University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2005 *Paper* 194

A General Imputation Methodology for Nonparametric Regression with Censored Data

Dan Rubin* Mark J. van der Laan†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, daniel.rubin@fda.hhs.gov

†Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper194

A General Imputation Methodology for Nonparametric Regression with Censored Data

Dan Rubin and Mark J. van der Laan

Abstract

We consider the random design nonparametric regression problem when the response variable is subject to a general mode of missingness or censoring. A traditional approach to such problems is imputation, in which the missing or censored responses are replaced by well-chosen values, and then the resulting covariate/response data are plugged into algorithms designed for the uncensored setting. We present a general methodology for imputation with the property of double robustness, in that the method works well if either a parameter of the full data distribution (covariate and response distribution) or a parameter of the censoring mechanism is well approximated. These procedures can be used advantageously when something is known about the censoring mechanism (i.e. when the censoring variable is independent of the survival time and response, in survival analysis), while methods based on maximizing a likelihood ignore this relevant information. We show how the methodology can be applied to examples where the response variable is missing, corresponds to a counterfactual outcome in a point treatment study, is right censored, or is subject to censoring as in current status data. To deal with identifiability problems (i.e. the conditional mean survival time may not be available from right censored data because of a lack of information regarding the survival distribution's tails), we show for these examples how the response of interest can be transformed, so that nonparametric regression remains a worthwhile endeavor. We remark on how our imputation procedure can be implemented by using general tools from efficiency theory and semiparametric estimation. General results are presented demonstrating how imputation procedures can accurately approximate regression functions when the imputed responses are entered into commonly used nonparametric regression procedures, including least squares estimators, complexity regularized least squares estimators, penalized least squares estimators, locally weighted average estimators, and estimators selected with cross-validation.

1 Introduction

Random design nonparametric regression is a well studied problem in the statistical literature. Substantial progress has been made in recent decades, with theoretical advances such as the determination of minimax rates of convergence for a variety of smoothness classes and loss functions, approximation properties of neural networks and other elaborate function classes, and the application of empirical process techniques. See Györfi et al. (2002) for a recent overview. Many new algorithms for the applied regression setting have also been introduced and examined, such as MARS, random forests, and support vector machine regression, as discussed in Hastie et al. (2001). Regression is a popular subject of study because informally the regression function provides the best prediction of a *response* given *covariates*, and nonparametric methods are required because current data is high-dimensional or complicated enough so that any assumed parametric or semiparametric model would almost certainly be misspecified.

However, nonparametric regression approaches are often avoided or do not produce reliable answers when the response data is subject to censoring or missingness. For instance, problems can arise in data structures such as

- Survey sampling. Here the response may simply be missing.
- Drug studies. Here the desired response might be a whole set of *counterfactuals*. Each subject may have been given a different dosage of a drug, but we can think of the full (uncensored) data for each subject as the set of responses they would have obtained from each different dosage level. Determining conditional counterfactual mean responses is an example of *causal inference*, and is a generalization of the missing data problem.
- Survival analysis. Here the response is a time until an event (e.g. death or relapse in a clinical trial), and the interest could be in predicting the conditional mean (truncated) survival (or log survival) time given baseline covariates. If patients are only followed up until certain random times, they are subject to *right censoring*.
- Cross-sectional data. Here the response again might be a survival time. Instead of being monitored until a certain random time, each subject might be examined at only one random time, and it would noted whether or not the survival time exceeds the random monitoring time. This would be an example of *current status data*.

We should note that the regression function may not actually be identifiable from the observed data for such censored data structures, particularly for the right censored and current status data as described above. With right censored data a small censoring time may prevent knowledge of the survival distribution's tails, and hence of its mean or conditional mean. Consequently, we will typically need to perform regression on some sort of transformed response, such as only considering counterfactual responses that have high probabilities of being observed, considering the conditional mean truncated (log) survival time with right censored data, or the conditional mean interval truncated

(log) survival time with current status data. Interest in nonparametric estimation for such data structures has typically been in estimating the conditional cumulative distribution function at points where it is identifiable, precisely to avoid the problems that can arise in regression, as discussed in Beran (1981) and Dabrowska (1989). Given new covariates, a statistician will often report or plot an entire estimated distribution for the response, rather than a single prediction as would be the case for regression. Nonparametric regression on a transformed response can be considered as a way of condensing this conditional distribution into a single number, facilitating comparisons between groups of subjects with different covariate values.

One popular method for performing regression with missing data is *imputation*, or replacing the missing responses with well-chosen values, and then plugging the covariate/response data into a full data (non-missing) regression algorithm. Imputation methods can be applied to any algorithm or black box method that could have been used with full data. This is advantageous if one feels that a particular type of fit will be useful or interpretable for the problem of interest, and also makes for a trivial implementation, modulo estimation of the imputation mapping. A benefit of imputation methods is that they can rely on a large body of literature and software designed for the full data nonparametric regression problem.

Unfortunately, there is no consensus as to how this imputation should be performed for general types of censored data structures. In this paper, we propose a general imputation methodology with the following desirable properties.

- The proposed procedure is *doubly robust*, in that estimating *either* of the F (full data) or G (censoring mechanism) parts of the likelihood will lead to good results, essentially giving the statistician two chances to estimate the desired parameter. This is an advantage over all methods relying on maximum likelihood, which will ignore any information about the censoring mechanism due to the factorization of the likelihood under coarsening at random (to be described in the sequel), as shown in Gill et al. (1997). Double robustness is particularly relevant when one can assume a lot of information about the censoring mechanism, such as missingness being completely at random, or a censoring time in survival analysis being unrelated a patient's baseline covariates and survival time.

- The methodology can be applied to any coarsening at random data structure where one would conceivably want to perform regression with uncensored data, through estimating the explicit imputation mapping to be given in (15). This generality across censored data structures partially ameliorates the problem of finding clever representations to handle censoring, such as proportional hazards modeling in survival analysis, which might rarely be used for other types of censoring (or full data).

- Our method generalizes commonly implemented *inverse probability of censoring weighted* estimators, as reviewed in Rotnitzky and Robins (2003) in the survival analysis setting. Inverse weighting applies in situations where there is positive probability of the full data being observed. The response is set to zero if there is any censoring, and otherwise it is inverse weighted by the probability that it is

observed. This simple method does take advantage of knowledge of the censoring mechanism, but is not doubly robust, and does not apply to all censored data structures. In section 3, we show how inverse probability of censoring weighted imputation methods emerge as special cases of the proposed doubly robust procedure.

In section 2 we formally define the problem of interest and introduce our estimator. In section 3 we give examples of the estimator for the different censored data structures listed previously, and show the implications of the double robustness property. In section 4 we provide a consistent estimator of the imputation mapping. Although the estimator of this section may not be practically useful, we discuss how it shows that the *irregular* problem of estimating the conditional mean response can be treated with efficiency theory (i.e. solving the *regular* problem of estimating the unconditional mean response). In section 5 we provide theoretical results for our imputation method, when imputed responses are plugged into least squares estimators, complexity regularized least squares estimators, penalized least squares estimators, locally weighted average estimators, and estimators selected with cross-validation. In this section show that the squared L^2 error of the regression estimator can typically be decomposed into a sharp full data bound (based on the regression error of using any covariate/response data for which the response has the same conditional mean as the censored response of interest) added to an imputation remainder (which can be made small if either a parameter of the full data distribution F or the censoring mechanism G is well approximated). All proofs appear in the appendix.

2 Formal Setup, Problem, and Estimator

Consider the triplet of random variables (X, C, O) defined on a probability space $(\Omega, \mathcal{F}, \mu)$, with support $\mathcal{X} \times \mathcal{C} \times \mathcal{O}$. Here X will denote the *full data*, or random variable we would have observed had there been no censoring. In the regression context, this will typically mean that $\sigma(Z), \sigma(Y) \subset \sigma(X)$, for Y a real-valued square integrable response, Z a vector-valued set of covariates that we would like to regress on the response, and $\sigma(\cdot)$ denoting the sigma field generated by a random variable. Our interest will be in estimating the regression function

$$m : z \rightarrow E[Y|Z = z]. \tag{1}$$

Here C is a *censoring variable* that determines how much of the full data we can actually observe. The *observed data* is defined by $O \equiv \Phi(X, C)$, for Φ a known measurable mapping of the full data and censoring variable. It is this data structure O that is assumed available to the statistician, based on i.i.d. copies $D_n \equiv \{O_1, ..., O_n\}$. If only the response is censored, we will assume that $\sigma(Z) \subset \sigma(O)$, so that the covariates are available from the observed data when estimating the desired regression function. We will let F denote the distribution of X, and P denote the distribution of O. We will further assume a regular conditional distribution G for the distribution of O given X, and recall that the regular conditional distribution always exists when (X, O) is defined

on a nice measurable space. It is clear that the distribution P is determined by the pair (F, G), so we will write $O \sim P = P_{F,G}$ to denote the distribution of the observed data. While we will not assume P to be known (that would defeat the purpose of estimating the regression function), we may specify a statistical model $\mathcal{M}$ (set of distributions) known to contain P, and likewise models $\mathcal{M}_{\mathcal{F}}$ and $\mathcal{M}_{\mathcal{G}}$ known to contain F and G. We will work with the standard L^2 risk, in that our interest will be in finding an estimator $m_n(\cdot) = m_n(\cdot|D_n) : \mathcal{Z} \to \mathcal{R}$ built from the observed data D_n to minimize

$$R(m_n, P_{F,G}) = E_{F,G}|m_n(Z) - m(Z)|^2, \tag{2}$$

where Z is drawn independently of $\{X_i, C_i, O_i\}_{i=1}^n$.

For the regression function $m : z \to E[Y|Z = z]$ to even be identifiable from the distribution P of the observed data O, we will generally need the assumption of *coarsening at random*. This notion was introduced for discrete random variables in Heitjan and Rubin (1991) and generalized in Gill et al. (1997). Our definition in this section is based on the latter reference, to which we refer for a more detailed discussion. For $o \in \mathcal{O}$, we let $\alpha(o)$ denote the restricted support of X implied by O being a coarsening of X. That is, we define

$$\alpha(o) \equiv \{x \in \mathcal{X} : o = \Phi(x, c) \text{ for some } c \in \mathcal{C}\} \tag{3}$$

and assume

$$(x, o) \to I(x \in \alpha(o)) \text{ is jointly measurable in } (x, o). \tag{4}$$

We then say that the regular conditional distribution $G_{O|X}$ of O given X satisfies coarsening at random if a version of G can be chosen so that for F-almost all $x, x' \in \mathcal{X}$

$$G_{O|X=x}(do) = G_{O|X=x'}(do) \text{ on } \{o : x \in \alpha(o)\} \cap \{o : x' \in \alpha(o)\}. \tag{5}$$

In words, this means that the conditional distribution of the observed data O given the full data value $X = x$ does not depend on the specific $x \in \mathcal{X}$, other than the requirement imposed by O being a coarsening of X. Gill et al. (1997) show that under coarsening at random, the likelihood factorizes into a part involving the full data distribution F and another part involving the censoring mechanism G. An implication is that maximum likelihood methods or procedures based on the likelihood principle will ignore any information about the censoring mechanism G, sometimes leading to less than optimal procedures. Such methods are a commonly used for nonparametric estimation with censored data, through specifying a sieve of increasingly large models for the data generating distribution, and fitting a member of the sieve depending on n with maximum likelihood.

For a simple example of how maximum likelihood can be less than optimal in censored data problems, consider the missing response example to be discussed in section 3. Here the full data is covariate/response $X = (Z, Y)$, but the observed data is $O = (Z, C, CY)$, for $C \in \{0, 1\}$ a missingness indicator. Coarsening at random is implied by $\{Y \perp C|Z\}$. Let $1 - \pi(w) = P(C = 0|W = w)$ denote the conditional probability of missingness. If $\pi(\cdot)$ is known and bounded away from zero (this might be the case for missingness *completely at random*, where $C \perp (Z, Y)$, so $\pi(\cdot)$ would

be a constant that could be easily estimated) a crude inverse probability of censoring weighted mapping would be

$$A_{\text{IPCW}}(O) = \frac{CY}{\pi(W)}. \tag{6}$$

It is easy to verify that $E[A_{\text{IPCW}}(O)|Z] = E[Y|Z]$ a.s., so that the inverse weighted responses would be valid imputed responses to be plugged into a nonparametric regression procedure. A procedure based on the likelihood principle would completely ignore the known mapping $\pi(\cdot)$, sometimes leading to poor estimators, as described in Robins and Ritov (1997).

We now present the proposed doubly robust imputation mapping. Consider the Hilbert spaces $L^2(F)$ and $L^2(P_{F,G})$ consisting of all measurable real-valued square integrable functions of X and O respectively, endowed with the inner products

$$< s_1(X), s_2(X) >_{L^2(F)} = E_F[s_1(X)s_2(X)] \tag{7}$$

$$< h_1(O), h_2(O) >_{L^2(P_{F,G})} = E_{F,G}[h_1(O)h_2(O)]. \tag{8}$$

We define the *score operator* $l_{F,G} : L^2(F) \to L^2(P_{F,G})$ as

$$l_{F,G}(s(X)) = E_{F,G}[s(X)|O]. \tag{9}$$

Its adjoint is clearly $l_G^T : L^2(P_{F,G}) \to L^2(F)$, given by

$$l_G^T(h(O)) = E_G[h(O)|X]. \tag{10}$$

Finally, we define the *information operator* $I_{F,G} : L^2(F) \to L^2(F)$ as the composition

$$I_{F,G} = l_G^T \circ l_{F,G}. \tag{11}$$

Recalling that Y is the full data response variable, we say that the *experimental censoring assumption* holds for $P_{F,G}$ whenever

$$\begin{aligned} &I_{F,G} : L^2(F) \to L^2(F) \text{ is one-to-one,} \\ &\text{(up to null sets, in that } I_{F,G}(s_1(X)) = I_{F,G}(s_2(X)) \text{ implies } s_1(X) = s_2(X) \text{ a.s.)} \\ &\text{and there exists } s(X) \in L^2(F) \text{ such that } I_{F,G}(s(X)) = Y \text{ a.s.} \end{aligned} \tag{12}$$

Sufficient conditions for this experimental censoring assumption are shown through the proof of Lemma 3.3 in (van der Laan, 1998). Specifically, if $\|h\|_{L^2(F)} > 0$ implies $\|l_{F,G}(h)\|_{L^2(P_{F,G})} > 0$ then the information operator is one-to-one. If there exists an $\epsilon > 0$ such that $\|l_{F,G}(h)\|_{L^2(P_{F,G})} \geq \epsilon \|h\|_{L^2(F)}$ then the information operator is onto. Hence, these two conditions together imply (12), and the inverse of the information operator is given by the Neumann series

$$I_{F,G}^{-1} = \sum_{i=0}^{\infty} (J - I_{F,G})^i, \tag{13}$$

for J the identity mapping. Whenever $\sigma(\Delta X) \subset \sigma(O)$ for $\Delta \in \{0, 1\}$, it follows immediately from this result that the experimental censoring assumption holds if there is an $\epsilon > 0$ such that

$$P(\Delta = 1|X) \geq \epsilon > 0 \text{ a.s.} \tag{14}$$

In words, this is a simple condition to check when the coarsening mechanism allows for the entire full data structure X to be part of the observed data, as is the case for several important examples discussed in the sequel, such as regression with a missing or right censored response.

Whenever the experimental censoring assumption (12) is satisfied we can define $I^{-1}_{F,G}(Y)$ as the unique (up to null sets) element of $L^2(F)$ that $I_{F,G}$ maps to Y. Our central object of study can now be defined as the random variable

$$A_{\mathrm{DR}}(O|F,G) \equiv l_{F,G} \circ I^{-1}_{F,G}(Y) \in L^2(P_{F,G}), \tag{15}$$

which we term the *double robust mapping* of the response Y, for reasons that will become apparent. We can now introduce our method for performing regression with censored data. *From the observed data $D_n = \{O_1, .., O_n\}$, our proposal is to form an estimate $A_n(\cdot) = A_n(\cdot|F_n, G_n)$ of the double robust mapping $A_{DR}(\cdot|F,G)$, and impute responses $\{A_n(O_1), ..., A_n(O_n)\}$ into standard nonparametric regression algorithms.*

We now attempt to informally motivate this proposal. We first define the set of *oracle imputation mappings* as the functions of the observed data with the same conditional mean as the response Y,

$$\mathcal{A}_{F,G} = \{h(O) : h \in L^2(P_{F,G}),\, E_{F,G}[h(O)|Z] = E_F[Y] \equiv m(Z) \text{ a.s.}\}. \tag{16}$$

Many nonparametric regression procedures have been introduced that make minimal assumptions on the covariate/response data generating distribution. Procedures such as nearest neighbor methods, certain types of spline smoothers, neural networks, decision trees, and other *black box* algorithms might estimate the regression function equally well when applied to the data

$$D_{n,\mathrm{FULL}} = \{(Z_i, Y_i)\}_{i=1}^n \tag{17}$$

or

$$D_{n,A} \equiv \{Z_i, A(O_i)\}_{i=1}^n,\, A(\cdot) \in \mathcal{A}_{F,G}, \tag{18}$$

in a sense we now describe.

Instead of making assumptions about ancillary features of the covariate/response distribution, nonparametric procedures typically directly target the regression function $m : z \to E[Y|Z = z] = E[A(O)|Z = z]$. Even though the laws $\mathcal{L}(\{Z, Y\})$ and $\mathcal{L}(\{Z, A(O)\})$ may be very different, the regression function is the same for both laws, so nonparametric procedures should ideally work well with either dataset as input. This would not necessarily be the case for non-adaptive procedures, in that a Gaussian linear model for $D_{n,\mathrm{FULL}}$ might not hold for $D_{n,A}$, so that the method would fail even if we would let the sample size n tend to infinity. Note that there might be no telling a priori whether a given nonparametric regression estimator will attain superior performance when fed the data $D_{n,\mathrm{FULL}}$ or $D_{n,A}$.

Regression procedures are often studied by their minimax performance with respect to a loss function, in that their risk is guaranteed to decrease at least by a certain rate with sample size n whenever the regression function $m(\cdot)$ belongs to a smoothness class. This rate is then compared to the best possible rate that can be guaranteed by any

procedure. For regression estimators guaranteed to achieve certain rates of convergence when the regression function lies in a specific smoothness classes, we can guarantee these rates regardless of whether we have the data $D_{n,\text{FULL}}$ or D_n. Consequently, results from the (uncensored) nonparametric regression literature can often be used to form sharp *full data bounds* $T_{n,A}$ for the risk of regression estimators built from covariates and *any* (unavailable) oracle imputations $A(\cdot) \in \mathcal{A}_{F,G}$.

The following lemma proven in the appendix shows that $A_{\text{DR}}(\cdot|F_1, G_1) \in \mathcal{A}_{F,G}$ under general conditions if either $F = F_1$ or $G = G_1$. Hence, the double robust mapping is a natural target for imputation, because $A_n(\cdot) = A_n(\cdot|F_n, G_n)$ can often be made close to an oracle imputation mapping $A \in \mathcal{A}_{F,G}$ if either F_n is close to F or G_n is close to G. Then, as we discuss, decent performance can be guaranteed.

Lemma 2.1. *Consider distributions $P_{F,G}$ and P_{F_1,G_1} on the observed data satisfying coarsening at random as in (5), with full data X, censoring variable C, observed data O, covariates Z, and response Y as specified in this section, such that*

i) P_{F_1,G_1} satisfies the experimental censoring assumption (12).

ii) $G_1(\cdot|X)$ satisfies $G(\cdot|X = x) << G_1(\cdot|X = x)$ for F-almost all $x \in \mathcal{X}$, so we can define the Radon-Nikodym derivative $\frac{dG}{dG_1}(\cdot|X = x)$ for F-almost all $x \in \mathcal{X}$. By Gill et al. (1997), this is can be written a.s. as a function of O. Suppose further that $\frac{dG}{dG_1}(O|X) \in L^2(P_{F_1,G_1})$.

Let the score operator l_{F_1,G_1}, its adjoint $l^T_{G_1}$, and the double robust mapping $A_{DR}(\cdot|F_1, G_1)$ be as in (9), (10), and (15). Suppose that $\sigma(Z) \subset \sigma(X)$ and $\sigma(Z) \subset \sigma(O)$, so that the random variable (covariate) Z is part of both the full and observed data. Then we have

$$E_{F,G}[A_{DR}(O|F_1, G_1)|Z] = E[Y|Z] = m(Z) \text{ a.s.} \tag{19}$$

if either $F = F_1$ or $G = G_1$. □

Our results in section 5 show that for commonly implemented types of regression estimators, the risk associated with using imputed response mapping $A_n(\cdot) = A_n(\cdot|D_n = \{O_1, ..., O_n\})$ can typically be bounded by a full data bound $T_{n,A}$ added to an *imputation remainder* $R_{n,A}$. The full data bound will be the risk associated with applying the nonparametric regression procedures to the oracle imputation data $D_{n,A}$. As previously discussed, this risk may be comparable in some sense to the risk associated with having uncensored data $D_{n,\text{FULL}}$. The imputation remainder will measure some distance between $A_n(\cdot)$ and $A(\cdot)$. Because these bounds hold for any oracle imputation mapping $A(\cdot) \in \mathcal{A}_{F,G}$, they can be applied to the oracle imputation mapping that is somehow the closest to A_n. Double robustness is then beneficial because the estimate $A_n(\cdot) = A_n(\cdot|F_n, G_n)$ of $A_{\text{DR}}(\cdot|F, G)$ can be made close to an oracle imputation mapping if either F_n is close to F or G_n is close to G. We can then control the imputation remainder, and consequently the risk of the imputation-based regression procedure.

3 Examples

In this section we show the double robust mappings for several censored data structures. Formal derivations for these mappings can be found in the book of van der Laan and Robins (2003), which was motivated by estimation problems in semiparametric models.

3.1 Regression with a Missing Response

The full data is given by $X = (W, Y)$ where W is a set of covariates and Y is a real-valued square-integrable response of interest. The regression function of interest is $m : z \to E[Y|Z = z]$ for $\sigma(Z) \subset \sigma(W)$. The observed data is given by $O = (W, C, CY)$, for $C \in \{0, 1\}$ an indicator of missingness. Such a data structure might exist in survey sampling problems, where a subject did not answer the question pertaining to the response variable. Coarsening at random is satisfied if

$$\{C \perp Y|W\}. \tag{20}$$

By (14), the experimental censoring assumption is satisfied if in addition,

$$P(C = 1|W) \geq \epsilon > 0 \text{ a.s.} \tag{21}$$

It can be show as in chapter six of van der Laan and Robins (2003) that the double robust mapping of the response Y is given by

$$A_{\text{DR}}(O|F, G) = \frac{YC}{\pi(W)} + (1 - \frac{C}{\pi(W)})Q(W) \tag{22}$$

for

$$\pi(W) = P(C = 1|W) \tag{23}$$

$$Q(W) = E[Y|C = 1, W]. \tag{24}$$

The nuisance parameters needed to evaluate the double robust mapping are the functions $\pi(\cdot)$ and $Q(\cdot)$, which respectively are determined by F and G. Clearly we can estimate π by π_n from binary regression, and estimate Q with Q_n from regressing Y on C and W. Our estimate of the double robust mapping is then

$$A_n(O) = \frac{YC}{\pi_n(W)} + (1 - \frac{C}{\pi_n(W)})Q_n(W). \tag{25}$$

With the fit $Q_n : W \to 0$, we obtain as a special case the estimated *inverse probability of censoring weighted* (IPCW) mapping

$$A_{n,\text{IPCW}}(O) = \frac{YC}{\pi_n(W)}. \tag{26}$$

Suppose that $|Y| \leq \beta_n$ a.s., that we force $|Q_n(W)| \leq \beta_n$ a.s., and we also force π_n to satisfy the boundedness condition in (21). Letting F_n and G_n denote fitted distributions agreeing with Q_n and π_n respectively, it is then easy to verify that

$$|A_n(O) - A_{\text{DR}}(O|F, G_n)| \leq R_{n,1} \equiv (1 + \epsilon^{-1})|Q_n(W) - Q(W)| \text{ a.s.} \tag{27}$$

$$|A_n(O) - A_{\text{DR}}(O|F_n, G)| \leq R_{n,2} \equiv 2\beta_n\epsilon^{-2}|\pi_n(W) - \pi(W)| \text{ a.s.} \tag{28}$$

In light of Lemma 2.1, it follows that there exists an oracle imputation mapping $A \in \mathcal{A}_{F,G}$ such that

$$E_{F,G}|A_n(O) - A(O)|^2 \leq \min(E_{F,G}[R_{n,1}^2], E_{F,G}[R_{n,2}^2]). \tag{29}$$

As $R_{n,1}$ measures how well F_n approximates F and $R_{n,2}$ measures how well G_n approximates G, we have demonstrated that there is an oracle imputation mapping with a small imputation remainder if we can accurately estimate either the full data distribution F or the censoring distribution G.

3.2 Causal Inference in Point-Treatment Studies

The full data is given by $X = (W, \{Y_c : c \in \mathcal{C}\})$, where W are covariates and $\{Y_c : c \in \mathcal{C}\}$ are real-valued counterfactual responses, for $\mathcal{C}$ an index set. The regression function of interest is

$$m : z \to E[\int_{c\in\mathcal{C}} \phi(c) Y_c \mu(dc) | Z = z] = \int_{c\in\mathcal{C}} \phi(c) E[Y_c | Z = z] \mu(dc) \tag{30}$$

where $\sigma(Z) \subset \sigma(W)$, for μ a measure on the subsets of $\mathcal{C}$ with the appropriate sigma-field, and $\phi(\cdot)$ is a known (not necessarily nonnegative) weight function. The observed data is given by $O = (W, C, Y_C)$, for $C \in \mathcal{C}$. Here $\{Y_c : c \in \mathcal{C}\}$ determines all responses that we would have liked to observe about a subject. For instance, $\mathcal{C}$ could represent different treatment choices for a subject in a medical study. In reality the subject will only be exposed to one treatment, but the counterfactual Y_c would represent the outcome that would have been observed had a subject (contrary to fact) taken treatment c. Coarsening at random is satisfied if

$$\{C \perp \{Y_c : c \in \mathcal{C}\} | W\}. \tag{31}$$

If $\mathcal{C}$ is a discrete set, μ corresponds to a counting measure, and $\phi(c) = I(c = c_0)$, then the desired regression function is the conditional mean of the counterfactual response Y_{c_0}. This framework also allows us to model the conditional mean of contrasts of counterfactual responses, which could be useful in the hypothetical medical study described above if we wanted to examine differences between two treatments. In such a setting, ϕ might be nonzero on $c \notin \{c_0, c_1\}$, with $\phi(c_0) = -\phi(c_1) = 1$. The experimental censoring assumption is satisfied if in addition C has a regular conditional distribution G given W, assumed a.s. mutually absolutely continuous with respect to μ, such that for some $\epsilon > 0$,

$$\phi(c) \neq 0 \text{ implies } \frac{dG}{d\mu}(c|W) \geq \epsilon > 0 \text{ a.s.} \tag{32}$$

It can be show as in chapter six of van der Laan and Robins (2003) that the double robust mapping of the unavailable response $\int \phi(c) Y_c \mu(dc)$ is given by

$$A_{\mathrm{DR}}(O|F, G) = \phi(C) \frac{Y_C - Q(C, W)}{dG/d\mu(C|W)} + \int \phi(c) Q(c, W) d\mu(c) \tag{33}$$

for the nuisance parameters

$$(c, w) \to \frac{dG}{d\mu}(c|W = w) \text{ as given above,} \tag{34}$$

$$Q(c, W) \equiv E[Y_C | C = c, W] \ . \tag{35}$$

We must estimate $\frac{dG}{d\mu}$ by $\frac{dG_n}{d\mu}$ through fitting a conditional distribution for C given W. When $\mathcal{C}$ is a finite set, this can be done using polychotomous regression. We can estimate Q by Q_n using any standard technique for regressing Y on C and W. Our estimate of the double robust mapping is then

$$A_n(O) = \phi(C) \frac{Y_C - Q_n(C, W)}{dG_n/d\mu(C|W)} + \int \phi(c) Q_n(c, W) \mu(dc). \tag{36}$$

With the fit $Q_n : (c, W) \to 0$, we obtain as a special case the estimated inverse probability of censoring weighted mapping

$$A_{n,\text{IPCW}}(O) = \phi(C)\frac{Y_C}{dG_n/d\mu(C|W)}. \tag{37}$$

Suppose that $|Y_C| \leq \beta_n$ a.s., that we force $|Q_n(c, W)| \leq \beta_n$ a.s., and we also force $\frac{dG_n}{d\mu}$ to satisfy the boundedness condition in (32). Letting F_n and G_n denote fitted distributions that are consistent with Q_n and $\frac{dG_n}{d\mu}$ respectively, it is then easy to verify that

$$|A_n(O) - A_{\text{DR}}(O|F, G_n)| \leq R_{n,1} \equiv (1 + \epsilon^{-1}) \int_{c\in\mathcal{C}} \phi(c)|Q_n(c, W) - Q(c, W)|\mu(dc) \text{ a.s.} \tag{38}$$

$$|A_n(O) - A_{\text{DR}}(O|F_n, G)| \leq R_{n,2} \equiv 2\beta_n\epsilon^{-2} \int_{c\in\mathcal{C}} \phi(c)|dG_n(c|W) - dG(c|W)| \text{ a.s.} \tag{39}$$

In light of Lemma 2.1, it follows that there exists an oracle imputation mapping $A \in \mathcal{A}_{F,G}$ such that

$$E_{F,G}|A_n(O) - A(O)|^2 \leq \min(E_{F,G}[R^2_{n,1}], E_{F,G}[R^2_{n,2}]). \tag{40}$$

As $R_{n,1}$ measures how well F_n approximates F and $R_{n,2}$ measures how well G_n approximates G, we have demonstrated that there is an oracle imputation mapping with a small imputation remainder if we can accurately estimate either the full data distribution F or the censoring distribution G.

3.3 Right Censored Data

The full data is given by $X = (W, Y)$, where W are covariates such that $\sigma(Z) \subset \sigma(W)$, and Y is a real-valued response of interest. The regression function of interest is $m : z \to E[Y|Z = z]$. For C a real-valued censoring variable with cumulative distribution function G and survival function $\bar{G} \equiv 1 - G$, we observe

$$O = (W,\ \tilde{Y} = \min(Y, C),\ \Delta = I(Y \leq C)). \tag{41}$$

Coarsening at random is satisfied if

$$\{C \perp Y|W\}. \tag{42}$$

Such data arise when Y represents the time until an event, such as a death or relapse in a medical study. In such studies, C normally represents the length of time that a subject is followed, to see if the event of interest has occurred. When regression is of interest, the response Y is frequently taken to measure the logarithm of a time variable. By (14), the experimental censoring assumption is satisfied if in addition to coarsening at random,

$$Y \leq \tau < \infty \text{ a.s.} \tag{43}$$

$$\bar{G}(\tau|W) \geq \epsilon > 0 \text{ a.s.} \tag{44}$$

Unfortunately, these conditions may not hold in practice if the censoring time has a tendency to be small, such as when a study is only carried out for a fixed period of time, and many survival times exceed the study length. We thus typically have to work with the truncated survival time

$$Y' = YI(Y \leq \tau) + \tau I(Y > \tau) \tag{45}$$

for regression to be possible, or even to accurately estimate the entire conditional distribution. It is our experience that function $z \to E[Y'|Z = z]$ remains interpretable, and that this parameter is a useful simplification of the conditional c.d.f for the survival time Y. It can be shown as in section 3.4 of van der Laan and Robins (2003) that the double robust mapping of the response is given by

$$A_{\mathrm{DR}}(O|F, G) = \frac{Y\Delta}{\bar{G}(\tilde{Y}_-|W)} + \int \frac{E_F[Y|Y > u, W]}{\bar{G}(u_-|W)} dM_G(u) \tag{46}$$

where $M_G(u)$ is the martingale

$$M_G(u) = I(C \leq u, \Delta = 0) - \int_\infty^u I(\tilde{Y} \geq s) \frac{dG(s|W)}{G(s_-|W)}. \tag{47}$$

The nuisance parameters needed to compute the double robust mappings are clearly the conditional distribution function $G(\cdot|W)$ of C on $[-\infty, \tau]$, and the function

$$Q : (u, W) \to E_F[Y|Y > u, W]. \tag{48}$$

With fits G_n and Q_n for G and Q, we can estimate the double robust mapping by

$$A_n(O) = \frac{Y\Delta}{\bar{G}_n(\tilde{Y}_-|W)} + \int \frac{Q_n(u, W)}{\bar{G}_n(u_-|W)} dM_{G_n}(u). \tag{49}$$

With the fit $Q_n : (u, W) \to 0$, we obtain as a special case the estimated inverse probability of censoring weighted mapping

$$A_{n,\mathrm{IPCW}}(O) = \frac{Y\Delta}{\bar{G}_n(\tilde{Y}_-|W)}. \tag{50}$$

If $|Y| \leq \beta_n$ a.s. (note that τ is only an upper bound on Y) so that we force $|Q(u, W)| \leq \beta_n$ a.s., and we force the fit G_n to satisfy the boundedness constraint of (44), it is easy to check that with probability one

$$|A_n(O) - A_{\mathrm{DR}}(O|F, G_n)| \leq R_{n,1} \equiv \epsilon^{-1} \left| \int (Q_n(u, W) - Q(u, W)) dM_{G_n}(u) \right| \tag{51}$$

and

$$|A_n(O) - A_{\mathrm{DR}}(O|F_n, G)| \leq R_{n,2} \equiv \beta_n \epsilon^{-2} |G_n(\tilde{Y}_-|W) - G(\tilde{Y}_-|W)|$$
$$+\beta_n \epsilon^{-2} \left| \int (G_n(u_-|W) - G(u_-|W)) dM_G(u) \right| + \beta_n \epsilon^{-2} \int d|M_{G_n}(u) - M_G(u)|. \tag{52}$$

From Lemma 2.1, it follows that there exists an oracle imputation mapping $A \in \mathcal{A}_{F,G}$ such that

$$E_{F,G}|A_n(O) - A(O)|^2 \leq \min(E_{F,G}[R_{n,1}^2], E_{F,G}[R_{n,2}^2]). \tag{53}$$

As $R_{n,1}$ measures how well F_n approximates F and $R_{n,2}$ measures how well G_n approximates G, we have demonstrated that there is an oracle imputation mapping with a small imputation remainder if we can accurately estimate either the conditional distribution of $\{Y|W\}$ or the conditional distribution $\{C|W\}$. We should note that the conditional distribution has traditionally served as the parameter of interest in nonparametric survival analysis, as discussed Beran (1981) and Dabrowska (1989). This parameter is more general than the regression function. Software such as the HARE function in R based on Kooperberg et al. (1995) can estimate the conditional c.d.f. $F(\cdot|Z)$, through a spline modeling procedure that can be shown to obey the likelihood principle. Knowledge of the censoring mechanism G will not affect the estimate, as we summarized in section 2. This holds even though knowledge of the censoring mechanism would make the inverse probability of censoring weighted imputation mapping given by (50) an *oracle imputation* mapping, because the inverse weights would be known. More realistically, it may be the case that the censoring variable C is completely independent of the full data (W, Y), so that the Kaplan-Meier estimator could be used to accurately estimate the inverse weights in (50). With the doubly robust imputation scheme we can use standard software such as HARE to model both F and G, leading to a reasonable answer if either of the two fits are accurate.

3.4 Current Status Data

The full data is given by $X = (W, Y)$, where W are covariates such that $\sigma(Z) \subset \sigma(W)$, and Y is a real-valued response of interest with conditional survival function $\bar{F}(\cdot|W)$. The regression function is $m : z \to E[Y|Z = z]$. For C a real-valued censoring variable with conditional Lebesgue density $g(\cdot|W)$, we observe $O = (W, C, \Delta = I(Y \leq C))$. Such data structures arise in cross-sectional studies, where Y is a survival time of interest, C is a single monitoring time, and the study records whether the survival time for a subject exceeds the monitoring time. When predicting survival is of interest, Y is often measured on the logarithm of the time scale. Coarsening at random is satisfied if

$$\{C \perp Y|W\} \tag{54}$$

The experimental censoring assumption is satisfied if in addition,

$$a \leq Y \leq b \text{ a.s.} \tag{55}$$

$$\inf\{g(c|W) : c \in [a, b]\} \geq \epsilon > 0 \text{ a.s.} \tag{56}$$

As with the right censored data, these identifiability assumptions may appear too strong to be practically useful. To handle this difficulty, we can consider performing regression on the interval truncated response

$$Y' = aI(Y < a) + YI(a \leq Y \leq b) + bI(Y > b). \tag{57}$$

It is our experience that the regression function $z \to E[Y'|Z = z]$ remains a worthwhile object of study. As can be found in chapter four of van der Laan and Robins (2003) that the double robust mapping of the response is given by

$$A_{\mathrm{DR}}(O|F, G) = \frac{(1 - \Delta) - \bar{F}(C|W)}{g(C|W)} + E_F[Y|W]. \tag{58}$$

The nuisance parameters required to compute the double robust mapping are the conditional density function of $\{C|W\}$ (part of G), the conditional distribution of $\{Y|W\}$ (part of F), and the regression function $Q(W) = E[Y|W]$ (part of F, and a function of the conditional distribution function). The density function g can be estimated with standard software. A crude method for estimating the conditional c.d.f. is based on the observation that $F(y|W) = P(Y \leq y|W) = P(\Delta = 1|C = y, W)$, showing that $F(y|W)$ can be fit with the binary regression of Δ on C and W. Some smoothing may be required to ensure the monotonicity of $y \to F(y|W = w)$. With fits g_n, F_n, and Q_n for the nuisance parameters stated above, we can estimate the double robust mapping with

$$A_n(O) = \frac{(1 - \Delta) - \bar{F}_n(C|W)}{g_n(C|W)} + Q_n(W) \tag{59}$$

With the fit of the conditional distribution function F_n and the regression function Q_n corresponding to Y being a point mass at a, we obtain as a special case the estimated inverse probability of censoring weighted mapping

$$A_{n,\mathrm{IPCW}}(O) = I(a \leq C \leq b)\frac{1 - \Delta}{g_n(C|W)} + a. \tag{60}$$

When the fits g_n, F_n and Q_n are consistent with the bounds of (56), and the distribution G agrees with g, it is simple to check that

$$|A_n(O) - A_{\mathrm{DR}}(O|F, G_n)| \leq R_{n,1} \equiv \epsilon^{-1}|F_n(C|W) - F(C|W)| + |Q_n(W) - Q(W)| \tag{61}$$

and

$$|A_n(O) - A_{\mathrm{DR}}(O|F_n, G)| \leq R_{n,2} \equiv \epsilon^{-2}|g_n(C|W) - g(C|W)|. \tag{62}$$

Because of Lemma 2.1, we have that there exists an oracle imputation mapping $A \in \mathcal{A}_{F,G}$ such that

$$E_{F,G}|A_n(O) - A(O)|^2 \leq \min(E_{F,G}[R_{n,1}^2], E_{F,G}[R_{n,2}^2]). \tag{63}$$

Because $R_{n,1}$ measures how well F_n approximates F and $R_{n,2}$ measures how well the fit g_n approximates the conditional density of g of $\{C|W\}$, we have demonstrated that there is an oracle imputation mapping with a small imputation remainder if we can accurately estimate at least one of the F or G distributions.

3.5 Cumulative Distribution Functions and Binary Regression

For the right censored and current status data structures, we previously mentioned how the imputation method can be used on transformed responses to avoid identifiability problems. We can further utilize transformed responses, if we are genuinely interested in an alternative parameter to the regression function. For instance, if the interest is in estimating the conditional c.d.f. of the response Y at a point t, we can use the imputation method with transformed responses

$$Y' = I(Y \leq t), \tag{64}$$

noting that $E_F[Y'|Z] = P_F(Y \leq t|Z)$. Hence, the imputed covariate/response data could then be plugged into a full data nonparametric regression or binary regression algorithm. Some additional smoothing might be required to ensure the $[0, 1]$ range restraint and the monotonicity of the estimated conditional c.d.f in $y \to P_F(Y' \leq y|Z)$. The imputation technique can also be used with binary regression algorithms for the counterfactual response data structure described in this section, when the counterfactual responses are binary outcomes.

Full data procedures for binary outcomes are attractive because using a general smoother to regress the imputed responses on Z may lead to a regression fit well outside of the $[0, 1]$ range restraint. Unfortunately, one potential problem with imputing responses into standard binary regression functions is that the new responses will not necessarily be $\{0, 1\}$ random variables, so that software written for the full data binary regression problem may consider the input data to be invalid. However, many standard binary regression programs are written to maximize the (possibly penalized) log likelihood

$$\beta \to \sum_{i=1}^{n} [Y_i \log m_\beta(Z_i) + (1 - Y_i) \log(1 - m_\beta(Z_i))], \tag{65}$$

for a parameter β indexing the fits of $m : z \to E[Y|Z = z]$. It still may be possible to carry out whichever maximization technique solved the original binary regression problem, only now solving (65) when $\{Y_i\}_{i=1}^n$ are imputed responses, and not necessarily $\{0, 1\}$ random variables. If β is a Euclidean parameter, it is easy to surmise that the Newton-Raphson algorithm can be implemented as in the original binary regression problem, because the changed Y_i responses enter linearly into the log likelihood as a function of β, so should not complicate the score or Hessian.

4 Efficiency Theory and Estimation of the Double Robust Mapping

We now describe how to consistently estimate the double robust mapping $A(\cdot|F, G)$ defined in (15), based on an efficient estimator of the mean response $\theta(F) \equiv E_F[Y]$. Although not practically useful as we discuss shortly, this imputation technique shows

that we can solve the irregular censored data regression problem by solving a regular censored data problem, to which we can apply results from the extensive semiparametric literature on efficient estimation. See Bickel et al. (1998) for a survey of efficiency theory, as well as rigorous definitions of terms we will use in the sequel such as *regular parametric model*, *score*, *pathwise differentiable*, *regular estimator*, *efficient estimator*, and *efficient influence curve*. Our construction is based on that of Klaassen (1987), and we borrow the notation from this source in the following lemma, when it does not conflict with that we previously introduced. Consider an estimator sequence $\{T_n\}$ for the mean response $\theta = E_F[Y]$. Let $\psi : \mathcal{R} \to \mathcal{R}$ denote a measurable, odd, twice differentiable function with first and second derivatives ψ', ψ'' satisfying

$$|\psi| \leq 1 \tag{66}$$

$$0 < \psi' \leq 1 \tag{67}$$

$$|\psi''| \leq 2. \tag{68}$$

Consider integer sequences $k_n \to \infty$ and $m_n \to \infty$ with $\lim_{n\to\infty} k_n^{-1} m_n = 0$, with $n = m_n^2 + k_n m_n$. Of course, in this section m_n is not the same thing as the regression estimate introduced in previous sections, and we hope this temporary change of notation is not unduly confusing. Relabel the i.i.d. data $D_n = \{O_1, ..., O_n\}$ as $\tilde{O}_{i,j}$, $i = 1, ..., k_n$, $j = 1, ..., m_n$ and $\hat{O}_{i,j}$, $i = 1, ..., m_n$, $j = 1, ..., m_n$. Define the random variables

$$\hat{\gamma}_n(\theta(F)) = m_n^{-1/5} + m_n^{-1} \sum_{i=1}^{m_n} \psi'(\sqrt{n}(t_n(\hat{O}_{i,1}, ..., \hat{O}_{i,m_n}) - \theta(F)))$$

$$J_n^{\psi}(O, \theta(F)) = k_n^{-1} \sum_{i=1}^{k_n} \{m_n^{-1/2} \sum_{j=1}^{m_n} \psi(\sqrt{m_n}(t_n(\tilde{O}_{i,1}, ..., \tilde{O}_{i,j-1}, O, \tilde{O}_{i,j+1}, ..., \tilde{O}_{i,m_n}) - \theta(F)))$$

$$- \sqrt{m_n}\psi(\sqrt{m_n}(t_n(\tilde{O}_{i,1}, ..., \tilde{O}_{i,m_n}) - \theta(F)))\}. \tag{69}$$

Klaassen's influence curve estimator (at the unknown $\theta(F) = E_F[Y]$) is defined by

$$\tilde{J}_n(O, \theta(F)) = J_n^{\psi}(O, \theta(F))\hat{\gamma}_n^{-1}(\theta(F)). \tag{70}$$

Our estimate of the double robust mapping $A(\cdot|F, G)$ is then

$$A_n(\cdot) \equiv \tilde{J}_n(\cdot, T_n) + T_n. \tag{71}$$

When $\{T_n\}$ is an efficient estimator sequence, then our estimator $A_n(\cdot)$ in (71) is often consistent in the $L^2(P_{F,G})$ sense for the double robust mapping $A(\cdot|F, G)$ in (15), as formalized in the following lemma, and proven in the appendix.

Lemma 4.1. *In the setting of this section, suppose that $\{T_n\}$ is a locally regular and asymptotically linear estimator sequence for $\theta(F) \equiv E_F[Y]$ at $P_{F,G} \in \mathcal{M}$, with influence curve*

$$J_{F,G}(O, \theta(F)) \equiv A_{DR}(O|F, G) - \theta(F). \tag{72}$$

Here $A_{DR}(O|F, G) \in L^2(P_{F,G})$ denotes the double robust mapping defined in (15), assumed to be well defined as in (15) because the experimental censoring assumption (12)

holds. That is, $T_n = t_n(O_1, ..., O_n)$ is a measurable map from $\mathcal{O} \times ... \times \mathcal{O}$ to $\mathcal{R}$, such that the regular sequence satisfies

$$T_n - \theta(F) = \frac{1}{n}\sum_{i=1}^{n} A_{DR}(O_i|F,G) - \theta(F) + o_{F,G}(n^{-1/2}). \tag{73}$$

Note: When the full data tangent space is locally saturated, the condition (73) is satisfied if and only $\{T_n\}$ is an efficient estimator for $\theta(F)$, as will be discussed in the sequel.

Assume further that

$$E_{F,G}|T_n - \theta(F)|^2 = o(n^{-7/10}). \tag{74}$$

Then for the estimator $A_n(\cdot)$ defined in (71),

$$\lim_{n\to\infty} E_{F,G}|A_n(O) - A_{DR}(O|F,G)|^2 = 0. \tag{75}$$

□

We now make several notes on the lemma's assumptions. By a locally saturated full data tangent space, we refer to the situation where the tangent space for $F \in \mathcal{M}_{\mathcal{F}}$ (the linear closure in the Hilbert space $L_0^2(F) = \{s(X) : E_F s(X) = 0,\ E_F^2 s^2(X) < \infty\}$ generated by all scores at F of regular parametric submodels of $\mathcal{M}(\mathcal{F})$) is equal to all of $L_0^2(F)$. In this case, the result given on pages 67-68 of Bickel et al. (1998) shows that the efficient influence curve for $\theta(F) = E_F[Y]$ in the full data model is $Y - \theta(F)$. The comments following (2.50) in van der Laan and Robins (2003) then show that the efficient influence curve in the observed data model is equal to $l_{F,G} \circ I_{F,G}^{-1}(Y - \theta(F))$, which is equal to $l_{F,G} \circ I_{F,G}^{-1}(Y) - \theta(F) = A_{\mathrm{DR}}(O|F,G) - \theta(F)$ by the fact that the score and information operators are clearly linear, and map any constant to itself. Because a regular estimator of a pathwise differentiable Euclidean parameter is efficient if and only if it is asymptotically linear with an influence curve equal to the efficient influence curve, the preceding lemma tells us that we can construct a consistent estimator of the desired double robust mapping whenever we can construct an efficient estimator of the mean response $\theta = E_F[Y]$. Note further that $J(O, \theta(F))$ is a legitimate influence curve (has mean zero and finite variance) if $A_{\mathrm{DR}}(O|F,G) \in L^2(P_{F,G})$ because

$$\begin{aligned} E[J(O,\theta(F))] = E[A_{\mathrm{DR}}(O|F,G)] - \theta(F) &= E[E[A_{\mathrm{DR}}(O|F,G)|Z]] - E[Y] \\ &= E[E[Y|Z]] - E[Y] = E[Y] - E[Y] = 0. \end{aligned} \tag{76}$$

Finally, we note asymptotic linearity implies that T_n is converging to $\theta(F)$ at the $\sqrt{n}$ rate, so it is not unreasonable to assume that the squared error rate of convergence behaves like the $O(n^{-1}) = o(n^{-7/10})$ rate of a sample mean in (74).

We caution that the estimator $A_n(\cdot)$ defined in the preceding lemma 4.1 may not be particularly practical, due to the sample splitting required in the Klaassen estimator of the influence curve. In addition, the result only applies when there is an efficient estimator of the mean response $\theta(F) = E_F[Y]$, and it might still be possible to perform

regression if this assumption is violated. Further, estimators for regular parameters typically already depend on estimators of the influence function, such as those estimators built to solve estimating equations. In the situation of the preceding lemma 4.1, this would entail estimating the desired function $A(\cdot|F,G)$ *before* constructing the estimator sequence $\{T_n\}$, which was then used in the lemma to automatically estimate this same imputation mapping. Rather than leading to a practical imputation strategy, we view the lemma as important because it shows that we can approximate the irregular regression function by estimating the same object (the double robust mapping $A_{\mathrm{DR}}(\cdot|F,G)$) as would be needed to efficiently estimate the regular mean response $\theta(F) \equiv E_F[Y]$. This demonstrates that general tools from the semiparametric theory for censored data problems, as discussed in Bickel et al. (1998) and van der Laan and Robins (2003), can apply to more exotic problems such as the nonparametric estimation of a regression function.

5 Upper Bounds for Nonparametric Regression with an Imputed Response

In this section we examine the procedure of imputing censored responses by estimating a function of the observed data that possesses the same conditional mean (given the predictors of interest) as the censored response, and then proceeding to build a nonparametric regression estimator as if there had been no censoring. We consider five commonly implemented varieties of estimators: least squares estimators, complexity regularized least squares estimators, penalized least squares estimators, locally weighted average estimators, and estimators selected with cross-validation. We derive distribution-free inequalities by bounding the expected squared error of these estimators through a full-data bound (which is typically sharp even if there is no censoring) added to an imputation remainder (which measures the quality of the imputation mapping). While this section is motivated by our general imputation methodology based on the doubly robust mapping, the results given here apply to any scheme based on imputing censored responses. We will be interested in bounds of the form

$$\int |m_n(z) - m(z)|^2 \mu(dz) \leq a_1 T_{n,A} + a_2 R_{n,A} \text{ a.s.}, \tag{77}$$

or the weaker bound

$$E|m_n(Z) - m(Z)|^2 = E[\int |m_n(z) - m(z)|^2 \mu(dz)] \leq a_1 E[T_{n,A}] + a_2 E[R_{n,A}], \tag{78}$$

where both bounds hold for scalars a_1 and a_2, and any oracle imputation mapping $A \in \mathcal{A}_{F,G}$ as in (16) (perhaps satisfying certain boundedness constraints). Here integration with respect to μ refers to integration with respect to the distribution of Z. Here $T_{n,A}$ will denote a *full-data bound*. That is, $E[T_{n,A}]$ will typically be a fairly useful upper bound for nonparametric regression based on observations $\{(Z_1, A(O_1)), ..., (Z_n, A(O_n))\}$, which we will be able to bound for *any* $A \in \mathcal{A}_{F,G}$ when the regression function m lies in certain broad function classes. $R_{n,A}$ will denote the imputation remainder, or measure

the distance between the estimated imputation mapping A_n and an oracle imputation mapping $A \in \mathcal{A}_{F,G}$. The imputation remainders we derive are typically based on empirical L^1 or squared L^2 distances between A_n and A.

Our upper bounds have two desirable properties. First, because the bounds will hold for any bounded function $A(\cdot)$ of the observed data in the class $\mathcal{A}_{F,G}$ of functions with the same regression function as the censored response, the infimum can be taken over these functions. This is important for doubly robust imputation schemes, where a sharp full data bound $T_{n,A}$ applies to any $A \in \mathcal{A}_{F,G}$, and the imputation remainder $R_{n,A}$ can be made small if either the F or G part of the likelihood is well approximated. Second, the upper bounds are additive in the full-data bound and the imputation remainder, so if we can argue that the imputation remainder is negligible (as would often occur in censored data problems if something specific were known about the censoring mechanism) then the bounds show that censored nonparametric regression is not considerably harder than uncensored nonparametric regression. The statements given in this section are proven in the appendix, and there are few technical difficulties involved in their derivation, with the needed decompositions and essential details being mostly found in Györfi et al. (2002). Expectations of empirical process terms defined over uncountable function classes should technically be thought of as outer expectations as in the exposition of Chapter 1 from van der Vaart and Wellner (1996), although we ignore this distinction in the following. In the sequel, (X, O, Z) are variables drawn from the same distribution as $\{X_i, O_i, Z_i\}_{i=1}^n$, but assumed independent of these variables.

5.1 An Elementary Bound for Imputed Response Regression

Consider any $m_n^\star : \mathcal{R}^d \to \mathcal{R}$, possibly built from the observed data D_n, the full data $\{X_1, ..., X_n\}$, or unavailable oracle imputations $\{A(O_1), ..., A(O_n)\}$. If $m_n^\star$ maps this potentially unavailable data to a measurable square integrable function of Z, then $(a+b)^2 \leq 2a + 2b$ implies that with probability one

$$\begin{aligned}\int |m_n(z) - m(z)|^2 \mu(dz) &= \int |m_n(z) - m_n^\star(z) + m_n^\star(z) - m(z)|^2 \mu(dz) \\ &\leq 2 \int |m_n^\star(z) - m(z)|^2 \mu(dz) + 2 \int |m_n(z) - m_n^\star(z)|^2 \mu(dz). \end{aligned} \quad (79)$$

This bound is useful when we consider $m_n^\star$ built from the covariates $\{Z_1, ..., Z_n\}$ and unavailable responses $\{A(O_1), ..., A(O_n)\}$ from an oracle imputation mapping $A \in \mathcal{A}_{F,G}$, and we can bound expectation of the first term $2E \int |m_n^\star(z) - m(z)|^2 \mu(dz)$ using results from the (uncensored) nonparametric regression literature. To bound the squared error risk of m_n, we then only have to argue that m_n approximates $m_n^\star$ provided we can control the difference between the estimated imputation mapping A_n and some oracle imputation mapping $A \in \mathcal{A}_{F,G}$.

5.2 Least Squares Estimators

For $\mathcal{F}_n$ a class of measurable functions $f : \mathcal{R}^d \to \mathcal{R}$. The *least squares estimator* is defined by

$$\tilde{m}_n = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(Z_i) - A_n(O_i)|^2 \tag{80}$$

Here we assume the a.s. existence and the measurability of the minimizing function. The truncated least squares estimator is given by

$$m_n \equiv T_{\beta_n} \tilde{m}_n = \text{sgn}(\tilde{m}_n) \min(\tilde{m}_n, \beta_n) \tag{81}$$

Several examples of least squares estimators are as follows.

- Popular types of least squares estimates are basis expansion methods, such as taking $\mathcal{F}_n = \{\sum_{i=1}^{k_n} c_i \phi_i(\cdot) : c_i \in \mathcal{R}, \phi : \mathcal{R}^d \to \mathcal{R}, k_n \in \mathcal{N}, \sum_{i=1}^{k_n} |c_i| \leq \beta_n\}$. Common choices for the ϕ include (tensor products of) polynomial, spline, or wavelets basis. The linearity of these functions in $\{\phi_1, ..., \phi_{k_n}\}$ simplifies estimation of the coefficients $\{c_1, ..., c_{k_n}\}$, where in this case the least squares estimator could be fit using the lasso algorithm.

- More elaborate methods can also be considered least squares estimates, even if the functions are not linear combinations of known bases functions. For example, using the least squares fit with $\mathcal{F}_n = \{\sum_{i=1}^{k_n} c_i \phi(a_i^T \cdot + b_i) + c_0 : k_n \in \mathcal{N}, a_i \in \mathcal{R}^d, b_i, c_i \in \mathcal{R}, \sum_{i=1}^{k_n} |c_i| \leq \beta_n \in \mathcal{R}, \phi(z) = \frac{1}{1+\exp(-z)}\}$ defines a type of *neural network* estimator.

We will give two bounds for the least squares estimator based on the imputation mapping A_n, proven in the appendix. The first is often useful for establishing consistency, while the second is helpful for finding rates of convergence when it is known that the regression function m belongs to a certain function class.

Lemma 5.1. *Let $\tilde{m}_n$ denote the least squares estimator defined in (80) with respect to a function class $\mathcal{F}_n$, and m_n the truncated version as in (81). Then for any $A \in \mathcal{A}_{F,G}$, and imputation remainder $R_{n,A} \equiv 4\beta_n \frac{1}{n} \sum_{i=1}^{n} |A_n(O_i) - A(O_i)|$ we have that with probability one*

$$\begin{aligned} \int |m_n(z) - m(z)|^2 \mu(dz) &\leq T_{n,A} + R_{n,A} \\ &\equiv \inf_{\{f : f \in \mathcal{F}_n, \|f\|_\infty \leq \beta_n\}} \int |f(z) - m(z)|^2 \mu(dz) \\ +2 \sup_{\{f : f \in \mathcal{F}_n, \|f\|_\infty \leq \beta_n\}} &|\frac{1}{n} \sum_{i=1}^{n} |f(Z_i) - A(O_i)|^2 - E|f(Z) - A(O)|^2| \\ &+ R_{n,A} \end{aligned} \tag{82}$$

□

Lemma 5.2. *Let m_n denote the truncated least squares estimator defined in (81) with respect to a function class $\mathcal{F}_n$. Suppose that $|A(O)| \leq \beta_n$ a.s. for some $A \in \mathcal{A}_{F,G}$, and let $R_{n,A}$ denote the imputation remainder*

$$R_{n,A} \equiv 4\beta_n \frac{1}{n}\sum_{i=1}^{n} |A_n(Z_i) - A(O_i)| + \frac{1}{n}\sum_{i=1}^{n} |A_n(O_i) - A(O_i)|^2. \tag{83}$$

Then we have with probability one that

$$\int |m_n(z) - m(z)|^2 \mu(dz) \leq T_{n,A} + 4R_{n,A} \equiv T_{n,1} + 2T_{n,2} + 4R_{n,A} \tag{84}$$

where

$$T_{n,1} \equiv \sup_{\{T_{\beta_n}(f): f \in \mathcal{F}_n\}} E|f(Z)-A(O)|^2 - E|m(Z)-A(O)|^2 - \frac{2}{n}\sum_{i=1}^{n}(|f(Z_i)-A(O_i)|^2 - |m(Z_i)-A(O_i)|^2) \tag{85}$$

and

$$ET_{n,2} \leq \inf_{\{f: f \in \mathcal{F}_n, \|f\|_\infty \leq \beta_n\}} \int |f(z) - m(z)|^2 \mu(dz) \tag{86}$$

If there is a finite VC-dimension of the subgraphs of $\mathcal{F}_n$, denoted by $V_{\mathcal{F}_n^+}$, then the proof of Theorem 11.5 in Györfi et al. (2002) implies

$$ET_{n,1} \leq \frac{c_1}{n} + \frac{c_2 + c + 3\log(n))}{n} V_{\mathcal{F}_n^+} \tag{87}$$

for $c_1 = 5136(1 + \log(42))\beta_n^4$, $c_2 = 10272\beta_n^4(\log(480\beta_n^2) + 1)$, and $c_3 = 10272\beta_n^4$. □

We refer to Györfi et al. (2002) for applications of Lemma 5.1 to proving consistency of least squares estimates based on piecewise polynomials, neural networks, radial basis function networks, data-dependent partitioning estimates, and B-spline estimates. In the same text, Lemma 5.2 is used to control the rate of convergence of the full data bound $T_{n,A}$ for this same group of estimators, when the regression function $m(\cdot)$ belongs to Hölder smoothness classes.

We now contrast this imputation approach to regression with the *unified loss based methodology* of van der Laan and Dudoit (2003). In that work, the full data squared error loss function was replaced with a general loss function

$$L_{\text{FULL}} : \mathcal{F}_n \times \mathcal{X} \to \mathcal{R}, \tag{88}$$

leading to the (full data) empirical risk minimizer

$$m_{n,\text{FULL}} = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^{n} L_{\text{FULL}}(f, X_i). \tag{89}$$

Of course, this could not be implemented with censored data, so the proposal was to somehow map the full data loss function L_{FULL} into an observed data loss function

$$L_{\text{OBSERVED}} : \mathcal{F}_n \times \mathcal{O} \to \mathcal{R} \tag{90}$$

such that $E_F[L_{\text{FULL}}(f,X)] = E_{F,G}[L_{\text{OBSERVED}}(f,O)]$ for each $f \in \mathcal{F}_n$, and then use the observed data empirical risk minimizer

$$m_n = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} L_{\text{OBSERVED}}(f, O_i). \tag{91}$$

van der Laan and Dudoit (2003) considered mapping the full data loss function to the observed data loss function through applying the double robust mapping we did in section 2, that is by composing the score operator with the inverse of the information operator, and applying this mapping to $L(f,X)$ for each $f \in \mathcal{F}_n$. Estimators built in such a fashion have double robustness properties similar to the imputation methods of this paper. Because we have assumed that $\sigma(Z) \subset \sigma(X)$ and $\sigma(Z) \subset \sigma(O)$, it is easy to verify from the linearity of $l_{F,G} \circ I_{F,G}^{-1} : L^2(F) \to L^2(P_{F,G})$ that under the experimental censoring assumption,

$$\begin{aligned}
& l_{F,G} \circ I_{F,G}^{-1}(|Y - f(Z)|^2) \\
& l_{F,G} \circ I_{F,G}^{-1}(f^2(Z)) - 2 l_{F,G} \circ I_{F,G}^{-1}(f(Z)Y) + l_{F,G} \circ I_{F,G}^{-1}(Y^2) \\
& = f^2(Z) - 2 f(Z) l_{F,G} \circ I_{F,G}^{-1}(Y) + l_{F,G} \circ I_{F,G}^{-1}(Y^2) \\
& = f^2(Z) - 2 f(Z) A_{\text{DR}}(O|F,G) + l_{F,G} \circ I_{F,G}^{-1}(Y^2) \\
& = f^2(Z) - 2 f(Z) A_{\text{DR}}(O|F,G) + A_{\text{DR}}(O|F,G)^2 - A_{\text{DR}}(O|F,G)^2 + l_{F,G} \circ I_{F,G}^{-1}(Y^2) \\
& = |f(Z) - A_{\text{DR}}(O|F,G)|^2 + [l_{F,G} \circ I_{F,G}^{-1}(Y^2) - A_{\text{DR}}(O|F,G)^2].
\end{aligned} \tag{92}$$

Because second term $[l_{F,G} \circ I_{F,G}^{-1}(Y^2) - A_{\text{DR}}(O|F,G)^2]$ does not depend on the candidate estimator $f \in \mathcal{F}_n$, it then follows that empirical risk minimization using the double robust mapping of the squared error loss function is equivalent to using the least squares estimator after applying doubly robust imputation to the censored response variable. So in this case, our procedure reduces to a previously introduced technique.

In spite of this reduction, there are many differences between our imputation method and the loss based ideas of van der Laan and Dudoit (2003). While the latter procedure can apply to any type of empirical risk minimization, the imputation method described in this paper is designed specifically for regression. Also, the reduction of one procedure to the other appears specific to using the double robust mapping for the response, as the two techniques do not coincide when using inverse probability of censoring weighted mappings for the imputations or the loss functions. And while the estimators considered by van der Laan and Dudoit (2003) are strictly empirical risk minimizers (possibly using the same mapping of the loss function to select tuning parameters with cross-validation), an advantage of the imputation methodology described here is that it can be applied to any black box nonparametric algorithm, and make use of existing full data software.

5.3 Least Squares Estimators with Complexity Regularization

Let $\mathcal{P}_n$ denote a finite set of parameters. For $p \in \mathcal{P}_n$ let $\mathcal{F}_{n,p}$ denote a set of measurable functions $f : \mathcal{R}^d \to \mathcal{R}$ and let $pen_n(p) \in \mathcal{R}_+$ denote a complexity penalty for $\mathcal{F}_{n,p}$. The

least squares estimator for each function class is defined by

$$\tilde{m}_{n,p} = \arg\min_{f \in \mathcal{F}_{n,p}} \frac{1}{n} \sum_{i=1}^{n} |f(Z_i) - A_n(O_i)|^2 \tag{93}$$

where we assume the a.s. existence and measurability of the minimizers. Instead of choosing the estimator $\tilde{m}_{n,p}$ minimizing the empirical risk, we now penalize the complexity of the function classes over which the least squares estimators are defined. The truncated *least squares estimator with complexity regularization* m_n is defined by setting

$$p^\star = \arg\min_{p \in \mathcal{P}_n} \frac{1}{n} \sum_{i=1}^{n} |\tilde{m}_{n,p}(Z_i) - A_n(O_i)|^2 + pen_n(p) \tag{94}$$

$$\tilde{m}_n = m_{n,p^\star} \tag{95}$$

$$m_n = T_{\beta_n}(\tilde{m}_n) \equiv \mathrm{sgn}(\tilde{m}_n) \min(\tilde{m}_n, \beta_n) \tag{96}$$

The following lemma is useful for separating the risk of the estimator into that of an imputation remainder and a manageable full data bound.

Lemma 5.3. *Let m_n denote the truncated least squares estimator with complexity regularization as given in (96), with respect to a parameter set $\mathcal{P}_n$ and function classes $\mathcal{F}_{n,p}$. For any oracle imputation mapping $A \in \mathcal{A}_{F,G}$, assume that $|A(O)| \leq \beta_n$ a.s. and define the imputation remainder of A_n by*

$$R_{n,A} \equiv 4\beta_n \frac{1}{n} \sum_{i=1}^{n} |A_n(Z_i) - A(O_i)| + \frac{1}{n} \sum_{i=1}^{n} |A_n(O_i) - A(O_i)|^2. \tag{97}$$

Then with probability one

$$\int |m_n(z) - m(z)|^2 \leq T_{n,A} + 4R_{n,A} \equiv T_{n,1} + 2T_{n,2} + 4R_{n,A} \tag{98}$$

where

$$\begin{aligned} T_{n,1} = \sup_{p \in \mathcal{P}_n} \sup_{\{T_{\beta_n}(f): f \in \mathcal{F}_{n,p}\}} \{ & E|f(Z) - A(O)|^2 - E|m(Z) - A(O)|^2 \\ & - \frac{2}{n} \sum_{i=1}^{n} (|f(Z_i) - A(O_i)|^2 - |m(Z_i) - A(O_i)|^2) \\ & - 2pen_n(p)\} \end{aligned}$$

$$ET_{n,2} \leq \inf_{p \in \mathcal{P}_n} \{ \inf_{\{f: f \in \mathcal{F}_{n,p}, \|f\|_\infty \leq \beta_n\}} \int |f(z) - m(z)|^2 \mu(dz) + pen_n(p)\} \tag{99}$$

By Theorem 12.1 in Györfi et al. (2002), if $pen_n(p) \geq \frac{5136\beta_n^4(1+\log(120\beta_n^4 n))V_{F_{n,p}^+} + c_p/2}{n}$ for all $p \in \mathcal{P}_n$ and some $c_p \in \mathcal{R}$ such that $\sum_{p \in \mathcal{P}_n} \exp(-c_p) \leq 1$, where $V_{F_{n,p}^+}$ is the VC-dimension of the subgraphs of $\mathcal{F}_{n,p}$, then we have the bound

$$ET_{n,1} \leq \frac{12840\beta_n^4}{n} \tag{100}$$

□

See Györfi et al. (2002) for examples of where the bounds on $T_{n,1}$ and $T_{n,2}$ in the previous lemma can be used to control rates of convergence for penalized least squares estimators (to be discussed in the following section), orthogonal series estimators, and piecewise polynomial least squares estimators. The advantage of Lemma 5.3 over Lemma 5.2 is that the infimum can be taken over $p \in \mathcal{P}_n$ in the bound on $T_{n,1}$, which can lead to sharper results.

5.4 Penalized Least Squares Estimators

Let J_n denote a penalty functional, mapping all measurable functions $f : \mathcal{R}^d \to \mathcal{R}$ to $[0, \infty] \equiv [0, \infty) \cup \{\infty\}$. Normally J_n is meant to penalize a lack of smoothness, or other measure of function complexity. Let

$$\tilde{m}_n = \arg\min_f \frac{1}{n} |f(Z_i) - A_n(O_i)|^2 + J_n(f) \tag{101}$$

where the minimization is over all measurable functions $f : \mathcal{R}^d \to \mathcal{R}$. We assume the a.s. existence and the measurability of the resulting minimizer. The *penalized least squares estimator* we consider is formed by truncating $\tilde{m}_n$, that is

$$m_n = T_{\beta_n}(\tilde{m}_n) \equiv \text{sgn}(\tilde{m}_n) \min(|\tilde{m}_n|, \beta_n) \tag{102}$$

Several examples of penalized least squares estimators are as follows.

- When Z is univariate ($d = 1$) a popular choice for J_n is $J_n(f) = \lambda_n \int_{\mathcal{R}} |f''(z)|^2 dz$ for $\lambda_n \in \mathcal{R}_+$, with $J_n(f)$ infinite if f does not possess an integrable second derivative. The penalized least squares estimator (without truncation) is then a natural cubic spline with knots at the observed predictors $\{Z_1, ..., Z_n\}$. Although the minimization problem in (101) is infinite-dimensional, it is remarkable that the solution $\tilde{m}_n$ can be easily computed. Typically λ_n is chosen with cross-validation from a held-out portion of the data.

- A more general class of penalties J_n are those corresponding to a *reproducing kernel Hilbert space*. Here a Hilbert space of functions $f : \mathcal{R}^d \to \mathcal{R}$ is defined by the eigen-expansion of a kernel function $K : \mathcal{R}^d \times \mathcal{R}^d \to \mathcal{R}$, and the penalty $J_n(f)$ is a constant λ_n multiplied by the norm of f in this Hilbert space, while $J_n(f)$ is infinite if f is not contained in the Hilbert space. Again, even though the minimization problem in (101) is infinite-dimensional, the solution $\tilde{m}_n$ has the finite dimensional form $\tilde{m}_n(z) = \sum_{i=1}^n c_i K(z, Z_i)$ for $c_1, ..., c_n \in \mathcal{R}$.

We will give two bounds for the penalized least squares estimator based on the imputation mapping A_n. The first is often useful for establishing consistency, while the second is helpful for finding rates of convergence when it is known that the regression function m belongs to a certain smoothness class.

Lemma 5.4. *Consider the penalized least squares estimator $\tilde{m}_n$ and the truncated version m_n as defined in (101) and (102), corresponding to penalty functional J_n as defined*

previously. For any $A \in \mathcal{A}_{F,G}$, suppose that $|A_n|, |A| \leq \beta_n$. Define the imputation remainder $R_{n,A}$ as

$$R_{n,A} \equiv 4\beta_n \frac{1}{n}\sum_{i=1}^{n} |A_n(Z_i) - A(O_i)| + \frac{1}{n}\sum_{i=1}^{n} |A_n(O_i) - A(O_i)|^2 \tag{103}$$

and the function class $\mathcal{F}_n$ as

$$\mathcal{F}_n \equiv \{f : J_n(f) \leq \beta_n^2 + J_n(0)\}. \tag{104}$$

Then with probability one

$$\int |m_n(z) - m(z)|^2 \mu(dz) \leq T_{n,A} + 2R_{n,A} \equiv J_n(m) + T_{n,1} + T_{n,2} + 2R_{n,A} \tag{105}$$

where

$$E[T_{n,1}] = 0 \tag{106}$$

and

$$T_{n,2} = \sup_{\{T_{\beta_n}(f) : f \in \mathcal{F}_n\}} |\frac{1}{n}\sum_{i=1}^{n} |f(Z_i) - A(O_i)|^2 - E|f(Z) - A(O)|^2| \tag{107}$$

□

Lemma 5.5. *Consider the penalized least squares estimator $\tilde{m}_n$ and the truncated version m_n as defined in (101) and (102), corresponding to penalty functional J_n as defined previously. For any $A \in \mathcal{A}_{F,G}$, suppose that $|A_n|, |A| \leq \beta_n$ a.s. Define the imputation remainder $R_{n,A}$ as*

$$R_{n,A} \equiv 4\beta_n \frac{1}{n}\sum_{i=1}^{n} |A_n(Z_i) - A(O_i)| + \frac{1}{n}\sum_{i=1}^{n} |A_n(O_i) - A(O_i)|^2 \tag{108}$$

and the function class $\mathcal{F}_n$ as

$$\mathcal{F}_n \equiv \{f : J_n(f) \leq \beta_n^2 + J_n(0)\}. \tag{109}$$

Then with probability one

$$\int |m_n(z) - m(z)|^2 \mu(dz) \leq T_{n,A} + 4R_{n,A} \equiv 2J_n(m) + T_{n,1} + 4R_{n,A} \tag{110}$$

where

$$T_{n,1} = \sup_{\{T_{\beta_n}(f) : f \in \mathcal{F}_n\}} |E|T_{\beta_n} f(Z) - A(O)|^2 - E|m(Z) - A(O)|^2$$
$$-\frac{2}{n}\sum_{i=1}^{n}(|T_{\beta_n} f(Z_i) - A(O_i)|^2 - |m(Z_i) - A(O_i)|^2) - 2J_n(f)| \tag{111}$$

□

The previous two lemmas should really be thought of as lemma templates, because we have not bounded $E[T_{n,1}]$ in either lemma. Chapters 20-21 of Györfi et al. (2002) show how Lemma 5.4 can be used to establish consistency and Lemma 5.5 can establish rates of convergence for m in Hölder smoothness classes, when the penalty functional corresponds to a scaled integrated squared derivative of the candidate function.

5.5 Locally Weighted Average Estimators

The following estimator may not be very practical because it relies on sample splitting, and we do not know whether similar results can be derived without this device. In a practical setting, we would advise against splitting the data, and instead opt to use locally weighted average estimators built from all n observations for both the imputation and regression steps. To derive the theoretical results, we consider splitting the data into a learning set $\{O_1, ..., O_{n_l}\}$ and a test set $\{O_{n_l+1}, ..., O_{n_l+n_t}\}$ for $n = n_l + n_t$. The imputation mapping is built only from the test set, and is denoted by $A_{n_t}(\cdot) = A_{n_t}(\cdot, O_{n_l+1}, ..., O_{n_l+n_t}) : \mathcal{O} \to \mathcal{R}$. We also define weight functions $W_{n_l,i}(\cdot) = W_{n_l,i}(\cdot, Z_1, ..., Z_{n_l}) : \mathcal{R}^d \to \mathcal{R}$, $i = 1, ..., n_l$ built from the learning set predictors $\{Z_1, ..., Z_{n_l}\}$. The *locally weighted average estimator* is then given by

$$m_n(Z) = \sum_{i=1}^{n_l} W_{n_l,i}(Z) A_{n_t}(O_i) \tag{112}$$

For $A \in \mathcal{A}_{F,G}$, we also consider an oracle locally weighted average estimator $m^\star_{n_l,A}(Z)$ defined by

$$m^\star_{n_l,A}(Z) = \sum_{i=1}^{n_l} W_{n_l,i}(Z) A(O_i) \tag{113}$$

Several examples of locally weighted average estimators are as follows.

- If $\{B_{n,1}, B_{n,2}, ...\}$ is a partition of $\mathcal{R}^d$ and $B_n(z)$ denotes the partition cell containing $z \in \mathcal{R}^d$, then $W_{n,i}(z) = \frac{I(Z_i \in B_n(z))}{\sum_{j=1}^n I(Z_j \in B_n(z))}$ (where by convention $0/0 = 0$) defines a *histogram regression*.
- For $K : \mathcal{R}^d \to \mathcal{R}$ a *kernel function* and $h_n \in \mathcal{R}$ a *bandwidth*, $W_{n,i} = \frac{K((z-Z_i)/h_n)}{\sum_{j=1}^n K((z-Z_i)/h_n)}$ (where by convention $0/0 = 0$) defines a *kernel estimator*.
- For $k_n \in \mathcal{N}$ and $d(\cdot, \cdot)$ a distance function on $\mathcal{R}^d$, the *k_n-nearest neighbor estimator* is defined by $W_{n,i}(z) = \frac{1}{k_n} I(\sum_{j=1}^n I(d(Z_j, z) \le d(Z_i, z)) \le k_n)$.

The following lemma is useful for separating the risk of the estimator into that of an imputation remainder, and the risk of an oracle estimator that has access to the unavailable oracle imputation mapping $A(\cdot) \in \mathcal{A}_{F,G}$.

Lemma 5.6. *For any $A \in \mathcal{A}_{F,G}$ let the regression estimator m_n and the unavailable (based on an the oracle imputation) estimator $m^\star_{n_l,A}$ be as defined in (112) and (113) for weights $W_{n_l,i}$, $i = 1, ..., n_l$, and estimated imputation mapping $A_{n_t}(\cdot)$ as given previously. Define the imputation remainder as $R_{n_t,A} = |A_{n_t}(O) - A(O)|^2$. Assume that*

(i) There is a constant c such that for every nonnegative measurable function $f : \mathcal{R}^d \to \mathcal{R}$ satisfying $Ef(Z) < \infty$ and any $n \in \mathcal{N}$ that $E \sum_{i=1}^n |W_{n,i}(Z)| f(Z_i) \le cEf(Z)$.

(ii) There is a constant D such that $\sum_{i=1}^n |W_{n,i}(Z)| \le D$ with probability one.

Then we have

$$\begin{aligned} E|m_n(Z) - m(Z)|^2 &\leq 2E[T_{n_l,A}] + (2cD)E[R_{n_t,A}] \\ &\equiv 2E|m^\star_{n_l,A}(Z) - m(Z)|^2 + (2cD)E[R_{n_t,A}] \end{aligned} \tag{114}$$

□

See chapters 4,5, and 6 of Györfi et al. (2002) for the weak conditions under which assumptions (i) and (ii) of the lemma hold for histogram regression, kernel, and nearest neighbor estimators. The same chapters also give rates of convergence for the oracle estimator's risk $E[T_{n_l,A}]$ when it is known that the function $m : z \to E[Y|Z = z]$ is a Lipschitz function, and show that these rates can be made the best possible in the minimax sense.

5.6 Estimators Selected with Cross-validation

We divide the data $D_n = \{O_1, ..., O_n\}$ into a learning sample $D_{n_l} = \{O_1, ..., O_{n_l}\}$ of size n_l and a test sample $\{O_{n_l+1}, ..., O_{n_l+n_t}\}$ of size n_t, where $n = n_l + n_t$. Suppose there is a finite set $\mathcal{Q}_n$ such that for each $h \in \mathcal{Q}_n$ there is a regression estimator $m^{(h)}_{n_l}(\cdot) = m^{(h)}_{n_l}(\cdot, D_{n_l}) : \mathcal{R}^d \to \mathcal{R}$ built only from the learning set. Let the *oracle selector* $\hat{h} = \hat{D}_n$ denote the random index $h \in \mathcal{Q}_n$ such that $m^{(h)}_{n_l}$ minimizes the $L^2(\mu)$ distance to the regression function m. That is,

$$\hat{h} = \arg\min_{h \in \mathcal{Q}_n} \int |m^{(h)}_{n_l}(z) - m(z)|^2 \mu(dz). \tag{115}$$

We estimate the unknown $\hat{h}$ with $H = H(D_n)$ minimizing the empirical risk of the candidate estimators over the training sample, with imputed responses. Formally,

$$H = H(D_n) = \arg\min_{h \in \mathcal{Q}_n} \frac{1}{n_t} \sum_{i=n_l+1}^{n} |m^{(h)}_{n_l} - A_n(O_i)|^2. \tag{116}$$

The corresponding estimator selected by cross-validation is then given by

$$m_n = m^{(H)}_{n_l}. \tag{117}$$

The following lemma can be used to separate the risk of the cross-validated estimator into the risk of the estimator $m^{(\hat{h})}_{n_l}$ selected by an oracle, an error term growing only logarithmically with the number of candidate estimators, and an imputation remainder.

Lemma 5.7. *Let m_n be the cross-validation selector as in (117) with $\mathcal{Q}_n, \hat{h}, H$ defined as given previously. For $A \in \mathcal{A}_{F,G}$, let $R_{n,A}$ denote the imputation remainder*

$$R_{n,A} = 4\beta_n \frac{1}{n_t} \sum_{i=n_l+1}^{n} |A_n(O_i) - A(O_i)| + \frac{1}{n_t} \sum_{i=n_l+1}^{n} |A_n(O_i) - A(O_i)|^2 \tag{118}$$

Assume that $\|m_n^{(h)}\|_\infty \leq \beta_n$ *for* $h \in \mathcal{Q}_n$ *(in practice this can be achieved by truncation) and that* $|A_n|, |A| \leq \beta_n$ *a.s. Then for any* $\delta > 0$ *and* $c(\delta) \equiv \beta_n^2(16/\delta + 35 + 19\delta)$*, we have that with probability one*

$$\int |m_n(z) - m(z)|^2 \mu(dz) \leq T_{n,A} + 2(1+\delta)E[R_{n,A}|D_{n_l}] \equiv T_{n,1} + (1+\delta)T_{n,2} + 2(1+\delta)E[R_{n,A}|D_{n_l}] \tag{119}$$

where

$$\begin{aligned} T_{n,2} &= \int |m_{n_l}^{(\hat{h})}(z) - m(z)|^2 \mu(dz) \\ T_{n,1} &= E[|m_{n_l}^{(H)}(Z) - A(O)|^2 | D_{n_l}] - E|m(Z) - A(O)|^2 \\ -(1+\delta)\frac{1}{n_t} &\sum_{i=n_l+1}^{n} (|m_{n_l}^{(H)}(Z_i) - A(O_i)|^2 - |m(Z_i) - A(O_i)|^2) \end{aligned} \tag{120}$$

By (7.6) in Györfi et al. (2002), $E[T_{n,1}|D_{n_l}] \leq c(\delta)\frac{1+\log(|\mathcal{Q}_n|)}{n_t}$ *with probability one.* □

It may appear plausible to obtain such results by simply conditioning on the training data, treating the cross-validation selector as a least squares estimator, and then appealing to Lemmas 5.1 or 5.2. Unfortunately, this does not give bounds as sharp as those appearing the lemma. The additional sharpness is a result of Q_n being a finite set, so that Bernstein's inequality can be utilized. We refer to the appendix for the proof, and chapter seven on Györfi et al. (2002). The same argument given in section 5.2 shows the equivalence of the imputation procedure described here with the estimator selection scheme of van der Laan and Dudoit (2003), where one performs cross-validation after applying the double robust mapping $l_{F,G} \circ I_{F,G}^{-1}(\cdot)$ to the full data squared error loss function.

6 Discussion

We have stated that the benefits of our proposed methodology include double robustness, generality across different censored data structures, and generality to any full data (black box) procedure due to the method's basis in imputation. At first sight, there may also appear to be several potential drawbacks to the proposed procedure. Under severe censoring, the regression function may not be identifiable from the observed data (i.e. unknown tails with right censored data), meaning the statistician must adjust or truncate the response (change the parameter of interest) for regression to be an interesting problem. Another issue is that for the examples presented in this paper, the double robust mapping relies on some type of inverse weighting, potentially leading to outliers among the imputed responses. However, this can usually be dealt with by artificially bounding any inverse weights (denominators) in the estimated mapping away from zero.

The double robust mapping can sometimes appear recursive, in that the regression parameter of interest is itself a nuisance parameter in the mapping, such as the functions denoted by Q in section 3. In these cases, the double robust imputation method can still

improve on poor estimators if G is accurately approximated, and it may be advisable to first use a likelihood-based approach to plug the nuisance parameter estimate into the double robust mapping. The imputation scheme can also be beneficial in such circumstances if the nuisance parameter is the regression function $w \to E[Y|W = w]$, but the interest is in a lower-dimensional regression function $z \to E[Y|Z = z]$ for $\sigma(Z) \subset \sigma(W)$.

Finally, although we presented an automatic estimate of the double robust mapping in (71), we pointed out the estimator was impractical. In fact, we have given no general way to estimate the double robust mapping, although we have determined the necessary nuisance parameters for the examples of section 3 and suggested how they could be estimated. A skeptic might conclude that we have simply replaced the *raw function approximation problem* (estimating the regression function $z \to E[Y|Z = z]$ with censored data) with another function estimation problem (estimating the double robust mapping $O \to A(O|F, G)$), and therefore that we have not presented an explicit procedure. Our counterargument is that double robustness implies that actually approximating $A(O|F, G)$ is unnecessary because the estimate $A_n(O) = A(O|F_n, G_n)$ will suffice if either F_n is close to F or G_n is close to G. Hence, we have replaced a single function estimation problem with an imputation task that lets us solve either of two function estimation problems, potentially without even knowing which one we have solved. Adding to the optimism, it was often a straightforward problem in the examples of section 3 to estimate the nuisance parameters (at least the parameter corresponding to the censoring mechanism G) that was involved in the double robust mapping $A(O|F, G)$, unlike the raw function approximation problem.

Appendix

proof of Lemma (2.1): Note that condition (i) is only needed to imply that $A_{\mathrm{DR}}(O|F_1, G_1) = l^T_{G_1} \circ I^{-1}_{F_1,G_1}(Y)$ is well defined, as in (15). We first prove the lemma for $G = G_1$. We use the facts that $l^T_G(s(O)) = E_G[s(O)|X]$ for $s(O) \in L^2(P_{F_1,G})$ by the definition of the adjoint l^T_G, the definition of the information operator as the adjoint l^T_G composed with the score operator $l_{F_1,G}$, and the fact that the information operator composed with its inverse is of course the identity mapping. That is,

$$E_G[A_{\mathrm{DR}}(O|F_1, G)|X] = l^T_G A(O|F_1, G) = l^T_G \circ l_{F_1,G} \circ I^{-1}_{F_1,G}(Y) = I_{F_1,G} \circ I^{-1}_{F_1,G}(Y) = Y \text{ a.s.} \tag{121}$$

As $\sigma(Z) \subset \sigma(X)$, we conclude that

$$E_{F,G}[A_{\mathrm{DR}}(O|F_1, G)|Z] = E_F[E_G[A_{\mathrm{DR}}(O|F_1, G)|X]|Z] = E_F[Y|Z] = m(Z) \text{ a.s.} \tag{122}$$

We now consider the case of $F = F_1$. For some fixed value z in the support $\mathcal{Z}$ of Z, define inner products on $L^2(F)$ and $L^2(P_{F,G_1})$ by

$$\begin{aligned} < s_1, s_2 >_{\mathcal{X}} &\equiv E_F[s_1(X)s_2(X)|Z = z] \\ < h_1, h_2 >_{\mathcal{O}} &\equiv E_{F,G_1}[h_1(O)h_2(O)|Z = z] \end{aligned} \tag{123}$$

(It is trivial to check that these do indeed define inner products, which endow Hilbert spaces). Then because of the inclusions $\sigma(Z) \subset \sigma(X)$ and $\sigma(Z) \subset \sigma(O)$, conditioning gives that for almost all $z \in \mathcal{Z}$,

$$\begin{aligned} &< h(O), l_{F,G_1}(s) >_{\mathcal{O}} \\ &= E_{F,G_1}[h(O)E[s(X)|O]|Z = z] \\ &= E_F[E_{F,G_1}[h(O)s(X)|O, Z = z]|Z = z] \\ &= E_{F,G_1}[h(O)s(X)|Z = z] \\ &= E_F[E_{G_1}[s(O)h(X)|X, Z = z]|Z = z] \\ &= E_F[s(X)E_{G_1}[h(O)|X, Z = z]|Z = z] \\ &= E_F[s(X)E_{G_1}[h(O)|X]|Z = z] \\ &= E_{F,G_1}[s(X)l^T_{G_1}(h)|Z = z] \\ &=< l^T_{G_1}(h), s(X) >_{\mathcal{X}} \end{aligned} \tag{124}$$

That is, $l^T_{G_1}$ is still the adjoint of the score operator l_{F,G_1} when we take the inner products for the Hilbert spaces $L^2(F)$ and $L^2(P_{F,G_1})$ by first conditioning on the covariate Z. Consequently, we note that if

$$s \in \{s : s \in L^2(P_{F,G_1}) \ , \ E_{G_1}[s(O)|X] = 0 \ P_{F,G_1}\text{-a.s.}\}, \tag{125}$$

then (124) implies that with probability one,

$$\begin{aligned} &E_{F,G_1}[s(O)A_{\mathrm{DR}}(O|F, G_1)|Z = z] \\ &=< s(O), A_{\mathrm{DR}}(O|F, G_1) >_{\mathcal{O}} \\ &=< s(O), l_{F,G_1} \circ I^{-1}_{F,G_1}(Y) >_{\mathcal{O}} \\ &=< l^T_{G_1} \circ s(O), I^{-1}_{F,G_1}(Y) >_{\mathcal{O}} \\ &=< E_{G_1}[s(O)|X], I^{-1}_{F,G_1}(Y) >_{\mathcal{O}} \\ &=< 0, I^{-1}_{F,G_1}(Y) >_{\mathcal{O}} \\ &= 0 \end{aligned} \tag{126}$$

In fact $\frac{dG}{dG_1}(O|X) - 1$ satisfies (125) by (ii) because formula (8) in Gill et al. (1997) show that the Radon-Nikodym derivative $\frac{dG}{dG_1}(O|X)$ can be written as a function of O when both $P_{F,G}$ and P_{F,G_1} both satisfy coarsening at random, so that for F-almost all $x \in \mathcal{X}$,

$$E_{F,G_1}[\frac{dG}{dG_1}(O|X)|X = x] = \int_{\mathcal{O}} \frac{dG}{dG_1}(o|x)dG_1(o|x) = \int_{\mathcal{O}} dG(o|x) = 1. \tag{127}$$

Hence, (126) implies that for F-almost all $z \in \mathcal{Z}$,

$$E_{F,G_1}[(\frac{dG_1}{dG}(O|X) - 1)A_{\mathrm{DR}}(O|F, G_1)|Z = z] = 0. \tag{128}$$

which together with the implication of (122) that $E_{F,G_1}[A(O|F,G_1)] = m(Z)$ a.s. provides the desired result that

$$\begin{aligned} m(Z) &= E_{F,G_1}[A_{\mathrm{DR}}(O|F,G_1)|Z] \\ &= E_{F,G_1}[A_{\mathrm{DR}}(O|F,G_1)\frac{dG}{dG_1}(O|X)|Z] \\ &= \int_{\mathcal{X}}\{\int_{\mathcal{O}} A_{\mathrm{DR}}(o|F,G_1)\frac{dG}{dG_1}(o|x)dG_1(o|x)\}dF(x|Z) \\ = \int_{\mathcal{X}}\{\int_{\mathcal{O}} A_{\mathrm{DR}}(o|F,G_1)dG(o|x)\}dF(x|Z) &= E_{F,G}[A_{\mathrm{DR}}(O|F,G_1)|Z] \text{ a.s.} \end{aligned} \tag{129}$$

□

proof of lemma 4.1: It is clear from the bounds on ψ and its first two derivatives that for any $O \in \mathcal{O}$ we have

$$\begin{aligned} J_n^{\psi}(O,\theta(F)) &\leq 2\sqrt{m_n} \\ |\frac{d}{d\theta}J_n^{\psi}(O,\theta(F))| &\leq \sqrt{m_n} \\ m_n^{-1/5} \leq \hat{\gamma}_n(\theta(F)) &\leq m_n^{-1/5} + 1 \\ |\frac{d}{d\theta}\hat{\gamma}_n(\theta(F))| &\leq 2 \end{aligned} \tag{130}$$

so that consequently

$$\begin{aligned} |\frac{d}{d\theta}\tilde{J}_n(O,\theta(F))| &= |\frac{d}{d\theta}[J_n^{\psi}(O,\theta(F))\hat{\gamma}_n^{-1}(\theta(F))]| \\ &= \frac{|[\frac{d}{d\theta}J_n^{\psi}(O,\theta(F))][\hat{\gamma}_n(\theta(F))] - [J_n^{\psi}(O,\theta(F))][\frac{d}{d\theta}\hat{\gamma}_n(\theta(F))]}{\hat{\gamma}^2(\theta(F))} \\ &\leq \frac{[\sqrt{m_n}][m_n^{-1/5}+1] + [2\sqrt{m_n}][2]}{m_n^{-2/5}} \\ &= O(m_n^{7/10}) \end{aligned} \tag{131}$$

A first-order Taylor expansion thus gives that for any $\theta_n \in \mathcal{R}$ and any $O \in \mathcal{O}$, for some $\theta^\star$ between $\theta(F)$ and θ_n we have

$$|\tilde{J}_n(O,\theta_n) - \tilde{J}_n(O,\theta(F))| = |\theta_n - \theta(F)||\frac{d}{d\theta}\tilde{J}_n(O,\theta^\star)| \leq O(m_n^{7/10})|\theta_n - \theta(F)| \tag{132}$$

Because $n = m_n^2 + k_n m_n \geq m_n^2$ implies that $m_n^{7/10} \leq n^{7/20}$, the preceding formula implies

$$|\tilde{J}_n(O,\theta_n) - \tilde{J}_n(O,\theta(F))| \leq O(n^{7/20})|\theta_n - \theta(F)| \tag{133}$$

Now, Klaassen (1987) proves in formulas (3.18) and (3.19) that under the assumed condition (73) (note that Klaassen's uniform integrability condition (3.4) trivially holds because the influence curve is linear in θ), the estimator $\tilde{J}_n$ given in (70) satisfies

$$E_{F,G}|\tilde{J}_n(O,\theta(F)) + \theta(F) - A_{\mathrm{DR}}(O|F,G)|^2 = E_{F,G}|\tilde{J}_n(O,\theta(F)) - J(O,\theta(F))|^2 = o(1). \tag{134}$$

We thus obtain from the elementary result $(a+b+c)^2 \leq 3a^2+3b^2+3c^2$, (133), and (134) that if $E_{F,G}|T_n - \theta(F)|^2 = o(n^{-7/10})$ as assumed,

$$\begin{aligned}
&E_{F,G}|A_n(O) - A_{\mathrm{DR}}(O|F,G)|^2 \\
&= E_{F,G}|\tilde{J}_n(O,T_n) + T_n - A_{\mathrm{DR}}(O|F,G)|^2 \\
&= E_{F,G}|\tilde{J}_n(O,T_n) - \tilde{J}_n(O,\theta(F)) + T_n - \theta(F) + \tilde{J}_n(O,\theta(F)) + \theta(F) - A_{\mathrm{DR}}(O|F,G)|^2 \\
&\leq 3E_{F,G}|\tilde{J}_n(O,T_n) - \tilde{J}_n(O,\theta(F))|^2 \\
&\quad +3E_{F,G}|T_n - \theta(F)|^2 \\
&\quad +3E_{F,G}|\tilde{J}_n(O,\theta(F)) + \theta(F) - A_{\mathrm{DR}}(O|F,G)|^2 \\
&\leq 3O(n^{7/10})E_{F,G}|T_n - \theta(F)|^2 + 3E_{F,G}|T_n - \theta(F)|^2 + o(1) \\
&= 3O(n^{7/10})o(n^{-7/10}) + o(n^{-7/10}) + o(1) \\
&= o(1)
\end{aligned} \tag{135}$$

leading to the desired result. □

We now recall the following elementary results. It is easy to check that for any square integrable random variable $f(Z)$ and any $A \in \mathcal{A}_{F,G}$, with probability one

$$\begin{aligned}
E|f(Z) - m(Z)|^2 &= \int |f(z) - m(z)|^2 \mu(dz) \\
&= E|f(Z) - A(O)|^2 - E|m(Z) - A(O)|^2,
\end{aligned} \tag{136}$$

and

$$\begin{aligned}
E[|m_n(Z) - m(Z)|^2|D_n] &= \int |m_n(z) - m(z)|^2 \mu(dz) \\
&= E[|m_n(Z) - A(O)|^2|D_n] - E|m(Z) - A(O)|^2.
\end{aligned} \tag{137}$$

proof of Lemma 5.1: By (137),

$$\begin{aligned}
\int |m_n(z)-m(z)|^2\mu(dz) &= E[|m_n(Z)-A(O)|^2|D_n] - E|m(Z)-A(O)|^2 \\
&= E[|m_n(Z)-A(O)|^2|D_n] - \inf_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\le\beta_n\}} E|f(Z)-A(O)|^2 \\
&\quad + \inf_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\le\beta_n\}} E|f(Z)-A(O)|^2 - E|m(Z)-A(O)|^2 \\
&= \sup_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\le\beta_n\}} \{E[|m_n(Z)-A(O)|^2|D_n] - \frac{1}{n}\sum_{i=1}^n |m_n(Z_i)-A(O_i)|^2 \\
&\quad + \frac{1}{n}\sum_{i=1}^n |m_n(Z_i)-A(O_i)|^2 - \frac{1}{n}\sum_{i=1}^n |f(Z_i)-A(O_i)|^2 \\
&\quad + \frac{1}{n}\sum_{i=1}^n |f(Z_i)-A(O_i)|^2 - E|f(Z)-A(O)|^2\} \\
&\quad + \inf_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\le\beta_n\}} E|f(Z)-A(O)|^2 - E|m(Z)-A(O)|^2 \\
&= \sup_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\le\beta_n\}} [T_{n,1}+T_{n,2}(f)+T_{n,3}(f)] + T_{n,4}(f) \qquad (138)
\end{aligned}$$

By (136),

$$\begin{aligned}
T_{n,4} &\equiv \inf_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\le\beta_n\}} E|f(Z)-A(O)|^2 - E|m(Z)-A(O)|^2 \\
&= \inf_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\le\beta_n\}} \int |f(z)-m(z)|^2\mu(dz) \qquad (139)
\end{aligned}$$

Clearly $T_{n,1}$ and $T_{n,3}(f)$ are bounded above by

$$\sup_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\le\beta_n\}} |\frac{1}{n}\sum_{i=1}^n |f(Z_i)-A(O_i)|^2 - E|f(Z)-A(O)|^2| \qquad (140)$$

Hence, it remains to show that if $\|f\|_\infty \le \beta_n$ for $f \in \mathcal{F}_n$ then $T_{n,2}(f) \le R_{n,A} \equiv 4\beta_n\frac{1}{n}|A_n(O_i)-A(O_i)|$. This follows from the definition of m_n as a truncated least

squares estimator because

$$
\begin{aligned}
T_{n,2}(f) &\equiv \frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)-A(O_i)|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A_n(O_i)+A_n(O_i)-A(O_i)|^2 \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)-A_n(O_i)+A_n(O_i)-A(O_i)|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A_n(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)-A_n(O_i)|^2 \\
&\quad +\frac{2}{n}\sum_{i=1}^{n}[m_n(Z_i)-A_n(O_i)-f(Z_i)+A_n(O_i)](A_n(O_i)-A(O_i)) \\
&\quad +\frac{1}{n}\sum_{i=1}^{n}|A_n(O_i)-A(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|A_n(O_i)-A(O_i)|^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n}[m_n(Z_i)-A_n(O_i)-f(Z_i)+A_n(O_i)](A_n(O_i)-A(O_i)) \\
&\leq \frac{2}{n}\sum_{i=1}^{n}[|m_n(Z_i)|+|f(Z_i)|]|A_n(O_i)-A(O_i)| \\
&\leq \frac{4\beta_n}{n}\sum_{i=1}^{n}|A_n(O_i)-A(O_i)| \qquad (141)
\end{aligned}
$$

□

We now note that if

$$\|m_n\|_\infty, |A(O)|, \|A_n\|_\infty, \|f\|_\infty \leq \beta_n \qquad (142)$$

with probability one for functions (A, f), then $(m_n(Z_i)-A_n(O_i))(A_n(O_i)-A(O_i)) \leq |m_n(Z_i)-A_n(O_i)||A_n(O_i)-A(O_i)| \leq (|m_n(Z_i)|+|A_n(O_i)|)|A_n(O_i)-A(O_i)| \leq 2\beta_n|A_n(O_i)-A(O_i)|$ with probability one. Similarly, $(f(Z_i)-A(O_i))(A(O_i)-A_n(O_i)) \leq (|f(Z_i)|+|A(O_i)|)|A_n(O_i)-A(O_i)| \leq 2\beta_n|A_n(O_i)-A(O_i)|$ with probability one. It follows immediately that for the *imputation remainder*

$$R_{n,A} \equiv 4\beta_n\frac{1}{n}\sum_{i=1}^{n}|A_n(Z_i)-A(O_i)|+\frac{1}{n}\sum_{i=1}^{n}|A_n(O_i)-A(O_i)|^2 \qquad (143)$$

then we have that with probability one

$$\frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A(O_i)|^2=\frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A_n(O_i)+A_n(O_i)-A(O_i)|^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A_n(O_i)|^2$$

$$+\frac{2}{n}\sum_{i=1}^{n}(m_n(Z_i)-A_n(O_i))(A_n(O_i)-A(O_i))+\frac{1}{n}\sum_{i=1}^{n}|A_n(O_i)-A(O_i)|^2$$

$$\leq\frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A_n(O_i)|^2+R_{n,A}, \qquad (144)$$

and

$$\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)-A_n(O_i)|^2=\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)-A(O_i)+A(O_i)-A_n(O_i)|^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)-A(O_i)|^2$$

$$+\frac{2}{n}\sum_{i=1}^{n}(f(Z_i)-A(O_i))(A(O_i)-A_n(O_i))+\frac{1}{n}\sum_{i=1}^{n}|A_n(O_i)-A(O_i)|^2$$

$$\leq\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)-A(O_i)|^2+R_{n,A}. \qquad (145)$$

proof of Lemma 5.2: By (137) we make the decomposition

$$\int|m_n(z)-m(z)|^2\mu(dz)=E[|m_n(Z)-A(O)|^2|D_n]-E|m(Z)-A(O)|^2$$

$$=E[|m_n(Z)-A(O)|^2|D_n]-E|m(Z)-A(O)|^2-\frac{2}{n}\sum_{i=1}^{n}(|m_n(Z_i)-A(O_i)|^2-|m(Z_i)-A(O_i)|^2)$$

$$+\frac{2}{n}\sum_{i=1}^{n}|m_n(Z_i)-A(O_i)|^2-\frac{2}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2 \qquad (146)$$

Clearly the first term on the right side of this decomposition is bounded above by $T_{n,1}$, so it remains to bound $\frac{2}{n}\sum_{i=1}^{n}|m_n(Z_i)-A(O_i)|^2-\frac{2}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2$. Using

(144), the definition of m_n as a truncated least squares estimator, and (145) we observe

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}|m_n(Z_i)-A(O_i)|^2-\frac{2}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n}|m_n(Z_i)-A_n(O_i)|^2-\frac{2}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+2R_{n,A} \\
&\leq \inf_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\leq\beta_n\}}\frac{2}{n}\sum_{i=1}^{n}|f(Z_i)-A_n(O_i)|^2-\frac{2}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+2R_{n,A} \\
&\leq \inf_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\leq\beta_n\}}\frac{2}{n}\sum_{i=1}^{n}|f(Z_i)-A(O_i)|^2-\frac{2}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+4R_{n,A} \quad (147)
\end{aligned}$$

Thus, the lemma is proven by taking

$$T_{n,2}\equiv\inf_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\leq\beta_n\}}\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)-A(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2 \quad (148)$$

after noting that (136) implies

$$\begin{aligned}
ET_{n,2}&=E[\inf_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\leq\beta_n\}}\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)-A(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2] \\
&\leq \inf_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\leq\beta_n\}}[\frac{1}{n}\sum_{i=1}^{n}(E|f(Z_i)-A(O_i)|^2-E|m(Z_i)-A(O_i)|^2)] \\
&=\inf_{\{f:f\in\mathcal{F}_n,\|f\|_\infty\leq\beta_n\}}\int|f(z)-m(z)|^2\mu(dz) \quad (149)
\end{aligned}$$

□

proof of Lemma 5.3: By (137) we make the decomposition

$$\int|m_n(z)-m(z)|^2\mu(dz)=T'_{n,1}+2[\frac{1}{n}|m_n(Z_i)-A(O_i)|^2-\frac{1}{n}|m(Z_i)-A(O_i)|^2+pen_n(p^\star)] \quad (150)$$

where

$$\begin{aligned}
T'_{n,1}&=E[|m_n(Z)-A(O)|^2|D_n]-E|m(Z)-A(O)|^2 \\
&-\frac{2}{n}\sum_{i=1}^{n}[|m_n(Z_i)-A(O_i)|^2-|m(Z_i)-A(O_i)|^2]-2pen_n(p^\star) \quad (151)
\end{aligned}$$

Clearly $T'_{n,1}\leq T_{n,1}$, so it remains to bound the second term in our decomposition. By (144), the definition of m_n as a truncated complexity regularized least squares

estimator, and (145) we have that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+pen_n(p^\star)\\
&\leq \frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A_n(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+pen_n(p^\star)+R_{n,A}\\
&\leq \inf_{p\in\mathcal{P}_n}\{\inf_{\{f:f\in\mathcal{F}_{n,p},\|f\|_\infty\leq\beta_n\}}\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)-A_n(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+pen_n(p)\}+R_{n,A}\\
&\leq \inf_{p\in\mathcal{P}_n}\{\inf_{\{f:f\in\mathcal{F}_{n,p},\|f\|_\infty\leq\beta_n\}}\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)-A(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+pen_n(p)\}+2R_{n,A}
\end{aligned}\tag{152}$$

The lemma now follows immediately by letting $T_{n,2}\equiv\inf_{p\in\mathcal{P}_n}\{\inf_{\{f:f\in\mathcal{F}_{n,p},\|f\|_\infty\leq\beta_n\}}\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)-A(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+pen_n(p)\}$ and noting that (136) implies

$$\begin{aligned}
ET_{n,2}&=E[\inf_{p\in\mathcal{P}_n}\{\inf_{\{f:f\in\mathcal{F}_{n,p},\|f\|_\infty\leq\beta_n\}}\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)-A(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+pen_n(p)\}]\\
&\leq \inf_{p\in\mathcal{P}_n}\{\inf_{\{f:f\in\mathcal{F}_{n,p},\|f\|_\infty\leq\beta_n\}}\frac{1}{n}\sum_{i=1}^{n}E|f(Z_i)-A(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}E|m(Z_i)-A(O_i)|^2+pen_n(p)\}\\
&=\inf_{p\in\mathcal{P}_n}\{\inf_{\{f:f\in\mathcal{F}_{n,p},\|f\|_\infty\leq\beta_n\}}\int|f(z)-m(z)|^2\mu(dz)+pen_n(p)\}
\end{aligned}\tag{153}$$

□

The definition of $\tilde{m}_n$ in (101) as a penalized least squares estimator implies that

$$J_n(\tilde{m}_n)\leq\frac{1}{n}\sum_{i=1}^{n}|\tilde{m}_n(Z_i)-A_n(O_i)|^2+J_n(\tilde{m}_n)\leq\frac{1}{n}\sum_{i=1}^{n}|A_n(O_i)|^2+J_n(0)\leq\beta_n^2+J_n(0)\tag{154}$$

For m_n as defined in (102), we conclude that with probability one

$$\tilde{m}_n\in\mathcal{F}_n\tag{155}$$

$$m_n\in T_{\beta_n}\mathcal{F}_n\equiv\{T_{\beta_n}(f):f\in\mathcal{F}_n\}\tag{156}$$

for

$$\mathcal{F}_n\equiv\{f:J_n(f)\leq\beta_n^2+J_n(0)\}.\tag{157}$$

proof of Lemma 5.4: By (136) we can write

$$\begin{aligned}\int |m_n(z)-m(z)|^2\mu(dz) &= E[|m_n(Z)-m(Z)|^2|D_n] - E|m(Z)-A(O)|^2 \\ &= E[|m_n(Z)-m(Z)|^2|D_n] - \frac{1}{n}\sum_{i=1}^n |m_n(Z_i)-A(O_i)|^2 \\ &\quad + \frac{1}{n}\sum_{i=1}^n |m_n(Z_i)-A(O_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(Z_i)-A(O_i)|^2 \\ &\quad + \frac{1}{n}\sum_{i=1}^n |m(Z_i)-A(O_i)|^2 - E|m(Z)-A(O)|^2 \\ &= \frac{1}{n}\sum_{i=1}^n |m(Z_i)-A(O_i)|^2 - E|m(Z)-A(O)|^2 \\ &\quad + E[|m_n(Z)-m(Z)|^2|D_n] - \frac{1}{n}\sum_{i=1}^n |m_n(Z_i)-A(O_i)|^2 \\ &\quad + \frac{1}{n}\sum_{i=1}^n |m_n(Z_i)-A(O_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(Z_i)-A(O_i)|^2 \\ &\equiv T_{n,1} + T'_{n,2} + \frac{1}{n}\sum_{i=1}^n |m_n(Z_i)-A(O_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(Z_i)-A(O_i)|^2 \qquad (158)\end{aligned}$$

Clearly $E[T_{n,1}] = 0$. Because $m_n \in T_{\beta_n}\mathcal{F}_n$ with probability one by (156), it follows that $T'_{n,2} \le T_{n,2}$ a.s., for $T_{n,2}$ as defined in the lemma statement. Hence, it remains to show that with probability one

$$\frac{1}{n}\sum_{i=1}^n |m_n(Z_i)-A(O_i)|^2 - \frac{1}{n}\sum_{i=1}^n |m(Z_i)-A(O_i)|^2 \le J_n(m) + 2R_{n,A} \qquad (159)$$

First note that under our assumption that $|A(O_i)| \le \beta_n$ a.s., the truncation of $\tilde{m}_n$ to m_n can only improve empirical risk. That is,

$$\frac{1}{n}\sum_{i=1}^n |m_n(Z_i)-A(O_i)|^2 \le \frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(Z_i)-A(O_i)|^2 \qquad (160)$$

Using this fact, (144), the definition of $\tilde{m}_n$ as a penalized least squares estimator, the assumed nonnegativity of $J_n(\cdot)$, and (145) we establish (159). Note that (145) can be applied for $f(Z) \equiv m(Z)$ because the bound $|A(O)| \le \beta_n$ a.s. implies that the regression function satisfies $|m(Z)| \le \beta_n$ a.s. Therefore, (159) (and hence the desired

lemma) follows because with probability one

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2\\
&\leq \frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A_n(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+R_{n,A}\\
&\leq \frac{1}{n}\sum_{i=1}^{n}|\tilde{m}_n(Z_i)-A_n(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+R_{n,A}\\
&\leq \frac{1}{n}\sum_{i=1}^{n}|\tilde{m}_n(Z_i)-A_n(O_i)|^2+J_n(\tilde{m}_n)-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+R_{n,A}\\
&\leq \frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A_n(O_i)|^2+J_n(m)-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+R_{n,A}\\
&\leq \sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+J_n(m)-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+2R_{n,A}\\
&= J_n(m)+2R_{n,A} \qquad (161)
\end{aligned}$$

□

proof of Lemma 5.5: By (137) we can form the decomposition

$$\begin{aligned}
\int|m_n(z)-m(z)|^2\mu(dz) &= E[|m_n(Z)-A(O)|^2|D_n]-E|m(Z)-A(O)|^2\\
&= E[|m_n(Z)-A(O)|^2|D_n]-E|m(Z)-A(O)|^2\\
&-\frac{2}{n}\sum_{i=1}^{n}|m_n(Z_i)-A(O_i)|^2+\frac{2}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2-2J_n(\tilde{m}_n)\\
&+\frac{2}{n}\sum_{i=1}^{n}|m_n(Z_i)-A(O_i)|^2-\frac{2}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+2J_n(\tilde{m}_n)\\
&= T'_{n,1}+\frac{2}{n}\sum_{i=1}^{n}|m_n(Z_i)-A(O_i)|^2-\frac{2}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+2J_n(\tilde{m}_n) \qquad (162)
\end{aligned}$$

Because $\tilde{m}_n \in \mathcal{F}_n$ and $m_n \in T_{\beta_n}\mathcal{F}_n$ a.s. by (155) and (156), it immediately follows that $T'_{n,1} \leq T_{n,1}$ a.s. Hence, to prove the lemma it remains to show that with probability one

$$\frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A(O_i)|^2-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+J_n(\tilde{m}_n)\leq J_n(m)+2R_{n,A} \quad (163)$$

Recall from (160) that truncating $\tilde{m}_n$ to m_n can only improve the empirical risk with the imputed responses. We can prove (163) using this fact, (144), the definition of $\tilde{m}_n$ as a penalized least squares estimator, and (145). Note that (145) can be applied

for $f(Z) \equiv m(Z)$ because the bound $|A(O)| \leq \beta_n$ a.s. implies that the regression function satisfies $|m(Z)| \leq \beta_n$ a.s. Therefore, (159) (and hence the desired lemma) follows because with probability one

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A(O_i)|^2+J_n(\tilde{m}_n)-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2 \\
\leq\ &\frac{1}{n}\sum_{i=1}^{n}|m_n(Z_i)-A_n(O_i)|^2+J_n(\tilde{m}_n)-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+R_{n,A} \\
\leq\ &\frac{1}{n}\sum_{i=1}^{n}|\tilde{m}_n(Z_i)-A_n(O_i)|^2+J_n(\tilde{m}_n)-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+R_{n,A} \\
\leq\ &\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A_n(O_i)|^2+J_n(m)-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+R_{n,A} \\
\leq\ &\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+J_n(m)-\frac{1}{n}\sum_{i=1}^{n}|m(Z_i)-A(O_i)|^2+2R_{n,A} \\
=\ &J_n(m)+2R_{n,A} \qquad (164)
\end{aligned}$$

□

proof of Lemma 5.6: From (79) it remains to bound $E|m_n(Z)-m^{\star}_{n_l,A}(Z)|^2$ by $(cD)E|A_{n_t}(O)-A(O)|^2$. By the Cauchy-Schwarz inequality and (ii), with probability one

$$\begin{aligned}
|m_n(Z)-m^{\star}_{n_l,A}(Z)|^2 &= |\sum_{i=1}^{n_l}W_{n_l,i}(Z)(A_{n_t}(O_i)-A(O_i))|^2 \\
&\leq |\sum_{i=1}^{n_l}\sqrt{|W_{n_l,i}(Z)|}\sqrt{|W_{n_l,i}(Z)|}|A_{n_t}(O_i)-A(O_i)||^2 \\
&\leq \sum_{i=1}^{n_l}|W_{n_l,i}|\sum_{i=1}^{n_l}|W_{n_l,i}(Z)||A_{n_t}(O_i)-A(O_i)|^2 \\
&\leq D\sum_{i=1}^{n_l}|W_{n_l,i}(Z)||A_{n_t}(O_i)-A(O_i)|^2 \qquad (165)
\end{aligned}$$

Now, using that $(A(O), A_{n_t})$ is independent of $\{Z_1, ..., Z_{n_l}, Z\}$, we see that $E[|A_{n_t}(O_i)-A(O_i)|^2|Z_1, ..., Z_{n_l}, Z] = E[|A_{n_t}(O_i)-A(O_i)|^2|Z_i] \equiv f(Z_i)$ for $i = 1, ..., n_l$. Using this

fact, (i), and the preceding inequality

$$E|m_n(Z) - m^{\star}_{n_l,A}(Z)|^2 \leq DE \sum_{i=1}^{n} |W_{n_l,i}(Z)||A_{n_t}(O_i) - A(O_i)|^2$$
$$= DE \sum_{i=1}^{n} E[|W_{n_l,i}(Z)||A_{n_t}(O_i) - A(O_i)|^2||Z_1, ..., Z_{n_l}, Z]$$
$$= DE \sum_{i=1}^{n} |W_{n_l,i}(Z)|E[|A_{n_t}(O_i) - A(O_i)|^2||Z_1, ..., Z_{n_l}, Z]$$
$$= (cD)E \sum_{i=1}^{n} |W_{n_l,i}(Z)|f(Z_i)$$
$$\leq (cD)Ef(Z) = (cD)E[E[|A_{n_t}(O) - A(O)|^2|Z]] = (cD)E|A_{n_t}(O) - A(O)|^2 \quad (166)$$

From our previous comments, this implies the desired result. □

proof of Lemma 5.7: By (137) and the fact that m_n is built only from D_{n_l}, we use the decomposition

$$\int |m_n(z) - m(z)|^2 \mu(dz) = E[|m_n(Z) - A(O)|^2|D_n] - E|m(Z) - A(O)|^2$$
$$= E[|m_n(Z) - A(O)|^2|D_{n_l}] - E|m(Z) - A(O)|^2$$
$$= E[|m^{(H)}_{n_l}(Z) - A(O)|^2|D_{n_l}] - E|m(Z) - A(O)|^2$$
$$-(1+\delta)\frac{1}{n_t} \sum_{i=n_l+1}^{n} (|m^{(H)}_{n_l}(Z_i) - A(O_i)|^2 - |m(Z_i) - A(O_i)|^2)$$
$$+(1+\delta)\frac{1}{n_t} \sum_{i=n_l+1}^{n} (|m^{(H)}_{n_l}(Z_i) - A(O_i)|^2 - |m(Z_i) - A(O_i)|^2)$$
$$= T_{n,1} + (1+\delta)\frac{1}{n_t} \sum_{i=n_l+1}^{n} (|m^{(H)}_{n_l}(Z_i) - A(O_i)|^2 - |m(Z_i) - A(O_i)|^2) \quad (167)$$

Thus, it remains to show that with probability one

$$E[\frac{1}{n_t} \sum_{i=n_l+1}^{n} (|m^{(H)}_{n_l}(Z_i) - A(O_i)|^2 - |m(Z_i) - A(O_i)|^2)|D_{n_l}]$$
$$\leq \int |m^{(\hat{h})}_{n_l}(z) - m(z)|^2 \mu(dz) + 2E[R_{n,A}|D_{n,l}] \quad (168)$$

This follows immediately from (144), the definition of $m^{(H)}_{n_l}$ as a test sample empirical

risk minimizer, and (145) because

$$\begin{aligned}
&\frac{1}{n_t}\sum_{i=n_l+1}^{n}|m_{n_l}^{(H)}(Z_i)-A(O_i)|^2-\frac{1}{n_t}\sum_{i=n_l+1}^{n}|m(Z_i)-A(O_i)|^2\\
&\leq\frac{1}{n_t}\sum_{i=n_l+1}^{n}|m_{n_l}^{(H)}(Z_i)-A_n(O_i)|^2-\frac{1}{n_t}\sum_{i=n_l+1}^{n}|m(Z_i)-A(O_i)|^2+R_{n,A}\\
&\leq\frac{1}{n_t}\sum_{i=n_l+1}^{n}|m_{n_l}^{(\hat{h})}(Z_i)-A_n(O_i)|^2-\frac{1}{n_t}\sum_{i=n_l+1}^{n}|m(Z_i)-A(O_i)|^2+R_{n,A}\\
&\leq\frac{1}{n_t}\sum_{i=n_l+1}^{n}|m_{n_l}^{(\hat{h})}(Z_i)-A(O_i)|^2-\frac{1}{n_t}\sum_{i=n_l+1}^{n}|m(Z_i)-A(O_i)|^2+2R_{n,A}. \qquad (169)
\end{aligned}$$

Taking the conditional expectation given D_{n_l} now clearly implies (168) by (137), and thus the desired result. □

Acknowledgements

Mark van der Laan receieved support from NIH grant NIH R01 GM07 1397.

References

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical Report, University of California, Berkeley.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J.A (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.

Dabrowska, D.M. (1989). Uniform consistency of the kernel conditional Kapaln-Meier estimate. *Annals of Statistics*. 17:1157-1167.

Gill, R.D., van der Laan, M.J., and Robins, J.M. (1997). Coarsening at random: characterizations, conjectures and counter-examples. *Proceedings of the First Seattle Symposium in Biostatistics*, 1995. D.Y. Lin and T.R Fleming (editors), Springer Lecture Notes in Statistics, 255-294. math.uu.nl/people/gill/Preprints/car0.pdf

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York.

Hastie, T., Tibshirani, R.J., and Friedman, J.H. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and coarse data. *Annals of Statistics*. 19:2244-2253.

Klaassen, C.A.J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimates. *Annals of Statistics*. 15:1548-1562.

Kooperberg C., Stone, C.J., and Troung Y.K. (1995). Hazard regression. *Journal of the American Statistical Association*. 90:78-74.

Robins, J.M. and Ritov, Y. (1997). Towards a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*. 16:285-319. biostat.harvard.edu/ robins/coda.pdf

Rotnitzky A. and Robins J.M. (2003). Inverse probability weighted estimation in survival analysis. *Encyclopedia of Biostatistics*. biostat.harvard.edu/∼robins/publications/IPW-survival-encyclopedia-submitted-corrected.pdf

van der Laan, M.J. (1998). Identity for NPMLE in censored data models. *Lifetime Data Models*. 4:83-102.

van der Laan, M.J. and Dudoit, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 130. bepress.com/ucbbiostat/paper130

van der Laan, M.J. and Robins, J.M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York.