**Research Notes**

# Nonparametric Estimation using SOS-Convexity

Prof. John Lafferty Students: YJ Choe, Max Cytrynbaum, Wei Hu

## 1 Introduction

(YJ will write out this section summerizing our first-day discussion when he has time.)

## 2 SOS-Convex Regression

Given $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ for $i = 1, \dots, n$, recall that we have the equivalence between the following optimization problems:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2 \\ \text{s.t.} \quad & f \text{ is convex.} \end{aligned} \tag{1}$$

$$\begin{aligned} \underset{\mathbf{z},\boldsymbol{\beta}}{\text{minimize}} \quad & \sum_{i=1}^n (y_i - z_i)^2 \\ \text{s.t.} \quad & z_j \geq z_i + \boldsymbol{\beta}_i^T(\mathbf{x}_j - \mathbf{x}_i) \qquad \forall\ i, j = 1, \dots, n. \end{aligned} \tag{2}$$

In particular, we can reduce the infinite-dimensional problem (1) into a finite-dimensional quadratic program (QP) (2), which can be efficiently solved. The solution to (2) can be viewed as a piecewise-linear convex function.

Here, we attempt to derive the analogous equivalence, i.e. find an equivalent convex optimization problem to the following infinite-dimensional problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2 \\ \text{s.t.} \quad & f \text{ is an SOS-convex polynomial of degree } 2d. \end{aligned} \tag{3}$$

Denote the vector of basis monomials up to degree $d$ by $\mathbf{v}_d(\mathbf{x}) = (1, x_1, \dots, x_p, x_1x_2, \dots, x_p^d)$, where $\mathbf{x} = (x_1, \dots, x_p)$. Also, let $s = \binom{2d+p}{p}$ denote the length of this vector for degree $2d$. Then, we may replace $f$ with a coefficient vector $\boldsymbol{\theta} \in \mathbb{R}^s$, such that

$$f(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{v}_{2d}(\mathbf{x}).$$

Note the one-to-one correspondence between $f$ and $\boldsymbol{\theta}$. Further, as done with the convex program, we introduce the auxiliary variable $\mathbf{z} = (z_1, \dots, z_n)$ so that

$$f(\mathbf{x}_i) = \boldsymbol{\theta}^T \mathbf{v}_{2d}(\mathbf{x}_i) = z_i \qquad \forall\ i = 1, \dots, n. \tag{4}$$

We can write this more concisely by introducing the matrix

$$V = V(\mathbf{x}_1, \dots, \mathbf{x}_n) = \begin{bmatrix} \mathbf{v}_d(\mathbf{x}_1) \\ \vdots \\ \mathbf{v}_d(\mathbf{x}_n) \end{bmatrix}_{n \times s}$$

so that (4) simply becomes

$$V\boldsymbol{\theta} = \mathbf{z}. \tag{5}$$

So we have a linear constraint on the coefficient $\theta$ that is equivalent to saying that the polynomial interpolates the points $\{(\mathbf{x}_i, z_i)\}_{i=1}^n$. Analogously, we can rewrite the objective to be

$$\sum_{i=1}^{n}(y_i - z_i)^2 = \|\mathbf{y} - \mathbf{z}\|^2 \tag{6}$$

where $\mathbf{y} = (y_1, \dots, y_n)$ and $\mathbf{z} = (z_1, \dots, z_n)$.

Now we want to rewrite the constraint that $f(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{v}_d(\mathbf{x})$ is SOS-convex. Recall that $p$ is SOS-convex if and only if

$$\mathbf{u}^T H_{\boldsymbol{\theta}}(\mathbf{x})\mathbf{u} = \mathbf{v}_d(\mathbf{x}, \mathbf{u})^T Q \mathbf{v}_d(\mathbf{x}, \mathbf{u}) \tag{7}$$

and

$$Q \succeq 0 \tag{8}$$

where $Q$ is a symmetric $r \times r$ matrix with $r = \binom{d+2p}{2p}$, $H_\theta(\mathbf{x})$ is the Hessian polynomial matrix of $p(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{v}_d(\mathbf{x})$, and $\mathbf{u} = (u_1, \dots, u_p)$.

Note first that (7) is an equality in the space of *polynomials* on $2p$ variables: $\mathbf{x}$ and $\mathbf{u}$. Also, (8) **suggests that the convex optimization program that we attempt to build is a semidefinite program (SDP).**

It is important to note that, **in this case where $f$ is a polynomial, the Hessian $H_\theta(\mathbf{x})$ is easy to compute – in fact, the coefficients of each entry of the Hessian are a linear function of the coefficient $\boldsymbol{\theta}$ of the original polynomial $f$.** For a multinomial index $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p) \in \mathbb{N}^p$ on $\mathbf{x}$ such that $\sum_{j=1}^p \alpha_j \leq d$, recall that the Hessian of $\mathbf{x}^{\boldsymbol{\alpha}}$ with respect to $x_i$ and $x_j$ is

$$\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}\mathbf{x}^{\boldsymbol{\alpha}} = \alpha_i \alpha_j \mathbf{x}^{\boldsymbol{\alpha}'_{i,j}}$$

where $\boldsymbol{\alpha}'_{i,j} = (\alpha_1, \dots, \alpha_i - 1, \dots, \alpha_j - 1, \dots, \alpha_p)$ for $i \neq j$ and $\boldsymbol{\alpha}'_{i,i} = (\alpha_1, \dots, \alpha_i - 2, \dots, \alpha_p)$. Note that if $\alpha_i = 0$ for any $i$ then the Hessian is zero anyway.

(7) is not a valid semidefinite constraint yet, because it is an equality between two polynomials. This means we want to equate the *coefficients* of the two polynomials on $(\mathbf{x}, \mathbf{y})$. Note first that

$$\mathbf{u}^T H_{\boldsymbol{\theta}}(\mathbf{x})\mathbf{u} = \sum_{i,j=1}^{p}\left(\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}f(\mathbf{x})\right)u_i u_j. \tag{9}$$

Then, there is at most one term (zero iff the partial derivative is zero) for each $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p) \in \mathbb{N}^p$ such that $\sum_{k=1}^p \alpha_k \leq d$, and for each $i, j = 1, \dots, p$. For each of these cases, we give a multinomial index $\boldsymbol{\gamma}_{\boldsymbol{\alpha},i,j} \in \mathbb{N}^{2p}$ on $(\mathbf{x}, \mathbf{u})$, and $\boldsymbol{\gamma}_{\boldsymbol{\alpha},i,j}$ will have a few specific properties: the first $p$ coordinates are precisely $\boldsymbol{\alpha}'_{i,j}$, which sum up to at most $d-2$, and the last $p$ coordinates are all zero except at the $(p+i)$th and the $(p+j)$th coordinate (only at the $(p+i)$th if $i = j$). The first corresponds to the fact that the Hessian can have at most degree $d-2$, and the second to the fact that each term has exactly one degree on $u_i$ and $u_j$. Thus, we have

$$\mathbf{u}^T H_{\boldsymbol{\theta}}(\mathbf{x})\mathbf{u} = \sum_{\boldsymbol{\alpha}}\sum_{i,j} H_{\boldsymbol{\theta}}(\boldsymbol{\alpha}, i, j)(\mathbf{x}, \mathbf{u})^{\boldsymbol{\gamma}_{\boldsymbol{\alpha},i,j}} \tag{10}$$

where $H_{\boldsymbol{\theta}}(\boldsymbol{\alpha}, i, j)$ represents the scalar coefficient for the term $(\mathbf{x}, \mathbf{u})^{\boldsymbol{\gamma}_{\boldsymbol{\alpha},i,j}}$.

Further, we can express the right-hand side in terms of their coordinates in the following way. First define the coordinate matrix $B_{\boldsymbol{\gamma}}$ for each multi-index $\boldsymbol{\gamma} \in \mathbb{N}^{2p}$ up to degree $2d$ such that

$$\mathbf{v}_d(\mathbf{x}, \mathbf{u})\mathbf{v}_d(\mathbf{x}, \mathbf{u})^T = \sum_{\boldsymbol{\gamma}} B_{\boldsymbol{\gamma}}(\mathbf{x}, \mathbf{u})^{\boldsymbol{\gamma}}.$$

Note that the matrices $B_{\boldsymbol{\gamma}}$ are simply "constants", i.e. they do not depend on the data or the program variables. With this, the right-hand side becomes

$$\begin{aligned}
\mathbf{v}_d(\mathbf{x}, \mathbf{u})^T Q \mathbf{v}_d(\mathbf{x}, \mathbf{u}) &= \operatorname{tr}(Q\mathbf{v}_d(\mathbf{x}, \mathbf{u})\mathbf{v}_d(\mathbf{x}, \mathbf{u})^T) \\
&= \langle Q, \mathbf{v}_d(\mathbf{x}, \mathbf{u})\mathbf{v}_d(\mathbf{x}, \mathbf{u})^T \rangle \\
&= \left\langle Q, \sum_{\boldsymbol{\gamma}} B_{\boldsymbol{\gamma}}(\mathbf{x}, \mathbf{u})^{\boldsymbol{\gamma}} \right\rangle \\
&= \sum_{\boldsymbol{\gamma}} \langle Q, B_{\boldsymbol{\gamma}} \rangle (\mathbf{x}, \mathbf{u})^{\boldsymbol{\gamma}}
\end{aligned} \tag{11}$$

where $\langle A, B\rangle = \operatorname{tr}(A^T B) = \sum_{i,j} A_{ij}B_{ij}$ is the matrix inner product. Note that $Q$ is symmetric.

Then, we can equate the coefficients of (10) and (11) to obtain:

$$\left\langle Q, B_{\boldsymbol{\gamma}_{\boldsymbol{\alpha},i,j}} \right\rangle = H_{\boldsymbol{\theta}}(\boldsymbol{\alpha}, i, j) \qquad \forall\ \boldsymbol{\alpha}, i, j \tag{12}$$

$$\langle Q, B_{\boldsymbol{\gamma}} \rangle = 0 \qquad \text{for all other } \boldsymbol{\gamma} \tag{13}$$

Putting (5), (6), (8), (12), and (13) together, (3) can be restated as the following problem:

$$\begin{aligned}
\underset{\mathbf{z},\boldsymbol{\theta},Q}{\text{minimize}} \quad & \|\mathbf{y} - \mathbf{z}\|^2 \\
\text{s.t.} \quad & V\boldsymbol{\theta} = \mathbf{z} \\
& \langle Q, B_{\boldsymbol{\gamma}_{\boldsymbol{\alpha},i,j}} \rangle = H_{\boldsymbol{\theta}}(\boldsymbol{\alpha}, i, j) \qquad \forall\ \boldsymbol{\alpha}, i, j \\
& \langle Q, B_{\boldsymbol{\gamma}} \rangle = 0 \qquad \text{for all other } \boldsymbol{\gamma} \\
& Q \succeq 0
\end{aligned} \tag{14}$$

(14) is almost an SDP, except that the objective is quadratic. But in general, we can introduce another auxiliary variable $t$ to restate the problem as

$$\begin{aligned}
\underset{t,\mathbf{z},\boldsymbol{\theta},Q}{\text{minimize}} \quad & t \\
\text{s.t.} \quad & \|\mathbf{y} - \mathbf{z}\|^2 \leq t \\
& V\boldsymbol{\theta} = \mathbf{z} \\
& \langle Q, B_{\boldsymbol{\gamma}_{\boldsymbol{\alpha},i,j}} \rangle = H_{\boldsymbol{\theta}}(\boldsymbol{\alpha}, i, j) \qquad \forall\ \boldsymbol{\alpha}, i, j \\
& \langle Q, B_{\boldsymbol{\gamma}} \rangle = 0 \qquad \text{for all other } \boldsymbol{\gamma} \\
& Q \succeq 0
\end{aligned} \tag{15}$$

Then, we are left with a quadratic inequality constraint. Fortunately, the following allows us to convert this into a semidefinite constraint.

**Fact 2.1** *For any* $\mathbf{x}, \mathbf{q} \in \mathbb{R}^p$ *and* $r \in \mathbb{R}$, $\mathbf{x}^T\mathbf{x} + \mathbf{q}^T\mathbf{x} + r \leq 0$ *if and only if* $\begin{bmatrix} I & -\mathbf{x} \\ -\mathbf{x}^T & -\mathbf{q}^T\mathbf{x} - r \end{bmatrix} \succeq 0.$

**Proof:** For any $\mathbf{y} \in \mathbb{R}^p$ and $z \in \mathbb{R}$,

$$\begin{aligned} \begin{bmatrix} \mathbf{y}^T & z \end{bmatrix} \begin{bmatrix} I & -\mathbf{x} \\ -\mathbf{x}^T & -\mathbf{q}^T\mathbf{x} - r \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ z \end{bmatrix} &= \mathbf{y}^T\mathbf{y} - 2z\mathbf{x}^T\mathbf{y} - z^2(\mathbf{q}^T\mathbf{x} + r) \\ &= \|\mathbf{y} - z\mathbf{x}\|^2 - z^2(\mathbf{x}^T\mathbf{x} + \mathbf{q}^T\mathbf{x} + r). \end{aligned}$$

If $\mathbf{x}^T\mathbf{x} + \mathbf{q}^T\mathbf{x} + r \leq 0$, then this is nonnegative for all $\mathbf{y} \in \mathbb{R}^p$ and $z \in \mathbb{R}$. Otherwise, one can find $\mathbf{y} \in \mathbb{R}^p$ and $z \in \mathbb{R}$ such that this is strictly negative. ■

Thus,

$$\begin{aligned} \|\mathbf{y} - \mathbf{z}\|^2 \leq t &\iff \mathbf{z}^T\mathbf{z} - 2\mathbf{y}^T\mathbf{z} + (\mathbf{y}^T\mathbf{y} - t) \leq 0 \\ &\iff \begin{bmatrix} I & -\mathbf{z} \\ -\mathbf{z}^T & 2\mathbf{y}^T\mathbf{z} - \mathbf{y}^T\mathbf{y} + t \end{bmatrix} \succeq 0. \end{aligned}$$

Note that the last relation is a linear matrix inequality (LMI), i.e. it says that a linear combination of symmetric matrices is positive semidefinite.

Thus, we can now write (15) into a semidefinite program:

$$\begin{aligned} \underset{t,\mathbf{z},\boldsymbol{\theta},Q}{\text{minimize}} \quad & t \\ \text{s.t.} \quad & \begin{bmatrix} I & -\mathbf{z} \\ -\mathbf{z}^T & 2\mathbf{y}^T\mathbf{z} - \mathbf{y}^T\mathbf{y} + t \end{bmatrix} \succeq 0 \\ & V\boldsymbol{\theta} = \mathbf{z} \\ & \langle Q, B_{\boldsymbol{\gamma}_{\boldsymbol{\alpha},i,j}} \rangle = H_{\boldsymbol{\theta}}(\boldsymbol{\alpha}, i, j) \qquad \forall\ \boldsymbol{\alpha}, i, j \\ & \langle Q, B_{\boldsymbol{\gamma}} \rangle = 0 \qquad \text{for all other } \boldsymbol{\gamma} \\ & Q \succeq 0 \end{aligned} \tag{16}$$

where the two semidefinite constraints can be restated – if necessary – into one semidefinite constraint

$$\begin{bmatrix} I & -\mathbf{z} & \\ -\mathbf{z}^T & 2\mathbf{y}^T\mathbf{z} - \mathbf{y}^T\mathbf{y} + t & \\ & & Q \end{bmatrix} \succeq 0.$$

Finally, note that the entire program depends on the degree of the SOS-convex polynomial that we started off with: $2d$.

## Further Questions

1. What is the program size? Is it tractable?
2. Can the zero constraints be simplified?
3. For any given $d$, is the program feasible? What is the behavior of the objective $t_d$?
4. How can SDP hierarchy (e.g. by Lasserre) help choosing/removing $d$?

## 3 Convexity Pattern Problem

We now consider a more restricted family of distributions that are hopefully more tractable and also have interesting applications.

With the familiar regression setting as in (1), first consider the additional constraint that $f$ is not only convex but also a function of only a few variables from $\mathbf{x} = (x_1, \ldots, x_p)$. For example, we may have

$$f(x_1, \ldots, x_p) = f(x_1, x_2) \qquad \forall\, \mathbf{x} \in \mathbb{R}^p$$

as one of the possibilities.

In [DCM], Qi, Xu, and Lafferty shows a way to approximate the solution to the above problem *additively*. Specifically, this is

$$\begin{aligned} \underset{f_1,\ldots,f_p}{\text{minimize}} \quad & \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} f_j(x_{ij}))^2 \\ \text{s.t.} \quad & f_1, \ldots, f_p \text{ convex} \end{aligned} \tag{17}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$. In other words, we have the model

$$Y = \sum_{j=1}^{p} f_j(X_j) + \varepsilon$$

in the population, with random variables $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ and $Y \in \mathbb{R}$.

We can view this as a problem of *sparsity patterns*, i.e. whether each variable is "relevant" ($f_j \not\equiv 0$) or not ($f_j \equiv 0$), and it is clear that there are $2^p$ sparsity patterns with $p$ variables.

**Here, we consider an analogous problem of choosing whether each $f_j$ is convex or concave.** Naturally, there are $2^p$ *convexity patterns*. We can write this problem as the following optimization problem:

$$\begin{aligned} \underset{\mathbf{Z},\mathbf{f},\mathbf{g}}{\text{minimize}} \quad & \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} [Z_j f_j(x_{ij}) + (1 - Z_j) g_j(x_{ij})] \right)^2 \\ \text{s.t.} \quad & Z_1, \ldots, Z_p \in \{0, 1\} \\ & f_1, \ldots, f_p \text{ convex} \\ & g_1, \ldots, g_p \text{ concave} \end{aligned} \tag{18}$$

Note that $Z_1, \ldots, Z_p$ are 0/1-boolean variables and $f_1, \ldots, f_p, g_1, \ldots, g_p$ are univariate functions.

In order to make the problem more tractable, we first give extra constraints: namely, that $f_1, \ldots, f_p, g_1, \ldots, g_p$ are *polynomials*. It is important to note that a univariate polynomial is convex if and only if it is SOS-convex. [Problem: Is the set of convex polynomials dense in the set of convex functions? Is this relevant?] We can rewrite the program as follows:

$$
\begin{aligned}
\underset{\mathbf{Z},\mathbf{f},\mathbf{g}}{\text{minimize}} \quad & \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \left[ Z_j f_j(x_{ij}) + (1 - Z_j) g_j(x_{ij}) \right] \right)^2 \\
\text{s.t.} \quad & Z_1, \ldots, Z_p \in \{0, 1\} \\
& f_1, \ldots, f_p \text{ are (SOS-)convex polynomials of degree at most } d \\
& g_1, \ldots, g_p \text{ are (SOS-)concave polynomials of degree at most } d
\end{aligned}
\tag{19}
$$

Using the similar trick as above, we hope to convert the constraints on $f_1, \ldots, f_p, g_1, \ldots, g_p$ into linear or semidefinite ones.

A more important feature of this program is the use of 0-1 variables. It is well-known that, in general, solving a 0-1 integer linear program is NP-hard, and one of the standard procedures in theoretical computer science in dealing with this problem is to relax it such that the boolean constraint is replaced by $Z_1, \ldots, Z_p \in [0, 1]$, or equivalently the quadratic constraint $Z_j^2 - Z_j \leq 0 \ \forall j = 1, \ldots, p$.

With this relaxation comes a family of LP/SDP hierarchies, such as the ones developed by Lovász-Schrijver, Sherali-Adams, and Lasserre. [Prof. Madhur Tulsiani's Survey] These hierarchies are all a sequence of convex programs (LPs or SDPs) whose objective approaches the actual 0-1 solution.

A good way to think about the hierarchies for 0-1 programs is to consider the $Z_j$'s the marginals of a distribution over a set of 0-1 solutions. Specifically, in the initial "round", consider $Z_j$ to be the marginal of the solution whose $j$th entry is 1 and all others are zero. Then, in consecutive rounds, the goal is to add the *joint probabilities* between these variables – in the $r$th round, we consider the joint random variables $Z_S$ for each $S \subseteq \{1, \ldots, p\}$ such that $|S| \leq r$. One can think of these "big variables" as $Z_S = \mathbb{E}\left[\prod_{j \in S} Z_j\right]$, i.e. the probability that all variables in $S$ are 1.

Our hope is to use one of the hierarchies to solve a set of relaxations of (19) that approximates the actual solution efficiently.

# 4 Log-SOS-Concave Density Estimation

## References

[BV] Boyd, S. and Vandenberghe, L. (2009). *Convex Optimization.* Cambridge University Press.

[Lasserre] Lasserre, J. B. (2009). *Moments, Positive Polynomials and Their Applications.* Vol. 1. World Scientific.

[DCM] Qi, Y., Xu, M., and Lafferty, J. (2014). *Learning High-Dimensional Concave Utility Functions for Discrete Choice Models.* NIPS Submission.