Estimating High Dimensional Monotone Index Models by Iterative Convex Optimization[1]

Shakeeb Khan, Xiaoying Lan, Elie Tamer

September 2021

Abstract

In this paper we propose a new approach to estimating large dimensional monotone index models. This class of models has been popular in the applied and theoretical econometrics literatures as they include discrete choice, nonparametric transformation, and duration models. The main advantage of our approach is computational: in comparison, rank estimation procedures such as proposed in Han (1987) and Cavanagh and Sherman (1998) optimize a nonsmooth, non convex objective function, and finding a global maximum gets increasingly difficult with a large number of regressors. This makes such procedures particularly unsuitable for "big data" models. For our semiparametric model of increasing dimension, we propose a new algorithm based estimator involving the method of sieves and establish asymptotic its properties. The algorithm uses an iterative procedure where the key step exploits its strictly convex objective function. Our main results here generalize those in, e.g. Dominitz and Sherman (2005) and Toulis and Airoldi (2017), who consider algorithmic based estimators for models of fixed dimension.

Key Words Monotone Index models, Convex Optimization, Sieve Projection Pursuit.

[1]We are grateful to conference participants at the BC/BU 2020 Econometrics Workshop, the 2019 Midwestern Econometrics Study group, the 2021 NASM of the Econometric Society, and seminar participants from UC Berkeley, UC Louvain, University of Bristol and University of Warwick for helpful comments.

1. Introduction

Monotone index models have received a great deal of attention in both the theoretical and applied econometrics literature, as many economic variables of interest are of a limited or qualitative nature. A leading special case in this class is the binary choice model which is usually represented by some variation of the following equation:

$$y_i = I[x_i'\beta_0 - \epsilon_i \geq 0] \tag{1.1}$$

where $I[\cdot]$ is the usual indicator function, y_i is the observed response variable, taking the values 0 or 1 and x_i is an observed vector of covariates which effect the behavior of y_i. Both the disturbance term ϵ_i, and the vector β_0 are unobserved, the latter often being the parameter estimated from a random sample of (y_i, x_i') $i = 1, 2, ...n$.

The disturbance term ϵ_i is restricted in ways that ensure identification of β_0. Parametric restrictions specify the distribution of ϵ_i up to a finite number of parameters and assume it is distributed independently of the covariates x_i. Under such a restriction, β_0 can be estimated (up to scale) using maximum likelihood or nonlinear least squares. However, except in special cases, these estimators are inconsistent if the distribution of ϵ_i is misspecified or conditionally heteroskedastic. Semiparametric, or "distribution free" restrictions have also been imposed in the literature, resulting in a variety of estimation procedures for β_0. The first was the "maximum score" estimator proposed in Manski (1975). Identification of β_0 was based on a conditional median restriction, and based on that Manski (1975), Manski (1984) established the estimator's consistency. Kim and Pollard (1990) established its rate of convergence and limiting distribution, which were $n^{-1/3}$ and non-Gaussian, respectively.

A main drawback of the estimator is its computational difficulty. This arises from the objective function in (**??**) being nonsmooth and nonconvex. This makes finding a global optimum a formidable task. Furthermore the problem becomes more difficult the larger the dimension of x_i.

Alternative semiparametric restrictions used in the literature were based independence/index restrictions. Estimation procedures under this restriction include those proposed by Han (1987), Ichimura (1993), Klein and Spady (1993b). These also have the robustness advantage over parametric approaches, but like maximum score are difficult to compute due to nonconvexity of their respective objective functions, and once again the difficulty increases with the dimension of x_i. Recent work which is motivated by computational concerns is Ahn et al. (2018). However, their two step procedure involves a fully nonparametric estimator in the first stage, so is also not suitable for models with a large number of regressors.

Consequently, a related drawback of all these procedures is that they are designed to estimate parameters in models of a small and *fixed* dimension. A relatively recent and thriving literature in econometrics and machine learning is recognizing the many advantages of allowing for large dimensional models. Such models have a particularly well empirical motivation in binary or discrete choice models. For example, in the decision whether or not to purchase a particular good, explanatory variables would include prices of of other goods which are substitutes or compliments, which could be a large set.

This is a special case of models that consider the situation when the dimension of x_i is large, and this is now often modeled with its dimension increasing with the sample size. Due primarily to its empirical relevance there has been a burgeoning literature on estimation and inference in certain econometric and statistics models with a large number of regressors or a large number of moment conditions. Examples include work in Belloni et al. (2018), Belloni et al. (2014b), Caner (2014), Cattaneo et al. (2018a), Cattaneo et al. (2018b), Chernozhukov et al. (2017), Van de Geer et al. (2014), Han and Phillips (2006), Mammen (1989), Mammen (1993), Newey and Windmeijer (2009), Portnoy (1984), Portnoy (1985).

Particularly related to the work in our paper of estimating large dimensional binary choice or monotone index models are the recent contributions in Sur and Candès (2019), Fan et al. (2020), and Dominitz and Sherman (2005). Sur and Candès (2019) considers inference in a large dimensional logit model, relying on the logistic distribution of the disturbance term. As is the case with all parametric approaches, estimates and inference results are not robust to such a rigid distributional specification.

In contrast, the approach in Fan et al. (2020) is semiparametric, and robust to distributional misspecification. They estimate parameters by optimizing at the objective function introduced in Han (1987), but with the number parameters increasing with the sample size. But unfortunately, such and related estimation procedures cannot be implemented in large dimensional models. This is still the case even with recent developments in algorithms and search methods for optimizing non smooth and/or non convex objective functions. See for example important recent work based on mixed integer programming (MIP) as in, e.g. Fan et al. (2020) and Shin and Todorov (2021).

Also related is the work in Dominitz and Sherman (2005), who consider an algorithmic based estimator for parameters in a class of monotonic index models. Like in our paper the motivation of their approach over existing methods is computational. But they focus on the fixed dimension case, and impose a shape restriction on the disturbance term which restricts the class of models compared to the existing semiparametric literature.

Therefore, in light of the drawbacks in the existing literature, this paper proposes a new estimation procedure to address this omission in this literature. Specifically we aim to construct a computationally feasible estimator for a semiparametric binary choice and monotone index models with *increasing* dimension and establish its asymptotic properties. As we will discuss in detail in the next section, our algorithm uses an iterative estimator based on a stochastic gradient descent method(SGD). and we show how to use the method of sieves (Chen (2007)) to approximate the distribution in each stage of the iteration.

The rest of the paper is organized in follows. In the next section we further discuss the models and parameters we wish to estimate and provide a brief literature review, highlighting important related work in the econometrics, computational and computer science literatures. In doing so we will compare the relative advantages and disadvantages from both theoretical and computational viewpoints. Section 3 then introduces our algorithmic based estimators. Section 4 then explores the asymptotic properties of this procedure, and provides detailed regularity conditions on the sieve space and basis functions, as well as those on the dimension space of the regressors. Section 5 further explores the finite sample properties of the estimator via a simulation study. Section 6 concludes by summarizing and future work, such as discussing other models for which similar algorithm based estimators can be applied to.

2. Model and Related Literature

. Generally speaking, the class of models we will consider estimating are often referred to as monotonic transformation models. One such variant was introduced in Han (1987). We express this model as the equation:

$$y_i = T(x_i'\beta_0, \epsilon_i) \tag{2.1}$$

Where y_i is an observed scalar dependent variable, x_i is an observed vector of covariates of fixed dimension p, and ϵ_i is an unobserved scalar disturbance term. $T(\cdot, \cdot)$ is an unknown transformation function assumed to be monotonic in each of its arguments. β_0 is an unknown p dimensional vector of regression coefficients, often the parameter of interest to identify and estimate from a random sample of (y_i, x_i). The popularity of class of models is that that it nests many special cases that arise in the literature. This includes binary choice models discussed in the previous section but also , censored regression models and duration models with unknown baseline hazard functions. Identification of β_0 is usually based on the assumption that ϵ_i has an unknown distribution that is is independent of x_i. To estimate

β_0, Han (1987) proposed the maximum rank correlation estimator. This involved optimizing the objective function:

$$G_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} I[y_i > y_j] I[x_i'\beta > x_j'\beta] \tag{2.2}$$

He showed the optimizer, subject to a scale normalization was consistent and Sherman (1993) established root-n consistency and asymptotic normality, under standard regularity conditions. Variants of the model and the estimator include work in Abrevaya (1999), Abrevaya (2000), Khan and Tamer (2007), Abrevaya et al. (2010), and more recently Khan et al. (2019) and Fan et al. (2020). While a desirable feature of the original MRC estimator was the generality of the class of modes that could be estimated, a major drawback was its implementability. The objective function is nonsmooth and non concave, making finding a global maximum virtually impossible when the dimension of x_i is sufficiently large. Recent advances in optimization routines such as mixed integer programing, used in, e.g. Fan et al. (2020) and Shin and Todorov (2021) are very valuable, they do not completely solve the problem as it still the case that optimization is np "hard", where n is the sample size and p is the dimension of x_i- see, e.g. Shin and Todorov (2021) for a detailed discussion on this. Other estimation procedures for this model that are not rank based include Cosslett (1983), Ichimura (1993), Ahn et al. (2018). As is the case with the MRC estimator they are not well suited for x_i having a moderately large dimension.

Cosslett (1983) proposes an algorithmic estimator based on MLE and includes include two steps. First he approximates the distribution of the error using basic distribution functions, second he estimates β via MLE and repeats the process until convergence. However, the estimators involves finding the maximum of a non-concave function. This is computationally hard because while one can use grid search to find the maximum, with more than just a few regressors it's almost impossible to implement those methods in practice. As mentioned previously, more modern methods such as MIP (Shin and Todorov (2021)) alleviate but do not completely solve this problem. Ichimura (1993) also involves a non convexity objective function in the iterative NLLS procedure. Ahn et al. (2018) involves two steps each of which is closed form. However, it also cannot be used in large dimensional models due to the fully nonparametric procedure in the first stage.

In this paper to address this omission in the literature, we propose a new iterative estimation procedure that is based on the stochastic gradient descent method(SGD). Furthermore we establish its asymptotic properties, specifically its convergence rate and limiting distribution. One requirement of existing SGD algorithms is that the error distribution is known so we instead use a sieve method to approximate the distribution. As we will explain, first we use their algorithm to estimate β as if the error is logit distributed, second we use a sieve method which we call Series Logit Estimator(SLE) to get the estimation of the distribution of error. Finally we use their algorithm to estimate β again using the estimated distribution and repeat until convergence. As we explain in detail below, we use the gradient method to get the maximum of each iteration, since our SLE is based on logit MLE that is globally convex in the parameters.

Algorithmic based approaches to estimate β in parametric models can be found in the computer science literature. Kalai and Sastry (2009) used monotonic regression. While their method is simple and fast in programming, they do not prove convergence. Agarwal et al. (2013) propose an estimator based on Kalai and Sastry (2009). They proved consistency but the estimator required the underlying distribution function be known.

Our iterative estimator is distinct from, but relates to Agarwal et al. (2013) and SGD, which (unlike ours) requires the knowledge of the error distribution. In their setting the SGD estimator is easy to compute because the algorithm of updating β is linear since the objective function is convex. It is one type of a Newton Raphson procedure and an example of the stochastic approximation method of Robbins and Monro (1951). In related work to that, Toulis and Airoldi (2017) propose implicit SGD estimator and derived its the limiting distribution.

What makes our iterative procedure distinct from all of these is it is not based on the assumption of a known error distribution[2]. Instead, our iterative method uses the method of sieve to estimate the unknown distribution. The method of sieves, proposed in, e.g. Grenander (1981) uses a sequence of finite-dimensional spaces, which is called the sieve space, to approximate an unknown infinite-dimensional space. The complexity of sieve space should increase with the number of observations and the sieves should be dense in the unknown space.

In our algorithm, we will use Series Logit Estimator(SLE), which is also used in Hirano et al. (2003) when they estimate the propensity score function in a treatment effect model. It is a special case of sieve MLE proposed by Geman and Hwang (1982), and they proved the consistency of sieve MLE with i.i.d data. For dependent and heterogeneous data, White (1991) provide a more detailed analysis. Hirano et al. (2003) use logistic model with power series. They only require some smoothness properties of the unknown distribution. Our estimator is similar to their two-step sieve estimator, but is iterative. It starts with modeling the unknown function nonparametrically and then estimates the parametric part with GMM or MLE. Under some regularity conditions, the parametric part of their two-step sieve estimator can get $\sqrt{n}-$asymptotic normality, see Chen (2007), Chen et al. (2003) for more discussion. As for the nonparametric part of sieve estimator, like Chen (2007) pointed out rates of convergence and limiting distribution theory for smooth functionals can be established.

Our estimator can extend to high dimensional cases. By high dimension we mean as the sample size increases to infinity, the number of regressors can also increase to infinity. Fan et al. (2020) propose general rank estimators in high dimensions. They apply the estimator to Han's MRC and obtain consistency if $p_n/n \to 0$ is satisfied, where p_n is the number of covariates and is growing with the number of observations n. Under a the more restrictive

[2]Distribution free algorithmic approaches distinct from what we propose in this paper and based on different assumptions include work by Dominitz and Sherman (2005), Gamarnik and Gaudio (2020), Lanteri et al. (2020).

condition that ${p_n}^2/n \to 0$, they attain asymptotic normality of the estimator. However, for implementation they use the algorithm by Wang (2007), which still suffers from the computational problems when the dimension is large, like many simplex search based algorithms, such as in Nelder and Mead (1965).

Sur and Candès (2019) consider logistic regression in high dimension. They find an area in the parameter space where MLE exists and they also explore what they call the 'average' behavior of the MLE, i.e, the true parameters are centered around a multiple of true parameter and the asymptotic variance of the MLE are also centered. As our estimator involves logit MLE inside the iteration we can apply some useful results from Sur and Candès (2019).

3. Estimation Procedure

In this section we introduce our algorithmic based estimator, establish its asymptotic properties and state the assumptions the theory is based on. For ease of illustration, we will focus on the binary choice model, but the algorithm based estimator and its asymptotic properties we discuss below easily carry over to the general monotone index model.

$$y_i = \mathbb{1}\{x_i^T \beta_0 > \epsilon_i\} \quad i = 1, 2, ...n \tag{3.1}$$

x_i is a p dimensional regressors whose transpose denoted by x_i^T, β_0 is a vector of length of p, $\mathbb{1}$ is an indicator function and ϵ is an unobserved random variable. The distribution of ϵ must satisfy some assumptions to make the estimator consistent. Specifically, we assume that it is J-Lipschitz condition, i.e, $0 \leq g(b) - g(a) \leq J * (b - a)$ for all $a \leq b$, where $g : \mathbb{R} \to \mathbb{R}$ is the CDF of error. Also here we x_i are some continuous random variables. Suppose we have n observations, each observation x_i is a $p * 1$ estimator.

First, we introduce the explicit stochastic gradient descent estimator(SGD), the estimator is a special example of stochastic approximation Robbins and Monro (1951), The following is the SGD algorithm, where we denote iterations by the letter k, $k = 1, 2, ...K(n)$, recalling n denotes the sample size.

Algorithm 1 SGD estimator $g(\cdot)$ known

1: Starting with initial guess $\hat{\beta}_0$, and starting with $k = 1$, set $\hat{\Sigma}_k = C_k$, where C_k is a $p * p$ matrix.
2: Set $\hat{D}_k = (g(x_k^T \hat{\beta}_{k-1}) - y_k) * x_k^T$
3: Update $\hat{\beta}_k = \hat{\beta}_{k-1} - \gamma_k \hat{\Sigma}_k \hat{D}_k$, where γ_k is a "learning parameter", whose properties we discuss below.
4: Go back to Step 1 and set $k = k + 1$.
5: Repeat until you get $\hat{\beta}_K$.

We alter the SGD algorithm to find a minimum value for a convex loss function.

$g(.)$ is a non-decreasing function, then according to Lemma 1, there exists a function $G : \mathbb{R} \to \mathbb{R}$ such that $G' = g$ and G is a convex function.

$$\zeta(\beta; (x, y)) = G(x^T \beta) - yx^T \beta \tag{3.2}$$

The loss function is similar to that proposed by Agarwal et al. (2013). Notice that the loss function is convex in β since G is convex. Now the k_{th} SGD updating for $\hat{\beta}$ becomes:

$$\hat{\beta}_k = \hat{\beta}_{k-1} - \gamma_k C_k \nabla \zeta(\beta_{k-1}; (x_k, y_k)) \tag{3.3}$$

So replacing Step 3 in the above algorithm this way, our SGD estimator at the K^{th} iteration as $\hat{\beta}_K$.

But this algorithm is for the case with known error distribution. In our model since it is unknown, we use the method of sieves to get a feasible semiparametric estimator. The following is the k_{th} sieve SGD **group** updating for β, with $k = 1, 2, ...K$.

$$\tilde{\beta}_k = \tilde{\beta}_{k-1} - \gamma_k C_k \frac{1}{n} \sum_{i=1}^{n} \nabla \tilde{\zeta}_{k-1}(\beta_{k-1}; (x_i, y_i)) \tag{3.4}$$

where $\tilde{\zeta}_{k-1}(\tilde{\beta}_{k-1}; (x_i, y_i))$ is the estimation for $\zeta(\tilde{\beta}_{k-1}; (x_i, y_i))$ using logistic series estimation.

The following details each step of this algorithm :

Algorithm 2 SieveSGD group estimator

1: Denote initial estimate of β_0 and $g(.)$ as $\tilde{\beta}_0$ and $\tilde{g}_0(.)$, and recall ^{T} denote transpose of a vector; calculate $\nabla\tilde{\zeta}_0(\tilde{\beta}_0;(x_i,y_i)) = (\tilde{g}_0(x_i^T\tilde{\beta}_0) - y_i)x_i$ for each i.
2: In first iteration, use group SGD updating in (3.4) to update $\tilde{\beta}_0$ to $\tilde{\beta}_1$
3: Calculate $z_{1i} = \tilde{\beta}_1 * x_i$ for each $i = 1, 2, ...n$
4: Using the full sample of n observations, calculate logistic regression of y_i on index $\tilde{\pi}_0^1 + z_{1i}\tilde{\pi}_1^1 + z_{1i}^2\tilde{\pi}_2^1 + ... + z_{1i}^q\tilde{\pi}_q^1$ to get estimation of error distribution $g(.)$(here q relates to order of sieve approximation)
5: Calculate $\nabla\tilde{\zeta}_1(\tilde{\beta}_1;(x_i,y_i)) = (L(\tilde{\pi}_0^1 + z_{1i}\tilde{\pi}_1^1 + z_{1i}^2\tilde{\pi}_2^1 + ... + z_{1i}^q\tilde{\pi}_q^1) - y_i)x_i$ for each i, where $L(.)$ denotes the CDF of logistic distribution.
6: Go back to 2 to get next iteration and repeat. So in general, in k_{th} iteration, use group SGD updating (3.4) to calculate $\tilde{\beta}_k$.
7: Calculate $z_{ki} = \tilde{\beta}_k * x_i$ for each i, and calculate logistic regression of y_i on index $\tilde{\pi}_0^k + z_{ki}\tilde{\pi}_1^k + z_{ki}^2\tilde{\pi}_2^k + ... + z_{ki}^q\tilde{\pi}_q^k$ to get estimation of error distribution $g(.)$, calculate $\nabla\tilde{\zeta}_k(\tilde{\beta}_k;(x_i,y_i)) = (L(\tilde{\pi}_0^k + z_{ki}\tilde{\pi}_1^k + z_{ki}^2\tilde{\pi}_2^k + ... + z_{ki}^q\tilde{\pi}_q^k) - y_i)x_i$ for each i.
8: Set $k = k + 1$ and repeat step 5 and 6 until you get to K and $\tilde{\beta}_K$.

We denote the SSGD estimator as $\tilde{\beta}_K$.

Finally we introduce a third algorithmic based estimator, also using the method of sieves. Basically this just averages all the K estimates computed in the previous algorithm.

Algorithm 3 SieveSGD average estimator

1: Initially guess β_0 and $g(.)$ as $\tilde{\beta}_0$ and $\tilde{g}_0(.)$, calculate $\nabla\tilde{\zeta}_0(\tilde{\beta}_0;(x_i,y_i)) = (\tilde{g}_0(x_i^T\tilde{\beta}_0) - y_i)x_i$ for each $i = 1, 2, ...n$.
2: In first iteration, use group SGD updating 3.4 to update $\tilde{\beta}_0$ to $\tilde{\beta}_1$
3: Calculate $z_{1i} = \tilde{\beta}_1 * x_i$ for each $i = 1, 2, ..n$
4: Calculate logistic regression of y_i on index $\tilde{\pi}_0^1 + z_{1i}\tilde{\pi}_1^1 + z_{1i}^2\tilde{\pi}_2^1 + ... + z_{1i}^q\tilde{\pi}_q^1$ to get estimation of error distribution $g(.)$(here q is the tuning parameter), calculate $\nabla\tilde{\zeta}_1(\tilde{\beta}_1;(x_i,y_i)) = (L(\tilde{\pi}_0^1 + z_{1i}\tilde{\pi}_1^1 + z_{1i}^2\tilde{\pi}_2^1 + ... + z_{1i}^q\tilde{\pi}_q^1) - y_i)x_i$ for each i.($L(.)$ is CDF of logistic distribution)
5: Go back to 2 to update $\tilde{\beta}_1$ to $\tilde{\beta}_2$ and repeat. In k_{th} iteration, use group SGD updating 3.4 calculate $\tilde{\beta}_k$.
6: Calculate $z_{ki} = \tilde{\beta}_k * x_i$ for each i, calculate logistic regression of y_i on index $\tilde{\pi}_0^k + z_{ki}\tilde{\pi}_1^k + z_{ki}^2\tilde{\pi}_2^k + ... + z_{ki}^q\tilde{\pi}_q^k$ to get updated estimation of error distribution $g(.)$, calculate $\nabla\tilde{\zeta}_k(\tilde{\beta}_k;(x_i,y_i)) = (L(\tilde{\pi}_0^k + z_{ki}\tilde{\pi}_1^k + z_{ki}^2\tilde{\pi}_2^k + ... + z_{ki}^q\tilde{\pi}_q^k) - y_i)x_i$ for each i.
7: Repeat step 5 and 6 until you get $\tilde{\beta}_K$.
8: Lastly, calculate the average of the K estimates, $\tilde{\beta}_k, k = 1, 2, ...K$.
$\bar{\beta}_K = \frac{1}{K-t}\sum_{k=1}^{k=K-t}\tilde{\beta}_k$

We denote this averaged estimator, ASSGD, as $\bar{\beta}_K$.

To establish the validity of our algorithmic based estimators we use the assumptions that are similar to Toulis and Airoldi (2017).

Assumption 3.1. $\{\gamma_k\} = \gamma_1 k^{-\gamma}$, *where* $\gamma_1 > 1$ *is the learning parameter,* $\gamma \in (0.5, 1]$.

Assumption 3.2. *function* $g(.)$ *satisfies J-Lipschitz conditions, i.e,* $0 \leq g(b) - g(a) \leq J * (b - a)$ *and* $g(.)$ *is non-decreasing and differentiable almost surely.*

Assumption 3.3. *The matrix* $\hat{I}_i(\beta) \equiv x_i x_i^T$ *has nonvanishing trace, that is , there exists constant* $b > 0$ *such that trace*$(\hat{I}_i(\beta)) \geq b$ *almost surely, for all* β. *The matrix* $I(\beta_0) = E(\hat{I}_i(\beta_0))$, *has minimum eigenvalue* $\underline{\lambda}_f > 0$ *and maximum eigenvalue* $\overline{\lambda}^f < \infty$. *(These are standard conditions- see, e.g. Lehmann and Casella (2006), Theorem 5.1,page 463).*

Assumption 3.4. C_k *is a fixed positive-definite matrix, such that* $C_k = C + O(\gamma_n)$, *where*$||C|| = 1$, $C \succ 0$ *and symmetric, and* C *commutes with* $I(\beta)$. *Every* C_k *has a greatest eigenvalue* $\overline{\lambda}_c$ *and smallest eigenvalue* $\underline{\lambda}_c$.

Our first theoretical result is for the SGD algorithm, which is for the parametric model as it is based on knowing the error distribution.

Theorem 1. *Under assumptions 3.1-3.4, assume $K = n$, use SGD algorithm 1 we get*

$$\mathbb{E}||\hat{\beta}_K - \beta_0||^2 \leq \frac{8\overline{\lambda}_c^2 \sigma_x^2 (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f1})}{2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f1}} n^{-1} + exp(-log(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f1})\phi(n))[||\hat{\beta}_0 - \beta_0|| + (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f1})^{n_0} A]$$

with n sufficiently large, where $A = 4\overline{\lambda}_c^2 \sum_i \gamma_i^2 < \infty$ and $\phi(n) = n^{1-\gamma}$ if $\gamma \in (0.5, 1]$ and $\phi(n) = \log n$ if $\gamma = 1$. n_0 is some constant.

Remark 1. *Thus the above theorem establishes that estimator based on the first algorithm is consistent and can converge at the parametric rate. While interesting as it can apply to any parametric model, and not just logit or probit to yield a computationally tractable estimator for a wide class of models, it is limited in scope when compared to distribution free estimators discussed earlier in the paper.*

To establish asymptotic properties of our SSGD algorithm based estimator for semiparametric models, we impose the following additional conditions. They are primarily for the sieve component in our algorithm and similar to those in Hirano et al. (2003)

Assumption 3.5. *the support $\mathbf{X}$ of X is a compact subset of $\mathbb{R}^p$.*

Assumption 3.6. *the cdf $g(.)$ is s times continuously differentiable, with $s \geq 4$.*

Assumption 3.7. *the cdf $g(.)$ is bounded away from zero and one on $\mathbf{X}$.*

Assumption 3.8. *the density of X is bounded away from zero on $\mathbf{X}$.*

Assumption 3.9. *$q \to \infty$ as $n \to \infty$ and $q^3/n \to 0$.*

With these assumptions we have the following result for our algorithmic based estimator for the semiparametric binary choice and monotone index models:

Theorem 2. *Under assumptions 3.1-3.4 and 3.6-3.10, assume* $\gamma_0 = 0$*. By setting* $n^{\frac{1}{2\gamma}} \leq K(n) \leq n^{\frac{1}{\gamma}}$*,using sieve SGD group algorithm 2 we get*

$$\mathbb{E}||\tilde{\beta}_{K(n)} - \beta_0||^2 \leq \frac{2(C_1\sqrt{C_2} + 4\overline{\lambda}_c^2\sigma_x^2)(1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2})}{2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2}}(K(n))^{-\gamma}$$
$$+ exp(-log(1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2})\phi(K(n)))[||\beta_0 - \beta_0|| + (1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2})^{n_0}A]$$

with n *sufficiently large, where* $A = (C_1\sqrt{C}_2 + 4\overline{\lambda}_c^2\sigma_x^2)\sum_i \gamma_i^2 < \infty$,$\phi(K(n)) = K(n)^{1-\gamma}$ *if* $\gamma < 1$ *and* $\phi(K(n)) = log(K(n))$ *if* $\gamma = 1$*.* n_0 *is some constant.*

Remark 2. *Thus we can conclude that our algorithmic based estimator for the regression coefficients in the semiparametric models are consistent and can indeed converge at the parametric rate. This is a main advantage of our approach compared to the existing literature, as our algorithm is designed to be implementable with many regressors, in contrast to rank based estimators and closed form estimators which require nonparametric estimation in the first stage. The result shows that this does not come at a cost of a slower rate of convergence.*

The next theorem establishes limiting distribution for the algorithmic estimator for models of fixed dimension.

Theorem 3. *Under assumptions 3.1-3.4 and 3.6-3.10, assume* $\gamma_0 = 0$*. By setting* $K(n) = n$ *and* $\gamma \in (0.5, 1)$*,using sieve SGD average algorithm 3 we get*

$$\sqrt{n}(\bar{\beta}_K - \beta_0) \rightarrow N(0, \Sigma_2^{-1}\Sigma_1\Sigma_2^{-1})$$

where $\Sigma_1 = \mathbb{E}g(x_k^T\beta_0)(1-g(x_k^T\beta_0))x_kx_k^T$ *and* $\Sigma_2 = \mathbb{E}g'(x_k^T\beta_0)x_kx_k^T - f(x_k^T\beta_0)$*, where* $f(x_k^T\beta_0)) = \lim_{q\to\infty} x_kR^q(x_k^T\beta_0)^T\mathbb{E}R^{q\prime}(x_i^T\beta^*)g'(x_i^T\beta_0)x_i^T$ *and* $R^q(x_k^T\beta_0)$ *is orthogonal polynomial function of* $x_k^T\beta_0$*, and* $R^{q\prime}(x_k^T\beta_0)$ *denotes its derivative.*

While the previous result is desirable it is limited in the sense that it is based on models of fixed dimension. This is in contrast to some of the recent literature designed for big data sets which are modeled as the dimension increasing with the sample size. To attain a result for these models, we impose the following additional assumptions on p, the number of regressors, which now depend on n.:

Assumption 3.10. *$var(x^T\beta_0)$ is bounded.*

Assumption 3.11. *$p \to \infty$ as $n \to \infty$ and $p/n \to 0$ where ρ is any positive number.*

Assumption 3.12. *$p \to \infty$ as $n \to \infty$ and $p^2/n \to 0$ where ρ is any positive number.*

With these additional conditions our next result establishes rates of convergence for the algorithmic estimator.

Theorem 4. *Under assumption 3.1-3.4 and 3.6-3.12, using sieve SGD group algorithm 2 and $\gamma_0 = 0$ and by setting $n^{\frac{1}{2\gamma}} \le K(n) \le n^{\frac{1}{\gamma}}$ with $pK(n)^{-\gamma} \to 0$,we get*

$$\begin{aligned}\mathbb{E}||\tilde{\beta}_{K(n)} - \beta_0||^2 \le \; & \frac{2(C_3\sqrt{C_4}C_5 + 4\overline{\lambda}_c^2\sigma_x^2)(1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2})}{2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2}} pK(n)^{-\gamma} \\ & + exp(-log(1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2})\phi(K(n)))[||\tilde{\beta}_0 - \beta_0|| + (1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2})^{n_0}A]\end{aligned}$$

with n sufficiently large, where $A = (C_3\sqrt{C}_4C_5 + 4\overline{\lambda}_c^2\sigma_x^2)\sum_i \gamma_i^2 < \infty$ and $\phi(K(n)) = (K(n))^{1-\gamma}$if $1-\gamma > 0$ and $\phi(K(n)) = log(K(n))$ if $1-\gamma = 0$. $\gamma \in (0.5, 1]$. n_0 is some constant.

Next, we state conditions for the limiting distribution theory of sieve based algorithm estimator

Theorem 5. *Under assumption 3.1-3.4 and 3.6-3.13, by setting $K(n) = n$ and choosing $\gamma \in (0.5, 1)$ and $\frac{p^2}{n^{2\gamma-1}} \to 0 \to 0$,using sieve SGD average algorithm 3, assuming $\gamma_0 = 0$ and x_k are independent across each regressor, for any $\varsigma \in \mathbb{R}^p$ with $||\varsigma|| = 1$ we get $||\bar{\beta}_K - \beta_0|| = o_p(\sqrt{\frac{p}{n}})$, and*

$$\sqrt{n}\frac{\varsigma'(\bar{\beta}_K - \beta_0)}{(\varsigma'\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}\varsigma)^{\frac{1}{2}}} \to N(0, 1)$$

where $\Sigma_1 = \mathbb{E}g(x_k^T\beta_0)(1-g(x_k^T\beta_0))x_kx_k^T$ *and* $\Sigma_2 = \mathbb{E}g'(x_k^T\beta_0)x_kx_k^T - f(x_k^T\beta_0)$*, where* $f(x_k^T\beta_0)) = \lim_{q\to\infty} x_kR^q(x_k^T\beta_0)^T\mathbb{E}R^{q\prime}(x_i^T\beta_0)g'(x_i^T\beta_0)x_i^T$ *and* $R^q(x_k^T\beta_0)$ *is orthogonal polynomial function of* $x_k^T\beta_0$*, and* $R^{q\prime}(x_k^T\beta_0)$ *denotes its derivative.*

4. Simulation Study

In this section we explore the relative finite sample properties of our estimation procedure by presenting the results from a series of Monte Carlo experiments. In the simulation study we focus on the binary choice model:

$$y_i = \mathbb{1}\{x_i^T\beta_0 > \epsilon\}$$

x_i and β_0 is a vector with length 9, the true value of β_0 is $\{1, 1, 2, 4, 5, -1, -2, -4, -5\}$. ϵ follows either standard normal distribution or cauchy distribution with location equivalent to 0 and scale equivalent to 1. We set $q = 2$, which means we use z, z^2 and z^3 to estimate the underlying distribution. The number of observations were 5000 or 10000. We calculate the average time of each experiment, mean bias and root mean square error with 500 experiments.

MRC estimator and MS estimator are not feasible in the binary choice model with more than 3 regressors. We compare our estimator (KLT) with Dominitz and Sherman (2005) (DS), where they use iterative least square with kernel estimation of the distribution of error, which in one sense is similar to ours. One major problem with theirs is that there are 3 tuning parameters in the process and no clear way to choose them in computing the estimator.

Table 1. computation time(second)

	KLT		DS	
Sample size	Normal error	Cauchy error	Normal error	Cauchy error
5000	349.896	201.324	758.784	746.196
10000	642.756	400.62		

We can see from Table 1 that our estimator requires much less time to compute than the estimator of Dominitz and Sherman (2005). For the sample size of 10000, the time of our estimator is around 10 min, which is reasonable and feasible for empirical studies.

Table 2. NORMAL DISTRIBUTION COMPARISON

	KLT				DS	
	N=5000		N=10000		N=5000	
Beta	Bias	RMSE	Bias	RMSE	Bias	RMSE
1	-0.00245	0.074159	-0.00431	0.051545	0.089473	-0.00759
2	-0.00528	0.116748	-0.00543	0.085119	0.128051	-0.00852
4	-0.00383	0.215743	-0.01637	0.156179	0.236556	-0.01471
5	-0.00334	0.264365	-0.02086	0.194141	0.291545	-0.01225
-1	0.001095	0.073209	0.003431	0.051931	0.089551	-0.00076
-2	0.00202	0.119057	0.008036	0.086456	0.128528	0.00513
-4	0.001222	0.214129	0.016845	0.158186	0.236176	0.009738
-5	0.003662	0.263349	0.018584	0.19901	0.289038	0.013762

Table 3. CAUCHY DISTRIBUTION COMPARISON

	KLT				DS	
	N=5000		N=10000		N=5000	
Beta	Bias	RMSE	Bias	RMSE	Bias	RMSE
1	0.009164	0.141747	0.007211	0.102666	0.192022	-0.16365
2	0.011969	0.230573	0.010308	0.167794	0.379245	-0.3429
4	0.028555	0.422073	0.0264	0.289527	0.743516	-0.68575
5	0.046395	0.532154	0.028414	0.359989	0.919298	-0.84691
-1	-0.01341	0.14509	-0.00531	0.103166	0.194638	0.163287
-2	-0.00424	0.228275	-0.01507	0.161979	0.379526	0.345423
-4	-0.03135	0.419305	-0.02808	0.295145	0.735462	0.681659
-5	-0.04541	0.530584	-0.02518	0.371865	0.921918	0.853838

Table 2 and Table 3 are the mean bias and root mean square error (RMSE) of our estimator and their estimator. The mean bias does not decrease with the number of observations may due to the constant complexity of sieve space. The RMSE of our estimator decrease with size. The bias and RMSE of their estimator is high and this may be because it's hard to select the many required tuning parameters with their procedure.

5. Conclusions

In this paper, we proposed a new estimation procedures for binary choice and monotonic index models with increasing dimensions. From an empirical perspective the model can be motivated by models of consumer demand with large consideration sets so prices of many compliments and substitutes are explanatory variables. Existing estimation procedures for this model cannot be implemented in practice when the number of regressors is large. In contrast, our algorithmic based procedure can be used for many regressor models as it involves convex optimization at each iteration of the procedure. We show this iterative procedure also has desirable asymptotic properties when the number of regressors increases with the sample size in ways that are standard in "big data" literature.

Our work here leaves areas for future research. This paper focused on a single equation binary choice model. It would be interesting to see how the proposed algorithmic estimator can be extended to nonbinary and/or systems of simultaneous equation models with a large number of regressors in each equation in each model. For example rank estimators were proposed for the multinomial choice model was proposed in Khan et al. (2019), but was difficult computationally when there were many regressors. We aim to see how our approach in this paper can be adapted to estimate that model and what its asymptotic properties would be. Similarly, Khan and Tamer (2007) propose a rank estimator for duration models with general forms of censoring, that was also difficult computationally for large dimensional models, We conjecture now and aim to show in future work that our approach here is adaptable for that class of models.

Finally, our results here concern high-dimensional models where the number of covariates is at most the same order as the sample size. A recent related literature concerns ultra-high-dimensional models where the number of covariates is much larger than the sample size. In those models some form of (approximate) sparsity is imposed in the model- see, e.g., Belloni et al. (2014a), Belloni et al. (2017). In that setting, inference is conducted after covariate selection, where the resulting number of selected covariates is much smaller. It would be of interest to investigate if such an approach for that type of design can be considered using our method here for large dimensional monotone index models.

References

Abrevaya, J. (1999): "Leapfrog Estimation of a Generalized Fixed-Effects Model with Unknown Transformation of the Dependent Variable," *Journal of Econometrics*, 93, 203–228.

——— (2000): "Rank Estimation of a Generalized Fixed-Effects Regression Model," *Journal of Econometrics*, 95, 1–23.

Abrevaya, J., J. Hausman, and S. Khan (2010): "Testing for Causal Effects in a Generalized Regression Model with Endogenous Regressors," *Econometrica*, 78, 2043–2061.

Agarwal, A., S. M. Kakade, N. Karampatziakis, L. Song, and G. Valiant (2013): "Least squares revisited: Scalable approaches for multi-class prediction," *arXiv preprint arXiv:1310.1949*.

Ahn, H., H. Ichimura, J. L. Powell, and P. A. Ruud (2018): "Simple estimators for invertible index models," *Journal of Business & Economic Statistics*, 36, 1–10.

Ahn, H. and J. Powell (1993): "Semiparametric Estimation of Censored Selection Models," *Journal of Econometrics*, 58, 3–29.

Andersen, E. (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society*, 32, 283–301.

Belloni, A., V. Chernozhukov, D. Chetverikov, and Y. Wei (2018): "Uniformly valid post-regularization confidence regions for many functional parameters in Z-estimation framework." *Annals of Statistics*, 46, 3643–3675.

Belloni, A., V. Chernozhukov, I. Fernandez-Val, and C. Hansen (2017): "Program Evaluation and Causal Inference with High-Dimensional Data," *Econometrica*, 85, 233–298.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650.

Belloni, A., V. Chernozhukov, and K. Kato (2014b): "Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems." *Biometrika*, 102, 77–94.

Cameron, A. C. and P. K. Trivedi (2005): *Microeconometrics: methods and applications*, Cambridge university press.

Caner, M. (2014): "Near exogeneity and weak identification in generalized empirical likelihood estimators: Many moment asymptotics." *Journal of Econometrics*, 182, 247–268.

Cattaneo, M., M. Jansson, and W. Newey (2018a): "Alternative asymptotics and the partially linear model with many regressors." *Econometric Theory*, 34, 277–301.

——— (2018b): "Inference in linear regression models with many covariates and heteroskedasticity." *Journal of the American Statistical Association*, 113, 1350–1361.

Cavanagh, C. and R. P. Sherman (1998): "Rank estimators for monotonic index models," *Journal of Econometrics*, 84, 351–382.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6, 5549–5632.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71, 1591–1608.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017): "Central limit theorems and bootstrap in high dimensions." *Annals of Probability*, 45, 2309–2352.

Cosslett, S. R. (1983): "Distribution-free maximum likelihood estimator of the binary choice model," *Econometrica: Journal of the Econometric Society*, 765–782.

Cox, D. (1972): "Regression Models and Life Tables," *Journal of the Royal Statistical Society Series B*, 34, 187–220.

——— (1975): "Partial Likelihood," *Biometrika*, 623, 269–276.

Dominitz, J. and R. P. Sherman (2005): "Some convergence theory for iterative estimation procedures with an application to semiparametric estimation," *Econometric Theory*, 21, 838–863.

Fan, Y., F. Han, W. Li, and X.-H. Zhou (2020): "On rank estimators in increasing dimensions," *Journal of Econometrics*, 214, 379–412.

Gamarnik, D. and J. Gaudio (2020): "Estimation of Monotone Multi-Index Models," *arXiv preprint arXiv:2006.02806v1*.

Gao, W. and M. Li (2019): "Robust Semiparametric Estimation in Panel Multinomial Choice Models," Working Paper.

Geman, S. and C.-R. Hwang (1982): "Nonparametric maximum likelihood estimation by the method of sieves," *The Annals of Statistics*, 401–414.

Grenander, U. (1981): "Abstract inference," Tech. rep.

Han, A. K. (1987): "Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator," *Journal of Econometrics*, 35, 303–316.

Han, C. and P. Phillips (2006): "GMM with many moment conditions." *Econometrica*, 74, 147–192.

He, X. and Q.-M. Shao (2000): "On parameters of increasing dimensions." *Journal of Multivariate Analysis*, 73, 120–135.

Heckman, J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–694.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

Ichimura, H. (1993): "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58, 71–120.

Kakade, S. M., V. Kanade, O. Shamir, and A. Kalai (2011): "Efficient learning of generalized linear and single index models with isotonic regression," in *Advances in Neural Information Processing Systems*, 927–935.

Kalai, A. T. and R. Sastry (2009): "The Isotron Algorithm: High-Dimensional Isotonic Regression." in *COLT*, Citeseer.

Khan, S., T. Komarova, and D. Nekipelov (2020): "On Optimal Set Estimation for Partially Identified Binary Choice Models," *Boston College working paper*.

Khan, S., F. Ouyang, and E. Tamer (2019): "Inference on Semiparametric Multinomial Response Models," *Quantitative Economics*, forthcoming.

Khan, S. and E. Tamer (2007): "Partial Rank Estimation of Transformation Models with General forms of Censoring," *Journal of Econometrics*, 136, 251–280.

——— (2018): "Discussion of Simple Estimators for Invertible Index Models by H. Ahn, H. Ichimura, J. Powell, and P. Ruud," *Journal of Business & Economic Statistics*, 36, 11–15.

Kim, J. and D. Pollard (1990): "Cube root asymptotics," *The Annals of Statistics*, 18, 191–219.

Klein, R. and R. Spady (1993a): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61, 387–421.

Klein, R. W. and R. H. Spady (1993b): "An efficient semiparametric estimator for binary response models," *Econometrica: Journal of the Econometric Society*, 387–421.

Komarova, T. (2013): "Binary choice models with discrete regressors: Identification and misspecification," *Journal of Econometrics*, 177, 14–33.

Lakkaraju, H., J. Kleinberg, J. Leskovec, J. Ludwig, and S. Mullainathan (2017): "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables," KDD Research Paper.

Lancaster, T. (1979): "Econometric Methods for the Duration of Unemployment," *Econometrica*, 47, 939–956.

Lanteri, A., M. Maggioni, and S. Vigogna (2020): "Conditional regression for single-index models," *arXiv preprint arXiv:2002.10008v2*.

Lee, L.-F. (1995): "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models," *Journal of Econometrics*, 65, 381–428.

Lehmann, E. L. and G. Casella (2006): *Theory of point estimation*, Springer Science & Business Media.

Lei, L., P. Bickel, and N. Karoui (2018): "Asymptotics for high dimensional regression m-estimates: Fixed design results." *Probability Theory Related Fields*, 172, 983–1079.

Lorentz, G. (1986): "Approximation of functions. The second edition ed," *American Mathematical Society, Rhode Island*.

Mammen, E. (1989): "Asymptotics with increasing dimension for robust regression with applications to the bootstrap." *Annals of Statistics*, 17, 382–400.

——— (1993): "Bootstrap and wild bootstrap for high dimensional linear models," *Annals of Statistics*, 21, 255–285.

Manski, C. F. (1975): "Maximum score estimation of the stochastic utility model of choice," *Journal of econometrics*, 3, 205–228.

——— (1984): "Semiparametric Analysis Of Discrete Response: Asymptotic Properties Of The Maximum Score Estimator," University of Wisconsin manuscript.

——— (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55, 357–362.

Nelder, J. and R. Mead (1965): "A Simplex Method for Function Minimization," *The Computer Journal*, 7, 308–313.

Newey, W. and F. Windmeijer (2009): "Generalized method of moments with many weak moment conditions," *Econometrica*, 77, 687–719.

Newey, W. K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica: Journal of the Econometric Society*, 1349–1382.

——— (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of econometrics*, 79, 147–168.

Pagan, A. and A. Ullah (1999): *Nonparametric econometrics*, Cambridge university press.

Portnoy, S. (1984): "Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large: Consistency," *Annals of Statistics*, 12, 1298–1309.

——— (1985): "Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large: Normal Approzimation," *Annals of Statistics*, 13, 1403–1417.

Ravikumar, P., M. Wainwright, and B. Yu (2008): "Single index convex experts: Efficient estimation via adapted bregman losses," in *Snowbird learning workshop*, Citeseer.

Robbins, H. and S. Monro (1951): "A stochastic approximation method," *The annals of mathematical statistics*, 400–407.

Ruud, P. A. (1983): "Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models," *Econometrica: Journal of the Econometric Society*, 225–228.

——— (1986): "Consistent estimation of limited dependent variable models despite misspecification of distribution," *Journal of Econometrics*, 32, 157–187.

Sherman, R. P. (1993): "The limiting distribution of the maximum rank correlation estimator," *Econometrica: Journal of the Econometric Society*, 123–137.

Shi, X., M. Shum, and W. Song (2018): "Estimating Semi-Parametric Panel Multinomial Choice Models using Cyclic Monotonicity," *Econometrica*, 86, 737–761.

Shin, Y. and Z. Todorov (2021): "Exact Computation of the Maximum Rank Correlation Estimator," *Forthcoming, Econometrics Journal.*

Sur, P. and E. J. Candès (2019): "A modern maximum-likelihood theory for high-dimensional logistic regression," *Proceedings of the National Academy of Sciences*, 116, 14516–14525.

Toulis, P. and E. M. Airoldi (2017): "Asymptotic and finite-sample properties of estimators based on stochastic gradients," *The Annals of Statistics*, 45, 1694–1727.

Van de Geer, S., P. Buhlmann, Y. Ritov, and D. R. (2014): "On asymptotically optimal confidence regions and tests for high-dimensional models." *Annals of Statistics*, 42, 1166–1202.

Wang, H. (2007): "A note on iterative marginal optimization: a simple algorithm for maximum rank correlation estimation," *Computational statistics & data analysis*, 51, 2803–2812.

White, H. (1991): "Some Results for Sieve Estimation With Dependent Observations," *Nonparametric and Semiparametric Methods in Econometrics and Statistics.*

A. Appendix

Lemma 1. *Suppose $g : R \to R$ is a non-decreasing function, then there exists a convex function $G : R \to R$ such that $G' = g$.*

Proof. Define $G(x) = \int_d^x g(t)dt$, where d is a constant. Then $G(x)$ is convex since $G'(x) = g(x) \geq 0$. □

Lemma 2. *Suppose X is a $v*1$ vector of random variables $X_1, X_2...X_v$ on product probability space $(\Omega, \mathcal{F}, P)$. P is the product of measures $P_1, P_2...P_v$. The domain of at least one of random variables is $\mathbb{R}$ and the measure of it is continuous. $\mathbb{E}(X^TX)$ is positive definite matrix. $g(.)$ is a non-negative continuous function on $\mathbb{R}$. $\mathbb{E}g(X^T\beta) > 0$ for constant vector β with length v. Then $\mathbb{E}g(X^T\beta)(X^TX)$ is positive definite matrix.*

Proof. We know $\mathbb{E}(X^TX)$ and $\mathbb{E}g(X^T\beta)(X^TX)$ are semi-positive definite matrix. If det$\mathbb{E}(X^TX) = 0$ if and only if there is linear relation between $X_1, X_2...X_v$, then there is no linear relation between $g(X^T\beta)X_1, g(X^T\beta)X_2...g(X^T\beta)X_v$ and we finish the proof.The sufficiency is obvious and we only prove the necessity. There exists a linear relation among columns of $\mathbb{E}(X^TX)$ since $\det\mathbb{E}(X^TX) = 0$. Denote $\mathbb{E}(X^TX)$ as $[A_1, A_2...A_v]$. Suppose $A_1 = a_2 * A_2 + a_3 * A_3 + ... + a_v * A_v$, where $a_1, a_2...a_v$ are constant, and at least one of them is not zero.By changing the second column into $a_2 * A_2 + a_3 * A_3 + ... + a_v * A_v$, we get a new matrix denoted as $[B_1, B_2...B_v]^2$, By changing the second rows into $a_2*B_2+a_3*B_3+...+a_v*B_v$ we get the new matrix, and the first $2*2$ elements are the following:

$$\begin{bmatrix} \mathbb{E}(X_1^2) & \mathbb{E}(X_1(a_2X_2 + a_3X_3 + ... + a_vX_v)) \\ \mathbb{E}(X_1(a_2X_2 + a_3X_3 + ... + a_vX_v)) & \mathbb{E}(a_2X_2 + a_3X_3 + ... + a_vX_v)^2 \end{bmatrix}$$

Then the determinant of the above matrix is 0, then by Hölder's inequality, $X_1 = a_2 * X_2 + a_3 * X_3 + ... + a_v * X_v$. □

Theorem 1. *Under assumptions 3.1-3.4, assume $K = n$ we get*

$$\mathbb{E}||\beta_K - \beta_0||^2 \leq \frac{8\overline{\lambda}_c^2\sigma_x^2(1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f1})}{2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f1}} n^{-1} + exp(-log(1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f1})\phi(n))[||\beta_0 - \beta_0|| + (1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f1})^{n_0}A]$$

with n sufficiently large, where $A = 4\overline{\lambda}_c^2\sum_i \gamma_i^2 < \infty$ and $\phi(n) = n^{1-\gamma}$ if $\gamma \in (0.5, 1]$ and $\phi(n) = logn$ if $\gamma = 1$. n_0 is some constant.

Proof. We start from Eq. (3) and k is the iterative times,

$$\beta_k - \beta_0 = \beta_{k-1} - \beta_0 - \gamma_k C_k \nabla\zeta(\beta_{k-1}; (x_k, y_k))$$

then,

$$\begin{aligned} ||\beta_k - \beta_0||^2 =& ||\beta_{k-1} - \beta_0||^2 \\ & - 2\gamma_k(\beta_{k-1} - \beta_0)^T C_k \nabla\zeta(\beta_{k-1}; (x_k, y_k)) \\ & + \gamma_k^2||C_k\nabla\zeta(\beta_{n-1}; (x_k, y_k))||^2 \end{aligned} \tag{A.1}$$

for the third term,

$$\gamma_k^2||C_k\nabla\zeta(\beta_{k-1};(x_k,y_k))||^2$$
$$\leq 4\gamma_k^2\overline{\lambda}_c^2\sigma_x^2$$

its expectation is bounded as

$$\mathbb{E}(\gamma_k^2||C_k\nabla\zeta(\beta_{k-1};(x_k,y_k))||^2)$$
$$\leq 4\gamma_k^2\overline{\lambda}_c^2\sigma_x^2$$

for the second term,

$$\begin{aligned}
&\mathbb{E}(-2\gamma_k(\beta_{k-1}-\beta_0)^T C_k\nabla\zeta(\beta_{k-1};(x_k,y_k)))\\
&=-2\gamma_k\mathbb{E}((\beta_{k-1}-\beta_0)^T C_k\nabla\zeta(\beta_{k-1};(x_l,y_k)))\\
&=-2\gamma_k\mathbb{E}((\beta_{k-1}-\beta_0)^T C_k\nabla h(\beta_{k-1};(x_k,y_k))) \quad [where \nabla h(\beta_{k-1};(x_k,y_k))=\mathbb{E}(\nabla\zeta(\beta_{k-1};(x_k,y_k)|\mathcal{F}_{k-1})]\\
&=-2\gamma_k\mathbb{E}((\beta_{k-1}-\beta_0)^T C_k(\nabla h(\beta_{k-1};(x_k,y_k))-\nabla h(\beta_0;(x_k,y_k)))\\
&\leq -2\gamma_k\underline{\lambda}_c\underline{\lambda}_{f1}\mathbb{E}||\beta_{k-1}-\beta_0||^2
\end{aligned}$$

Where $\underline{\lambda}_{f1}$ is the least eigenvalue of $\mathbb{E}g(x_k^T\beta_0)x_k^Tx_k$. The last inequality comes from strong convexity by Assumption 3.3 and 2. $\nabla h(\beta_0;(x_k,y_k))=0$ is implied by Eq.3.1

$$\begin{aligned}
&g(x_k^T\beta_0)-\mathbb{E}(y_k|x_k)=0\\
\implies\ &g(x_k^T\beta_0)x_k-\mathbb{E}(y_k|x_k)x_k=0\\
\implies\ &\mathbb{E}(\nabla\zeta(\beta_0;(x_k,y_k))=0\\
\implies\ &\nabla h(\beta_0;(x_k,y_k))=0
\end{aligned}$$

Then we can rewrite Eq. A.1 as

$$\begin{aligned}
\mathbb{E}||\beta_k-\beta_0||^2 \leq &(1-2\gamma_k\underline{\lambda}_c\underline{\lambda}_{f1})\mathbb{E}||\beta_{k-1}-\beta_0||^2+4\gamma_k^2\overline{\lambda}_c^2\sigma_x^2\\
&\frac{1}{(1+2\gamma_k\underline{\lambda}_c\underline{\lambda}_{f1})}\mathbb{E}||\beta_{k-1}-\beta_0||^2+4\gamma_k^2\overline{\lambda}_c^2\sigma_x^2
\end{aligned}$$

By corollary 2.1 in Toulis and Airoldi (2017) with $a_k=4\gamma_k^2\overline{\lambda}_c^2\sigma_x^2$ and $b_k=2\gamma_k\underline{\lambda}_c\underline{\lambda}_{f1}$,and

$K = n$ we get

$$\mathbb{E}||\beta_K - \beta_0||^2 \leq \frac{8\overline{\lambda}_c^2 \sigma_x^2 (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f1})}{2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f1}} n^{-1} + exp(-log(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f1})\phi(n))[||\beta_0 - \beta_0|| + (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f1})^{n_0} A]$$

with n sufficiently large, where $A = 4\overline{\lambda}_c^2 \sum_i \gamma_i^2 < \infty$ and $\phi(n) = n^{1-\gamma}$ if $\gamma \in (0.5, 1]$ and $\phi(n) = logn$ if $\gamma = 1$. n_0 is some constant. □

Theorem 2. *Under assumptions 3.1-3.4 and 3.6-3.10, assume $\gamma_0 = 0$.By setting $n^{\frac{1}{2\gamma}} \leq K(n) \leq n^{\frac{1}{\gamma}}$,using sieve SGD group algorithm 2 we get*

$$\mathbb{E}||\tilde{\beta}_{K(n)} - \beta_0||^2 \leq \frac{2(C_1\sqrt{C_2} + 4\overline{\lambda}_c^2 \sigma_x^2)(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f2})}{2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f2}} (K(n))^{-\gamma}$$
$$+ exp(-log(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f2})\phi(K(n)))[||\beta_0 - \beta_0|| + (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f2})^{n_0} A]$$

with n sufficiently large, where $A = (C_1\sqrt{C}_2 + 4\overline{\lambda}_c^2 \sigma_x^2) \sum_i \gamma_i^2 < \infty$,$\phi(K(n)) = K(n)^{1-\gamma}$ if $\gamma < 1$ and $\phi(K(n)) = log(K(n))$ if $\gamma = 1$. n_0 is some constant.

Proof. the following are notations and definitions from Hirano et al. (2003) with some changes; we use matrix norm $||A|| = \sqrt{tr(A'A)}$. Define

$$L_N(\pi) = \frac{1}{n}\sum_{i=1}^{n}(y_i lnL(R_q^{\hat{\beta}}(x_i)'\pi) + (1 - y_i)lnL(1 - R_q^{\hat{\beta}}(x_i)'\pi))$$

$R_q^{\hat{\beta}}(x_i) \equiv R^q(x_i^T\hat{\beta})$, $R_q^{\hat{\beta}}(x) \equiv R^q(x^T\hat{\beta})$, $R_q^{\beta_0}(x) \equiv R^q(x^T\beta_0)$,$R^q(.)$ is the basis functions in Hirano et al. (2003) with order q. $\mathbb{E}R_q^{\hat{\beta}}(x_i) = 0$ for non-constant term and $\mathbb{E}R_q^{\hat{\beta}}(x_i)'R_q^{\hat{\beta}}(x_i) = 1$. $\iota(q) = sup_{x\in X}||R_q^{\hat{\beta}}(x_i)||$, where $\iota(q) \leq Cq$ for some constant C. $L(.)$ is logistic distribution. $g^*(x) \equiv g(x^T\beta_0)$. $L_N(\pi)$ is the MLE of y_i on $x_i^T\hat{\beta}$. Define

$$\hat{\pi}_q = \underset{\pi}{argmax} L_N(\pi)$$

then, we have

$$||\beta_k - \beta_0||^2 = ||\beta_{k-1} - \beta_0||^2 - 2\gamma_k \frac{1}{n}\sum_{i=1}^{n}(\beta_{k-1} - \beta_0)^T C_k \nabla\hat{\zeta}(\beta_{k-1}; (x_i, y_i)) + \gamma_k^2 \frac{1}{n}\sum_{i=1}^{n}||C_k \nabla\hat{\zeta}(\beta_{k-1}; (x_i, y_i))||^2$$

where $\nabla\hat{\zeta}(\beta_{k-1}; (x_i, y_i) = (L(R_q^{\beta_{k-1}}(x_i)'\hat{\pi}_q) - y_i)x_i$.

for the second term, by maximize $L_N(\pi)$, we get

$$\frac{1}{n}\sum_{i=1}^{n} L(R_q^{\beta_{k-1}}(x_i)'\hat{\pi}_q) - y_i)R_q^{\beta_{k-1}}(x_i) = 0. \tag{A.2}$$

then,

$$\mathbb{E}(L(R_q^{\beta_{k-1}}(x_k)'\hat{\pi}_q) - g(x_k^T\beta_0))R_q^{\beta_{k-1}}(x_k)|\beta_{k-1}, \hat{\pi}_q) = O(\sqrt{1/n}). \tag{A.3}$$

We can approximate $L(R_q^{\beta_{k-1}}(x_k)'\hat{\pi}_q)$ and $g(x_k^T\beta_0)$ with $R_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q$ and $R_q^{\beta_0}(x_k)'\tilde{\pi}_q^*$, according to Lorentz (1986), assuming the second term is increasing[3], then equation becomes

$$\mathbb{E}((R_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q - R_q^{\beta_0}(x_k)'\tilde{\pi}_q^*)R_q^{\beta_{k-1}}(x_k)) = O(\sqrt{1/n}) + O(q^{-s}). \tag{A.4}$$

then we can get $\tilde{\pi}_q^w$

$$\tilde{\pi}_q = \mathbb{E}(R_q^{\beta_{k-1}}(x_k)R_q^{\beta_0}(x_k)')\tilde{\pi}_q^* + O(\sqrt{1/n}) + O(q^{-s}). \tag{A.5}$$

[3]we can relax this to allow some portion of the function is not increasing, but it will not change the result here.

then,

$$
\begin{aligned}
&\mathbb{E}(2\gamma_k \frac{1}{n}\sum_{i=1}^{n}(\beta_{k-1}-\beta_0)^T C_k \nabla\hat{\zeta}(\beta_{k-1};(x_i,y_i)))\\
=&2\gamma_k \underline{\lambda}_c \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(L(R_q^{\beta_{k-1}}(x_i)'\hat{\pi}_q)-y_i)(x_i^T\beta_{k-1}-x_i^T\beta_0)\\
\geq&2\gamma_k \underline{\lambda}_c \mathbb{E}_{\beta_{k-1}}\mathbb{E}((L(R_q^{\beta_{k-1}}(x_k)'\hat{\pi}_q)-g(x_k^T\beta_0))(x_i^T\beta_{k-1}-x_k^T\beta_0)|\beta_{k-1})\\
&-(O(\iota(q)q^{-s}\frac{1}{\sqrt{n}})+O(\frac{\iota(q)^2}{n})+O(\frac{1}{\sqrt{n}}))(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1}-\beta_0||^2)^{\frac{1}{2}}\\
&-O(\frac{1}{\sqrt{n}})\mathbb{E}_{\beta_{k-1}}||\beta_{k-1}-\beta_0||^2\\
\geq&2\gamma_k \underline{\lambda}_c \mathbb{E}_{\beta_{k-1}}\mathbb{E}((R_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q-R_q^{\beta_0}(x_k)'\tilde{\pi}_q^*))(x_i^T\beta_{k-1}-x_k^T\beta_0)|\beta_{k-1})\\
&-\gamma_k(O(q^{-s})+O(\frac{\iota(q)^2}{n})+O(\frac{1}{\sqrt{n}}))(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1}-\beta_0||^2)^{\frac{1}{2}}-O(\frac{1}{\sqrt{n}})\mathbb{E}_{\beta_{k-1}}||\beta_{k-1}-\beta_0||^2\\
\geq&2\gamma_k \underline{\lambda}_c \mathbb{E}_{\beta_{k-1}}\mathbb{E}((\tilde{R}_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q^*-R_q^{\beta_0}(x_k)'\tilde{\pi}_q^*))(x_i^T\beta_{k-1}-x_k^T\beta_0)|\beta_{k-1})\\
&-\gamma_k(O(\sqrt{1/n})+O(q^{2-s})+O(\frac{\iota(q)^2}{n}))(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1}-\beta_0||^2)^{\frac{1}{2}}-O(\iota(q)^2\sqrt{\frac{q}{n}})\mathbb{E}_{\beta_{k-1}}||\beta_{k-1}-\beta_0||^2\\
\geq&2\gamma_k \underline{\lambda}_c \mathbb{E}_{\beta_{k-1}}\mathbb{E}((\tilde{R}_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q^*-R_q^{\beta_0}(x_k)'\tilde{\pi}_q^*))(\tilde{g}^{-1}(\tilde{R}_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q^*)-x_k^T\beta_0)|\beta_{k-1})\\
&-\gamma_k(O(\sqrt{1/n})+O(q^{2-s})+O(\frac{\iota(q)^2}{n}))(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1}-\beta_0||^2)^{\frac{1}{2}}-O(\iota(q)^2\sqrt{\frac{q}{n}})\mathbb{E}_{\beta_{k-1}}||\beta_{k-1}-\beta_0||^2
\end{aligned}
$$

where $\tilde{R}_q^{\beta_{k-1}}(x_k)' \equiv R_q^{\beta_{k-1}}(x_k)'\mathbb{E}(R_q^{\beta_{k-1}}(x_k)R_q^{\beta_0}(x_k)')$ and $\tilde{g}(x_k^T\beta_0) \equiv R_q^{\beta_0}(x_k)'\tilde{\pi}_q^*$.

The first inequality is coming from

$$
\begin{aligned}
&\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(L(R_q^{\beta_{k-1}}(x_i)'\hat{\pi}_q) - y_i)(x_i^T\beta_{k-1} - x_i^T\beta_0)\\
=&\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(L(R_q^{\beta*}(x_i)'\hat{\pi}_q^*) - g(x_k^T\beta_0))(x_i^T\beta_{k-1} - x_i^T\beta_0)\\
&+\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(L(R_q^{\beta_{k-1}}(x_i)'\hat{\pi}_q) - L(R_q^{\beta*}(x_i)'\hat{\pi}_q^*))(x_i^T\beta_{k-1} - x_i^T\beta_0)\\
&-\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(y_i - g(x_k^T\beta_0))(x_i^T\beta_{k-1} - x_i^T\beta_0)\\
=&\mathbb{E}_{\beta_{k-1}}\mathbb{E}((L(R_q^{\beta_0}(x_k)'\hat{\pi}_q^*) - g(x_k^T\beta_0))(x_i^T\beta_{k-1} - x_k^T\beta_0)|\beta_{k-1}) + (O(\iota(q)q^{-s}\frac{1}{\sqrt{n}}) + O(\frac{\iota(q)^2}{n}))(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0\\
&+O(\frac{1}{\sqrt{n}})\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2 + \mathbb{E}_{\beta_{k-1}}\mathbb{E}((L(R_q^{\beta_{k-1}}(x_i)'\hat{\pi}_q) - L(R_q^{\beta_0}(x_k)'\hat{\pi}_q^*))(x_i^T\beta_{k-1} - x_k^T\beta_0)|\beta_{k-1})\\
&+O(\frac{1}{\sqrt{n}})(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2)^{\frac{1}{2}}\\
=&\mathbb{E}_{\beta_{k-1}}\mathbb{E}((L(R_q^{\beta_{k-1}}(x_k)'\hat{\pi}_q) - g(x_k^T\beta_0))(x_i^T\beta_{k-1} - x_k^T\beta_0)|\beta_{k-1})\\
&+(O(\iota(q)q^{-s}\frac{1}{\sqrt{n}}) + O(\frac{\iota(q)^2}{n}) + O(\frac{1}{\sqrt{n}}))(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2)^{\frac{1}{2}}\\
&+O(\frac{1}{\sqrt{n}})\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2
\end{aligned}
$$

where $\hat{\pi}_q^*$ is the value of $\hat{\pi}_q$ when $\beta_{k-1} = \beta_0$ in equation A.2. The proof is similar to the bound on (5) in the addendum of Hirano et al. (2003).

then,

$$\mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)' O(\frac{1}{\sqrt{n}})|\beta_{k-1}) = \mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)' \mathbb{E}(L(R_q^{\beta_{k-1}}(x_k)'\hat{\pi}_q) - g(x_k^T \beta_0)) R_q^{\beta_{k-1}}(x_k)|\beta_{k-1}, \hat{\pi}_q)|\beta_{k-1})$$

$$=\mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)'(\tilde{\pi}_q - \tilde{\pi}_q^*)|\beta_{k-1}) + \mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)'\mathbb{E}(R_q^{\beta_{k-1}}(x_k) O(q^{-s})|\beta_{k-1}, \hat{\pi}_q)|\beta_{k-1})$$

$$+ \mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)'\mathbb{E}(R_q^{\beta_{k-1}}(x_k)(R_q^{\beta_{k-1}}(x_k) - R_q^{\beta_0}(x_k)\tilde{\pi}_q^*)|\beta_{k-1}, \hat{\pi}_q)|\beta_{k-1})$$

$$=\mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1}(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)(y_i - g(x_i^T\beta_0)))|\beta_{k-1})$$

$$+ \mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1}(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)(g(x_i^T\beta_0) - R_q^{\beta_0}(x_i)\tilde{\pi}_q^*))|\beta_{k-1})$$

$$+ \mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1}$$

$$* (\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)((R_q^{\beta_0}(x_i) - R_q^{\beta_{k-1}}(x_i)'\mathbb{E}(R_q^{\beta_{k-1}}(x_k)' R_q^{\beta_0}(x_k))\tilde{\pi}_q^*))|\beta_{k-1})$$

$$+ \{\mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1}$$

$$* (\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)(R_q^{\beta_{k-1}}(x_i)'(\mathbb{E}(R_q^{\beta_{k-1}}(x_k)' R_q^{\beta_0}(x_k) - R_q^{\beta_{k-1}}(x_i)')\tilde{\pi}_q^*))|\beta_{k-1})$$

$$+ \mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)'\mathbb{E}(R_q^{\beta_{k-1}}(x_k)(R_q^{\beta_{k-1}}(x_k) - R_q^{\beta_0}(x_k)\tilde{\pi}_q^*)))|\beta_{k-1}, \hat{\pi}_q)|\beta_{k-1})\}$$

$$+ \mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1}(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i) O(q^{-s}))|\beta_{k-1})$$

$$+ \mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)'\mathbb{E}(R_q^{\beta_{k-1}}(x_k) O(q^{-s})|\beta_{k-1}, \hat{\pi}_q)|\beta_{k-1})$$

By requiring $s \geq 4.5$ and we consider $q = n^d$, $d < 1/5$ and $d > \frac{1}{2(s-2)}$the bound become

$$O(\frac{1}{\sqrt{n}}) + O(\iota(q)^2 \frac{1}{\sqrt{n}})||\beta_{k-1} - \beta_0||$$

see Appendix B for more information.

$O(\sqrt{1/n})$ is invariant to k if $\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2$ is convergent for k and n sufficient large. We will address this issue later.

The last inequality comes from approximate continuous function $\tilde{g}^{-1}(\tilde{R}_q^{\beta_{k-1}}(x_k))$ by A.4. $f(x_k^T \beta^{k-1}) \equiv \mathbb{E}((\tilde{R}_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q^* - R_q^{\beta_0}(x_k)'\tilde{\pi}_q^*))(\tilde{g}^{-1}(\tilde{R}_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q^*) - x_k^T\beta_0)|\beta_{k-1})$. Denote $x_k^T\beta_{k-1}$

as z, we can rewrite $\tilde{R}_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q^* - R_q^{\beta_0}(x_k)'\tilde{\pi}_q^*$ as $\mathbb{E}(R_q^{\beta_0}(x_k)'\tilde{\pi}_q^*|z) - R_q^{\beta_0}(x_k)'\tilde{\pi}_q^*$, so $f(.) \geq 0$ with equality if and only if $\beta_{k-1} = \beta_0$. $f'(x_k^T\beta_0) = 0$ and $f''(x_k^T\beta_0) = \mathbb{E}(\tilde{g}^{-1})'(\tilde{R}_q^{\beta_0}(x_k))(\frac{\partial \tilde{R}_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q^*}{\partial(x_k^T\beta_{k-1})}|_{\beta^{k-1}=\beta_0})^2$ 0 since $\tilde{g}(.)$ is increasing.

$$\frac{\partial \tilde{R}_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q^*}{\partial(x_k^T\beta_{k-1})}|_{\beta^{k-1}=\beta_0} = \frac{\partial R_q^{\beta_0}(x_k)'}{\partial(x_k^T\beta_0)}\tilde{\pi}_q + R_q^{\beta_0}(x_k)\mathbb{E}(\frac{\partial R_q^{\beta_0}(x_k)}{\partial(x_k^T\beta_0)}R_q^{\beta_0}(x_k))\tilde{\pi}_q$$

We know that $\mathbb{E}R_q^{\beta_0}(x_k)\mathbb{E}(\frac{\partial R_q^{\beta_0}(x_k)}{\partial(x_k^T\beta_0)}R_q^{\beta_0}(x_k))\tilde{\pi}_q = 0$ and $R_q^{\beta_0}(x_k)\mathbb{E}(\frac{\partial R_q^{\beta_0}(x_k)}{\partial(x_k^T\beta_0)}R_q^{\beta_0}(x_k))\tilde{\pi}_q$ is continuons in $x_k^T\beta_0$ and $(\tilde{g}^{-1})'(\tilde{R}_q^{\beta_0}(x_k)) = 1/(g'(x_k^T\beta_0) + O(q^{-s}))$, so $f''(x_k^T\beta_0) > 0$. then, by 2 we have

$$\begin{aligned}
&\mathbb{E}(2\gamma_k\frac{1}{n}\sum_{i=1}^{n}(\beta_{k-1}-\beta_0)^T C_k \nabla\hat{\zeta}(\beta_{k-1};(x_i, g(x_i^T\beta_0)))))\\
\geq& 2\gamma_k\underline{\lambda}_c\underline{\lambda}_{f2}\mathbb{E}_{\beta_{k-1}}||\beta_{k-1}-\beta_0||^2\\
&-\gamma_k(\sqrt{1/n}))(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1}-\beta_0||^2)^{\frac{1}{2}} - \gamma_k O(\iota(q)^2\frac{1}{\sqrt{n}})\mathbb{E}_{\beta_{k-1}}||\beta_{k-1}-\beta_0||^2
\end{aligned}$$

where λ_{f2} is the smallest eigenvalue of $\mathbb{E}(\tilde{g}^{-1})'(\tilde{R}_q^{\beta_0}(x_k))(\frac{\partial \tilde{R}_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q^*}{\partial(x_k^T\beta_{k-1})}|_{\beta^{k-1}=\beta_0})^2 x_k^T x_k$.

for the third term,

$$\begin{aligned}
&\gamma_k^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}||C_k\nabla\hat{\zeta}(\beta_{k-1};(x_i, g(x_i^T\beta_0)))||^2\\
\leq& 4\gamma_k^2\overline{\lambda}_c^2\sigma_x^2
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}||\beta_k-\beta_0||^2 \leq& (1-2\gamma_k\underline{\lambda}_c\underline{\lambda}_f + \gamma_k O(\iota(q)^2\frac{1}{\sqrt{n}})\mathbb{E}||\beta_{k-1}-\beta_0||^2\\
&+\gamma_k(O(\sqrt{1/n}))(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1}-\beta_0||^2)^{\frac{1}{2}} + 4\gamma_k^2\overline{\lambda}_c^2\sigma_x^2
\end{aligned}$$

then, if n is sufficiently large, $n \geq n_1$,

$$\begin{aligned}\mathbb{E}||\beta_k - \beta_0||^2 \leq &(1 - 2\gamma_k \underline{\lambda}_c \underline{\lambda}_{f2})\mathbb{E}||\beta_{k-1} - \beta_0||^2 \\ &+ \gamma_k (O(\sqrt{1/n}))(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2)^{\frac{1}{2}} + 4\gamma_k^2 \overline{\lambda}_c^2 \sigma_x^2 \\ \leq &\frac{1}{1 + 2\gamma_k \underline{\lambda}_c \underline{\lambda}_{f2}} \mathbb{E}||\beta_{k-1} - \beta_0||^2 \\ &+ \gamma_k (O(\sqrt{1/n}))(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2)^{\frac{1}{2}} + 4\gamma_k^2 \overline{\lambda}_c^2 \sigma_x^2\end{aligned}$$

Here we can treat $(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2)^{\frac{1}{2}} \leq \mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2 + 1$. Then we can see from the bound for $\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2$ of each iteration that $\mathbb{E}_{\beta_{K(n)-1}}||\beta_{K(n)-1} - \beta_0||^2$ is convergent for $K(n) = n^{\frac{1}{\gamma}}$ and n sufficient large even if $O(\sqrt{1/n})$ is variant to k. So we can choose the supremum of $O(\sqrt{1/n})$ among each iteration. then, assume $\gamma_0 = 0$, there exists a constant C_1 such that $O(\sqrt{1/n}) \leq C_1 n^{\frac{1}{2}}$

$$\mathbb{E}||\beta_k - \beta_0||^2 \leq \frac{1}{1 + 2\gamma_k \underline{\lambda}_c \underline{\lambda}_{f2}} \mathbb{E}||\beta_{k-1} - \beta_0||^2 + \gamma_k C_1 n^{\frac{1}{2}} (\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2)^{\frac{1}{2}} + 4\gamma_k^2 \overline{\lambda}_c^2 \sigma_x^2$$

Since $\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2$ converges, by guess $\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2 \leq C_2 k^{-\gamma}$ we can solve the inequality easily, we are done if the guessing is right in the aggregate inequality. By setting $K(n) = n^{\frac{1}{\gamma}}$ and corollary 2.1 in Toulis and Airoldi (2017) with $a_k = (C_1\sqrt{C_2} + 4\overline{\lambda}_c^2 \sigma_x^2)\gamma_k^2$ and $b_k = 2\gamma_k \underline{\lambda}_c \underline{\lambda}_{f2}$, By setting $n^{\frac{1}{2\gamma}} \leq K(n) \leq n^{\frac{1}{\gamma}}$,using sieve SGD group algorithm 2 we get

$$\begin{aligned}\mathbb{E}||\tilde{\beta}_{K(n)} - \beta_0||^2 \leq &\frac{2(C_1\sqrt{C_2} + 4\overline{\lambda}_c^2 \sigma_x^2)(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f2})}{2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f2}} (K(n))^{-\gamma} \\ &+ exp(-log(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f2})\phi(K(n)))[||\tilde{\beta}_0 - \beta_0|| + (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f2})^{n_0} A]\end{aligned}$$

with n sufficiently large, where $A = (C_1\sqrt{C}_2 + 4\overline{\lambda}_c^2 \sigma_x^2)\sum_i \gamma_i^2 < \infty$,$\phi(K(n)) = K(n)^{1-\gamma}$ if $\gamma < 1$ and $\phi(K(n)) = log(K(n))$ if $\gamma = 1$. n_0 is some constant. We can choose C large enough so that $\frac{2(C_1\sqrt{C_2} + 4\overline{\lambda}_c^2 \sigma_x^2)(1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f2})}{2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f2}} + ||\tilde{\beta}_0 - \beta_0|| + (1 + 2\gamma_1 \underline{\lambda}_c \underline{\lambda}_{f2})^{n_0} A \leq C_2$, then $\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2 \leq C_2 k^{-\gamma}$. □

Theorem 3. *Under assumptions 3.1-3.4 and 3.6-3.10, assume* $\gamma_0 = 0$*. By setting* $K(n) = n$ *and* $\gamma \in (0.5, 1)$*,using sieve SGD average algorithm 3 we get*

$$\sqrt{n}(\bar{\beta}_K - \beta_0) \to N(0, \Sigma_2^{-1}\Sigma_1\Sigma_2^{-1})$$

where $\Sigma_1 = \mathbb{E}g(x_k^T\beta_0)(1 - g(x_k^T\beta_0))((\mathbb{E}x_kx_k^T)^{-\frac{1}{2}}x_k + x_k^T\beta_0 l_{\beta_0})((\mathbb{E}x_kx_k^T)^{-\frac{1}{2}}x_k + x_k^T\beta_0 l_{\beta_0})^T$ *and* $\Sigma_2 = \mathbb{E}g'(x_k^T\beta_0)(I + l_{\beta_0}(\beta_0)')(\mathbb{E}x_kx_k^T)^{-\frac{1}{2}}x_kx_k^T$*.* $l_{\beta_0} = [1/l_1, 1/l_2, ...1/l_p]$*, and* l_i *is the ith element of* $(\mathbb{E}x_kx_k^T)^{\frac{1}{2}}\beta_0$*.*

Proof. W.L.O.G, we set $\mathbb{E}x_kx_k^T = I_p$ and then we calculate the variance without this assumption by using $(\mathbb{E}x_kx_k^T)^{-\frac{1}{2}}x_k$ and $(\mathbb{E}x_kx_k^T)^{\frac{1}{2}}(\bar{\beta}_K - \beta_0)$ to replace x_k and $(\bar{\beta}_K - \beta_0)$ respectively.

First, we write equation 3.4 as $\frac{1}{n}\sum_{i=1}^n \nabla\tilde{\zeta}_{k-1}(\tilde{\beta}_{k-1}; (x_i, y_i)) = \frac{1}{\gamma_k}(\tilde{\beta}_{k-1} - \tilde{\beta}_k)$. By Theorem 2, Taylor expansion on $\frac{1}{n}\sum_{i=1}^n \nabla\tilde{\zeta}_{k-1}(\tilde{\beta}_{k-1}; (x_i, y_i))$ we get $\frac{1}{n}\sum_{i=1}^n \nabla\tilde{\zeta}_{k-1}(\tilde{\beta}_{k-1}; (x_i, y_i)) = \frac{1}{n}\sum_{i=1}^n \nabla\tilde{\zeta}_{k-1}(\beta_0; (x_i, y_i)) + \frac{1}{n}\sum_{i=1}^n \frac{\partial\nabla\tilde{\zeta}_{k-1}(\beta_0;(x_i,y_i))}{\partial\beta}(\tilde{\beta}_{k-1} - \beta_0)$. We know that $\frac{1}{n}\sum_{i=1}^n \nabla\tilde{\zeta}_{k-1}(\beta_0; (x_i, y_i)) - \frac{1}{n}\sum_{i=1}^n \nabla\zeta(\beta_0; (x_i, y_i)) - \frac{1}{n}\sum_i x_i^T\beta_0 l_{\beta_0}(y_i - g(x_i^T\beta_0))$ is negeligible from the similar argument in theorem 2, then if we prove $\frac{1}{n}\sum_{k=1}^n \frac{1}{\gamma_k}(\tilde{\beta}_{k-1} - \tilde{\beta}_k)$ is negligible $o(1/\sqrt{n})$ and

$$\frac{1}{n}\sum_{i=1}^n \frac{\partial\nabla\tilde{\zeta}_{k-1}(\beta_0; (x_i, y_i))}{\partial\beta} \xrightarrow{p} (\frac{1}{n}\sum_{i=1}^n \frac{\partial\nabla\zeta(\beta_0; (x_i, y_i))}{\partial\beta} + \lim_{q\to\infty} \mathbb{E}x_kR_q^{\beta_0}(x_k)'\mathbb{E}(R_q^{\beta_0}(x_k)g'(x_k^T\beta_0)x_k^T))$$

is negligible $o(1/\sqrt{n})$. then $\frac{1}{n}\sum_{k=1}^n(\tilde{\beta}_k - \beta_0)$ behaves like

$$(\frac{1}{n}\sum_{i=1}^n \frac{\partial\nabla\zeta(\beta_0; (x_i, y_i))}{\partial\beta} + \lim_{q\to\infty} \mathbb{E}x_kR_q^{\beta_0}(x_k)'\mathbb{E}(R_q^{\beta_0}(x_k)g'(x_k^T\beta_0)x_k^T))^{-1}$$
$$*(\frac{1}{n}\sum_{i=1}^n \nabla\zeta(\beta_0; (x_i, y_i)) + \frac{1}{n}\sum_i x_i^T\beta_0 l_{\beta_0}(y_i - g(x_i^T\beta_0)))$$
$$\to N(0, \Sigma_{22}^{-1}\Sigma_{11}(\Sigma_{22}^{-1})^T)$$

where $\Sigma_{11} = \mathbb{E}g(x_k^T\beta_0)(1 - g(x_k^T\beta_0))(x_k + x_k^T\beta_0 l_{\beta_0}^t)(x_k + x_k^T\beta_0 l_{\beta_0}^t)^T$ and $\Sigma_{22} = \mathbb{E}g'(x_k^T\beta_0)(I +$

$l^t_{\beta_0}(\beta_0)')x_kx_k^T$ and $l^t_{\beta_0} = [1/\beta_0^{(1)}, 1/\beta_0^{(2)}, , , 1/\beta_0^{(p)}]'$.

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\tilde{\beta}_{k-1}-\tilde{\beta}_k) &\leq \frac{1}{n}(-\frac{1}{\gamma_n}(\tilde{\beta}_n-\beta_0)+\sum_{k=1}^{n-1}|(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}})(\tilde{\beta}_k-\beta_0)|+\frac{1}{\gamma_1}(\tilde{\beta}_0-\beta_0))\\ &\leq \frac{1}{n}(-\frac{1}{\gamma_n}(\tilde{\beta}_n-\beta_0)+C\sum_{k=1}^{n-1}\frac{1}{\sqrt{k}}+\frac{1}{\gamma_1}(\tilde{\beta}_0-\beta_0))\\ &= o(1/\sqrt{n})\end{aligned}$$

This means $\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\tilde{\beta}_{k-1}-\tilde{\beta}_k)$ is negligible.

For Σ_1

$$\begin{aligned}&\lim_{n\to\infty}(\frac{1}{n}\sum_{i=1}^{n}\nabla\zeta(\beta_0;(x_i,y_i))+\frac{1}{n}\sum_i x_i^T\beta_0 l_{\beta_0}(y_i-g(x_i^T\beta_0)))\\ =&\mathbb{E}g(x_k^T\beta_0)(1-g(x_k^T\beta_0))(x_k+x_k^T\beta_0 l_{\beta_0})(x_k+x_k^T\beta_0 l_{\beta_0})^T\end{aligned}$$

For the second term in Σ_2, if we use the similar argument in theorem 2, we know that the second term is negligible.

$$\lim_{q\to\infty}\mathbb{E}x_kR_q^{\beta_0}(x_k)'\mathbb{E}(R_q^{\beta_0}(x_k)g'(x_k^T\beta_0)x_k^T) = \mathbb{E}g'(x_k^T\beta_0)x_k^T\beta_0 l_{\beta_0}x_k^T = \mathbb{E}g'(x_k^T\beta_0)l_{\beta_0}(\beta_0)'x_kx_k^T$$

since $(\mathbb{E}x_kR_q^{\beta_0}(x_k)')R_q^{\beta_0}(x_k) = x_k^T\beta_0 l_{\beta_0}$ by getting fitted value of x_k regressing on $R_q^{\beta_0}(x_k)$.

At last we drop the independent assumption $\mathbb{E}x_kx_k^T = I_p$.Then $N(0,\Sigma_{22}^{-1}\Sigma_{11}(\Sigma_{22}^{-1})^T)$ becomes $N(0,\Sigma_2^{-1}\Sigma_1(\Sigma_2^{-1})^T)$. where $\Sigma_1 = \mathbb{E}g(x_k^T\beta_0)(1-g(x_k^T\beta_0))((\mathbb{E}x_kx_k^T)^{-\frac{1}{2}}x_k+x_k^T\beta_0 l_{\beta_0})((\mathbb{E}x_kx_k^T)^{-\frac{1}{2}}x_k+x_k^T\beta_0 l_{\beta_0})^T$ and $\Sigma_2 = \mathbb{E}g'(x_k^T\beta_0)(I+l_{\beta_0}(\beta_0)')(\mathbb{E}x_kx_k^T)^{-\frac{1}{2}}x_kx_k^T$. $l_{\beta_0} = [1/l_1, 1/l_2, ...1/l_p]$, and l_i is the ith element of $(\mathbb{E}x_kx_k^T)^{\frac{1}{2}}\beta_0$.

□

Theorem 4. *Under assumption 3.1-3.4 and 3.6-3.12, using sieve SGD group algorithm 2 and* $\gamma_0 = 0$ *and by setting* $n^{\frac{1}{2\gamma}} \leq K(n) \leq n^{\frac{1}{\gamma}}$ *with* $pK(n)^{-\gamma} \to 0$*,we get*

$$\mathbb{E}||\tilde{\beta}_{K(n)} - \beta_0||^2 \leq \frac{2(C_3\sqrt{C_4}C_5 + 4\overline{\lambda}_c^2\sigma_x^2)(1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2})}{2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2}} pK(n)^{-\gamma}$$
$$+ exp(-log(1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2})\phi(K(n)))[||\beta_0 - \beta_0|| + (1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2})^{n_0}A]$$

with n *sufficiently large, where* $A = (C_3\sqrt{C}_4C_5 + 4\overline{\lambda}_c^2 p\sigma_x^2)\sum_i \gamma_i^2 = O(p)$ *and* $\phi(K(n)) = (K(n))^{1-\gamma}$*if* $1 - \gamma > 0$ *and* $\phi(K(n)) = log(K(n))$ *if* $1 - \gamma = 0$*.* $\gamma \in (0.5, 1]$*.* n_0 *is some constant.*

Proof. with assumption 3.11, we only have two changes here. The first one is

$$\begin{aligned}
&\mathbb{E}(2\gamma_k \frac{1}{n}\sum_{i=1}^{n}(\beta_{k-1} - \beta_0)^T C_k \nabla\hat{\zeta}(\beta_{k-1}; (x_i, y_i))) \\
=&2\gamma_k\underline{\lambda}_c\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(L(R_q^{\beta_{k-1}}(x_i)'\hat{\pi}_q) - y_i)(x_i^T\beta_{k-1} - x_i^T\beta_0) \\
\geq&2\gamma_k\underline{\lambda}_c\mathbb{E}_{\beta_{k-1}}\mathbb{E}((\tilde{R}_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q^* - R_q^{\beta_0}(x_k)'\tilde{\pi}_q^*))(\tilde{g}^{-1}(\tilde{R}_q^{\beta_{k-1}}(x_k)'\tilde{\pi}_q^*) - x_k^T\beta_0)|\beta_{k-1}) \\
&- \gamma_k(O(\sqrt{p/n}) + O(\sqrt{p}q^{2-s}) + O(\frac{\sqrt{p}\iota(q)^2}{n}))(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2)^{\frac{1}{2}} - O(\iota(q)^2\sqrt{\frac{q}{n}})\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2
\end{aligned}$$

The second one is

$$\begin{aligned}
&\gamma_k^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}||C_k\nabla\hat{\zeta}(\beta_{k-1}; (x_i, g(x_i^T\beta_0)))||^2 \\
\leq&4p\gamma_k^2\overline{\lambda}_c^2\sigma_x^2
\end{aligned}$$

then, if n is sufficiently large, $n \geq n_1$,

$$\begin{aligned}
\mathbb{E}||\beta_k - \beta_0||^2 \leq&(1 - 2\gamma_k\underline{\lambda}_c\underline{\lambda}_{f2})\mathbb{E}||\beta_{k-1} - \beta_0||^2 \\
&+ \gamma_k(O(\sqrt{p/n}))(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2)^{\frac{1}{2}} + 4p\gamma_k^2\overline{\lambda}_c^2\sigma_x^2 \\
\leq&\frac{1}{1 + 2\gamma_k\underline{\lambda}_c\underline{\lambda}_{f2}}\mathbb{E}||\beta_{k-1} - \beta_0||^2 \\
&+ \gamma_k(O(\sqrt{p/n}))(\mathbb{E}_{\beta_{k-1}}||\beta_{k-1} - \beta_0||^2)^{\frac{1}{2}} + 4p\gamma_k^2\overline{\lambda}_c^2\sigma_x^2
\end{aligned}$$

By corollary 2.1 in Toulis and Airoldi (2017) and setting $n^{\frac{1}{2\gamma}} \leq K(n) \leq n^{\frac{1}{\gamma}}$, we get

$$\mathbb{E}||\beta_{K(n)} - \beta_0||^2 \leq \frac{2(C_3\sqrt{C_4}C_5 + 4\overline{\lambda}_c^2\sigma_x^2)(1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2})}{2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2}} pK(n)^{-\gamma}$$
$$+ exp(-log(1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2})\phi(K(n)))[||\tilde{\beta}_0 - \beta_0|| + (1 + 2\gamma_1\underline{\lambda}_c\underline{\lambda}_{f2})^{n_0}A]$$

with n sufficiently large, where $A = (C_3\sqrt{C}_4C_5 + 4\overline{\lambda}_c^2 p\sigma_x^2)\sum_i \gamma_i^2 = O(p)$ and $\phi(K(n)) = (K(n))^{1-\gamma}$if $1 - \gamma > 0$ and $\phi(K(n)) = log(K(n))$ if $1 - \gamma = 0$. $\gamma \in (0.5, 1]$. n_0 is some constant. □

Theorem 5. *Under assumption 3.1-3.4 and 3.6-3.13, by setting $K(n) = n$ and choosing $\gamma \in (0.5, 1)$ and $\frac{p^2}{n^{2\gamma-1}} \to 0 \to 0$,using sieve SGD average algorithm 3, assuming $\gamma_0 = 0$ and x_k are independent across each regressor, for any $\varsigma \in \mathbb{R}^p$ with $||\varsigma|| = 1$ we get $||\bar{\beta}_K - \beta_0|| = o_p(\sqrt{\frac{p}{n}})$, and*

$$\sqrt{n}\frac{\varsigma'(\bar{\beta}_K - \beta_0)}{(\varsigma'\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}\varsigma)^{\frac{1}{2}}} \to N(0,1)$$

where $\Sigma_1 = \mathbb{E}g(x_k^T\beta_0)(1 - g(x_k^T\beta_0))((\mathbb{E}x_kx_k^T)^{-\frac{1}{2}}x_k + x_k^T\beta_0 l_{\beta_0})((\mathbb{E}x_kx_k^T)^{-\frac{1}{2}}x_k + x_k^T\beta_0 l_{\beta_0})^T$ and $\Sigma_2 = \mathbb{E}g'(x_k^T\beta_0)(I + l_{\beta_0}(\beta_0)')(\mathbb{E}x_kx_k^T)^{-\frac{1}{2}}x_kx_k^T$. $l_{\beta_0} = [1/l_1, 1/l_2, ...1/l_p]$, and l_i is the ith element of $(\mathbb{E}x_kx_k^T)^{\frac{1}{2}}\beta_0$.

Proof. There are two differences compared to the proof when p is fixed. The first it the following:

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}\varsigma'(\tilde{\beta}_{k-1} - \tilde{\beta}_k) &\le \frac{1}{n}(-\frac{1}{\gamma_n}\varsigma'(\tilde{\beta}_n - \beta_0) + \sum_{k=1}^{n-1}|(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}})\varsigma'(\tilde{\beta}_k - \beta_0)| + \frac{1}{\gamma_1}\varsigma'(\tilde{\beta}_0 - \beta_0)) \\
&< \frac{1}{n}(-\frac{1}{\gamma_n}(\tilde{\beta}_n - \beta_0) + \sum_{k=1}^{n-1}|(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}||\varsigma'||||\tilde{\beta}_k - \beta_0|| + \frac{1}{\gamma_1}\varsigma'(\tilde{\beta}_0 - \beta_0)) \\
&< \frac{1}{n}(-\frac{1}{\gamma_n}(\tilde{\beta}_n - \beta_0) + \sum_{k=1}^{n-1}|(k - (k-1))|C\sqrt{\frac{p}{k}} + \frac{1}{\gamma_1}\varsigma'(\tilde{\beta}_0 - \beta_0)) \\
&= o(\sqrt{\frac{p}{n}})
\end{aligned}$$

this means $\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k}(\tilde{\beta}_{k-1} - \tilde{\beta}_k)$ is negligible.The second difference is the following:

The second-order term of Taylor expansion of $\nabla\tilde{\zeta}_{k-1}(\beta_0; (x_i, y_i))$ is $\frac{\partial^2\nabla\tilde{\zeta}_{k-1}(\tilde{\beta}_k^*;(x_i,y_i))}{\partial\beta^2}$, where $\tilde{\beta}_k^* = \psi\tilde{\beta}_k + (1-\psi)\beta_0$ and $\psi \in [0,1]$. $\frac{\partial^2\nabla\tilde{\zeta}_{k-1}(\tilde{\beta}_k^*;(x_i,y_i))}{\partial\beta^2}$ is bounded since $\tilde{\beta}_K = \beta_0 + o(1)$ and Σ_{22} has bounded derivatives. Then the second-order term of Taylor expansion of $\frac{1}{n^2}\sum_{k=1}^{n}\sum_{i=1}^{n}\varsigma'\nabla\tilde{\zeta}_{k-1}(\beta_0; (x_i, y_i))$ is bounded by $C\frac{1}{n}\sum_{k=1}^{n}||\mathbb{E}\varsigma' x_k||\frac{p}{k^\gamma} \le C\frac{1}{n}\sum_{k=1}^{n}\frac{p^{\frac{3}{2}}}{k^\gamma}$, which is $o(\sqrt{\frac{p}{n}})$ if $\frac{p^2}{n^{2\gamma-1}} \to 0$.

then $\frac{1}{n}\sum_{k=1}^{n}(\tilde{\beta}_k - \beta_0)$ behaves like

$$(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\nabla\zeta(\beta_0;(x_i,y_i))}{\partial\beta} + \lim_{q\to\infty}\mathbb{E}x_k R_q^{\beta_0}(x_k)'\mathbb{E}(R_q^{\beta_0}(x_k)g'(x_k^T\beta_0)x_k^T))^{-1}$$

$$*(\frac{1}{n}\sum_{i=1}^{n}\nabla\zeta(\beta_0;(x_i,y_i)) + \frac{1}{n}\sum_i x_i^T\beta_0 l_{\beta_0}(y_i - g(x_i^T\beta_0)))$$

then for any $\varsigma \in \mathbb{R}^p$ we get $||\beta_K - \beta_0|| = o_p(\sqrt{\frac{p}{n}})$, and

$$\sqrt{n}\frac{\varsigma'(\beta_K - \beta_0)}{(\varsigma'\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}\varsigma)^{\frac{1}{2}}} \to N(0,1)$$

where $\Sigma_1 = \mathbb{E}g(x_k^T\beta_0)(1-g(x_k^T\beta_0))((\mathbb{E}x_k x_k^T)^{-\frac{1}{2}}x_k + x_k^T\beta_0 l_{\beta_0})((\mathbb{E}x_k x_k^T)^{-\frac{1}{2}}x_k + x_k^T\beta_0 l_{\beta_0})^T$ and $\Sigma_2 = \mathbb{E}g'(x_k^T\beta_0)(I+l_{\beta_0}(\beta_0)')(\mathbb{E}x_k x_k^T)^{-\frac{1}{2}}x_k x_k^T$. $l_{\beta_0} = [1/l_1, 1/l_2, ...1/l_p]$, and l_i is the ith element of $(\mathbb{E}x_k x_k^T)^{\frac{1}{2}}\beta_0$.

□

B. Appendix B

$$\mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)' O(\frac{1}{\sqrt{n}})|\beta_{k-1}) = \mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)' \mathbb{E}(L(R_q^{\beta_{k-1}}(x_k)'\hat{\pi}_q) - g(x_k^T\beta_0)) R_q^{\beta_{k-1}}(x_k)|\beta_{k-1}, \hat{\pi}_q)|\beta_{k-1})$$

$$= \mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)'(\tilde{\pi}_q - \tilde{\pi}_q)|\beta_{k-1}) + \mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)' \mathbb{E}(R_q^{\beta_{k-1}}(x_k) O(q^{-s})|\beta_{k-1}, \hat{\pi}_q)|\beta_{k-1})$$

$$+ \mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)' \mathbb{E}(R_q^{\beta_{k-1}}(x_k)(R_q^{\beta_{k-1}}(x_k) - R_q^{\beta_0}(x_k)\tilde{\pi}_q)|\beta_{k-1}, \hat{\pi}_q)|\beta_{k-1})$$

$$= \mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1}(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)(y_i - g(x_i^T\beta_0)))|\beta_{k-1}) \tag{B.1}$$

$$+ \mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1}(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)(g(x_i^T\beta_0) - R_q^{\beta_0}(x_i)\tilde{\pi}_q))|\beta_{k-1}) \tag{B.2}$$

$$+ \mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1}$$

$$* (\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)((R_q^{\beta_0}(x_i) - R_q^{\beta_{k-1}}(x_i)' \mathbb{E}(R_q^{\beta_{k-1}}(x_k)' R_q^{\beta_0}(x_k))\tilde{\pi}_q))|\beta_{k-1}) \tag{B.3}$$

$$+ \{\mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1}$$

$$* (\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)(R_q^{\beta_{k-1}}(x_i)' \mathbb{E}(R_q^{\beta_{k-1}}(x_k)' R_q^{\beta_0}(x_k) - R_q^{\beta_{k-1}}(x_i)')\tilde{\pi}_q))|\beta_{k-1})$$

$$+ \mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)'(\mathbb{E}(R_q^{\beta_{k-1}}(x_k)(R_q^{\beta_{k-1}}(x_k) - R_q^{\beta_0}(x_k))\tilde{\pi}_q)|\beta_{k-1}, \hat{\pi}_q)|\beta_{k-1})\} \tag{B.4}$$

$$+ \mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1}(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i) O(q^{-s}))|\beta_{k-1}) \tag{B.5}$$

$$+ \mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)' \mathbb{E}(R_q^{\beta_{k-1}}(x_k) O(q^{-s})|\beta_{k-1}, \hat{\pi}_q)|\beta_{k-1}) \tag{B.6}$$

For B.1

$$
\begin{aligned}
&\mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1}(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)(y_i - g(x_i^T\beta_0)))|\beta_{k-1}) \\
&\le \mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)(y_i - g(x_i^T\beta_0)))|\beta_{k-1}) + O(\iota(q)^3 q^{-s}\frac{\sqrt{q}}{\sqrt{n}}) \\
&\le \frac{1}{n}\sum_i x_i^T \beta_{k-1} l_{\beta_{k-1}}(y_i - g(x_i^T\beta_0)) + O(\iota(q)^3 q^{-s}\frac{\sqrt{q}}{\sqrt{n}}) \\
&\le O(\frac{1}{\sqrt{n}}) + O(\iota(q)^3 q^{-s}\frac{\sqrt{q}}{\sqrt{n}})
\end{aligned}
$$

where we use $||\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1} - I_q|| = O_p(\iota(q)\sqrt{\frac{q}{n}})$ by Newey (1997). The next-to-last equation needs independence assumption of x_i across each regressor. Even without the independence assumption we still can get the last equation.

For B.2

$$
\begin{aligned}
&\mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1}(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)(g(x_i^T\beta_0) - R_q^{\beta_0}(x_i)\tilde{\pi}_q))|\beta_{k-1}) \\
=&O(\iota(q)^2 q^{-s})(1 + O(\iota(q)\sqrt{\frac{q}{n}}))
\end{aligned}
$$

For B.3

$$
\begin{aligned}
&\mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1} \\
&* (\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)((R_q^{\beta_0}(x_i) - R_q^{\beta_{k-1}}(x_i)'\mathbb{E}(R_q^{\beta_{k-1}}(x_k)' R_q^{\beta_0}(x_k))\tilde{\pi}_q))|\beta_{k-1}) \\
=&O(\iota(q)^2\frac{1}{\sqrt{n}})(1 + O(\iota(q)\sqrt{\frac{q}{n}}))||\beta_{k-1} - \beta_0||
\end{aligned}
$$

Here $\mathbb{E}R_q^{\beta_{k-1}}(x_i)((R_q^{\beta_0}(x_i) - R_q^{\beta_{k-1}}(x_i)'\mathbb{E}(R_q^{\beta_{k-1}}(x_k)' R_q^{\beta_0}(x_k))\tilde{\pi}_q) = 0$ because by regressing $R_q^{\beta_0}(x_i))'\tilde{\pi}_q$ on $R_q^{\beta_0}(x_i)$ we get the residual $((R_q^{\beta_0}(x_i) - R_q^{\beta_{k-1}}(x_i)'\mathbb{E}(R_q^{\beta_{k-1}}(x_k)' R_q^{\beta_0}(x_k))\tilde{\pi}_q)$, which is orthogonal to $R_q^{\beta_{k-1}}(x_i)$.

For B.4

$$
\begin{aligned}
&\{\mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1} \\
&* (\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)(R_q^{\beta_{k-1}}(x_i)'(\mathbb{E}(R_q^{\beta_{k-1}}(x_k)' R_q^{\beta_0}(x_k) - R_q^{\beta_{k-1}}(x_i)'))\tilde{\pi}_q))|\beta_{k-1}) \\
&+ \mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)'\mathbb{E}(R_q^{\beta_{k-1}}(x_k)(R_q^{\beta_{k-1}}(x_k) - R_q^{\beta_0}(x_k)\tilde{\pi}_q)|\beta_{k-1}, \hat{\pi}_q)|\beta_{k-1})\}|\beta_{k-1}) \\
=&0
\end{aligned}
$$

For B.5

$$
\begin{aligned}
&\mathbb{E}((x_k^T R_q^{\beta_{k-1}}(x_k)')(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i)' R_q^{\beta_{k-1}}(x_i))^{-1}(\frac{1}{n}\sum_i R_q^{\beta_{k-1}}(x_i) O(q^{-s}))|\beta_{k-1}) \\
=&O(\iota(q)^2 q^{-s})(1 + O(\iota(q)\sqrt{\frac{q}{n}}))
\end{aligned}
$$

For B.6

$$
\begin{aligned}
&\mathbb{E}(x_k^T R_q^{\beta_{k-1}}(x_k)'\mathbb{E}(R_q^{\beta_{k-1}}(x_k) O(q^{-s})|\beta_{k-1}, \hat{\pi}_q)|\beta_{k-1}) \\
=&O(\iota(q)^2 q^{-s})
\end{aligned}
$$

Adding the bound together, we get

$$
\begin{aligned}
&O(\frac{1}{\sqrt{n}}) + O(\iota(q)^3 q^{-s}\frac{\sqrt{q}}{\sqrt{n}}) + O(\iota(q)^2 q^{-s})(1 + O(\iota(q)\sqrt{\frac{q}{n}})) \\
+&O(\iota(q)^2\frac{1}{\sqrt{n}})(1 + O(\iota(q)\sqrt{\frac{q}{n}}))||\beta_{k-1} - \beta_0|| + 0 \\
+&O(\iota(q)^2 q^{-s})(1 + O(\iota(q)\sqrt{\frac{q}{n}})) + O(\iota(q)^2 q^{-s}) \\
\leq&O(\iota(q)^2 q^{-s}) + O(\frac{1}{\sqrt{n}}) + O(\iota(q)^2\frac{\sqrt{q}}{n}) + O(\iota(q)^2\frac{1}{\sqrt{n}})||\beta_{k-1} - \beta_0||
\end{aligned}
$$

By requiring $s \geq 4.5$ and we consider $q = n^d$ and $d < 1/5$ the bound become

$$
O(\frac{1}{\sqrt{n}}) + O(\iota(q)^2\frac{1}{\sqrt{n}})||\beta_{k-1} - \beta_0||
$$