IDENTIFICATION AND ESTIMATION OF NONPARAMETRIC STRUCTURAL MODELS BY INSTRUMENTAL VARIABLES METHOD[*]

Woocheol Kim[†]
Korea Institute of Public Finance
and
Institute of Statistics and Econometrics,
Humboldt University zu Berlin

March, 2004 (First draft: November 2003)

Abstract

This paper concerns a new statistical approach to instrumental variables (IV) method for nonparametric structural models with additive errors. A general identifying condition of the model is proposed, based on richness of the space generated by marginal discretizations of joint density functions. For consistent estimation, we develop statistical regularization theory to solve a random Fredholm integral equation of the first kind. A minimal set of conditions are given for consistency of a general regularization method. Using an abstract smoothness condition, we derive some optimal bounds, given the accuracies of preliminary estimates, and show the convergence rates of various regularization methods, including (the ordinary/iterated/generalized) Tikhonov and Showalter's methods. An application of the general regularization theory is discussed with a focus on a kernel smoothing method. We show an exact closed form, as well

[*] I would like to thank W. Härdle, O. Linton, W. Newey, P. Phillips, and V. Spokoiny for helpful discussions and comments.

[†] Address correspondence to: Woocheol Kim, Korea Institute of Public Finance, 79-6 Garak-Dong, Songpa-Gu, Seoul, Republic of Korea, 138-774. E-mail address: wkim@kipf.re.kr

as the optimal convergence rate, of the kernel IV estimates of various regularization methods. The finite sample properties of the estimates are investigated via a small-scale Monte Carlo experiment.

1 Introduction

In econometric models, explanatory variables are often presumed to be endogenous (i.e., correlated with error terms), when their relation to dependent variables represents optimizing behaviors of individuals or market equilibrium. The relations thereof are called 'structural' to be differentiated from a reduced form that comes out of a pure statistical underpinning. The literature abounds in studies of various structural models with different sources of endogeneity, including linear simultaneous equations, measurement errors, heterogenous treatment effects, random effects in panel data, and sample selection, etc. Common to the previous studies, however, is a restrictive assumption that the true structural relation is known *a priori* up to some parametric class. When misspecification is one's main concern, nonparametric methods can be a useful alternative, whose development in a structural setup is of only recent interest. This paper, specializing in structural models with additive errors, contributes to nonparametric instrumental variables method, by providing new results for both identification and estimation. Suppose that the random variables (Y_i, X_i) are generated by the following regression models, with X_i including some endogenous variables, say, Z_i;

$$Y_i = m(X_i) + \varepsilon_i, \tag{1}$$

where ε_i is $\text{iid}(0, \sigma^2)$ and $m(\cdot)$ is an unknown function. Due to endogeneity in X, the structural function $m(\cdot)$ needs to be identified by postulating a set of instrumental variables (IV) W that satisfy certain stochastic restrictions w.r.t. the errors ε. W is allowed to have common elements with X;

$$X = (Z, W_1), \text{ and } W = (W_1, W_2). \tag{2}$$

With infinite dimensional parameters to be identified, the instrumental variables are required to satisfy stronger restrictions than for parametric models. For example, Roehrig (1988) assumed that W is independent of ε, to identify $m(\cdot)$. Alternatively, one may prefer to assume weaker restrictions on W in form of conditional moments so that the models can afford more general features such as conditional heteroscedasticity. Along this line, two different methods have recently been considered. One is the *instrumental variable* method of Newey and Powell (1988, 2002) and Darolles, Florens, and Renault (2001) that use a restriction

$$E(\varepsilon|W) = 0, \tag{3}$$

and the other is the *control function* method by Newey, Powell and Vella (1999) that assume $E(\varepsilon|X, \eta) = E(\varepsilon|\eta)$, where $\eta = X - E(X|W)$. Although either restriction does not imply the other, the moment condition (3) is the one that is more familiar and easily interpretable. Including the standard regression as a special case (for $W = X$), (1) through (3) give rise to nonparametric generalization of various structural models analyzed by the IV methods of Sargan (1958) and Amemiya (1974) as well as the GMM of Hansen (1982); henceforth the name comes.[1] Simplicity of the models, however, comes only at a cost of nonstandard identification and estimation. Since the seminal work by Newey and Powell (1988), it has been well noted that, under (3), $m(\cdot)$ is characterized only implicitly via an integral equation. For identification of m, it is crucial to know the algebraic properties of an integral operator which is defined on an infinite dimensional functional space. For estimation, we need to solve a random Fredholm integral equation of the first kind. Since there may exist *no* or *more than one* solutions to the random integral equation, a natural estimator can be defined by using a generalized inverse. However, as will be shown later in section 3.1, such estimator is not consistent in general. Inconsistency of such naive estimator, called as ill-posedness of inverse problems, is related to discontinuity of the underlying mapping from a reduced-form to a structural function. For these reasons, adequate statistical theory has not as yet been fully developed for an IV estimator of the nonparametric structural model in (1) through (3). Only a consistency result was shown for a general case of common elements by Newey and Powell (1988, 2002), who also suggested a primitive condition for identification under a parametric assumption of exponential family. For a special case of disjoint X and W, Darolles, Florens, and Renault (2001) made some important improvements, succeeding in deriving a lower bound on the convergence rate of their estimates.[2]

This paper, trying to improve upon the previous works, provides a more general approach to identification and estimation of the structural model in (1) through (3). We give a new identification result that does not rely on any parametric assumption. The suggested identifying condition is closely related to richness of the linear space that are generated by marginal discretizations of the

[1] As an alternative generalization of the linear 2SLS, the control function approach treats endogeniety of X as an omitted variable problem, and corrects endogeneity bias by inclusion of some 'control' variables (η in the above), as in Heckman(1979)'s two-step estimator for selcetivity bias. See Blundell and Powell (2001) for a detailed comparison of two methods.

[2] In a more recent work, Hall and Horowitz (2003) tried to provide deeper results on the convergence rates for nonparametric IV estimation.

joint density function. Under continuity of the density function, the condition is also shown to be necessary for identification. For consistent estimation, a general theory of statistical regularization is developed to find a stable solution to a random Fredholm integral equation of the first kind. In contrast to the ad-hoc approaches in the previous works, we give more systematic analysis of regularization to resolve the ill-posedness of statistical inverse problems. For example, applying random operator theory, we show a minimal set of conditions under which a large class of regularized estimators are consistent. Also, the optimal bounds of the convergence rates, given the accuracies of the preliminary estimates, are derived, using a notion of the modulus of stochastic equicontinuity of a random operator. For comparison of asymptotic properties of various regularization, we calculate the convergence rates of the ordinary/iterated/generalized Tikhonov and Showalter's methods. According to our results, Showalter's method can attain the optimal bounds in a general case, while three types of Tikhonov methods are suboptimal in some cases. A specific example of kernel IV estimates is considered to illustrate how the general theory can be applied in practice. Unlike the previous works, we show an exact closed form of the regularized kernel estimates explicit in a regularization parameter. Computations of the regularized estimates only require standard finite-dimensional matrix operations. The convergence rates of those estimates are derived, based on the general theory of statistical regularization.

There are many works on nonparametric estimation of other structural models. An extensive list can be found in a recent survey by Blundell and Powell (2001a) and the references therein. Some of them, among others, are Altonji and Matzkin (2001), Imbens and Newey (2001), and Chesher (2002) that develop nonparametric methods for nonseparable structural models. Ai and Chen (2001), who consider semiparametric GMM estimation of structural models, show $\sqrt{n}$-consistency of parametric terms as well as the semiparametric asymptotic efficiency. In Blundell and Powell (2001b), a control function approach is used for a semiparametric binary response model with endogenous variables. In statistics literature, there are some earlier works on ill-posed inverse problems, such as deconvolution (Fan, 1991) and noisy integral equations (Nychka and Cox, 1989); see the survey by O'Sullivan (1986) and van Rooij and Ruymgaart (1999), for more results. Those works, however, are different from our approach in that they assume a known integral operator. For nonparametric estimation of additive models, Mammen, Linton, and Nielsen (1999) and Linton and Mammen (2003) work with random integral equations, but their inverse problems are well-posed.

The rest of paper proceeds as follows. Section 2 concerns an identification issue, and suggest a general identification condition. In section 3, we first show ill-posedness of the IV estimation problem, and develop general theory of statistical regularization for consistent estimation, including a discussion for optimal bounds. Section 4 is devoted to derivation of the convergence rates for various regularization methods. Section 5 applies the general results to a specific example of kernel IV estimation. Both closed forms and asymptotic properties of the estimates are shown. The finite sample properties of the estimates are investigated via a small-scale Monte Carlo experiment. All the technical proofs for the theorems are collected in the appendices.

Notations: *w.p.1 (*or *a.s.)* stands for 'with probability one', and *w.p.a.1*, for 'with probability approaching to one'.

2 Identification

Throughout the paper, we assume that the sample observations $\{(Y_i, Z_i, W_i)$: $i = 1, 2, .., n\}$ are randomly drawn out from a distribution $F_{Y,Z,W}(Y, Z, W)$ defined on $\mathcal{Y} \times \mathcal{Z} \times \mathcal{W}$ $(\subset \mathbb{R} \times \mathbb{R}^{d_z} \times \mathbb{R}^{d_2})$, where $\mathcal{W} = \mathcal{W}_1 \times \mathcal{W}_2 \subset \mathbb{R}^{d_{w_1}} \times \mathbb{R}^{d_{w_2}}$ and $d_2 = d_{w_1} + d_{w_2}$. The support of X is given by $\mathcal{X} \equiv \mathcal{Z} \times \mathcal{W}_1$ $\subset \mathbb{R}^{d_1}$, where $d_1 = d_z + d_{w_1}$. $F_{Y,Z,W}(\cdot)$ is assumed to be absolutely continuous with density $f_{Y,Z,W}(\cdot)$. The joint density function for (Y, W) and (Z, W) is denoted by $f_{Y,W}(\cdot)$ and $f_{Z,W}(\cdot)$, respectively. Let $L^2(\mathcal{X})$ be the infinite-dimensional Hilbert space of square-integrable functions defined on $\mathcal{X}$, with norm given by $||m(\cdot)||^2_{L^2(\mathcal{X})} = \int_{\mathcal{X}} m^2(x)dx$. Below, we give a precise definition for identification of the model. When (3) holds for $m(\cdot)$ and $\widetilde{m}(\cdot)$, the two functions are called 'observationally equivalent'.

Definition 2.1 The structural function $m(\cdot)$ in (1) is identified in $L^2(\mathcal{X})$ by instrumental variables W, if and only if (3) holds for $m(\cdot) \in L^2(\mathcal{X})$, and any observationally equivalent functions, $m(\cdot)$ and $\widetilde{m}(\cdot)$, are identical in the sense that $m(X) = \widetilde{m}(X)$, *w.p.1.*

Given $F_{Y,Z,W}(\cdot)$ in a class of distributions $\mathcal{F}$, we define

$$h_F(w) = \int_{\mathcal{Y}} y f_{Y,W}(y, w)dy,$$

and an integral operator by

$$T_F : L^2(\mathcal{X}) \to L^2(\mathcal{W}), \quad \text{with } (T_F m)(w) = \int_{\mathcal{Z}} m(z, w_1) f_{Z,W}(z, w)dz,$$

where the subscript of T and h means that they are defined by the underlying distribution F. The subscript will be omitted unless confusion arises. For a linear operator $T : L^2(\mathcal{X}) \to L^2(\mathcal{W})$, its operator norm is defined by $||T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})} \equiv \sup_{m\in L^2(\mathcal{X}),(m\neq 0)} ||Tm||_{L^2(\mathcal{W})}/||m||_{L^2(\mathcal{X})}$. Throughout the paper, we assume that the joint density is square-integrable so that the linear operator T is bounded in the sense of $||T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})} < \infty$. $\mathcal{N}(T)$ and $\mathcal{R}(T)$ denotes the null space and the range, respectively, of T. Since two conditions in Def.2.1 are equivalent to existence of a unique solution to '$T_F(m)(w) = h_F(w)$', we get the following result.

Proposition 2.1 Let $\mathcal{F}^*$ be such that

$$\mathcal{F}^* = \{F_{Y,Z,W}(\cdot) \in \mathcal{F} : \text{(a) } \mathcal{N}(T_F) = \{0\} \text{ and (b) } h_F \in \mathcal{R}(T_F)\}. \tag{4}$$

Given a model (1)-(3) with a true distribution $F^0_{Y,Z,W}(\cdot)$, $m(\cdot)$ is identified to be $T_F^{-1}(h_F)$ in $L^2(\mathcal{X})$ by the instrumental variables W, if and only if $F^0_{Y,Z,W}(\cdot) \in \mathcal{F}^*$.

Proposition 2.1 makes clear that some distributional assumptions are needed for valid identification. In the previous works, most considerations are given to the uniqueness condition, with less known about existence. For example, as Newey and Powell (2002, p3) note, the injectivity condition in (4)-(a) is equivalent to statistical completeness of $F_{Z|W}(\cdot|\cdot)$ in the 'parameter' W. Using a parametric distributional assumption that $F^0_{Z,W}$ is in a class of exponential family, they derived some primitive condition for identifiability. A more flexible nonparametric approach was made by Darolles, Florens, and Renault (2001), under an assumption that there is no common element between X and W. Both conditions of uniqueness and existence are discussed in detail, based on singular-values expansion of a compact operator T. It is shown that an equivalent characterization of (4)-(a) can be given in terms of nonlinear canonical correlations of X and W, and the existence condition in (4)-(b) translates into imposing some smoothness on a reduced form function.[3] Unlike Newey and Powell (2002), their results, however, are delimited by a strong restriction that prevents an element of explanatory variables being used as an instrument. One more comment deserves note, concerning the scope of generality of the conditions in (4). Obviously, certain distributional assumptions on (Z, W) will suffice for identifiability, i.e., for T to be one-to-one. Existence, however, cannot be ensured in a similar manner, since an operator acting on an infinite-dimensional space does not necessarily possess a closed range, i.e., T is not onto in general. It is well known in functional analysis (Kress, 1989, p20) that, given a compact operator T, $\mathcal{R}(T)$ is closed if and only if $\dim(\mathcal{R}(T)) < \infty$. Therefore, what one can expect as most favorable to (4)-(b) will be that T has a dense range in $L^2(\mathcal{W})$. In the rest of this section, we provide an alternative but more general identification result, showing when T is one-to-one or has a dense range. Noncompactness of T is allowed.

[3] That is, the generalized Fourier coefficients of h (w.r.t. the singular functions) decay fast enough relative to the singular values.

For convenience of exposition, we first consider a case with disjoint X and W, and then give an extension to a common-element case. Given $\{\omega_l\}_{l=1}^L \subset \mathcal{W}$, we define a marginal discretization (w.r.t. W) of a joint density function, by

$$f_L^{\omega}(x) = [f_{X,W}(x,\omega_1), .., f_{X,W}(x,\omega_L)]',$$

and let $\text{lin}(\{f_L(\cdot,\omega_l)\}_{l=1}^L)$ be the linear space generated by $\{f_L(\cdot,\omega_l)\}_{l=1}^L$. For a sequence $\{\omega_l\}_{l=1}^\infty \subset \mathcal{W}$, let $\overline{\text{lin}}\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^\infty$ be the closure of $\text{lin}\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^\infty$ in $L^2(\mathcal{X})$, and $[\text{lin}\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^\infty]^\perp$ the orthogonal compliment of $\text{lin}\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^\infty$, i.e.,

$$L^2(\mathcal{X}) = \overline{\text{lin}}\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^\infty \oplus [\text{lin}\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^\infty]^\perp.$$

Our identification results make use of the following conditions.

C.2.1 For a sequence $\overline{\mathcal{W}} = \{\omega_l\}_{l=1}^\infty \subset \mathcal{W}$, $\text{lin}\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^\infty$ is dense in $L^2(\mathcal{X})$; i.e.,

$$\overline{\text{lin}}\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^\infty = L^2(\mathcal{X}).$$

C.2.2 For a sequence $\overline{\mathcal{X}} = \{\varkappa_l\}_{l=1}^\infty \subset \mathcal{X}$, $\text{lin}\{f_{X,W}(\varkappa_l,\cdot)\}_{l=1}^\infty$ is dense in $L^2(\mathcal{W})$; i.e.,

$$\overline{\text{lin}}\{f_{X,W}(\varkappa_l,\cdot)\}_{l=1}^\infty = L^2(\mathcal{W}).$$

Both conditions address richness of the linear spaces that are generated by marginal discretizations of the joint density function. C.2.1 will hold, if a complete orthogonal basis of $L^2(\mathcal{X})$ is generated by linear combinations of $\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^\infty$. We below show sufficiency of C.2.1 and C.2.2 for T to be one-to-one and have a dense range, respectively.

Theorem 2.2 Suppose that a structural model is given by (1)-(3) with W_1 empty.

(i) If C.2.1 holds, then, the integral operator $T : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ is one-to-one, i.e., $m(\cdot)$ is identifiable in $L^2(\mathcal{X})$.

(ii) If C.2.2 holds, then, T has a dense range in $L^2(\mathcal{W})$; i.e., at least one $m(\cdot) \in L^2(\mathcal{X})$ satisfies the given structural model, for $h(\cdot)$ in some 'dense' subspace.

Symmetry of two conditions in Theorem 2.2 can be explained easily by introducing an adjoint operator. If $G : M \to H$ is a bounded linear operator from a Hilbert space M to a Hilbert space

H, the adjoint of G is the operator $G^* : H \to M$ satisfying $< Gm, h >_M = < m, G^*h >_H$, for all $m \in M$ and $h \in H$, where $< \cdot, \cdot >_M$is the inner product of M. In the present case, the adjoint of $T : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ is $T^* : L^2(\mathcal{W}) \to L^2(\mathcal{X})$ with $(T^*h)(x) = \int f_{X,W}(x,w)h(w)dw$. From the relation '$\overline{\mathcal{R}}(T) = \mathcal{N}^{\perp}(T^*)$', T has a dense range, if and only if T^* is one-to-one. In consequence, the second assertion of Theorem 2.2 follows as a mirror image, once the first is true. The suggested identifying condition seems rather abstract, partly because we do not use any parametric assumptions. Roughly speaking, identifiability depends on the way that the density function of X, conditional on $W = \omega_l$, varies over different values of ω_l's.[4] For example, the model is identifiable, if some sequence of the conditional density functions, $\{f_{X|W}(\cdot|\omega_l)\}_{l=1}^{\infty}$, includes (or spans) a complete basis of $L^2(\mathcal{X})$. Although it is not easy to check the condition in a practical case, it allows for a useful finite-dimensional approximation of the underlying structural function. Let $P_{f_L^\omega}$ be the orthogonal projection onto $\text{lin}(\{f_{X,W}(x,\omega_l)\}_{l=1}^{L})$;

$$(P_{f_L^\omega} m)(x) = f_L^\omega(x)' Q_\omega^{*\dagger} \int_{\mathcal{X}} f_L^\omega(z) m(z) dz,$$

where $Q_\omega^* = \int_{\mathcal{X}} f_L^\omega(z) f_L^\omega(z)' dz$. Suppose that the joint distribution $F_{Y,Z,W}$ is known, i.e., we know both the density function $f_{X,W}(\cdot,\cdot)$ and the reduced form $h(\cdot)$. Then, from $h_L^\omega = [h(\omega_1), .., h(\omega_L)]' = \int_{\mathcal{X}} f_L^\omega(z) m(z) dz$, we can calculate the exact form of the projection of m via

$$(P_{f_L^\omega} m)(x) = f_L^\omega(x)' Q_\omega^{*\dagger} h_L^\omega. \tag{5}$$

If C.2.1 holds, the above projection delivers a valid approximation of m, since $||P_{f_L^\omega} m - m||_{L^2(\mathcal{X})} \to 0$, as $L \to \infty$, under denseness of the linear span of $\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^{\infty}$. Identifiability of m is now obvious from uniqueness of the limit of a convergent sequence in a Hilbert space. In mathematical literature, the method of moment collocation uses (5) to find a numerical solution to an integral equation-see Kress (1989, p. 267), for example. The following theorem concerns necessariness of C.2.1 for identifiability. Under a weak condition on a joint density function, it shows that C.2.1 should hold for any 'dense' discretization points, when the model is identified.

Theorem 2.3 Suppose that a structural model is given by (1)-(3) with W_1 empty. In addition, assume that $f_{X,W}(\cdot,\cdot)$ is continuous on $\mathcal{X} \times \mathcal{W}$.

(i) If $m(\cdot)$ is identifiable, then, C.2.1 holds for any dense subset $\overline{\mathcal{W}}$ of $\mathcal{W}$.

(ii) If T has a dense range in $L^2(\mathcal{W})$, then, C.2.2 holds for any dense subset $\overline{\mathcal{X}}$ of $\mathcal{X}$.

[4] Here, $f_{X,W}(\cdot,\omega_l)$ is normalized implicitly by $f_W(\omega_l)$ so that $\int_{\mathcal{X}} f_{X,W}(x,\omega_l)/f_W(\omega_l)dx = 1$.

The symmetry argument also applies to Theorem 2.3. We remark that the conclusions of Theorem 2.3 are not stronger, in its context, than the assumptions of Theorem 2.2, since the former does not extend to general discretization points other than a dense subset of $\mathcal{W}$. For an immediate application of the above results, one may consider a joint density function, $f_{X,W}(x,w) = \sum_{k=1}^{K} p_k(x)q_k(w)$. This includes a trivial case with X independent of W. Since $\mathrm{lin}\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^{\infty}$ is at most of K-dimension, C.2.1 is violated for any $\overline{\mathcal{W}}$ in $\mathcal{W}$, so the model is not identifiable. Another implication of Theorem 2.3 concerns validity of $\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^{\infty}$ as approximating functions. It shows, under identifiability of m, that the finite-dimensional approximation in (5) will be consistent, if the sequence $\{\omega_l\}_{l=1}^{L}$ becomes dense in $\mathcal{W}$ as $L \to \infty$. The moment collocation method showed a similar but more general result by applying theory of a reproducing kernel Hilbert space; see Nashed and Wahba (1974). Our development here is much simpler, only based on Theorem 2.3.

We close the section by extending the above results to a case where X and W have some common elements of W_1. A slight modification of C.2.1 and 2.2 is enough to obtain the same conclusions.

C.2.3 For all $w_1 \in \mathcal{W}_1$, there exists a sequence $\{\omega_{2l}\}_{l=1}^{\infty} \subset \mathcal{W}_2$ such that $\mathrm{lin}\{f_{Z,W_1,W_2}(\cdot,w_1,\omega_{2l})\}_{l=1}^{\infty}$ is dense in $L^2_{\mathcal{Z}}$.

C.2.4 For all $w_1 \in \mathcal{W}_1$, there exists a sequence $\{\xi_l\}_{l=1}^{\infty} \subset \mathcal{Z}$ such that $\mathrm{lin}\{f_{Z,W_1,W_2}(\xi_l,w_1,\cdot)\}_{l=1}^{\infty}$ is dense in $L^2_{\mathcal{W}_2}$.

Theorem 2.4 Suppose that a structural model is given by (1)-(3) with W_1 possibly not empty.
(i) If C.2.3 holds, then, the integral operator $T : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ is one-to-one; i.e., $m(\cdot)$ is identifiable in $L^2(\mathcal{X})$.

(ii) If C.2.4 holds, then, T has a dense range in $L^2(\mathcal{W})$; i.e., at least one $m(\cdot) \in L^2(\mathcal{X})$ satisfies the given structural model, for $h(\cdot)$ in some 'dense' subspace.

Extension of Theorem 2.3 is done in a similar way. In this case, C.2.3 and C.2.4 holds for any dense subset $\{\omega_{2l}\}_{l=1}^{\infty}$ and $\{\xi_l\}_{l=1}^{\infty}$, respectively. We omit the details, since they are straightforward.

3 Statistical Theory of Regularization for Ill-Posed Problems

This section considers statistical estimation of a structural function which we assume to be identified by $m_0 = T^{-1}(h_0)$, where $h_0(w) = \int_{\mathcal{Y}} y f^0_{Y,W}(y,w)dy$ and $(Tm)(w) = \int_{\mathcal{Z}} m(z,w_1) f^0_{Z,W}(z,w)dz$. In mathematics theory for integral equations, it has been a central issue how to estimate m_0 from an approximate of h_0, when T is known. Ill-posedness of such inverse problems is now well known in the literature, and can be treated by regularization theory. Our statistical problems are more complicated, since the operator itself needs to be estimated. Some additional works are required, if one wants to apply regularization theory for solving an integral equation of a random operator. Newey and Powell (1988, 2002), Darolles, Florens, and Renault (2001), and Hall and Horowitz (2003) have recently attacked the issue, showing consistency and the convergence rates of their estimators. Those methods, categorized as (the classical or ordinary) Tikhonov regularization, possess a common form of ridge estimation which turns out be suboptimal in some cases. This paper takes a more general approach to the statistical inverse problems, trying to extend regularization theory into random integral equations. The statistical issues, such as consistency, optimal bounds, and the convergence rates, are discussed with no limitation on specific estimation of h_0 or T.

3.1 Generalized Inverse and Ill-Posed Problems

In a practical case where h_0 and $f^0_{Z,W}$ are unknown, approximate characterization of m_0 relies necessarily on some preliminary estimates, $\widehat{h}_{0,n}$ and $\widehat{f}_{Z,W}$, given an observed sample $\{(Y_i, Z_i, W_i)\}_{i=1}^n$. The actual problem to be solved is a 'random' Fredholm integral equation of the first kind,

$$(\widehat{T}_n m)(w) \equiv \int_{\mathcal{Z}} m(z,w_1)\widehat{f}_{Z,W,n}(z,w)dz = \widehat{h}_{0,n}(w). \tag{6}$$

Like mathematical inverse problems, several difficulties arise in estimating m_0 by inverting $\widehat{h}_{0,n}$ through $\widehat{T}_n$. Estimation of T is usually carried out by a certain discretization scheme-i.e., by determining finitely many unknowns. Since, in that case, $\widehat{T}_n$ is generally of finite rank, it is likely that $\widehat{h}_{0,n} \notin \mathcal{R}(\widehat{T}_n)$, or $\widehat{T}_n$ is not invertible.[5] The integral equation in (6) may possess *no* or *more than one* solutions. A common practice in econometric theory is to extend the notion of solution to the idea

[5] An operator G has a finite rank, if $\dim[\mathcal{R}(G)] < \infty$. It is obvious that any integral operator with a degenerate kernel is of finite rank. Throughout the paper, it will be assumed implicitly that $\dim[\mathcal{R}(\widehat{T}_n)] < \infty$, for n fixed, unless otherwise stated.

of the best approximation, based on minimum-distance. Given $\widehat{h}_{0,n}$ and $\widehat{T}_n$, the minimum-distance estimator of m_0 is defined by

$$\widehat{m}_n^\dagger = \underset{m(\cdot)\in L^2(\mathcal{X})}{\arg\min} ||\widehat{T}_n m - \widehat{h}_{0,n}||^2_{L^2(\mathcal{W})}, \tag{7}$$

where $\widehat{m}_n^\dagger$ is the solution of minimum norm, unless the minimum-distance estimator is unique. The underlying mapping from $\widehat{h}_{0,n}$ to $\widehat{m}_n^\dagger$ is so called the (Moore-Penrose) generalized inverse of $\widehat{T}_n$;

$$\widehat{m}_n^\dagger = \widehat{T}_n^\dagger(\widehat{h}_{0,n}),$$

where $\widehat{T}_n^\dagger$ is such that $\widehat{T}_n\widehat{T}_n^\dagger\widehat{T}_n = \widehat{T}_n$. From the first order condition of (7), it follows that $\widehat{T}_n^*\widehat{T}_n\widehat{m}_n^\dagger = \widehat{T}_n^*\widehat{h}_{0,n}$, leading to $\widehat{T}_n^\dagger = (\widehat{T}_n^*\widehat{T}_n)^\dagger\widehat{T}_n^*$, where $\widehat{T}_n^* : L^2(\mathcal{W}) \to L^2(\mathcal{X})$ is the adjoint of $\widehat{T}_n : L^2(\mathcal{X}) \to L^2(\mathcal{W})$. The generalized inverse of $\widehat{T}_n$ has a domain given by $\mathcal{R}^\perp(\widehat{T}_n) \oplus \mathcal{R}(\widehat{T}_n)$; see Groetsch (1993, p.80), for example. Since $\mathcal{R}(\widehat{T}_n)$ is finite-dimensional and thus closed, it holds that $\mathcal{R}^\perp(\widehat{T}_n)\oplus\mathcal{R}(\widehat{T}_n) = L^2(\mathcal{W})$, for any fixed n. That is, $\widehat{m}_n^\dagger$ is well defined for any $\widehat{h}_{0,n} \in L^2(\mathcal{W})$. When $\widehat{T}_n^*\widehat{T}_n$ is one-to-one on the range space of $\widehat{T}_n^*$, the minimum distance estimator is simplified to $\widehat{m}_n^\dagger = (\widehat{T}_n^*\widehat{T}_n)^{-1}\widehat{T}_n^*\widehat{h}_{0,n}$.

Remark 3.1 (*Closed form of solutions*) Although $\widehat{m}_n^\dagger$ serves as an approximate solution to (6), it may seem elusive to find the exact functional form of $\widehat{m}_n^\dagger$. Below in section 5, we show that it is possible to derive the exact closed form of $\widehat{m}_n^\dagger$, when h_0 and $f_{Z,W}$ are estimated by the kernel smoothing method. Instead of the exact form of $\widehat{m}_n^\dagger$, one may try to define an alternative minimum-distance estimator, by discretizing (6) on collocation points, say, $\{\omega_l\}_{l=1}^L \subset \mathcal{W}$. With $\widehat{T}_{n,L}^\omega : L^2(\mathcal{X}) \to \mathbb{R}^L$ defined by $\widehat{T}_{n,L}^\omega(m) = [(\widehat{T}_n m)(\omega_1), ..(\widehat{T}_n m)(\omega_L)]'$, the relevant minimization problem is

$$\widetilde{m}_{n,L}^\dagger = \underset{m(\cdot)\in L^2(\mathcal{X})}{\arg\min} ||\widehat{T}_{n,L}^\omega m - \widehat{h}_L^\omega||^2_{\mathbb{R}^L} = \underset{m(\cdot)\in L^2(\mathcal{X})}{\arg\min} \sum_{l=1}^L [(\widehat{T}_n m)(\omega_l) - \widehat{h}_{0,n}(\omega_l)]^2,$$

where $\widehat{h}_L^\omega = [\widehat{h}_{0,n}(\omega_1), .., \widehat{h}_{0,n}(\omega_L)]$. The closed form of $\widetilde{m}_{n,L}^\dagger$ is available, regardless of an estimation method used for h_0 and $f_{Z,W}$. When X and W are disjoint, the minimum distance estimator is exactly of the same form as (5) with h_0 and $f_{Z,W}$ replaced by their estimates

$$\widetilde{m}_{n,L}^\dagger(x) \equiv \widehat{T}_{n,L}^{\omega\dagger}(\widehat{h}_L^\omega) = \widehat{f}_L^\omega(x)'\widehat{Q}_\omega^{*\dagger}\widehat{h}_L^\omega,$$

where $\widehat{f}_L^\omega(x) = [\widehat{f}_{X,W}(x,\omega_1), .., \widehat{f}_{X,W}(x,\omega_L)]$, and $Q_\omega^* = \int_\mathcal{X} f_L^\omega(u) f_L^\omega(u)' du$. ■

Consistency of the natural estimator $\widehat{m}_n^\dagger$, however, is not ensured by consistency of the preliminary estimates $\widehat{h}_{0,n}$ and $\widehat{T}_n$. For clarity of the statement, we need to define statistical properties of random operators. As usual, $\widehat{h}_{0,n}$ is said to be (L^2-) consistent for h_0, if and only if $||\widehat{h}_{0,n} - h_0||_{L^2(\mathcal{W})} \xrightarrow{p} 0$, as $n \to \infty$.

Definition 3.1 (i) A random operator $\widehat{T}_n : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ is consistent for $T : L^2(\mathcal{X}) \to L^2(\mathcal{W})$, if and only if $||\widehat{T}_n m - Tm||_{L^2(\mathcal{W})} \xrightarrow{p} 0$, for all $m \in L^2(\mathcal{X})$, i.e., $\widehat{T}_n$ converges pointwise to T (in $L^2(\mathcal{X})$) in probability. (ii) $\widehat{T}_n$ is uniformly consistent for T on $\mathcal{M}_X \subset L^2(\mathcal{X})$, if and only if $\text{plim}_{n\to\infty} \sup_{m\in\mathcal{M}_X,\, m\neq 0} ||\widehat{T}_n m - Tm||_{L^2(\mathcal{W})}/||m||_{L^2(\mathcal{X})} = 0$, i.e., $\widehat{T}_n$ converges to T uniformly on $\mathcal{M}_X$, in probability.

Recalling the definition of $||\widehat{T}_n - T||_{\mathcal{M}_X \to L^2(\mathcal{W})}$, uniform convergence in Definition 3.1(ii) with $\mathcal{M}_X = L^2(\mathcal{X})$ is equivalent to convergence in operator norm. Asymptotic properties of $\widehat{m}_n^\dagger$ are explained from the decomposition

$$\widehat{T}_n^\dagger(\widehat{h}_{0,n}) - m_0 = \widehat{T}_n^\dagger[\widehat{h}_{0,n} - \widehat{T}_n m_0] + [(\widehat{T}_n^*\widehat{T}_n)^\dagger \widehat{T}_n^*\widehat{T}_n - I]m_0. \tag{8}$$

The term in the first square bracket, written as $(\widehat{h}_{0,n} - h_0) - (\widehat{T}_n - T)m_0$, represents the composite errors associated with estimation of h_0 and T. The term will converge to zero in L^2-norm, if $\widehat{h}_{0,n}$ and $\widehat{T}_n$ are consistent. The second term, due to non-invertibility of $\widehat{T}_n$, reduces to $-P_{\mathcal{N}(\widehat{T}_n)}m_0$, by the identity $(\widehat{T}_n^*\widehat{T}_n)^\dagger \widehat{T}_n^*\widehat{T}_n = I - \widehat{P}_{\mathcal{N}(\widehat{T}_n)}$, where $P_{\mathcal{N}(\widehat{T}_n)}$ is the orthogonal projector onto $\mathcal{N}(\widehat{T}_n)$; see Nashed (1976). or Groetsch (1977). Applying Lemma 3.1 in the appendix, we can show that the second term converges to zero in probability, under consistency of $\widehat{T}_n$ and invertibility of T. For consistency of $\widehat{m}_n^\dagger$, it remains crucial to know whether $\widehat{T}_n^\dagger$ is bounded uniformly in n. If $||\widehat{T}_n^\dagger||_{L^2(\mathcal{W})\to L^2(\mathcal{X})} = O_p(1)$, consistency of $\widehat{m}_n^\dagger$ will follow from a direct extension of the Slutzky Theorem to infinite-dimensional spaces. The following result, however, shows that uniform boundedness of $\widehat{T}_n^\dagger$ does not obtain in a fairly regular situation.

Proposition 3.1 Suppose that $T : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ is one-to-one, and $\widehat{T}_n$ has a finite rank. Assume that $\widehat{T}_n : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ is uniformly consistent for T on $\mathcal{M}_X \subset L^2(\mathcal{X})$ s.t. $\dim(\mathcal{M}_X) = \infty$. Then,

$$\text{plim}_{n\to\infty} ||\widehat{T}_n^\dagger||_{L^2(\mathcal{W})\to L^2(\mathcal{X})} = \infty.$$

The only binding condition in Proposition 3.1 is uniform convergence of $\widehat{T}_n$ to T on some infinite-dimensional subspace of $L^2(\mathcal{X})$, which in fact holds under a quite regular condition such that $\widehat{f}_{Z,W,n}(\cdot,\cdot)$ converges to the truth in L^2-norm. To see this, we just observe that, by Cauchy-Schwartz inequality, for all $m(\cdot) \in L^2(\mathcal{X})$,

$$\begin{aligned} ||(\widehat{T}_n - T)m||^2_{L^2(\mathcal{W})} &\leq ||||\widehat{f}_{Z,W,n}(z,w) - f_{Z,W}(z,w)||_{L^2(\mathcal{Z})}||m(z,w_1)||_{L^2(\mathcal{Z})}||^2_{L^2(\mathcal{W})} \\ &\leq ||\widehat{f}_{Z,W}(\cdot) - f_{Z,W}(\cdot)||^2_{L^2(\mathcal{Z}\times\mathcal{W})}||m||^2_{L^2(\mathcal{X})}, \end{aligned} \tag{9}$$

i.e., $||\widehat{T}_n - T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})} \leq ||\widehat{f}_{Z,W}(\cdot) - f_{Z,W}(\cdot)||^2_{L^2(\mathcal{Z}\times\mathcal{W})}$. With no strong restrictions imposed, Proposition 3.1 characterizes asymptotic unboundedness of $\widehat{T}_n^\dagger$ as a generic property. It dose not mean inconsistency of $\widehat{m}_n^\dagger$ automatically, since uniform boundedness of $\widehat{T}_n^\dagger$ is not a necessary condition for consistency.[6] However, the estimator, in general, lacks stability w.r.t. the statistical errors in $\widehat{T}_n$ or $\widehat{h}_{0,n}$. Even small perturbations of $\widehat{T}_n$ or $\widehat{h}_{0,n}$ may result in unacceptably large errors in $\widehat{m}_n^\dagger = \widehat{T}_n^\dagger(\widehat{h}_{0,n})$. Since $\widehat{T}_n^\dagger$ becomes more explosive as $n \to \infty$, the approximate solutions may get worse, as more observations (and thus more discretizations) are used in estimating $\widehat{T}_n$ and $\widehat{h}_{0,n}$. In this sense, the estimation problem in (6) is called *statistically ill-posed.* It needs to be pointed out that such ill-posedness occurs, because the underlying mapping from a reduced form to a structural function is not continuous. By Bounded Inverse Theorem, the inverse operator $T^{-1} : \mathcal{R}(T) \to L^2(\mathcal{X})$ is bounded (i.e., continuous), if and only if $\mathcal{R}(T)$ is closed. In infinite dimensional Hilbert spaces, $\mathcal{R}(T)$ is closed, only when T has a degenerate kernel such as $f_{Z,W}(z,w) = \sum_{k=1}^K p_k(z)q_k(w)$. We showed that such density functions are excluded by identifiability. It then appeals to our intuition that $\widehat{T}_n^\dagger$ will be unbounded, as $\widehat{T}_n$ gets close to T.

3.2 Consistent Estimation by Regularization

Difficulties in the estimation problem (6) are closely related to the smoothing effects of the (estimated) integral operator. Since nonsmooth components, like cusps or edges, in m are smoothed out by integration, the reverse operation will amplify any high-frequency parts of $\widehat{h}_{0,n}$, just as simple differentiation does. Considering that the estimation errors in $\widehat{h}_{0,n}$ correspond to such high-frequency parts, naive inversion of $\widehat{h}_{0,n}$ may end up with extremely large errors in estimating m. From this observation, two points are essential in dealing with an ill-posed problem. Firstly, to cure the instability problem, one needs to be able to filter out the high-frequency components of $\widehat{h}_{0,n}$, in a controllable way. Secondly, such filtering should not make substantial loss of information in restoring a true solution. The first problem is resolved by 'regularization' that amounts to a bounded approximation of the unbounded inverse operator. For the second, certain 'smoothness' needs to be imposed on a true

[6] Consistency of $\widehat{m}_n^\dagger$ depends on whether $||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})}$ converges to zero at a faster rate than the square-root of the minimum eigenvalue of $\widehat{T}_n\widehat{T}_n^*$ decays.

solution, since the loss of information due to a bounded approximation concentrates on nonsmooth components. The following example shows how one can make use of additional knowledge about smoothness of $m(\cdot)$ to regularize an ill-posed problem.

Example 3.1 (*the classical Tikhonov regularization; compactification)* Suppose that the true solution m_0 is continuously differentiable, having square-integrable derivatives. A set of admissible solutions is now given by $\mathcal{M}_X^B = \{m(\cdot) \in L^2(\mathcal{X})\text{: } ||m||_{L^2(\mathcal{X})} + ||m'||_{L^2(\mathcal{X})} \leq B$, for some $B > 0\}$, where m is of bounded Sobolev norm. Since $\mathcal{M}_X^B$ is compact in $L^2(\mathcal{X})$ by the Sobolev Imbedding Theorem, the (injective) operator T, when restricted to $\mathcal{M}_X^B$, has a 'bounded' inverse, say, $T^{-1}_{|\mathcal{M}_X^B}$.[7] That is, an ill-posed problem can be regularized via compactification. Stable approximation of m_0 is possible, if the 'bounded' operator, $T^{-1}_{|\mathcal{M}_X^B}$ can be estimated consistently. Letting $\widehat{T}_{|\mathcal{M}_X^B}$ be a restriction of $\widehat{T}_n$ on $\mathcal{M}_X^B$, we define an estimator by $\widehat{m}_n^B = \widehat{T}^{\dagger}_{|\mathcal{M}_X^B}(\widehat{h}_{0,n})$, where $\widehat{T}^{\dagger}_{|\mathcal{M}_X^B}$ is the Moore-Penrose generalized inverse of $\widehat{T}_{|\mathcal{M}_X^B}$. Comparing to (7), $\widehat{m}_n^B$ comes from solving constrained minimum-distance; for $B > 0$,

$$\widehat{m}_n^B = \underset{m(\cdot) \in L^2(\mathcal{X})}{\arg\min} ||\widehat{T}_n m - \widehat{h}_{0,n}||_{L^2(\mathcal{W})} \quad \text{s.t. } ||m||_{L^2(\mathcal{X})} + ||m'||_{L^2(\mathcal{X})} \leq B. \tag{10}$$

For fixed B, $\widehat{T}^{\dagger}_{|\mathcal{M}_X^B}$ is uniformly bounded, and hence consistency of $\widehat{T}_{|\mathcal{M}_X^B}$ and $\widehat{h}_{0,n}$ is sufficient for consistency of $\widehat{m}_n^B$. (10) shows clearly that a filtering effect is achieved by damping out highly-oscillating parts of the approximate solutions. In Newey and Powell (2002), (10) was combined with orthogonal series expansions, to define a regularized nonparametric 2SLS estimator. ■

An implicit regularization effect is used in the compactification method by imposing integral bounds on the derivatives. An alternative but more general class of regularization methods are generated by a direct way of bounded approximation for the inverse operator T^{-1}. As a sensible modification of $\widehat{T}_n^{\dagger} = (\widehat{T}_n^* \widehat{T}_n)^{\dagger} \widehat{T}_n^*$, we suggest a family of bounded operators

$$\widehat{R}_{\alpha,n} = U_\alpha(\widehat{T}_n^* \widehat{T}_n)\widehat{T}_n^*, \quad \alpha > 0 \tag{11}$$

that satisfies:

(a) $U_\alpha(\widehat{T}_n^* \widehat{T}_n)$ is close to $(\widehat{T}_n^* \widehat{T}_n)^{\dagger}$, for small α, in the sense that $\widehat{R}_{\alpha,n}\widehat{T}_n$ converges pointwise to the identity, I, in $L^2(\mathcal{X})$, and

(b) $U_\alpha(\widehat{T}_n^* \widehat{T}_n)$ is uniformly bounded (in n) by a known function of α, say, $1/\alpha$.

[7] Let $T_{|\mathcal{K}} : \mathcal{K} \subset L^2(\mathcal{X}) \to L^2(\mathcal{W})$ be a restriction of a bounded injective operator, T: $L^2(\mathcal{X}) \to L^2(\mathcal{W})$. If $\mathcal{K}$ is compact in $L^2(\mathcal{X})$, then, $T^{-1}_{|\mathcal{K}}$: $\mathcal{R}(T_{|\mathcal{K}}) \to \mathcal{K}$, is continuous, by *Tikhonov's theorem*-see Groetsch (1993, p.79).

By the former condition, $\widehat{R}_{\alpha,n}$ (with α small) serves as an approximation of T^{-1}, as $\widehat{T}_n^\dagger$ does. The second condition means that $\widehat{R}_{\alpha,n}$, unlike $\widehat{T}_n^\dagger$, is stabilized through a newly-introduced term α called a regularization parameter. Note that, in contrast to $\widehat{T}^\dagger_{|\mathcal{M}_X^B}$, boundedness of $\widehat{R}_{\alpha,n}$ is controlled in an explicit way, via the regularization parameter. To guarantee the properties of (a) and (b), we will need the following conditions on $U_\alpha(\cdot)$ that are borrowed from mathematical regularization theory.

Condition 3.1 Let $\overline{\lambda} \equiv \sup_{n\geq n_0} ||\widehat{T}_n^*\widehat{T}_n||_{L^2(\mathcal{X})\to L^2(\mathcal{X})}$. A parameter dependent family of continuous functions, $\{U_\alpha(\cdot)\}_{\alpha>0}$, defined on $(0,\overline{\lambda}]$, satisfy that (i) $\sup_{\lambda\in(0,\overline{\lambda}]} |U_\alpha(\lambda)\lambda| \leq C < \infty$, for $\alpha > 0$, (ii) $\lim_{\alpha\to 0^+} U_\alpha(\lambda) = \frac{1}{\lambda}$, for all $\lambda \in (0,\overline{\lambda}]$, and (iii) $\sup_{\lambda\in(0,\overline{\lambda}]} |U_\alpha(\lambda)| = O(\frac{1}{\alpha})$, as $\alpha \to 0^+$.[8]

From the fact that $\widehat{T}_n^*\widehat{T}_n$ is self-adjoint, $U_\alpha(\widehat{T}_n^*\widehat{T}_n)$ is well defined based on spectral theory for self-adjoint linear operators, as long as the real-valued function $U_\alpha(\cdot)$ is defined on the spectrum of $\widehat{T}_n^*\widehat{T}_n$. Since the random operator $\widehat{T}_n$ in practice is of finite rank and thereby compact, it is sufficient to define $U_\alpha(\cdot)$ on a bounded interval, $(0,\overline{\lambda}]$, where $\overline{\lambda} = \sup_{n\geq n_0} ||\widehat{T}_n^*\widehat{T}_n||_{L^2(\mathcal{X})\to L^2(\mathcal{X})}$. For such $U_\alpha(\cdot)$, an approximate solution to (6) is defined by

$$\widehat{m}_{\alpha,n} = \widehat{R}_{\alpha,n}\widehat{h}_{0,n} = U_\alpha(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{h}_{0,n}, \tag{12}$$

which we call a regularized IV estimator of m_0. In Lemma 3.2 of the appendix, we show that, for $U_\alpha(\cdot)$ satisfying C.3.1, two properties of (a) and (b) in the above hold with $||\widehat{R}_{\alpha,n}||_{L^2(\mathcal{W})\to L^2(\mathcal{X})} = O_{as}(\alpha^{-1/2})$, whenever $\widehat{T}_n$ converges pointwise to the true (injective) operator T in $L^2(\mathcal{X})$. To see the implications, we consider an error decomposition of the regularized estimates, which is given, similar to (8), by

$$\widehat{m}_{\alpha,n} - m_0 = \widehat{s}_\alpha + \widehat{b}_\alpha \equiv \widehat{R}_{\alpha,n}(\widehat{h}_{0,n} - \widehat{T}_n m_0) + [U_\alpha(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{T}_n - I]m_0. \tag{13}$$

The first term corresponds to propagation of the composite errors, and the second, an extra error term due to regularization. From uniform boundedness of $\widehat{R}_{\alpha,n}$ (by $1/\sqrt{\alpha}$), it follows that $\widehat{s}_\alpha$ converges to zero in probability, if the decaying rate of $\sqrt{\alpha}$ is slower than the convergence rates of $\widehat{h}_{0,n}$ and $\widehat{T}_n$. Negligibility of $\widehat{b}_\alpha$ (as $\alpha \to 0$) is obvious from the property (a), i.e., from pointwise convergence of $U_\alpha(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{T}_n$ to I in $L^2(\mathcal{X})$. In sum, given consistency of $\widehat{T}_n$ and $\widehat{h}_{0,n}$, the regularization methods in (11), with $U_\alpha(\cdot)$ satisfying C.3.1, yield consistent estimation of m_0, for some choices of a regularization parameter, $\alpha = \alpha(n)$, converging to zero.

Theorem 3.2 Suppose that $U_\alpha(\cdot)$ satisfies C.3.1, and the linear operator $\widehat{T}_n : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ has a finite rank. Also, assume that $||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} \xrightarrow{p} 0$, and $\widehat{T}_n$ is a consistent estimator for

[8] Section 4 gives a detailed discussion about several examples of $U_\alpha(\cdot)$ satisfying C.3.1.

the true operator $T : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ which is bounded and injective. If $\alpha = \alpha(n)$ is such that $\alpha(n) \to 0$ and $||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})}/\sqrt{\alpha(n)} \xrightarrow{p} 0$, as $n \to \infty$, then, $||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} \xrightarrow{p} 0$, as $n \to \infty$, for all $m_0 \in L^2(\mathcal{X})$.

In contrast to Example 3.1, the consistency result of Theorem 3.2 applies to any square-integrable function, $m_0 \in L^2(\mathcal{X})$, with no constraint on smoothness of m_0. It instead requires implicitly to know the convergence rate of the composite errors, $\zeta_n \equiv \widehat{h}_{0,n} - \widehat{T}_n m_0$. For standard nonparametric procedures, such rate will be available under some smoothness conditions on h_0 and $f_{Z,W}$. An immediate choice of $\alpha = \alpha(n)$ follows, for example, from (9) which, together with the triangle inequality, leads to

$$||\zeta_n||_{L^2(\mathcal{W})} \le ||\widehat{h}_{0,n} - h_0||_{L^2(\mathcal{W})} + C||\widehat{f}_{Z,W} - f_{Z,W}||^2_{L^2(\mathcal{Z}\times\mathcal{W})},$$

for $m_0 \in L^2(\mathcal{X})$ with $||m_0|| \le C < \infty$. Under the identification relation ($h_0 = Tm_0$), a reduced form function is equivalent to an integral of the underlying structural function, so, it will satisfy some smoothness automatically, if $f_{Z,W}(\cdot)$ does.

Remark 3.2 (*strong L^2-consistency of $\widehat{m}_{\alpha,n}$*) It is possible to show a strong form of Theorem 3.2 with an replacement of '$\xrightarrow{p}$' by '$\xrightarrow{as}$', through a straightforward extension of the proofs for Lemma 3.1 and 3.2 in the appendix. Suppose that $\widehat{h}_{0,n}$ and $\widehat{f}_{Z,W}$ are strongly L^2-consistent for h_0 and $f_{Z,W}$, respectively, in the sense that $||\widehat{h}_{0,n} - h_0||_{L^2(\mathcal{W})} = o_{as}(1)$, and $||\widehat{f}_{Z,W} - f_{Z,W}||^2_{L^2(\mathcal{Z}\times\mathcal{W})} = o_{as}(1)$. Again, from (9), the latter condition implies $||\widehat{T}_n - T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})} = o_{as}(1)$, which is sufficient for Lemma 3.1 as well as Lemma 3.2.(iii) to hold almost surely. In consequence, strong L^2-consistency of $\widehat{m}_{\alpha,n}$ obtains, under $\alpha = \alpha(n) \to 0$ such that $||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})}/\sqrt{\alpha(n)} \xrightarrow{as} 0$. ■

3.3 Smoothness Condition and Optimal Bounds

Smoothness of m_0 needs to be assumed for further asymptotic properties of the estimators. Following mathematical regularization theory, this section introduces an abstract smoothness condition, based on a sourcewise-representation of m_0. Use of such condition is illustrated by deriving some lower bounds on the convergence rates for the estimators in (12). The same smoothness condition turns out to play a crucial role in defining new optimal bounds. More analyses of the convergence rates will be given in the next section.

In Theorem 3.2, we already discussed the convergence rate of the first term in (13). The L^2-norm of $\widehat{s}_\alpha$ is determined by the noise level, $||\zeta_n||_{L^2(\mathcal{W})}$, multiplied by the condition number,

$||\widehat{R}_{\alpha,n}||_{L^2(\mathcal{W})\to L^2(\mathcal{X})}$. More careful investigation reveals that ζ_n consists of stochastic errors from estimating h_0 and a bias from estimating m_0-see Proposition 5.3 in Section 5. For the convergence rate of $\widehat{m}_{\alpha,n}$, it remains to calculate the asymptotic order of the regularization errors, $\widehat{b}_\alpha = (\widehat{\Gamma}_\alpha - I)m_0$, where $\widehat{\Gamma}_\alpha = U_\alpha(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{T}_n$. Unlike the consistency result in Theorem 3.2, the convergence rate of $\widehat{b}_\alpha$ cannot be fixed, for arbitrary $m_0 \in L^2(\mathcal{X})$. This is because $\widehat{\Gamma}_\alpha$ does not converge 'uniformly' to I, on $L^2(\mathcal{X})$, for any choice of $U_\alpha(\cdot)$.[9] A meaningful question then will be whether the convergence rate of $\widehat{b}_\alpha$ is available, still on a large subset of $L^2(\mathcal{X})$, by strengthening some of the conditions in C.3.1 appropriately.

Condition 3.2 Given $U_\alpha : (0, \overline{\lambda}] \to \mathbb{R}$, it holds for any $\mu \in (0, \overline{\mu}]$ that $\sup_{\lambda\in(0,\overline{\lambda}]} \lambda^\mu |U_\alpha(\lambda)\lambda - 1| \leq C\alpha^\mu$, for any $\alpha \in (0, \alpha_0)$, where $\alpha_0 > 0$.

It is clear that C.3.1(ii) follows from C.3.2. The latter condition also implies C.3.1(i), by Principle of Uniform Boundedness; see Taylor and Lay (1980, p.190). In Lemma 3.3 of the appendix, we show that, for $U_\alpha(\cdot)$ satisfying C.3.2,

$$||(\widehat{\Gamma}_\alpha - I)(\widehat{T}_n^*\widehat{T}_n)^\mu||_{L^2(\mathcal{X})\to L^2(\mathcal{X})} \leq C\alpha^{\min(\mu,\overline{\mu})}, \text{ a.s,} \tag{14}$$

implying that $\widehat{b}_\alpha$ decays at the rate of $\alpha^{\min(\mu,\overline{\mu})}$, if m_0 lies in the range space of $(\widehat{T}_n^*\widehat{T}_n)^\mu$, for all n sufficiently large. Suppose that $\widehat{T}_n$ converges uniformly T in $L^2(\mathcal{X})$, then, it makes a sense that a similar argument will hold on the range space of $(T^*T)^\mu$. Below, we will use this observation to calculate the convergence rate of $\widehat{m}_{\alpha,n}$, by assuming additional information about the true solution such that

$$m_0 \in \mathcal{M}_\mu \equiv \mathcal{R}((T^*T)^\mu), \quad \text{for } \mu > 0. \tag{15}$$

Remark 3.3 (i) Since T is an integral operator of smoothing effects, the sourcewise representation of m_0 in (15) can be understood as an *abstract smoothness condition.* The alternative definition of smoothness is indeed one of the features that distinguish regularization theory from the standard nonparametric methods. To get some idea of the condition, suppose that bivariate r.v. (X, W), supported by $[0, 1] \times [0, 1]$, have uniform distributions such that $f_{X,W}(x, w) = 1$, for $0 \leq x \leq w \leq 1$, and $f_{X,W}(x, w) = 0$, otherwise. For $\mu = 1$, (15) reduces to '$m_0(x) = \int_x^1 \int_0^w m_\mu(u)dudw$, for some

[9] Such property of $\widehat{\Gamma}_\alpha$, in fact, obtains only on a finite-dimensional subspace of $L^2(\mathcal{X})$. For this, we only remark that (i) the limit (in operator norm) of a sequence of compact operators is also compact; and (ii) the identity operator $I : \mathcal{M}_X \to \mathcal{M}_X$ is compact if and only if $\dim(\mathcal{M}_X) < \infty$; see Kress (1989, p.18, Theorem 2.16 and 2.19).

$m_\mu \in L^2[0,1]$', implying that m_0 has square-integrable (generalized) second-derivatives. In general, the abstract smoothness condition imposes stronger smoothness on m_0, as the kernel of the integral operator becomes smoother, or μ increases.

(ii) By definition, $\mathcal{M}_\mu \subset \mathcal{M}_{\mu'}$, for $\mu' \leq \mu$. Also, from $\overline{\mathcal{R}}((T^*T)^\mu) = \mathcal{N}^\perp((T^*T)^\mu)$ and $\mathcal{N}((T^*T)^\mu) = \mathcal{N}(T)$, it follows that $\mathcal{M}_\mu$ is dense in $L^2(\mathcal{X})$, if T is one-to-one, i.e., m_0 is identifiable.

(ii) As indicated by (14), the decaying rate of the pure regularization bias can be quite slow for μ close to zero, and cannot exceed $\alpha^{\overline{\mu}}$ even for $\mu > \overline{\mu}$. The latter phenomenon, known as saturation of regularization, depends on the way that $\widehat{R}_{\alpha,n}$ approximates the inverse operator, T^{-1}. We define the *qualification* of a regularization method to be $\overline{\mu}(> 0)$, if and only if C.3.2 holds only for $\mu \in (0, \overline{\mu}]$, but not for $\mu > \overline{\mu}$. ■

For $m_0 \in \mathcal{M}_\mu$, the regularization errors separate into two parts

$$\widehat{b}_\alpha = (\widehat{\Gamma}_\alpha - I)(\widehat{T}_n^*\widehat{T}_n)^\mu m_\mu - (\widehat{\Gamma}_\alpha - I)[(\widehat{T}_n^*\widehat{T}_n)^\mu - (T^*T)^\mu]m_\mu, \tag{16}$$

where $m_\mu = (T^*T)^{-\mu}m_0 \in L^2(\mathcal{X})$. For a benchmark case of $m_0 \in \mathcal{R}(T^*)$ or $\mathcal{R}(T^*T)$, the convergence rate of $\widehat{b}_\alpha$ follows easily from (14), since the second term is quite simple in this case. We remark that $\mathcal{R}(T^*) = \mathcal{R}((T^*T)^{1/2})$, since, by polar decomposition, $T^* = (T^*T)^{1/2}U$, where U is a unitary operator such that $U^*U = I_{\mathcal{D}(T^*)}$; see Taylor and Lay (1980, p.379). Let C and C_i denote a generic constant that is a finite real number.

Theorem 3.3 Let $U_\alpha(\cdot)$ satisfy C.3.1 and C.3.2, with $\overline{\mu} \geq 1$. Then, (i) for $m_0 \in \mathcal{R}((T^*T)^{1/2})$, it holds that, for any n,

$$||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} \leq \frac{C_1}{\sqrt{\alpha}}||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} + C_2\alpha^{1/2}||h_1||_{L^2(\mathcal{W})} + C_3||(\widehat{T}_n^* - T^*)h_1||_{L^2(\mathcal{X})}, \text{ a.s.} \tag{17}$$

where $h_1 = T^{*-1}(m_0)$. And, (ii) for any $m_0 \in \mathcal{M}_1 = \mathcal{R}(T^*T)$, it holds that, for any n,

$$\begin{aligned} ||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} \quad \leq \quad & \frac{C_1}{\sqrt{\alpha}}||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} + C_2\alpha||m_1||_{L^2(\mathcal{X})} \\ & + C_3\sqrt{\alpha}||(\widehat{T}_n - T)m_1||_{L^2(\mathcal{W})} + C_4||(\widehat{T}_n^* - T^*)h_1||_{L^2(\mathcal{X})}\}, \quad \text{a.s.} \end{aligned} \tag{18}$$

where $m_1 = (T^*T)^{-1}m_0$, and $h_1 = Tm_1$.

Theorem 3.3 shows sufficiency of C.3.1 and 3.2 for derivation of the convergence rates for the general regularization method in (11)-at least, for specific orders of smoothness.[10] The first term in (17)

[10] We need to know the shape of $U_\alpha(\cdot)$, for the convergence rates in a more general case of μ; see the analyses in section 4.

or (18) is already explained. The remaining terms correspond to the asymptotic orders of the regularization errors $(\widehat{b}_\alpha)$. The second term represents the decaying rate of the (stochastic) pure regularization bias, which is given by (14). It decays to zero at a faster rate, as m_0 is smoother. The last terms reflect how the estimation of the unknown operator affects the convergence rates. If m_0 is further restricted to be in $\mathcal{M}_{\mu,\rho} = (T^*T)^\mu(B_\rho)$, Theorem 4.3 can be expressed more conveniently, by means of the uniform convergence rate of $\widehat{T}_n$ and $\widehat{T}_n^*$, where B_ρ is the sphere (with radius ρ) in $L^2(\mathcal{X})$.

Corollary 3.4 Assume the conditions of Theorem 3.3. If $\widehat{T}_n$ (and $\widehat{T}_n^*$) converges uniformly to T on $L^2(\mathcal{X})$ (and T^* on $L^2(\mathcal{W})$, respectively), then, (i) for $m_0 \in \mathcal{M}_{1/2,\rho}$,

$$||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} \leq C\{\frac{1}{\sqrt{\alpha}}||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} + \alpha^{1/2} + ||\widehat{T}_n^* - T^*||_{L^2(\mathcal{X})\to L^2(\mathcal{W})}\}, \text{ a.s.}$$

and (ii) for $m_0 \in \mathcal{M}_{1,\rho}$,

$$\begin{aligned} ||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} \leq\ & C\{\frac{1}{\sqrt{\alpha}}||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} + \alpha \\ & + \sqrt{\alpha}||\widehat{T}_n - T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})} + ||\widehat{T}_n^* - T^*||_{T(\mathcal{M}_{0,\rho})\to L^2(\mathcal{X})}\}, \text{ a.s.} \end{aligned}$$

Once the estimators for $\widehat{h}_{0,n}$ and $\widehat{T}_n$ are fixed, the asymptotic order of each term appearing in Corollary 3.4 can be calculated from the standard results on nonparametric estimation. To be rigorous, the above results give only a lower bound on the convergence rates of $\widehat{m}_{\alpha,n}$. In Theorem 4.4 of the next section, we show that a sharper bound in fact is available for a specific regularization method, through improvements upon the last term in (18). Related to this issue, an interesting question concerns the best-possible convergence rate attainable by approximate solutions to (6).

In mathematical inverse problems with T known, Tautenhahn (1998) showed that the best-possible convergence rate for $m_0 \in \mathcal{M}_{\mu,\rho}$ is given by $O(\overline{\delta}_n^{\frac{2\mu}{2\mu+1}})$, where $\overline{\delta}_n$ denotes the (deterministic) errors in estimating h_0; i.e., $\overline{\delta}_n = ||\widehat{h}_{0,n} - h_0||_{L^2(\mathcal{W})}$. In the rest of the section, we extend the argument of Tautenhahn (1998) to a statistical inverse problem in (6). To this effect, we first need to set up a meaningful criterion for optimal bounds. Assuming that any reasonable estimation of m_0 makes use of the relation in (6), proper optimal bounds may well depend on accuracies of the preliminary estimates, $\widehat{h}_{0,n}$ or $\widehat{T}_n$, or both. As will be made clear shortly, our characterization of the best-possible convergence rate is closely related to the composite error bound, $||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})}$. Let a large class of estimation methods, $\mathfrak{R}$, consist of a (possibly nonlinear) mapping $R : L^2(\mathcal{W}) \to L^2(\mathcal{X})$ such that $R(0) = 0$, and the resulting estimate for m_0 is defined by $R(\widetilde{h})$, where $\widetilde{h}$ is a given estimate of h_0. Given preliminary estimates, $\widehat{h}_{0,n}$ and $\widehat{T}_n$, such that $||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} = O_p(\delta_n)$, we define the worst-case convergence rate of $R \in \mathfrak{R}$, for $m_0 \in \mathcal{M} \subset L^2(\mathcal{X})$, by

$$\Xi(\{\delta_k\}, \mathcal{M}, R) = \sup_{m_0 \in \mathcal{M},\ ||\widehat{h}_{0,k} - \widehat{T}_k m_0||_{L^2(\mathcal{W})} = O_p(\delta_k)} \mathrm{E}(||R(\widehat{h}_{0,n}) - m_0||_{L^2(\mathcal{X})}).$$

A rate-optimal method R^* in $\mathfrak{R}$ is the one for which there exists N (≥ 1) such that

$$\Xi(\{\delta_k\}, \mathcal{M}, R^*) \leq C \inf_{R \in \mathfrak{R}} \Xi(\{\delta_k\}, \mathcal{M}, R),$$

for all $n \geq N$. In the appendix (the proof of Theorem 3.5), it is shown that the best-possible convergence rates of any estimation method (in minimax sense) is bounded by the modulus of stochastic equicontinuity of $\widehat{T}_n^{\dagger}$. This generalizes the result by Ivanov et al (1978) for a deterministic case with T known. An explicit form of such bound can be calculated, especially when $\mathcal{M}$ is given by $\mathcal{M}_{\mu,\rho}$. The following theorem, in this way, establishes the best-possible convergence rate for $m \in \mathcal{M}_{\mu,\rho}$, given some consistent estimates, $\widehat{h}_n$ and $\widehat{T}_n$.

Theorem 3.5 Assume that $\widehat{T}_n$ converges pointwise to T in $L^2(\mathcal{X})$, and that $\{\delta_k^2/\rho^2\}_{k \geq N}^{\infty} \in \sigma((T^*T)^{1+2\mu})$, for some $N \geq 1$, where $\sigma(T^*T)$ denotes the spectrum of the self-adjoint operator T^*T. Then,

$$\inf_{R \in \mathfrak{R}} \Xi(\{\delta_k\}, \mathcal{M}_{\mu,\rho}, R) = O_p(\delta_n^{\frac{2\mu}{2\mu+1}}). \qquad (19)$$

According to Theorem 3.5, the optimal bound is determined jointly by the composite error bound (δ_n), and the order of smoothness (μ). A faster convergence rate is possible, as both T and h are estimated with more accuracies, and m_0 becomes smoother (i.e., μ increases). The only difference of Theorem 3.5 from the result of Tautenhahn (1998) is the replacement of the error bound, $\overline{\delta}_n$ $= ||\widehat{h}_{0,n} - h_0||_{L^2(\mathcal{W})}$, by $\delta_n = ||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})}$. The extension is somewhat natural, since T also has to be estimated in the statistical inverse problem. It should be pointed out that (19) cannot be used in the same way as the usual statistical bounds, since optimality in (19) is only relative to the accuracies of $\widehat{h}_{0,n}$ and $\widehat{T}_n$. Without additional assumptions, it does not seem possible to tell which δ_n is minimal, while Stone (1982)'s bounds directly applies to $\overline{\delta}_n$. In this sense, we will call (19) as the *quasi-optimal bounds*. An important application of the quasi-optimal bounds concerns derivation of the actual convergence rate of a regularization method. In the rest of the paper, we will use δ_n to denote the convergence rate of the composite errors, given some preliminary estimates, $\widehat{h}_{0,n}$ and $\widehat{T}_n$; $||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} = O_p(\delta_n)$.

Remark 3.4 Let $\widehat{m}_{\alpha,n}$ be a regularized estimator of $m_0 \in \mathcal{R}(T^*)$, defined by (12), where $U_\alpha(\cdot)$ satisfies C.3.1 and C.3.2, with $\overline{\mu} \geq 1$. We suppose, as a side condition, that the given preliminary estimates, $\widehat{h}_{0,n}$ and $\widehat{T}_n$, satisfy

$$||\widehat{T}_n^* - T^*||_{L^2(\mathcal{X}) \to L^2(\mathcal{W})} \leq C||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})}^{1/2}.$$

In section 5, we will show that no strong restrictions are imposed by the side condition. From Theorem 3.5 and Corollary 3.4(i), we obtain both lower and upper bounds on the convergence rate of $\widehat{m}_{\alpha,n}$;

$$O_p(\delta_n^{1/2}) \leq ||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} \leq O_p(\delta_n/\sqrt{\alpha}) + O_p(\sqrt{\alpha}) + O_p(\delta_n^{1/2}).$$

If we choose a regularization parameter such that $\alpha = \alpha_n^* = C\delta_n$, the actual convergence rate of $\widehat{m}_{\alpha,n}$ is given by $\delta_n^{1/2}$. In other words, the above lower bound, which is in fact sharp, attains quasi-optimality in (19), given the side condition and $\alpha = \alpha_n^*$. ■

Remark 3.5 For derivation of the optimal bound, we do not assume special properties of the estimates for $\widehat{h}_n$ or $\widehat{T}_n$, except that the sequence, $\{||\widehat{h}_n - \widehat{T}_n m||_{L^2(\mathcal{W})}^2\}$, lies in the spectrum of the operator, $(T^*T)^{1+2\mu}$. Such assumption does not seem so strong in general, since zero is always an accumulation point in the spectrum of T^*T, when T has a non-closed range. If there is additional information about the preliminary estimates, or the side condition of Theorem 3.5 is violated, one may possibly get a faster convergence rate. ■

4 Optimal Convergence Rates of Various Regularization Methods

While Theorem 3.3 sheds light on the asymptotic properties of the general regularization methods, it is not clear how those results extend to a more general case of $\mu > 0$. The main difficulties are involved with finding a sharp bound of the term, $||(\widehat{T}_n^*\widehat{T}_n)^\mu - (T^*T)^\mu||_{L^2(\mathcal{X}) \to L^2(\mathcal{X})}$.[11] In this section,

[11] By Vainikko and Veretennikov (1986), an obvious bound is available;

$$||(\widehat{T}_n^*\widehat{T}_n)^\mu - (T^*T)^\mu||_{L^2(\mathcal{X}) \to L^2(\mathcal{X})} \leq C \max\{||\widehat{T}_n^* - T^*||_{L^2(\mathcal{X}) \to L^2(\mathcal{W})}^{\min(\mu,1)}, ||\widehat{T}_n - T||_{L^2(\mathcal{X}) \to L^2(\mathcal{W})}^{\min(\mu,1)}\}, \text{ a.s.}$$

Unfortunately, the resulting bound ends up only with a quite slower convergence rate, when μ is arbitrarily close to zero.

we use an alternative decomposition of $\widehat{b}_\alpha$, to derive the convergence rates for $m_0 \in \mathcal{M}_\mu$ (with $\mu > 0$);

$$\widehat{b}_\alpha = b_{1\alpha} + \widehat{b}_{2\alpha} \equiv (\Gamma_\alpha - I)(T^*T)^\mu m_\mu + (\widehat{\Gamma}_\alpha - \Gamma_\alpha)(T^*T)^\mu m_\mu, \tag{20}$$

where $\Gamma_\alpha = U_\alpha(T^*T)T^*T$. The first term ($b_{1\alpha}$) stands for a (deterministic) pure regularization bias, whose asymptotic behaviors have been analyzed in detail by mathematical regularization theory. Under C.3.2, the decaying rate of $b_{1\alpha}$ is the same as (14), from Lemma 3.3. The remaining error term ($\widehat{b}_{2\alpha}$), specific to statistical inverse problems, arises from use of estimated operators. Asymptotic properties of $\widehat{b}_{2\alpha}$, in general, depend on a particular shape of $U_\alpha(\cdot)$ as well as given estimates of T. In mathematics literature, various regularization methods have been suggested, that satisfy the conditions in C.3.1 and C.3.2. We select some of popular methods that are different in qualification, and show how special features of $U_\alpha(\cdot)$ affect the statistical properties of $\widehat{R}_{\alpha,n}$.

Ordinary Tikhonov Method With a choice of $U_{1,\alpha}(\lambda) = (\alpha+\lambda)^{-1}$, (12) leads to the ordinary Tikhonov regularization method (OTR) such that

$$\widehat{m}_{1,\alpha} = \widehat{R}_{1,\alpha}\widehat{h}_{0,n} = U_{1,\alpha}(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{h}_{0,n} = (\alpha I + \widehat{T}_n^*\widehat{T}_n)^{-1}\widehat{T}_n^*\widehat{h}_{0,n}, \tag{21}$$

By applying differential calculus in Hilbert space, one can show that $\widehat{m}_{1,\alpha}$ is a unique minimizer of the Tikhonov functional, i.e.,

$$\widehat{m}_{1,\alpha} = \underset{m(\cdot)\in L^2(\mathcal{X})}{\arg\min} \; ||\widehat{T}_n m - \widehat{h}_{0,n}||^2_{L^2(\mathcal{W})} + \alpha||m||^2_{L^2(\mathcal{X})}, \tag{22}$$

see Tikhonov and Arsenin (1977). OTR cures for instability of the generalized inverse, via penalization of (7), comparing to constrained minimum-distance in the compactification method.[12] If the constraint in (10) is specified in L^2-norm rather than the Sobolev norm, both types of minimum-distance are in a dual relation. They will yield the same estimates, if the regularization parameter of OTR (α) is equal to the Lagrange multiplier implied by (10). It is straightforward to check that C.3.1 and C.3.2 are satisfied by $U_{1,\alpha}(\cdot)$. The latter condition holds for $\mu \leq 1$, but not for $\mu > 1$; namely, the qualification of OTR is $\overline{\mu}_{OTR} = 1$. Consistency of OTR is obvious from Theorem 3.2. For a limited case of μ (equal to 1/2 or 1), the convergence rate of OTR is also available by Theorem 3.3. The following theorem, coinciding with Theorem 3.3 in the limited case, shows how the latter theorem extends to a general value of μ (> 0), at least for OTR.

[12] In the classical papers on ill-posed problems, Tikhonov (1963) and Phillips (1962) used, as a penalty term, Sobolev norm of m and L^2-norm of its derivatives, instead of $||m||^2_{L^2(\mathcal{X})}$ in (22).

Theorem 4.1 (i) For $m_0 \in \mathcal{R}((T^*T)^\mu)$ with $\mu > 0$, it holds that

$$\begin{aligned}||\widehat{m}_{1,\alpha} - m_0||_{L^2(\mathcal{X})} \;\leq\; & \frac{C_1}{\sqrt{\alpha}}||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} + C_2\alpha^{\min(\mu,1)}||\overline{m}_\mu||_{L^2(\mathcal{X})} \\ & + C_3\alpha^{\min(\mu-1/2,1/2)}||(\widehat{T}_n - T)\overline{m}_{\alpha,\mu}||_{L^2(\mathcal{W})} + C_4\alpha^{\min(\mu-1/2,0)}||(\widehat{T}_n^* - T^*)\overline{h}_{\alpha,\mu}||_{L^2(\mathcal{X})}, \quad \text{a.s.}\end{aligned}$$

where $\overline{m}_\mu = (T^*T)^{\max(\mu-1,0)}m_\mu$, $\overline{m}_{\alpha,\mu} = \alpha^{\min(1-\mu,0)}(\alpha I + T^*T)^{-1}(T^*T)^\mu m_\mu \in \mathcal{M}_{\max(\mu-1,0)}$, and $\overline{h}_{\alpha,\mu}$ $= \alpha^{\min(1/2-\mu,0)}T(\alpha I + T^*T)^{-1}(T^*T)^\mu m_\mu \in T^{*-1}(\mathcal{M}_{\max(\mu,1/2)})$.

(ii) For $m_0 \in \mathcal{M}_{\mu,\rho}$, with $\mu > 0$, it holds that

$$\begin{aligned}||\widehat{m}_{1,\alpha} - m_0||_{L^2(\mathcal{X})} \;\leq\; & C\{\frac{1}{\sqrt{\alpha}}||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} + \alpha^{\min(\mu,1)} + \alpha^{\min(\mu-1/2,1/2)}||\widehat{T}_n - T||_{\mathcal{M}_{\max(\mu-1,0)}\to L^2(\mathcal{W})} \\ & + \alpha^{\min(\mu-1/2,0)}||\widehat{T}_n^* - T^*||_{T^{*-1}(\mathcal{M}_{\max(\mu,1/2)})\to L^2(\mathcal{X})}\}, \quad \text{a.s.} \qquad (23)\end{aligned}$$

Owing to the unit qualification ($\overline{\mu}_{OTR} = 1$), the decaying rate of the regularization bias of OTR ($b_{1\alpha}$), which is given in the second term of (23), cannot be faster than α, regardless of smoothness of m_0. Similar saturation effects take place in the error term corresponding to $\widehat{b}_{2\alpha}$. Its relevant bounds, given in the last two terms of (23), cannot be improved beyond the benchmark case of $\mu = 1$. Applying an argument used in Remark 3.4, we can show that the lower bounds in Theorem 4.1 leads to the actual convergence rate of OTR, in some cases. Let $\mu_q = \min\{\mu, q\}$ and $\mu_q^\dagger = \max(\mu_q, 1/2)$, where q is a positive integer.

Remark 4.1 (*Suboptimality of OTR*) (a) Assume a side condition such that $\max\{||\widehat{T}_n - T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})}, ||\widehat{T}_n^* - T^*||_{L^2(\mathcal{W})\to L^2(\mathcal{X})}\} \leq O_p(\delta_n^{2\mu_1^\dagger/(2\mu_1+1)})$. If we choose a regularization parameter, $\alpha = \alpha_n^* = C\delta_n^{2/(2\mu_1+1)}$ so that $\alpha^{-1/2}||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} \simeq \alpha^{\mu_1}$, then, the last two terms in (23) are of order, not greater than $O_p(\alpha^{\mu_1})$. Consequently, by Theorem 3.5 and Theorem 4.1, we have, for $m_0 \in \mathcal{M}_\mu$ ($\mu > 0$), that

$$O_p(\delta_n^{\frac{2\mu}{2\mu+1}}) \leq ||\widehat{m}_{1,\alpha} - m_0||_{L^2(\mathcal{X})} \leq O_p(\delta_n^{\frac{2\mu_1}{2\mu_1+1}}).$$

The given choice of $\alpha = \alpha^*(n)$ ensures quasi-optimality of $\widehat{m}_{1,\alpha}$, for $m_0 \in \mathcal{M}_\mu$ with $\mu \leq 1$, but not for $\mu > 1$. Due to early saturation, the optimal bounds $\delta_n^{2\mu/(2\mu+1)}$, with $\mu > 1$, are in facts not attainable by OTR with any choice of $\alpha = \alpha(n)$, for a similar reason in Groetsch (1983, Proposition 2.2).

(b) If we note that constrained minimum-distance in (10) is dual to OTR (with penalization by the Sobolev norm), suboptimality of the compactification method can be understood in a similar way. ■

Iterated Tikhonov Regularization A direct improvement upon OTR can be made by bias-reduction in $\widehat{m}_{1,\alpha}$. Noting that the regularization bias of OTR ($\widehat{b}_\alpha$) can be estimated consistently by $(\widehat{\Gamma}_\alpha - I)\widehat{m}_{1,\alpha}$, a bias-corrected version of $\widehat{m}_{1,\alpha}$ is given by $\widehat{m}_{2,\alpha} = (I + \widehat{E}_\alpha)\widehat{m}_{1,\alpha}$, where $\widehat{E}_\alpha = I - \widehat{\Gamma}_\alpha = \alpha(\alpha I + \widehat{T}_n^*\widehat{T}_n)^{-1}$. In form of (12), the estimator is written as $\widehat{m}_{2,\alpha} = \widehat{R}_{2,\alpha}\widehat{h}_{0,n} = U_{2,\alpha}(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{h}_{0,n}$, where $U_{2,\alpha}(\lambda) = [1+\alpha(\alpha+\lambda)^{-1}](\alpha+\lambda)^{-1}$. Rewriting $U_{2,\alpha}(\lambda)$ as $[(\lambda+\alpha)^2-\alpha^2]/[\lambda(\lambda+\alpha)^2]$, it is easy to check that C.3.1 as well as C.3.2 (with $\overline{\mu} = 2$) hold. By means of the larger qualification of $\widehat{R}_{2,\alpha}$, the pure regularization bias of $\widehat{m}_{2,n}^\alpha$ decays at the rate of $O(\alpha^{\min(\mu,2)})$, leading to a faster convergence rate than that of OTR, for $m_0 \in \mathcal{R}((T^*T)^\mu)$ with $\mu > 1$. For $\mu > 2$, further improvements are possible, by applying a similar argument repeatedly. Letting $U_{q,\alpha}(\lambda) = [(\lambda+\alpha)^q - \alpha^q]/[\lambda(\lambda+\alpha)^q]$, we define the iterated Tikhonov regularization of order q (hereafter, ITR(q)) by

$$\widehat{m}_{q,\alpha} = \widehat{R}_{q,\alpha}\widehat{h}_{0,n} = U_{q,\alpha}(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{h}_{0,n} = \sum_{j=1}^{q}\widehat{E}_\alpha^{j-1}(\alpha I + \widehat{T}_n^*\widehat{T}_n)^{-1}\widehat{T}_n^*\widehat{h}_{0,n},$$

where the last equality comes from $U_{q,\alpha}(\lambda) = \sum_{j=1}^q(\frac{\alpha}{\alpha+\lambda})^{j-1}(\frac{1}{\alpha+\lambda})$. Straightforward calculations show that both C.3.1 and C.3.2 are satisfied by $U_{q,\alpha}(\cdot)$, with the qualification of ITR(q) equal to q. In an alternative way, $\widehat{m}_{q,\alpha}$ can be induced from an iterative procedure

$$(\alpha I + \widehat{T}_n^*\widehat{T}_n)\widehat{m}_{q,\alpha} = \widehat{T}_n^*\widehat{h}_{0,n} + \alpha\widehat{m}_{q-1,\alpha}, \text{ with } \widehat{m}_{0,\alpha} = 0. \tag{24}$$

The initial condition shows that OTR is equivalent to ITR(1). In (24), each step of iteration requires the same operator to be inverted, and thus the computational costs of ITR(q) is almost the same as that of OTR. For a variational characterization of $\widehat{m}_{q,\alpha}$, we remark that (24) is the normal equation of the penalized minimum-distance

$$\min_{m\in L^2(\mathcal{X})} ||\widehat{T}_n m - \widehat{h}_{0,n}||^2_{L^2(\mathcal{W})} + \alpha||m - \widehat{m}_{q-1,\alpha}||^2_{L^2(\mathcal{X})}.$$

When T is known, the asymptotic properties of ITR are studied by King and Chillingworth (1979) and Engl (1987). The following theorem gives an extension to a statistical inverse problem.

Theorem 4.2 For $m_0 \in \mathcal{M}_{\mu,\rho}$, with $\mu > 0$, it holds that

$$\begin{aligned} ||\widehat{m}_{q,\alpha} - m_0||_{L^2(\mathcal{X})} \leq\ & C\{\frac{1}{\sqrt{\alpha}}||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} + \alpha^{\min(\mu,q)} + \alpha^{\min(\mu-1/2,1/2)}||\widehat{T}_n - T||_{\mathcal{M}_{\max(\mu-1,0)}\to L^2(\mathcal{W})} \\ & +\alpha^{\min(\mu-1/2,0)}||\widehat{T}_n^* - T^*||_{T^{*-1}(\mathcal{M}_{\max(\mu,1/2)})\to L^2(\mathcal{X})}\}, \quad \text{a.s.} \end{aligned}$$

where q is any (finite) positive integer.

The only difference of Theorem 4.2 from Theorem 4.1 lies in the faster convergence rate of the pure regularization bias (the second term in the above bound), improving upon OTR for $\mu > 1$.

Remark 4.2 For $m_0 \in \mathcal{M}_\mu$ with $\mu \leq q$, the quasi-optimality of ITR(q) is proved in the same way as Remark 4.1. Here, the relevant side condition to be assumed is $\max\{||\widehat{T}_n - T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})}, ||\widehat{T}_n^* - T^*||_{L^2(\mathcal{W})\to L^2(\mathcal{X})}\} \leq O_p(\delta_n^{2\mu_q^\dagger/(2\mu_q+1)})$. Under a choice of $\alpha = \alpha_n^* \simeq \delta_n^{2/(2\mu_q+1)}$, the actual convergence rate of ITR(q) is quasi-optimal, i.e., $||\widehat{m}_{q,\alpha} - m_0||_{L^2(\mathcal{X})} = O_p(\delta_n^{2\mu_q/(2\mu_q+1)})$. ■

Generalized Tikhonov Regularization Another extension of OTR has been suggested, by Plato and Vainikko (1990) and Tautenhahn (1998), to overcome a disadvantage due to early saturation. As in ITR, their method generalizes OTR by choosing an alternative penalty term, but the motivation is rather different. Suppose that the true solution is known to be sufficiently smooth, say, $m \in \mathcal{R}((T^*T)^\mu)$ with $\mu \geq (q-1)/2$, for positive integer q. Then, one may try to penalize variability of m through $||(T^*T)^{-(q-1)/2}m||_{L^2(\mathcal{X})}$, instead of the standard L^2-norm of m. From T being an integral operator, $(T^*T)^{-(q-1)/2}$ behaves like a differential operator, implying that $||(T^*T)^{-(q-1)/2}m||_{L^2(\mathcal{X})}$ serves as L^2-norm of a generalized derivative of m. The differential norm will be useful, especially for control over highly-oscillating behaviors of a function, just like the Sobolev norm. Those considerations give rise to the generalized Tikhonov regularization method of order q (hereafter, GTR(q)), defined as

$$\widehat{m}_{q,\alpha}^g = \min_{m \in L^2(\mathcal{X})} ||\widehat{T}_n m - \widehat{h}_{0,n}||_{L^2(\mathcal{W})}^2 + \alpha^q ||(\widehat{T}_n^* \widehat{T}_n)^{-(q-1)/2} m||_{L^2(\mathcal{X})}^2, \quad \text{for } q \geq 1.$$

Applying differential calculus, we can show that $\widehat{m}_{q,\alpha}^g$ is the solution to the normal equation, $\widehat{T}_n^*(\widehat{T}_n \widehat{m}_{q,\alpha}^g - \widehat{h}_{0,n}) + \alpha^q (\widehat{T}_n^* \widehat{T}_n)^{-(q-1)} \widehat{m}_{q,\alpha}^g = 0$. Hence, using $U_{q,\alpha}^g(\lambda) = (\alpha^q + \lambda^q)^{-1} \lambda^{(q-1)}$, we can represent GTR$(q)$ in form of (12)

$$\widehat{m}_{q,\alpha}^g = \widehat{R}_{q,\alpha}^g \widehat{h}_{0,n} = U_{q,\alpha}^g (\widehat{T}_n^* \widehat{T}_n) \widehat{T}_n^* \widehat{h}_{0,n} = [\alpha^q I + (\widehat{T}_n^* \widehat{T}_n)^q]^{-1} (\widehat{T}_n^* \widehat{T}_n)^{(q-1)} \widehat{T}_n^* \widehat{h}_{0,n}.$$

Obviously, GTR(1) reduces to OTR. All the conditions in C.3.1 and C.3.2 are satisfied by $U_{q,\alpha}(\cdot)$, with the qualification of GTR(q) equal to q. The theorem below shows that the convergence rate of GTR(q) is the same as that of ITR(q). Following Remark 4.2, we also can establish the quasi-optimality of GTR(q), for $m_0 \in \mathcal{M}_{\mu,\rho}$ with $\mu \leq q$.

Theorem 4.3 For $m_0 \in \mathcal{M}_{\mu,\rho}$, (with $\mu > 0$), the same bounds as in Theorem 4.2, apply to $\widehat{m}_{q,\alpha}^g$.

Showalter's Regularization The analyses so far have been confined to regularization methods of finite qualification. To give an example of infinite-qualification regularization, we consider Showalter's integral formula for the generalized inverse of $\widehat{T}_n$

$$\widehat{T}_n^\dagger = \int_0^\infty \exp(-s\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^* ds.$$

Showalter (1967) showed that the above equality holds precisely on the domain of $\widehat{T}_n^\dagger$, which, by the argument above Remark 3.1, is equal to $L^2(\mathcal{W})$, for any finite n. A bounded approximation of $\widehat{T}_n^\dagger$ is obtained by replacing the infinite interval of integral by a finite one, say, $[0, 1/\alpha]$. Using

$$U_\alpha^s(\lambda) = \int_0^{1/\alpha} \exp(-s\lambda)ds = \begin{cases} \lambda^{-1}[1-\exp(-\frac{\lambda}{\alpha})], & \text{for } \lambda > 0, \\ \quad\alpha^{-1} & \text{otherwise} \end{cases}, \tag{25}$$

we define Showalter's regularization (SW) by

$$\widehat{m}_{\alpha,n}^s = \widehat{R}_\alpha^s \widehat{h}_{0,n} = U_\alpha^s(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{h}_{0,n} = [\int_0^{1/\alpha} \exp(-s\widehat{T}_n^*\widehat{T}_n)ds]\widehat{T}_n^*\widehat{h}_{0,n}.$$

From $\sup_{x>0} x^{-1}[1-\exp(-x)] \leq 1$, C.3.1(i) holds for $U_\alpha^s(\cdot)$. The rest of conditions of C.3.1 and C.3.2 follow from $\sup_{x\geq 0} \exp(-x)x^\mu \leq e^{-\mu}\mu^\mu$, and $\lambda^\mu|U_\alpha^s(\lambda)\lambda - 1| = \alpha^\mu \exp(-\frac{\lambda}{\alpha})\left(\frac{\lambda}{\alpha}\right)^\mu$, for *any* $\mu > 0$. The latter condition implies that the qualification of SW is infinite. For mathematical inverse problems with T known, the convergence rate of Showalter's regularization was studied by Schock (1985) and Engl and Gfrerer (1988). Below we extend those results to a stochastic case.

Theorem 4.4 For $m_0 \in \mathcal{M}_{\mu,\rho}$, with $\mu > 0$, it holds that,

$$\begin{aligned} ||\widehat{m}_{\alpha,n}^s - m_0||_{L^2(\mathcal{X})} \quad \leq \quad & C\{\frac{1}{\sqrt{\alpha}}||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} + \alpha^\mu \\ & +\alpha^{\mu-1/2}||\widehat{T}_n - T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})} + \alpha^{\mu-1/2}||\widehat{T}_n^* - T^*||_{L^2(\mathcal{W})\to L^2(\mathcal{X})}\}, \quad \text{a.s.}(26) \end{aligned}$$

By means of the infinite qualification, Showalter's method does not suffer from any saturation effects, accounting for simplicity of the convergence rate in (26), which is free of other nature of the regularization scheme. The second term in (26) indicates that smoothness of m_0 is sufficient to determine the decaying rate of the pure regularization bias. The last two terms reflect additional gains of SW, by sharpening the corresponding bounds in the previous theorems. For example, when $\mu > 1/2$, the last term in (26) is of smaller order than those in Theorem 3.3(b) and Theorem 4.1 through 4.3, since, in that case, the former decays at the rate of $o_{a.s}(||\widehat{T}_n^* - T^*||_{L^2(\mathcal{W})\to L^2(\mathcal{X})})$.

Remark 4.3 Advantages of SW over other methods are highlighted in attaining the optimal bounds in (19), for an arbitrary order of smoothness in m_0. Under a simple side condition such that $\max\{||\widehat{T}_n - T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})}, ||\widehat{T}_n^* - T^*||_{L^2(\mathcal{W})\to L^2(\mathcal{X})}\} \leq O_p(\delta_n^{1/(2\mu+1)})$, it follows from Theorem 3.5 and 4.4 that

$$O_p(\delta_n^{2\mu/(2\mu+1)}) \leq ||\widehat{m}_{\alpha,n}^s - m_0||_{L^2(\mathcal{X})} \leq O_p(\delta_n/\sqrt{\alpha}) + O_p(\alpha^\mu) + O_p(\alpha^{\mu-1/2}\delta_n^{1/(2\mu+1)}).$$

For $\alpha = \alpha_n^* \simeq \delta_n^{\frac{2}{2\mu+1}}$, the actual convergence rate of SW is given by $||\widehat{m}_{\alpha,n}^s - m_0||_{L^2(\mathcal{X})} = O_p(\delta_n^{2\mu/(2\mu+1)})$, which ensures quasi-optimality of $\widehat{m}_{\alpha,n}^s$, for any $\mu > 0$. Note that the necessary side condition is weaker than the previous ones. ■

5 Nonparametric Kernel IV Estimates

Various types of regularized estimates are conceivable, according to different nonparametric procedures for estimating h_0 and T. As a preeminent example, this section applies a kernel smoothing method to obtain the preliminary estimates. A general class of regularized kernel estimators for m_0 then follow from (12). Those estimators include, as a special case, the kernel estimator (regularized by OTR) in Darolles, Florens, and Renault (2001), although the latter depends on a slightly different definition for h_0 and T. Their estimator, lacking an exact closed form, can be computed only approximately, via an additional discretization method, such as the collocation method in Remark 3.1. A lower bound on the convergence rate was shown for the estimator, under a simplifying condition on the bandwidth parameters. Such a lower bound, however, turns out to be too rough to evaluate the actual convergence rate, not allowing for an optimal choice of bandwidth and regularization parameters. Moreover, early saturation of OTR prevents their estimator attaining the optimality bounds in (19), for relatively smooth functions. In this section, we develop more advanced results for kernel IV estimation. Using spectral theory for compact self-adjoint operators, we figure out the closed form of the kernel estimator which is defined by the general regularization method in (12). Consistency as well as the 'actual' convergence rates of those estimators are shown by applying the statistical results in section 3 or 4. Quasi-optimal bounds play a crucial role in our developments for the optimal choice of smoothing parameters.

5.1 Closed Form of Kernel IV Estimates

We start with a case where there is no common element between X and W. Assume that the underlying structural function is identified by $m_0 = T^{-1}h_0$, where $h_0(w) = \int_{\mathcal{Y}} y f_{Y,W}(y,w)dy$, and $T : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ is given by $(Tm)(w) = \int_{\mathcal{X}} m(x) f_{X,W}(x,w)dx$. Let $\widehat{f}_{Y,W,n}(\cdot,\cdot)$ and $\widehat{f}_{X,W,n}(\cdot,\cdot)$ be a typical kernel estimator for $f_{Y,W}(\cdot,\cdot)$ and $f_{X,W}(\cdot,\cdot)$, respectively, from the observations $\{(Y_i, Z_i, W_i)\}_{i=1}^n$

$$\begin{aligned}
\widehat{f}_{Y,W,n}(y,w) &= n^{-1}\sum_{i=1}^n K_{g_0}(y_i - y)K_{g_2}(W_i - w),\\
\widehat{f}_{X,W,n}(x,w) &= n^{-1}\sum_{i=1}^n K_{g_1}(X_i - x)K_{g_2}(W_i - w),
\end{aligned}$$

where $K_g(s) = \Pi_{r=1}^d \frac{1}{g}K(s/g)$, with $K(\cdot)$ being a symmetric function defined on the real line, and $d = \dim(s)$. The preliminary estimates for h_0 and $T : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ are defined by

$$\widehat{h}_{0,n}(w) = \int_{\mathcal{Y}} y \widehat{f}_{Y,W,n}(y,w)dy = n^{-1}\sum_{i=1}^n K_{g_2}(W_i - w)Y_i, \tag{27}$$

and $\widehat{T}_n : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ such that

$$(\widehat{T}_n m)(w) = \int \widehat{f}_{X,W,n}(x,w)m(x)dx = \int \left[n^{-1}\sum_{i=1}^n K_{g_1}(X_i - x)K_{g_2}(W_i - w)\right] m(x)dx. \tag{28}$$

Also, define $\widehat{T}_n^* : L^2(\mathcal{W}) \to L^2(\mathcal{X})$, by

$$(\widehat{T}_n^* h)(x) = \int \widehat{f}_{X,W,n}(x,w)h(w)dw = \int \left[n^{-1}\sum_{i=1}^n K_{g_1}(X_i - x)K_{g_2}(W_i - w)\right] h(w)dw.$$

By Fubini's Theorem, $< \widehat{T}_n m, h >_{L^2(\mathcal{W})} = < m, \widehat{T}_n^* h >_{L^2(\mathcal{X})}$, a.s.; two random operators, $\widehat{T}_n$ and $\widehat{T}_n^*$, are adjoint to each other. The integral operator $\widehat{T}_n$ has a degenerate kernel, i.e., $\widehat{f}_{X,W,n}(\cdot,\cdot)$ is a finite sum of products of kernel weights on each observation (X_i, W_i). Thus, $\widehat{T}_n$ has a finite rank, with $\dim(\mathcal{R}(\widehat{T}_n)) \leq n$, from which follow boundedness as well as compactness, of $\widehat{T}_n$ and the self-adjoint operator $\widehat{T}_n^*\widehat{T}_n : L^2(\mathcal{X}) \to L^2(\mathcal{X})$.[13] Applying spectral theory for compact self-adjoint operators, a regularized kernel estimator of m_0 is now well defined by (12), with $\widehat{T}_n$ and $\widehat{h}_{0,n}$ given as above, as long as $U_\alpha(\cdot)$ is defined on a bounded interval, $(0, \overline{\lambda}]$, where $\overline{\lambda} = \sup_{n \geq n_0} ||\widehat{T}_n^*\widehat{T}_n||_{L^2(\mathcal{X}) \to L^2(\mathcal{X})}$.

[13] Namely, $||\widehat{T}_n||_{L^2(\mathcal{X}) \to L^2(\mathcal{W})} \overset{a.s.}{\leq} B_n$, for some $B_n < \infty$, and for $\mathcal{M}_\rho = \{m \in L^2(\mathcal{X}) : ||m||_{L^2(\mathcal{X})} \leq \rho\}$, $\widehat{T}_n(\mathcal{M}_\rho)$ is compact in $L^2(\mathcal{W})$, a.s.

To show the closed form of the kernel IV estimator, we need the following definitions. Letting $K_n^X(x) = [K_{g_1}(X_1 - x), .., K_{g_1}(X_n - x)]'$, and $K_n^W(w) = [K_{g_2}(W_1 - w), .., K_{g_1}(W_n - w)]'$, we define

$$M_X = \int_{\mathcal{X}} K_n^X(x) K_n^X(x)' dx, \text{ and } M_W = \int_{\mathcal{W}} K_n^W(w) K_n^W(w)' dw.$$

Using integration-by-substitution, the (i, j)-th element of M_W, for example, is written more compactly, via a convolution-kernel function, as

$$M_{ij}^W = \int_{\mathcal{W}} K_{g_2}(W_i - w) K_{g_2}(W_j - w) dw = K_{g_2}^c(W_i - W_j),$$

where $K_{g_2}^c(w) = (1/g_2) \int_{\mathcal{W}} K(w/g_2 - s)K(s)ds$. A straightforward calculation shows that M_W is a $(n \times n)$ symmetric nonnegative semi-definite matrix, for which the square-root matrix $M_W^{1/2}$ is well-defined, satisfying $M_W = M_W^{1/2} M_W^{1/2}$.[14] Letting $Q_{X,W} = n^{-2} M_W^{1/2} M_X M_W^{1/2}$, $Q_{X,W}$ is also a $(n \times n)$ symmetric nonnegative semi-definite matrix, whose eigenvalues are all real and positive. We denote, by $\lambda_{\max}(Q_{X,W})$, the maximum of those eigenvalues.

Theorem 5.1 Let $\widehat{h}_{0,n}$ and $\widehat{T}_n$ be defined by (27) and (28), respectively, and $\widehat{T}_n^*$ be the adjoint of $\widehat{T}_n$. Assume that $U_\alpha(\cdot)$ is any real-valued function defined on a bounded interval, $(0, \overline{\lambda}]$ where $\overline{\lambda}$ $\geq \sup_{n \geq n_0} \lambda_{\max}(Q_{X,W})$. Then, for any $n \geq n_0$,

$$\widehat{m}_{\alpha,n}(x) = [U_\alpha(\widehat{T}_n^* \widehat{T}_n)(\widehat{T}_n^* \widehat{h}_{0,n})](x) = n^{-2} K_n^X(x)' M_W^{1/2} U_\alpha(Q_{X,W}) M_W^{1/2} \mathbf{y}, \tag{29}$$

where $\mathbf{y} = (Y_1, ..Y_n)'$.

By Theorem 5.1, the abstract operator-form of the kernel IV estimator translates into a concrete matrix-form. With $U_\alpha(Q_{X,W})$ calculated by the standard eigenvalues decomposition, computations of $\widehat{m}_{\alpha,n}$ only involve simple operation of finite-dimensional matrices, when the convolution-kernel weights in M_X and M_W are given. For example, the kernel IV estimates, regularized by Showalter's method, are computed by

$$\widehat{m}_{\alpha,n}^s = [\int_0^{1/\alpha} \exp(-s\widehat{T}_n^* \widehat{T}_n)ds] \widehat{T}_n^* \widehat{h}_{0,n} = n^{-2} K_n^X(x)' M_W^{1/2} U_\alpha^s(Q_{X,W}) M_W^{1/2} \mathbf{y},$$

with $U_\alpha^s(Q_{X,W}) = F_n U_\alpha^s(\Lambda_n) F_n'$, where Λ_n is a diagonal matrix consisting of eigenvalues of $Q_{X,W}$, and F_n is a matrix of corresponding eigenvectors.[15]

[14] $a' M_W W a = \sum_{1 \leq i,,j \leq n} a_i M_{ij}^W a_j = \int_{\mathcal{W}} \left[\sum_{i=1}^n a_i K_h(W_i - w)\right]^2 dw \geq 0$, for any $a (\neq 0) \in \mathbb{R}^n$. For positive-definiteness of M_W, it suffices to assume that $\{K_h(W_i - \cdot)\}_{i=1}^n$ is linearly independent.

[15] Letting $\lambda_{i,n}$ be the i-th eigenvalue of $Q_{X,W}$, $[U_\alpha^s(\Lambda_n)]_{(i,i)}$ is equal to $\lambda_{i,n}^{-1}[1 - \exp(-\lambda_{i,n}/\alpha)]$, for $\lambda_{i,n} > 0$, and equal to α^{-1}, otherwise; see (25) in section 4.

Remark 5.1. (i) Suppose that $K(\cdot)$ is a density function from a stable distribution, say, a gaussian kernel. Then, a further simplification of the convolution-kernel weight is available;

$$M_{ij}^W = K_{g_2}^c(W_i - W_j) = K_{\sqrt{2}g_2}(W_i - W_j),$$

from $K^c(s) = K(s/\sqrt{2})/\sqrt{2}$, since, by the stability assumption, the shape of a convoluted density function is not changed, except that the variance doubles. In that case, all the matrices in (29) are calculated in a straightforward way. In general, when there is no explicit form for the convolution kernel, we can compute $K^c(\cdot)$ by numerical integration.

(ii) By Theorem 5.1, the naive minimum-distance estimator in section 3.1 has a closed form

$$\widehat{m}_n^\dagger(x) = K_n^X(x)' M_W^{1/2}(M_W^{1/2} M_X M_W^{1/2})^\dagger M_W^{1/2}\mathbf{y}.$$

If both $K_n^X(\cdot)$ and $K_n^W(\cdot)$ are assumed to be linearly independent, then, M_W and M_X are positive definite, from which we get $\widehat{m}_n^\dagger(x) = K_n^X(x)' M_X^{-1}\mathbf{y}$. From

$$\begin{aligned}(\widehat{T}_n\widehat{m}_n^\dagger)(w) &= \int \widehat{f}_{X,W,n}(x,w)\widehat{m}_n^\dagger(x)dx = n^{-1}K_n^W(w)' < K_n^X(\cdot), K_n^{X\prime}(\cdot) >_{L^2(\mathcal{X})} M_X^{-1}\mathbf{y} \\ &= n^{-1}K_n^W(w)'\mathbf{y} = \widehat{h}_{0,n}(w),\end{aligned}$$

$\widehat{m}_n^\dagger(\cdot)$ is confirmed to be one of the exact solutions to the integral equation, $\widehat{T}_n m = \widehat{h}_{0,n}$, where $\widehat{T}_n$ is in general not invertible. By definition of the generalized inverse, $\widehat{m}_n^\dagger(\cdot)$ will be the solution of minimum-norm. Instability of $\widehat{m}_n^\dagger$ is obvious from the minimum eigenvalue of M_X converging to zero, as $n \to \infty$, since a pair of elements in $K_n^X(\cdot)$ should become arbitrarily close to each other.

(iii) In Darolles, Florens, and Renault (2001), an alternative kernel estimator of m_0 is defined, based on OTR, by $\widetilde{m}_{\alpha,n} = [\alpha I + \widetilde{T}_n^*\widetilde{T}_n]^{-1}\widetilde{T}_n^*\widetilde{h}_n$, where $(\widetilde{T}_n m)(w) = \int \widehat{f}_{X|W,n}(x|w)m(x)dx$, and $\widetilde{h}_n(w) = \int_{\mathcal{Y}} y\widehat{f}_{Y|W,n}(y|w)dy$. Unlike $\widehat{m}_{\alpha,n}$ in (29), their estimator does not possess an exact closed form.

(iv) Let $\widehat{f}_{X,W,n}^c(x,w) = n^{-1}\sum_{i=1}^n K_{g_1}(X_i - x)K_{g_2}^c(W_i - w)$, where $K^c(\cdot)$ be a convolution kernel function in the above. Denote, by $\widehat{f}_{X,W,n}^c(x,\mathbf{W})$, the column vector of the joint density estimates, $[\widehat{f}_{X,W,n}^c(x,W_1),..,\widehat{f}_{X,W,n}^c(x,W_n)]'$. In a matrix form, $\widehat{f}_{X,W,n}^{c\prime}(x,\mathbf{W}) = n^{-1}K_n^{X\prime}(x)M_W$. From $\widehat{T}_n^*\widehat{h}_n = n^{-2}K_n^{X\prime}(\cdot)M_W\mathbf{y} = n^{-1}\widehat{f}_{X,W,n}^{c\prime}(\cdot,\mathbf{W})\mathbf{y}$, we rewrite $\widehat{m}_{\alpha,n}(\cdot)$ in (29) as

$$[R_n^\alpha(\widehat{h}_{0,n})](x) = [U_\alpha(\widehat{T}_n^*\widehat{T}_n)(n^{-1}\widehat{f}_{X,W}^{c\prime}(\cdot,\mathbf{W})\mathbf{y})](x) = n^{-1}\sum_{i=1}^n \left[U_\alpha(\widehat{T}_n^*\widehat{T}_n)\widehat{f}_{XW}^c(\cdot,W_i)\right](x)y_i.$$

This shows that $\widehat{m}_{\alpha,n}(\cdot)$ includes, as a special case with $U_\alpha(\lambda) = (\alpha+\lambda)^{-1}$, the kernel estimator suggested by Hall and Horowitz (2003). ■

We turn to an extension to a common-element case, where $X = (Z, W_1)$ and $W = (W_1, W_2)$. Let m_0 be identified by $T^{-1}h_0$, where $h_0(w) = \int_{\mathcal{Y}} y f_{Y,W}(y, w)dy$, and $T : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ is given by $(Tm)(w) = \int_{\mathcal{Z}} m(z, w_1) f_{Z,W}(z, w)dz$. We will use the same estimate of h_0 as (27). Using

$$\widehat{f}_{Z,W_1,W_2}(z, w_1, w_2) = n^{-1} \sum_{i=1}^{n} K_{g_1}(Z_i - x)K_{g_2}(W_{1i} - w_1)K_{g_2}(W_{2i} - w_2),$$

the preliminary estimates of T and its adjoint are defined, in a similar way to (28), by

$$\begin{aligned} (\widehat{T}_n m)(w_1, w_2) &= \int_{\mathcal{Z}} m(z, w_1)\widehat{f}_{Z,W_1,W_2}(z, w_1, w_2)dz, \text{ and} \\ (\widehat{T}_n^* h)(z, w_1) &= \int_{\mathcal{W}_2} h(w_1, w_2)\widehat{f}_{Z,W_1,W_2}(z, w_1, w_2)dw_2, \end{aligned} \tag{30}$$

respectively. A regularized kernel estimator of m_0, in the presence of common elements between X and W, is defined by (12), with $\widehat{T}_n$ and $\widehat{T}_n^*$ are modified as above. Letting $K_n^X(z, w_1) = [K_{g_1}(Z_1 - z)K_{g_2}(W_{11} - w_1), .., K_{g_1}(Z_n - z)K_{g_2}(W_{1n} - w_1)]'$, and $K_n^{W_2}(w_2) = [K_{g_2}(W_{21} - w_2), .., K_{g_1}(W_{2n} - w_2)]'$, we define

$$M_{(Z,W_1)}(w_1) = \int_{\mathcal{Z}} K_n^X(z, w_1)K_n^X(z, w_1)'dz, \;\; M_{W_2} = \int_{\mathcal{W}_2} K_n^{W_2}(w_2)K_n^{W_2}(w_2)'dw_2,$$

and

$$Q_{Z,W}(w_1) = \; n^{-2}M_{W_2}^{1/2}M_{(Z,W_1)}(w_1)M_{W_2}^{1/2},$$

where $M_{W_2}^{1/2}$ is the square-root of M_{W_2}. With $A \odot B$ denoting the matrix Hadamard product (i.e., element-by-element multiplication), $M_{(Z,W_1)}(w_1)$ is equivalent to $[M_Z \odot \mathcal{K}_{W_1}(w_1)]$, where the (i, j)-th element of M_Z and $\mathcal{K}_{W_1}(w_1)$ is given by $M_{ij}^Z \;= K_{g_1}^c(Z_i - Z_j)$ and $\mathcal{K}_{ij}^{W_1}(w_1) = K_{g_2}(W_{1i} - W_{1j})$, respectively. Note that $Q_{Z,W}(w_1)$, a function of w_1, is symmetric and nonnegative semi-definite, for any $w_1 \in \mathcal{W}_1$.

Theorem 5.2 Let $\widehat{h}_{0,n}$ and $\widehat{T}_n$ be defined by (27) and (30), respectively, and $\widehat{T}_n^*$ be the adjoint of $\widehat{T}_n$. Assume that $U_\alpha(\cdot)$ is any real-valued function defined on a bounded interval, $(0, \overline{\lambda}]$ where $\overline{\lambda}$ $\geq \sup_{w_1 \in \mathcal{W}_1} \sup_{n \geq n_0} \lambda_{\max}(Q_{Z,W}(w_1))$. Then, for any $n \geq n_0$,

$$\widehat{m}_{\alpha,n}(z, w_1) = [U_\alpha(\widehat{T}_n^*\widehat{T}_n)(\widehat{T}_n^*\widehat{h}_{0,n})](z, w_1) = n^{-2}K_n^X(z, w_1)'M_{W_2}^{1/2}U_\alpha(Q_{Z,W}(w_1))M_{W_2}^{1/2}[K_n^{W_1}(w_1) \odot \mathbf{y}]. \tag{31}$$

Using the matrix Hadamard product, we may rewrite $\widehat{m}_{\alpha,n}(z, w_1)$ in Theorem 5.2 as

$$n^{-2}[K_n^Z(z) \odot K_n^{W_1}(w_1)]'M_{W_2}^{1/2}U_\alpha(n^{-2}M_{W_2}^{1/2}[M_Z \odot \{K_n^{W_1}(w_1)K_n^{W_1}(w_1)'\}]M_{W_2}^{1/2})M_{W_2}^{1/2}[K_n^{W_1}(w_1) \odot \mathbf{y}],$$

which shows how the kernel IV estimator in (29) is generalized by the presence of W_1, the common elements between X and W. No additional difficulties arise in computing $\widehat{m}_{\alpha,n}(z, w_1)$, compared to $\widehat{m}_{\alpha,n}(x)$ in Theorem 5.1.

5.2 Optimal Convergence Rates

We continue to analyze asymptotic properties of the kernel estimators in the previous section. Considering that (29) is a special case of (31), our asymptotic derivations will focus on a common-element case, as specified by the first condition below.

C.5.1 (a) The random vector (Y_i, Z_i, W_i) is independent and identically distributed, satisfying (1)-(3), with m_0 identified by $T^{-1}h_0$, where $h_0(w) = \int_{\mathcal{Y}} y f_{Y,W}(y, w)dy$, and the injective operator $T : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ is such that $(Tm)(w) = \int_{\mathcal{Z}} m(z, w_1) f_{Z,W}(z, w)dz$. We assume that $d_2 \geq d_1$. (b) $E(Y^2|W = w)$ is bounded uniformly in w, a.s.

C.5.2 Let $K(\cdot) \in \mathbb{K}_{p^*}$, where $\mathbb{K}_{p^*}$ is the class of all Borel measurable symmetric real-valued functions $K(s)$ such that (a)

$$\int |K(s)|ds < \infty, \quad \int K(s)ds = 1, \quad \int K^2(s)ds < \infty, \quad \sup |K(s)| < \infty,$$

and (b) $\int s^j K(s)ds = 0$, for $j = 1, .., p^* - 1$, and $\mu_{p^*}(K) = \int s^{p^*} K(s)ds < \infty$, where p^* is an even integer.

C.5.3 The joint density functions $f_{Z,W}(\cdot, \cdot)$ is square-integrable and bounded;

$$\int_{\mathcal{W}} \int_{\mathcal{Z}} f^2_{Z,W}(z, w)dzdw < \infty, \text{ and } \sup_{(z,w)\in\mathcal{Z}\times\mathcal{W}} f_{Z,W}(z, w) \leq C < \infty.$$

C.5.4 $f_{Z,W}(\cdot, \cdot)$ and $m_0(\cdot)$ have continuous p_0-th and p_1-th partial derivatives, respectively, that are square-integrable, where $p_0 \geq d_1/2$.

C.5.5 (a) The bandwidth parameters $(g_1,\, g_2)$ satisfy that $\max(g_1,\, g_2) \to 0$, $ng_2^{d_2} \to \infty$. (a) The regularization parameter α satisfies that $\alpha \to 0$, $ng_2^{d_2}\alpha \to \infty$, and $g_1^{p_0}/\sqrt{\alpha} \to 0$, as $n \to \infty$.

All the technical conditions in C.5.2 through C.5.4 are standard in nonparametric kernel estimation. The joint density function is not required to have a compact support, nor restricted to be bounded away from zero. The square-integrability condition in C.5.3 entails boundedness of the linear operator T. C.5.5(b), which is rather stronger than C.5.5(a), is necessary for consistency of the regularized kernel estimates. Let $\widehat{h}_{0,n}$ and $\widehat{T}_n$ be given by (27) and (30), respectively. Our first result concerns sufficiency of the above conditions for derivation of the basic properties of the preliminary estimates, including consistency and the convergence rates.

Proposition 5.3 Suppose that C.5.1 through C.5.3, and C.5.5(a) hold. Then,

(i) $\widehat{T}_n$ is uniformly consistent for T, i.e., $||\widehat{T}_n - T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})} \xrightarrow{p} 0$, as $n \to \infty$.

Assume additionally that C.5.4 holds and $K(\cdot) \in \mathbb{K}_{p^*}$, with $p^* \geq \overline{p} = \max(p_0, p_1)$. Then,

$$\begin{aligned} \text{(ii)}\ \ ||\widehat{T}_n - T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})} &= O_p(1/\sqrt{ng_2^{d_2}} + g_1^{p_0} + g_2^{p_0}\}), \\ ||\widehat{T}_n^* - T^*||_{L^2(\mathcal{W})\to L^2(\mathcal{X})} &= O_p(1/\sqrt{ng_1^{d_1}} + g_1^{p_0} + g_2^{p_0}\}), \quad \text{and} \end{aligned}$$

$$\text{(iii)}\ \ ||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} = O_p(1/\sqrt{ng_2^{d_2}} + g_1^{\overline{p}}).$$

Let $\widehat{m}_{\alpha,n}$ be the kernel estimates defined by (31). When $U_\alpha(\cdot)$ satisfies C.3.1 and 3.2, the asymptotic properties of the general kernel estimates can be shown from Proposition 5.3, applied to Theorem 3.2 and Theorem 3.3 (or Corollary 3.4).

Theorem 5.4 Assume that C.5.1 through C.5.5 hold, with $p_0 = p_1$, and $U_\alpha(\cdot)$ satisfies C.3.1. Then,

(i) $||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} \xrightarrow{p} 0$, as $n \to \infty$, for all $m_0 \in L^2(\mathcal{X})$.

Assume additionally that $U_\alpha(\cdot)$ satisfies C.3.2, with $\overline{\mu} \geq 1$. Then,

$$\text{(ii)}\ ||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} \leq O_p(\frac{1}{\sqrt{\alpha}}[\frac{1}{\sqrt{ng_2^{d_2}}} + g_1^{p_0}]) + O_p(\frac{1}{\sqrt{ng_1^{d_1}}} + g_2^{p_0}) + \begin{cases} O_p(\sqrt{\alpha}), & \text{for } m_0 \in \mathcal{M}_{1/2,\rho}, \\ O_p(\alpha), & \text{for } m_0 \in \mathcal{M}_{1,\rho}, \end{cases}$$

Using the argument in Remark 3.4, we can show that, under additional conditions on (g_1, g_2, α), the lower bounds in Theorem 5.4 gives rise to the actual convergence rates of $\widehat{m}_{\alpha,n}$. Throughout this section, a vector of smoothing parameters (g_1, g_2, α) is called quasi-optimal, if it allows for $\widehat{m}_{\alpha,n}$ in (31) to attain the bounds in (19).

Theorem 5.5 Assume that C.5.1 through C.5.5 hold, with $p_0 = p_1$, and $U_\alpha(\cdot)$ satisfies C.3.1 and C.3.2, with $\overline{\mu} \geq 1$.

(i) Let m_0 be any function in $\mathcal{M}_{1/2,\rho}$. Suppose that the bandwidth parameters (g_1, g_2) satisfy a side condition such that $(ng_1^{d_1})^{-1/2} \leq O(g_1^{p_0/2})$, and $g_2^{2p_0} \leq O([ng_2^{d_2}]^{-1/2})$. Then, the optimal convergence rate of $\widehat{m}_{\alpha,n}$ is given by $||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} = O_p(n^{-\frac{p_0}{4p_0+d_2}})$, under the of smoothing parameters such that $g_{1n}^* = C_0 n^{-\frac{1}{p_0+d_1}}$, $g_{2n}^* = C_1 n^{-\frac{1}{4p_0+d_2}}$, and $\alpha_n^* = C_2 n^{-\frac{2p_0}{4p_0+d_2}}$.

(ii) Let m_0 be any function in $\mathcal{M}_{1,\rho}$. Suppose a side condition on (g_1, g_2) such that $(ng_1^{d_1})^{-1/2}$ $\leq O(g_1^{2p_0/3})$, and $g_2^{3p_0/2} \leq O([ng_2^{d_2}]^{-1/2})$. Then, the optimal convergence rate of $\widehat{m}_{\alpha,n}$ is given by $||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} = O_p(n^{-\frac{p_0}{3p_0+d_2}})$, under the optimal choice of smoothing parameters such that g_{1n}^* $= C_0 n^{-\frac{1}{(4/3)p_0+d_1}}$, $g_{2n}^* = C_1 n^{-\frac{1}{3p_0+d_2}}$, and $\alpha_n^* = C_2 n^{-\frac{p_0}{3p_0+d_2}}$.

Remark 5.2 (i) Minimizing the lower bounds in Theorem 5.4 w.r.t. (g_1, g_2, α) can provide an alternative way to find the optimal choice of the smoothing parameters. Due to trade-off between the variance and bias terms, the lower bounds corresponding to $m_0 \in \mathcal{M}_{1,\rho}$, for example, are minimized by $(g_{1n}^{**}, g_{2n}^{**}, \alpha_n^{**})$ such that $1/\sqrt{n\alpha_n^{**} g_{2n}^{**d_2}} \simeq g_{2n}^{**p_0}$, $1/\sqrt{ng_{1n}^{**d_1}} \simeq g_{1n}^{**p_0}/\sqrt{\alpha_n^{**}}$, and $\alpha_n^{**} \simeq (1/\sqrt{ng_{2n}^{**d_2}} + g_{1n}^{**p_0})/\sqrt{\alpha_n^{**}}$. From $g_{2n}^{**} \simeq (n\alpha_n^{**})^{-1/(2p_0+d_2)}$ and $g_{1n}^{**} \simeq (n/\alpha_n^{**})^{-1/(2p_0+d_1)}$, it follows that $g_1^{**p_0}\sqrt{ng_{2n}^{**d_2}}$ $= o(\alpha_n^{**(p_0-d_2/2)/(2p_0+d_2)}) = o(1)$, by the assumptions of $d_1 \leq d_2$ and $d_2/2 \leq p_0$, implying that $\alpha_n^{**} \simeq$ $1/\sqrt{n\alpha_n^{**} g_{2n}^{**d_2}}$. As a consequence, $\alpha_n^{**} \simeq \alpha_n^* \simeq n^{-\frac{p_0}{3p_0+d_2}}$ and $g_{2n}^{**} \simeq g_{2n}^* \simeq n^{-\frac{1}{3p_0+d_2}}$, which leads to the same convergence rate as in Theorem 5.5.(ii). Difference of g_{1n}^{**} from g_{1n}^* only affects the terms of second order.

(ii) When $p_0 = 2$ and $d_2 = 1$, we get, from Theorem 5.5.(ii), $||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} = O_p(n^{-\frac{2}{7}})$, which is faster than the rate $O_p(n^{-\frac{1}{4}})$ of Darolles, Florens, and Renault (2001), but slower than $O_p(n^{-\frac{2}{5}})$ available for kernel estimation of reduced forms. Roughly speaking, the optimal choice (g_{1n}^*, g_{2n}^*) requires undersmoothing in the direction of Z and oversmoothing in the direction of W, compared to the standard kernel estimation of joint density functions. ■

Results in Theorem 5.4 and 5.5 have been derived only for the benchmark case of $m_0 \in \mathcal{M}_{1/2}$ or $\mathcal{M}_1$, although no specific form of $U_\alpha(\cdot)$ is assumed except C.3.1 and C.3.2. For a general case of $m_0 \in \mathcal{M}_\mu$(with $\mu > 0$), we will use Theorem 4.1 through 4.4 to show the convergence rates of $\widehat{m}_{\alpha,n}$, regularized by (the ordinary/iterated/generalized) Tikhonov and Showalter's methods.

Theorem 5.6 Assume that C.5.1 through C.5.5 hold with $p_0 = p_1$, and $m_0(\cdot) \in \mathcal{M}_{\mu,\rho}$, with $\mu > 0$.

(i) Let $\widehat{m}^s_{\alpha,n}$ be given by (31), with $U_\alpha(\cdot) = U^s_\alpha(\cdot)$ in (25). Then, it holds

$$||\widehat{m}^s_{\alpha,n} - m_0||_{L^2(\mathcal{X})} \leq O_p(\frac{1}{\sqrt{\alpha}}[\frac{1}{\sqrt{ng_2^{d_2}}} + g_1^{p_0}]) + O_p(\alpha^{\mu}) + O_p(\alpha^{\mu-1/2}[\frac{1}{\sqrt{ng_1^{d_1}}} + g_2^{p_0}]).$$

Assume a side condition on (g_1, g_2) such that

$$(ng_{1n}^{d_1})^{-1/2} \leq O(g_{1n}^{p_0/(2\mu+1)}), \text{ and } g_{2n}^{p_0(2\mu+1)} \leq O(1/\sqrt{ng_{2n}^{d_2}}).$$

Then, the optimal convergence rate of $\widehat{m}_{\alpha,n}$ is given by

$$||\widehat{m}^s_{\alpha,n} - m_0||_{L^2(\mathcal{X})} = O_p(n^{-\frac{2\mu p_0}{2(2\mu+1)p_0+d_2}}),$$

under the choice of smoothing parameters such that

$$g^*_{1n} = C_0 n^{-\frac{(2\mu+1)}{2p_0+(2\mu+1)d_1}}, g^*_{2n} = C_1 n^{-\frac{1}{2(2\mu+1)p_0+d_2}}, \text{ and } \alpha^*_n = C_2 n^{-\frac{2p_0}{2(2\mu+1)p_0+d_2}}.$$

(ii) Let $\widehat{m}^q_{\alpha,n}$ be given by (31) with $U_\alpha(\cdot) = U_{q,\alpha}(\cdot)$ or $U^g_{q,\alpha}(\cdot)$, as defined in section 4, where $q \geq 1$. Then, it holds

$$||\widehat{m}^q_{\alpha,n} - m_0||_{L^2(\mathcal{X})} \leq O_p(\frac{1}{\sqrt{\alpha}}[\frac{1}{\sqrt{ng_2^{d_2}}} + g_1^{p_0}]) + O_p(\alpha^{\min(\mu,q)}) + O_p(\alpha^{\min(\mu-1/2,0)}[\frac{1}{\sqrt{ng_1^{d_1}}} + g_2^{p_0}]).$$

Assume a side condition on (g_1, g_2) such that

$$(ng_{1n}^{d_1})^{-1/2} \leq O(g_{1n}^{2\mu_q^\dagger p_0/(2\mu_q+1)}), \text{ and } g_{2n}^{p_0(2\mu_q+1)/2\mu_q^\dagger} \leq O(1/\sqrt{ng_{2n}^{d_2}}),$$

where $\mu_q = \min(\mu, q)$ and $\mu_q^\dagger = \max(\mu_q, 1/2)$. Then, the optimal convergence rate is given by

$$||\widehat{m}^q_{\alpha,n} - m_0||_{L^2(\mathcal{X})} = O_p(n^{-\frac{2\mu_q p_0}{2(2\mu_q+1)p_0+2\mu_q^\dagger d_2}}),$$

under the choice of smoothing parameters such that

$$g^*_{1n} = C_0 n^{-\frac{(2\mu_q+1)}{4\mu_q^\dagger p_0+(2\mu_q+1)d_1}}, g^*_{2n} = C_1 n^{-\frac{2\mu_q^\dagger}{2(2\mu_q+1)p_0+2\mu_q^\dagger d_2}}, \text{ and } \alpha^*_n = C_2 n^{-\frac{2p_0}{2(2\mu_q+1)p_0+2\mu_q^\dagger d_2}}.$$

The optimal rates of convergence in Theorem 5.6 can be obtained by minimizing (w.r.t. g_1, g_2, and α) the lower bounds given in Theorem 5.6. As in Remark 5.2, two methods give rise to the same choice of (g^*_{2n}, α^*_n), with different g_1's of only second-order effect.

Remark 5.3 (i) Note that the lower bounds in Theorem 5.6.(i) are sharper than that of Theorem 5.4.(ii), at least for $m_0 \in \mathcal{M}_{1,\rho}$. The improvement occurs because the former, unlike the latter, has been derived under a specific feature of Showalter's method. By means of a weaker side condition, Showalter's method can possibly give the faster optimal rate of convergence than are allowed by Theorem 5.5.(ii), which is based on a general regularization method of C.3.1 and C.3.3. The optimal rates of convergence of $\widehat{m}^s_{\alpha,n}$ and $\widehat{m}^q_{\alpha,n}$ are the same, only for the case with $\mu \leq 1/2$, where $\mu_q = \mu$ and $\mu^\dagger_q = 1/2$. Otherwise, the former is better. For $q < \mu$, the convergence rates of $\widehat{m}^q_{\alpha,n}$ do not improve, as μ increases. This confirms the fact that the three variants of Tikhonov methods are not free from the saturation effects, due to finite-qualification.

(ii) Theorem 5.6 shows that the convergence rate of $\widehat{m}^s_{\alpha,n}$ gets faster, as $p_1 (= p_0)$ or μ increase, i.e., m_0 becomes smoother. Kernel estimation of structural functions also suffers from the curse of dimensionality. Here, the dimensionality is determined by W, rather than X. This may seem natural, if we consider that statistical properties of $\widehat{m}_{\alpha,n}$ depend crucially on the accuracies of the preliminary estimates $\widehat{h}_0$ and $\widehat{T}_n$. Assuming $\dim(W) \geq \dim(X)$ as a regularity condition, the optimal convergence rate of $\widehat{m}^s_{\alpha,n}$ will deteriorate as $\dim(X)$ increases. Owing to ill-posedness of the problems, Stones's bounds are not attainable by $\widehat{m}^s_{\alpha,n}$, when $p_0 > (d_2/2)(2\mu + 1)$. We think that the condition is not too strong, since greater μ is generally accompanied by higher order of differentiability. ■

5.3 Numerical Example

In this section, we carry out a small scale Monte Carlo experiment to investigate the finite sample properties of the kernel IV estimators studied in the previous sections. The design for simulation is as follows. Assuming that $(X, W, \varepsilon)' \sim N(0, \Sigma)$, samples $\{(Y_i, X_i, W_i)\}_{i=1}^n$ of size $n = 200$ are generated from a bivariate model,

$$Y_i = \sqrt{2}\cos(X_i) + \varepsilon_i, \tag{32}$$

where

$$\Sigma = \begin{bmatrix} 1 & 1/\sqrt{2} & 1/\sqrt{8} \\ 1/\sqrt{2} & 1 & 0 \\ 1/\sqrt{8} & 0 & 1/2 \end{bmatrix}.$$

Our interest is in applying the kernel IV estimates ($\widehat{m}_{\alpha,n}$) in (29) to estimate the regression function ($m_0(x) \equiv \sqrt{2}\cos(x)$) of the model (32). To see how different regularization methods perform in finite samples, we will consider the kernel estimates regularized by OTR/ITR(2)/GTR(2) and SW. The specific forms of the preliminary estimates in (27) and (28) are fixed by the gaussian kernel function, together with the common bandwidth parameters;

$$(g_1, g_2) = (g, g), \text{ with } g \in G = \{0.3, 0.4, 0.5, 0.6\}.$$

For practical reasons, various regularization parameters are used in calculating $\widehat{m}_{\alpha,n}$ such that

$$\alpha \in A = \{0.001, 0.005, 0.01, 0.015, 0.02\}.$$

As argued in Remark 5.1, no numerical integration is necessary for computing M_W or $Q_{X,W}$, in this (gaussian kernel) case. For each simulated data, we compute $\widehat{m}_{\alpha,n}(x)$ at the 19 quantiles (from 5% through 95%) of x, obtaining a sample pointwise MSE (mean squared errors) of $\widehat{m}_{\alpha,n}$. The same procedure is repeated 1000 times for the whole experiment, which allows us to approximate the true MSE by averaging the sample MSE's over all repetitions. The simulation results are summarized in Table 1, showing the estimated MSE of the various regularized estimates, as well as its decomposition into the squared-bias and variance terms (the two numbers in the parenthesis). The bias term is computed by comparing the true function ($m_0(\cdot)$) and an average (over repetitions) of the estimates ($\widehat{m}_{\alpha,n}$) at each fixed quantile of x. The variance term is defined by the rest of MSE from the squared-bias. Figure 1(a) through (d) display the averaged estimates of the four regularization methods over various α's, with g set to be a representative value of 0.4. Figure 2 collects some of those averaged estimates that correspond to the optimal choice of α (with $g = 0.4$), where the optimality criterion is to minimize MSE. Our interpretation of the results is as follows.

(i) Both Table 1 and Figure 1 show that the regularized IV estimates perform reasonably well under the given sample size, as long as α is not too small (i.e., for $\alpha \geq 0.005$). To one's expectation, the naive kernel estimates (corresponding to $\alpha = 0$, not shown) turn out to suffer from drastically large MSE's, indicative of the instability problem due to ill-posedness of the IV estimation.

(ii) When $\alpha = 0.005$ is chosen, four different regularization methods show similar values of MSE. We achieve slight improvements in the bias term from using ITR(2) or SW rather than OTR. However, the gains are blurred by increases in the variance term, implying that the overall performances of the four methods are similar to each other.

(iii) For other values of α (≥ 0.01), the MSE of OTR is much larger than that of other regularization methods. As α increases, the OTR estimates are getting worse, while the estimates from ITR(2), GTR(2) and SW are still performing well or even better. The bias-variance decompositions in Table 1 reveal that such deterioration in the statistical errors of OTR is attributable to a larger increase in the (regularization) bias term. This also can be corroborated from looking at Fig. 1 which depicts different bias-characteristics of OTR and ITR(2)/GTR(2)/SW by varying a regularization parameter α. Roughly speaking, our simulation results partly support the asymptotic results in section 4 that the refined regularization methods of ITR, GTR and SW have an advantage in bias reduction over OTR.

(iv) To summarize, given the simulation design in (32), we get similar minimum MSE's from the regularization methods of OTR/ITR(2)/GTR(2) and SW, applied to the kernel IV estimates in (29); see the numbers with * in Table 1. That is, the four methods show no significant differences in statistical accuracies, when the smoothing parameters are chosen optimally.[16] Fig. 2 highlights the similarities in the bias terms of different methods for that case. As argued in (iii), the finite sample properties of OTR, however, are quite different from the other methods in that the bias term of the former is highly sensitive to a small change from the optimal regularization parameter.

A Appendices

A.1 Section 2

Proof of Theorem 2.2 (i) Suppose that T is not one-to-one, i.e., there exists a nonzero function $m^* \in L^2(\mathcal{X})$ such that $(Tm^*)(w) = 0$, for all $w \in \mathcal{W}$. From

$$(Tm^*)(\omega_l) = \int_{\mathcal{X}} m^*(x) f_{X,W}(x,\omega_l)dx = < m^*(\cdot),\ f_{X,W}(\cdot,\omega_l) >_{L^2(\mathcal{X})} = 0, \text{ for any } \omega_l \in \overline{\mathcal{W}},$$

it follows that m^* is orthogonal to any linear combination of $\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^{\infty}$, i.e.,

$$m^* \in [\text{lin}\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^{\infty}]^{\perp}.$$

Since the orthogonal complement of $\text{lin}\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^{\infty}$ includes a nonzero function, $\overline{\text{lin}}\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^{\infty}$ is a proper subset of $L^2(\mathcal{X})$, contradicting to denseness of $\text{lin}\{f_{X,W}(\cdot,\omega_l)$ in $L^2(\mathcal{X})$.

(ii) Noting that $\overline{\mathcal{R}}(T) = \mathcal{N}^{\perp}(T^*)$, it suffices to show that T^* is one-to-one from $L^2(\mathcal{W})$ to $L^2(\mathcal{X})$, under C.2.2. The proof is direct from symmetry of the argument used in (i). ■

[16] This occurs when the degree of the abstract smoothness of m_0 does not exceed one. See the arguments in Theorem 3.3, for example.

Proof of Theorem 2.3 (i) Suppose that C.2.1 is violated for some dense subset $\overline{\mathcal{W}}$ of $\mathcal{W}$, i.e., $[\text{lin}\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^{\infty}]^{\perp}$ is not empty. Then, there exists nonzero m^*in $L^2(\mathcal{X})$, which is orthogonal to any linear combination of $\{f_{X,W}(\cdot,\omega_l)\}_{l=1}^{\infty}$. Letting $h^*(w) \equiv (Tm^*)(w)$, this implies that

$$h^*(w) = \int_{\mathcal{X}} m^*(x) f_{X,W}(x,w)dx = < m^*(\cdot),\ f_{X,W}(\cdot,w) >_{L^2(\mathcal{X})} = 0, \text{ for all } w \in \overline{\mathcal{W}}.$$

Note that $h^*(\cdot)$ is continuous in w, due to continuity of $f_{X,W}(\cdot,\cdot)$ on $\mathcal{X} \times \mathcal{W}$. Since $h^*(\cdot) = 0$, on a dense subset of $\mathcal{W}$, it follows from continuity of $h(\cdot)$ that $h^*(w) = (Tm^*)(w) = 0$, for all $w \in \mathcal{W}$, which contradicts to the assumption that T is one-to-one.

(ii) From $\overline{\mathcal{R}}(T) = \mathcal{N}^{\perp}(T^*)$, the proof is direct from symmetry of the argument for showing part (i). ■

Proof of Theorem 2.4 We only show the first assertion, since the second is clear by symmetry. Suppose that there exists nonzero $m_0(\cdot) \in L^2(\mathcal{X})$ with $Tm_0 = 0$. This means that there exists a subset $\overline{\mathcal{W}}_1$ of $\mathcal{W}_1$ (with $\overline{\mathcal{W}}_1$ not measure zero) such that $m_0(z,\omega_1)$ is a nonzero function of z, for all $\omega_1 \in \overline{\mathcal{W}}_1$, but $(T_{\omega_1} m_0)(w_2) = 0$, for all $\omega_1 \in \overline{\mathcal{W}}_1$, where $T_{\omega_1} : L^2(\mathcal{Z}) \to L^2(\mathcal{W}_2)$ is given by $(T_{\omega_1} m)(w_2) = \int_{\mathcal{Z}} m(z,\omega_1) f_{Z,W_1,W_2}(z,\omega_1,w_2)dz$. Since for all $\omega_1 \in \overline{\mathcal{W}}_1$, $\text{lin}\{f_{Z,W_1,W_2}(\cdot,\omega_1,\omega_{2l})\}_{l=1}^{\infty}$ is dense in $L^2(\mathcal{Z})$, it follows that $||P_{f_L^{\omega}} m(\cdot,\omega_1) - m(\cdot,\omega_1)||_{L^2(\mathcal{Z})} \to 0$, as $L \to \infty$, for any $m(\cdot,\omega_1) \in L^2(\mathcal{Z})$. Consequently, there exists L^*(depending on m_0) such that $||P_{f_L^{\omega}} m_0(\cdot,\omega_1)||_{L^2(\mathcal{Z})} \geq ||m_0(\cdot,\omega_1)||_{L^2(\mathcal{Z})}/2 > 0$. This is a contradiction, since, from $T_{\omega_1} m_0 = 0$,

$$(P_{f_L^{\omega}} m_0(\cdot,\omega_1))(z,\omega_1) = f_L^{\omega}(z)' Q_{\omega\omega}^{*\dagger} [\{T_{\omega} m_0(\cdot,\omega_1)\}(\omega_1,\omega_{2l})]_{l=1}^{L} = 0, \quad \text{for any } L(\geq 1).$$

■

A.2 Section 3

The following lemmas are useful for showing the main results of Section 3 and 4.

Definition A sequence of linear random operators $\widehat{T}_n : \mathcal{M} \to \mathcal{H}$ is asymptotically one-to-one, if and only if $P_{\mathcal{N}(\widehat{T}_n)}$ converges pointwise to zero, in probability; i.e., for any $m \in \mathcal{M}$, $||P_{\mathcal{N}(\widehat{T}_n)} m||_{\mathcal{M}} \xrightarrow{p} 0$.

Lemma 3.1 Assume that a sequence of random operators $\widehat{T}_n : \mathcal{M} \to \mathcal{H}$ converges pointwise, in probability, to a bounded operator $T : \mathcal{M} \to \mathcal{H}$ which is one-to-one, where $\mathcal{M}$ and $\mathcal{H}$ are a Hilbert space. Then, $\widehat{T}_n : \mathcal{M} \to \mathcal{H}$ is asymptotically one-to-one.

Proof. From $\mathcal{N}(T) = \mathcal{N}(T^*T) = \mathcal{R}^{\perp}(T^*T)$, injectivity of $T : \mathcal{M} \to \mathcal{H}$ is equivalent to that T^*T has a dense range in $\mathcal{M}$. That is, for any arbitrary element m_0 in $\mathcal{M}$, there exists a sequence $\{m_l\}_{l=1}^{\infty}$ such that $T^*Tm_l \to m_0$, as $l \to \infty$. By the triangle inequality, for any l,

$$||(\widehat{T}_n^*\widehat{T}_n - T^*T)m_l||_{\mathcal{M}} \ \leq ||\widehat{T}_n^*(\widehat{T}_n - T)m_l||_{\mathcal{M}} \ + ||(\widehat{T}_n^* - T^*)Tm_l||_{\mathcal{M}}, \text{ a.s.} \tag{33}$$

From $(\widehat{T}_n^* - T^*) = (\widehat{T}_n - T)^*$, pointwise convergence of $\widehat{T}_n$ to T implies that, for any $h \in \mathcal{H}$,

$$\begin{aligned} ||(\widehat{T}_n^* - T^*)h||_{\mathcal{M}}^2 &= \ < (\widehat{T}_n^* - T^*)h, (\widehat{T}_n^* - T^*)h >_{\mathcal{M}} \\ &= \ < h, \ (\widehat{T}_n - T)\widetilde{m} >_{\mathcal{H}} \xrightarrow{p} 0, \ \text{ as } n \to \infty, \end{aligned}$$

by continuity of the inner product, where $\widetilde{m} = (\widehat{T}_n^* - T^*)h$. This shows negligibility of the second term in the righthand-side of (33). In addition, for any fixed $h \in \mathcal{H}$, $||\widehat{T}_n^*h||_{\mathcal{M}} \leq ||(\widehat{T}_n^* - T^*)h||_{\mathcal{M}} + ||T^*h||_{\mathcal{M}} \leq C||T^*h||_{\mathcal{M}}$, for n sufficiently large, which, by boundedness of T, gives

$$\sup_n ||\widehat{T}_n^*h||_{\mathcal{M}} = O_p(1), \text{ for each } h \in \mathcal{H}.$$

By the Principle of Uniform Boundedness-see Taylor and Lay (1980, p.190), the above implies that the sequence $\{\widehat{T}_n^*\}$ is bounded uniformly in n, i.e., $\sup_n ||\widehat{T}_n^*||_{\mathcal{H}\to\mathcal{M}} = O_p(1)$.[17] Negligibility of the first term in the righthand-side of (33) follows from pointwise convergence of $\widehat{T}_n$ to T, since $||\widehat{T}_n^*(\widehat{T}_n - T)m_l||_{\mathcal{M}} \leq ||\widehat{T}_n^*||_{\mathcal{H}\to\mathcal{M}}||(\widehat{T}_n - T)m_l||_{\mathcal{M}} = O_p(1)||(\widehat{T}_n - T)m_l||_{\mathcal{M}} \xrightarrow{p} 0$, as $n \to \infty$, for any $m_l \in \mathcal{M}$. In consequence, $\widehat{T}_n^*\widehat{T}_n$ converges pointwise to T^*T in $\mathcal{M}$, which, together with the definition of $\{m_l\}_{l=1}^{\infty}$, leads to

$$||(\widehat{T}_n^*\widehat{T}_n)m_l - m_0||_{\mathcal{M}} \leq ||(\widehat{T}_n^*\widehat{T}_n - T^*T)m_l||_{\mathcal{M}} + ||(T^*T)m_l - m_0||_{\mathcal{M}} \xrightarrow{p} 0, \ \text{ as } \max(n, \ l) \to \infty.$$

Using $\mathcal{M} = \mathcal{R}^{\perp}(\widehat{T}_n^*\widehat{T}_n) \oplus \overline{\mathcal{R}}(\widehat{T}_n^*\widehat{T}_n)$, we have, by the orthogonal projection in Hilbert space, that $m_0 = P_{\mathcal{R}^{\perp}(\widehat{T}_n^*\widehat{T}_n)}m_0 + P_{\overline{\mathcal{R}}(\widehat{T}_n^*\widehat{T}_n)}m_0$, yielding

$$\begin{aligned} ||P_{\mathcal{N}(\widehat{T}_n^*\widehat{T}_n)}m_0||_{\mathcal{M}} &= \ ||P_{\mathcal{R}^{\perp}(\widehat{T}_n^*\widehat{T}_n)}m_0||_{\mathcal{M}} = ||P_{\overline{\mathcal{R}}(\widehat{T}_n^*\widehat{T}_n)}m_0 - m_0||_{\mathcal{M}} \\ &= \inf_{m\in\overline{\mathcal{R}}(\widehat{T}_n^*\widehat{T}_n)} ||m - m_0||_{\mathcal{M}} \leq ||(\widehat{T}_n^*\widehat{T}_n)m_l - m_0||_{\mathcal{M}} \xrightarrow{p} 0, \ \text{ as } n \to \infty, \end{aligned}$$

[17] Suppose that $\mathcal{M}$ and $\mathcal{H}$ are normed linear spaces, and $\mathcal{M}$ is complete. Let $\{T_n\}_{n=1}^{\infty}$ be a sequence of linear bounded operators, $T_n : \mathcal{M} \to \mathcal{H}$, such that

$$\sup_{n\geq 1} ||T_nm||_{\mathcal{H}} < \infty, \text{ for each } m \in \mathcal{M}.$$

Then, $\sup_{n\geq 1} ||T_n|| < \infty$.

where the last inequality holds from $(\widehat{T}_n^*\widehat{T}_n)m_l \in \overline{\mathcal{R}}(\widehat{T}_n^*\widehat{T}_n)$. Since this result holds for any $m_0 \in \mathcal{M}$, and $\mathcal{N}(\widehat{T}_n^*\widehat{T}_n) = \mathcal{N}(\widehat{T}_n)$, the assertion is proved. ■

Lemma 3.2 Suppose that $U_\alpha(\cdot)$ satisfies C.3.1, and $\widehat{T}_n : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ has a finite rank. If $\alpha = \alpha(n) \to 0$ as $n \to \infty$, then,

(i) $||U_\alpha(\widehat{T}_n^*\widehat{T}_n)||_{L^2(\mathcal{X})\to L^2(\mathcal{X})} = O_{a.s}(\alpha^{-1})$,

(ii) $||\widehat{R}_{\alpha,n}||_{L^2(\mathcal{W})\to L^2(\mathcal{X})} = O_{a.s}(\sqrt{\alpha}^{-1})$,

Assume additionally that $\widehat{T}_n : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ converges pointwise, in probability, to $T : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ which is bounded and one-to-one. Then,

(iii) $||[U_\alpha(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{T}_n - I]m||_{L^2(\mathcal{X})} = o_p(1)$, for all $m \in L^2(\mathcal{X})$.

Proof From finite rank of $\widehat{T}_n$, the self-adjoint operator $\widehat{T}_n^*\widehat{T}_n$ is compact and thereby has a spectral representation, such as $\widehat{T}_n^*\widehat{T}_n(\cdot) = \sum_{j=1}^{J_n} \lambda_j P_{v_j}$, where λ_j's and P_{v_j}'s denote the eigenvalues of $\widehat{T}_n^*\widehat{T}_n$ and the orthogonal projection onto the eigenspace generated by the eigenfunction, v_j, that corresponds to λ_j, respectively. With $\overline{\lambda} = \sup_{n\geq n_0}\{\, ||\widehat{T}_n^*\widehat{T}_n||_{L^2(\mathcal{X})\to L^2(\mathcal{X})}\}$, we get, by spectral calculus, that

$$\begin{aligned} ||U_\alpha(\widehat{T}_n^*\widehat{T}_n)m||^2_{L^2(\mathcal{X})} &= \sum_{j=1}^{J_n} U_\alpha^2(\lambda_j)||P_{v_j}m||^2_{L^2(\mathcal{X})} \leq \sup_{\lambda\in(0,\overline{\lambda}]} |U_\alpha(\lambda)|^2 \sum_{j=1}^{J_n} ||P_{v_j}m||^2_{L^2(\mathcal{X})} \\ &\leq \alpha^{-2}||\sum_{j=1}^{J_n} P_{v_j}m||^2_{L^2(\mathcal{X})}, \text{ a.s, for } \alpha \to 0^+, \end{aligned}$$

where the last inequality comes from C.3.1(iii) and orthogonality of v_j and $v_{j'}$, for $j \neq j'$. Since $\sum_{j=1}^{J_n} P_{v_j}$ is itself a projection operator, it holds that $||\sum_{j=1}^{J_n} P_{v_j}||_{L^2(\mathcal{X})\to L^2(\mathcal{X})} \leq 1$, implying

$$||U_\alpha(\widehat{T}_n^*\widehat{T}_n)||^2_{L^2(\mathcal{X})\to L^2(\mathcal{X})} = \sup_{m\in L^2(\mathcal{X})} ||U_\alpha(\widehat{T}_n^*\widehat{T}_n)m||^2_{L^2(\mathcal{X})}/||m||^2_{L^2(\mathcal{X})} \leq \alpha^{-2}||\sum_{j=1}^{J_n} P_{v_j}||^2_{L^2(\mathcal{X})} = \alpha^{-2}, \text{ a.s.}$$

This completes the proof for (i). In a similar way, letting Q_j be the projection onto the space generated by $\widehat{T}_n v_j \in L^2(\mathcal{W})$, the singular values decomposition of $\widehat{T}_n^*\widehat{T}_n$ yields

$$\begin{aligned} ||U_\alpha(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*h||^2_{L^2(\mathcal{X})} &= \sum_{j=1}^{J_n} U_\alpha^2(\lambda_j)\lambda_j||Q_jh||^2_{L^2(\mathcal{W})} \leq \sup_\lambda |U_\alpha(\lambda)\lambda| \sup_\lambda |U_\alpha(\lambda)| \sum_j ||Q_jh||^2_{L^2(\mathcal{W})} \\ &\leq \sup_\lambda |U_\alpha(\lambda)\lambda| \sup_\lambda |U_\alpha(\lambda)|||h||^2_{L^2(\mathcal{W})}, \text{ a.s.} \end{aligned}$$

From C.3.1(i) and (iii), we get

$$\sup_{h\in L^2(\mathcal{W})} \frac{||U_\alpha(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^* h||^2_{L^2(\mathcal{X})}}{||h||^2_{L^2(\mathcal{W})}} \leq C \sup_\lambda |U_\alpha(\lambda)| \leq C\alpha^{-1}, \text{ a.s,}$$

proving (ii). For a proof of (iii), we let $J_{n,1} = \{j \in I_+ : j \leq J_n, \lambda_j > 0\}$, and $P_{\mathcal{N}(\widehat{T}_n^*\widehat{T}_n)}$ the orthogonal projection onto the null space of $\widehat{T}_n^*\widehat{T}_n$. From $U_\alpha(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{T}_n = \sum_{j\in J_{n,1}} U_\alpha(\lambda_j)\lambda_j P_{v_j}$, and $I = \sum_{j\in J_{n,1}} P_{v_j} + P_{\mathcal{N}(\widehat{T}_n^*\widehat{T}_n)}$, it follows that

$$\begin{aligned} & ||[U_\alpha(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{T}_n - I]m||^2_{L^2(\mathcal{X})} \\ = & \sum_{j\in J_{n,1}} [U_\alpha(\lambda_j)\lambda_j - 1]^2 ||P_{v_j} m||^2_{L^2(\mathcal{X})} + ||P_{\mathcal{N}(\widehat{T}_n^*\widehat{T}_n)} m||^2_{L^2(\mathcal{X})}. \end{aligned}$$

By the Dominated Convergence Theorem, the first term converges to zero, a.s, as $n \to 0$, since, by C.3.1(ii),

$$\lim_{\alpha\to 0} \sum_{j\in J_{n,1}} [U_\alpha(\lambda_j)\lambda_j - 1]^2 ||P_{v_j} m||^2_{L^2(\mathcal{X})} = \sum_j^{J_n} \lim_{\alpha\to 0} [U_\alpha(\lambda_j)\lambda_j - 1]^2 ||P_{v_j} m||^2_{L^2(\mathcal{X})} = 0, \text{ a.s.}$$

Negligibility of the second term, i.e., $||P_{\mathcal{N}(\widehat{T}_n^*\widehat{T}_n)} m||^2_{L^2(\mathcal{X})} = o_p(1)$, is immediate from Lemma 3.1, since $\mathcal{N}(\widehat{T}_n^*\widehat{T}_n) = \mathcal{N}(\widehat{T}_n)$. ■

The following lemma is well known in mathematical theory of inverse problems, see ,for example. We introduce the proof, just for completeness of arguments.

Lemma 3.3 Let $G : L^2(\mathcal{X}) \to L^2(\mathcal{W})$ be a linear bounded operator and $G^* : L^2(\mathcal{W}) \to L^2(\mathcal{X})$ be adjoint to G. If U_α satisfies C.3.2, then, for all $m \in L^2(\mathcal{X})$,

$$\text{(i)} \quad ||[U_\alpha(G^*G)G^*G - I](G^*G)^\mu||_{L^2(\mathcal{X})\to L^2(\mathcal{X})} \leq C\alpha^{\min(\mu,\bar{\mu})}, \text{ for } \mu > 0.$$

$$\text{(ii)} \quad ||[U_\alpha(G^*G)G^*G - I]G^*||_{L^2(\mathcal{X})\to L^2(\mathcal{X})} \leq C\alpha^{1/2}, \text{ for } \mu > 0.$$

Proof (i) Let $\{E_\lambda\}$ be a spectral family for G^*G such that

$$[U_\alpha(G^*G)G^*G - I](G^*G)^\mu = \int_0^\infty (U_\alpha(\lambda)\lambda - 1)\lambda^\mu dE_\lambda.$$

By spectral theory, for $m \in L^2(\mathcal{X})$,

$$||[U_\alpha(G^*G)G^*G - I](G^*G)^\mu m||^2_{L^2(\mathcal{X})} = \int_0^\infty [(U_\alpha(\lambda)\lambda - 1)\lambda^\mu]^2 d||E_\lambda m||^2_{L^2(\mathcal{X})}.$$

From $\int_0^\infty d||E_\lambda m||^2_{L^2(\mathcal{X})} = ||m||^2 < \infty$, we have, by C.3.2, that

$$||[U_\alpha(G^*G)G^*G - I](G^*G)^\mu m||_{L^2(\mathcal{X})} \leq C \sup_{\lambda \in (0,\overline{\lambda}]} \lambda^\mu |U_\alpha(\lambda)\lambda - 1| \leq C'\alpha^\mu.$$

The proof for (ii) is immediate from (i), since $\mathcal{R}(G^*) = \mathcal{R}((G^*G)^{1/2})$, for any linear bounded operator G. ■

Proof of Proposition 3.1 Suppose that $\sup_n ||\widehat{T}_n^\dagger||_{L^2(\mathcal{W})\to L^2(\mathcal{X})} = O_p(1)$. Then, from uniform convergence of $\widehat{T}_n$ to T on $\mathcal{M}_X$, it follows that

$$||\widehat{T}_n^\dagger(\widehat{T}_n - T)||_{\mathcal{M}_X\to L^2(\mathcal{X})} \leq ||\widehat{T}_n^\dagger||_{L^2(\mathcal{W})\to L^2(\mathcal{X})}||\widehat{T}_n - T||_{\mathcal{M}_X\to L^2(\mathcal{W})} = O_p(1)||\widehat{T}_n - T||_{\mathcal{M}_X\to L^2(\mathcal{W})} \xrightarrow{p} 0,$$

i.e., $\widehat{T}_n^\dagger\widehat{T}_n$ converges $\widehat{T}_n^\dagger T$ uniformly on $\mathcal{M}_X$, in probability. From the identity, $I - \widehat{T}_n^\dagger T = P_{\mathcal{N}(\widehat{T}_n)}$-see Groetsch (1977), we get, by Lemma 3.1, that $||\widehat{T}_n^\dagger\widehat{T}_n - I||_{\mathcal{M}_X\to L^2(\mathcal{X})} \xrightarrow{p} 0$. Consequently, by the triangle inequality,

$$||\widehat{T}_n^\dagger T - I||_{\mathcal{M}_X\to L^2(\mathcal{X})} \leq ||\widehat{T}_n^\dagger(\widehat{T}_n - T)||_{\mathcal{M}_X\to L^2(\mathcal{X})} + ||\widehat{T}_n^\dagger\widehat{T}_n - I||_{\mathcal{M}_X\to L^2(\mathcal{X})} \xrightarrow{p} 0.$$

That is, for any $h \in T(\mathcal{M}_X)$, $\widehat{T}_n^\dagger h$ converges to $T^{-1}h$, in probability, which, by the Principle of Uniform Boundedness, implies that $\sup_n ||\widehat{T}_n^\dagger - T^{-1}||_{L^2(\mathcal{W})\to L^2(\mathcal{X})} = O_p(1)$. From

$$||T^{-1}||_{L^2(\mathcal{W})\to L^2(\mathcal{X})} \leq ||\widehat{T}_n^\dagger - T^{-1}||_{L^2(\mathcal{W})\to L^2(\mathcal{X})} + ||\widehat{T}_n^\dagger||_{L^2(\mathcal{W})\to L^2(\mathcal{X})} = O_p(1),$$

follows boundedness of the mapping, $T^{-1} : T(\mathcal{M}_X) \to \mathcal{M}_X$. Since $\{\widehat{T}_n\}$ is a sequence of compact operators (from $\dim[\mathcal{R}(\widehat{T}_n)] < \infty$), the limit of $\{\widehat{T}_n\}$, i.e., T, is also compact on $\mathcal{M}_X$. By injectiveness of T, $\dim[T(\mathcal{M}_X)] = \dim[\mathcal{M}_X] = \infty$, which contradicts to the fact that a compact operator cannot have a bounded inverse, when its range space is infinite-dimensional; see Kress (1989, p20). ■

Proof of Theorem 3.2 The result is direct from the triangle inequality and application of Lemma 3.2 (ii) and (iii) to (13).

Proof of Theorem 3.3 (i) Since $\mathcal{R}(T^*) = \mathcal{R}((T^*T)^{1/2})$, it follows from (13) that the error decomposition for $m_0 \in \mathcal{M}_{1/2}$ is given by

$$\widehat{R}_{\alpha,n}(\widehat{h}_{0,n} - \widehat{T}_n m_0) + [(\widehat{\Gamma}_\alpha - I)\widehat{T}_n^*]h_1 - (\widehat{\Gamma}_\alpha - I)(\widehat{T}_n^* - T^*)h_1,$$

where $h_1 = T^{*-1}(m_0)$. By Lemma 3.2(ii), L^2-norm of the first term is bounded by $\frac{C_1}{\sqrt{\alpha}}||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})}$, almost surely. By Lemma 3.3(i) and (ii), $(\widehat{\Gamma}_\alpha - I)$ and $(\widehat{\Gamma}_\alpha - I)\widehat{T}_n^*$ are uniformly bounded by C_3 and $C_2\alpha^{1/2}$, almost surely, respectively, which proves the first assertion.

(ii) For the case with $m_0 \in \mathcal{M}_1$, the error decomposition takes form of

$$\widehat{R}_{\alpha,n}(\widehat{h}_n - \widehat{T}_n m_0) + [(\widehat{\Gamma}_\alpha - I)(\widehat{T}_n^* \widehat{T}_n)]m_1 - [(\widehat{\Gamma}_\alpha - I)\widehat{T}_n^*](\widehat{T}_n - T)m_1 - (\widehat{\Gamma}_\alpha - I)(\widehat{T}_n^* - T^*)Tm_1,$$

where $m_1 = (T^*T)^{-1}m_0$. The proof for the second assertion follows immediately, if we additionally apply the same uniform-boundedness argument to $(\widehat{\Gamma}_\alpha - I)(\widehat{T}_n^* \widehat{T}_n)$, again based on Lemma 3.3(i). ■

Proof of Corollary 3.4 From $m_0 \in \mathcal{M}_{1/2,\rho}$, $||h_1||_{L^2(\mathcal{W})} \leq \rho$, which, by the definition of operator norm, implies that $||(\widehat{T}_n^* - T^*)h_1||_{L^2(\mathcal{W})} \leq \rho ||\widehat{T}_n^* - T^*||_{L^2(\mathcal{W}) \to L^2(\mathcal{X})}$. This proves the first assertion. In a similar way, for $m_0 \in \mathcal{M}_{1,\rho}$, we have, by definition, that $||m_1||_{L^2(\mathcal{X})} \leq \rho$, and $||(\widehat{T}_n - T)m_1||_{L^2(\mathcal{W})} \leq \rho ||\widehat{T}_n - T||_{L^2(\mathcal{X}) \to L^2(\mathcal{W})}$. From $||T||_{L^2(\mathcal{X}) \to L^2(\mathcal{W})} \leq C^*$, it follows that $||h_1||_{L^2(\mathcal{W})} = ||Tm_1||_{L^2(\mathcal{W})} \leq C^* \rho$. This, together with $h_1 \in T(L^2(\mathcal{X}))$, gives $||(\widehat{T}_n^* - T^*)h_1||_{L^2(\mathcal{X})} \leq C^* \rho ||\widehat{T}_n^* - T^*||_{T(\mathcal{M}_{0,\rho}) \to L^2(\mathcal{X})}$. ■

Proof of Theorem 3.5 Let $\Omega(\{\delta_k\}, \mathcal{M})$ be the modulus of stochastic equicontinuity for $\widehat{T}_n^\dagger$ on $\mathcal{M}$, i.e.,

$$\Omega(\{\delta_k\}, \mathcal{M}) = \sup_{m \in \mathcal{M},\ ||\widehat{T}_k m||_{L^2(\mathcal{W})} = O_p(\delta_k)} ||m||_{L^2(\mathcal{X})}.$$

From the definition of the worst-case convergence rate, it holds for any $R \in \mathfrak{R}$ that

$$\begin{aligned}
\Xi(\{\delta_k\}, \mathcal{M}, R) &= \sup_{m \in \mathcal{M},\ ||\widehat{h}_k - \widehat{T}_k m||_{L^2(\mathcal{W})} = O_p(\delta_k)} \mathrm{E}(||R(\widehat{h}_n) - m||_{L^2(\mathcal{X})}) \\
&\geq \sup_{m \in \mathcal{M},\ ||\widehat{T}_k m||_{L^2(\mathcal{W})} = O_p(\delta_k)} \mathrm{E}(||R(0) - m||^2_{L^2(\mathcal{X})}) \\
&= \sup_{m \in \mathcal{M},\ ||\widehat{T}_k m||_{L^2(\mathcal{W})} = O_p(\delta_k)} ||m||^2_{L^2(\mathcal{X})},
\end{aligned}$$

where the inequality trivially holds for $\widehat{h}_n = 0$, and the last equality is due to the assumption, $R(0) = 0$. Hence, the modulus of stochastic equicontinuity imposes a bound on the best-possible convergence rate (among the class, $\mathfrak{R}$) in the sense that

$$\inf_{R \in \mathfrak{R}} \Xi(\{\delta_k\}, \mathcal{M}, R) \geq C\Omega(\{\delta_k\}, \mathcal{M}),$$

for some $C > 0$. It suffices to show the explicit form of the bound, for $\mathcal{M} = \mathcal{M}_{\mu,\rho}$;

$$\Omega(\{\delta_n\}, \mathcal{M}_{\mu,\rho}) = O(\delta_n^{\frac{2\mu}{2\mu+1}}).$$

For $m = (T^*T)^\mu m_\mu \in \mathcal{M}_{\mu,\rho}$, we have, by the Hölder inequality, that $||m||_{L^2(\mathcal{X})} = ||(T^*T)^\mu m_\mu||_{L^2(\mathcal{X})}$ $\leq ||(T^*T)^{\mu+1/2}m_\mu||_{L^2(\mathcal{X})}^{\frac{2\mu}{2\mu+1}}||m_\mu||_{L^2(\mathcal{X})}^{\frac{1}{2\mu+1}}$. By definition, $(T^*T)^{\mu+1/2}m_\mu = (T^*T)^{1/2}m$, and, from T^* being the adjoint of T, it follows that

$$\begin{aligned} ||(T^*T)^{1/2}m||_{L^2(\mathcal{X})}^2 &= \quad < (T^*T)^{1/2}m, (T^*T)^{1/2}m >_{L^2(\mathcal{X})} = < m, (T^*T)m >_{L^2(\mathcal{X})} \\ &= \ < Tm, Tm >_{L^2(\mathcal{X})} = ||Tm||_{L^2(\mathcal{X})}^2, \end{aligned}$$

That is,

$$||m||_{L^2(\mathcal{X})} \leq ||(T^*T)^{1/2}m||_{L^2(\mathcal{X})}^{\frac{2\mu}{2\mu+1}}||m_\mu||_{L^2(\mathcal{X})}^{\frac{1}{2\mu+1}} \leq ||Tm||_{L^2(\mathcal{W})}^{\frac{2\mu}{2\mu+1}}\rho^{\frac{1}{2\mu+1}}. \tag{34}$$

By the triangle inequality and pointwise convergence of $\widehat{T}_n$ to T in $L^2(\mathcal{X})$, it holds for any $m \in L^2(\mathcal{X})$ that

$$\begin{aligned} ||Tm||_{L^2(\mathcal{W})} &\leq ||\widehat{T}_n m||_{L^2(\mathcal{W})} + ||\widehat{T}_n m - Tm||_{L^2(\mathcal{W})} = ||\widehat{T}_n m||_{L^2(\mathcal{W})}(1 + \frac{||\widehat{T}_n m - Tm||_{L^2(\mathcal{W})}}{||\widehat{T}_n m||_{L^2(\mathcal{W})}}) \\ &\leq ||\widehat{T}_n m||_{L^2(\mathcal{W})}(2 + \frac{||Tm||_{L^2(\mathcal{W})}}{||\widehat{T}_n m||_{L^2(\mathcal{W})}}) \leq C||\widehat{T}_n m||, \quad \textit{w.p.a.1}, \end{aligned}$$

where C (> 3) does not depend on m. This, together with (34), implies that, for any $m \in \mathcal{M}_{\mu,\rho}$,

$$||m||_{L^2(\mathcal{X})} \leq C||\widehat{T}_n m||_{L^2(\mathcal{W})}^{\frac{2\mu}{2\mu+1}}, \ \textit{w.p.a.1},$$

leading to

$$\Omega(\{\delta_k\}, \mathcal{M}) = \sup_{m\in\mathcal{M},\ ||\widehat{T}_n m||_{L^2(\mathcal{W})} = O_p(\delta_n)} ||m||_{L^2(\mathcal{X})} \leq O_p(\delta_n^{\frac{2\mu}{2\mu+1}}). \tag{35}$$

It only remains to show that the bound in (35) is sharp, i.e., there are some cases that

$$\Omega(\{\delta_n\}, \mathcal{M}_{\mu,\rho}) = O_p(\delta_n^{\frac{2\mu}{2\mu+1}}).$$

Let δ_k^2/ρ^2 be given by an eigenvalue of the operator $(T^*T)^{1+2\mu}$ and v_k be the corresponding eigenfunction with $||v_k|| = \rho$. For $m_k \equiv (T^*T)^\mu v_k \in \mathcal{M}_{\mu,\rho}$, it holds that $||Tm_k||_{L^2(\mathcal{W})}^2 = ||T(T^*T)^\mu v_k||_{L^2(\mathcal{W})}^2 =$ $< (T^*T)^{1+2\mu}v_k,\ v_k >_{L^2(\mathcal{X})} = \delta_k^2$, since $(T^*T)^{1+2\mu}v_k = (\delta_k^2/\rho^2)v_k$, from the definition of an eigenvalue. By double use of the triangle inequality, we obtain

$$\begin{aligned} ||\widehat{T}_n m_k||_{L^2(\mathcal{W})}^2 &\leq ||Tm_k||_{L^2(\mathcal{W})}^2 + ||(\widehat{T}_n - T)m_k||_{L^2(\mathcal{W})}^2 \leq ||Tm_k||_{L^2(\mathcal{W})}^2(1 + \frac{||(\widehat{T}_n - T)m_k||_{L^2(\mathcal{W})}^2}{||Tm_k||_{L^2(\mathcal{W})}^2}) \\ &\leq ||Tm_k||_{L^2(\mathcal{W})}^2(2 + \frac{||\widehat{T}_n m_k||_{L^2(\mathcal{W})}^2}{||Tm_k||_{L^2(\mathcal{W})}^2}) = O_p(\delta_k^2), \end{aligned}$$

where the last equality comes from pointwise convergence of $\widehat{T}_n$ to T (in $L^2(\mathcal{X})$) in probability. From $m_k \in \mathcal{M}_{\mu,\rho}$ and $||\widehat{T}_n m_k||_{L^2(\mathcal{W})} = O_p(\delta_k)$, it follows that $\Omega(\{\delta_n\}, \mathcal{M}_{\mu,\rho}) \geq ||m_k||_{L^2(\mathcal{X})} = \{< (T^*T)^{2\mu} v_k, v_k >_{L^2(\mathcal{X})}\}^{1/2} = \{(\delta_k^2/\rho^2)^{\frac{2\mu}{2\mu+1}} < v_k, v_k >_{L^2(\mathcal{X})}\}^{1/2} = (\delta_k/\rho)^{\frac{2\mu}{2\mu+1}} \rho \geq C\delta^{\frac{2\mu}{2\mu+1}}$. If $\delta_k^2/\rho^2 \in \sigma((T^*T)^{1+2\mu})$ is not an eigenvalue, then δ_k^2/ρ^2 belongs to the continuous spectrum of $(T^*T)^{1+2\mu}$ and there exists a sequence $\{v_{k,j}\}_{j=1}^{\infty}$ satisfying $||(T^*T)^{1+2\mu} v_{k,j} - (\delta_k^2/\rho^2) v_{k,j}||_{L^2(\mathcal{X})} \to 0$, and $||v_{k,j}||_{L^2(\mathcal{X})} = \rho$. In this case, too, we can show $\Omega(\{\delta_n\}, \mathcal{M}_{\mu,\rho}) \geq C\delta_k^{\frac{2\mu}{2\mu+1}}$, with a slight modification of the above argument. ■

A.3 Section 4

In the proofs below, we use the following error decomposition for each regularization

$$\widehat{m}_{\alpha,n} - m_0 = \widehat{R}_{\alpha,n}(\widehat{h}_n - \widehat{T}_n m_0) + (\Gamma_\alpha - I)(T^*T)^\mu m_\mu + (\widehat{\Gamma}_\alpha - \Gamma_\alpha)(T^*T)^\mu m_\mu, \qquad (36)$$

where $\widehat{R}_{\alpha,n}$, $\widehat{\Gamma}_\alpha$, and Γ_α have the same definition as in section 3, which, of course, vary over regularization methods.

Proof of Theorem 4.1 (a) Letting $\Gamma_{1,\alpha} = U_{1,\alpha}(T^*T)T^*T$ and $\widehat{\Gamma}_{1,\alpha} = U_{1,\alpha}(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{T}_n$, the error decomposition of OTR is given by (36), under (15). Since both C.3.1 and C.3.2 hold for $U_{1,\alpha}(\cdot)$, the uniform bound for $\widehat{R}_{\alpha,n}$ and $(\Gamma_\alpha - I)(T^*T)^\mu$ follows from Lemma 3.2 and 3.3, respectively. From the qualification of OTR equal to one, the second term is bounded by $C\alpha^{\min(\mu.1)}$, see Lemma 3.3(i). For the last term in (36), we use $A^{-1} - B^{-1} = -A^{-1}(A-B)B^{-1}$ to obtain

$$\begin{aligned}(\widehat{\Gamma}_{1,\alpha} - \Gamma_{1,\alpha}) &= -\alpha[(\alpha I + \widehat{T}_n^*\widehat{T}_n)^{-1} - (\alpha I + T^*T)^{-1}] \\ &= \{\alpha(\alpha I + \widehat{T}_n^*\widehat{T}_n)^{-1}\widehat{T}_n^*\}(\widehat{T}_n - T)(\alpha I + T^*T)^{-1} \\ &\quad + \alpha(\alpha I + \widehat{T}_n^*\widehat{T}_n)^{-1}(\widehat{T}_n^* - T^*)T(\alpha I + T^*T)^{-1}.\end{aligned}$$

By Lemma 3.2(ii), $\alpha^{1/2}(\alpha I + \widehat{T}_n^*\widehat{T}_n)^{-1}\widehat{T}_n^*$ is bounded (uniform in n), and thus

$$\begin{aligned}N_{1,1} &= ||\{\alpha(\alpha I + \widehat{T}_n^*\widehat{T}_n)^{-1}\widehat{T}_n^*\}(\widehat{T}_n - T)\left\{(\alpha I + T^*T)^{-1}(T^*T)^\mu m_\mu\right\}||_{L^2(\mathcal{X})} \\ &\leq C\alpha^{\min(\mu-1/2,1/2)}||(\widehat{T}_n - T)\left\{\alpha^{\min(1-\mu,0)}(\alpha I + T^*T)^{-1}(T^*T)^\mu m_\mu\right\}||_{L^2(\mathcal{W})}, \\ &= C\alpha^{\min(\mu-1/2,1/2)}||(\widehat{T}_n - T)\overline{m}_{\alpha,\mu}||_{L^2(\mathcal{W})},\end{aligned}$$

where $\overline{m}_{\alpha,\mu} = \alpha^{\min(1-\mu,0)}(\alpha I + T^*T)^{-1}(T^*T)^\mu m_\mu$, with $m_\mu \in L^2(\mathcal{X})$. Note that $\overline{m}_{\alpha,\mu} \in \mathcal{M}_{\max(\mu-1,0)}$.

In a similar way, by uniform boundedness of $\alpha(\alpha I + \widehat{T}_n^* \widehat{T}_n)^{-1}$ (from Lemma 3.2(i)),

$$\begin{aligned} N_{1,2} &= ||\alpha(\alpha I + \widehat{T}_n^* \widehat{T}_n)^{-1}(\widehat{T}_n^* - T^*)T(\alpha I + T^*T)^{-1}(T^*T)^{\mu} m_{\mu}||_{L^2(\mathcal{X})} \\ &\leq C\alpha^{\min(\mu-1/2,0)}||(\widehat{T}_n^* - T^*)\{\alpha^{\min(1/2-\mu,0)}T(\alpha I + T^*T)^{-1}(T^*T)^{\mu} m_{\mu}\}||_{L^2(\mathcal{X})} \\ &= C\alpha^{\min(\mu-1/2,0)}||(\widehat{T}_n^* - T^*)\overline{h}_{\alpha,\mu}||_{L^2(\mathcal{X})}, \end{aligned}$$

where $\overline{h}_{\alpha,\mu} = \alpha^{\min(1/2-\mu,0)}T(\alpha I + T^*T)^{-1}(T^*T)^{\mu} m_{\mu}$, and so $\overline{h}_{\alpha,\mu} \in T^{*-1}(\mathcal{M}_{\max(\mu,1/2)})$.

(b) For $m_0 \in \mathcal{M}_{\mu,\rho}$, both $\overline{m}_{\alpha,\mu}$ and $\overline{h}_{\alpha,\mu}$ are bounded by $C\rho$, for some $C > 0$, implying

$$\begin{aligned} N_{1,1} &\leq C\rho\alpha^{\min(\mu-1/2,1/2)}||\widehat{T}_n - T||_{\mathcal{M}_{\max(\mu-1,0)} \to L^2(\mathcal{W})}, \text{ and} \\ N_{1,2} &\leq C\rho\alpha^{\min(\mu-1/2,0)}||\widehat{T}_n^* - T^*||_{T^{*-1}(\mathcal{M}_{\max(\mu,1/2)}) \to L^2(\mathcal{X})}, \end{aligned}$$

which completes the proof. ■

Proof of Theorem 4.2 Let $\Gamma_{q,\alpha} = U_{q,\alpha}(T^*T)T^*T$ and $\widehat{\Gamma}_{q,\alpha} = U_{q,\alpha}(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{T}_n$. We use the error decomposition in (36) whose first and second terms are analyzed in the same as before, since $U_{q,\alpha}(\cdot)$ satisfies C.3.1 and C.3.2. The uniform bound of $(\Gamma_{\alpha} - I)(T^*T)^{\mu}$, which is equal to $C\alpha^{\min(\mu,q)}$, follows again from Lemma 3.3, since the qualification of $U_{q,\alpha}(\cdot)$ is equal to q. Using

$$\lambda U_{q,\alpha}(\lambda) = \frac{(\alpha+\lambda)^q - \alpha^q}{(\alpha+\lambda)^q} = 1 - [\alpha(\alpha+\lambda)^{-1}]^q,$$

we have, by spectral calculus, that

$$\widehat{\Gamma}_{q,\alpha} - \Gamma_{q,\alpha} = -(\widehat{E}_{\alpha}^q - E_{\alpha}^q),$$

where $\widehat{E}_{\alpha} = \alpha(\alpha + \widehat{T}_n^*\widehat{T}_n)^{-1}$ and $E_{\alpha} = \alpha(\alpha + T^*T)^{-1}$. By adding and subtracting $\widehat{E}_{\alpha}E_{\alpha}^{q-1}(T^*T)^{\mu}$, the last term of (36) is equivalent to

$$\begin{aligned} \Delta_q &\equiv (\widehat{\Gamma}_{q,\alpha} - \Gamma_{q,\alpha})(T^*T)^{\mu} = -(\widehat{E}_{\alpha}^q - E_{\alpha}^q)(T^*T)^{\mu} \\ &= -[\widehat{E}_{\alpha}(\widehat{E}_{\alpha}^{q-1} - E_{\alpha}^{q-1})(T^*T)^{\mu} + (\widehat{E}_{\alpha} - E_{\alpha})E_{\alpha}^{q-1}(T^*T)^{\mu}] \\ &= \widehat{E}_{\alpha}\Delta_{q-1} + \Delta_1 E_{\alpha}^{q-1}, \end{aligned}$$

where the last equality comes from $[\alpha(\alpha I + T^*T)]^q(T^*T)^{\mu} = (T^*T)^{\mu}[\alpha(\alpha I + T^*T)]^q$. Using backward induction, we rewrite Δ_q as a weighted sum of Δ_1

$$\Delta_q m_{\mu} = \sum_{j=0}^{q-1} \widehat{E}_{\alpha}^j \Delta_1 E_{\alpha}^{q-1-j} m_{\mu},$$

which, by the triangle inequality and uniform boundedness of $\widehat{E}_\alpha$ and E_α, yields

$$||\Delta_q m_\mu||_{L^2(\mathcal{X})} \leq C||\Delta_1 m_\mu||_{L^2(\mathcal{X})}, \text{ for } q \text{ finite.}$$

By the results on $\Delta_1 m_\mu \equiv (\widehat{\Gamma}_{1,\alpha} - \Gamma_{1,\alpha})(T^*T)^\mu m_\mu$ given in the proof of Theorem 4.1, the proof is completed. ■

Proof of Theorem 4.3 Since $U^g_{q,\alpha}(\cdot)$ satisfies C.3.1 and C.3.2, with $\overline{\mu}_{GTR_q} = q$, we get the same conclusion as in the proof of ITR(q), for the first two terms in the relevant error decomposition for GTR(q) according to (36). It suffices to show the order of $(\widehat{\Gamma}^g_{q,\alpha} - \Gamma^g_{q,\alpha})(T^*T)^\mu m_\mu$, where $\Gamma^g_{q,\alpha} = U^g_{q,\alpha}(T^*T)T^*T$ and $\widehat{\Gamma}^g_{q,\alpha} = U^g_{q,\alpha}(\widehat{T}^*_n\widehat{T}_n)\widehat{T}^*_n\widehat{T}_n$. From

$$\lambda U^g_{q,\alpha}(\lambda) = 1 - \alpha^q(\alpha^q + \lambda^q)^{-1},$$

a straightforward calculation gives

$$\begin{aligned}
\widehat{\Gamma}^g_{q,\alpha} - \Gamma^g_{q,\alpha} &= -\alpha^q[\widehat{V}_{q,\alpha} - V_{q,\alpha}] = \alpha^q\widehat{V}_{q,\alpha}[(\widehat{T}^*_n\widehat{T}_n)^q - (T^*T)^q]V_{q,\alpha} \\
&= \alpha^q\widehat{V}_{q,\alpha}\widehat{T}^*_n\widehat{T}_n[(\widehat{T}^*_n\widehat{T}_n)^{q-1} - (T^*T)^{q-1}]V_{q,\alpha} + \alpha^q\widehat{V}_{q,\alpha}(\widehat{T}^*_n\widehat{T}_n - T^*T)(T^*T)^{q-1}V_{q,\alpha},
\end{aligned}$$

where $\widehat{V}_{q,\alpha} = [\alpha^q I + (\widehat{T}^*_n\widehat{T}_n)^q]^{-1}$ and $V_{q,\alpha} = [\alpha^q I + (T^*T)^q]^{-1}$. If we multiply and divide simultaneously the first term of $\widehat{\Gamma}^g_{q,\alpha} - \Gamma^g_{q,\alpha}$ by both $\widehat{V}_{q-1,\alpha}$ and $V_{q-1,\alpha}$, and the second term, both by $\widehat{V}_{1,\alpha}$ and $V_{1,\alpha}$, we get

$$\begin{aligned}
(\widehat{\Gamma}^g_{q,\alpha} - \Gamma^g_{q,\alpha})(T^*T)^\mu &= \widehat{V}_{q,\alpha}\widehat{T}^*_n\widehat{T}_n\widehat{V}^{-1}_{q-1,\alpha}\left\{\alpha^{q-1}\widehat{V}_{q-1,\alpha}[(\widehat{T}^*_n\widehat{T}_n)^{q-1} - (T^*T)^{q-1}]V_{q-1,\alpha}(T^*T)^\mu\right\}\alpha V^{-1}_{q-1,\alpha}V_{q,\alpha} \\
&\quad + \alpha^{q-1}\widehat{V}_{q,\alpha}\widehat{V}^{-1}_{1,\alpha}\left\{\alpha\widehat{V}_{1,\alpha}(\widehat{T}^*_n\widehat{T}_n - T^*T)V_{1,\alpha}(T^*T)^\mu\right\}V^{-1}_{1,\alpha}(T^*T)^{q-1}V_{q,\alpha} \\
&= D_{1,q}[(\widehat{\Gamma}^g_{q-1,\alpha} - \Gamma^g_{q-1,\alpha})(T^*T)^\mu]D_{2,q} + D_{3,q}[(\widehat{\Gamma}^g_{1,\alpha} - \Gamma^g_{1,\alpha})(T^*T)^\mu]D_{4,q},
\end{aligned}$$

where

$$\begin{aligned}
D_{1,q} &= [\alpha^q I + (\widehat{T}^*_n\widehat{T}_n)^q]^{-1}[\alpha^{q-1}\widehat{T}^*_n\widehat{T}_n + (\widehat{T}^*_n\widehat{T}_n)^q], \\
D_{2,q} &= [\alpha^q I + (T^*T)^q]^{-1}[\alpha^q I + \alpha(T^*T)^{q-1}], \\
D_{3,q} &= [\alpha^q I + (\widehat{T}^*_n\widehat{T}_n)^q]^{-1}[\alpha^q I + \alpha^{q-1}\widehat{T}^*_n\widehat{T}_n], \\
D_{4,q} &= [\alpha^q I + (T^*T)^q]^{-1}[\alpha(T^*T)^{q-1} + (T^*T)^q].
\end{aligned}$$

We claim that, $D^*_{(p,q)} \equiv [\alpha^q I + (T^*T)^q]^{-1}\alpha^p(T^*T)^{q-p}$ is bounded, for any p and q such that $q \geq p \geq 0$. To see this, we only need to observe that

$$\begin{aligned}
D^*_{(p,q)} &= \alpha^{-(q-p)}\{\alpha^q[\alpha^q I + (T^*T)^q]^{-1}\}(T^*T)^{q-p} \\
&= \alpha^{-(q-p)}\{[I - U^g_{q,\alpha}(T^*T)T^*T](T^*T)^{q-p}\} \\
&\leq C,
\end{aligned}$$

where the last inequality follows by Lemma 3.3(i). For the same reason, $\widehat{D}^*_{(p,q)} \equiv [\alpha^q I + (\widehat{T}_n^*\widehat{T}_n)^q]^{-1}\alpha^p(\widehat{T}_n^*\widehat{T}_n)^{q-p}$ is bounded uniformly in n. Consequently, each $D_{i,q}$ (for $i = 1, .., 4$), which is a linear combination of $D^*_{(p,q)}$'s or $\widehat{D}^*_{(p,q)}$'s, is also bounded, implying

$$||\Delta_q^g m_\mu|| \equiv ||(\widehat{\Gamma}^g_{q,\alpha} - \Gamma^g_{q,\alpha})(T^*T)^\mu m_\mu||_{L^2(\mathcal{X})} \leq C_1||\Delta^g_{q-1}m_\mu||_{L^2(\mathcal{X})} + C_2||\Delta^g_1 m_\mu||_{L^2(\mathcal{X})},$$

which, by backward induction, leads to

$$||(\widehat{\Gamma}^g_{q,\alpha} - \Gamma^g_{q,\alpha})(T^*T)^\mu m_\mu||_{L^2(\mathcal{X})} \leq C||(\widehat{\Gamma}^g_{1,\alpha} - \Gamma^g_{1,\alpha})(T^*T)^\mu m_\mu||_{L^2(\mathcal{X})}, \text{ for } q \text{ finite.}$$

This completes the proof, since GTR of order one is equivalent to OTR, i.e., $\widehat{\Gamma}^g_{1,\alpha} - \Gamma^g_{1,\alpha} = \widehat{\Gamma}_{1,\alpha} - \Gamma_{1,\alpha}$. ■

Proof of Theorem 4.4 With C.3.1 satisfied by $U^s_\alpha(\cdot)$, the first term of (36) is analyzed in the same as above. Due to the infinite qualification of Showalter's method, the pure regularization bias is of order $O(\alpha^\mu)$, for any $\mu > 0$. Let $\Gamma^s_\alpha = U^s_\alpha(\cdot)(T^*T)T^*T$ and $\widehat{\Gamma}^s_\alpha = U^s_\alpha(\cdot)(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{T}_n$. Using $\lambda U^s_\alpha(\cdot)(\lambda) = 1 - \exp(-\lambda/\alpha)$, and $\exp(x) = \sum_{j=0}^\infty \frac{x^j}{j!}$, we obtain, by spectral calculus,

$$\begin{aligned}
\widehat{\Gamma}^s_\alpha - \Gamma^s_\alpha &= -[\exp(-\widehat{T}_n^*\widehat{T}_n/\alpha) - \exp(-T^*T/\alpha)] \\
&= \exp(-\widehat{T}_n^*\widehat{T}_n/\alpha)[\exp(\widehat{T}_n^*\widehat{T}_n/\alpha) - \exp(T^*T/\alpha)]\exp(-T^*T/\alpha) \\
&= \exp(-\widehat{T}_n^*\widehat{T}_n/\alpha)\sum_{j=0}^\infty \frac{1}{j!}(1/\alpha)^j[(\widehat{T}_n^*\widehat{T}_n)^j - (T^*T)^j]\exp(-T^*T/\alpha) \\
&= \exp(-\widehat{T}_n^*\widehat{T}_n/\alpha)\sum_{j=1}^\infty \frac{1}{j!}(1/\alpha)^j[(\widehat{T}_n^*\widehat{T}_n)^j - (T^*T)^j]\exp(-T^*T/\alpha),
\end{aligned}$$

since $(\widehat{T}_n^*\widehat{T}_n)^0 = (T^*T)^0 = I$. From the identity

$$(\widehat{T}_n^*\widehat{T}_n)^j - (T^*T)^j = \sum_{k=0}^{j-1}(\widehat{T}_n^*\widehat{T}_n)^k(\widehat{T}_n^*\widehat{T}_n - T^*T)(T^*T)^{j-1-k},$$

follows

$$\begin{aligned}
&(\widehat{\Gamma}^s_\alpha - \Gamma^s_\alpha)(T^*T)^\mu \\
= \; &\exp(-\widehat{T}_n^*\widehat{T}_n/\alpha)\sum_{j=1}^\infty \frac{1}{j!}(1/\alpha)^j\sum_{k=0}^{j-1}(\widehat{T}_n^*\widehat{T}_n)^k(\widehat{T}_n^*\widehat{T}_n - T^*T)(T^*T)^{j-1-k}\exp(-T^*T/\alpha)(T^*T)^\mu \\
= \; &\sum_{j=1}^\infty \frac{1}{j!}(1/\alpha)^j\sum_{k=0}^{j-1}\exp(-\widehat{T}_n^*\widehat{T}_n/\alpha)(\widehat{T}_n^*\widehat{T}_n)^k\widehat{T}_n^*(\widehat{T}_n - T)(T^*T)^{j-1-k}\exp(-T^*T/\alpha)(T^*T)^\mu \\
&+\sum_{j=1}^\infty \frac{1}{j!}(1/\alpha)^j\sum_{k=0}^{j-1}\exp(-\widehat{T}_n^*\widehat{T}_n/\alpha)(\widehat{T}_n^*\widehat{T}_n)^k(\widehat{T}_n^* - T^*)T(T^*T)^{j-1-k}\exp(-T^*T/\alpha)(T^*T)^\mu.
\end{aligned}$$

Applying a similar argument used in Lemma 3.3, we can calculate the bound of each term in the above infinite sum. For example,

$$\begin{aligned}
&||T(T^*T)^{j-1-k}\exp(-T^*T/\alpha)(T^*T)^{\mu}||_{L^2(\mathcal{X})\to L^2(\mathcal{W})}\\
\le\ & \sup_{0<\lambda\le\overline{\lambda}} |\exp(-\lambda/\alpha)\lambda^{j-1/2-k+\mu}| \ \le C\alpha^{j-1/2-k+\mu},
\end{aligned}$$

and likewise,

$$|| \exp(-\widehat{T}_n^*\widehat{T}_n/\alpha)(\widehat{T}_n^*\widehat{T}_n)^k\widehat{T}_n^*||_{L^2(\mathcal{X})\to L^2(\mathcal{X})} \le C\alpha^{k+1/2}.$$

Hence, by the triangle inequality, the asymptotic order of the last term in (36) is given by

$$\begin{aligned}
&||(\widehat{\Gamma}_\alpha^s-\Gamma_\alpha^s)(T^*T)^{\mu}||_{L^2(\mathcal{X})\to L^2(\mathcal{W})}\\
\le\ & C\sum_{j=1}^{\infty}\frac{1}{j!}(1/\alpha)^j\sum_{k=0}^{j-1}[\alpha^{k+1/2}\alpha^{j-1-k+\mu}||\widehat{T}_n-T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})}+\alpha^k\alpha^{j-1/2-k+\mu}||\widehat{T}_n^*-T^*||_{L^2(\mathcal{W})\to L^2(\mathcal{X})}]\\
=\ & C\left(\sum_{j=1}^{\infty}\frac{1}{j!}j\right)\alpha^{\mu-1/2}[||\widehat{T}_n-T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})}+||\widehat{T}_n^*-T^*||_{L^2(\mathcal{W})\to L^2(\mathcal{X})}]\\
=\ & C'\alpha^{\mu-1/2}[||\widehat{T}_n-T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})}+||\widehat{T}_n^*-T^*||_{L^2(\mathcal{W})\to L^2(\mathcal{X})}],
\end{aligned}$$

where we used $\sum_{j=0}^{\infty}\frac{1}{j!}j \le \sum_{j=0}^{\infty}\frac{1}{j!}2^j = e^2 < \infty$. ■

A.4 Section 5

Proof of Theorem 5.1

Step I (a matrix form of $\widehat{T}_n^*\widehat{T}_n$): Let $\widehat{g}_{XX}(\cdot,\cdot)$ be the kernel of the self-adjoint operator $\widehat{T}_n^*\widehat{T}_n : L^2(\mathcal{X})\to L^2(\mathcal{X})$, i.e.,

$$\widehat{g}_{XX}(x,u)=\int_{\mathcal{W}}\widehat{f}_{X,W}(x,w)\widehat{f}_{X,W}(u,w)dw.$$

By a straightforward calculation, $\widehat{g}_{XX}(\cdot,\cdot)$ is written, in a matrix form, as

$$\widehat{g}_{XX}(x,u)=n^{-2}K_n^X(x)'M_WK_n^X(u).$$

Plugging in $\widehat{g}_{XX}(\cdot,\cdot)$ into the operator $\widehat{T}_n^*\widehat{T}_n$ yields

$$\begin{aligned}
(\widehat{T}_n^*\widehat{T}_nm)(x) &= \int\widehat{g}_{XX}(x,u)m(u)du = n^{-2}K_n^X(x)'M_W\int_{\mathcal{X}}K_n^X(u)m(u)du\\
&= n^{-2}K_n^X(x)'M_W < K_n^X(\cdot),\ m(\cdot) >_{L^2(\mathcal{X})} . \qquad (37)
\end{aligned}$$

Step II (the spectral representation of $\widehat{T}_n^*\widehat{T}_n$): Let $\{(\lambda_s, e_s)\}_{s=1}^{n'}$ denote all the nonzero eigenvalues and the corresponding eigenvectors of $Q_{X,W} = n^{-2}M_W^{1/2}M_X M_W^{1/2}$, where $n' = \text{rank}(Q_{X,W}) \leq n$. Define

$$v_s(x) = K_n^X(x)' M_W^{1/2} e_s. \tag{38}$$

We claim that the spectral representation of the compact self-adjoint operator $\widehat{T}_n^*\widehat{T}_n$ is given by

$$\widehat{T}_n^*\widehat{T}_n = \sum_{s=1}^{n'} \lambda_s P_{v_s},$$

where P_{v_s} denotes the orthogonal projection on the subspace generated by the function v_s. To prove the claim, it suffices to show that all the nonzero eigenvalues and the corresponding eigenfunctions of $\widehat{T}_n^*\widehat{T}_n$ are given by $\{(\lambda_s, v_s)\}_{s=1}^{n'}$. From the definition of (λ_s, e_s), it follows that

$$\begin{aligned}
(\widehat{T}_n^*\widehat{T}_n v_s)(x) &= n^{-2}K_n^X(x)' M_W < K_n^X,\ v_s >_{L^2(\mathcal{X})} \\
&= n^{-2}K_n^X(x)' M_W < K_n^X,\ K_n^{X\prime} >_{L^2(\mathcal{X})} M_W^{1/2} e_s \\
&= K_n^X(x)' M_W^{1/2}(n^{-2}M_W^{1/2}M_X M_W^{1/2}) e_s \\
&= K_n^X(x)' M_W^{1/2}(\lambda_s e_s) = \lambda_s v_s,
\end{aligned}$$

implying that $\{(\lambda_s, v_s)\}_{s=1}^{n'}$ is a subset of the eigensystem of $\widehat{T}_n^*\widehat{T}_n$ corresponding to the nonzero eigenvalues. From $\dim(\mathcal{R}(\widehat{T}_n^*\widehat{T}_n)) = \min[\dim(\text{lin}\{K_n^X(\cdot)\}), \dim(\text{lin}\{K_n^W(\cdot)\})] = \text{rank}(Q_{X,W})$, the number of nonzero eigenvalues of $\widehat{T}_n^*\widehat{T}_n$ is equal to n', completing the proof for the claim.

Step III (the spectral representation of $r(\widehat{T}_n^*\widehat{T}_n)$): From the theorem on spectral calculus-see, Taylor and Lay (1980, p.368, for example), we obtain the spectral representation of $r(\widehat{T}_n^*\widehat{T}_n)$

$$r(\widehat{T}_n^*\widehat{T}_n)(\cdot) = \sum_{s=1}^{n} r(\lambda_s) P_{v_s}(\cdot) = \sum_{s=1}^{n} r(\lambda_s) v_s (< v_s, v_s >_{L^2(\mathcal{X})})^{-1} < v_s, \cdot >_{L^2(\mathcal{X})} .$$

By plugging in (38) into the above equation,

$$\begin{aligned}
&[r(\widehat{T}_n^*\widehat{T}_n)m](x) \\
= &\sum_{s=1}^{n} r(\lambda_s) K_n^X(x)' M_W^{1/2} e_s (e_s' M_W^{1/2} M_X M_W^{1/2} e_s)^{-1} e_s' M_W^{1/2} < K_n^X,\ m >_{L^2(\mathcal{X})} ,
\end{aligned}$$

which, by definition of e_s, reduces to

$$n^{-2}K_n^X(x)' M_W^{1/2} \left[\sum_{s=1}^{n} r(\lambda_s) \lambda_s^{-1} P_{e_s} \right] M_W^{1/2} < K_n^X,\ m >_{L^2(\mathcal{X})}$$

$$= n^{-2}K_n^X(x)'M_W^{1/2}r(Q_{X,W})Q_{X,W}^{-1}M_W^{1/2} < K_n^X,\ m >_{L^2(\mathcal{X})} .$$

Step IV (Closed form of $R_n^\alpha(\widehat{h}_n)$): From $\widehat{h}_n(w) = n^{-1}K_n^W(w)\mathbf{y}$,

$$\begin{aligned}(\widehat{T}_n^*\widehat{h}_n)(x) &= \int_{\mathcal{W}} \widehat{h}_n(w)\widehat{f}_{X,W}(x,w)dw = n^{-2}K_n^X(x)'[\int_{\mathcal{W}} K_n^W(w)K_n^W(w)'dw]\mathbf{y} \\ &= n^{-2}K_n^X(x)'M_W\mathbf{y}.\end{aligned}$$

Therefore,

$$\begin{aligned}R_n^\alpha(\widehat{h}_n)(x) &= [U_\alpha(\widehat{T}_n^*\widehat{T}_n)\widehat{T}_n^*\widehat{h}_n](x) \\ &= n^{-2}K_n^X(x)'M_W^{1/2}U_\alpha(Q_{X,W})Q_{X,W}^{-1}M_W^{1/2} < K_n^X(\cdot),\ n^{-2}K_n^X(\cdot)'M_W >_{L^2(\mathcal{X})} \mathbf{y} \\ &= n^{-2}K_n^X(x)'M_W^{1/2}U_\alpha(Q_{X,W})M_W^{1/2}\mathbf{y}.\end{aligned}$$

■

Proof of Theorem 5.2

Given the common element fixed to be $w_1 \in \mathcal{W}_1$, we define $\widehat{T}_{w_1,n} : L^2_{w_1}(\mathcal{Z}) \to L^2_{w_1}(\mathcal{W}_2)$ by

$$\widehat{T}_{w_1,n}(m)(w_2) = \int_{\mathcal{Z}} m(u,w_1)\widehat{f}_{Z,W_1,W_2}(u,w_1,w_2)du,$$

and likewise, $\widehat{T}^*_{w_1,n} : L^2_{w_1}(\mathcal{W}_2) \to L^2_{w_1}(\mathcal{Z})$ by $\widehat{T}^*_{w_1,n}(h)(z) = \int_{\mathcal{W}_2} h(w_1,w_2)\widehat{f}_{Z,W_1,W_2}(z,w_1,w_2)dw_2$, where $L^2_{w_1}(\mathcal{Z}) = \{m(\cdot,w_1) \in L^2(\mathcal{Z})\}$ and $L^2_{w_1}(\mathcal{W}_2) = \{h(w_1,\cdot) \in L^2(\mathcal{W}_1)\}$. Note that the kernel of the self-adjoint operator $\widehat{T}^*_{w_1,n}\widehat{T}_{w_1,n} : L^2_{w_1}(\mathcal{Z}) \to L^2_{w_1}(\mathcal{Z})$ is given by

$$\begin{aligned}\widehat{g}_{ZZ}(u,z;w_1) &= \int_{\mathcal{W}_2} \widehat{f}_{Z,W_1,W_2}(u,w_1,w_2)\widehat{f}_{Z,W_1,W_2}(z,w_1,w_2)dw_2 \\ &= n^{-2}K_n^X(u,w_1)'M_{W_2}K_n^X(z,w_1),\end{aligned}$$

yielding

$$\begin{aligned}(\widehat{T}^*_{w_1,n}\widehat{T}_{w_1,n}m)(z) &= \int_{\mathcal{Z}} \widehat{g}_{ZZ}(u,z;w_1)m(u,w_1)du \\ &= n^{-2}K_n^X(z,w_1)'M_{W_2} < K_n^X(\cdot,w_1),\ m(\cdot,w_1) >_{L^2(\mathcal{Z})} .\end{aligned}$$

Let $\{(\lambda_{w_1,s}, e_{w_1,s})\}_{s=1}^n$ be the nonzero eigenvalues and the corresponding eigenvectors of $Q_{Z,W}(w_1)$. From a similar argument to Step II in the proof of Theorem 5.1, we obtain the spectral representation of $\widehat{T}^*_{w_1,n}\widehat{T}_{w_1,n}$, which is given by

$$\widehat{T}^*_{w_1,n}\widehat{T}_{w_1,n} = \sum_{s=1}^{n'} \lambda_{w_1,s}P_{v_{w_1,s}},$$

where $P_{v_{w_1,s}}$ is the orthogonal projection on the eiegenspace generated by the eigenfunction $v_{w_1,s}(z)$ $= n^{-1}K_n^X(z,w_1)'M_{W_2}^{1/2}e_{w_1,s}$. Applying spectral calculus, we have that, for $m(\cdot,w_1) \in L^2_{w_1}(\mathcal{Z})$,

$$\begin{aligned}
& [r(\widehat{T}^*_{w_1,n}\widehat{T}_{w_1,n})m](z) \\
= & \sum_{s=1}^{n} r(\lambda_{w_1,s})P_{v_{w_1,s}}(m)(z) \\
= & n^{-2}K_n^X(z,w_1)'M_{W_2}^{1/2}r(Q_{Z,W}(w_1))Q_{Z,W}^{-1}(w_1)M_{W_2}^{1/2} < K_n^X(\cdot,w_1),\ m(\cdot,w_1) >_{L^2(\mathcal{Z})}.
\end{aligned}$$

Observing that

$$\begin{aligned}
(\widehat{T}^*_{w_1,n}\widehat{h}_n)(z) &= \int_{\mathcal{W}_2}\widehat{f}_{Z,W_1,W_2}(z,w_1,w_2)\widehat{h}_n(w_1,w_2)dw_2 \\
&= n^{-2}\sum_{1\le i\le n}\sum_{1\le j\le n}K_h(Z_i-z)K_h(W_{1i}-w_1)M_{ij}^{W_2}K_h(W_{1j}-w_1)y_j \\
&= n^{-2}K_n^X(z,w_1)'M_{W_2}[K_n^{W_1}(w_1)\odot\mathbf{y}],
\end{aligned}$$

we now get

$$\begin{aligned}
& \widehat{m}_{\alpha,n}(z,w_1) \\
= & [U_\alpha(\widehat{T}^*_n\widehat{T}_n)(\widehat{T}^*_n\widehat{h}_n)](z,w_1) \\
= & U_\alpha(\widehat{T}^*_{w_1,n}\widehat{T}_{w_1,n})(\widehat{T}^*_{w_1,n}\widehat{h}_n)(z) \\
= & n^{-2}K_n^X(z,w_1)'M_{W_2}^{1/2}U_\alpha(Q_{Z,W}(w_1))Q_{Z,W}^{-1}(w_1)M_{W_2}^{1/2} < K_n^X(\cdot,w_1),\ (\widehat{T}^*_{w_1,n}\widehat{h}_n)(\cdot) >_{L^2(\mathcal{Z})} \\
= & n^{-2}K_n^X(z,w_1)'M_{W_2}^{1/2}U_\alpha(Q_{Z,W}(w_1))M_{W_2}^{1/2}[K_n^{W_1}(w_1)\odot\mathbf{y}].
\end{aligned}$$

■

Proof of Proposition 5.3 With $a*b$ denoting convolution of a and b, we define

$$m_{c(g_1)}(z,w_1) \equiv (K_{g_1}*m)(z,w_1) = \int K_{g_1}(z-s)m(s,w_1)ds, \tag{39}$$

and

$$\begin{aligned}
f_{Z,W}^{c(g)}(z,w) &\equiv (K_{(g_1,g_2)}*f_{Z,W})(z,w) = \int_{\mathcal{W}}\int_{\mathcal{Z}}K_{g_1}(s_1-z)K_{g_2}(s_2-w)f_{Z,W}(s_1,s_2)ds_1ds_2 \\
&= \mathrm{E}[K_{g_1}(Z_i-z)K_{g_2}(W_i-w)].
\end{aligned}$$

By adding and subtracting $\int f_{Z,W}^{c(g)}(z,w)m(z,w_1)dz$, the estimation errors of $\widehat{T}_n$ are decomposed into

$$\begin{aligned}
& (\widehat{T}_n m - Tm)(w) \\
= \;& n^{-1}\sum_{i=1}^{n}\int [K_{g_1}(Z_i - z)K_{g_2}(W_i - w) - f_{Z,W}^{c(g)}(z,w)]m(z,w_1)dz \\
& + \int [f_{Z,W}^{c(g)}(z,w) - f_{Z,W}(z,w)]m(z,w_1)dz \\
\equiv \;& \mathrm{s}_n(w) + \mathrm{B}_n(w),
\end{aligned}$$

from which we obtain the MISE of $\widehat{T}_n m$, given by

$$\mathrm{E}\int_{\mathcal{W}}\left[(\widehat{T}_n - T)m\right]^2(w)dw = \int_{\mathcal{W}}\left\{\mathrm{Var}\left[\mathrm{s}_n(w)\right] + \mathrm{E}^2\left[\mathrm{B}_n(w)\right]\right\}dw.$$

Noting that

$$\mathrm{s}_n(w) = n^{-1}\sum_{i=1}^{n}\{K_{g_2}(W_i - w)m_{c(g_1)}(Z_i,w_1) - \mathrm{E}[K_{g_2}(W_i - w)m_{c(g_1)}(Z_i,w_1)]\},$$

the standard calculation of the variance term (under the i.i.d. assumption in C.5.1) yields

$$\begin{aligned}
\mathrm{Var}\left[s_n(w)\right] &= n^{-1}\mathrm{Var}\left[K_{g_2}(W_i - w)m_{c(g_1)}(Z_i,w_1)\right] \\
&= n^{-1}g_2^{-d_2}\int_{\mathcal{Z}} m_{c(g_1)}^2(z,w_1)\left[\int_{\mathcal{W}} K^2(u)f_{Z,W}(z,w+g_2u)du\right]dz + O(n^{-1}),
\end{aligned}$$

implying that

$$\int_{\mathcal{W}}\mathrm{Var}\left[s_n(w)\right]dw \leq \frac{C}{ng_2^{d_2}}||K||_2^2||m||_{L^2(\mathcal{X})}^2,$$

where the last inequality is due to the dominated convergence theorem and boundedness of $f_{Z,W}(\cdot)$ in C.5.3. To calculate the bias term, we observe that, by Cauchy-Schwartz inequality,

$$\begin{aligned}
\int_{\mathcal{W}}\mathrm{B}_n^2(w)dw &= \int_{\mathcal{W}}[\int_{\mathcal{Z}}\{f_{Z,W}^c(z,w) - f_{Z,W}(z,w)\}m(z,w_1)dz]^2dw \\
&\leq ||f_{Z,W}^c - f_{Z,W}||_{L^2(\mathcal{Z}\times\mathcal{W})}^2||m||_{L^2(\mathcal{X})}^2,
\end{aligned}$$

leading to

$$\mathrm{E}\int_{\mathcal{W}}[(\widehat{T}_n m - Tm)(w)]^2dw \leq ||m||_{L^2(\mathcal{X})}^2(||f_{Z,W}^c - f_{Z,W}||_{L^2(\mathcal{Z}\times\mathcal{W})}^2 + \frac{C}{ng_2^{d_2}}||K||_2^2),$$

i.e.,

$$\mathrm{E}||\widehat{T}_n - T||_{L^2(\mathcal{X})\to L^2(\mathcal{W})}^2 = \sup_{m(\neq 0)\in L^2(\mathcal{X})}\frac{\mathrm{E}\int_{\mathcal{W}}\left[(\widehat{T}_n - T)m\right]^2(w)dw}{||m||_{L^2(\mathcal{X})}^2} \leq ||f_{Z,W}^c - f_{Z,W}||_{L^2(\mathcal{Z}\times\mathcal{W})}^2 + \frac{C}{ng_2^{d_2}}||K||_2^2.$$

Under C.5.2 (i.e., $\int |K(s)|ds < \infty$ and $\sup |K(s)| < \infty$), the convolution error ($||f^c_{Z,W} - f_{Z,W}||^2_{L^2(\mathcal{Z}\times\mathcal{W})}$) converges to zero, as g_1 and g_2 go to zero, for any square integrable $f_{Z,W}(\cdot,\cdot)$. This, together with the bandwidth condition, $ng_2^{d_2} \to 0$, gives rise to uniform consistency of $\widehat{T}_n$, proving part (i). When there exist p_0-th partial derivatives of $f_{Z,W}(\cdot,\cdot)$ that are continuous and square integrable-i.e., C.5.4 holds, we have, by application of the standard Taylor expansion, that

$$\mathrm{E}||(\widehat{T}_n - T)m||^2_{L^2(\mathcal{X})\to L^2(\mathcal{W})} = O(g_1^{2p_0} + g_2^{2p_0}) + O(\frac{1}{ng_2^{d_2}}).$$

By symmetry of the above arguments, we also get the convergence rate of $\widehat{T}^*_n$.

It remains to prove part (iii). Let $r(Z_i, W_{1i}) = m(Z_i, W_{1i}) - m_{c(g_1)}(Z_i, w_1)$, where $m_{c(g_1)}(\cdot,\cdot)$ is defined by (39). From $(\widehat{T}_n m_0)(w) = n^{-1}\sum_{i=1}^n K_{g_2}(W_i - w)m_{c(g_1)}(Z_i, w_1)$, we get

$$\begin{aligned} & (\widehat{h}_{0,n} - \widehat{T}_n m_0)(w) \\ = \;& n^{-1}\sum_{i=1}^n K_{g_2}(W_i - w)\varepsilon_i + n^{-1}\sum_{i=1}^n K_{g_2}(W_i - w)r(Z_i, W_{1i}) \\ = \;& n^{-1}\sum_{i=1}^n K_{g_2}(W_i - w)\varepsilon_i + n^{-1}\sum_{i=1}^n K_{g_2}(W_i - w)(\nu_i - \nu_i^c) + n^{-1}\sum_{i=1}^n K_{g_2}(W_i - w)\mathrm{E}(r(Z_i, W_{1i})|W_i) \\ \equiv \;& s_{1,n}(w) + s_{2,n}(w) + \mathrm{B}_n(w), \end{aligned}$$

where $\nu_i = m(Z_i, W_{1i}) - \mathrm{E}(m(Z_i, W_{1i})|W_i)$, and $\nu_i^c = m_{c(g_1)}(Z_i, w_1) - \mathrm{E}(m_{c(g_1)}(Z_i, w_1)|W_i)$. As a consequence,

$$\mathrm{E}\int_{\mathcal{W}} (\widehat{h}_{0,n} - \widehat{T}_n m_0)^2(w)dw = \int_{\mathcal{W}} \left\{\mathrm{Var}\left[s_{1,n}(w) + s_{2,n}(w)\right] + \mathrm{E}^2[\mathrm{B}_n(w)]\right\} dw.$$

By the standard argument in kernel regression, the variance of the main stochastic term is calculated in a straightforward way;

$$\mathrm{Var}\left[s_{1,n}(w)\right] = \frac{1}{ng_2^{d_2}}||K||_2^2\mathrm{E}(\varepsilon_i^2|W_i = w)f_W(w)(1 + o(1)),$$

leading to

$$\begin{aligned} \int_{\mathcal{W}} \mathrm{Var}\left[s_{1,n}(w)\right] dw &= \frac{1}{ng_2^{d_2}}||K||_2^2 \left[\int_{\mathcal{W}} \mathrm{E}(\varepsilon_i^2|W_i = w)f_W(w)dw\right](1 + o(1)) \\ &= \frac{1}{ng_2^{d_2}}||K||_2^2\sigma_\varepsilon^2(1 + o(1)) = O(\frac{1}{ng_2^{d_2}}), \end{aligned}$$

where $\sigma_\varepsilon^2 = \mathrm{E}(\varepsilon_i^2)$. From $\mathrm{E}[\nu_i|W_i] = \mathrm{E}[\nu_i^c|W_i] = 0$, it follows that

$$\mathrm{E}[(\nu_i - \nu_i^c)^2|W_i] = \mathrm{Var}[r(Z_i, W_{1i})|W_i] \le \mathrm{E}[r^2(Z_i, W_{1i})|W_i],$$

implying, by the iid assumption and the law of iterated expectation, that

$$\begin{aligned}\mathrm{Var}\left[s_{2,n}(w)\right] &= n^{-1}\mathrm{E}\{[K_{g_2}(W_i-w)(\nu_i-\nu_i^c)]^2\} \leq n^{-1}\mathrm{E}\{[K_{g_2}(W_i-w)]^2\mathrm{E}[r^2(Z_i,W_{1i})|W_i]\}\\ &= n^{-1}\mathrm{E}\{[K_{g_2}(W_i-w)r(Z_i,W_{1i})]^2\}\\ &= \frac{1}{ng_2^{d_2}}||K||_2^2\int_{\mathcal{Z}} r^2(z,w_1)f_{Z,W}(z,w)dz(1+o(1)).\end{aligned}$$

From boundedness of $f_{Z,W}(\cdot,\cdot)$,

$$\int_{\mathcal{W}}\mathrm{Var}\left[s_{2,n}(w)\right]dw \leq \frac{C}{ng_2^{d_2}}||K||_2^2||m(\cdot,\cdot)-m_{c(g_1)}(\cdot,\cdot)||^2_{L^2(\mathcal{X})} = o(\frac{1}{ng_2^{d_2}}),$$

since the convolution error, $||m(\cdot,\cdot)-m_{c(g_1)}(\cdot,\cdot)||_{L^2(\mathcal{X})}$, converges to zero, as $g_1\to 0$. To calculate the bias term, we note, by the dominated convergence theorem, that

$$\begin{aligned}\mathrm{E}[\mathrm{B}_n(w)] &= \mathrm{E}[K_{g_2}(W_i-w)r(Z_i,W_{1i})]\\ &= \int_{\mathcal{Z}}[m(z,w_1)-m_{c(g_1)}(z,w_1)]f_{Z,W}(z,w)dz(1+o(1)). \qquad (40)\end{aligned}$$

Letting $f_{Z,W}^{c(g_1)}(u,w)=\int_{\mathcal{Z}}K_{g_1}(z-u)f_{Z,W}(z,w)dz$, we obtain an alternative form of the bias such that

$$\mathrm{E}[\mathrm{B}_n(w)] = \{\int_{\mathcal{Z}}\left[f_{Z,W}(u,w)-f_{Z,W}^{c(g_1)}(u,w)\right]m(u,w_1)du\}(1+o(1)), \qquad (41)$$

since $\int_{\mathcal{Z}}m_{c(g_1)}(z,w_1)f_{Z,W}(z,w)dz = \int_{\mathcal{Z}}f_{Z,W}^{c(g_1)}(u,w)m(u,w_1)du,$, by Fubini's Theorem. By Cauchy-Schwartz inequality, it follows from (40) and (41), together with square-integrability of $f_{Z,W}(\cdot,\cdot)$ and $m(\cdot,\cdot)$, that

$$\int_{\mathcal{W}}\mathrm{E}^2\left[\mathrm{B}_n(w)\right]dw \leq C\min\{||m(\cdot,\cdot)-m_{c(g_1)}(\cdot,\cdot)||^2_{L^2(\mathcal{X})},\ ||f_{Z,W}(\cdot,\cdot)-f_{Z,W}^{c(g_1)}(\cdot,\cdot)||^2_{L^2(\mathcal{Z}\times\mathcal{W})}\},$$

which, by the standard method of Taylor expansion (under C.5.4) gives

$$\int_{\mathcal{W}}\mathrm{E}^2\left[\mathrm{B}_n(w)\right]dw = O(g_1^{\max\{2p_0,2p_1\}}).$$

Letting $\overline{p}=\max(p_0,p_1)$, we finally get

$$\begin{aligned}\mathrm{E}\int_{\mathcal{W}}(\widehat{h}_{0,n}-\widehat{T}_n m_0)^2(w)dw &= \int_{\mathcal{W}}\left\{\mathrm{Var}\left[s_{1,n}(w)+s_{2,n}(w)\right]+\mathrm{E}^2\left[\mathrm{B}_n(w)\right]\right\}dw\\ &= O(\frac{1}{ng_2^{d_2}})+O(g_1^{2\overline{p}}),\end{aligned}$$

i.e.,

$$||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} = O_p(\frac{1}{\sqrt{ng_2^{d_2}}} + g_1^{\overline{p}}).$$

■

Proof of Theorem 5.4 Under C.5.1 through C.5.4, all the conditions of Theorem 3.2 follow from Proposition 5.3.(i) and (iii), proving the consistency result for $\widehat{m}_{\alpha,n}$. Also, Proposition 5.3.(ii) and (iii) applied to Corollary 3.4.(i), leads to, for $m_0 \in \mathcal{M}_{1/2,\rho}$,

$$\begin{aligned} ||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} &\leq O_p(\frac{1}{\sqrt{\alpha}}[\frac{1}{\sqrt{ng_2^{d_2}}} + g_1^{\overline{p}}]) + O_p(\sqrt{\alpha}) + O_p(\frac{1}{\sqrt{ng_1^{d_1}}} + g_1^{p_0} + g_2^{p_0}) \\ &= O_p(\frac{1}{\sqrt{\alpha}}[\frac{1}{\sqrt{ng_2^{d_2}}} + g_1^{p_0}]) + O_p(\sqrt{\alpha}) + O_p(\frac{1}{\sqrt{ng_1^{d_1}}} + g_2^{p_0}), \end{aligned}$$

since $g_1^{p_0} = o(g_1^{\overline{p}}/\sqrt{\alpha})$, from $\alpha_n = o(1)$ and $\overline{p} = p_0 = p_1$, by assumption. In a similar way, Proposition 5.3 and Corollary 3.4.(ii) yield, for $m_0 \in \mathcal{M}_{1,\rho}$,

$$\begin{aligned} ||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} &\leq O_p(\frac{1}{\sqrt{\alpha}}[\frac{1}{\sqrt{ng_2^{d_2}}} + g_1^{\overline{p}}]) + O_p(\alpha) + O_p(\sqrt{\alpha}[\frac{1}{\sqrt{ng_2^{d_2}}} + g_1^{p_0} + g_2^{p_0}]) \\ &\quad + O_p(\frac{1}{\sqrt{ng_1^{d_1}}} + g_1^{p_0} + g_2^{p_0}) \\ &= O_p(\frac{1}{\sqrt{\alpha}}[\frac{1}{\sqrt{ng_2^{d_2}}} + g_1^{p_0}]) + O_p(\alpha) + O_p(\frac{1}{\sqrt{ng_1^{d_1}}} + g_2^{p_0}), \end{aligned}$$

since $\sqrt{\alpha}[1/\sqrt{ng_2^{d_2}} + g_1^{p_0} + g_2^{p_0}] = o(1/\sqrt{n\alpha g_2^{d_2}}) + o(g_1^{p_0} + g_2^{p_0})$. ■

Proof of Theorem 5.5 We only give a proof for part(ii), since part (i) is shown in the same way. We first show that, under the given side condition, the profile of quasi-optimal smoothing parameters are given by $\{(g_1, g_2, \alpha^*)\}$ with $\alpha^* = [\max(1/\sqrt{ng_2^{d_2}}, g_1^{p_0})]^{2/3}$. When the regularization parameter is fixed by $\alpha^* = [\max(1/\sqrt{ng_2^{d_2}}, g_1^{p_0})]^{2/3}$, it follows from the given side condition that $(1/\sqrt{ng_1^{d_1}} + g_2^{p_0}) \leq O(\max\{g_1^{2p_0/3}, [ng_2^{d_2}]^{-1/3}) = O(\alpha^*)$. Consequently, the lower bounds in Theorem 5.4.(ii), corresponding to $m_0 \in \mathcal{M}_{1,\rho}$, reduces to

$$||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} \leq O_p(\frac{1}{\sqrt{\alpha^*}}[1/\sqrt{ng_2^{d_2}} + g_1^{p_0}]) + O_p(\alpha^*) = O_p([\max(1/\sqrt{ng_2^{d_2}}, g_1^{p_0})]^{2/3}).$$

Since $||\widehat{h}_{0,n} - \widehat{T}_n m_0||_{L^2(\mathcal{W})} \equiv O_p(\delta_n) = O_p(\max\{1/\sqrt{ng_2^{d_2}}, g_1^{p_0}\})$, from Proposition 5.3, it also holds by Theorem 3.5 that

$$O_p([\max(1/\sqrt{ng_2^{d_2}}, g_1^{p_0})]^{2/3}) = O_p(\delta_n^{2/3}) \leq ||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})},$$

implying that the actual convergence rate of $\widehat{m}_{\alpha,n}$, given by $O_p([\max(1/\sqrt{ng_2^{d_2}}, g_1^{p_0})]^{2/3})$, is quasi-optimal for $m_0 \in \mathcal{M}_{1,\rho}$. This confirms the quasi-optimality of $\{(g_1, g_2, \alpha^*)\}$. It is not difficult to show suboptimality of regularization parameters other than $\alpha^* = [\max(1/\sqrt{ng_2^{d_2}}, g_1^{p_0})]^{2/3}$. For example, when the regularization parameter is of greater order than α^*, the convergence rate of $\widehat{m}_{\alpha,n}$ is determined, under the side condition, by the dominant term $O_p(\alpha)$ which is greater than $O_p(\delta_n^{2/3})$. We next decide on the fastest possible rate of convergence, out of the quasi-optimal profile $\{(g_1, g_2, \alpha^*)\}$. From $||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} = O_p([\max(1/\sqrt{ng_2^{d_2}}, g_1^{p_0})]^{2/3})$, it is possible to improve the convergence rate of $\widehat{m}_{\alpha,n}$ by making g_2 larger and g_1 smaller, as long as they satisfy the side condition. Obviously, the most favorable choice of (g_1, g_2) is the one under which the side condition hold as an equality. That is, the optimal choice of (g_1, g_2) is given by (g_{1n}^*, g_{2n}^*) such that $(ng_{1n}^{*d_1})^{-1/2} \simeq g_{1n}^{*2p_0/3}$ and $g_{2n}^{*3p_0/2} \simeq (ng_{2n}^{*d_2})^{-1/2}$, leading to $g_{1n}^* = C_0 n^{-\frac{1}{(4/3)p_0+d_1}}$, $g_{2n}^* = C_1 n^{-\frac{1}{3p_0+d_2}}$, and $\alpha_n^* = C_2 n^{-\frac{p_0}{3p_0+d_2}}$. Note that all the basic conditions in C.5.5 are satisfied by $(g_{1n}^*, g_{2n}^*, \alpha_n^*)$. Since $g_{1n}^{*p_0} = C_0 n^{-\frac{3p_0}{4p_0+3d_1}}$ is of smaller order than $1/\sqrt{ng_{2n}^{*d_2}} = O(n^{-\frac{3p_0}{6p_0+2d_2}})$, by the assumptions of $d_1/2 \leq p_0$ and $d_1 \leq d_2$, we now obtain the optimal convergence rate of $\widehat{m}_{\alpha,n}$, given by $||\widehat{m}_{\alpha,n} - m_0||_{L^2(\mathcal{X})} = O_p(n^{-\frac{p_0}{3p_0+d_2}})$. ■

Proof of Theorem 5.6 (i) After plugging in the results of Proposition 5.3 into Theorem 4.4, we get

$$||\widehat{m}_{\alpha,n}^s - m_0||_{L^2(\mathcal{X})} \leq O_p(\frac{1}{\sqrt{\alpha}}[\frac{1}{\sqrt{ng_2^{d_2}}} + g_1^{p_0}]) + O_p(\alpha^\mu) + O_p(\alpha^{\mu-1/2}[\frac{1}{\sqrt{ng_1^{d_1}}} + g_2^{p_0}]),$$

from $p_0 = p_1$ and $\alpha^{\mu-1/2}(1/\sqrt{ng_2^{d_2}} + g_1^{p_0}) = o(\alpha^\mu)$, since $n\alpha g_2^{d_2} \to \infty$ and $\alpha \to 0$, as $n \to \infty$; see the assumption in C.5.5(b). Following the same arguments in the proof of Theorem 5.5, we can show that the profile of quasi-optimal smoothing parameters are given by $\{(g_1, g_2, \alpha^*)\}$ with $\alpha^* = [\max(1/\sqrt{ng_{2n}^{d_2}}, g_{1n}^{p_0})]^{2/(2\mu+1)}$, under which the actual convergence rate of $\widehat{m}_{\alpha,n}^s$ reduces to

$$||\widehat{m}_{\alpha,n}^s - m_0||_{L^2(\mathcal{X})} = [\max(1/\sqrt{ng_{2n}^{d_2}}, g_{1n}^{p_0})]^{2\mu/(2\mu+1)}.$$

The fastest possible rate of convergence, among the quasi-optimal ones, is achieved by choosing $(g_1, g_2) = (g_{1n}^*, g_{2n}^*)$ such that $(ng_{1n}^{*d_1})^{-1/2} \simeq g_{1n}^{*p_0/(2\mu+1)}$, and $g_{2n}^{*p_0(2\mu+1)} \simeq (ng_{2n}^{*d_2})^{-1/2}$. By letting

$g_{1n}^* = C_0 n^{-\frac{(2\mu+1)}{2p_0+(2\mu+1)d_1}}$, $g_{2n}^* = C_1 n^{-\frac{1}{2(2\mu+1)p_0+d_2}}$, and $\alpha_n^* = C_2 n^{-\frac{2p_0}{2(2\mu+1)p_0+d_2}}$, we obtain the optimal convergence rate of $\widehat{m}_{\alpha,n}^s$, given as

$$||\widehat{m}_{\alpha,n}^s - m_0||_{L^2(\mathcal{X})} = O_p(n^{-\frac{2\mu p_0}{2(2\mu+1)p_0+d_2}}),$$

since $g_{1n}^{*p_0} = C_0 n^{-\frac{(2\mu+1)p_0}{2p_0+(2\mu+1)d_1}}$ is of smaller order than $1/\sqrt{ng_{2n}^{*d_2}} = O(n^{-\frac{(2\mu+1)p_0}{2(2\mu+1)p_0+d_2}})$, by the assumption that $d_1/2 \leq p_0$ and $d_1 \leq d_2$. All the basic conditions in C.5.5 hold for $(g_{1n}^*, g_{2n}^*, \alpha_n^*)$.

(ii) Application of Proposition 5.3 to Theorem 4.1 through Theorem 4.3 gives

$$||\widehat{m}_{\alpha,n}^q - m_0||_{L^2(\mathcal{X})} \leq O_p(\frac{1}{\sqrt{\alpha}}[\frac{1}{\sqrt{ng_2^{d_2}}} + g_1^{p_0}]) + O_p(\alpha^{\min(\mu,q)}) + O_p(\alpha^{\min(\mu-1/2,0)}[\frac{1}{\sqrt{ng_1^{d_1}}} + g_2^{p_0}]),$$

since $\alpha^{\min(\mu-1/2,1/2)}(1/\sqrt{ng_2^{d_1}} + g_1^{p_0}) = o(1/\sqrt{n\alpha g_2^{d_2}}) + o(g_1^{p_0}/\sqrt{\alpha})$, and $\alpha^{\min(\mu-1/2,0)}(g_1^{p_0}) = o(g_1^{p_0}/\sqrt{\alpha})$. Following the same arguments in the proof of Theorem 5.5, we can show that the profile of quasi-optimal smoothing parameters are given by $\{(g_1, g_2, \alpha^*)\}$ with $\alpha^* = [\max(1/\sqrt{ng_{2n}^{d_2}}, g_{1n}^{p_0})]^{2/(2\mu_q+1)}$, under which the actual convergence rate of $\widehat{m}_{\alpha,n}^q$ is

$$||\widehat{m}_{\alpha,n}^q - m_0||_{L^2(\mathcal{X})} = [\max(1/\sqrt{ng_{2n}^{d_2}}, g_{1n}^{p_0})]^{2\mu_q/(2\mu_q+1)}.$$

The fastest possible rate of convergence, among the quasi-optimal ones, is achieved by choosing $(g_1, g_2) = (g_{1n}^*, g_{2n}^*)$ such that $(ng_{1n}^{*d_1})^{-1/2} \simeq g_{1n}^{2\mu_q^\dagger p_0/(2\mu_q+1)}$, and $g_{2n}^{p_0(2\mu_q+1)/2\mu_q^\dagger} \simeq (ng_{2n}^{*d_2})^{-1/2}$. By letting $g_{1n}^* = C_0 n^{-\frac{(2\mu_q+1)}{4\mu_q^\dagger p_0+(2\mu_q+1)d_1}}$, $g_{2n}^* = C_1 n^{-\frac{2\mu_q^\dagger}{2(2\mu_q+1)p_0+2\mu_q^\dagger d_2}}$, and $\alpha_n^* = C_2 n^{-\frac{2p_0}{2(2\mu_q+1)p_0+2\mu_q^\dagger d_2}}$, we obtain the optimal convergence rate of $\widehat{m}_{\alpha,n}^q$, given as

$$||\widehat{m}_{\alpha,n}^q - m_0||_{L^2(\mathcal{X})} = O_p(n^{-\frac{2\mu_q p_0}{2(2\mu_q+1)p_0+2\mu_q^\dagger d_2}}),$$

since $g_{1n}^{*p_0} = C_0 n^{-\frac{(2\mu_q+1)p_0}{4\mu_q^\dagger p_0+(2\mu_q+1)d_1}}$ is of smaller order than $1/\sqrt{ng_{2n}^{*d_2}} = O(n^{-\frac{(2\mu_q+1)p_0}{2(2\mu_q+1)p_0+2\mu_q^\dagger d_2}})$, by the assumption that $d_1/2 \leq p_0$ and $d_1 \leq d_2$. ■

References

[1] Ai, C. and X. Chen (2001): "Efficient estimation of models with conditional moment restrictions containing unknown functions," Working Paper, LSE.

[2] Altonji, J. G. and R. Matzkin (2001): "Panel data estimators for nonseparable models with endogenous regressors," NBER Working Paper No. TO267.

[3] Amemiya, T. (1974): "The nonlinear two-stage least-squares estimator," *Journal of Econometrics* 2, 105-110.

[4] Birman, M. and M. Solomjak (1980): *Quantitative Analysis in Sobolev Imbedding Theorems and Applications to Spectral Theory*. American Math. Soc. Tran., series 2, Vol 114.

[5] Blundell, R. and J. L. Powell (2001a): "Endogeneity in nonparametric and semiparametric regression models," forthcoming in *Advances in Econometrics, Proceedings of the World Meetings, 2000*, ed. by L. Hansen, North Holland.

[6] Blundell, R. and J. L. Powell (2001b): "Endogeneity in semiparametric binary response models," Center for Microdata Methods and Practice, Working Paper, 05/01.

[7] Brown, D. J. and R. Matzkin (1998): "Estimation of nonparametric functions in simultaneous equations models, with an application to consumer demand," Cowles Foundation Working Paper.

[8] Chesher, A. (2002): "Local identification in nonseparable models," Center for Microdata Methods and Practice, Working Paper, 05/02.

[9] Darolles, S., J-P. Florens, and E. Renault (2001): "Nonparametric instrumental regression, Working Paper, GREMAQ, University of Toulouse.

[10] Das, M. (1999): "Instrumental variable estimation of models with discrete endogenous regressors, " presented at 2000 World Congress of the Econometric Society.

[11] Engl, H. (1987): "On the choice of the regularization parameter for iterated Tikhonov regularization of ill-posed problems," *Journal of Approximation Theory 49, 55-63.*

[12] Engl, H. and H. Gfrerer (1988): "A posteriori parameter choice for general regularization methods for solving linear ill-posed problems," *Applied Numerical Mathematics 4, 395-417.*

[13] Engle, H.W., M. Hanke, and A. Neubauer (2000): *Regularization of Inverse Problems.* Dordrecht: Kluwer Academic Press.

[14] Fan, J. (1991): “Global behavior of deconcovolution kernel estimates,” *Statistica Sinica* 1, 541-551.

[15] Groetsch, C.W. (1993): *Inverse Problems in the Mathematical Sciences.* Braunschweig: Vieweg.

[16] Groetsch, C.W. (1983): “On the Asymptotic Order of Accuracy of Tikhonov Regularization,” *Journal of Optimization Theory and Applications 41, 293-298.*

[17] Groetsch, C.W. (1977): *Generalized Inverses of Linear Operators: Representation and Approximation.* New York: Dekker.

[18] Hall P. and J. L. Horowitz (2003): “Nonparametric methods for inference in the presence of instrumental variables,” Working Paper, Center for Microdata Methods and Practice.

[19] Hansen L. P. (1982): “Large sample properties of generalized method of moments estimators,” *Econometrica* 50, 1029-1054.

[20] Hutson, V. and J. S. Pym (1980): *Applications of Functional Analysis and Operator Theory.* London: Academic Press.

[21] Imbens, G. W. and W. K. Newey (2001): “Identification and estimation of triangular simultaneous equations models without additivity, Working Paper.

[22] Ivanov, V.K., V.V. Vasin, and V.P. Tanana (1978): *Theory of Linear Ill-Posed Problems.* Moscow: Nauka.

[23] King, J.T. and D. Chillingworth (1979): “Approximation of generalized inverses by iterated regularization,” *Numerical Functional Analyses and Optimizations 1, 499*-513.

[24] Kirsch, A. (1996): *An Introduction to the Mathematical Theory of Inverse Problems.* New York: Springer Verlag.

[25] Kress, R. (1989): *Linear Integral Equations.* Berlin: Springer Verlag.

[26] Linton, O. B. and E. Mammen (2003): “Estimating semiparametric ARCH(∞) models by kernel smoothing methods,” Working Paper, LSE.

[27] Mammen, E., O. B. Linton, and J. Nielsen (1999): "The existence and asymptotic properties of a backfitting projection algorithm under weak conditions," *The Annals of Statistics,* 27(5), 1443-1490.

[28] Nashed, M.Z. (1976): *Generalized Inverses and Applications.* New York: Academic Press.

[29] Nashed, M.Z. and G. Wahba (1974): "Convergence rates of approximate least squares solutions of linear integral and operator equations of the first kind," *Mathematics of Computation* 28(125), 69-80.

[30] Newey W. K. and J. L. Powell (1988, 2002): "Nonparametric instrumental variables estimation, MIT Working Paper.

[31] Newey W. K., J. L. Powell, and F. Vella (1999): "Nonparametric estimation of triangular simultaneous equations models," *Econometrica* 67, 565-603.

[32] Nychka, D. and D. Cox (1989): "Convergence rates for regularized solutions of integral equations from discrete noisy data," *The Annals of Statistics,* 17(5), 556-572.

[33] O'Sullivan, F. (1986): "Ill-posed inverse problems (with Discussion)," *Statistical Science,* 4, 503-527.

[34] Phillips, D.L. (1962): "A technique for the numerical solution of certain integral equations of the first kind," *Journal of the Association for Computing Machinery* 9, 84-97.

[35] Plato, R. and G. Vainikko (1990): "On the regularization of projection methods for solving ill-posed problems," *Numerische Mathematik* 57, 63-79.

[36] Roehrig, C. S. (1988): "Conditions for Identification in Nonparametric and Parametric Models," *Econometrica*, 55, 875-891.

[37] Sargang, J. D. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica* 26, 393-415.

[38] Schock, E. (1985): "Approximate solution of ill-posed equations: Arbitrarily slow convergence vs superconvergence, in *Constructive Methods for the Practical Treatment of Integral Equations* (edited by G. Hämmerlin and K.H. Hoffmann), pp. 234-243, Basel: Birkhäuser.

[39] Showalter D. (1967): “Representation and computation of the pseudoinverse,” *Proc. Amer. Math. Soc. 18*, 584-586.

[40] Silverman, B.W. (1978): “Weak and strong uniform consistency of the kernel estimate of a density and its derivatives,” *Annals of Statistics* Vol.6, No.1, 177-184.

[41] Tautenhahn, U. (1998): “Optimality for ill-posed problems under general source conditions,” *Numerical Functional Analyses and Optimizations 19, 377*-398.

[42] Theil, H. (1953): Repeated Least Squares Applied to Complete Equation Systems, The Hague; Central Planning Bureau.

[43] Tikhonov, A.N. and V. Arsenin (1977): *Solutions of Ill-Posed Problems.* New York: Wiley.

[44] Tikhonov, A.N (1963): “Solution of incorrectly formulated problems and the regularization method,” *Soviet Mathematics Doklady 4*, 1035-1038.

[45] Vainikko, G.M. and A.Y. Veretennikov (1986): *Iteration Procedures in Ill-Posed problems.* 1st Ed. Moscow: Nauka.

[46] van Rooij, A.C.M., and F.H. Ruymgaart (1999): “On inverse estimation,” in *Asymptotics, Nonparametrics, and Time Series*, ed. by S. Ghosh. New York: Marcel Dekker, 579-613.

Table 1. MSE (Squared Bias + Variance) of Various Regularized-Kernel Estimates

α	g	OTR	ITR(2)	GTR(2)	SW
.001	.3	.063 (.012 + .051)	.117 (.020 + .097)	.075 (.024 + .051)	.091 (.023 + .069)
	.4	.049 (.014 + .035)	.096 (.035 + .060)	.058 (.020 + .039)	.075 (.027 + .047)
	.5	.039 (.012 + .027)	.086 (.040 + .046)	.042 (.012 + .030)	.058 (.022 + .036)
	.6	.034 (.012 + .022)	.072 (.035 + .037)	.042 (.017 + .025)	.047 (.017 + .030)
.005	.3	.033 (.013 + .020)	.042 (.009 + .033)	.033 (.013 + .021)	.035 (.009 + .026)
	.4	$.029^{*}$ (.014 + .015)	.033 (.009 + .024)	.034 (.018 + .016)	.030 (.010 + .020)
	.5	.030 (.018 + .012)	.030 (.010 + .020)	.039 (.026 + .013)	.031 (.015 + .016)
	.6	.037 (.027 + .010)	.034 (.017 + .017)	.047 (.037 + .011)	.038 (.025 + .013)
.01	.3	.042 (.027 + .014)	.032 (.009 + .023)	$.033^{*}$ (.017 + .016)	.030 (.011 + .019)
	.4	.041 (.030 + .011)	$.028^{*}$ (.010 + .017)	.034 (.022 + .012)	$.028^{*}$ (.014 + .014)
	.5	.045 (.036 + .009)	.029 (.015 + .014)	.039 (.029 + .009)	.032 (.021 + .011)
	.6	.053 (.046 + .007)	.036 (.024 + .012)	.047 (.039 + .008)	.041 (.031 + .009)
.015	.3	.056 (.044 + .012)	.030 (.011 + .019)	.034 (.020 + .014)	.029 (.013 + .016)
	.4	.057 (.048 + .009)	.028 (.014 + .014)	.035 (.025 + .011)	.029 (.017 + .012)
	.5	.063 (.056 + .007)	.031 (.019 + .011)	.041 (.032 + .009)	.034 (.024 + .010)
	.6	.073 (.067 + .006)	.038 (.029 + .010)	.049 (.041 + .008)	.042 (.033 + .008)
.02	.3	.073 (.062 + .011)	.031 (.014 + .017)	.038 (.024 + .013)	.030 (.015 + .015)
	.4	.076 (.068 + .008)	.029 (.017 + .013)	.039 (.029 + .010)	.030 (.019 + .011)
	.5	.084 (.078 + .006)	.033 (.023 + .010)	.045 (.037 + .008)	.035 (.026 + .009)
	.6	.097 (.091 + .005)	.041 (.032 + .009)	.054 (.046 + .007)	.044 (.036 + .008)

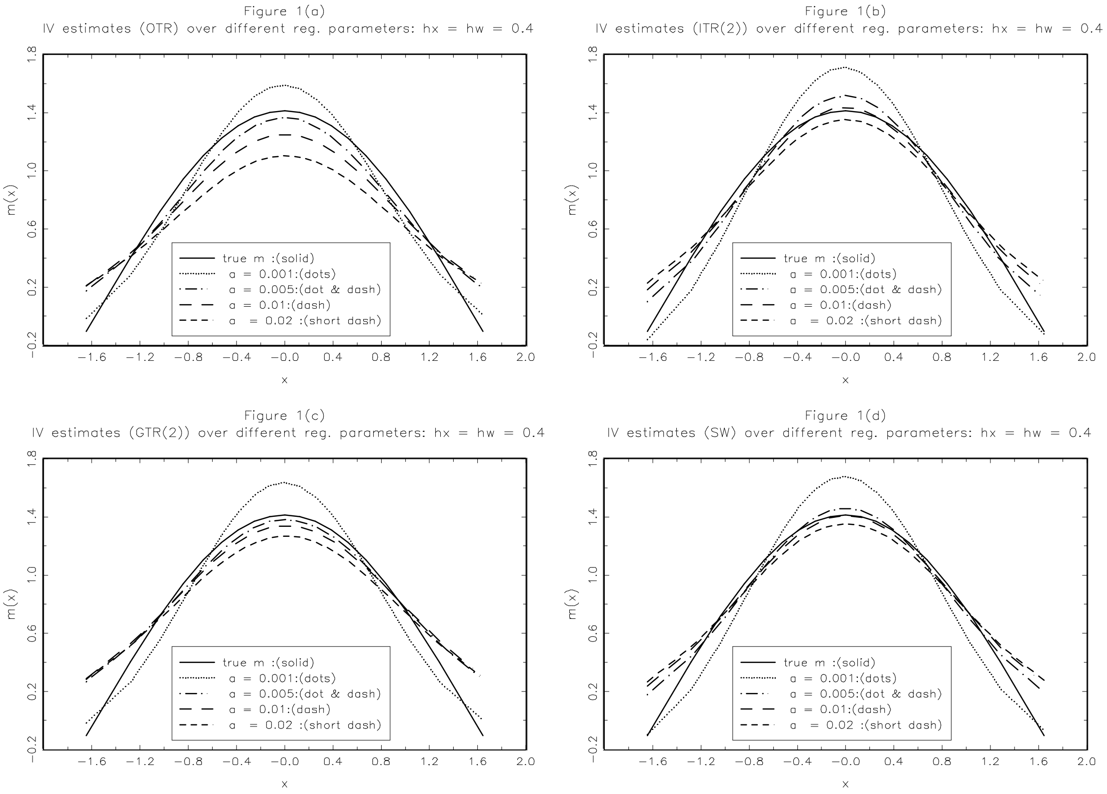

Figure 1: Averaged IV estimates over different regularization parameters: with g = 0.4

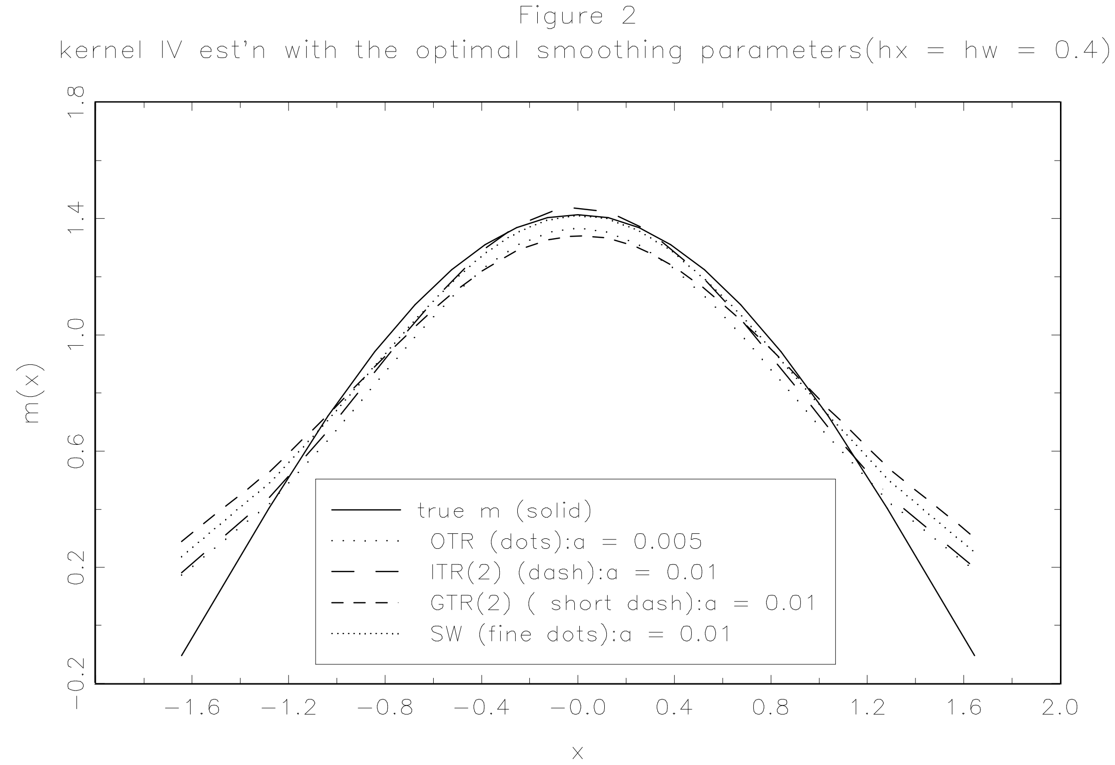

Figure 2: Averaged IV estimates for the optimal choice of regularization parameter (with g = 0.4)