A Practical Guide to Compact Infinite Dimensional Parameter Spaces*

Joachim Freyberger† Matthew A. Masten‡

May 16, 2017

Abstract

We gather and review general compactness results for many commonly used parameter spaces in nonparametric estimation, and we provide several new results. We consider three kinds of functions: (1) functions with bounded domains which satisfy standard norm bounds, (2) functions with bounded domains which do not satisfy standard norm bounds, and (3) functions with unbounded domains. In all three cases we provide two kinds of results, compact embedding and closedness, which together allow one to show that parameter spaces defined by a $\|\cdot\|_s$-norm bound are compact under a norm $\|\cdot\|_c$. We illustrate how these results are typically used in econometrics by considering two common settings: nonparametric mean regression and nonparametric instrumental variables estimation.

JEL classification: C14, C26, C51

Keywords: Nonparametric Estimation, Sieve Estimation, Trimming, Nonparametric Instrumental Variables

*This paper was presented at Duke, the 2015 Triangle Econometrics Conference, and the 2016 North American and China Summer Meetings of the Econometric Society. We thank audiences at those seminars as well as Bruce Hansen, Kengo Kato, Jack Porter, Yoshi Rai, and Andres Santos for helpful conversations and comments.

†Department of Economics, University of Wisconsin-Madison, jfreyberger@ssc.wisc.edu

‡Department of Economics, Duke University, matt.masten@duke.edu

1 Introduction

Compactness is a widely used assumption in econometrics, for both finite and infinite dimensional parameter spaces. It can ensure the existence of extremum estimators and is an important step in many consistency proofs (e.g. Wald 1949). Even for noncompact parameter spaces, compactness results are still often used en route to proving consistency. For finite dimensional parameter spaces, the Heine-Borel theorem provides a simple characterization of which sets are compact. For infinite dimensional parameter spaces the situation is more delicate. In finite dimensional spaces, all norms are equivalent: convergence in any norm implies convergence in all norms. This is not true in infinite dimensional spaces, and hence the choice of norm matters. Even worse, unlike in finite dimensional spaces, closed balls in infinite dimensional spaces *cannot* be compact. Specifically, if $\|\cdot\|$ is a norm on a function space $\mathscr{F}$, then a $\|\cdot\|$-ball is $\|\cdot\|$-compact if and only if $\mathscr{F}$ is finite dimensional. This suggests that compactness and infinite dimensionality are mutually exclusive. A key insight from functional analysis, however, is that both concepts can be retained by using *two* norms—define the parameter space using one and obtain compactness in the other one. This idea goes back to at least the 1930's, and is a motivation for the weak* topology; see the Banach-Alaoglu theorem. In econometrics, this idea has been used by Gallant and Nychka (1987) and subsequent authors in the sieve estimation literature. Compactness assumptions are widely used in this literature. Specifically, it is common to define the parameter space as a ball with the norm $\|\cdot\|_s$ and obtain compactness under a norm $\|\cdot\|_c$. This result can then be used to prove consistency of a function estimator in the norm $\|\cdot\|_c$.

In the present paper, we make two main contributions. First, we gather and review many of these compactness results. Unlike much of the previous mathematics literature, we focus on norms and parameter spaces most useful in econometrics to provide a unified and easily accessible treatment. The results are particularly complicated for functions defined on an unbounded Euclidean domain, where various commonly used choices of norms imply very different parameter spaces (for a specific example, see our discussion in section 6). Second, we provide several new compactness results, which relax important restrictions on parameter spaces in the previous literature. For example, for functions on unbounded Euclidean domains our results allow for parameter spaces which include polynomials of arbitrary degree.

We organize the results into two main parts, depending on the domain of the function of interest: bounded or unbounded. We first consider functions on bounded Euclidean domains which satisfy a norm bound, such as having a bounded Sobolev integral or sup-norm. Second, we consider functions defined on an unbounded Euclidean domain, where we build on and extend the important work of Gallant and Nychka (1987). Finally, we return to functions on a bounded Euclidean domain, but now suppose they do *not* directly satisfy a norm bound. One example is the quantile function $Q_X : (0,1) \to \mathbb{R}$ for a random variable X with full support. Since $Q_X(\tau)$ asymptotes to $\pm\infty$ as τ approaches 0 or 1, the derivatives of Q_X are unbounded. Nonetheless, we show that compactness results may apply if we replace unweighted norms with weighted norms.

In general there are two steps to showing that a parameter space defined as a ball under $\|\cdot\|_s$ is

compact under $\|\cdot\|_c$. First we prove a compact embedding result, which means that the $\|\cdot\|_c$-closure of the parameter space is $\|\cdot\|_c$-compact. Second, we show that the parameter space is actually $\|\cdot\|_c$-closed, and hence equals its closure and hence is compact. We show that some choices of the pair $\|\cdot\|_s$ and $\|\cdot\|_c$ satisfy the first step, but not the closedness step.

For functions on unbounded Euclidean domains, we follow the approach of Gallant and Nychka (1987) and introduce weighted norms. Gallant and Nychka (1987) showed how to extend compact embedding proofs for bounded domains to unbounded domains. We review and extend their result and show how it applies to a general class of weighting functions, as well as many choices of $\|\cdot\|_s$ and $\|\cdot\|_c$, such as Sobolev L_2 norms, Sobolev sup-norms, and Hölder norms. In particular, unlike existing results, our result allows for many kinds of exponential weight functions. This allows, for example, parameter spaces for regression functions which include polynomials of arbitrary degree. We also discuss additional commonly used weighting functions, such as polynomial upweighting and polynomial downweighting. We explain how the choice of weight function constrains the parameter space. In a typical analysis, the choice of norm in which we prove consistency also has implications on how strong other regularity conditions are, such as those for obtaining asymptotic normality, and how easy these conditions are to check. Such considerations may also affect the choice of norms.

We illustrate these considerations with two simple applications. First, we consider estimation of mean regression functions with full support regressors. We give low level conditions for consistency of a sieve least squares estimator, and discuss how the choice of norm is used in this result. We also show that weighted norms can be interpreted as a generalization of trimming. Second, we discuss the nonparametric instrumental variables model. We again give conditions for consistency of a sieve NPIV estimator and discuss the role of the norm in this result.

The rest of this paper is organized as follows. We first conclude this section by placing our results in the context of the related literature. Then in section 2 we review the definitions of the norms and function spaces used throughout the paper. Our main results are in sections 3, 4, and 5, where we consider each of the three cases discussed above. In section 6 we discuss our applications. Section 7 concludes. Some formal definitions, additional lemmas, and proofs of all results are all given in a supplemental appendix.

Related literature

All of our compact embedding results for unweighted function spaces are well known in the mathematics literature (see, for example, Adams and Fournier 2003). For weighted Sobolev spaces, Kufner (1980) was one of the earliest studies. He focuses on functions with bounded domains, and proves several general embedding theorems for a large class of weight functions. These are not, however, compact embedding results. Schmeisser and Triebel (1987) also study weighted function spaces, but do not prove compact embedding results. As discussed above, Gallant and Nychka (1987) prove an important compact embedding result for functions with unbounded domains. Haroske and Triebel (1994a) prove a general compact embedding result for a large class of weighted spaces. This result, as well as the followup work by Triebel and coauthors, such as

Haroske and Triebel (1994b) and Edmunds and Triebel (1996), relies on assumptions which hold for polynomial weights, but not for exponential weights (see pages 14 and 17 for details). Moreover, as we show, these results also do not apply to functions with bounded domain. Hence, except in one particular case (see our discussion of Brown and Opic 1992 below), our compact embedding results using weighted norms for functions on bounded domains are the first that we are aware of. Likewise, except in one particular case (again see our Brown and Opic 1992 discussion below), our compact embedding results for functions on unbounded domains allow for a much larger class of weight functions than previously allowed. In particular, we allow for exponential weight functions. Note, however, that the results by Triebel and coauthors allow for more general function spaces, including Besov spaces and many others. We focus on Sobolev spaces, Hölder spaces, and spaces of continuously differentiable bounded functions because these are by far the most commonly used function spaces in econometrics.

Brown and Opic (1992) give high level conditions on the weight functions for a compact embedding result similar to that in Gallant and Nychka (1987), for both bounded and unbounded domains. Similar to Gallant and Nychka (1987), this result is only for compact embeddings of a Sobolev L_p space into a space of bounded continuous functions. This result allows for many kinds of exponential weights. In these cases, our results provide simpler lower level conditions on the weight functions, although these conditions are less general. Importantly, we also provide seven further compact embedding results that they do not consider. See pages 17 and 24 for more details.

Just seven years after Wald's (1949) consistency proof, Kiefer and Wolfowitz (1956) extended his ideas to apply to nonparametric maximum likelihood estimators.[1] Their results rely on the well-known fact that the space of cdfs is compact under the weak convergence topology. In econometrics, their results have been applied by Cosslett (1983), Heckman and Singer (1984), and Matzkin (1992). More recently, Fox and Gandhi (2015) and Fox, Kim, and Yang (2015) have used similar ideas, relying on this particular compactness result. This compactness result is certainly powerful when the cdf is our object of interest. We are often interested in other functions, however, like pdfs or regression functions. The results in this paper can be applied in these cases. Wong and Severini (1991) extended the analysis of nonparametric MLE even further. They still make a compact parameter space assumption, but do not restrict attention to cdfs.

Compactness results like those we here are used throughout the sieve literature. For example, see Elbadawi, Gallant, and Souza (1983), Gallant and Nychka (1987), Gallant and Tauchen (1989), Fenton and Gallant (1996), Newey and Powell (2003), Ai and Chen (2003), Chen, Hong, and Tamer (2005), Chen, Fan, and Tsyrennikov (2006), Brendstrup and Paarsch (2006), Chernozhukov, Imbens, and Newey (2007), Hu and Schennach (2008), Chen, Hansen, and Scheinkman (2009a), Freyberger (2012), Santos (2012), and Khan (2013). Chen (2007) gives additional references to sieve estimation in the literature. Appendix A in the supplement to Chen and Pouzo (2012) provides a brief overview of some of the compactness results we discuss.

[1] Wald (1949) did attempt to generalize his results to the infinite dimensional case in his final section. His approach, however, is to assume that closed balls are compact (his assumption 9(iv)). As we've discussed, this implies the parameter space is actually finite dimensional.

An alternative approach in the sieve literature to assuming a compact parameter space is to use penalization methods. In this case, it is often assumed that the penalty function is lower semicompact. For example, see Chen and Pouzo (2012) theorem 3.2 and Chen and Pouzo (2015) assumption 3.2(iii). For the penalty function $\text{pen}(\cdot) = \|\cdot\|_s$ and consistency norm $\|\cdot\|_c$, lower semicompactness of $\text{pen}(\cdot)$ means that $\|\cdot\|_s$-balls are $\|\cdot\|_c$-compact. This is precisely the conclusion of a compact embedding and closedness result combined. Hence our results are useful even if one does not want to assume the parameter space itself is compact.

Even when neither compactness nor penalization is necessary for consistency, such as in theorem 3.1 of Chen (2007), an 'identifiable uniqueness' or 'well separated' point of maximum assumption is needed. Also see van der Vaart (2000) theorem 5.7, van der Vaart and Wellner (1996) lemma 3.2.1, and the discussion in section 2.6 of Newey and McFadden (1994). Compactness combined with continuity of the population objective function provide simple sufficient conditions for this assumption, as Chen (2007) discusses via her condition 3.1″. Similarly, compactness may be used to verify uniform convergence assumptions, such as condition 3.5 of theorem 3.1 in Chen (2007); we discuss this further on page 28.

2 Norms for functions

Since the choice of norm for infinite dimensional function spaces matters, we begin with a brief survey of the three kinds of norms most commonly used in econometrics: Sobolev sup-norms, Sobolev integral norms, and Hölder norms. These norms are defined for functions $f : \mathcal{D} \to \mathbb{R}$ where the domain $\mathcal{D}$ is an open subset of $\mathbb{R}^{d_x}$, possibly the entire space $\mathbb{R}^{d_x}$, for an integer $d_x \geq 1$.[2] For these functions, denote the differential operator by

$$\nabla^\lambda = \frac{\partial^{|\lambda|}}{\partial x_1^{\lambda_1} \cdots \partial x_{d_x}^{\lambda_{d_x}}} = \frac{\partial^{\lambda_1}}{\partial x_1^{\lambda_1}} \cdots \frac{\partial^{\lambda_{d_x}}}{\partial x_{d_x}^{\lambda_{d_x}}},$$

where $\lambda = (\lambda_1, \ldots, \lambda_{d_x})$ is a multi-index, a d_x-tuple of non-negative integers, and $|\lambda| = \lambda_1 + \cdots + \lambda_{d_x}$. Note that $\nabla^0 f = f$.

The first space we consider are continuously differentiable functions whose derivatives are uniformly bounded. Let m be a nonnegative integer. For an m-times differentiable function $f : \mathcal{D} \to \mathbb{R}$, define the *weighted Sobolev sup-norm* of f as

$$\|f\|_{m,\infty,\mu} = \max_{0 \leq |\lambda| \leq m} \sup_{x \in \mathcal{D}} |\nabla^\lambda f(x)| \mu(x).$$

Here $\mu : \mathcal{D} \to \mathbb{R}_+$ is a continuous nonnegative weight function. Let $\|\cdot\|_{m,\infty}$ denote the unweighted Sobolev sup-norm; that is, the weighted Sobolev sup-norm with the identity weight $\mu(x) \equiv 1$. For

[2] Restricting ourselves to open subsets avoids the problem of defining derivatives at the boundary. For functions with closed domains, our results can be extended under a continuity at the boundary assumption; see lemma S3 in the supplemental appendix.

the identity weight and $m = 0$, $\|\cdot\|_{m,\infty,\mu}$ is just the usual sup-norm. Relatedly, notice that

$$\|f\|_{m,\infty,\mu} = \max_{0\leq|\lambda|\leq m} \|\nabla^\lambda f\|_{0,\infty,\mu}.$$

Let $\mathscr{C}_m(\mathcal{D})$ denote the space of m-times continuously differentiable functions $f : \mathcal{D} \to \mathbb{R}$. Let

$$\mathscr{C}_{m,\infty,\mu}(\mathcal{D}) = \{f \in \mathscr{C}_m(\mathcal{D}) : \|f\|_{m,\infty,\mu} < \infty\}.$$

The normed vector space $(\mathscr{C}_{m,\infty,\mu}(\mathcal{D}), \|\cdot\|_{m,\infty,\mu})$ is $\|\cdot\|_{m,\infty,\mu}$-complete[3], and hence it is a $\|\cdot\|_{m,\infty,\mu}$-Banach space. When $\mu(x) \equiv 1$, define $\mathscr{C}_{m,\infty}(\mathcal{D}) = \mathscr{C}_{m,\infty,1}(\mathcal{D})$.

The next space we consider replaces the sup-norm with an L_p norm. Let p satisfy $1 \leq p < \infty$. To ensure completeness of this space, we must allow for functions which are only weakly differentiable, rather than just classically differentiable functions. When both the classical and weak derivatives exist, they are equal. We denote both kinds of derivatives by $\nabla^\lambda f$. Adams and Fournier (2003) formally define the weak derivative on page 22, and we briefly discuss the completeness issue in section J of the supplemental appendix.

Let $\mathscr{W}_m(\mathcal{D})$ denote the set of all functions $f : \mathcal{D} \to \mathbb{R}$ which have weak derivatives of all orders $0 \leq |\lambda| \leq m$. For $f \in \mathscr{W}_m(\mathcal{D})$, define the *weighted Sobolev L_p norm* of f as

$$\|f\|_{m,p,\mu} = \left(\sum_{0\leq|\lambda|\leq m} \int_{\mathcal{D}} |\nabla^\lambda f(x)|^p \mu(x)\, dx \right)^{1/p}.$$

μ is a weight function as above. We also call this a Sobolev integral norm. Let $\|\cdot\|_{m,p}$ denote the unweighted Sobolev L_p norm. For the identity weight and $m = 0$, $\|\cdot\|_{m,p,\mu}$ is just the usual L_p norm. Relatedly, notice that

$$\|f\|^p_{m,p,\mu} = \sum_{0\leq|\lambda|\leq m} \|\nabla^\lambda f\|^p_{0,p,\mu}.$$

$\|\cdot\|_{0,p,\mu}$ is called the *weighted L_p norm*. While the Sobolev sup-norm measures functions in terms of the pointwise largest values of the function and its derivatives, the Sobolev L_p norm measures functions in terms of the average values of the function and its derivatives.

Define the *weighted Sobolev space* by

$$\mathscr{W}_{m,p,\mu}(\mathcal{D}) = \{f \in \mathscr{W}_m(\mathcal{D}) : \|f\|_{m,p,\mu} < \infty\}.$$

The normed vector space $(\mathscr{W}_{m,p,\mu}(\mathcal{D}), \|\cdot\|_{m,p,\mu})$ is $\|\cdot\|_{m,p,\mu}$-complete[4], and hence it is a $\|\cdot\|_{m,p,\mu}$-Banach space. When $\mu(x) \equiv 1$, define $\mathscr{W}_{m,p}(\mathcal{D}) = \mathscr{W}_{m,p,1}(\mathcal{D})$. For both of the weighted Sobolev norms, there is a less common alternative approach to incorporating the weighting function, which we discuss in section 4.3.

[3]Under assumption 6 below. For example, see theorem 5.1 of Rodríguez, Álvarez, Romera, and Pestana (2004).

[4]Under assumption 6 below. Again see theorem 5.1 of Rodríguez et al. (2004).

The final space of functions we consider is similar to the space of functions with bounded unweighted Sobolev sup-norms. Define the *Hölder coefficient* of a function $f : \mathcal{D} \to \mathbb{R}$ by

$$[f]_\nu = \sup_{x,y \in \mathcal{D}, x \neq y} \frac{|f(x) - f(y)|}{\|x - y\|_e^\nu}$$

for some $\nu \in (0, 1]$, called the *Hölder exponent*, where $\|\cdot\|_e$ is the $\mathbb{R}^{d_x}$-Euclidean norm.[5] A function with $[f]_\nu < \infty$ is Hölder continuous since

$$|f(x) - f(y)| \leq [f]_\nu \cdot \|x - y\|_e^\nu$$

holds for all $x, y \in \mathcal{D}$. Define the *Hölder norm* of f as

$$\begin{aligned}\|f\|_{m,\infty,\mathbb{1},\nu} &= \|f\|_{m,\infty} + \max_{|\lambda|=m} [\nabla^\lambda f]_\nu \\ &= \max_{|\lambda| \leq m} \sup_{x \in \mathcal{D}} |\nabla^\lambda f(x)| + \max_{|\lambda|=m} \sup_{x,y \in \mathcal{D}, x \neq y} \frac{|\nabla^\lambda f(x) - \nabla^\lambda f(y)|}{\|x - y\|_e^\nu},\end{aligned}$$

where recall that $\|\cdot\|_{m,\infty}$ is the unweighted Sobolev sup-norm. The Hölder coefficient generalizes the supremum over the derivative; for differentiable functions f we have

$$[f]_1 = \sup_{x \in \mathcal{D}} |\nabla f(x)|.$$

The Hölder exponent $[f]_1$, however, is also defined for nondifferentiable functions f. Define the *Hölder space* with exponent ν by

$$\mathscr{C}_{m,\infty,\mathbb{1},\nu}(\mathcal{D}) = \{f \in \mathscr{C}_m(\mathcal{D}) : \|f\|_{m,\infty,\mathbb{1},\nu} < \infty\}.$$

The normed vector space $(\mathscr{C}_{m,\infty,\mathbb{1},\nu}(\mathcal{D}), \|\cdot\|_{m,\infty,\mathbb{1},\nu})$ is $\|\cdot\|_{m,\infty,\mathbb{1},\nu}$-complete. We discuss weighted Hölder spaces, along with an alternative approach to weighted Sobolev spaces, in section 4.3. For all of these function spaces, we omit the domain $\mathcal{D}$ from the notation when it is understood.

3 Functions on bounded domains

Let $(\mathscr{F}, \|\cdot\|_s)$ and $(\mathscr{G}, \|\cdot\|_c)$ be Banach spaces with $\mathscr{F} \subseteq \mathscr{G}$. These could be any of the spaces mentioned in the previous section. Our main goal is to understand when the space

$$\Theta = \{f \in \mathscr{F} : \|f\|_s \leq B\} \tag{1}$$

is $\|\cdot\|_c$-compact, for various choices of the two norms, where $B > 0$ is a finite constant. $\|\cdot\|_s$ is called the *strong* norm, since it will be stronger than $\|\cdot\|_c$ in the sense that $\|\cdot\|_c \leq M\|\cdot\|_s$ for a finite constant M. Because we cannot obtain compactness of Θ in the strong norm without

[5] $\nu > 1$ is excluded since $[f]_\nu < \infty$ for a $\nu > 1$ implies that f is constant.

reducing it to a finite dimensional set, we instead obtain compactness under $\|\cdot\|_c$, which is called the *consistency* or *compactness* norm. In econometrics applications, we obtain consistency of our function estimators in this latter norm (see section 6).

The general approach to obtaining $\|\cdot\|_c$-compactness of Θ has two steps. First, we prove that Θ is *relatively* $\|\cdot\|_c$-compact, meaning that the $\|\cdot\|_c$-closure of Θ is $\|\cdot\|_c$-compact. This is essentially what it means for the space $(\mathscr{F}, \|\cdot\|_s)$ to be *compactly embedded* in the space $(\mathscr{G}, \|\cdot\|_c)$, which is denoted with $\mathscr{F} \hookrightarrow \mathscr{G}$. See appendix A for a precise definition. Next, we show that Θ is actually $\|\cdot\|_c$-closed, and hence its $\|\cdot\|_c$-closure is just Θ itself. Consequently, Θ itself is $\|\cdot\|_c$-compact.

Thus our first result concerns compact embeddings.

Theorem 1 (Compact Embedding)**.** Let $\mathcal{D} \subseteq \mathbb{R}^{d_x}$ be a bounded open set, where $d_x \geq 1$ is some integer. Let $m, m_0 \geq 0$ be integers. Let $\nu \in (0, 1]$. Then the following embeddings are compact:

1. $\mathscr{W}_{m+m_0,2} \hookrightarrow \mathscr{C}_{m,\infty}$, if $m_0 > d_x/2$ and $\mathcal{D}$ satisfies the cone condition.
2. $\mathscr{W}_{m+m_0,2} \hookrightarrow \mathscr{W}_{m,2}$, if $m_0 > d_x/2$ and $\mathcal{D}$ satisfies the cone condition.
3. $\mathscr{C}_{m+m_0,\infty} \hookrightarrow \mathscr{C}_{m,\infty}$, if $m_0 \geq 1$ and $\mathcal{D}$ is convex.
4. $\mathscr{C}_{m+m_0,\infty} \hookrightarrow \mathscr{W}_{m,2}$, if $m_0 > d_x/2$, and $\mathcal{D}$ satisfies the cone condition.
5. $\mathscr{C}_{m+m_0,\infty,\mathbb{1},\nu} \hookrightarrow \mathscr{C}_{m,\infty}$, for $m_0 \geq 0$.

The cone condition is formally defined in appendix A. As we cite in the proof, all of these results are well known in mathematics. Result 5 shows that sets bounded under the Hölder norm are relatively compact under the Sobolev sup-norm, even with the same number of derivatives; the extra Hölder coefficient piece is sufficient to yield relative compactness. Result 3 shows that sets bounded under Sobolev sup-norms are relatively compact under Sobolev sup-norms using fewer derivatives. Result 2 shows that sets bounded under Sobolev L_2 norms are relatively compact under Sobolev L_2 norms with fewer derivatives, where the number of derivatives we have to drop depends on the dimension d_x of the domain. Finally, results 1 and 5 show the relationship between the Sobolev sup-norm and the Sobolev L_2 norm. Sets bounded under one are relatively compact under the other with fewer derivatives, where again the number of derivatives we must drop depends on d_x. Results 1, 2, and 4 require $\mathcal{D}$ to satisfy the cone condition, which is a geometric regularity condition on the shape of $\mathcal{D}$. When $d_x = 1$, a sufficient condition for the cone condition is that $\mathcal{D}$ is a finite union of open intervals. When $d_x > 1$, a sufficient condition is that $\mathcal{D}$ is the product of such finite unions.

By combining cases 4 and 5 and applying lemma 4, we also obtain compact embedding of Hölder spaces into Sobolev L_2 spaces. Here and throughout the paper, however, we focus only on the function space combinations which are most commonly used in econometrics.

Theorem 1 only shows that sets bounded under the norm $\|\cdot\|_s$ on the left hand side of the $\hookrightarrow$ are relatively compact under the norm $\|\cdot\|_c$ on the right hand side of the $\hookrightarrow$. As mentioned earlier, this means that their $\|\cdot\|_c$-closure is $\|\cdot\|_c$-compact. The following theorem shows that in some of these cases, $\|\cdot\|_s$-closed balls are $\|\cdot\|_c$-closed as well.

Theorem 2 (Closedness)**.** Let $\mathcal{D} \subseteq \mathbb{R}^{d_x}$ be a bounded open set, where $d_x \geq 1$ is some integer. Let $m, m_0 \geq 0$ be integers. Let $\nu \in (0, 1]$. Let $(\mathscr{F}, \|\cdot\|_s)$ and $(\mathscr{G}, \|\cdot\|_c)$ be Banach spaces with $\mathscr{F} \subseteq \mathscr{G}$, where $\|f\|_s < \infty$ for all $f \in \mathscr{F}$ and $\|f\|_c < \infty$ for all $f \in \mathscr{G}$. Define Θ as in equation (1). Then the results in table 1 hold. For cases (1) and (2) we also assume $m_0 > d_x/2$ and $\mathcal{D}$ satisfies the cone condition. For cases (3) and (4) we also assume $m_0 \geq 1$. For case (5) we also assume $\mathcal{D}$ satisfies the cone condition.

	$\|\cdot\|_s$	$\|\cdot\|_c$	Θ is $\|\cdot\|_c$-closed?
(1)	$\|\cdot\|_{m+m_0,2}$	$\|\cdot\|_{m,\infty}$	Yes
(2)	$\|\cdot\|_{m+m_0,2}$	$\|\cdot\|_{m,2}$	Yes
(3)	$\|\cdot\|_{m+m_0,\infty}$	$\|\cdot\|_{m,\infty}$	No
(4)	$\|\cdot\|_{m+m_0,\infty}$	$\|\cdot\|_{m,2}$	No
(5)	$\|\cdot\|_{m+m_0,\infty,\mathbb{1},\nu}$	$\|\cdot\|_{m,\infty}$	Yes

Table 1

Results 1, 2, and 5 of theorem 2 combined with results 1, 2, and 5 of theorem 1 give pairs of strong and consistency norms such that the $\|\cdot\|_s$-ball Θ defined in equation (1) is $\|\cdot\|_c$-compact. We illustrate how to apply these results in section 6. We also discuss additional implications of the choice of norms in that section.

For results 3 and 4, however, we see that Θ is *not* $\|\cdot\|_c$-closed. We could nonetheless proceed by simply agreeing to just work with the $\|\cdot\|_c$-closure $\overline{\Theta}$ of Θ instead. Theorem 1 then ensures that this $\|\cdot\|_c$-closure is $\|\cdot\|_c$-compact. Moreover, by the very definition of the closure, every element in the closure can be approximated arbitrarily by an element in the original set. Hence, as is needed in econometrics applications, we can construct sequences of approximations that still satisfy any necessary rate conditions. In sieve estimation, the choice of sieve space in practice also will not be affected by whether we use the closure or not.[6] Working with the closure is precisely what Gallant and Nychka (1987) did, until Santos' (2012) lemma A.1 showed that their parameter space was actually closed, thus proving result 2 in theorem 2 above.

Nonetheless, as with Santos' (2012) result, it is informative to know when the closure can be characterized. In case 3, a simple characterization is possible. Here the strong norm is the Sobolev sup-norm. It turns out that the $\|\cdot\|_c$-closure is precisely a Hölder space with exponent $\nu = 1$, as we show in the supplemental appendix L. Hence, there is no difference between working with the $\|\cdot\|_c$-closure in case 3 or just using case 5 with $\nu = 1$ and one fewer derivative (the closure in case 3 will contain functions whose $m + m_0$'th derivatives do not exist). This is one reason why we sometimes use the Hölder norm rather than the conceptually simpler Sobolev sup-norm. We are unaware of any simple characterizations of the closure in case 4.

[6] This extension to the closure may, however, affect other assumptions in one's analysis. For example, an estimation criterion function will now be defined over the closure and hence any assumptions on it typically must be satisfied over this extended domain.

4 Functions on unbounded domains

Gallant and Nychka (1987) extended the first compact embedding result from theorem 1 to spaces of functions on $\mathcal{D} = \mathbb{R}^{d_x}$. In this section, we show how to further extend their result in several ways. In particular, our results allow for exponential weighting functions, as well as the standard polynomial weighting functions used by Gallant and Nychka and subsequent authors. We also extend results 2–4 of theorem 1 as well as the closedness results of theorem 2 to $\mathcal{D} = \mathbb{R}^{d_x}$. All of these results use *weighted* norms, as introduced in section 2. There are at least two reasons to use weighted norms for functions on $\mathbb{R}^{d_x}$. The first is that many functions do not satisfy unweighted norm bounds. For example, the linear function $f(x) = x$ on $\mathbb{R}$ has $\|f\|_{0,\infty} = \infty$. By sufficiently *downweighting* the tails of f, however, the linear function can have a finite weighted sup-norm. The second reason is that even when a function f satisfies an unweighted norm, we can *upweight* the tails of f, which yields a stronger norm than the unweighted norm. This makes our concept of convergence finer. As in Gallant and Nychka's application, this is often the case with probability density functions, since they must converge to zero in their tails.

A further subtlety is that we actually need to use two different weighting functions—one for the strong norm $\|\cdot\|_s$, denoted by μ_s, and another for the consistency norm $\|\cdot\|_c$, denoted by μ_c. The reason comes from the main step in Gallant and Nychka's compact embedding argument. Their idea was to truncate the domain $\mathcal{D} = \mathbb{R}^{d_x}$ by considering a ball centered at the origin and its complement. Inside the ball, we can apply one of the results from theorem 1. The piece outside the ball, which depends on tail values of the functions and their weights, is made small by swapping out one weight function for another, and then using the properties of these two weight functions.

In the following subsection 4.1, we discuss the various classes of weight functions we will use. In many cases, these weight functions are more general than those considered in Gallant and Nychka (1987) and elsewhere in the literature. In subsection 4.2 we give the main compact embedding and closedness results for functions on $\mathcal{D} = \mathbb{R}^{d_x}$.

4.1 Weight functions

Throughout this section we let $\mu, \mu_c, \mu_s : \mathcal{D} \to \mathbb{R}_+$ be nonnegative functions and $m, m_0 \geq 0$ be integers. We first discuss some general properties of weight functions. We then examine several specific examples. We conclude by discussing general assumptions on the classes of weight functions we use in our main compact embedding and closedness results, and show that these hold for specific examples.

Our first result is simple, but important.

Proposition 1. Suppose there are constants M_1 and M_2 such that

$$0 < M_1 \leq \mu(x) \leq M_2 < \infty$$

for all $x \in \mathcal{D}$. Then

1. $\|\cdot\|_{m,\infty,\mu}$ and $\|\cdot\|_{m,\infty}$ are equivalent norms.

2. $\|\cdot\|_{m,2,\mu}$ and $\|\cdot\|_{m,2}$ are equivalent norms.

Proposition 1 says that weight functions which are bounded away from zero and infinity are trivial in the sense that they do not actually generate a new topology. Consequently, any nontrivial weight function must either diverge to infinity (upweighting) or converge to zero (downweighting) for some sequence of points in $\mathcal{D}$. These are the only two kinds of weight functions we must consider.

The next result shows that upweighting only allows for functions which go to zero in their tails.

Proposition 2. Let $\mathcal{D} = \mathbb{R}^{d_x}$. Suppose $\mu(x) \to \infty$ as $\|x\|_e \to \infty$. Suppose that for some constant $B < \infty$, either (a) $\|f\|_{0,\infty,\mu} \leq B$ or (b) $\|f\|_{0,2,\mu} \leq B$ holds. Then $f(x) \to 0$ as $\|x\|_e \to \infty$.

This result implies that derivatives of f must go to zero in the tails when f is bounded in one of the upweighted Sobolev norms $\|\cdot\|_{m,\infty,\mu}$ or $\|\cdot\|_{m,2,\mu}$ with $m > 0$. Proposition 2 implies that the choice between upweighting and downweighting will primarily depend on whether we want to study spaces with functions f that do not go to zero at infinity. For example, for spaces of regression functions, we typically will choose downweighting.[7] For spaces of probability density functions, we typically will choose upweighting as in Gallant and Nychka (1987).

Polynomial weights

The most common weight function used in econometrics is the polynomial weighting function,

$$\begin{aligned}\mu(x) &= (1 + x'x)^{\delta} \\ &= (1 + \|x\|_e^2)^{\delta},\end{aligned}$$

where $\delta \in \mathbb{R}$ is a constant. If $\delta > 0$ then this function upweights for large values of x, while if $\delta < 0$ then this function downweights for large values of x. These possibilities are illustrated in figure 1.

One reason that polynomial weights are ubiquitous is that the well-known compact embedding result of Gallant and Nychka (1987) applies specifically to polynomial weights. In our theorem 3 below, we restate this result and show how to generalize it to allow for additional classes of weight functions. There, as in section 3, we want to understand when spaces of functions

$$\Theta = \{f \in \mathscr{F} : \|f\|_s \leq B\}$$

are $\|\cdot\|_c$-compact, where $(\mathscr{F}, \|\cdot\|_s)$ is a Banach space and $B < \infty$ is a constant. To allow for the space $\mathscr{F}$ to contain functions with domain $\mathcal{D} = \mathbb{R}^{d_x}$, we will choose $\|\cdot\|_s$ and $\|\cdot\|_c$ to be weighted norms, with corresponding weights μ_s and μ_c, respectively.

[7] See, however, Newey and Powell (2003) page 1569, who use upweighting for spaces of regression functions, but include a parametric component to their function spaces to allow for certain unbounded functions. We discuss this further in section 6.

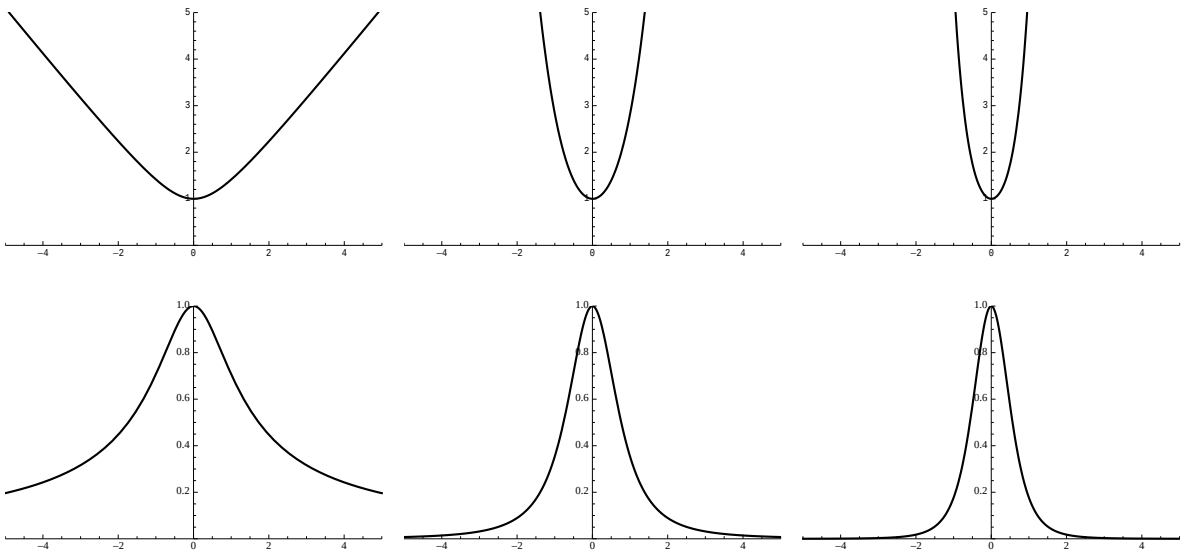

Figure 1: Polynomial weighting functions $\mu(x) = (1+x^2)^{\delta}$. Top: Upweighting, with $\delta = 0.5, 1.5, 2.5$ from left to right. Bottom: Downweighting, with $\delta = -0.5, -1.5, -2.5$ from left to right.

To understand what it means for a function to have a bounded weighted norm, consider the Sobolev sup-norm case where $\|\cdot\|_s = \|\cdot\|_{m+m_0,\infty,\mu_s}$ with polynomial weights $\mu_s(x) = (1+x'x)^{\delta_s}$. Then $f \in \Theta$ implies that

$$\sup_{x\in\mathbb{R}^{d_x}} |\nabla^{\lambda} f(x)|(1+x'x)^{\delta_s} \leq B$$

for every $0 \leq |\lambda| \leq m + m_0$. Consider the upweighting case, $\delta_s > 0$. We have already pointed out that upweighting implies the levels of f and its derivatives must go to zero in their tails. But here, with the specific polynomial form on the weight function, we know the precise rate at which the tails must go to zero:

$$|\nabla^{\lambda} f(x)| = O\big(\mu_s(x)^{-1}\big) = O\big((1+x'x)^{-\delta_s}\big) \tag{2}$$

as $\|x\|_e \to \infty$, for each $0 \leq |\lambda| \leq m+m_0$. For example, with $d_x = 1$ and $\delta_s = 1$, $|f(x)|$ can go to zero at the same rate as $\mu_s(x)^{-1} = 1/(1+x^2) = O(x^{-2})$. But it cannot go to zero any slower, because that would violate the norm bound. Recall that the t-distribution with one degree of freedom has pdf $C/(1+x^2)$ where C is a normalizing constant. So the fattest tails $|f(x)|$ can have are these t-like tails.

Next consider the downweighting case, $\delta_s < 0$. Then $|f(x)|$ no longer has to converge to zero in the tails. But it also cannot diverge too quickly. The norm bound tells us exactly how fast it can diverge, which is given exactly as in equation (2). For example, with $d_x = 1$ and $\delta_s = -1$, $|f(x)|$ can diverge at the rate $\mu_s(x)^{-1} = 1 + x^2 = O(x^2)$. This point implies that with polynomial weights, the choice of δ_s determines the maximum order polynomial that is in Θ. In general, for $\delta_s = -n$ where n is a natural number, $\mu_s(x)^{-1} = O(x^{2n})$ is the highest order polynomial allowed. Similar analysis applies for the Sobolev L_2 norm, for both downweighting and upweighting.

Exponential weights

An alternative to polynomial weighting are the exponential weights

$$\begin{aligned}\mu(x) &= [\exp(x'x)]^{\delta} \\ &= \exp(\delta \|x\|_e^2),\end{aligned}$$

where $\delta \in \mathbb{R}$ is a constant. $\delta > 0$ corresponds to upweighting, while $\delta < 0$ corresponds to downweighting. These possibilities have similar qualitative appearances to the polynomial weights in figure 1.

As with polynomial weights, we want to understand what it means for a function to be in the $\|\cdot\|_s$-ball Θ, where $\|\cdot\|_s$ is a weighted norm. Consider the Sobolev sup-norm case $\|\cdot\|_s = \|\cdot\|_{m+m_0,\infty,\mu_s}$ with $\mu_s(x) = \exp[\delta_s(x'x)]$. Then $f \in \Theta$ implies that

$$\sup_{x \in \mathbb{R}^{d_x}} |\nabla^\lambda f(x)| \exp[\delta_s(x'x)] \leq B$$

for every $0 \leq |\lambda| \leq m + m_0$. Hence

$$|\nabla^\lambda f(x)| = O\big(\mu_s(x)^{-1}\big) = O\big(\exp[-\delta_s(x'x)]\big)$$

as $\|x\|_e \to \infty$, for each $0 \leq |\lambda| \leq m+m_0$. Consider the downweighting case $\delta_s < 0$. Then we see that by using exponential weights we can allow for $|\nabla^\lambda f(x)|$ to diverge to infinity at an exponential rate. In particular, $|\nabla^\lambda f(x)|$ can diverge at *any* polynomial rate. More precisely, $|\nabla^\lambda f(x)|$ proportional to x^n for any natural number $n > 0$ will satisfy the specified rate, regardless of the value of $\delta_s < 0$. In contrast, using a polynomial downweighting function requires specifying a maximum order of polynomial allowed.

Consider the upweighting case, $\delta_s > 0$. We have already pointed out that upweighting implies the levels of f and its derivatives must go to zero in their tails. But here, with the specific polynomial form on the weight function, we know the precise rate at which the tails must go to zero: $O\big(\exp[-\delta_s(x'x)]\big)$. In applications, this is likely to be very restrictive. For example, it rules out t-distribution like tails. For this reason, we do not discuss exponential upweighting any further. Similar analysis applies for the Sobolev L_2 norm, for both downweighting and upweighting.

While we focus on the weights $\mu(x) = \exp(\delta\|x\|_e^2)$ throughout this paper, one could consider a wide variety of exponential weight functions, such as $\exp(\delta\|x\|_e^\kappa)$ where $\kappa \in \mathbb{R}$ is an additional weight function parameter. Another possibility is to use a different finite dimensional norm, like the ℓ_1-norm $\|x\|_1 = \sum_{k=1}^{d_x} |x_k|$. This yields the weight function $\exp(\delta\|x\|_1^\kappa)$.

Assumptions on weight functions

With these two main classes of weight functions in mind, we state our main results for the two general weight functions μ_s and μ_c used in defining the strong and consistency norms. We will,

however, make several assumptions on these weight functions. We then verify that these assumptions hold for either polynomial or exponential weights, or both. The first assumption states that the consistency norm weight goes to zero faster than the strong norm weight as we go further out in the tails.

Assumption 1.

$$\frac{\mu_c(x)}{\mu_s(x)} \to 0$$

as $\|x\|_e \to \infty$ (for $\mathcal{D} = \mathbb{R}^{d_x}$) or as $\text{dist}(x, \text{Bd}(\overline{\mathcal{D}})) \to 0$ (for bounded $\mathcal{D}$).

Here $\text{dist}(x, \text{Bd}(\overline{\mathcal{D}})) = \min_{y \in \text{Bd}(\overline{\mathcal{D}})} \|x - y\|_e \to 0$ where $\text{Bd}(\overline{\mathcal{D}})$ denotes the boundary of the closure of $\mathcal{D}$. As discussed earlier, the key idea to prove compact embedding is to truncate the domain $\mathbb{R}^{d_x}$, and then ensure that the norm outside the truncated region is small. Assumption 1 is one part of ensuring that this step works. Both polynomial weights

$$\mu_c(x) = (1 + x'x)^{\delta_c} \quad \text{and} \quad \mu_s(x) = (1 + x'x)^{\delta_s}$$

and exponential weights

$$\mu_c(x) = \exp[\delta_c(x'x)] \quad \text{and} \quad \mu_s(x) = \exp[\delta_s(x'x)]$$

have the form $\rho(x)^\delta$ where $\rho(x) \to \infty$ as $\|x\| \to \infty$. Hence for both kinds of weights the ratio is

$$\frac{\mu_c(x)}{\mu_s(x)} = \rho(x)^{\delta_c - \delta_s}$$

and so assumption 1 holds by choosing $\delta_c < \delta_s$.

The following assumption, which bounds the ratio for all x, not just x's in the limit, plays a similar role in the proof.

Assumption 2. There is a finite constant $M_5 > 0$ such that

$$\frac{\mu_c(x)}{\mu_s(x)} \leq M_5$$

for all $x \in \mathcal{D}$.

As above, assumption 2 holds for both polynomial and exponential weights with $\delta_c < \delta_s$. The next assumptions bounds the derivatives of the (square root) strong norm weight function by its (square root) levels.

Assumption 3. There is a finite constant $K > 0$ such that

$$|\nabla^\lambda \mu_s^{1/2}(x)| \leq K \mu_s^{1/2}(x)$$

for all $|\lambda| \leq m + m_0$ and for all $x \in \mathcal{D}$.

This assumption is precisely what Gallant and Nychka (1987) used in their analysis. This assumption was also used by Schmeisser and Triebel (1987) page 246 equation 2, and followup work including Haroske and Triebel (1994a,b) and Edmunds and Triebel (1996). Gallant and Nychka's lemma A.2 proves the following result.

Proposition 3. Let $\mu_s(x) = (1 + x'x)^{\delta_s}$ and $\mathcal{D} = \mathbb{R}^{d_x}$. Then assumption 3 holds for any integers $m, m_0 \geq 0$ and any $\delta_s \in \mathbb{R}$.

Assumption 3 also holds for certain kinds of exponential weights. For example, for $d_x = 1$ and $\delta_s = -1$ consider $\mu_s(x) = \exp(-|x|)$. Then the weak derivative of $\sqrt{\mu_s(x)}$ with respect to x is $-\sqrt{\mu_s(x)}\text{sign}(x)$, and hence

$$\frac{\left|\frac{\partial}{\partial x}\sqrt{\mu_s(x)}\right|}{\sqrt{\mu_s(x)}} = |-\text{sign}(x)| \leq 1.$$

Assumption 3 does *not* allow for many other kinds of exponential weights, however. For example, consider $d_x = 1$ and $\delta_s = -1$ again but now using the Euclidean norm for x:

$$\mu_s(x) = \exp(-x^2).$$

Then

$$\frac{\partial}{\partial x}\sqrt{\mu_s(x)} = -x\sqrt{\mu_s(x)}$$

and hence

$$\frac{\left|\frac{\partial}{\partial x}\sqrt{\mu_s(x)}\right|}{\sqrt{\mu_s(x)}} = |x|.$$

The function $|x|$ is unbounded on $\mathbb{R}$ and so assumption 3 fails. The function $|x|$ is, however, bounded for any compact subset of $\mathbb{R}$. This motivates the following weaker version of assumption 3.

Assumption 4. For every compact subset $\mathcal{C} \subseteq \mathcal{D}$, there is a constant $K_{\mathcal{C}} < \infty$ such that

$$|\nabla^{\lambda}\mu_s^{1/2}(x)| \leq K_{\mathcal{C}}\mu_s^{1/2}(x)$$

for all $|\lambda| \leq m + m_0$ and for all $x \in \mathcal{C}$.

This relaxation of assumption 3 will also be important in section 5 when we consider weighted norms for functions with bounded domains. The following proposition shows that exponential weights using the Euclidean norm satisfy assumption 4. Also note that polynomial weights immediately satisfy it since they satisfy the stronger assumption 3.

Proposition 4. Let $\mu_s(x) = \exp[\delta_s(x'x)]$ and $\mathcal{D} = \mathbb{R}^{d_x}$. Then assumption 4 holds for any integers $m, m_0 \geq 0$ and any $\delta_s \in \mathbb{R}$.

Finally, for one of our results we use the following assumption.

Assumption 5. There is a function $g(x)$ such that the following hold.

1. $g(x) \to \infty$ as $\|x\|_e \to \infty$ (for $\mathcal{D} = \mathbb{R}^{d_x}$) or as $\text{dist}(x, \text{Bd}(\overline{\mathcal{D}})) \to 0$ (for bounded $\mathcal{D}$).
2. For $\tilde{\mu}_c^{1/2}(x) \equiv g(x)\mu_c^{1/2}(x)$ there is a constant $M < \infty$ such that

$$\max_{0 \leq |\lambda| \leq m_0} |\nabla^\lambda \tilde{\mu}_c^{1/2}(x)| \leq M \mu_s^{1/2}(x)$$

for all $x \in \mathcal{D}$.

In the supplemental appendix K we give some intuitive discussion of assumption 5. The main purpose of considering assumption 5 is similar to our motivation for assumption 4: it allows for cases where assumption 3 does not hold. In particular, in the following proposition we show that assumption 5 holds for exponential weights.

Proposition 5. Let $\mu_c(x) = \exp[\delta_c(x'x)]$, $\mu_s(x) = \exp[\delta_s(x'x)]$, and $\mathcal{D} = \mathbb{R}^{d_x}$. Then assumption 5 holds for any $\delta_s, \delta_c \in \mathbb{R}$ such that $\delta_c < \delta_s$.

Our final assumption on the weight functions ensures that the weighted spaces are complete. See Kufner and Opic (1984) and more recently Rodríguez et al. (2004) for more details. This assumption is a minor modification of the first part of assumption H in Brown and Opic (1992).[8]

Assumption 6. Let $\mathcal{M} = \{x \in \mathcal{D} : \mu_c(x) \neq 0\}$. Then for any bounded open subset $\mathcal{O} \subseteq \mathcal{M}$, (1) μ_c is continuous on $\mathcal{O}$ and (2) μ_c is bounded above and below by positive constants on $\mathcal{O}$.

For $\mathcal{D} = \mathbb{R}^{d_x}$, assumption 6 rules out weights like $\mu_c(x) = (x'x)^2$ since then $\mu_c(x)$ is not bounded away from zero on $(0, 1)$, for example. This assumption is satisfied by $\mu_c(x) = (1 + x'x)^2$, however, and more generally for $\mu_c(x) = (1 + x'x)^{\delta_c}$, $\delta_c \in \mathbb{R}$. It is also satisfied by the exponential weights $\mu_c(x) = \exp[\delta_c(x'x)]$. This assumption is also satisfied by indicator weight functions like $\mu_c(x) = \mathbb{1}(\|x\|_e \leq M)$ for some constant M.

4.2 Compact embeddings and closedness results

As in the bounded domain case, we begin with a compact embedding result.

Theorem 3 (Compact Embedding)**.** Let $\mathcal{D} = \mathbb{R}^{d_x}$ for some integer $d_x \geq 1$. Let $\mu_c, \mu_s : \mathcal{D} \to \mathbb{R}_+$ be nonnegative, $m + m_0$ times continuously differentiable functions. $m, m_0 \geq 0$ are integers. Suppose assumptions 1, 2, 4, and 6 hold. Then the following embeddings are compact:

1. $\mathscr{W}_{m+m_0,2,\mu_s} \hookrightarrow \mathscr{C}_{m,\infty,\mu_c^{1/2}}$, if $m_0 > d_x/2$ and either of assumption 3 or 5 holds.
2. $\mathscr{W}_{m+m_0,2,\mu_s} \hookrightarrow \mathscr{W}_{m,2,\mu_c}$, if $m_0 > d_x/2$.
3. $\mathscr{C}_{m+m_0,\infty,\mu_s} \hookrightarrow \mathscr{C}_{m,\infty,\mu_c}$, if $m_0 \geq 1$.

[8] As discussed in the proof of theorem 3, assumption 6 could be weakened slightly to a local integrability assumption.

4. $\mathscr{C}_{m+m_0,\infty,\mu_s} \hookrightarrow \mathscr{W}_{m,2,\mu_c}$, if $m_0 > d_x/2$, μ_s is bounded away from zero for any compact subset of $\mathbb{R}^{d_x}$, and $\int_{\|x\|_e>J} \mu_c(x)/\mu_s^2(x)\, dx < \infty$ for some J.

Using the stronger assumption 3, Gallant and Nychka (1987) showed case (1) in their lemma A.4. Case (1) with polynomial weights was used, for example, by Newey and Powell (2003) and Santos (2012).[9] Under the stronger assumption 3, Haroske and Triebel (1994a) show cases (1)–(4) as special cases of their theorem on page 136. Haroske and Triebel furthermore assume via their definition 1(ii) on page 133 that the weight functions have at most polynomial growth. Their results therefore do not allow for any exponential weights. For example, for $d_x = 1$, they do not allow for either $\mu(x) = \exp(\delta|x|)$ or $\mu(x) = \exp(\delta x^2)$. Brown and Opic (1992) give high level conditions for a compact embedding result similar to case (1), with $m_0 = 1$ and $m = 0$. They do not study the other cases we consider. They do, however, allow for a large class of weight functions, which includes the exponential weight functions we discussed earlier (for example, see their example 5.5 plus remark 5.2).

To our best knowledge, cases (2)–(4) with any kind of exponential weight function have not been shown in the literature. The proof for these cases is similar to that for case (1), which is a modification of Gallant and Nychka's original proof. Our result for case (1) gives lower level conditions on the weight functions compared to Brown and Opic (1992), although these conditions are less general. Finally, note that the results by Triebel and coauthors allow for more general function spaces, including Besov spaces and many others, although again, they restrict attention to weight functions with at most polynomial growth.

Theorem 4 (Closedness)**.** Let $\mathcal{D} = \mathbb{R}^{d_x}$ where $d_x \geq 1$ is some integer. Let $m, m_0 \geq 0$ be integers. Let $(\mathscr{F}, \|\cdot\|_s)$ and $(\mathscr{G}, \|\cdot\|_c)$ be Banach spaces with $\mathscr{F} \subseteq \mathscr{G}$, where $\|f\|_s < \infty$ for all $f \in \mathscr{F}$ and $\|f\|_c < \infty$ for all $f \in \mathscr{G}$. Define Θ as in equation (1). Suppose assumptions 1, 2, and 4 hold. Then the results of table 2 hold. For cases (1) and (2) we also assume $m_0 > d_x/2$ and that assumption 6 holds, and in case (1) also that assumption 5 holds. For cases (3) and (4) we also assume $m_0 \geq 1$.

	$\|\cdot\|_s$	$\|\cdot\|_c$	Θ is $\|\cdot\|_c$-closed?
(1)	$\|\cdot\|_{m+m_0,2,\mu_s}$	$\|\cdot\|_{m,\infty,\mu_c^{1/2}}$	Yes
(2)	$\|\cdot\|_{m+m_0,2,\mu_s}$	$\|\cdot\|_{m,2,\mu_c}$	Yes
(3)	$\|\cdot\|_{m+m_0,\infty,\mu_s}$	$\|\cdot\|_{m,\infty,\mu_c}$	No
(4)	$\|\cdot\|_{m+m_0,\infty,\mu_s}$	$\|\cdot\|_{m,2,\mu_c}$	No

Table 2

Case (1) generalizes Santos (2012) lemma A.2, which only considered polynomial upweighting. Case (2) was also shown in the proof of Santos (2012) lemma A.2, again only for polynomial upweighting.

[9] Santos (2012) allowed for a general unbounded domain $\mathcal{D}$ rather than $\mathcal{D} = \mathbb{R}^{d_x}$ specifically. We restrict attention to functions with full support merely for simplicity.

Just as in section 3, theorems 3 and 4 can be combined to show that the $\|\cdot\|_s$-ball Θ is $\|\cdot\|_c$-compact by choosing various combinations of strong and consistency norms given in table 2. All of our remarks in that section apply here as well. The only new point is that in addition to the choice of norm, one must also choose the weight functions μ_s and μ_c.

4.3 Alternative approaches to defining weighted norms

Thus far we have defined weighted Sobolev and Hölder norms by weighting each derivative piece equally. For example, with $m = 1$ and $d_x = 1$, the weighted Sobolev sup-norm is

$$\|f\|_{1,\infty,\mu} = \max\left\{\sup_{x\in\mathcal{D}} |f(x)|\mu(x), \quad \sup_{x\in\mathcal{D}} |f'(x)|\mu(x)\right\}.$$

The Sobolev integral norms were defined similarly, with each derivative using the same weight function. Call this the *equal weighting* approach. While this is the most common approach to defining weighting norms in econometrics, it is not the only reasonable way to define weighted norms. The next most common alternative is to convert any unweighted norm $\|\cdot\|$ into a weighted norm $\|\cdot\|_\mu$ by first weighting the function and then applying the unweighted norm:

$$\|f\|_\mu = \|\mu f\|.$$

Call this the *product weighting* approach. For example, suppose we start with the unweighted Sobolev sup-norm, with $m = 1$ and $d_x = 1$. Assume μ is differentiable. Then

$$\begin{aligned}\|\mu f\|_{1,\infty} &= \max\left\{\sup_{x\in\mathcal{D}} |f(x)|\mu(x), \quad \sup_{x\in\mathcal{D}} |f'(x)\mu(x) + f(x)\mu'(x)|\right\} \\ &\leq \max\left\{\sup_{x\in\mathcal{D}} |f(x)|\mu(x), \quad \sup_{x\in\mathcal{D}} |f'(x)|\mu(x), \quad \sup_{x\in\mathcal{D}} |f(x)|\mu'(x)\right\}.\end{aligned}$$

Here we see that, compared to equal weighting, product weighting picks up an extra term involving the derivative of the weight function $\mu'(x)$. Notice that when $m = 0$, the product and equal weighting approaches to defining weighted Sobolev integral and sup-norms are equivalent.

The following result shows that, for a class of weight functions including polynomial weighting, these two approaches to defining Sobolev norms are equivalent. Consequently, it is irrelevant which one we use.

Proposition 6. Define the norms

$$\|\cdot\|_{m,2,\mu^{1/2},\text{ALT}} = \|\mu^{1/2}\cdot\|_{m,2} \qquad \text{and} \qquad \|\cdot\|_{m,\infty,\mu,\text{ALT}} = \|\mu\cdot\|_{m,\infty}.$$

Suppose assumption 3 holds for μ. Then

1. $\|\cdot\|_{m,2,\mu^{1/2},\text{ALT}}$ and $\|\cdot\|_{m,2,\mu}$ are equivalent norms.
2. $\|\cdot\|_{m,\infty,\mu,\text{ALT}}$ and $\|\cdot\|_{m,\infty,\mu}$ are equivalent norms.

As discussed earlier, assumption 3 does not hold for all feasible weight functions. So these two approaches to defining weighted norms are not necessarily equivalent for any given choice of weight function. The theorem in section 5.1.4 of Schmeisser and Triebel (1987) gives a result related to proposition 6 for a large class of weighted function spaces.[10]

A main reason to consider product weighting is that it easily applies when it is not clear how to define an equally weighted norm. In particular, it allows us to define the *weighted Hölder norm* by

$$\|\cdot\|_{m,\infty,\mu,\nu} = \|\mu \cdot\|_{m,\infty,\mathbb{1},\nu}$$

for $\nu \in (0,1]$. Let $\mathscr{C}_{m,\infty,\mu,\nu}(\mathcal{D}) = \{f \in \mathscr{C}_m(\mathcal{D}) : \|f\|_{m,\infty,\mu,\nu} < \infty\}$ denote the weighted Hölder space with exponent ν. The difficulty in defining an equally weighted Hölder norm comes from the Hölder coefficient piece, which is a supremum over two different points in the domain, unlike the sup-norm part.[11] The product weighted Hölder norm is commonly used in econometrics, as in Ai and Chen (2003) example 2.1[12], Chen et al. (2005), Hu and Schennach (2008), and Khan (2013).

If $\mathcal{D}$ is bounded, then compact embedding and closedness results for product weighted norms follow immediately from our results on bounded $\mathcal{D}$ with unweighted norms. For unbounded $\mathcal{D}$, we provide the following two results.

Theorem 5 (Compact Embedding)**.** Let $\mathcal{D} = \mathbb{R}^{d_x}$ for some integer $d_x \geq 1$. Let $\mu_c, \mu_s : \mathcal{D} \to \mathbb{R}_+$ be nonnegative, $m + m_0$ times continuously differentiable functions. Define $\tilde{\mu}(x) = (1 + x'x)^{-\delta}$ for some $\delta > 0$ and assume that $\mu_c(x) = \mu_s(x)\tilde{\mu}(x)$. Then the following embeddings are compact:

1. $\mathscr{W}_{m+m_0,2,\mu_s,\text{ALT}} \hookrightarrow \mathscr{C}_{m,\infty,\mu_c,\text{ALT}}$, if $m_0 > d_x/2$.
2. $\mathscr{W}_{m+m_0,2,\mu_s,\text{ALT}} \hookrightarrow \mathscr{W}_{m,2,\mu_c,\text{ALT}}$, if $m_0 > d_x/2$.
3. $\mathscr{C}_{m+m_0,\infty,\mu_s,\text{ALT}} \hookrightarrow \mathscr{C}_{m,\infty,\text{ALT}}$, if $m_0 \geq 1$.
4. $\mathscr{C}_{m+m_0,\infty,\mu_s,\nu} \hookrightarrow \mathscr{C}_{m,\infty,\mu_c,\text{ALT}}$, if $m_0 \geq 0$.

Under the stronger assumption 3, the product and equal weighted norms are equivalent, by proposition 6. Schmeisser and Triebel (1987) showed this equivalence and Haroske and Triebel (1994a) used it to to prove cases (1)–(4) of theorem 5 under assumption 3 and the further assumption that the weight functions have at most polynomial growth (definition 1(ii) on page 133 of Haroske

[10] This result is cited and applied in much of Triebel and coauthor's followup work. In particular, as Haroske and Triebel (1994a) show in the proof of their theorem 2.3 (page 145 step 1), this equivalence result can be used to prove compact embedding results. This proof strategy does not apply when the norms are not equivalent, which is why we rely on the more primitive approach of Gallant and Nychka (1987).

[11] See, however, Brown and Opic (1992) equations (2.8) and (2.9), who suggest one way to define equally weighted Hölder norms.

[12] In this example the parameter space is an unweighted Hölder space for functions with unbounded domain, but the consistency norm is a downweighted sup-norm. Hence this is an example of case 4 in theorems 5 and 6. Also, as we discuss in section 6, this kind of unweighted parameter space assumption rules out linear functions. Note that in other examples using an unweighted Hölder space on $\mathbb{R}^{d_x}$ is less restrictive, since the functions of interest are naturally bounded. For example, Chen, Hu, and Lewbel (2009b) and Carroll, Chen, and Hu (2010) consider spaces of pdfs while Blundell, Chen, and Kristensen (2007) (assumption 2(i)) consider spaces of Engel curves.

and Triebel 1994a). Our result relaxes assumption 3 and does not impose a polynomial growth bound on the weight functions. Our cases (1)–(4) of theorem 5 are therefore the first we are aware of to allow for exponential weight functions when using product weighted norms.

We use our previous results in theorem 3 to prove cases (1)–(3). We adapt the proof of theorem 3 to prove case (4).

Theorem 6 (Closedness)**.** Let $\mathcal{D} = \mathbb{R}^{d_x}$ where $d_x \geq 1$ is some integer. Let $m, m_0 \geq 0$ be integers. Let $\nu \in (0, 1]$. Let $(\mathscr{F}, \|\cdot\|_s)$ and $(\mathscr{G}, \|\cdot\|_c)$ be Banach spaces with $\mathscr{F} \subseteq \mathscr{G}$, where $\|f\|_s < \infty$ for all $f \in \mathscr{F}$ and $\|f\|_c < \infty$ for all $f \in \mathscr{G}$. Define Θ as in equation (1). Define $\tilde{\mu}(x) = (1 + x'x)^{-\delta}$ for some $\delta > 0$ and assume that $\mu_c(x) = \mu_s(x)\tilde{\mu}(x)$. Then the results of table 3 hold. For cases (1) and (2) we also assume $m_0 > d_x/2$.

	$\|\cdot\|_s$	$\|\cdot\|_c$	Θ is $\|\cdot\|_c$-closed?
(1)	$\|\cdot\mu_s\|_{m+m_0,2}$	$\|\cdot\mu_c\|_{m,\infty}$	Yes
(2)	$\|\cdot\mu_s\|_{m+m_0,2}$	$\|\cdot\mu_c\|_{m,2}$	Yes
(3)	$\|\cdot\mu_s\|_{m+m_0,\infty}$	$\|\cdot\mu_c\|_{m,\infty}$	No
(4)	$\|\cdot\mu_s\|_{m+m_0,\infty,\mathbb{1},\nu}$	$\|\cdot\mu_c\|_{m,\infty}$	Yes

Table 3

As mentioned above, we do not impose assumption 3 on the strong norm in either theorem 5 or theorem 6. We also do not impose the weaker assumption 4. We do, however, strengthen assumptions 1 and 2 by assuming a particular rate of convergence on the ratio μ_c/μ_s, namely, that it is polynomial:

$$\frac{\mu_c(x)}{\mu_s(x)} = \frac{1}{(1 + x'x)^{\delta}}$$

for some $\delta > 0$. This assumption is satisfied when both μ_c and μ_s are polynomial weight functions themselves. This case has been used in the previous literature which chooses the weighted Hölder norm, such as in Chen et al. (2005). This assumption is also, however, satisfied by the choice

$$\mu_s(x) = \exp(\delta_s \|x\|_e^2) \qquad \text{and} \qquad \mu_c(x) = \frac{\exp(\delta_s \|x\|_e^2)}{(1 + x'x)^{\delta}}$$

for $\delta > 0$ and $\delta_s < 0$. Hence theorems 5 and 6 can still be applied if we want our parameter space Θ to contain for polynomial functions of all orders, as discussed earlier. Finally, note that a compact embedding result under the norm μ_c yields a compact embedding result under any weaker norm, by lemma 4. For example, with $m = 0$, μ_c defined as the ratio of an exponential and polynomial as above, and $\tilde{\mu}_c = \exp(\delta_c \|x\|_e^2)$ for $\delta_c < \delta_s$, $\|\cdot\|_{0,\infty,\tilde{\mu}_c}$ is weaker than $\|\cdot\|_{0,\infty,\mu_c}$. Theorem 5 part 4 then implies that $\mathscr{C}_{0,\infty,\mu_s,\nu}$ is compactly embedded in $\mathscr{C}_{0,\infty,\tilde{\mu}_c}$.

5 Weighted norms for bounded domains

In section 3 we showed that when the domain $\mathcal{D}$ is bounded, sets of functions f that satisfy a norm bound $\|f\|_s \leq B$ are $\|\cdot\|_c$-compact for three possible choices of norm pairs—see table 1. In this section we consider functions with a bounded domain, but which do *not* satisfy a norm bound $\|\cdot\|_s \leq B$ for any of the choices in table 1.

Example 5.1 (Quantile function)**.** *Let X be a scalar random variable with full support on $\mathbb{R}$ and absolutely continuous distribution with respect to the Lebesgue measure. Let $Q_X : (0,1) \to \mathbb{R}$ denote its quantile function. Since the derivative of Q_X asymptotes to $\pm\infty$ as $\tau \to 0$ or 1, $\|Q_X\|_{0,\infty} = \infty$. Hence, although the domain $\mathcal{D} = (0,1)$ is bounded, Q_X is not in any Sobolev sup-norm space or Hölder space. Indeed, Csörgö (1983, page 5) notes that*

$$\|\widehat{Q}_X - Q_X\|_{0,\infty} \to \infty \qquad a.s.$$

as $n \to \infty$ where

$$\widehat{Q}_X(\tau) = \inf\{x : \widehat{F}_X(x) \geq \tau\}$$

is the sample quantile function for an iid sample $\{x_1, \ldots, x_n\}$, and $\widehat{F}_X$ is the empirical cdf. Also see page 322 of van der Vaart (2000).

On the other hand, it is certainly possible for such a quantile function Q_X to be bounded in a weighted *Sobolev sup-norm space or a weighted Hölder space. In fact, by examining the Bahadur representation of $\widehat{Q}_X$ it can be shown that $\widehat{Q}_X$ converges in the weighted sup-norm over $\tau \in (0,1)$ with weight function*

$$f_X(F_X^{-1}(\tau)) = \frac{\partial Q_X(t)}{\partial t}\Big|_{t=\tau}.$$

Note that this weight function depends on how fast the quantile function diverges as $\tau \to 0$ or $\tau \to 1$.

More generally, we may want to estimate quantile functions in settings more complicated than simply taking a sample quantile. In such settings, the compact embedding and closedness results developed in this section can be useful.

Example 5.2 (Transformation models)**.** *Consider the model*

$$T(Y) = \alpha + X\beta + U, \qquad U \perp\!\!\!\perp X.$$

where Y, X, and U are continuously distributed scalar random variables. T is an unknown strictly increasing transformation function. Let F_U and f_U be the (unknown) cdf and pdf of U, respectively.

Suppose Y has compact support $\text{supp}(Y) = [y_L, y_U]$*. If we allow distributions of U to have full support, like $\mathcal{N}(0,1)$, then the transformation function $T(y)$ must diverge to infinity as $y \to y_U$ or to negative infinity as $y \to y_L$. We are again in a situation like the quantile function above, where because the derivatives of T diverge, it is not in any unweighted Sobolev sup-norm or Hölder space.*

Horowitz (1996) constructs an estimator $\widehat{T}(y)$ of $T(y)$ and shows, among other results, that

$$\sup_{y\in[a,b]} |\widehat{T}(y) - T(y)| \xrightarrow{p} 0,$$

where a and b are such that $T(y)$ and $T'(y)$ are bounded on $[a,b]$. These bounds on T and T' imply that $[a,b]$ is a strict subset of $\text{supp}(Y)$ *when* $\text{supp}(Y)$ *is compact and U has full support. Chiappori, Komunjer, and Kristensen (2015) extend the arguments in Horowitz (1996) to allow for a nonparametric regression function and endogenous regressors. Also see Chen and Shen (1998), who study a transformation model assuming Y has bounded support in their example 3, and example 3 on page 618 of Wong and Severini (1991).*

As with the quantile function, the compact embedding and closedness results developed in this section may be useful for proving consistency of estimators of T in weighted norms uniformly over its entire domain.

These examples show that sometimes our functions of interest do not satisfy standard unweighted norm bounds. Hence the compactness and closedness results theorems 1 and 2 do not apply. In this section, we show that we can, however, recover compactness by using weighted norms. As in section 4, we focus on equal weighting norms.[13]

5.1 Weight functions

Proposition 1 applies for bounded domains, and hence again we see that only weight functions that go to zero or infinity at the boundary are nontrivial. Since our main motivation for considering weighted norms is to expand the set of functions which can have a bounded norm, we will restrict attention to downweighting. For simplicity we will also focus on the one dimensional case $d_x = 1$ with $\mathcal{D} = (0,1)$, as in the quantile function example. As before, there are two natural classes of weight functions. First, we consider polynomial weights

$$\mu(x) = [x^\alpha (1-x)^\beta]^\delta$$

for $\alpha, \beta \geq 0$ and $\delta \in \mathbb{R}$. $\alpha > 1$, $\beta > 1$, and $\delta > 0$ ensure that $\mu(x) \to 0$ as $x \to 0$ or $x \to 1$. Next, we consider exponential weights,

$$\mu(x) = \exp[\delta x^\alpha (1-x)^\beta].$$

For example, with $\delta = \alpha = \beta = -1$,

$$\mu(x) = \exp\left[\frac{-1}{x(1-x)}\right].$$

[13] Compactness and closedness results for product weighting norms with bounded domains follow immediately from theorems 1 and 2 regarding unbounded domains.

If we had $\alpha > 0$ and $\beta < 0$ then this allows for asymmetric weights where the tail goes to zero at one boundary point but not the other. Figure 2 illustrates some of these weight functions.

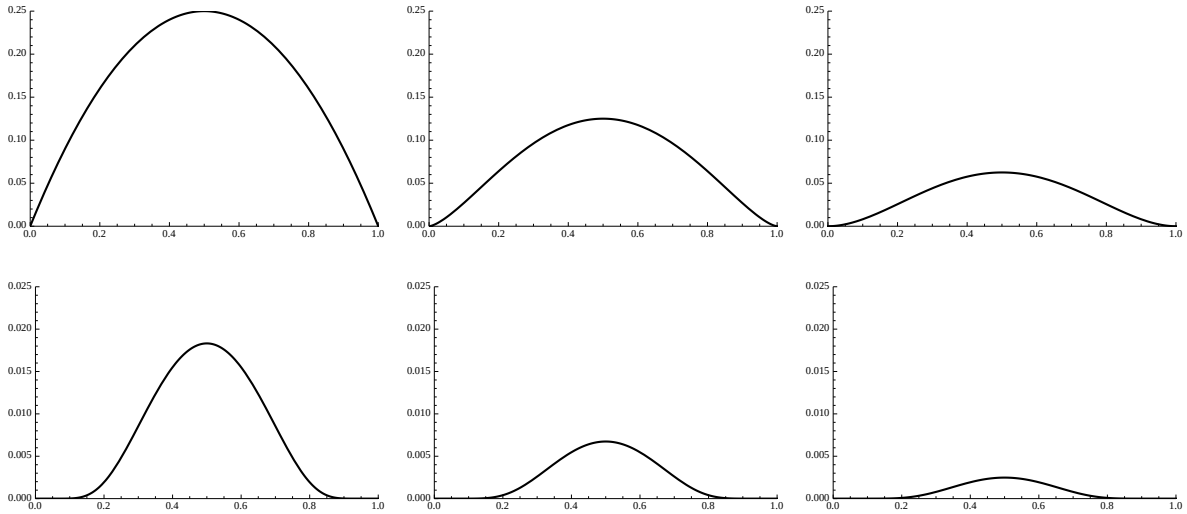

Figure 2: Top: Polynomial weighting functions $\mu(x) = [x(1-x)]^{\delta}$ for $\delta = 1, 1.5, 2$, from left to right. Bottom: Exponential weighting functions $\mu(x) = \exp[\delta x^{-1}(1-x)^{-1}]$ with $\delta = -1, -1.25, -1.5$, from left to right.

The interpretation of $\|f\|_s \leq B$ for a weighted norm $\|\cdot\|_s$ with $\mathcal{D}$ bounded is similar to the interpretation when $\mathcal{D} = \mathbb{R}^{d_x}$ discussed in section 4.1. This norm bound places restrictions on the tail behavior of $f(x)$ as x approaches the boundary of $\overline{\mathcal{D}}$. For example, let $\mathcal{D} = (0,1)$ and consider the Sobolev sup-norm $\|\cdot\|_s = \|\cdot\|_{m+m_0,\infty,\mu_s}$ with polynomial weights $\mu_s(x) = [x(1-x)]^{\delta_s}$, $\delta_s > 0$. Then $f \in \Theta = \{f \in \mathscr{F} : \|f\|_s \leq B\}$ implies that

$$\sup_{x \in \mathcal{D}} |\nabla^{\lambda} f(x)| x^{\delta_s} (1-x)^{\delta_s} \leq B$$

for every $0 \leq |\lambda| \leq m + m_0$. For example,

$$|f(x)| = O(x^{-\delta_s})$$

as $x \to 0$. That is, the function $|f(x)|$ can diverge to ∞ as $x \to 0$, but it can't do so faster than the polynomial $1/x^{\delta_s}$ diverges to ∞ as $x \to 0$. A similar tail constraint holds as $x \to 1$, and also for the derivatives of f up to order $m + m_0$. A similar interpretation of Θ applies when $\|\cdot\|_s$ is the weighted Sobolev L_2 norm, like the discussion of section 4.1.

The analysis now proceeds similarly as in the unbounded domain case. One important difference is that assumption 3 *cannot* hold for nontrivial weight functions on bounded domains, as the following proposition shows.

Proposition 7. There does not exist a function $\mu : (0,1) \to \mathbb{R}_+$ such that

1. $\mu(x) \to 0$ as $x \to 0$ or $x \to 1$.

2. $|\mu'(x)| \leq K\mu(x)$ for all $x \in (0,1)$.

The weaker assumption 4, however, can still hold. The following proposition verifies this for both polynomial and exponential weights.

Proposition 8. Assumption 4 holds for both $\mu_s(x) = [x(1-x)]^{\delta_s}$ and $\mu_s(x) = \exp[\delta_s x^{-1}(1-x)^{-1}]$, for any $\delta_s \in \mathbb{R}$.

The following result illustrates that assumption 5 can also hold for exponential weights. It can be generalized to $d_x > 1$, $\alpha, \beta \neq -1$, and arbitrary bounded $\mathcal{D}$.

Proposition 9. Let $\mu_c(x) = \exp[\delta_c x^{-1}(1-x)^{-1}]$, $\mu_s(x) = \exp[\delta_s x^{-1}(1-x)^{-1}]$, and $\mathcal{D} = (0,1)$. Then assumption 5 holds for any $\delta_s, \delta_c \in \mathbb{R}$ such that $\delta_c < \delta_s$.

It can be shown that such exponential weight functions also satisfy the other weight function assumptions discussed in section 4, for appropriate choices of δ_c and δ_s.

5.2 Compact embeddings and closedness results

As in the previous cases, we begin with a compact embedding result.

Theorem 7 (Compact Embedding)**.** Let $\mathcal{D} \subseteq \mathbb{R}^{d_x}$ be a bounded open set, where $d_x \geq 1$ is some integer. Let $\mu_c, \mu_s : \mathcal{D} \to \mathbb{R}_+$ be nonnegative, $m + m_0$ times continuously differentiable functions. $m, m_0 \geq 0$ are integers. Suppose assumptions 1, 2, 4, and 6 hold. Then the following embeddings are compact:

1. $\mathscr{W}_{m+m_0,2,\mu_s} \hookrightarrow \mathscr{C}_{m,\infty,\mu_c^{1/2}}$, if assumption 5 holds, $m_0 > d_x/2$, and $\mathcal{D}$ satisfies the cone condition.
2. $\mathscr{W}_{m+m_0,2,\mu_s} \hookrightarrow \mathscr{W}_{m,2,\mu_c}$, if $m_0 > d_x/2$ and $\mathcal{D}$ satisfies the cone condition.
3. $\mathscr{C}_{m+m_0,\infty,\mu_s} \hookrightarrow \mathscr{C}_{m,\infty,\mu_c}$, if $m_0 \geq 1$ and $\mathcal{D}$ is convex.
4. $\mathscr{C}_{m+m_0,\infty,\mu_s} \hookrightarrow \mathscr{W}_{m,2,\mu_c}$, if $m_0 > d_x/2$, $\mathcal{D}$ satisfies the cone condition, μ_s is bounded away from zero for any compact subset of $\mathcal{D}$, and $\int_{A^c} \mu_c(x)/\mu_s^2(x)\, dx < \infty$ for some open set $A \subseteq \mathcal{D}$ with $\overline{A} \cap \mathrm{Bd}(\overline{\mathcal{D}}) = \emptyset$.

Because of proposition 7, none of the results from Schmeisser and Triebel (1987) or the followup work by Triebel and coauthors applies to weighted norms on bounded domains. As in the unbounded domain case, however, Brown and Opic (1992) give high level conditions for a compact embedding result similar to case (1) of theorem 7, with $m_0 = 1$ and $m = 0$. Again, they do not study the other cases we consider, and they allow for a large class of weight functions which includes exponential weights. Hence, to our best knowledge, cases (2)–(4) of theorem 7 are new. The proof is similar to the proof of theorem 3, which in turn is a generalization of the proof of lemma A.4 in Gallant and Nychka (1987). We end this section with a corresponding closedness result.

Theorem 8 (Closedness)**.** Let $\mathcal{D} \subseteq \mathbb{R}^{d_x}$ be a bounded open set, where $d_x \geq 1$ is some integer. Let $m, m_0 \geq 0$ be integers. Let $(\mathscr{F}, \|\cdot\|_s)$ and $(\mathscr{G}, \|\cdot\|_c)$ be Banach spaces with $\mathscr{F} \subseteq \mathscr{G}$, where $\|f\|_s < \infty$ for all $f \in \mathscr{F}$ and $\|f\|_c < \infty$ for all $f \in \mathscr{G}$. Define Θ as in equation (1). Suppose assumptions 1, 2 and 4 hold. Then the results of table 2 hold. For cases (1) and (2) we also assume $m_0 > d_x/2$, that $\mathcal{D}$ satisfies the cone condition, and that assumption 6 holds, and in case (1) also that assumption 5 holds. For cases (3) and (4) we also assume $m_0 \geq 1$.

	$\|\cdot\|_s$	$\|\cdot\|_c$	Θ is $\|\cdot\|_c$-closed?
(1)	$\|\cdot\|_{m+m_0,2,\mu_s}$	$\|\cdot\|_{m,\infty,\mu_c^{1/2}}$	Yes
(2)	$\|\cdot\|_{m+m_0,2,\mu_s}$	$\|\cdot\|_{m,2,\mu_c}$	Yes
(3)	$\|\cdot\|_{m+m_0,\infty,\mu_s}$	$\|\cdot\|_{m,\infty,\mu_c}$	No
(4)	$\|\cdot\|_{m+m_0,\infty,\mu_s}$	$\|\cdot\|_{m,2,\mu_c}$	No

Table 4

6 Applications

In this section we illustrate how the compact embedding and closedness results discussed in this paper are applied to nonparametric estimation problems in econometrics. We discuss how the choice of norms affects the parameter space, the strength of the conclusions one obtains, and how other assumptions like moment conditions depend on this choice. In the first example we consider mean regression functions for full support regressors. We show that weighted norms can be interpreted as a generalization of trimming. In the second example, we discuss nonparametric instrumental variable estimation. In each example we focus on consistency of a sieve estimator of a function of interest, but similar considerations arise for inference. While it may be possible to prove some of these consistency results without using compact embeddings, our goal is merely to provide simple illustrations of the general ideas which are widely used in the sieve estimation literature.

We show consistency by verifying the conditions of a general consistency result stated below. Denote the data by $\{Z_i\}_{i=1}^n$ where $Z_i \in \mathbb{R}^{d_Z}$. Throughout this section we assume the data are independent and identically distributed. The parameter of interest is $\theta_0 \in \Theta$, where Θ is the parameter space. Θ may be finite or infinite dimensional. Let $Q(\theta)$ be a population objective function such that

$$\theta_0 = \operatorname*{argmax}_{\theta \in \Theta} Q(\theta).$$

Let Θ_{k_n} be a sieve space as described in the examples below. A sieve extremum estimator $\widehat{\theta}_n$ of θ_0 is defined by

$$\widehat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta_{k_n}} \widehat{Q}_n(\theta).$$

$\widehat{Q}_n$ is the sample objective function, which depends on the data. Our assumptions ensure that θ_0

and $\widehat{\theta}_n$ are well defined.[14] Let $d(\cdot,\cdot)$ be a pseudo-metric on Θ. Typically $d(\theta_1,\theta_2) = \|\theta_1 - \theta_2\|_c$ for some norm $\|\cdot\|_c$ on Θ. We now have the following result.

Proposition 10 (Consistency of sieve extremum estimators)**.** Suppose the following assumptions hold.

1. Θ and Θ_{k_n} are compact under $d(\cdot,\cdot)$.
2. $Q(\theta)$ and $\widehat{Q}_n(\theta)$ are continuous under $d(\cdot,\cdot)$ on Θ and Θ_{k_n}, respectively.
3. $Q(\theta) = Q(\theta_0)$ implies $d(\theta,\theta_0) = 0$ for all $\theta \in \Theta$. $Q(\theta_0) > -\infty$.
4. $\Theta_k \subseteq \Theta_{k+1} \subseteq \cdots \subseteq \Theta$ for all $k \geq 1$. There exists a sequence $\pi_k\theta_0 \in \Theta_k$ such that $d(\pi_k\theta_0,\theta_0) \to 0$ as $k \to \infty$.
5. $k_n \to \infty$ as $n \to \infty$ and $\sup_{\theta\in\Theta_{k_n}} |\widehat{Q}_n(\theta) - Q(\theta)| \xrightarrow{p} 0$.

Then $d(\widehat{\theta}_n,\theta_0) \xrightarrow{p} 0$ as $n \to \infty$.

Proposition 10 is a slight modification of lemma A1 in Newey and Powell (2003). The assumptions require a compact parameter space, which we can obtain by choosing a strong norm $\|\cdot\|_s$ and a consistency norm $\|\cdot\|_c$, letting $d(\theta_1,\theta_2) = \|\theta_1 - \theta_2\|_c$, and constructing the parameter space as explained in sections 3, 4, and 5. The strong norm should be chosen such the parameter space is large enough to contain θ_0. The consistency norm not only needs to be selected carefully to ensure compactness, but it will also affect the remaining assumptions, such as conditions needed for continuity of Q and $\widehat{Q}_n$ (assumption 2). Similarly, a larger parameter space usually requires stronger assumptions to ensure uniform convergence of the sample objective function (assumption 5). Assumption 3 is an identification condition, which allows $Q(\theta) = Q(\theta_0)$ for $\theta \neq \theta_0$ as long as $d(\theta,\theta_0) = 0$. Assumption 4 is a standard approximation condition on the sieve space.

6.1 Mean regression functions and trimming

Let Y and X be scalar random variables and define $g_0(x) \equiv \mathbb{E}(Y \mid X = x)$. Suppose $g_0 \in \Theta$, where Θ is the parameter space defined below. Suppose X is continuously distributed with density $f_X(x) > 0$ for all $x \in \mathbb{R}$. Hence $\text{supp}(X) = \mathbb{R}$. Notice that

$$\begin{aligned}\mathbb{E}((Y - g(X))^2) &= \mathbb{E}((Y - g_0(X))^2) + \mathbb{E}((g_0(X) - g(X))^2) \\ &\geq \mathbb{E}((Y - g_0(X))^2).\end{aligned}$$

[14] Alternatively, we can define $\widehat{\theta}_n$ as any estimator that satisfies $\widehat{Q}_n(\widehat{\theta}_n) = \sup_{\theta\in\Theta_{k_n}} \widehat{Q}_n(\theta) + o_p(1)$. Assuming $\widehat{\theta}_n$ exists, we would then not have to assume that $\widehat{Q}$ is continuous or that Θ_{k_n} is compact. We use the more restrictive definition because in our examples below these assumptions are satisfied.

The inequality is strict whenever $\mathbb{E}((g(X) - g_0(X))^2) > 0$, which holds unless $g(x) = g_0(x)$ almost everywhere. This result suggests the sieve least squares estimator

$$\widehat{g}(x) = \underset{g \in \Theta_{k_n}}{\operatorname{argmax}} \; -\frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i))^2,$$

where Θ_{k_n} is a sieve space for Θ. For example, let $p_j : \mathbb{R} \to \mathbb{R}$ be a sequence of basis functions for Θ. Then we could choose the linear sieve space

$$\Theta_{k_n} = \left\{ g \in \Theta : g(x) = \sum_{j=1}^{k_n} b_j p_j(x) \text{ for some } b_1, \ldots, b_{k_n} \in \mathbb{R} \right\}.$$

Let $\|\cdot\|_c$ denote the consistency norm and let $\|\cdot\|_s$ be a strong norm. The parameter space Θ is a $\|\cdot\|_s$-ball as explained in sections 3, 4, and 5. Intuitively, the unweighted L_2 or sup-norms on $\mathbb{R}$ are too strong to be a consistency norm because the data provides no information about $g_0(x)$ for x larger than the largest observation. In fact, to apply any of the compactness results with such a choice of $\|\cdot\|_c$, we would have to use a strong norm with upweighting. By proposition 2, this implies that we would have to assume that $g(x) \to 0$ as $|x| \to \infty$. Since this assumption would rule out the linear regression model, we instead use the downweighted sup-norm

$$\|g\|_c = \|g\|_{0,\infty,\mu_c} = \sup_{x \in \mathbb{R}} |g(x)| \mu_c(x),$$

where $\mu_c(x)$ is nonnegative and $\mu_c(x) \to 0$ as $|x| \to \infty$. As a parameter space we can then either use a weighted Hölder space (by theorems 5 and 6) or a weighted Sobolev space (by theorems 3 and 4). As an example, we choose a weighted Sobolev L_2 parameter space, and give low level conditions under which $\|\widehat{g} - g_0\|_c \xrightarrow{p} 0$ in the following proposition.

Proposition 11 (Consistency of sieve least squares)**.** Suppose the following assumptions hold.

1. Let $\|\cdot\|_c = \|\cdot\|_{0,\infty,\mu_c}$, $\|\cdot\|_s = \|\cdot\|_{1,2,\mu_s}$, and

   $$\Theta = \{ g \in \mathscr{W}_{1,2,\mu_s} : \|g\|_{1,2,\mu_s} \leq B \}.$$

   The weight functions $\mu_c, \mu_s : \mathbb{R} \to \mathbb{R}_+$ are nonnegative and continuously differentiable. μ_c^2 and μ_s satisfy assumptions 1, 2, 4, 5, and 6. μ_c and μ_s satisfy assumption 1. g_0 is continuous.

2. $\mathbb{E}(\mu_c(X)^{-2}) < \infty$ and $\mathbb{E}(Y^2) < \infty$.

3. Θ_k is $\|\cdot\|_c$-closed for all k. $\Theta_k \subseteq \Theta_{k+1} \subseteq \cdots \subseteq \Theta$ for all $k \geq 1$. For any $M > 0$, there exists $g_k \in \Theta_k$ such that $\sup_{x:|x| \leq M} |g_k(x) - g_0(x)| \to 0$ as $k \to \infty$.

4. $k_n \to \infty$ as $n \to \infty$.

Then $\|\widehat{g} - g_0\|_c \xrightarrow{p} 0$ as $n \to \infty$.

The proof of proposition 11 follows by verifying the conditions of proposition 10. Assumption 1 of proposition 11 ensures that Θ is $\|\cdot\|_c$-compact by our compact embedding and closedness results, parts 1 of theorems 3 and 4. Together with the moment assumptions of proposition 11, this compactness provides a simple and primitive sufficient condition for the uniform convergence assumption 5 of proposition 10.

As mentioned earlier, we must use downweighting—$\mu_s(x) \to 0$ as $|x| \to \infty$—in the strong norm to allow g_0 to be linear. The faster μ_s converges to 0, the larger is the parameter space. However, allowing for a larger parameter space has several consequences. First, by our assumptions on the relationship between μ_s and μ_c, faster convergence of μ_s to zero implies faster convergence of μ_c to zero. This weakens the consistency norm. Consequently, both continuity and uniform convergence are harder to verify. In proposition 11 we ensure these two assumptions hold by requiring $\mathbb{E}(\mu_c(X)^{-2}) < \infty$. But here we see that the faster μ_c converges to 0, the more moments of X we assume exist. For example, suppose $\mu_s(x) = (1+x^2)^{-\delta_s}$ and $\mu_c(x) = (1+x^2)^{-\delta_c}$ with $\delta_s > 0$. The conditions on the weight functions require that $\delta_s < 2\delta_c$ and the moment condition is $\mathbb{E}\left((1+X^2)^{2\delta_c}\right) < \infty$. Thus larger δ_s's yield larger parameter spaces, but imply δ_c must also be larger, and hence we need more moments of X. Next suppose $\mu_s(x) = \exp(-\delta_s x^2)$ and $\mu_c(x) = \exp(-\delta_c x^2)$ with $0 < \delta_s < 2\delta_c$. Then the moment condition is $\mathbb{E}[\exp(\delta_c X^2)] < \infty$. This is equivalent to requiring that the tails of X are sub-Gaussian, $\mathbb{P}(|X| > t) \leq C\exp(-ct^2)$ for constants C and c, which in turn implies that all moments of X are finite.

The only remaining assumption is the condition on the sieve spaces. There are many choices of sieve spaces which satisfy this last condition because it only requires that g_0 can be approximated on any compact subset of $\mathbb{R}$. See Chen (2007) for examples.

Weakening the assumptions and generalized trimming

The assumption $\mathbb{E}(\mu_c(X)^{-2}) < \infty$ in proposition 11 rules out indicator weight functions, like $\mu_c(x) = \mathbb{1}(|x| \leq M)$. The need for this moment condition arises because while we weigh down large values of X in the consistency norm, we do not weigh them explicitly in the objective function. Assuming the existence of moments imposes the weight implicitly. It ensures that outliers of the regressor, which can affect the estimator in regions where $\mu_c(x)$ is large, occur with small probability. This discussion suggests that using a weighted objective function may lead to weaker assumptions. That is, let

$$\widehat{g}_w(x) = \underset{g \in \Theta_{k_n}}{\operatorname{argmax}} \; -\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))^2 \mu_c(X_i)^2.$$

Indeed, we obtain the following proposition.

Proposition 12 (Consistency of sieve least squares)**.** Suppose the following assumptions hold.

1. Let $\|\cdot\|_c = \|\cdot\|_{0,\infty,\mu_c}$, $\|\cdot\|_s = \|\cdot\|_{1,2,\mu_s}$, and

$$\Theta = \{g \in \mathscr{W}_{1,2,\mu_s} : \|g\|_{1,2,\mu_s} \leq B\}.$$

The weight functions $\mu_c, \mu_s : \mathbb{R} \to \mathbb{R}_+$ are nonnegative and continuously differentiable. μ_c^2 and μ_s satisfy assumptions 1, 2, 4, 5, and 6. μ_c and μ_s satisfy assumptions 1 and 2. $\mu_c(x) > 0$ implies $\mathbb{P}(\mu_c(X) > 0 \mid |X - x| \leq \varepsilon) > 0$ for any $\varepsilon > 0$. g_0 is continuous.

2. $\mathbb{E}(Y^2) < \infty$, $\mathbb{E}(Y^2 \mu_c(X)^2) < \infty$, and $\mathbb{E}((Y - g_0(X))^2) < \infty$.

3. Θ_k is $\|\cdot\|_c$-closed for all k. $\Theta_k \subseteq \Theta_{k+1} \subseteq \cdots \subseteq \Theta$ for all $k \geq 1$. For any $M > 0$, there exists $g_k \in \Theta_k$ such that $\sup_{x:|x|\leq M} |g_k(x) - g_0(x)| \to 0$ as $k \to \infty$.

4. $k_n \to \infty$ as $n \to \infty$.

Then $\|\widehat{g}_w - g_0\|_c \xrightarrow{p} 0$ as $n \to \infty$.

We can interpret this proposition as a generalized version of trimming, where by trimming we mean using the weight function $\mu_c(x) = \mathbb{1}(|x| \leq M)$ for a fixed constant M. With this weight function we only obtain convergence of $\widehat{g}_w(x)$ to $g_0(x)$ uniformly over x in the compact subset $[-M, M]$ of the support of the regressor. Even with this weight function, however, if we omit the weight from the objective function as in proposition 11, then outliers of X affect $\widehat{g}(x)$ even for $x \in [-M, M]$. Trimming simply discards the outliers. The more general result proposition 12 simply gives these observations less weight. The advantage of using a weight function such as $\mu_c(x) = (1 + x^2)^{-\delta_c}$ rather than the trimming weight $\mu_c(x) = \mathbb{1}(|x| \leq M)$ is that it implies uniform convergence over *any* compact subset of $\mathbb{R}$.

Finally, in related prior work, Chen and Christensen (2015b) derive sup-norm consistency rates for a sieve least squares estimator when the regressors have full support by using a sequence of trimming functions. Also see Andrews (1995), Newey (1997), Hansen (2014), and Belloni, Chernozhukov, Chetverikov, and Kato (2015). Unlike these papers, we prove consistency in a general weighted sup-norm with regressors on $\mathbb{R}$.

6.2 Nonparametric instrumental variables estimation

In this section we apply our results to the nonparametric instrumental variable model

$$Y = g_0(X) + U, \qquad \mathbb{E}(U \mid Z) = 0,$$

where Y, X, and Z are continuously distributed scalar random variables and $f_X(x) > 0$ for all $x \in \mathbb{R}$. Assume $g_0 \in \Theta$, where Θ is the parameter space defined below. Since $\mathbb{E}\left(\mathbb{E}(Y - g_0(X) \mid Z)^2\right) = 0$, Newey and Powell (2003) suggest estimating g_0 in two steps. First, for any $g \in \Theta$ estimate $\rho(z, g) \equiv \mathbb{E}(Y - g(X) \mid Z = z)$ using a series estimator. Call this estimator $\widehat{\rho}(z, g)$. Then let

$$\widehat{g}(x) = \underset{g \in \Theta_{k_n}}{\operatorname{argmax}} \; -\frac{1}{n} \sum_{i=1}^{n} \widehat{\rho}(Z_i, g)^2.$$

where Θ_{k_n} is a sieve space for function in Θ, as before. See Newey and Powell (2003) for more estimation details.

Define

$$\widetilde{\Theta} = \{g \in \mathscr{W}_{m+m_0,2,\mu_s} : \|g\|_{m+m_0,2,\mu_s} \leq \tilde{B}\},$$

where $\mu_s(x) = (1+x^2)^{\delta_s}$, $\delta_s > 0$, and $m, m_0 \geq 0$. Let $a(x) \in \mathbb{R}^{d_a}$ be a vector of known functions of x. Newey and Powell (2003) define the parameter space by

$$\Theta_{\text{NP}} = \{a(\cdot)'\beta + g_1(\cdot) : \beta'\beta \leq B_\beta,\, g_1 \in \widetilde{\Theta}\}.$$

Proposition 2 implies that for any $g_1 \in \widetilde{\Theta}$, it holds that $|g_1(x)| \to 0$ as $|x| \to \infty$. The term $a(x)'\beta$ ensures that the tails of g_0 are not required to converge to 0, but it requires the tails of g_0 to be modeled parametrically. As a consistency norm Newey and Powell (2003) use $\|\cdot\|_{m,\infty,\mu_c}$, where μ_c upweights the tails of the functions as well. Also see Santos (2012) for a similar parameter space.

In this section we modify the arguments of Newey and Powell (2003) to allow for nonparametric tails of the function g_0. In particular, we let $\mu_s(x) \to 0$ as $|x| \to \infty$. Consequently we allow for a larger parameter space. The main cost of allowing for a larger parameter space is that we obtain consistency in a weaker norm.

The population objective function is

$$Q(g) = -\mathbb{E}(\mathbb{E}(Y - g(X) \mid Z)^2).$$

The generalization of trimming used in the previous section is generally not possible here because although $\mathbb{E}(Y - g_0(X) \mid Z = z) = 0$ for all z, usually $\mathbb{E}((Y - g_0(X))\mu_c(X) \mid Z = z) \neq 0$ for some z. Instead we follow the approach of proposition 11.

The following proposition provides low level conditions under which $\|\widehat{g} - g_0\|_c \xrightarrow{p} 0$. As in the previous subsection, $\|\cdot\|_c$ is a weighted sup-norm and the parameter space is a weighted Sobolev L_2 space.[15] The arguments can easily be adapted to allow for higher order derivatives in the consistency norm or a weighted Hölder space as the parameter space.

Proposition 13 (Consistency of sieve NPIV estimator)**.** Suppose the following assumptions hold.

1. For all $g \in \Theta$, $\mathbb{E}(Y - g(X) \mid Z = z) = 0$ for almost all z implies $g(x) = g_0(x)$ for almost all x.
2. Let $\|\cdot\|_c = \|\cdot\|_{0,\infty,\mu_c}$, $\|\cdot\|_s = \|\cdot\|_{1,2,\mu_s}$, and

   $$\Theta = \{g \in \mathscr{W}_{1,2,\mu_s} : \|g\|_{1,2,\mu_s} \leq B\}.$$

   The weight functions $\mu_c, \mu_s : \mathbb{R} \to \mathbb{R}_+$ are nonnegative and continuously differentiable. μ_c^2 and μ_s satisfy assumptions 1, 2, 4, 5, and 6. μ_c and μ_s satisfy assumptions 1 and 2. g_0 is continuous.
3. $\mathbb{E}(Y^2) < \infty$, $\mathbb{E}(\mu_c(X)^{-2}) < \infty$, and $\mathbb{E}\left((\text{var}(Y - g(X) \mid Z))^2\right) < \infty$ for all $g \in \Theta$.
4. For any $b(z)$ with $\mathbb{E}[b(Z)^2] < \infty$ there is $g_k \in \Theta_k$ with $\mathbb{E}[(b(Z) - g_k(Z))^2] \to 0$ as $k \to \infty$.

[15] Chen and Christensen (2015a) derive the rate of convergence in the sup-norm when X has compact support.

5. Θ_k is $\|\cdot\|_c$-closed for all k. $\Theta_k \subseteq \Theta_{k+1} \subseteq \cdots \subseteq \Theta$ for all $k \geq 1$. For any $M > 0$, there exists $g_k \in \Theta_k$ such that $\sup_{x:|x|\leq M} |g_k(x) - g_0(x)| \to 0$ as $k \to \infty$.

6. $k_n \to \infty$ as $n \to \infty$ such that $k_n/n \to 0$.

Then $\|\widehat{g} - g_0\|_c \xrightarrow{p} 0$.

Assumption 1 is the identification condition known as completeness. Besides this assumption and compared to the regression model in proposition 11, the additional assumptions are assumption 4 and the last part of assumption 3. These two conditions ensure that the first stage regression is sufficiently accurate and they are implied by assumption 3 of Newey and Powell (2003). We use the same sieve space to approximate $g_0(x)$ and $b(z)$, but the arguments can easily be generalized at the expense of additional notation. The last part of assumption 3 holds for example if either $\mathbb{E}(Y^4) < \infty$ and $\mathbb{E}(\mu_c(X)^{-4}) < \infty$ or $\text{var}(Y - g(X) \mid Z) \leq M$ for some $M > 0$ and all $g \in \Theta$.

Chen and Pouzo (2012) discuss convergence in a weighted sup-norm of a penalized estimator in the NPIV model as an example of their general consistency theorem. Chen and Christensen (2015a) derive many new and important results for the NPIV model. Among others, they derive minimax optimal sup-norm convergence rates and they describe an estimator which achieves those rates. Their results apply when X and Z have compact support.

Rescaling the regressors

An alternative to proving consistency using the previous proposition is to first transform X to the interval $[0, 1]$ and then apply consistency results for functions on compact support. For example, let $W = \Phi(X)$ where Φ denotes the standard normal cdf, and let $h_0(w) = g_0(\Phi^{-1}(w))$. Then

$$Y = h_0(W) + U, \qquad \mathbb{E}(U \mid Z) = 0$$

and knowledge of h_0 implies knowledge of g_0. Estimating h_0 might appear to be simpler because W has support on $[0, 1]$. However, notice that h_0 is unbounded if X has support on $\mathbb{R}$ and if g_0 is unbounded on $\mathbb{R}$. Thus, for example, to allow g_0 to be linear we have to use weighted norms. Specifically, notice that using the change of variables $w = \Phi(x)$ the unweighted Sobolev L_2 norm of h_0 with $m = 1$ is

$$\|h_0\|_{1,2} = \int_0^1 \left(h_0(w)^2 + h_0'(w)^2\right)\, dw = \int_{-\infty}^{\infty} \left(g_0(x)^2 + g_0'(x)^2\phi(x)^{-2}\right) \phi(x)\, dx,$$

where ϕ denotes the standard normal cdf. Therefore, $\|h_0\|_{1,2}$ is unbounded unless $|g_0(x)| \to 0$ as $|x| \to \infty$. Similarly, h_0 is generally not Hölder continuous. Hence any parameter space assumptions on h_0 must be imposed using weighted norms, such as those as discussed in section 5. Moreover, notice that

$$\sup_{w\in[0,1]} |h_0(w)| = \sup_{x\in\mathbb{R}} |g_0(x)|$$

and as argued in the previous subsection, the unweighted sup-norm on $\mathbb{R}$ is too strong to be a consistency norm unless we know that $|g_0'(x)| \to 0$ as $|x| \to \infty$. Finally, it holds that

$$\|h_0\|_{0,2} = \int_0^1 h_0(w)^2 \, dw = \int_{-\infty}^{\infty} g_0(x)^2 \phi(x) \, dx = \|g_0\|_{0,2,\phi}$$

Therefore convergence of an estimator of h_0 in the unweighted L_2 norm on $[0,1]$ is equivalent to convergence of the corresponding estimator of g_0 in a weighted L_2 norm on $\mathbb{R}$.

7 Conclusion

In this paper we have gathered and reviewed many previously known compact embedding results for convenient reference. Furthermore, we have proved several new compact embedding results which generalize the existing results and were not previously known. Unlike most previous results, our results allow for exponential weight functions. Our new results also allow for weighted norms on bounded domains, of which only one prior result existed, even for polynomial weights. We additionally gave closedness results, some of which were known and some of which are apparently new to the econometrics literature. Finally, we discussed the practical relevance of these results. We explained how the choice of norm and weight function affect the functions allowed in the parameter space. We also showed how to apply these results in two examples: nonparametric mean regression and nonparametric instrumental variables estimation.

After showing consistency of an estimator, the next step is to consider rates of convergence and inference. For these results, it is often helpful to have results on entropy numbers for the function space of interest. For functions with bounded domain satisfying standard norm bounds, many well known results exist. For example, van der Vaart and Wellner (1996) theorem 2.7.1 gives covering number rates for Hölder balls with the sup-norm as the consistency norm. Such results are refinements of compact embedding results, since totally bounded parameter spaces are relatively compact. For functions with full support, fewer entropy number results exist. For example, lemma A.3 of Santos (2012) generalizes van der Vaart and Wellner (1996) theorem 2.7.1 to the case where Θ is a polynomial-upweighted Sobolev L_2 ball and $\|\cdot\|_c$ is the Sobolev sup-norm. Note that a compact embedding result is used as the first step in his proof. Haroske and Triebel (1994a,b) and Haroske (1995) also provide similar results for a large class of weighted spaces, again restricting to a class of weight functions satisfying assumption 3 and which have at most polynomial growth. Since our results allow for more general weight functions, it would be useful to know whether these entropy number results generalize as well.

Finally, applying a result on sieve approximation rates is one step when deriving convergence rates of sieve estimators. For example, see theorem 3.2 of Chen (2007) and the subsequent discussion. Many approximation results for functions on the real line, such as those discussed in Mhaskar (1986), are for exponentially weighted sup-norms. Therefore, our extension of the compact embedding results to exponential weights should be useful when combined with these approximation

results to derive sieve estimator convergence rates.

References

Adams, R. A. and J. J. Fournier (2003): *Sobolev spaces*, vol. 140, Academic press, 2nd ed.

Ai, C. and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71, 1795–1843.

Andrews, D. W. (1995): "Nonparametric kernel estimation for semiparametric models," *Econometric Theory*, 11, 560–586.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "Some new asymptotic theory for least squares series: Pointwise and uniform results," *Journal of Econometrics*, 186, 345–366.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-nonparametric IV estimation of shape-invariant Engel curves," *Econometrica*, 75, 1613–1669.

Brendstrup, B. and H. J. Paarsch (2006): "Identification and estimation in sequential, asymmetric, english auctions," *Journal of Econometrics*, 134, 69–94.

Brown, R. and B. Opic (1992): "Embeddings of weighted Sobolev spaces into spaces of continuous functions," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 439, 279–296.

Carroll, R. J., X. Chen, and Y. Hu (2010): "Identification and estimation of nonlinear models using two samples with nonclassical measurement errors," *Journal of Nonparametric Statistics*, 22, 379–399.

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

Chen, X. and T. M. Christensen (2015a): "Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation," *Working paper*.

——— (2015b): "Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions," *Journal of Econometrics*.

Chen, X., Y. Fan, and V. Tsyrennikov (2006): "Efficient estimation of semiparametric multivariate copula models," *Journal of the American Statistical Association*, 101, 1228–1240.

Chen, X., L. P. Hansen, and J. Scheinkman (2009a): "Nonlinear principal components and long-run implications of multivariate diffusions," *The Annals of Statistics*, 4279–4312.

Chen, X., H. Hong, and E. Tamer (2005): "Measurement error models with auxiliary data," *Review of Economic Studies*, 72, 343–366.

Chen, X., Y. Hu, and A. Lewbel (2009b): "Nonparametric identification and estimation of nonclassical errors-in-variables models without additional information," *Statistica Sinica*, 19, 949–968.

Chen, X. and D. Pouzo (2012): "Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals," *Econometrica*, 80, 277–321.

——— (2015): "Sieve Wald and QLR inferences on semi/nonparametric conditional moment models," *Econometrica*, 83, 1013–1079.

Chen, X. and X. Shen (1998): "Sieve extremum estimates for weakly dependent data," *Econometrica*, 289–314.

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007): "Instrumental variable estimation of nonseparable models," *Journal of Econometrics*, 139, 4–14.

Chiappori, P.-A., I. Komunjer, and D. Kristensen (2015): "Nonparametric identification and estimation of transformation models," *Journal of Econometrics*, 188, 22–39.

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

Csörgö, M. (1983): *Quantile processes with statistical applications*, SIAM.

Edmunds, D. E. and H. Triebel (1996): *Function spaces, entropy numbers, differential operators*, Cambridge University Press.

Elbadawi, I., A. R. Gallant, and G. Souza (1983): "An elasticity can be estimated consistently without a priori knowledge of functional form," *Econometrica*, 1731–1751.

Fenton, V. M. and A. R. Gallant (1996): "Qualitative and asymptotic performance of SNP density estimators," *Journal of Econometrics*, 74, 77–118.

Fox, J. T. and A. Gandhi (2015): "Nonparametric identification and estimation of random coefficients in multinomial choice models," *RAND Journal of Economics, Forthcoming*.

Fox, J. T., K. I. Kim, and C. Yang (2015): "A simple nonparametric approach to estimating the distribution of random coefficients in structural models," *Working paper*.

Freyberger, J. (2012): "Nonparametric panel data models with interactive fixed effects," *Working paper*.

Gallant, A. and D. Nychka (1987): "Semi-nonparametric maximum likelihood estimation," *Econometrica*, 55, 363–390.

Gallant, A. R. and G. Tauchen (1989): "Seminonparametric estimation of conditionally constrained heterogeneous processes: Asset pricing applications," *Econometrica*, 1091–1120.

Hansen, B. E. (2014): "A unified asymptotic distribution theory for parametric and nonparametric least squares," *Working paper*.

Haroske, D. (1995): "Approximation numbers in some weighted function spaces," *Journal of Approximation Theory*, 83, 104–136.

Haroske, D. and H. Triebel (1994a): "Entropy numbers in weighted function spaces and eigenvalue distributions of some degenerate pseudodifferential operators I," *Mathematische Nachrichten*, 167, 131–156.

——— (1994b): "Entropy numbers in weighted function spaces and eigenvalue distributions of some degenerate pseudodifferential operators II," *Mathematische Nachrichten*, 168, 109–137.

Heckman, J. and B. Singer (1984): "A method for minimizing the impact of distributional assumptions in econometric models for duration data," *Econometrica*, 271–320.

Horowitz, J. L. (1996): "Semiparametric estimation of a regression model with an unknown transformation of the dependent variable," *Econometrica*, 103–137.

Hu, Y. and S. M. Schennach (2008): "Instrumental variable treatment of nonclassical measurement error models," *Econometrica*, 76, 195–216.

Khan, S. (2013): "Distribution free estimation of heteroskedastic binary response models using Probit/Logit criterion functions," *Journal of Econometrics*, 172, 168–182.

Kiefer, J. and J. Wolfowitz (1956): "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters," *The Annals of Mathematical Statistics*, 887–906.

Kufner, A. (1980): *Weighted Sobolev spaces*, BSB B. G. Teubner Verlagsgesellschaft.

Kufner, A. and B. Opic (1984): "How to define reasonably weighted Sobolev spaces," *Commentationes Mathematicae Universitatis Carolinae*, 25, 537–554.

Matzkin, R. L. (1992): "Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models," *Econometrica*, 239–270.

Mhaskar, H. N. (1986): "Weighted polynomial approximation," *Journal of Approximation Theory*, 46, 100–110.

Newey, W. K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147–168.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

Newey, W. K. and J. L. Powell (2003): "Instrumental variable estimation of nonparametric models," *Econometrica*, 71, 1565–1578.

Rodríguez, J. M., V. Álvarez, E. Romera, and D. Pestana (2004): "Generalized weighted Sobolev spaces and applications to Sobolev orthogonal polynomials I," *Acta Applicandae Mathematicae*, 80, 273–308.

Santos, A. (2012): "Inference in nonparametric instrumental variables with partial identification," *Econometrica*, 80, 213–275.

Schmeisser, H.-J. and H. Triebel (1987): *Topics in Fourier analysis and function spaces*, John Wiley & Sons.

van der Vaart, A. W. (2000): *Asymptotic statistics*, Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak convergence and empirical processes*, Springer.

Wald, A. (1949): "Note on the consistency of the maximum likelihood estimate," *The Annals of Mathematical Statistics*, 595–601.

Wong, W. H. and T. A. Severini (1991): "On maximum likelihood estimation in infinite dimensional parameter spaces," *The Annals of Statistics*, 603–632.