EFFICIENT AND ADAPTIVE LINEAR REGRESSION IN SEMI-SUPERVISED SETTINGS

By Abhishek Chakrabortty and Tianxi Cai*

Harvard University

We consider the linear regression problem under semi-supervised settings wherein the available data typically consists of: (i) a small or moderate sized 'labeled' data, and (ii) a *much larger* sized 'unlabeled' data. Such data arises naturally from settings where the outcome, unlike the covariates, is expensive to obtain, a frequent scenario in modern studies involving large databases like electronic medical records (EMR). Supervised estimators like the ordinary least squares (OLS) estimator utilize only the labeled data. It is often of interest to investigate if and when the unlabeled data can be exploited to improve estimation of the regression parameter in the adopted linear model.

In this paper, we propose a class of 'Efficient and Adaptive Semi-Supervised Estimators' (EASE) to improve estimation efficiency. The EASE are two-step estimators adaptive to model mis-specification, leading to improved (optimal in some cases) efficiency under model mis-specification, and equal (optimal) efficiency under a linear model. This adaptive property, often unaddressed in the existing literature, is crucial for advocating 'safe' use of the unlabeled data. The construction of EASE primarily involves a flexible 'semi-non-parametric' imputation, including a smoothing step that works well even when the number of covariates is not small; and a follow up 'refitting' step along with a cross-validation (CV) strategy both of which have useful practical as well as theoretical implications towards addressing two important issues: under-smoothing and over-fitting. We establish asymptotic results including consistency, asymptotic normality and the adaptive properties of EASE. We also provide influence function expansions and a 'double' CV strategy for inference. The results are further validated through extensive simulations, followed by application to an EMR study on auto-immunity.

1. Introduction. In recent years, semi-supervised learning (SSL) has emerged as an exciting new area of research in statistics and machine learning. A detailed discussion on SSL including its practical relevance, the primary question of interest in SSL, and the existing relevant literature can be found in Chapelle, Schölkopf and Zien (2006) and Zhu (2008). A typi-

*This research was partially supported by National Institutes of Health grants R01 HL089778, U54 HG007963, and U01 HL121518.

Keywords and phrases: Semi-supervised Linear Regression, Semi-parametric Inference, Model Mis-specification, Adaptive Estimation, Semi-non-parametric Imputation.

cal semi-supervised (SS) setting is characterized by two types of available data: (i) a small or moderate sized 'labeled' data, $\mathcal{L}$, containing observations for both an outcome Y and a set of covariates $\mathbf{X}$ of interest, and (ii) an 'unlabeled' data, $\mathcal{U}$, of *much larger* size but having observations *only* for the covariates $\mathbf{X}$. By virtue of its large size, $\mathcal{U}$ essentially gives us the distribution of $\mathbf{X}$, denoted henceforth by $\mathbb{P}_{\mathbf{X}}$. Such a setting arises naturally whenever the covariates are easily available so that unlabeled data is plentiful, but the outcome is costly or difficult to obtain, thereby limiting the size of $\mathcal{L}$. This scenario is directly relevant to a variety of practical problems, especially in the modern 'big data' era, with massive unlabeled datasets (often electronically recorded) becoming increasingly available and tractable. A few familiar examples include machine learning problems like text mining, web page classification, speech recognition, natural language processing etc.

Among biomedical applications, a particularly interesting problem where SSL can be of great use is the statistical analysis of electronic medical records (EMR) data. Endowed with a wealth of de-identified clinical and phenotype data for large patient cohorts, EMR linked with bio-repositories are increasingly gaining popularity as rich resources of data for discovery research (Kohane, 2011). Such large scale datasets obtained in a cost-effective and timely manner are of great importance in modern medical research for addressing important questions such as the biological role of genetic variants in disease susceptibility and progression (Kohane, 2011). However, one major bottleneck impeding EMR driven research is the difficulty in obtaining validated phenotype information (Liao et al., 2010) since they are labor intensive or expensive to obtain. Thus, gold standard labels and genomic measurements are typically available only for a small subset nested within a large cohort. In contrast, digitally recorded data on the clinical variables are often available on all subjects, highlighting the necessity and utility of developing robust SSL methods that can leverage such rich source of auxiliary information to improve phenotype definition and estimation precision.

SSL primarily distinguishes from standard supervised methods by making use of $\mathcal{U}$, an information that is ignored by the latter. The ultimate question of interest in SSL is to investigate if and when the information on $\mathbb{P}_{\mathbf{X}}$ in $\mathcal{U}$ can be exploited to improve the efficiency over a given supervised approach. In recent years, several graph based non-parametric SSL approaches have been proposed (Zhu, 2005; Belkin, Niyogi and Sindhwani, 2006) for regression or classification. These approaches essentially target non-parametric SS estimation of $\mathbb{E}(Y\,|\,\mathbf{X})$ and therefore, for provable improvement guarantees, must rely implicitly or explicitly on assumptions relating $\mathbb{P}_{\mathbf{X}}$ to $\mathbb{P}_{Y|\mathbf{X}}$ (the conditional distribution of Y given $\mathbf{X}$), as duly noted and characterized more formally in Lafferty and Wasserman (2007). For non-parametric classification problems, the theoretical underpinnings of SSL including its scope and the consequences of using $\mathcal{U}$ have been also studied earlier by Castelli and Cover (1995, 1996). More parametric SS approaches, still aimed mostly at prediction, have also been studied for classification, including the 'generative model' approach (Nigam et al., 2000; Nigam, 2001) which is based on modeling the joint distribution of $(Y, \mathbf{X})$ as an identifiable mixture of parametric models, thereby implicitly relating $\mathbb{P}_{Y|\mathbf{X}}$ and $\mathbb{P}_{\mathbf{X}}$. However, these approaches depend strongly on the validity of the assumed mixture model, violation of which can actually *degrade* their performance compared to the supervised approach (Cozman and Cohen, 2001; Cozman, Cohen and Cirelo, 2003).

However SS *estimation problems*, especially from a semi-parametric point of view, has been somewhat less studied in SSL. Such problems are generally aimed at estimating some (finite-dimensional) parameter $\theta_0 \equiv \theta_0(\mathbb{P})$, where $\mathbb{P} = (\mathbb{P}_{Y|\mathbf{X}}, \mathbb{P}_{\mathbf{X}})$, and the key to the potential usefulness of $\mathcal{U}$ in improving estimation of θ_0 lies in understanding when $\theta_0(\mathbb{P})$ relates to $\mathbb{P}_{\mathbf{X}}$. For simple parameters like $\theta_0(\mathbb{P}) = \mathbb{E}(Y)$, unless $\mathbb{E}(Y \,|\, \mathbf{X})$ is a constant, θ_0 clearly depends on $\mathbb{P}_{\mathbf{X}}$ and hence, improved SS estimation is possible compared to the supervised estimator $\overline{Y}_{\mathcal{L}}$, the sample mean of Y based on $\mathcal{L}$. The situation is however more subtle for other choices of θ_0, especially those where θ_0 is the target parameter corresponding to an underlying parametric *working* model for $\mathbb{P}_{Y|\mathbf{X}}$. This includes the least squares parameter, as studied in this paper, targeted by a working linear model for $\mathbb{E}(Y \,|\, \mathbf{X})$. Such models are often adopted due to their appealing simplicity and interpretability.

In general, for such cases, if the adopted working model for $\mathbb{P}_{Y|\mathbf{X}}$ is correct and θ_0 is not related to $\mathbb{P}_{\mathbf{X}}$, then one *cannot* possibly gain through SSL by using the knowledge of $\mathbb{P}_{\mathbf{X}}$ (Zhang and Oles, 2000; Seeger, 2002). On the other hand, under model mis-specification, θ_0 may inherently *depend* on $\mathbb{P}_{\mathbf{X}}$, and thus imply the potential utility of $\mathcal{U}$ in improving the estimation. However, inappropriate use of $\mathcal{U}$ may lead to degradation of the estimation precision. This therefore signifies the need for *robust* and efficient SS estimators that are *adaptive* to model mis-specification, so that they are as efficient as the supervised estimator under the correct model and more efficient under model mis-specification. To the best of our knowledge, work done along these lines is relatively scarce in the SSL literature, one notable exception being the recent work of Kawakita and Kanamori (2013), where they use a very different approach based on density ratio estimation, building on the more restrictive approach of Sokolovska, Cappé and Yvon (2008). However, as we observe in our simulation studies, the extent of the efficiency gain actually achieved by these approaches can be quite incremental, at least in finite samples. Further, the seemingly unclear choice of the ideal (nuisance) model to be used for density ratio estimation can also have a significant impact on the performance, both finite sample and asymptotic, of these estimators.

We propose here a class of Efficient and Adaptive Semi-Supervised Estimators (EASE) in the context of linear regression problems. We essentially adopt a semi-parametric perspective wherein the adopted linear 'working' model can be potentially mis-specified, and the goal is to obtain efficient and adaptive SS estimators of the regression parameter through robust usage of $\mathcal{U}$. The EASE are two-step estimators with a simple and scalable construction based on a first step of 'semi-non-parametric' (SNP) imputation which includes a smoothing step and a follow-up 'refitting' step. In the second step, we regress the imputed outcomes against the covariates using the unlabeled data to obtain our SNP imputation based SS estimator, and then further combine it optimally with the supervised estimator to obtain the final EASE. Dimension reduction methods are also employed in the smoothing step to accommodate higher dimensional $\mathbf{X}$, if necessary. Further, we extensively adopt cross-validation (CV) techniques in the imputation, leading to some useful theoretical properties (apart from practical benefits) typically not observed for smoothing based two-step estimators. We demonstrate that EASE is guaranteed to be efficient and adaptive in the sense discussed above, and also achieves semi-parametric optimality whenever the SNP imputation is 'sufficient' or the linear model holds. We also provide data adaptive methods to optimally select the directions for smoothing when dimension reduction is desired, and tools for inference with EASE.

The rest of this paper is organized as follows. In Section 2, we formulate the SS linear regression problem. In Section 3, we construct a family of SS estimators based on SNP imputation and establish all their properties, and further propose the EASE as a refinement of these estimators. For all our proposed estimators, we also address their associated inference procedures based on 'double' CV methods. In Section 4, we discuss a kernel smoothing based implementation of the SNP imputation and establish all its properties. In Section 5, we discuss SS dimension reduction techniques, useful for implementing the SNP imputation. Simulation results and an application to an EMR study are shown in Section 6, followed by concluding discussions in Section 7. Proofs of all theoretical results, and further numerical results and discussions are distributed in the Appendix and Supplementary Material.

2. Problem Set-up.

Data Representation. Let $Y \in \mathbb{R}$ denote the outcome random variable and $\mathbf{X} \in \mathbb{R}^p$ denote the covariate vector, where p is fixed, and let $\mathbf{Z} = (Y, \mathbf{X}')'$.

Then the entire data available for analysis can be represented as $\mathbb{S} = (\mathcal{L} \cup \mathcal{U})$, where $\mathcal{L} = \{\mathbf{Z}_i \equiv (Y_i, \mathbf{X}_i')' : i = 1, \ldots, n\}$ consists of n independent and identically distributed (i.i.d.) observations from the joint distribution $\mathbb{P}_{\mathbf{Z}}$ of $\mathbf{Z}$, $\mathcal{U} = \{\mathbf{X}_i : i = n+1, \ldots, n+N\}$ consists of N i.i.d. observations from $\mathbb{P}_{\mathbf{X}}$, and $\mathcal{L} \perp\!\!\!\perp \mathcal{U}$. Throughout, for notational convenience, we use the subscript 'j' to denote the unlabeled observations, and re-index without loss of generality (w.l.o.g.) the N observations in $\mathcal{U}$ as: $\mathcal{U} = \{\mathbf{X}_j : j = n+1, \ldots, n+N\}$.

Assumption 2.1 (Basic Assumptions). (a) We assume that $\mathbf{Z}$ has finite 2^{nd} moments and $\mathbf{\Sigma} \equiv \text{Var}(\mathbf{X})$ is positive definite, denoted as $\mathbf{\Sigma} \succ 0$. We also assume, for simplicity, that $\mathbf{X}$ has a compact support $\mathcal{X} \subseteq \mathbb{R}^p$.

(b) We assume $N \gg n$ i.e. $n/N \to 0$ as $n, N \to \infty$, and $\mathcal{L}$ and $\mathcal{U}$ arise from the same underlying distribution, i.e. $\mathbf{Z} \sim \mathbb{P}_{\mathbf{Z}}$ for all subjects in $\mathcal{S}$.

Notations. Let $\mathbf{\Gamma} = \mathbb{E}(\overrightarrow{\mathbf{X}}\overrightarrow{\mathbf{X}}') \succ 0$, where $\forall\ \mathbf{v} \in \mathbb{R}^p$, $\overrightarrow{\mathbf{v}} = (1, \mathbf{v}')' \in \mathbb{R}^{(p+1)}$. Let $\mathcal{L}_2(\mathbb{P}_{\mathbf{X}})$ denote the space of all $\mathbb{R}$-valued measurable functions of $\mathbf{X}$ having finite L_2 norm with respect to (w.r.t.) $\mathbb{P}_{\mathbf{X}}$, and for any $g(.) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}})$, let $\mathbf{\Sigma}(g) \succ 0$ denote the $(p+1) \times (p+1)$ matrix $\mathbf{\Gamma}^{-1}\mathbb{E}[\overrightarrow{\mathbf{X}}\overrightarrow{\mathbf{X}}'\{Y - g(\mathbf{X})\}^2]\mathbf{\Gamma}^{-1}$. Lastly, let $\|\cdot\|$ denote the L_2 vector norm, and for any integer $a \geq 1$, let I_a denote the identity matrix of order a, and $\mathcal{N}_a[\boldsymbol{\mu}, \mathbf{\Omega}]$ denote the a-variate Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^a$ and covariance matrix $\mathbf{\Omega}_{a \times a} \succ 0$.

Remark 2.1. Assumption 2.1 (b) enlists some fundamental characteristics of SS settings. Indeed, the condition of $\mathcal{L}$ and $\mathcal{U}$ being equally distributed has usually been an integral part of the *definition* of SS settings (Chapelle, Schölkopf and Zien, 2006; Kawakita and Kanamori, 2013). Interpreted in missing data terminology, it entails that Y in $\mathcal{U}$ are 'missing completely at random' (MCAR), with the missingness/labeling being typically by design. Interestingly, the crucial assumption of MCAR, although commonly required, has often stayed implicit in the SSL literature (Lafferty and Wasserman, 2007). It is important to note that while the SS set-up can be viewed as a missing data problem, it is quite *different* from standard ones, since with $n/N \to 0$ i.e. $|\mathcal{U}| \gg |\mathcal{L}|$, the proportion of Y observed in $\mathcal{S}$ tends to 0 in SSL. Hence, the 'positivity assumption' typical in missing data theory, requiring this proportion to be bounded away from 0, is violated here. It is also worth noting that owing to such violations, the analysis of SS settings under more general missingness mechanisms such as 'missing at random' (MAR) is considerably more complicated and to our knowledge, the literature for SS *estimation* problems under such settings is virtually nonexistent. Furthermore, for such problems, the traditional goal in SSL, that of improving upon a 'supervised' estimator, can become unclear without

MCAR, unless an appropriately weighted version of the supervised estimator is considered. Given these subtleties and the traditional assumptions (often implicit) in SSL, the MCAR condition is assumed for most of this paper, although a brief discussion on possible extensions of our proposed SS estimators to MAR settings is provided in the Supplementary Material.

2.1. *The Target Parameter and Its Supervised Estimator.* We consider the linear regression *working model* given by:

$$Y = \overrightarrow{\mathbf{X}}'\boldsymbol{\theta} + \epsilon, \quad \text{with} \quad \mathbb{E}(\epsilon \mid \mathbf{X}) = 0, \tag{2.1}$$

where, $\boldsymbol{\theta} \in \mathbb{R}^{(p+1)}$ is an unknown regression parameter. Accounting for the potential mis-specification of the working model (2.1), we define the target parameter of interest as a model free parameter, as follows:

Definition 2.1. The target parameter $\boldsymbol{\theta}_0$ for linear regression may be defined as the solution to the normal equations: $\mathbb{E}\{\overrightarrow{\mathbf{X}}(Y - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta})\} = \mathbf{0}$ in $\boldsymbol{\theta} \in \mathbb{R}^{(p+1)}$, or equivalently, $\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta} \in \mathbb{R}^{(p+1)}}{\text{argmin}}\ \mathbb{E}(Y - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta})^2$.

Existence and uniqueness of $\boldsymbol{\theta}_0$ in 2.1 is clear. Further, $\overrightarrow{\mathbf{X}}'\boldsymbol{\theta}_0$ is the L_2 projection of $\mathbb{E}(Y \,|\, \mathbf{X}) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}})$ onto the subspace of all linear functions of $\mathbf{X}$ and hence, is the best linear predictor of Y given $\mathbf{X}$. The linear model (2.1) is *correct* (else, *mis-specified*) if and only if $\mathbb{E}(Y \,|\, \mathbf{X})$ lies in this space (in which case, $\mathbb{E}(Y \,|\, \mathbf{X}) = \overrightarrow{\mathbf{X}}'\boldsymbol{\theta}_0$). When the model is correct, $\boldsymbol{\theta}_0$ depends only on $\mathbb{P}_{Y|\mathbf{X}}$, not on $\mathbb{P}_{\mathbf{X}}$. Hence, improved estimation of $\boldsymbol{\theta}_0$ through SSL is impossible in this case unless further assumptions relating $\boldsymbol{\theta}_0$ to $\mathbb{P}_{\mathbf{X}}$ are made. On the other hand, under model mis-specification, the normal equations defining $\boldsymbol{\theta}_0$ inherently depend on $\mathbb{P}_{\mathbf{X}}$, thereby implying the potential utility of SSL in improving the estimation of $\boldsymbol{\theta}_0$ in this case.

The usual supervised estimator of $\boldsymbol{\theta}_0$ is the OLS estimator $\widehat{\boldsymbol{\theta}}$, the solution in $\boldsymbol{\theta}$ to the equation: $n^{-1}\sum_{i=1}^{n} \overrightarrow{\mathbf{X}}_i(Y_i - \overrightarrow{\mathbf{X}}_i'\boldsymbol{\theta}) = \mathbf{0}$, the normal equations based on $\mathcal{L}$. Under Assumption 2.1 (a), it is well known that as $n \to \infty$,

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\psi}_0(\mathbf{Z}_i) + O_p\left(n^{-\frac{1}{2}}\right) \xrightarrow{d} \mathcal{N}_{(p+1)}[\mathbf{0}, \boldsymbol{\Sigma}(g_{\boldsymbol{\theta}_0})], \tag{2.2}$$

where $\boldsymbol{\psi}_0(\mathbf{Z}) = \boldsymbol{\Gamma}^{-1}\{\overrightarrow{\mathbf{X}}(Y - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta}_0)\}$ and $g_{\boldsymbol{\theta}}(\mathbf{X}) = \overrightarrow{\mathbf{X}}'\boldsymbol{\theta} \ \ \forall\, \boldsymbol{\theta} \in \mathbb{R}^{(p+1)}$.

Our primary goal is to obtain efficient SS estimators of $\boldsymbol{\theta}_0$ using the *entire* data $\mathcal{S}$ and compare their efficiencies to that of $\widehat{\boldsymbol{\theta}}$. It is worth noting that the estimation efficiency of $\boldsymbol{\theta}_0$ also relates to the predictive performance of the fitted linear model since its out-of-sample prediction error is directly related to the mean squared error (w.r.t. the $\boldsymbol{\Sigma}$ metric) of the parameter estimate.

3. A Family of Imputation Based Semi-Supervised Estimators. If Y in $\mathcal{U}$ were actually observed, then one would simply fit the working model to the entire data in $\mathcal{S}$ for estimating $\boldsymbol{\theta}_0$. Our general approach is precisely motivated by this intuition. We first attempt to impute the missing Y in $\mathcal{U}$ based on suitable training of $\mathcal{L}$ in step (I). Then in step (II), we fit the linear model (2.1) to $\mathcal{U}$ with the imputed outcomes. Clearly, the imputation is critical. Inaccurate imputation would lead to biased estimate of $\boldsymbol{\theta}_0$, while inadequate imputation would result in loss of efficiency. We next consider SS estimators constructed under two imputation strategies for step (I) including a fully non-parametric imputation based on kernel smoothing (KS), and a semi-non-parametric (SNP) imputation that involves a smoothing step and a follow up 'refitting' step. Although the construction of the final EASE is based on the SNP imputation strategy, it is helpful to begin with a discussion of the first strategy in order to appropriately motivate and elucidate the discussion on EASE and the SNP imputation strategy.

3.1. *A Simple SS Estimator via Fully Non-Parametric Imputation.* We present here an estimator based on a fully non-parametric imputation involving KS when p is small. For simplicity, we shall assume here that $\mathbf{X}$ is continuous with a density $f(\cdot)$. Let $m(\mathbf{x}) = \mathbb{E}(Y \,|\, \mathbf{X} = \mathbf{x})$ and $l(\mathbf{x}) = m(\mathbf{x})f(\mathbf{x})$. Consider the local constant KS estimator of $m(\mathbf{x})$,

$$\widehat{m}(\mathbf{x}) \;=\; \frac{\frac{1}{nh^p}\sum_{i=1}^{n}\{K_h(\mathbf{X}_i,\mathbf{x})\}Y_i}{\frac{1}{nh^p}\sum_{i=1}^{n}K_h(\mathbf{X}_i,\mathbf{x})} \;=\; \frac{\widehat{l}(\mathbf{x})}{\widehat{f}(\mathbf{x})}, \tag{3.1}$$

where $K_h(\mathbf{u},\mathbf{v}) = K\{(\mathbf{u}-\mathbf{v})/h\}$ with $K : \mathbb{R}^p \to \mathbb{R}$ being some suitable kernel function and $h = h(n) > 0$ being the bandwidth. With $\widehat{m}(\cdot)$ as defined in (3.1), we now fit (2.1) to the imputed unlabeled data: $[\{\widehat{m}(\mathbf{X}_j), \mathbf{X}_j'\}' : j = n+1, ..., n+N]$ and obtain a SS estimator $\widehat{\boldsymbol{\theta}}_{np}$ of $\boldsymbol{\theta}_0$ as the solution in $\boldsymbol{\theta}$ to:

$$\frac{1}{N}\sum_{j=n+1}^{n+N}\overrightarrow{\mathbf{X}}_j\{\widehat{m}(\mathbf{X}_j) - \overrightarrow{\mathbf{X}}_j'\boldsymbol{\theta}\} = \mathbf{0}. \tag{3.2}$$

Here and throughout in our constructions of SS estimators, $\mathcal{L}$ with either the true or the imputed Y is not included in the final fitting step mostly due to technical convenience in the asymptotic analysis of our estimators, and also due to the fact that the contribution of $\mathcal{L}$, included in any form, in the final fitting step is asymptotically negligible since $n/N \to 0$. In order to study the properties of $\widehat{\boldsymbol{\theta}}_{np}$, we would require uniform (in L_∞ norm) convergence of $\widehat{m}(\cdot)$ to $m(\cdot)$, a problem that has been extensively studied in the non-parametric statistics literature (Newey, 1994; Andrews, 1995; Masry, 1996;

Hansen, 2008) under fairly general settings and assumptions. In particular, we would assume the following regularity conditions to hold:

Assumption 3.1. (i) $K(\cdot)$ is a symmetric q^{th} order kernel for some integer $q \geq 2$. (ii) $K(\cdot)$ is bounded, Lipschitz continuous and has a bounded support $\mathcal{K} \subseteq \mathbb{R}^p$. (iii) $\mathbb{E}(|Y|^s) < \infty$ for some $s > 2$. $\mathbb{E}(|Y|^s \mid \mathbf{X} = \mathbf{x}) f(\mathbf{x})$ and $f(\mathbf{x})$ are bounded on $\mathcal{X}$. (iv) $f(\mathbf{x})$ is bounded away from 0 on $\mathcal{X}$. (v) $m(\cdot)$ and $f(\cdot)$ are q times continuously differentiable with bounded q^{th} derivatives on some open set $\mathcal{X}_0 \supseteq \mathcal{X}$. (vi) For any $\delta > 0$, let $A_\delta \subseteq \mathbb{R}^p$ denote the set $\{(\mathbf{x} - \mathbf{X})/\delta : \mathbf{x} \in \mathcal{X}\}$. Then, for small enough δ, $A_\delta \supseteq \mathcal{K}$ almost surely (a.s.).

Conditions (i)-(v) are fairly standard in the literature. In (v), the set $\mathcal{X}_0$ is needed mostly to make the notion of differentiability well-defined, with both $m(\cdot)$ and $f(\cdot)$ understood to have been analytically extended over $(\mathcal{X}_0 \backslash \mathcal{X})$. Condition (vi) implicitly controls the tail behaviour of $\mathbf{X}$, requiring that perturbations of $\mathbf{X}$ in the form of $(\mathbf{X} + \delta\boldsymbol{\phi})$ with $\boldsymbol{\phi} \in \mathcal{K}$ (bounded) and δ small enough, belong to $\mathcal{X}$ a.s. $[\mathbb{P}_{\mathbf{X}}]$. We now present our result on $\widehat{\boldsymbol{\theta}}_{np}$.

Theorem 3.1. *Suppose* $n^{\frac{1}{2}} h^q \to 0$ *and* $(\log n)/(n^{\frac{1}{2}} h^p) \to 0$ *as* $n \to \infty$, *and let* $r_n = n^{\frac{1}{2}} h^q + (\log n)/(n^{\frac{1}{2}} h^p) + (n/N)^{\frac{1}{2}}$. *Then, under Assumption 3.1,*

$$n^{\frac{1}{2}} \left(\widehat{\boldsymbol{\theta}}_{np} - \boldsymbol{\theta}_0 \right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\psi}_{\mathit{eff}}(\mathbf{Z}_i) + O_p(r_n) \quad \xrightarrow{d} \mathcal{N}_{(p+1)}[\mathbf{0}, \boldsymbol{\Sigma}(m)], \tag{3.3}$$

where $\boldsymbol{\psi}_{\mathit{eff}}(\mathbf{Z}) = \boldsymbol{\Gamma}^{-1}[\overrightarrow{\mathbf{X}}\{Y - m(\mathbf{X})\}]$.

Remark 3.1. Theorem 3.1 establishes the efficient and adaptive nature of $\widehat{\boldsymbol{\theta}}_{np}$. The asymptotic variance $\boldsymbol{\Sigma}(m)$ of $\widehat{\boldsymbol{\theta}}_{np}$ satisfies $\boldsymbol{\Sigma}(g) - \boldsymbol{\Sigma}(m) \succeq 0 \ \forall \ g(\cdot) \in \mathcal{L}^2(\mathbf{X})$ and the inequality is strict unless $g(\cdot) = m(\cdot)$ a.s. $[\mathbb{P}_{\mathbf{X}}]$. Hence, $\widehat{\boldsymbol{\theta}}_{np}$ is asymptotically *optimal* among the class of all regular and asymptotically linear (RAL) estimators of $\boldsymbol{\theta}_0$ with influence function (IF) of the form: $\boldsymbol{\Gamma}^{-1}[\overrightarrow{\mathbf{X}}\{Y - g(\mathbf{X})\}]$ with $g(\cdot) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}})$. In particular, $\widehat{\boldsymbol{\theta}}_{np}$ is more efficient than $\widehat{\boldsymbol{\theta}}$ whenever (2.1) is mis-specified, and equally efficient when (2.1) is correct i.e. $m(\cdot) = g_{\boldsymbol{\theta}_0}(\cdot)$. Further, it can also be shown that $\boldsymbol{\psi}_{\mathrm{eff}}(\mathbf{Z})$ is the 'efficient' IF for estimating $\boldsymbol{\theta}_0$ under the semi-parametric model $\mathcal{M}_{\mathbf{X}} \equiv \{(\mathbb{P}_{Y|\mathbf{X}}, \mathbb{P}_{\mathbf{X}}) : \mathbb{P}_{\mathbf{X}}$ is known, $\mathbb{P}_{Y|\mathbf{X}}$ is unrestricted upto Assumption 2.1 (a)$\}$. Thus, $\widehat{\boldsymbol{\theta}}_{np}$ also globally *achieves* the semi-parametric efficiency bound under $\mathcal{M}_{\mathbf{X}}$. Lastly, note that at any parametric sub-model in $\mathcal{M}_{\mathbf{X}}$ that corresponds to (2.1) being correct, $\widehat{\boldsymbol{\theta}}$ also achieves optimality, thus showing that under $\mathcal{M}_{\mathbf{X}}$, it is not possible to improve upon $\widehat{\boldsymbol{\theta}}$ if the linear model is correct.

Remark 3.2. The asymptotic results in Theorem 3.1 require a kernel of order $q > p$ and h smaller in order than the 'optimal' bandwidth order $h_{opt} = O(n^{-1/(2q+p)})$. This *under-smoothing* requirement, often encountered in two-step estimators involving a first-step smoothing (Newey, Hsieh and Robins, 1998), generally results in sub-optimal performance of $\widehat{m}(.)$. The optimal under-smoothed bandwidth order for Theorem 3.1 is given by: $O(n^{-1/(q+p)})$.

3.2. *SS Estimators Based on Semi-Non-Parametric (SNP) Imputation.*

The simple and intuitive imputation strategy in Section 3.1 based on a fully non-parametric p-dimensional KS is however often undesirable in practice owing to the curse of dimensionality. In order to accommodate larger p, we now propose a more flexible SNP imputation method involving a dimension reduction, if needed, followed by a non-parametric calibration. An additional 'refitting' step is proposed to reduce the impact of bias from non-parametric estimation and possibly inadequate imputation due to dimension reduction. We also introduce some flexibility in terms of the smoothing methods, apart from KS, that can be used for the non-parametric calibration.

Let $r \leq p$ be a fixed positive integer and let $\mathbf{P}_r = [\mathbf{p}_1, .., \mathbf{p}_r]_{p\times r}$ be any rank r transformation matrix. Let $\mathbf{X}_{\mathbf{P}_r} = \mathbf{P}_r'\mathbf{X}$. Given $(r, \mathbf{P}_r)$, we may now consider approximating the regression function $\mathbb{E}(Y\,|\,\mathbf{X})$ by smoothing Y over the r dimensional $\mathbf{X}_{\mathbf{P}_r}$ instead of the original $\mathbf{X} \in \mathbb{R}^p$. In general, $\mathbf{P}_r$ can be user-defined and data dependent. A few reasonable choices of $\mathbf{P}_r$ are discussed in Section 5. If $\mathbf{P}_r$ depends only on the distribution of $\mathbf{X}$, it may be assumed to be known given the SS setting considered. If $\mathbf{P}_r$ also depends on the distribution of Y, then it needs to be estimated from $\mathcal{L}$ and the smoothing needs to be performed using the estimated $\mathbf{P}_r$.

For approximating $\mathbb{E}(Y\,|\,\mathbf{X})$, we may consider *any* reasonable smoothing technique $\mathcal{T}$. Some examples of $\mathcal{T}$ include KS, kernel machine regression and smoothing splines. Let $m(\mathbf{x}; \mathbf{P}_r)$ denote the 'target function' for smoothing Y over $\mathbf{X}_{\mathbf{P}_r}$ using $\mathcal{T}$. For notational simplicity, the dependence of $m(\mathbf{x}; \mathbf{P}_r)$ and other quantities on $\mathcal{T}$ is suppressed throughout. For $\mathcal{T} :=$ KS, the appropriate target function is given by: $m(\mathbf{x}; \mathbf{P}_r) = m_{\mathbf{P}_r}(\mathbf{P}_r'\mathbf{x})$, where $m_{\mathbf{P}_r}(\mathbf{z}) \equiv \mathbb{E}(Y \mid \mathbf{X}_{\mathbf{P}_r} = \mathbf{z})$. For basis function expansion based methods, $m(\mathbf{x}; \mathbf{P}_r)$ will typically correspond to the L_2 projection of $m(\mathbf{x}) \equiv \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}})$ onto the functional space spanned by the basis functions associated with $\mathcal{T}$. The results in this section apply to any choice of $\mathcal{T}$ that satisfies the required conditions. In Section 4, we provide more specific results for the implementation of our methods using $\mathcal{T} :=$ KS.

Note that we do *not* assume $m(\mathbf{x}; \mathbf{P}_r) = m(\mathbf{x})$ anywhere, and hence the name 'semi-non-parametric'. Clearly, with $\mathbf{P}_r = I_p$ and $\mathcal{T} :=$ KS, it reduces to a fully non-parametric approach. We next describe the two sub-steps involved in step (I) of the SNP imputation: (Ia) smoothing, and (Ib) refitting.

(Ia) Smoothing Step. With $\mathbf{P}_r$ and $m(\mathbf{x}; \mathbf{P}_r)$ as defined above, let $\widehat{\mathbf{P}}_r$ and $\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$ respectively denote their estimators based on $\mathcal{L}$. In order to address potential overfitting issues in the subsequent steps, we further consider generalized versions of these estimators based on $\mathbb{K}$-fold CV for a given fixed integer $\mathbb{K} \geq 1$. For any $\mathbb{K} \geq 2$, let $\{\mathcal{L}_k\}_{k=1}^{\mathbb{K}}$ denote a random partition of $\mathcal{L}$ into $\mathbb{K}$ disjoint subsets of equal sizes, $n_{\mathbb{K}} = n/\mathbb{K}$, with index sets $\{\mathcal{I}_k\}_{k=1}^{\mathbb{K}}$. Let $\mathcal{L}_k^-$ denote the set excluding $\mathcal{L}_k$ with size $n_{\mathbb{K}}^- = n - n_{\mathbb{K}}$ and respective index set $\mathcal{I}_k^-$. Let $\widehat{\mathbf{P}}_{r,k}$ and $\widehat{m}_k(\mathbf{x}; \widehat{\mathbf{P}}_{r,k})$ denote the corresponding estimators based on $\mathcal{L}_k^-$. Further, for notational consistency, we define for $\mathbb{K} = 1$, $\mathcal{L}_k = \mathcal{L}_k^- = \mathcal{L}$; $\mathcal{I}_k = \mathcal{I}_k^- = \{1, ..., n\}$; $n_{\mathbb{K}} = n_{\mathbb{K}}^- = n$; $\widehat{\mathbf{P}}_{r,k} = \widehat{\mathbf{P}}_r$ and $\widehat{m}_k(\mathbf{x}; \widehat{\mathbf{P}}_{r,k}) = \widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$.

(Ib) Refitting Step. In this step, we fit the linear model to $\mathcal{L}$ using $\mathbf{X}$ as predictors and the estimated $m(\mathbf{X}; \mathbf{P}_r)$ as an *offset*. To motivate this, we recall that the fully non-parametric imputation given in Section 3.1 consistently estimates $\mathbb{E}(Y|\mathbf{X})$, the L_2 projection onto a space that always contains the working model space, i.e. the linear span of $\overrightarrow{\mathbf{X}}$. This need not be true for the SNP imputation, since we do not assume $m(\mathbf{X}; \mathbf{P}_r) = m(\mathbf{X})$ necessarily. The refitting step essentially 'adjusts' for this so that the final imputation, combining the predictions from these two steps, targets a space that contains the working model space. In particular, for $\mathcal{T} :=$ KS with $r < p$, this step is critical to remove potential bias due to inadequate imputation.

Interestingly, it turns out that the refitting step should *always* be performed, *even when* $m(\mathbf{X}; \mathbf{P}_r) = m(\mathbf{X})$. It plays a crucial role in reducing the bias of the resulting SS estimator due to the inherent bias from non-parametric curve estimation. In particular, for $\mathcal{T} :=$ KS with *any* $r \leq p$, it ensures that a bandwidth of the optimal order can be used, thereby *eliminating the under-smoothing issue* as encountered in Section 3.1. The target parameter for the refitting step is simply the regression coefficient obtained from regressing the residual $Y - m(\mathbf{X}; \mathbf{P}_r)$ on $\mathbf{X}$ and may be defined as: $\boldsymbol{\eta}_{\mathbf{P}_r}$, the solution in $\boldsymbol{\eta} \in \mathbb{R}^{(p+1)}$ to the equation: $\mathbb{E}[\overrightarrow{\mathbf{X}}\{Y - m(\mathbf{X}; \mathbf{P}_r) - \overrightarrow{\mathbf{X}}'\boldsymbol{\eta}\}] = \mathbf{0}$. For any $\mathbb{K} \geq 1$, we estimate $\boldsymbol{\eta}_{\mathbf{P}_r}$ as $\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r, \mathbb{K})}$, the solution in $\boldsymbol{\eta}$ to the equation:

$$n^{-1} \sum_{k=1}^{\mathbb{K}} \sum_{i \in \mathcal{I}_k} \overrightarrow{\mathbf{X}}_i \{Y_i - \widehat{m}_k(\mathbf{X}_i; \widehat{\mathbf{P}}_{r,k}) - \overrightarrow{\mathbf{X}}_i' \boldsymbol{\eta}\} = \mathbf{0}. \tag{3.4}$$

For $\mathbf{X}_i \in \mathcal{L}_k$, the estimate of $m(\mathbf{X}_i; \mathbf{P}_r)$ to be used as an offset is obtained from $\widehat{m}_k(\cdot\, ; \widehat{\mathbf{P}}_{r,k})$ that is based on data in $\mathcal{L}_k^-$. For $\mathbb{K} \geq 2$, with $\mathcal{L}_k^- \perp\!\!\!\perp \mathcal{L}_k$, the residuals are thus estimated in a cross-validated manner. For $\mathbb{K} = 1$ however,

$\widehat{m}(\cdot\ ;\widehat{\mathbf{P}}_r)$ is estimated using the entire $\mathcal{L}$ which can lead to considerable underestimation of the true residuals owing to over-fitting and consequently, substantial finite sample bias in the resulting SS estimator of $\boldsymbol{\theta}_0$. This bias can be effectively reduced by using the CV approach with $\mathbb{K} \geq 2$. We next estimate the *target function* for the SNP imputation given by:

$$\begin{aligned} &(3.5) \qquad & \mu(\mathbf{x};\mathbf{P}_r) \;&=\; m(\mathbf{x};\mathbf{P}_r) + \overrightarrow{\mathbf{x}}'\boldsymbol{\eta}_{\mathbf{P}_r} \;\; \text{as:} \\ &(3.6) \qquad & \widehat{\mu}(\mathbf{x};\widehat{\mathcal{P}}_{r,\mathbb{K}}) \;&=\; \mathbb{K}^{-1}\sum_{k=1}^{\mathbb{K}} \widehat{m}_k(\mathbf{x};\widehat{\mathbf{P}}_{r,k}) + \overrightarrow{\mathbf{x}}'\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})}, \end{aligned}$$

where $\widehat{\mathcal{P}}_{r,\mathbb{K}} = \{\widehat{\mathbf{P}}_{r,k}\}_{k=1}^{\mathbb{K}}$. For notational simplicity, we suppress throughout the inherent dependence of $\widehat{\mu}(\cdot\ ;\ \cdot)$ itself on $\mathbb{K}$ and $\{\mathcal{L}_k^-\}_{k=1}^{\mathbb{K}}$. Note that similar to $m(\mathbf{X};\mathbf{P}_r)$, we also do *not* assume $\mu(\mathbf{X};\mathbf{P}_r) = m(\mathbf{X})$. Apart from the geometric motivation for the refitting step and its technical role in bias reduction, it also generally ensures the condition: $\mathbb{E}[\overrightarrow{\mathbf{X}}\{Y - \mu(\mathbf{X};\mathbf{P}_r\}] = \mathbf{0}$, *regardless* of the true underlying $m(\mathbf{X})$. This condition is a key requirement for the asymptotic expansions, in Theorem 3.2, of our resulting SS estimators. Using $\widehat{\mu}(\cdot\ ;\widehat{\mathcal{P}}_{r,\mathbb{K}})$, we now construct our final SS estimator as follows.

SS Estimator from SNP Imputation. In step (II), we fit the linear model to the SNP imputed unlabeled data: $[\{\widehat{\mu}(\mathbf{X}_j;\widehat{\mathcal{P}}_{r,\mathbb{K}}), \mathbf{X}_j'\}', j = n+1, ..., n+N]$ and obtain a SS estimator $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ of $\boldsymbol{\theta}_0$ given by:

$$(3.7) \;\; \widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} \text{ is the solution in } \boldsymbol{\theta} \text{ to } \frac{1}{N}\sum_{j=n+1}^{n+N} \overrightarrow{\mathbf{X}}_j\{\widehat{\mu}(\mathbf{X}_j;\widehat{\mathcal{P}}_{r,\mathbb{K}}) - \overrightarrow{\mathbf{X}}_j'\boldsymbol{\theta}\} = \mathbf{0}.$$

For convenience of further discussion, let us define: $\forall\ k \in \{1,\ldots,\mathbb{K}\}$,

$$\begin{aligned} &(3.8) \qquad \widehat{\Delta}_k(\mathbf{x};\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k}) = \widehat{m}_k(\mathbf{x};\widehat{\mathbf{P}}_{r,k}) - m(\mathbf{x};\mathbf{P}_r) \;\;\forall\, \mathbf{x} \in \mathcal{X}, \quad \text{and} \\ &(3.9) \qquad \widehat{\mathbf{G}}_k(\mathbf{x}) = \overrightarrow{\mathbf{x}}\widehat{\Delta}_k(\mathbf{x};\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k}) - \mathbb{E}_{\mathbf{X}}\{\overrightarrow{\mathbf{X}}\widehat{\Delta}_k(\mathbf{X};\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k})\} \;\;\forall\, \mathbf{x} \in \mathcal{X}, \end{aligned}$$

where $\mathbb{E}_{\mathbf{X}}(\cdot)$ denotes expectation w.r.t. $\mathbf{X} \in \mathcal{U}$. The dependence of $\widehat{\mathbf{G}}_k(\cdot)$ on $(\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k})$ and $\mathbb{P}_{\mathbf{X}}$ is suppressed here for notational simplicity. We now present our main result summarizing the properties of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$.

THEOREM 3.2. *Suppose that $\mathcal{T}$ satisfies: (i) sup$_{\mathbf{x}\in\mathcal{X}}|m(\mathbf{x};\mathbf{P}_r)| < \infty$ and (ii) sup$_{\mathbf{x}\in\mathcal{X}}|\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r) - m(\mathbf{x};\mathbf{P}_r)| = O_p(c_n)$ for some $c_n = o(1)$. With $\widehat{\mathbf{G}}_k(.)$ as in (3.9), define $\mathbb{G}_{n,\mathbb{K}} = n^{-\frac{1}{2}}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k}\widehat{\mathbf{G}}_k(\mathbf{X}_i)$. Then, for any $\mathbb{K} \geq 1$,*

$$(3.10) \qquad n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\theta}_0\right) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\psi}(\mathbf{Z}_i;\mathbf{P}_r) - \boldsymbol{\Gamma}^{-1}\mathbb{G}_{n,\mathbb{K}} + O_p(c^*_{n,\mathbb{K}}),$$

where $\boldsymbol{\psi}(\mathbf{Z};\mathbf{P}_r) = \boldsymbol{\Gamma}^{-1}[\overrightarrow{\mathbf{X}}\{Y - \mu(\mathbf{X};\mathbf{P}_r)\}]$ *and* $c^*_{n,\mathbb{K}} = c_{n_{\mathbb{K}}^-} + n^{-\frac{1}{2}} + (n/N)^{\frac{1}{2}} = o(1)$. *Further, for any fixed* $\mathbb{K} \geq 2$, $\mathbb{G}_{n,\mathbb{K}} = O_p(c_{n_{\mathbb{K}}^-})$, *so that*

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\theta}_0\right) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\psi}(\mathbf{Z}_i;\mathbf{P}_r) + O_p(c_{n_{\mathbb{K}}^-} + c^*_{n,\mathbb{K}}), \tag{3.11}$$

which converges in distribution to $\mathcal{N}_{(p+1)}[\mathbf{0}, \boldsymbol{\Sigma}\{\mu(\cdot\,;\mathbf{P}_r)\}]$.

Remark 3.3. If the imputation is ‘sufficient’ so that $\mu(\mathbf{x};\mathbf{P}_r) = m(\mathbf{x})$, then $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$, for any $\mathbb{K} \geq 2$, enjoys the same set of optimality properties as those noted in Remark 3.1 for $\widehat{\boldsymbol{\theta}}_{np}$ (while requiring less stringent assumptions about $K(\cdot)$ and h, if KS is used). If $\mu(\mathbf{x};\mathbf{P}_r) \neq m(\mathbf{x})$, then it is however unclear whether $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ is always more efficient than $\widehat{\boldsymbol{\theta}}$. This will be addressed in Section 3.3 where we develop the final EASE.

Remark 3.4. Apart from the fairly mild condition (i), Theorem 3.2 *only* requires uniform consistency of $\widehat{m}(\cdot\,;\widehat{\mathbf{P}}_r)$ w.r.t. $m(\cdot\,;\mathbf{P}_r)$ for establishing the $n^{\frac{1}{2}}$-consistency and asymptotic normality (CAN) of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ for any $\mathbb{K} \geq 2$. The uniform consistency typically holds for a wide range of smoothing methods $\mathcal{T}$ under fairly general conditions. For $\mathcal{T} :=$ KS in particular, we provide explicit results in Section 4 under mild regularity conditions that allow the use of any kernel order and the associated optimal bandwidth order. This is a notable relaxation from the stringent requirements for Theorem 3.1 that necessitate under-smoothing and the use of higher order kernels.

Remark 3.5. The CAN property of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$ has *not* yet been established. The term $\mathbb{G}_{n,\mathbb{K}}$ in (3.10) behaves quite differently when $\mathbb{K} = 1$ compared to $\mathbb{K} \geq 2$. We derive the properties of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$ in Section 4 for $\mathcal{T} :=$ KS.

3.3. *Efficient and Adaptive Semi-Supervised Estimators (EASE).*

To ensure adaptivity even when $\mu(\mathbf{x};\mathbf{P}_r) \neq m(\mathbf{x})$, we now define the final EASE as an optimal linear combination of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$. Specifically, for any fixed $(p+1)\times(p+1)$ matrix $\boldsymbol{\Delta}$, $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}(\boldsymbol{\Delta}) = \widehat{\boldsymbol{\theta}} + \boldsymbol{\Delta}(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} - \widehat{\boldsymbol{\theta}})$ is a CAN estimator of $\boldsymbol{\theta}_0$ whenever $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ are, and an optimal $\boldsymbol{\Delta}$ can be selected easily to minimize the asymptotic variance of the combined estimator. For simplicity, we focus here on $\boldsymbol{\Delta}$ being a diagonal matrix with $\boldsymbol{\Delta} = \mathrm{diag}(\delta_1, ..., \delta_{p+1})$. Then the EASE is defined as $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})} \equiv \widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}(\widehat{\boldsymbol{\Delta}})$ with $\widehat{\boldsymbol{\Delta}}$ being any consistent estimator (see Section 3.4 for details) of the minimizer $\overline{\boldsymbol{\Delta}} = \mathrm{diag}(\overline{\delta}_1, ..., \overline{\delta}_{p+1})$, where $\forall\ 1 \leq l \leq (p+1)$,

$$\overline{\delta}_l = -\lim_{\epsilon \downarrow 0} \frac{\mathrm{Cov}\left\{\boldsymbol{\psi}_{0[l]}(\mathbf{Z}),\ \boldsymbol{\psi}_{[l]}(\mathbf{Z}; \mathbf{P}_r) - \boldsymbol{\psi}_{0[l]}(\mathbf{Z})\right\}}{\mathrm{Var}\left\{\boldsymbol{\psi}_{[l]}(\mathbf{Z}; \mathbf{P}_r) - \boldsymbol{\psi}_{0[l]}(\mathbf{Z})\right\}\ +\ \epsilon}, \tag{3.12}$$

and for any vector $\mathbf{a}$, $\mathbf{a}_{[l]}$ denotes its l^{th} component. Note that in (3.12), the ϵ and the limit outside are included to formally account for the case: $\boldsymbol{\psi}_{0[l]}(\mathbf{Z}) = \boldsymbol{\psi}_{[l]}(\mathbf{Z}, \mathbf{P}_r)$ a.s. $[\mathbb{P}_{\mathbf{Z}}]$, when we define $\overline{\delta}_l = 0$ for identifiability.

It is straightforward to show that $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}(\overline{\boldsymbol{\Delta}})$ are asymptotically equivalent, so that $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ is a RAL estimator of $\boldsymbol{\theta}_0$ satisfying:

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\theta}_0\right) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\psi}(\mathbf{Z}_i; \mathbf{P}_r, \overline{\boldsymbol{\Delta}}) + o_p(1) \ \xrightarrow{d} \mathcal{N}_{(p+1)}[\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{P}_r}(\overline{\boldsymbol{\Delta}})],$$

as $n \to \infty$, where $\boldsymbol{\psi}(\mathbf{Z}; \mathbf{P}_r, \overline{\boldsymbol{\Delta}}) = \boldsymbol{\psi}_0(\mathbf{Z}) + \overline{\boldsymbol{\Delta}}\{\boldsymbol{\psi}(\mathbf{Z}; \mathbf{P}_r) - \boldsymbol{\psi}_0(\mathbf{Z})\}$ and $\boldsymbol{\Sigma}_{\mathbf{P}_r}(\overline{\boldsymbol{\Delta}}) = \mathrm{Var}\{\boldsymbol{\psi}(\mathbf{Z}; \mathbf{P}_r, \overline{\boldsymbol{\Delta}})\}$. Note that when either the linear model holds or the SNP imputation is sufficient, then $\boldsymbol{\psi}(\mathbf{Z}; \mathbf{P}_r, \overline{\boldsymbol{\Delta}}) = \boldsymbol{\psi}_{\mathrm{eff}}(\mathbf{Z})$, so that $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ is asymptotically optimal in the sense of Remark 3.1. Further, when neither cases hold, $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ is no longer optimal, but is *still* efficient and adaptive compared to $\widehat{\boldsymbol{\theta}}$. Lastly, if the imputation is certain to be sufficient (e.g. if $r = p$ and $\mathcal{T} :=$ KS), we may simply define $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})} = \widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$.

Remark 3.6. It can be shown that under $\mathcal{M}_{\mathbf{X}}$, defined in Remark 3.1, the class of all possible IFs achievable by RAL estimators of $\boldsymbol{\theta}_0$ is given by: $\mathcal{IF}_{\boldsymbol{\theta}_0,\mathcal{M}_{\mathbf{X}}} = \{\boldsymbol{\phi}_g(\mathbf{Z}) \equiv \boldsymbol{\Gamma}^{-1}\overrightarrow{\mathbf{X}}\{Y - g(\mathbf{X})\} : g(\cdot) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}}),\, \mathbb{E}\{\boldsymbol{\phi}_g(\mathbf{Z})\} = \mathbf{0}\}$. The IFs achieved by $\widehat{\boldsymbol{\theta}}$, $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ are clearly members of this class. The SNP imputation, for various choices of the imputation function $\mu(\cdot\,; \mathbf{P}_r)$, therefore equips us with a *family* of RAL estimator pairs $\{\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}, \widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}\}$ for estimating $\boldsymbol{\theta}_0$. The IF of $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ is further guaranteed to dominate that of $\widehat{\boldsymbol{\theta}}$, and when $\mu(\cdot\,; \mathbf{P}_r) = m(\cdot)$, it also dominates all other IFs $\in \mathcal{IF}_{\boldsymbol{\theta}_0,M_{\mathbf{X}}}$.

3.4. *Inference for EASE and the SNP Imputation Based SS Estimators.*

We now provide procedures for making inference about $\boldsymbol{\theta}_0$ based on $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ obtained using $\mathbb{K} \geq 2$. We also employ a 'double' CV to overcome bias in variance estimation due to over-fitting. A key step involved in the variance estimation is to obtain reasonable estimates of $\{\mu(\mathbf{X}_i; \mathbf{P}_r)\}_{i=1}^{n}$.

Although $\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})}$ in (3.4) was constructed via CV, the corresponding estimate, $\widehat{\mu}(\mathbf{x};\widehat{\mathcal{P}}_{r,\mathbb{K}})$ in (3.6), of $\mu(\mathbf{x};\mathbf{P}_r)$ is likely to be over-fitted for $\mathbf{X}_i \in \mathcal{L}$. To construct bias corrected estimates of $\mu(\mathbf{X}_i;\mathbf{P}_r)$, we first obtain $\mathbb{K}$ separate *doubly cross-validated* estimates of $\boldsymbol{\eta}_{\mathbf{P}_r}$, $\{\widehat{\boldsymbol{\eta}}^k_{(\mathbf{P}_r,\mathbb{K})} : k = 1,...,\mathbb{K}\}$, with $\widehat{\boldsymbol{\eta}}^k_{(\mathbf{P}_r,\mathbb{K})}$, for each k, being the solution in $\boldsymbol{\eta}$ to $\sum_{k'\neq k} \mathcal{S}_{k'}(\boldsymbol{\eta}) = \mathbf{0}$, where

$$\mathcal{S}_{k'}(\boldsymbol{\eta}) = \textstyle\sum_{i\in\mathcal{I}_{k'}} \overrightarrow{\mathbf{X}}_i\{Y_i - \widehat{m}_{k'}(\mathbf{X}_i;\widehat{\mathbf{P}}_{r,k'}) - \overrightarrow{\mathbf{X}}'_i\boldsymbol{\eta}\} \quad \forall\, k' \in \{1,\ldots,\mathbb{K}\}.$$

For each k and $k' \neq k$, $\mathcal{S}_{k'}(\boldsymbol{\eta})$ is constructed such that $\{\mathbf{Z}_i : i \in \mathcal{I}_{k'}\}$ used for obtaining $\widehat{\boldsymbol{\eta}}^k_{(\mathbf{P}_r,\mathbb{K})}$ is independent of $\widehat{m}_{k'}(\cdot\,;\widehat{\mathbf{P}}_{r,k'})$ that is based on $\mathcal{L}^-_{k'} \perp\!\!\!\perp \mathcal{L}_{k'}$. Then, for each $\mathbf{X}_i \in \mathcal{L}_k$ and $k \in \{1,\ldots,\mathbb{K}\}$, we may estimate $\mu(\mathbf{X}_i;\mathbf{P}_r)$ as:

$$\widehat{\mu}_k(\mathbf{X}_i;\widehat{\mathcal{P}}_{r,\mathbb{K}}) = \widehat{m}_k(\mathbf{X}_i;\widehat{\mathbf{P}}_{r,k}) + \overrightarrow{\mathbf{X}}'_i\widehat{\boldsymbol{\eta}}^k_{(\mathbf{P}_r,\mathbb{K})}.$$

We exclude $\mathcal{S}_k(\boldsymbol{\eta})$ in the construction of $\widehat{\boldsymbol{\eta}}^k_{(\mathbf{P}_r,\mathbb{K})}$ to reduce over-fitting bias in the residuals $\{Y_i - \widehat{\mu}_k(\mathbf{X}_i;\widehat{\mathcal{P}}_{r,\mathbb{K}})\}$ which we now use for estimating the IFs.

For each $\mathbf{Z}_i \in \mathcal{L}_k$ and $k \in \{1,..,\mathbb{K}\}$, we estimate $\boldsymbol{\psi}_0(\mathbf{Z}_i)$ and $\boldsymbol{\psi}(\mathbf{Z}_i;\mathbf{P}_r)$, the corresponding IFs of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$, respectively as:

$$\widehat{\boldsymbol{\psi}}_0(\mathbf{Z}_i) = \widehat{\boldsymbol{\Gamma}}^{-1}\{\overrightarrow{\mathbf{X}}_i(Y_i - \overrightarrow{\mathbf{X}}'_i\widehat{\boldsymbol{\theta}})\} \text{ and } \widehat{\boldsymbol{\psi}}_k(\mathbf{Z}_i;\mathbf{P}_r) = \widehat{\boldsymbol{\Gamma}}^{-1}[\overrightarrow{\mathbf{X}}_i\{Y_i - \widehat{\mu}_k(\mathbf{X}_i;\widehat{\mathcal{P}}_{r,\mathbb{K}})\}],$$

where $\widehat{\boldsymbol{\Gamma}}$ denotes any consistent estimator of $\boldsymbol{\Gamma}$ from $\mathcal{L}$ and/or $\mathcal{U}$, e.g. $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}_n \equiv n^{-1}\sum_{i=1}^n \overrightarrow{\mathbf{X}}_i\overrightarrow{\mathbf{X}}'_i$ based on $\mathcal{L}$, or $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}_N \equiv N^{-1}\sum_{j=n+1}^{n+N} \overrightarrow{\mathbf{X}}_j\overrightarrow{\mathbf{X}}'_j$ based on $\mathcal{U}$. Then, $\boldsymbol{\Sigma}\{\mu(\cdot\,;\mathbf{P}_r)\}$ in (3.11) may be consistently estimated as:

$$\widehat{\boldsymbol{\Sigma}}\{\mu(\cdot\,;\mathbf{P}_r)\} = n^{-1}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k} \widehat{\boldsymbol{\psi}}_k(\mathbf{Z}_i;\mathbf{P}_r)\widehat{\boldsymbol{\psi}}'_k(\mathbf{Z}_i;\mathbf{P}_r).$$

To estimate the combination matrix $\overline{\boldsymbol{\Delta}}$ in (3.12) and the asymptotic variance, $\boldsymbol{\Sigma}_{\mathbf{P}_r}(\overline{\boldsymbol{\Delta}})$, of EASE consistently, let us define, $\forall\, 1 \leq l \leq (p+1)$,

$$\widehat{\sigma}_{l,12} = -\, n^{-1}\textstyle\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k} \widehat{\boldsymbol{\psi}}_{0[l]}(\mathbf{Z}_i)\{\widehat{\boldsymbol{\psi}}_{k[l]}(\mathbf{Z}_i;\mathbf{P}_r) - \widehat{\boldsymbol{\psi}}_{0[l]}(\mathbf{Z}_i)\},$$
$$\widehat{\sigma}_{l,22} = n^{-1}\textstyle\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k} \{\widehat{\boldsymbol{\psi}}_{k[l]}(\mathbf{Z}_i;\mathbf{P}_r) - \widehat{\boldsymbol{\psi}}_{0[l]}(\mathbf{Z}_i)\}^2,$$

and $\widehat{\delta}_l = \widehat{\sigma}_{l,12}/(\widehat{\sigma}_{l,22} + \epsilon_n)$ for some sequence $\epsilon_n \to 0$ with $n^{\frac{1}{2}}\epsilon_n \to \infty$. Then, we estimate $\overline{\boldsymbol{\Delta}}$ and $\boldsymbol{\Sigma}_{\mathbf{P}_r}(\overline{\boldsymbol{\Delta}})$ respectively as: $\widehat{\boldsymbol{\Delta}} = \operatorname{diag}(\widehat{\delta}_1,...,\widehat{\delta}_{p+1})$ and

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{P}_r}(\widehat{\boldsymbol{\Delta}}) = n^{-1}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k} \widehat{\boldsymbol{\psi}}_k(\mathbf{Z}_i;\mathbf{P}_r,\widehat{\boldsymbol{\Delta}})\widehat{\boldsymbol{\psi}}'_k(\mathbf{Z}_i;\mathbf{P}_r,\widehat{\boldsymbol{\Delta}}),$$

where $\widehat{\boldsymbol{\psi}}_k(\mathbf{Z};\mathbf{P}_r,\widehat{\boldsymbol{\Delta}}) = \widehat{\boldsymbol{\psi}}_0(\mathbf{Z}) + \widehat{\boldsymbol{\Delta}}\{\widehat{\boldsymbol{\psi}}_k(\mathbf{Z};\mathbf{P}_r) - \widehat{\boldsymbol{\psi}}_0(\mathbf{Z})\}\ \forall\ k \in \{1,\ldots,\mathbb{K}\}$. Normal confidence intervals (CIs) for the parameters of interest can also be constructed accordingly based on these variance estimates.

4. Implementation Based on KS. We next detail the specific implementation of the SNP imputation based on KS estimators. With $\mathcal{T} := \text{KS}$, the target function for the smoothing is given by: $m(\mathbf{x}; \mathbf{P}_r) = m_{\mathbf{P}_r}(\mathbf{P}_r'\mathbf{x}) \equiv \mathbb{E}(Y \mid \mathbf{X}_{\mathbf{P}_r} = \mathbf{P}_r'\mathbf{x})$. For simplicity, we assume that $\mathbf{X}_{\mathbf{P}_r}$ is continuous with a density $f_{\mathbf{P}_r}(\cdot)$ and support $\mathcal{X}_{\mathbf{P}_r} \equiv \{\mathbf{P}_r'\mathbf{x} : \mathbf{x} \in \mathcal{X}\} \subseteq \mathbb{R}^r$. Let us now consider the following class of local constant KS estimators for $m(\mathbf{x}; \mathbf{P}_r)$:

$$(4.1) \quad \widehat{m}_k(\mathbf{x}; \widehat{\mathbf{P}}_{r,k}) = \frac{\frac{1}{n_{\mathbb{K}}^- h^r} \sum_{i \in \mathcal{I}_k^-} \{K_h(\widehat{\mathbf{P}}'_{r,k}\mathbf{X}_i, \widehat{\mathbf{P}}'_{r,k}\mathbf{x})\} Y_i}{\frac{1}{n_{\mathbb{K}}^- h^r} \sum_{i \in \mathcal{I}_k^-} K_h(\widehat{\mathbf{P}}'_{r,k}\mathbf{X}_i, \widehat{\mathbf{P}}'_{r,k}\mathbf{x})} \quad \forall\ 1 \leq k \leq \mathbb{K},$$

where $K_h(\cdot)$ and h are as in Section 3.1 with $K(\cdot)$ now being a suitable kernel on $\mathbb{R}^r$. In the light of Theorem 3.2, we focus primarily on establishing the uniform consistency of $\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r) \equiv \widehat{m}_1(\mathbf{x}; \widehat{\mathbf{P}}_{r,1})$ in (4.1) with $\mathbb{K} = 1$, *accounting* for the additional estimation error from $\widehat{\mathbf{P}}_r$. For establishing the desired result, we shall assume the following regularity conditions to hold:

Assumption 4.1. (i) $K(\cdot)$ is a symmetric kernel of order $q \geq 2$ with finite q^{th} moments. (ii) $K(\cdot)$ is bounded, integrable and is either Lipschitz continuous with a compact support or, has a bounded derivative $\boldsymbol{\nabla} K(\cdot)$ which satisfies: $\|\boldsymbol{\nabla} K(\mathbf{z})\| \leq \Lambda \|\mathbf{z}\|^{-\rho}\ \forall\ \mathbf{z} \in \mathbb{R}^r$ with $\|\mathbf{z}\| > L$, where $\Lambda > 0$, $L > 0$ and $\rho > 1$ are some fixed constants, and $\|.\|$ denotes the standard L_2 vector norm. (iii) $\mathcal{X}_{\mathbf{P}_r} \subseteq \mathbb{R}^r$ is compact. $\mathbb{E}(|Y|^s) < \infty$ for some $s > 2$. $\mathbb{E}(|Y|^s \mid \mathbf{X}_{\mathbf{P}_r} = \mathbf{z}) f_{\mathbf{P}_r}(\mathbf{z})$ and $f_{\mathbf{P}_r}(\mathbf{z})$ are bounded on $\mathcal{X}_{\mathbf{P}_r}$. (iv) $f_{\mathbf{P}_r}(\mathbf{z})$ is bounded away from 0 on $\mathcal{X}_{\mathbf{P}_r}$. (v) $m_{\mathbf{P}_r}(\mathbf{z})$ and $f_{\mathbf{P}_r}(\mathbf{z})$ are both q times continuously differentiable with bounded q^{th} derivatives on some open set $\mathcal{X}_{0,\mathbf{P}_r} \supseteq \mathcal{X}_{\mathbf{P}_r}$. *Additional Conditions (required only when* $\mathbf{P}_r$ *needs to be estimated):* (vi) $K(\cdot)$ has a bounded and integrable derivative $\boldsymbol{\nabla} K(\cdot)$. (vii) $\boldsymbol{\nabla} K(\cdot)$ satisfies: $\|\boldsymbol{\nabla} K(\mathbf{z}_1) - \boldsymbol{\nabla} K(\mathbf{z}_2)\| \leq \|\mathbf{z}_1 - \mathbf{z}_2\|\ \phi(\mathbf{z}_1)\ \forall\ \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^r$ such that $\|\mathbf{z}_1 - \mathbf{z}_2\| \leq L^*$, for some fixed constant $L^* > 0$, and some bounded and integrable function $\phi : \mathbb{R}^r \to \mathbb{R}^+$. (viii) $\boldsymbol{\nabla} K(\cdot)$ is Lipschitz continuous on $\mathbb{R}^r$. (ix) $\mathbb{E}(\mathbf{X} \mid \mathbf{X}_{\mathbf{P}_r} = \mathbf{z})$ and $\mathbb{E}(\mathbf{X}Y \mid \mathbf{X}_{\mathbf{P}_r} = \mathbf{z})$ are both continuously differentiable with bounded first derivatives on $\mathcal{X}_{0,\mathbf{P}_r} \supseteq \mathcal{X}_{\mathbf{P}_r}$.

Assumption 4.1, mostly adopted from Hansen (2008), imposes some mild smoothness and moment conditions most of which are fairly standard, except perhaps the conditions on $K(\cdot)$ in (vi)-(viii) all of which are however satisfied by the Gaussian kernel among others. We now propose the following result.

Theorem 4.1. *Suppose* $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$ *for some* $\alpha_n = o(1)$ *with* $\alpha_n = 0$ *identically if* $\mathbf{P}_r$ *is known. Let* q *be the order of the kernel* $K(.)$ *in*

(4.1) for some integer $q \geq 2$. Define:

$$a_{n,1} = \alpha_n \left(\frac{\log n}{nh^{r+2}} \right)^{\frac{1}{2}} + \alpha_n^2 h^{-(r+2)} + \alpha_n, \quad a_{n,2} = \left(\frac{\log n}{nh^r} \right)^{\frac{1}{2}} + h^q$$

and assume that each of the terms involved in $a_{n,1} = o(1)$ and $a_{n,2} = o(1)$. Then, under Assumption 4.1, $\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$, based on (4.1), satisfies:

$$sup_{\mathbf{x} \in \mathcal{X}} \; |\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r) - m(\mathbf{x}; \mathbf{P}_r)| = O_p(a_{n,1} + a_{n,2}). \tag{4.2}$$

REMARK 4.1. Theorem 4.1 establishes the L_∞ error rate of $\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$ under mild regularity conditions and restrictions on h. Among its various implications, the rate also ensures uniform consistency of $\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$ at the optimal bandwidth order: $h_{opt} = O(n^{-1/(2q+r)})$ for any kernel order $q \geq 2$ and any $r \leq p$, as long as $\alpha_n = o(n^{-(r+2)/(4q+2r)})$ which always includes: $\alpha_n = O(n^{-\frac{1}{2}})$ and $\alpha_n = 0$. These two cases are particularly relevant in practice as $\mathbf{P}_r$ being finite dimensional, $n^{\frac{1}{2}}$-consistent estimators of $\mathbf{P}_r$ should typically exist. For both cases, using h_{opt} results in $a_{n,1}$ to be of lower order (for $q > 2$) or the same order (for $q = 2$) compared to that of the main term $a_{n,2}$, so that the usual optimal rate prevails as the overall error rate.

Properties of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r, \mathbb{K})}$ for $\mathbb{K} = 1$. We now address the CAN property of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r, \mathbb{K})}$ for $\mathbb{K} = 1$ under the KS framework. Based on (3.10) and Remark 3.5, the only step required for this is to effectively control the term $\mathbb{G}_{n,\mathbb{K}}$ in (3.10). We propose the following result in this regard.

THEOREM 4.2. *Let $\mathbb{K} = 1$, $\mathcal{T} :=$ KS, $\mathbb{G}_{n,\mathbb{K}}$ be as in (3.10), and $\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$ be the KS estimator based on (4.1). Let α_n, $a_{n,1}$ and $a_{n,2}$ be as in Theorem 4.1 with $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$. Assume that $a^*_{n,1}$ and $a^*_{n,2}$ are $o(1)$, where*

$$a^*_{n,1} = \alpha_n + \frac{\alpha_n}{n^{\frac{1}{2}} h^{(r+1)}} + n^{\frac{1}{2}} \alpha_n^2 h^{-2} + n^{\frac{1}{2}} a_{n,1}^2 + n^{\frac{1}{2}} a_{n,1} a_{n,2} \quad \textit{and} \quad a^*_{n,2} = n^{\frac{1}{2}} a_{n,2}^2.$$

Then, under Assumption 4.1, $\mathbb{G}_{n,\mathbb{K}} = O_p(a^*_{n,1} + a^*_{n,2}) = o_p(1)$. Further, let $c^*_{n,\mathbb{K}}$ be as in Theorem 3.2 with $c_n = (a_{n,1} + a_{n,2})$. Then, using (3.10),

$$n^{\frac{1}{2}} \left(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r, \mathbb{K})} - \boldsymbol{\theta}_0 \right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{Z}_i, \mathbf{P}_r) + O_p(c^*_{n,\mathbb{K}} + d_n), \tag{4.3}$$

where $d_n = a^*_{n,1} + a^*_{n,2}$. Hence, $n^{\frac{1}{2}} (\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r, \mathbb{K})} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_{(p+1)}[\mathbf{0}, \boldsymbol{\Sigma}\{\mu(\cdot\,; \mathbf{P}_r)\}]$.

REMARK 4.2. Note that the term $a^*_{n,2}$ *always* requires $q > r/2$ in order to converge to 0, thus showing the contrasting behavior of the case $\mathbb{K} = 1$ compared to $\mathbb{K} \geq 2$ where no such higher order kernel restriction is required. Nevertheless, when $\alpha_n = O(n^{-\frac{1}{2}})$ or $\alpha_n = 0$, the optimal bandwidth order: $h_{opt} = O(n^{-1/(2q+r)})$ can indeed be *still* used as long as $q > r/2$ is satisfied. Despite these facts and all the theoretical guarantee in Theorem 4.2, empirical evidence however seems to suggest that $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$ can be substantially *biased* in finite samples, in part due to over-fitting.

REMARK 4.3. *Technical benefits of refitting and CV:* Suppose that $\mathbf{P}_r = I_p$, so that the SNP imputation with $\mathcal{T} :=$ KS is indeed sufficient. Further, assume that all of Theorems 3.1-4.2 hold, so that the estimators $\widehat{\boldsymbol{\theta}}_{np}$, $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$, and $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ ($\mathbb{K} \geq 2$) are comparable and all asymptotically optimal. However, their constructions are quite different which can significantly affect their finite sample performances. $\widehat{\boldsymbol{\theta}}_{np}$ is based on KS only, and requires stringent under-smoothing and a kernel of order $q > p$ (Remark 3.2); $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$ is based on KS and refitting (*although* the KS itself is certain to be sufficient), and requires no under-smoothing but needs a (weaker) kernel order condition ($q > p/2$) (Remark 4.2); while $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ ($\mathbb{K} \geq 2$) additionally involves CV, and requires no under-smoothing or higher order kernel conditions (Remark 3.4). This highlights the critical role played by refitting and CV, apart from their primary roles in the SNP imputation, in removing any under-smoothing and/or higher order kernel restrictions when $\mathcal{T} :=$ KS, and this continues to hold for any other $(r, \mathbf{P}_r)$ as well. In particular, it shows, rather surprisingly, that refitting should be performed in order to avoid under-smoothing *even if* the smoothing is known to be sufficient.

REMARK 4.4. As mentioned in Section 3.2, $\mathcal{T} :=$ KS along with possible dimension reduction is just *one* reasonable choice of $\mathcal{T}$ for implementing the SNP imputation, all technical requirements for which have been thoroughly established in Section 4. In general, other smoothing methods, as long as the requirements are satisfied, can also be equally used as choices of $\mathcal{T}$. One such choice could be kernel machine (KM) regression (with possibly no use of dimension reduction, as KM uses penalization to effectively regularize the target even with $\mathbf{P}_r = I_p$). We leave its implementation details to the reader as they are readily available in a multitude of references, and also skip any theoretical treatment, considering the primary goal and scope of this paper. However, detailed numerical results are presented in Section 6 for this choice of $\mathcal{T}$ as well to illustrate the wider applicability of our proposed methods.

5. Choices of $\mathbf{P}_r$: Dimension Reduction Techniques.

We next discuss choosing and estimating the matrix $\mathbf{P}_r$ $(r < p)$ to be used for dimension reduction, if required, in the SNP imputation, and which can play an important role in the sufficiency of the imputation. Simple choices of $\mathbf{P}_r$ include r leading principal component directions of $\mathbf{X}$ or any r canonical directions of $\mathbf{X}$. Note that under the SS setting, $\mathbf{P}_r$ is effectively known if it only involves the distribution of $\mathbf{X}$, as is true for these choices. We now focus primarily on the case where $\mathbf{P}_r$ also depends on the distribution of Y and hence, is unknown. Such a choice of $\mathbf{P}_r$ is often desirable to ensure that the imputation is as 'sufficient' as possible for predicting Y. Several reasonable choices of such $\mathbf{P}_r$ and their estimation are possible based on sufficient dimension reduction (s.d.r.) methods like Sliced Inverse Regression (SIR) (Li, 1991), Principal Hessian Directions (PHD) (Li, 1992; Cook, 1998), Sliced Average Variance Estimation (SAVE) (Cook and Weisberg, 1991; Cook and Lee, 1999) etc.

In particular, we focus here on SIR where the choice of $\mathbf{P}_r$ is given by: $\mathbf{P}_r^0 = \boldsymbol{\Sigma}^{-\frac{1}{2}}\overline{\mathbf{P}}_r$, with $\overline{\mathbf{P}}_r$ being the r leading eigenvectors of $\mathbb{M} = \mathrm{Var}\{\mathbb{E}(\mathbb{X} \mid Y)\}$, where $\mathbb{X} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu})$, with $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$, denotes the standardized version of $\mathbf{X}$. It is well known (Li, 1991) that these directions lead to an optimal (in some appropriate sense) r-dimensional linear transformation of $\mathbf{X}$ that can be predicted by Y. Apart from these general optimality, they also have deeper implications in the context of s.d.r. We refer the reader to Li (1991) and other relevant references in the s.d.r. literature for further details.

For estimating $\mathbf{P}_r^0$, we consider the SIR algorithm of Li (1991) and further propose a SS modification to it. With $\mathbb{K}$ and $\{\mathcal{L}_k^-, \mathcal{I}_k^-, \widehat{\mathbf{P}}_{r,k}\}_{k=1}^{\mathbb{K}}$ as before, let $(\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k)$ denote the estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ based on $\mathcal{L}_k^-$ and define $\mathbb{X}^{(k)} = \widehat{\boldsymbol{\Sigma}}_k^{-\frac{1}{2}}(\mathbf{X} - \widehat{\boldsymbol{\mu}}_k)$. Then, the original SIR algorithm estimates $\mathbf{P}_r^0$ based on $\mathcal{L}_k^-$ as follows: (i) Divide the range of $\{Y_i\}_{i \in \mathcal{I}_k^-}$ into H slices $\{I_1, .., I_H\}$, where H may depend on $n_{\mathbb{K}}^-$. For $1 \leq h \leq H$, let $\widehat{p}_{h,k}$ denote the proportion of $\{Y_i\}_{i \in \mathcal{I}_k^-}$ in slice I_h; (ii) For each I_h, let $\widehat{\mathbb{M}}_{h,k}$ denote the sample average of the set: $\{\mathbb{X}_i^{(k)} \in \mathcal{L}_k^- : Y_i \in I_h\}$; (iii) Estimate $\mathbb{M}$ as: $\widehat{\mathbb{M}}_k = \sum_{h=1}^{H} \widehat{p}_{h,k} \widehat{\mathbb{M}}_{h,k} \widehat{\mathbb{M}}_{h,k}'$ and $\mathbf{P}_r^0$ as: $\widehat{\mathbf{P}}_{r,k}^0 = \widehat{\boldsymbol{\Sigma}}_k^{-\frac{1}{2}} \widehat{\mathbf{P}}_{r,k}$, where $\widehat{\mathbf{P}}_{r,k}$ denotes the r leading eigenvectors of $\widehat{\mathbb{M}}_k$. However, the SIR algorithm often tends to give unstable estimates of $\mathbf{P}_r^0$, especially for the directions corresponding to the smaller eigenvalues of $\mathbb{M}$. To improve the efficiency in estimating $\mathbf{P}_r^0$, we now propose a semi-supervised SIR (SS-SIR) algorithm as follows.

SS-SIR Algorithm. Given $\{\mathcal{L}_k^-, \mathcal{I}_k^-, \widehat{\mathbf{P}}_{r,k}\}_{k=1}^{\mathbb{K}}$, let $(\widehat{\boldsymbol{\mu}}_k^*, \widehat{\boldsymbol{\Sigma}}_k^*)$ denote the estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ based on $\mathcal{L}_k^- \cup \mathcal{U}$ and define $\mathbb{X}^{(k*)} = \widehat{\boldsymbol{\Sigma}}_k^{*-\frac{1}{2}}(\mathbf{X} - \widehat{\boldsymbol{\mu}}_k)$. Then the SS-SIR proceeds as follows. Step (i) stays the same as in SIR. In step (ii), for each k, and each $j \in \{n+1, ..., n+N\}$, we impute Y_j as $Y^*_{j,k} = Y_{\widehat{i}_{j,k}}$, where $\widehat{i}_{j,k} = \mathrm{argmin}_{i \in \mathcal{I}^-_k} \|\mathbb{X}^{(k*)}_i - \mathbb{X}^{(k*)}_j\|^2$. For each I_h, let $\widehat{\mathbb{M}}^*_{h,k}$ be the sample average of the set: $\{\mathbb{X}^{(k*)}_i \in \mathcal{L}^-_k : Y_i \in I_h\} \cup \{\mathbb{X}^{(k*)}_j \in \mathcal{U} : Y^*_{j,k} \in I_h\}$. Then in step (iii), we estimate $\mathbb{M}$ as: $\widehat{\mathbb{M}}^*_k = \sum_{h=1}^H \widehat{p}_{h,k} \widehat{\mathbb{M}}^*_{h,k} \widehat{\mathbb{M}}^{*'}_{h,k}$ and then, $\mathbf{P}^0_r$ as: $\widehat{\mathbf{P}}^{0*}_{r,k} = \widehat{\boldsymbol{\Sigma}}^{*-\frac{1}{2}}_k \widehat{\mathbf{P}}^*_{r,k}$, where $\widehat{\mathbf{P}}^*_{r,k}$ denotes the r leading eigenvectors of $\widehat{\mathbb{M}}^*_k$.

The SS-SIR algorithm aims to improve the estimation of $\mathbf{P}^0_r$ by making use of $\mathcal{U}$ in step (ii) through a nearest neighbour approximation for the unobserved Y in $\mathcal{U}$ using $\mathcal{L}^-_k$. With $n^-_{\mathbb{K}}$ large enough and $m(\cdot)$ smooth enough, the imputed and the true underlying Y should belong to the same slice with a high probability. Thus, the set of $\mathbb{X}$'s belonging to a particular slice is now 'enriched' and consequently, improved estimation of $\mathbb{M}$ and $\mathbf{P}^0_r$ is expected. The proposed method based on a nearest neighbor approximation is also highly scalable and while other smoothing based approximations may be used, they can be computationally intensive. The SS-SIR algorithm is fairly robust to the choice of H, and $H = O(n^{\frac{1}{2}} \log n)$ seems to give fairly satisfactory performance. The slices may be chosen to have equal width or equal number of observations. For SIR, $n^{\frac{1}{2}}$-consistency of the estimates are well established (Li, 1991; Duan and Li, 1991; Zhu and Ng, 1995) for various formulations under fairly general settings (without any model based assumptions). The theoretical properties of SS-SIR, although not derived here, are expected to follow similarly. Our simulation results (not shown here) further suggest that SS-SIR significantly outperforms SIR, leading to substantially improved estimation of $\boldsymbol{\theta}_0$ from the proposed methods.

6. Numerical Studies.

6.1. *Simulation Studies.* We conducted extensive simulation studies to examine the finite sample performance of our proposed point and interval estimation procedures as well as to compare with existing methods. Throughout we let $n = 500$, $N = 10000$, and considered $p = 2, 10$ and 20. For our CV based methods, we let $\mathbb{K} = 5$. The true values of the target parameter $\boldsymbol{\theta}_0$ were estimated via monte carlo with a large sample size of $50,000$. For each configuration, the results were summarized based on 500 replications. Results for $p = 2$ are summarized in the Supplementary Material, and the discussions below focus primarily on $p = 10$ and 20.

We generated $\mathbf{X} \sim \mathcal{N}_p[\mathbf{0}, I_p]$ and restricted $\mathbf{X}$ to $[-5, 5]^p$ to ensure its boundedness. Given $\mathbf{X} = \mathbf{x}$, we generated $Y \sim \mathcal{N}_1[m(\mathbf{x}), 1]$, where we considered four different choices of $m(\mathbf{x})$:

(i) *Linear*: $m(\mathbf{x}) = \mathbf{x}'\mathbf{b}_p$;
(ii) *Non-linear one component (NL1C):* $m(\mathbf{x}) = (\mathbf{x}'\mathbf{b}_p) + (\mathbf{x}'\mathbf{b}_p)^2$;
(iii) *Non-linear two component (NL2C):* $m(\mathbf{x}) = (\mathbf{x}'\mathbf{b}_p)(1 + \mathbf{x}'\boldsymbol{\delta}_p)$; and
(iv) *Non-linear three component (NL3C):* $m(\mathbf{x}) = (\mathbf{x}'\mathbf{b}_p)(1+\mathbf{x}'\boldsymbol{\delta}_p)+(\mathbf{x}'\boldsymbol{\omega}_p)^2$;

where, for each setting, we considered $\mathbf{b}_p = \mathbf{b}_p^{(1)} \equiv (\mathbf{1}'_{p/2}, \mathbf{0}'_{p/2})'$ and $\mathbf{b}_p = \mathbf{b}_p^{(2)} \equiv \mathbf{1}_p$, and set $\boldsymbol{\delta}_p = (\mathbf{0}'_{p/2}, \mathbf{1}'_{p/2})'$ and $\boldsymbol{\omega}_p = (1, 0, 1, 0, \ldots, 1, 0)'_{p\times 1}$, where for any a, $\mathbf{1}_a = (1, \ldots, 1)'_{a\times 1}$ and $\mathbf{0}_a = (0, \ldots, 0)'_{a\times 1}$. Through appropriate choices of $\mathbf{b}_p$, $\boldsymbol{\delta}_p$ and $\boldsymbol{\omega}_p$, as applicable, these models can incorporate commonly encountered linear, quadratic and interaction effects.

For each setting, we used two choices of the smoothing method: (a) $\mathcal{T} := \mathrm{KS}_{2,\mathbf{P}_2}$ denoting KS with 2-dimensional smoothing over $\mathbf{P}'_r\mathbf{X} \equiv \mathbf{P}'_2\mathbf{X}$, where $\mathbf{P}_2$ was estimated via SIR with $H = 100$ slices of equal width, following which $\{\widehat{m}_k(\mathbf{x}, \widehat{\mathbf{P}}_{r,k})\}_{k=1}^{\mathbb{K}}$ were obtained via KS using a Gaussian kernel; (b) $\mathcal{T} :=$ KM where we let $\mathbf{P}_r = I_p$ and then estimated $\{\widehat{m}_k(\mathbf{x}; I_p)\}_{k=1}^{\mathbb{K}}$ using kernel machine (KM) regression based on a radial basis function (RBF) kernel. Throughout, h for KS, and all tuning parameters for KM were selected via least squares CV. For (a) with $\mathbf{X} \sim \mathcal{N}_p[\mathbf{0}, I_p]$, results from Li (1991) imply that the SNP imputation with $r = 2$ is sufficient for models (i)-(iii), and insufficient for model (iv). For comparison, we also implemented two other SS estimators: the density ratio based "DRESS" estimator of Kawakita and Kanamori (2013) and the estimator of Sokolovska, Cappé and Yvon (2008) called "MSSL" by Kawakita and Kanamori (2013). The density ratio estimation for the DRESS estimator was implemented using either (i) linear bases $\{1, (\mathbf{X}_{[j]})_{j=1}^{p}\}$ (DRESS$_1$); or (ii) cubic bases $\{1, (\mathbf{X}_{[j]}^{d})_{j=1,d=1}^{p,\,3}\}$ (DRESS$_3$).

First, we compare the various estimators with respect to their efficiencies based on empirical mean squared error. In Table 1, we present the efficiencies of the proposed SNP and EASE estimators as well as other SS estimators relative to the OLS. As expected, under model mis-specification, our estimators are substantially more efficient than the OLS with the relative efficiency (RE) as high as near 5 fold when $p = 10$ and 3 fold when $p = 20$, for the non-linear models. The efficiency gain is generally lower for $p = 20$ than for $p = 10$, likely a consequence of overfitting of the non-parametric estimators involved in the SNP imputation for larger p. Comparing EASE to SNP, the EASE generally perform better for both linear and non-linear settings, as expected. Comparing the two smoothers, it appears that $\mathcal{T} :=$ KM generally attains higher efficiency compared to that of $\mathcal{T} := \mathrm{KS}_{2,\mathbf{P}_2}$. This is in part due to the high variability in the SIR direction estimation which impacts performance of the resulting SS estimator in finite samples. Interestingly, none of the existing SS estimators perform well with REs ranging only from about 0.9 to 1.1 across all settings.

(a) $p = 10$

Setting	Models	OLS	SNP	EASE	SNP	EASE	Other SS Estimators		
		(Ref.)	($\mathcal{T} := \mathrm{KS}_{2,\mathbf{P}_2}$)		($\mathcal{T} :=$ KM)		DRESS$_1$	DRESS$_3$	MSSL
(I)	Linear	1	0.895	0.983	0.772	0.985	0.982	0.927	0.982
	NL1C	1	4.481	4.424	4.501	5.543	1.136	1.110	1.135
	NL2C	1	2.683	2.700	4.268	5.055	1.120	1.016	1.119
	NL3C	1	2.772	2.795	4.481	5.560	1.102	1.025	1.103
(II)	Linear	1	0.841	0.989	0.657	0.993	0.981	0.924	0.981
	NL1C	1	4.511	4.585	4.416	5.471	1.132	1.030	1.130
	NL2C	1	3.596	3.634	4.405	5.497	1.127	1.042	1.128
	NL3C	1	3.280	3.301	4.636	5.566	1.110	1.079	1.109

(b) $p = 20$

Setting	Models	OLS	SNP	EASE	SNP	EASE	Other SS Estimators		
		(Ref.)	($\mathcal{T} := \mathrm{KS}_{2,\mathbf{P}_2}$)		($\mathcal{T} :=$ KM)		DRESS$_1$	DRESS$_3$	MSSL
(I)	Linear	1	0.673	0.986	0.740	0.981	0.956	0.866	0.956
	NL1C	1	2.256	2.288	2.680	3.630	1.035	0.920	1.035
	NL2C	1	1.414	1.388	2.661	3.544	1.032	0.922	1.033
	NL3C	1	1.539	1.531	2.605	3.510	1.049	0.931	1.051
(II)	Linear	1	0.519	0.991	0.609	0.989	0.958	0.872	0.958
	NL1C	1	2.290	2.346	2.669	3.660	1.032	0.908	1.031
	NL2C	1	1.899	1.917	2.766	3.963	1.036	0.917	1.036
	NL3C	1	1.937	1.949	2.682	3.702	1.046	0.958	1.046

Table 1: Efficiencies of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ (SNP) and $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ (EASE) using $\mathcal{T} := \mathrm{KS}_{2,\mathbf{P}_2}$ or $\mathcal{T} :=$ KM, as well as DRESS$_1$, DRESS$_3$ and MSSL, relative to $\widehat{\boldsymbol{\theta}}$ (OLS) with respect to the empirical mean squared error (MSE) under models (i), (ii), (iii) and (iv) each with: (I) $\mathbf{b}_p = \mathbf{b}_p^{(1)}$ or, (II) $\mathbf{b}_p = \mathbf{b}_p^{(2)}$.

We next examine the performance of the proposed inference procedures. In Table 2(a) and (b), we present the bias, empirical standard error (ESE), the average of the estimated standard error (ASE) and the coverage probability (CovP) of the 95% CIs for each component of $\boldsymbol{\theta}_0$ when $p = 10$ under the linear and NL2C models. In general, the EASE with both the KS and the KM smoothers have negligible biases although the KM based estimator appears to have slightly lower biases. The ASEs are close to the ESEs and the CovPs are close to the nominal level, suggesting that the variance estimators work well in practice with $\mathbb{K} = 5$. As shown in Table 2(c), the other SS estimators tend to have slightly larger biases and substantially larger standard errors (SEs) compared to our estimators under the NL2C model.

(a) OLS and EASE for the linear model.

Parameter	OLS ($\widehat{\boldsymbol{\theta}}$)		EASE ($\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}; \mathcal{T} := \mathrm{KS}_{2,\mathbf{P}_2}$)				EASE ($\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}; \mathcal{T} := \mathrm{KM}$)			
	Bias	ESE	Bias	ESE	ASE	CovP	Bias	ESE	ASE	CovP
$\alpha_0 = 0$	-0.001	0.043	-0.001	0.043	0.044	0.95	0.000	0.043	0.044	0.96
$\beta_{01} = 1$	0.002	0.044	-0.003	0.045	0.044	0.94	0.004	0.047	0.044	0.93
$\beta_{02} = 1$	-0.001	0.044	-0.005	0.044	0.044	0.94	0.000	0.045	0.044	0.95
$\beta_{03} = 1$	-0.001	0.046	-0.005	0.046	0.044	0.95	-0.004	0.045	0.044	0.94
$\beta_{04} = 1$	-0.002	0.045	-0.006	0.045	0.044	0.94	0.001	0.047	0.044	0.94
$\beta_{05} = 1$	-0.004	0.048	-0.008	0.049	0.044	0.92	-0.001	0.046	0.044	0.95
$\beta_{06} = 0$	-0.000	0.045	-0.001	0.045	0.044	0.94	0.001	0.045	0.044	0.95
$\beta_{07} = 0$	0.003	0.046	0.003	0.046	0.044	0.93	0.001	0.043	0.044	0.96
$\beta_{08} = 0$	-0.001	0.045	-0.001	0.045	0.044	0.95	-0.000	0.048	0.044	0.94
$\beta_{09} = 0$	-0.002	0.047	-0.002	0.048	0.044	0.94	0.000	0.045	0.045	0.95
$\beta_{010} = 0$	0.003	0.045	0.003	0.045	0.044	0.94	-0.002	0.045	0.045	0.94

(b) OLS and EASE for the NL2C model.

Parameter	OLS ($\widehat{\boldsymbol{\theta}}$)		EASE ($\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}; \mathcal{T} := \mathrm{KS}_{2,\mathbf{P}_2}$)				EASE ($\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}; \mathcal{T} := \mathrm{KM}$)			
	Bias	ESE	Bias	ESE	ASE	CovP	Bias	ESE	ASE	CovP
$\alpha_0 = 0$	-0.015	0.239	-0.016	0.146	0.136	0.93	0.013	0.105	0.096	0.93
$\beta_{01} = 1$	0.000	0.260	0.015	0.159	0.160	0.96	-0.004	0.124	0.112	0.93
$\beta_{02} = 1$	-0.004	0.269	0.017	0.173	0.158	0.93	0.010	0.127	0.113	0.93
$\beta_{03} = 1$	-0.015	0.249	0.018	0.156	0.158	0.95	-0.000	0.118	0.113	0.95
$\beta_{04} = 1$	-0.001	0.267	0.016	0.164	0.159	0.94	0.007	0.124	0.113	0.93
$\beta_{05} = 1$	0.013	0.260	0.019	0.164	0.158	0.94	0.002	0.120	0.113	0.94
$\beta_{06} = 0$	-0.010	0.281	0.008	0.164	0.155	0.94	0.005	0.119	0.112	0.94
$\beta_{07} = 0$	0.006	0.277	0.002	0.166	0.155	0.93	0.011	0.116	0.111	0.95
$\beta_{08} = 0$	-0.008	0.277	-0.004	0.167	0.156	0.94	-0.001	0.120	0.112	0.95
$\beta_{09} = 0$	0.002	0.279	0.003	0.160	0.157	0.95	0.007	0.118	0.113	0.95
$\beta_{010} = 0$	-0.008	0.272	0.002	0.160	0.155	0.95	0.004	0.130	0.111	0.91

(c) All other SS estimators for the models in (a) and (b) above.

Linear Model						NL2C Model					
DRESS$_1$		DRESS$_3$		MSSL		DRESS$_1$		DRESS$_3$		MSSL	
Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE
-0.001	0.043	-0.001	0.044	-0.001	0.043	-0.004	0.223	-0.003	0.226	-0.004	0.223
-0.002	0.044	-0.001	0.046	-0.002	0.044	-0.014	0.266	-0.009	0.279	-0.014	0.266
0.000	0.045	0.001	0.046	0.000	0.045	0.005	0.257	0.006	0.266	0.006	0.257
0.006	0.045	0.006	0.047	0.006	0.045	-0.013	0.256	-0.019	0.281	-0.011	0.256
0.003	0.045	0.003	0.046	0.003	0.045	-0.005	0.262	-0.007	0.274	-0.005	0.262
-0.004	0.047	-0.004	0.049	-0.004	0.047	0.002	0.250	-0.007	0.266	0.002	0.252
-0.001	0.045	-0.001	0.046	-0.001	0.045	-0.017	0.239	-0.009	0.247	-0.017	0.239
-0.000	0.048	-0.001	0.050	-0.000	0.048	-0.022	0.260	-0.019	0.270	-0.022	0.260
-0.004	0.043	-0.003	0.044	-0.004	0.043	-0.011	0.241	-0.013	0.261	-0.010	0.241
-0.001	0.048	-0.001	0.049	-0.001	0.048	-0.020	0.256	-0.019	0.259	-0.020	0.256
-0.003	0.047	-0.003	0.049	-0.003	0.047	-0.020	0.252	-0.022	0.269	-0.020	0.252

Table 2: Coordinate-wise bias, ESE, ASE and CovP of EASE, obtained using $\mathcal{T} := \mathrm{KS}_{2,\mathbf{P}_2}$ or $\mathcal{T} := \mathrm{KM}$, for estimating $\boldsymbol{\theta}_0$ under the linear and NL2C models with $p = 10$ and $\mathbf{b}_p = \mathbf{b}_p^{(1)}$. Shown also are the corresponding bias and ESE of the OLS, as well as the DRESS$_1$, DRESS$_3$ and MSSL estimators.

6.2. *Application to EMR Data.* We applied our proposed SS procedures to an EMR study of rheumatoid arthritis (RA), a systemic auto-immune disease (AD), conducted at the Partners HealthCare (Liao et al., 2010). The study cohort consists of 3854 RA patients with blood samples stored. The outcome of interest is the (logarithm of) anti-CCP (antibodies to cyclic citrullinated polypeptide), a biomarker that is often used to determine subtypes of RA. Due to cost constraints, anti-CCP was measured only for a random subset of $n = 355$ patients, thereby leading to a SS set-up. To investigate the validity of the MCAR assumption, we report in Table II in the Supplementary Material summary measures of the distributions in the labeled and unlabeled data for each of the predictors, as well as p-values from various tests for assessing equality of those distributions. The results suggest that the MCAR assumption is appropriate in this study.

We relate the log anti-CCP level to a set of $p = 24$ clinical variables $\mathbf{X}$ related to ADs, including age, gender, race; total counts of codified and/or narrative mentions extracted from physicians' notes via natural language processing (NLP) for various RA related conditions including RA, Lupus, Polymyalgiarheumatica (PmR), Spondyloarthritis (SpA), as well as various RA medications; indicators of seropositivity and radiological evidence of erosion; mentions of rheumatoid factor (RF), as well as anti-CCP positivity from prior medical history. Since the tests for RF and anti-CCP were not always ordered, missing indicators for these variables were also included. All count variables were transformed as: $x \to \log(1 + x)$ to increase stability of the model fitting. All predictors were normalized to have unit variance.

We obtained the OLS, the EASE using both $\mathcal{T} := \mathrm{KS}_{2,\mathbf{P}_2}$ and $\mathcal{T} := \mathrm{KM}$ in the smoothing step, as well as the DRESS_1 estimator for comparison. For EASE, we again used $\mathbb{K} = 5$ and for the $\mathrm{KS}_{2,\mathbf{P}_2}$ smoother, $\mathbf{P}_2$ was obtained using SIR with $H = 80$ slices of equal width. In Table 3, we present the coordinate-wise estimates of the regression parameters along with their estimated SEs and the corresponding p-values based on these estimates. Overall, the point estimators from all methods are quite close to each other. Our proposed EASE, with both KS and KM smoothers, is substantially more efficient than the OLS across all coordinates with efficiency ranging from about 1.4 to 2.4. The DRESS_1 estimator improved estimation for a few coordinates but the efficiency remains comparable to the OLS for most coordinates. This again suggests the advantage of our proposed estimators compared to both OLS and other SS estimators.

Predictors	OLS ($\widehat{\boldsymbol{\theta}}$)			EASE ($\mathrm{KS}_{2,\mathbf{P}_2}$)			RE	DRESS_1			RE	EASE (KM)			RE
	Est	SE	Pval	Est	SE	Pval		Est	SE	Pval		Est	SE	Pval	
Age	.105	.076	.168	.106	.064	.099	1.40	.094	.073	.199	1.09	.104	.064	.103	1.42
Gender	-.032	.059	.589	-.028	.050	.570	1.41	-.027	.058	.638	1.04	-.031	.049	.524	1.44
Race	-.041	.065	.534	-.042	.055	.452	1.40	-.044	.067	.511	.95	-.040	.055	.462	1.41
Lupus	.038	.066	.563	.048	.052	.359	1.59	.021	.063	.731	1.11	.037	.051	.464	1.70
PmR	-.075	.044	.088	-.076	.031	.013	2.07	-.074	.031	.016	2.10	-.075	.030	.014	2.04
RA	.015	.089	.862	.012	.076	.879	1.37	.008	.080	.923	1.23	.016	.075	.832	1.30
SpA	-.137	.102	.177	-.133	.072	.063	2.02	-.128	.075	.089	1.82	-.136	.066	.038	2.37
Other ADs	-.022	.078	.775	-.024	.058	.679	1.79	-.018	.067	.792	1.35	-.022	.056	.692	1.93
Erosion	.076	.070	.278	.078	.059	.184	1.44	.085	.069	.221	1.03	.076	.058	.189	1.47
Seropositivity	.056	.062	.370	.054	.053	.310	1.37	.041	.061	.496	1.05	.055	.052	.296	1.41
Anti-CCP$_{prior}$	.572	.136	.000	.557	.110	.000	1.54	.567	.123	.000	1.23	.568	.107	.000	1.60
Anti-CCP$_{miss}$	.527	.123	.000	.520	.097	.000	1.61	.508	.115	.000	1.15	.523	.096	.000	1.64
RF	.128	.081	.113	.125	.066	.059	1.49	.149	.079	.059	1.05	.127	.066	.054	1.49
RF$_{miss}$	.085	.085	.316	.084	.070	.233	1.46	.137	.080	.088	1.12	.084	.070	.231	1.48
Azathioprine	-.080	.071	.263	-.074	.056	.185	1.62	-.075	.062	.225	1.33	-.079	.053	.132	1.83
Enbrel	.138	.070	.048	.133	.058	.021	1.48	.136	.073	.064	.91	.137	.057	.017	1.49
Gold salts	.138	.050	.006	.136	.043	.002	1.37	.147	.050	.003	1.01	.137	.042	.001	1.40
Humira	-.051	.068	.453	-.049	.057	.391	1.43	-.057	.067	.389	1.03	-.051	.056	.360	1.49
Infliximab	.003	.069	.968	.008	.057	.887	1.50	.000	.067	.994	1.07	.003	.055	.959	1.57
Leflunomide	-.027	.069	.697	-.023	.058	.693	1.40	-.031	.071	.660	.93	-.026	.057	.644	1.45
Methotrexate	-.021	.073	.775	-.024	.061	.699	1.42	-.025	.073	.728	1.01	-.022	.060	.720	1.46
Plaquenil	-.043	.069	.540	-.038	.057	.503	1.47	-.044	.070	.532	.98	-.042	.057	.460	1.48
Sulfasalazine	-.114	.074	.125	-.116	.063	.064	1.39	-.105	.072	.145	1.06	-.113	.061	.065	1.45
Other meds.	-.042	.074	.570	-.052	.060	.385	1.52	-.052	.071	.466	1.10	-.042	.059	.473	1.59

Table 3: Estimates (Est) of the regression coefficients based on OLS, EASE obtained using either $\mathcal{T} := \mathrm{KS}_{2,\mathbf{P}_2}$ or $\mathcal{T} := \mathrm{KM}$, as well as DRESS_1, along with their estimated standard errors (SE) and the corresponding p-values (Pval.) for testing the null effect of each predictor. Shown also are the relative efficiencies (RE) of all the estimators compared to the OLS.

We also estimated the prediction errors (PEs) for each of the fitted linear models based on the aforementioned estimation methods via CV. To remove potential randomness in the CV partitions, we averaged over 10 replications of leave-5-out CV estimates. The PE was about 1.28 for EASE with both smoothers, 1.29 for OLS and 1.30 for DRESS_1. For prediction purposes, we may also directly employ non-parametric estimates of the conditional mean rather than the fitted linear models. The PE in fact is slightly larger when we use $\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$ or $\widehat{\mu}(\mathbf{x}; \widehat{\mathcal{P}}_{r,\mathbb{K}})$. The PE was 1.34 for KS and 1.33 for KM based on $\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$, and 1.30 for KS and 1.28 for KM based on $\widehat{\mu}(\mathbf{x}; \widehat{\mathcal{P}}_{r,\mathbb{K}})$. This confirms that while the linear model may be mis-specified, it may often be preferable to non-parametric models in practice as it may achieve simplicity without substantial loss in prediction performance.

7. Discussion. We have developed in this paper an efficient and adaptive estimation strategy for the SS linear regression problem. The adaptive property possessed by the proposed EASE is crucial for advocating ‘safe’ use of the unlabeled data and is often unaddressed in the existing literature. In general, the magnitude of the efficiency gain with EASE depends on the inherent degree of non-linearity in $\mathbb{E}(Y\,|\,\mathbf{X})$ and the extent of sufficiency of the underlying SNP imputation. In particular, if the imputation is sufficient or the working linear model is correct, $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ is further optimal among a wide class of estimators. We obtained theoretical results along with IF expansions for $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ substantiating all our claims and also validated them based on numerical studies. The double CV method further facilitates accurate inference, overcoming potential over-fitting issues in finite samples due to smoothing. An R code for implementing EASE is available upon request.

The proposed SNP imputation, the key component of EASE, apart from being flexible and scalable, enjoys several useful properties. The refitting step and CV play a crucial role in reducing the bias of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$, and for $\mathcal{T} :=$ KS in particular, eradicate any under-smoothing or higher order kernel requirements: two undesirable, yet often inevitable, conditions required for $n^{\frac{1}{2}}$-consistency of two-step estimators based on a first step of smoothing. Theorem 4.2, apart from showing the distinct behaviour of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$ compared to $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ for $\mathbb{K} \geq 2$, also highlights the key role of CV in completely removing kernel order restrictions, apart from addressing over-fitting issues. The error rates in the results of Theorems 4.1-4.2 are quite sharp and account for any estimation error from $\widehat{\mathbf{P}}_r$. The regularity conditions required are also fairly mild and standard in the literature. The continuity assumption on $\mathbf{X}$ in Sections 3.1 and 4 is mostly for the convenience of proofs, and the results continue to hold for more general $\mathbf{X}$. Lastly, while we have focussed here on linear regression for simplicity, our methods can indeed be easily adapted to other regression problems such as logistic regression for binary outcomes.

When the goal is solely that of prediction, one obviously does not have to employ linear regression models, and models that incorporate non-linear effects can be helpful. For such settings, the estimators $\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$ or $\widehat{\mu}(\mathbf{x}; \widehat{\mathcal{P}}_{r,\mathbb{K}})$, obtained as by-products of our SNP imputation, can themselves serve as potentially useful non-linear predictors. These SNP estimators may substantially outperform naive non-parametric estimators such as a p-dimensional KS estimator, as demonstrated in Table III of the Supplementary Material for the models considered in our simulation studies. In practice, when the covariates are substantially correlated and the dimension of p is not small as in the EMR example, it is unclear whether non-linear models necessarily provide better prediction performance than the linear models. Under such settings, the linear model also has a clear advantage due to its simplicity. Furthermore, while prediction is a vitally important goal of predictive modeling, association analysis under interpretable models is key to clinical studies for discovery research and efficient estimation of the corresponding model parameters remains an important task.

We end with a comment on the *choice of* $\mathbb{K} \geq 2$ in $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$. While (3.11) holds for any $\mathbb{K} \geq 2$, the error term in (3.11) depends on $\mathbb{K}$ through $c_{n_{\mathbb{K}}^-}$ and more precisely, through $\widetilde{c}_{n_{\mathbb{K}}^-} = \mathbb{K}^{\frac{1}{2}} c_{n_{\mathbb{K}}^-}$. Since $\mathbb{K}$ is fixed, $c_{n_{\mathbb{K}}^-}$ and $\widetilde{c}_{n_{\mathbb{K}}^-}$ are asymptotically equivalent. But for a given n, $c_{n_{\mathbb{K}}^-}$ is expected to decrease with $\mathbb{K}$, while $\widetilde{c}_{n_{\mathbb{K}}^-}$ is likely to increase. It is however desirable that both are small since $c_{n_{\mathbb{K}}^-}$ inherently controls the efficiency of the SNP imputation, while $\widetilde{c}_{n_{\mathbb{K}}^-}$ directly controls the bias of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$. Hence, a reasonable choice of $\mathbb{K} \geq 2$ may be based on minimizing: $(c^2_{n_{\mathbb{K}}^-} + \lambda \widetilde{c}^2_{n_{\mathbb{K}}^-})$ for some $\lambda \geq 0$. Since the (first order) asymptotic variance of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ is independent of $\mathbb{K}$, this is equivalent to a penalized minimization of the asymptotic MSE of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ with λ denoting the weightage of the (lower order) bias relative to the (first order) variance. In general, the optimal $\mathbb{K}$ should be inversely related to λ. Conversely, choice of any $\mathbb{K}$ may be viewed to have an associated regularization effect (through λ) resulting in a 'variance-bias trade-off' with smaller $\mathbb{K}$ leading to lower bias at the cost of some efficiency, and higher $\mathbb{K}$ leading to improved efficiency in lieu of some bias. In practice, we find that $\mathbb{K} = 5$ works well, and $\mathbb{K} = 10$ tends to give slightly smaller MSE at the cost of increased bias.

APPENDIX A

A.1. Preliminaries. The following Lemmas A.1-A.3 would be useful in the proofs of the main theorems. The proofs of these lemmas as well as Theorems 3.1, 4.1 and 4.2 can be found in the Supplementary Material.

LEMMA A.1. *Let $\mathbf{Z} \in \mathbb{R}^l$ be any random vector and $\mathbf{g}(\mathbf{Z}) \in \mathbb{R}^d$ be any measurable function of $\mathbf{Z}$, where l and d are fixed. Let $\mathbb{S}_n = \{\mathbf{Z}_i\}_{i=1}^n \perp\!\!\!\perp \mathbb{S}_m = \{\mathbf{Z}_j\}_{j=1}^m$ be two random samples of n and m i.i.d. observations of $\mathbf{Z}$ respectively. Let $\widehat{\mathbf{g}}_n(\cdot)$ be any estimator of $\mathbf{g}(\cdot)$ based on $\mathbb{S}_n$ such that the random sequence: $\widehat{T}_n \equiv sup_{\mathbf{z}\in\boldsymbol{\chi}} \|\widehat{\mathbf{g}}_n(\mathbf{z})\|$ is $O_p(1)$, where $\boldsymbol{\chi} \subseteq \mathbb{R}^l$ denotes the support of $\mathbf{Z}$. Let $\widehat{\mathbf{G}}_{n,m}$ denote the (double) random sequence: $m^{-1}\sum_{\mathbf{Z}_j \in \mathbb{S}_m} \widehat{\mathbf{g}}_n(\mathbf{Z}_j)$, and let $\overline{\mathbf{G}}_n$ denote the random sequence: $\mathbb{E}_{\mathbb{S}_m}(\widehat{\mathbf{G}}_{n,m}) = \mathbb{E}_{\mathbf{Z}}\{\widehat{\mathbf{g}}_n(\mathbf{Z})\}$, where $\mathbb{E}_{\mathbf{Z}}(.)$ denotes expectation w.r.t. $\mathbf{Z} \in \mathbb{S}_m \perp\!\!\!\perp \mathbb{S}_n$, and all expectations involved are assumed to be finite almost surely (a.s.) $[\mathbb{S}_n]$ $\forall$ n. Then: (a) $\mathbf{G}_{n,m} - \overline{\mathbf{G}}_n = O_p(m^{-\frac{1}{2}})$, and (b) as long as $g(.)$ has finite 2^{nd} moments, $m^{-1}\sum_{\mathbf{Z}_j \in \mathbb{S}_m} \mathbf{g}(\mathbf{Z}_j) - \mathbb{E}_{\mathbf{Z}}\{\mathbf{g}(\mathbf{Z})\} = O_p(m^{-\frac{1}{2}})$.*

The next two lemmas would be useful in the proof of Theorem 4.2. They may also be of more general use in other applications that involve controlling empirical process terms indexed by KS estimators. Suppose that Assumption 2.1 (a) holds, and consider the KS framework introduced in Section 4. Let

$l_{\mathbf{P}_r}(\mathbf{w}) = m_{\mathbf{P}_r}(\mathbf{w}) f_{\mathbf{P}_r}(\mathbf{w})$ and $\widetilde{\varphi}^{(\varrho)}_{\mathbf{P}_r}(\mathbf{w}) = (nh^r)^{-1}\sum_{i=1}^n K_h(\mathbf{w}, \mathbf{P}_r'\mathbf{X}_i)Y_i^{\varrho}$, for $\varrho = 0, 1$. Let $\widetilde{f}_{\mathbf{P}_r} = \widetilde{\varphi}^{(0)}_{\mathbf{P}_r}$, $\widetilde{l}_{\mathbf{P}_r} = \widetilde{\varphi}^{(1)}_{\mathbf{P}_r}$, $\widetilde{m}_{\mathbf{P}_r} = \widetilde{l}_{\mathbf{P}_r}/\widetilde{f}_{\mathbf{P}_r}$, $\varphi^{(0)}_{\mathbf{P}_r} = f_{\mathbf{P}_r}$ and $\varphi^{(1)}_{\mathbf{P}_r} = l_{\mathbf{P}_r}$. Let $\varphi^{(\varrho)}(\mathbf{x}; \mathbf{P}_r) = \varphi^{(\varrho)}_{\mathbf{P}_r}(\mathbf{P}_r'\mathbf{x})$ and $\widetilde{\varphi}^{(\varrho)}(\mathbf{x}; \mathbf{P}_r) = \widetilde{\varphi}^{(\varrho)}_{\mathbf{P}_r}(\mathbf{P}_r'\mathbf{x})$ $\forall\ \varrho = 0, 1$. Let $\widetilde{f} = \widetilde{\varphi}^{(0)}$, $\widetilde{l} = \widetilde{\varphi}^{(1)}$ and $\widetilde{m} = \widetilde{l}/\widetilde{f}$. Now, let $\mathbb{P}_n$ denote the empirical probability measure on $\mathbb{R}^p$ based on $\{\mathbf{X}_i\}_{i=1}^n$, and for any measurable function $\boldsymbol{\gamma}(\cdot)$ (possibly vector valued) of $\mathbf{X}$, let $\mathbb{G}_n^*(\boldsymbol{\gamma}) = n^{\frac{1}{2}}\int \boldsymbol{\gamma}(\mathbf{x})(\mathbb{P}_n - \mathbb{P}_{\mathbf{X}})(d\mathbf{x})$.

Lemma A.2. *Consider the set-up introduced above. For any fixed integer $d \geq 1$, let $\boldsymbol{\lambda}(\cdot)$ be any $\mathbb{R}^d$-valued measurable function of $\mathbf{X}$ that is bounded a.s. $[\mathbb{P}_{\mathbf{X}}]$. Define: $b_n^{(1)} = n^{-\frac{1}{2}}h^{-r} + h^q$ and $a_{n,2} = (\log n)^{\frac{1}{2}}(nh^r)^{-\frac{1}{2}} + h^q$. Assume $b_n^{(1)} = o(1)$ for (A.1) and $n^{\frac{1}{2}}a_{n,2}^2 = o(1)$ for (A.2) below. Then, under Assumption 4.1 (i)-(v), and $\forall\ \varrho \in \{0, 1\}$,*

$$\mathbb{G}_n^*[\boldsymbol{\lambda}(\cdot)\{\widetilde{\varphi}^{(\varrho)}(\cdot\,; \mathbf{P}_r) - \varphi^{(\varrho)}(\cdot\,; \mathbf{P}_r)\}] = O_p(b_n^{(1)}) = o_p(1), \quad \text{and} \tag{A.1}$$

$$\mathbb{G}_n^*[\boldsymbol{\lambda}(\cdot)\{\widetilde{m}(\cdot\,; \mathbf{P}_r) - m(\cdot\,; \mathbf{P}_r)\}] = O_p(n^{\frac{1}{2}}a_{n,2}^2) = o_p(1). \tag{A.2}$$

Let $\widehat{\varphi}^{(\varrho)}(\mathbf{x}; \widehat{\mathbf{P}}_r) = (nh^r)^{-1}\sum_{i=1}^n K_h(\widehat{\mathbf{P}}_r'\mathbf{x}, \widehat{\mathbf{P}}_r'\mathbf{X}_i)Y_i^{\varrho}$ $\forall\ \varrho \in \{0, 1\}$, where $\widehat{\mathbf{P}}_r$ is as in section 3.2 and all other notations are the same as in the set-up of Lemma A.2. Let $\widehat{f}(\mathbf{x}; \widehat{\mathbf{P}}_r) = \widehat{\varphi}^{(0)}(\mathbf{x}; \widehat{\mathbf{P}}_r)$ and $\widehat{l}(\mathbf{x}; \widehat{\mathbf{P}}_r) = \widehat{\varphi}^{(1)}(\mathbf{x}; \widehat{\mathbf{P}}_r)$. Then:

Lemma A.3. *Consider the set-up of Lemma A.2. Let $\widehat{\varphi}^{(\varrho)}(\mathbf{x}; \widehat{\mathbf{P}}_r)$ be as above, and let $\boldsymbol{\lambda}(\cdot)$ be as in Lemma A.2. Suppose $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$ for some $\alpha_n = o(1)$. Assume $b_n^{(2)} = o(1)$, where $b_n^{(2)} = \alpha_n + n^{-\frac{1}{2}}\alpha_n h^{-(r+1)} + n^{\frac{1}{2}}\alpha_n^2(h^{-2} + n^{-1}h^{-(r+2)})$. Then, under Assumption 4.1,*

$$\mathbb{G}_n^*[\boldsymbol{\lambda}(\cdot)\{\widehat{\varphi}^{(\varrho)}(\cdot\,; \widehat{\mathbf{P}}_r) - \widetilde{\varphi}^{(\varrho)}(\cdot\,; \mathbf{P}_r)\}] = O_p(b_n^{(2)}) = o_p(1)\ \ \forall\ \varrho \in \{0, 1\}. \tag{A.3}$$

A.2. Proof of Theorem 3.2.

Let $\boldsymbol{\Gamma}_n = \frac{1}{n}\sum_{i=1}^n \overrightarrow{\mathbf{X}}_i\overrightarrow{\mathbf{X}}_i'$, and

$$\mathbf{T}_n^{(1)} = \frac{1}{n}\sum_{i=1}^n \overrightarrow{\mathbf{X}}_i\left\{Y_i - \mu(\mathbf{X}_i; \mathbf{P}_r)\right\}, \mathbf{T}_{n,\mathbb{K}}^{(2)} = \frac{1}{n}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k}\overrightarrow{\mathbf{X}}_i\widehat{\Delta}_k(\mathbf{X}_i; \mathbf{P}_r, \widehat{\mathbf{P}}_{r,k}).$$

Then, using (3.4)-(3.8), it is straightforward to see that:

$$\mathbb{E}[\overrightarrow{\mathbf{X}}\{Y - \mu(\mathbf{X}; \mathbf{P}_r)\}] \equiv \mathbb{E}[\overrightarrow{\mathbf{X}}\{Y - m(\mathbf{X}; \mathbf{P}_r) - \overrightarrow{\mathbf{X}}'\boldsymbol{\eta}_{\mathbf{P}_r}\}] = \mathbf{0}, \quad \text{and} \tag{A.4}$$

$$\boldsymbol{\Gamma}_n\left(\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\eta}_{\mathbf{P}_r}\right) = \mathbf{T}_n^{(1)} - \mathbf{T}_{n,\mathbb{K}}^{(2)}. \tag{A.5}$$

Under (A.4), Assumptions 2.1 (a) and (i), it follows from Lemma A.1 (b) that $\mathbf{T}_n^{(1)} = O_p(n^{-\frac{1}{2}})$. Next, due to assumption (ii) and boundedness of $\mathbf{X}$,

$$\|\mathbf{T}_{n,\mathbb{K}}^{(2)}\| \leq n^{-1} \sum_{k=1}^{\mathbb{K}} \sum_{i \in \mathcal{I}_k} \sup_{\mathbf{x} \in \mathcal{X}} \{\|\overrightarrow{\mathbf{x}}\| \, |\widehat{\Delta}_k(\mathbf{x}; \mathbf{P}_r, \widehat{\mathbf{P}}_{r,k})|\} = O_p(c_{n_{\mathbb{K}}^-}).$$

Finally, under Assumption 2.1 (a), we have: $\mathbf{\Gamma}_n = \mathbf{\Gamma} + O_p(n^{-\frac{1}{2}})$ using Lemma A.1 (b). Further, since $\mathbf{\Gamma}_n$ is invertible a.s., $\mathbf{\Gamma}_n^{-1} = \mathbf{\Gamma}^{-1} + O_p(n^{-\frac{1}{2}})$. Using all these facts, we then have: $(\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\eta}_{\mathbf{P}_r}) = \mathbf{\Gamma}_n^{-1}(\mathbf{T}_n^{(1)} - \mathbf{T}_{n,\mathbb{K}}^{(2)}) = \mathbf{\Gamma}^{-1}(\mathbf{T}_n^{(1)} - \mathbf{T}_{n,\mathbb{K}}^{(2)}) + O_p\{n^{-\frac{1}{2}}(n^{-\frac{1}{2}} + c_{n_{\mathbb{K}}^-})\}$. Thus,

$$(\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\eta}_{\mathbf{P}_r}) = \mathbf{\Gamma}^{-1}(\mathbf{T}_n^{(1)} - \mathbf{T}_{n,\mathbb{K}}^{(2)}) + O_p(n^{-1} + n^{-\frac{1}{2}} c_{n_{\mathbb{K}}^-}). \tag{A.6}$$

Next, let us define:

$$\mathbf{\Gamma}_N = N^{-1} \sum_{j=n+1}^{n+N} \overrightarrow{\mathbf{X}}_j \overrightarrow{\mathbf{X}}'_j, \quad \mathbf{R}_N^{(1)} = N^{-1} \sum_{j=n+1}^{n+N} \overrightarrow{\mathbf{X}}_j \{\mu(\mathbf{X}_j; \mathbf{P}_r) - \overrightarrow{\mathbf{X}}'_j \boldsymbol{\theta}_0\},$$

$$\text{and } \widehat{\mathbf{R}}_{N,n}^{(\mathbb{K})} = N^{-1} \sum_{j=n+1}^{n+N} \overrightarrow{\mathbf{X}}_j \{\widehat{\mu}(\mathbf{X}_j; \widehat{\mathcal{P}}_{r,\mathbb{K}}) - \mu(\mathbf{X}_j; \mathbf{P}_r)\}.$$

Then, using (3.7), we have:

$$\mathbf{\Gamma}_N(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\theta}_0) = N^{-1} \sum_{j=n+1}^{n+N} \overrightarrow{\mathbf{X}}_j [\widehat{\mu}(\mathbf{X}_j; \widehat{\mathcal{P}}_{r,\mathbb{K}}) - \overrightarrow{\mathbf{X}}'_j \boldsymbol{\theta}_0] = \mathbf{R}_N^{(1)} + \widehat{\mathbf{R}}_{N,n}^{(\mathbb{K})}.$$

Next, using (3.4)-(3.8), we have: $\widehat{\mathbf{R}}_{N,n}^{(\mathbb{K})} = \mathbf{\Gamma}_N(\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\eta}_{\mathbf{P}_r}) + \widehat{\mathbf{S}}_{N,n}^{(\mathbb{K})}$, where

$$\widehat{\mathbf{S}}_{N,n}^{(\mathbb{K})} = \mathbb{K}^{-1} \textstyle\sum_{k=1}^{\mathbb{K}} \{N^{-1} \sum_{j=n+1}^{n+N} \overrightarrow{\mathbf{X}}_j \widehat{\Delta}_k(\mathbf{X}_j; \mathbf{P}_r, \widehat{\mathbf{P}}_{r,k})\}.$$

Hence, we have: $\mathbf{\Gamma}_N(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\theta}_0) = \mathbf{\Gamma}_N(\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\eta}_{\mathbf{P}_r}) + \mathbf{R}_N^{(1)} + \widehat{\mathbf{S}}_{N,n}^{(\mathbb{K})}$. Now, under assumptions (i)-(ii) and Assumption 2.1 (a), we have:

$$\text{(I)} \quad \textstyle\sum_{k=1}^{\mathbb{K}} \sup_{\mathbf{x} \in \mathcal{X}} \|\overrightarrow{\mathbf{x}} \widehat{\Delta}_k(\mathbf{x}; \mathbf{P}_r, \widehat{\mathbf{P}}_{r,k})\| = O_p(1),$$

so that using Lemma A.1 (a), $\widehat{\mathbf{S}}_{N,n}^{(\mathbb{K})} = \mathbb{K}^{-1} \sum_{k=1}^{\mathbb{K}} \widehat{\mathbf{S}}_{n,k}^* + O_p(N^{-\frac{1}{2}})$, where $\widehat{\mathbf{S}}_{n,k}^* = \mathbb{E}_{\mathbf{X}}\{\overrightarrow{\mathbf{X}} \widehat{\Delta}_k(\mathbf{X}; \mathbf{P}_r, \widehat{\mathbf{P}}_{r,k})\} \;\forall\, 1 \leq k \leq \mathbb{K}$;

$$\text{(II)} \quad \mathbf{R}_N^{(1)} = \mathbb{E}[\overrightarrow{\mathbf{X}}\{\mu(\mathbf{X}; \mathbf{P}_r) - \overrightarrow{\mathbf{X}}' \boldsymbol{\theta}_0\}] + O_p(N^{-\frac{1}{2}}) = O_p(N^{-\frac{1}{2}})$$

from Lemma A.1 (b) and $\mathbb{E}[\overrightarrow{\mathbf{X}}\{\mu(\mathbf{X};\mathbf{P}_r)-\overrightarrow{\mathbf{X}}'\boldsymbol{\theta}_0\}]=\mathbf{0}$ due to (A.4) and 2.1; and lastly, (III) $\boldsymbol{\Gamma}_N^{-1}=\boldsymbol{\Gamma}^{-1}+O_p(N^{-\frac{1}{2}})$. It then follows from (I)-(III) that

$$\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}-\boldsymbol{\theta}_0=(\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})}-\boldsymbol{\eta}_{\mathbf{P}_r})+\mathbb{K}^{-1}\boldsymbol{\Gamma}^{-1}\sum_{k=1}^{\mathbb{K}}\widehat{\mathbf{S}}^*_{n,k}+O_p(N^{-\frac{1}{2}}). \tag{A.7}$$

Using (A.6) and (3.9) in (A.7), we then have:

$$\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}-\boldsymbol{\theta}_0=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\mathbf{Z}_i;\mathbf{P}_r)-\ \boldsymbol{\Gamma}^{-1}\frac{1}{\mathbb{K}}\sum_{k=1}^{\mathbb{K}}\Big\{\frac{1}{n_{\mathbb{K}}}\sum_{i\in\mathcal{I}_k}\widehat{\mathbf{G}}_k(\mathbf{X}_i)\Big\}+O_p(b_{n,\mathbb{K}}),$$

where $b_{n,\mathbb{K}}=n^{-1}+n^{-\frac{1}{2}}c_{n^-_{\mathbb{K}}}+N^{-\frac{1}{2}}$. It follows, as claimed in (3.10), that

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}-\boldsymbol{\theta}_0)=n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\psi}(\mathbf{Z}_i;\mathbf{P}_r)-\boldsymbol{\Gamma}^{-1}\mathbb{G}_{n,\mathbb{K}}+O_p(c^*_{n,\mathbb{K}})\quad\blacksquare \tag{A.8}$$

We next show that $\mathbb{G}_{n,\mathbb{K}}=O_p(c_{n^-_{\mathbb{K}}})$ for any fixed $\mathbb{K}\geq 2$. To this end, let $\mathbb{T}^{(n)}_k=(n_{\mathbb{K}})^{-\frac{1}{2}}\sum_{i\in\mathcal{I}_k}\widehat{\mathbf{G}}_k(\mathbf{X}_i)$, $\widehat{D}_k=\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{\Delta}_k(\mathbf{x};\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k})|$ and $C=\sup_{\mathbf{x}\in\mathcal{X}}\|\overrightarrow{\mathbf{x}}\|<\infty$. For any subset $\mathcal{A}\subseteq\mathcal{L}$, let $\mathbb{P}_{\mathcal{A}}$ denote the joint distribution of the observations in $\mathcal{A}$, and let $\mathbb{E}_{\mathcal{A}}(\cdot)$ denote expectation w.r.t. $\mathbb{P}_{\mathcal{A}}$. By definition, $\mathbb{G}_{n,\mathbb{K}}=\mathbb{K}^{-\frac{1}{2}}\sum_{k=1}^{\mathbb{K}}\mathbb{T}^{(n)}_k=O_p(c_{n^-_{\mathbb{K}}})$ if and only if given any $\epsilon>0$, $\exists\ M_\epsilon>0$ such that $\mathbb{P}\left(\|\mathbb{G}_{n,\mathbb{K}}\|>M_\epsilon c_{n^-_{\mathbb{K}}}\right)\leq\epsilon\ \forall\ n$. Note that for any $M>0$,

$$\begin{aligned}\mathbb{P}\left(\|\mathbb{G}_{n,\mathbb{K}}\|>Mc_{n^-_{\mathbb{K}}}\right)&\leq\ \mathbb{P}\left(\mathbb{K}^{-\frac{1}{2}}\sum_{k=1}^{\mathbb{K}}\|\mathbb{T}^{(n)}_k\|>Mc_{n^-_{\mathbb{K}}}\right)\\ &\leq\sum_{k=1}^{\mathbb{K}}\mathbb{P}\left(\mathbb{K}^{-\frac{1}{2}}\|\mathbb{T}^{(n)}_k\|>\frac{Mc_{n^-_{\mathbb{K}}}}{\mathbb{K}}\right)\leq\sum_{k=1}^{\mathbb{K}}\sum_{l=1}^{p+1}\mathbb{P}\left\{|\mathbb{T}^{(n)}_{k[l]}|>\frac{Mc_{n^-_{\mathbb{K}}}}{\mathbb{K}^{\frac{1}{2}}(p+1)^{\frac{1}{2}}}\right\}\\ &=\sum_{k=1}^{\mathbb{K}}\sum_{l=1}^{p+1}\mathbb{E}_{\mathcal{L}^-_k}\left[\mathbb{P}_{\mathcal{L}_k}\left\{|\mathbb{T}^{(n)}_{k[l]}|>\frac{Mc_{n^-_{\mathbb{K}}}}{\mathbb{K}^{\frac{1}{2}}(p+1)^{\frac{1}{2}}}\ \middle|\ \mathcal{L}^-_k\right\}\right],\end{aligned} \tag{A.9}$$

where the steps follow from repeated use of Bonferroni's inequality and other standard arguments. Now, conditional on $\mathcal{L}^-_k$ $(\perp\!\!\!\perp\ \mathcal{L}_k,$ with $\mathbb{K}\geq 2)$, $n^{\frac{1}{2}}_{\mathbb{K}}\mathbb{T}^{(n)}_k$ is a centered sum of the i.i.d. random vectors $\{\overrightarrow{\mathbf{X}}_i\widehat{\Delta}_k(\mathbf{X}_i;\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k})\}_{i\in\mathcal{I}_k}$ which, due to assumption (ii) and the compactness of $\mathcal{X}$, are bounded by: $C\widehat{D}_k<\infty$

a.s. $[\mathbb{P}_{\mathcal{L}_k^-}]\ \forall\, k, n$. Hence, applying Hoeffding's inequality to $\mathbb{T}^{(n)}_{k[l]}\ \forall\, l$, we have:

$$\mathbb{P}_{\mathcal{L}_k}\left\{|\mathbb{T}^{(n)}_{k[l]}| > \frac{M c_{n_{\mathbb{K}}^-}}{\mathbb{K}^{\frac{1}{2}}(p+1)^{\frac{1}{2}}} \,\middle|\, \mathcal{L}_k^-\right\} \ \le\ 2\ \exp\left\{-\ \frac{M^2 c^2_{n_{\mathbb{K}}^-}}{2(p+1)\mathbb{K}C^2\widehat{D}_k^2}\right\} \tag{A.10}$$
$$\text{a.s. } [\mathbb{P}_{\mathcal{L}_k^-}]\ \forall\, n;\ \text{for each } k \in \{1,...,\mathbb{K}\} \text{ and } \forall\ 1 \le l \le (p+1).$$

Now, since $\widehat{D}_k = O_p(c_{n_{\mathbb{K}}^-})$, $(c_{n_{\mathbb{K}}^-}/\widehat{D}_k) \ge 0$ is stochastically bounded away from 0. Thus, $\forall\, k$, and for any given $\epsilon > 0$, $\exists\ \delta(k,\epsilon) > 0$ (independent of n) such that: $\mathbb{P}_{\mathcal{L}_k^-}\{(c_{n_{\mathbb{K}}^-}/\widehat{D}_k) \le \delta(k,\epsilon)\} \le \epsilon^*\ \forall\, n$, where $\epsilon^* = \epsilon/\{4\mathbb{K}(p+1)\} > 0$. Let $\widetilde{\delta}(\mathbb{K},\epsilon) = \min\{\delta(k,\epsilon) : k = 1,...,\mathbb{K}\} > 0$ (as $\mathbb{K}$ is fixed). Let $\mathbb{A}(k,\epsilon)$ denote the event: $\{(c_{n_{\mathbb{K}}^-}/\widehat{D}_k) \le \widetilde{\delta}(\mathbb{K},\epsilon)\}$, and let $\mathbb{A}^c(k,\epsilon)$ be its complement. Then, $\mathbb{P}_{\mathcal{L}_k^-}\ \{\mathbb{A}(k,\epsilon)\} \le \epsilon^*$, while on $\mathbb{A}^c(k,\epsilon)$, $(c_{n_{\mathbb{K}}^-}/\widehat{D}_k) > \widetilde{\delta}(\mathbb{K},\epsilon)$. Thus, the bound in (A.10) is dominated by: 2 $\exp[-M^2\widetilde{\delta}^2(\mathbb{K},\epsilon)/\{2(p+1)\mathbb{K}C^2\}]$ on $\mathbb{A}^c(k,\epsilon)$, and trivially by 2 on $\mathbb{A}(k,\epsilon)\ \forall\, k$. Plugging the bound of (A.10) into (A.9) and using all these facts, we then have:

$$\begin{aligned}
\mathbb{P}\left(\|\mathbb{G}_{n,\mathbb{K}}\| > M c_{n_{\mathbb{K}}^-}\right) \ &\le\ \sum_{k=1}^{\mathbb{K}}\sum_{l=1}^{p+1}\mathbb{E}_{\mathcal{L}_k^-}\left[2\ \exp\left\{-\ \frac{M^2 c^2_{n_{\mathbb{K}}^-}}{2(p+1)\mathbb{K}C^2\widehat{D}_k^2}\right\}\right] \\
&= \sum_{k=1}^{\mathbb{K}}\sum_{l=1}^{p+1}\mathbb{E}_{\mathcal{L}_k^-}\left[2\ \exp\left\{-\ \frac{M^2 c^2_{n_{\mathbb{K}}^-}}{2(p+1)\mathbb{K}C^2\widehat{D}_k^2}\right\}\left\{1_{\mathbb{A}^c(k,\epsilon)} + 1_{\mathbb{A}(k,\epsilon)}\right\}\right] \\
&\le \sum_{k=1}^{\mathbb{K}}\sum_{l=1}^{p+1}\left[2\ \exp\left\{-\ \frac{M^2\widetilde{\delta}^2(\mathbb{K},\epsilon)}{2(p+1)\mathbb{K}C^2}\right\}\mathbb{P}_{\mathcal{L}_k^-}\ \{\mathbb{A}^c(k,\epsilon)\} + 2\ \mathbb{P}_{\mathcal{L}_k^-}\ \{\mathbb{A}(k,\epsilon)\}\right] \\
&\le\ 2\mathbb{K}(p+1)\left[\exp\left\{-\ \frac{M^2\widetilde{\delta}^2(\mathbb{K},\epsilon)}{2(p+1)\mathbb{K}C^2}\right\} + \epsilon^*\right]
\end{aligned}$$
$$\le\ \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\ \ \text{(with some suitable choice } M_\epsilon \text{ for } M\text{)}, \tag{A.11}$$

where the last step follows from noting the definition of ϵ^* and choosing M_ϵ to be any M large enough such that 4 $\exp[-\ M^2\widetilde{\delta}^2(\mathbb{K},\epsilon)/\{2(p+1)\mathbb{K}C^2\}] \le \epsilon/\{\mathbb{K}(p+1)\}$. Thus, (A.11) shows $\mathbb{G}_{n,\mathbb{K}} = O_p(c_{n_{\mathbb{K}}^-})$ for any fixed $\mathbb{K} \ge 2$. This further establishes (3.11) and all its associated implications. The proof of Theorem 3.2 is now complete. ■

ACKNOWLEDGEMENTS

The authors would like to thank Dr. James Robins and Dr. Eric Tchetgen Tchetgen for many helpful discussions throughout the progress of this paper, as well as the editor, the anonymous associate editor and the two referees for their useful comments and suggestions that helped significantly in improving and revising the original version of this article.

SUPPLEMENTARY MATERIAL

Supplement to "Efficient and Adaptive Linear Regression in Semi-Supervised Settings". The supplement includes: (i) Supplementary results for the simulation studies and the real data analysis, (ii) Brief discussions on generalization of the proposed SS estimators to MAR settings, (iii) Proof of Lemma A.1, (iv) Proof of Theorem 3.1, (v) Proof of Theorem 4.1, and (vi) Proofs of Lemmas A.2-A.3 and Theorem 4.2.

REFERENCES

Andrews, D. W. K. (1995). Nonparametric Kernel Estimation for Semiparametric Models. *Econometric Theory* **11** 560-586.

Belkin, M., Niyogi, P. and Sindhwani, V. (2006). Manifold Regularization : A Geometric Framework for Learning from Labeled and Unlabeled Examples. *The Journal of Machine Learning Research* **7** 2399-2434.

Castelli, V. and Cover, T. M. (1995). The Exponential Value of Labeled Samples. *Pattern Recognition Letters* **16** 105-111.

Castelli, V. and Cover, T. M. (1996). The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition with an Unknown Mixing Parameter. *IEEE Transactions on Information Theory* **42** 2102-2117.

Chapelle, O., Schölkopf, B. and Zien, A. (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA, USA.

Cook, R. D. (1998). Principal Hessian Directions Revisited (with Discussion). *Journal of the American Statistical Association* **93** 84-100.

Cook, R. D. and Lee, H. (1999). Dimension Reduction in Binary Response Regression. *Journal of the American Statistical Association* **94** 1187-1200.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced Inverse Regression" by K.-C. Li. *Journal of the American Statistical Association* **86** 328-332.

Cozman, F. G. and Cohen, I. (2001). Unlabeled Data Can Degrade Classification Performance of Generative Classifiers. Technical Report No. HPL-2001-234, HP Laboratories, Palo Alto, CA, USA.

Cozman, F. G., Cohen, I. and Cirelo, M. C. (2003). Semi-Supervised Learning of Mixture Models. In *Proceedings of the Twentieth ICML* 99-106.

Duan, N. and Li, K.-C. (1991). Sliced Regression: A Link-Free Regression Method. *The Annals of Statistics* **19** 505-530.

Hansen, B. E. (2008). Uniform Convergence Rates for Kernel Estimation with Dependent Data. *Econometric Theory* **24** 726-748.

Kawakita, M. and Kanamori, T. (2013). Semi-Supervised Learning with Density-Ratio Estimation. *Machine Learning* **91** 189-209.

Kohane, I. S. (2011). Using Electronic Health Records to Drive Discovery in Disease Genomics. *Nature Reviews Genetics* **12** 417-428.

Lafferty, J. D. and Wasserman, L. (2007). Statistical Analysis of Semi-Supervised Regression. *Advances in Neural Information Processing Systems* **20** 801-808.

Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association* **86** 316-327.

Li, K.-C. (1992). On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma. *Journal of the American Statistical Association* **87** 1025-1039.

Liao, K. P., Cai, T., Gainer, V. et al. (2010). Electronic Medical Records for Discovery Research in Rheumatoid Arthritis. *Arthritis Care and Research* **62** 1120-1127.

Masry, E. (1996). Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates. *Journal of Time Series Analysis* **17** 571-600.

Newey, W. K. (1994). Kernel Estimator of Partial Means and a Generalized Variance Estimator. *Econometric Theory* **10** 1-21.

Newey, W. K., Hsieh, F. and Robins, J. (1998). Undersmoothing and Bias Corrected Functional Estimation. Technical Report No. 98-17, Dept. of Economics, MIT, USA.

Newey, W. K. and McFadden, D. (1994). Large Sample Estimation and Hypothesis Testing. *Handbook of Econometrics* **4** 2111–2245.

Nigam, K. P. (2001). Using Unlabeled Data to Improve Text Classification. PhD thesis, Carnegie Mellon University, USA. CMU-CS-01-126.

Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents Using EM. *Machine Learning* **39** 103-134.

Seeger, M. (2002). Learning with Labeled and Unlabeled Data. Technical Report No. EPFL-REPORT-161327, University of Edinburgh, UK.

Sokolovska, N., Cappé, O. and Yvon, F. (2008). The Asymptotics of Semi-Supervised Learning in Discriminative Probabilistic Models. In *Proceedings of the Twenty Fifth ICML* 984-991.

Zhang, T. and Oles, F. J. (2000). The Value of Unlabeled Data for Classification Problems. In *Proceedings of the Seventeenth ICML* 1191-1198.

Zhu, X. (2005). Semi-Supervised Learning through Graphs. PhD thesis, Carnegie Mellon University, USA. CMU-LTI-05-192.

Zhu, X. (2008). Semi-Supervised Learning Literature Survey. Technical Report No. 1530, Computer Sciences, University of Wisconsin-Madison, USA.

Zhu, L.-X. and Ng, K. W. (1995). Asymptotics of Sliced Inverse Regression. *Statistica Sinica* **5** 727-736.

SUPPLEMENT TO "EFFICIENT AND ADAPTIVE LINEAR REGRESSION IN SEMI-SUPERVISED SETTINGS"

BY ABHISHEK CHAKRABORTTY AND TIANXI CAI

Harvard University

I. NUMERICAL STUDIES: SUPPLEMENTARY RESULTS

I.1. Simulation Results for $\mathbf{p = 2}$. For $p = 2$, we investigated three choices of $m(\mathbf{x})$ as follows:

(Linear): $m(\mathbf{x}) = x_1 + x_2$;
(NL-$\mathrm{I}_{\lambda^{(k)}}$): $m(\mathbf{x}) = x_1 + x_2 + \lambda^{(k)} x_1 x_2$ for $\lambda^{(1)} = 0.5$ and $\lambda^{(2)} = 1$; and
(NL-$\mathrm{Q}_{\gamma^{(k)}}$): $m(\mathbf{x}) = x_1 + x_2 + \gamma^{(k)}(x_1^2 + x_2^2)$ for $\gamma^{(1)} = 0.3$ and $\gamma^{(2)} = 1$.

Since the dimension is low, we implemented EASE using the KS and KM smoothers with $\mathbf{P}_2 = I_2$ for both, i.e. without any dimension reduction. For comparison, the other SS estimators were also obtained. In Table I, we summarize the efficiencies of all the estimators relative to OLS, based on the empirical mean squared error (MSE), where for any estimator $\widetilde{\boldsymbol{\theta}}$, the empirical MSE is summarized as $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2$ averaged over the 500 replications.

Models	OLS	SNP	EASE	SNP	EASE	Other SS Estimators		
	(Ref.)	($\mathcal{T}$:= KS)		($\mathcal{T}$:= KM)		DRESS$_1$	DRESS$_3$	MSSL
Linear	1	0.897	0.995	0.920	0.988	0.993	0.963	0.993
NL-$\mathrm{I}_{\lambda^{(1)}}$	1	1.229	1.243	1.338	1.355	1.072	1.039	1.072
NL-$\mathrm{I}_{\lambda^{(2)}}$	1	2.261	2.261	2.301	2.267	1.217	1.181	1.216
NL-$\mathrm{Q}_{\gamma^{(1)}}$	1	2.241	2.215	2.500	2.550	1.187	2.063	1.187
NL-$\mathrm{Q}_{\gamma^{(2)}}$	1	4.096	4.144	4.612	4.641	1.352	3.217	1.352

Table I: Efficiencies of SNP and EASE, obtained using $\mathcal{T}$:= KS or KM, as well as DRESS$_1$, DRESS$_3$ and MSSL, relative to OLS with respect to the empirical MSE under the various models considered with $p = 2$.

For this setting, all estimators have comparable efficiency under the linear model, as expected. Under the non-linear models, the EASE estimators are substantially more efficient than the OLS and also more efficient than the other SS estimators. For the non-linear models with quadratic effects, the DRESS$_3$ is also substantially more efficient than the OLS while our EASE estimator performs even better. For the non-linear models with interaction effects, the efficiency gain was very modest when employing existing SS estimation procedures while it was quite substantial for EASE.

I.2. Supplementary Results for the Data Example. We present in Table II some summary measures of the distributions in the labeled and unlabeled data for each of the predictors in the data example, and also report p-values for diagnostic tests aimed at detecting any possible differences in the labeled and unlabeled data distributions for each of the predictors.

Predictors	Labeled Data		Unlabeled Data		P-values from Diagnostic Tests		
	Mean	Sd	Mean	Sd	T-test	Wilcoxon Test	PS Model
Age	4.090	0.241	4.070	0.272	0.151	0.373	0.475
Gender	0.786	0.411	0.799	0.401	0.556	0.547	0.574
Race	0.696	0.461	0.673	0.469	0.371	0.378	0.456
Lupus	0.230	0.520	0.251	0.600	0.461	0.877	0.689
PmR	0.057	0.336	0.078	0.382	0.269	0.326	0.255
RA	4.171	1.071	4.084	1.079	0.144	0.055	0.255
SpA	0.073	0.343	0.066	0.313	0.716	0.932	0.780
Other ADs	0.251	0.642	0.271	0.690	0.582	0.780	0.540
Erosion	0.577	0.495	0.567	0.494	0.709	0.701	0.743
Seropositivity	0.369	0.483	0.395	0.489	0.331	0.335	0.231
Anti-CCP$_{prior}$	0.386	0.629	0.405	0.718	0.592	0.471	0.564
Anti-CCP$_{miss}$	0.645	0.479	0.610	0.488	0.192	0.198	0.089
RF	0.949	0.772	0.897	0.846	0.227	0.184	0.199
RF$_{miss}$	0.307	0.462	0.324	0.468	0.516	0.520	0.410
Azathioprine	0.121	0.397	0.137	0.419	0.463	0.449	0.734
Enbrel	0.738	0.858	0.722	0.822	0.740	0.921	0.574
Gold salts	0.346	0.568	0.336	0.555	0.729	0.791	0.962
Humira	0.856	0.833	0.917	0.835	0.193	0.190	0.177
Infliximab	0.386	0.673	0.400	0.672	0.711	0.607	0.789
Leflunomide	0.549	0.740	0.555	0.743	0.895	0.912	0.973
Methotrexate	1.417	0.638	1.389	0.669	0.435	0.584	0.549
Plaquenil	0.248	0.464	0.273	0.496	0.331	0.463	0.398
Sulfasalazine	0.535	0.752	0.554	0.734	0.661	0.458	0.724
Other meds.	0.163	0.380	0.189	0.403	0.223	0.192	0.322
(Intercept)	–	–	–	–	–	–	0.000

Table II: Comparison of the means and standard deviations (sd) from the labeled and unlabeled data for each predictor in the data example. Shown also are the p-values obtained from various diagnostic tests, testing for possible differences in the distributions of each of the predictors in the labeled and unlabeled data, including a two-sample T-test (with possibly unequal variances in the two populations), a Wilcoxon rank sum test, and a test obtained by fitting a parametric logistic regression model for the propensity score (PS) of missingness, with all the predictors included as covariates, and then testing for the null effect of each of the predictors in the fitted model.

I.3. Simulation Results on the Prediction Error. In Table III, we present the out-of-sample mean squared prediction error for various SNP imputation estimators under all the models considered in the simulation studies for $p = 2$, 10 and 20. The results suggest that the SNP imputation estimators for both KS and KM based smoothers perform substantially better than the naive non-parametric estimator based on a p-dimensional kernel smoothing.

(a) $p = 2$

Models	$\mathcal{T} := \text{KS}$		$\mathcal{T} := \text{KM}$	
	$\widehat{m}_{\mathcal{T}}$	$\widehat{\mu}_{\mathcal{T}}$	$\widehat{m}_{\mathcal{T}}$	$\widehat{\mu}_{\mathcal{T}}$
Linear	0.36	0.35	0.35	0.35
NL-$\text{I}_{\lambda^{(1)}}$	0.34	0.33	0.32	0.32
NL-$\text{I}_{\lambda^{(2)}}$	0.29	0.28	0.27	0.27
NL-$\text{Q}_{\gamma^{(1)}}$	0.29	0.29	0.27	0.27
NL-$\text{Q}_{\gamma^{(2)}}$	0.20	0.20	0.17	0.18

(b) $p = 10$ and 20

		$p = 10$					$p = 20$				
	Models	$\mathcal{T} := \text{KS}_{2,\mathbf{P}_2}$		$\mathcal{T} := \text{KM}$		KS_p	$\mathcal{T} := \text{KS}_{2,\mathbf{P}_2}$		$\mathcal{T} := \text{KM}$		KS_p
		$\widehat{m}_{\mathcal{T}}$	$\widehat{\mu}_{\mathcal{T}}$	$\widehat{m}_{\mathcal{T}}$	$\widehat{\mu}_{\mathcal{T}}$		$\widehat{m}_{\mathcal{T}}$	$\widehat{\mu}_{\mathcal{T}}$	$\widehat{m}_{\mathcal{T}}$	$\widehat{\mu}_{\mathcal{T}}$	
(I)	Linear	0.186	0.174	0.214	0.204	0.260	0.143	0.099	0.130	0.119	0.698
	NL1C	0.143	0.131	0.147	0.154	0.290	0.337	0.332	0.280	0.317	0.921
	NL2C	0.241	0.228	0.150	0.156	0.554	0.557	0.566	0.271	0.299	1.000
	NL3C	0.275	0.254	0.138	0.146	0.458	0.543	0.552	0.282	0.327	0.974
(II)	Linear	0.106	0.096	0.133	0.126	0.313	0.095	0.054	0.083	0.073	0.755
	NL1C	0.147	0.131	0.135	0.143	0.658	0.341	0.336	0.283	0.331	1.095
	NL2C	0.179	0.163	0.135	0.142	0.487	0.423	0.422	0.272	0.302	0.980
	NL3C	0.218	0.199	0.137	0.144	0.527	0.424	0.424	0.282	0.318	0.996

Table III: Mean squared prediction errors (PEs), relative to $\text{Var}(Y)$, of the SNP imputation estimators $\widehat{m}_{\mathcal{T}} = \widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$ and $\widehat{\mu}_{\mathcal{T}} = \widehat{\mu}(\mathbf{x}; \widehat{\mathcal{P}}_{r,\mathbb{K}})$, with $\mathcal{T} := \text{KS}_{2,\mathbf{P}_2}$ or $\mathcal{T} := \text{KM}$, under the various models discussed in the simulation studies. Shown also are the corresponding PEs for the fully non-parametric KS estimator, KS_p, for comparison, in the case of $p = 10$ and 20.

II. GENERALIZATION TO THE MISSING AT RANDOM (MAR) CASE

Our SS estimation methods proposed so far assume that the underlying Y for subjects in $\mathcal{U}$ are MCAR, a standard (and often implicit) assumption made in SSL. In this section, we provide some discussions on possible generalizations of our SS methods to the MAR case. Such generalizations might be desirable for settings where the availability of Y is not determined by design. To this end, let $\bar{N} = N + n$ denote the sample size of the entire data $\mathbb{S} = \mathcal{L} \cup \mathcal{U}$. Then, $\mathbb{S} = \{\mathbf{Z}_i \equiv (R_i, R_i Y_i, \mathbf{X}_i) : i = 1, \ldots, \bar{N}\}$ consists of $\bar{N}$ i.i.d. realizations of $\mathbf{Z} = (R, RY, \mathbf{X})$, where $R \in \{0, 1\}$ denotes the indicator of Y being observed. As opposed to the stronger MCAR setting with the assumption $R \perp\!\!\!\perp (Y, \mathbf{X})$ and the probability law of $\mathbb{S}$ being determined by the law of $(Y, \mathbf{X})$, we now have: under the MAR setting, $R \perp\!\!\!\perp Y \mid \mathbf{X}$ and the probability law of $\mathbb{S}$ is determined by $\mathbb{P}_{\mathbf{Z}}$, the law of $\mathbf{Z}$. For notational ease, we also let $\mathbb{P}_{\bar{N}}$ denote the empirical measure for $\mathbb{S}$, and for any function $\mathbf{e}(\cdot)$ of $\mathbf{Z}$, possibly random and vector-valued, we let $\mathbb{P}_{\bar{N}}(\mathbf{e}) = \bar{N}^{-1} \sum_{i=1}^{\bar{N}} \mathbf{e}(\mathbf{Z}_i)$, and $\mathbb{P}_{\mathbf{Z}}(\mathbf{e}) = \mathbb{E}_{\mathbf{Z}}\{\mathbf{e}(\mathbf{Z})\} = \int \mathbf{e}(\mathbf{z}) d\mathbb{P}_{\mathbf{Z}}(\mathbf{z})$.

Under a SS set-up as above, we have: $n = \sum_{i=1}^{\bar{N}} R_i$ is a random quantity and $n/\bar{N} \to 0$ in probability. It is important to note that $\pi_{\bar{N}} \equiv \mathbb{P}(R = 1)$ must depend on $\bar{N}$. Let $\pi_{\bar{N}}(\mathbf{X}) = \mathbb{P}(R = 1 \mid \mathbf{X})$ be the "propensity score", assumed to be strictly greater than 0 almost surely (a.s.) for any given $\bar{N}$ and let $b_{\bar{N}} = [\mathbb{E}\{\pi_{\bar{N}}^{-2}(\mathbf{X})\}]^{-\frac{1}{2}}$. Then, under the above set-up, we assume that $\mathbb{E}\{\pi_{\bar{N}}(\mathbf{X})\} = \pi_{\bar{N}} \to 0$, $b_{\bar{N}} \to 0$ and $\bar{N} b_{\bar{N}} \to \infty$ as $\bar{N} \to \infty$. This decaying sampling probability is the main factor that distinguishes SSL from standard missing data problems and contributes to the complexity of devising and analyzing an SS estimator which, even for the MCAR setting was seen to have a convergence rate of $(\bar{N} b_{\bar{N}})^{-\frac{1}{2}}$, rather than $\bar{N}^{-\frac{1}{2}}$, with $b_{\bar{N}} = (n/\bar{N})$. For simplicity, we shall first assume that $\pi_{\bar{N}}(\mathbf{X})$ is known and next detail how we may extend our proposed procedures to obtain SS estimators of $\boldsymbol{\theta}_0$, the solution to: $\boldsymbol{\phi}(\boldsymbol{\theta}) \equiv \mathbb{E}\{\overrightarrow{\mathbf{X}}(Y - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta})\} = \mathbf{0}$, under the MAR setting.

To derive an efficient SS estimator of $\boldsymbol{\theta}_0$ based on $\mathbb{S}$ under MAR, we first note that our proposed SNP estimator $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r, \mathbb{K})}$ in Section 3.2 remains valid even under the MAR setting, whenever the imputation is sufficient i.e. $\mu(\mathbf{X}; \mathbf{P}_r)$ equals the true conditional mean $m(\mathbf{X}) = \mathbb{E}(Y \mid \mathbf{X})$. For the general case allowing for insufficient imputation, we need to modify our SNP imputation to account for the MAR setting. To this end, we note that

$$\boldsymbol{\phi}(\boldsymbol{\theta}) \equiv \mathbb{E}\{\overrightarrow{\mathbf{X}}(Y - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta})\} = \mathbb{E}\left\{\frac{R}{\pi_{\bar{N}}(\mathbf{X})} \overrightarrow{\mathbf{X}}(Y - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta})\right\}, \tag{1}$$

$$= \mathbb{E}[\overrightarrow{\mathbf{X}}\{m(\mathbf{X}) - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta}\}], \tag{2}$$

$$= \mathbb{E}[\overrightarrow{\mathbf{X}}\{m(\mathbf{X}) - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta}\}] + \mathbb{E}\left[\frac{R}{\pi_{\bar{N}}(\mathbf{X})} \overrightarrow{\mathbf{X}}\{Y - m(\mathbf{X})\}\right]. \tag{3}$$

More generally, for any $\mu(\cdot) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}})$ satisfying:

$$\mathbb{E}\left[\frac{R}{\pi_{\bar{N}}(\mathbf{X})} \overrightarrow{\mathbf{X}}\{Y - \mu(\mathbf{X})\}\right] \equiv \mathbb{E}[\overrightarrow{\mathbf{X}}\{m(\mathbf{X}) - \mu(\mathbf{X})\}] = \mathbf{0}, \tag{4}$$

it is easy to see that the following representation of $\boldsymbol{\phi}(\cdot)$ holds under MAR:

$$\boldsymbol{\phi}(\boldsymbol{\theta}) = \mathbb{E}[\overrightarrow{\mathbf{X}}\{\mu(\mathbf{X}) - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta}\}] + \mathbb{E}\left[\frac{R}{\pi_{\bar{N}}(\mathbf{X})} \overrightarrow{\mathbf{X}}\{Y - \mu(\mathbf{X})\}\right]. \tag{5}$$

Let $\widehat{\mu}(\cdot)$ be any estimator of $\mu(\cdot)$ based on $\mathbb{S}$. Motivated by (5), we may then devise an SSL estimator of $\boldsymbol{\theta}_0$, $\widehat{\boldsymbol{\theta}}_{\mathrm{MAR}} \equiv \widehat{\boldsymbol{\theta}}_{\mathrm{MAR},\, \mu(\cdot)}$, as the solution to:

$$(6) \qquad \widehat{\boldsymbol{\phi}}_{\bar{N}}(\boldsymbol{\theta}) \equiv \bar{N}^{-1} \sum_{i=1}^{\bar{N}} \left[\overrightarrow{\mathbf{X}}_i \{\widehat{\mu}(\mathbf{X}_i) - \overrightarrow{\mathbf{X}}_i' \boldsymbol{\theta}\} + \frac{R_i}{\pi_{\bar{N}}(\mathbf{X}_i)} \overrightarrow{\mathbf{X}}_i \{Y_i - \widehat{\mu}(\mathbf{X}_i)\} \right].$$

Then, letting $\boldsymbol{\Gamma}_{\bar{N}} = \bar{N}^{-1} \sum_{i=1}^{\bar{N}} \overrightarrow{\mathbf{X}}_i \overrightarrow{\mathbf{X}}_i'$, it is straightforward to show that:

$$(7) \qquad \boldsymbol{\Gamma}_{\bar{N}} \left(\widehat{\boldsymbol{\theta}}_{\mathrm{MAR}} - \boldsymbol{\theta}_0 \right) = \mathbb{P}_{\bar{N}}(\mathbf{T}) + \mathbb{P}_{\bar{N}}(\mathbf{S}) - \mathbb{P}_{\bar{N}}(\widehat{\mathbf{e}}), \text{ where}$$

where $\mathbf{T}(\mathbf{Z}) = \overrightarrow{\mathbf{X}}\{\mu(\mathbf{X}) - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta}_0\}$, $\mathbf{S}(\mathbf{Z}) = \{R/\pi_{\bar{N}}(\mathbf{X})\}\overrightarrow{\mathbf{X}}\{Y - \mu(\mathbf{X})\}$, and

$$\widehat{\mathbf{e}}(\mathbf{Z}) \equiv \left\{ \frac{R}{\pi_{\bar{N}}(\mathbf{X})} - 1 \right\} \overrightarrow{\mathbf{X}}\{\widehat{\mu}(\mathbf{X}) - \mu(\mathbf{X})\}.$$

Convergence rates of the terms in (7) need careful analysis as the asymptotics here is *non-standard*, with the dominating rate being slower than $N^{-\frac{1}{2}}$. To this end, note that $\mathbb{P}_{\bar{N}}(\mathbf{T})$ is a simple centered i.i.d. average of variables with bounded variance. Hence, $\mathbb{T}_{\bar{N}} = O(N^{-\frac{1}{2}})$ indeed. On the other hand, $\mathbb{P}_{\bar{N}}(\mathbf{S})$ has a slower convergence rate since the variance of $\mathbf{S}$, $\mathbf{V}_{\bar{N}}$, diverges due to the $\pi_{\bar{N}}(\cdot) \downarrow 0$ appearing in the denominator. Under mild moment conditions, it can be shown that $b_{\bar{N}}\mathbf{V}_{\bar{N}}$ converges to a positive definite matrix $\mathbf{V}$ with $\|\mathbf{V}\| < \infty$. Hence, using concentration inequalities, and assuming $\bar{N}b_{\bar{N}} \to \infty$, it can be shown that the convergence rate of $\mathbb{P}_{\bar{N}}(\mathbf{S})$ is $O\{(Nb_{\bar{N}})^{-\frac{1}{2}}\}$. Further, using CLT for triangular arrays, it can be shown under suitable conditions that $(Nb_{\bar{N}})^{\frac{1}{2}}\mathbb{P}_{\bar{N}}(\mathbf{S}) \xrightarrow{d} \mathcal{N}_{(p+1)}[\mathbf{0}, \mathbf{V}]$. Lastly, to control the term $\mathbb{P}_{\bar{N}}(\widehat{\mathbf{e}})$, note that $\mathbb{P}_{\mathbf{Z}}(\widehat{\mathbf{e}}) = \mathbf{0}$. Therefore, $\mathbb{G}_{\bar{N}}$ is a *centered* empirical process indexed by $\widehat{\mu}(\cdot) - \mu(\cdot)$. Hence, as long as $\mathbb{E}_{\mathbf{X}}[\{\widehat{\mu}(\mathbf{X}) - \mu(\mathbf{X})\}^2] \xrightarrow{P} 0$, and $\widehat{\mu}(\cdot) - \mu(\cdot)$ lies in a $\mathbb{P}-$Donsker class with probability $\to 1$, it can be shown using results from empirical process theory that $(Nb_{\bar{N}})^{\frac{1}{2}}\mathbb{P}_{\bar{N}}(\widehat{\mathbf{e}}) = o_p(1)$. Finally, note that $\boldsymbol{\Gamma}_{\bar{N}} \succ 0$ a.s., and $\boldsymbol{\Gamma}_{\bar{N}}^{-1} = \boldsymbol{\Gamma}^{-1} + O_p(N^{-\frac{1}{2}})$. Hence, under suitable regularity conditions, we have:

$$\begin{aligned} (\bar{N}b_{\bar{N}})^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{\mathrm{MAR}} - \boldsymbol{\theta}_0) &= (\bar{N}b_{\bar{N}})^{\frac{1}{2}}\boldsymbol{\Gamma}^{-1}\frac{1}{\bar{N}}\sum_{i=1}^{\bar{N}} \mathbf{S}(\mathbf{Z}_i) + o_p(1) \\ &\xrightarrow{d} \mathcal{N}_{(p+1)}[\mathbf{0}, \boldsymbol{\Gamma}^{-1}\mathbf{V}\boldsymbol{\Gamma}^{-1}], \quad \text{with } \mathbf{V} \text{ as defined above.} \end{aligned}$$

Having now provided an abstract sketch of the construction of the estimators and their properties, we next briefly discuss the *choice* of the 'imputation' function $\mu(\cdot)$, and its estimator $\widehat{\mu}(\cdot)$ inherent in the construction of $\widehat{\boldsymbol{\theta}}_{\text{MAR}} \equiv \widehat{\boldsymbol{\theta}}_{\text{MAR},\,\mu(\cdot)}$. With $(r, \mathbf{P}_r, \widehat{\mathbf{P}}_r)$ as defined in Section 3.2 and $\{K(\cdot), h, K_h(\cdot,\cdot)\}$ as in Section 4, we may modify the SNP estimator in Section 3.2 under the MAR setting as follows: consider $\mu(\mathbf{X}) \equiv \mu(\mathbf{X};\mathbf{P}_r) = m(\mathbf{X};\mathbf{P}_r) + \overrightarrow{\mathbf{X}}'\boldsymbol{\eta}_{\mathbf{P}_r}$, where $m(\mathbf{X};\mathbf{P}_r) = \mathbb{E}(Y\,|\,\mathbf{P}_r'\mathbf{X})$, and $\boldsymbol{\eta}_{\mathbf{P}_r}$ satisfies

$$\mathbb{E}\left[\frac{R}{\pi_{\bar{N}}(\mathbf{X})}\overrightarrow{\mathbf{X}}\{Y - m(\mathbf{X};\mathbf{P}_r) - \overrightarrow{\mathbf{X}}'\boldsymbol{\eta}_{\mathbf{P}_r}\}\right] = \mathbf{0}.$$

This will ensure that (4) holds. Then, we may estimate $\mu(\mathbf{X};\mathbf{P}_r)$ as $\widehat{\mu}(\mathbf{X};\widehat{\mathbf{P}}_r) = \widehat{m}(\mathbf{X};\widehat{\mathbf{P}}_r) + \overrightarrow{\mathbf{X}}'\widehat{\boldsymbol{\eta}}_{\mathbf{P}_r}$, where

$$\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r) \;=\; \frac{\sum_{i=1}^{\bar{N}} \frac{R_i}{\pi_{\bar{N}}(\mathbf{X}_i)} Y_i K_h(\widehat{\mathbf{P}}_r'\mathbf{X}_i, \widehat{\mathbf{P}}_r'\mathbf{x})}{\sum_{i=1}^{\bar{N}} \frac{R_i}{\pi_{\bar{N}}(\mathbf{X}_i)} K_h(\widehat{\mathbf{P}}_r'\mathbf{X}_i, \widehat{\mathbf{P}}_r'\mathbf{x})}, \quad \text{and}$$

$$\widehat{\boldsymbol{\eta}}_{\mathbf{P}_r} \text{ satisfies: } \bar{N}^{-1}\sum_{i=1}^{\bar{N}} \frac{R_i}{\pi_{\bar{N}}(\mathbf{X}_i)} \overrightarrow{\mathbf{X}}_i\{Y_i - \widehat{m}(\mathbf{X}_i;\widehat{\mathbf{P}}_r) - \overrightarrow{\mathbf{X}}_i'\widehat{\boldsymbol{\eta}}_{\mathbf{P}_r}\} = \mathbf{0}.$$

Thus, to accommodate MAR, one essentially needs to implement appropriately *weighted* versions of both the smoothing and the re-fitting steps in our original SNP imputation. Of course, while we have chosen the smoothing method $\mathcal{T}$ to be the weighted KS here for illustration, other reasonable choices of $\mathcal{T}$ such as an appropriately weighted KM may also be used. Under MCAR, with $\pi_{\bar{N}}(\mathbf{X}) \equiv \pi_{\bar{N}} \equiv n/\bar{N}$ and $b_{\bar{N}} = (n/\bar{N})$, the estimator $\widehat{\boldsymbol{\theta}}_{\text{MAR}}$ indeed becomes (asymptotically) equivalent to the SNP estimators obtained earlier in Section 3.2. Further, with various choices of $\mu(\cdot)$, the SNP imputation strategy again equips us with a *family* of SS estimators of $\boldsymbol{\theta}_0$ under the MAR setting, with $\mu(\cdot) = m(\cdot)$ leading to the optimal estimator.

The above estimator $\widehat{\boldsymbol{\theta}}_{\text{MAR}}$ is derived with a known $\pi_{\bar{N}}(\cdot)$, for simplicity. In practice, $\pi_{\bar{N}}(\cdot)$ is typically unknown and a consistent estimator $\widehat{\pi}_{\bar{N}}(\cdot)$ may be constructed. Then, one may modify $\widehat{\boldsymbol{\theta}}_{\text{MAR}}$ by replacing $\pi_{\bar{N}}(\cdot)$ with $\widehat{\pi}_{\bar{N}}(\cdot)$ in all the steps. The resulting estimator will have an expansion similar to (7) but with extra error terms accounting for the variability in $\widehat{\pi}_{\bar{N}}(\cdot)$, which need to be properly controlled. The theoretical analysis will be more involved since establishing the convergence rates and asymptotic expansion for $\widehat{\pi}_{\bar{N}}(\cdot) - \pi_{\bar{N}}(\cdot)$ is also non-standard due to $b_{\bar{N}} \to 0$. Lastly, the score equation (5) used to construct $\widehat{\boldsymbol{\theta}}_{\text{MAR}}$ has the additional benefit of *'double robustness'*, in the sense that *even if* $\widehat{\pi}_{\bar{N}}(\cdot)$ is inconsistent for $\pi_{\bar{N}}(\cdot)$, as long as $\widehat{\mu}(\cdot)$ estimates the true $m(\cdot)$, $\widehat{\boldsymbol{\theta}}_{\text{MAR},\,\mu(\cdot)}$ is consistent for $\boldsymbol{\theta}_0$. On the other hand, as long as $\widehat{\pi}_{\bar{N}}(\cdot)$ targets the true $\pi_{\bar{N}}(\cdot)$, then for *any* choice of $\mu(\cdot)$,

$\widehat{\boldsymbol{\theta}}_{\text{MAR},\,\mu(\cdot)}$ is consistent for $\boldsymbol{\theta}_0$. For the MCAR case, $\widehat{\pi}_{\bar{N}}(\cdot) \equiv \widehat{\pi}_{\bar{N}} = n/\bar{N}$ is always consistent for $\pi_{\bar{N}}(\cdot)$ and in fact this is exactly what allowed us to achieve a *family* of SNP estimators, all consistent for $\boldsymbol{\theta}_0$, for various choices of the SNP imputation function $\mu(\cdot)$.

III. PROOF OF LEMMA A.1

Firstly, since d is fixed, it suffices to prove the result for any arbitrary scalar coordinate $\widehat{\mathbf{G}}^{(j)}_{n,m} \equiv \widehat{\mathcal{G}}_{n,m}$ (say) and $\overline{\mathbf{G}}^{(j)}_{n} \equiv \overline{\mathcal{G}}_{n}$ (say) of $\widehat{\mathbf{G}}_{n,m}$ and $\overline{\mathbf{G}}_n$ respectively, for any $j \in \{1, \ldots, d\}$. For any data $\mathbb{S}$ and $\mathbb{S}^*$, we let $\mathbb{P}_{\mathbb{S}}$ and $\mathbb{P}_{\mathbb{S},\mathbb{S}^*}$ denote the joint probability distributions of the observations in $\mathbb{S}$ and $(\mathbb{S}, \mathbb{S}^*)$ respectively, $\mathbb{E}_{\mathbb{S}}(\cdot)$ denote the expectation w.r.t $\mathbb{P}_{\mathbb{S}}$, and $\mathbb{P}_{\mathbb{S}|\mathbb{S}^*}$ denote the conditional probability distribution of the observations in $\mathbb{S}$ given $\mathbb{S}^*$.

To show that $\widehat{\mathcal{G}}_{n,m} - \overline{\mathcal{G}}_n = O_p(m^{-\frac{1}{2}})$, we first note that since $\mathbb{S}_n \perp\!\!\!\perp \mathbb{S}_m$,

$$\mathbb{P}_{\mathbb{S}_n,\mathbb{S}_m}\left(|\widehat{\mathcal{G}}_{n,m} - \overline{\mathcal{G}}_n| > m^{-\frac{1}{2}}t\right) \;=\; \mathbb{E}_{\mathbb{S}_n}\left\{\mathbb{P}_{\mathbb{S}_m}\left(|\widehat{\mathcal{G}}_{n,m} - \overline{\mathcal{G}}_n| > m^{-\frac{1}{2}}t \,\middle|\, \mathbb{S}_n\right)\right\},$$

for any $t > 0$. Now, conditional on $\mathbb{S}_n$, $\widehat{\mathbf{G}}_{n,m} - \overline{\mathbf{G}}_n$ is a centered average of $\{\widehat{\mathbf{g}}_n(\mathbf{Z}_j)\}_{j=1}^m$ which are i.i.d. and bounded by $\widehat{T}_n < \infty$ a.s. $[\mathbb{P}_{\mathbb{S}_n}]$ $\forall\, n$. Hence, applying Hoeffding's inequality, we have for any n and m,

$$(8) \qquad \mathbb{P}_{\mathbb{S}_m}\left(|\widehat{\mathcal{G}}_{n,m} - \overline{\mathcal{G}}_n| > m^{-\frac{1}{2}}t \,\middle|\, \mathbb{S}_n\right) \;\leq\; 2\,\exp\left(-\,\frac{2m^2t^2}{4m^2\widehat{T}_n^2}\right) \quad \text{a.s. } [\mathbb{P}_{\mathbb{S}_n}].$$

Now, since $\widehat{T}_n \geq 0$ is $O_p(1)$, we have: for any given $\epsilon > 0$, $\exists\; \delta(\epsilon) > 0$ such that: $\mathbb{P}_{\mathbb{S}_n}\{\widehat{T}_n > \delta(\epsilon)\} \leq \epsilon/4 \;\forall\, n$. Let $\mathbb{A}(\epsilon)$ denote the event: $\{\widehat{T}_n > \delta(\epsilon)\}$ and let $\mathbb{A}^c(\epsilon)$ denote its complement. Then, using (8), we have: $\forall\, n$ and m,

$$\begin{aligned}
&\mathbb{P}_{\mathbb{S}_n,\mathbb{S}_m}\left(|\widehat{\mathcal{G}}_{n,m} - \overline{\mathcal{G}}_n| > m^{-\frac{1}{2}}t\right) \;\leq\; \mathbb{E}_{\mathbb{S}_n}\left\{2\,\exp\left(-\,\frac{2m^2t^2}{4m^2\widehat{T}_n^2}\right)\right\}\\
&= \mathbb{E}_{\mathbb{S}_n}\left\{2\,\exp\left(-\,\frac{t^2}{2\widehat{T}_n^2}\right)\right\} = \mathbb{E}_{\mathbb{S}_n}\left[2\,\exp\left(-\,\frac{t^2}{2\widehat{T}_n^2}\right)\left\{1_{\mathbb{A}^c(\epsilon)} + 1_{\mathbb{A}(\epsilon)}\right\}\right]\\
&\leq \left[2\,\exp\left\{-\,\frac{t^2}{2\delta^2(\epsilon)}\right\}\mathbb{P}_{\mathbb{S}_n}\left\{\mathbb{A}^c(\epsilon)\right\} + 2\,\mathbb{P}_{\mathbb{S}_n}\left\{\mathbb{A}(\epsilon)\right\}\right]\\
&\leq 2\,\exp\left\{-\,\frac{t^2}{2\delta^2(\epsilon)}\right\} + \frac{\epsilon}{2} \quad \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \text{ (for some suitable choice of } t),
\end{aligned}$$

where the last step follows by choosing $t \equiv t_\epsilon$ to be any large enough t such that $\exp\{-t^2/2\delta^2(\epsilon)\} \leq \epsilon/4$. Such a choice of t_ϵ clearly exists. This establishes the first claim (a) in Lemma A.1. The second claim (b) in Lemma A.1 is a trivial consequence of the Central Limit Theorem (CLT). ∎

IV. PROOF OF THEOREM 3.1

To show Theorem 3.1, we first note that under Assumption 3.1 (i)-(v), and letting $a_n = (\log n)^{\frac{1}{2}}(nh^p)^{-\frac{1}{2}} + h^q$, the following holds:

$$\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{m}(\mathbf{x}) - m(\mathbf{x})| = O_p(a_n) = \sup_{\mathbf{x}\in\mathcal{X}}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})|. \tag{9}$$

(9) is a fairly standard result and we only provide a sketch of its proof as follows. Under Assumption 3.1 (ii)-(iii), using Theorem 2 of Hansen (2008), $\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{l}(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\widehat{l}(\mathbf{x})\}| = O_p(a_n^*) = \sup_{\mathbf{x}\in\mathcal{X}}|\widehat{f}(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\widehat{f}(\mathbf{x})\}|$, where $a_n^* = (\log n)^{\frac{1}{2}}(nh^p)^{-\frac{1}{2}}$. Next, using standard arguments based on Taylor series expansions of $l(\cdot)$ and $m(\cdot)$ under their assumed smoothness, and noting that $K(\cdot)$ is a q^{th} order kernel having finite q^{th} moments, we obtain:

$$\sup_{\mathbf{x}\in\mathcal{X}}|\mathbb{E}_{\mathcal{L}}\{\widehat{l}(\mathbf{x})\} - l(\mathbf{x})| = O(h^q) = \sup_{\mathbf{x}\in\mathcal{X}}|\mathbb{E}_{\mathcal{L}}\{\widehat{f}(\mathbf{x})\} - f(\mathbf{x})|.$$

Combining these two results, and the definitions of $m(.)$ and $\widehat{m}(.)$ along with Assumption 3.1 (iv), we have (9). Next, note that using (3.2), we have:

$$\begin{aligned}\mathbf{\Gamma}_N(\widehat{\boldsymbol{\theta}}_{np} - \boldsymbol{\theta}_0) &= \mathbb{E}_{\mathcal{U}}[N^{-1}\sum_{j=n+1}^{n+N}\overrightarrow{\mathbf{X}}_j\{\widehat{m}(\mathbf{X}_j) - \overrightarrow{\mathbf{X}}_j'\boldsymbol{\theta}_0\}] + O_p(N^{-\frac{1}{2}}) \\ &= \mathbb{E}_{\mathbf{X}}[\overrightarrow{\mathbf{X}}\{\widehat{m}(\mathbf{X}) - m(\mathbf{X})\}] + O_p(N^{-\frac{1}{2}}),\end{aligned}$$

where the first step is due to Lemma A.1 (a) with $\sup_{\mathbf{x}\in\mathcal{X}}\|\overrightarrow{\mathbf{x}}\{\widehat{m}(\mathbf{x}) - \overrightarrow{\mathbf{x}}'\boldsymbol{\theta}_0\}\| \leq \sup_{\mathbf{x}\in\mathcal{X}}[\|\overrightarrow{\mathbf{x}}\|\{|\widehat{m}(\mathbf{x}) - m(\mathbf{x})| + |m(\mathbf{x}) - \overrightarrow{\mathbf{x}}'\boldsymbol{\theta}_0|\}] = O_p(1)$ due to (9) and the boundedness of $\mathbf{X}$ and $m(\cdot)$, while the last step uses: $\mathbb{E}_{\mathbf{X}}[\overrightarrow{\mathbf{X}}\{m(\mathbf{X}) - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta}_0\}] = \mathbf{0}$ which follows from the definitions of $\boldsymbol{\theta}_0$ and $m(\cdot)$. It then follows further, using $\mathbf{\Gamma}_N^{-1} = \mathbf{\Gamma}^{-1} + O_p(N^{-\frac{1}{2}})$, that

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{np} - \boldsymbol{\theta}_0) \;=\; n^{\frac{1}{2}}\,\mathbf{\Gamma}^{-1}\mathbb{E}_{\mathbf{X}}[\overrightarrow{\mathbf{X}}\{\widehat{m}(\mathbf{X}) - m(\mathbf{X})\}] + O_p\left(\frac{n}{N}\right)^{\frac{1}{2}}.$$

Letting $\phi_n(\mathbf{X}) = (nh^p)^{-1}\sum_{i=1}^n K\{(\mathbf{X}-\mathbf{X}_i)/h\}\{Y_i - m(\mathbf{X})\}$, and expanding the first term in the above equation, we now obtain:

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}}_{np} - \boldsymbol{\theta}_0\right) \;=\; \mathbf{\Gamma}^{-1}\left(\mathbf{T}_{n,1}^{(1)} + \mathbf{T}_{n,1}^{(2)}\right) + O_p\left(\frac{n}{N}\right)^{\frac{1}{2}}, \tag{10}$$

where $\mathbf{T}_{n,1}^{(1)} = n^{\frac{1}{2}}\,\mathbb{E}_{\mathbf{X}}\{\overrightarrow{\mathbf{X}}\phi_n(\mathbf{X})/f(\mathbf{X})\}$ and

$$\begin{aligned}\mathbf{T}_{n,1}^{(2)} &= n^{\frac{1}{2}}\,\mathbb{E}_{\mathbf{X}}\left[\overrightarrow{\mathbf{X}}\phi_n(\mathbf{X})\{\widehat{f}(\mathbf{X})^{-1} - f(\mathbf{X})^{-1}\}\right] \\ &= n^{\frac{1}{2}}\,\mathbb{E}_{\mathbf{X}}[\overrightarrow{\mathbf{X}}\,\{\widehat{m}(\mathbf{X}) - m(\mathbf{X})\}\,\{f(\mathbf{X}) - \widehat{f}(\mathbf{X})\}/f(\mathbf{X})] \\ &\leq n^{\frac{1}{2}}\,\sup_{\mathbf{x}\in\mathcal{X}}\left\{\|\overrightarrow{\mathbf{x}}\|\,|\widehat{m}(\mathbf{x}) - m(\mathbf{x})|\left|\widehat{f}(\mathbf{x})/f(\mathbf{x}) - 1\right|\right\} \;= O_p\left(n^{\frac{1}{2}}a_n^2\right),\end{aligned} \tag{11}$$

where the last step in (11) follows from (9), Assumption 3.1 (iv) and the boundedness of $\mathbf{X}$. For $\mathbf{T}_{n,1}^{(1)}$, we have:

$$\begin{aligned}\mathbf{T}_{n,1}^{(1)} &= n^{\frac{1}{2}} \int_{\mathcal{X}} \overrightarrow{\mathbf{x}} \phi_n(\mathbf{x}) d\mathbf{x} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_{\mathcal{X}} \overrightarrow{\mathbf{x}} h^{-p} K_h(\mathbf{x}-\mathbf{X}_i) \left\{Y_i - m(\mathbf{x})\right\} d\mathbf{x} \\ &= n^{\frac{1}{2}} \sum_{i=1}^{n} n^{-1} \int_{\mathcal{A}_{i,n}} \overrightarrow{(\mathbf{X}_i + h\boldsymbol{\psi}_i)}\, K\left(\boldsymbol{\psi}_i\right) \left\{Y_i - m(\mathbf{X}_i + h\boldsymbol{\psi}_i)\right\} d\boldsymbol{\psi}_i, \qquad (12)\end{aligned}$$

where $\boldsymbol{\psi}_i = (\mathbf{x} - \mathbf{X}_i)/h$ and $\mathcal{A}_{i,n} = \{\boldsymbol{\psi}_i \in \mathbb{R}^p : (\mathbf{X}_i + h\boldsymbol{\psi}_i) \in \mathcal{X}\}$. Now, since $K(\cdot)$ is zero outside the bounded set $\mathcal{K}$, the i^{th} integral in (12) only runs over $(\mathcal{A}_{i,n} \cap \mathcal{K})$. Further, since $h = o(1)$, using Assumption 3.1 (vi), $\mathcal{A}_{i,n} \supseteq \mathcal{K}$ a.s. $[\mathbb{P}_{\mathcal{L}}]$ or, $(\mathcal{A}_{i,n} \cap \mathcal{K}) = \mathcal{K}$ a.s. $[\mathbb{P}_{\mathcal{L}}]$ $\forall\ 1 \leq i \leq n$ with n large enough. Thus, for large enough n, (12) can be written as:

$$\begin{aligned}\mathbf{T}_{n,1}^{(1)} &= n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_{\mathcal{K}} \overrightarrow{(\mathbf{X}_i + h\boldsymbol{\psi}_i)}\, K\left(\boldsymbol{\psi}_i\right) \left\{Y_i - m(\mathbf{X}_i + h\boldsymbol{\psi}_i)\right\} d\boldsymbol{\psi}_i \ \text{ a.s. } [\mathbb{P}_{\mathcal{L}}] \\ &= n^{\frac{1}{2}} \sum_{i=1}^{n} n^{-1} \left[\overrightarrow{\mathbf{X}}_i \left\{Y_i - m(\mathbf{X}_i)\right\} + O_p(h^q)\right] \qquad (13) \\ &= n^{-\frac{1}{2}} \sum_{i=1}^{n} \overrightarrow{\mathbf{X}}_i \left\{Y_i - m(\mathbf{X}_i)\right\} + O_p\left(n^{\frac{1}{2}} h^q\right), \qquad (14)\end{aligned}$$

where (13), and hence (14), follows from standard arguments based on Taylor series expansions of $m(\mathbf{X}_i + h\boldsymbol{\psi}_i)$ around $m(\mathbf{X}_i)$ under the assumed smoothness of $m(\cdot)$, and using the fact that $K(\cdot)$ is a q^{th} order kernel. Combining (10), (11) and (14), and noting that under our assumptions, $(n^{\frac{1}{2}} a_n^2 + n^{\frac{1}{2}} h^q)$ $= O\{n^{\frac{1}{2}} h^q + (\log n)(n^{\frac{1}{2}} h^p)^{-1}\}$, the result of Theorem 3.1 now follows. ■

V. PROOF OF THEOREM 4.1

Let $a_{n,2} = (\log n)^{\frac{1}{2}} (nh^r)^{-\frac{1}{2}} + h^q$. Then, we first note that

$$\sup_{\mathbf{w} \in \mathcal{X}_{\mathbf{P}_r}} |\widetilde{\varphi}_{\mathbf{P}_r}^{(\varrho)}(\mathbf{w}) - \varphi_{\mathbf{P}_r}^{(\varrho)}(\mathbf{w})| = O_p(a_{n,2}), \quad \forall\ \varrho \in \{0, 1\}. \tag{15}$$

To see this, note that under Assumption 4.1 (ii)-(iii), Theorem 2 of Hansen (2008) applies, and we have for $d_n = (\log n)^{\frac{1}{2}} (nh^r)^{-\frac{1}{2}}$,

$$\sup_{\mathbf{w} \in \mathcal{X}_{\mathbf{P}_r}} |\widetilde{\varphi}_{\mathbf{P}_r}^{(\varrho)}(\mathbf{w}) - \mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_{\mathbf{P}_r}^{(\varrho)}(\mathbf{w})\}| = O_p(d_n) \quad \forall\ \varrho \in \{0, 1\}.$$

Next, using standard arguments based on a q^{th} order Taylor series expansion of $\varphi_{\mathbf{P}_r}^{(\varrho)}(\cdot)$ and noting that $K(\cdot)$ is a q^{th} order kernel, we obtain:

$$\sup_{\mathbf{w} \in \mathcal{X}_{\mathbf{P}_r}} |\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_{\mathbf{P}_r}^{(\varrho)}(\mathbf{w})\} - \varphi_{\mathbf{P}_r}^{(\varrho)}(\mathbf{w})| = O(h^q) \quad \forall\ \varrho \in \{0, 1\}.$$

Combining these two results gives (15). Further,

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}} \ |\widetilde{m}(\mathbf{x};\mathbf{P}_r) - m(\mathbf{x};\mathbf{P}_r)| \ = \ \sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}} \ |\widetilde{m}_{\mathbf{P}_r}(\mathbf{w}) - m_{\mathbf{P}_r}(\mathbf{w})| \\
&\leq \ \sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}} \ \left| \frac{\widetilde{l}_{\mathbf{P}_r}(\mathbf{w}) - l_{\mathbf{P}_r}(\mathbf{w})}{\widetilde{f}_{\mathbf{P}_r}(\mathbf{w})} \right| + \sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}} \ \left\{ \left| \frac{|l_{\mathbf{P}_r}(\mathbf{w})|}{f_{\mathbf{P}_r}(\mathbf{w})} - \frac{|l_{\mathbf{P}_r}(\mathbf{w})|}{\widetilde{f}_{\mathbf{P}_r}(\mathbf{w})} \right| \right\} \\
&(16) = \ O_p(a_{n,2}),
\end{aligned}$$

where the last step follows from repeated use of (15) and Assumption 4.1 (iii)-(iv). Next, we aim to bound $\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{\varphi}^{(\varrho)}(\mathbf{x};\widehat{\mathbf{P}}_r) - \widetilde{\varphi}^{(\varrho)}(\mathbf{x};\mathbf{P}_r)|$ to account for the potential estimation error of $\widehat{\mathbf{P}}_r$. Using a first order Taylor series expansion of $K(.)$ under Assumption 4.1 (vi), we have: $\forall \ \varrho \in \{0, 1\}$,

$$\begin{aligned}
\widehat{\varphi}^{(\varrho)}(\mathbf{x};\widehat{\mathbf{P}}_r) - \widetilde{\varphi}^{(\varrho)}(\mathbf{x};\mathbf{P}_r) \ &= \ \frac{1}{nh^r}\sum_{i=1}^{n} \boldsymbol{\nabla} K'(\mathbf{w}_{i,\mathbf{x}})(\widehat{\mathbf{P}}_r - \mathbf{P}_r)' \left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right) Y_i^{\varrho} \\
(17) \qquad &= \ \mathrm{trace}\left\{(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}}\right\} + \mathrm{trace}\left\{(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'\widehat{\mathbf{M}}^{(2)}_{n,\varrho,\mathbf{x}}\right\},
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}} &= \frac{1}{nh^{r+1}}\sum_{i=1}^{n}(\mathbf{x}-\mathbf{X}_i)\left\{\boldsymbol{\nabla} K'\left(\frac{\mathbf{P}_r'\mathbf{x} - \mathbf{P}_r'\mathbf{X}_i}{h}\right)\right\} Y_i^{\varrho} \ \ \text{and} \\
\widehat{\mathbf{M}}^{(2)}_{n,\varrho,\mathbf{x}} &= \frac{1}{nh^{r+1}}\sum_{i=1}^{n}(\mathbf{x}-\mathbf{X}_i)\left\{\boldsymbol{\nabla} K'(\mathbf{w}_{i,\mathbf{x}}) - \boldsymbol{\nabla} K'\left(\frac{\mathbf{P}_r'\mathbf{x} - \mathbf{P}_r'\mathbf{X}_i}{h}\right)\right\} Y_i^{\varrho},
\end{aligned}$$

with $\mathbf{w}_{i,\mathbf{x}} \in \mathbb{R}^r$ being 'intermediate' points satisfying: $\|\mathbf{w}_{i,\mathbf{x}} - \mathbf{P}_r'(\mathbf{x}-\mathbf{X}_i)h^{-1}\| \leq \|\widehat{\mathbf{P}}_r'(\mathbf{x}-\mathbf{X}_i)h^{-1} - \mathbf{P}_r'(\mathbf{x}-\mathbf{X}_i)h^{-1}\| \leq O_p(\alpha_n h^{-1})$. The last bound, based on $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$ and the compactness of $\mathcal{X}$, is uniform in $(i, \mathbf{x})$. For any matrix $\mathbf{A} = [a_{\mathrm{ij}}]$, let $\|\mathbf{A}\|_{\max}$ denote the max-norm of $\mathbf{A}$, and $|\mathbf{A}|$ denote the matrix $[|a_{\mathrm{ij}}|]$. Now, Assumption 4.1 (viii) implies: $\|\boldsymbol{\nabla} K(\mathbf{w}_1) - \boldsymbol{\nabla} K(\mathbf{w}_2)\| \leq B\|\mathbf{w}_1 - \mathbf{w}_2\| \ \forall \ \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^r$, for some constant $B < \infty$. Then using the above arguments, we note that $\forall \ \varrho \in \{0, 1\}$, $\| \sup_{\mathbf{x}\in\mathcal{X}} |\widehat{\mathbf{M}}^{(2)}_{n,\varrho,\mathbf{x}}| \|_{\max}$ is bounded by:

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}} \left\{ \frac{B}{nh^{r+1}}\sum_{i=1}^{n}\|\mathbf{x}-\mathbf{X}_i\| \left\| \mathbf{w}_{i,\mathbf{x}} - \frac{\mathbf{P}_r'\mathbf{x} - \mathbf{P}_r'\mathbf{X}_i}{h} \right\| |Y_i^{\varrho}| \right\} \\
&\leq \ \sup_{\mathbf{x}\in\mathcal{X}} \left\{ \frac{B}{nh^{r+1}}\sum_{i=1}^{n}\|\mathbf{x}-\mathbf{X}_i\| \left\| \frac{(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'(\mathbf{x}-\mathbf{X}_i)}{h} \right\| |Y_i^{\varrho}| \right\} \\
&\leq \sup_{\mathbf{x}\in\mathcal{X},\mathbf{X}\in\mathcal{X}} \left\{ \|\mathbf{x}-\mathbf{X}\| \|(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'(\mathbf{x}-\mathbf{X})\| \right\} \frac{B}{nh^{r+2}}\sum_{i=1}^{n}|Y_i^{\varrho}| \ \leq O_p\left(\frac{\alpha_n}{h^{r+2}}\right).
\end{aligned}$$

The first two steps above use the triangle inequality, the Lipschitz continuity of $\boldsymbol{\nabla} K(\cdot)$ and the definition of $\mathbf{w}_{i,\mathbf{x}}$, while the next two use the compactness of $\mathcal{X}$, the uniform bound obtained in the last paragraph, the Law of Large Numbers (LLN), and that $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$. Thus, we have:

$$\sup_{\mathbf{x}\in\mathcal{X}} \left|\text{trace}\left\{(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'\widehat{\mathbf{M}}^{(2)}_{n,\varrho,\mathbf{x}}\right\}\right| = O_p\left(\frac{\alpha_n^2}{h^{r+2}}\right) \quad \forall\ \varrho \in \{0,1\}. \tag{18}$$

Now for bounding $\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}}$, let us first write it as: $\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}} = \widehat{\mathbf{M}}^{(1,1)}_{n,\varrho,\mathbf{x}} - \widehat{\mathbf{M}}^{(1,2)}_{n,\varrho,\mathbf{x}}$, where $\widehat{\mathbf{M}}^{(1,1)}_{n,\varrho,\mathbf{x}} = (nh^{r+1})^{-1}\sum_{i=1}^n \mathbf{x}\boldsymbol{\nabla} K'\{\mathbf{P}_r'(\mathbf{x}-\mathbf{X}_i)/h\}Y_i^\varrho$ and $\widehat{\mathbf{M}}^{(1,2)}_{n,\varrho,\mathbf{x}} = (nh^{r+1})^{-1}\sum_{i=1}^n \mathbf{X}_i\boldsymbol{\nabla} K'\{\mathbf{P}_r'(\mathbf{x}-\mathbf{X}_i)/h\}Y_i^\varrho\ \forall\ \varrho \in \{0,1\}$. Then, under Assumption 4.1 (iii), (vi) and (vii), using Theorem 2 of Hansen (2008) along with the compactness of $\mathcal{X}$, we have: for each $s \in \{1,2\}$ and $\varrho \in \{0,1\}$,

$$\left\|\sup_{\mathbf{x}\in\mathcal{X}}\left|\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,\mathbf{x}} - \mathbb{E}_{\mathcal{L}}\left(\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,\mathbf{x}}\right)\right|\right\|_{\max} \le O_p\left(\frac{\log n}{nh^{r+2}}\right)^{\frac{1}{2}}. \tag{19}$$

Now, $\forall\ \varrho \in \{0,1\}$, let $\nu^{(\varrho)}(\mathbf{w}) = \mathbb{E}\{Y^\varrho \mid \mathbf{X}_{\mathbf{P}_r} = \mathbf{w}\}f_{\mathbf{P}_r}(\mathbf{w})$ and $\boldsymbol{\xi}^{(\varrho)}(\mathbf{w}) = \mathbb{E}\{\mathbf{X}Y^\varrho \mid \mathbf{X}_{\mathbf{P}_r} = \mathbf{w}\}f_{\mathbf{P}_r}(\mathbf{w})$. Further, let $\{\boldsymbol{\nabla}\nu^{(\varrho)}(\mathbf{w})\}_{r\times 1}$ and $\{\boldsymbol{\nabla}\boldsymbol{\xi}^{(\varrho)}(\mathbf{w})\}_{p\times r}$ denote their respective first order derivatives. Then, $\forall\ \varrho \in \{0,1\}$, we have:

$$\begin{aligned}
&\left\|\sup_{\mathbf{x}\in\mathcal{X}}\left|\mathbb{E}_{\mathcal{L}}\left(\widehat{\mathbf{M}}^{(1,1)}_{n,\varrho,\mathbf{x}}\right)\right|\right\|_{\max} \\
&= \left\|\sup_{\mathbf{x}\in\mathcal{X}}\left|\frac{\mathbf{x}}{h^{r+1}}\int \nu^{(\varrho)}(\mathbf{w})\ \boldsymbol{\nabla} K'\left(\frac{\mathbf{P}_r'\mathbf{x}-\mathbf{w}}{h}\right)d\mathbf{w}\right|\right\|_{\max} \\
&= \left\|\sup_{\mathbf{x}\in\mathcal{X}}\left|\mathbf{x}\int \boldsymbol{\nabla}\nu^{(\varrho)'}\left(\mathbf{P}_r'\mathbf{x}+h\boldsymbol{\psi}\right)\ K(\boldsymbol{\psi})d\boldsymbol{\psi}\right|\right\|_{\max} = O(1),
\end{aligned} \tag{20}$$

$$\begin{aligned}
&\left\|\sup_{\mathbf{x}\in\mathcal{X}}\left|\mathbb{E}_{\mathcal{L}}\left(\widehat{\mathbf{M}}^{(1,2)}_{n,\varrho,\mathbf{x}}\right)\right|\right\|_{\max} \\
&= \left\|\sup_{\mathbf{x}\in\mathcal{X}}\left|h^{-(r+1)}\int \boldsymbol{\xi}^{(\varrho)}(\mathbf{w})\ \boldsymbol{\nabla} K'\left(\frac{\mathbf{P}_r'\mathbf{x}-\mathbf{w}}{h}\right)d\mathbf{w}\right|\right\|_{\max} \\
&= \left\|\sup_{\mathbf{x}\in\mathcal{X}}\left|\mathbf{D}(\mathbf{x})\int \boldsymbol{\nabla}\boldsymbol{\xi}^{(\varrho)}\left(\mathbf{P}_r'\mathbf{x}+h\boldsymbol{\psi}\right)\ K(\boldsymbol{\psi})d\boldsymbol{\psi}\right|\right\|_{\max} = O(1),
\end{aligned} \tag{21}$$

where, $\forall\ \mathbf{x} \in \mathcal{X}$, $\mathbf{D}(\mathbf{x})$ denotes the $p \times p$ diagonal matrix: $\text{diag}(\mathbf{x}_{[1]}, \ldots, \mathbf{x}_{[p]})$. In both (20) and (21), the first step follows from definition, the second from standard arguments based on integration by parts (applied coordinate-wise) and change of variable, while the last one is due to compactness of $\mathcal{X}$ and a medley of the conditions in Assumption 4.1 namely, boundedness and integrability of $K(\cdot)$ and $\boldsymbol{\nabla} K(\cdot)$, (iii) and (v) for (20) so that $\boldsymbol{\nabla}\nu^{(\varrho)}(\cdot)$ is bounded on $\mathcal{X}_{\mathbf{P}_r}$, and (ix) for (21). It now follows that for each $\varrho \in \{0,1\}$,

$$\left\| \sup_{\mathbf{x}\in\mathcal{X}} \left| \mathbb{E}_{\mathcal{L}}\left(\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}}\right)\right| \right\|_{\max} = O(1). \tag{22}$$

Letting $d_n^* = (\log n)^{\frac{1}{2}}(nh^{r+2})^{-\frac{1}{2}}$, we now have from (19) and (22):

$$\sup_{\mathbf{x}\in\mathcal{X}} \left| \operatorname{trace}\left\{ (\widehat{\mathbf{P}}_r' - \mathbf{P}_r')\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}} \right\}\right| = O_p\left(\alpha_n d_n^* + \alpha_n\right) \quad \forall\ \varrho \in \{0,1\}. \tag{23}$$

Applying (23) and (18) to (17) using the triangle inequality, we have $\forall\ \varrho$,

$$\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{\varphi}^{(\varrho)}(\mathbf{x};\widehat{\mathbf{P}}_r) - \widetilde{\varphi}^{(\varrho)}(\mathbf{x};\mathbf{P}_r)| = O_p\left\{ \frac{\alpha_n^2}{h^{r+2}} + \alpha_n \frac{(\log n)^{\frac{1}{2}}}{(nh^{r+2})^{\frac{1}{2}}} + \alpha_n \right\}. \tag{24}$$

Finally, note that $\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r) = \widehat{l}(\mathbf{x};\widehat{\mathbf{P}}_r)/\widehat{f}(\mathbf{x};\widehat{\mathbf{P}}_r) = \widehat{\varphi}^{(1)}(\mathbf{x};\widehat{\mathbf{P}}_r)/\widehat{\varphi}^{(0)}(\mathbf{x};\widehat{\mathbf{P}}_r)$. Repeated use of (24), along with (16) and Assumption 4.1 (iii)-(iv), leads to:

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}} \left| \widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r) - m(\mathbf{x};\mathbf{P}_r)\right| \\
&\leq \sup_{\mathbf{x}\in\mathcal{X}} \left| \widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r) - \widetilde{m}(\mathbf{x};\mathbf{P}_r)\right| + \sup_{\mathbf{x}\in\mathcal{X}} \ |\widetilde{m}(\mathbf{x};\mathbf{P}_r) - m(\mathbf{x};\mathbf{P}_r)| \\
&\leq \sup_{\mathbf{x}\in\mathcal{X}} \left\{ \left| \frac{\widehat{l}(\mathbf{x};\widehat{\mathbf{P}}_r) - \widetilde{l}(\mathbf{x};\mathbf{P}_r)}{\widehat{f}(\mathbf{x};\widehat{\mathbf{P}}_r)} \right| + \left| \frac{\widetilde{l}(\mathbf{x};\mathbf{P}_r)}{\widetilde{f}(\mathbf{x};\mathbf{P}_r)} - \frac{\widetilde{l}(\mathbf{x};\mathbf{P}_r)}{\widehat{f}(\mathbf{x};\widehat{\mathbf{P}}_r)} \right| \right\} + O_p(a_{n,2}) \\
&\leq O_p\left\{ \frac{\alpha_n^2}{h^{r+2}} + \alpha_n \frac{(\log n)^{\frac{1}{2}}}{(nh^{r+2})^{\frac{1}{2}}} + \alpha_n \right\} + O_p(a_{n,2}) \quad = O_p(a_{n,1} + a_{n,2}).
\end{aligned} \tag{25}$$

The proof of Theorem 4.1 is now complete. ■

VI. PROOFS OF LEMMAS A.2-A.3 AND THEOREM 4.2

VI.1. Proof of Lemma A.2. First note that for each $\varrho \in \{0,1\}$,

$$\int \widetilde{\varphi}^{(\varrho)}(\mathbf{x};\mathbf{P}_r)\mathbb{P}_n(d\mathbf{x}) = n^{-2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} \mathbf{H}^{(n,\varrho)}_{i_1,i_2}$$

is a V-statistic, where $\mathbf{H}^{(n,\varrho)}_{i_1,i_2} = h^{-r}\boldsymbol{\lambda}(\mathbf{X}_{i_1})Y_{i_2}^{\varrho}K\{\mathbf{P}_r'(\mathbf{X}_{i_1} - \mathbf{X}_{i_2})/h\}$. Using the V-statistic projection result given in Lemma 8.4 of Newey and McFadden (1994), it then follows that for each $\varrho \in \{0,1\}$,

$$\begin{aligned}
&\mathbb{G}_n^*\left\{ \boldsymbol{\lambda}(\cdot)[\widetilde{\varphi}^{(\varrho)}(\cdot\,;\mathbf{P}_r) - \mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}^{(\varrho)}(\cdot\,;\mathbf{P}_r)\}]\right\} \\
&= n^{-\frac{1}{2}}O_p\left[\mathbb{E}(\|\mathbf{H}^{(n,\varrho)}_{i_1,i_1}\|) + \{\mathbb{E}(\|\mathbf{H}^{(n,\varrho)}_{i_1,i_2}\|^2)\}^{\frac{1}{2}} \right] = O_p\left(n^{-\frac{1}{2}}h^{-r}\right),
\end{aligned} \tag{26}$$

The last step follows from $K(\cdot)$ and $\boldsymbol{\lambda}(\cdot)$ being bounded and Y^{ϱ} having finite 2^{nd} moments. Now, observe that $n^{\frac{1}{2}}\mathbb{G}_n^*\left\{\boldsymbol{\lambda}(\cdot)[\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_\star^{(\varrho)}(\cdot\,;\mathbf{P}_r)\}]\right\}$ is a centered sum of i.i.d. random vectors bounded by:

$$D_{n,\varrho} = \sup_{\mathbf{x}\in\mathcal{X}}\left\{\|\boldsymbol{\lambda}(\mathbf{x})\|\,|\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_\star^{(\varrho)}(\mathbf{x};\mathbf{P}_r)\}|\right\} = O(h^q) \quad \forall\ \varrho\in\{0,1\},$$

where throughout, for any estimator $\tilde{\xi}(\cdot)$ with population limit $\xi(\cdot)$, we use the notation $\tilde{\xi}_\star(\cdot)$ to denote its centered version given by: $\tilde{\xi}_\star(\cdot)=\tilde{\xi}(\cdot)-\xi(\cdot)$. Here, $D_{n,\varrho}=O(h^q)$ since $\boldsymbol{\lambda}(\cdot)$ is bounded and $\sup_{\mathbf{x}\in\mathcal{X}}|\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_\star(\mathbf{x};\mathbf{P}_r)\}| = \sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}}|\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}^{(\varrho)}_{\mathbf{P}_r}(\mathbf{w})\} - \varphi^{(\varrho)}_{\mathbf{P}_r}(\mathbf{w})| = O(h^q)$, as argued while proving (15). Hence, $\exists$ a constant $\kappa_\varrho > 0$ such that $h^q/D_{n,\varrho} \geq \kappa_\varrho\ \forall\ n$. Then, using Hoeffding's Inequality, we have: $\forall\ n$, given any $\epsilon > 0$ and any $M = M(\epsilon)$ large enough,

$$\sum_{l=1}^{d}\mathbb{P}\left[\left|\mathbb{G}_n^*\left\{\boldsymbol{\lambda}_{[l]}(\cdot)[\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_\star^{(\varrho)}(\cdot\,;\mathbf{P}_r)\}]\right\}\right| > \frac{Mh^q}{d^{\frac{1}{2}}}\right] \leq 2d\exp\left(-\frac{M^2h^{2q}}{2dD_{n,\varrho}^2}\right) \Rightarrow$$

$$\mathbb{P}\left[\left\|\mathbb{G}_n^*\left\{\boldsymbol{\lambda}(\cdot)[\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_\star^{(\varrho)}(\cdot\,;\mathbf{P}_r)\}]\right\}\right\| > Mh^q\right] \leq 2d\exp\left(-\frac{M^2\kappa_\varrho^2}{2d}\right) \leq \epsilon \quad \Rightarrow$$

$$\mathbb{G}_n^*\left\{\boldsymbol{\lambda}(\cdot)[\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_\star^{(\varrho)}(\cdot\,;\mathbf{P}_r)\}]\right\} = O_p(h^q) \quad \forall\ \varrho\in\{0,1\}. \tag{27}$$

Combining (26) and (27) using the linearity of $\mathbb{G}_n^*(\cdot)$, we then have (A.1). ■

Next, to show (A.2), let $f(\mathbf{x};\mathbf{P}_r) = \varphi^{(0)}(\mathbf{x};\mathbf{P}_r)$ and $l(\mathbf{x};\mathbf{P}_r) = \varphi^{(1)}(\mathbf{x};\mathbf{P}_r)$. Then, we write

$$\mathbb{G}_n^*[\boldsymbol{\lambda}(\cdot)\{\widetilde{m}_\star(\cdot\,;\mathbf{P}_r)\}] \;=\; \mathbb{G}_n^*[\boldsymbol{\lambda}(\cdot)\{\widetilde{\mathbf{T}}^{(1)}_{n,\mathbf{P}_r}(\cdot) - \widetilde{\mathbf{T}}^{(2)}_{n,\mathbf{P}_r}(\cdot) - \widetilde{\mathbf{T}}^{(3)}_{n,\mathbf{P}_r}(\cdot) + \widetilde{\mathbf{T}}^{(4)}_{n,\mathbf{P}_r}(\cdot)\}],$$

where

$$\widetilde{\mathbf{T}}^{(1)}_{n,\mathbf{P}_r}(\mathbf{x}) = \frac{\widetilde{l}_\star(\mathbf{x};\mathbf{P}_r)}{f(\mathbf{x};\mathbf{P}_r)}, \qquad \widetilde{\mathbf{T}}^{(2)}_{n,\mathbf{P}_r}(\mathbf{x}) = \frac{\widetilde{f}_\star(\mathbf{x};\mathbf{P}_r)l(\mathbf{x};\mathbf{P}_r)}{f(\mathbf{x};\mathbf{P}_r)^2},$$

$$\widetilde{\mathbf{T}}^{(3)}_{n,\mathbf{P}_r}(\mathbf{x}) = \frac{\widetilde{l}_\star(\mathbf{x};\mathbf{P}_r)\widetilde{f}_\star(\mathbf{x};\mathbf{P}_r)}{\widetilde{f}(\mathbf{x};\mathbf{P}_r)f(\mathbf{x};\mathbf{P}_r)}, \text{ and } \widetilde{\mathbf{T}}^{(4)}_{n,\mathbf{P}_r}(\mathbf{x}) = \frac{l(\mathbf{x};\mathbf{P}_r)\widetilde{f}_\star(\mathbf{x};\mathbf{P}_r)^2}{\widetilde{f}(\mathbf{x};\mathbf{P}_r)f(\mathbf{x};\mathbf{P}_r)^2}. \tag{28}$$

Since $\boldsymbol{\lambda}^{(1)}_{\mathbf{P}_r}(\mathbf{x}) \equiv \boldsymbol{\lambda}(\mathbf{x})f(\mathbf{x};\mathbf{P}_r)^{-1}$ and $\boldsymbol{\lambda}^{(2)}_{\mathbf{P}_r}(\mathbf{x}) \equiv \boldsymbol{\lambda}(\mathbf{x})l(\mathbf{x};\mathbf{P}_r)f(\mathbf{x};\mathbf{P}_r)^{-2}$ are bounded a.s. $[\mathbb{P}_{\mathbf{X}}]$ due to Assumption 4.1 (iii)-(iv) and the boundedness of $\boldsymbol{\lambda}(\cdot)$, using these as choices of '$\boldsymbol{\lambda}(\cdot)$' in (A.1), we have:

$$\mathbb{G}_n^*\{\boldsymbol{\lambda}^{(1)}_{\mathbf{P}_r}(\cdot)\widetilde{l}_\star(\cdot\,;\mathbf{P}_r)\} = \mathbb{G}_n^*\{\boldsymbol{\lambda}(\cdot)\widetilde{\mathbf{T}}^{(1)}_{n,\mathbf{P}_r}(\cdot)\} = O_p(b_n^{(1)}),$$
$$\mathbb{G}_n^*\{\boldsymbol{\lambda}^{(2)}_{\mathbf{P}_r}(\cdot)\widetilde{f}_\star(\cdot\,;\mathbf{P}_r)\} = \mathbb{G}_n^*\{\boldsymbol{\lambda}(\cdot)\widetilde{\mathbf{T}}^{(2)}_{n,\mathbf{P}_r}(\cdot)\} = O_p(b_n^{(1)}).$$

Further, for each $s \in \{3,4\}$, $\sup_{\mathbf{x}\in\mathcal{X}} \|\widetilde{\mathbf{T}}^{(s)}_{n,\mathbf{P}_r}(\mathbf{x})\| \leq O_p(a^2_{n,2})$ which follows from repeated use of (15) along with Assumption 4.1 (iii)-(iv). Consequently, with $\boldsymbol{\lambda}(\cdot)$ bounded a.s. $[\mathbb{P}_{\mathbf{X}}]$, for each $s \in \{3,4\}$, $\mathbb{G}^*_n\{\boldsymbol{\lambda}(\cdot)\widetilde{\mathbf{T}}^{(s)}_{n,\mathbf{P}_r}(\cdot)\}$ is bounded by: $O_p(n^{\frac{1}{2}}a^2_{n,2})$. Combining all these results using the linearity of $\mathbb{G}^*_n(\cdot)$, we finally obtain: $\mathbb{G}^*_n\{\boldsymbol{\lambda}(\cdot)\widetilde{m}_\star(\cdot\,;\mathbf{P}_r)\} = O_p(b^{(1)}_n + n^{\frac{1}{2}}a^2_{n,2}) = O_p(n^{\frac{1}{2}}a^2_{n,2})$, thus leading to (A.2). The proof of the lemma is now complete. ■

VI.2. Proof of Lemma A.3. Throughout this proof, all additional notations introduced, if not explicitly defined, are understood to have been adopted from the proof of Theorem 4.1 in Section V. Now, using (17), $\widehat{\varphi}^{(\varrho)}(\mathbf{x};\widehat{\mathbf{P}}_r) - \widetilde{\varphi}^{(\varrho)}(\mathbf{x};\mathbf{P}_r) = \text{trace}\{(\widehat{\mathbf{P}}'_r - \mathbf{P}'_r)\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}}\} + \text{trace}\{(\widehat{\mathbf{P}}'_r - \mathbf{P}'_r)\widehat{\mathbf{M}}^{(2)}_{n,\varrho,\mathbf{x}}\}$, and $\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}} = \widehat{\mathbf{M}}^{(1,1)}_{n,\varrho,\mathbf{x}} - \widehat{\mathbf{M}}^{(1,2)}_{n,\varrho,\mathbf{x}}$, as defined in Section V. Thus,

$$\mathbb{G}^*_n[\boldsymbol{\lambda}(\cdot)\{\widehat{\varphi}^{(\varrho)}(\cdot\,;\widehat{\mathbf{P}}_r) - \widetilde{\varphi}^{(\varrho)}(\cdot\,;\mathbf{P}_r)\}] = \mathbb{G}^*_n\left\{\widehat{\boldsymbol{\zeta}}^{(1,1)}_{n,\varrho,\boldsymbol{\lambda}}(\cdot) - \widehat{\boldsymbol{\zeta}}^{(1,2)}_{n,\varrho,\boldsymbol{\lambda}}(\cdot) + \widehat{\boldsymbol{\zeta}}^{(2)}_{n,\varrho,\boldsymbol{\lambda}}(\cdot)\right\},$$

where $\forall\ (\omega) \in \{(1,1),(1,2),(2)\}$, $\varrho \in \{0,1\}$, and $\mathbf{x} \in \mathcal{X}$,

$$\widehat{\boldsymbol{\zeta}}^{(\omega)}_{n,\varrho,\boldsymbol{\lambda}}(\mathbf{x}) = \boldsymbol{\lambda}(\mathbf{x})\ \text{trace}\left\{(\widehat{\mathbf{P}}'_r - \mathbf{P}'_r)\widehat{\mathbf{M}}^{(\omega)}_{n,\varrho,\mathbf{x}}\right\}. \tag{29}$$

Then, $\forall\ s \in \{1,2\}$ and $l \in \{1,\ldots,d\}$, each element of

$$\int \boldsymbol{\lambda}_{[l]}(\mathbf{x})\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,\mathbf{x}}\mathbb{P}_n(d\mathbf{x}) = n^{-2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\mathbb{H}^{(n,s)}_{l,\varrho}(i_1,i_2)$$

is a V-statistic, where

$$\mathbb{H}^{(n,s)}_{l,\varrho}(i_1,i_2) = h^{-(r+1)}\boldsymbol{\lambda}_{[l]}(\mathbf{X}_{i_1})Y^{\varrho}_{i_2}\mathbf{U}^{(s)}(i_1,i_2)\boldsymbol{\nabla}K'\{\mathbf{P}'_r(\mathbf{X}_{i_1} - \mathbf{X}_{i_2})/h\}$$

with $\mathbf{U}^{(1)}(i_1,i_2) = \mathbf{X}_{i_1}$ and $\mathbf{U}^{(2)}(i_1,i_2) = \mathbf{X}_{i_2}$. Hence, similar to the proof of (26), using Lemma 8.4 of Newey and McFadden (1994) with $\mathcal{X}$ compact, $\boldsymbol{\nabla}K(\cdot)$ and $\boldsymbol{\lambda}(\cdot)$ bounded, and Y^{ϱ} having finite 2^{nd} moments, we have: for each $l \in \{1,\ldots,d\}$, $s \in \{1,2\}$ and $\varrho \in \{0,1\}$,

$$\left\|\mathbb{G}^*_n\left[\boldsymbol{\lambda}_{[l]}(\cdot)\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,(\cdot)} - \mathbb{E}_{\mathcal{L}}\left\{\boldsymbol{\lambda}_{[l]}(\cdot)\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,(\cdot)}\right\}\right]\right\|_{\max} = O_p\left(n^{-\frac{1}{2}}h^{-(r+1)}\right).$$

It then follows from $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$ that for each s and ϱ,

$$\mathbb{G}^*_n\left[\widehat{\boldsymbol{\zeta}}^{(1,s)}_{n,\varrho,\boldsymbol{\lambda}}(\cdot) - \mathbb{E}_{\mathcal{L}}\left\{\widehat{\boldsymbol{\zeta}}^{(1,s)}_{n,\varrho,\boldsymbol{\lambda}}(\cdot)\right\}\right] = O_p\left(\alpha_n n^{-\frac{1}{2}}h^{-(r+1)}\right). \tag{30}$$

Next, for any given l, s and ϱ, each element of $n^{\frac{1}{2}}\mathbb{G}_n^*[\mathbb{E}_{\mathcal{L}}\{\boldsymbol{\lambda}_{[l]}(\cdot)\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,(\cdot)}\}]$ is a centered sum of i.i.d. random variables which are bounded by:

$$\left\|\sup_{\mathbf{x}\in\mathcal{X}}\left\{\|\boldsymbol{\lambda}(\mathbf{x})\|\;|\mathbb{E}_{\mathcal{L}}(\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,\mathbf{x}})|\right\}\right\|_{\max} = O(1),$$

where the order follows from (20), (21) and the boundedness of $\boldsymbol{\lambda}(\cdot)$. Hence, similar to the proof of (27), using Hoeffding's inequality and that $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$, we have: $\forall\ l \in \{1,\ldots,d\}$, $s \in \{1,2\}$ and $\varrho \in \{0,1\}$,

$$\left\|\mathbb{G}_n^*\left[\mathbb{E}_{\mathcal{L}}\left\{\boldsymbol{\lambda}_{[l]}(\cdot)\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,(\cdot)}\right\}\right]\right\|_{\max} = O_p(1) \Rightarrow \mathbb{G}_n^*\left[\mathbb{E}_{\mathcal{L}}\left\{\widehat{\boldsymbol{\zeta}}^{(1,s)}_{n,\varrho,\boldsymbol{\lambda}}(\cdot)\right\}\right] = O_p(\alpha_n). \tag{31}$$

For any matrix $\mathbf{A}$, let us denote by $\mathbf{A}_{[a,b]}$ the $(a,b)^{th}$ element of $\mathbf{A}$. Now, to control $\mathbb{G}_n^*\{\widehat{\boldsymbol{\zeta}}^{(2)}_{n,\varrho,\boldsymbol{\lambda}}(.)\}$ in (29), note that $\|\mathbb{G}_n^*\{\widehat{\boldsymbol{\zeta}}^{(2)}_{n,\varrho,\boldsymbol{\lambda}}(\cdot)\}\|$ is bounded by:

$$\begin{aligned}
& n^{\frac{1}{2}}\sup_{\mathbf{x}\in\mathcal{X}}\|\boldsymbol{\lambda}(\mathbf{x})\|\sum_{a,b}\int\left|(\widehat{\mathbf{P}}'_r - \mathbf{P}'_r)_{[b,a]}\left(\widehat{\mathbf{M}}^{(2)}_{n,\varrho,\mathbf{x}}\right)_{[a,b]}\right|(\mathbb{P}_n + \mathbb{P}_{\mathbf{X}})(d\mathbf{x}) \\
& \leq n^{\frac{1}{2}}rp\sup_{\mathbf{x}\in\mathcal{X},\mathbf{X}\in\mathcal{X}}\{\|\boldsymbol{\lambda}(\mathbf{x})\|\,\|\mathbf{x}-\mathbf{X}\|\}\left\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\right\|_{\max}\widehat{\mathbb{Z}}^{\varrho *}_n \\
& \leq O_p\left(n^{\frac{1}{2}}\alpha_n\right)\widehat{\mathbb{Z}}^{\varrho *}_n,
\end{aligned} \tag{32}$$

where the last step follows from $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$ and the boundedness of $\mathcal{X}$ and $\boldsymbol{\lambda}(\cdot)$, and $\widehat{\mathbb{Z}}^{\varrho *}_n = \int \widehat{\mathbb{Z}}^{(\varrho)}_n(\mathbf{x})\,(\mathbb{P}_n + \mathbb{P}_{\mathbf{X}})(d\mathbf{x})$ with

$$\widehat{\mathbb{Z}}^{(\varrho)}_n(\mathbf{x}) = n^{-1}\sum_{i=1}^{n}\frac{|Y_i^{\varrho}|}{h^{r+1}}\left\|\boldsymbol{\nabla}K(\mathbf{w}_{i,\mathbf{x}}) - \boldsymbol{\nabla}K\left\{\frac{\mathbf{P}'_r(\mathbf{x}-\mathbf{X}_i)}{h}\right\}\right\|.$$

Now, $\|\mathbf{w}_{i,\mathbf{x}} - \mathbf{P}'_r(\mathbf{x}-\mathbf{X}_i)h^{-1}\| \leq \|(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'(\mathbf{x}-\mathbf{X}_i)h^{-1}\| \leq O_p(\alpha_n h^{-1})$ uniformly in $(i,\mathbf{x})$, as noted while proving (18). Further, with L^*, as defined in Assumption 4.1 (vii), let $\mathbb{A}_n$ denote the event: $\{\|(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'(\mathbf{x}-\mathbf{X}_i)h^{-1}\| \leq L^*\ \ \forall\,\mathbf{x}\in\mathcal{X},\ i = 1,..,n\}$. Then, with $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$, $\mathcal{X}$ compact and $\alpha_n h^{-1} = o(1)$ since $n^{\frac{1}{2}}\alpha_n^2 h^{-2} = o(1)$ as assumed, it follows that $\mathbb{P}(\mathbb{A}_n) \to 1$. Using these along with Assumption 4.1 (vii) and the function $\phi(.)$ defined therein, we have: on $\mathbb{A}_n$ with $\mathbb{P}(\mathbb{A}_n) \to 1$,

$$\begin{aligned}
\widehat{\mathbb{Z}}^{(\varrho)}_n(\mathbf{x}) & \leq \sum_{i=1}^{n}\frac{|Y_i^{\varrho}|}{nh^{r+1}}\left\|\frac{(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'(\mathbf{x}-\mathbf{X}_i)}{h}\right\|\phi\left\{\frac{\mathbf{P}'_r(\mathbf{x}-\mathbf{X}_i)}{h}\right\} \\
& \leq \sqrt{rp}\sup_{\mathbf{x}\in\mathcal{X},\mathbf{X}\in\mathcal{X}}\|\mathbf{x}-\mathbf{X}\|\left\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\right\|_{\max}\sum_{i=1}^{n}\frac{|Y_i^{\varrho}|}{nh^{r+2}}\phi\left\{\frac{\mathbf{P}'_r(\mathbf{x}-\mathbf{X}_i)}{h}\right\}.
\end{aligned}$$

Thus, $\widehat{\mathbb{Z}}_n^{\varrho *} \leq O_p\left(\alpha_n \widetilde{\mathbb{Z}}_n^{\varrho *}\right)$, where $\widetilde{\mathbb{Z}}_n^{\varrho *} = \int \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x})(\mathbb{P}_n + \mathbb{P}_{\mathbf{X}})(d\mathbf{x})$,

$$\widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x};\mathbf{Z}_i), \text{ and } \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x};\mathbf{Z}) = \frac{|Y^{\varrho}|}{h^{r+2}}\phi\left\{\frac{\mathbf{P}_r'(\mathbf{x}-\mathbf{X})}{h}\right\}.$$

Let $\mathbf{Z}^0 \equiv (Y^0, \mathbf{X}^{0'})' \sim \mathbb{P}_{\mathbf{Z}}$ be generated independent of $\mathcal{L}$, and define:

$$\begin{aligned}
\widetilde{\mathbb{U}}_{n,\varrho}^{(1)} &= n^{-1}\sum_{i=1}^{n} \mathbb{E}_{\mathbf{X}^0}\{\widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{X}^0;\mathbf{Z}_i)\}, & \widetilde{\mathbb{U}}_{n,\varrho}^{(2)} &= n^{-1}\sum_{i=1}^{n} \mathbb{E}_{\mathbf{Z}^0}\{\widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{X}_i;\mathbf{Z}^0)\},\\
\widetilde{\mathbb{U}}_{n,\varrho}^{(1,1)} &= \mathbb{E}\{\widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{X}^0;\mathbf{Z}^0)\}, \quad \text{and} & \widetilde{\mathbb{V}}_{n,\varrho}^{(k)} &= \mathbb{E}\{\widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{X}^0;\mathbf{Z})^k\} \text{ for } k = 1, 2.
\end{aligned}$$

Then, first note that: $\int \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x})\mathbb{P}_{\mathbf{X}}(d\mathbf{x}) = \widetilde{\mathbb{U}}_{n,\varrho}^{(1)}$. Further, since

$$\int \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x})\mathbb{P}_n(d\mathbf{x}) = n^{-2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{X}_{i_1};\mathbf{Z}_{i_2})$$

is a V-statistic, we have:

$$\int \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x})\mathbb{P}_n(d\mathbf{x}) = \widetilde{\mathbb{U}}_{n,\varrho}^{(1)} + \widetilde{\mathbb{U}}_{n,\varrho}^{(2)} - \widetilde{\mathbb{V}}_{n,\varrho}^{(1)} + O_p\{n^{-1}\widetilde{\mathbb{U}}_{n,\varrho}^{(1,1)} + n^{-1}(\widetilde{\mathbb{V}}_{n,\varrho}^{(2)})^{\frac{1}{2}}\}$$

using Lemma 8.4 of Newey and McFadden (1994). Then, with all notations as above, we have:

$$n^{-1}\widetilde{\mathbb{U}}_{n,\varrho}^{(1,1)} + n^{-1}(\widetilde{\mathbb{V}}_{n,\varrho}^{(2)})^{\frac{1}{2}} \leq O_p\left(n^{-1}h^{-(r+2)}\right), \tag{33}$$

$$\begin{aligned}
\text{and} \qquad \widetilde{\mathbb{U}}_{n,\varrho}^{(1)} &= \frac{1}{nh^{r+2}}\sum_{i=1}^{n}|Y_i^{\varrho}|\int_{\mathcal{X}_{\mathbf{P}_r}} \phi\left(\frac{\mathbf{w}-\mathbf{P}_r'\mathbf{X}_i}{h}\right) f_{\mathbf{P}_r}(\mathbf{w})d\mathbf{w}\\
&\leq \frac{B_{\mathbf{P}_r}}{nh^2}\sum_{i=1}^{n}\left\{|Y_i^{\varrho}|\int_{A_{\mathbf{X}_i}^n} \phi(\boldsymbol{\psi}_i)d\boldsymbol{\psi}_i\right\},\\
&\leq \frac{B_{\mathbf{P}_r}}{h^2}\left\{\int_{\mathbb{R}^r}\phi(\boldsymbol{\psi})d\boldsymbol{\psi}\right\}\left\{n^{-1}\sum_{i=1}^{n}|Y_i^{\varrho}|\right\} \leq O_p\left(h^{-2}\right), \qquad (34)
\end{aligned}$$

where $\boldsymbol{\psi}_i = h^{-1}(\mathbf{w} - \mathbf{P}_r'\mathbf{X}_i)\ \forall\ i$, $A_{\mathbf{x}}^n = \{\boldsymbol{\psi} : (\mathbf{P}_r'\mathbf{x} + h\boldsymbol{\psi}) \in \mathcal{X}_{\mathbf{P}_r}\}\ \forall\ \mathbf{x} \in \mathcal{X}$, and $B_{\mathbf{P}_r} = \sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}} f_{\mathbf{P}_r}(\mathbf{w}) < \infty$. The error rate in (33) follows since $\phi(\cdot)$ is bounded and Y^{ϱ} has finite 2^{nd} moments, while that of $\widetilde{\mathbb{U}}_{n,\varrho}^{(1)}$ follows from Assumption 4.1 (iii), integrability of $\phi(\cdot)$, and LLN applied to the sequence

$\{Y_i^\varrho\}_{i=1}^n$ having finite 2^{nd} moments. Now, note that $\widetilde{\mathbb{U}}_{n,\varrho}^{(2)} - \widetilde{\mathbb{V}}_{n,\varrho}^{(1)}$ is a centered average of $[\mathbb{E}_{\mathbf{Z}^0}\{\widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{X}_i;\mathbf{Z}^0)\}]_{i=1}^n$ which are i.i.d. and bounded by:

$$\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbf{Z}}\{\widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x};\mathbf{Z})\} = \sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{h^{r+2}} \int_{\mathcal{X}_{\mathbf{P}_r}} \phi\left(\frac{\mathbf{P}_r'\mathbf{x}-\mathbf{w}}{h}\right) \overline{m}_{\mathbf{P}_r}^{(\varrho)}(\mathbf{w}) f_{\mathbf{P}_r}(\mathbf{w})d\mathbf{w},$$

where $\overline{m}_{\mathbf{P}_r}^{(\varrho)}(\mathbf{w}) = \mathbb{E}(|Y|^\varrho \mid \mathbf{X}_{\mathbf{P}_r} = \mathbf{w})\ \forall\ \varrho \in \{0,1\}$ and $\mathbf{w} \in \mathcal{X}_{\mathbf{P}_r}$. Using the integrability of $\phi(\cdot)$, we then have:

$$\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbf{Z}}\{\widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x};\mathbf{Z})\} \ \le\ \sup_{\mathbf{x}\in\mathcal{X}} \frac{C_{\mathbf{P}_r}^{(\varrho)}}{h^{r+2}} \int_{\mathcal{X}_{\mathbf{P}_r}} \phi\left(\frac{\mathbf{P}_r'\mathbf{x}-\mathbf{w}}{h}\right) d\mathbf{w}$$
$$\le \sup_{\mathbf{x}\in\mathcal{X}} \frac{C_{\mathbf{P}_r}^{(\varrho)}}{h^2} \int_{A_{\mathbf{x}}^n} \phi(-\boldsymbol{\psi})\, d\boldsymbol{\psi} \le \frac{C_{\mathbf{P}_r}^{(\varrho)}}{h^2} \left\{\int_{\mathbb{R}^r} \phi(\boldsymbol{\psi})d\boldsymbol{\psi}\right\} \ =\ O\left(h^{-2}\right),$$

where $C_{\mathbf{P}_r}^{(\varrho)} = \sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}} \overline{m}_{\mathbf{P}_r}^{(\varrho)}(\mathbf{w}) f_{\mathbf{P}_r}(\mathbf{w}) < \infty$ due to Assumption 4.1 (iii), and $A_{\mathbf{x}}^n = \{\boldsymbol{\psi} : (\mathbf{P}_r'\mathbf{x} + h\boldsymbol{\psi}) \in \mathcal{X}_{\mathbf{P}_r}\}$, as before. It then follows, similar to the proof of (27), from a simple application of Hoeffding's inequality that

$$\widetilde{\mathbb{U}}_{n,\varrho}^{(2)} - \widetilde{\mathbb{V}}_{n,\varrho}^{(1)} = O_p\left(n^{-\frac{1}{2}}h^{-2}\right). \tag{35}$$

Using (33)-(35), we finally have: $\widetilde{\mathbb{Z}}_n^{\varrho*} = O_p(h^{-2} + n^{-1}h^{-(r+2)})$. Hence,

$$\widehat{\mathbb{Z}}_n^{\varrho*} = \int \widehat{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x})\,(\mathbb{P}_n + \mathbb{P}_{\mathbf{X}})(d\mathbf{x}) \le O_p\left(\alpha_n \widetilde{\mathbb{Z}}_n^{\varrho*}\right) = O_p\left(\frac{\alpha_n}{h^2} + \frac{\alpha_n}{nh^{r+2}}\right), \tag{36}$$

$$\text{and} \quad \left\|\mathbb{G}_n^*\left\{\widehat{\boldsymbol{\zeta}}_{n,\varrho,\boldsymbol{\lambda}}^{(2)}(\cdot)\right\}\right\| \le O_p\left(\frac{n^{\frac{1}{2}}\alpha_n^2}{h^2} + \frac{n^{\frac{1}{2}}\alpha_n^2}{nh^{r+2}}\right) \quad \forall\ \varrho \in \{0,1\}, \tag{37}$$

where the final bound in (37) follows from (32). The desired result in (A.3) now follows by applying (30), (31) and (37) to (29) using the linearity of $\mathbb{G}_n^*(\cdot)$. The proof of the lemma is now complete. (Note that conditions (i), (iv) and (viii) in Assumption 4.1 were actually not used in this proof). ■

VI.3. Proof of Theorem 4.2. Finally, to establish the result of Theorem 4.2, let $\boldsymbol{\lambda}_0(\mathbf{x}) = \overrightarrow{\mathbf{x}}$ which is measurable and bounded on $\mathcal{X}$. Further, with $\mathbb{G}_n^*(\cdot)$ as defined in Appendix A.1, note that $\mathbb{G}_{n,\mathbb{K}}$ for $\mathbb{K} = 1$ is given by:

$$\mathbb{G}_{n,\mathbb{K}} = \mathbb{G}_n^*\{\boldsymbol{\lambda}_0(\cdot)\widetilde{m}_\star(\cdot\,;\mathbf{P}_r)\} + \mathbb{G}_n^*[\boldsymbol{\lambda}_0(\cdot)\{\widehat{m}(\cdot\,;\widehat{\mathbf{P}}_r) - \widetilde{m}(\cdot\,;\mathbf{P}_r)\}], \tag{38}$$

due to linearity of $\mathbb{G}_n^*(\cdot)$. Now, using Lemma A.2, we have:

$$\mathbb{G}_n^*\{\boldsymbol{\lambda}_0(\cdot)\widetilde{m}_\star(\cdot\,;\mathbf{P}_r)\} = O_p(n^{\frac{1}{2}}a_{n,2}^2) = O_p(a_{n,2}^*). \tag{39}$$

The second term $\mathbb{G}_n^*[\boldsymbol{\lambda}_0(\cdot)\{\widehat{m}(\cdot\,;\widehat{\mathbf{P}}_r) - \widetilde{m}(\cdot\,;\mathbf{P}_r)\}]$ in (38) can be written as:

$$\begin{aligned} &\mathbb{G}_n^*[\boldsymbol{\lambda}_0(\cdot)\{\widehat{\mathbf{T}}^{(1)}_{n,\mathbf{P}_r}(\cdot) - \widehat{\mathbf{T}}^{(2)}_{n,\mathbf{P}_r}(\cdot) - \widehat{\mathbf{T}}^{(3)}_{n,\mathbf{P}_r}(\cdot) + \widehat{\mathbf{T}}^{(4)}_{n,\mathbf{P}_r}(\cdot)\}] \\ &= O_p\left(b_n^{(2)} + n^{\frac{1}{2}}a_{n,1}^2 + n^{\frac{1}{2}}a_{n,1}a_{n,2}\right) = O_p\left(a_{n,1}^*\right), \end{aligned} \tag{40}$$

where with slight abuse of notation,

$$\begin{aligned} \widehat{\mathbf{T}}^{(1)}_{n,\mathbf{P}_r}(\mathbf{x}) &= \frac{\widehat{a} - \widetilde{a}}{b}, \quad \widehat{\mathbf{T}}^{(2)}_{n,\mathbf{P}_r}(\mathbf{x}) = \frac{a(\widehat{b} - \widetilde{b})}{b^2}, \\ \widehat{\mathbf{T}}^{(3)}_{n,\mathbf{P}_r}(\mathbf{x}) &= \frac{(\widehat{a} - \widetilde{a})(\widetilde{b} - b)}{b\,\widetilde{b}} + \frac{(\widehat{a} - \widetilde{a})(\widehat{b} - \widetilde{b})}{\widetilde{b}\,\widehat{b}}, \quad \text{and} \\ \widehat{\mathbf{T}}^{(4)}_{n,\mathbf{P}_r}(\mathbf{x}) &= \frac{\widetilde{a}(\widehat{b} - \widetilde{b})^2}{\widehat{b}\,b^2} - \frac{(\widetilde{a} - a)(\widehat{b} - \widetilde{b})}{b^2} + \frac{a(\widehat{b} - \widetilde{b})(\widetilde{b} - b)(b + \widetilde{b})}{(b\,\widetilde{b})(b\,\widehat{b})}, \end{aligned}$$

with $(a, b) = \{l(\mathbf{x};\mathbf{P}_r), f(\mathbf{x};\mathbf{P}_r)\}$, $(\widetilde{a}, \widetilde{b}) = \{\widetilde{l}(\mathbf{x};\mathbf{P}_r), \widetilde{f}(\mathbf{x};\mathbf{P}_r)\}$ and $(\widehat{a}, \widehat{b}) = \{\widehat{l}(\mathbf{x};\widehat{\mathbf{P}}_r), \widehat{f}(\mathbf{x};\widehat{\mathbf{P}}_r)\}\}$.

For (40), the starting expansion is due to a linearization similar to (28), while the final rate is due to the following: note that $\boldsymbol{\lambda}^{(1)}_{0,\mathbf{P}_r}(\cdot) \equiv b^{-1}\boldsymbol{\lambda}_0(\cdot)$ and $\boldsymbol{\lambda}^{(2)}_{0,\mathbf{P}_r}(\cdot) \equiv ab^{-2}\boldsymbol{\lambda}_0(\cdot)$ are both bounded a.s. $[\mathbb{P}_{\mathbf{X}}]$ due to Assumption 4.1 (iii)-(iv) and the boundedness of $\boldsymbol{\lambda}_0(\cdot)$. Hence, using these as choices of '$\boldsymbol{\lambda}(\cdot)$' in Lemma A.3, we have: $\mathbb{G}_n^*\{(\widehat{a} - \widetilde{a})\boldsymbol{\lambda}^{(1)}_{0,\mathbf{P}_r}(\cdot)\} = \mathbb{G}_n^*[\boldsymbol{\lambda}_0(\cdot)\{\widehat{\mathbf{T}}^{(1)}_{n,\mathbf{P}_r}(\cdot)\}] = O_p(b_n^{(2)})$ and $\mathbb{G}_n^*\{(\widehat{b} - \widetilde{b})\boldsymbol{\lambda}^{(2)}_{0,\mathbf{P}_r}(\cdot)\} = \mathbb{G}_n^*[\boldsymbol{\lambda}_0(\cdot)\{\widehat{\mathbf{T}}^{(2)}_{n,\mathbf{P}_r}(\cdot)\}] = O_p(b_n^{(2)})$ respectively. Further, note that for each $s \in \{3, 4\}$, $\sup_{\mathbf{x}\in\mathcal{X}}\|\widehat{\mathbf{T}}^{(s)}_{n,\mathbf{P}_r}(\mathbf{x})\| \leq O_p(a_{n,1}^2 + a_{n,1}a_{n,2})$ which follows from repeated use of (15), (24) along with Assumption 4.1 (iii)-(iv). Consequently, with $\boldsymbol{\lambda}_0(\mathbf{x})$ bounded a.s. $[\mathbb{P}_{\mathbf{X}}]$, for each $s \in \{3, 4\}$, $\mathbb{G}_n^*[\boldsymbol{\lambda}_0(\cdot)\{\widehat{\mathbf{T}}^{(s)}_{n,\mathbf{P}_r}(\cdot)\}]$ is bounded by: $O_p(n^{\frac{1}{2}}a_{n,1}^2 + n^{\frac{1}{2}}a_{n,1}a_{n,2})$. Combining all these results using the linearity of $\mathbb{G}_n^*(\cdot)$ and noting that with $a_{n,2}^* = o(1)$, $(b_n^{(2)} + n^{\frac{1}{2}}a_{n,1}^2 + n^{\frac{1}{2}}a_{n,1}a_{n,2}) = O(a_{n,1}^*)$, (40) now follows and, along with (39) and (38), implies: $\mathbb{G}_{n,\mathbb{K}} = O_p(a_{n,1}^* + a_{n,2}^*)$ as claimed in Theorem 4.2. Lastly, using this in (3.10), the expansion in (4.3) and its associated implications follow. The proof of Theorem 4.2 is now complete. ■

Abhishek Chakrabortty
Department of Biostatistics
Harvard University
Boston, MA 02115, USA.
E-mail: achakrabortty@mail.harvard.edu

Tianxi Cai
Department of Biostatistics
Harvard University
Boston, MA 02115, USA.
E-mail: tcai@hsph.harvard.edu