Freyberger, Joachim; Masten, Matthew

Working Paper

Compactness of infinite dimensional parameter spaces

cemmap working paper, No. CWP01/16

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Freyberger, Joachim; Masten, Matthew (2015) : Compactness of infinite dimensional parameter spaces, cemmap working paper, No. CWP01/16, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2016.0116

This Version is available at:
https://hdl.handle.net/10419/130087

cemmap
centre for microdata methods and practice

Compactness of infinite dimensional parameter spaces

Joachim Freyberger
Matthew Masten

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP01/16

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Compactness of Infinite Dimensional Parameter Spaces*

Joachim Freyberger† Matthew A. Masten‡

December 23, 2015

Abstract

We provide general compactness results for many commonly used parameter spaces in nonparametric estimation. 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 apply these results to 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 and the 2015 Triangle Econometrics Conference. We thank audiences at those seminars as well as Bruce Hansen, 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. The solution to this problem is to use *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, which says that $\|\cdot\|$-balls are compact under the weak* topology (but not under $\|\cdot\|$, otherwise the space would be finite dimensional).

In econometrics, this idea has been used by Gallant and Nychka (1987) and subsequent authors in the sieve estimation literature. There we 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 gather all of these compactness results together, along with several new ones. We organize our results into three 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 all of these cases, 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. Consequently, if one nevertheless wants to use these choices, then one should allow for parameters in the closure.

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 applications. First, we consider estimation of mean regression functions with full support regressors. We give low level conditions for consistency of both a sieve least squares and a penalized sieve least squares estimator, and discuss how the choice of norm is used in these results. 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.

We conclude this section with a brief review of the 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 16 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 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 review 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), 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.

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″.

The rest of this paper is organized as follows. In section 2 we review the definitions of the

[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.

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. Definitions, statements of lemmas, and some proofs are in the appendix. All other results and proofs are given in a supplemental appendix.

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 $\|f\|_{m,\infty}$ denote the unweighted Sobolev sup-norm; that is, the weighted Sobolev sup-norm with the identity weight $\mu(x) \equiv 1$. For 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.

The next space we consider replaces the sup-norm with an L_p norm. Let p satisfy $1 \leq p < \infty$.

[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.

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

For an m-times differentiable function $f : \mathcal{D} \to \mathbb{R}$, 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 $\|f\|_{m,p}$ 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

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

$\|\cdot\|_{0,p,\mu}$ is called the *weighted L_p norm*. Let $\mathscr{L}_{p,\mu}(\mathcal{D})$ denote the space of functions $f : \mathcal{D} \to \mathbb{R}$ with $\|f\|_{0,p,\mu} < \infty$.

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. The space

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

equipped with the norm $\|\cdot\|_{m,p,\mu}$ is *not* $\|\cdot\|_{m,p,\mu}$-complete. For unweighted spaces, $\mu(x) \equiv 1$, we instead consider the *completion* of this space, denoted by $\mathscr{H}_{m,p,1}(\mathcal{D})$. An important result from functional analysis known as the 'H=W theorem' states that this completion equals the *Sobolev space* $\mathscr{W}_{m,p,1}(\mathcal{D})$, which is the set of all $\mathscr{L}_{p,1}(\mathcal{D})$ functions f which have weak derivatives and whose weak partial derivatives $\nabla^\lambda f$ are in $\mathscr{L}_{p,1}(\mathcal{D})$ for all $0 \leq |\lambda| \leq m$.[4] For weighted spaces, the H=W theorem does not necessarily hold; see Zhikov (1998).[5] For this reason, we follow the literature by defining the weighted Sobolev space $\mathscr{W}_{m,p,\mu}$ as the set of all $\mathscr{L}_{p,\mu}(\mathcal{D})$ functions f which have weak derivatives and whose weak partial derivatives $\nabla^\lambda f$ are in $\mathscr{L}_{p,\mu}(\mathcal{D})$ for all $0 \leq |\lambda| \leq m$. 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.

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.[6] A function

[4] See theorem 3.17 in Adams and Fournier (2003).

[5] Similar results sometimes obtain, however. For example, see Kufner and Opic (1984) remark 4.8 and also the discussion in Zhikov (1998). Also see remark 4.1 of Kufner and Opic (1984).

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

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 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 $m_0 \geq 0$.

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 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 $\mathcal{D}$. It is formally defined in appendix A. 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. 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 H. 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.

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 subtly is that we actually 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 probability density functions, we typically will choose upweighting as in Gallant and Nychka (1987). For spaces of regression functions, we typically will choose downweighting.[7]

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.

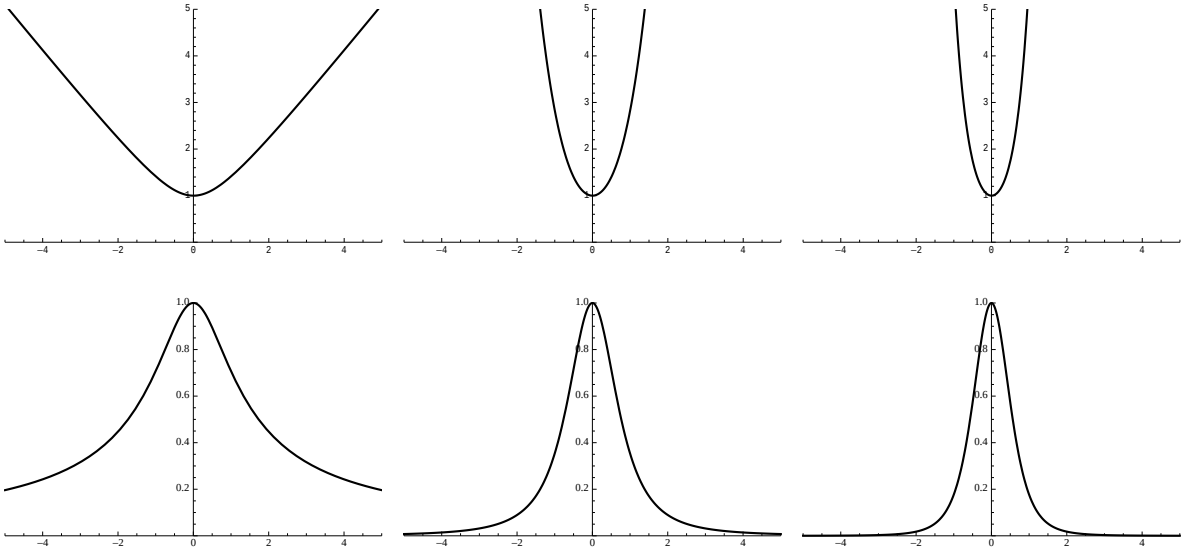

Figure 1: Polynomial weighting functions $\mu(x) = (1+x^2)^{\delta}$. Top: Upweighting, with $\delta = 0.5, 1.5, 2.5$ from left to right. Bottom: Downweighting, with $\delta = -0.5, -1.5, -2.5$ from left to right.

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

[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.

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.

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 G 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 $m_0 \geq 1$.
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

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

[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.

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.

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 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.

[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.

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 $\operatorname{supp}(Y) = [y_L, y_U]$*. 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 $\delta = \alpha = \beta = -1$,

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

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.

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

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

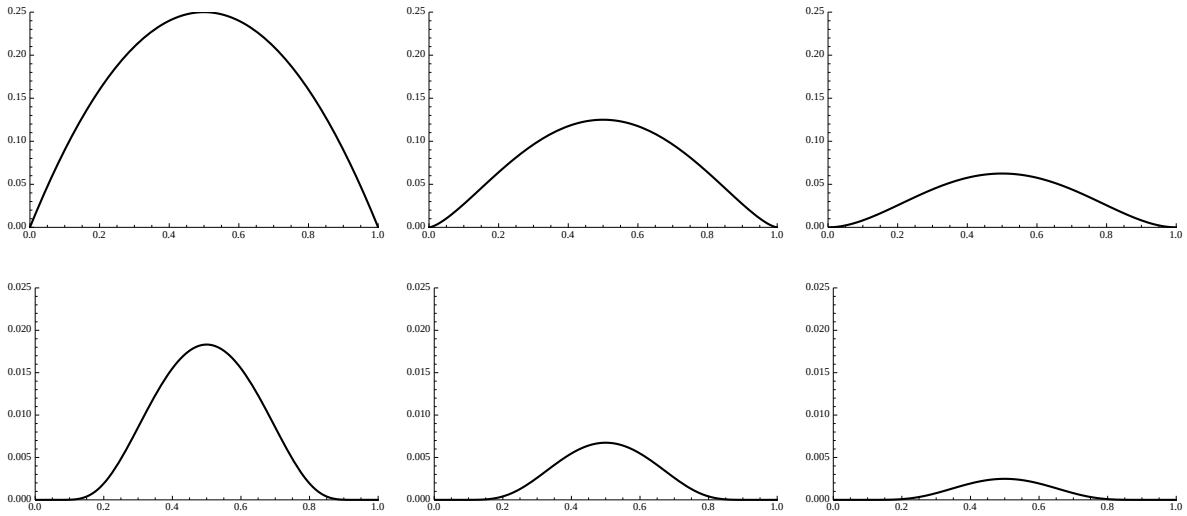

Figure 2: Top: Polynomial weighting functions $\mu(x) = [x(1-x)]^{\delta}$ for $\delta = 1, 1.5, 2$, from left to right. Bottom: Exponential weighting functions $\mu(x) = \exp[\delta x^{-1}(1-x)^{-1}]$ with $\delta = -1, -1.25, -1.5$, from left to right.

the 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 or alternative estimators.

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.

[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.

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 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

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

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

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

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 $n \to \infty$.

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. They also discuss the possibility of using polynomial or exponential weights, but do not derive any results for these weight functions. Also, their results apply to iid and non-iid data and they develop inference results for functionals of the mean regression function.

Penalized sieve least squares

An alternative to assuming a compact parameter space as in proposition 12 is to add a penalty term to the objective function. That is, suppose $g_0 \in \mathscr{W}_{1,2,\mu_s}$, but we do not want to impose an a priori known bound on $\|g_0\|_s = \|g_0\|_{1,2,\mu_s}$. Instead, let

$$\widetilde{\Theta}_{k_n} = \left\{ g \in \mathscr{W}_{1,2,\mu_s} : g(x) = \sum_{j=1}^{k_n} b_j p_j(x) \text{ for some } b_1, \ldots, b_{k_n} \in \mathbb{R} \text{ and } \|g\|_s \leq B_n \right\}$$

for some sequence of constants $B_n \to \infty$. Define the penalized sieve least squares estimator

$$\tilde{g}_w(x) = \underset{g \in \widetilde{\Theta}_{k_n}}{\operatorname{argmax}} \; - \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i))^2 \mu_c(X_i)^2 + \lambda_n \|g\|_s \right).$$

λ_n is a penalty parameter that converges to zero as the sample size grows. The following proposition uses arguments from Chen and Pouzo (2012) to show that $\tilde{g}_w$ is consistent for g_0.

Proposition 13 (Consistency of penalized 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

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

   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 $\varepsilon > 0$. $\sup_{x \in \mathbb{R}} \mu_c(x) < \infty$. g_0 is continuous.

2. $\mathbb{E}(Y^2) < \infty$, $\mathbb{E}(Y^2 \mu_c(X)^2) < \infty$, and $\mathbb{E}((Y - g_0(X))^4) < \infty$.

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 $n \to \infty$.

5. $\lambda_n > 0$, $\lambda_n = o(1)$ and $\max\{1/\sqrt{n}, \|g_{k_n} - g_0\|_c\} = O(\lambda_n)$.

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

Proposition 13 allows for a noncompact parameter space. The additional assumption needed is assumption 5, which imposes an upper bound on the rate of convergence of λ_n. Assumption 3 implies that $\|g_{k_n} - g_0\|_c$ converges to 0 and assumption 5 then imposes that λ_n cannot converge at a faster rate.

In propositions 11 and 12 we used the compact embedding and closedness results of sections 3, 4, and 5 directly to pick norms such that the compact parameter space assumption holds. In proposition 13 this is no longer an issue because we do not need a compact parameter space. However, the results of sections 3, 4, and 5 are still used in the proof, and hence the choice of norm here is still important, as discussed in section 3.2.1 of Chen and Pouzo (2012). Essentially, our proof of proposition 13 first uses lemma A.3 in Chen and Pouzo (2012) to show that for some finite $M_0 > 0$

$$\tilde{g}_w \in \{g \in \mathscr{W}_{1,2,\mu_s} : \|g\|_{1,2,\mu_s} \leq M_0\}$$

with probability arbitrarily close to 1 for all large n. We then use the arguments from the proof of proposition 12 to prove that $\|\tilde{g}_w - g_0\|_c \xrightarrow{p} 0$. It's at this step where the compact embedding and closedness results help.

An alternative proof can be obtained by showing that our low level sufficient conditions imply the assumptions of theorem 3.2 in Chen and Pouzo (2012), which is a general consistency theorem, applies when X has compact support, and allows for both nonsmooth residuals and a noncompact parameter space. One of the assumptions of theorem 3.2 is that the penalty function is lower semicompact, which here means that $\|\cdot\|_s$-balls are $\|\cdot\|_c$-compact. This is precisely the kind of result we have discussed throughout this paper.

Finally, we note that while both of these approaches—assuming a compact parameter space, or using a penalty function—lead to easy-to-interpret sufficient conditions, one could also use theorem 3.1 in Chen (2007), which may avoid both compactness and penalty functions.

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 before 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 14 (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.

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

3. $\mathbb{E}(Y^2) < \infty$, $\mathbb{E}(\mu_c(X)^{-2}) < \infty$, and $\mathbb{E}\left((\text{var}(Y_i - g(X_i) \mid Z_i))^2\right) < \infty$ 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 $k \to \infty$.

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 $g \in \Theta$.

We can use a penalty function instead of compact parameter space under some additional assumptions very similar to those in proposition 13. 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 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 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.

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.

Folland, G. B. (1999): *Real analysis: Modern techniques and their applications*, John Wiley & Sons, Inc., 2nd ed.

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*.

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.

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. 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.

Zhikov, V. V. (1998): "Weighted Sobolev spaces," *Sbornik: Mathematics*, 189, 1139–1170.

A Some formal definitions and useful lemmas

In this appendix we first state some formal definitions. These are primarily used as background for the various compact embedding results. We then give some brief lemmas we use elsewhere. Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed vector spaces. Then we use the following definitions.

- $A \subseteq X$ is $\|\cdot\|_X$-*bounded* if there is a scalar $R > 0$ such that $\|x\|_X \leq R$ for all $x \in A$. Equivalently, if A is contained in a $\|\cdot\|_X$-ball of radius R: $A \subseteq \{x \in X : \|x\|_X \leq R\}$.

- $A \subseteq X$ is $\|\cdot\|_X$-*relatively compact* if its $\|\cdot\|_X$-closure is $\|\cdot\|_X$-compact.

- $(X, \|\cdot\|_X)$ is *embedded* in $(Y, \|\cdot\|_Y)$ if

  1. X is a vector subspace of Y, and
  2. the identity operator $I : X \to Y$ defined by $Ix = x$ for all $x \in X$ is continuous.

  This is also sometimes called being *continuously embedding*, since the identity operator is required to be continuous. Since I is linear, part (2) is equivalent to the existence of a constant M such that

  $$\|x\|_Y \leq M\|x\|_X \qquad \text{for all } x \in X.$$

  Write $X \hookrightarrow Y$ to denote that $(X, \|\cdot\|_X)$ is embedded in $(Y, \|\cdot\|_Y)$.

- $T : X \to Y$ is a *compact operator* if it maps $\|\cdot\|_X$-bounded sets to $\|\cdot\|_Y$-relatively compact sets. That is, $T(A)$ is $\|\cdot\|_Y$-relatively compact whenever A is $\|\cdot\|_X$-bounded.

- $(X, \|\cdot\|_X)$ is *compactly embedded* in $(Y, \|\cdot\|_Y)$ if it is embedded and if the identity operator I is compact.

- A *cone* is a set $C = C(v, a, h, \kappa) = \{v + x \in \mathbb{R}^n : 0 \leq \|x\|_e \leq h, \angle(x, a) \leq \theta\}$. This cone is defined by four parameters: The cone's vertex $v \in \mathbb{R}^n$, an axis direction vector $a \in \mathbb{R}^n$, a height $h \in [0, \infty]$, and an angle parameter $\theta \in (0, 2\pi]$. $\angle(x, a)$ denotes the angle between x and a (let $\angle(x, x) = 0$). $\theta > 0$ ensures that the cone has volume. If $h < \infty$ then we say C is a *finite cone*.

- A set A satisfies the *cone condition* if there is some finite cone C such that for every $x \in A$ the cone C can be moved by rigid motions to have x as its vertex; that is, there is a finite cone C_x with vertex at x which is congruent to C. A sufficient condition for this is that A is a product of intervals, or that A is a ball.

Lemma 1. If all $\|\cdot\|_X$-balls are $\|\cdot\|_Y$-relatively compact, then $(X, \|\cdot\|_X)$ is compactly embedded in $(Y, \|\cdot\|_Y)$.

Lemma 1 states that, for proving compact embeddedness, it suffices to show that any $\|\cdot\|_X$-ball is $\|\cdot\|_Y$-relatively compact.

Lemma 2. Let $\|\cdot\|_X$ and $\|\cdot\|_Y$ be norms on a vector space A. Suppose A is $\|\cdot\|_X$-closed and $\|\cdot\|_X \leq C\|\cdot\|_Y$ for $C < \infty$. Then A is $\|\cdot\|_Y$-closed.

Corollary 1. Let $(\mathscr{F}_j, \|\cdot\|_j)$ be Banach spaces for all $j \in \mathbb{N}$ such that $\mathscr{F}_{j+1} \subseteq \mathscr{F}_j$ and $\|f\|_j \leq C_j\|f\|_{j+1}$ for all $f \in \mathscr{F}_{j+1}$, where $C_j < \infty$. Let

$$\Theta_j = \{f \in \mathscr{F}_j : \|f\|_j \leq C\}.$$

Assume Θ_k is $\|\cdot\|_1$-closed. Then Θ_k is $\|\cdot\|_j$-closed for all $1 \leq j < k$.

Lemma 2 says that closedness in a weaker norm can always be converted to closedness in a stronger norm. Lemma 3 is from Santos (2012) and gives conditions under which the reverse is true: when we can take closedness in a stronger norm and convert that to closedness in a weaker norm.

Lemma 3 (Lemma A.1 of Santos 2012)**.** Let $(H_1, \|\cdot\|_1)$ and $(H_2, \|\cdot\|_2)$ be separable Hilbert spaces. Suppose $(H_1, \|\cdot\|_1)$ is compactly embedded in $(H_2, \|\cdot\|_2)$. Let $B < \infty$ be a constant. Then the $\|\cdot\|_1$-ball

$$\Omega = \{h \in H_1 : \|h\|_1 \leq B\}$$

is $\|\cdot\|_2$-closed.

Lemma 4. Let $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$, and $(Z, \|\cdot\|_Z)$ be Banach spaces. Suppose

1. $(X, \|\cdot\|_X)$ is compactly embedded in $(Z, \|\cdot\|_Z)$.
2. $(Z, \|\cdot\|_Z)$ is embedded in $(Y, \|\cdot\|_Y)$.

Then $(X, \|\cdot\|_X)$ is compactly embedded in $(Y, \|\cdot\|_Y)$.

Note that assumption 2 implies

$$\{g : \|g\|_Z < \infty\} \subseteq \{g : \|g\|_Y < \infty\}.$$

B Norm inequality lemmas

Lemma 5. Let $\mu : \mathcal{D} \to \mathbb{R}_+$ be a nonnegative function. Let $m_0, m \geq 0$ be integers. Suppose assumption 4 holds for $\mu = \mu_s$. Then for every compact subset $\mathcal{C} \subseteq \mathcal{D}$, there is a constant $M_{\mathcal{C}} < \infty$ such that

$$\|\mu^{1/2} f\|_{m+m_0,2,\mathbb{1}_{\mathcal{C}}} \leq M_{\mathcal{C}}\|f\|_{m+m_0,2,\mu\mathbb{1}_{\mathcal{C}}}$$

for all f such that these norms are defined. If the stronger assumption 3 holds, then this result holds for $\mathcal{C} = \mathcal{D}$ too.

Lemma 5 generalizes lemma A.1 part (a) of Gallant and Nychka (1987) to allow for more general weight functions, as discussed in section 4.1. Note that Gallant and Nychka's (1987) lemma A.1

stated $\sup_{x\in\mathcal{D}} \mu(x) < \infty$ as an additional assumption. This condition is not used in our proof, nor was it used in their proof, which is fortunate since it is violated when μ upweights.

Lemma 6. Let $\mu : \mathcal{D} \to \mathbb{R}_+$ be a nonnegative function. Let $m \geq 0$ be an integer. Suppose assumption 4 holds for $\mu = \mu_s$. Then for every compact subset $\mathcal{C} \subseteq \mathcal{D}$, there is a constant $M_{\mathcal{C}} < \infty$ such that

$$\|f\|_{m,\infty,\mu^{1/2}\mathbb{1}_{\mathcal{C}}} \leq M_{\mathcal{C}} \|\mu^{1/2} f\|_{m,\infty,\mathbb{1}_{\mathcal{C}}}.$$

for all f such that these norms are defined. If the stronger assumption 3 holds, then this result holds for $\mathcal{C} = \mathcal{D}$ too.

Lemma 6 generalizes lemma A.1 part (d) of Gallant and Nychka (1987) to allow for the weaker assumption 4. Lemma 7 below is analogous to lemma 6, except now using the Sobolev L_2 norm instead of the Sobolev sup-norm. One difference, though, is that the norm on the left hand side now has μ instead of $\mu^{1/2}$.

Lemma 7. Let $\mu : \mathcal{D} \to \mathbb{R}_+$ be a nonnegative function. Let $m \geq 0$ be an integer. Suppose assumption 4 holds for $\mu = \mu_s$. Then for every compact subset $\mathcal{C} \subseteq \mathcal{D}$, there is a constant $M_{\mathcal{C}} < \infty$ such that

$$\|f\|_{m,2,\mu\mathbb{1}_{\mathcal{C}}} \leq M_{\mathcal{C}} \|\mu^{1/2} f\|_{m,2,\mathbb{1}_{\mathcal{C}}}$$

for all f such that these norms are defined. If the stronger assumption 3 holds, then this result holds for $\mathcal{C} = \mathcal{D}$ too.

Lemma 8. Let $\mu : \mathcal{D} \to \mathbb{R}_+$ be a nonnegative function. Let $m \geq 0$ be an integer. Then there is a constant $M < \infty$ such that

$$\|\mu f\|_{m,\infty} \leq M \|f\|_{m,\infty,\mu}$$

for all functions f such that these norms are defined.

C Proof of the compact embedding theorems 1 and 3

In this section we prove theorems 1 and 3. The general outline of the proof of theorem 3 follows the proof of Gallant and Nychka's (1987) lemma A.4, which is a proof of theorem 3 case (1) under the stronger assumption 3.

Proof of theorem 1 (Compact embedding).

1. This follows by the Rellich-Kondrachov theorem (Adams and Fournier (2003) theorem 6.3 part II, equation 5), since m_0 is a positive integer, and since $m_0 > d_x/2$ and $\mathcal{D}$ satisfies the cone condition. In applying the theorem, their j is our m. Their m is our m_0. Moreover, in their notation, we set $p = 2$ and $k = n = d_x$.

2. This follows by the Rellich-Kondrachov theorem (Adams and Fournier (2003) theorem 6.3, part II, equation 6), since m_0 is a positive integer, and since $m_0 > d_x/2$ and $\mathcal{D}$ satisfies the

cone condition. In applying the theorem, as in the previous part above, their j is our m and their m is our m_0. We set also $q = p = 2$ and $k = n = d_x$.

3. This follows by Adams and Fournier (2003) theorem 1.34 equation 3, and their subsequent remark at the end of that theorem statement.

4. This follows since $\|\cdot\|_{m+m_0,2} \leq M\|\cdot\|_{m+m_0,\infty}$ for some constant $0 < M < \infty$ and hence $\|\cdot\|_{m+m_0,\infty}$ bounded sets are also $\|\cdot\|_{m+m_0,2}$ bounded sets. Then apply part (2), which shows that these bounded sets are $\|\cdot\|_{m,2}$-relatively compact.

5. This follows by applying the Ascoli-Arzela theorem; see Adams and Fournier (2003) theorem 1.34 equation 4.

□

Proof of theorem 3 (Compact embedding for unbounded domains with equal weighting). We split the proof into several steps. For each of the cases, define the norms $\|\cdot\|_s$ and $\|\cdot\|_c$ as in table 5.

	$\|\cdot\|_s$	$\|\cdot\|_c$
(1)	$\|\cdot\|_{m+m_0,2,\mu_s}$	$\|\cdot\|_{m,\infty,\mu_c^{1/2}}$
(2)	$\|\cdot\|_{m+m_0,2,\mu_s}$	$\|\cdot\|_{m,2,\mu_c}$
(3)	$\|\cdot\|_{m+m_0,\infty,\mu_s}$	$\|\cdot\|_{m,\infty,\mu_c}$
(4)	$\|\cdot\|_{m+m_0,\infty,\mu_s}$	$\|\cdot\|_{m,2,\mu_c}$

Table 5

1. **Only look at balls**. By lemma 1, it suffices to show that for any $B > 0$, the $\|\cdot\|_s$-ball Θ of radius B is $\|\cdot\|_c$-relatively compact.

   (Cases 1 and 2.) $\Theta = \{f \in \mathscr{W}_{m+m_0,2,\mu_s}(\mathcal{D}) : \|f\|_{m+m_0,2,\mu_s} \leq B\}$.

   (Cases 3 and 4.) $\Theta = \{f \in \mathscr{C}_{m+m_0,\infty,\mu_s}(\mathcal{D}) : \|f\|_{m+m_0,\infty,\mu_s} \leq B\}$.

2. **Stop worrying about the closure**. We need to show that the $\|\cdot\|_c$-closure of Θ is $\|\cdot\|_c$-compact. Let $\{\bar{f}_n\}_{n=1}^{\infty}$ be a sequence from the $\|\cdot\|_c$-closure of Θ. It suffices to show that $\{\bar{f}_n\}$ has a convergent subsequence. By the definition of the closure, there exists a sequence $\{f_n\}$ from Θ with

   $$\lim_{n\to\infty} \|f_n - \bar{f}_n\|_c = 0.$$

   By the triangle inequality it suffices to show that $\{f_n\}$ has a convergent subsequence. The space

   (Case 1.) $\mathscr{C}_{m,\infty,\mu_c^{1/2}}$

   (Cases 2 and 4.) $\mathscr{W}_{m,2,\mu_c}$

(Case 3.) $\mathscr{C}_{m,\infty,\mu_c}$

is complete, so it suffices to show that $\{f_n\}$ has a Cauchy subsequence. The proof of completeness of these spaces is as follows. Recall that a function $f : \mathcal{D} \to \mathbb{R}$ on the Euclidean domain $\mathcal{D} \subseteq \mathbb{R}^{d_x}$ is locally integrable if for every compact subset $\mathcal{C} \subseteq \mathcal{D}$, $\int_{\mathcal{C}} |f(x)|\, dx < \infty$. Assumption 6 implies that both $\mu_c^{-1/2}$ (as needed in cases 1, 2, and 4) and μ_c^{-1} (as needed in case 3) are locally integrable on the support of μ_c. Next:

(Case 1) Follows by local integrability of $\mu_c^{-1/2}$ and applying theorem 5.1 of Rodríguez et al. (2004). To see this, using their notation, assumption 6′ ensures that $\Omega_1 = \cdots = \Omega_k = \mathbb{R}$ (defined in definition 4 on their page 277) and $\Omega^{(0)} = \mathbb{R}$ (defined on their page 280), and hence by their remark on page 303, the conditions of theorem 5.1 hold. This result is not specific to the one dimensional domain case; for example, see Brown and Opic (1992). The reason we use the power $-1/2$ of μ_c in assumption 6′ is by the $p = \infty$ case in definition 2 on page 277 of Rodríguez et al. (2004).

(Cases 2 and 4.) Follows by local integrability of $\mu_c^{-1/2}$, and theorem 1.11 of Kufner and Opic (1984) and their remark 4.10 (which extends their theorem to allow for higher order derivatives). The reason we use the power $-1/2$ of μ_c in assumption 6′ is by the $p = 2 < \infty$ case in definition 2 on page 277 of Rodríguez et al. (2004), or equivalently, equation (1.5) on page 538 of Kufner and Opic (1984).

(Case 3.) Follows by local integrability of μ_c^{-1} and then the same argument as case 1. The reason we use the power -1 of μ_c in assumption 6′ is by the $p = \infty$ case in definition 2 on page 277 of Rodríguez et al. (2004).

This step is important because functions in the closure may not be differentiable, in which case their norm might not be defined. Even when their norm is defined, functions in the closure do not necessarily satisfy the norm bound. Also, note that if μ_c does not have full support, such as $\mu_c(x) = \mathbb{1}(\|x\|_e \leq M)$ for some constant $M > 0$, then we simply restrict the domain to $\mathcal{D} \cap \{x \in \mathbb{R}^{d_x} : \|x\|_e \leq M\}$ and then proceed as in the bounded support case.

3. **Truncate the domain.** The key idea to deal with the unbounded domain is to partition $\mathbb{R}^{d_x}$ into the open Euclidean ball about the origin

$$\Omega_J = \{x \in \mathbb{R}^{d_x} : x'x < J\} = \{x \in \mathbb{R}^{d_x} : \|x\| < J^2\}$$

and its complement Ω_J^c. As we show in step 9 below, the norm on $\mathbb{R}^{d_x}$ can be split into two pieces: one on Ω_J and another on its complement. We will then show that each of these pieces is small. Restricting ourselves to Ω_J, we will apply existing embedding theorems for bounded domains. We then eventually pick J large enough so that the truncation error is small, which is possible because our weight functions get small as $\|x\|$ gets large.

Let $\mathbb{1}_{\Omega_J}(x) = 1$ if $x \in \Omega_J$ and equal zero otherwise.

4. **Switch to the unweighted norm** so that we can apply an existing compact embedding result for unweighted norms (on bounded domains). Since the f_n are in Θ, we know their weighted norm $\|\cdot\|_s$ is bounded by B. We show that a modified version of the sequence is bounded in an unweighted norm.

   (Cases 1 and 2.) The unweighted norm we work with here is $\|\cdot\|_{m+m_0,2,\mathbb{1}_{\Omega_J}}$. For all n,

$$\begin{aligned}\|\mu_s^{1/2} f_n\|_{m+m_0,2,\mathbb{1}_{\Omega_J}} &\leq M_J \|f_n\|_{m+m_0,2,\mu_s \mathbb{1}_{\Omega_J}} \\ &\leq M_J \|f_n\|_{m+m_0,2,\mu_s} \\ &\leq M_J B.\end{aligned}$$

   The first inequality follows by lemma 5, which can be applied by using our assumed bound

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

   for all $x \in \mathcal{C}$, where $\mathcal{C}$ is any compact subset of $\mathbb{R}^{d_x}$. Here and below we let M_J denote the constant from lemma 5 corresponding to the compact set Ω_J. The third inequality follows since $f_n \in \Theta$ and by the definition of Θ. Thus, for each J, $\{\mu_s^{1/2} f_n\}$ is $\|\cdot\|_{m+m_0,2,\mathbb{1}_{\Omega_J}}$-bounded. Notice that in this step we picked up a power $1/2$ of the weight function.

   (Case 3.) The unweighted norm we work with here is $\|\cdot\|_{m+m_0,\infty,\mathbb{1}_{\Omega_J}}$. For all n,

$$\begin{aligned}\|\mu_s f_n\|_{m+m_0,\infty,\mathbb{1}_{\Omega_J}} &\leq M \|f_n\|_{m+m_0,\infty,\mu_s \mathbb{1}_{\Omega_J}} \\ &\leq M \|f_n\|_{m+m_0,\infty,\mu_s} \\ &\leq MB.\end{aligned}$$

   The first inequality follows by lemma 8. The third inequality follows since $f_n \in \Theta$ and by the definition of Θ. Thus, for each J, $\{\mu_s f_n\}$ is $\|\cdot\|_{m+m_0,\infty,\mathbb{1}_{\Omega_J}}$-bounded.

   (Case 4.) The unweighted norm we work with here is $\|\cdot\|_{m+m_0,\infty,\mathbb{1}_{\Omega_J}}$. For all n,

$$\begin{aligned}\|\mu_s^{1/2} f_n\|_{m+m_0,\infty,\mathbb{1}_{\Omega_J}} &\leq M \|f_n\|_{m+m_0,\infty,\mu_s^{1/2}\mathbb{1}_{\Omega_J}} \\ &= M \max_{0\leq|\lambda|\leq m+m_0} \sup_{x\in\mathcal{D}} |\nabla^\lambda f_n(x)| \mu_s^{1/2}(x) \mathbb{1}_{\Omega_J}(x) \\ &= M \max_{0\leq|\lambda|\leq m+m_0} \sup_{x\in\mathcal{D}} |\nabla^\lambda f_n(x)| \mu_s(x) \mu_s^{-1/2}(x) \mathbb{1}_{\Omega_J}(x) \\ &\leq M \left(\max_{0\leq|\lambda|\leq m+m_0} \sup_{x\in\mathcal{D}} |\nabla^\lambda f_n(x)| \mu_s(x) \right) \sup_{\|x\|_e > J^2} \mu_s^{-1/2}(x) \\ &= M \|f_n\|_{m+m_0,\infty,\mu_s^{1/2}} \sup_{\|x\|_e > J^2} \mu_s^{-1/2}(x) \\ &\leq MBK_J.\end{aligned}$$

The first inequality follows by lemma 8. The final inequality follows since $f_n \in \Theta$ and by the definition of Θ, as well as by assumption that μ_s is bounded away from zero for any compact subset of $\mathbb{R}^{d_x}$. Thus, for each J, $\{\mu_s^{1/2} f_n\}$ is $\|\cdot\|_{m+m_0,\infty,\mathbb{1}_{\Omega_J}}$-bounded.

5. **Apply an embedding theorem for bounded domains**.

   (Case 1.) By theorem 1 part 1, $\mathscr{W}_{m+m_0,2,\mathbb{1}_{\Omega_J}}$ is compactly embedded in $\mathscr{C}_{m,\infty,\mathbb{1}_{\Omega_J}}$. Thus, since $\{\mu_s^{1/2} f_n\}$ is $\|\cdot\|_{m+m_0,2,\mathbb{1}_{\Omega_J}}$-bounded, it is relatively compact in $\mathscr{C}_{m,\infty,\mathbb{1}_{\Omega_J}}$.

   (Case 2.) By theorem 1 part 2, $\mathscr{W}_{m+m_0,2,\mathbb{1}_{\Omega_J}}$ is compactly embedded in $\mathscr{W}_{m,2,\mathbb{1}_{\Omega_J}}$. Thus, since $\{\mu_s^{1/2} f_n\}$ is $\|\cdot\|_{m+m_0,2,\mathbb{1}_{\Omega_J}}$-bounded, it is relatively compact in $\mathscr{W}_{m,2,\mathbb{1}_{\Omega_J}}$.

   (Case 3.) By theorem 1 part 3, $\mathscr{C}_{m+m_0,\infty,\mathbb{1}_{\Omega_J}}$ is compactly embedded in $\mathscr{C}_{m,\infty,\mathbb{1}_{\Omega_J}}$. Thus, since $\{\mu_s f_n\}$ is $\|\cdot\|_{m+m_0,\infty,\mathbb{1}_{\Omega_J}}$-bounded, it is relatively compact in $\mathscr{C}_{m,\infty,\mathbb{1}_{\Omega_J}}$.

   (Case 4.) By theorem 1 part 4, $\mathscr{C}_{m+m_0,\infty,\mathbb{1}_{\Omega_J}}$ is compactly embedded in $\mathscr{W}_{m,2,\mathbb{1}_{\Omega_J}}$. Thus, since $\{\mu_s^{1/2} f_n\}$ is $\|\cdot\|_{m+m_0,\infty,\mathbb{1}_{\Omega_J}}$-bounded, it is relatively compact in $\mathscr{W}_{m,2,\mathbb{1}_{\Omega_J}}$.

   In cases 1, 2, and 4 we used that $m_0 > d_x/2$, and note that Ω_J satisfies the cone condition. In case 3 we used that Ω_J is convex and $m_0 \geq 1$.

6. **Extract a subsequence**. Set $J = 1$. By the previous step, there is a subsequence

   (Case 1.) $\{\mu_s^{1/2} f_j^{(1)}\}_{j=1}^{\infty}$ and a ψ_1 in $\mathscr{C}_{m,\infty,\mathbb{1}_{\Omega_1}}$ such that

   $$\lim_{j\to\infty} \|\mu_s^{1/2} f_j^{(1)} - \psi_1\|_{m,\infty,\mathbb{1}_{\Omega_1}} = 0.$$

   (Cases 2 and 4.) $\{\mu_s^{1/2} f_j^{(1)}\}_{j=1}^{\infty}$ and a ψ_1 in $\mathscr{W}_{m,2,\mathbb{1}_{\Omega_1}}$ such that

   $$\lim_{j\to\infty} \|\mu_s^{1/2} f_j^{(1)} - \psi_1\|_{m,2,\mathbb{1}_{\Omega_1}} = 0.$$

   (Case 3.) $\{\mu_s f_j^{(1)}\}_{j=1}^{\infty}$ and a ψ_1 in $\mathscr{C}_{m,\infty,\mathbb{1}_{\Omega_1}}$ such that

   $$\lim_{j\to\infty} \|\mu_s f_j^{(1)} - \psi_1\|_{m,\infty,\mathbb{1}_{\Omega_1}} = 0.$$

7. **Do it for all** J. Repeating this argument for all J, we have a bunch of nested subsequences

   (Cases 1, 2, and 4.)

   $$\{\mu_s^{1/2} f_n\} \supset \{\mu_s^{1/2} f_j^{(1)}\} \supset \{\mu_s^{1/2} f_j^{(2)}\} \supset \cdots$$

   each with

(Case 1.)

$$\lim_{j\to\infty} \|\mu_s^{1/2} f_j^{(J)} - \psi_J\|_{m,\infty,\mathbb{1}_{\Omega_J}} = 0.$$

(Cases 2 and 4.)

$$\lim_{j\to\infty} \|\mu_s^{1/2} f_j^{(J)} - \psi_J\|_{m,2,\mathbb{1}_{\Omega_J}} = 0.$$

(Case 3.)

$$\{\mu_s f_n\} \supset \{\mu_s f_j^{(1)}\} \supset \{\mu_s f_j^{(2)}\} \supset \cdots$$

each with

$$\lim_{j\to\infty} \|\mu_s f_j^{(J)} - \psi_J\|_{m,\infty,\mathbb{1}_{\Omega_J}} = 0.$$

The reason we have to extract a further subsequence from

(Cases 1, 2, and 4.) $\{\mu_s^{1/2} f_1^{(1)}\}$ is that $\{\mu_s^{1/2} f_1^{(1)}\}$
(Case 3.) $\{\mu_s f_1^{(1)}\}$ is that $\{\mu_s f_1^{(1)}\}$

only converges in the norm with $J = 1$; it may not converge in the norm with $J = 2$. So we extract a further subsequence which does converge in the norm with $J = 2$, and so on.

8. **Define the main subsequence**. Set $f_j = f_j^{(j)}$. Then $\{f_j\}$ is a subsequence of $\{f_n\}$. Our goal is to show that $\{f_j\}$ is $\|\cdot\|_c$-Cauchy. Let $\varepsilon > 0$ be given. This is a kind of diagonalization argument.

9. **Split the consistency norm into two pieces**.

(Cases 1 and 3.) For any weight μ_c and any set Ω, we have

$$\begin{aligned}
\|f\|_{m,\infty,\mu_c} &\equiv \max_{0\le|\lambda|\le m} \sup_{x\in\mathbb{R}^{d_x}} |\nabla^\lambda f(x)|\mu_c(x)\big(\mathbb{1}_\Omega(x) + \mathbb{1}_{\Omega^c}(x)\big) \\
&= \max_{0\le|\lambda|\le m} \sup_{x\in\mathbb{R}^{d_x}} \Big(|\nabla^\lambda f(x)|\mu_c(x)\mathbb{1}_\Omega(x) + |\nabla^\lambda f(x)|\mu_c(x)\mathbb{1}_{\Omega^c}(x)\Big) \\
&\le \max_{0\le|\lambda|\le m} \sup_{x\in\mathbb{R}^{d_x}} |\nabla^\lambda f(x)|\mu_c(x)\mathbb{1}_\Omega(x) + \max_{0\le|\lambda|\le m} \sup_{x\in\mathbb{R}^{d_x}} |\nabla^\lambda f(x)|\mu_c(x)\mathbb{1}_{\Omega^c}(x) \\
&= \|f\|_{m,\infty,\mu_c\mathbb{1}_\Omega} + \|f\|_{m,\infty,\mu_c\mathbb{1}_{\Omega^c}},
\end{aligned}$$

where Ω^c is the complement of Ω. Hence, for any J, and for any f_j and f_k in our main subsequence $\{f_j\}$ we have

(Case 1.)

$$\|f_j - f_k\|_{m,\infty,\mu_c^{1/2}} \le \|f_j - f_k\|_{m,\infty,\mu_c^{1/2}\mathbb{1}_{\Omega_J}} + \|f_j - f_k\|_{m,\infty,\mu_c^{1/2}\mathbb{1}_{\Omega_J^c}}.$$

(Case 3.)

$$\|f_j - f_k\|_{m,\infty,\mu_c} \le \|f_j - f_k\|_{m,\infty,\mu_c\mathbb{1}_{\Omega_J}} + \|f_j - f_k\|_{m,\infty,\mu_c\mathbb{1}_{\Omega_J^c}}.$$

(Cases 2 and 4.) We want to show that

$$\|f\|_{m,2,\mu_c} \leq \|f\|_{m,2,\mu_c \mathbb{1}_{\Omega_J}} + \|f\|_{m,2,\mu_c \mathbb{1}_{\Omega_J^c}}.$$

We have

$$\begin{aligned}
\|f\|_{m,2,\mu_c}^2 &= \sum_{0\leq|\lambda|\leq m} \int [\nabla^\lambda f(x)]^2 \mu_c(x)\, dx \\
&= \sum_{0\leq|\lambda|\leq m} \left[\int [\nabla^\lambda f(x)]^2 \mu_c(x) \mathbb{1}_{\Omega_J}(x)\, dx + \int [\nabla^\lambda f(x)]^2 \mu_c(x) \mathbb{1}_{\Omega_J^c}(x)\, dx \right] \\
&= \sum_{0\leq|\lambda|\leq m} \int [\nabla^\lambda f(x)]^2 \mu_c(x) \mathbb{1}_{\Omega_J}(x)\, dx + \sum_{0\leq|\lambda|\leq m} \int [\nabla^\lambda f(x)]^2 \mu_c(x) \mathbb{1}_{\Omega_J^c}(x)\, dx \\
&= \|f\|_{m,2,\mu_c \mathbb{1}_{\Omega_J}}^2 + \|f\|_{m,2,\mu_c \mathbb{1}_{\Omega_J^c}}^2.
\end{aligned}$$

Hence

$$\begin{aligned}
\|f\|_{m,2,\mu_c} &= \sqrt{\|f\|_{m,2,\mu_c \mathbb{1}_{\Omega_J}}^2 + \|f\|_{m,2,\mu_c \mathbb{1}_{\Omega_J^c}}^2} \\
&\leq \|f\|_{m,2,\mu_c \mathbb{1}_{\Omega_J}} + \|f\|_{m,2,\mu_c \mathbb{1}_{\Omega_J^c}},
\end{aligned}$$

where the last line follows by $\sqrt{a^2+b^2} \leq a+b$ for $a, b \geq 0$. Hence, for any J, and for any f_j and f_k in our main subsequence $\{f_j\}$ we have

$$\|f_j - f_k\|_{m,2,\mu_c} \leq \|f_j - f_k\|_{m,2,\mu_c \mathbb{1}_{\Omega_J}} + \|f_j - f_k\|_{m,2,\mu_c \mathbb{1}_{\Omega_J^c}},$$

where recall that Ω_J^c is the complement of Ω_J.

Now we just need to show that if j, k are sufficiently far out in the sequence, and J is large enough, that both of these pieces on the right hand side are small.

10. **Outside truncation piece is small**.

(Case 1.) Since $f_j \in \Theta$ for all j, $\|f_j\|_{m+m_0,2,\mu_s} \leq B$ for all j. This combined with assumption 5 let us apply lemma 9 to find a large enough J such that

$$\|f_j\|_{m,\infty,\mu_c^{1/2} \mathbb{1}_{\Omega_J^c}} < \frac{\varepsilon}{4}$$

for all j. By the triangle inequality,

$$\|f_j - f_k\|_{m,\infty,\mu_c^{1/2} \mathbb{1}_{\Omega_J^c}} < 2\frac{\varepsilon}{4} = \frac{\varepsilon}{2}.$$

(Case 2.) For this case,

$$\begin{aligned}\|f_j\|_{m,2,\mu_s\mathbb{1}_{\Omega_J^c}} &\leq \|f_j\|_{m,2,\mu_s} \\ &\leq \|f_j\|_{m+m_0,2,\mu_s} \\ &\leq B,\end{aligned}$$

where the last line follows since $f_j \in \Theta$. Next, by assumption 1,

$$\frac{\mu_c(x)}{\mu_s(x)} \to 0 \qquad \text{as} \qquad x'x \to \infty.$$

So we can choose J large enough that

$$\left(\frac{\mu_c(x)}{\mu_s(x)}\right)^{1/2} < \frac{\varepsilon^2}{4^2B^2}$$

for all $x'x > J$; i.e., for all $x \in \Omega_J^c$. Next, we have

$$\begin{aligned}\|f_j\|^2_{m,2,\mu_c\mathbb{1}_{\Omega_J^c}} &= \sum_{0\leq|\lambda|\leq m}\int_{\Omega_J^c} |\nabla^\lambda f_j(x)|^2\mu_c(x)\,dx \\ &= \sum_{0\leq|\lambda|\leq m}\int_{\Omega_J^c} |\nabla^\lambda f_j(x)|^2\mu_s(x)\frac{\mu_c(x)}{\mu_s(x)}\,dx \\ &\leq \sum_{0\leq|\lambda|\leq m}\int_{\Omega_J^c} |\nabla^\lambda f_j(x)|^2\mu_s(x)\frac{\varepsilon^2}{4^2M}\,dx \\ &= \frac{\varepsilon^2}{4^2B^2}\sum_{0\leq|\lambda|\leq m}\int_{\Omega_J^c} |\nabla^\lambda f_j(x)|^2\mu_s(x)\,dx \\ &= \frac{\varepsilon^2}{4^2B^2}\|f_j\|^2_{m,2,\mu_s\mathbb{1}_{\Omega_J^c}} \\ &\leq \frac{\varepsilon}{4^2B^2}B^2 \\ &= \frac{\varepsilon^2}{4^2}.\end{aligned}$$

(Case 4.) For this case,

$$\begin{aligned}
\|f_j\|^2_{m,2,\mu_c\mathbb{1}_{\Omega_J^c}} &= \sum_{0\leq|\lambda|\leq m}\int_{\Omega_J^c}|\nabla^\lambda f_j(x)|^2\mu_c(x)\,dx \\
&= \sum_{0\leq|\lambda|\leq m}\int_{\Omega_J^c}|\nabla^\lambda f_j(x)|^2\mu_s^2(x)\frac{\mu_c(x)}{\mu_s^2(x)}\,dx \\
&\leq C\|f\|^2_{m,\infty,\mu_s}\int_{\Omega_J^c}\frac{\mu_c(x)}{\mu_s^2(x)}\,dx \\
&\leq CB^2\int_{\Omega_J^c}\frac{\mu_c(x)}{\mu_s^2(x)}\,dx \\
&\leq \frac{\varepsilon^2}{4^2},
\end{aligned}$$

where in the last step we choose J large enough so that[16]

$$\int_{\Omega_J^c}\frac{\mu_c(x)}{\mu_s^2(x)}\,dx \leq \frac{\varepsilon^2}{4^2CB^2}.$$

This is possible by our assumption that the integral on the left hand side is finite for at least some J. That implies, by the monotone convergence theorem for sequences of pointwise decreasing functions (e.g., Folland (1999) exercise 15 on page 52), that the integral converges to zero as $J\to\infty$.

(Cases 2 and 4). Take the square root of both sides to get

$$\|f_j\|_{m,2,\mu_c\mathbb{1}_{\Omega_J^c}} \leq \frac{\varepsilon}{4}.$$

By the triangle inequality,

$$\|f_j - f_k\|_{m,2,\mu_c\mathbb{1}_{\Omega_J^c}} < 2\frac{\varepsilon}{4} = \frac{\varepsilon}{2}.$$

[16]Here we see that we could weaken our assumption on the integral to merely that $\int_{\Omega_J^c}\mu_c(x)/\mu_s(x)\,dx<\infty$ for some J if we switched to using the weight $\mu_s^{1/2}$ instead of μ_s in defining the parameter space.

(Case 3.) We have

$$\begin{aligned}
\|f_j\|_{m,\infty,\mu_c\mathbb{1}_{\Omega_J}^c} &= \max_{0\le|\lambda|\le m}\sup_{x\in\mathcal{D}}|\nabla^\lambda f_j(x)|\mu_c(x)\mathbb{1}_{\Omega_J^c}(x)\\
&= \max_{0\le|\lambda|\le m}\sup_{x\in\mathcal{D}}|\nabla^\lambda f_j(x)|\mu_s(x)\frac{\mu_c(x)}{\mu_s(x)}\mathbb{1}_{\Omega_J^c}(x)\\
&\le \left(\max_{0\le|\lambda|\le m}\sup_{x\in\mathcal{D}}|\nabla^\lambda f_j(x)|\mu_s(x)\right)\sup_{\|x\|_e\ge J^2}\frac{\mu_c(x)}{\mu_s(x)}\\
&= \|f_j\|_{m,\infty,\mu_s}\sup_{\|x\|_e\ge J^2}\frac{\mu_c(x)}{\mu_s(x)}\\
&\le \|f_j\|_{m+m_0,\infty,\mu_s}\sup_{\|x\|_e\ge J^2}\frac{\mu_c(x)}{\mu_s(x)}\\
&\le B\sup_{\|x\|_e\ge J^2}\frac{\mu_c(x)}{\mu_s(x)}\\
&\le B\frac{\varepsilon}{4B}\\
&= \frac{\varepsilon}{4}.
\end{aligned}$$

The second to last line follows by choosing J large enough, and using assumption 1. By the triangle inequality,

$$\|f_j - f_k\|_{m,\infty,\mu_c\mathbb{1}_{\Omega_J^c}} < 2\frac{\varepsilon}{4} = \frac{\varepsilon}{2}.$$

11. **Inside truncation piece is small**. In the previous step we chose a specific value of J, so here we take J as fixed. $\{f_j\}_{j=J}^\infty = \{f_j^{(j)}\}_{j=J}^\infty$ (equality follows by definition of f_j) is a subsequence from $\{f_j^{(J)}\}$. This follows since the subsequences are nested:

    (Cases 1, 2, and 4.) $\{\mu_s^{1/2}f_n\} \supset \{\mu_s^{1/2}f_j^{(1)}\} \supset \{\mu_s^{1/2}f_j^{(2)}\} \supset \cdots$.

    (Case 3.) $\{\mu_s f_n\} \supset \{\mu_s f_j^{(1)}\} \supset \{\mu_s f_j^{(2)}\} \supset \cdots$.

    (Case 1.) Since $\{\mu_s^{1/2}f_j^{(J)}\}$ converges in the norm $\|\cdot\|_{m,\infty,\mathbb{1}_{\Omega_J}}$ it is also Cauchy in that norm. Thus there is some K large enough (take $K > J$) such that

    $$\|\mu_s^{1/2}(f_j - f_k)\|_{m,\infty,\mathbb{1}_{\Omega_J}} < \frac{\varepsilon}{2M_5^{1/2}M_J'}$$

    for all $k, j > K$. Here M_J' is the constant from applying lemma 6 to $\mathcal{C} = \Omega_J$. Notice that this constant is different from M_J, which comes from applying lemma 5.

Hence

$$\begin{aligned}
\|f_j - f_k\|_{m,\infty,\mu_c^{1/2}\mathbb{1}_{\Omega_J}} &\leq M_5^{1/2}\|f_j - f_k\|_{m,\infty,\mu_s^{1/2}\mathbb{1}_{\Omega_J}} \\
&\leq M_5^{1/2} M_J' \|\mu_s^{1/2}(f_j - f_k)\|_{m,\infty,\mathbb{1}_{\Omega_J}} \qquad \text{by lemma 6} \\
&< M_5^{1/2} M_J' \frac{\varepsilon}{2M_5^{1/2} M_J'} \\
&= \frac{\varepsilon}{2}.
\end{aligned}$$

Applying lemma 6 uses assumption 4. The first line follows since

$$\begin{aligned}
\|f\|_{m,\infty,\mu_c^{1/2}\mathbb{1}_{\Omega_J}} &= \max_{0\leq|\lambda|\leq m} \sup_{x\in\mathbb{R}^{d_x}} |\nabla^\lambda f(x)|\mu_c^{1/2}(x)\mathbb{1}_{\Omega_J}(x) \\
&= \max_{0\leq|\lambda|\leq m} \sup_{x\in\mathbb{R}^{d_x}} |\nabla^\lambda f(x)|\mu_s^{1/2}(x)\left(\frac{\mu_c(x)}{\mu_s(x)}\right)^{1/2}\mathbb{1}_{\Omega_J}(x) \\
&\leq \max_{0\leq|\lambda|\leq m} \sup_{x\in\mathbb{R}^{d_x}} |\nabla^\lambda f(x)|\mu_s^{1/2}(x) M_5^{1/2}\mathbb{1}_{\Omega_J}(x) \\
&= M_5^{1/2}\|f\|_{m,\infty,\mu_s^{1/2}\mathbb{1}_{\Omega_J}},
\end{aligned}$$

where we used our assumption 2 that

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

for all $x \in \mathbb{R}^{d_x}$.

(Cases 2 and 4.) Since $\{\mu_s^{1/2} f_j^{(J)}\}$ converges in the norm $\|\cdot\|_{m,2,\mathbb{1}_{\Omega_J}}$ it is also Cauchy in that norm. Thus there is a K large enough (take $K > J$) such that

$$\|\mu_s^{1/2}(f_j - f_k)\|_{m,2,\mathbb{1}_{\Omega_J}} < \frac{\varepsilon}{2M_5^{1/2} M_J'}$$

for all $j, k > K$. Here M_J' is the constant from applying lemma 7 to $\mathcal{C} = \Omega_J$. Applying this lemma uses assumption 4. We need to show that this implies

$$\|f_j - f_k\|_{m,2,\mu_c\mathbb{1}_{\Omega_J}}$$

is small ($\leq \varepsilon/2$) for all $j, k > K$. We have

$$\begin{aligned}
\|f\|_{m,2,\mu_c \mathbb{1}_{\Omega_J}} &= \left(\sum_{0 \leq |\lambda| \leq m} \int_{\mathcal{D}} [\nabla^\lambda f(x)]^2 \mu_c(x) \mathbb{1}_{\Omega_J}(x) \, dx \right)^{1/2} \\
&= \left(\sum_{0 \leq |\lambda| \leq m} \int_{\mathcal{D}} [\nabla^\lambda f(x)]^2 \mu_s(x) \frac{\mu_c(x)}{\mu_s(x)} \mathbb{1}_{\Omega_J}(x) \, dx \right)^{1/2} \\
&\leq \left(\sup_{x \in \mathbb{R}^{d_x}} \frac{\mu_c(x)}{\mu_s(x)} \sum_{0 \leq |\lambda| \leq m} \int_{\mathcal{D}} [\nabla^\lambda f(x)]^2 \mu_s(x) \mathbb{1}_{\Omega_J}(x) \, dx \right)^{1/2} \\
&\leq M_5^{1/2} \left(\sum_{0 \leq |\lambda| \leq m} \int_{\mathcal{D}} [\nabla^\lambda f(x)]^2 \mu_s(x) \mathbb{1}_{\Omega_J}(x) \, dx \right)^{1/2} \\
&= M_5^{1/2} \|f\|_{m,2,\mu_s \mathbb{1}_{\Omega_J}},
\end{aligned}$$

where the fourth line follows by assumption 2, which said that

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

for all $x \in \mathbb{R}_{d_x}$. This shows us how to switch from weighting with μ_c to weighting with μ_s. By lemma 7,

$$\|f\|_{m,2,\mu_s \mathbb{1}_{\Omega_J}} \leq M_J' \|\mu_s^{1/2} f\|_{m,2,\mathbb{1}_{\Omega_J}}.$$

Thus we are done since

$$\begin{aligned}
\|f_j - f_k\|_{m,2,\mu_c \mathbb{1}_{\Omega_J}} &\leq M_5^{1/2} \|f_j - f_k\|_{m,2,\mu_s \mathbb{1}_{\Omega_J}} \\
&\leq M_5^{1/2} M_J' \|\mu_s^{1/2}(f_j - f_k)\|_{m,2,\mathbb{1}_{\Omega_J}} \\
&\leq M_5^{1/2} M_J' \frac{\varepsilon}{2 M_5^{1/2} M_J'} \\
&= \frac{\varepsilon}{2}.
\end{aligned}$$

(Case 3.) Since $\{\mu_s f_j^{(J)}\}$ converges in the norm $\|\cdot\|_{m,\infty,\mathbb{1}_{\Omega_J}}$ it is also Cauchy in that norm. Thus there is some K large enough (take $K > J$) such that

$$\|\mu_s(f_j - f_k)\|_{m,\infty,\mathbb{1}_{\Omega_J}} < \frac{\varepsilon}{2 M_5 M_J'}$$

for all $k, j > K$. Here M_J' is the constant from applying lemma 6 to $\mathcal{C} = \Omega_J$. Notice that this constant is different from M_J, which comes from applying lemma 5.

Hence

$$\begin{aligned}
\|f_j - f_k\|_{m,\infty,\mu_c \mathbb{1}_{\Omega_J}} &\leq M_5 \|f_j - f_k\|_{m,\infty,\mu_s \mathbb{1}_{\Omega_J}} \\
&\leq M_5 M_J' \|\mu_s (f_j - f_k)\|_{m,\infty,\mathbb{1}_{\Omega_J}} \quad \text{by lemma 6 applied with } \mu = \mu_s^2 \\
&< M_5 M_J' \frac{\varepsilon}{2 M_5 M_J'} \\
&= \frac{\varepsilon}{2}.
\end{aligned}$$

Applying lemma 6 uses assumption 4. The first line follows since

$$\begin{aligned}
\|f\|_{m,\infty,\mu_c \mathbb{1}_{\Omega_J}} &= \max_{0 \leq |\lambda| \leq m} \sup_{x \in \mathcal{D}} |\nabla^\lambda f(x)| \mu_c(x) \mathbb{1}_{\Omega_J}(x) \\
&= \max_{0 \leq |\lambda| \leq m} \sup_{x \in \mathcal{D}} |\nabla^\lambda f(x)| \mu_s(x) \frac{\mu_c(x)}{\mu_s(x)} \mathbb{1}_{\Omega_J}(x) \\
&\leq \max_{0 \leq |\lambda| \leq m} \sup_{x \in \mathcal{D}} |\nabla^\lambda f(x)| \mu_s(x) M_5 \mathbb{1}_{\Omega_J}(x) \\
&= M_5 \|f\|_{m,\infty,\mu_s \mathbb{1}_{\Omega_J}},
\end{aligned}$$

where the third line follows by assumption 2.

12. **Put previous two steps together**. We now have

$$\|f_j - f_k\|_c \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

for all $k, j > K$. The constants only depend on the choice of weight functions, not J or any other variable that changes along the sequence. Thus we have shown that $\{f_j\}$ is $\|\cdot\|_c$-Cauchy.

□

Lemma 9. Let $\mu_c, \mu_s : \mathcal{D} \to \mathbb{R}_+$ be nonnegative functions. Let $m, m_0 \geq 0$ be integers. Let Ω_J be defined as in the proof of either theorem 3 or 5. Suppose assumption 5 holds and $\|f\|_{m+m_0,2,\mu_s} \leq B$. Then there is a function $K(J)$ such that

$$\|f\|_{m,\infty,\mu_c^{1/2} \mathbb{1}_{\Omega_J^c}} \leq K(J)$$

where $K(J) \to 0$ as $J \to \infty$.

Proof of lemma 9. For all $0 \leq |\lambda| \leq m$,

$$\begin{aligned}
\|\nabla^\lambda f\|_{0,\infty,\mu_c^{1/2}\mathbb{1}_{\Omega_J^c}} &= \sup_{x\in\Omega_J^c} |\nabla^\lambda f(x)|\mu_c^{1/2}(x) \\
&= \sup_{x\in\Omega_J^c} |\nabla^\lambda f(x)|\tilde{\mu}_c^{1/2}(x)\frac{1}{g(x)} \\
&\leq \sup_{x\in\Omega_J^c} |\nabla^\lambda f(x)|\tilde{\mu}_c^{1/2}(x) \sup_{x\in\Omega_J^c} \frac{1}{g(x)} \\
&= \|\tilde{\mu}_c^{1/2}\nabla^\lambda f\|_{0,\infty,\mathbb{1}_{\Omega_J^c}} \sup_{x\in\Omega_J^c} \frac{1}{g(x)} \\
&\leq \|\tilde{\mu}_c^{1/2}\nabla^\lambda f\|_{0,\infty} \sup_{x\in\Omega_J^c} \frac{1}{g(x)}.
\end{aligned}$$

By the Sobolev embedding theorem (Adams and Fournier 2003, theorem 4.12, part 1, case A, equation 1) there is a constant $M_2 < \infty$ such that

$$\|g\|_{0,\infty} \leq M_2\|g\|_{m_0,2}$$

for all g in $\mathscr{W}_{m_0,2}$ where $m_0 > d_x/2$. This inequality implies

$$\begin{aligned}
\|\tilde{\mu}_c^{1/2}\nabla^\lambda f\|_{0,\infty} &\leq M_2\|\tilde{\mu}_c^{1/2}\nabla^\lambda f\|_{m_0,2} \\
&\leq M_2 M\|\nabla^\lambda f\|_{m_0,2,\mu_s} \\
&\equiv M_3\|\nabla^\lambda f\|_{m_0,2,\mu_s}.
\end{aligned}$$

The second line follows by using assumption 5 in arguments as in the proof of lemma 5. Hence

$$\begin{aligned}
\|\nabla^\lambda f\|_{0,\infty,\mu_c^{1/2}\mathbb{1}_{\Omega_J^c}} &\leq M_3\|\nabla^\lambda f\|_{m_0,2,\mu_s} \sup_{x\in\Omega_J^c} \frac{1}{g(x)} \\
&\leq M_3 \left(\sum_{0\leq|\eta|\leq|\lambda|+m_0} \|\nabla^\eta f\|_{0,2,\mu_s} \right) \sup_{x\in\Omega_J^c} \frac{1}{g(x)} \\
&\leq M_3 \left(\sum_{0\leq|\eta|\leq|\lambda|+m_0} \|f\|_{m+m_0,2,\mu_s} \right) \sup_{x\in\Omega_J^c} \frac{1}{g(x)} \\
&\leq M_3 \left(\sum_{0\leq|\eta|\leq|\lambda|+m_0} B \right) \sup_{x\in\Omega_J^c} \frac{1}{g(x)} \\
&\leq M_3 \left(\sum_{0\leq|\eta|\leq m+m_0} B \right) \sup_{x\in\Omega_J^c} \frac{1}{g(x)} \\
&\equiv K(J).
\end{aligned}$$

The second line uses $\sqrt{a_1^2 + \cdots + a_n^2} \leq a_1 + \cdots + a_n$ and the definition of the Sobolev L_2 norm. The

third line uses $|a_i| \leq \sqrt{a_1^2 + \cdots + a_n^2}$ for $i = 1, \ldots, n$. By the definition of Ω_J, and since $g(x) \to \infty$ as $\|x\|_e \to \infty$ (for $\mathcal{D} = \mathbb{R}^{d_x}$) or as x approaches $\mathrm{Bd}(\overline{\mathcal{D}})$ (for bounded $\mathcal{D}$),

$$\sup_{x \in \Omega_J^c} \frac{1}{g(x)} \to 0.$$

Hence $K(J) \to 0$ as $J \to \infty$. Finally,

$$\begin{aligned} \|f\|_{m,\infty,\mu_c^{1/2}\mathbb{1}_{\Omega_J^c}} &= \max_{0 \leq |\lambda| \leq m} \|\nabla^\lambda f\|_{0,\infty,\mu_c^{1/2}\mathbb{1}_{\Omega_J^c}} \\ &\leq K(J). \end{aligned}$$

□

Supplemental Appendix to "Compactness of Infinite Dimensional Parameter Spaces"

Joachim Freyberger* Matthew A. Masten†

December 23, 2015

Abstract

This supplemental appendix provides proofs for all results not already proven in the appendix of the main paper. We also provide several additional results discussed in the main paper.

A Some useful lemmas: Proofs

Proof of lemma 1. Let $A \subseteq X$ be $\|\cdot\|_X$-bounded. Then it is contained in a $\|\cdot\|_X$-ball. That ball is $\|\cdot\|_Y$-relatively compact by assumption. So A is a subset of a $\|\cdot\|_Y$-relatively compact set. Containment is preserved by taking closures of both sets, and hence the $\|\cdot\|_Y$-closure of A is a subset of a $\|\cdot\|_Y$-compact set, and is also $\|\cdot\|_Y$-compact since it is a closed subset of a compact set. □

Proof of lemma 2. Let $\{a_n\}$ be a sequence in A. Since A is $\|\cdot\|_X$-closed, any element a such that $\|a_n - a\|_X \to 0$ must be in A. Let a be such that $\|a_n - a\|_Y \to 0$. Then $\|a_n - a\|_X \to 0$ by our norm inequality. Hence $a \in A$. □

Proof of corollary 1. Follows by repeatedly applying lemma 2. □

Proof of lemma 3. This proof is given in lemma A.1 of Santos (2012) and we therefore omit it. □

Proof of lemma 4. Since $(X, \|\cdot\|_X)$ is embedded in $(Z, \|\cdot\|_Z)$, there exists a constant $M_1 > 0$ such that

$$\|\cdot\|_Z \leq M_1 \|\cdot\|_X.$$

Likewise, by assumption 2, there is a constant constant $M_2 > 0$ such that $\|\cdot\|_Y \leq M_2 \|\cdot\|_Z$. Hence

$$\|\cdot\|_Y \leq M_1 M_2 \|\cdot\|_X.$$

Thus $(X, \|\cdot\|_X)$ is embedded in $(Y, \|\cdot\|_Y)$. Next we need to show that this embedding is compact. Let $A \subseteq X$ be $\|\cdot\|_X$-bounded. Let $\{a_n\}$ be a sequence in A. By assumption 1 there is a subsequence

*Department of Economics, University of Wisconsin-Madison, jfreyberger@ssc.wisc.edu
†Department of Economics, Duke University, matt.masten@duke.edu

$\{a_{n_k}\}$ that $\|\cdot\|_Z$-converges. But by assumption 2, $\|\cdot\|_Z$ is a stronger norm than $\|\cdot\|_Y$ and hence this subsequence $\|\cdot\|_Y$-converges. Thus every sequence in A has a $\|\cdot\|_Y$-convergent subsequence and so A is $\|\cdot\|_Y$-compact. □

B Norm inequality lemmas: Proofs

In the proof of lemma 5 and other lemmas, we use the following: The product rule tells us how to differentiate functions like $h(x)g(x)$. The generalization of this rule is called *Leibniz's formula* or the *General Leibniz rule*. For functions u and v that are $|\alpha|$ times continuously differentiable near x, it is

$$[\nabla^\alpha(uv)](x) = \sum_{\{\beta:\beta\leq\alpha\}} \begin{bmatrix}\alpha\\ \beta\end{bmatrix} \nabla^\beta u(x)\nabla^{\alpha-\beta}v(x).$$

Here $\beta \leq \alpha$ is interpreted as being component-wise: $\beta \leq \alpha$ if $\beta_j \leq \alpha_j$ for $1 \leq j \leq d_x$, where d_x is the number of components in the multi-indices β and α, and is also equal to the dimension of the argument x of the functions u and v. Also,

$$\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = \prod_{j=1}^{d_x} \binom{\alpha_j}{\beta_j}$$

where

$$\binom{\alpha_j}{\beta_j} = \frac{\alpha_j!}{\beta_j!(\alpha_j-\beta_j)!}$$

is the binomial coefficient. For a reference on this formula, see Adams and Fournier (2003), page 2.

Proof of lemma 5. Applying Leibniz's formula to the function $\mu(x)^{1/2}f(x)$ we have

$$\nabla^\lambda(\mu^{1/2}f) = \sum_{\{\beta:\beta\leq\lambda\}} \begin{bmatrix}\lambda\\ \beta\end{bmatrix} (\nabla^\beta f)(\nabla^{\lambda-\beta}\mu^{1/2}),$$

for $|\lambda| \leq m + m_0$. By the triangle inequality, this implies

$$\|\nabla^\lambda(\mu^{1/2}f)\|_{0,2,\mathbb{1}_\mathcal{C}} \leq \sum_{\{\beta:\beta\leq\lambda\}} \begin{bmatrix}\lambda\\ \beta\end{bmatrix} \|\nabla^{\lambda-\beta}\mu^{1/2}\nabla^\beta f\|_{0,2,\mathbb{1}_\mathcal{C}}.$$

Using the bound on the derivatives of $\mu^{1/2}$ we have

$$\begin{aligned}
\|\nabla^{\lambda-\beta}\mu^{1/2}\nabla^{\beta}f\|_{0,2,\mathbb{1}_{\mathcal{C}}} &= \left(\int_{\mathcal{C}}[\nabla^{\lambda-\beta}\mu^{1/2}(x)\nabla^{\beta}f(x)]^2\,dx\right)^{1/2} \\
&= \left(\int_{\mathcal{C}}|\nabla^{\lambda-\beta}\mu^{1/2}(x)|^2[\nabla^{\beta}f(x)]^2\,dx\right)^{1/2} \\
&\leq \left(\int_{\mathcal{C}}|K_{\mathcal{C}}\mu^{1/2}(x)|^2[\nabla^{\beta}f(x)]^2\,dx\right)^{1/2} \\
&= K_{\mathcal{C}}^2\left(\int_{\mathcal{C}}[\nabla^{\beta}f(x)]^2\mu(x)\,dx\right)^{1/2} \\
&= K_{\mathcal{C}}^2\|\nabla^{\beta}f\|_{0,2,\mu\mathbb{1}_{\mathcal{C}}} \\
&\leq K_{\mathcal{C}}^2\|f\|_{m+m_0,2,\mu\mathbb{1}_{\mathcal{C}}},
\end{aligned}$$

where the last line follows since $m+m_0 \geq 0$. Thus, for $|\lambda| \leq m+m_0$,

$$\|\nabla^{\lambda}(\mu^{1/2}f)\|_{0,2,\mathbb{1}_{\mathcal{C}}} \leq \left(\sum_{\{\beta:\beta\leq\lambda\}}\begin{bmatrix}\lambda\\ \beta\end{bmatrix}\right)K_{\mathcal{C}}^2\|f\|_{m+m_0,2,\mu\mathbb{1}_{\mathcal{C}}}.$$

Next,

$$\begin{aligned}
\|\mu^{1/2}f\|_{m+m_0,2,\mathbb{1}_{\mathcal{C}}}^2 &= \sum_{0\leq|\lambda|\leq m+m_0}\|\nabla^{\lambda}(\mu^{1/2}f)\|_{0,2,\mathbb{1}_{\mathcal{C}}}^2 \\
&\leq \sum_{0\leq|\lambda|\leq m+m_0}\left[\left(\sum_{\{\beta:\beta\leq\lambda\}}\begin{bmatrix}\lambda\\ \beta\end{bmatrix}\right)K_{\mathcal{C}}^2\|f\|_{m+m_0,2,\mu\mathbb{1}_{\mathcal{C}}}\right]^2 \\
&= \|f\|_{m+m_0,2,\mu\mathbb{1}_{\mathcal{C}}}^2\left[K_{\mathcal{C}}^2\sum_{0\leq|\lambda|\leq m+m_0}\left(\sum_{\{\beta:\beta\leq\lambda\}}\begin{bmatrix}\lambda\\ \beta\end{bmatrix}\right)\right]^2 \\
&\equiv \|f\|_{m+m_0,2,\mu\mathbb{1}_{\mathcal{C}}}^2M_{\mathcal{C}}^2
\end{aligned}$$

and hence

$$\|\mu^{1/2}f\|_{m+m_0,2,\mathbb{1}_{\mathcal{C}}} \leq M_{\mathcal{C}}\|f\|_{m+m_0,2,\mu\mathbb{1}_{\mathcal{C}}}$$

as desired. When assumption 3 holds, the same proof above applies, but the constants now hold over all $\mathcal{D}$. □

Proof of lemma 6. We use induction. The inequality holds for $m = 0$ with $M_{\mathcal{C}} = 1$ since

$$\begin{aligned}\|f\|_{0,\infty,\mu^{1/2}\mathbb{1}_{\mathcal{C}}} &= \sup_{x\in\mathcal{D}} |f(x)|\mu^{1/2}(x)\mathbb{1}_{\mathcal{C}}(x) \\ &= \sup_{x\in\mathcal{D}} |\mu^{1/2}(x)f(x)|\mathbb{1}_{\mathcal{C}}(x) \\ &= \|\mu^{1/2}f\|_{0,\infty,\mathbb{1}_{\mathcal{C}}}.\end{aligned}$$

Suppose the inequality holds for m and let $0 < |\lambda| \leq m + 1$. By Leibniz's formula,

$$\nabla^{\lambda}(\mu^{1/2}f) = (\nabla^{\lambda}f)\mu^{1/2} + \sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}} \begin{bmatrix}\lambda\\ \beta\end{bmatrix} (\nabla^{\lambda-\beta}\mu^{1/2})(\nabla^{\beta}f),$$

which implies that

$$\begin{aligned}|(\nabla^{\lambda}f)\mu^{1/2}| &\leq |\nabla^{\lambda}(\mu^{1/2}f)| + \left|\sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}} \begin{bmatrix}\lambda\\ \beta\end{bmatrix} (\nabla^{\lambda-\beta}\mu^{1/2})(\nabla^{\beta}f)\right| \\ &\leq |\nabla^{\lambda}(\mu^{1/2}f)| + K_{\mathcal{C}} \sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}} \begin{bmatrix}\lambda\\ \beta\end{bmatrix} \mu^{1/2}|\nabla^{\beta}f|.\end{aligned}$$

The second line follows by assumption 4, assuming we only evaluate this inequality at $x \in \mathcal{C}$. Taking the supremum over x in $\mathcal{C}$ and the maximum over $|\lambda| \leq m + 1$ gives

$$\|f\|_{m+1,\infty,\mu^{1/2}\mathbb{1}_{\mathcal{C}}} \leq \|\mu^{1/2}f\|_{m+1,\infty,\mathbb{1}_{\mathcal{C}}} + K'_{\mathcal{C}}\|f\|_{m,\infty,\mu^{1/2}\mathbb{1}_{\mathcal{C}}},$$

by the definition of the norms, and since λ isn't included in the sum we get only m derivatives in this last term on the right hand side. Moreover, we picked up an extra $\leq$ since we moved the max and supremum inside the summation in the second term, and then were left with the constant

$$K'_{\mathcal{C}} \equiv K_{\mathcal{C}} \sum_{|\lambda|\leq m} \sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}} \begin{bmatrix}\lambda\\ \beta\end{bmatrix} < \infty.$$

By the induction hypothesis there is an $M'_{\mathcal{C}} < \infty$ such that

$$\|f\|_{m,\infty,\mu^{1/2}\mathbb{1}_{\mathcal{C}}} \leq M'_{\mathcal{C}}\|\mu^{1/2}f\|_{m,\infty,\mathbb{1}_{\mathcal{C}}}.$$

Moreover,

$$\|\mu^{1/2}f\|_{m,\infty,\mathbb{1}_{\mathcal{C}}} \leq \|\mu^{1/2}f\|_{m+1,\infty,\mathbb{1}_{\mathcal{C}}}.$$

Thus

$$\|f\|_{m,\infty,\mu^{1/2}\mathbb{1}_{\mathcal{C}}} \leq M'_{\mathcal{C}}\|\mu^{1/2}f\|_{m+1,\infty,\mathbb{1}_{\mathcal{C}}}.$$

Plugging this into our expression from earlier yields

$$\begin{aligned}\|f\|_{m+1,\infty,\mu^{1/2}\mathbb{1}_{\mathcal{C}}} &\leq \|\mu^{1/2}f\|_{m+1,\infty,\mathbb{1}_{\mathcal{C}}} + K'_{\mathcal{C}}\|f\|_{m,\infty,\mu^{1/2}\mathbb{1}_{\mathcal{C}}} \\ &\leq \|\mu^{1/2}f\|_{m+1,\infty,\mathbb{1}_{\mathcal{C}}} + K'_{\mathcal{C}}M'_{\mathcal{C}}\|\mu^{1/2}f\|_{m+1,\infty,\mathbb{1}_{\mathcal{C}}} \\ &= (1 + K'_{\mathcal{C}}M'_{\mathcal{C}})\|\mu^{1/2}f\|_{m+1,\infty,\mathbb{1}_{\mathcal{C}}} \\ &\equiv M_{\mathcal{C}}\|\mu^{1/2}f\|_{m+1,\infty,\mathbb{1}_{\mathcal{C}}}.\end{aligned}$$

When assumption 3 holds, the same proof above applies, but the constants now hold over all $\mathcal{D}$. □

Proof of lemma 7. We will modify the proof of lemma 6 as appropriate. As there, we use proof by induction. For the base case, set $m = 0$. Then

$$\begin{aligned}\|f\|_{0,2,\mu\mathbb{1}_{\mathcal{C}}} &= \left(\int_{\mathcal{C}}[f(x)]^2\mu(x)\,dx\right)^{1/2} \\ &= \left(\int_{\mathcal{C}}[\mu^{1/2}(x)f(x)]^2\,dx\right)^{1/2} \\ &= \|\mu^{1/2}f\|_{0,2,\mathbb{1}_{\mathcal{C}}}.\end{aligned}$$

Thus the result holds for $m = 0$. Now suppose it holds for m. Let $|\lambda|$ be such that $0 < |\lambda| \leq m+1$. Then, as in the proof of lemma 6, we have

$$\nabla^{\lambda}(\mu^{1/2}f) = (\nabla^{\lambda}f)\mu^{1/2} + \sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}}\begin{bmatrix}\lambda\\ \beta\end{bmatrix}(\nabla^{\lambda-\beta}\mu^{1/2})(\nabla^{\beta}f)$$

by Leibniz's formula. As in that proof, applying our bound on the derivative of the weight function, we get

$$|\nabla^{\lambda}f|\mu^{1/2} \leq |\nabla^{\lambda}(\mu^{1/2}f)| + K_{\mathcal{C}}\sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}}\begin{bmatrix}\lambda\\ \beta\end{bmatrix}|\nabla^{\beta}f|\mu^{1/2}.$$

Now we square both sides and integrate over $\mathcal{C}$ to obtain

$$\begin{aligned}
\int_{\mathcal{C}} |\nabla^\lambda f(x)|^2 \mu(x)\, dx \leq & \int_{\mathcal{C}} |[\nabla^\lambda(\mu^{1/2} f)](x)|^2\, dx \\
& + \int_{\mathcal{C}} K_{\mathcal{C}}^2 \sum_{\{\tilde{\beta}:\tilde{\beta}\leq\lambda,\tilde{\beta}\neq\lambda\}} \sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}} \begin{bmatrix}\lambda\\ \tilde{\beta}\end{bmatrix}\begin{bmatrix}\lambda\\ \beta\end{bmatrix} |\nabla^{\tilde{\beta}} f(x)| \cdot |\nabla^\beta f(x)| \mu(x)\, dx \\
& + \int_{\mathcal{C}} 2|[\nabla^\lambda(\mu^{1/2} f)](x)| K_{\mathcal{C}} \sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}} \begin{bmatrix}\lambda\\ \beta\end{bmatrix} |\nabla^\beta f(x)| \mu^{1/2}(x)\, dx \\
= & \int_{\mathcal{C}} |[\nabla^\lambda(\mu^{1/2} f)](x)|^2\, dx \\
& + K_{\mathcal{C}}^2 \sum_{\{\tilde{\beta}:\tilde{\beta}\leq\lambda,\tilde{\beta}\neq\lambda\}} \sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}} \begin{bmatrix}\lambda\\ \tilde{\beta}\end{bmatrix}\begin{bmatrix}\lambda\\ \beta\end{bmatrix} \int_{\mathcal{C}} |\nabla^{\tilde{\beta}} f(x)| \cdot |\nabla^\beta f(x)| \mu(x)\, dx \\
& + 2K_{\mathcal{C}} \sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}} \begin{bmatrix}\lambda\\ \beta\end{bmatrix} \int_{\mathcal{C}} |[\nabla^\lambda(\mu^{1/2} f)](x)| \cdot |\nabla^\beta f(x)| \mu^{1/2}(x)\, dx \\
\equiv & (1) + (2) + (3).
\end{aligned}$$

In the third term, we can apply Leibniz's formula again,

$$|\nabla^\beta f|\mu^{1/2} \leq |\nabla^\beta(\mu^{1/2} f)| + K_{\mathcal{C}} \sum_{\{\eta:\eta\leq\beta,\eta\neq\beta\}} \begin{bmatrix}\beta\\ \eta\end{bmatrix} |\nabla^\eta f|\mu^{1/2}$$

to get

$$\begin{aligned}
(3) &\equiv 2K_{\mathcal{C}} \sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}} \begin{bmatrix}\lambda\\ \beta\end{bmatrix} \int_{\mathcal{C}} |[\nabla^\lambda(\mu^{1/2} f)](x)| \cdot |\nabla^\beta f(x)| \mu^{1/2}(x)\, dx \\
&\leq 2K_{\mathcal{C}} \sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}} \begin{bmatrix}\lambda\\ \beta\end{bmatrix} \Bigg(\int_{\mathcal{C}} |[\nabla^\lambda(\mu^{1/2} f)](x)| \cdot |[\nabla^\beta(\mu^{1/2} f)](x)|\, dx \\
&\qquad + K_{\mathcal{C}} \sum_{\{\eta:\eta\leq\beta,\eta\neq\beta\}} \begin{bmatrix}\beta\\ \eta\end{bmatrix} \int_{\mathcal{C}} |[\nabla^\lambda(\mu^{1/2} f)](x)| \cdot |\nabla^\eta f(x)| \mu^{1/2}(x)\, dx \Bigg).
\end{aligned}$$

We can apply Leibniz's formula again to eliminate the $|\nabla^\eta f(x)|\mu^{1/2}(x)$ term. Continuing in this manner, we get a sum solely of integrals of the form

$$\int_{\mathcal{C}} |[\nabla^\lambda(\mu^{1/2} f)](x)| \cdot |[\nabla^\beta(\mu^{1/2} f)](x)|\, dx.$$

Now replace one of the two absolute value terms in the integrand with whichever one is largest.

Suppose its the λ piece. This yields

$$\int_{\mathcal{C}} |[\nabla^{\lambda}(\mu^{1/2} f)](x)| \cdot |[\nabla^{\beta}(\mu^{1/2} f)](x)| \, dx \leq \int_{\mathcal{C}} |[\nabla^{\lambda}(\mu^{1/2} f)](x)|^2 \, dx.$$

Thus the third piece is now a sum of terms like this one, where the multi-index in the differential operator can go as high as $|\lambda|$. Summing (3) over $|\lambda|$ with $0 \leq |\lambda| \leq m+1$ we obtain a sum of many unweighted integrals over $\mathcal{C}$ with integrands of the form $|[\nabla^{\lambda}(\mu^{1/2} f)](x)|^2$. Now all we have to do is group all these integrals such that our entire expression (3) is a multiple of

$$\sum_{0 \leq |\lambda| \leq m+1} \int_{\mathcal{C}} |[\nabla^{\lambda}(\mu^{1/2} f)](x)|^2 \, dx = \|\mu^{1/2} f\|^2_{m+1,2,\mathbb{1}_{\mathcal{C}}}.$$

If there are any 'missing' integrals, we can just add on the missing ones (which will give us another inequality, but that's ok since we only need an upper bound). Thus we see that, after summing over $0 \leq |\lambda| \leq m+1$, the term (3) is bounded above by

$$C_{3,\mathcal{C}} \|\mu^{1/2} f\|^2_{m+1,2,\mathbb{1}_{\mathcal{C}}}$$

for some constant $C_{3,\mathcal{C}} > 0$.

Consider now the second piece. It is a sum of integrals of the form

$$\int_{\mathcal{C}} |\nabla^{\tilde{\beta}} f(x)| \cdot |\nabla^{\beta} f(x)| \mu(x) \, dx.$$

Basically the same argument from third piece applies. We can replace one of the absolute values here with whichever is the largest, thus obtaining an integral of the form

$$\int_{\mathcal{C}} |\nabla^{\beta} f(x)|^2 \mu(x) \, dx.$$

Now summing these terms over $0 \leq |\lambda| \leq m+1$ we see that after grouping all the integrals and adding any missing terms, the entire expression (2) is a multiple of

$$\sum_{0 \leq |\lambda| \leq m} \int_{\mathcal{C}} |\nabla^{\lambda} f(x)|^2 \, dx = \|f\|^2_{m,2,\mathbb{1}_{\mathcal{C}}}.$$

It is important here that the sum only goes up to m, not $m+1$. This is because, in the term (2), the β and $\tilde{\beta}$ pieces are always strictly smaller than λ, and λ itself can only go up to $m+1$. Hence β and $\tilde{\beta}$ can only go up to m. Thus we see that the term (2) is bounded above by

$$C_{2,\mathcal{C}} \|f\|^2_{m,2,\mathbb{1}_{\mathcal{C}}}$$

for some constant $C_{2,\mathcal{C}} > 0$. Finally, consider the term (1). This term is easy because when we sum

over $0 \leq |\lambda| \leq m+1$ this term exactly equals

$$\|\mu^{1/2} f\|^2_{m+1,2,\mathbb{1}_{\mathcal{C}}}$$

without having to add any extra terms or mess with the integrands. Combining all these results, we see (by also summing over the left hand side of our original inequality) that

$$\|f\|^2_{m+1,2,\mu\mathbb{1}_{\mathcal{C}}} \leq (1 + C_{3,\mathcal{C}})\|\mu^{1/2} f\|^2_{m+1,2,\mathbb{1}_{\mathcal{C}}} + C_{2,\mathcal{C}}\|f\|^2_{m,2,\mathbb{1}_{\mathcal{C}}}.$$

Now apply the induction hypothesis to the last term to get

$$\begin{aligned}\|f\|^2_{m+1,2,\mu\mathbb{1}_{\mathcal{C}}} &\leq (1 + C_{3,\mathcal{C}})\|\mu^{1/2} f\|^2_{m+1,2,\mathbb{1}_{\mathcal{C}}} + C_{2,\mathcal{C}}\|\mu^{1/2} f\|^2_{m+1,2,\mathbb{1}_{\mathcal{C}}} \\ &= (1 + C_{3,\mathcal{C}} + C_{2,\mathcal{C}})\|\mu^{1/2} f\|^2_{m+1,2,\mathbb{1}_{\mathcal{C}}}.\end{aligned}$$

Finally, take the square root of both sides to get

$$\|f\|_{m+1,2,\mu\mathbb{1}_{\mathcal{C}}} \leq (1 + C_{3,\mathcal{C}} + C_{2,\mathcal{C}})^{1/2}\|\mu^{1/2} f\|_{m+1,2,\mathbb{1}_{\mathcal{C}}}$$

as desired. When assumption 3 holds, the same proof above applies, but the constants now hold over all $\mathcal{D}$. □

Proof of lemma 8. As in the proof of lemma 6, we have

$$\nabla^\lambda(\mu^{1/2} f) = (\nabla^\lambda f)\mu^{1/2} + \sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}} \begin{bmatrix}\lambda \\ \beta\end{bmatrix} (\nabla^{\lambda-\beta}\mu^{1/2})(\nabla^\beta f).$$

Hence

$$|\nabla^\lambda(\mu^{1/2} f)| \leq |(\nabla^\lambda f)\mu^{1/2}| + \sum_{\{\beta:\beta\leq\lambda,\beta\neq\lambda\}} \begin{bmatrix}\lambda \\ \beta\end{bmatrix} |(\nabla^\beta f)\mu^{1/2}|.$$

Take the sup over x and the max over $|\lambda| \leq m+1$ to get

$$\|\mu^{1/2} f\|_{m+1,\infty} \leq \|f\|_{m+1,\infty,\mu^{1/2}} + K'\|f\|_{m,\infty,\mu^{1/2}}.$$

Since $\|f\|_{m,\infty,\mu^{1/2}} \leq \|f\|_{m+1,\infty,\mu^{1/2}}$ we get

$$\|\mu^{1/2} f\|_{m+1,\infty} \leq (1 + K')\|f\|_{m+1,\infty,\mu^{1/2}}.$$

The result follows by evaluating this inequality with the weight μ^2. □

C Proofs of the compact embedding theorems 5 and 7

Proof of theorem 5 (Compact embedding for unbounded domains with product weighting). For cases 1–3, we apply lemma S1 below, which allows us to convert our previous compact embedding and closedness results for equal weighting to results for product weighting. For case 4, we do not have such a prior result because it's not clear how to define equal weighted Hölder norms, as discussed in the main paper. Hence for this case we instead modify the proof of the previous compact embedding and closedness results.

Cases 1–3: Theorem 3 (case 1: part 1 with the s weight equal to the constant 1 and the c weight equal to $\tilde{\mu}^2$) (case 2: part 2 with the s weight equal to 1 and the c weight equal to $\tilde{\mu}$) (case 3: part 3, with weights chosen as in case 2) implies that (cases 1 and 2: $\mathscr{W}_{m+m_0,2,\mathbb{1}}$) (case 3: $\mathscr{C}_{m+m_0,\infty,\mathbb{1}}$) is compactly embedded in (cases 1 and 3: $\mathscr{C}_{m,\infty,\tilde{\mu}}$) (case 2: $\mathscr{W}_{m,2,\tilde{\mu}}$). Note that both the constant weight function, $\tilde{\mu}$, and $\tilde{\mu}^2$ satisfy the local integrability assumptions $6'$ and $6''$ as well as assumption 3.

By proposition 6, (cases 1 and 3: $\|\cdot\|_{m,\infty,\tilde{\mu}}$) (case 2: $\|\cdot\|_{m,2,\tilde{\mu}}$) and (cases 1 and 3: $\|\cdot\|_{m,\infty,\tilde{\mu},\text{ALT}}$) (case 2: $\|\cdot\|_{m,2,\tilde{\mu},\text{ALT}}$) are equivalent norms. Therefore (cases 1 and 2: $\mathscr{W}_{m+m_0,2,\mathbb{1}} = \mathscr{W}_{m+m_0,2,\mathbb{1},\text{ALT}}$) (case 3: $\mathscr{C}_{m+m_0,\infty,\mathbb{1},\text{ALT}}$) is compactly embedded in (cases 1 and 3: $\mathscr{C}_{m,\infty,\tilde{\mu},\text{ALT}}$) (case 2: $\mathscr{W}_{m,2,\tilde{\mu},\text{ALT}}$). Lemma S1 part 1 now implies that (cases 1 and 2: $\mathscr{W}_{m+m_0,2,\mu_s,\text{ALT}}$) (case 3: $\mathscr{C}_{m+m_0,\infty,\mu_s,\text{ALT}}$) is compactly embedded in (cases 1 and 3: $\mathscr{C}_{m,\infty,\mu_c,\text{ALT}}$) (case 2: $\mathscr{W}_{m,2,\mu_c,\text{ALT}}$).

Case 4: The proof is similar to the proof of theorem 3. Since we have already given a detailed proof of that theorem, here we only comment on the nontrivial modifications to that proof. The numbers here refer to the steps in that proof.

1. $\Theta = \{f \in \mathscr{C}_{m+m_0,\infty,\mu_s,\nu} : \|\mu_s f\|_{m+m_0,\infty,\mathbb{1},\nu} \leq B\}$.

2. Completeness of the function spaces under product weighting follows by completeness of the unweighted spaces.

4. This step is not necessary since, by definition of the product weighted norms, $f_n \in \Theta$ for all n implies

   $\{\mu_s f_n\}$ is $\|\cdot\|_{m+m_0,\infty,\mathbb{1},\nu}$-bounded. In particular, this implies it is $\|\cdot\|_{m+m_0,\infty,\mathbb{1}_{\Omega_J},\nu}$-bounded for each J, where here

   $$\|g\|_{m+m_0,\infty,\mathbb{1}_{\Omega_J},\nu} = \|g\|_{m+m_0,\infty,\mathbb{1}_{\Omega_J}} + \max_{|\lambda|=m+m_0} \sup_{x,y\in\Omega_J,x\neq y} \frac{|\nabla^\lambda f(x) - \nabla^\lambda f(y)|}{\|x-y\|_e^\nu}.$$

   Generally, in this proof indicators in the weight function placeholder denote the set over which integration or suprema are taken.

5. Apply theorem 1 part 5. Since $\{\mu_s f_n\}$ is $\|\cdot\|_{m+m_0,\infty,\mathbb{1}_{\Omega_J},\nu}$-bounded, it is $\|\cdot\|_{m,\infty,\mathbb{1}_{\Omega_J}}$-relatively compact.

9. By identical calculations as before, we have

$$\|f_j - f_k\|_{m,\infty,\mu_c,\text{ALT}} \leq \|f_j - f_k\|_{m,\infty,\mu_c \mathbb{1}_{\Omega_J},\text{ALT}} + \|f_j - f_k\|_{m,\infty,\mu_c \mathbb{1}_{\Omega_J^c},\text{ALT}}.$$

10. For $f_j \in \Theta$ we have

$$\begin{aligned}
\|f_j\|_{m,\infty,\mu_c \mathbb{1}_{\Omega_J^c},\text{ALT}} &= \|\mu_c f_j\|_{m,\infty,\mathbb{1}_{\Omega_J^c}} \\
&= \|\mu_s \tilde{\mu} f_j\|_{m,\infty,\mathbb{1}_{\Omega_J^c}} \\
&\leq M \|\mu_s f_j\|_{m,\infty,\tilde{\mu}\mathbb{1}_{\Omega_J^c}} \\
&= M \max_{0 \leq |\lambda| \leq m} \sup_{x \in \Omega_J^c} |\nabla^\lambda (\mu_s(x) f_j(x))| \tilde{\mu}(x) \\
&\leq M \max_{0 \leq |\lambda| \leq m} \sup_{x \in \mathbb{R}^{d_x}} |\nabla^\lambda (\mu_s(x) f_j(x))| \sup_{x \in \Omega_J^c} \tilde{\mu}(x) \\
&\leq M \|\mu_s f_j\|_{m+m_0,\infty,\mathbb{1},\nu} \sup_{x \in \Omega_J^c} \tilde{\mu}(x) \\
&\leq MB \sup_{x \in \Omega_J^c} \tilde{\mu}(x).
\end{aligned}$$

The third line follows by lemma 8. The last line follows since $f_j \in \Theta$. Now since $\tilde{\mu}(x) = (1 + x'x)^{-\delta}$, $\delta > 0$, converges to zero in the tails, we can choose J large enough such that

$$\sup_{x \in \Omega_J^c} \tilde{\mu}(x) < \frac{\varepsilon}{4MB}.$$

Hence, by the triangle inequality,

$$\|f_j - f_k\|_{m,\infty,\mu_c \mathbb{1}_{\Omega_J^c}} < \frac{\varepsilon}{2}.$$

11. Since $\{\mu_s f_j^{(J)}\}$ converges in the norm $\|\cdot\|_{m,\infty,\mathbb{1}_{\Omega_J}}$ it is also Cauchy in that norm. Thus there is some K large enough (take $K > J$) such that

$$\|\mu_s (f_j - f_k)\|_{m,\infty,\mathbb{1}_{\Omega_J}} < \frac{\varepsilon}{2M}$$

for all $k, j > K$, where M is a constant given below. Hence

$$\begin{aligned}
\|f_j - f_k\|_{m,\infty,\mu_c \mathbb{1}_{\Omega_J},\text{ALT}} &= \|\mu_c(f_j - f_k)\|_{m,\infty,\mathbb{1}_{\Omega_J}} \\
&= \|\mu_s \tilde{\mu}(f_j - f_k)\|_{m,\infty,\mathbb{1}_{\Omega_J}} \\
&\leq M\|\mu_s(f_j - f_k)\|_{m,\infty,\tilde{\mu}\mathbb{1}_{\Omega_J}} \\
&\leq M\|\mu_s(f_j - f_k)\|_{m,\infty,\mathbb{1}_{\Omega_J}} \\
&< M\frac{\varepsilon}{2M} \\
&= \frac{\varepsilon}{2}.
\end{aligned}$$

The third line follows by lemma 8. The fourth line follows since $\tilde{\mu}(x) = (1 + x'x)^{-\delta} \leq 1$ for all x.

□

Lemma S1. Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be Banach spaces where $\|f\|_X < \infty$ for all $f \in X$ and $\|f\|_Y < \infty$ for all $f \in Y$. Moreover, suppose that for all $f \in X$

$$\|f\|_X = \|f\|_s$$

and for all $f \in Y$

$$\|f\|_Y = \|f\tilde{\mu}\|_c$$

where $\|\cdot\|_s$ and $\|\cdot\|_c$ are norms and $\tilde{\mu}$ is a weight function. Let $(\tilde{X}, \|\cdot\|_{\tilde{X}})$ and $(\tilde{Y}, \|\cdot\|_{\tilde{Y}})$ be Banach spaces where $\|f\|_{\tilde{X}} < \infty$ for all $f \in \tilde{X}$ and $\|f\|_{\tilde{Y}} < \infty$ for all $f \in \tilde{Y}$. Moreover, suppose that for all $f \in \tilde{X}$

$$\|f\|_{\tilde{X}} = \|f\mu_s\|_s$$

and for all $f \in \tilde{Y}$

$$\|f\|_{\tilde{Y}} = \|f\mu_s\tilde{\mu}\|_c$$

for some weight function μ_s.

1. (Compact embedding) Suppose $(X, \|\cdot\|_X)$ is compactly embedded in $(Y, \|\cdot\|_Y)$. Then $(\tilde{X}, \|\cdot\|_{\tilde{X}})$ is compactly embedded in $(\tilde{Y}, \|\cdot\|_{\tilde{Y}})$.

2. (Closedness) Suppose

   $$\Omega = \{f \in X : \|f\|_X \leq B\}$$

   is $\|\cdot\|_Y$-closed. Then

   $$\tilde{\Omega} = \{f \in \tilde{X} : \|f\|_{\tilde{X}} \leq B\}$$

   is $\|\cdot\|_{\tilde{Y}}$-closed.

Proof of lemma S1.

1. Let $f \in \tilde{X}$. By definition, $\|f\|_{\tilde{X}} = \|f\mu_s\|_s < \infty$. Define $h = f\mu_s$ and notice that $h \in X$. Since $(X, \|\cdot\|_X)$ is compactly embedded in $(Y, \|\cdot\|_Y)$, $X \subseteq Y$ and there exists a constant C such that $\|h\|_Y \leq C\|h\|_X$. First note that $h \in X$ implies $\|h\|_Y < \infty$ and hence $\|h\tilde{\mu}\|_c = \|f\mu_s\tilde{\mu}\|_c < \infty$. So $f \in \tilde{Y}$ and thus $\tilde{X} \subseteq \tilde{Y}$. Next, note that

$$\begin{aligned}\|h\|_Y \leq C\|h\|_X &\Leftrightarrow \|h\tilde{\mu}\|_c \leq C\|h\|_s \\ &\Leftrightarrow \|f\mu_s\tilde{\mu}\|_c \leq C\|f\mu_s\|_s \\ &\Leftrightarrow \|f\|_{\tilde{Y}} \leq C\|f\|_{\tilde{X}}.\end{aligned}$$

   Next let $\{f_n\}$ be a sequence in the $\|\cdot\|_{\tilde{Y}}$-closure of

$$\tilde{\Omega} = \{f \in \tilde{X} : \|f\|_{\tilde{X}} \leq B\} = \{f \in \tilde{X} : \|f\mu_s\|_s \leq B\}.$$

   Let $h_n = f_n\mu_s$. Then by definition of the norms, h_n is a sequence in the $\|\cdot\|_Y$-closure of

$$\Omega = \{h \in X : \|h\|_X \leq B\}.$$

   Since $(X, \|\cdot\|_X)$ is compactly embedded in $(Y, \|\cdot\|_Y)$, there exists a subsequence $h_{n_j} = f_{n_j}\mu_s$, which is $\|\cdot\|_Y$-Cauchy. That is, for any $\varepsilon > 0$, there exists an N such that $\|h_{n_j} - h_{n_k}\|_Y \leq \varepsilon$ for all $j, k > N$. But

$$\|h_{n_j} - h_{n_k}\|_Y = \|(h_{n_j} - h_{n_k})\tilde{\mu}\|_c = \|(f_{n_j} - f_{n_k})\mu_s\tilde{\mu}\|_c = \|f_{n_j} - f_{n_k}\|_{\tilde{Y}}.$$

   Therefore, f_{n_j} is a subsequence of f_n which is $\|\cdot\|_{\tilde{Y}}$-Cauchy. Since $(\tilde{Y}, \|\cdot\|_{\tilde{Y}})$ is Banach, f_j converges to a point in $\tilde{Y}$. Hence $(\tilde{X}, \|\cdot\|_{\tilde{X}})$ is compactly embedded in $(\tilde{Y}, \|\cdot\|_{\tilde{X}})$.

2. Let f_n be a sequence in $\tilde{\Omega}$ such that for some $f \in \tilde{X}$, $\|f_n - f\|_{\tilde{Y}} \to 0$ as $n \to \infty$. Since $f_n \in \tilde{\Omega}$ we have $\|f_n\mu_s\|_s = \|f\|_{\tilde{X}} \leq B$. Let $h_n = f_n\mu_s$ and $h = f\mu_s$. Since

$$\|h_n\|_X = \|h_n\|_s = \|f_n\mu_s\|_s = \|f\|_{\tilde{X}} \leq B$$

   we have $h_n \in \Omega$. Moreover,

$$\|h_n - h\|_Y = \|(h_n - h)\tilde{\mu}\|_c = \|f_n - f\|_{\tilde{Y}} \to 0.$$

   Since $\Omega = \{f \in X : \|f\|_X \leq B\}$ is $\|\cdot\|_Y$-closed, $h \in \Omega$. That is, $f\mu_s \in \Omega$, which implies that

$$\|f\|_{\tilde{X}} = \|f\mu_s\|_X \leq B.$$

   Hence $f \in \tilde{\Omega}$. So $\tilde{\Omega}$ is $\|\cdot\|_{\tilde{Y}}$-closed.

□

Proof of theorem 7 (Compact embedding for weighted norms on bounded domains). The proof is similar to the proof of theorem 3. Since we have already given a detailed proof of that theorem, here we only comment on the nontrivial modifications to that proof. The numbers here refer to the steps in that proof.

2. For case 1, $\Omega_1 = \cdots = \Omega_k = \mathcal{D}$ and $\Omega^{(0)} = \mathcal{D}$ when applying Rodríguez, Álvarez, Romera, and Pestana (2004).

3. We use the following more general domain truncation: Let $\{\Omega_J\}$ be a sequence of open subsets of $\mathcal{D}$ such that

   (a) $\Omega_J \subseteq \Omega_{J+1}$ for any J,

   (b) $\bigcup_{J=1}^{\infty} \Omega_J = \mathcal{D}$, and

   (c) The closure of Ω_J does not contain the boundary of the closure of $\mathcal{D}$ for any J. That is, $\text{Boundary}(\overline{\mathcal{D}}) \cap \overline{\Omega}_J = \emptyset$ for all J.

   Roughly speaking, the sets Ω_J are converging to $\mathcal{D}$ from the inside. They do this in such a way that for any J, the boundary points of $\overline{\mathcal{D}}$ are well separated from Ω_J.

The rest of the steps go through with very minor modifications. □

D Proofs of closedness theorems

Proof of theorem 2 (Closedness for bounded domains). For this proof we let $d_x = 1$ to simplify the notation. All arguments generalize to $d_x > 1$.

1. We want to show that the $\|\cdot\|_s = \|\cdot\|_{m+m_0,2}$-ball Θ is $\|\cdot\|_c = \|\cdot\|_{m,\infty}$-closed. $(\mathscr{W}_{m+m_0,2}, \|\cdot\|_{m+m_0,2})$ is compactly embedded in $(\mathscr{W}_{m,2}, \|\cdot\|_{m,2})$ by part 2 of theorem 1, which applies since we assumed $\mathcal{D}$ satisfies the cone condition and $m_0 > d_x/2$. Lemma A.1 in Santos (2012) (reproduced in the main paper's appendix on page 38 for convenience) then implies that that the $\|\cdot\|_{m+m_0,2}$-ball Θ is $\|\cdot\|_{m,2}$-closed, because the Sobolev L_2 spaces are separable Hilbert spaces (theorem 3.6 of Adams and Fournier 2003). Finally, since $\|\cdot\|_{m,2} \leq \|\cdot\|_{m,\infty}$ corollary 1 implies that Θ is $\|\cdot\|_{m,\infty}$-closed.

2. We want to show that the $\|\cdot\|_s = \|\cdot\|_{m+m_0,2}$-ball Θ is $\|\cdot\|_c = \|\cdot\|_{m,2}$-closed. We already showed this in the proof of part 1.

3. We want to show that the $\|\cdot\|_s = \|\cdot\|_{m+m_0,\infty}$-ball Θ is not $\|\cdot\|_c = \|\cdot\|_{m,\infty}$-closed. Consider the case $m = 0$ and $m_0 = 1$, so that Θ is the set of continuously differentiable functions whose levels and first derivatives are uniformly bounded by B. We will show that this set is not closed in the ordinary sup-norm $\|\cdot\|_{0,\infty}$.

   Suppose $\mathcal{D} = (-1, 1)$. Define

   $$g_k(x) = \sqrt{x^2 + 1/k}.$$

for integers $k \geq 1$. These are smooth approximations to the absolute value function: For each $x \in \mathcal{D}$, $g_k(x) \to \sqrt{x^2} = |x|$ as $k \to \infty$. g_k is continuous and differentiable, with first derivative

$$\begin{aligned} g_k'(x) &= \frac{1}{2}(x^2 + 1/k)^{-1/2} \cdot 2x \\ &= \frac{x}{\sqrt{x^2 + 1/k}}. \end{aligned}$$

So

$$|g_k'(x)| \leq \frac{|x|}{\sqrt{x^2 + 1/k}} \leq \frac{|x|}{\sqrt{x^2}} = 1$$

for all k. Also,

$$|g_k(x)| = \sqrt{x^2 + 1/k} \leq \sqrt{1 + 1/k} \leq \sqrt{1 + 1} = \sqrt{2}$$

for all k. Hence $g_k \in \Theta = \{f \in \mathscr{C}_1(\mathcal{D}) : \|f\|_{1,\infty} \leq B\}$ for each k, where $B = 1 + \sqrt{2}$. But, letting $f(x) = |x|$,

$$\|g_k - f\|_{0,\infty} = \sup_{x \in \mathcal{D}} |g_k(x) - f(x)| \to 0$$

as $k \to \infty$. Since f is not differentiable at 0, $f \notin \Theta$. This implies that Θ is not closed under $\|\cdot\|_{0,\infty}$.

4. We want to show that the $\|\cdot\|_s = \|\cdot\|_{m+m_0,\infty}$-ball Θ is not $\|\cdot\|_c = \|\cdot\|_{m,2}$-closed. The same counterexample from part 4 applies here as well. Letting $m = 0$ and $m_0 = 1$, we will show that the $\|\cdot\|_{1,\infty}$-ball Θ is not closed in the ordinary L_2 norm $\|\cdot\|_{0,2}$. From part 4, we constructed a sequence g_k in Θ such that

$$\|g_k - f\|_{0,\infty} \to 0$$

as $k \to \infty$, for $f \notin \Theta$. Convergence in $\|\cdot\|_{0,\infty}$ implies convergence in $\|\cdot\|_{0,2}$ and hence

$$\|g_k - f\|_{0,2} \to 0$$

as $k \to \infty$. Therefore Θ is not closed under $\|f\|_{0,2}$.

5. We want to show that $\|\cdot\|_{m+m_0,\infty,\mathbb{1},\nu}$-balls are $\|\cdot\|_{m,\infty}$-closed, where $m_0 \geq 0$. Since $\|\cdot\|_{0,\infty} \leq \|\cdot\|_{m,\infty}$, corollary 1 shows that it is sufficient to prove the result for $m = 0$. That is, it is sufficient to prove that the $\|\cdot\|_{m_0,\infty,\mathbb{1},\nu}$-ball

$$\Theta_{m_0} \equiv \{f \in \mathscr{C}_{m_0,\infty,\mathbb{1},\nu} : \|f\|_{m_0,\infty,\mathbb{1},\nu} \leq B\}$$

is $\|\cdot\|_{0,\infty}$-closed, for all $m_0 \geq 0$. We proceed by induction on m_0.

Step 1 (Base Case): Let $m_0 = 0$. We want to show that Θ_0 is $\|\cdot\|_{0,\infty}$-closed, so we will show that its complement $\Theta_0^c = \mathscr{C}_{0,\infty} \setminus \Theta_0$ is $\|\cdot\|_{0,\infty}$-open. That is, for any $f \in \Theta_0^c$ there

exists an $\varepsilon > 0$ such that

$$\{g \in \mathscr{C}_{0,\infty} : \|f - g\|_{0,\infty} \leq \varepsilon\} \subseteq \Theta_0^c.$$

So take an arbitrary $f \in \Theta_0^c$. Since f is outside the Hölder ball Θ_0, its Hölder norm is larger than B,

$$\sup_{x \in \mathcal{D}} |f(x)| + \sup_{x_1, x_2 \in \mathcal{D}, x_1 \neq x_2} \frac{|f(x_1) - f(x_2)|}{|x_1 - x_2|^\nu} > B.$$

Hence there exist points $\bar{x}, \bar{x}_1, \bar{x}_2$ in the Euclidean closure of $\mathcal{D}$ with $\bar{x}_1 \neq \bar{x}_2$ such that

$$|f(\bar{x})| + \frac{|f(\bar{x}_1) - f(\bar{x}_2)|}{|\bar{x}_1 - \bar{x}_2|^\nu} > B.$$

Define

$$\delta = |f(\bar{x})| + \frac{|f(\bar{x}_1) - f(\bar{x}_2)|}{|\bar{x}_1 - \bar{x}_2|^\nu} - B > 0.$$

Our goal is find a $\|\cdot\|_{0,\infty}$-ball around f with some positive radius ε such that all functions g in that ball are also not in the Hölder ball Θ_0. So we need these functions g to have a large Hölder norm (larger than B). Let's examine that. For all $g \in \mathscr{C}_{0,\infty}$,

$$\begin{aligned}
\|g\|_{0,\infty,\mathbb{1},\nu} &= \sup_{x \in \mathcal{D}} |g(x)| + \sup_{x_1, x_2 \in \mathcal{D}, x_1 \neq x_2} \frac{|g(x_1) - g(x_2)|}{|x_1 - x_2|^\nu} \\
&\geq |g(\bar{x})| + \frac{|g(\bar{x}_1) - g(\bar{x}_2)|}{|\bar{x}_1 - \bar{x}_2|^\nu} \\
&\geq |f(\bar{x})| - |f(\bar{x}) - g(\bar{x})| + \frac{|g(\bar{x}_1) - g(\bar{x}_2)|}{|\bar{x}_1 - \bar{x}_2|^\nu} \\
&= |f(\bar{x})| - |f(\bar{x}) - g(\bar{x})| \\
&\quad + \frac{|f(\bar{x}_1) - f(\bar{x}_2)|}{|\bar{x}_1 - \bar{x}_2|^\nu} - \frac{|f(\bar{x}_1) - f(\bar{x}_2)|}{|\bar{x}_1 - \bar{x}_2|^\nu} + \frac{|g(\bar{x}_1) - g(\bar{x}_2)|}{|\bar{x}_1 - \bar{x}_2|^\nu} \\
&\geq |f(\bar{x})| - |f(\bar{x}) - g(\bar{x})| \\
&\quad + \frac{|f(\bar{x}_1) - f(\bar{x}_2)|}{|\bar{x}_1 - \bar{x}_2|^\nu} - \frac{|\,(f(\bar{x}_1) - g(\bar{x}_1)) - (f(\bar{x}_2) - g(\bar{x}_2))\,|}{|\bar{x}_1 - \bar{x}_2|^\nu} \\
&= B + \delta - \left(|f(\bar{x}) - g(\bar{x})| + \frac{|\,(f(\bar{x}_1) - g(\bar{x}_1)) - (f(\bar{x}_2) - g(\bar{x}_2))\,|}{|\bar{x}_1 - \bar{x}_2|^\nu} \right).
\end{aligned}$$

The third and fifth lines follow by the reverse triangle inequality. The last line follows by the definition of δ. If we can make this last piece in parentheses small enough, we'll be done. For any $\varepsilon > 0$,

$$g \in \{g \in \mathscr{C}_{0,\infty} : \|f - g\|_{0,\infty} \leq \varepsilon\}$$

implies

$$|f(\bar{x}) - g(\bar{x})| + \frac{|(f(\bar{x}_1) - g(\bar{x}_1)) - (f(\bar{x}_2) - g(\bar{x}_2))|}{|\bar{x}_1 - \bar{x}_2|^\nu} \leq \varepsilon + \frac{2\varepsilon}{|\bar{x}_1 - \bar{x}_2|^\nu}$$

by the triangle inequality. So suppose we choose ε so that

$$\varepsilon + \frac{2\varepsilon}{|\bar{x}_1 - \bar{x}_2|^\nu} \leq \frac{\delta}{2}.$$

Note that this choice of ε depends on the particular $f \in \Theta_0^c$ chosen at the beginning, via δ and $\bar{x}_1$ and $\bar{x}_2$. Then for all $g \in \mathscr{C}_{0,\infty}$ with $\|f - g\|_{0,\infty} \leq \varepsilon$ we have

$$\begin{aligned} \|g\|_{0,\infty,\mathbb{1},\nu} &\geq B + \delta - \frac{\delta}{2} \\ &= B + \frac{\delta}{2} \\ &> B. \end{aligned}$$

Hence $g \in \Theta_0^c$ for all such g. Thus Θ_0^c is $\|\cdot\|_{0,\infty}$-open and hence Θ_0 is $\|\cdot\|_{0,\infty}$-closed.

Step 2 (Induction Step): Next we suppose that Θ_{m_0} is $\|\cdot\|_{0,\infty}$-closed for some integer $m_0 \geq 0$. We will show that this implies Θ_{m_0+1} is $\|\cdot\|_{0,\infty}$-closed.

Since Θ_{m_0} is $\|\cdot\|_{0,\infty}$-closed, we have that for all f in $\Theta_{m_0}^c = \mathscr{C}_{0,\infty} \setminus \Theta_{m_0}$ there exists an $\varepsilon > 0$ such that for all $g \in \mathscr{C}_{0,\infty}$ with

$$\|f - g\|_{0,\infty} \leq \varepsilon,$$

it holds that $g \in \Theta_{m_0}^c$. As in the base case, we will show that $\Theta_{m_0+1}^c$ is $\|\cdot\|_{0,\infty}$-open. So take an arbitrary $f \in \Theta_{m_0+1}^c$. We will show that there exists an $\varepsilon > 0$ such that for all $g \in \mathscr{C}_{0,\infty}$ with $\|f - g\|_{0,\infty} \leq \varepsilon$ we have $g \in \Theta_{m_0+1}^c$. We have to consider several cases, depending on the properties of the f we're given. First, $\Theta_{m_0+1} \subsetneq \Theta_{m_0}$ implies

$$\Theta_{m_0}^c \subsetneq \Theta_{m_0+1}^c.$$

So it might be the case that $f \in \Theta_{m_0}^c$. This is case (a) below. Moreover, it is possible that $f \in \Theta_{m_0+1}^c$ but $f \notin \Theta_{m_0}^c$. This case could occur for several reasons. It might be that $f \in \mathscr{C}_{m_0+1,\infty,\mathbb{1},\nu}$, so $\|f\|_{m_0+1,\infty,\mathbb{1},\nu} \leq D$ for some constant $D < \infty$, but that this norm, while finite, is still too big:

$$\|f\|_{m_0+1,\infty,\mathbb{1},\nu} > B.$$

This is case (b) below. Another possibility is that $f \notin \mathscr{C}_{m_0+1,\infty,\mathbb{1},\nu}$. But $f \notin \Theta_{m_0}^c$, $f \in \Theta_{m_0}$ and hence its m_0'th derivative exists and is Hölder continuous. So there are three reasons why $f \notin \mathscr{C}_{m_0+1,\infty,\mathbb{1},\nu}$ could occur: Either the (m_0+1)'th derivative does not exist (case (c) below), the (m_0+1)'th derivative exists but is not $\|\cdot\|_{0,\infty}$-bounded (i.e., the first piece of the Hölder norm $\|f\|_{m_0+1,\infty,\mathbb{1},\nu}$ is infinite) (case (d) below), or the (m_0+1)'th derivative exists and is $\|\cdot\|_{0,\infty}$-bounded, but is not Hölder continuous (i.e., the first piece of the Hölder norm $\|f\|_{m_0+1,\infty,\mathbb{1},\nu}$ is finite, but the second piece is infinite) (case (e) below).

(a) Suppose $f \in \Theta_{m_0}^c$. But we already know from the induction assumption that $\Theta_{m_0}^c$ is

open. Hence there exists an $\varepsilon > 0$ such that for all $g \in \mathscr{C}_{0,\infty}$ with $\|f - g\|_{0,\infty} \leq \varepsilon$ it holds that $g \in \Theta^c_{m_0} \subsetneq \Theta^c_{m_0+1}$.

(b) Suppose $f \notin \Theta^c_{m_0}$ and $f \in \mathscr{C}_{m_0+1,\infty,\mathbb{1},\nu}$ with

$$B < \|f\|_{m_0+1,\infty,\mathbb{1},\nu} \leq D$$

for some constant $D < \infty$. Since $f \notin \Theta^c_{m_0}$, $f \in \Theta_{m_0}$ and hence

$$\|f\|_{m_0,\infty,\mathbb{1},\nu} \leq B.$$

Let $g \in \mathscr{C}_{0,\infty}$ be such that $\|f - g\|_{0,\infty} \leq \varepsilon$. Remember that our goal is to find an $\varepsilon > 0$ such that all of these g are in $\Theta^c_{m_0+1}$. Regardless of the value of ε, if $g \notin \mathscr{C}_{m_0+1,\infty,\mathbb{1},\nu}$ (in which case $g \notin \Theta_{m_0+1}$ and so $g \in \Theta^c_{m_0+1}$) or if $\|g\|_{m_0+1,\infty,\mathbb{1},\nu} \geq C$ for some finite constant $C > B$, then $g \in \Theta^c_{m_0+1}$. So suppose that $g \in \mathscr{C}_{m_0+1,\infty,\mathbb{1},\nu}$ and

$$\|g\|_{m_0+1,\infty,\mathbb{1},\nu} \leq C.$$

We will show that although this norm is smaller than C, it is still larger than B. For each $x \in \mathcal{D}$ and $\delta > 0$ with $x + \delta \in \mathcal{D}$,[1] the mean value theorem implies that there exists an $x_g \in [x, x + \delta]$ such that

$$g'(x_g) = \frac{g(x + \delta) - g(x)}{\delta}$$

and hence

$$\begin{aligned} g'(x) &= g'(x_g) + (g'(x) - g'(x_g)) \\ &= \frac{g(x + \delta) - g(x)}{\delta} + (g'(x) - g'(x_g)). \end{aligned}$$

Note that g is differentiable because $g \in \mathscr{C}_{m_0+1,\infty,\mathbb{1},\nu}$. Likewise, there exists an $x_f \in [x, x + \delta]$ such that

$$f'(x) = \frac{f(x + \delta) - f(x)}{\delta} + (f'(x) - f'(x_f)).$$

[1] The cone condition implies that there exists a single $\delta > 0$ such that, for all $x \in \mathcal{D}$, at least one of $x + \delta \in \mathcal{D}$ or $x - \delta \in \mathcal{D}$ holds.

It follows that

$$
\begin{aligned}
&\|f' - g'\|_{0,\infty} \\
&= \sup_{x\in\mathcal{D}} |f'(x) - g'(x)| \\
&= \sup_{x\in\mathcal{D}} \left| \left(\frac{f(x+\delta) - f(x)}{\delta} + (f'(x) - f'(x_f)) \right) - \left(\frac{g(x+\delta) - g(x)}{\delta} + (g'(x) - g'(x_g)) \right) \right| \\
&= \sup_{x\in\mathcal{D}} \left| \frac{f(x+\delta) - g(x+\delta)}{\delta} - \frac{f(x) - g(x)}{\delta} + (f'(x) - f'(x_f)) + (g'(x) - g'(x_g)) \right| \\
&\leq \sup_{x\in\mathcal{D}} \left(\frac{|f(x+\delta) - g(x+\delta)|}{\delta} + \frac{|f(x) - g(x)|}{\delta} + |f'(x) - f'(x_f)| + |g'(x) - g'(x_g)| \right) \\
&\leq \frac{2\varepsilon}{\delta} + D\delta^\nu + C\delta^\nu
\end{aligned}
$$

The fourth line follows by the triangle inequality. The last line by $\|f - g\|_{0,\infty} \leq \varepsilon$, $x_f \in [x, x+\delta]$, $x_g \in [x, x+\delta]$, and since f' and g' are both Hölder continuous with Hölder constants D and C, respectively (which follows because $\|f\|_{m_0+1,\infty,\mathbb{1},\nu} \leq D$ and $\|g\|_{m_0+1,\infty,\mathbb{1},\nu} \leq C$).

Let $\varepsilon_1 > 0$ be arbitrary. Choose $\delta > 0$ such that $D\delta^\nu \leq \varepsilon_1/3$ and $C\delta^\nu \leq \varepsilon_1/3$. After choosing δ, choose ε such that $2\varepsilon/\delta \leq \varepsilon_1/3$. Thus

$$\|f' - g'\|_{0,\infty} \leq \varepsilon_1.$$

We have shown that if the first derivatives of f and g are Hölder continuous, we can make the derivatives for all g with $\|f - g\|_{0,\infty} \leq \varepsilon$ arbitrarily close to the derivative of f by choosing ε small enough. An analogous argument shows that if $\|f' - g'\|_{0,\infty} \leq \varepsilon_1$ and if the second derivatives are Hölder continuous, then we can make the second derivatives arbitrarily close. Applying this argument recursively to higher order derivative shows that for any $\varepsilon_{m_0+1} > 0$, we can pick an $\varepsilon > 0$ such that for all g with $\|g\|_{m_0+1,\infty,\mathbb{1},\nu} \leq C$ and $\|f - g\|_{0,\infty} \leq \varepsilon$,

$$\|\nabla^{m_0+1} f - \nabla^{m_0+1} g\|_{0,\infty} \leq \varepsilon_{m_0+1}.$$

Our argument from the base case (step 1) now implies that if ε_{m_0+1} is small enough, then $\|g\|_{m_0+1,\infty,\mathbb{1},\nu} > B$ for all $g \in \mathscr{C}_{0,\infty}$ with $\|f - g\|_{0,\infty} \leq \varepsilon$. Hence $g \in \Theta^c_{m_0+1}$. Note that we use $\|f\|_{m_0+1,\infty,\mathbb{1},\nu} > B$ when applying the base case argument.

(c) Suppose that for some $\bar{x} \in \mathcal{D}$, $\nabla^{m_0+1} f(\bar{x})$ does not exist. Then $f \notin \mathscr{C}_{m_0+1,\infty,\mathbb{1},\nu}$. But since $f \notin \Theta^c_{m_0}$, we know that the m_0'th derivative of f exists and is Hölder continuous. As in case (b), take $g \in \mathscr{C}_{0,\infty}$ such that $\|f - g\|_{0,\infty} \leq \varepsilon$ and suppose that $g \in \mathscr{C}_{m_0+1,\infty,\mathbb{1},\nu}$ $\|g\|_{m_0+1,\infty,\mathbb{1},\nu} \leq C$ for $C > B$ (remember from part (b) that otherwise we know $g \in \Theta^c_{m_0+1}$ already). Since the m_0'th derivative of f exists and is Hölder continuous, we know that the only way for the derivative $\nabla^{m_0+1} f(\bar{x})$ to not exist is if it has a kink—its right hand side derivative does not exist, its left hand side derivative does not exist, or

both exist but are not equal. So we consider each of these three cases separately.

i. Suppose the right hand side derivative of $\nabla^{m_0} f$ at $\bar{x}$ does not exist. That is,

$$\lim_{h \searrow 0} \frac{\nabla^{m_0} f(\bar{x}+h) - \nabla^{m_0} f(\bar{x})}{h}$$

does not exist. Then there exists a $\delta > 0$ such that for any $\eta > 0$ we can find an h with $0 < h < \eta$ and

$$\left| \frac{\nabla^{m_0} f(\bar{x}+h) - \nabla^{m_0} f(\bar{x})}{h} - \nabla^{m_0+1} g(\bar{x}) \right| > \delta.$$

If such a δ did not exists, then

$$\lim_{h \searrow 0} \frac{\nabla^{m_0} f(\bar{x}+h) - \nabla^{m_0} f(\bar{x})}{h} = \nabla^{m_0+1} g(\bar{x})$$

by definition of the limit. For such a fixed h, we have

$$\begin{aligned}
\delta &< \left| \frac{\nabla^{m_0} f(\bar{x}+h) - \nabla^{m_0} f(\bar{x})}{h} - \nabla^{m_0+1} g(\bar{x}) \right| \\
&\leq \left| \frac{\nabla^{m_0} f(\bar{x}+h) - \nabla^{m_0} g(\bar{x}+h) + \nabla^{m_0} g(\bar{x}) - \nabla^{m_0} f(\bar{x})}{h} \right| \\
&\quad + \left| \frac{\nabla^{m_0} g(\bar{x}+h) - \nabla^{m_0} g(\bar{x})}{h} - \nabla^{m_0+1} g(\bar{x}) \right| \\
&\leq \left| \frac{\nabla^{m_0} f(\bar{x}+h) - \nabla^{m_0} g(\bar{x}+h) + \nabla^{m_0} g(\bar{x}) - \nabla^{m_0} f(\bar{x})}{h} \right| \\
&\quad + \left| \nabla^{m_0+1} g(\tilde{x}) - \nabla^{m_0+1} g(\bar{x}) \right| \\
&\leq \left| \frac{\nabla^{m_0} f(\bar{x}+h) - \nabla^{m_0} g(\bar{x}+h) + \nabla^{m_0} g(\bar{x}) - \nabla^{m_0} f(\bar{x})}{h} \right| + Ch^{\nu}.
\end{aligned}$$

The second line follows by the triangle inequality. The third line by the mean value theorem, since $\nabla^{m_0} g$ is differentiable, and here $\tilde{x} \in [\bar{x}, \bar{x}+h]$. The fourth line follows since $\nabla^{m_0+1} g$ is Hölder continuous with constant C, and since $\tilde{x} \in [\bar{x}, \bar{x}+h]$ so that $\|\tilde{x} - \bar{x}\| \leq h$. Now choose h small enough such that $Ch^{\nu} \leq \delta/2$. For this fixed h, pick ε small enough such that

$$\|\nabla^{m_0} f - \nabla^{m_0} g\|_{0,\infty} \leq \frac{\delta h}{4}.$$

Then

$$\delta < \left| \frac{\nabla^{m_0} f(\bar{x}+h) - \nabla^{m_0} f(\bar{x})}{h} - \nabla^{m_0+1} g(\bar{x}) \right| \leq \delta,$$

a contraction.

ii. Suppose the left hand side derivative of $\nabla^{m_0} f$ at $\bar{x}$ does not exist. That is,

$$\lim_{h \searrow 0} \frac{\nabla^{m_0} f(\bar{x}) - \nabla^{m_0} f(\bar{x} - h)}{h}$$

does not exist. This case proceeds analogously to the previous case.

iii. Both the left hand and right hand side derivatives of $\nabla^{m_0} f$ at $\bar{x}$ exist, but they are not equal:

$$\lim_{h \searrow 0} \frac{\nabla^{m_0} f(\bar{x} + h) - \nabla^{m_0} f(\bar{x})}{h} \neq \lim_{h \searrow 0} \frac{\nabla^{m_0} f(\bar{x}) - \nabla^{m_0} f(\bar{x} - h)}{h}.$$

Considering the distance between the right hand side and left hand side secant lines, for any $h > 0$ such that $[\bar{x} - h, \bar{x} + h] \subseteq \mathcal{D}$, we obtain

$$\begin{aligned}
&\left| \left(\frac{\nabla^{m_0} f(\bar{x} + h) - \nabla^{m_0} f(\bar{x})}{h} \right) - \left(\frac{\nabla^{m_0} f(\bar{x}) - \nabla^{m_0} f(\bar{x} - h)}{h} \right) \right| \\
&\quad \leq 4 \frac{\varepsilon_{m_0}}{h} + \left| \left(\frac{\nabla^{m_0} g(\bar{x} + h) - \nabla^{m_0} g(\bar{x})}{h} \right) - \left(\frac{\nabla^{m_0} g(\bar{x}) - \nabla^{m_0} g(\bar{x} - h)}{h} \right) \right| \\
&\quad = 4 \frac{\varepsilon_{m_0}}{h} + \left| \left(\nabla^{m_0+1} g(\tilde{x}_1) - \nabla^{m_0+1} g(\tilde{x}_2) \right) \right| \\
&\quad \leq 4 \frac{\varepsilon_{m_0}}{h} + C(2h)^{\nu}.
\end{aligned}$$

For the first line, we used the triangle inequality plus the fact that for any $\varepsilon_{m_0} > 0$, there exists an $\varepsilon > 0$ not depending on g such that $\|f - g\|_{0,\infty} \leq \varepsilon$ implies

$$\|\nabla^{m_0} f - \nabla^{m_0} g\|_{0,\infty} \leq \varepsilon_{m_0}.$$

This follows from our argument in part (b), since $\nabla^{m_0} f$ and $\nabla^{m_0} g$ are Hölder continuous.

In the second line, we used the mean value theorem, since $g \in \mathscr{C}_{m_0+1,\infty,\mathbb{1},\nu}$, where $\tilde{x}_1 \in [\bar{x}, \bar{x} + h]$ and $\tilde{x}_2 \in [\bar{x} - h, \bar{x}]$. In the third line we used Hölder continuity of $\nabla^{m_0+1} g$ since $\|g\|_{m_0+1,\infty,\mathbb{1},\nu} \leq C$, plus the fact that $|\tilde{x}_1 - \tilde{x}_2| \leq 2h$.

Since

$$\lim_{h \searrow 0} \frac{\nabla^{m_0} f(\bar{x} + h) - \nabla^{m_0} f(\bar{x})}{h} \neq \lim_{h \searrow 0} \frac{\nabla^{m_0} f(\bar{x}) - \nabla^{m_0} f(\bar{x} - h)}{h}$$

there exists a $\delta > 0$ such that for an arbitrarily small h

$$\left| \left(\frac{\nabla^{m_0} f(\bar{x} + h) - \nabla^{m_0} f(\bar{x})}{h} \right) - \left(\frac{\nabla^{m_0} f(\bar{x}) - \nabla^{m_0} f(\bar{x} - h)}{h} \right) \right| > \delta.$$

Choose h such that $C(2h)^{\nu} \leq \delta/2$. Then for this fixed h, pick ε small enough such

that $4\varepsilon_{m_0}/h \leq \delta/2$. Then

$$\delta < \left| \left(\frac{\nabla^{m_0} f(\bar{x}+h) - \nabla^{m_0} f(\bar{x})}{h} \right) - \left(\frac{\nabla^{m_0} f(\bar{x}) - \nabla^{m_0} f(\bar{x}-h)}{h} \right) \right| \leq \delta,$$

a contraction.

In all three cases where $\nabla^{m_0+1} f(\bar{x})$ does not exist, we have derived a contradiction. Hence there does not exist a $g \in \mathscr{C}_{0,\infty}$ with $\|g\|_{m_0+1,\infty,\mathbb{1},\nu} \leq C$ and $\|f-g\|_{0,\infty} \leq \varepsilon$. This implies that for all $g \in \mathscr{C}_{0,\infty}$ with $\|f-g\|_{0,\infty} \leq \varepsilon$ it holds that $g \in \Theta^c_{m_0+1}$.

(d) Suppose $\nabla^{m_0+1} f(x)$ exists for all $x \in \mathcal{D}$ but

$$\sup_{x \in \mathcal{D}} |\nabla^{m_0+1} f(x)| = \infty.$$

For example, this happens with $f(x) = \sqrt{x}$ when $\mathcal{D} = (0,1)$ and $m_0 = 0$. Then there exists a $\bar{x} \in \mathcal{D}$ such that

$$C < |\nabla^{m_0+1} f(\bar{x})| < \infty$$

for some constant $C > B$. Thus, for all $\|g\|_{m_0+1,\infty,\mathbb{1},\nu} \leq C$,

$$\begin{aligned}
|\nabla^{m_0+1} g(\bar{x})| &\geq |\nabla^{m_0+1} f(\bar{x})| - \left|\nabla^{m_0+1} g(\bar{x}) - \nabla^{m_0+1} f(\bar{x})\right| \\
&= |\nabla^{m_0+1} f(\bar{x})| - \left| \lim_{h \to 0} \frac{\nabla^{m_0} g(\bar{x}+h) - \nabla^{m_0} g(\bar{x})}{h} - \lim_{h \to 0} \frac{\nabla^{m_0} f(\bar{x}+h) - \nabla^{m_0} f(\bar{x})}{h} \right| \\
&= |\nabla^{m_0+1} f(\bar{x})| - \lim_{h \to 0} \left| \frac{\nabla^{m_0} g(\bar{x}+h) - \nabla^{m_0} f(\bar{x}+h)}{h} - \frac{\nabla^{m_0} g(\bar{x}) - \nabla^{m_0} f(\bar{x})}{h} \right|.
\end{aligned}$$

The first line follows by the reverse triangle inequality. Since the limit in the last line exists and is finite, for any $\delta > 0$, we can find an $\bar{h} > 0$ with $[\bar{x}, \bar{x}+\bar{h}] \subseteq \mathcal{D}$ such that the difference between the limit and the term we're taking the limit of evaluated at $\bar{h}$ is smaller than δ. Hence

$$\begin{aligned}
|\nabla^{m_0+1} g(\bar{x})| &\geq |\nabla^{m_0+1} f(\bar{x})| - \left| \frac{\nabla^{m_0} g(\bar{x}+\bar{h}) - \nabla^{m_0} f(\bar{x}+\bar{h})}{\bar{h}} - \frac{\nabla^{m_0} g(\bar{x}) - \nabla^{m_0} f(\bar{x})}{\bar{h}} \right| - \delta \\
&\geq C - \delta - \left| \frac{\nabla^{m_0} g(\bar{x}+\bar{h}) - \nabla^{m_0} f(\bar{x}+\bar{h})}{\bar{h}} - \frac{\nabla^{m_0} g(\bar{x}) - \nabla^{m_0} f(\bar{x})}{\bar{h}} \right|.
\end{aligned}$$

As in part (b), for any $\varepsilon_{m_0} > 0$, there is an $\varepsilon > 0$ such that $\|f-g\|_{0,\infty} \leq \varepsilon$ implies

$$\|\nabla^{m_0} f - \nabla^{m_0} g\|_{0,\infty} \leq \varepsilon_{m_0}.$$

Let ε_{m_0} such that

$$\left| \frac{\nabla^{m_0} g(\bar{x}+\bar{h}) - \nabla^{m_0} f(\bar{x}+\bar{h})}{\bar{h}} - \frac{\nabla^{m_0} g(\bar{x}) - \nabla^{m_0} f(\bar{x})}{\bar{h}} \right| \leq \delta.$$

Then

$$\begin{aligned}|\nabla^{m_0+1} g(\bar{x})| &\geq C - 2\delta \\ &> B\end{aligned}$$

where the last line follows if we choose $\delta > 0$ such that $C - 2\delta > B$, that is, $\delta < (C-B)/2$, which is possible since $C > B$. We have shown that the first piece of the Hölder norm $\|g\|_{m_0+1,\infty,\mathbb{1},\nu}$ is larger than B, and so the entire norm is larger than B and hence $g \in \Theta^c_{m_0+1}$.

(e) Finally, suppose

$$\sup_{x \in \mathcal{D}} |\nabla^{m_0+1} f(x)| \leq D < \infty$$

but $\nabla^{m_0+1} f$ is not Hölder continuous:

$$\sup_{x_1, x_2 \in \mathcal{D}, x_1 \neq x_2} \frac{|\nabla^{m_0+1} f(x_1) - \nabla^{m_0+1} f(x_2)|}{|x_1 - x_2|^\nu} = \infty.$$

Again take $g \in \mathscr{C}_{0,\infty}$ such that $\|f - g\|_{0,\infty} \leq \varepsilon$ and suppose that $\|g\|_{m_0+1,\infty,\mathbb{1},\nu} \leq C$ for $C > B$. Since $\nabla^{m_0+1} f$ is not Hölder continuous, there exist x_1 and x_2 in $\mathcal{D}$, $x_1 \neq x_2$, such that

$$\left| \frac{\nabla^{m_0+1} f(x_1) - \nabla^{m_0+1} f(x_2)}{|x_1 - x_2|^\nu} \right| > B + C.$$

Moreover, by the triangle inequality,

$$\begin{aligned}&\left| \frac{\nabla^{m_0+1} f(x_1) - \nabla^{m_0+1} f(x_2)}{|x_1 - x_2|^\nu} \right| \\ &\leq \left| \frac{\nabla^{m_0+1} g(x_1) - \nabla^{m_0+1} g(x_2)}{|x_1 - x_2|^\nu} \right| + \\ &+ \lim_{h \to 0} \left| \frac{(\nabla^{m_0} g(x_1 + h) - \nabla^{m_0} g(x_1)) - (\nabla^{m_0} f(x_1 + h) - \nabla^{m_0} f(x_1))}{h} \Big/ |x_1 - x_2|^\nu \right| \\ &+ \lim_{h \to 0} \left| \frac{(\nabla^{m_0} g(x_2 + h) - \nabla^{m_0} g(x_2)) - (\nabla^{m_0} f(x_2 + h) - \nabla^{m_0} f(x_2))}{h} \Big/ |x_1 - x_2|^\nu \right|.\end{aligned}$$

As in part (b), for any $\varepsilon_{m_0} > 0$, there is an $\varepsilon > 0$ such that $\|f - g\|_{0,\infty} \leq \varepsilon$ implies

$$\|\nabla^{m_0} f - \nabla^{m_0} g\|_{0,\infty} \leq \varepsilon_{m_0}.$$

Returning to our previous inequality, we see that since the limits on the right hand side are finite and since $\nabla^{m_0+1} g$ is Hölder continuous, for any $\delta > 0$ there is an $\bar{h} > 0$ which

does not depend on g such that

$$\begin{aligned}
&\left|\frac{\nabla^{m_0+1}f(x_1)-\nabla^{m_0+1}f(x_2)}{|x_1-x_2|^\nu}\right| \\
&\leq \left|\frac{\nabla^{m_0+1}g(x_1)-\nabla^{m_0+1}g(x_2)}{|x_1-x_2|^\nu}\right| \\
&+ \left|\frac{\left(\nabla^{m_0}g(x_1+\bar{h})-\nabla^{m_0}g(x_1)\right)-\left(\nabla^{m_0}f(x_1+\bar{h})-\nabla^{m_0}f(x_1)\right)}{\bar{h}}\Bigg/|x_1-x_2|^\nu\right| \\
&+ \left|\frac{\left(\nabla^{m_0}g(x_2+\bar{h})-\nabla^{m_0}g(x_2)\right)-\left(\nabla^{m_0}f(x_2+\bar{h})-\nabla^{m_0}f(x_2)\right)}{\bar{h}}\Bigg/|x_1-x_2|^\nu\right| + \delta \\
&\leq C + \frac{4\varepsilon_{m_0}}{\bar{h}|x_1-x_2|^\nu} + \delta.
\end{aligned}$$

This is the same argument we used in part (d). In the last line we used $\|g\|_{m_0+1,\infty,\mathbb{1},\nu} \leq C$, the triangle inequality, and $\|\nabla^{m_0}f - \nabla^{m_0}g\|_{0,\infty} \leq \varepsilon_{m_0}$. Choose $\delta = B/2$. Then choose ε_{m_0} small enough so that

$$\frac{4\varepsilon_0}{\bar{h}|x_1-x_2|^\nu} < \frac{B}{2}.$$

Combining our results, we have shown

$$C + B < \left|\frac{\nabla^{m_0+1}f(x_1)-\nabla^{m_0+1}f(x_2)}{|x_1-x_2|^\nu}\right| \leq C + B,$$

a contradiction.

□

Proof of theorem 4 (Closedness under equal weightings).

1. We want to show that the $\|\cdot\|_s = \|\cdot\|_{m+m_0,2,\mu_s}$-ball Θ is $\|\cdot\|_c = \|\cdot\|_{m,\infty,\mu_c^{1/2}}$-closed. Part 1 of our compact embedding result theorem 3 says that $\mathscr{W}_{m+m_0,2,\mu_s}$ is compactly embedded in $\mathscr{C}_{m,\infty,\mu_c^{1/2}}$. Now consider the space $(\mathscr{W}_{m,2,\mu_a}, \|\cdot\|_{m,2,\mu_a})$ where μ_a is such that

$$\int_{\mathbb{R}^{d_x}} \frac{\mu_a(x)}{\mu_c(x)}\, dx \leq C_1.$$

Then for any $f \in \mathscr{C}_{m,\infty,\mu_c^{1/2}}$,

$$\begin{aligned}
\|f\|^2_{m,2,\mu_a} &= \sum_{0\le|\lambda|\le m} \int_{\mathbb{R}^{d_x}} |\nabla^\lambda f(x)|^2 \mu_a(x)\, dx \\
&= \sum_{0\le|\lambda|\le m} \int_{\mathbb{R}^{d_x}} |\nabla^\lambda f(x)|^2 \mu_c(x) \frac{\mu_a(x)}{\mu_c(x)}\, dx \\
&\le C\|f\|^2_{m,\infty,\mu_c^{1/2}} \int_{\mathbb{R}^{d_x}} \frac{\mu_a(x)}{\mu_c(x)}\, dx \\
&\le CC_1 \|f\|_{m,\infty,\mu_c^{1/2}}.
\end{aligned}$$

Hence

$$\mathscr{C}_{m,\infty,\mu_c^{1/2}} \subseteq \mathscr{W}_{m,2,\mu_a}.$$

But we also know that $\mathscr{W}_{m+m_0,2,\mu_s}$ is compactly embedding in $\mathscr{C}_{m,\infty,\mu_c^{1/2}}$. Therefore, by lemma 4, $\mathscr{W}_{m+m_0,2,\mu_s}$ is compactly embedded in $\mathscr{W}_{m,2,\mu_a}$. Both of these are separable Hilbert spaces by arguments as in the proof of theorem 3.6 in Kufner (1980), which is analogous to Adams and Fournier (2003) theorem 3.6. Hence lemma A.1 of Santos (2012) implies that Θ is $\|\cdot\|_{m,2,\mu_a}$-closed. But now lemma 2 and the inequality $\|\cdot\|_{m,2,\mu_a} \le (CC_1)^{1/2}\|\cdot\|_{m,\infty,\mu_c^{1/2}}$ imply that Θ is $\|\cdot\|_{m,\infty,\mu_c^{1/2}}$-closed.

2. We want to show that the $\|\cdot\|_s = \|\cdot\|_{m+m_0,2,\mu_s}$-ball Θ is $\|\cdot\|_c = \|\cdot\|_{m,2,\mu_c}$-closed. Part 2 of our compact embedding result theorem 3 says that $\mathscr{W}_{m+m_0,2,\mu_s}$ is compactly embedded in $\mathscr{W}_{m,2,\mu_c}$. Both of these are separable Hilbert spaces, as discussed in the previous part. Hence lemma A.1. of Santos (2012) implies that Θ is $\|\cdot\|_{m,2,\mu_c}$-closed.

3. We want to show that the $\|\cdot\|_s = \|\cdot\|_{m+m_0,\infty,\mu_s}$-ball Θ is not $\|\cdot\|_c = \|\cdot\|_{m,\infty,\mu_c}$-closed. The same counterexample from the proof of part 3 of theorem 2 can be adapted here as well, by smoothly extending its domain definition to $\mathcal{D} = \mathbb{R}$.

4. We want to show that the $\|\cdot\|_s = \|\cdot\|_{m+m_0,\infty,\mu_s}$-ball Θ is not $\|\cdot\|_c = \|\cdot\|_{m,2,\mu_c}$-closed. As in the previous part, this can be shown by extending the same counterexample from theorem 2.

□

Proof of theorem 6 (Closedness under product weightings). **Cases 1 and 2**. This follows exactly as in the proof of theorem 5, except we apply theorem 4 and then lemma S1 part 2

Case 3. As in theorem 4, we can adapt the counterexample from theorem 2 by smoothly extending its domain to $\mathcal{D} = \mathbb{R}$.

Case 4. Assume $d_x = 1$ for simplicity. This proof is a close modification to the corresponding proof of theorem 2 for bounded domains. As in that proof, it suffices to prove the result for $m = 0$. For any $g \in \mathscr{C}_{m_0,\infty,\mu_s,\nu}$ define $g_s(x) = \mu_s(x)g(x)$ and $g_c(x) = \mu_c(x)g(x)$. We want to prove that

$$\Theta_{m_0} \equiv \{g \in \mathscr{C}_{m_0,\infty,\mu_s,\nu} : \|g\|_{m_0,\infty,\mu_s,\nu} \le B\}$$

is $\|\cdot\|_{m_0,\infty,\mu_c}$-closed, for all $m_0 \geq 0$. We proceed by induction on m_0.

Step 1 (Base Case): Let $m_0 = 0$. We want to show that Θ_0 is $\|\cdot\|_{0,\infty,\mu_c}$-closed, so we will show that its complement $\Theta_0^c = \mathscr{C}_{0,\infty,\mu_c} \setminus \Theta_0$ is $\|\cdot\|_{0,\infty,\mu_c}$-open. So take an arbitrary $f \in \Theta_0^c$. We will show that there exists an $\varepsilon > 0$ such that

$$\{g \in \mathscr{C}_{0,\infty,\mu_c} : \|f - g\|_{0,\infty,\mu_c} \leq \varepsilon\} \subseteq \Theta_0^c.$$

Since f is outside the weighted Hölder ball Θ_0, its weighted Hölder norm is larger than B,

$$\sup_{x\in\mathbb{R}} |f_s(x)| + \sup_{x_1,x_2\in\mathbb{R}} \frac{|f_s(x_1) - f_s(x_2)|}{|x_1 - x_2|^\nu} > B.$$

Hence there exist points $\bar{x}, \bar{x}_1, \bar{x}_2 \in \mathbb{R}$ with $\bar{x}_1 \neq \bar{x}_2$ such that

$$|f_s(\bar{x})| + \frac{|f_s(\bar{x}_1) - f_s(\bar{x}_2)|}{|\bar{x}_1 - \bar{x}_2|^\nu} > B.$$

Define

$$\delta = |f_s(\bar{x})| + \frac{|f_s(\bar{x}_1) - f_s(\bar{x}_2)|}{|\bar{x}_1 - \bar{x}_2|^\nu} - B > 0.$$

Next, for all $g \in \mathscr{C}_{0,\infty,\mu_c}$,

$$\begin{aligned}
\|g\|_{0,\infty,\mu_s,\nu} &\geq |g_s(\bar{x})| + \frac{|g_s(\bar{x}_1) - g_s(\bar{x}_2)|}{|\bar{x}_1 - \bar{x}_2|^\nu} \\
&\geq |f_s(\bar{x})| - |f_s(\bar{x}) - g_s(\bar{x})| \\
&\quad + \frac{|f_s(\bar{x}_1) - f_s(\bar{x}_2)|}{|\bar{x}_1 - \bar{x}_2|^\nu} - \frac{|(f_s(\bar{x}_1) - g_s(\bar{x}_1)) - (f_s(\bar{x}_2) - g_s(\bar{x}_2))|}{|\bar{x}_1 - \bar{x}_2|^\nu} \\
&= B + \delta - \left(|f_s(\bar{x}) - g_s(\bar{x})| + \frac{|(f_s(\bar{x}_1) - g_s(\bar{x}_1)) - (f_s(\bar{x}_2) - g_s(\bar{x}_2))|}{|\bar{x}_1 - \bar{x}_2|^\nu} \right) \\
&= B + \delta - \left(|f_c(\bar{x}) - g_c(\bar{x})| \frac{\mu_s(\bar{x})}{\mu_c(\bar{x})} + \frac{\left|(f_c(\bar{x}_1) - g_c(\bar{x}_1))\frac{\mu_s(\bar{x}_1)}{\mu_c(\bar{x}_1)} - (f_c(\bar{x}_2) - g_c(\bar{x}_2))\frac{\mu_s(\bar{x}_2)}{\mu_c(\bar{x}_2)}\right|}{|\bar{x}_1 - \bar{x}_2|^\nu} \right).
\end{aligned}$$

For all $g \in \mathscr{C}_{0,\infty,\mu_c}$ with

$$\|f - g\|_{0,\infty,\mu_c} = \|f_c - g_c\|_{0,\infty} \leq \varepsilon$$

we have

$$|f_c(\bar{x}) - g_c(\bar{x})| \frac{\mu_s(\bar{x})}{\mu_c(\bar{x})} + \frac{\left|(f_c(\bar{x}_1) - g_c(\bar{x}_1))\frac{\mu_s(\bar{x}_1)}{\mu_c(\bar{x}_1)} - (f_c(\bar{x}_2) - g_c(\bar{x}_2))\frac{\mu_s(\bar{x}_2)}{\mu_c(\bar{x}_2)}\right|}{|\bar{x}_1 - \bar{x}_2|^\nu} \leq \varepsilon \frac{\mu_s(\bar{x})}{\mu_c(\bar{x})} + \frac{\varepsilon\frac{\mu_s(\bar{x}_1)}{\mu_c(\bar{x}_1)} + \varepsilon\frac{\mu_s(\bar{x}_2)}{\mu_c(\bar{x}_2)}}{|\bar{x}_1 - \bar{x}_2|^\nu}$$

by the triangle inequality. So suppose we choose ε small enough that the right hand side is $\leq \delta/2$.

Then for all $g \in \mathscr{C}_{0,\infty,\mu_c}$ with $\|f-g\|_{0,\infty,\mu_c} \leq \varepsilon$ we have

$$\begin{aligned}\|g\|_{0,\infty,\mu_s,\nu} &\geq B + \delta - \frac{\delta}{2} \\ &> B.\end{aligned}$$

Hence $g \in \Theta_0^c$ for all such g. Thus Θ_0^c is $\|\cdot\|_{0,m,\mu_c}$-open and hence Θ_0 is $\|\cdot\|_{0,m,\mu_c}$-closed.

Step 2 (Induction Step): This step follows the same arguments as those with bounded support. As in step 1, the main idea is simply to replace g with either g_c or g_s, as appropriate. □

Proof of theorem 8 (Closedness for weighted norms on bounded domains). This proof is identical to the proof of theorem 6, except that now we use the compact embedding results of theorem 7 when necessary. □

E Proofs of propositions from section 4

Proof of proposition 1. This proof is straightforward and we therefore omit it. □

Proof of proposition 2. This proof is straightforward and we therefore omit it. □

Proof of proposition 3. This proof is given in Gallant and Nychka (1987) as lemma A.2, and hence we omit it. □

Proof of proposition 4. This proof is similar to the proof of proposition 3, which was shown in lemma A.2 of Gallant and Nychka (1987). Let $\mathcal{C} \subseteq \mathcal{D}$ be compact. We prove the proposition by induction on m (letting $m_0 = 0$, since it is irrelevant for the present result). For the base case, $m = 0$, the result holds trivially by letting $K_{\mathcal{C}} = 1$. Next suppose it holds for $m-1$. Choose λ such that $|\lambda| = m$ and let $\nabla^\lambda = \nabla^\beta \nabla^\alpha$ where $|\alpha| = 1$ and $|\beta| = m-1$. The result holds trivially if $\delta_s = 0$, so let $\delta_s \neq 0$. Then

$$\begin{aligned}\nabla^\lambda[\mu_s^{1/2}(x)] &= \nabla^\lambda \left[\exp\left(\frac{\delta_s}{2}(x'x)\right)\right] \\ &= \nabla^\beta \left(\nabla^\alpha \left[\exp\left(\frac{\delta_s}{2}(x'x)\right)\right]\right) \\ &= \nabla^\beta \left(\frac{\delta_s}{2}\exp\left(\frac{\delta_s}{2}(x'x)\right)\cdot \nabla^\alpha(x'x)\right) \\ &= \frac{\delta_s}{2}\sum_{\gamma\leq\beta}\begin{bmatrix}\beta\\ \gamma\end{bmatrix} \nabla^\gamma \left[\exp\left(\frac{\delta_s}{2}(x'x)\right)\right] \nabla^{\alpha+\beta-\gamma}(x'x) \\ &= \frac{\delta_s}{2}\sum_{\gamma\leq\beta}\begin{bmatrix}\beta\\ \gamma\end{bmatrix} [\nabla^\gamma \mu_s^{1/2}(x)]\nabla^{\alpha+\beta-\gamma}(x'x).\end{aligned}$$

In the fourth line we used Leibniz's formula. Next,

$$\begin{aligned}|\nabla^{\alpha+\beta-\gamma}(x'x)| &\leq \sum_{i=1}^{d_x}(x_i^2 + 2|x_i| + 2) \\ &\leq 4(1 + x'x).\end{aligned}$$

Hence

$$\begin{aligned}|\nabla^{\lambda}[\mu_s^{1/2}(x)]| &\leq \frac{|\delta_s|}{2}\sum_{\gamma\leq\beta}\begin{bmatrix}\beta\\ \gamma\end{bmatrix}|\nabla^{\gamma}\mu_s^{1/2}(x)|\cdot|4(1+x'x)| \\ &\leq 2|\delta_s|\sum_{\gamma\leq\beta}\begin{bmatrix}\beta\\ \gamma\end{bmatrix}K_{\mathcal{C},m-1}\mu_s^{1/2}(x)\cdot|1+x'x| \\ &\leq 2|\delta_s|\sum_{\gamma\leq\beta}\begin{bmatrix}\beta\\ \gamma\end{bmatrix}K_{\mathcal{C},m-1}\mu_s^{1/2}(x)\cdot M_{\mathcal{C}} \\ &= \mu_s^{1/2}(x)\left(2|\delta_s|\sum_{\gamma\leq\beta}\begin{bmatrix}\beta\\ \gamma\end{bmatrix}K_{\mathcal{C},m-1}\cdot M_{\mathcal{C}}\right).\end{aligned}$$

Here $M_{\mathcal{C}} = \sup_{x\in\mathcal{C}}|1 + x'x|$, which is finite since $\mathcal{C}$ is compact. The second line follows by the induction hypothesis. □

Proof of proposition 5. Pick $g(x) = 1 + x'x$. Notice that $g(x) \to \infty$ as $\|x\|_e \to \infty$. We prove the result by showing that for any $0 \leq |\lambda| \leq m_0$,

$$\nabla^{\lambda}\tilde{\mu}_c^{1/2}(x) = \exp\left[\frac{\delta_c}{2}(x'x)\right]\cdot p_{\lambda}(x) \qquad (*)$$

for some polynomial $p_\lambda(x)$. Consequently, dividing by $\mu_s^{1/2}(x)$ yields

$$\frac{\nabla^{\lambda}\tilde{\mu}_c^{1/2}(x)}{\mu_s^{1/2}(x)} = \exp\left[\frac{\delta_c - \delta_s}{2}(x'x)\right]\cdot p_{\lambda}(x).$$

Since $\delta_c < \delta_s$,

$$\left|\frac{\nabla^{\lambda}\tilde{\mu}_c^{1/2}(x)}{\mu_s^{1/2}(x)}\right|$$

converges to zero as $\|x\|_e \to \infty$. This implies there is a J such that for all x with $\|x\|_e > J$, this ratio is smaller than M_1. For all x with $\|x\|_e \leq J$, this ratio is a continuous function (the product of an exponential and a polynomial) on a compact set, and hence achieves a maximum M_2. Let $M = \max\{M_1, M_2\}$. Thus the ratio is bounded by M for all $x \in \mathbb{R}^{d_x}$.

So it suffices to show equation $(*)$. We proceed by induction. For the base case, $|\lambda| = 0$,

$$\begin{aligned}\nabla^0 \tilde{\mu}_c^{1/2}(x) &= \exp[\delta_c(x'x)/2] \cdot g(x) \\ &= \exp[\delta_c(x'x)/2] \cdot (1 + x^2).\end{aligned}$$

So the base case holds with $p_0(x) = g(x) = 1 + x^2$. Next, suppose it holds for $|\lambda| = m - 1$. Choose λ such that $|\lambda| = m$ and let $\nabla^\lambda = \nabla^\beta \nabla^\alpha$ where $|\alpha| = 1$ and $|\beta| = m - 1$. Then

$$\begin{aligned}\nabla^\lambda[\tilde{\mu}_c^{1/2}(x)] &= \nabla^\alpha[\nabla^\beta \tilde{\mu}_c^{1/2}(x)] \\ &= \nabla^\alpha[\exp[\delta_c(x'x)/2] \cdot p_\beta(x)] \\ &= \exp[\delta_c(x'x/2)](\delta_c/2)p_\beta(x)\nabla^\alpha(x'x) + \exp[\delta_c(x'x)/2]\nabla^\alpha p_\beta(x) \\ &= \exp[\delta_c(x'x)/2]\left((\delta_c/2)p_\beta(x)\nabla^\alpha(x'x) + \nabla^\alpha p_\beta(x)\right).\end{aligned}$$

Since the derivative of a polynomial is a polynomial, we're done. □

Proof of proposition 6.

1. This follows immediately from lemmas 5 and 7:

$$\|\mu^{1/2} f\|_{m,2} \leq M_1 \|f\|_{m,2,\mu} \leq M_1 M \|\mu^{1/2} f\|_{m,2}.$$

2. This follows immediately from lemmas 6 and 8.

□

F Proofs of propositions from section 5

Proof of proposition 7. Suppose such a function μ existed. Define $g : (0,1) \to \mathbb{R}$ by $g(x) = \log \mu(x)$. Then (1) implies that $g(x) \to -\infty$ as $x \to 0$. (2) implies that

$$g'(x) = \frac{1}{\mu(x)}\mu'(x) \leq K.$$

Hence $|g'(x)| \leq K$ for all $x \in (0,1)$. This is a contradiction to $g(x) \to -\infty$ as $x \to 0$. □

Proof of proposition 8. First consider the polynomial weight case, $\mu_s(x) = [x(1-x)]^{\delta_s}$. The proof is similar to the proof of propositions 3. We proceed by induction. For the base case $m = 0$, the result holds trivially by letting $K_{\mathcal{C}} = 1$. Next suppose it holds for $m - 1$. If $\delta_s = 0$ the result holds

trivially, so let $\delta_s \neq 0$. We have

$$\begin{aligned}
\nabla^m[\mu_s^{1/2}(x)] &= \nabla^m\left([x(1-x)]^{\delta_s/2}\right) \\
&= \nabla^{m-1}\nabla^1\left([x(1-x)]^{\delta_s/2}\right) \\
&= \nabla^{m-1}\left(\frac{\delta_s}{2}[x(1-x)]^{\delta_s/2-1}\nabla^1[x(1-x)]\right) \\
&= \frac{\delta_s}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}\nabla^\gamma\left([x(1-x)]^{\delta_s/2-1}\right)\nabla^{1+(m-1)-\gamma}[x(1-x)] \\
&= \frac{\delta_s}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}\nabla^\gamma\left(\mu_{s,\tilde{\delta}}^{1/2}(x)\right)\nabla^{m-\gamma}[x(1-x)].
\end{aligned}$$

Here $\tilde{\delta} = \delta_s - 1/2$. $\nabla^n[x(1-x)]$ is either $x - x^2$ for $n = 0$, $1 - 2x$ for $n = 1$, -2 for $n = 2$, and 0 for $n > 2$. Hence

$$\begin{aligned}
M_{\mathcal{C}} &\equiv \sup_{x\in\mathcal{C}}|\nabla^{m-\gamma}[x(1-x)]| \\
&< \infty
\end{aligned}$$

since $\mathcal{D}$ is bounded. So for all $x \in \mathcal{C}$,

$$\begin{aligned}
|\nabla^m[\mu_s^{1/2}(x)]| &\leq \frac{|\delta_s|}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}|\nabla^\gamma[\mu_{s,\tilde{\delta}}^{1/2}(x)]|\cdot|\nabla^{m-\gamma}[x(1-x)]| \\
&\leq \frac{|\delta_s|}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}K_{\mathcal{C},m-1}\mu_{s,\tilde{\delta}}^{1/2}(x)\cdot M_{\mathcal{C}} \\
&= \mu_{s,\tilde{\delta}}^{1/2}(x)\left(\frac{|\delta_s|}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}K_{\mathcal{C},m-1}M_{\mathcal{C}}\right) \\
&= [x(1-x)]^{\delta_s/2-1}\left(\frac{|\delta_s|}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}K_{\mathcal{C},m-1}M_{\mathcal{C}}\right) \\
&= \mu_s^{1/2}(x)\frac{1}{x(1-x)}\left(\frac{|\delta_s|}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}K_{\mathcal{C},m-1}M_{\mathcal{C}}\right) \\
&\leq \mu_s^{1/2}(x)M'_{\mathcal{C}}\left(\frac{|\delta_s|}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}K_{\mathcal{C},m-1}M_{\mathcal{C}}\right).
\end{aligned}$$

The second line follows by our $M_{\mathcal{C}}$ bound from above, and by the induction hypothesis with constant $K_{\mathcal{C},m-1}$. The last line follows since $\mathcal{C} \subseteq (0,1)$ is compact, and hence x is bounded away from zero and one. So

$$M'_{\mathcal{C}} \equiv \sup_{x\in\mathcal{C}}\frac{1}{x(1-x)} < \infty.$$

Next consider the exponential weight case, $\mu_s(x) = \exp[\delta_s x^{-1}(1-x)^{-1}]$. The proof for this case is similar to the proofs of propositions 3 and 4. Let $\mathcal{C} \subseteq \mathcal{D}$ be compact. We prove the proposition by induction on m (letting $m_0 = 0$, since it is irrelevant for the present result). For the base case, $m = 0$, the result holds trivially by letting $K_{\mathcal{C}} = 1$. Next suppose it holds for $m-1$. The result holds trivially if $\delta_s = 0$, so let $\delta_s \neq 0$. Then

$$\begin{aligned}
\nabla^m[\mu_s^{1/2}(x)] &= \nabla^m \left[\exp\left(\frac{\delta_s}{2}\frac{1}{x(1-x)}\right)\right] \\
&= \nabla^{m-1}\left(\nabla^1\left[\exp\left(\frac{\delta_s}{2}\frac{1}{x(1-x)}\right)\right]\right) \\
&= \nabla^{m-1}\left(\frac{\delta_s}{2}\exp\left(\frac{\delta_s}{2}\frac{1}{x(1-x)}\right)\cdot\nabla^1\left(\frac{1}{x(1-x)}\right)\right) \\
&= \frac{\delta_s}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}\nabla^\gamma\left[\exp\left(\frac{\delta_s}{2}\frac{1}{x(1-x)}\right)\right]\nabla^{1+(m-1)-\gamma}\left(\frac{1}{x(1-x)}\right) \\
&= \frac{\delta_s}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}[\nabla^\gamma\mu_s^{1/2}(x)]\nabla^{m-\gamma}\left(\frac{1}{x(1-x)}\right).
\end{aligned}$$

In the fourth line we used Leibniz's formula. Next, for any natural number n,

$$\nabla^n\left(\frac{1}{x(1-x)}\right) = n!\sum_{j=0}^{n}\frac{(-1)^{n-j}}{(1-x)^{j+1}x^{n+1-j}}.$$

Hence

$$\begin{aligned}
|\nabla^m[\mu_s^{1/2}(x)]| &\leq \frac{|\delta_s|}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}|\nabla^\gamma\mu_s^{1/2}(x)|\cdot\left|n!\sum_{j=0}^{n}\frac{(-1)^{n-j}}{(1-x)^{j+1}x^{n+1-j}}\right| \\
&\leq \frac{|\delta_s|}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}|\nabla^\gamma\mu_s^{1/2}(x)|\cdot M_{\mathcal{C}} \\
&\leq \frac{|\delta_s|}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}K_{\mathcal{C},m-1}\mu_s^{1/2}(x)\cdot M_{\mathcal{C}} \\
&= \mu_s^{1/2}(x)\left(\frac{|\delta_s|}{2}\sum_{\gamma\leq m-1}\begin{bmatrix} m-1 \\ \gamma \end{bmatrix}K_{\mathcal{C},m-1}\cdot M_{\mathcal{C}}\right).
\end{aligned}$$

Here

$$M_{\mathcal{C}} = \sup_{x\in\mathcal{C}}\left|n!\sum_{j=0}^{n}\frac{(-1)^{n-j}}{(1-x)^{j+1}x^{n+1-j}}\right|,$$

which is finite since $\mathcal{C} \subseteq (0,1)$ is compact, and hence x is bounded away from zero and one. The third line follows by the induction hypothesis. □

Proof of proposition 9. Let $g(x) = x^{-1}(1-x)^{-1}$. Then $g(x) \to \infty$ as $x \to 0$ or $x \to 1$. Note that

$\mathrm{Bd}(\overline{\mathcal{D}}) = \{0, 1\}$. The rest of the proof is similar to that of proposition 5. It suffices to show that for any $0 \leq |\lambda| \leq m_0$,

$$\nabla^\lambda \tilde{\mu}_c^{1/2}(x) = \mu_c(x) \cdot r_\lambda(x) \qquad (*)$$

for some rational function r_λ. Dividing $(*)$ by $\mu_s^{1/2}(x)$ yields

$$\frac{\nabla^\lambda \tilde{\mu}_c^{1/2}(x)}{\mu_s^{1/2}(x)} = \exp[(\delta_c - \delta_s)g(x)] \cdot r_\lambda(x).$$

Since $\delta_c < \delta_s$, the absolute value of this expression converges to zero as $x \to 0$ or 1. This proves part 2 of assumption 5. The proof of equation $(*)$ is as in the proof of 5: The base case holds immediately with $r_0(x) = g(x)$. The induction step follows since the derivative of a rational function is still rational. □

G Discussion of assumption 5

To get some intuition for assumption 5, consider the one dimensional case $d_x = 1$. In this case, we can usually take $m_0 = 1$, since $m_0 > d_x/2$ is then satisfied (see theorem 3 below). Then

$$\begin{aligned}\frac{|\nabla^0 \tilde{\mu}_c^{1/2}(x)|}{\mu_s^{1/2}(x)} &= \left|\frac{\nabla^0[\mu_c^{1/2}(x)g(x)]}{\mu_s^{1/2}(x)}\right| \\ &\leq \left(\frac{\mu_c(x)}{\mu_s(x)}\right)^{1/2} |g(x)|\end{aligned}$$

and

$$\begin{aligned}\frac{|\nabla^1 \tilde{\mu}_c^{1/2}(x)|}{\mu_s^{1/2}(x)} &= \left|\frac{\nabla^1[\mu_c^{1/2}(x)g(x)]}{\mu_s^{1/2}(x)}\right| \\ &= \left|\frac{\nabla^1 \mu_c^{1/2}(x)}{\mu_s^{1/2}(x)} g(x) + \frac{\mu_c^{1/2}(x)}{\mu_s^{1/2}(x)} \nabla^1 g(x)\right| \\ &\leq \frac{|\nabla^1 \mu_c^{1/2}(x)|}{\mu_s^{1/2}(x)} |g(x)| + \left(\frac{\mu_c(x)}{\mu_s(x)}\right)^{1/2} |\nabla^1 g(x)|.\end{aligned}$$

So when $d_x = 1$ with $m_0 = 1$, a sufficient condition for 5 is that there is a function g that diverges to infinity in the tails, but whose levels diverge slow enough that

$$|g(x)| = o\left(\left[\frac{\mu_c(x)}{\mu_s(x)}\right]^{-1/2}\right) \qquad \text{and} \qquad |g(x)| = o\left(\left[\frac{|\nabla^1 \mu_c^{1/2}(x)|}{\mu_s^{1/2}(x)}\right]^{-1}\right)$$

and whose first derivative also satisfies

$$|\nabla^1 g(x)| = o\left(\left[\frac{\mu_c(x)}{\mu_s(x)}\right]^{-1/2}\right).$$

For further intuition, suppose assumption 3 held for μ_c. Then for all $x \in \mathbb{R}^{d_x}$ and any $0 \leq |\lambda| \leq m_0$,

$$\begin{aligned}|\nabla^\lambda \mu_c^{1/2}(x)| &\leq K\mu_c^{1/2}(x) \\ &= K\left(\frac{\mu_c(x)}{\mu_s(x)}\right)^{1/2}\mu_s^{1/2}(x)\end{aligned}$$

and hence

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

Now suppose assumption 1 holds. Then the right hand side converges to zero as $\|x\|_e \to \infty$. Thus, in this special case, a sufficient condition for assumption 5 is that $|g(x)|$ and its derivative $|\nabla^1 g(x)|$ do not diverge faster than $\sqrt{\mu_c(x)/\mu_s(x)}$ converges to zero.

H Closure of differentiable functions

The following lemma shows that the Sobolev sup-norm closure of a Sobolev sup-norm (with more derivatives) ball is a Hölder space with exponent 1. We assume $d_x = 1$ for notational simplicity, but the result can be extended to $d_x > 1$.

Lemma S2. Let $\mathcal{D}$ be a convex open subset of $\mathbb{R}$. Let $m, m_0 \geq 0$ be integers. Define

$$\Theta_D = \{f \in \mathscr{C}_{m+m_0+1}(\mathcal{D}) : \|f\|_{m+m_0+1,\infty} \leq B\}$$

and

$$\Theta_L = \{f \in \mathscr{C}_{m+m_0}(\mathcal{D}) : \|f\|_{m+m_0,\infty,\mathbb{1},1} \leq B\}.$$

Let $\bar{\Theta}_D$ be the $\|\cdot\|_{m,\infty}$-closure of Θ_D. Then $\bar{\Theta}_D = \Theta_L$.

Proof. We prove equality by showing that $\bar{\Theta}_D \subseteq \Theta_L$ and $\Theta_L \subseteq \bar{\Theta}_D$.

1. ($\bar{\Theta}_D \subseteq \Theta_L$). Let $f \in \bar{\Theta}_D$. We will show that $f \in \Theta_L$. By the definition of the $\|\cdot\|_{m,\infty}$-closure, there exists a sequence $f_n \in \Theta_D$ such that

$$\|f_n - f\|_{m,\infty} \to 0.$$

Since $f_n \in \Theta_D$,

$$\|f_n\|_{m+m_0+1,\infty} = \max_{0\leq|\lambda|\leq m+m_0+1} \sup_{x\in\mathcal{D}} |\nabla^\lambda f_n(x)| \leq B.$$

Also notice that for all $x, y \in \mathcal{D}$,

$$\frac{|\nabla^{m+m_0} f_n(x) - \nabla^{m+m_0} f_n(y)|}{|x-y|} \leq |\nabla^{m+m_0+1} f_n(\tilde{x})| \leq \sup_{x \in \mathcal{D}} |\nabla^{m+m_0+1} f_n(x)|$$

where $\tilde{x}$ is between x and y, by the mean value theorem and convexity of $\mathcal{D}$. It follows that

$$\max_{|\lambda| \leq m+m_0} \sup_{x \in \mathcal{D}} |\nabla^\lambda f_n(x)| + \max_{|\lambda| = m+m_0} \sup_{x,y \in \mathcal{D}, x \neq y} \frac{|\nabla^\lambda f_n(x) - \nabla^\lambda f_n(y)|}{|x-y|} \leq \|f_n\|_{m+m_0+1,\infty} \leq B$$

and therefore $f_n \in \Theta_L$. But from part 5 of Theorem 2 we know that Θ_L is $\|\cdot\|_{m,\infty}$-closed and since $\|f_n - f\|_{m,\infty} \to 0$ it follows that $f \in \Theta_L$.

2. ($\Theta_L \subseteq \bar{\Theta}_D$) Let $f \in \Theta_L$. We will show that $f \in \bar{\Theta}_D$. Specifically, we will show how to $\|\cdot\|_{m,\infty}$-approximate f by a sequence of functions $\tilde{f}_n$ in Θ_D. Define

$$M_1 = \max_{|\lambda| \leq m+m_0} \sup_{x,y \in \mathcal{D}, x \neq y} \frac{|\nabla^\lambda f(x) - \nabla^\lambda f(y)|}{|x-y|} < \infty$$

and

$$M_2 = \sup_{x,y \in \mathcal{D}, x \neq y} \frac{|\nabla^{m+m_0} f(x) - \nabla^{m+m_0} f(y)|}{|x-y|} < \infty.$$

If $\mathcal{D} \neq \mathbb{R}$, then since $\nabla^{m+m_0} f$ is Lipschitz, the Kirszbraun theorem (e.g., theorem 6.1.1 on page 189 of Dudley 2002) allows us to extend $\nabla^{m+m_0} f$ to a function "$\nabla^{m+m_0} F$" on $\mathbb{R}$ with the same Lipschitz constant. Define F to be the $m+m_0$ times antiderivative of $\nabla^{m+m_0} F$. Then F is $(m+m_0)$-times differentiable, $\nabla^{m+m_0} F$ is Lipschitz with constant M_2, and $F|_{\mathcal{D}} = f$. In particular, for this extension F,

$$\max_{|\lambda| \leq m+m_0} \sup_{x,y \in \mathbb{R}, x \neq y} \frac{|\nabla^\lambda F(x) - \nabla^\lambda F(y)|}{|x-y|} = M_1$$

and

$$\sup_{x,y \in \mathbb{R}, x \neq y} \frac{|\nabla^{m+m_0} F(x) - \nabla^{m+m_0} F(y)|}{|x-y|} = M_2.$$

From here on we let $f(x) = F(x)$ denote the value of this extension of f if $x \notin \mathcal{D}$. The main issue is that f is only $(m+m_0)$-times differentiable, but we want to approximate it by functions that are just a little bit smoother—functions that are $(m+m_0+1)$-times differentiable. To do this, we convolve f with a smoother function:

$$f_n(x) = [f * \psi_{\varepsilon_n}](x) = \int_{\mathbb{R}} f(x + \varepsilon_n y) \psi(y)\, dy.$$

Here $*$ denotes convolution. ε_n is a sequence with $\varepsilon_n \to 0$ as $n \to \infty$. ψ_{ε_n} is an approximation to the identity: a function $\psi_{\varepsilon_n}(u) = \psi(u/\varepsilon_n)/\varepsilon_n$ where $\psi : \mathbb{R} \to \mathbb{R}$ is a $(m+m_0+1)$-times continuously differentiable function such that $\psi(y) \geq 0$ for all $y \in \mathbb{R}$, $\psi(y) = 0$ if $|y| \geq 1$, and

$\int_{-1}^{1} \psi(y)\, dy = 1$. For example,

$$\psi(y) = B_k(1-y^2)^k \mathbb{1}(|y| \leq 1).$$

where $k > m + m_0 + 1$ and B_k is such that the function integrates to 1. Note that f_n is $(m + m_0 + 1)$-times differentiable.

For all $\lambda \leq m + m_0$,

$$\begin{aligned}
[\nabla^\lambda f_n](x) &= [\nabla^\lambda f * \psi_{\varepsilon_n}](x) \\
&= \int_{\mathbb{R}} [\nabla^\lambda f](x-z) \frac{1}{\varepsilon_n} \psi\left(\frac{z}{\varepsilon_n}\right) dz \\
&= \int_{-1}^{1} [\nabla^\lambda f](x - \varepsilon_n y)\psi(y)\, dy.
\end{aligned}$$

The last line follows by a change of variables and since ψ is zero outside $[-1, 1]$. Hence

$$\begin{aligned}
|\nabla^\lambda f_n(x) - \nabla^\lambda f(x)| &\leq \int_{-1}^{1} |\nabla^\lambda f(x - \varepsilon_n y) - \nabla^\lambda f(x)| \psi(y)\, dy \\
&\leq \int_{-1}^{1} |M_1 \varepsilon_n y| \psi(y)\, dy \\
&= \varepsilon_n M_1 \int_{-1}^{1} |y| \psi(y)\, dy \\
&\equiv \delta_n
\end{aligned}$$

for all $\lambda \leq m + m_0$. The first line follows since ψ integrates to 1. Since $\delta_n \to 0$, it follows that

$$\|f_n - f\|_{m+m_0,\infty} \to 0.$$

Moreover,

$$\begin{aligned}
|\nabla^{m+m_0} f_n(x_1) - \nabla^{m+m_0} f_n(x_2)| &\leq \int |\nabla^{m+m_0} f(x_1 - \varepsilon_n y) - \nabla^{m+m_0} f(x_2 - \varepsilon_n y)| \psi(y)\, dy \\
&\leq M_2 |x_1 - x_2|.
\end{aligned}$$

Since f_n is $(m + m_0 + 1)$-times continuously differentiable,

$$|\nabla^{m+m_0+1} f_n(x)| = \lim_{h \to 0} \frac{|\nabla^{m+m_0} f_n(x+h) - \nabla^{m+m_0} f_n(x)|}{h} \leq M_2$$

for each $x \in \mathbb{R}$. Recall that

$$M_2 = \sup_{x,y \in \mathcal{D}, x \neq y} \frac{|\nabla^{m+m_0} f(x) - \nabla^{m+m_0} f(y)|}{|x - y|}.$$

This implies that

$$\begin{aligned}
\|f_n\|_{m+m_0+1,\infty} &\leq \|f_n\|_{m+m_0,\infty} + \sup_{x\in\mathcal{D}} |\nabla^{m+m_0+1} f_n(x)| \\
&\leq \|f\|_{m+m_0,\infty} + \|f_n - f\|_{m+m_0,\infty} + \sup_{x\in\mathcal{D}} |\nabla^{m+m_0+1} f_n(x)| \\
&\leq \|f\|_{m+m_0,\infty} + \delta_n + \sup_{x\in\mathcal{D}} |\nabla^{m+m_0+1} f_n(x)| \\
&\leq \left(\|f\|_{m+m_0,\infty} + \sup_{x,y\in\mathcal{D}, x\neq y} \frac{|\nabla^{m+m_0} f(x) - \nabla^{m+m_0} f(y)|}{|x-y|} \right) + \delta_n \\
&\leq B + \delta_n.
\end{aligned}$$

The last line follows since $f \in \Theta_L$. Thus f_n is *almost* in Θ_D, but not quite. But we can just rescale f_n to put it inside Θ_D: Let

$$\tilde{f}_n(x) = \frac{B}{B+\delta_n} f_n(x).$$

Then $\|\tilde{f}_n\|_{m+m_0+1,\infty} \leq B$ and so $\tilde{f}_n \in \Theta_D$. Moreover,

$$\begin{aligned}
\|\tilde{f}_n - f\|_{m,\infty} &\leq \|\tilde{f}_n - f\|_{m+m_0,\infty} \\
&\leq \|\tilde{f}_n - f_n\|_{m+m_0,\infty} + \|f_n - f\|_{m+m_0,\infty} \\
&= \max_{0\leq|\lambda|\leq m+m_0} \sup_{x\in\mathcal{D}} \left| \nabla^\lambda \left(\frac{B}{B+\delta_n} f_n(x) \right) - \nabla^\lambda f_n(x) \right| + \|f_n - f\|_{m+m_0,\infty} \\
&= \left| \frac{B}{B+\delta_n} - 1 \right| \|f_n\|_{m+m_0,\infty} + \|f_n - f\|_{m+m_0,\infty} \\
&= \frac{\delta_n}{B+\delta_n} \|f_n\|_{m+m_0,\infty} + \|f_n - f\|_{m+m_0,\infty}.
\end{aligned}$$

Since $\|f_n\|_{m+m_0,\infty} \leq \|f_n\|_{m+m_0+1,\infty} \leq B + \delta_n$,

$$\frac{\delta_n}{B+\delta_n} \|f_n\|_{m+m_0,\infty} \to 0.$$

We also know that $\|f_n - f\|_{m+m_0,\infty} \to 0$. It follows that

$$\|\tilde{f}_n - f\|_{m,\infty} \to 0.$$

But remember that $\tilde{f}_n \in \Theta_D$. So, by definition of the $\|\cdot\|_{m,\infty}$-closure, $f \in \bar{\Theta}_D$.

□

I Sup-norm convergence over closed domains $\mathcal{D}$

Throughout the paper we have focused on functions with open domains $\mathcal{D}$. In practice we may also be interested in functions with closed domains $\mathcal{D}$. First, note that convergence of a sequence of functions in a Sobolev L_p norm where the integral is taken over the interior of $\mathcal{D}$ implies convergence in the Sobolev L_p norm where the integral is taken over the entire $\mathcal{D}$. This follows since $\mathcal{D}$ is a subset of $\mathbb{R}^{d_x}$ and hence its boundary has measure zero. So the value of the integral is not affected by its values on the boundary. For Sobolev sup-norms, however, convergence over the interior of $\mathcal{D}$ does not automatically imply convergence over all of $\mathcal{D}$. In the following lemma, we illustrate how to do this extension for sequences from a Hölder ball which are known to converge in the ordinary sup-norm over the interior. Similar results can be obtained with different parameter spaces and for convergence in general Sobolev sup-norms.

Lemma S3. Let $\mathcal{D} \subseteq \mathbb{R}^{d_x}$ be closed and convex. Let $f_n : \mathcal{D} \to \mathbb{R}$ be a sequence of functions in

$$\Theta = \{f \in \mathscr{C}_0(\mathcal{D}) : \|f\|_{0,\infty,\mathbb{1},\nu} \leq B\}.$$

Suppose

$$\sup_{x \in \text{int}\mathcal{D}} |f_n(x) - f(x)| \to 0.$$

for some function f. Suppose f is continuous at each boundary point in $\mathcal{D}$. Then

$$\sup_{x \in \mathcal{D}} |f_n(x) - f(x)| \to 0.$$

Proof of lemma S3. We want to show that for any $\varepsilon > 0$, there is an N such that

$$|f_n(x) - f(x)| \leq \varepsilon$$

for all $n \geq N$, for all $x \in \mathcal{D}$. For each $x \in \mathcal{D}$, choose an element $z_x \in \text{int}\mathcal{D}$ such that $\|x - z_x\|_e^\nu \leq \varepsilon/(3B)$ and

$$|f(x) - f(z_x)| \leq \frac{\varepsilon}{3}.$$

This is possible since f is continuous on all of $\mathcal{D}$, including at boundary points, and by convexity of $\mathcal{D}$. By the triangle inequality,

$$\begin{aligned} |f_n(x) - f(x)| &= |f_n(x) - f(x) - f_n(z_x) + f_n(z_x) - f(z_x) + f(z)| \\ &\leq |f_n(x) - f_n(z_x)| + |f(x) - f(z_x)| + |f_n(z_x) - f(z_x)|. \end{aligned}$$

By the definition of this parameter space we have

$$\sup_{x \in \mathcal{D}} |f_n(x) - f_n(z_x)| \leq B\|x - z_x\|_e^\nu \leq \frac{\varepsilon}{3}.$$

By uniform convergence of f_n to f on the interior of $\mathcal{D}$, there is an N such that

$$|f_n(z_x) - f(z_x)| \leq \frac{\varepsilon}{3}$$

for all $n \geq N$. Thus we're done. □

J Proofs for section 6

Proof of proposition 10. We omit this proof because it is almost identical to the proof of lemma A1 in Newey and Powell (2003). □

Proof of proposition 11. We verify the conditions of proposition 10.

1. The parameter space is $\|\cdot\|_c$-compact by part 1 of theorems 3 and 4. Since the sieve space is a $\|\cdot\|_c$-closed subset of the $\|\cdot\|_c$-compact set Θ, it is also $\|\cdot\|_c$-compact.

2. Define $Q(g) = -\mathbb{E}((Y - g(X))^2)$. Then for $g_1, g_2 \in \Theta$,

$$\begin{aligned}
&|Q(g_1) - Q(g_2)| \\
&= \left|\mathbb{E}(g_2(X)^2 - g_1(X)^2) + \mathbb{E}(2Y(g_1(X) - g_2(X)))\right| \\
&\leq \left|\mathbb{E}(g_2(X)^2 - g_1(X)^2)\right| + |\mathbb{E}(2Y(g_1(X) - g_2(X)))| \\
&= |\mathbb{E}(g_2(X) - g_1(X))(g_2(X) + g_1(X))| + 2\,|\mathbb{E}(Y(g_1(X) - g_2(X)))| \\
&\leq \sqrt{\mathbb{E}\left((g_2(X) - g_1(X))^2\right)\mathbb{E}\left((g_2(X) + g_1(X))^2\right)} + 2\sqrt{\mathbb{E}\left(Y^2\right)\mathbb{E}\left((g_1(X) - g_2(X))^2\right)} \\
&\leq \sqrt{\mathbb{E}\left((g_2(X) - g_1(X))^2\right)\mathbb{E}\left(2g_2(X)^2 + 2g_1(X)^2\right)} + 2\sqrt{\mathbb{E}\left(Y^2\right)\mathbb{E}\left((g_1(X) - g_2(X))^2\right)}.
\end{aligned}$$

The fourth line follows from the Cauchy-Schwarz inequality and the last line from $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b, \in \mathbb{R}$. Next,

$$\mathbb{E}((g_1(X) - g_2(X))^2) \leq \left(\sup_{x\in\mathbb{R}} |g_1(x) - g_2(x)|\mu_c(x)\right)^2 \mathbb{E}(\mu_c(X)^{-2}) = \|g_1 - g_2\|_c^2 \cdot \mathbb{E}(\mu_c(X)^{-2}).$$

Moreover, for all $g \in \Theta$,

$$\begin{aligned}
\mathbb{E}(g(X)^2) &= \mathbb{E}(g(X)^2\mu_c(X)^2\mu_c(X)^{-2}) \\
&\leq \|g\|_c^2 \cdot \mathbb{E}(\mu_c(X)^{-2}) \\
&\leq C^2\|g\|_s^2 \cdot \mathbb{E}(\mu_c(X)^{-2}) \\
&\leq C^2B^2\mathbb{E}(\mu_c(X)^{-2}).
\end{aligned}$$

The third line follows by the compact embedding result, part 1 of theorem 3. Therefore

$$|Q(g_1) - Q(g_2)| \leq 2\left(BC\mathbb{E}(\mu_c(X)^{-2}) + \sqrt{\mathbb{E}\left(Y^2\right)\mathbb{E}\left(\mu_c(X)^{-2}\right)}\right)\|g_1 - g_2\|_c.$$

Since $\mathbb{E}(Y^2) < \infty$ and $\mathbb{E}(\mu_c(X)^{-2}) < \infty$, Q is $\|\cdot\|_c$-continuous. Similarly, let $\widehat{Q}_n(g) = -\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i))^2$. Identical arguments imply that

$$|\widehat{Q}_n(g_1) - \widehat{Q}_n(g_2)| \leq 2\left(BC\frac{1}{n}\sum_{i=1}^n \mu_c(X_i)^{-2} + \sqrt{\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)\left(\frac{1}{n}\sum_{i=1}^n \mu_c(X_i)^{-2}\right)}\right)\|g_1 - g_2\|_c.$$

Hence $\widehat{Q}$ is $\|\cdot\|_c$-continuous.

3. Suppose $Q(g) = Q(g_0)$. Then $\mathbb{E}((Y - g(X))^2) = \mathbb{E}((Y - g_0(X))^2)$, which implies that $g(X) = g_0(X)$ almost everywhere. If $g(\bar{x}) \neq g_0(\bar{x})$ for some $\bar{x}$, then $g(\bar{x}) \neq g_0(\bar{x})$ in a neighborhood of $\bar{x}$ by continuity of g_0, a contradiction. Hence $g(x) = g_0(x)$ for all $x \in \mathbb{R}$. Thus $\|g - g_0\|_c = \sup_{x\in\mathbb{R}} |g(x) - g_0(x)|\mu_c(x) = 0$. Moreover,

$$Q(g_0) = -\mathbb{E}((Y - g_0(X))^2) > -\mathbb{E}(2Y^2 + 2g_0(X)^2) > -\infty.$$

4. For any $g_k \in \Theta_k$

$$\|g_k - g_0\|_c \leq \sup_{|x|\leq M} |g_k(x) - g(x)| \sup_{|x|\leq M} \mu_c(x) + \sup_{|x|\geq M} |(g_k(x) - g(x))\mu_s(x)| \sup_{|x|\geq M} \frac{\mu_c(x)}{\mu_s(x)}.$$

Let $\varepsilon > 0$. Since g_k and g_0 are in Θ,

$$\sup_{|x|\geq M} |(g_k(x) - g(x))\mu_s(x)| \leq \|g_k - g\|_s \leq 2B.$$

Thus, since μ_c and μ_s satisfy assumption 1, we can choose M such that

$$\sup_{|x|\geq M} |(g_k(x) - g(x))\mu_s(x)| \sup_{|x|\geq M} \frac{\mu_c(x)}{\mu_s(x)} \leq \frac{\varepsilon}{2}.$$

By assumption, for a fixed M, we can pick k large enough to make $\sup_{|x|\leq M} |g_k(x) - g(x)|$ arbitrarily small. By μ_c^2 satisfying the integrability assumption 6′ and continuity of μ_c, $\sup_{|x|\leq M} \mu_c(x) < \infty$. Hence we can pick k large enough so that

$$\sup_{|x|\leq M} |g_k(x) - g(x)| \sup_{|x|\leq M} \mu_c(x) \leq \frac{\varepsilon}{2}.$$

Thus $\|g_k - g_0\|_c \leq \varepsilon$. Hence we have shown that $\|g_k - g_0\|_c \to 0$ as $k \to 0$.

5. For all $g \in \Theta_{k_n} \subseteq \Theta$,

$$(Y - g(X))^2 \leq 2Y^2 + g(X)^2 \leq 2Y^2 + 2B^2C^2\mu_c(X)^{-2}.$$

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

$$\mathbb{E}\left(\sup_{g\in\Theta}(Y - g(X))^2\right) < \infty.$$

Hence Jennrich's uniform law of large numbers implies that

$$\sup_{g\in\Theta_{k_n}} |\widehat{Q}_n(g) - Q(g)| \xrightarrow{p} 0.$$

□

Proof of proposition 12. The proof is similar to the one of proposition 11 and verifies the conditions of proposition 10.

1. This step is identical to the corresponding step in the proof of proposition 11.

2. Define $Q(g) = -\mathbb{E}((Y - g(X))^2\mu_c(X)^2)$. Then for $g_1, g_2 \in \Theta$,

$$\begin{aligned}
|Q(g_1) - Q(g_2)| &= \left|\mathbb{E}\left((g_2(X)^2 - g_1(X)^2)\mu_c(X)^2\right) + \mathbb{E}\left(2Y(g_1(X) - g_2(X))\mu_c(X)^2\right)\right| \\
&\leq \sqrt{\mathbb{E}\left((g_2(X) - g_1(X))^2\mu_c(X)^2\right)\mathbb{E}\left((g_2(X) + g_1(X))^2\mu_c(X)^2\right)} \\
&\quad + 2\sqrt{\mathbb{E}\left(Y^2\mu_c(X)^2\right)\mathbb{E}\left((g_1(X) - g_2(X))^2\mu_c(X)^2\right)} \\
&\leq \sqrt{\mathbb{E}\left((g_2(X) - g_1(X))^2\mu_c(X)^2\right)\mathbb{E}\left(2g_2(X)^2\mu_c(X)^2 + 2g_1(X)^2\mu_c(X)^2\right)} \\
&\quad + 2\sqrt{\mathbb{E}\left(Y^2\mu_c(X)^2\right)\mathbb{E}\left((g_1(X) - g_2(X))^2\mu_c(X)^2\right)}.
\end{aligned}$$

   Next,

$$\mathbb{E}\left((g_1(X) - g_2(X))^2\mu_c(X)^2\right) \leq \|g_1 - g_2\|_c^2.$$

   Moreover, for all $g \in \Theta$,

$$\mathbb{E}\left(g(X)^2\mu_c(X)^2\right) \leq B^2M_5^2.$$

   Therefore

$$|Q(g_1) - Q(g_2)| \leq 2\left(BM_5 + \sqrt{\mathbb{E}\left(Y^2\mu_c(X)^2\right)}\right)\|g_1 - g_2\|_c.$$

   Since $\mathbb{E}\left(Y^2\mu_c(X)^2\right) < \infty$, Q is continuous. Similarly, let $\widehat{Q}_n(g) = -\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i))^2\mu_c(X_i)^2$. Identical arguments imply that

$$|\widehat{Q}_n(g_1) - \widehat{Q}_n(g_2)| \leq 2\left(BM_5 + \sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2\mu_c(X_i)^2}\right)\|g_1 - g_2\|_c.$$

   Hence $\widehat{Q}$ is continuous.

3. As before, $\mathbb{E}((Y - g(X))^2\mu_c(X)^2) = \mathbb{E}((Y - g_0(X))^2\mu_c(X)^2)$ implies $g(X)\mu_c(X) = g_0(X)\mu_c(X)$ almost everywhere. If $g(\bar{x}) \neq g_0(\bar{x})$ for some $\bar{x}$, then $g(\bar{x}) \neq g_0(\bar{x})$ in a neighborhood of $\bar{x}$

by continuity of g_0. Moreover if $\mu_c(\bar{x}) > 0$, then $\mu_c(x) > 0$ with positive probability in a neighborhood of $\bar{x}$, which contradicts that $g(X)\mu_c(X) = g_0(X)\mu_c(X)$ almost everywhere. Thus, $g(\bar{x}) \neq g_0(\bar{x})$ implies $\mu_c(\bar{x}) = 0$. Therefore $\|g - g_0\|_c = 0$. Moreover,

$$Q(g_0) = -\mathbb{E}((Y - g_0(X))^2\mu_c(X)^2) > -\mathbb{E}(2Y^2\mu_c(X)^2 + 2g_0(X)^2\mu_c(X)^2) > -\infty.$$

4. This step is identical to the corresponding step in the proof of proposition 11.

5. For all $g \in \Theta_{k_n} \subseteq \Theta$,

$$(Y - g(X))^2\mu_c(X)^2 \leq 2Y^2\mu_c(X)^2 + 2g(X)^2\mu_c(X)^2 \leq 2Y^2\mu_c(X)^2 + 2B^2M_5^2.$$

This combined with $\mathbb{E}(Y^2\mu_c(X)^2) < \infty$ let us apply Jennrich's uniform law of large numbers, which gives

$$\sup_{\theta\in\Theta_{k_n}} |\widehat{Q}_n(\theta) - Q(\theta)| \xrightarrow{p} 0.$$

□

Proof of proposition 13. Let $g_{k_n} \in \widetilde{\Theta}_{k_n}$ such that $\|g_{k_n} - g_0\|_c \to 0$. Then $\|g_{k_n}\|_c \leq \|g_0\|_c + 1$ for n large enough. Moreover, $\|g_0\|_c \leq C\|g_0\|_s < \infty$. From the proof of proposition 12 we know that

$$|Q(g_{k_n}) - Q(g_0)| \leq 2\left(M_5(\|g_0\|_c + 1) + \sqrt{\mathbb{E}\left(Y^2\mu_c(X)^2\right)}\right)\|g_{k_n} - g_0\|_c$$

and

$$|\widehat{Q}_n(g_{k_n}) - \widehat{Q}_n(g_0)| \leq 2\left(M_5(\|g_0\|_c + 1) + \sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2\mu_c(X_i)^2}\right)\|g_{k_n} - g_0\|_c.$$

Now write

$$\widehat{Q}_n(g_{k_n}) - Q(g_{k_n}) = \left(\widehat{Q}_n(g_{k_n}) - \widehat{Q}_n(g_0)\right) + \left(\widehat{Q}_n(g_0) - Q(g_0)\right) + \left(Q(g_0) - Q(g_{k_n})\right).$$

$\widehat{Q}_n(g_0) - Q(g_0) = O_p(1/\sqrt{n})$ by the central limit theorem, which applies since $\mathbb{E}((Y - g_0(X))^4) < \infty$ and μ_c is uniformly bounded above. Thus,

$$\widehat{Q}_n(g_{k_n}) - Q(g_{k_n}) = O_p(\|g_{k_n} - g_0\|_c + 1/\sqrt{n}).$$

Since $\max\{1/\sqrt{n}, \|g_{k_n} - g_0\|_c\} = O(\lambda_n)$, lemma A.3 in Chen and Pouzo (2012) implies that for some $M_0 > 0$ it holds that $\|g_0\|_s \leq M_0$ and

$$\tilde{g}_w \in \{g \in \mathscr{W}_{1,2,\mu_s} : \|g\|_{1,2,\mu_s} \leq M_0\}$$

with probability arbitrarily close to 1 for all large n. Hence it suffices to prove that $\|\bar{g}_w - g_0\|_c \xrightarrow{p} 0$,

where

$$\bar{g}_w(x) = \operatorname*{argmax}_{g \in \tilde{\Theta}_{k_n}^{M_0}} -\left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i))^2 \mu_c(X_i)^2 + \lambda_n \|g\|_s \right)$$

and $\tilde{\Theta}_{k_n}^{M_0} = \{g \in \widetilde{\Theta}_{k_n} : \|g\|_s \leq M_0\}$.

Consistency now follows from proposition 12 under two additional arguments:

1. First, $\sup_{g \in \tilde{\Theta}_{k_n}^{M_0}} \lambda_n \|g\|_s \leq \lambda_n M_0 \to 0$ and therefore the sample objective function (including the penalty) still converges to Q uniformly over $g \in \tilde{\Theta}_{k_n}^{M_0}$.

2. Second, since $\tilde{\Theta}_{k_n}^{M_0}$ is finite dimensional, for any $g_1, g_2 \in \tilde{\Theta}_{k_n}^{M_0}$ there exists $D > 0$ such that $|\|g_1\|_s - \|g_2\|_s| \leq D|\|g_1\|_c - \|g_2\|_c| \leq D\|g_1 - g_2\|_c$. Hence the sample objective function (including the penalty) is still continuous on $\tilde{\Theta}_{k_n}^{M_0}$.

All other assumptions of proposition 10 hold using the same arguments as those in the proof of proposition 12. Thus $\|\bar{g}_w - g_0\|_c \xrightarrow{p} 0$ and hence $\|\tilde{g}_w - g_0\|_c \xrightarrow{p} 0$. □

Proof of proposition 14. The proof is adapted from the proof of theorem 4.3 in Newey and Powell (2003). Again we verify the conditions of proposition 10.

1. This step is identical to the corresponding step in the proof of proposition 11.

2a. Define $Q(g) = -\mathbb{E}(\mathbb{E}(Y - g(X) \mid Z)^2)$. For $g_1, g_2 \in \Theta$,

$$\begin{aligned}
&|\mathbb{E}(Y - g_1(X) \mid Z)^2 - \mathbb{E}(Y - g_2(X) \mid Z)^2| \\
&= |\mathbb{E}(2Y \mid Z)\mathbb{E}(g_2(X) - g_1(X) \mid Z) + \mathbb{E}(g_2(X) - g_1(X) \mid Z)\mathbb{E}(g_2(X) + g_1(X) \mid Z)| \\
&\leq |\mathbb{E}(2Y + g_2(X) + g_1(X) \mid Z)| \cdot |\mathbb{E}(g_2(X) - g_1(X) \mid Z)| \\
&= |\mathbb{E}((2g_0(X) + g_2(X) + g_1(X))\mu_c(X)\mu_c(X)^{-1} \mid Z)| \cdot |\mathbb{E}((g_2(X) - g_1(X))\mu_c(X)\mu_c(X)^{-1} \mid Z)| \\
&\leq 4BM_5|\mathbb{E}(\mu_c(X)^{-1} \mid Z)| \cdot M_5\|g_1 - g_2\|_c \cdot |\mathbb{E}(\mu_c(X)^{-1} \mid Z)| \\
&= 4BM_5^2\mathbb{E}(\mu_c(X)^{-1} \mid Z)^2\|g_1 - g_2\|_c \\
&\leq 4BM_5^2\mathbb{E}(\mu_c(X)^{-2} \mid Z)\|g_1 - g_2\|_c.
\end{aligned}$$

The fourth line uses $\mathbb{E}(U \mid Z) = 0$ and the last uses Jensen's inequality. Therefore

$$\begin{aligned}
|Q(g_1) - Q(g_2)| &\leq \mathbb{E}\left(|\mathbb{E}(Y - g_1(X) \mid Z)^2 - \mathbb{E}(Y - g_2(X) \mid Z)^2|\right) \\
&\leq 4BM_5^2\mathbb{E}(\mu_c(X)^{-2})\|g_1 - g_2\|_c.
\end{aligned}$$

Hence, Q is continuous.

2b. Let

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

Define P_Z as the $n \times k_n$ matrix with (i,j)th element $p_j(X_i)$. Let $Q_Z = P_Z(P_Z'P_Z)^- P_Z'$ where $(P_Z'P_Z)^-$ denotes the Moore-Penrose generalized inverse of $(P_Z'P_Z)$. Let Y and $g(X)$ be the $n \times 1$ vectors with elements Y_i and $g(X_i)$, respectively. Define $\widehat{Q}_n(g) = -\frac{1}{n}\|Q_Z(Y - g(X))\|^2$. Then for $g_1, g_2 \in \Theta$,

$$
\begin{aligned}
&|\widehat{Q}_n(g_1) - \widehat{Q}_n(g_2)| \\
&= \left| \frac{1}{n}\|Q_Z(Y - g_1(X))\|^2 - \frac{1}{n}\|Q_Z(Y - g_2(X))\|^2 \right| \\
&\leq \frac{1}{n}\|Q_Z(g_1(X) - g_2(X))\| \cdot \|Q_Z(2Y - g_1(X) - g_2(X))\| \\
&\leq \frac{1}{n}\|g_1(X) - g_2(X)\| \cdot \|2Y - g_1(X) - g_2(X)\| \\
&= \sqrt{\frac{1}{n}\sum_{i=1}^n (g_1(X_i) - g_2(X_i))^2 \mu_c(X_i)^2 \mu_c(X_i)^{-2}} \sqrt{\frac{1}{n}\sum_{i=1}^n (2Y_i - g_1(X_i) - g_2(X_i))^2} \\
&\leq \left(\sqrt{\frac{1}{n}\sum_{i=1}^n \mu_c(X_i)^{-2}} \sqrt{\frac{1}{n}\sum_{i=1}^n 4Y_i^2 + 4B^2 M_5^2 \mu_c(X_i)^{-2}} \right) \|g_1 - g_2\|_c.
\end{aligned}
$$

The second line follows because, by the Cauchy-Schwarz inequality,

$$|(a'a) - (b'b)| = |(a-b)'(a+b)| \leq \sqrt{(a-b)'(a-b)}\sqrt{(a+b)'(a+b)}$$

for all $a, b \in \mathbb{R}^n$. The third line follows because Q_Z is idempotent and thus $\|Q_Z b\| \leq \|b\|$ for all $b \in \mathbb{R}^n$. Hence $\widehat{Q}_n$ is continuous.

3. By completeness, $Q(g) = -\mathbb{E}(\mathbb{E}(Y - g(X) \mid Z)^2) = 0$ implies that $g(x) = g_0(x)$ almost everywhere. Identical arguments as those in the proof of proposition 11 then imply that $\|g - g_0\|_c = 0$, by continuity of g_0. Moreover,

$$Q(g_0) = -\mathbb{E}(\mathbb{E}(U \mid Z)^2) = 0 > -\infty.$$

4. Assumption 4 of proposition 10 holds using identical arguments as those in the proof of proposition 11.

5. Assumption 5 of proposition 10 requires convergence of $\widehat{Q}_n$ to Q uniformly over the sieve spaces. We show this by applying corollary 2.2 in Newey (1991). Θ is $\|\cdot\|_c$-compact, which is Newey's assumption 1. Q is $\|\cdot\|_c$-continuous, which is Newey's equicontinuity assumption. Next, define

$$B_n = \left(\sqrt{\frac{1}{n}\sum_{i=1}^n \mu_c(X_i)^{-2}} \sqrt{\frac{1}{n}\sum_{i=1}^n 4Y_i^2 + 4B^2 M_5^2 \mu_c(X_i)^{-2}} \right)$$

and recall that

$$|\widehat{Q}_n(g_1) - \widehat{Q}_n(g_2)| \leq B_n \|g_1 - g_2\|_c.$$

By Kolmogorov's strong law of large numbers and the existence of the relevant moments, $B_n = O_p(1)$. Hence Newey's assumption 3A holds. All that remains is to show Newey's assumption 2, pointwise convergence: $|\widehat{Q}(g) - Q(g)| = o_p(1)$ for all $g \in \Theta$. First write

$$\begin{aligned}|\widehat{Q}(g) - Q(g)| = &\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y - g(X) \mid Z = Z_i)^2 - \mathbb{E}\left(\mathbb{E}(Y - g(X) \mid Z)^2\right) \\ &+ \frac{1}{n}\sum_{i=1}^{n} \left(\widehat{\mathbb{E}}(Y - g(X) \mid Z = Z_i)^2 - \mathbb{E}(Y - g(X) \mid Z = Z_i)^2\right),\end{aligned}$$

where $\widehat{\mathbb{E}}(Y - g(X) \mid Z = Z_i)$ is the series estimator of the conditional expectation evaluated at Z_i. For the first part notice that $\mathbb{E}(Y - g(X) \mid Z = Z_i)^2$ is iid and

$$\begin{aligned}\mathbb{E}\left(\mathbb{E}(Y - g(X) \mid Z)^2\right) &\leq \mathbb{E}\left(\mathbb{E}((Y - g(X))^2 \mid Z)\right) \\ &\leq \mathbb{E}(2Y^2 + 2g(X)^2) \\ &\leq 2\mathbb{E}(Y^2) + 2\mathbb{E}(\mu_c(X)^{-1})\|g\|_c^2 \\ &< \infty.\end{aligned}$$

It follows from Kolmogorov's strong law of large numbers that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y - g(X) \mid Z = Z_i)^2 - \mathbb{E}\left(\mathbb{E}(Y - g(X) \mid Z)^2\right) \xrightarrow{p} 0.$$

Next, following Newey (1991), define ρ as the $n \times 1$ vector containing $Y_i - g(X_i)$ and h as the $n \times 1$ vector containing $\mathbb{E}(Y - g(X) \mid Z = Z_i)$. Then

$$\left|\frac{1}{n}\sum_{i=1}^{n} \left(\widehat{E}(Y - g(X) \mid Z = Z_i)^2 - \mathbb{E}(Y - g(X) \mid Z = Z_i)^2\right)\right| = \left|\|Q_Z\rho\|^2 - \|h\|^2\right|/n.$$

Since for all $a, b \in \mathbb{R}^n$ it holds that $a'a - b'b = (a-b)'(a-b) + 2b'(a-b)$,

$$\left|\|Q_Z\rho\|^2 - \|h\|^2\right|/n \leq \left(\|Q_Z\rho - h\|^2 + 2\|h\| \cdot \|Q_Z\rho - h\|\right)/n.$$

Since

$$\|h\|^2/n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y - g(X) \mid Z = Z_i)^2,$$

the previous arguments imply that $\|h\|^2/n = O_p(1)$. It therefore suffices to prove that $\|Q_Z\rho - h\|^2/n = o_p(1)$, which by Markov's inequality is implied by

$$\mathbb{E}\left(\|Q_Z\rho - h\|^2\right)/n \to 0.$$

as $n \to 0$. Newey (1991) shows

$$\mathbb{E}\left(\|Q_Z\rho - h\|^2\right)/n \leq \mathbb{E}\left(\operatorname{trace}(Q_Z \operatorname{var}(h \mid Z))\right)/n + o(1).$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left(\|Q_Z\rho - h\|^2\right)/n &\leq \mathbb{E}\left(\sum_{i=1}^n (Q_Z)_{ii} \operatorname{var}(Y_i - g(X_i) \mid Z_i)\right)/n + o(1) \\
&\leq \mathbb{E}\left(\sqrt{\frac{1}{n}\sum_{i=1}^n (Q_Z)_{ii}^2 \frac{1}{n}\sum_{i=1}^n \operatorname{var}(Y_i - g(X_i) \mid Z_i)^2}\right) + o(1) \\
&\leq \mathbb{E}\left(\sqrt{\frac{1}{n}\operatorname{trace}(Q_Z' Q_Z)\frac{1}{n}\sum_{i=1}^n \operatorname{var}(Y_i - g(X_i) \mid Z_i)^2}\right) + o(1) \\
&= \mathbb{E}\left(\sqrt{\frac{1}{n}\operatorname{trace}(Q_Z)\frac{1}{n}\sum_{i=1}^n \operatorname{var}(Y_i - g(X_i) \mid Z_i)^2}\right) + o(1) \\
&\leq \sqrt{\frac{k_n}{n}}\mathbb{E}\left(\sqrt{\frac{1}{n}\sum_{i=1}^n \operatorname{var}(Y_i - g(X_i) \mid Z_i)^2}\right) + o(1) \\
&\leq \sqrt{\frac{k_n}{n}}\sqrt{\mathbb{E}\left(\operatorname{var}(Y_i - g(X_i) \mid Z_i)^2\right)} + o(1).
\end{aligned}$$

The second line follows from the Cauchy-Schwarz inequality. The third line from the definition of the trace. The fourth line because Q_Z is idempotent. The fifth line because $\operatorname{trace}(Q_Z) \leq k_n$. The last line by Jensen's inequality. Since $\mathbb{E}\left(\left(\operatorname{var}(Y_i - g(X_i) \mid Z_i)\right)^2\right) < \infty$ and $k_n/n \to 0$, it follows that

$$\mathbb{E}\left(\|Q_Z\rho - h\|^2\right)/n \to 0.$$

□

References

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

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

Dudley, R. M. (2002): *Real analysis and probability*, vol. 74, Cambridge University Press.

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

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

Newey, W. K. (1991): "Uniform convergence in probability and stochastic equicontinuity," *Econometrica*, 1161–1167.

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.