Shape-Constrained Regression using Sum of Squares Polynomials

Mihaela Curmei* Georgina Hall†

Abstract

We consider the problem of fitting a polynomial function to a set of data points, each data point consisting of a feature vector and a response variable. In contrast to standard polynomial regression, we require that the polynomial regressor satisfy shape constraints, such as monotonicity, Lipschitz-continuity, or convexity. We show how to use semidefinite programming to obtain polynomial regressors that have these properties. We then prove that, under some assumptions on the generation of the data points, the regressor obtained is a consistent estimator of the underlying shape-constrained function. We follow up with a thorough empirical comparison of our regressor to the convex least squares estimator introduced in [Hildreth 1954, Holloway 1979] and show that our regressor can be very valuable in settings where the number of data points is large and where new predictions need to be made quickly and often. We also propose a method that relies on linear and second-order cone programs to quickly update our regressor when a new batch of data points is provided. We conclude with two novel applications. The first application aims to approximate the function that maps a conic program's data to its optimal value. This enables us to obtain quick estimations of the optimal value without solving the conic program, which can be useful for real-time decision-making. We illustrate this on an example in inventory management contract negotiation. In the second application, we compute optimal transport maps using shape constraints as regularizers following [Paty 2020], and show, via a color transfer example, that this is a setting in which our regressor significantly outperforms other methods.

Keywords: Polynomial regression, convex regression, semidefinite programming, consistent estimators, sum of squares polynomials, optimal transport

AMS classification: 90C23, 90C22, 62J02, 90C90, 90B05, 49Q22

1 Introduction

As in most regression frameworks, we consider in this paper a feature vector $X \in \mathbb{R}^n$ and a response variable $Y \in \mathbb{R}$, and posit the existence of a relationship between X and Y, which we model by an unknown function f. The goal of regression is then to provide, given observed feature vector-response variable pairings $(X_1, Y_1), \dots, (X_m, Y_m)$, an estimator or regressor $\hat{f}$ of f. In *parametric* regression, which is our setting of interest, $\hat{f}$ is taken to be the minimizer of an optimization problem of the type

$$\min_{g \in \Theta} \mathcal{L}((g(X_1), Y_1), \dots, (g(X_m), Y_m)),$$

*Mihaela Curmei is with the department of Electrical Engineering and Computer Science at the University of California, Berkeley. Email: mcurmei@eecs.berkeley.edu

†Georgina Hall is with the department of Decision Sciences at INSEAD. Email: georgina.hall@insead.edu

where Θ is a class of parametric functions mapping $\mathbb{R}^n$ to $\mathbb{R}$ and $\mathcal{L}(\cdot)$ is a loss function quantifying the distance between the vectors $(g(X_1),\ldots,g(X_m))$ and $(Y_1,\ldots,Y_m)$. Here, we restrict ourselves to least-squares polynomial regression, taking Θ to be the set $P_{n,d}$ of polynomials in n variables and of degree d and $\mathcal{L}(\cdot)$ to be the 2-norm squared, i.e., $\mathcal{L}((g(X_1),Y_1),\ldots,(g(X_m),Y_m)) = \sum_{i=1}^m (g(X_i)-Y_i)^2$. We also assume moving forward that the feature space (and thus the domain of f and $\hat{f}$) is given by a full-dimensional box $B \subset \mathbb{R}^n$ defined as

$$B = \{(x_1,\ldots,x_n) \in \mathbb{R}^n \mid l_i \leq x_i \leq u_i,\ \forall i = 1,\ldots,n\}, \tag{1}$$

where $l_i < u_i$ for all $i = 1,\ldots,n$. This is not a restrictive assumption in practice as each feature tends to have a natural range in which it lies (which can possibly be quite large). Our approach also has natural extensions to semialgebraic feature spaces.

We deviate from standard least-squares polynomial regression in this paper by assuming that additional knowledge pertaining to the shape of f is given to us and then requiring that the estimator $\hat{f}$ be of the same shape. In other words, we consider least-squares *shape-constrained* polynomial regression. The shape constraints we consider are of two types: convexity constraints and K-bounded-derivative constraints, as defined below.

Definition 1 (Convexity over a box). A function f is convex over a box B if for any $x, y \in B$ and for any $\lambda \in [0,1]$, we have

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

This is equivalent, for twice differentiable functions f and full-dimensional boxes B, to requiring that

$$H_f(x) \succeq 0, \forall x \in B,$$

where $H_f(x)$ is the Hessian of f.

A proof of the equivalence of the two characterizations can be found, e.g., in [10, section 1.1.4].

Definition 2 (K-bounded derivatives over a box). Given K_1^-, K_1^+, …, K_n^-, $K_n^+ \in \mathbb{R} \cup \{-\infty, +\infty\}$ such that $K_i^- \leq K_i^+$ for $i = 1,\ldots,n$, and the associated vector

$$K := (K_1^-, K_1^+, \ldots, K_n^-, K_n^+),$$

a continuously-differentiable function f is said to have *K-bounded derivatives* over a box B if, for $i = 1,\ldots,n$,

$$K_i^- \leq \frac{\partial f(x)}{\partial x_i} \leq K_i^+, \forall x \in B. \tag{2}$$

Requiring that these conditions hold over B is less restrictive than requiring that they hold over the whole of $\mathbb{R}^n$. Furthermore, the K-bounded derivative condition subsumes many other important conditions, such as monotonicity and Lipschitz-continuity, when f is continuously differentiable. For example, f being increasing[1] in the variable x_i is equivalent to $\frac{\partial f(x)}{\partial x_i} \geq 0, \forall x \in B$, which corresponds to taking $K_i^- = 0$ and $K_i^+ = +\infty$ in Definition 2. Likewise, f being M-Lipschitz, i.e., $|f(x) - f(y)| \leq M||x-y||$ for a fixed positive scalar M, some norm $||.||$, and any $x, y \in B$, is equivalent to $-M \leq \frac{\partial f(x)}{\partial x_i} \leq M$ for all $x \in B$ and any $i = 1,\ldots,n$. Combined, the shape constraints given in Definitions 1 and 2 cover the wide majority of shape constraints actually

[1]As a reminder, f is increasing in variable x_i if $f(x_1,\ldots,x_i,\ldots,x_n) \leq f(x_1,\ldots,x_i+h,\ldots,x_n)$ for all $(x_1,\ldots,x_n) \in B$ and $h > 0$ such that $(x_1,\ldots,x_i+h,\ldots,x_n) \in B$.

considered in applications. A short and non-exhaustive list of areas where convex-constrained regression appears, e.g., includes economics [41], psychology [19], electrical engineering [25]), and medicine [51]. Similarly, monotone-constrained regression occurs in medicine [29], biology and environmental engineering [54], electrical and computer engineering [42, 43], economics [6], and civil engineering [14].

The rest of the paper thus focuses on two estimators: a convex-constrained regressor $\bar{g}_{m,d} : \mathbb{R}^n \to \mathbb{R}$, solution to

$$\begin{aligned} \bar{g}_{m,d} := \arg\min_{g \in P_{n,d}} \quad & \sum_{i=1}^{m} (Y_i - g(X_i))^2 \\ \text{s.t.} \quad & H_g(x) \succeq 0, \forall x \in B, \end{aligned} \tag{3}$$

and a bounded-derivatives regressor $\bar{h}_{m,d} : \mathbb{R}^n \to \mathbb{R}$, solution to

$$\begin{aligned} \bar{h}_{m,d} := \arg\min_{h \in P_{n,d}} \quad & \sum_{i=1}^{m} (Y_i - h(X_i))^2 \\ \text{s.t.} \quad & K_i^- \leq \frac{\partial h(x)}{\partial x_i} \leq K_i^+, \ i = 1, \ldots, n, \forall x \in B. \end{aligned} \tag{4}$$

While we consider each shape constraint in a separate optimization problem, it is of course possible to require that both type of constraints hold simultaneously. We further note that both of these optimization problems are feasible and, if m is large enough and the data points are linearly independent, have unique solutions, and so the regressors $\bar{g}_{m,d}$ and $\bar{h}_{m,d}$ are well defined. They are functions of d, the degree of the polynomial allowed by the user, and the datapoints $(X_1, Y_1), \ldots, (X_m, Y_m)$: we use the subscripts d and m respectively to denote these dependencies. If the shape constraints were removed in (3) and (4), these problems would be least-squares polynomial regression problems, which can be cast as linear systems and solved efficiently. The question is then whether the addition of these constraints changes the complexity of solving the problem; if it does, how one can hope to obtain an approximate solution; and whether the approximate solution obtained is a good candidate estimator. We answer these questions and many others in the rest of the paper. Before doing so, we briefly review the literature on the topic of shape-constrained regression.

1.1 Literature Review

Shape-constrained regression is a fundamental problem in statistics and, consequently, there exists a vast literature on the topic. We mention here but do not review the approaches described in [2, 3, 58, 30, 16, 17, 46] as they are too orthogonal to ours. We focus instead on three literature streams which are particularly close to our work.

The first one is a set of papers revolving around the *convex least-squares estimator (CLSE)*, introduced in [27, 28], which produces a convex estimator (though it has also been adapted to cater to monotonocity and Lipschitz-continuity [34]). It is obtained, as can be inferred from its name, by searching for a function among the set of convex functions that minimizes the least squares error between predicted values and measured values. Perhaps unexpectedly, this problem is tractable and can be reduced to a quadratic program (QP). The estimator thus obtained is a piecewise linear function and computing a prediction from a new feature vector can be done by solving a linear program (LP); see [31]. The CLSE is arguably the most prevalent shape-constrained regressor in the literature, and as a consequence, we provide a complete analysis of the relative pros and cons of the CLSE compared to our regressor in Section 4.1. Other works in this area involve either showing statistical properties of the CLSE [34, 56] or designing improved algorithms for its computation [26, 37, 40, 15, 35].

The second line of work that relates to ours is [38]. The setting considered there is that of multivariate polynomial regression and convexity of the regressor is enforced via the use of sum of squares polynomials, which is also what we do here. There are many differences between the two contributions however. From a theoretical perspective, we propose a proof of consistency of our estimators. We also enrich the original results by considering new shape constraints and providing a comprehensive complexity classification of the problem of solving (3) and (4) as a function of the degree d. From a computational perspective, we give a thorough comparison to the extant approach, the CLSE, showing that our approach is dominant in many regimes of interest. We also provide a version of our estimator, which is based on linear and second-order cone programming, that avoids resolving the initial semidefinite program from scratch when few data points are added to the model. Finally, we identify unorthodox areas of application of our sum of squares-based regressors, such as real-time optimal value prediction for conic programs and approximation of optimal transport maps, where our regressors perform particularly well.

The third line of work which has commonalities with ours is [59]. It is quite similar to [38] in terms of techniques as it uses semidefinite programming and sum of squares to enforce monotonicity, this time, of the polynomial regressor. Contrarily to [38], it also has a consistency result for this set-up. The major difference between our work and [59] is that [59] only considers the univariate setting.

1.2 Outline and Contributions

Our first contribution is to provide a complete complexity classification of the problem of solving (3) and (4) in terms of the degree d (Section 2.1). In particular, we prove that when $d \geq 3$, these are hard to solve. In light of this, we show how to use sum of squares (sos) polynomials—a brief review of which is given in Section 2.2—to construct arbitrarily accurate approximations of the solutions to problems (3) and (4) (Section 2.3). We call these estimators the Sum of Squares-based Estimators (SOSE). In Section 3, we provide a proof of consistency of the SOSEs: this is a fundamental statistical property of estimators, guaranteeing that we can recover the true underlying function as the number m of observations grows. We then show in Section 4.1 that the SOSE outperforms the CLSE introduced in Section 1.1, both in terms of training/prediction time and mean-squared test error in many regimes of interest (parametrized by m and n). In Section 4.2, we apply our methodology to a a well-known dataset (the KLEMS database) which appears in economics and relates production of a sector back to capital, labor, energy, materials, and services. We observe that we are able to outperform the traditional approach (which uses Cobb-Douglas functions) on 50 out of the 65 sectors listed in the KLEMS database. We follow this up, in Section 4.3, with an offline-online procedure which enables us to quickly update our original estimator when a small number of new data points are added to the existing set: instead of solving the full semidefinite program from scratch, we instead solve a linear or second-order cone program. Section 5 presents two interesting outlets for shape-constrained regression. In Section 5.1, we show that predicting the optimal value of a conic program from past (data, optimal value) pairs can be couched as a shape-constrained regression problem in some parameters of the program: we use this insight to construct a tool for contract negotiations in inventory management, though the approach is more widely applicable. In Section 5.2, we describe how our techniques can be used to obtain an optimal transport map with certain desirable properties. Focusing on a color transfer application, we show that using the SOSE (as compared to the CLSE) enables us to dispense with image pre-processing steps which were previously necessary. All code can be found at `https://tinyurl.com/9cs9rbtw`.

2 Complexity Classification Based on d and Sum of Squares Approximations

It is natural to wonder whether problems (3) and (4) can be solved in time polynomial in the input m,n, B, and $\{(X_i, Y_i)\}_i$, for certain ranges of d. To this effect, we provide a complete characterization of the complexity of solving (3) and (4) based on d in Section 2.1. We then propose an sos-based approximation of (3) and (4) in Section 2.3 after a quick review of sos theory in Section 2.2.

2.1 Complexity Classification Based on d

In this subsection, we work in the standard Turing model of computation (see, e.g., [57]), where the input to every problem instance must be defined by a finite number of bits. As a consequence, the input to all decisions problems we consider is always assumed to be rational. We now introduce the decision versions of the optimization problems (3) and (4).

Definition 3. Let BD-DER-REG-d be the following decision problem. Given a box B as in (1), i.e., (l_i, u_i), $i = 1, \ldots, n$, data points $(X_i, Y_i) \in B \times \mathbb{R}, i = 1, \ldots, m$, a vector K as defined in Definition 2, and a rational number t, decide whether there exists a polynomial p of degree d with K-bounded derivatives over B such that

$$\sum_{i=1}^{m} (Y_i - p(X_i))^2 \leq t.$$

Definition 4. Let CONV-REG-d be the following decision problem. Given a box B as in (1), i.e., (l_i, u_i), $i = 1, \ldots, n$, data points $(X_i, Y_i) \in B \times \mathbb{R}, i = 1, \ldots, m$, and a rational number t, decide whether there exists a polynomial p of degree d that is convex over B such that

$$\sum_{i=1}^{m} (Y_i - p(X_i))^2 \leq t.$$

Theorem 2.1. *BD-DER-REG-d is strongly NP-hard for $d \geq 3$ and is in P for $d \leq 2$. CONV-REG-d is strongly NP-hard for $d \geq 3$, can be reduced to a semidefinite program of polynomial size for $d = 2$, and is in P for $d = 1$.*

The proof of Theorem 2.1 is given in Appendix A.

2.2 Review of Sum of Squares Polynomials

We say that a polynomial p of degree $2d$ and in n variables is a *sum of squares* (sos) polynomial if p can be written as

$$p(x_1, \ldots, x_n) = \sum_{i=1}^{r} q_i^2(x_1, \ldots, x_n)$$

for some polynomials q_i of degree d and in n variables. We denote by $\Sigma_{n,2d}$ the set of sos polynomials in n variables and of degree at most $2d$. Sos polynomials combine a few characteristics that make them very useful in practice. First, testing membership to $\Sigma_{n,2d}$ can be reduced to a semidefinite program of polynomial size. Indeed, a polynomial $p(x_1, \ldots, x_n)$ of degree $2d$ is sos if and only if there exists a positive semidefinite matrix Q such that $p(x) = z(x)^T Q z(x)$, where $z(x) = (1, x_1, \ldots, x_n, \ldots, x_n^d)^T$. Second, sum of squares polynomials can be used to algebraically certify nonnegativity of a polynomial over a *basic semialgebraic set*, i.e., a set defined by a finite number of polynomial inequalities. We make use of such an algebraic certificate here, or *Positivstellensatz*, due to Putinar.

Theorem 2.2 (Putinar)**.** *Let $g_1, \dots, g_s$ be polynomials in n variables such that the set*

$$\Omega := \{x \in \mathbb{R}^n \mid g_1(x) \geq 0, \dots, g_n(x) \geq 0\}$$

is Archimedean. If a polynomial p is positive on Ω, then there exist sos polynomials $s_0, \dots, s_s$ such that

$$p(x) = s_0(x) + s_1(x)g_1(x) + \ldots + s_s(x)g_s(x). \tag{5}$$

The sets which we consider are boxes, and thus the Archimedean condition (which is a stronger requirement than compactness [32]) is trivially satisfied in our case. When the degree of the sos polynomials involved in (5) is fixed, searching for a certificate of positivity of a polynomial over Ω is a semidefinite program. Furthermore, such a certificate is guaranteed to exist if the degree of the sos polynomials is large enough. The caveat of course is that one does not know a priori how high the degree of the sum of squares polynomials must be.

Another concept that will be useful to us in Section 2.3 is that of *sum of squares matrices*. Recall that a polynomial matrix is a matrix with entries that are polynomials. We say that a $t \times t$ polynomial matrix $M(x)$ is an *sos matrix* if there exists a $t' \times t$ polynomial matrix $V(x)$ such that $M(x) = V(x)^T V(x)$. This is equivalent to requiring that, for $y \in \mathbb{R}^n$, $y^T M(x) y$ be a sum of squares (polynomial) in x and y. As a consequence, testing whether a given polynomial matrix is an sos matrix can again be done by solving a semidefinite program. We denote by $\Sigma^M_{n,2d,t}$ the set of sos matrices of size $t \times t$ and with entries that are polynomials of degree at most $2d$ and in n variables. Scherer and Hol [55] generalized (2.2) to this setting.

Theorem 2.3 (Scherer and Hol)**.** *Let $g_1, \dots, g_s$ and Ω be as defined in (2.2). If a symmetric polynomial matrix $H(x)$ is positive definite on Ω (i.e., if $H(x) \succ 0, \forall x \neq 0$ in Ω), then there exist sos matrices $S_0(x), S_1(x), \dots, S_s(x)$ such that*

$$H(x) = S_0(x) + g_1(x) \cdot S_1(x) + \ldots + g_s(x) \cdot S_s(x).$$

2.3 Sum of Squares Approximations

Theorem 2.1 shows that optimization problems (3) and (4) are hard to solve when $d \geq 3$. We make use of sum of squares theory to construct sum of squares-based approximations to (3) and (4). For (3), we replace the constraint $H_g(x) \succeq 0, \forall x \in B$, by a sum of squares-based condition as indicated in Theorem 2.3. Likewise, for (4), we replace the constraints $K_i^- \leq \frac{\partial g(x)}{\partial x_i} \leq K_i^+$ for all $i = 1, \dots, m$ and $x \in B$ by sum of squares-based constraints as indicated in Theorem 2.2.

When using both Theorems 2.2 and 2.3, we consider $\Omega = B$. Note that both theorems depend not only on the *set* that Ω defines, but on the *way* it is defined. We choose to use the following representation of B,

$$B = \{(x_1, \dots, x_n) \mid (u_i - x_i)(x_i - l_i) \geq 0, i = 1, \dots, n\}, \tag{6}$$

which is different, but equivalent, to that given in (1). This is because this particular representation enables us to take $g_i(x) = (u_i - x_i)(x_i - l_i), i = 1, \dots, n$, thus leading to only n defining inequalities of B rather than $2n$ as we would have had, had we used the representation given in (1). Moving forward, we also assume that m is large enough and that the data points $\{(X_i, Y_i)\}_i$ are linearly independent so that the solutions to (3) and (4) are unique.

Definition 5. We define $\tilde{g}_{m,d,r}$ to be the solution to the optimization problem:

$$\begin{aligned} \tilde{g}_{m,d,r} := \arg & \min_{g \in P_{n,d}, S_0, \ldots, S_n \in \Sigma^M_{n,2r,n}} \sum_{i=1}^{m} (Y_i - g(X_i))^2 \\ \text{s.t.} \quad & H_g(x) = S_0(x) + g_1(x) S_1(x) + \ldots + g_n(x) S_n(x). \end{aligned} \tag{7}$$

Definition 6. We define $\tilde{h}_{m,d,r}$ to be the solution to the optimization problem:

$$\begin{aligned} \tilde{h}_{m,d,r} := \arg & \min_{h \in P_{n,d}, s^+_{ij}, s^-_{ij} \in \Sigma_{n,2r}} \sum_{i=1}^{m} (Y_i - h(X_i))^2 \\ \text{s.t.} \quad & K^+_i - \frac{\partial h(x)}{\partial x_i} = s^+_{i0}(x) + s^+_{i1}(x) g_1(x) + \ldots + s^+_{in}(x) g_n(x), i = 1, \ldots, n, \\ & \frac{\partial h(x)}{\partial x_i} - K^-_i = s^-_{i0}(x) + s^-_{i1}(x) g_1(x) + \ldots + s^-_{in}(x) g_n(x), i = 1, \ldots, n. \end{aligned} \tag{8}$$

We refer to $\tilde{g}_{m,d,r}$ and $\tilde{h}_{m,d,r}$ as the Sum of Squares Estimators (SOSEs) of the underlying shape-constrained function f.

Theorem 2.4. *Let $\bar{g}_{m,d}$ and $\bar{h}_{m,d}$ be as defined in (3) and (4). For fixed d and m, we have*

$$\lim_{r \to \infty} \sup_{x \in B} |\bar{g}_{m,d}(x) - \tilde{g}_{m,d,r}(x)| \to 0 \tag{9}$$

$$\text{and} \ \lim_{r \to \infty} \sup_{x \in B} |\bar{h}_{m,d}(x) - \tilde{h}_{m,d,r}(x)| \to 0. \tag{10}$$

The proof of this theorem is more complex than one would think and is given in Appendix B.

Remark 2.1. The approximations $\tilde{g}_{m,d,r}$ and $\tilde{h}_{m,d,r}$ of $\bar{g}_{m,d}$ and $\bar{h}_{m,d}$ have some appreciable characteristics. For instance, as they are polynomials, they are smooth functions. They are also consistent estimators of f, as we show in Section 3, under certain assumptions on the way the data points $(X_i, Y_i), i = 1, \ldots, m$ are generated. Furthermore, one can produce sum of squares certificates to certify that $\tilde{g}_{m,d,r}$ is indeed convex over B and $\tilde{h}_{m,d,r}$ has bounded derivatives. In terms of computation, the size of the semidefinite programs that need to be solved to obtain $\tilde{g}_{m,d,r}$ and $\tilde{h}_{m,d,r}$ scale polynomially in the number n of features. Adding additional data points to the problem does not impact the size of the semidefinite program as it only adds terms to the objective. We provide numerical evidence of these characteristics in Section 4.1. We also show that while Theorem 2.4 theoretically holds when r goes to the limit, in practice, even small values of r give rise to good-quality estimators.

Remark 2.2. In some cases, problems (3) and (4) and problems (7) and (8) are equivalent for some value of r, known a priori. This is the case when $d = 1$ and when $d = 2$, as shown in Theorem 2.1. Another such case is when $\bar{g}_{m,d}$ and $\bar{h}_{m,d}$ are separable functions. To see this, one can use the results of Polya and Szego, see, e.g. [50], to rewrite (3) and (4) as sum of squares programs of low degree.

3 Consistency of the Sum of Squares-Based Estimators

Up until now, we have simply assumed that we are given m data points (X_i, Y_i), where $X_i \in \mathbb{R}^n$ and $Y_i \in \mathbb{R}$, without making explicit assumptions on how X_i and Y_i are obtained. In this section, we make three new assumptions.

Assumption 1. The vectors $X_1, \dots, X_m$ are randomly generated and are independently and identically distributed (iid) with $E[||X_1||^2] < \infty$.

Assumption 2. The support of the random vectors $X_1, \dots, X_m$ is a full- dimensional box $B \subseteq \mathbb{R}^n$ defined as in (1). In other words, $P(X_i \in B) = 1$. Furthermore, we assume that for any full-dimensional set $C \subseteq B$, $P(X_i \in C) > 0$.

Assumption 3. There exists a continuous function $f : B \to \mathbb{R}$ such that

$$Y_i = f(X_i) + \nu_i, \forall i = 1, \dots, m,$$

where ν_i are random variables with support $\mathbb{R}$ and the following characteristics:

$$P(\nu_1 \in dz_1, \dots, \nu_m \in dz_n | X_1, \dots, X_m) = \prod_{i=1}^{m} P(\nu_i \in dz_i | X_i), \forall z_1, \dots, z_m \in \mathbb{R}, \tag{11}$$

$$E[\nu_i | X_i] = 0 \text{ almost surely (a.s.)}, \forall i = 1, \dots, m, \tag{12}$$

$$E[\nu_i^2] =: \sigma^2 < \infty \ \forall i = 1, \dots, m. \tag{13}$$

Assumptions 1 and 3 imply that the sequence $\{(X_i, Y_i)\}_{i=1,\dots,m}$ is iid, that $E[\nu_1] = 0$, and that $E[Y_1^2] < \infty$.

Using these three assumptions, we show that the regressors $\tilde{g}_{m,d,r}$ and $\tilde{h}_{m,d,r}$ are *consistent* estimators of f, provided that f is respectively twice continuously differentiable and convex over B (Theorem 3.1) or has K-bounded derivatives over B (Theorem 3.2). This is a key property of estimators and shows that as the number of observations grows, we are able to recover the true underlying relationship between X_i and Y_i.

Theorem 3.1. *Let C be any compact full-dimensional subset of B such that no point on the boundary of B is in C. Assuming that f is twice continuously differentiable and convex over B, that $\tilde{g}_{m,d,r}$ is as defined in (7), and that Assumptions 1 through 3 hold, we have*

$$\sup_{x \in C} |\tilde{g}_{m,d,r}(x) - f(x)| \to 0 \text{ almost surely (a.s.)} \tag{14}$$

as $d, m, r \to \infty$.

Theorem 3.2. *Let C be any compact full-dimensional subset of B such that no point on the boundary of B is in C. Let $K = (K_1^-, K_1^+, \dots, K_n^-, K_n^+)$ be a vector of finite scalars with $K_i^- < K_i^+$ for all $i = 1, \dots, n$. Assuming that f has K-bounded derivatives over B, that $\tilde{h}_{m,d,r}$ is as defined in (8), and that Assumptions 1 through 3 hold, we have*

$$\sup_{x \in C} |\tilde{h}_{m,d,r}(x) - f(x)| \to 0 \text{ a.s.} \tag{15}$$

as $d, m, r \to \infty$, i.e., $P(\lim_{d,m,r\to\infty} \sup_{x\in C} |\tilde{h}_{m,d,r}(x) - f(x)| \to 0) = 1$.

Remark 3.1. Consistency is a property that by definition involves limits. This does not mean that one need necessarily take large d or r for $\tilde{g}_{m,d,r}$ and $\tilde{h}_{m,d,r}$ to be good approximations of f. In fact, in (4.1), we see that in practice r and d can be taken to be small without this adversely affecting the mean squared error on the test set. The one parameter which it may be of interest to increase to improve the fit is m, which is a non-issue in our case as the size of our semidefinite programs does not scale with m.

Remark 3.2. One could extend these theorems to the box B itself, provided that we assume stronger assumptions on the sampling of the pairs of points $(X_i, Y_i)_{i=1,\ldots,m}$. Namely, we would need to assume that a non-negligeable fraction of the sample is located at the vertices of B. As this is unlikely to occur in practice, we have chosen to instead show this version of the theorem, which comes with much more reasonable assumptions on the sampling of the data.

The proofs of Theorems 3.1 and 3.2 follow immediately from Theorem 2.4, the two lemmas below, and the triangle inequality.

Lemma 3.3. *Let C be any compact full-dimensional subset of B such that no point on the boundary of B is in C. Assuming that f is twice continuously differentiable and convex over B, that $\bar{g}_{m,d}$ is as defined in (3), and that Assumptions 1 through 3 hold, we have*

$$\sup_{x \in C} |\bar{g}_{m,d}(x) - f(x)| \to 0 \text{ a.s.} \tag{16}$$

as $m \to \infty$ and $d \to \infty$.

Lemma 3.4. *Let C be any compact full-dimensional subset of B such that no point on the boundary of B is in C and let $K = (K_1^-, K_1^+, \ldots, K_n^-, K_n^+)$ be a vector of finite scalars with $K_i^- < K_i^+$ for all $i = 1, \ldots, n$. Assuming that f has K-bounded derivatives, that $\bar{h}_{m,d}$ is as defined in (4), and that Assumptions 1 through 3 hold, we have*

$$\sup_{x \in C} |\bar{h}_{m,d}(x) - f(x)| \to 0 \text{ a.s.} \tag{17}$$

as $m \to \infty$ and $d \to \infty$.

For ease of reading, we defer their proofs to Appendix C. We comment here, however, on the similarities with the proof of consistency of the CLSE in [34], from which our proofs are inspired. The major differences between both proofs are artefacts of the differences between the estimators under consideration. The CLSE is a piecewise linear function, as mentioned before, and only depends on one parameter, m. The estimators appearing in (3.3) and (3.4) are polynomial functions relying on two different parameters, m and d. Thus, before we can even apply proof techniques given in [34], we have to work our way from $\bar{g}_{m,d}$ and $\bar{h}_{m,d}$ to estimators that only vary with m. This requires us to show, among other results, Weierstrass-type theorems for appropriately smooth functions that are convex or have bounded-derivatives over a box. Even when this work is done, we cannot directly apply the results in [34]. Indeed, [34] assumes that $X_1, \ldots, X_m$ are sampled from $\mathbb{R}^n$, which is an assumption that we cannot make (in light, e.g., of our Weierstrass-type results). This requires us to rework parts of the proof given in [34] as detailed in Appendix C. Finally, we also include a proof of consistency of the bounded-derivative case which is only partially included in [34], as [34] considers the monotonous case only.

4 Computation and Numerical Performance of the SOSEs

Obtaining the SOSEs requires solving a semidefinite program. In this section, we show experimentally that neither the degree d of the SOSEs, nor the degree r of the multipliers need be very large to obtain good results on the test set. We also provide a thorough experimental comparison to the CLSE and highlight the settings where it is preferable to use the SOSEs. We follow this up with a short application of the SOSE to one of the poster children of shape-constrained regression: fitting a production function to data. We show that we outperform the prevalent approach in economics (relying on Cobb-Douglass functions) on a large majority of industries under consideration. We then conclude with a method for updating an existing SOSE when a small number of new data points are added to the initial set, which does not involve solving a semidefinite program.

4.1 Performance of the SOSEs

Our experimental setup for this section is the following: we generate m data-points X_i uniformly from the n-dimensional unit box $[0,1]^n$. To obtain the corresponding responses Y_i, we let

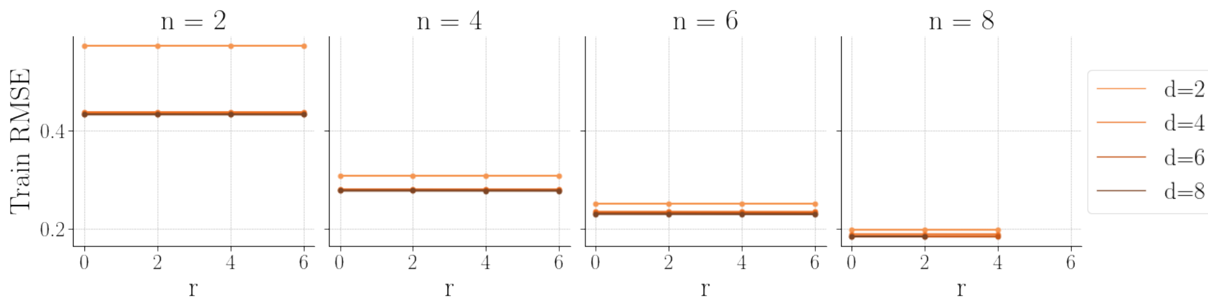
$$f(z) = f(z_1, \ldots, z_n) = (z_1 + \ldots + z_n) \cdot \log(z_1 + \ldots + z_n)$$

and take $Y_i = f(X_i) + \nu_i$ for all $i = 1, \ldots, m$, where ν_i is randomly generated following a standard Gaussian. We remark that f is purposefully chosen to *not* be a polynomial so as to not give an unfair advantage to our method. Furthermore, f is convex (but not monotonous) so we fit a convex-constrained polynomial $\tilde{g}_{m,d,r}$ to the data. Similar observations to the ones we make here can be obtained for different f and for the bounded-derivatives SOSE, $\tilde{h}_{m,d,r}$. For each experiment, we specify the values taken on by m, n, d, and r. The train root mean squared error (RMSE) corresponds to the square root of the objective value of the optimization problem (7) and the optimization solver time is given in seconds. The test RMSE is obtained by generating an additional 1000 points X_i and evaluating the predicted values against the true values $f(X_i)$ (without adding noise). All experiments were performed on a MacBook Pro (2.6 GHz 6-Core Intel Core i7 processor and 16 GB RAM). To solve the semidefinite programs obtained, we used MATLAB R2020b, MOSEK [1], and YALMIP (release R20200930) [36]. If we further wished to speed up the solving times of our semidefinite programs, we could use recent approaches to semidefinite programming [47, 39] which are more scalable, such as first-order methods [45, 33]. We do not do so here as our results are already competitive using MOSEK.

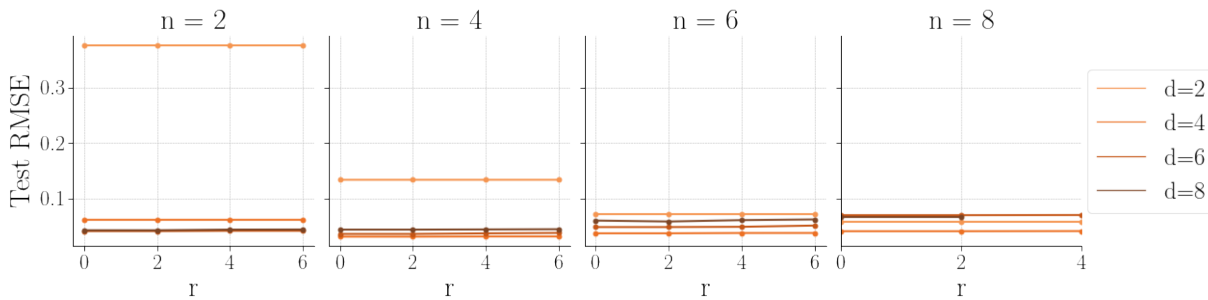
We start this subsection by considering the effect of r and d on the quality of approximation of f. As mentioned in Remark 2.1, $\tilde{g}_{m,d,r}$ can be obtained via a semidefinite program whose size is not impacted by the number m of training points. Its size scales polynomially however with the number n of features (when d is fixed) and with d and r (when n is fixed). While the value of n is typically an artifact of the application, the values of d and r can be tuned as needed so that they provide a good regressor. Results such as Theorem 3.1 may seem to advocate for large d and r to obtain a good regressor, leading to semidefinite programs of a problematic size. Fortunately, our experimental results (both in this experimental setup and throughout the rest of the paper) show that neither d nor r need be large to obtain good-quality regressors. Figure 1 illustrates that increasing r seems to have little to no impact on the test RMSE whereas increasing d has diminishing returns, if not negative effects, on the test RSME with the best test RMSE achieved typically for $d = 4$ or $d = 6$. These results are further borne out by our comparison to the CLSE below. In fact, one can view the choice of r and d as a way of regularizing the shape-constrained regression problem: the smaller the values of r and d, the more regularized our problem is.

We now compare the SOSE to the CLSE introduced in [27, 28]. Obtaining the SOSE involves solving SDP (7), which is done as explained above—in particular, we do not resort to algorithms tailored to the problem. Obtaining the CLSE involves solving a QP, on the other hand; see, e.g., [34]. In a naive implementation, this QP has a number of variables that scales linearly and a number of constraints that scale quadratically with the number m of data points[2]. We do not use a naive implementation to compute the CLSE however, but make use of the companion code of [15] available at `https://github.com/wenyuC94/ConvexRegression`. This corresponds to a tailored algorithm for the problem which incorporates heuristics to reduce the number of constraints and improve the scalability of the method. Despite the disparity in methods used to solve the two problems, the SOSE is nevertheless much faster to compute in settings where the number of data points is moderate to large ($m = 2000$ to $m = 10,000$); see Figure 2. In fact, for most of the

[2] As mentioned previously, the size of the SDP does not scale with m. Both the SDP and the QP scale polynomially in n.



(a) Train RMSE.

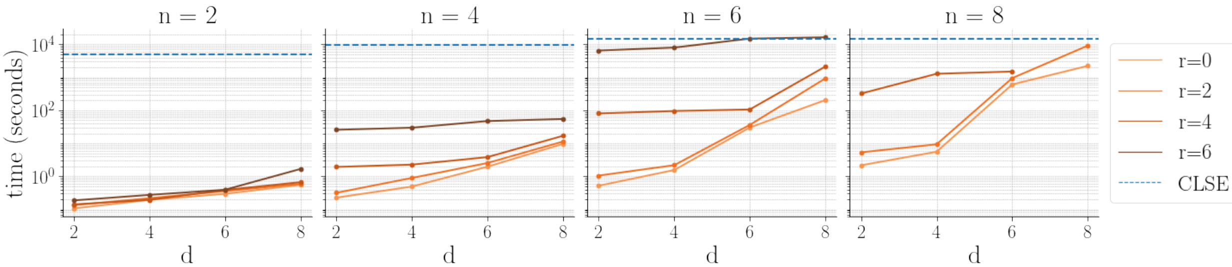


(b) Test RMSE.

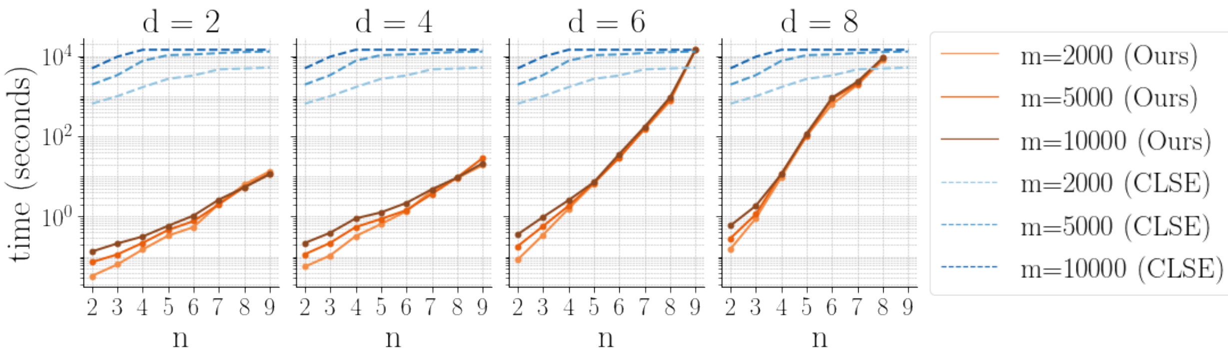
Figure 1: Train and Test RMSE for the convex-constrained regression problem in (7) with $m = 10,000$ data points generated as described in Section 4.1, as the number n of features, the degree d of the polynomials, and the degree r of the multipliers vary. Lighter color corresponds to lower d.

QPs solved for these values of m, the max run-time of 4 hours has been reached. The results in Figure 2 thus suggest that the SOSE should be prioritized in settings where m is large but n is small, whereas the CLSE should be prioritized in settings where m is small but n is larger. The two applications in Section 5 correspond to the former setting and so are particularly adapted to the SOSE.

In terms of predictions, the SOSE method simply evolves evaluating a polynomial function (the SOSE) at the new data point, whereas obtaining a new prediction from the CLSE involves solving an LP whose size scales quadratically in m. A comparison of prediction times for the SOSE and the CLSE is given in Table 2: it takes typically less than 1/100 of a second to compute a prediction with the SOSE, whereas it can take up to 300 seconds to compute a prediction with the CLSE (for $m = 500$). For m larger than 1000, it becomes infeasible to do so. Two additional issues can occur when evaluating the CLSE on new data points, i.e., solving the LP. The first issue is that the CLSE estimator is defined only for points within the convex hull of the train data points: solving the LP for a point outside of this set results in an unbounded problem. If one cannot design the training set, this can be problematic if the number n of features increases, as the number m of data points cannot follow suit; see the last column in Table 1 where, in some cases, more than 50% of the test points fall outside of the convex hull. In contrast, the SOSE is well-defined over the full box B and can compute predictions for any point in B, regardless of the location of the points in the training set. The second issue is that, if the solution to the QP is only approximate (e.g., if it is obtained via first-order methods), then the LP becomes infeasible. In other words, if the CLSE slightly violates its convexity constraints, obtaining a new prediction becomes impossible. This is again in contrast to the SOSE: while using first-order solvers could lead to a solution which



(a) Solver time of (7) for different numbers n of features and polynomials of different degree d and helper degree r when using $m = 10,000$ training samples. The dashed blue line corresponds to the solver time for the CLSE QP.



(b) Solver time needed to compute the CLSE and the SOSE in (7) with respect to number n of features. For the SOSE, we take $r = 2$ and a range of degrees d. Lighter colors correspond to fewer training points. All run-times capped at 4 hours per setup.

Figure 2: Runtime comparison of our method and CLSE.

slightly violates the shape constraints in this case as well, it does not affect the ability to obtain new predictions.

Comparing the quality of the SOSE against that of the CLSE requires us to compute the test RMSE for each, which involves computing many new predictions. To circumvent the prediction issues for the CLSE described above, we take m to be in $\{100, 200, 500\}$. We also remove all unbounded predictions obtained and introduce slack variables to the LP so as to avoid infeasibility. Train and test RMSEs are then given in Table 1 for both methods. We see that the CLSE tends to overfit the training data as across all of the (m, n) settings, the CLSE has a lower train RMSE than the SOSE, yet has higher test RMSE. Furthemore, the test RMSE of our method improves with the sample size, which is not the case for the CLSE. We note that these predictions are obtained for low values of d and for $r = 2$, further illustrating that these need not be large for the SOSE to be of good-quality.

4.2 Fitting a Production Function to Data

Fitting a production function to data is one of the best-known applications of shape-constrained regression and thus provides a good (first) testing ground for the SOSE. The goal is to estimate the functional relationship between the yearly inputs of Capital (K), Labor (L), and Intermediate goods (I) to an industry, and the yearly gross-output production Out of that industry. As Out is assumed to be a decreasing function in K, L, and I, as well as concave in K, L, and I by virtue of diminishing returns, an estimator constrained to have this shape is desirable. Traditionally, in

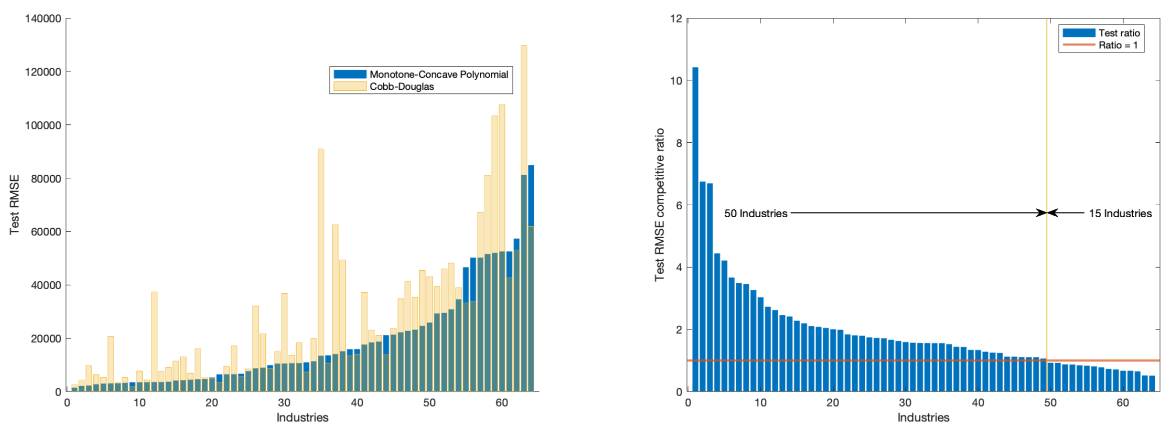
		SOSE						CLSE		
		train RMSE			test RMSE			train	test	% un-
m	n	$d=2$	$d=4$	$d=6$	$d=2$	$d=4$	$d=6$	RMSE	RMSE	bounded
100	2	0.537	0.415	0.399	0.343	**0.134**	0.154	0.346	0.352	4
	3	0.404	0.371	0.363	0.178	**0.105**	0.192	0.258	0.194	26
	4	0.228	0.191	0.177	**0.176**	0.206	0.211	0.099	0.244	52
	5	0.279	0.216	0.187	**0.213**	0.228	0.463	0.093	0.313	76
	6	0.196	0.145	0.109	**0.151**	0.215	0.652	0.046	0.329	84
200	2	0.551	0.442	0.429	0.327	**0.080**	0.115	0.379	0.415	2
	3	0.407	0.362	0.359	0.169	**0.062**	0.086	0.288	0.373	16
	4	0.267	0.227	0.218	0.134	**0.134**	0.174	0.135	0.315	38
	5	0.272	0.241	0.219	**0.185**	0.205	0.323	0.104	0.299	66
	6	0.224	0.186	0.163	**0.130**	0.174	0.348	0.059	0.285	74
500	2	0.610	0.445	0.437	0.315	**0.076**	0.082	0.394	0.626	0
	3	0.414	0.351	0.349	0.160	0.035	**0.031**	0.297	0.543	6
	4	0.294	0.267	0.263	0.132	**0.062**	0.066	0.205	0.482	22
	5	0.269	0.248	0.241	0.158	**0.110**	0.120	0.162	0.466	40
	6	0.252	0.226	0.216	0.112	**0.080**	0.093	0.128	0.370	56

Table 1: Comparison of the train and test RMSEs for the SOSE $\tilde{g}_{m,d,r}$ computed in (7) (with $r = 2$) and for the CLSE, for different values of m and n. Also includes the percentage of randomly generated test points that give rise to an unbounded prediction; see Section 4.1. Best test RMSE marked in bold font.

	SOSE			CLSE		
n	d=2	d=4	d=6	m=100	m=200	m=500
2	0.0015	0.0068	0.0178	8.99	35.59	239.7
3	0.0018	0.0073	0.0206	9.22	36.25	239.15
4	0.0019	0.0086	0.0237	9.18	36.36	247.49
5	0.0021	0.0087	0.0296	9.27	36.68	247.77
6	0.0027	0.0096	0.0355	9.35	37.39	255.97

Table 2: Comparison of the time it takes (in seconds) to make one new prediction for the SOSE (as a function of n and d) and for the CLSE (as a function of n and m).

the economics literature, this is done by fitting a *Cobb-Douglas production function* to the data, i.e., by finding (a, b, c, d) such that the function $Out = a \cdot K^b \cdot L^c \cdot I^d$ is as close as possible to the observed data. The advantage of such an approach is that it can be couched as a linear regression problem by working in log-space, with the shape constraints being imposed via the constraints $b, c, d \geq 0$, $b + c + d \leq 1$ and $a \geq 0$. We therefore compare the SOSE to the Cobb-Douglas estimator in this subsection. To fit the estimators, we consider the USA KLEMS data (available at `http://www.worldklems.net/data.htm`), which contains yearly gross-output production data Out for 65 industries in the US, from 1947 to 2014 as well as yearly inputs of Capital K, Labor L, and Intermediate goods I, adjusted for inflation. Since the data is temporal, we perform a temporal split for our training-testing splits. We then fit the Cobb-Douglas estimator and the SOSE with degree $d = 4$ and $r = 2$ and the aforementioned shape constraints to the data. The results obtained are given in Figure 3. As can be seen, our method outperforms the traditional Cobb-Douglass technique on 50 out of the 65 industries, sometimes quite significantly. This is in spite of d and r being very low, which once again illustrates that r and d need not be very high to obtain competitive results against other methodologies.



(a) (b)

Figure 3: Comparison of the Test RMSE for the Cobb-Douglass production functions and the SOSE with the same shape constraints (concavity and monotonicity) across 65 industries; see Section 4.2. In Figure 3a, the test RMSE values obtained, and in Figure 3b, the ratio of the Cobb-Douglass RMSE over the SOSE RMSE across industries.

4.3 An Offline-Online Method for Computing the SOSEs

We now consider a setting that can occur in practice. Suppose we are given an initial training set $(X_1, Y_1), \ldots, (X_M, Y_M)$, containing M data point pairs. We obtain via the sum of squares program in (7) or (8) the SOSEs, for appropriate values of d and r. We are then provided with M' new data point pairs (X_{M+1}, Y_{M+1}), $\ldots$, $(X_{M+M'}, Y_{M+M'})$, $M' \ll M$ which we would like to incorporate to our existing model as training data, without having to resolve the sos program from scratch and incur too much additional computational cost. We propose a method to do so, which leverages some concepts given in [4]. The idea is that, while we would infrequently re-solve the SDP, and we would so offline, we would use this method more frequently to update our estimator (e.g., for each arrival of a batch of M' new data points) in an *online* fashion.

For exposition purposes, we present this method in the context of imposing convexity only on the regressor, and *global* convexity rather than convexity over a box. The exact same techniques hold for convexity over a box and other shape constraints. In the global convexity setting, finding a convex-constrained regressor $\tilde{g}^{global}_{n,2d,n}$ simply involves solving the following semidefinite program:

$$\begin{aligned} \tilde{g}^{global}_{n,2d,n} := \arg \quad & \min_{g \in P_{n,2d}} \sum_{i=1}^{M} (Y_i - g(X_i))^2 \\ \text{s.t.} \quad & H_g(x) \in \Sigma^M_{n,2d-2,n}. \end{aligned} \tag{18}$$

We assume that (18) has already been solved and that we have an optimal polynomial $\tilde{g}^{global}_{n,2d,n}$. Following Section 2.2, this implies that $y^T H_{\tilde{g}^{global}_{n,2d,n}}(x) y$ is sos, and that there exists a matrix $Q^* \succeq 0$ such that

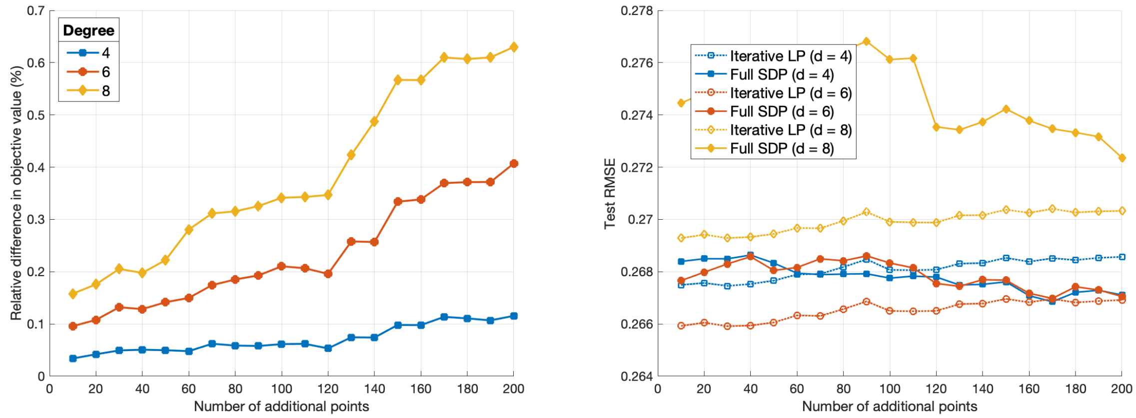
$$y^T H_{\tilde{g}^{global}_{n,2d,n}}(x) y = w(x,y)^T Q^* w(x,y),$$

where $w(x,y)$ is the vector of all monomials of degree 1 in $(y_1, \ldots, y_n)$ and up to $d-1$ in $(x_1, \ldots, x_n)$. We now consider the arrival of M' new training points to the system: this only changes the coefficients of the objective function. If $M' \ll M$ and the data points are generated from the same process, one can expect these coefficients to change very little and, thus, the new optimal solution (i.e., the updated regressor) should be quite close to the previous solution. Resolving (18) from scratch with a marginally changed objective seems computationally inefficient in consequence, while not updating the regressor to take into consideration the new data is suboptimal. An intermediary to these two methods, corresponding to our online step, is to solve the following optimization problem:

$$\begin{aligned} \tilde{g}^{updated}_{n,2d,n} := \arg \quad & \min_{g \in P_{n,2d},\ Q \in DD_N} \sum_{i=1}^{M+M'} (Y_i - g(X_i))^2 \\ \text{s.t.} \quad & H_g(x) = w(x,y)^T U^* Q U^* w(x,y), \end{aligned} \tag{19}$$

where $N = n \cdot \binom{n+d-1}{d-1}$, U^* is the matrix square root of Q^*, and the set DD_N refers to the set of symmetric diagonally dominant matrices of size $N \times N$. (Recall that a matrix M is diagonally dominant if $M_{ii} \geq \sum_{j \neq i} |M_{ij}|$ for all $i = 1, \ldots, n$.) As any diagonal dominant matrix is positive semidefinite by virtue of Gershgorin's circle theorem [21], $\tilde{g}^{updated}_{n,2d,n}$ will have the required shape constraint. Furthermore, by selecting $Q = I$, we are able to recover the solution to (18): it can thus only be that $\tilde{g}^{updated}_{n,2d,n}$ improves on $\tilde{g}^{global}_{n,2d,n}$. The polynomial $\tilde{g}^{updated}_{n,2d,n}$ is obtained however at considerably less expense than what we would have incurred if we had replaced the constraint $Q \in DD_N$ by the constraint $Q \succeq 0$ in (19), or equivalently, added the term $\sum_{i=M+1}^{M'} (Y_i - g(X_i))^2$ to (18). Indeed, (19) is a *linear program* (LP), not a semidefinite program. Furthermore, starting from $Q = I$ should position us quite close to the true optimal solution. When a new batch of M' datapoints arrives, one simply recomputes a new U^*, obtained as the matrix square root of the previous solution Q^*, and resolves.

We illustrate the approach in Figure 4 with $n = 4$. There, we solve the initial problem (18) with $M = 1000$ points. We then receive batches of $M' = 10$ points and compute new regressors at each arrival, either by solving (18) from scratch (solid line) or by using the iterative process described above (dashed line). From the figures, we can see that the optimality gap between the two methods is very small throughout (Figure 4a), even when adding a considerable number of data points (though the optimality gap grows as expected as the number of points increases). The differences in terms of test RMSE are in the one thousandths (Figure 4b), which suggests that little



(a) (b)

Figure 4: Comparison of the quality of the regressors obtained by re-solving the semidefinite program (18) from scratch or solving the linear program (19) iteratively; see Section 4.3. Figure 4a plots for varying degrees d the relative difference (%) in the objective values of (18) and (19) solved iteratively. Figure 4b shows for varying degrees d the test RMSE of (18) (solid lines) and the test RMSE of (19) solved iteratively (dashed lines).

is lost in terms of quality via these updates, though much is to be gained in terms of solving times (e.g., when $n = 8$ and $d = 6$, the speedup is of order 10; when $n = 6$ and $d = 8$, the speedup is of order 20). If a higher-quality solution is required at the expense of slightly increased running times, one can replace DD_n by the set SDD_n of scaled diagonally dominant matrices to end up with a second-order cone program (SOCP); see [4] for the definition of SDD_n and more details.

5 Applications of Shape-Constrained Regression to Optimization and Optimal Transport

In this section, we provide two original applications of shape-constrained regression, where SOSEs are particularly valuable. The first one involves approximating the function that maps a conic program's data to its optimal value by an SOSE. The second application focuses on optimal transport.

5.1 Predicting the Optimal Value of a Conic Program

Let $\mathcal{K} \subseteq \mathbb{R}^n$ be a proper cone, $\langle \cdot, \cdot \rangle$ be an inner product on $\mathbb{R}^n \times \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $c \in \mathbb{R}^n$. We denote by $\mathcal{K}^*$ the dual cone associated to $(\mathcal{K}, \langle \cdot, \cdot \rangle)$ and by A^* the adjoint operator of A. We also use $x \succeq_{\mathcal{K}} y$ to mean that $x - y \in \mathcal{K}$ and we say that $f : \mathbb{R}^n \mapsto \mathbb{R}$ is $\mathcal{K}$-nonincreasing if $x \succeq_{\mathcal{K}} y \Rightarrow f(x) \leq f(y)$. We now consider the pair of primal-dual conic programs (see, e.g., [12, Exercise 5.42] for a proof of duality):

$$v_P(b, c) := \inf_{x \in \mathbb{R}^n} \ \langle c, x \rangle \quad \text{s.t.} \ Ax \preceq_{\mathcal{K}} b \qquad (P)$$

$$v_D(b, c) := \sup_{y \in \mathbb{R}^m} \ -\langle b, y \rangle \quad \text{s.t.} \ A^* y + c = 0, \quad y \succeq_{\mathcal{K}^*} 0. \qquad (D)$$

We assume that strong duality holds, which implies that $v_P(b, c) = v_D(b, c) =: v(b, c)$. As it turns out, $v(b, c)$ satisfies a number of shape constraints.

Proposition 5.1. *Assume that strong duality holds for the primal-dual pair (P)-(D). Then, the function $v(b, c)$ is*

(i) convex in b

(ii) concave in c

(iii) $\mathcal{K}$-nondecreasing in b.

Proof. We prove each statement separately.

1. For fixed $y \in \mathbb{R}^m$, $b \mapsto -\langle b, y\rangle$ is linear. Following [12, Section 3.2.3], $v_D(b, c)$ is convex in b. As $v_D(b, c) = v(b, c)$ by strong duality, the result follows.

2. For fixed $x \in \mathbb{R}^n$, $c \mapsto \langle c, x\rangle$ is linear. Following [12, Section 3.2.3], $v_P(b, c)$ is convex in c. As $v_P(b, c) = v(b, c)$ by strong duality, the result follows.

3. We refer here to (P) by $P_{b,c}$ to reflect the dependency of (P) on b and c. Let $b \preceq_{\mathcal{K}} b'$. Any feasible solution to $P_{b,c}$ is feasible to $P_{b',c}$. Indeed, if x is a feasible solution to $P_{b,c}$, i.e., $b - Ax \in \mathcal{K}$, then $b' - Ax = b' - b + b - Ax \in \mathcal{K}$. It follows that the feasible set of $P_{b,c}$ is a subset of that of $P_{b',c}$, and so $v(b', c) \leq v(b, c)$.

□

Proposition 5.1 suggests a possible use of the SOSE, and shape-constrained regression more generally: one can compute the SOSE of $v(b, c)$ and then use it to obtain quick predictions of the optimal value of (P) for any new set of parameters (b, c), without actually having to solve (P) explicitly. This can prove useful for real-time decision making and we consider an application to real-time inventory management contract negotiation below.

5.1.1 An Application to Real-Time Inventory Management Contract Negotiation

In a single-product inventory with a finite-time horizon, the state of the inventory at time $t = 1, 2, \ldots, T$ is specified by the amount $x_t \in \mathbb{R}$ of product in the inventory at the beginning of period t. During the period, the inventory manager (or "retailer") orders $q_t \geq 0$ units of product from the supplier (we assume it arrives immediately) and satisfies external demand for $d_t \geq 0$ units of the product. Thus, the state equation of the inventory is given by $x_{t+1} = x_t + q_t - d_t$. We assume $x_1 = 0$ and allow for $x_t \leq 0$ (backlogged demand). We further enforce a minimum amount L that the retailer needs to buy, i.e., $\sum_{t=1}^{T} q_t \geq L$. The retailer wishes to minimize the overall inventory management cost. To this effect, we let $h, p, c, s \geq 0$ with $h + p \geq s$ be the respective costs[3] per period and per unit, of storing the product, backlogged demand, replenishing the inventory, and salvaging the product [7]. Following [8], we further assume that the supplier and retailer agree on a *flexible commitment contract*: at time $t = 0$, the retailer must commit to projected future orders, $w_t \in \mathbb{R}, 1 \leq t \leq T$. These do not have to be fully respected, but any deviation from them will result in penalty fees. A penalty $\alpha^{\pm}$ will be incurred per unit of excess/recess of the actual orders q_t as compared to commitments w_t and a penalty $\beta^{\pm}$ will be incurred for variations in the commitment w_t across periods. The problem that the inventory manager has to solve to obtain the

[3]For simplicity, these quantities do not vary with t. Our results do not change if they do.

minimum-cost inventory is thus:

$$\begin{aligned}\min_{x_t,q_t,w_t} \quad & \sum_{t=1}^{T}\left(h\max\{x_{t+1},0\}+p\max\{0,-x_{t+1}\}+\alpha^{+}\max\{q_t-w_t,0\}+\alpha^{-}\max\{w_t-q_t,0\}\right)\\ & +\sum_{t=1}^{T}cq_t-s\max\{x_{T+1},0\}+\sum_{t=2}^{T}\left(\beta^{+}\max\{w_t-w_{t-1},0\}+\beta^{-}\max\{w_{t-1}-w_t,0\}\right) \\ \text{s.t.} \quad & x_{t+1}=x_t+q_t-d_t, t=1,\ldots,T, \quad x_1=0, \quad \sum_{t=1}^{T}q_t\geq L, \quad q_t\geq 0,\ t=1,\ldots,T.\end{aligned} \tag{20}$$

Typically however, the demand $d_t, t=1,\ldots,T$ is assumed to be uncertain. Writing $d:=(d_1,\ldots,d_T)^T$, we assume that d belongs either to a box $\mathcal{S}_B\subseteq\mathbb{R}^T$ or to an ellipsoid $\mathcal{S}_E\subseteq\mathbb{R}^T$ as done in [8]. We also make a distinction between the variables $\{w_t\}$, which by definition have to be set upfront, and the variables $\{q_t\}$ that can be set in each time period, by letting q_t depend on prior demands $d_1,\ldots,d_{t-1}$. More specifically, we assume an *affine* dependency between q_t and d_t, as done in [8], i.e., we write:

$$q_t=q_t^0+\sum_{\tau=1}^{t-1}q_t^{\tau}d_{\tau}, t=1,\ldots,T,$$

where $\{q_t^{\tau}\}_{t,\tau}$ are additional variables. We now focus on writing the uncertain formulation of (20) in a way which is more convenient to determine shape constraints. To do so, we first eliminate the variables $\{x_t\}_t$ in (20) by noting that $x_{t+1}=\sum_{\tau=1}^{t}(q_\tau-d_\tau)$, $t=1,\ldots,T$. We then introduce new variables: $C\in\mathbb{R}$ equals the objective function of (20),

$$\begin{aligned} z_t^{\pm}&=\max\{\pm(w_t-w_{t-1}),0\}, & y_t&=\bar{h}_t\max\{x_t+1,0\}+p\max\{0,-x_{t+1}\},\\ u_t&=\max\{q_t-w_t,0\}, & v_t&=\max\{w_t-q_t,0\},\ \text{for } t=1,\ldots,T,\end{aligned}$$

where $\bar{h}_t=h$ for $t=1,\ldots,T-1$ and $\bar{h}_T=h-s$. Note that $\{y_t\},\{u_t\},\{v_t\}$ all depend on d_t and thus require the introduction of auxiliary variables $\{y_t^{\tau}\}_{t,\tau}$, $\{v_t^{\tau}\}_{t,\tau}$, $\{u_t^{\tau}\}_{t,\tau}$ to reflect their affine dependency on d. Thus, we can rewrite the worst-case formulation of (20) as the following semi-infinite program:

$$\begin{aligned}\min_{C,w_t,z_t^{\pm},y_t^{\tau},u_t^{\tau},v_t^{\tau},q_t^{\tau}} \quad & \beta^{+}\cdot\sum_{t=2}^{T}z_t^{+}+\beta^{-}\sum_{t=2}^{T}z_t^{-}+C\\ \text{s.t.} \quad & z_t^{+}\geq w_t-w_{t-1},\ z_t^{+}\geq 0, \quad z_t^{-}\geq w_{t-1}-w_t,\ z_t^{-}\geq 0, \quad t=2,\ldots,T,\\ \forall d\in\mathcal{S}_{B/E}: \quad & C\geq\sum_{t=1}^{T}\left(y_0^t+cq_0^t+\alpha^{+}u_0^t+\alpha^{-}v_0^t+\sum_{\tau=1}^{t-1}\left(y_\tau^t+cq_\tau^t+\alpha^{+}u_\tau^t+\alpha^{-}v_\tau^t\right)d_\tau\right)\\ & y_0^t+\sum_{\tau=1}^{t-1}y_\tau^t d_\tau\geq\bar{h}_t\sum_{\tau=1}^{t}\left(q_0^{\tau}+\sum_{\sigma=1}^{\tau-1}q_\sigma^{\tau}d_\sigma-d_\tau\right),\ t=1,\ldots,T,\\ & y_0^t+\sum_{\tau=1}^{t-1}y_\tau^t d_\tau\geq p\sum_{\tau=1}^{t}\left(-q_0^{\tau}-\sum_{\sigma=1}^{\tau-1}q_\sigma^{\tau}d_\sigma+d_\tau\right),\ t=1,\ldots,T,\\ & u_0^t+\sum_{\tau=1}^{t-1}u_\tau^t d_\tau\geq q_0^t+\sum_{\tau=1}^{t-1}q_\tau^t d_\tau-w_t,\ u_0^t+\sum_{\tau=1}^{t-1}u_\tau^t d_\tau\geq 0,\ t=1,\ldots,T,\\ & v_0^t+\sum_{\tau=1}^{t-1}v_\tau^t d_\tau\geq w_t-q_0^t-\sum_{\tau=1}^{t-1}q_\tau^t d_\tau,\ v_0^t+\sum_{\tau=1}^{t-1}v_\tau^t d_\tau\geq 0,\ t=1,\ldots,T,\\ & \sum_{t=1}^{T}\left(q_0^t+\sum_{\tau=1}^{t-1}q_\tau^t d_\tau\right)\geq L,\ q_0^t+\sum_{\tau=1}^{t-1}q_\tau^t d_\tau\geq 0,\ t=1,\ldots,T,\end{aligned} \tag{21}$$

Depending on whether $d \in \mathcal{S}_B$ or $\mathcal{S}_E$, (21) can be cast as either an LP or an SOCP by rewriting the constraints involving d. For example, if $\mathcal{S}_B$ is given by

$$\mathcal{S}_B = \{d \in \mathbb{R}^T \mid |d_t - \bar{d}_t| \leq \rho_t, t = 1, \ldots, T\},$$

the constraint

$$\sum_{t=1}^{T} \left(q_0^t + \sum_{\tau=1}^{t-1} q_\tau^t d_\tau \right) \geq L, \ \forall d \in \mathcal{S}_B$$

is equivalent to

$$\sum_{t=1}^{T} q_t^0 + \sum_{\tau=1}^{T} \left(\sum_{t=1}^{T} q_t^\tau \right) \bar{d}_\tau - \sum_{\tau=1}^{T} \gamma_\tau \rho_\tau \geq L, \quad -\gamma_\tau \leq \sum_{t=1}^{T} (-q_t^\tau) \leq \gamma_\tau, \ \tau = 0, \ldots, T, \tag{22}$$

where $\{\gamma_\tau\}_\tau$ are new variables; see [8, Section 3.3] for more details. We consider now the following scenario: at time $t = 0$, before the inventory starts, the retailer and the supplier negotiate the details of their contract, that is, they need to agree on values of $\alpha^\pm$, $\beta^\pm$, and L. To do so effectively, the retailer needs to know in real time the worst-case cost incurred for different choices of these values. This cannot be done by solving the LP/SOCP as, e.g., running one such LP with $T = 100$ takes ≈ 15 minutes, whilst running the SOCP with $T = 100$ takes more than 60 minutes. Thus, a tool which produces an approximation of the minimum inventory cost as a function of $(\alpha^\pm, \beta^\pm, L)$ which can easily be evaluated for new parameter values would be valuable. We use shape-constrained regression to do this as explained above.

Proposition 5.2. *Let $v(\alpha^+, \alpha^-, \beta^+, \beta^-, L)$ be the optimal value of the LP obtained by taking $d \in \mathcal{S}_B$ in (21). We have that:*

- *(i) v is convex in L*
- *(ii) v is concave jointly in (β^+, β^-)*
- *(iii) v is nondecreasing in $\alpha^+, \alpha^-, \beta^+, \beta^-$, and L.*

The same results hold for the SOCP obtained by taking $d \in \mathcal{S}_E$ in 21, assuming that strong duality for the SOCP holds.

Proof. Property (i) follows from Proposition 5.1 (i) and (22). Proprety (ii) follows from Proposition 5.1 (ii). For (iii), it is straightforward to see that v does not decrease with β^+ and β^- as $\sum_{t=2}^{T} z_t^\pm \geq 0$. It is also straightforward to see that v does not decrease with L as, when L increases, the left-hand side inequality in (22) becomes harder to satisfy, and so the feasible set shrinks. We give a proof by contradiction that v does not increase with α^+ (a similar reasoning applies for α^-). Let $\alpha_1 < \alpha_2$ and let v_1 (resp. v_2) be the optimal value of (21) when $\alpha^+ = \alpha_1$ (resp α_2). By way of contradiction, we assume that $v_1 > v_2$. We denote an optimal solution to (21) when $\alpha^+ = \alpha_2$ by $\{\tilde{w}_t\}$, $\{\tilde{z}_t^+\}$, $\{\tilde{z}_t^-\}$,$\tilde{C}$,$\{\tilde{y}_\tau^t\}$,$\{\tilde{q}_\tau^t\}$, $\{\tilde{v}_\tau^t\}$, $\{\tilde{u}_\tau^t\}$ and take $\forall t, \tau$,

$$w_t = \tilde{w}_t, z_t^\pm = \tilde{z}_t^\pm, \ C = \tilde{C}, \ y_\tau^t = \tilde{y}_\tau^t, \ q_\tau^t = \tilde{q}_\tau^t, \ v_\tau^t = \tilde{v}_\tau^t, \ u_\tau^t = \frac{\alpha_2}{\alpha_1} \tilde{u}_\tau^t. \tag{23}$$

One can check that as $\alpha_2/\alpha_1 \geq 1$, the solution in (23) is feasible to (21) when $\alpha^+ = \alpha_1$. Furthermore, it is easy to see that the corresponding objective value is equal to v_2. As $v_1 > v_2$, this contradicts optimality of v_1. □

m	d	SOSE Ave (%) Train	SOSE Ave (%) Test	SOSE Max (%) Test	UPR Ave (%) Train	UPR Ave (%) Test	UPR Max (%) Test
100	2	2.522	2.487	**6.329**	1.545	1.899	6.723
	4	1.018	1.027	**3.506**	0.000	4.366	31.656
	6	0.853	0.867	**3.102**	0.000	3.781	24.169
200	2	2.530	2.451	6.060	1.662	1.885	**4.400**
	4	1.111	1.065	**3.191**	0.227	0.645	4.636
	6	0.939	0.907	**1.849**	0.000	3.174	19.809
400	2	2.550	2.490	5.892	1.698	1.803	**4.325**
	4	1.065	1.026	**2.731**	0.304	0.477	3.706
	6	0.898	0.870	**1.889**	0.000	9.34	225.441
600	2	2.577	2.478	5.849	1.788	1.752	**4.209**
	4	1.074	1.023	2.314	0.322	0.412	**1.618**
	6	0.904	0.869	**1.652**	0.109	0.615	9.521

Table 3: Relative accuracy of the values predicted by the SOSE (with $r = 2$) and the UPR against those obtained by the LP for the contract negotiation application given in (5.1). Best test max deviation marked in bold font (smaller is better).

Thus, when preparing for the negotiation, we generate m training points by sampling uniformly at random m tuples $(\alpha^+, \alpha^-, \beta^+, \beta^-, L)$ and computing the minimum inventory cost v by solving (21) with a fixed set $\mathcal{S}_B$. We then fit the degree-d SOSE of v (with $r = 2$) to these points. During the negotiation, to obtain the minimum inventory cost for new values of $(\alpha^\pm, \beta^\pm, L)$, we simply evaluate the SOSE at these values. This task only takes milliseconds, in contrast to solving the LPs or SOCPs which takes minutes if not hours. In Table 3, we give the relative accuracy of the optimal values predicted by the SOSE, both on average (train and test) and in the worst-case (test only), as compared to the true optimal values obtained via the LP. (Similar results hold for $\mathcal{S}_E$.) As can be seen, the accuracy of the predictions is 1-2% on average. We also compare the SOSE against the *Unconstrained* Polynomial Regressor (UPR). We note that the SOSE performs best for moderate values of m (200 or 400) and larger values of d (4 or 6). This implies that when using the SOSE as compared to the UPR, we need not solve as many LPs offline to obtain good-quality predictions. Furthermore, the SOSE appears to be more robust across variations in the data, with consistent average performance over training and testing sets and best worst-case performance, particularly for higher d. Note that for these values of m, the CLSE could not be used for real-time predictions due to the time it takes to compute new predictions; see Table 2.

5.2 Shape-Constrained Optimal Transport Maps and Color Transfer

An *optimal transport map* is, at a high level, a function that maps one probability measure to another while incurring minimum cost. In many applications (such as image processing and computer graphics), it is of interest to determine an optimal transport map given two measures and a cost function [49]. Interestingly, the problem of computing an optimal transport map can be related back to shape-constrained regression, as optimal transport maps are known to have specific shapes when the cost function under consideration or the measures they are defined over have certain properties. For example, if the cost function is the l_2-norm and one of the measures is continuous with respect to the Lebesgue measure, the Brenier theorem states that the optimal transport map is uniquely defined as the gradient of a convex function [13]. Following [48], rather than observing these properties of the map a posteriori, we use these shape constraints as regularizers when computing the optimal transport maps. This gives rise to shape-constrained regression problems. To solve these, [48] propose an approach that can be viewed as a CLSE-based approach. We propose to use instead the SOSE, which we show is particularly well-suited to this application.

To better illustrate our method, we focus on the concrete application of *color transfer*, though our methodology is applicable more widely to, e.g., the other applications mentioned in [48] and voice transfer. The color transfer problem is defined by two images, the *input image* and the *target image* [52]. The goal is to transfer the colors of the target image to the input image, thereby creating a third image, the *output image*; see Figure 5. We now describe in detail how the color transfer problem can be reformulated as a sequence of shape-constrained regression problems, following [48]. Each pixel in the input and target images is associated to an RGB color triple in $[0,1]^3$, once each entry is normalized by 256. We thus define a discrete measure

$$\mu = \sum_{i=1}^{\tilde{N}} a_i \delta_{x_i} \quad \left(\text{resp. } \nu = \sum_{j=1}^{\tilde{M}} b_j \delta_{y_j}\right)$$

over the space of colors in the input (resp. target) image by taking $x_i \in [0,1]^3, i = 1, \ldots \tilde{N}$, (resp. $y_j \in [0,1]^3, j = 1, \ldots, \tilde{M},$) to be the distinct color triples in the input (resp. target) image, with $\tilde{N}$ (resp. $\tilde{M}$) being less than or equal to the number of pixels in the input (resp. target) image, and a_i (resp. b_j) to be the ratio of number of pixels of color x_i (resp. y_j) to the total number of pixels. The idea is then to search for a function $f^* : [0,1]^3 \to \mathbb{R}$ and a coupling $P^* \in \mathbb{R}^{\tilde{N} \times \tilde{M}}$, solutions to the following optimization problem:

$$\begin{aligned} \inf_{f, P \in \mathbb{R}^{\tilde{N} \times \tilde{M}}} \quad & \sum_{i,j} P_{ij} ||\nabla f(x_i) - y_j||^2 \\ \text{s.t.} \quad & P \geq 0, \; P 1_{\tilde{M}} = a, \; P^T 1_{\tilde{N}} = b \\ & \nabla f \text{ is } L\text{-lipschitz and } f \text{ is } \ell\text{-strongly convex over } [0,1]^3, \end{aligned} \tag{24}$$

where L and ℓ are parameters of the problem[4]. We derive from the optimal solution f^* to (24), the optimal transport map (or color transfer map in this case) given by $\nabla f^* : [0,1]^3 \to [0,1]^3$. To obtain the output image, we simply apply ∇f^* to the RGB triple of each pixel in the input image, obtain a new RGB triple (i.e., the new color of the pixel), and plot the resulting image. By enforcing the shape constraints given in (24), we are requiring that the optimal transport map have desirable properties such as "two dark colors in the input image map to two dark colors in the output image"; see [48] for more details.

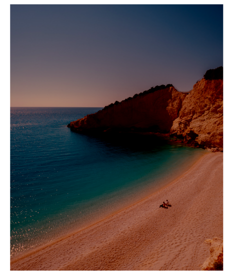
In its current form however, problem (24) is not quite a shape-constrained regression problem of the type (7) or (8). This is due to the matrix variable P which makes the problem non-convex. To circumvent this issue, we solve (24) using alternate minimization: we fix f and solve for P using, e.g., Sinkhorn's algorithm (see, e.g., [49]). We then fix P and solve for f. If we parametrize f as a polynomial (with P fixed), we obtain a shape-constrained polynomial regression problem:

$$\begin{aligned} \inf_{f \in P_{3,d}} \quad & \sum_{i,j} P_{ij} ||\nabla f(x_i) - y_j||^2 \\ \text{s.t.} \quad & \ell \cdot I \preceq H_f(x) \leq L \cdot I, \; \forall x \in [0,1]^3, \end{aligned} \tag{25}$$

[4] As a reminder, a function $g : \mathbb{R}^m \to \mathbb{R}^n$ is L-Lipschitz over a box B as defined in (1), if $||g(x) - g(y)|| \leq L \cdot ||x - y||$, for all $x, y \in B$, where $|| \cdot ||$ is some norm. Similarly, a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is ℓ-strongly convex over B if $H_f(x) \succeq \ell I$, for all $x \in B$.

(a) Input image

(b) Target image



(c) Output image

Figure 5: The color transfer problem in (5.2) with $\ell = 1$ and $L = 10$.

which we solve using the sum of squares techniques described in Section 2.3. We iterate this process until convergence.

An example of the output images obtained via this process is given in Figure 5. Additional illustrations can be found in Figure 6 for different values of l and L with $d = 4$ and $r = 3$. The color transfer application works particularly well for the SOSE as the number of features is small (equal to 3), the number of data points is very large, as it corresponds to the number of pixels in the images, and as a large number of new predictions need to be made (one per pixel of the input image). In contrast, the CLSE approach considered in [48] requires the authors to segment the images via k-means clustering to compute f^* in a reasonable amount of time. Pre-processing of this type can lead to undesirable artifacts in the output image and grainy texture, which our method avoids. Furthermore, predictions in [48] are obtained via a quadratically-constrained quadratic program (rather than a linear program as is the case for standard CLSE) due to the nature of the shape constraints, which makes each prediction quite expensive to compute. This is in opposition to the SOSE, where predictions are simply point evaluations of a polynomial.

A Proof of Theorem 2.1

We make use of three lemmas.

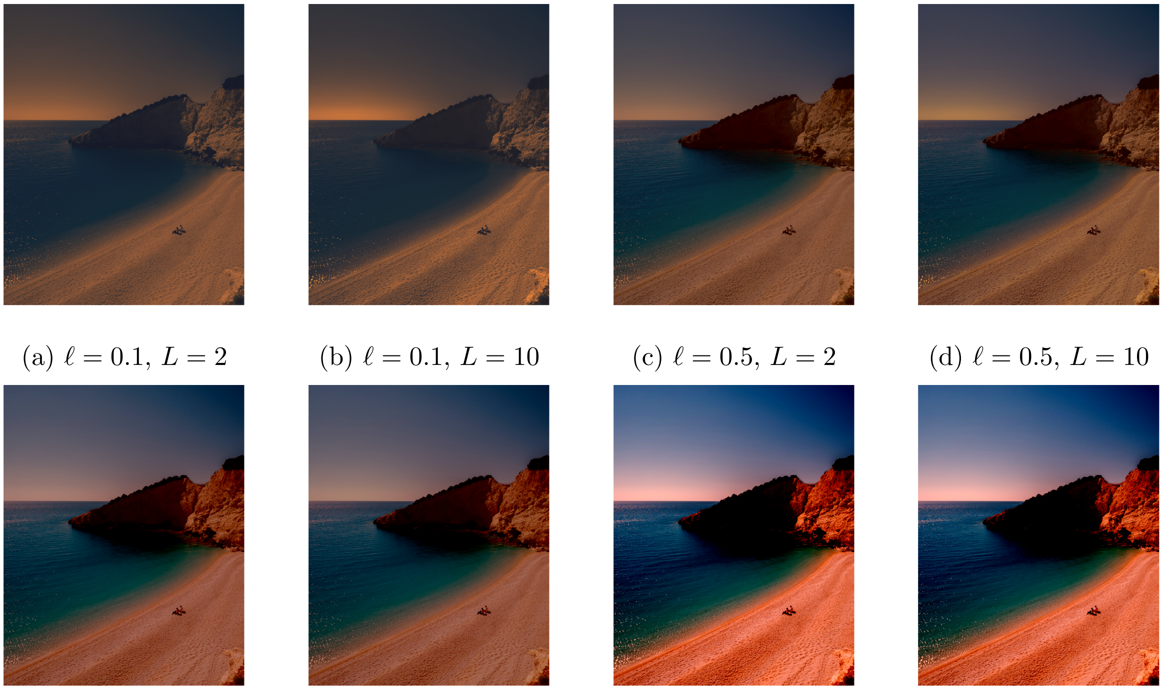
Lemma A.1. *Given a polynomial p of degree $d \geq 3$, a vector K as defined in Definition 2, and a box B as defined in (1), it is strongly NP-hard to test whether p has K-bounded derivatives over B.*

Proof. We first show the result for $d = 3$ by providing a reduction from MAX-CUT. Recall that in an unweighted undirected graph $G = (V, E)$ with no self-loops, a *cut* partitions the n nodes of the graph into two sets, S and $\bar{S}$. The size of the cut is given by the number of edges connecting a node in S to a node in $\bar{S}$. MAX-CUT is then the following problem: given a graph G and an integer k, test whether G has a cut of size at least k. It is well-known that MAX-CUT is NP-hard [20].

We denote by A the adjacency matrix of the graph, i.e., A is an $n \times n$ matrix such that $A_{ij} = 1$ if $\{i, j\} \in E$ and $A_{ij} = 0$ otherwise, and by $\gamma := \max_i \{A_{ii} + \sum_{j \neq i} |A_{ij}|\}$. Note that γ is an integer (it is the maximum degree in the graph) and is an upper bound on the largest eigenvalue of A from Gershgorin's circle theorem [21].

We show that G does not have a cut of size $\geq k$ if and only if the cubic polynomial

$$p(x_1, \ldots, x_n) = \sum_{j=2}^{n} \frac{x_1^2 A_{1j} x_j}{4} + \frac{x_1}{2} \sum_{1<i<j\leq n} x_i A_{ij} x_j - \frac{\gamma x_1^3}{12} - \frac{\gamma x_1}{4} \sum_{i=2}^{n} x_i^2 + x_1 \left(k + \frac{n\gamma}{4} - \frac{e^T A e}{4} \right) \tag{26}$$



(a) $\ell = 0.1,\ L = 2$ (b) $\ell = 0.1,\ L = 10$ (c) $\ell = 0.5,\ L = 2$ (d) $\ell = 0.5,\ L = 10$

(e) $\ell = 1,\ L = 2$ (f) $\ell = 1,\ L = 10$ (g) $\ell = 2,\ L = 5$ (h) $\ell = 2,\ L = 10$

Figure 6: Color transfer outputs for different regularization parameters ℓ and L; see Section 5.2.

has K-bounded derivatives on $B = [-1, 1]^n$, where

$$K_1^- = 0,\ K_1^+ = n^2 + k,\ K_2^- = \ldots = K_n^- = -n,\ K_2^+ = \ldots = K_n^+ = n + 1. \tag{27}$$

Letting $x = (x_1, \ldots, x_n)^T$, we compute the partial derivatives of p:

$$\frac{\partial p(x)}{\partial x_1} = \frac{1}{4} x^T (A - \gamma I) x + (k + \frac{n\gamma}{4} - \frac{1}{4} e^T A e),$$
$$\frac{\partial p(x)}{\partial x_i} = \frac{1}{4} x_1^2 A_{1i} + \frac{1}{2} x_1 \cdot \sum_{j \neq i, j > 1} x_i A_{ij} - \frac{\gamma}{2} x_1 \cdot x_i,\ i = 2, \ldots, n.$$

As $-1 \leq x_i \leq 1$ for $i = 1, \ldots, n$ and $\gamma \leq n$, it is straightforward to check that

$$\frac{\partial p(x)}{\partial x_1} \leq K_1^+ \text{ and } K_i^- \leq \frac{\partial p(x)}{\partial x_i} \leq K_i^+,\ i = 1, \ldots, n,\ \forall x \in B.$$

The statement to show thus becomes: G does not have a cut of size $\geq k$ if and only if

$$\frac{\partial p(x)}{\partial x_1} = \frac{1}{4} x^T (A - \gamma I) x + k + \frac{n\gamma}{4} - \frac{1}{4} e^T A e \geq K_1^- = 0,\ \forall x \in B.$$

The converse implication is easy to prove: if $\frac{\partial p(x)}{\partial x_1} \geq 0$ for all $x \in B$, then, in particular, $\frac{\partial p(x)}{\partial x_1} \geq 0$ for $x \in \{-1, 1\}^n$. When $x \in \{-1, 1\}^n$, $\gamma x^T x = \gamma n$, and so

$$k \geq \frac{1}{4} e^T A e - \frac{1}{4} x^T A x,\ \forall x \in \{-1, 1\}^n. \tag{28}$$

Any cut in G can be encoded by a vector $x \in \{-1,1\}^n$ with $x_i = 1$ if node i is on one side of the cut and with $x_i = -1$ if node i is on the other side of the cut. The size of the cut is then given by $\frac{1}{4}e^T Ae - \frac{1}{4}x^T Ax$ [22]. Hence, (28) is equivalent to stating that all cuts in G are of size less than or equal to k.

For the implication, if G does not have a cut of size greater than or equal to k, then (28) holds, which is equivalent to

$$\frac{1}{4}x^T(A-\gamma I)x \geq -k - \frac{n\gamma}{4} + \frac{1}{4}e^T Ae,\ \forall x \in \{-1,1\}^n. \tag{29}$$

Now, by definition of γ, $A-\gamma I \preceq 0$, i.e., $x^T(A-\gamma I)x$ is concave. Let $y \in B$. We have $y = \sum_{i=1}^{2^n} \lambda_i x_i$ where x_i are the corners of B, which are in $\{-1,1\}^n$, $\lambda_i \geq 0$, $i = 1,\dots,2^n$, and $\sum_{i=1}^{2^n} \lambda_i = 1$. As $y \mapsto y^T(A-\gamma I)y$ is concave and (29) holds,

$$\frac{1}{4}y^T(A-\gamma I)y \geq \sum_{i=1}^{2^n} \lambda_i x_i^T(A-\gamma I)x_i \geq \sum_{i=1}^{2^n} \lambda_i(-k - \frac{n\gamma}{4} + \frac{1}{4}e^T Ae) = -k - \frac{n\gamma}{4} + \frac{1}{4}e^T Ae.$$

This concludes the proof for $d = 3$. For $d \geq 4$, we define

$$\tilde{p}(x_1,\dots,x_n,x_{n+1}) := p(x_1,\dots,x_n) + x_{n+1}^d \in P_n^d,\ \tilde{K} := (K,0,1),\ \text{and}\ \tilde{B} := B \times [0,1],$$

with p as in (26), K as in (27), and $B = [-1,1]^n$. We compute:

$$\frac{\partial \tilde{p}(x_1,\dots,x_{n+1})}{\partial x_i} = \frac{\partial p(x_1,\dots,x_n)}{\partial x_i}\ \text{for}\ i = 1,\dots,n,\ \text{and}\ \frac{\partial \tilde{p}(x_1,\dots,x_{n+1})}{\partial x_{n+1}} = d \cdot x_{n+1}^{d-1}.$$

As $d \cdot x_{n+1}^{d-1} \in [0,1]$ when $x_{n+1} \in [0,1]$, it follows that $\tilde{p}$ has $\tilde{K}$-bounded derivatives over $\tilde{B}$ if and only if p has K-bounded derivatives over B. □

Lemma A.2 ([5, 23])**.** *Given a polynomial p of degree $d \geq 3$ and a box B as defined in (1), it is strongly NP-hard to test whether p is convex over B.*

Lemma A.3 ([44])**.** *Let S_n be a closed full-dimensional simplex in $\mathbb{R}^n$ with vertices $t_0,\dots,t_n$. Let $d \geq 1$ be an integer and let Z_d denote the set of numbers $\{0, 1/d, 2/d, \dots, 1\}$. We associate with S_n the discrete point set $\Gamma(n,d)$:*

$$\Gamma(n,d) := \{x \in \mathbb{R}^n \mid x = \sum_{i=0}^{n} \lambda_i t_i, \lambda_i \in Z_d, \sum_{i=0}^{n} \lambda_i = 1\}. \tag{30}$$

Let $w_i, i = 1,\dots,\binom{n+d}{d}$ be the points contained in $\Gamma(n,d)$ and let $f_i, i = 1,\dots,\binom{n+d}{d}$ be arbitrary numbers in $\mathbb{R}$, then there is exactly one polynomial $p \in P_{n,d}$ such that $p(w_i) = f_i, i = 1,\dots \binom{n+d}{d}$.

Proof of Theorem 2.1. Let $d \geq 3$. We only show NP-hardness of BD-DER-REG-d (the proof of NP-hardness of CONV-REG-d is analogous) and this is done via a reduction from the problem in Lemma A.1. To this effect, consider an instance of the problem, i.e., a polynomial $\tilde{p}$ of degree d, a vector $\tilde{K}$, and a box $\tilde{B}$. We construct our reduction thus: we take $K = \tilde{K}$, $B = \tilde{B}$, and $t = 0$. Recalling that $\tilde{B} = [\tilde{l}_1, \tilde{u}_1] \times \ldots \times [\tilde{l}_n, \tilde{u}_n]$, we set

$$t_0 = (\tilde{l}_1, \tilde{l}_2, \dots, \tilde{l}_n), t_1 = (\tilde{u}_1, \tilde{l}_2, \dots, \tilde{l}_n), \dots, t_n = (\tilde{l}_1, \tilde{l}_2, \dots, \tilde{u}_n)$$

and take $X_1,\dots,X_m$, where $m = \binom{n+d}{d}$, be the points contained in $\Gamma_{n,d}$ as defined in (30). Note that $X_1,\dots,X_m$ are rational, can be computed in polynomial time, and belong to B as $s_0,\dots,s_n \in B$.

We then take $Y_i = \tilde{p}(X_i), i = 1, \ldots, m$. It is then easy to see that the answer to BD-DER-REG-d is YES if and only if $\tilde{p}$ has $\tilde{K}$-bounded derivatives on $\tilde{B}$. The converse is immediate by taking $p = \tilde{p}$. For the implication, if the answer to BD-DER-REG-d is YES, then there exists a polynomial p of degree d with K-bounded derivatives over B such that $\sum_{i=1}^m (Y_i - p(X_i))^2 = 0$, i.e., $p(X_i) = Y_i$ for $i = 1, \ldots, m$. From Lemma A.3, as $\tilde{p} \in P_{n,d}$, it must be the case that $p = \tilde{p}$ and so $\tilde{p}$ has K-bounded derivatives over B.

Now, let $d = 1$ and denote by $\mathbf{K}^+ = (K_1^+, \ldots, K_n^+)$ and by $\mathbf{K}^- = (K_1^-, \ldots, K_n^-)$. The answer to BD-DER-REG-d is YES if and only if the optimal value of

$$\begin{aligned} &\min_{a_0 \in \mathbb{R}, a \in \mathbb{R}^n} \sum_{i=1}^m (Y_i - a^T X_i - a_0)^2 \\ &\text{s.t. } \mathbf{K}^- \leq a \leq \mathbf{K}^+ \end{aligned}$$

is less than or equal to t (here the inequalities are component-wise). As this is a quadratic program (QP), BD-DER-REG-d is in P when $d = 1$. Similarly, the answer to CONV-REG-d is YES if and only if the optimal value of $\min_{a_0 \in \mathbb{R}, a \in \mathbb{R}^n} \sum_{i=1}^m (Y_i - a^T X_i - a_0)^2$ is less than or equal to t (as any affine function is convex). Computing this optimal value simply involves solving a system of linear equations, which implies that CONV-REG-d is in P for $d = 1$.

Let $d = 2$. We denote by S^n the set of symmetric matrices in $\mathbb{R}^{n \times n}$ and parametrize p as $p(x) = x^T Q x + b^T x + c$, where $Q \in S^n$, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$. The answer to BD-DER-REG-d is YES if and only if the optimal value of

$$\begin{aligned} &\min_{Q \in S^n, b \in \mathbb{R}^n, c \in \mathbb{R}} \sum_{i=1}^m (Y_i - X_i^T Q X_i - b^T X_i - c)^2 \\ &\text{s.t. } \mathbf{K}^- \leq 2Qx + b \leq \mathbf{K}^+, \forall x \in B \end{aligned} \tag{31}$$

is less than or equal to t. As B is a compact, convex, and full-dimensional polyhedron, one can use [24, Proposition I.1.] to rewrite the condition $2Qx + b \leq \mathbf{K}^+, \forall x \in B$ equivalently as

$$\mathbf{K}^+ - 2Qx - b = \lambda + \Lambda^-(x - l) + \Lambda^+(u - x),\ \lambda \geq 0,\ \Lambda^+, \Lambda^- \geq 0$$

where $\lambda \in \mathbb{R}^n$, $\Lambda^\pm \in \mathbb{R}^{n \times n}$ are additional variables. A similar technique can be used for $2Qx + b \geq \mathbf{K}^-, \forall x \in B$. Thus, (31) is equivalent to a QP and so testing whether its objective value is less than or equal to t can be done in polynomial time. Now, the answer to CONV-REG-d is YES if and only if the optimal value of

$$\begin{aligned} &\min_{Q \in S^n, b \in \mathbb{R}^n, c \in \mathbb{R}} \sum_{i=1}^m (Y_i - X_i^T Q X_i - b^T X_i - c)^2 \\ &\text{s.t. } Q \succeq 0 \end{aligned}$$

is less than or equal to t. As this is a polynomial-size semidefinite program, the result follows. □

B Proof of Theorem 2.4

We show the following equivalent result.

Theorem B.1. *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a strictly convex quadratic function and let x^* be its (unique) minimizer*[5] *over a closed convex set $S \subseteq \mathbb{R}^n$. Suppose that a sequence $\{x_j\}_j \subseteq S$ is such that $\lim_{j \to \infty} f(x_j) = f(x^*)$. Then, $\lim_j ||x_j - x^*|| = 0$.*

[5] Note that x^* must exist as f is coercive.

The proof of Theorem 2.4 follows by identifying polynomials in $P_{n,d}$ with their coefficients[6] in $\mathbb{R}^{\binom{n+d}{d}}$. We then apply Theorem B.1 with $f(c) = \sum_{i=1}^{m}(Y_i - c(X_i))^2$, which is a strictly convex quadratic function by virtue of our assumptions on m and on the data. For (9), we consider $S = \{c \in \mathbb{R}^{\binom{n+d}{d}} \mid H_c(x) \succeq 0, \forall x \in B\}$, with $\bar{g}_{m,d}$ playing the role of x^* and $\{\tilde{g}_{m,d,r}\}_r$ playing the role of $\{x_j\}_j$. Convergence of $\{f(\tilde{g}_{m,d,r})\}_r$ to $f(\bar{g}_{m,d})$ follows from Theorem 2.3. For (10), we take

$$S = \{c \in \mathbb{R}^{\binom{n+d}{d}} \mid K_i^- \leq \frac{\partial c(x)}{\partial x_i} \leq K_i^+, \forall x \in B, \forall i = 1, \ldots, n\},$$

with $\bar{h}_{m,d}$ playing the role of x^* and $\{\tilde{h}_{m,d,r}\}_r$ playing the role of $\{x_j\}_j$. Convergence of $\{f(\tilde{h}_{m,d,r})\}_r$ to $f(\bar{h}_{m,d})$ follows from Theorem 2.2.

The proof of Theorem B.1 requires the following lemma.

Lemma B.2. *Let $x^* \in \mathbb{R}^n$ and $\{S_j\}_j \subset \mathbb{R}^n$ be a sequence of sets satisfying*

(i) $S_j, j = 1, 2, \ldots$ is nonempty and compact,

(ii) $S_{j+1} \subseteq S_j$, for $j = 1, 2, \ldots$,

(iii) $\cap_j S_j = \{x^*\}$

then, $\forall \epsilon > 0$, $\exists k_0$ such that $S_k \subseteq B(x^*, \epsilon)$, $\forall k \geq k_0$, where $B(x^*, \epsilon)$ is the ball centered at x^* and with radius $\epsilon > 0$.

Proof. We first claim that:

$$\text{if } d_k = diam(S_k) := \max_{y,z \in S_k} ||y - z|| \text{ then } \lim_{k \to \infty} d_k = 0.$$

(We note that d_k is finite for all k by virtue of (i).) To see this, observe that, due to (ii), $\{d_k\}_k$ is nonincreasing and bounded below by zero. Thus, it converges to some limit d^*, with $d_k \geq d^*, \forall k$. Suppose for the sake of contradiction that $d^* > 0$. As S_k is compact, for all k, there exist y_k and z_k in S_k such that $||y_k - z_k|| = d_k$. As $\{y_k\}_k$ is a bounded sequence, it has a convergent subsequence $\{y_{\theta(k)}\}$. Likewise, $\{z_{\theta(k)}\}$ is a bounded sequence, so it has a convergent subsequence $\{z_{\phi(k)}\}$. We now let:

$$\bar{y} := \lim_{k \to \infty} \{y_{\phi(k)}\} \text{ and } \bar{z} := \lim_{k \to \infty} \{z_{\phi(k)}\}.$$

As $||y_{\phi(k)} - z_{\phi(k)}|| = d_{\phi(k)} \geq d^*, \forall k$, and $|| \cdot ||$ is continuous, we have that $|\bar{y} - \bar{z}|| \geq d^*$. Now, for any $i \in \mathbb{Z}_+$, (ii) implies that $y_{\phi(k)}, z_{\phi(k)} \in S_i$ for $\phi(k) \geq i$. Since S_i is closed by virtue of (i), we must have $\bar{y}, \bar{z} \in S_i$. As the i chosen before was arbitrary and using (iii), it follows that $\bar{y} \in \cap_k S_k = \{x^*\}$ and $\bar{z} \in \cap_k S_k = \{x^*\}$, which contradicts $||\bar{y} - \bar{z}|| \geq d^* > 0$. This proves the claim.

From the claim, we know that $\forall \epsilon > 0$, $\exists k_0$ such that $d_k < \epsilon$, $\forall k \geq k_0$. As

$$d_k = \max_{y,z \in S_k} ||y - z|| \geq \max_{z \in S_k} ||x^* - z||,$$

$\max_{z \in S_k} ||x^* - z|| < \epsilon$, $\forall k \geq k_0$. This implies that $S_k \subseteq B(x^*, \epsilon), \forall k \geq k_0$. □

Proof of Theorem B.1. Let $S_k := \left\{x \in S \mid f(x) \leq f(x^*) + \frac{1}{k}\right\}$. The assumptions of Lemma B.2 are met: (i) holds as S_k is nonempty (it contains x^*) and compact (it is contained in the sublevel set of a coercive function); (ii) holds as $S_{k+1} \subseteq S_k$ by definition; and (iii) holds as x^* is contained in

[6] In the following, with some abuse, we use the same notation for both the polynomials and their coefficients.

S_k for all k and it is the unique minimizer of f over S. We can thus apply Lemma B.2. Let $\epsilon > 0$. From Lemma B.2,

$$\exists k_0 \text{ s.t. } \forall k \geq k_0, x \in S \text{ and } f(x) \leq f(x^*) + \frac{1}{k} \Rightarrow ||x - x^*|| \leq \epsilon.$$

Now, as $\lim_{j\to\infty} f(x_j) = f(x^*)$, then $\exists j_0$ such that $\forall j \geq j_0, |f(x_j) - f(x^*)| \leq \frac{1}{k_0}$. As $x_j \in S, \forall j$, it follows that $||x_j - x^*|| \leq \epsilon$, which concludes the proof. □

C Proofs of Lemmas 3.3 and 3.4

The proofs of Lemmas 3.3 and 3.4 are given in Appendix C.3. Before giving them, we first derive Weierstrass-type results for functions with shape constraints (Appendix C.1) and show that our estimators $\bar{g}_{m,d}$ and $\bar{h}_{m,d}$ are uniformly bounded and Lipschitz continuous, with constants independent of the data points (Appendix C.2). We denote by $C_{n,d}$ (resp. $K_{n,d}$) the set of polynomials of degree d in n variables that are convex over the box B (resp. have K-bounded derivatives over B).

C.1 Weierstrass-type results for shape-constrained functions

We show here results that are similar to the Weierstrass theorem: namely that we can approximate, over a box B, any convex-constrained function or function with bounded derivatives arbitrarily well by a polynomial with the same characteristics. The presence of shape constraints prevents us from leveraging the Weierstrass theorem as is.

Proposition C.1 (e.g., Theorem 6.7 in [18])**.** *Consider the Bernstein multivariate polynomial of degree d and in n variables, defined over $[0,1]^n$:*

$$B_d(f,x) = \sum_{j_1+\ldots+j_n=d} f\left(\frac{j_1}{d}, \ldots, \frac{j_n}{d}\right) C_d^{j_1} \ldots C_d^{j_n} x_1^{j_1}(1-x_1)^{d-j_1} \ldots x_n^{j_n}(1-x_n)^{d-j_n},$$

where $C_d^{j_i} = \frac{d!}{j_i!(d-j_i)!}$. Let m be an integer and assume that f is m times continuously differentiable. Let $k = (k_1, \ldots, k_n)$ be a multi-index such that $\sum_{i=1}^n |k_i| \leq m$ and denote by

$$\partial^k f = \frac{\partial^k f(x)}{\partial x_1^{k_1} \ldots \partial x^{k_n}}.$$

Then, $\forall k$ such that $\sum_{i=1}^n |k_i| \leq m$, $\lim_{d\to\infty} \sup_{x\in[0,1]^n} |\partial^k B_d(f,x) - \partial^k f(x)| = 0$.

This result can easily be extended to hold over any box $B \subset \mathbb{R}^n$ by simply scaling and translating the variables in the Bernstein polynomials.

Proposition C.2 (Approximating a function that is convex over a box)**.** *Let f be a twice continuously differentiable function defined over B such that $H_f(x) \succeq 0$, for all $x \in B$ (i.e., f is convex over B). Define*

$$g_d := \arg \min_{g \in C_{n,d}} \sup_{x\in B} |f(x) - g(x)|.$$

For any $\epsilon > 0$, there exists d such that $\sup_{x\in B} |g_d(x) - f(x)| < \epsilon$.

Appendix A in [9], e.g., guarantees the existence of g_d (as the objective function is coercive in the coefficients of g and the set $C_{n,2d}$ is closed) but not its uniqueness. Thus, g_d is one of the existing minimizers.

Proof. Let $\epsilon > 0$ and $M := \max_{x\in B} \frac{1}{2}\sum_{i=1}^{n} x_i^2$. From Proposition C.1, as f is twice continuously differentiable, there exists a polynomial q of degree d such that

$$\sup_{x\in B} |f(x) - q(x)| \leq \frac{\epsilon}{2(1+2nM)} \text{ and } \sup_{x\in B} \left| \frac{\partial^2 f(x)}{\partial x_i \partial x_j} - \frac{\partial^2 q(x)}{\partial x_i \partial x_j} \right| \leq \frac{\epsilon}{2(1+2nM)}, \forall i, j. \tag{32}$$

Let $\Delta H(x) = H_q(x) - H_f(x)$. As f and q are twice continuously differentiable, the entries of $\Delta H(x)$ are continuous in x. This implies that $x \mapsto \lambda_{\min}(\Delta H(x))$ is continuous [11, Corollary VI.1.6]. Hence, if we let $\Lambda := \min_{x\in B} \lambda_{\min}(\Delta H(x))$, it follows that there exists $x_0 \in B$ such that $\Lambda = \lambda_{\min}\Delta H(x_0)$. We now bound this quantity. Recall that for a symmetric $n \times n$ real-valued matrix M, $||M||_{\max}$ is the max-norm of M, i.e., its largest entry in absolute value,

$$||M||_2 = \max\{|\lambda_{\min}(M)|, |\lambda_{\max}(M)|\},$$

and $||M||_2 \leq n||M||_{\max}$. From (32), we have that

$$||\Delta H(x_0)||_{\max} \leq \frac{\epsilon}{2(1+2nM)},$$

which implies that $\max\{|\lambda_{\min}(\Delta H(x_0))|, |\lambda_{\max}(\Delta H(x_0))|\} \leq \frac{n\epsilon}{2(1+2nM)}$, and so

$$-\frac{n\epsilon}{2(1+2nM)} \leq \Lambda \leq \frac{n\epsilon}{2(1+2nM)}.$$

By definition of Λ, we thus have $\Delta H(x) \succeq -\frac{n\epsilon}{2(1+2nM)}$ for all $x \in B$. Now, consider

$$p(x) := q(x) + \frac{n\epsilon}{2(1+2nM)} x^T x.$$

For any $x \in B$, we have

$$|f(x) - p(x)| \leq |f(x) - q(x)| + |q(x) - p(x)| \leq \frac{\epsilon}{2(1+2nM)} + \frac{n\epsilon}{2(1+2nM)} \cdot 2M < \epsilon.$$

Using our previous result on $\Delta H(x)$, the definition of p, and the fact that $H_f(x) \succeq 0$,

$$H_p(x) = H_p(x) - H_q(x) + H_q(x) - H_f(x) + H_f(x) \succeq \frac{2n\epsilon}{2(1+2nM)} I - \frac{n\epsilon}{2(1+2nM)} I \succ 0.$$

From this, it follows that there exists a degree d and a polynomial $p \in C_{n,d}$ such that $\sup_{x\in B} |f(x) - p(x)| < \epsilon$. The definition of g_d as the minimizer of $\sup_{x\in B} |f(x) - g(x)|$ for any $g \in C_{n,d}$ enables us to obtain the result. □

Proposition C.3 (Approximating a function with K-bounded derivatives)**.** *Let $K = (K_1^-, K_1^+, \ldots, K_n^-, K_n^+)$ be a vector of finite scalars with $K_i^- < K_i^+$ for all $i = 1, \ldots, n$ and let f be a continuously differentiable function defined over B with K-bounded derivatives. Define*

$$h_d := \arg\min_{g\in K_{n,d}} \sup_{x\in B} |f(x) - g(x)|.$$

For any $\epsilon > 0$, there exists d such that $\sup_{x\in B} |h_d(x) - f(x)| < \epsilon$.

Proof. Let $\epsilon > 0$, take $M = \max_{x\in B} ||x||_\infty$ and $M' = \max_{x\in B} |f(x)|$. From Proposition C.1, there exists a polynomial q of degree d such that

$$\max_{x\in B} |f(x) - q(x)| \leq \epsilon' \text{ and } \max_{x\in B} \left| \frac{\partial f(x)}{\partial x_i} - \frac{\partial q(x)}{\partial x_i} \right| \leq \epsilon',$$

where ϵ' is a positive scalar implicitly defined as $\epsilon = \phi(\epsilon')$ with

$$\phi(\epsilon') = \epsilon' \cdot \left(1 + \frac{nM \max_j(|K_j^+| + |K_j^-|) + 2M' + 2\epsilon'}{\min_j |K_j^+ - K_j^-|}\right).$$

Note that ϵ' is guaranteed to exist from the intermediate value theorem as $\phi(\cdot)$ is increasing, maps 0 to 0, and infinity to infinity. Now consider

$$p(x) := q(x) \cdot \left(1 - \frac{2\epsilon'}{\min_j |K_j^+ - K_j^-| + 2\epsilon'}\right) + \sum_i \epsilon' \cdot \frac{K_i^+ + K_i^-}{\min_j |K_j^+ - K_j^-| + 2\epsilon'} \cdot x_i.$$

We show that p has K-bounded derivatives and that $\sup_{x\in B} |p(x) - f(x)| \leq \epsilon$. It follows from the definition of h_d that Proposition C.3 holds. Let $x \in B$. We have

$$\begin{aligned} |p(x) - f(x)| &\leq \left| q(x) - f(x) + \epsilon' \cdot \frac{\sum_i x_i(K_i^+ + K_i^-) - 2 \cdot q(x)}{\min_j |K_j^+ - K_j^-| + 2\epsilon'} \right| \\ &\leq \epsilon' \cdot \left(1 + \frac{nM \max_j(|K_j^+| + |K_j^-|) + 2M' + 2\epsilon'}{\min_j |K_j^+ - K_j^-|}\right) = \epsilon, \end{aligned}$$

where we have used the fact that $|q(x)| \leq |f(x)| + \epsilon'$ for any $x \in B$ in the second inequality. We now show that p thus defined has K-bounded derivatives. Again, let $x \in B$ and $i \in \{1, \dots, n\}$. We have

$$\frac{\partial p(x)}{\partial x_i} = \frac{\partial q(x)}{\partial x_i} \cdot \left(1 - \frac{2\epsilon'}{\min_j |K_j^+ - K_j^-| + 2\epsilon'}\right) + \epsilon' \cdot \frac{K_i^+ + K_i^-}{\min_j |K_j^+ - K_j^-| + 2\epsilon'}.$$

As $\frac{\partial f(x)}{\partial x_i} \leq K_i^+$ and $\frac{\partial q(x)}{\partial x_i} \leq \frac{\partial f(x)}{\partial x_i} + \epsilon'$, it follows that

$$\begin{aligned} \frac{\partial p(x)}{\partial x_i} &\leq (K_i^+ + \epsilon') \cdot \left(1 - \frac{2\epsilon'}{\min_j |K_j^+ - K_j^-| + 2\epsilon'}\right) + \epsilon' \cdot \frac{K_i^+ + K_i^-}{\min_j |K_j^+ - K_j^-| + 2\epsilon'} \\ &= K_i^+ - \epsilon' \cdot \frac{K_i^+ - K_i^- - \min_j |K_j^+ - K_j^-|}{\min_j |K_j^+ - K_j^-| + 2\epsilon'} \leq K_i^+. \end{aligned}$$

Likewise, as $\frac{\partial f(x)}{\partial x_i} \geq K_i^-$ and $\frac{\partial q(x)}{\partial x_i} \geq \frac{\partial f(x)}{\partial x_i} - \epsilon'$, it follows that

$$\begin{aligned} \frac{\partial p(x)}{\partial x_i} &\geq (K_i^- - \epsilon') \cdot \left(1 - \frac{2\epsilon'}{\min_j |K_j^+ - K_j^-| + 2\epsilon'}\right) + \epsilon' \cdot \frac{K_i^+ + K_i^-}{\min_j |K_j^+ - K_j^-| + 2\epsilon'} \\ &= K_i^- + \epsilon' \cdot \frac{K_i^+ - K_i^- - \min_j |K_j^+ - K_j^-|}{\min_j |K_j^+ - K_j^-| + 2\epsilon'} \geq K_i^-. \end{aligned}$$

Hence, p has K-bounded derivatives. □

C.2 Boundedness and Lipschitz continuity of $\bar{g}_{m,d}$ and $\bar{h}_{m,d}$

In this section, we prove that $\bar{g}_{m,d}$ and $\bar{h}_{m,d}$ are uniformly upper bounded and Lipschitz continuous (with Lipschitz constants that do not depend on the data) over certain boxes contained within B. For this purpose, we introduce the following notation: let η be a scalar such that

$$0 < \eta < \min_{i=1\dots,m} \frac{u_i - l_i}{2} \tag{33}$$

and let $B_\eta := \{x \mid l_i + \eta \leq x_i \leq u_i - \eta,\ i = 1, \ldots, n\}$. If (33) holds, we have that B_η is full-dimensional and a strict subset of B ($B = B_\eta$ when $\eta = 0$), and conversely, if B_η is full-dimensional and a strict subset of B, then (33) must hold. Where possible, we use the results derived in [34], specifying where they can be found.

Proposition C.4. *Let g_d be defined as in Proposition C.2 and $\bar{g}_{m,d}$ defined as in (3). Furthermore, let η be a scalar such that (33) holds. We have the following properties:*

(i) $\exists c_\eta > 0$, which is independent of the data $(X_1, Y_1), \ldots (X_m, Y_m)$, such that $|\bar{g}_{m,d}(x)| \leq c_\eta$ a.s. for all $x \in B_{3\eta/4}$.

(ii) $\exists M_\eta > 0$, which is independent of the data $(X_1, Y_1), \ldots (X_m, Y_m)$, such that $|\bar{g}_{m,d}(x) - \bar{g}_{m,d}(y)| \leq M_\eta ||x - y||$ a.s. for all $x, y \in B_\eta$, i.e., $\bar{g}_{m,d}$ is M_η-Lipschitz over B_η.

(iii) $\exists N_\eta > 0$, which is independent of the data $(X_1, Y_1), \ldots (X_m, Y_m)$, such that $|g_d(x) - g_d(y)| \leq N_\eta ||x - y||$ for all $x, y \in B_\eta$, i.e., g_d is N_η-Lipschitz over B_η.

Proof. (i) The idea here is to control the value of $\bar{g}_{m,d}$ at the corners and the analytic center of B. Convexity of $\bar{g}_{m,d}$ enables us to conclude that $\bar{g}_{m,d}$ is upper and lower bounded a.s. over $B_{3\eta/4}$. We start by using Step 2 in [34] with $\hat{g}_n = g_{m,d}$ and $g^* = g_d$. This provides us with the following a.s. bound on $\frac{1}{m}\sum_{i=1}^m \bar{g}_{m,d}^2(X_i)$:

$$\frac{1}{m}\sum_{i=1}^m \bar{g}_{m,d}^2(X_i) \leq 9E[(Y_1 - g_d(X_1))^2] + 3E[g_d^2(X_1)] =: \beta \text{ a.s.}$$

We use this bound to show the existence of sample points X_i in the "corners" and around the analytic center of B such that $|\bar{g}_{m,d}(X_i)|$ is uniformly bounded (in m). To do this, we define for each vertex $i, i = 1, \ldots, 2^n$, of B, a box B_i^v which is included in B, has vertex i as a vertex, and has edges of length $\eta/4$. In other words, if vertex i_0 of B is given by $(l_1, u_1, u_2, \ldots, u_n)$, then the corresponding box $B_{i_0}^v$ is defined as

$$B_{i_0}^v := \{x \in \mathbb{R}^n \mid l_1 \leq x_1 \leq l_1 + \frac{\eta}{4}, u_2 - \frac{\eta}{4} \leq x_2 \leq u_2, \ldots, u_n - \frac{\eta}{4} \leq x_n \leq u_n\}.$$

We further define

$$B_0^v := \{x \in \mathbb{R}^n \mid \frac{u_i + l_i}{2} - \frac{\eta}{8} \leq x_i \leq \frac{u_i + l_i}{2} + \frac{\eta}{8}, i = 1, \ldots, n\}.$$

We refer the reader to Figure 7 for illustrations of these boxes and their relationships to other boxes appearing in the proof. Note that, for all $i = 0, \ldots, 2^n$, $B_i^v \subset B$ and is full dimensional. However, when $i \geq 1$, $B_i^v \cap B_{\eta/2} = \emptyset$ whereas $B_0^v \subseteq B_\eta \subseteq B_{3\eta/4}$. Let $\gamma_i := P(X \in B_i^v), i = 0, \ldots, 2^n$, and $\gamma := \min\{\gamma_0, \ldots, \gamma_{2^n}\}$. As B_i^v is full-dimensional for all i, it follows that $\gamma > 0$. Leveraging (16) in [34, Step 4], we obtain, for each $i \in \{0, \ldots, 2^n\}$ and for a positive scalar r such that $\frac{\beta}{r^2} \leq \frac{\gamma}{2}$, when m is large enough $\frac{1}{m}\sum_{j=1}^m P(X_j \in B_i^v, |\bar{g}_{m,d}(X_j)| \leq r) = \frac{\gamma}{2} > 0$. We use this to obtain upper and lower bounds on $\bar{g}_{m,d}(x)$ over $B_{3\eta/4}$ which only depend on the probability distribution of X_i and B_η (i.e., these bounds do not depend on the number of data points, nor on the data points themselves). The proof of the lower bound requires us to show that $\bar{g}_{m,d}$ is actually upper bounded over $B_{\eta/2}$. As $B_{\eta/2}$ is a superset of $B_{3\eta/4}$, this will imply that $\bar{g}_{m,d}$ is upper bounded over $B_{3\eta/4}$.

Upper bound: We show that $B_{\eta/2}$ is a subset of the convex hull of $X_{I(1)}, \ldots, X_{I(2^n)}$. This then implies that any x in $B_{\eta/2}$ can be written as a convex combination of these points, and so, using convexity of $\bar{g}_{m,d}$, we can conclude that $\bar{g}_{m,d}(x) \leq r$. To see that $B_{\eta/2}$ is a subset of the convex

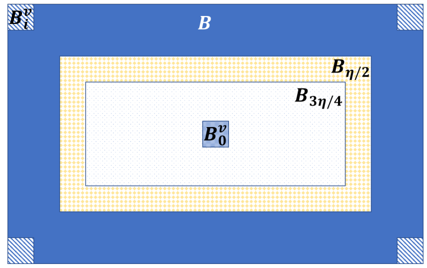


Figure 7: An illustration of the boxes that appear in the proof of Proposition C.4.

hull of $X_{I(1)}, \ldots, X_{I(2^n)}$, first note that $X_{I(i)} \notin B_{\eta/2}$ for all $i = 1, \ldots, 2^n$ as $B_i^v \cap B_{\eta/2} = \emptyset$. Hence, either $B_{\eta/2}$ is a subset of convex hull of $X_{I(1)}, \ldots, X_{I(2^n)}$ or the two sets are disjoint. We show that the former has to hold. This follows from the fact that $X_0 = \frac{1}{2^n}\sum_{i=1}^{2^n} X_{I(i)}$, which is in the convex hull of $X_{I(1)}, \ldots, X_{I(2^n)}$, is also in $B_{\eta/2}$. To see this, note that for a fixed component k of the vectors $\{X_{I(i)}\}_i$, there are exactly 2^{n-1} of these components that belong to $[l_k, l_k + \frac{\eta}{4}]$ and 2^{n-1} that belong to $[u_k - \frac{\eta}{4}, u_k]$. This implies that the k-th component of X_0 belongs to the interval $[\frac{u_k+l_k}{2} - \frac{\eta}{8}; \frac{u_k+l_k}{2} + \frac{\eta}{8}]$. As $\frac{u_k+l_k}{2} + \frac{\eta}{8} \leq u_k - \frac{\eta}{2}$ and $\frac{u_k+l_k}{2} - \frac{\eta}{8} \geq l_k + \frac{\eta}{2}$ by consequence of (33), we get that X_0 is in $B_{\eta/2}$.

Lower bound: Let $x \in B_{3\eta/4}$. We use the lower bound proof in [34, Step 4] to obtain that $\bar{g}_{m,d}(x) \geq -3r$. Taking $c_\eta = \max\{r, 3r\} = 3r$ gives us the expected result.

(ii) Similarly to [34, Step 5], as $\bar{g}_{m,d}$ is convex over B and a.s. bounded on $B_{3\eta/4}$ by c_η from (i), there exists a constant $M_\eta = \frac{8c_\eta}{\eta}$ which is independent of the data, such that $\bar{g}_{m,d}$ is M_η-Lipschitz over B_η; for a proof of this, see [53, Theorem A].

(iii) As g_d is continuous over B, g_d has a maximum over B. Furthermore, g_d is convex over B. The result follows, using a similar argument to (ii).

□

Proposition C.5. *Let h_d be defined as in Proposition C.3 and $\bar{h}_{m,d}$ defined as in (4). The following properties hold:*

(i) $\exists M'_\eta > 0$, which is independent of the data $(X_1, Y_1), \ldots (X_m, Y_m)$, such that $|\bar{h}_{m,d}(x) - \bar{h}_{m,d}(y)| \leq M'_\eta ||x - y||$ a.s. for all $x, y \in B_\eta$, i.e., $\bar{h}_{m,d}$ is M'_η-Lipschitz over B_η.

(ii) $\exists N'_\eta > 0$, which is independent of the data $(X_1, Y_1), \ldots (X_m, Y_m)$, such that $|h_d(x) - h_d(y)| \leq N'_\eta ||x - y||$ for all $x, y \in B_\eta$, i.e., h_d is N'_η-Lipschitz over B_η.

Proof. This follows immediately from the fact that both h_d and $\bar{h}_{m,d}$ have K-bounded derivatives, with K being a vector of finite scalars. □

C.3 Proof of Lemmas 3.3 and 3.4

Proof of Lemma 3.3. We define $C_{n,d}$ and g_d as previously. Let $\epsilon > 0$. This proof has three steps: the first step establishes that one can obtain an arbitrarily good approximation of f by a family of convex polynomials $\{g_d\}_d$ and that showing consistency of $\bar{g}_{m,d}$ over any compact set C in B can be done by showing consistency of $\bar{g}_{m,d}$ over B_η for some η such that (33) holds. In the second step, we show that g_d and $\bar{g}_{m,d}$ are "close" on the random samples X_i; this is then used in the third step to show that the two functions are uniformly close and hence that f and $\bar{g}_{m,d}$ are also uniformly close. Step 1 is new compared to [34], Step 2 follows quite closely parts of the proof of [34] but has to account for the box constraints among other minor details, Step 3 is close to parts of [34] but has to account for the presence of g_d.

Step 1: approximating f by a convex polynomial g_d. From Proposition C.2, there exists $d := d(\epsilon)$ such that

$$\sup_{x \in B} |g_d(x) - f(x)| \leq \frac{\epsilon}{2}, \tag{34}$$

where g_d is defined as in Proposition C.2. Henceforth, we fix d to this value.

We now prove that the problem of showing consistency of $\bar{g}_{m,d}$ over any compact subset C of B can be reduced to showing consistency of $\bar{g}_{m,d}$ over some box B_η where η is such that (33) holds. Let C be any full-dimensional compact subset of B such that no point of the boundary of B is in C. As $C \cap int(B) = C$, there exists $\eta_C > 0$ such that $C \subseteq B_{\eta_C}$. Furthermore, there exists $\eta_\epsilon > 0$ such that

$$2\sqrt{2E[((Y_1 - g_d(X_1))^2 \mathbf{1}(X_1 \notin B_{\eta_\epsilon})]} \cdot \sqrt{5E[(Y_1 - g_d(X_1))^2]} \leq \epsilon. \tag{35}$$

To see this, note that as $\eta \to 0$, $P(X_1 \notin B_\eta) \to 0$ with $P(X_1 \notin B) = 0$ (this is a consequence of $P(X \in A) > 0$ for any full-dimensional set A). Existence of η_ϵ then follows by expanding out the expression and using Assumptions 1 and 3 together with the fact that both f and g_d are continuous over B and so bounded over B. We let $\eta := \min\{\eta_C, \eta_\epsilon\}$. Thus defined, η is such that (33) holds as C is full-dimensional and a subset of B_η. As a consequence, in the rest of the proof, we restrict ourselves to showing that $\sup_{x \in B_\eta} |\bar{g}_{m,d}(x) - f(x)| \to 0$ a.s. when $m, d \to \infty$ instead of (16). Indeed, this implies (16) as $C \subseteq B_\eta$ but B_η is simpler than C to work with.

Step 2: showing that, for fixed d, $\frac{1}{m}\sum_{i=1}^m (\bar{g}_{m,d}(X_i) - g_d(X_i))^2 \to 0$ a.s. as $m \to \infty$ Following Step 1 in [34], we have:

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m (\bar{g}_{m,d}(X_i) - g_d(X_i))^2 \leq & \frac{2}{m}\sum_{i=1}^m (Y_i - g_d(X_i))(\bar{g}_{m,d}(X_i) - g_d(X_i))\mathbf{1}(X_i \in B_\eta) \\ & + \frac{2}{m}\sum_{i=1}^m (Y_i - g_d(X_i))(\bar{g}_{m,d}(X_i) - g_d(X_i))\mathbf{1}(X_i \notin B_\eta)\end{aligned} \tag{36}$$

Step 3 of [34] combined to (35) enables us to show that for large enough m,

$$\frac{2}{m}\sum_{i=1}^m (Y_i - g_d(X_i))(\bar{g}_{m,d}(X_i) - g_d(X_i))\mathbf{1}(X_i \notin B_\eta) \leq \epsilon \text{ a.s..} \tag{37}$$

We now focus on the term that includes the sample points inside of B_η. As done in [34], we replace $\bar{g}_{m,d}$ by a deterministic approximation and then apply the strong law of large numbers. Let

$$\mathcal{C} = \{\text{polynomials } p : B_\eta \mapsto \mathbb{R} \text{ of degree } d,\ M_\eta\text{-Lipschitz with } |p(x)| \leq c_\eta, \forall x \in B_\eta\},$$

where M_η and c_η are the constants given in Proposition C.4, which do not depend on the data $(X_1, Y_1), \ldots, (X_m, Y_m)$. Proposition C.4 implies that $\bar{g}_{m,d}$ belongs to $\mathcal{C}$ for large enough m. Furthermore, given that $\mathcal{C}$ is a subset of the set of continuous functions over the box B_η and given that all functions in $\mathcal{C}$ are uniformly bounded and Lipschitz, it follows from Ascoli-Arzelá's theorem that

$\mathcal{C}$ is compact in the metric $d(f,g) = \sup_{x \in B_\eta} |f(x) - g(x)|$. As a consequence, $\mathcal{C}$ has a finite ϵ-net: we denote by $p_1, \ldots, p_R$ the polynomials belonging to it. Hence, for large enough m, there exists $r \in \{1, \ldots, R\}$ such that $\sup_{x \in B_\eta} |p_r(x) - \bar{g}_{m,d}(x)| < \epsilon$, which is the deterministic approximation of $\bar{g}_{m,d}$. Following Step 7 of [34], we show that:

$$\begin{aligned}
&\frac{2}{m} \sum_{i=1}^{m} (Y_i - g_d(X_i))(\bar{g}_{m,d}(X_i) - g_d(X_i)) \mathbf{1}(X_i \in B_\eta) \\
&\leq \frac{2}{m} \cdot \epsilon \sum_{i=1}^{m} |Y_i - g_d(X_i)| + \max_{j=1,\ldots,R} \frac{2}{m} \sum_{i=1}^{m} (Y_i - g_d(X_i))(p_j(X_i) - g_d(X_i)) \mathbf{1}(X_i \in B_\eta).
\end{aligned}$$

As g_d is bounded over B, we use the strong law of large numbers to obtain

$$\frac{1}{m} \sum_{i=1}^{m} |Y_i - g_d(X_i)| \leq 2E[|Y_1 - g_d(X_1)|] \text{ a.s., for large enough } m. \tag{38}$$

We also have for any $j \in \{1, \ldots, R\}$, using C.2 and 3,

$$\begin{aligned}
&\frac{2}{m} \sum_{i=1}^{m} (Y_i - g_d(X_i))(p_j(X_i) - g_d(X_i)) \mathbf{1}(X_i \in B_\eta) \\
&= \frac{2}{m} \sum_{i=1}^{m} (Y_i - f(X_i) + f(X_i) - g_d(X_i))(p_j(X_i) - g_d(X_i)) \mathbf{1}(X_i \in B_\eta) \\
&\leq \frac{2}{m} \sum_{i=1}^{m} \nu_i (p_j(X_i) - g_d(X_i)) \mathbf{1}(X_i \in B_\eta) + \frac{2}{m} \epsilon \sum_{i=1}^{m} |p_j(X_i) - g_d(X_i)| \mathbf{1}(X_i \in B_\eta).
\end{aligned}$$

Given Assumption 1 and the fact that h_j is uniformly bounded over B_η, from the strong law of large numbers, for any $j \in \{1, \ldots, r\}$ and for large enough m, we have

$$\frac{2}{m} \sum_{i=1}^{m} \nu_i (p_j(X_i) - g_d(X_i)) \mathbf{1}(X_i \in B_\eta) \leq \epsilon \text{ a.s.} \tag{39}$$

Similarly, using the strong law of large numbers again, for large enough m,

$$\frac{1}{m} \sum_{i=1}^{m} |p_j(X_i) - g_d(X_i)| \mathbf{1}(X_i \in B_\eta) \leq 2E[|p_j(X_1) - g_d(X_1)| \mathbf{1}(X_1 \in B_\eta)] \text{ a.s.} \tag{40}$$

Combining (38), (39), and (40), we conclude that for large enough m,

$$\begin{aligned}
&\frac{2}{m} \sum_{i=1}^{m} (Y_i - g_d(X_i))(\bar{g}_{m,d}(X_i) - g_d(X_i)) \mathbf{1}(X_i \in B_\eta) \\
&\leq 2\epsilon \left(2E[|Y_1 - g_d(X_1)|] + \max_{j=1,\ldots,R} 2E[|p_j(X_1) - g_d(X_1)| \mathbf{1}(X_1 \in B_\eta)] \right) + \epsilon.
\end{aligned}$$

Combining this with (37) in (36), we obtain our conclusion for fixed d.

Step 3: showing that $\sup_{x \in B_\eta} |f(x) - \bar{g}_{m,d}(x)| \to 0$ **a.s. when** $m \to \infty$ **and** $d \to \infty$ We fix d as previously and m to be as large as needed. Let $x \in B_\eta$ and let δ be a fixed positive scalar such that $(M_\eta + N_\eta)\delta \leq \epsilon/4$. As B_η is compact, there exists a finite partition $C_1, \ldots, C_K$ of B_η such that the diameter of C_k, $k = 1, \dots, K$, is less than δ (i.e., $\sup_{x,y \in C_k} ||x - y|| \leq \delta$) and C_k is full dimensional. It follows from Assumption 2 that each C_k contains at least one X_i. Furthermore, as $x \in B_\eta$, $x \in C_k$ for some k_0. Let $i_{k_0} = \arg\min_{\{i | X_i \in C_{k_0}\}} |g_d(X_i) - \bar{g}_{m,d}(X_i)|$. For d fixed and large m, we have

$$\begin{aligned} |f(x) - \bar{g}_{m,d}(x)| &\leq |f(x) - g_d(x)| + |g_d(x) - g_d(X_{i_{k_0}})| + |g_d(X_{i_{k_0}}) - \bar{g}_{m,d}(X_{i_{k_0}})| \\ &\quad + |\bar{g}_{m,d}(X_{i_{k_0}}) - \bar{g}_{m,d}(x)| \\ &\leq \frac{\epsilon}{2} + N_\eta \cdot \delta + \frac{\sum_{i=1}^m |g_d(X_i) - \bar{g}_{m,d}(X_i)| I(X_i \in C_{k_0})}{\sum_{i=1}^m I(X_i \in C_{k_0})} + M_\eta \cdot \delta, \end{aligned}$$

where we have used the fact that both g_d and $\bar{g}_{m,d}$ are Lipschitz (for m large enough in the case of $\bar{g}_{m,d}$) with Lipschitz constants N_η and M_η respectively, which do not depend on the data (see Proposition C.4), together with the fact that the minimum of a vector is less than or equal to its average. This implies that

$$\sup_{x \in B_\eta} |f(x) - \bar{g}_{m,d}(x)| \leq \frac{\epsilon}{2} + \delta(M_\eta + N_\eta) + \frac{1}{m} \sum_{i=1}^m |g_d(X_i) - \bar{g}_{m,d}(X_i)| \cdot \max_{k=1,\dots,K} \frac{m}{\sum_{i=1}^m I(X_i \in C_k)}$$

We leverage [34, Step 8] and Step 2 to bound, for large enough m,

$$\frac{1}{m} \sum_{i=1}^m |g_d(X_i) - \bar{g}_{m,d}(X_i)| \leq \frac{\epsilon \cdot \min_{k=1,\dots,K} P(X_1 \in C_k)}{8} \text{ a.s.}$$

Thus, using the definition of δ, $\sup_{x \in B_\eta} |f(x) - \bar{g}_{m,d}(x)| \leq \epsilon$ for very large m. □

Proof of Lemma 3.4. The same proof goes through providing that $C_{n,d}$ is replaced by $K_{n,d}$, g_d by h_d, $\bar{g}_{m,d}$ by $\bar{h}_{m,d}$, Proposition C.2 by Proposition C.3, and Proposition C.4 by Proposition C.5. □

Acknowledgments

We would like to thank Amir Ali Ahmadi for bringing the problem of shape-constrained regression to our attention as well as for his suggestions and feedback. We would also like to thank Ioana Popescu and Johan Löfberg for their helpful comments and Rahul Mazumder and Dimitrije Ruzic for providing access to relevant datasets.

References

[1] *MOSEK reference manual*, 2020. Version 9. Available at `http://www.mosek.com/`.

[2] N. E. Aguilera and P. Morin. Approximating optimization problems over convex functions. *Numerische Mathematik*, 111(1):1–34, 2008.

[3] N. E. Aguilera and P. Morin. On convex functions and the finite element method. *SIAM Journal on Numerical Analysis*, 47(4):3139–3157, 2009.

[4] A. A. Ahmadi and G. Hall. Sum of squares basis pursuit with linear and second order cone programming. *Algebraic and Geometric Methods in Discrete Mathematics*, 685:27–53, 2017.

[5] A. A. Ahmadi and G. Hall. On the complexity of detecting convexity over a box. *Mathematical Programming*, pages 1–15, 2019.

[6] Y. Ait-Sahalia and J. Duarte. Nonparametric option pricing under shape restrictions. *Journal of Econometrics*, 116(1-2):9–47, 2003.

[7] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

[8] A. Ben-Tal, B. Golany, A. Nemirovski, and J.-P. Vial. Retailer-supplier flexible commitments contracts: A robust optimization approach. *Manufacturing & Service Operations Management*, 7(3):248–271, 2005.

[9] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific Belmont, 1995.

[10] D. P. Bertsekas. *Convex Optimization Theory*. Athena Scientific Belmont, 2009.

[11] R. Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013.

[12] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[13] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[14] V. Charpentier, S. Adriaenssens, F. Meggers, and O. Baverel. Comfort driven optimization and control for adaptive shading applications. *Building Information and Research*, 2017.

[15] W. Chen and R. Mazumder. Multivariate convex regression at scale. *arXiv preprint arXiv:2005.11588*, 2020.

[16] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in Neural Information Processing Systems*, pages 472–478, 2001.

[17] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia. Incorporating functional knowledge in neural networks. *Journal of Machine Learning Research*, 10(Jun):1239–1262, 2009.

[18] A. Fellhauer. Approximation of smooth functions using Bernstein polynomials in multiple variables. *arXiv preprint arXiv:1609.01940*, 2016.

[19] C. R. Gallistel, S. Fairhurst, and P. Balsam. The learning curve: implications of a quantitative analysis. *Proceedings of the National Academy of Sciences*, 101(36):13124–13131, 2004.

[20] M. R. Garey and D. S. Johnson. *Computers and Intractability*. W. H. Freeman and Co., San Francisco, Calif., 1979.

[21] S. Gershgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et naturelles*, 6:749–754, 1931.

[22] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

[23] G. Hall. *Optimization over nonnegative and convex polynomials with and without semidefinite programming*. PhD thesis, Princeton University, 2018.

[24] D. Handelman. Representing polynomials by positive linear functions on compact convex polyhedra. *Pacific Journal of Mathematics*, 132(1):35–62, 1988.

[25] L. Hannah and D. Dunson. Ensemble methods for convex regression with applications to geometric programming based circuit design. *arXiv preprint arXiv:1206.4645*, 2012.

[26] L. A. Hannah and D. B. Dunson. Multivariate convex regression with adaptive partitioning. *The Journal of Machine Learning Research*, 14(1):3261–3294, 2013.

[27] C. Hildreth. Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, 49(267):598–619, 1954.

[28] C. A. Holloway. On the estimation of convex functions. *Operations Research*, 27(2):401–407, 1979.

[29] T. L. Jernigan, S. L. Archibald, C. Fennema-Notestine, A. C. Gamst, J. C. Stout, J. Bonner, and J. R. Hesselink. Effects of age on tissues and regions of the cerebrum and cerebellum. *Neurobiology of aging*, 22(4):581–594, 2001.

[30] H. Kay and L. H. Ungar. Estimating monotonic functions and their bounds. *AIChE Journal*, 46(12):2426–2434, 2000.

[31] T. Kuosmanen. Representation theorem for convex nonparametric least squares. *The Econometrics Journal*, 11(2):308–325, 2008.

[32] J. B. Lasserre. Convexity in semialgebraic geometry and polynomial optimization. *SIAM Journal on Optimization*, 19(4):1995–2014, 2009.

[33] X. Li, D. Sun, and K.-C. Toh. QSDPNAL: A two-phase augmented Lagrangian method for convex quadratic semidefinite programming. *Mathematical Programming Computation*, 10(4):703–743, 2018.

[34] E. Lim and P. W. Glynn. Consistency of multidimensional convex regression. *Operations Research*, 60(1):196–208, 2012.

[35] M. Lin, D. Sun, and K.-C. Toh. Efficient algorithms for multivariate shape-constrained convex regression problems. *arXiv preprint arXiv:2002.11410*, 2020.

[36] J. Löfberg. YALMIP : A toolbox for modeling and optimization in MATLAB. In *In Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

[37] A. Magnani and S. P. Boyd. Convex piecewise-linear fitting. *Optimization and Engineering*, 10(1):1–17, 2009.

[38] A. Magnani, S. Lall, and S. Boyd. Tractable fitting with convex polynomials via sum of squares. In *Proceedings of the 44^{th} IEEE Conference on Decision and Control*, 2005.

[39] A. Majumdar, G. Hall, and A. A. Ahmadi. Recent scalability improvements for semidefinite programming with applications in machine learning, control, and robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:331–360, 2020.

[40] R. Mazumder, A. Choudhury, G. Iyengar, and B. Sen. A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association*, (just-accepted), 2017.

[41] R. F. Meyer and J. W. Pratt. The consistent assessment and fairing of preference functions. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):270–278, 1968.

[42] D. R. Miller and A. Sofer. Completely monotone regression estimates of software failure rates. In *Proceedings of the 8th international conference on Software engineering*, pages 343–348. IEEE Computer Society Press, 1985.

[43] H. Narasimhan, A. Cotter, M. Gupta, and S. L. Wang. Pairwise fairness for ranking and regression. In *33rd AAAI Conference on Artificial Intelligence*, 2020.

[44] R. Nicolaides. On a class of finite elements generated by Lagrange interpolation. *SIAM Journal on Numerical Analysis*, 9(3):435–445, 1972.

[45] B. O'donoghue, E. Chu, N. Parikh, and S. Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, 2016.

[46] D. Papp and F. Alizadeh. Shape-constrained estimation using nonnegative splines. *Journal of Computational and graphical Statistics*, 23(1):211–231, 2014.

[47] D. Papp and S. Yildiz. Sum-of-squares optimization without semidefinite programming. *SIAM Journal on Optimization*, 29(1):822–851, 2019.

[48] F.-P. Paty, A. d'Aspremont, and M. Cuturi. Regularity as regularization: Smooth and strongly convex Brenier potentials in optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1222–1232. PMLR, 2020.

[49] G. Peyré and M. Cuturi. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[50] V. Powers and B. Reznick. Polynomials that are positive on an interval. *Transactions of the American Mathematical Society*, 352(10):4677–4692, 2000.

[51] J. L. Prince and A. S. Willsky. Reconstructing convex sets from support line measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(4):377–389, 1990.

[52] J. Rabin, S. Ferradans, and N. Papadakis. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4852–4856. IEEE, 2014.

[53] A. W. Roberts and D. E. Varberg. Another proof that convex functions are locally Lipschitz. *The American Mathematical Monthly*, 81(9):1014–1016, 1974.

[54] M. J. Schell and B. Singh. The reduced monotonic regression method. *Journal of the American Statistical Association*, 92(437):128–135, 1997.

[55] C. W. Scherer and C. W. J. Hol. Matrix sum-of-squares relaxations for robust semi-definite programs. *Mathematical Programming*, 107(1):189–211, 2006.

[56] E. Seijo and B. Sen. Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39(3):1633–1657, 2011.

[57] M. Sipser. Introduction to the theory of computation. *ACM Sigact News*, 27(1):27–29, 1996.

[58] S. Wang. Neural network techniques for monotonic nonlinear models. *Computers & Operations Research*, 21(2):143–154, 1994.

[59] Y. Wang, L. Li, and C. Dang. Calibrating classification probabilities with shape-restricted polynomial regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.