Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Semiparametric Efficiency in Convexity Constrained Single-Index Model

Arun K. Kuchibhotla, Rohit K. Patra & Bodhisattva Sen

To cite this article: Arun K. Kuchibhotla, Rohit K. Patra & Bodhisattva Sen (2023) Semiparametric Efficiency in Convexity Constrained Single-Index Model, Journal of the American Statistical Association, 118:541, 272-286, DOI: 10.1080/01621459.2021.1927741

To link to this article: https://doi.org/10.1080/01621459.2021.1927741

View supplementary material

Published online: 26 Jul 2021.

Submit your article to this journal

Article views: 573

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 118, NO. 541, 272–286: Theory and Methods
https://doi.org/10.1080/01621459.2021.1927741

Semiparametric Efficiency in Convexity Constrained Single-Index Model

Arun K. Kuchibhotla[a], Rohit K. Patra[b], and Bodhisattva Sen[c]

[a]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA; [b]Department of Statistics, University of Florida, Gainesville, FL; [c]Department of Statistics, Columbia University, New York, NY

ABSTRACT
We consider estimation and inference in a single-index regression model with an unknown convex link function. We introduce a convex and Lipschitz constrained least-square estimator (CLSE) for both the parametric and the nonparametric components given independent and identically distributed observations. We prove the consistency and find the rates of convergence of the CLSE when the errors are assumed to have only $q \geq 2$ moments and are allowed to depend on the covariates. When $q \geq 5$, we establish $n^{-1/2}$-rate of convergence and asymptotic normality of the estimator of the parametric component. Moreover, the CLSE is proved to be semiparametrically efficient if the errors happen to be homoscedastic. We develop and implement a numerically stable and computationally fast algorithm to compute our proposed estimator in the R package `simest`. We illustrate our methodology through extensive simulations and data analysis. Finally, our proof of efficiency is geometric and provides a general framework that can be used to prove efficiency of estimators in a wide variety of semiparametric models even when they do not satisfy the efficient score equation directly. Supplementary files for this article are available online.

ARTICLE HISTORY
Received October 2019
Accepted January 2021

KEYWORDS
Bundled parameter; Errors with finite moments; Geometric proof of semiparametric efficiency; Lipschitz constrained least squares; Shape restricted function estimation

1. Introduction

Suppose we have n iid observations $\{(X_i, Y_i) \in \chi \times \mathbb{R}, 1 \leq i \leq n\}$ from the following single-index regression model:

$$Y = m_0(\theta_0^\top X) + \epsilon, \tag{1}$$

where $X \in \chi \subset \mathbb{R}^d$ $(d \geq 1)$ is the predictor, $Y \in \mathbb{R}$ is the response variable, and ϵ satisfies $\mathbb{E}(\epsilon|X) = 0$ and $\mathbb{E}(\epsilon^2|X) < \infty$ almost everywhere (a.e.) P_X, the distribution of X. We assume that the real-valued link function m_0 and $\theta_0 \in \mathbb{R}^d$ are the unknown parameters of interest.

Single-index models are ubiquitous in regression because they provide convenient dimension reduction and interpretability. The single-index model circumvents the curse of dimensionality encountered in estimating the fully nonparametric regression function $\mathbb{E}(Y|X = \cdot)$ by assuming that the link function depends on X only through a one-dimensional projection, that is, $\theta_0^\top X$; see, for example, Powell, Stock, and Stoker (1989). Moreover, the coefficient vector θ_0 provides interpretability Li and Racine (2007) and the one-dimensional nonparametric link function m_0 offers some flexibility in modeling. The above model has received a lot of attention in statistics in the last few decades; see, for example, Powell, Stock, and Stoker (1989), Li and Duan (1989), Ichimura (1993), Härdle, Hall, and Ichimura (1993), Hristache, Juditsky, and Spokoiny (2001), Delecroix, Hristache, and Patilea (2006), Cui, Härdle, and Zhu (2011), Kuchibhotla and Patra (2020), and the references therein. The above articles propose estimators for the single-index model under the assumption that m_0 is smooth (i.e., two or three times differentiable).

However, quite often in the context of a real application, qualitative assumptions on m_0 may be available. For example, in microeconomics, production and utility functions are often assumed to be concave and nondecreasing; concavity indicates decreasing marginal returns/utility (Varian 1984; Matzkin 1991; Li and Racine 2007). In finance, the relationship between call option prices and strike price is often known to be convex and decreasing (Aït-Sahalia and Duarte 2003); in stochastic control, value functions are often assumed to be convex (Keshavarz, Wang, and Boyd 2011). The following two real-data examples further illustrate that convexity/concavity constraints arise naturally in many applications.

Example 1.1 (Boston housing data). Harrison and Rubinfeld (1978) studied the effect of different covariates on real estate price in the greater Boston area. The response variable Y was the log-median value of homes in each of the 506 census tracts in the Boston standard metropolitan area. A single-index model is appropriate for this dataset; see, for example, Gu and Yang (2015), Wang and Yang (2009), Wang et al. (2010), and Yu, Mammen, and Park (2011). The above articles considered the following covariates in their analysis: average number of rooms per dwelling, full-value property-tax rate per 10000 U.S.D., pupil–teacher ratio by town school district, and proportion of population that is of "lower (economic) status" in percentage points. In the left panel of Figure 1, we provide the scatterplot of $\{(Y_i, \hat{\theta}^\top X_i)\}_{i=1}^{506}$, where $\hat{\theta}$ is the estimate of θ_0 obtained in Wang and Yang (2009). We also plot estimates of m_0 obtained from Kuchibhotla and Patra (2020) and Wang and Yang (2009). The plot suggests a convex and nondecreasing

CONTACT Rohit K. Patra ✉ *rohitpatra@ufl.edu* Department of Statistics, University of Florida, Gainesville, FL 32611-7011.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

© 2021 American Statistical Association

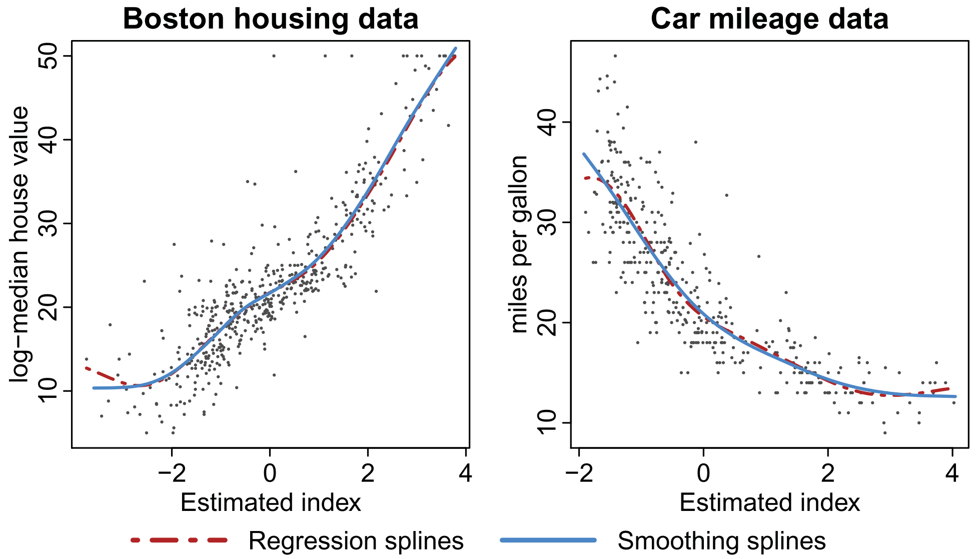


Figure 1. Scatterplots of $\{(X_i^\top\hat{\theta}, Y_i)\}_{i=1}^n$, where $\hat{\theta}$ is the estimator of θ_0 proposed in Wang and Yang (2009). NOTE: The plot is overlaid with the smoothing and regression spline-based function estimators of m_0 proposed in Kuchibhotla and Patra (2020) and Wang and Yang (2009), respectively. Left panel: Boston housing data (see Section 6.1); right panel: the car mileage data (see Section 6.2).

relationship between the log-median home prices and the index, but the fitted link functions satisfy these shape constraints only approximately.

Example 1.2 (Car mileage data). Donoho and Ramos (1983) considered a dataset containing mileages of different cars. The data contains mileages of 392 cars as well as the following covariates: displacement, weight, acceleration, and horsepower. Cheng, Zhao, and Li (2012) and Kuchibhotla and Patra (2020) had fit a partial linear model and a single-index model, respectively. In the right panel of Figure 1, we plot the estimators proposed in Kuchibhotla and Patra (2020) and Wang and Yang (2009). Both of these works consider estimation in the single-index model under only smoothness assumptions. The “law of diminishing returns” suggests m_0 should be convex and nonincreasing. However, as observed in Figure 1, the estimators based only on smoothness assumptions satisfy this shape constraint only approximately.

In both of the examples, the smoothing-based estimators do not incorporate the known shape of the nonparametric function. Thus, the estimators are not guaranteed to be convex (or monotone) in finite samples. Moreover, the choice of the tuning parameter in smoothness-based estimators is tricky as different values for the tuning parameter lead to very different shapes. This unpredictable behavior makes the smoothness-based estimators of m_0 less interpretable, and motivates the study of a convexity constrained single-index model. We discuss these two datasets and our analysis in more detail in Sections 6.1 and 6.2.

In this article, we propose constrained least-square estimators for m_0 and θ_0 that is guaranteed to satisfy the inherent convexity constraint in the link function everywhere. The proposed methodology is appealing for two main reasons: (i) the estimator is interpretable and takes advantage of naturally occurring qualitative constraints; and (ii) unlike smoothness-based estimators, the proposed estimator is highly robust to the choice of the tuning parameter without sacrificing efficiency.

In the following, we conduct a systematic study of the computation, consistency, and rates of convergence of the estimators, under mild assumptions on the covariate and error distributions. We further prove that the estimator for the finite-dimensional parameter θ_0 is asymptotically normal. Moreover, this estimator is shown to be semiparametrically efficient if the errors happen to be homoscedastic, that is, when $\mathbb{E}(\epsilon^2|X) \equiv \sigma^2$ a.e. for some constant σ^2. It should be noted that in the examples above the link function is also known to be monotone. To keep things simple, we focus on only convexity constrained single-index model. However, *all* our results continue to hold under the additional monotonicity assumption, that is, our conclusions hold for convex/concave and nondecreasing/nonincreasing m_0. More generally, our results continue to hold under *any* additional shape constraints; see Remarks 3.6, 4.3, and S.1.1 and Section 6 in the article for more details.

One of the main contributions of this article is our novel geometric proof of the semiparametric efficiency of the constrained least-square estimator. Note that proving semiparametric efficiency of constrained (and/or penalized) least-square estimators often requires a delicate use of the structure of the estimator of the nonparametric component (say $\hat{m}$) to construct *least favorable paths*; see, for example, Murphy, van der Vaart, and Wellner (1999), (Bolthausen, Perkins, and van der Vaart 2002, chap. 9.3), and Huang (1996) (also see Example 4.1). In contrast, our approach is based on the following simple observation. For a traditional smoothness-based estimator $\hat{m}$, the path $t \mapsto \hat{m} + ta$ will belong to the (function) parameter space for *any* smooth “perturbation” a (for small enough $t \in (-1, 1)$). However, this is no longer true when the underlying parameter space is constrained. But, observe that the projection of $\hat{m} + ta$ onto the constrained function space certainly yields a “valid" path. Our proof technique is based on differentiability properties of the path $t \mapsto \Pi(\hat{m} + ta)$, where Π denotes the L_2-projection onto the (constrained) function space. This general principle is applicable to other shape constrained semiparametric models, because differentiability of the projection operator is well-studied in the context of constrained optimization algorithms; see Section 1.1 for a more detailed discussion. Also see Example 4.1, where we discuss the applicability of our technique in (re)proving the semiparametric efficiency of the nonparametric maximum likelihood estimator in the Cox proportional hazard model under current status censoring Huang (1996). To be more specific, we study the following Lipschitz constrained convex least-square estimator (CLSE):

$$(\check{m}_L, \check{\theta}_L) := \underset{(m,\theta)\in\mathcal{M}_L\times\Theta}{\arg\min} \quad Q_n(m,\theta), \tag{2}$$

where

$$Q_n(m,\theta) := \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(\theta^\top X_i)\}^2$$

and $\mathcal{M}_L$ denotes the class of all L-Lipschitz real-valued convex functions on $\mathbb{R}$ and

$$\Theta := \{\eta = (\eta_1, \dots, \eta_d) \in \mathbb{R}^d : |\eta| = 1 \text{ and } \eta_1 \geq 0\} \subset S^{d-1}.$$

Here $|\cdot|$ denotes the usual Euclidean norm, and S^{d-1} is the Euclidean unit sphere in $\mathbb{R}^d$. The norm-1 and the positivity constraints are necessary for identifiability of the model.[1]

The Lipschitz constraint in Equation (2) is not restrictive as all convex functions are Lipschitz in the interior of their domains. Furthermore in shape-constrained single-index models, the Lipschitz constraint is known to lead to computational advantages (Kalai and Sastry 2009; Kakade et al. 2011; Lim 2014; Ganti et al. 2015; Mazumder et al. 2019). Additionally on the theoretical side, the Lipschitzness assumption allows us to control the behavior of the estimator near the boundary of its domain. This control is crucial for establishing semiparametric efficiency. To the best of our knowledge, this is the first work proving semiparametric efficiency for an estimator in a *bundled parameter* problem (where the parametric and nonparametric components are intertwined; see Huang and Wellner 1997) where the nonparametric estimate is shape constrained and non-smooth. Note that the convexity constraint in Equation (2) leads to a convex piecewise affine estimator $\check{m}_L$ for the link function m_0; see Section 3 for a detailed discussion.

Our theoretical and methodological study can be split in two broad categories. In Section 3, we find the rate of convergence of the CLSE as defined in Equation (2), whereas in Section 4 we establish the asymptotic normality and semiparametric efficiency of $\check{\theta}_L$. Suppose that m_0 is L_0-Lipschitz, that is, $m_0 \in \mathcal{M}_{L_0}$. If the tuning parameter L is chosen such that $L \geq L_0$, then under mild distributional assumptions on X and ϵ, we show that $\check{m}_L$ and $\check{m}_L(\check{\theta}_L^\top \cdot)$ are minimax rate optimal for estimating m_0 and $m_0(\theta_0^\top \cdot)$, respectively; see Theorems 3.1 and 3.4. We also allow for the tuning parameter L to depend on the data and show that the rate of convergence of $\check{m}_L(\check{\theta}_L \cdot)$ is uniform in $L \in [L_0, nL_0]$, up to a $\sqrt{\log\log n}$ multiplicative factor; see Theorem 3.2. This result justifies the usage of a data-dependent choice of L, such as cross-validation. Additionally, in Theorem 3.5, we find the rate of convergence of $\check{m}_L'$. In Section 4, we establish that if $L \geq L_0$, then $\check{\theta}_L$ is $\sqrt{n}$-consistent and $n^{1/2}(\check{\theta}_L - \theta_0)$ is asymptotically normal with mean 0 and finite variance; see Theorem 4.1. The asymptotic normality of $\check{\theta}_L$ can be readily used to construct confidence intervals for θ_0. Further, we show that if the errors happen to be homoscedastic, then $\check{\theta}_L$ is semiparametrically efficient.

Our contributions on the computational side are 2-fold. In Section S.1 of the supplementary file, we propose an alternating descent algorithm for estimation in the single-index model (1). Our descent algorithm works as follows: when θ is fixed, the m update is obtained by solving a quadratic program with linear constraints, and when m is fixed, we update θ by taking a small step on the Stiefel manifold Θ with a guarantee of descent. We implement the proposed algorithm in the R package `simest`. Through extensive simulations (see Section 5 and Section S.4 of the supplementary file), we show that the finite sample performance of our estimators is robust to the choice of the tuning parameter L. Thus, we think the practitioner can choose L to be very large without sacrificing any finite sample performance. Even though the minimization problem is non-convex, we illustrate that the proposed algorithm (when used with multiple random starting points) performs well in a variety of simulation scenarios when compared to existing methods.

1.1. Semiparametric Efficiency and Shape Constraints

Although estimation in single-index models under smoothness assumptions is well-studied (see, e.g., Li and Duan 1989; Powell, Stock, and Stoker 1989; Ichimura 1993; Härdle, Hall, and Ichimura 1993; Hristache, Juditsky, and Spokoiny 2001; Delecroix, Hristache, and Patilea 2006; Wang and Yang 2009; Cui, Härdle, and Zhu 2011 and the references therein), estimation and efficiency in shape-restricted single-index models have not received much attention. The earliest reference on this topic we could find was the work of Murphy, van der Vaart, and Wellner (1999), where the authors considered a penalized likelihood approach in the current status regression model (which is similar to the single-index model) with a monotone link function. Chen and Samworth (2016) considered maximum likelihood estimation in a generalized additive index model (a more general model than Equation (1)) and only prove consistency of the proposed estimators. In Balabdaoui, Durot, and Jankowski (2019), the authors studied model (1) under monotonicity constraint and prove $n^{1/3}$-consistency of the LSE of θ_0; however they do not obtain the limiting distribution of the estimator of θ_0. Balabdaoui, Groeneboom, and Hendrickx (2019) proposed a tuning parameter-free $\sqrt{n}$-consistent (but not semiparametrically efficient) estimator for the index parameter in the monotone single-index model.

In this article, we show that $\check{\theta}_L$ is semiparametrically efficient under homoscedastic errors. Our proof of the semiparametric efficiency is novel and can be applied to other semiparametric models when the estimator does not readily satisfy the efficient score equation. In fact, we provide a new and general technique for establishing semiparametric efficiency of an estimator when the nuisance tangent set is not the space of all square integrable functions. The basic idea is as follows. Suppose $\ell_{\theta_0,m_0}(y,x)$

[1] Without any sign or scale constraint on Θ no (m_0, θ_0) will be identifiable. To see this, fix any (m_0, θ_0) and define $m_1(t) := m_0(-2t)$ and $\theta_1 = -\theta_0/2$, then $m_0(\theta_0^\top \cdot) \equiv m_1(\theta_1^\top \cdot)$; see Carroll et al. (1997), Cui, Härdle, and Zhu (2011), and Gaïffas and Lecué (2007) for identifiability of the model (1). Also see Section 2.2 for further discussion.

represents the semiparametrically efficient influence function, meaning that the "best" estimator $\tilde{\theta}$ of θ_0 satisfies the following asymptotic linear expansion:

$$\eta^\top(\tilde{\theta} - \theta_0) = \frac{1}{n}\sum_{i=1}^{n} \eta^\top \ell_{\theta_0,m_0}(Y_i, X_i) + o_p(n^{-1/2}), \tag{3}$$

for every $\eta \in \mathbb{R}^d$. A crucial step in establishing that $\check{\theta}_L$ satisfies Equation (3) is to show for any $\eta \in \mathbb{R}^d$,

$$n^{-1}\sum_{i=1}^{n} \eta^\top \ell_{\check{\theta}_L,\check{m}_L}(Y_i, X_i) = o_p(n^{-1/2}),$$

that is, $\check{\theta}_L$ is an *approximate zero* of the efficient score equation (Bolthausen, Perkins, and van der Vaart 2002, theor.6.20). Because $(\check{m}_L, \check{\theta}_L)$ minimizes $(m, \theta) \mapsto Q_n(m, \theta)$ over $\mathcal{M}_L \times \Theta$, the traditional way to prove the approximate zero property is to use the fact that $\partial Q_n(\check{m}_L + ta, \check{\theta}_L + t\eta)/\partial t|_{t=0} = 0$ for all perturbation "directions" (a, η) and find an a such that the derivative of $t \mapsto Q_n(\check{m}_L + ta, \check{\theta}_L + t\eta)$ at $t = 0$ is $n^{-1}\sum_{i=1}^{n} \eta^\top \ell_{\check{\theta}_L,\check{m}_L}(Y_i, X_i)$; see, for example, Newey and Stoker (1993). In fact, using this method one can often show that the estimator satisfies the efficient score equation *exactly*. If $\check{m}_L + ta$ is a valid path (i.e., $\check{m}_L + ta \in \mathcal{M}_L$ for all t in some neighborhood of zero) for an arbitrary but "smooth" a then it is relatively straightforward to establish the approximate zero property Newey and Stoker (1993).[2] However, this approach does not work when the nonparametric function m_0 is constrained. This is because under constraints, $\check{m}_L + ta$ might not be a valid path for arbitrary but smooth a. The novelty of our proposed approach lies in observing that in contrast to $t \mapsto \check{m}_L + ta$, $t \mapsto \Pi_{\mathcal{M}_L}(\check{m}_L + ta)$ is always a valid path for every smooth a; here $\Pi_{\mathcal{M}_L}(f)$ is the L_2-projection of f onto $\mathcal{M}_L$. Thus, if $t \mapsto \Pi_{\mathcal{M}_L}(\check{m}_L + ta)$ is differentiable, then $\partial Q_n(\Pi_{\mathcal{M}_L}(\check{m}_L + ta), \check{\theta}_L + t\eta)/\partial t|_{t=0} = 0$ for any perturbation (a, η). Then establishing that $\check{\theta}_L$ is an approximate zero boils down to finding an a such that

$$\frac{\partial}{\partial t} Q_n(\Pi_{\mathcal{M}_L}(\check{m}_L + ta), \check{\theta}_L + t\eta)\Big|_{t=0} = n^{-1}\sum_{i=1}^{n} \eta^\top \ell_{\check{\theta}_L,\check{m}_L}(Y_i, X_i) + o_p(n^{-1/2}).$$

Differentiability of projection operators is well-studied; see, for example, Dharanipragada and Arun (1996), Fitzpatrick and Phelps (1982), McCormick and Tapia (1972), Shapiro (1994), and Sokolowski and Zolesio (1992) for sufficient conditions for a general projection operator to be differentiable. The generality and the usefulness of our technique can be understood from the fact that no specific structure of $\check{m}_L$ or $\mathcal{M}_L$ is used in the previous discussion; we elaborate on this in Section 4.2. On the other hand, existing methods (see, e.g., Murphy, van der Vaart, and Wellner 1999) require delicate (and not generalizable) use of the structure of the nonparametric estimator to create valid paths around the nonparametric function; see, for example, Murphy, van der Vaart, and Wellner (1999) for semiparametric efficiency in current status regression, and (Bolthausen, Perkins, and van der Vaart 2002, chap. 9.3) and Huang (1996) for efficiency in the Cox proportional hazard model with current status data; see Example 4.1.

1.2. Organization of the Exposition

Our exposition is organized as follows: in Section 2, we introduce some notation and formally define the CLSE. In Section 3, we state our assumptions, prove consistency, and give rates of convergence for the CLSE. In Section 4, we detail our new method to prove semiparametric efficiency of the CLSE. We use this to prove $\sqrt{n}$-consistency, asymptotic normality, and efficiency (when the errors happen to be homoscedastic) of the CLSE of θ_0. We discuss an algorithm to compute the proposed estimator in Section S.1. In Section 5, we provide an extensive simulation study and compare the finite sample performance of the proposed estimator with existing methods in the literature. In Section 6, we analyze the Boston housing data Harrison and Rubinfeld (1978) and the car mileage data Donoho and Ramos (1983) introduced in Examples 1.1 and 1.2 in more details. In both of the cases, we show that the natural shape constraint leads to stable and interpretable estimates. Section 7 provides a brief summary of the article and discusses some open problems.

Section numbers in the supplementary file are prefixed with "S." Section S.2 of the supplementary file provides some insights into the proof of Theorem 4.1, one of our main results. Section S.4 provides further simulation studies. Section S.5 provides additional discussion on the identifiability of the parameters. Sections S.7–S.12 contain the proofs of our results. Section S.10 completes our novel proof of semiparametric efficiency sketched in Section 4.2.

2. Notation and Estimation

2.1. Preliminaries

In what follows, we assume that we have iid data $\{(X_i, Y_i)\}_{i=1}^{n}$ from Equation (1). We start with some notation. Let $\chi \subset \mathbb{R}^d$ denote the support of X and define

$$\begin{aligned} D &:= \operatorname{conv}\{\theta^\top x : x \in \chi, \theta \in \Theta\}, \\ D_\theta &:= \{\theta^\top x : x \in \chi\}, \quad \text{and} \quad D_0 := D_{\theta_0}, \end{aligned} \tag{4}$$

where $\operatorname{conv}(A)$ denotes the convex hull of the set A. Let $\mathcal{M}_L$ denote the class of real-valued convex functions on D that are uniformly Lipschitz with Lipschitz bound L. For any $m \in \mathcal{M}_L$, let m' denote the nondecreasing right derivative of the real-valued convex function m. Because m is a uniformly Lipschitz function with Lipschitz constant L, without loss of generality, we can assume that $|m'(t)| \le L$, for all $t \in D$. We use $\mathbb{P}$ to denote the probability of an event and $\mathbb{E}$ for the expectation of a random quantity. For any $\theta \in \Theta$, let $P_{\theta^\top X}$ denote the distribution of $\theta^\top X$. For $g : \chi \to \mathbb{R}$, define $||g||^2 := \int g^2(x)dP_X(x)$. Let $P_{\epsilon,X}$ denote the joint distribution of (ϵ, X) and let $P_{\theta,m}$ denote the joint distribution of (Y, X) when $Y = m(\theta^\top X) + \epsilon$, where ϵ is defined in (1). In particular, P_{θ_0,m_0} denotes the joint distribution of (Y, X) when $X \sim P_X$ and (Y, X) satisfies (1). For any set

[2] As $\theta \in \Theta$ is restricted to have norm 1, $\theta + t\eta$ does not belong to the parametric space for $t \neq 0$ and $\eta^\top\theta \neq 0$. However, this can be easily remedied by considering another path that is differentiable and has the same "direction;" we define such a path in Equation (11).

$I \subseteq \mathbb{R}^p$ $(p \geq 1)$ and any function $g : I \to \mathbb{R}$, we define $||g||_\infty := \sup_{u \in I} |g(u)|$ and $||g||_{I_1} := \sup_{u \in I_1} |g(u)|$, for $I_1 \subseteq I$. The notation $a \lesssim b$ is used to express that $a \leq Cb$ for some constant $C > 0$. For any function $f : \chi \to \mathbb{R}^r$, $r \geq 1$, let $\{f_i\}_{1 \leq i \leq r}$ denote each of the components of f, that is, $f(x) = (f_1(x), \ldots, f_r(x))$ and $f_i : \chi \to \mathbb{R}$. We define $||f||_{2,P_{\theta_0,m_0}} := \sqrt{\sum_{i=1}^r ||f_i||^2}$ and $||f||_{2,\infty} := \sqrt{\sum_{i=1}^r ||f_i||_\infty^2}$. For any function $g : D \to \mathbb{R}$ and $\theta \in \Theta$, we define $(g \circ \theta)(x) := g(\theta^\top x)$, for all $x \in \chi$. We use the following (standard) empirical process theory notation. For any function $f : \mathbb{R} \times \chi \to \mathbb{R}$, $\theta \in \Theta$, and $m : \mathbb{R} \to \mathbb{R}$, we define

$$P_{\theta,m} f := \int f(y,x) dP_{\theta,m}(y,x).$$

Note that $P_{\theta,m}f$ can be a random variable when θ or m or both are random. Moreover, for any function $f : \mathbb{R} \times \chi \to \mathbb{R}$, we define $\mathbb{P}_n f := n^{-1} \sum_{i=1}^n f(Y_i, X_i)$ and $\mathbb{G}_n f := \sqrt{n}(\mathbb{P}_n - P_{\theta_0,m_0})f$.

2.2. Identifiability

We now discuss the identifiability of $m_0 \circ \theta_0$ and (m_0, θ_0). Letting $Q(m, \theta) := \mathbb{E}[Y - m(\theta^\top X)]^2$, observe that (m_0, θ_0) minimizes $Q(\cdot, \cdot)$. In fact we can show in Section S.5.1, that

$$\inf_{\substack{\{(m,\theta):\, m \circ \theta \in L_2(P_X) \\ \text{and } \|m \circ \theta - m_0 \circ \theta_0\| > \delta\}}} \left[Q(m,\theta) - Q(m_0,\theta_0)\right] > \delta^2, \tag{5}$$

for any $\delta > 0$.

This implies that $m_0 \circ \theta_0$ is always identifiable and further, one can hope to consistently estimate $m_0 \circ \theta_0$ by minimizing the sample version of $Q(m, \theta)$; see (2).

Note that the identification of $m_0 \circ \theta_0$ does not guarantee that both m_0 and θ_0 are separately identifiable. Hence, in what follows, when dealing with the properties of separated parameters, we will directly assume:

(A0) The parameters $m_0 \in \mathcal{M}_{L_0}$ and $\theta_0 \in \Theta$ are separately identifiable, that is, $m \circ \theta = m_0 \circ \theta_0$ for some $(m, \theta) \in \mathcal{M}_{L_0} \times \Theta$ implies that $m = m_0$ and $\theta = \theta_0$.

Ichimura (1993) had found general sufficient conditions on the distribution of X under which 2.2 holds; these sufficient conditions allow for some components of X to be discrete, also see (Horowitz 1998, pp. 12–17) and (Li and Racine 2007, prop. 8.1). When X has a density with respect to Lebesgue measure, (Lin and Kulasekera 2007, theor. 1) find a simple sufficient condition for (A0). We discuss and compare these two sufficient conditions in Section S.5.2 of the supplementary file.

3. Convex and Lipschitz Constrained LSE

Recall that CLSE is defined as the minimizer of $(m, \theta) \mapsto Q_n(m, \theta)$ over $\mathcal{M}_L \times \Theta$. Because $Q_n(m, \theta)$ depends only on the values of the function at $\{\theta^\top X_i\}_{i=1}^n$, it is immediately clear that the minimizer $\check{m}_L$ is unique only at $\{\check{\theta}_L^\top X_i\}_{i=1}^n$. Since $\check{m}_L$ is restricted to be convex, we interpolate the function linearly between $\check{\theta}_L^\top X_i$'s and extrapolate the function linearly outside the data points.[3] Thus, $\check{m}$ is piecewise affine. In Section S.7 of the supplementary file, we prove the existence of the minimizer in Equation (2). The optimization problem (2) might not have a unique minimizer and the results that follow hold true for any global minimizer.

Remark 3.1. For every fixed θ, $m(\in \mathcal{M}_L) \mapsto Q_n(m, \theta)$ has a unique minimizer. The minimization over the class of uniformly Lipschitz functions is a quadratic program with linear constraints and can be computed easily; see Section S.11.

3.1. Asymptotic Analysis of the Regression Function Estimate

In this section, we study the asymptotic behavior of $\check{m}_L \circ \check{\theta}_L$. We will now list the assumptions under which we study the rates of convergence of the CLSE for the regression function.

(A1) The unknown convex link function m_0 is bounded by some constant M_0 (≥ 1) on D and is uniformly Lipschitz with Lipschitz constant L_0.

(A2) The support of X, χ, is a subset of $\mathbb{R}^d$ and $\sup_{x \in \chi} |x| \leq T$, for some finite $T \in \mathbb{R}$.

(A3) The error ϵ in model (1) has finite qth moment, that is, $K_q := \left[\mathbb{E}(|\epsilon|^q)\right]^{1/q} < \infty$ where $q \geq 2$. Moreover, $\mathbb{E}(\epsilon|X) = 0$, P_X a.e. and $\sigma^2(x) := \mathbb{E}(\epsilon^2|X = x) \leq \sigma^2 < \infty$ for all $x \in \chi$.

The above assumptions deserve comments. (A2) implies that the support of the covariates is bounded. In Assumption (A3), we allow ϵ to be heteroscedastic and ϵ can depend on X. Our assumption on ϵ is more general than those considered in the shape constrained literature, most works assume that all moments of ϵ are finite and "well-behaved," see, for example, Balabdaoui, Groeneboom, and Hendrickx (2019), Hristache, Juditsky, and Spokoiny (2001), and Xia et al. (2002).

Theorem 3.1 (proved in Section S.9.1) below provides an upper bound on the rate of convergence of $\check{m}_L \circ \check{\theta}_L$ to $m_0 \circ \theta_0$ under the $L_2(P_X)$ norm. The following result is a finite sample result and shows the explicit dependence of the rate of convergence on $L = L_n$, d, and q.

Theorem 3.1. Assume (A1)–(A3). Let $\{L_n\}_{n \geq 1}$ be a fixed sequence such that $L_n \geq L_0$ for all n and let

$$r_n := \min\left\{\frac{n^{2/5}}{d^{2/5} L_n}, \frac{n^{1/2-1/2q}}{L_n^{(3q+1)/(4q)}}\right\}. \tag{6}$$

Then for every $n \geq 1$ and $u \geq 1$, there exists a constant $\mathfrak{C} \geq 0$ depending only on σ, M_0, L_0, T, and K_q, and constant C depending only on K_q, σ, and q, such that

$$\sup_{\theta_0, m_0, \epsilon, X} \mathbb{P}\left(r_n ||\check{m}_{L_n} \circ \check{\theta}_{L_n} - m_0 \circ \theta_0|| \geq u\mathfrak{C}\right) \leq \frac{C}{u^q} + \frac{\sigma^2}{n},$$

where the supremum is taken over all $\theta_0 \in \Theta$ and all joint distributions of (ϵ, X) and parameters m_0 for which Assumptions (A1)–(A3) are satisfied with constants σ, M_0, L_0, T, and K_q. In particular, if $q \geq 5$, $d = O(1)$, and $L_n = O(1)$ as $n \to \infty$, then $||\check{m}_{L_n} \circ \check{\theta}_{L_n} - m_0 \circ \theta_0|| = O_p(n^{-2/5})$.

[3] Linear interpolation/extrapolation does not violate the convexity or the L-Lipschitz property.

Note that Equation (6) allows for the dimension d to grow with n and θ_0 to change with n. For example if $L_n \equiv L$ for some fixed $L \geq L_0$, then we have that $||\check{m}_{L_n} \circ \check{\theta}_{L_n} - m_0 \circ \theta_0|| = o_p(1)$ if $d = o(n^{1-1/q})$. In the rest of the article, we assume that d is fixed. In Proposition S.6.1 in Section 6.1, we find the minimax lower bound for the single-index model (1), and show that $\check{m}_L \circ \check{\theta}_L$ is minimax rate optimal when $q \geq 5$.

The next result shows that the rates in Theorem 3.1 are in fact uniform (up to a $\sqrt{\log\log n}$ factor) in $L \in [L_0, nL_0]$. This uniform-in-L result is important for the study of the estimator with a data-driven choice of L such as cross-validation or Lepski's method Lepski and Spokoiny (1997). Theorem 3.1 alone cannot provide such a rate guarantee because it requires L to be non-stochastic.

Theorem 3.2. Under the assumptions of Theorem 3.1, the CLSE satisfies

$$\sup_{L_0 \leq L \leq nL_0} \min\left\{\frac{n^{2/5}}{L}, \frac{n^{1/2-1/(2q)}}{\sqrt{L}}\right\} ||\check{m}_L \circ \check{\theta}_L - m_0 \circ \theta_0|| \\ = O_p\left(\sqrt{\log\log n}\right).$$

Remark 3.2 (Diverging L). The dependence on L in Theorems 3.1 and 3.2 suggest that the estimator may not be consistent if $L \equiv L_n$ diverges too quickly with the sample size. The simulation in Section 5.3 suggests that the estimation error has negligible dependence on L and that the dependence on L in Theorems 3.1 and 3.2 might be suboptimal. We believe this discrepancy is due to the lack of available technical tools to prove uniform boundedness of the estimator $\check{m}_{n,L}$ in terms of L. At present, we are only able to prove that with high probability, $||\check{m}_{n,L}||_\infty \leq LT + M_0 + 1$ for all $L \geq L_0$; see Lemma S.9.1. If one can prove $||\check{m}_{n,L}||_\infty \leq C$ for all $L \geq L_0$, with high probability, for a constant C independent of L, then our proofs can be modified to remove the dependence on L in Theorems 3.1 and 3.2.

3.2. Asymptotic Analysis of $\check{m}$ and $\check{\theta}$

In this section, we establish the consistency and find rates of convergence of $\check{m}_{L_n}$ and $\check{\theta}_{L_n}$ separately. In Theorem 3.1, we proved that $\check{m}_{L_n} \circ \check{\theta}_{L_n}$ converges in the $L_2(P_{\theta_0,m_0})$ norm but that does not guarantee that $\check{m}_{L_n}$ converges to m_0 in the $||\cdot||_{D_0}$ norm. A typical approach for proving consistency of $\check{m}_{L_n}$ is to prove that $\{\check{m}_{L_n}\}$ is precompact in the $||\cdot||_{D_0}$ norm (D_0 is defined in Equation (4)); see, for example, Balabdaoui, Durot, and Jankowski (2019), Murphy, van der Vaart, and Wellner (1999). The Arzelà-Ascoli theorem establishes that the necessary and sufficient condition for compactness (with respect to the uniform norm) of an arbitrary class of continuous functions on a bounded domain is that the function class be uniformly bounded and equicontinuous. However, if L_n is allowed to grow to infinity, then it is not clear whether the sequence of functions $\{\check{m}_{L_n}\}$ is equicontinuous. Thus, to study the asymptotic properties of $\check{m}_{L_n}$ and $\check{\theta}_{L_n}$, we assume that $L_n \equiv L \geq L_0$, is a fixed constant. For the rest of article, we will use $\check{m}$ and $\check{\theta}$ to denote $\check{m}_L$ (or $\check{m}_{L_n}$) and $\check{\theta}_L$ (or $\check{\theta}_{L_n}$), respectively. The next theorem (proved in Section S.9.4) establishes consistency of $\check{m}$ and $\check{\theta}$ separately. Recall that m_0' denotes the nondecreasing right derivative of the convex function m_0.

Theorem 3.3. Suppose the assumptions of Theorem 3.1 and (A0) hold. Then, for any fixed $L \geq L_0$ and any compact subset C in the interior of D_0, we have

$$|\check{\theta} - \theta_0| = o_p(1), \\ ||\check{m} - m_0||_{D_0} = o_p(1), \quad \text{and} \quad ||\check{m}' - m_0'||_C = o_p(1).$$

Fix an orthonormal basis $\{e_1, \dots, e_d\}$ of $\mathbb{R}^d$ such that $e_1 = \theta_0$. Define $H_{\theta_0} := [e_2, \dots, e_d] \in \mathbb{R}^{d\times(d-1)}$. We will use the following two additional assumptions to establish upper bounds on the rate of convergence of $\check{m}$ and $\check{\theta}$.

(A4) $H_{\theta_0}^\top \mathbb{E}\big[\mathrm{Var}(X|\theta_0^\top X)\{m_0'(\theta_0^\top X)\}^2\big] H_{\theta_0}$ is a positive-definite matrix.

(A5) The density of $\theta_0^\top X$ with respect to the Lebesgue measure is bounded above by $\overline{C}_d < \infty$.

Assumption (A4), is used to find the rate of convergence for $\check{\theta}$ and $\check{m}$ separately and is widely used in all works studying root-n consistent estimation of θ_0 in the single-index model, see, for example, Powell, Stock, and Stoker (1989), Ichimura (1993), Kuchibhotla and Patra (2020), and Balabdaoui, Groeneboom, and Hendrickx (2019); also see Remark 3.3. (A.5) is mild, and is satisfied if $X = (X_1, \dots, X_d)$ has a continuous covariate X_k such that (i) X_k has a bounded density; and (ii) $\theta_{0,k} > 0$. Compare Assumption (A5) with Ichimura (1993), Cui, Härdle, and Zhu (2011), and Balabdaoui, Groeneboom, and Hendrickx (2019), Wang and Yang (2009), and Wang and Wang (2015) where it is assumed that $\theta^\top X$ has a density bounded away from zero for all θ in a neighborhood of θ_0. Assumption (A5) is used to find rates of convergence of the derivative of the estimators of m_0. In Theorem 3.4, we only use the fact that $\theta_0^\top X$ is absolutely continuous with respect to Lebesgue measure. The following result (proved in Section S.9.5) establishes upper bounds on the rate of convergence of $\check{\theta}$ and $\check{m}$, respectively.

Theorem 3.4. If Assumptions (A0)–(A5) hold, $q \geq 5$, and $L \geq L_0$, then we have

$$|\check{\theta} - \theta_0| = O_p(n^{-2/5}) \quad \text{and} \\ \int (\check{m}(t) - m_0(t))^2 dP_{\theta_0^\top X}(t) dt = O_p(n^{-4/5}).$$

Remark 3.3. Note that, under homoscedastic errors in Equation (1), the efficient information for θ_0 is a scalar multiple of $H_{\theta_0}^\top \mathbb{E}\big[\mathrm{Var}(X|\theta_0^\top X)\{m_0'(\theta_0^\top X)\}^2\big] H_{\theta_0} =: \mathcal{I}_0$; see Section 4.1. If $\mathcal{I}_0$ is not positive definite, then there is zero information for θ_0 along some directions. In that case, we can show that $|\mathcal{I}_0^{1/2}(\check{\theta} - \theta_0)| = O_p(n^{-2/5})$; see (E.68) in the supplementary file.

A simple modification of the proof of Proposition S.6.1 will prove that $\check{m}$ is also minimax rate optimal. Under additional smoothness assumptions on m_0, in the following theorem (proved in Section S.9.7), we show that $\check{m}'$, the right derivative of $\check{m}$, converges to m_0' in both the L_2 and the supremum norms.

Theorem 3.5. Suppose assumptions of Theorem 3.4 hold and m_0' is $\frac{1}{2}$-Hölder continuous on D_0, then

$$||\check{m}' \circ \theta_0 - m_0' \circ \theta_0|| = O_p\big(n^{-2/15}\big) \quad \text{and}$$
$$||\check{m}' \circ \check{\theta} - m_0' \circ \check{\theta}|| = O_p\big(n^{-2/15}\big). \qquad (7)$$

Further, if m_0 is twice continuously differentiable and Assumption (A4) (in Section 4), then for any compact subset C in the interior of D_0, we have

$$\sup_{t \in C} |\check{m}(t) - m_0(t)| = O_p(n^{-8/(25+5\beta)}) \quad \text{and}$$
$$\sup_{t \in C} |\check{m}'(t) - m_0'(t)| = O_p(n^{-4/(25+5\beta)}). \qquad (8)$$

Remark 3.4. As in (7), (8) can also be proved under γ-Hölder continuity of m_0', but in this case the rate of convergence depends on γ explicitly. Assumption (B2) allows for the density of $\theta_0^\top X$ to be zero at some points in its support; see Section 4 for a detailed discussion. Further if the density of $\theta_0^\top X$ is bounded away from zero, then β can be taken to be 0.

Remark 3.5. The condition $q \geq 5$ in Theorems 3.4 and 3.5 can be relaxed at the expense of slower rates of convergence. In fact, by following the arguments in the proofs, we can show, with $p_n := \max\{n^{-2/5}, n^{-1/2+1/(2q)}\}$ for any $q \geq 2$, that $|\check{\theta} - \theta_0| = O_p(p_n)$, and

$$||\check{m} \circ \theta_0 - m_0 \circ \theta_0|| = O_p(p_n),$$
$$||\check{m}' \circ \theta_0 - m_0' \circ \theta_0|| = O_p(p_n^{1/3}) \quad \text{and}$$
$$||\check{m}' \circ \check{\theta} - m_0' \circ \check{\theta}|| = O_p(p_n^{1/3}).$$

Remark 3.6 (Additional shape constraints on the link function). It might often be the case that in addition to convexity, the practitioner is interested in imposing additional shape constraints (such as monotonicity, unimodality, or k-monotonicity Guntuboyina and Sen 2018) on m_0. For example, in the datasets considered in Examples 1.1 and 1.2, the link function is plausibly both convex *and* monotone; see Chen and Samworth (2016) for further motivation on additional shape constraints. The conclusions (and proofs) of Theorems 3.1 and 3.2–3.5 also hold for the CLSE under additional constraints on the link function. An intuitive explanation is that the parameter space $\mathcal{M}_L$ is only reduced by imposing additional constraints on the link function and this can only give better rates (if not the same). In case of an additional monotonicity constraint on m_0, one can modify the proof of Proposition S.6.1 to show that the rate obtained in Theorem 3.1 is in fact minimax optimal for the the CLSE (under further monotonicity constraint).

4. Semiparametric Inference for the CLSE

The main result in this section shows that $\check{\theta}$ is $\sqrt{n}$-consistent and asymptotically normal; see Theorem 4.1. Moreover, $\check{\theta}$ is shown to be semiparametrically efficient for θ_0 if the errors happen to be homoscedastic. The asymptotic analysis of $\check{\theta}$ is involved as $\check{m}$ is a piecewise affine function and hence not differentiable everywhere.

Before deriving the limit law of $\check{\theta}$, we introduce some notations and assumptions. Let $p_{\epsilon,X}$ denote the joint density (with respect to some dominating measure on $\mathbb{R} \times \chi$) of (ϵ, X). Let $p_{\epsilon|X}(\cdot, x)$ and $p_X(\cdot)$ denote the corresponding conditional probability density of ϵ given $X = x$ and the marginal density of X, respectively. In the following we list additional assumptions used in Theorem 4.1. Recall D and D_0 from Equation (4) and let Λ denote the Lebesgue measure.

(B1) $m_0 \in \mathcal{M}_{L_0}$ and m_0 is $(1+\gamma)$-Hölder continuous on D_0 for some $\gamma > 0$. Furthermore, m_0 is strongly convex on D, that is, there exists a $\kappa_0 > 0$ such that $m_0(t) - \kappa_0 t^2$ is convex.

(B2) There exists $\beta \geq 0$ and $\underline{C}_d > 0$ such that $\mathbb{P}(\theta_0^\top X \in I) \geq \underline{C}_d \, \Lambda(I)^{1+\beta}$, for all intervals $I \subset D_0$.

For every $\theta \in \Theta$, define $h_\theta(u) := \mathbb{E}[X | \theta^\top X = u]$.

(B3) The function $u \mapsto h_{\theta_0}(u)$ is 1/2-Hölder continuous and for a constant $\bar{M} > 0$ and every $\theta \in \Theta$,

$$\mathbb{E}\Big(|h_\theta(\theta_0^\top X) - h_{\theta_0}(\theta_0^\top X)|^2\Big) \leq \bar{M}|\theta - \theta_0|. \qquad (9)$$

(B4) The density $p_{\epsilon|X}(e, x)$ is differentiable with respect to e for all $x \in \chi$.

Assumptions (B1)–(B4) deserve comments. (B1) is much weaker than the standard assumptions used in semiparametric inference in single-index models (Murphy, van der Vaart, and Wellner 1999, theor. 3.2). Assumption (B2) is an improvement compared to the assumptions in the existing literature. Assumption (B2) pertains to the distribution of $\theta_0^\top X$ and is inspired by (Gaïffas and Lecué 2007, assump. (D)). In contrast, most existing works require the density of $\theta_0^\top X$ to be bounded away from zero (i.e., $\beta = 0$); see, for example, Ichimura (1993, assump. 5.3(II)), Cui, Härdle, and Zhu (2011, assump. (d)), Balabdaoui, Groeneboom, and Hendrickx (2019, lem. F.3), Wang and Yang (2009, assump. A2), and Wang and Wang (2015, assump. A2). Our assumption is significantly weaker because it allows the density of $\theta_0^\top X$ to be zero at some points in its support. For example, when $X \sim \text{Uniform}[0, 1]^d$, the density of $\theta_0^\top X$ might not be bounded away from zero (Gaïffas and Lecué 2007, fig. 1), but (B2) holds with $\beta = 1$. Assumption (B3) can be favorably compared to those in Murphy, van der Vaart, and Wellner (1999, theor. 3.2), Groeneboom and Hendrickx (2018, assump. (A5)), Balabdaoui, Groeneboom, and Hendrickx (2019, assump. (A5)), and Song (2014, assump. G2 (ii)). We use the smoothness assumption (B3) when establishing semiparametric efficiency of $\check{\theta}$. The Lipschitzness assumption (4.1) can be verified by using the techniques of Alonso and Brambila-Paz (1998), when $u \mapsto h_\theta(u)$ is 1/2-Hölder continuous for all θ in a neighborhood of θ_0 and the Hölder constants are uniformly bounded in θ.

In general, establishing semiparametric efficiency of an estimator proceeds in two steps. Let $\hat{\xi}$ and $\hat{\gamma}$ denote the estimators of a parametric component ξ_0 and a nuisance component γ_0 in a general semiparametric model. In a broad sense, the proof of semiparametric efficiency of $\hat{\xi}$ involves two main steps: (i) finding the efficient score of the model at the truth (call it ℓ_{ξ_0,γ_0}); and (ii) proving that $(\hat{\xi}, \hat{\gamma})$ satisfies $\mathbb{P}_n \ell_{\hat{\xi},\hat{\gamma}} = o_p(n^{-1/2})$;

see (Bolthausen, Perkins, and van der Vaart 2002, pp. 436–437) for a detailed discussion. In Sections 4.1 and 4.2, we discuss steps (i) and (ii) in our context, respectively.

4.1. Efficient Score

In this subsection, we calculate the efficient score for the model:

$$Y = m(\theta^\top X) + \epsilon, \tag{10}$$

where m, X, and ϵ satisfy assumptions (B1)–(B4). First, observe that the parameter space Θ is a closed subset of $\mathbb{R}^d$ and the interior of Θ in $\mathbb{R}^d$ is the empty set. Thus, to compute the score for model (10), we construct a path on the sphere. We use $\mathbb{R}^{d-1}$ to parameterize the paths for model (10) on Θ when $\theta_{0,1} > 0$. For each $\eta \in \mathbb{R}^{d-1}$, $s \in \mathbb{R}$, and $|s| \leq |\eta|^{-1}$, define the following path, with "direction" η, through θ (which lies on the unit sphere)

$$\zeta_s(\theta, \eta) := \sqrt{1 - s^2|\eta|^2}\,\theta + sH_\theta\eta, \tag{11}$$

where for every $\theta \in \Theta$, $H_\theta \in \mathbb{R}^{d\times(d-1)}$ is such that for every $\eta \in \mathbb{R}^{d-1}$, $|H_\theta\eta| = |\eta|$ and $H_\theta\eta$ is orthogonal to θ. Furthermore, we need $\theta \mapsto H_\theta$ to satisfy some smoothness properties; see Kuchibhotla and Patra (2020, lem. 1) for such a construction. Note that, if $\theta_{0,1} = 0$, then for any s in a neighborhood of zero, there exists an $\eta \in \mathbb{R}^{d-1}$ such that $\zeta_s(\theta_0, \eta) \notin \Theta$. Thus, if $\theta_{0,1} = 0$, then θ_0 lies on the "boundary" of Θ and the existing semiparametric theory breaks down. Therefore, for the rest of the article, we assume that $\theta_{0,1}$ is strictly positive.

The log-likelihood of model (10) is $l_{\theta,m}(y, x) = \log[p_{\epsilon|X}(y - m(\theta^\top x), x)p_X(x)]$. For any $\eta \in S^{d-2}$, consider the path defined as $s \mapsto \zeta_s(\theta, \eta)$. Note that by the definition of H_θ, $s \mapsto \zeta_s(\theta, \eta)$ is a valid path in Θ through θ; that is, $\zeta_0(\theta, \eta) = \theta$ and $\zeta_s(\theta, \eta) \in \Theta$ for every s in some neighborhood of 0. Thus, the score for the parametric submodel is

$$\left.\frac{\partial l_{\zeta_s(\theta,\eta),m}(y,x)}{\partial s}\right|_{s=0} = \eta^\top S_{\theta,m}(y,x), \tag{12}$$

where

$$S_{\theta,m}(y,x) := -\frac{p'_{\epsilon|X}\big(y - m(\theta^\top x), x\big)}{p_{\epsilon|X}\big(y - m(\theta^\top x), x\big)} m'(\theta^\top x)H_\theta^\top x.$$

The next step in computing the efficient score for model (10) at (m, θ) is to compute the nuisance tangent space of the model (here the nuisance parameters are $p_{\epsilon|X}, p_X$, and m). To do this defines a parametric submodel for the unknown nonparametric components:

$$\begin{aligned} m_{s,a}(t) &= m(t) - sa(t), \\ p_{\epsilon|X;s,b}(e,x) &= p_{\epsilon|X}(e,x)(1 + sb(e,x)), \\ p_{X;s,q}(x) &= p_X(x)(1 + sq(x)), \end{aligned}$$

where $s \in \mathbb{R}$, $b : \mathbb{R} \times \chi \to \mathbb{R}$ is a bounded function such that $\mathbb{E}(b(\epsilon, X)|X) = 0$ and $\mathbb{E}(\epsilon b(\epsilon, X)|X) = 0$, $q : \chi \to \mathbb{R}$ is a bounded function such that $\mathbb{E}(q(X)) = 0$, and $a \in \mathcal{D}_m$, with

$$\begin{aligned} \mathcal{D}_m := \Big\{f \in L_2(\Lambda) : \;& f'(\cdot) \text{ exists and} \\ & m_{s,f}(\cdot) \in \mathcal{M}_L \text{ for all} \\ & s \in B_0(\delta) \text{ for some } \delta > 0\Big\}. \end{aligned}$$

Note that when m satisfies (B1) then $\mathcal{D}_m$ reduces to $\mathcal{D}_m = \{f \in L_2(\Lambda) : f'(\cdot) \text{ exists}\}$. Thus, $\overline{\text{lin}}\,\mathcal{D}_m = L_2(\Lambda)$. Newey and Stoker (1993, theor. 4.1) (also see Ma and Zhu 2013, prop. 1) showed that when the parametric score is $\eta^\top S_{\theta,m}(\cdot,\cdot)$ and the nuisance tangent space corresponding to m is $L_2(\Lambda)$, then the efficient score for model (10) is

$$\frac{1}{\sigma^2(x)}(y - m(\theta^\top x))m'(\theta^\top x)H_\theta^\top \times \left\{x - \frac{\mathbb{E}(\sigma^{-2}(X)X|\theta^\top X = \theta^\top x)}{\mathbb{E}(\sigma^{-2}(X)|\theta^\top X = \theta^\top x)}\right\}. \tag{13}$$

Note that the efficient score depends on $p_{\epsilon|X}$ and p_X only through $\sigma^2(\cdot)$. However, if the errors happen to be homoscedastic (i.e., $\sigma^2(\cdot) \equiv \sigma^2$), then the *efficient score* is $\ell_{\theta,m}(x,y)/\sigma^2$, where

$$\ell_{\theta,m}(x,y) := (y - m(\theta^\top x))m'(\theta^\top x)H_\theta^\top[x - h_\theta(\theta^\top x)]. \tag{14}$$

As $\sigma^2(\cdot)$ is unknown we restrict ourselves to efficient estimation under homoscedastic error; see Remark 4.2 for a brief discussion.

4.2. Efficiency of the CLSE

The $\sqrt{n}$-consistency, asymptotic normality, and efficiency (when the errors are homoscedastic) of $\check{\theta}$ will be established *if* we could show that

$$\sqrt{n}\,\mathbb{P}_n\ell_{\check{\theta},\check{m}} = o_p(1) \tag{15}$$

and the class of functions $\ell_{\theta,m}$ indexed by (θ, m) in a "neighborhood" of (θ_0, m_0) satisfies some technical conditions; see, for example, Bolthausen, Perkins, and van der Vaart (2002, chap. 6.5). As discussed in Section 1.1, because $(\check{m}, \check{\theta})$ minimizes $(m, \theta) \mapsto Q_n(m, \theta)$ over $\mathcal{M}_L \times \Theta$, the traditional way to prove Equation (15) is to use the fact that $\partial Q_n(\check{m}_{s,a}, \zeta_s(\theta, \eta))/\partial s|_{s=0} = 0$ for any (a, η) such that $s \mapsto (\check{m}_{s,a}, \zeta_s(\theta, \eta))$ is a valid path (i.e., $a \in \overline{\text{lin}}\,\mathcal{D}_{\check{m}}$). One then finds $(a, \eta) \in \mathcal{D}_{\check{m}} \times \mathbb{R}^{d-1}$ such that the derivative of $s \mapsto Q_n(\check{m}_{s,a}, \zeta_s(\theta, \eta))$ at $s = 0$ is approximately $n^{-1}\sum_{i=1}^n \eta^\top \ell_{\check{\theta},\check{m}}(Y_i, X_i)$; such an (a, η) is called the (approximate) *least favorable submodel*; see Bolthausen, Perkins, and van der Vaart (2002, sec. 9.2). In Section 4.1, we saw that if m is strongly convex then $\overline{\text{lin}}\,\mathcal{D}_m = L_2(\Lambda)$. However, $\check{m}$ is piecewise affine and we can only show that $\overline{\text{lin}}\,\mathcal{D}_{\check{m}} \subset L_2(\Lambda)$. Thus, $s \mapsto \check{m}_{s,a}$ is valid path only if $a \in \mathcal{D}_{\check{m}}$; see Murphy, van der Vaart, and Wellner (1999) for another example where $\overline{\text{lin}}\,\mathcal{D}_{\check{m}} \neq L_2(\Lambda)$. In such cases, it is hard to find the least favorable submodel as often the step to compute the least favorable model involves computing projection onto $\overline{\text{lin}}\,\mathcal{D}_{\check{m}}$; see, for example, Newey (1990). Thus, when $\overline{\text{lin}}\,\mathcal{D}_{\check{m}}$ is not $L_2(\Lambda)$ (or a very simple subspace of $L_2(\Lambda)$), the standard linear path arguments fail to find the least favorable submodel. To overcome this, Murphy, van der Vaart, and Wellner (1999) used a very complicated and nonlinear path; see (Murphy, van der Vaart, and Wellner 1999, sec. 6.2); also see Kuchibhotla and Patra (2020).

Our proposed technique crucially relies on the observation that $s \mapsto \Pi_{\mathcal{M}_L}(\check{m}_{s,a})$ is a valid path for every $a \in L_2(\Lambda)$. Thus, if $s \mapsto \Pi_{\mathcal{M}_L}(\check{m}_{s,a})$ is differentiable, then establishing that $\check{\theta}$ is an approximate zero boils down to finding an $a \in L_2(\Lambda)$ such that

$$\frac{\partial}{\partial s} Q_n(\Pi_{\mathcal{M}_L}(\check{m}_{s,a}), \zeta_s(\theta,\eta))\Big|_{s=0} = n^{-1}\sum_{i=1}^{n} \eta^\top \ell_{\check{\theta},\check{m}}(Y_i, X_i) + o_p(n^{-1/2}). \tag{16}$$

for every $\eta \in \mathbb{R}^{d-1}$. In Section S.10, we show $s \mapsto \Pi_{\mathcal{M}_L}(\check{m}_{s,a})$ is differentiable if $a \in \mathcal{X}_{\check{m}}$, where

$$\mathcal{X}_{\check{m}} := \big\{a \in L_2(\Lambda) : a \text{ is a piecewise affine continuous function with kinks at } \{\check{t}_i\}_{i=1}^{\mathfrak{p}}\big\}, \tag{17}$$

and $\{\check{t}_i\}_{i=1}^{\mathfrak{p}}$ are the set of kinks of $\check{m}$. For a piecewise affine function, a kink is a point where the slope changes. Furthermore, in Theorem S.10.1, we find an $a \in \mathcal{X}_{\check{m}}$ that satisfies Equation (16). The advantage of the technique proposed here is that the construction of approximate least favorable submodel is analytic and does not rely on the ability of the user to "guess" the least favorable submodel; see, e.g., Bolthausen, Perkins, and van der Vaart (2002, secs. 9.2 and 9.3) and Murphy, van der Vaart, and Wellner (1999). The above discussion and Bolthausen, Perkins, and van der Vaart (2002, theor. 6.20) led to our main result (Theorem 4.1) of this section. Recall S_{θ_0,m_0} and $\ell_{\theta,m}$ defined in Equations (12) and (14), respectively.

Theorem 4.1. Assume (A0)–(A5) and (B1)–(B4) hold. Let $\theta_{0,1} > 0$, $q \geq 5$, and $L \geq L_0$. If $\gamma > 1/2 + \beta/8$ and $V_{\theta_0,m_0} := P_{\theta_0,m_0}(\ell_{\theta_0,m_0} S^\top_{\theta_0,m_0})$ is a nonsingular matrix in $\mathbb{R}^{(d-1)\times(d-1)}$, then

$$\sqrt{n}(\check{\theta} - \theta_0) \xrightarrow{d} N(0, H_{\theta_0} V^{-1}_{\theta_0,m_0} I_{\theta_0,m_0} (H_{\theta_0} V^{-1}_{\theta_0,m_0})^\top), \tag{18}$$

where $I_{\theta_0,m_0} := P_{\theta_0,m_0}(\ell_{\theta_0,m_0}\ell^\top_{\theta_0,m_0})$. Further, if $\sigma^2(\cdot) \equiv \sigma^2$, then $V_{\theta_0,m_0} = I_{\theta_0,m_0}$ and

$$\sqrt{n}(\check{\theta} - \theta_0) \xrightarrow{d} N(0, \sigma^4 H_{\theta_0} I^{-1}_{\theta_0,m_0} H^\top_{\theta_0}).$$

Remark 4.1. If m_0 is twice continuously differentiable, then $\gamma = 1$. Hence, $\gamma > 1/2 + \beta/8$ is equivalent to assuming $\beta \in [0, 4)$. Note that $\beta > 0$ allows for covariate distributions for which the density of $\theta_0^\top X$ can go to zero. In Theorem 4.1, to keep notations in the proof simple, we assume that $q \geq 5$. However, by using Remark 3.5, this condition can be weakened to $q \geq 4$. In Section S.3, we show that the limiting variances in Theorem 4.1 are unique and do not depend on the particular choice of $\theta \mapsto H_\theta$.

Sketch of the proof. The proof follows along the lines of (Bolthausen, Perkins, and van der Vaart 2002, theor. 6.20). The main novelty in the proof is a new mechanism to verify that the estimator satisfies the score Equation (15). However to simplify the algebra involved,[4] we will work with

$$\psi_{\theta,m}(x,y) := (y - m(\theta^\top x))m'(\theta^\top x)H_\theta^\top[x - h_{\theta_0}(\theta^\top x)], \tag{19}$$

a slight modification of $\ell_{\theta,m}$. The only difference between $\ell_{\theta,m}$ and $\psi_{\theta,m}$ is the last term $(h_\theta(\theta^\top X))$. In Section S.2 of the supplementary file, we show that

$$\sqrt{n}\,\mathbb{P}_n \psi_{\check{\theta},\check{m}} = o_p(1), \tag{20}$$

implies

$$\sqrt{n} V_{\theta_0,m_0} H^\top_{\theta_0}(\check{\theta} - \theta_0) = \mathbb{G}_n \psi_{\theta_0,m_0} + o_p(1 + \sqrt{n}|\check{\theta} - \theta_0|). \tag{21}$$

The conclusion of the proof follows by observing that $\psi_{\theta_0,m_0} = \ell_{\theta_0,m_0}$. We will now give a brief sketch of the proof of Equation (20). Define for every (m,θ), $\eta \in \mathbb{R}^{d-1}$, $a : D \to \mathbb{R}$, and $t \in \mathbb{R}$,

$$\zeta_t(\theta,\eta) := \sqrt{1 - t^2|\eta|^2}\,\theta + tH_\theta\eta \qquad \text{and}$$
$$\xi_t(u; a, m) := \Pi_{\mathcal{M}_L}(m - ta)(u).$$

Observe that $(\check{m}, \check{\theta})$ is the minimizer of $(m,\theta) \mapsto Q_n(m,\theta)$ and $t \mapsto (\zeta_t(\check{\theta},\eta), \xi_t(u; a, \check{m}))$ is a valid path in $\mathcal{M}_L \times \Theta$ through $(\check{\theta}, \check{m})$. Thus, $t = 0$ is the minimizer of $t \mapsto Q_n(\zeta_t(\check{\theta},\eta), \xi_t(\cdot; a, \check{m}))$ for every $\eta \in \mathbb{R}^{d-1}$ and $a : D \to \mathbb{R}$. Hence, if $t \mapsto Q_n(\zeta_t(\check{\theta},\eta), \xi_t(\cdot; a, \check{m}))$ is differentiable then

$$\frac{\partial}{\partial t} Q_n(\zeta_t(\check{\theta},\eta), \xi_t(\cdot; a, \check{m}))\Big|_{t=0} = 0.$$

Furthermore, if functions $a_1, a_2, \ldots, a_K$ (for some $K \geq 1$) are such that $t \mapsto Q_n(\zeta_t(\check{\theta},\eta), \xi_t(\cdot; a_j, \check{m}))$ is differentiable for all $1 \leq j \leq K$, then

$$\sum_{j=1}^{K} \alpha_j \frac{\partial}{\partial t} Q_n(\zeta_t(\check{\theta},\eta), \xi_t(\cdot; a_j, \check{m}))\Big|_{t=0} = 0,$$

for any $\alpha_1, \ldots, \alpha_K \in \mathbb{R}$. Note that the proof of Equation (20) will be complete, if we can show that for every $\eta \in S^{d-2}$, then there exist a $K \geq 1$ and functions $a_j : D \to \mathbb{R}, 1 \leq j \leq K$ such that $t \mapsto \Pi_{\mathcal{M}_L}(\check{m} - ta_j)(u)$ is differentiable and

$$\eta^\top \mathbb{P}_n \psi_{\check{\theta},\check{m}} = \sum_{j=1}^{K} \alpha_j \frac{\partial}{\partial t} Q_n(\zeta_t(\check{\theta},\eta), \xi_t(\cdot; a_j, \check{m}))\Big|_{t=0} + o_p(n^{-1/2}). \tag{22}$$

This means that it is enough to consider the approximation of $\eta^\top \mathbb{P}_n \psi_{\check{\theta},\check{m}}$ by the linear closure of $\{\partial Q_n(\zeta_t(\check{\theta},\eta), \xi_t(\cdot; a, \check{m}))/\partial t|_{t=0} : t \mapsto Q_n(\zeta_t(\check{\theta},\eta), \xi_t(\cdot; a, \check{m}))$ is differentiable at $t = 0\}$. Instead of fully characterizing the linear closure set, we find a large enough subset that suffices for our purpose using the following steps.

1. We find a set of perturbations a such that $t \mapsto \xi_t(\cdot; a, m)$ is differentiable. Recall $\mathcal{X}_{\check{m}}$ defined in Equation (17). In Lemma S.10.2 (stated and proved in the supplementary file), we show that $\mathcal{X}_{\check{m}} \subseteq \{a : D \to \mathbb{R} \mid t \mapsto \xi_t(\cdot; a, \check{m})$ is differentiable at $t = 0\}$.
2. For every such $a \in \mathcal{X}_{\check{m}}$, in Lemma S.10.3, we show that

$$-\frac{1}{2}\frac{\partial}{\partial t} Q_n(\zeta_t(\check{\theta},\eta), \xi_t(\cdot; a, \check{m}))\Big|_{t=0} = \mathbb{P}_n\Big[(y - \check{m}(\check{\theta}^\top x))\Big\{\eta^\top \check{m}'(\check{\theta}^\top x)H^\top_{\check{\theta}} x - a(\check{\theta}^\top x)\Big\}\Big].$$

[4] All the proofs will go through with $\ell_{\theta,m}$ instead of $\psi_{\theta,m}$. However, usage of $\ell_{\theta,m}$ will require more remainder terms to be controlled and thus will lead to more tedious proofs.

Thus, to prove Equation (22), it is enough to show that

$$\inf_{a\in\overline{\operatorname{lin}}(\mathcal{X}_{\check{m}})}\Big|\eta^\top\mathbb{P}_n\psi_{\check{\theta},\check{m}}- \mathbb{P}_n\big[(y-\check{m}(\check{\theta}^\top x))\{\eta^\top\check{m}'(\check{\theta}^\top x)H_{\check{\theta}}^\top x-a(\check{\theta}^\top x)\}\big]\Big| = o_p(n^{-1/2}),$$

where $\psi_{\theta,m}$ is defined in Equation (19). In more general constraint spaces, one might need to use the generality of $\overline{\operatorname{lin}}(\mathcal{X}_{\check{m}})$ but in our case, it suffices to work with $\mathcal{X}_{\check{m}}$; see Theorem S.10.1. □

Remark 4.2 (Efficiency under heteroscedasticity). It is important to note that Equation (13), the efficient score, depends on $\sigma^2(\cdot)$. Without additional assumptions, estimators of $\sigma^2(\cdot)$ will have poor finite sample performance (especially if d is large) which in turn will lead to poor finite sample performance of the weighted LSE; see (Tsiatis 2006, pp. 93–95).

Remark 4.3 (Efficiency under additional shape constraints). As discussed in Remark 3.6, it might be the case that the practitioner is interested in imposing additional shape constraints such as monotonicity, unimodality, or k-monotonicity (in addition to convexity). If m_0 satisfies these constraints in a strict sense (i.e., m_0 is strictly monotone or k-monotone), then the discussion in Section 4.1 implies that the efficient score (at the truth) is still (13) even under the additional shape constraints. This is true, because $\overline{\operatorname{lin}}\,\mathcal{D}_{m_0} = L_2(\Lambda)$ even under these additional shape constraints on link functions, as m_0 does not lie on the "boundary" of the parameter space. In fact, under these additional constraints, the proof of Theorem 4.1 can be used with minor modifications to show that CLSE of θ_0 satisfies Equation (18).

To further illustrate the usefulness of our new approach, we discuss the proof of semiparametric efficiency in the Cox proportional hazards model under current status censoring (Huang 1996; Bolthausen, Perkins, and van der Vaart 2002).

Example 4.1 (Cox proportional hazards model with current status data). Suppose that we observe a random sample of size n from the distribution of $X = (C, \Delta, Z)$, where $\Delta = 1\{T \le C\}$, such that the survival time T and the observation time C are independent given $Z \in \mathbb{R}^d$, and that T follows a Cox proportional hazards model with parameter θ_0 and cumulative hazard function Λ_0; see, for example, Huang (1996, sec. 2) for a more detailed discussion of this model. Huang (1996) showed that $\hat{\Lambda}$, the nonparametric maximum likelihood estimator (NPMLE) of Λ_0, is a right-continuous step function with possible discontinuities only at $C_1, \dots, C_n$ (the observed censoring/inspection times). Huang (1996) also proved that $\hat{\theta}$ (the NPMLE for θ_0) is an efficient estimator for θ_0. However, just as in the single-index model, the proof of efficiency is complicated due to the fact that $s \mapsto \hat{\Lambda} + sh$ will not necessarily be a valid hazard function for every smooth $h(\cdot)$.[5] To establish Equation (15) for the above model (Huang 1996, pp. 563 and 564), "guesses" an approximately least favorable path (also see Bolthausen, Perkins, and van der Vaart 2002, pp. 439–441). However, using the arguments above we can easily see that $s \mapsto \Pi(\hat{\Lambda} + sh)$ is differentiable if h is a piecewise constant function with possible discontinuities only at the points of discontinuities of $\hat{\Lambda}$. Then using the property that $\|\hat{\Lambda} - \Lambda_0\| = o_p(n^{-1/3})$, one can establish a result similar to Equation (16). A similar strategy can be used to establish efficiency in the current status regression model in Murphy, van der Vaart, and Wellner (1999).

4.3. Construction of Confidence Sets and Validating the Asymptotics

Theorem 4.1 shows that when the errors happen to be homoscedastic the CLSE of θ_0 is $\sqrt{n}$-consistent and asymptotically normal with covariance matrix:

$$\Sigma^0 := \sigma^4 H_{\theta_0} P_{\theta_0,m_0}[\ell_{\theta_0,m_0}(Y,X)\ell_{\theta_0,m_0}^\top(Y,X)]^{-1}H_{\theta_0}^\top, \qquad (23)$$

where ℓ_{θ_0,m_0} is defined in Equation (14). This result can be used to construct confidence sets for θ_0. However since Σ^0 is unknown, we propose using the following plug-in estimator of Σ^0:

$$\check{\Sigma} := \check{\sigma}^4 H_{\check{\theta}}\big[\mathbb{P}_n\big(\ell_{\check{\theta},\check{m}}(Y,X)\ell_{\check{\theta},\check{m}}^\top(Y,X)\big)\big]^{-1}H_{\check{\theta}}^\top,$$

where $\check{\sigma}^2 := \sum_{i=1}^n [Y_i - \check{m}(\check{\theta}^\top X_i)]^2/n$. Note that Theorems 3.4 and 3.5 imply consistency of $\check{\Sigma}$.

For example one can construct the following $1 - 2\alpha$ confidence interval for $\theta_{0,i}$:

$$\left[\max\left\{-1, \check{\theta}_i - \frac{z_\alpha}{\sqrt{n}}\left(\check{\Sigma}_{i,i}\right)^{1/2}\right\}, \min\left\{1, \check{\theta}_i + \frac{z_\alpha}{\sqrt{n}}\left(\check{\Sigma}_{i,i}\right)^{1/2}\right\}\right], \qquad (24)$$

where z_α denotes the upper αth quantile of the standard normal distribution. The truncation guarantees that confidence interval is a subset of the parameter set.

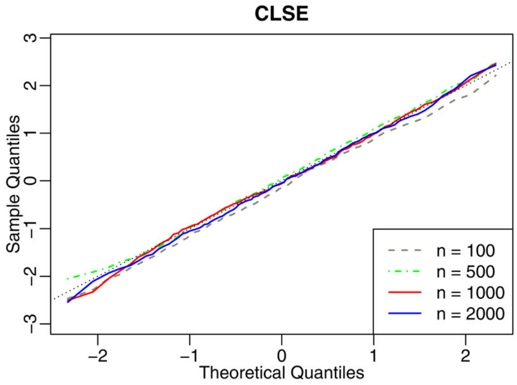
We now give an illustrative simulation example. We generate n iid observations from the model: $Y = (\theta_0^\top X)^2 + N(0, .3^2)$, where $X \sim \text{Uniform}[-1, 1]^3$ and $\theta_0 = (1, 1, 1)/\sqrt{3}$, for n increasing from 50 to 1000. For the above model, $\Sigma_{1,1}^0$ is 0.22.[6] In the left panel of Figure 2, we present the Q–Q plot of $\sqrt{n}[\Sigma_{1,1}^0]^{-1/2}(\check{\theta}_1 - \theta_{0,1})$ based on 800 replications; on the x-axis we have the quantiles of the standard normal distribution. The Q–Q plot validates the asymptotic normality and shows that the sample variance of the CLSE converges to the limiting variance found in Theorem 4.1. In the right panel of Figure 2, we present empirical coverages (from 800 replications) of 95% confidence intervals based on the CLSE constructed via Equation (24).

5. Simulation Study

In Section S.1 of the supplementary file, we develop an alternating minimization algorithm to compute the CLSE (2). In this section we illustrate the finite sample performance of the CLSE using the implementation in the R package `simest`. We

[5] $\hat{\Lambda} + sh$ is not guaranteed to be monotone as $\hat{\Lambda}$ is a nondecreasing piecewise constant function and not strictly increasing.

[6] To compute the limiting variance in Equation (23), we used a Monte Carlo approximation of $P_{\theta_0,m_0}[\ell_{\theta_0,m_0}(Y,X)\ell_{\theta_0,m_0}^\top(Y,X)]$ with sample size 2×10^5 and true (m_0, θ_0, P_X). The limiting covariance matrix $\Sigma^0 = 0.33I_3 - 0.11J_3$, where I_3 is the 3 × 3 identity matrix and J_3 is the 3 × 3 matrix of all ones.



n	CLSE	
	Coverage	Avg Length
50	0.92	0.30
100	0.91	0.18
200	0.92	0.13
500	0.94	0.08
1000	0.93	0.06

Figure 2. Summary of $\check{\theta}$ (over 800 replications) based on n iid observations from the model 4.3. Left panel: Q-Q plots for $\sqrt{n}\left[\Sigma^0_{1,1}\right]^{-1/2}(\check{\theta}_1 - \theta_{0,1})$ for $n \in \{100, 500, 1000, 2000\}$. The dotted black line corresponds to the $y = x$ line; right panel: estimated coverage probabilities and average lengths of nominal 95% confidence intervals for the first coordinate of θ_0.

also compare its performance with other existing estimators, namely the `EFM` estimator (the estimating function method; see Cui, Härdle, and Zhu 2011), the `EDR` estimator (effective dimension reduction; see Hristache, Juditsky, and Spokoiny 2001), and the estimator proposed in Kuchibhotla and Patra (2020) with the tuning parameter chosen by generalized cross-validation (Kuchibhotla and Patra 2020; we denote this estimator by `Smooth`). We use `CvxLip` to denote the CLSE.

5.1. Another Convex Constrained Estimator

Alongside these existing estimators, we also numerically study another natural estimator under the convexity shape constraint—the convex LSE—denoted by `CvxLSE` below. This estimator is obtained by minimizing the sum of squared errors subject to only the convexity constraint. Formally, the `CvxLSE` is

$$(m^{\dagger}_n, \theta^{\dagger}_n) := \underset{(m,\theta)\in\mathcal{C}\times\Theta}{\arg\min}\ Q_n(m, \theta) \qquad (25)$$

The computation of `CvxLSE` is discussed in Remark S.12 and is implemented in the R package `simest`. However, theoretical analysis of this estimator is difficult because of various reasons; see Section S.14 of the supplementary file for a brief discussion. In our simulation studies, we observe that the performance of `CvxLSE` is very similar to that of `CvxLip`.

In what follows, we will use $(\tilde{m}, \tilde{\theta})$ to denote a generic estimator that will help us describe the quantities in the plots; e.g., we use $||\tilde{m} \circ \tilde{\theta} - m_0 \circ \theta_0||_n = [\frac{1}{n}\sum_{i=1}^n(\tilde{m}(\tilde{\theta}^\top x_i) - m_0(\theta_0^\top x_i))^2]^{1/2}$ to denote the in-sample root-mean-squared estimation error of $(\tilde{m}, \tilde{\theta})$, for all the estimators considered. From the simulation study, it is easy to conclude that the proposed estimators have superior finite sample performance in the most sampling scenarios considered.

5.2. Increasing Dimension

To illustrate the behavior/performance of the estimators as d grows, we consider the following single-index model $Y = (\theta_0^\top X)^2 + t_6$, where $\theta_0 = (2, 1, \mathbf{0}_{d-2})^\top/\sqrt{5}$ and $X \in \mathbb{R}^d \sim \text{Uniform}[-1, 5]^d$, where t_6 denotes the Student's t-distribution with 6 degrees of freedom. In each replication, we observe $n = 100$ iid observations from the model. It is easy to see that the performance of all the estimators worsen as the dimension increases from 10 to 100 and `EDR` has the worst overall performance; see Figure 3. However, when $d = 100$, the convex constrained estimators have significantly better performance. This simulation scenario is similar to the one considered in Example 3 of (Cui, Härdle, and Zhu 2011, sec. 3.2).

5.3. Choice of L

In this subsection, we consider a simple simulation experiment to demonstrate that the finite sample performance of the CLSE is robust to the choice of tuning parameter. We generate an iid sample (of size $n = 500$) from the following model:

$$\begin{aligned} &Y = (\theta_0^\top X)^2 + N(0, .1^2), \quad \text{where} \\ &X \sim \text{Uniform}[-1, 1]^4 \text{ and} \\ &\theta_0 = (1, 1, 1, 1)^\top/2. \end{aligned} \qquad (26)$$

Observe that, we have $-2 \leq \theta^\top X \leq 2$ and $L_0 := \sup_{t\in[-2,2]} m_0'(t) = 4$ as $m_0(t) = t^2$. To understand the effect of L on the performance of the CLSE, we show the boxplot of $\sum_{i=1}^4 |\check{\theta}_i - \theta_{0,i}|/4$ as L varies from 3 $(< L_0)$ to 10 in Figure 4. Figure 4 also includes the `CvxLSE` which corresponds to $L = \infty$. The plot clearly show that the performance of `CvxLip` is not significantly affected by the particular choice of the tuning parameter. The observed robustness in the behavior of the estimators can be attributed to the stability endowed by the convexity constraint.

6. Real Data Analysis

In this following, we analyze the two real datasets discussed in Examples 1.1 and 1.2.

6.1. Boston Housing Data

We briefly recall the discussion in Example 1.1. The Boston housing dataset was collected by Harrison and Rubinfeld (1978)

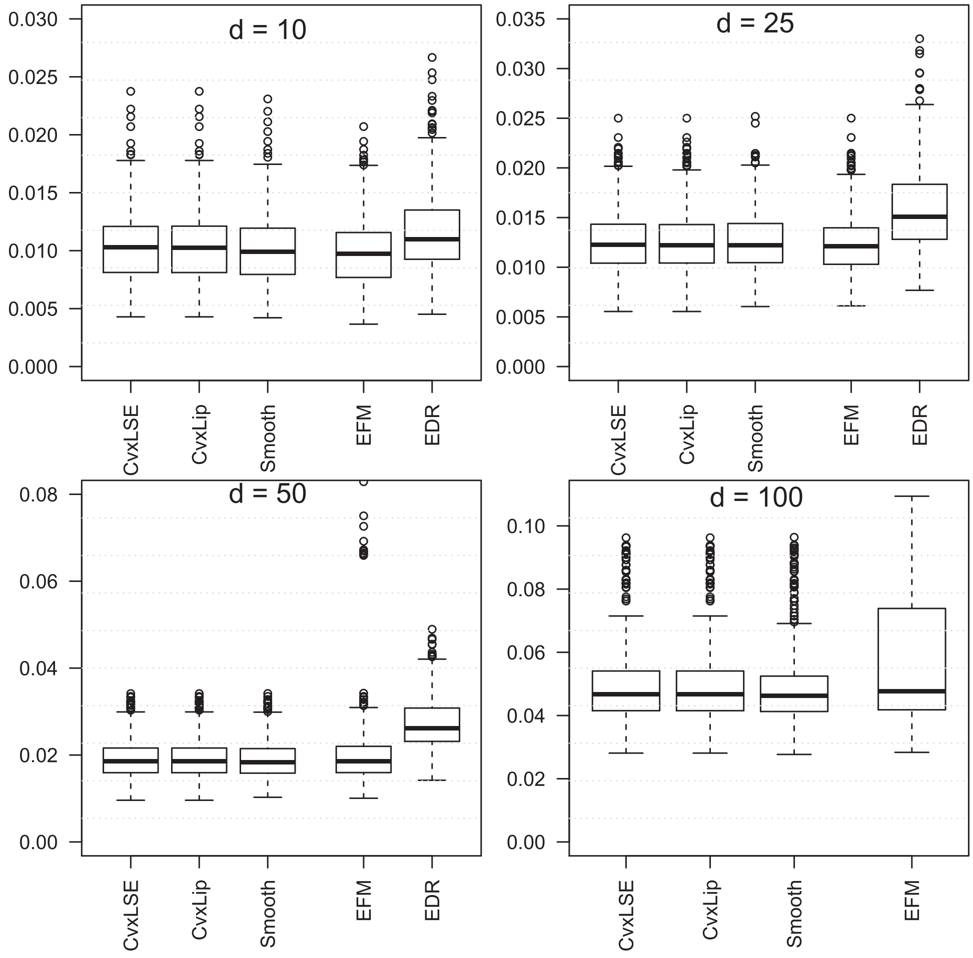


Figure 3. Boxplots of $\sum_{i=1}^{d} |\tilde{\theta}_i - \theta_{0,i}|/d$ (over 500 replications) based on 100 observations from the simulation setting in Section 5.2 for dimensions 10, 25, 50, and 100, shown in the top-left, the top-right, the bottom-left, and the bottom-right panels, respectively. The bottom-right panel does not include EDR as the R-package `EDR` does not allow for $d = 100$.

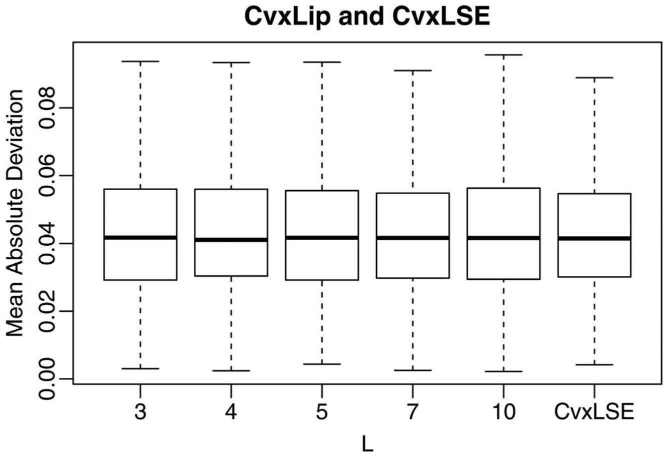


Figure 4. Boxplots of $\frac{1}{4}\sum_{i=1}^{4} |\tilde{\theta}_i - \theta_{0,i}|$ (over 1000 replications) for the model (26) ($d = 4$ and $n = 500$) `CvxLip` for $L = \{3, 4, 5, 7, 10\}$ and `CvxLSE` (i.e., $L = \infty$).

to study the effect of different covariates on the real estate price in the greater Boston area. The dependent variable Y is the log-median value of homes in each of the 506 census tracts in the Boston standard metropolitan area. Harrison and Rubinfeld (1978) observed 13 covariates and fit a linear model after taking log transformation for 3 covariates and power transformations for three other covariates; also see Wang et al. (2010) for a discussion of this dataset.

Breiman and Friedman (1985) did further analysis to deal with multi-collinearity of the covariates and selected four variables using a penalized stepwise method. The chosen covariates were average number of rooms per dwelling (RM), full-value property-tax rate per 10, 000 U.S.D (TAX), pupil–teacher ratio by town school district (PT), and proportion of population that is of "lower (economic) status" in percentage points (LS). Following Wang and Yang (2009) and Yu, Mammen, and Park (2011), we take logarithms of LS and TAX to reduce sparse areas in the dataset. Furthermore, we have scaled and centered each of the covariates to have mean 0 and variance 1. Wang and Yang (2009) fit a nonparametric additive regression model to the selected variables and obtained an R^2 (the coefficient of determination) of 0.64. Wang et al. (2010) fit a single-index model to this data using the set of covariates suggested in Chen and Li (1998). In Gu and Yang (2015), the authors created 95% uniform confidence band for the link function and reject the null hypothesis that the link function is linear. In both Gu and Yang (2015) and Wang et al. (2010), the fitted link function is approximately nondecreasing and convex; see (Wang et al. 2010, fig. 2) and (Gu and Yang 2015, fig. 5). This motivates us to fit a *nondecreasing* and convex single-index model to the

Boston housing dataset. In particular, we consider the following estimator:

$$(\hat{m}_L, \hat{\theta}_L) := \underset{(m,\theta)\in\mathcal{M}_L\cap\mathcal{N}\times\Theta}{\arg\min}\ Q_n(m,\theta) \tag{27}$$

where $\mathcal{N}$ is the set of real-valued nondecreasing functions on D. Following the discussions in Remarks 3.6 and 4.3, we observe that the results in this article also hold for $(\hat{m}_L, \hat{\theta}_L)$. The computation of the CLSE under the additional monotonicity constraint is discussed in Remark S.1.1 and implemented in the accompanying R package.

We summarize our results in Table 1. We call $(\hat{m}_L, \hat{\theta}_L)$, the `MonotoneCLSE`. In Figure 5, we plot the scatterplot of $\{(\hat{\theta}_L^\top X_i, Y_i)\}_{i=1}^{506}$ overlaid with the plot of $\hat{m}_L(\cdot)$ and the regression spline-based estimator of Wang and Yang (2009). For `MonotoneCLSE` and `CvxLip`, we chose $L = 30$ (an arbitrary but large number). We also observe that the R^2 for the monotonicity and convexity constrained (`MonotoneCLSE`) and just convexity constrained single-index models (`CvxLip` and `CvxLSE`), when using all the available covariates, is approximately 0.80. To further understand the predictive properties of the estimators under different smoothness and shape constraints, in Table 1, we report the 5-fold cross-validation error averaged over 100 random partitions. The large cross-validation error for the `CvxLSE` is due to over-fitting of $m_n^\dagger$ at the boundary of its support; see Figure S.1 (supplementary material) for an illustration of this boundary effect.

6.2. Car Mileage Data

First, we briefly recall the discussion in Example 1.2. We consider the car mileage dataset of Donoho and Ramos (1983) for a second application for the convex single-index model. We model the mileage (Y) of 392 cars using the covariates (X): displacement (Ds), weight (W), acceleration (A), and horsepower (H). Cheng, Zhao, and Li (2012) fit a partial linear model to this this dataset, while Kuchibhotla and Patra (2020) fit a single-index model (without any shape constraint). The "law of diminishing returns" suggests m_0 should be convex and nonincreasing. However, the estimators based only on smoothness assumptions satisfy these shape constraints only approximately. In the right panel of Figure 5, we fit a convex and nonincreasing single-index model.

We have scaled and centered each of covariates to have mean 0 and variance 1 for our analysis, just as in Section 6.1. We performed a test of significance for θ_0 using the plug-in variance estimate in Section 4.3. The covariates A, Ds, and H were found to be significant and each of them had p-value less than 10^{-5}. In the right panel of Figure 5, we have the scatterplot of $\{(\hat{\theta}_L^\top X_i, Y_i)\}_{i=1}^{392}$ overlaid with the plot of $\hat{m}_L(\cdot)$ and regression spline-based estimator obtained in Wang and Yang (2009); here $\hat{\theta}_L$ is defined as in Equation (27) but $\mathcal{N}$ now denotes the class of real-valued *nonincreasing* functions on D. Table 1 lists different estimators for θ_0 and their respective R^2 and cross-validation errors.

Table 1. Estimates of θ_0 and generalized R^2 for the datasets in Sections 6.1 and 6.2.

Method	Boston data						Car mileage data					
	RM	log(TAX)	PT	log(LS)	R^2	CV-error	Ds	W	A	H	R^2	CV-error
`LM`[a]	2.34	−0.37	−1.55	−5.11	0.73	20.75	−0.63	−4.49	−0.06	−1.68	0.71	18.61
`Smooth`	0.44	−0.18	−0.27	−0.83	0.77	17.80	0.42	0.18	0.11	0.88	0.76	15.29
`MonotoneCLSE`	0.49	−0.21	−0.25	−0.81	0.80	17.93	0.44	0.17	0.13	0.87	0.76	15.34
`CvxLip`	0.48	−0.23	−0.26	−0.80	0.80	17.93	0.44	0.18	0.12	0.87	0.76	15.22
`CvxLSE`	0.43	−0.20	−0.28	−0.84	0.80	21.44	0.39	0.14	0.12	0.90	0.77	16.38
`EFM`	0.48	−0.19	−0.21	−0.83	–	–	0.44	0.18	0.13	0.87	–	–
`EDR`	0.44	−0.14	−0.18	−0.87	–	–	0.33	0.11	0.15	0.93	–	–

[a] `LM` denotes the linear regression model.

`EFM` and `EDR` do not provide a function estimator and hence we do not show an R^2 value. CV-error denotes out of 5-fold cross-validation averaged over 100 random partitions.

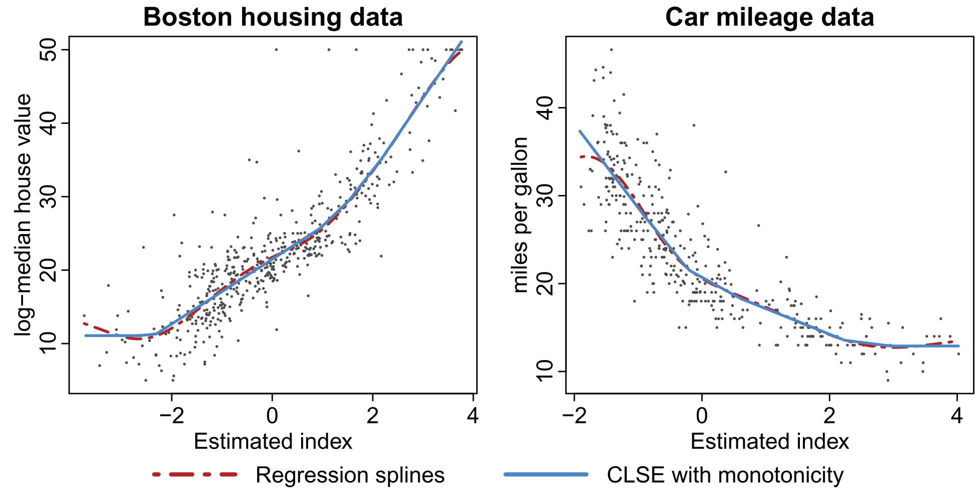


Figure 5. Scatterplots of $\{(X_i^\top \check{\theta}, Y_i)\}_{i=1}^n$ overlaid with the plots of function estimates proposed in Wang and Yang (2009) (red, dot-dashed) and monotonicity constrained CLSE proposed in this article (blue, solid) for the two real datasets considered. Left panel: Boston housing data (Section 6.1), nondecreasing CLSE; right panel: the car mileage data (Section 6.2), nonincreasing CLSE.

7. Discussion

In this article we have proposed and studied a Lipschitz constrained LSE in the convex single-index model. Our estimator of the regression function is minimax rate optimal (Proposition S.6.1) and the estimator of the index parameter is semiparametrically efficient when the errors happen to be homoscedastic (Theorem 4.1). This work represents the first in the literature of semiparametric efficiency of the LSE when the nonparametric function estimator is non-smooth and parameters are bundled. Our proof of semiparametric efficiency is geometric and provides a general framework that can be used to prove efficiency of estimators in a wide variety of semiparametric models even when the estimators do not satisfy the efficient score equation directly; see sketch of proof of Theorem 4.1 and Example 4.1 in Section 4.2.

Theorem 3.1 proves the worst-case rate of convergence for the CLSE. It is well known in convex regression that if the true regression function is piecewise linear, then the LSE converges at a much faster (near parametric) rate Guntuboyina and Sen (2018). This behavior is called the *adaptation* property of the LSE. It is natural to wonder if such a property also holds for $\check{m} \circ \check{\theta}$. In Section S.4.3 of the supplementary file, we investigate the behavior of $\check{m} \circ \check{\theta}$ and $\check{\theta}$ (as sample size increases) when m_0 is piecewise linear. The simulation suggests that $\check{m} \circ \check{\theta}$ converges at a near parametric rate when m_0 is piecewise linear. However, a formal proof of this is beyond the scope of this article as it requires different techniques. Furthermore, the asymptotic behavior of $\check{\theta}$ in this setting is an open problem.

Supplementary Material

The supplementary material contains a detailed discussion of the alternating minimization algorithm, some additional simulations, and the proofs of all the results in the article.

Funding

This work was supported by NSF (grants nos. DMS-17-12822 and AST-16-14743).

References

Aït-Sahalia, Y., and Duarte, J. (2003), "Nonparametric Option Pricing Under Shape Restrictions," *Journal of Econometrics*, 116, 9–47. [272]

Alonso, A., and Brambila-Paz, F. (1998), "Lp-Continuity of Conditional Expectations," *Journal of Mathematical Analysis and Applications*, 221, 161–176. [278]

Balabdaoui, F., Durot, C., and Jankowski, H. (2019), "Least Squares Estimation in the Monotone Single Index Model," *Bernoulli*, 25, 3276–3310. [274,277]

Balabdaoui, F., Groeneboom, P., and Hendrickx, K. (2019), "Score Estimation in the Monotone Single-Index Model," *Scandinavian Journal of Statistics*, 46, 517–544. [274,276,277,278]

Bolthausen, E., Perkins, E., and van der Vaart, A. (2002), "Semiparametric Statistics," in *Lectures on Probability Theory and Statistics (Saint-Flour, 1999)* (Vol. 1781), Lecture Notes in Mathematics, ed. P. Bernard, Berlin: Springer, pp. 331–457. [273,275,279,280,281]

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American statistical Association*, 80, 580–598. [283]

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489. [274]

Chen, C.-H., and Li, K.-C. (1998), "Can SIR be as Popular as Multiple Linear Regression?" *Statistica Sinica*, 8, 289–316. [283]

Chen, Y., and Samworth, R. J. (2016), "Generalized Additive and Index Models With Shape Constraints," *Journal of the Royal Statistical Society*, Series B, 78, 729–754. [274,278]

Cheng, G., Zhao, Y., and Li, B. (2012), "Empirical Likelihood Inferences for the Semiparametric Additive Isotonic Regression," *Journal of Multivariate Analysis*, 112, 172–182. [273,284]

Cui, X., Härdle, W. K., and Zhu, L. (2011), "The EFM Approach for Single-Index Models," *Annals of Statistics*, 39, 1658–1688. [272,274,277,278,282]

Delecroix, M., Hristache, M., and Patilea, V. (2006), "On Semiparametric M-estimation in Single-Index Regression," *Journal of Statistical Planning and Inference*, 136, 730–769. [272,274]

Dharanipragada, S., and Arun, K. (1996), "A Quadratically Convergent Algorithm for Convex-Set Constrained Signal Recovery," *IEEE Transactions on Signal Processing*, 44, 248–266. [275]

Donoho, D., and Ramos, E. (1983), "Cars Dataset–1983 ASA Data Exposition Dataset," available at: *http://lib.stat.cmu.edu/datasets/cars.data*. [273,275,284]

Fitzpatrick, S., and Phelps, R. R. (1982), "Differentiability of the Metric Projection in Hilbert Space," *Transactions of the American Mathematical Society*, 270, 483–501. [275]

Gaïffas, S., and Lecué, G. (2007), "Optimal Rates and Adaptation in the Single-Index Model Using Aggregation," *Electronic Journal of Statistics*, 1, 538–573. [274,278]

Ganti, R., Rao, N., Willett, R. M., and Nowak, R. (2015), "Learning Single Index Models in High Dimensions," arXiv:1506.08910. [274]

Groeneboom, P., and Hendrickx, K. (2018), "Current Status Linear Regression," *The Annals of Statistics*, 46, 1415–1444. [278]

Gu, L., and Yang, L. (2015), "Oracally Efficient Estimation for Single-Index Link Function With Simultaneous Confidence Band," *Electronic Journal of Statistics*, 9, 1540–1561. [272,283]

Guntuboyina, A., and Sen, B. (2018), "Nonparametric Shape-Restricted Regression," *Statistical Science*, 33, 568–594. [278,285]

Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single-Index Models," *Annals of Statistics*, 21, 157–178. [272,274]

Harrison, D., and Rubinfeld, D. L. (1978), "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, 5, 81–102. [272,275,282,283]

Horowitz, J. L. (1998), *Semiparametric Methods in Econometrics* (Vol. 131). Lecture Notes in Statistics, New York: Springer-Verlag. [276]

Hristache, M., Juditsky, A., and Spokoiny, V. (2001), "Direct Estimation of the Index Coefficient in a Single-Index Model," *Annals of Statistics*, 29, 595–623. [272,274,276,282]

Huang, J. (1996), "Efficient Estimation for the Proportional Hazards Model With Interval Censoring," *Annals of Statistics*, 24, 540–568. [273,274,275,281]

Huang, J., and Wellner, J. A. (1997), "Interval Censored Survival Data: A Review of Recent Progress," in *Proceedings of the First Seattle Symposium in Biostatistics*, pp. 123–169. [274]

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58, 71–120. [272,274,276,277,278]

Kakade, S. M., Kanade, V., Shamir, O., and Kalai, A. (2011), "Efficient Learning of Generalized Linear and Single Index Models With Isotonic Regression," in *Advances in Neural Information Processing Systems*, eds. J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, Red Hook, NY: Curran Associates, Inc., pp. 927–935. [274]

Kalai, A. T., and Sastry, R. (2009), "The Isotron Algorithm: High-Dimensional Isotonic Regression," in 22nd Conference on Learning Theory, Montreal, Quebec, Canada. [274]

Keshavarz, A., Wang, Y., and Boyd, S. (2011), "Imputing a Convex Objective Function," in 2011 IEEE International Symposium on Intelligent Control, Denver, CO, pp. 613–619. [272]

Kuchibhotla, A. K., and Patra, R. K. (2020), "Efficient Estimation in Single Index Models Through Smoothing Splines," *Bernoulli*, 26, 1587–1618. [272,273,277,279,282,284]

Lepski, O. V., and Spokoiny, V. G. (1997), "Optimal Pointwise Adaptive Methods in Nonparametric Estimation," *The Annals of Statistics*, 2512–2546. [277]

Li, K.-C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *Annals of Statistics*, 17, 1009–1052. [272,274]

Li, Q., and Racine, J. S. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton, NJ: Princeton University Press. . [272,276]

Lim, E. (2014), "On Convergence Rates of Convex Regression in Multiple Dimensions," *INFORMS Journal on Computing*, 26, 616–628. [274]

Lin, W., and Kulasekera, K. B. (2007), "Identifiability of Single-Index Models and Additive-Index Models," *Biometrika*, 94, 496–501. [276]

Ma, Y., and Zhu, L. (2013), "Doubly Robust and Efficient Estimators for Heteroscedastic Partially Linear Single-Index Models Allowing High Dimensional Covariates," *Journal of the Royal Statistical Society*, Series B, 75, 305–322. [279]

Matzkin, R. L. (1991), "Semiparametric Estimation of Monotone and Concave Utility Functions for Polychotomous Choice Models," *Econometrica*, 59, 1315–1327. [272]

Mazumder, R., Choudhury, A., Iyengar, G., and Sen, B. (2019), "A Computational Framework for Multivariate Convex Regression and Its Variants," *Journal of the American Statistical Association*, 114, 318–331. [274]

McCormick, G., and Tapia, R. (1972), "The Gradient Projection Method Under Mild Differentiability Conditions," *SIAM Journal on Control*, 10, 93–98. [275]

Murphy, S. A., van der Vaart, A. W., and Wellner, J. A. (1999), "Current Status Regression," *Mathematical Methods of Statistics*, 8, 407–425. [273,274,275,277,278,279,280,281]

Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135. [279]

Newey, W. K., and Stoker, T. M. (1993), "Efficiency of Weighted Average Derivative Estimators and Index Models," *Econometrica*, 61, 1199–1223. [275,279]

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430. [272,274,277]

Shapiro, A. (1994), "Existence and Differentiability of Metric Projections in Hilbert Spaces," *SIAM Journal on Optimization*, 4, 130–141. [275]

Sokolowski, J., and Zolesio, J.-P. (1992), "Shape Sensitivity Analysis of Variational Inequalities," in *Introduction to Shape Optimization*, eds. R. L. Graham, J. Stoer, and R. Varga, Berlin: Springer, pp. 163–239. [275]

Song, K. (2014), "Semiparametric Models With Single-Index Nuisance Parameters," *Journal of Econometrics*, 178, 471–483. [278]

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, Springer Series in Statistics, New York: Springer. [281]

Varian, H. R. (1984), "The Nonparametric Approach to Production Analysis," *Econometrica*, 52, 579–597. [272]

Wang, G., and Wang, L. (2015), "Spline Estimation and Variable Selection for Single-Index Prediction Models With Diverging Number of Index Parameters," *Journal of Statistical Planning and Inference*, 162, 1–19. [277,278]

Wang, J., and Yang, L. (2009), "Efficient and Fast Spline-Backfitted Kernel Smoothing of Additive Models," *Annals of Institute of Statistical Mathematics*, 61, 663–690. [272,273,274,277,278,283,284]

Wang, J.-L., Xue, L., Zhu, L., and Chong, Y. S. (2010), "Estimation for a Partial-Linear Single-Index Model," *Annals of Statistics*, 38, 246–274. [272,283]

Xia, Y., Tong, H., Li, W., and Zhu, L.-X. (2002), "An Adaptive Estimation of Dimension Reduction Space," *Journal of the Royal Statistical Society*, Series B, *64*(3), 363–410. [276]

Yu, K., Mammen, E., and Park, B. U. (2011), "Semi-Parametric Regression: Efficiency Gains From Modeling the Nonparametric Part," *Bernoulli*, 17, 736–748. [272,283]