k\|$ in basic proximal-gradient method is bounded as follows:*

$$\sum_{k=0}^{n-1}\|\xi_k\| \leq B\sum_{k=0}^{n-1}\frac{1}{S_k} + \frac{2M}{3}\log\frac{1}{\delta} + \sqrt{2V\big(\sum_{k=0}^{n-1}\frac{1}{S_k}\big)\log\frac{1}{\delta} + \frac{4M^2}{9}\log^2\frac{1}{\delta}}. \tag{4.9}$$

Proof. From Lemma 3.1 we have $\|\nabla F(\theta_k)\| \leq M$. Since we have $\nabla F(\theta_k) + \xi_k = \frac{1}{n}\sum_{i=1}^{n} v_d(x_i) - \frac{1}{t}\sum_{j=1}^{t} v_d(x^{(j)})$ for some samples $x^{(1)}, \cdots, x^{(t)} \in K$, we also have $\|\nabla F(\theta_k) + \xi_k\| \leq M$. Therefore $\|\xi_k\| \leq 2M$ for $k = 0, 1, \cdots$.

Let $\varphi = \sum_{k=0}^{n-1}\mathbb{E}[\|\xi_k\|\big|\mathcal{F}_k]$ and $\sigma^2 = \sum_{k=0}^{n-1}\mathbb{V}ar[\|\xi_k\|\big|\mathcal{F}_k]$. By (4.8) we have

$$\varphi \leq B\sum_{k=0}^{n-1}\frac{1}{S_k}$$

and

$$\sigma^2 \le V\sum_{k=0}^{n-1}\frac{1}{S_k}.$$

By the generalized Bernstein inequlaity for martingales proposed in [Freedman (1975)], we have for any $\varepsilon > 0$,

$$\mathbb{P}(\sum_{k=0}^{n-1}\|\xi_k\| \ge \varphi + \varepsilon) \le e^{-\frac{\varepsilon^2}{2(\sigma^2+2M\varepsilon/3)}} \le e^{-\frac{\varepsilon^2}{2V\sum_{k=0}^{n-1}\frac{1}{S_k}+\frac{4M\varepsilon}{3}}} \equiv \delta.$$

Solving ε in the last equality, we have

$$\varepsilon = \frac{2M}{3}\log\frac{1}{\delta} + \sqrt{2V\big(\sum_{k=0}^{n-1}\frac{1}{S_k}\big)\log\frac{1}{\delta} + \frac{4M^2}{9}\log^2\frac{1}{\delta}}.$$

Therefore, with probability at least $1-\delta$, we have

$$\sum_{k=0}^{n-1}\|\xi_k\| \le \varphi + \varepsilon \le B\sum_{k=0}^{n-1}\frac{1}{S_k} + \frac{2M}{3}\log\frac{1}{\delta} + \sqrt{2V\big(\sum_{k=0}^{n-1}\frac{1}{S_k}\big)\log\frac{1}{\delta} + \frac{4M^2}{9}\log^2\frac{1}{\delta}}.$$

□

Recall that in order to obtain the $O(1/n)$ rate for basic proximal-gradient method, the error term should be bounded by a constant. The above theorem indicates that for any fixed δ, this can be achieved with probability at least $1-\delta$ if $\{1/S_k\}$ is summable (eg. $S_k = \Omega(k^{1+\eta})(\eta > 0)$).

Similarly, we bound the error term in accelerated proximal-gradient method as follows:

Theorem 4.5. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, the error term $\sum_{k=0}^{n-1}(k+1)\|\xi_k\|$ in accelerated proximal-gradient method is bounded as follows:*

$$\sum_{k=0}^{n-1}(k+1)\|\xi_k\| \le B\sum_{k=0}^{n-1}\frac{k+1}{S_k} + \frac{2Mn}{3}\log\frac{1}{\delta} + \sqrt{2V\big(\sum_{k=0}^{n-1}\frac{(k+1)^2}{S_k}\big)\log\frac{1}{\delta} + \frac{4M^2n^2}{9}\log^2\frac{1}{\delta}}. \quad (4.10)$$

Proof. Let $X_k = (k+1)\|\xi_k\|(k = 0, 1, \cdots, n-1)$. For $k = 0, 1, \cdots, n-1$ we have

$$\|X_k\| \le n \cdot 2M = 2Mn.$$

Let $\varphi = \sum_{k=0}^{n-1}\mathbb{E}[X_k|\mathcal{F}_k]$ and $\sigma^2 = \sum_{k=0}^{n-1}\mathbb{V}ar[X_k|\mathcal{F}_k]$. By (4.8) we have

$$\varphi \le B\sum_{k=0}^{n-1}\frac{k+1}{S_k}$$

and

$$\sigma^2 \le V\sum_{k=0}^{n-1}\frac{(k+1)^2}{S_k}.$$

By the generalized Bernstein inequlaity for martingales, we have for any $\varepsilon > 0$,

$$\mathbb{P}(\sum_{k=0}^{n-1} X_k \geq \varphi + \varepsilon) \leq e^{-\frac{\varepsilon^2}{2(\sigma^2 + 2Mn\varepsilon/3)}} \leq e^{-\frac{\varepsilon^2}{2V\sum_{k=0}^{n-1}\frac{(k+1)^2}{S_k} + \frac{4Mn\varepsilon}{3}}} \equiv \delta.$$

Solving ε in the last equality, we have

$$\varepsilon = \frac{2Mn}{3}\log\frac{1}{\delta} + \sqrt{2V\big(\sum_{k=0}^{n-1}\frac{(k+1)^2}{S_k}\big)\log\frac{1}{\delta} + \frac{4M^2n^2}{9}\log^2\frac{1}{\delta}}.$$

Therefore, with probability at least $1-\delta$, we have

$$\sum_{k=0}^{n-1}(k+1)\|\xi_k\| = \sum_{k=0}^{n-1} X_k \leq \varphi + \varepsilon \leq B\sum_{k=0}^{n-1}\frac{k+1}{S_k} + \frac{2Mn}{3}\log\frac{1}{\delta} + \sqrt{2V\big(\sum_{k=0}^{n-1}\frac{(k+1)^2}{S_k}\big)\log\frac{1}{\delta} + \frac{4M^2n^2}{9}\log^2\frac{1}{\delta}}.$$

□

Note that for any δ, the upper bound in (4.10) is always $\Omega(n)$ for any fixed δ. Plugging this bound back to (4.7), we find that accelerated proximal-gradient is not guaranteed to converge.

5 Theoretical Properties of the Estimator

We discuss some of the statistical properties of our estimator. Here, we still let K be a compact convex set as in the previous sections. We make the following conjecture that we only know for sure in the case $p = 1$ (univariate).

Conjecture 5.1. *SOS-convex polynomials are dense in continuous convex functions in the sense of $\|\cdot\|_{K,\infty}$. In other words, for any continuous convex function f on K and any $\varepsilon > 0$, there exists an sos-convex polynomial g such that*

$$\sup_{x\in K}|f(x) - g(x)| < \varepsilon.$$

Let f^* be the maximum likelihood log-concave density as discussed in [Cule et al. (2010)]. It is shown in [Cule and Samworth (2010)] that f^* has attractable theoretical properties both when the true underlying density is log-concave and when this model is misspecified. This motivates us to study the convergence of log-sos-concave densities to f^*. In particular, assuming the validity of Conjecture 5.1, we show in this section that:

(1) There is a sequence of truncated log-sos-concave densities supported on K that converges in distribution to f^*;

(2) There is a sequence of log-sos-concave densities supported on $\mathbb{R}^p$ that converges in distribution to f^*;

(3) As degree $d \to \infty$, the likelihood value returned by program (3.2) converges to that of f^*.

5.1 Proof of the Denseness Conjecture in 1-Dimension

Now we will give the proof of Conjecture 5.1 for $p = 1$. Restricted on K, let $SOSX$ denote the set of sos-convex polynomials, $SMSX$ denote the set of smooth strictly convex functions, SMX denote the set of smooth convex functions, and CTX denote the set of continuous convex functions. Consider the chain of denseness relations (in the sense of $\|\cdot\|_{\infty,K}$):

$$SOSX \overset{dense}{\subset} SMSX \overset{dense}{\subset} SMX \overset{dense}{\subset} CTX \tag{5.1}$$

We will prove the leftmost relation only for $p = 1$, and the other two for all $p \geq 1$.

Denseness of $SOSX$ in $SMSX$

We are only able to give the proof for the case $p = 1$. Let $f \in SMSX$. In 1-dimension, we have $K = [a, b]$ for some $a < b$. Since f is smooth and strictly convex, $f'' > 0$ on K. By Stone-Weierstrass theorem, let $\{q_m\}_{m=1}^{\infty}$ be a sequence of polynomials such that $q_m \to f''$ in $\|\cdot\|_{\infty}$. Then there exists an M such that $m \geq M$ implies $q_m > 0$ on K.

Define polynomials

$$p_m(x) = \int_a^x \left(\int_a^t q_m(s)ds \right) dt \tag{5.2}$$

Then clearly we have $p_m''(x) = q_m(x)$. This means for $m \geq M$, p_m is strictly convex on K, so p_m is an sos-convex polynomial (for $p = 1$, all convex polynomials are sos-convex). Since $b - a < \infty$, it is easy to see that $p_m \to f$ uniformly on $[a, b]$ (one way to do this is to apply bounded convergence twice to the expression in (5.2) above).

Therefore, we complete the proof of $SOSX \overset{dense}{\subset} SMSX$.

Denseness of $SMSX$ in SMX

This one is very easy to show. Let $f \in SMX$. For any $\varepsilon > 0$, $f_\varepsilon(x) := f(x) + \varepsilon\|x\|^2$ is strictly convex and

$$|f_\varepsilon(x) - f(x)| \leq \varepsilon\|x\|^2 \leq \varepsilon \cdot Diam(K)^2, \forall x \in K.$$

The proof is done.

Denseness of SMX in CTX

Let $f \in CTX$. We first extend f to be a continuous convex function on $\mathbb{R}^p$.

It is easy to see that convolution with a positive function preserves convexity of a function. That is, for any function φ that satisfies $\varphi \geq 0$ on $\mathbb{R}^p$, the convolution

$$(\varphi * f)(x) = \int_{\mathbb{R}^p} \varphi(t) f(x - t)dt$$

is convex.

Let φ be a smooth positive function whose support is bounded (e.g. a bump function). We require that $\int_{\mathbb{R}^p} \varphi(x)dx = 1$. For any $\varepsilon > 0$, define $\varphi_\varepsilon(x) = \frac{1}{\varepsilon}\varphi(\frac{x}{\varepsilon})$. Then we have

$$\int_{\mathbb{R}^p} \varphi_\varepsilon(x)dx = \int_{\mathbb{R}^p} \varphi(\frac{x}{\varepsilon})d\frac{x}{\varepsilon} = 1.$$

Let $f_\varepsilon = (\varphi_\varepsilon * f)$. It suffices to show the following facts to complete the proof:

(i) f_ε is convex for any $\varepsilon > 0$.

(ii) f_ε is smooth for any $\varepsilon > 0$.

(iii) $\lim_{\varepsilon \to 0} \big(\sup_{x \in K} |f_\varepsilon(x) - f(x)| \big) \to 0$.

Note that (i) is true because φ_ε is positive and convolution with a positive function preserves convexity.

For (ii), we have

$$(\varphi * f)(x) = \int_{\mathbb{R}^p} \varphi(t) f(x-t) dt = \int_{\mathbb{R}^p} \varphi(x-t) f(t) dt. \tag{5.3}$$

Whenever we need to take a partial derivative $\frac{\partial}{\partial x_\alpha}(\varphi * f)(x)$, we just pass the differentiation operator through the integral (since φ is boundedly supported and f is continuous) and differentiate the integrand φ which is smooth by assumption. Thus (ii) is true.

For (iii), we will show that for any $\varepsilon > 0$, there exists $\eta > 0$ such that

$$|f_\eta(x) - f(x)| \leq \varepsilon, \forall x \in K.$$

By the uniform continuity of f on K, there exists $\delta_1 > 0$ such that $x, y \in K$ and $\|x - y\| \leq \delta_1$ imply $|f(x) - f(y)| < \varepsilon$. Define another compact set

$$K' = \{y \in \mathbb{R}^p | \exists x \in K, \|x - y\| \leq \delta_1\}.$$

(It is obvious that $K \subset K'$.) Then there exists $\delta \in (0, \delta_1]$ such that $x, y \in K'$ and $\|x - y\| \leq \delta_1$ imply $|f(x) - f(y)| < \varepsilon$.

For any $x \in K \subset K'$ and $y \in \mathbb{R}^p$, if $\|x - y\| \leq \delta$, we have $\|x - y\| \leq \delta_1$ and hence $y \in K'$, which imply $|f(x) - f(y)| < \varepsilon$ by the definition of δ. In other words, for any $x \in K$ and $\|t\| \leq \delta$, we have $|f(x) - f(x-t)| < \varepsilon$.

Finally, we choose $\eta > 0$ small enough such that $\varphi_\eta(x) = 0$ for any $\|x\| \geq \delta$. (This can be done because the support of φ is bounded.) Then for any $x \in K$, we have

$$\begin{aligned}
|f(x) - f_\eta(x)| &= |\int_{\mathbb{R}^p} f(x)\varphi_\eta(t)dt - \int_{\mathbb{R}^p} f(x-t)\varphi_\eta(t)dt| \\
&= |\int_{\mathbb{R}^p} (f(x) - f(x-t))\varphi_\eta(t)dt| \\
&\leq \int_{\|t\| \leq \delta} |f(x) - f(x-t)|\varphi_\eta(t)dt + |\int_{\|t\| \geq \delta} (f(x) - f(x-t))\varphi_\eta(t)dt| \\
&\leq \int_{\|t\| \leq \delta} \varepsilon \varphi_\eta(t)dt + |\int_{\|t\| \geq \delta} (f(x) - f(x-t)) \cdot 0 dt| \\
&= \varepsilon.
\end{aligned}$$

This concludes the proof of (iii).

Therefore, we complete the proof of $SMX \overset{dense}{\subset} CTX$.

5.2 Convergence in Distribution

Let C_n be the convex hull of data points $x_1, \cdots, x_n \in \mathbb{R}^p$. As shown in [Cule et al. (2010)], the maximum likelihood log-concave density $f^*(x) = e^{-h(x)}$ is supported on C_n and $h(x)$ is a convex piecewise affine function on C_n. To be more specific, there exists $y_1, \cdots, y_n \in \mathbb{R}$ such that $h(x)$ is the largest convex function such that $h(x_i) \leq y_i$ for $i = 1, \cdots, n$.

Assume that $C_n \subset int(K)$. We will prove the following

Theorem 5.2. *Assume Conjecture 5.1 is true, then:*

(i) There is a sequence of truncated log-sos-concave densities supported on K that converges in distribution to f^*;

(ii) There is a sequence of log-sos-concave densities supported on $\mathbb{R}^p$ that converges in distribution to f^*.

Proof. It suffices to prove the existence a sequence of sos-convex polynomials $\{s_m(x)\}_{m=1,2,\cdots}$ such that for any bounded integrable function G on $\mathbb{R}^p$,

$$\begin{aligned} \frac{\int_K G(x)e^{-s_m(x)}dx}{\int_K e^{-s_m(x)}dx} &\to \int_{C_n} G(x)e^{-h(x)}dx \quad (m \to \infty) \\ \frac{\int_{\mathbb{R}^p} G(x)e^{-s_m(x)}dx}{\int_{\mathbb{R}^p} e^{-s_m(x)}dx} &\to \int_{C_n} G(x)e^{-h(x)}dx \quad (m \to \infty), \end{aligned} \tag{5.4}$$

because for an sos-convex polynomial $s(x)$, $f_1(x) = \frac{e^{-s(x)}}{\int_K e^{-s(x)}dx}(x \in K)$ is a truncated log-sos-concave densities supported on K and $f_2(x) = \frac{e^{-s(x)}}{\int_{\mathbb{R}^p} e^{-s(x)}dx}(x \in \mathbb{R}^p)$ is a log-sos-concave density supported on $\mathbb{R}^p$.

Since C_n is a polytope in $\mathbb{R}^p$, we can write $C_n = \{x \in \mathbb{R}^p | a_i^T x \leq b_i, i = 1, \cdots, K\}$, where $a_i \in \mathbb{R}^p \setminus \{0\}, b_i \in \mathbb{R} (i = 1, \cdots, K)$. For $\varepsilon > 0$, let $C_n^\varepsilon = \{x \in \mathbb{R}^p | a_i^T x \leq b_i, i = 1, \cdots, K\}$. Then for any $\varepsilon > 0$, C_n^ε is a polytope and $C_n \subset C_n^\varepsilon$. It is easy to see

$$\lim_{\varepsilon \to 0} \big(\sup_{x \in C_n^\varepsilon} d(x, C_n) \big) = 0 \tag{5.5}$$

and

$$\lim_{\varepsilon \to 0} \mathcal{L}(C_n^\varepsilon) = \mathcal{L}(C_n) \tag{5.6}$$

where $d(x, S)$ denotes the Euclidean distance from the point x to the set S and $\mathcal{L}$ denotes Lebesgue measure. Because $C_n \subset int(K)$, when ε is sufficiently small we have $C_n^\varepsilon \subset int(K)$. We will only consider ε in such range.

Let $N > 0$ be a constant such that $|h(x)| \leq N$ for any $x \in C_n$. Now we extend h in the following way:

1. Choose $M > N$.
2. Choose $\varepsilon > 0$ small enough such that the following convex function $\overline{h}$ defined on $C_n^\varepsilon \subset K$ satisfies $\overline{h}(x) = h(x)$ for all $x \in C_n$:
   (a) Let $x_1', \cdots, x_l'$ be all the extreme points in C_n^ε.

(b) Let $\overline{h}(x)$ be the largest convex function such that

$$\overline{h}(x_i) \le y_i \qquad i = 1, \cdots, n$$

and

$$\overline{h}(x'_j) \le M \qquad j = 1, \cdots, l \tag{5.7}$$

We argue that such extension is always possible for $M > N$ if ε is sufficiently small. Note that $\overline{h}$ is a convex piecewise affine function on C_n^ε. Since $\{x'_j | j = 1, \cdots, l\}$ are the extreme points of C_n^ε, we have $\overline{h}(x'_j) = M$ for $j = 1, \cdots, l$ and thus $\overline{h}(x) = M$ for all $x \in \partial C_n^\varepsilon$. Since $M > N$, when ε gets small enough, the conditions (5.7) have no effect on the function $\overline{h}$ within C_n, which leads to $\overline{h} = h$ on C_n as we desire.

Now we envoke Conjecture 5.1 for continuous convex function $\overline{h}$ on ∂C_n^ε: for any $\eta > 0$, there exists an sos-convex polynomial $s(x)$ such that

$$|s(x) - \overline{h}(x)| < \eta, \qquad \forall x \in C_n^\varepsilon.$$

Since C_n is compact and $G(x)$ is bounded, we have

$$\lim_{\eta \to 0} \int_{C_n} G(x) e^{-s(x)} dx = \int_{C_n} G(x) e^{-\overline{h}(x)} dx = \int_{C_n} G(x) e^{-h(x)} dx. \tag{5.8}$$

Note that $s(x) \le \overline{h}(x) + \eta \le M + \eta$ on C_n^ε and $G(x)$ is bounded. From (5.6) we know that

$$\lim_{\varepsilon \to 0} \int_{C_n^\varepsilon \setminus C_n} G(x) e^{-s(x)} dx = 0. \tag{5.9}$$

Let $\gamma = \sup_{x \in C_n^\varepsilon} d(x, C_n)$. From (5.5) we know that $\gamma \to 0$ as $\varepsilon \to 0$. For any $x \in \mathbb{R}^p \setminus C_n^\varepsilon$, let $P(x)$ be the projection of x onto C_n and x' be the intersection of segment $(x, P(x))$ and ∂C_n^ε. Then $P(x)$ is also the projection of x' onto C_n. Since s is convex, we have

$$\|x - x'\| s(P(x)) + \|x' - P(x)\| s(x) \ge \|x - P(x)\| s(x'),$$

which means

$$s(x) \ge \frac{\|x - P(x)\| s(x') - \|x - x'\| s(P(x))}{\|x' - P(x)\|} = s(x') + \frac{\|x - x'\|}{\|P(x) - x'\|} (s(x') - s(P(x))).$$

Because

$$\begin{aligned} s(P(x)) &\le \overline{h}(P(x)) + \eta = h(P(x)) + \eta \le N + \eta \\ s(x') &\ge \overline{h}(x') - \eta = M - \eta \end{aligned}$$

and

$$\|P(x) - x'\| = d(x', C_n) \le \gamma,$$

we have

$$s(x) \ge M - \eta + \frac{\|x - x'\|}{\gamma} ((M - \eta) - (N - \eta)) \ge M - \eta + \frac{\max(0, \|x\| - T)}{\gamma} (M - N - 2\eta)$$

if $\eta < (M-N)/2$, where $T = \sup_{x\in K} \|x\|$.

Since $G(x)$ is bounded, let $Z = \sup_{x\in\mathbb{R}^p} |G(x)|$. Then we have

$$\begin{aligned}\int_{\mathbb{R}^p\setminus C_n^\varepsilon} |G(x)|e^{-s(x)}dx &\le \int_{\mathbb{R}^p\setminus C_n^\varepsilon} |G(x)|e^{-s(x)}dx \le Z\int_{\mathbb{R}^p\setminus C_n^\varepsilon} e^{-(M-\eta)-\frac{\max(0,\|x\|-T)}{\gamma}(M-N-2\eta)}dx\\ &\le Z\int_{\mathbb{R}^p} e^{-(M-\eta)-\frac{\max(0,\|x\|-T)}{\gamma}(M-N-2\eta)}dx\\ &= Z\int_{\|x\|\le T} e^{-(M-\eta)}dx + Z\int_{\|x\|\ge T} e^{-(M-\eta)-\frac{\|x\|-T}{\gamma}(M-N-2\eta)}dx.\end{aligned}$$

As $M\to\infty$, the first term above tends to 0; for fixed M and $\eta < (M-N)/2$, the second term tends to 0 as $\gamma\to 0$, or $\varepsilon\to 0$. Therefore

$$\lim_{M\to\infty}\lim_{\varepsilon\to 0}\int_{\mathbb{R}^p\setminus C_n^\varepsilon} |G(x)|e^{-s(x)}dx = 0$$

which implies

$$\lim_{M\to\infty}\lim_{\varepsilon\to 0}\int_{\mathbb{R}^p\setminus C_n^\varepsilon} G(x)e^{-s(x)}dx = 0 \tag{5.10}$$

and

$$\lim_{M\to\infty}\lim_{\varepsilon\to 0}\int_{K\setminus C_n^\varepsilon} G(x)e^{-s(x)}dx = 0. \tag{5.11}$$

Adding (5.8), (5.9) and (5.11) together, we obtain

$$\lim_{M\to\infty}\lim_{\varepsilon\to 0,\eta\to 0}\int_K G(x)e^{-s(x)}dx = \int_{C_n} G(x)e^{-s(x)}dx.$$

Similarly, adding (5.8), (5.9) and (5.10) together, we obtain

$$\lim_{M\to\infty}\lim_{\varepsilon\to 0,\eta\to 0}\int_{\mathbb{R}^p} G(x)e^{-s(x)}dx = \int_{C_n} G(x)e^{-s(x)}dx.$$

Therefore, there exists a sequence of sos-convex polynomials $\{s_m(x)\}_{m=1,2,\cdots}$ such that

$$\begin{aligned}&\int_K G(x)e^{-s_m(x)}dx \to \int_{C_n} G(x)e^{-h(x)}dx \quad (m\to\infty)\\ &\int_{\mathbb{R}^p} G(x)e^{-s_m(x)}dx \to \int_{C_n} G(x)e^{-h(x)}dx \quad (m\to\infty).\end{aligned} \tag{5.12}$$

Let $G(x)\equiv 1$ in (5.12), we obtain

$$\begin{aligned}&\int_K e^{-s_m(x)}dx \to \int_{C_n} e^{-h(x)}dx = 1 \quad (m\to\infty)\\ &\int_{\mathbb{R}^p} e^{-s_m(x)}dx \to \int_{C_n} e^{-h(x)}dx = 1 \quad (m\to\infty).\end{aligned} \tag{5.13}$$

Dividing the two limit relations in (5.12) with their corresponding limits of normalization constant in (5.13), we have (5.4). The proof is completed. □

5.3 Convergence of Objective Function Value

Now we can easily show the following

Theorem 5.3. *Assume Conjecture 5.1 is true. Let f_d^{sos} be the density of the maximum likelihood estimator given by* (3.2)*, then we have*

$$\sum_{i=1}^{n} \log f_d^{sos}(x_i) \to \sum_{i=1}^{n} \log f^*(x_i) \qquad (d \to \infty) \tag{5.14}$$

Proof. By Theorem 5.2, there exists a sequence of truncated log-sos-concave densities $\{f_m\}_{m=1,2,\cdots}$ supported on K that converges in distribution to f^*. This implies

$$\sup_{x \in K} |f_m(x) - f^*(x)| \to 0 \qquad (m \to \infty).$$

Since K is compact and f_m, f^* are continuous, we have

$$\sup_{x \in K} |\log f_m(x) - \log f^*(x)| \to 0 \qquad (m \to \infty),$$

which implies

$$|\log f_m(x_i) - \log f^*(x_i)| \to 0 \qquad (m \to \infty)$$

for $i = 1, \cdots, n$.

Therefore

$$\sum_{i=1}^{n} \log f_m(x_i) \to \sum_{i=1}^{n} \log f^*(x_i) \qquad (m \to \infty),$$

concluding the proof. □

6 Future Work

Some directions for future work on log-sos-concave density estimation are

1. Experiments.
   Simulations need to be done for testing the performance of the algorithm.

2. Theoretical properties.
   Except for the asymptotic equivalence with the estimator in [Cule et al. (2010)] as discussed in this paper, we could investigate the consistensy of our estimator and the behavior of this estimator with increasing degree.

Acknowledgments

We would like to thank Prof. John Lafferty for his wonderful mentorship in this summer, and to thank our colleagues Yo Joong Choe and Sabyasachi Chatterjee for helpful discussions and suggestions. We are also greatful to Prof. László Babai and Prof. Stuart Kurtz for organizing this wonderful REU program.

References

[Ahmadi and Parrilo (2011)] Ahmadi, A. A., and Parrilo, P. A. A convex polynomial that is not sos-convex. *Mathematical Programming*, 135(1-2), 275-292.

[Ahmadi and Parrilo (2013)] Ahmadi, A. A., and Parrilo, P. A. A complete characterization of the gap between convexity and sos-convexity. *SIAM Journal on Optimization*, Vol. 23, No. 2, 811-833.

[Ahmadi et al. (2013)] Ahmadi, A. A., Olshevsky, A., Parrilo, P. A., and Tsitsiklis, J. N. NP-hardness of deciding convexity of quartic polynomials and related problems. *Mathematical Programming*, 137(1-2), 453-476.

[Boyd and Vandenberghe (2004)] Boyd, S., and Vandenberghe, L. Convex optimization. *Cambridge University Press*, 2004.

[Cule and Samworth (2010)] Cule, M., and Samworth, R. Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Statist.* 4, 254-270.

[Cule et al. (2010)] Cule, M., Samworth, R., and Stewart M. Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Statist. Soc.* B 72, Part 5, 545-607.

[Freedman (1975)] Freedman, D. A. On tail probabilities for martingales. *Ann. Probab.* 3, 100-118.

[Helton and Nie (2010)] Helton, J. W., and Nie, J. Semidefinite representation of convex sets. *Mathematical Programming* 122(1, Ser. A):21-64, 2010.

[Honorio (2012)] Honorio, J. Convergence rates of biased stochastic optimization for learning sparse Ising models. *ICML*, 2012.

[Lovász and Vempala (2006)] Lovász, L., and Vempala, S. Fast algorithms for logconcave functions: sampling, rounding, integration and optimization. *FOCS*, 2006.

[Narayanan and Rakhlin (2013)] Narayanan, H., and Rakhlin, A. Efficient sampling from time-varying log-concave distributions. *aiXiv:1309.5977v1*, 2013.

[Nemirovski et al. (2009)] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 2009.

[Schmidt et al. (2011)] Schmidt, M., Le Roux, N., and Bach, F. Convergence rates of inexact proximal-gradient methods for convex optimization. *NIPS*, 2011.