

**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:

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

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

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 optimization problem:

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

Denote the vector of basis monomials up to degree  $k$  by  $\mathbf{v}_k(\mathbf{x}) = (1, x_1, \dots, x_p, x_1^2, x_1 x_2, \dots, x_p^k)^T$ , where  $\mathbf{x} = (x_1, \dots, x_p)$ . Then the length of  $\mathbf{v}_k(\mathbf{x})$  is  $\binom{k+p}{p}$ . Let

$$A_k = \left\{ \boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p) \in \mathbb{N}^p \left| \sum_{j=1}^p \alpha_j \leq k \right. \right\}.$$

Then, we may represent  $f$  by a coefficient vector  $\boldsymbol{\theta} \in \mathbb{R}^s$  ( $s = \binom{2d+p}{p}$ ), such that

$$f(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{v}_{2d}(\mathbf{x}) = \sum_{\boldsymbol{\alpha} \in A_{2d}} \theta_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}, \tag{4}$$

where  $\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \dots x_p^{\alpha_p}$ . 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

$$f(\mathbf{x}_i) = \boldsymbol{\theta}^T \mathbf{v}_{2d}(\mathbf{x}_i) = z_i \quad \forall i = 1, \dots, n. \tag{5}$$

We can write this more concisely by introducing the matrix

$$V = V(\mathbf{x}_1, \dots, \mathbf{x}_n) = \begin{bmatrix} \mathbf{v}_{2d}(\mathbf{x}_1)^T \\ \vdots \\ \mathbf{v}_{2d}(\mathbf{x}_n)^T \end{bmatrix}_{n \times s}$$

so that (5) simply becomes

$$V\boldsymbol{\theta} = \mathbf{z}. \quad (6)$$

So we have a linear constraint on the coefficient  $\boldsymbol{\theta}$  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

$$\sum_{i=1}^n (y_i - z_i)^2 = \|\mathbf{y} - \mathbf{z}\|^2 \quad (7)$$

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

Now we want to rewrite the constraint that  $f(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{v}_{2d}(\mathbf{x})$  is SOS-convex. Recall that  $f$  is SOS-convex if and only if the polynomial  $\mathbf{u}^T H_f(\mathbf{x}) \mathbf{u}$  is sos in  $(\mathbf{x}, \mathbf{u}) \in \mathbb{R}^{2p}$ , where  $H_f(\mathbf{x})$  is the Hessian of  $f$ .

For  $i, j \in \{1, \dots, p\}$  we have

$$H_f(\mathbf{x})_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} = \sum_{\boldsymbol{\alpha} \in A_{2d}} \theta_{\boldsymbol{\alpha}} \frac{\partial^2 \mathbf{x}^{\boldsymbol{\alpha}}}{\partial x_i \partial x_j} = \begin{cases} \sum_{\boldsymbol{\alpha} \in A_{2d}} \theta_{\boldsymbol{\alpha}} \alpha_i \alpha_j \mathbf{x}^{\boldsymbol{\beta}_{\boldsymbol{\alpha}, i, j}} & (i \neq j) \\ \sum_{\boldsymbol{\alpha} \in A_{2d}} \theta_{\boldsymbol{\alpha}} \alpha_i (\alpha_i - 1) \mathbf{x}^{\boldsymbol{\beta}_{\boldsymbol{\alpha}, i, i}} & (i = j) \end{cases} = \sum_{\boldsymbol{\alpha} \in A_{2d}} c_{\boldsymbol{\alpha}, i, j} \theta_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\beta}_{\boldsymbol{\alpha}, i, j}} \quad (8)$$

where

$$\boldsymbol{\beta}_{\boldsymbol{\alpha}, i, j} = \begin{cases} (\alpha_1, \dots, \max(\alpha_i - 1, 0), \dots, \max(\alpha_j - 1, 0), \dots, \alpha_p) & i \neq j \\ (\alpha_1, \dots, \max(\alpha_i - 2, 0), \dots, \alpha_p) & i = j \end{cases}$$

and

$$c_{\boldsymbol{\alpha}, i, j} = \begin{cases} \alpha_i \alpha_j & i \neq j \\ \alpha_i (\alpha_i - 1) & i = j. \end{cases}$$

Then we have

$$\mathbf{u}^T H_f(\mathbf{x}) \mathbf{u} = \sum_{i, j=1}^p \left( \sum_{\boldsymbol{\alpha} \in A_{2d}} c_{\boldsymbol{\alpha}, i, j} \theta_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\beta}_{\boldsymbol{\alpha}, i, j}} \right) u_i u_j,$$

which can be further written as

$$\mathbf{u}^T H_f(\mathbf{x}) \mathbf{u} = \sum_{1 \leq i \leq j \leq p} \sum_{\boldsymbol{\beta} \in A_{2d-2}} h_{\boldsymbol{\beta}, i, j}(\boldsymbol{\theta}) \mathbf{x}^{\boldsymbol{\beta}} u_i u_j \quad (9)$$

where

$$h_{\boldsymbol{\beta}, i, j}(\boldsymbol{\theta}) = \begin{cases} (\beta_i + 2)(\beta_i + 1) \theta_{(\beta_1, \dots, \beta_i + 2, \dots, \beta_p)} & i = j \\ 2(\beta_i + 1)(\beta_j + 1) \theta_{(\beta_1, \dots, \beta_i + 1, \dots, \beta_j + 1, \dots, \beta_p)} & i < j. \end{cases}$$

It is then easy to see that  $\mathbf{u}^T H_f(\mathbf{x}) \mathbf{u}$  is SOS if and only if there exists a matrix  $Q$  such that

$$\mathbf{u}^T H_f(\mathbf{x}) \mathbf{u} = \mathbf{v}'_d(\mathbf{x}, \mathbf{u})^T Q \mathbf{v}'_d(\mathbf{x}, \mathbf{u}) \quad (10)$$

$$Q \succeq 0 \quad (11)$$

where  $\mathbf{v}'_d(\mathbf{x}, \mathbf{u})$  is the vector of all monomials in  $(\mathbf{x}, \mathbf{u})$  in which the degrees of all  $x_i$ 's have sum at most  $d-1$  and there is exactly one  $u_i$ , i.e.,

$$\mathbf{v}'_d(\mathbf{x}, \mathbf{u}) = (u_1 \mathbf{v}_{d-1}(\mathbf{x})^T, u_2 \mathbf{v}_{d-1}(\mathbf{x})^T, \dots, u_p \mathbf{v}_{d-1}(\mathbf{x})^T)^T.$$

The length of  $\mathbf{v}'_d(\mathbf{x}, \mathbf{u})$  is  $r = p \binom{p+d-1}{p}$ .  $Q$  is a  $r \times r$  matrix.

(10) 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  $(\mathbf{x}, \mathbf{u})$ . The left-hand side is given by (9). Further, we can express the right-hand side in terms of their coordinates in the following way. First define the coordinate matrix  $B_{\beta, i, j}$  for each  $\beta \in A_{2d-2}, 1 \leq i \leq j \leq p$  such that

$$\mathbf{v}'_d(\mathbf{x}, \mathbf{u}) \mathbf{v}'_d(\mathbf{x}, \mathbf{u})^T = \sum_{1 \leq i \leq j \leq p} \sum_{\beta \in A_{2d-2}} B_{\beta, i, j} \mathbf{x}^\beta u_i u_j.$$

Note that the matrices  $B_{\beta, i, j}$ 's are simply "constants", i.e. they only depend on  $d$  (and  $p$ ). With this, the right-hand side of (10) becomes

$$\begin{aligned} \mathbf{v}'_d(\mathbf{x}, \mathbf{u})^T Q \mathbf{v}'_d(\mathbf{x}, \mathbf{u}) &= \text{tr}(Q \mathbf{v}'_d(\mathbf{x}, \mathbf{u}) \mathbf{v}'_d(\mathbf{x}, \mathbf{u})^T) \\ &= \langle Q, \mathbf{v}'_d(\mathbf{x}, \mathbf{u}) \mathbf{v}'_d(\mathbf{x}, \mathbf{u})^T \rangle \\ &= \left\langle Q, \sum_{1 \leq i \leq j \leq p} \sum_{\beta \in A_{2d-2}} B_{\beta, i, j} \mathbf{x}^\beta u_i u_j \right\rangle \\ &= \sum_{1 \leq i \leq j \leq p} \sum_{\beta \in A_{2d-2}} \langle Q, B_{\beta, i, j} \rangle \mathbf{x}^\beta u_i u_j \end{aligned} \quad (12)$$

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

Then, we can equate the coefficients of (9) and (12) to obtain:

$$\langle Q, B_{\beta, i, j} \rangle = h_{\beta, i, j}(\boldsymbol{\theta}) \quad \forall \beta \in A_{2d-2}, 1 \leq i \leq j \leq p \quad (13)$$

Putting (6), (7), (11), and (13) together, (3) can be restated as the following problem:

$$\begin{aligned} &\underset{\mathbf{z}, \boldsymbol{\theta}, Q}{\text{minimize}} && \|\mathbf{y} - \mathbf{z}\|^2 \\ &\text{s.t.} && V\boldsymbol{\theta} = \mathbf{z} \\ &&& \langle Q, B_{\beta, i, j} \rangle = h_{\beta, i, j}(\boldsymbol{\theta}) \quad \forall \beta \in A_{2d-2}, 1 \leq i \leq j \leq p \\ &&& Q \succeq 0 \end{aligned} \quad (14)$$

(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

$$\begin{aligned} &\underset{t, \mathbf{z}, \boldsymbol{\theta}, Q}{\text{minimize}} && t \\ &\text{s.t.} && \|\mathbf{y} - \mathbf{z}\|^2 \leq t \\ &&& V\boldsymbol{\theta} = \mathbf{z} \\ &&& \langle Q, B_{\beta, i, j} \rangle = h_{\beta, i, j}(\boldsymbol{\theta}) \quad \forall \beta \in A_{2d-2}, 1 \leq i \leq j \leq p \\ &&& Q \succeq 0 \end{aligned} \quad (15)$$

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

**Lemma 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:

$$\begin{aligned} &\underset{t, \mathbf{z}, \boldsymbol{\theta}, Q}{\text{minimize}} && t \\ &\text{s.t.} && \begin{bmatrix} I & -\mathbf{z} \\ -\mathbf{z}^T & 2\mathbf{y}^T \mathbf{z} - \mathbf{y}^T \mathbf{y} + t \end{bmatrix} \succeq 0 \\ &&& V\boldsymbol{\theta} = \mathbf{z} \\ &&& \langle Q, B_{\beta, i, j} \rangle = h_{\beta, i, j}(\boldsymbol{\theta}) \quad \forall \beta \in A_{2d-2}, 1 \leq i \leq j \leq p \\ &&& Q \succeq 0 \end{aligned} \tag{16}$$

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. For any given  $d$ , is the program feasible? What is the behavior of the objective  $t_d$ ?
3. 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  $\mathbf{x} = (x_1, \dots, x_p)$ . For example, we may have

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

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

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

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

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

in the population, with random variables  $X = (X_1, \dots, X_p) \in \mathbb{R}^p$  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:

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

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

In order to make the problem more tractable, we first give extra constraints: namely, that  $f_1, \dots, f_p, g_1, \dots, g_p$  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:

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

Using the similar trick as above, we hope to convert the constraints on  $f_1, \dots, f_p, g_1, \dots, g_p$  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  $Z_1, \dots, Z_p \in [0, 1]$ , or equivalently the quadratic constraint  $Z_j^2 - Z_j \leq 0 \ \forall j = 1, \dots, p$ .

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  $S \subseteq \{1, \dots, p\}$  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

### 4.1 Problem Formulation

Consider the family of log-sos-concave densities on  $K \subseteq \mathbb{R}^p$ :

$$p(\mathbf{x}) \propto \exp(-f(\mathbf{x}))$$

or

$$p(\mathbf{x}) = \frac{\exp(-f(\mathbf{x}))}{\int_K \exp(-f(\mathbf{t})) d\mathbf{t}}.$$

where  $f(\mathbf{x})$  is an sos-convex polynomial. If we restrict the degree of  $f$  to be at most  $2d$ , then we can express  $f$  same as (4):

$$f(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{v}_{2d}(\mathbf{x}) = \sum_{\alpha \in A_{2d}} \theta_{\alpha} \mathbf{x}^{\alpha}.$$

Given  $n$  i.i.d. samples  $\mathbf{x}_1, \dots, \mathbf{x}_n$  from distribution  $p(\mathbf{x}; \boldsymbol{\theta})$ , the likelihood is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n p(\mathbf{x}_i; \boldsymbol{\theta}) = \frac{\exp(-\boldsymbol{\theta}^T \sum_{i=1}^n \mathbf{v}_{2d}(\mathbf{x}_i))}{(\int_K \exp(-\boldsymbol{\theta}^T \mathbf{v}_{2d}(\mathbf{x})) d\mathbf{x})^n},$$

and then

$$-\frac{1}{n} \log L(\boldsymbol{\theta}) = \frac{1}{n} \boldsymbol{\theta}^T \sum_{i=1}^n \mathbf{v}_{2d}(\mathbf{x}_i) + \log \int_K \exp(-\boldsymbol{\theta}^T \mathbf{v}_{2d}(\mathbf{x})) d\mathbf{x}.$$

So the maximum likelihood estimation of  $f$  (or equivalently,  $\boldsymbol{\theta}$ ) can be summerized by the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{\theta}}{\text{minimize}} \quad & \frac{1}{n} \boldsymbol{\theta}^T \sum_{i=1}^n \mathbf{v}_{2d}(\mathbf{x}_i) + \log \int_K \exp(-\boldsymbol{\theta}^T \mathbf{v}_{2d}(\mathbf{x})) d\mathbf{x} \\ \text{s.t.} \quad & \boldsymbol{\theta}^T \mathbf{v}_{2d}(\mathbf{x}) \text{ is sos-convex.} \end{aligned} \tag{20}$$

Denote the above objective function by  $g(\boldsymbol{\theta})$ , which is a convex function. The gradient and Hessian of  $g$  are:

$$\begin{aligned} \nabla g(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^n \mathbf{v}_{2d}(\mathbf{x}_i) + \frac{\int_K \exp(-\boldsymbol{\theta}^T \mathbf{v}_{2d}(\mathbf{x})) (-\mathbf{v}_{2d}(\mathbf{x})) d\mathbf{x}}{\int_K \exp(-\boldsymbol{\theta}^T \mathbf{v}_{2d}(\mathbf{x})) d\mathbf{x}} \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{v}_{2d}(\mathbf{x}_i) - \mathbb{E}_{\boldsymbol{\theta}}(\mathbf{v}_{2d}(\mathbf{X})), \end{aligned} \tag{21}$$

$$\nabla^2 g(\boldsymbol{\theta}) = \mathbb{V}_{\boldsymbol{\theta}}(\mathbf{v}_{2d}(\mathbf{X})), \tag{22}$$

where  $\mathbf{X}$  is a random variable with distribution  $p(\mathbf{x}; \boldsymbol{\theta})$ .



## 4.2 Stochastic Gradient Method (Sketch)

One possible approach to solving (20) is stochastic gradient method, which generates a sequence  $\{\boldsymbol{\theta}_k\}_{k \geq 1}$  through the recursion:

$$\boldsymbol{\theta}_{k+1} \leftarrow P_{\text{sos}}(\boldsymbol{\theta}_k - \alpha_k(\nabla g(\boldsymbol{\theta}_k) + \xi_k)), k = 1, 2, \dots \quad (23)$$

where the initial point  $\boldsymbol{\theta}_1$  is feasible for (20),  $\{\alpha_k\}$  is a positive sequence of stepsizes which may be chosen in different ways,  $\xi_k$  is the (stochastic) error in the gradient evaluation, and  $P_{\text{sos}}(\boldsymbol{\gamma})$  is the projection of  $\boldsymbol{\gamma}$  onto the feasible set of (20), i.e.,

$$\begin{aligned} P_{\text{sos}}(\boldsymbol{\gamma}) : \quad & \underset{\boldsymbol{\theta}}{\text{minimize}} \quad \|\boldsymbol{\theta} - \boldsymbol{\gamma}\|^2 \\ & \text{s.t.} \quad \boldsymbol{\theta}^T \mathbf{v}_{2d}(\mathbf{x}) \text{ is sos-convex.} \end{aligned} \quad (24)$$

Refer to the sos-convex regression problem. We know that (24) is equivalent to

$$\begin{aligned} P_{\text{sos}}(\boldsymbol{\gamma}) : \quad & \underset{\boldsymbol{\theta}, Q}{\text{minimize}} \quad \|\boldsymbol{\theta} - \boldsymbol{\gamma}\|^2 \\ & \text{s.t.} \quad \langle Q, B_{\boldsymbol{\beta}, i, j} \rangle = h_{\boldsymbol{\beta}, i, j}(\boldsymbol{\theta}) \quad \forall \boldsymbol{\beta} \in A_{2d-2}, 1 \leq i \leq j \leq p \\ & \quad Q \succeq 0 \end{aligned} \quad (25)$$

Similar to the transformation from (14) to (16), (25) is equivalent to an SDP:

$$\begin{aligned} P_{\text{sos}}(\boldsymbol{\gamma}) : \quad & \underset{\boldsymbol{\theta}, Q, t}{\text{minimize}} \quad t \\ & \text{s.t.} \quad \begin{bmatrix} I & -\boldsymbol{\theta} \\ -\boldsymbol{\theta}^T & 2\boldsymbol{\gamma}^T \boldsymbol{\theta} - \boldsymbol{\gamma}^T \boldsymbol{\gamma} + t \end{bmatrix} \succeq 0 \\ & \quad \langle Q, B_{\boldsymbol{\beta}, i, j} \rangle = h_{\boldsymbol{\beta}, i, j}(\boldsymbol{\theta}) \quad \forall \boldsymbol{\beta} \in A_{2d-2}, 1 \leq i \leq j \leq p \\ & \quad Q \succeq 0 \end{aligned} \quad (26)$$

## 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.