cemmap

centre for microdata methods and practice

Towards a general large sample theory for regularized estimators

Michael Jansson
Demian Pouzo

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP63/19

Towards a General Large Sample Theory for Regularized Estimators *

Michael Jansson[†] and Demian Pouzo[‡]

November 18, 2019

Abstract

We present a general framework for studying regularized estimators; such estimators are pervasive in estimation problems wherein "plug-in" type estimators are either ill-defined or ill-behaved. Within this framework, we derive, under primitive conditions, consistency and a generalization of the asymptotic linearity property. We also provide data-driven methods for choosing tuning parameters that, under some conditions, achieve the aforementioned properties. We illustrate the scope of our approach by studying a wide range of applications, revisiting known results and deriving new ones.

Table of Contents

*We would like to thank Peter Bickel, Ivan Canay, Xiaohong Chen, Max Farrell, Eric Gautier, Joel Horowitz, Jack Porter as well as participants in numerous seminars and conferences for their helpful comments. Usual disclaimer applies.

†UC Berkeley, Dept. of Economics. E-mail: mjansson@berkeley.edu.

‡UC Berkeley, Dept. of Economics. E-mail: dpouzo@berkeley.edu (corresponding author).

1 Introduction

Most of econometrics and statistics is concerned with estimation or constructing inferential procedures like tests or confidence regions for some parameter of interest and relying on large sample theory for studying their properties. In general, one can think of a parameter as a mapping, ψ, from the probability distribution generating the data, P, to some parameter space, which could be of finite or infinite dimension depending on the application. A common method for estimating this parameter is to replace the unknown distribution P by an estimator of it like the empirical distribution of the data. This estimation technique is commonly known as "plug-in" estimation and encompasses many popular estimators like OLS, MLE, GMM and the Bootstrap, among others.

The complexity of modern datasets — especially since the advent of the so-called "big data revolution" — forces researchers to write more complex models (such as semi-/non-parametric models and high-dimensional models) to describe these datasets and to try to estimate more complex parameter of interest, e.g., a whole function like a density or a regression function or a high-dimensional vector. Unfortunately, in these cases it was noted as early as Stein [1956] that the mapping ψ defining the parameter of interest is typically ill-behaved or even ill-defined when evaluated at the empirical distribution. The widespread solution in these cases is to regularize the problem.

Regularization procedures are ubiquitous in economics and elsewhere. Examples of these include kernel-based estimators; series-based estimators and penalization-based estimators among many others.[1] Even though there has been an enormous amount of work in econometrics and other sciences studying the properties of these procedures, they are viewed, by and large, as separate and unrelated. In particular, results like consistency or large sample distribution theory, when they exists, they have only been derived in a case-by-case basis; to our knowledge, there is no general theory or systematic approach. The goal of this paper is to fill this gap by providing the basis for an unifying large sample theory for regularized estimators that will allow us to make systematic progress in studying their large sample properties.

Our point of departure is the general conceptual framework put forward by Bickel and Li (Bickel and Li [2006]), wherein the authors propose a general definition of regularization. According to their framework, a regularization can be viewed sequence of parameter mappings, $(\psi_k)_{k=1}^{\infty}$ that replaces the original parameter mapping, ψ, each element is well-behaved, and its limit coincides with the original mapping. The index of this sequence (denoted by k) represents what is often referred as the tuning (or regularization) parameter; e.g. it is the (inverse of the) bandwidth for kernels, the number of terms in a series expansion, or the (inverse of the) scale parameter in penalizations. We complement Bickel and Li's analysis, first, by providing additional examples, ranging from density estimation, bootstrapping parameters at the boundary and non-parametric IV regression. Second, and more importantly, we provide two set of general theorems that establish large sample properties for regularized estimators. One set of results establishes consistency and rate of convergence, and a data-driven method for choosing the tuning parameter that achieves these rates. Another set of results provide foundations for large sample distribution theory by deriving a generalization of the classical asymptotic linearity property.

Our approach to obtain consistency and convergence rate results is akin to the one used in the standard large sample theory for "plug-in" estimators, in the sense that it relies on continuity of the mapping used for estimation (see Wolfowitz [1957], Donoho and Liu [1991]). The key difference is that in plug-in estimation this mapping is ψ, but for regularized estimators the natural mapping is the (sequence of) *regularized* parameter mappings, $(\psi_k)_{k=1}^{\infty}$; this difference — in particular, the fact that we have a sequence of mappings — introduces nuances that are not present in the standard "plug-in" estimation case. We show that the key component of the convergence rate is the modulus of continuity of the regularized mapping, which, typically, will deteriorate as one moves further into the sequence of regularized mappings, thus yielding a generalized version of the well-known "noise-bias" trade-off. While

[1]Examples of regularizations are so ubiquitous that providing a thorough review is outside the scope of the paper; see e.g. Bickel and Li [2006], Bühlmann and Van De Geer [2011], Härdle and Linton [1994] and Chen [2007] for excellent reviews of several regularization methods.

this result, by itself, does not constitute a big leap from Bickel and Li's framework, we use the underlying insights to propose a data-driven method to choose the tuning parameter that under some conditions yields convergence rates proportional to the "oracle" ones, i.e., those implied by the choice that balances the "noise-bias" trade-off. This method is an extension of the Lepski method as presented in Pereverzev and Schock [2006] for ill-posed inverse problems.[2]

Our second set of results are concerned with obtaining a type of asymptotic linear representation for regularized estimators. The property of asymptotic linearity is well-known in the literature and is the cornerstone of large sample distribution theory. This property states that the estimator, once centered at the true parameter, is equal to a sample average of a mean zero function — referred as the influence function — plus an asymptotically negligible term.

In parametric models, asymptotic linearity is typically satisfied by commonly used estimators like the "plug-in" estimator. In more complex settings such semi-/non-parametric models, however, this is not longer true. In such cases, there are no estimators satisfying this property, because, for instance, the efficiency bound of the parameter of interest is infinite, or more generally, the parameter is not root-n estimable. For these situations, asymptotic representations analogous to asymptotic linearity have been obtained in specific examples for specific regularizations, but, to our knowledge, there is no general approach. This is specially problematic as there is no systematic method for properly standardizing the estimator in situations where the parameter is not root-n estimable.[3]

Our goal is to propose a systematic approach. For doing so, we introduce a generalization of asymptotic linearity that relaxes certain features of the standard property but still provides a useful asymptotic characterization of the estimator. The new property, which we call Generalized Asymptotic Linearity (GAL for short), relaxes the standard one in two dimensions: It allows for the location term to be different from the true parameter, and it

[2]Similar versions has been used in several particular applications. Closest to our examples are the work: by Pouzo [2015] for regularized M-estimators; by Chen and Christensen [2015] in non-parametric IV regressions; by Gine and Nickl [2008] for estimation of the integrated square density; by Gaillac and Gautier [2019] in a random coefficient model; by Lepski and Spokoiny [1997] for estimation of a function at a point.

[3]For density and regression estimation problems there is a large literature, especially for particular functionals like evaluation at a point; e.g. see Eggermont and LaRiccia [2001] Vol I and II for references and results. In more general contexts such as M-estimation and GMM-based models, to our knowledge, the literature is much more sparse with only a few papers allowing for slower than root-n parameters in particular settings. Closest to ours are the papers by Chen and Liao [2014], Chen et al. [2014] in the context of M-estimation models with series/sieve-based estimators; Newey [1994] in a two-stage moment model using kernel-based estimators; Chen and Pouzo [2015] in conditional moment models with sieve-based estimators; Cattaneo and Farrell [2013] in partitioning estimators of the conditional expectation function and its derivatives.

allows for this centering and the influence function to vary with the sample size. Each of these relaxations attempts to capture different nuances that commonly arise in the scattered set of existing results in the literature. Our results, which we now describe, will shed more light on the role and necessity of each.

We provide sufficient conditions for regularized estimators to satisfy GAL. Analogously to the theory of asymptotic linearity for plug-in estimators, our results rely on a notion of differentiability, but contrary to plug-in estimators, it relies on differentiability of each element in the sequence of regularized mappings, $(\psi_k)_{k=1}^{\infty}$ not on differentiability of the original mapping ψ.

As a consequence of this approach, GAL for regularized estimators exhibits two simplified features. First, the location term is given by $\psi_k(P)$ which can be interpreted as a psuedo-true parameter. The second simplified feature concerns the influence function and its dependence on the sample size. As in the location term, the dependence on the sample size of the influence function arises only through the dependence of the tuning parameter, k, on the sample size. Thus, the relevant object is a sequence of influence functions, each related to the derivative of the elements in $(\psi_k)_{k=1}^{\infty}$. We view this quantity as the natural departure from the traditional influence function as it is the sequence of regularized mappings, $(\psi_k)_{k=1}^{\infty}$, an not the original mapping, ψ, the one used for constructing the estimator. This last feature allows us to propose a natural and systematic way of standardizing the estimator regardless of whether root-n consistency holds. To explain this, we first note that in situations where asymptotic linearity holds, the proper standardization is given by square root of the sample size divided by the standard error of the value of the influence function. Under GAL the standarization of the regularized estimators turns out to be analogous except that in this case the influence function is indexed by the tuning parameter which at the same time depends on the sample size. Whether the standardization is root-n or slower depends on the behavior of the standard error of the value influence function as we move further into the sequence of regularized mappings (i.e., as k diverges).

The GAL property is established in cases where the parameter is of finite as well as of infinite dimension. While the former case is the prevalent one in most semi-parametric models, we show that the latter case is still of interest since it can be used as the basis for the construction of confidence bands for unknown functions.

Notation. The term "wpa1-P" is short for with probability approaching 1 under P, so for a generic sequence of IID random variables $(Z_n)_n$ with $Z_n \sim P$, the phrase "$Z_n \in A$ wpa1-P" formally means $P(Z_n \notin A) = o(1)$. For any random variables (X, Y) we use p_X and p_{XY} to denote the pdf (w.r.t. Lebesgue) corresponding to X and X, Y resp. For any linear normed spaces $(A, ||.||_A)$ and $(B, ||.||_B)$, let A^* be the dual of A, and for any continuous, homogeneous of degree 1 function $f : (A, ||.||_A) \mapsto (B, ||.||_B)$, $||f||_* = \sup_{a \in A:\, ||a||_A \neq 1} ||f(a)||_B$. For a Euclidean set S, we use $L^p(S)$ to denotes the set of L^p functions with respect to Lebesgue. For any other measure μ, we use $L^p(S, \mu)$ or $L^p(\mu)$. The norm $||.||$ denotes the Euclidean norm and when applied to matrices it corresponds to the operator norm. For any matrix A, let $e_{min}(A)$ denote the minimal eigenvalue. The symbol $\precsim$ denotes less or equal up to universal constants; $\succsim$ is defined analogously.

2 Examples

In this section we present several canonical examples which will be studied throughout the paper. We start, however, by presenting a classical example — density estimation at a point — so as to motivate and illustrate our results.

Example 2.1 (Density Evaluation)**.** The parameter of interest is the density function evaluated at a point, which can be formally viewed as a mapping from the space of probability distributions to $\mathbb{R}$, given by $P \mapsto \psi(P) = p(0)$, where p denotes the pdf of P. This mapping is only defined over the class of probabilities that admit a pdf for which $p(0)$ is well-defined and finite (e.g. continuous at 0); since the empirical probability distribution, P_n, does not belong to this class one cannot implement the standard "plug-in" estimator $\psi(P_n)$. To circumvent this shortcoming, the standard estimator used is given by $n^{-1} \sum_{i=1}^{n} \kappa_k(Z_i)$ where $\kappa_k(\cdot) = k\kappa(k\cdot)$, κ is a kernel (i.e., a smooth function over $\mathbb{R} \setminus \{0\}$ that ingrates to one); $1/k$ acts as the bandwidth of the kernel estimator. This estimator can be cast as $\psi_k(P_n)$ where

$$P \mapsto \psi_k(P) = (\kappa_k \star P)(0) \equiv \int_{\mathbb{R}} \kappa_k(z)\, P(dz),\ \forall k \in \mathbb{N}.$$

For any fixed k, ψ_k evaluated at P_n is well-defined and well-behaved; and as k diverges, $(\kappa_k \star P)(0)$ will approximate $p(0)$. That is, $(\psi_k)_{k \in \mathbb{N}}$ regularizes the parameter ψ. In Section 3 we provide a general definition of regularization that encompasses this case and many others and provides a conceptual framework to derive general asymptotic results.

Consistency. One such asymptotic result is consistency of the estimator, $\psi_{k(n)}(P_n)$, for some diverging sequence $(k(n))_{n \in \mathbb{N}}$. In this case, it is easy to see that it follows from ensuring that both the "sampling error" $|\psi_{k(n)}(P_n) - \psi_{k(n)}(P)|$ and the "approximation error" $|\psi_{k(n)}(P) - \psi(P)|$ vanish. While the convergence of the latter term follows directly from the construction of the regularization provided that $k(n) \to \infty$, convergence of the sampling error is more delicate. The main challenge stems from the fact that even though P_n is expected to converge to P as n diverges, the modulus of continuity of ψ_k increases as $k = k(n)$ diverges.

The desired result follows by having a good estimate of the modulus of continuity of the mapping of ψ_k as a function of k. In Section 4.1 we apply these ideas to postulate sufficient conditions for consistency and convergence rates for general regularizations.

For this example there are many ways of choosing the tuning parameter in a data-driven way (e.g. Härdle and Linton [1994] and reference therein). In order to obtain optimal (or close to optimal) rates of convergence, the choice has to balance the approximation and the sampling errors; see Birge and Massart [1998]. In Section 4.1 we present a data-driven way for choosing the tuning parameter inspired in the Lepski Method (e.g. Pereverzev and Schock [2006]) that achieves this balance and can be applied to a large class of regularizations.

Asymptotic Linearity. It is well-known that the parameter, $p(0) = \psi(P)$, is not root-n estimable and thus the proposed estimator is not asymptotically linear. The following representation, however, does hold:

$$\psi_k(P_n) - \psi_k(P) = n^{-1} \sum_{i=1}^{n} \{\kappa_k(Z_i) - E_P[\kappa_k(Z)]\}$$

which can be viewed as a generalization of asymptotic linearity in which the estimator $\psi_k(P_n)$ is centered at $\psi_k(P) = \int \kappa_k(z)p(z)dz$ instead of $\psi(P) = p(0)$. Moreover, by drawing an analogy with the standard approach for root-n estimable parameters (see Hampel et al. [2011], Bickel et al. [1998], Newey [1990]), for each k, the term in the curly brackets can be thought as an influence function. This term plays a crucial role on determining the asymptotic distribution of the estimator and on determining the proper way of standardizing it. For general regularized estimators, exact representations of this form are not always possible; however, in Section 5.2 we identify a class of regularizations — satisfying a certain differentiability notion (see Definition 5.2) — that admit, asymptotically, an analogous representation, with the influence function being a function of the derivative of the regularization; such representation can be viewed as a generalization of the asymptotic linearity.

It is well-known that the scaling is given by $\sqrt{n/k_n}$ which is slower (for some k_n that diverges with n) than the "standard" $\sqrt{n}$. The $\sqrt{k_n}$ correction arises because it is the correct order of the influence function, i.e., $\sqrt{Var_P\left(n^{-1/2}\sum_{i=1}^{n}\{\kappa_{k_n}(Z_i) - E_P[\kappa_{k_n}(Z)]\}\right)} = \sqrt{Var_P\left(\kappa_{k_n}(Z)\right)} \asymp \sqrt{k_n}$. Our results extend this simple observation to a large class of regularizations, thereby providing a canonical way for "standardizing" the estimator: By using $\sqrt{n}$ divided the standard deviation of the influence function, which it depends on n through the tuning parameter.

It is worth to point out that the feature of slower than root-n convergence rate is pervasive and it is shared by many regularized parameters, especially in semi-/non-parametric models. Our method can thus be viewed as extending the standard approach for root-n estimable parameters to a larger class of problems.

Our method, however, introduces a trade-off in which the choice of tuning parameter should ensure (generalized) asymptotic linearity, while also ensuring that the approximation error is small relative to the variability of the influence function. In Section 5.5 we present a general formalization of this trade-off and use the data-driven method discussed above to tackle it.

The result in Theorem 5.2 below focuses on asymptotic linear representations for finite dimensional parameters; e.g. vector-valued functionals of the density. Even though this representation is enough to cover this example and many others, in other instances the parameter of interest can be infinite dimensional. For instance, one could be interested on the asymptotic behavior of $\sup_z |(\kappa_k \star P)(z) - p(z))/\sigma_P(z)|$ for some $\sigma_P > 0$ (e.g. Bickel and Rosenblatt [1973]). To cover this case, the results in Theorem 5.2 must be extended since they rely on approximations that are valid *pointwise* on z, not *uniformly*. In Section 5.6 we extend our results to cover such cases, and, by using insights from dual spaces, show how the results can be used as the basis for constructing confidence bands for functions. △

We now provide additional examples that will be developed throughout the paper. These examples further illustrate the scope of our approach and additional nuances of the type of problems we can study.

Example 2.2 (Integrated Square Density)**.** Consider a similar setup as in example 2.1, but now the parameter of interest in this case is given by

$$P \mapsto \psi(P) = \int p(x)^2 dx.$$

This mapping is well-defined over probabilities with density in $L^2(\mathbb{R})$, but not when evaluated at the empirical probability distribution, P_n, since P_n does not have a density; it needs to be regularized. Bickel and Ritov [1990] showed that even though the efficiency bound is finite, no estimator converges at root-n rate; thereby illustrating that in some circumstances studying the local shape of ψ can be quite misleading. Our approach does not suffer from this criticism since it directly captures the (local) behavior of the estimator at hand. It is also general enough to encompass many of the proposed estimators in the literature, including "leave-one-out" types.

We believe our method complements the literature (e.g. Bickel and Ritov [1990], Bickel and Ritov [1988], Hall and Marron [1987] and Gine and Nickl [2008]) by providing a unifying framework that, among other things, allow us to better understand how certain aspects of the model/regularization affect the behavior of the estimator's convergence rate. Also, from a technical standpoint, this example illustrates how our method handles non-linear parameter mappings. △

Example 2.3 (Non-Parametric IV Regression (NPIV))**.** Consider the Non-Parametric IV Regression model characterized by

$$E[Y - h(W) \mid X] = 0, \tag{1}$$

where h is such that $E[|h(W)|^2] < \infty$, Y is the outcome variable, W is the endogenous regressor and X is the IV. It is well-known that the problem needs to be regularized; see Darolles et al. [2011], Hall and Horowitz [2005], Ai and Chen [2003, 2007b], Newey and Powell [2003], Florens [2003], Blundell et al. [2007] among others. We show how our method encompasses commonly used regularizations schemes such as sieves-based and penalized-based ones.

We focus on the case where the parameter of interest is a linear functional of h. For each regularization scheme, we derive the influence function of the regularization and show how its standard deviation can be used to appropriately scale the estimator to obtain a generalized asymptotic linear representation regardless of whether the parameter is root-n estimable or not. This last result, illustrates how our method can be used to generalize the approach proposed in Chen and Pouzo [2015] to general regularizations. As a by-product, we extend the results in Ackerberg et al. [2014] and link the influence function of the sieve-based regularization to simpler, fully parametric, *misspecified* GMM models. △

The next example is not really an example, it is rather a canonical estimation technique. It illustrates how our high level conditions translates to a particular estimation technique.

Example 2.4 (Regularized M-Estimators)**.** Given some model $\mathcal{M}$, the parameter mapping is defined as

$$\psi(P) = \arg\min_{\theta \in \Theta} E_P[\phi(Z, \theta)], \ \forall P \in \mathcal{M},$$

where Θ and $\phi : \mathbb{Z} \times \Theta \to \mathbb{R}_+$ are primitives of the problem and are such that the argmin is non-empty for any $P \in \mathcal{M}$. Many models of interest fit in this framework: High-dimensional linear and quantile regressions, non-parametric regression and likelihood-based models among others. In all of these cases, $\psi(P_n)$ is ill-defined or ill-behaved so it needs to be regularized.

We show how our results provide a general way of scaling the regularized estimator — even if the parameter is not root-n estimable —, and how they can be employed to get new limit theorems for confidence bands for general M-estimators, as well as a data-driven method to choose the tuning parameter. △

3 Setup

Let $\mathbb{Z} \subseteq \mathbb{R}^d$ and let $\boldsymbol{z} \equiv (z_1, z_2, ...) \in \mathbb{Z}^\infty$ denote a sequence of IID data drawn from some $P \in \mathcal{P}(\mathbb{Z}) \subset ca(\mathbb{Z})$, where $\mathcal{P}(\mathbb{Z})$ is the set of Borel probability measures over $\mathbb{Z}$ and $ca(\mathbb{Z})$ is the space of signed Borel measures of finite variation. For each $P \in \mathcal{P}(\mathbb{Z})$, let $\mathbf{P}$ be the induced probability over $\mathbb{Z}^\infty$. A **model** is defined as a subset of $\mathcal{P}(\mathbb{Z})$; and it will typically be denoted as $\mathcal{M}$.

Remark 3.1. Since we only consider IID random variables, it is enough to define a model as a family of probabilities over marginal probabilities. For richer data structures, one would have to define the model as a family of probabilities over $(Z_1, Z_2, ...)$. See Appendix A.2 for a discussion about how to extend our results to general stationary models. $\triangle$

A **parameter on model** $\mathcal{M}$ is a mapping $\psi : \mathcal{M} \to \Theta$ with $(\Theta, ||.||_\Theta)$ being a normed space.[4]

For the results in this paper, we need to endow $\mathcal{M}$ with some topology. For the results in Section 4.1 it suffices to work with a distance, d, under which the empirical distribution (defined below) converges to P. For the results in Section 5 and beyond, however, it is convenient to have more structure on the distance function, and thus, we work with a distance of the form

$$||P - Q||_{\mathcal{S}} \equiv \sup_{f \in \mathcal{S}} \left| \int f(z)P(dz) - \int f(z)Q(dz) \right|$$

where $\mathcal{S}$ is some class of Borel measurable and uniformly bounded functions (bounded by one). For instance, the total variation norm can be viewed as taking $\mathcal{S}$ as the class of indicator functions over Borel sets, and its denoted directly as $||.||_{TV}$; the weak topology over $\mathcal{P}(\mathbb{Z})$ is metricized by taking $\mathcal{S} = LB$ — the space of bounded Lipschitz functions — and its norm is denoted directly as $||.||_{LB}$; see van der Vaart and Wellner [1996] for a more thorough discussion.

3.1 Regularization

Let $\mathcal{D} \subseteq \mathcal{P}(\mathbb{Z})$ be the set of all discretely supported probability distributions. Let $P_n \in \mathcal{D}$ be the **empirical distribution**, where $P_n(A) = n^{-1} \sum_{i=1}^n 1\{Z_i \in A\}$ for any $A \subseteq \mathbb{Z}$.As illustrated by our examples, in many situations — especially in non-/semi-parametric models — the parameter mapping might be either ill-defined (e.g., if $P_n \notin \mathcal{M}$) or ill-behaved when evaluated at the empirical distribution P_n, so it has to be regularized.

[4] If the mapping does not point-identified an element of Θ, i.e., ψ is one-to-many, our results go through with minimal changes that account for the fact that $\psi(P)$ is a set in Θ.

The following definition of regularization is based on the first part of the definition in Bickel and Li [2006] p. 7. To state it, we define a **tuning set** as any subset of $\mathbb{R}_+$ that is unbounded from above, and the **approximation error** function as $k \mapsto B_k(P) \equiv ||\psi_k(P) - \psi(P)||_\Theta$.

Definition 3.1. *Given a model $\mathcal{M}$, a regularization of the parameter mapping ψ is a sequence $\boldsymbol{\psi} \equiv (\psi_k)_{k\in\mathbb{K}}$ such that $\mathbb{K}$ is a tuning set and*

1. *For any $k \in \mathbb{K}$, $\psi_k : \mathbb{D}_\psi \subseteq ca(\mathbb{Z}) \to \Theta$ where $\mathbb{D}_\psi \supseteq \mathcal{M} \cup \mathcal{D}$.*

2. *For any $P \in \mathcal{M}$, $\lim_{k\to\infty} B_k(P) = 0$.*

Condition 1 ensures that $\psi_k(P_n)$ and $\psi_k(P)$ are well-defined and that they are singletons for all $k \in \mathbb{K}$. Condition 2 ensures that, in the limit, the regularization approximates the original parameter mapping; the limit is warranted as the tuning set $\mathbb{K}$ is unbounded from above. In many applications the tuning set is given by $\mathbb{N}$ but there are applications such as kernel-based estimators, where it is more natural to use a (uncountable) subset of $\mathbb{R}_+$.

For each $k \in \mathbb{K}$, the implied estimator is given by $\psi_k(P_n)$ which — like the "plug-in" estimator — is permutation invariant. While, this restriction still encompasses a wide array of commonly used methods, it does rule out some estimation methods, notably those that rely on non-trivial sample-splitting procedures. We briefly discuss how to extend our framework to these cases in Appendix A.1.

Conditions 1 and 2 are not enough to obtain "nice" asymptotic properties of the regularized estimator such as consistency and asymptotic normality. In analogy to the standard asymptotic theory for "plug-in" estimators, these properties will be obtained by essentially imposing different degrees of smoothness on the regularization.

3.2 Examples (cont.)

The following examples illustrate that the Definition 3.1 encompasses a wide array of commonly used methods.

Example 3.1 (Integrated Square Density (cont.))**.** In this case $\Theta = \mathbb{R}$. The model is defined as the class of probability measures, P, with Lebesgue density, p, such that $p \in L^\infty(\mathbb{R})$ and

$$|p(x+t) - p(x)| \leq C(x)|t|^\varrho, \ \forall t, x \in \mathbb{R}, \tag{2}$$

with $C \in L^2(\mathbb{R})$ and $\varrho \in (0, 0.5)$. This restriction is rather mild and is similar to those used in the literature, e.g. Bickel and Ritov [1988], Hall and Marron [1987] and Powell et al. [1989].

We consider a class of regularizations given by

$$P \mapsto \psi_k(P) = \int (\kappa_k \star P)(x) P(dx), \ \forall k \in \mathbb{K}, \tag{3}$$

where κ is a kernel such that $\int |\kappa(u)||u|^{\varrho} du < \infty$, and $t \mapsto \kappa_k(t) \equiv k\kappa(kt)$. Thus, $1/k$ acts as the bandwidth for each $k \in \mathbb{K}$ which is a tuning set in $\mathbb{R}_{++}$.

Depending on the form of κ this regularization encompasses many estimators proposed in the literature. For instance, when $\kappa = \rho + \lambda(\rho - \rho \star \rho)$ with some $\lambda \in \mathbb{R}$ and some kernel ρ, and, for any $k > 0$, $z \mapsto \hat{p}_k(z) = \frac{1}{n}\sum_{i=1}^{n} \rho_k(Z_i - z)$, it follows that[5]

1. For $\lambda = 0$, the implied estimator is $n^{-1}\sum_{i=1}^{n} \hat{p}_k(Z_i) = n^{-2}\sum_{i,j} \rho_k(Z_i - Z_j)$.

2. For $\lambda = -1$, the implied estimator is $\int (\hat{p}_k(z))^2 dz = n^{-2}\sum_{i,j} (\rho \star \rho)_k(Z_i - Z_j)$.

3. For $\lambda = 1$, the implied estimator is $2n^{-1}\sum_{i=1}^{n} \hat{p}_k(Z_i) - \int (\hat{p}_k(z))^2 dz = n^{-2}\sum_{i,j}(2\rho_k - (\rho \star \rho)_k)(Z_i - Z_j)$.

The first two estimators are standard; the third estimator is inspired by the one considered in Newey et al. [2004], wherein κ is a twicing kernel. Moreover, the formalization in display (3) captures commonly used "leave-one-out" estimators by simply imposing $\kappa(0) = 0$. For instance, the "leave-one-out" versions of the estimators 1-3 are given by

1'. For $\lambda = 0$, the implied estimator is $n^{-2}\sum_{i \neq j} \rho_k(Z_i - Z_j)$.

2'. For $\lambda = -1$, the implied estimator is $n^{-2}\sum_{i \neq j} (\rho \star \rho)_k(Z_i - Z_j)$.

3'. For $\lambda = 1$, the implied estimator is $n^{-2}\sum_{i \neq j}(2\rho_k - (\rho \star \rho)_k)(Z_i - Z_j)$

These estimators are essentially the ones considered by Gine and Nickl [2008] and Hall and Marron [1987] (see also Powell and Stoker [1996] and references therein); the estimator 3' is also a somewhat simplified version of the one considered in Bickel and Ritov [1988].

Condition 1 in Definition 3.1 holds since we can set $\mathbb{D}_{\psi} = ca(\mathbb{R})$; Condition 2 of that definition follows from the next proposition, which establishes a bound for the approximation error.

Proposition 3.1. There exists a finite constant $C > 0$ such that for any $k \in \mathbb{K}$ and any $P \in \mathcal{M}$,

$$B_k(P) \leq C k^{-2\varrho} E_{|\kappa|}[|U|^{2\varrho}].$$

Proof. See Appendix B. □

[5] Details of the claims 1-3 and 1'-3' below are shown in the Appendix B.

△

Example 3.2 (NPIV (cont.))**.** For a given subspace of $L^2([0,1],p_W)$, Θ, the model $\mathcal{M}$ is defined as the class of probabilities over $Z=(Y,W,X)\in\mathbb{R}\times[0,1]^2$ with pdf with respect to Lebesgue, p, such that:[6] (1) $p_X = p_W = U(0,1)$, $E[|Y|^2]<\infty$ and $||p_{XW}||_{L^\infty}<\infty$; and (2) there exists a unique $h\in\Theta$ that satisfies 1. The restriction (1) can be relaxed and is made for simplicity so we can focus on the objects of interest that are h and P; it implies that $L^2([0,1],p_X)=L^2([0,1],p_W)=L^2([0,1])$ which simplifies the derivations.[7] The restriction (2) is what defines an IV non-parametric model. It implies that for any $P\in\mathcal{M}$, $r_P(\cdot)\equiv\int yP_{YX}(dy,\cdot)$ is well-defined and belongs to the range of the operator $T_P:\Theta\subseteq L^2([0,1])\to L^2([0,1])$ given by $T_P[h](\cdot)=\int h(w)p_{WX}(w,\cdot)dw$ for any $h\in L^2([0,1])$.[8] Thus, for any $P\in\mathcal{M}$, $\psi(P)$ is the (unique) solution of $r_P=T_P[h]$.

To illustrate our method, we consider the estimation of a linear functional of $\psi(P)$ of the form $\gamma(P)\equiv\int\pi(w)\psi(P)(w)dw$ for some $\pi\in L^2([0,1])$, which by the Riesz representation theorem covers any linear bounded functional on $L^2([0,1])$.

It is well-known that the estimation problem needs to be regularized. First, we need to regularize the "first stage parameters" — the operator T_P and r_P;second, given the regularization of T_P and r_P, the inverse problem for finding $\psi(P)$ typically needs to be regularized; e.g. when T_P is compact or when $\psi(P)$ is not a singleton.

By setting $\mathbb{K}=\mathbb{N}$, the regularization of the "first stage" is given by a sequence of mappings $(T_{k,P},r_{k,P})_{k\in\mathbb{N}}$ such that, for any $k\in\mathbb{N}$, $T_{k,P}:\Theta\to L^2([0,1])$ and $r_{k,P}\in L^2([0,1])$. The "second stage" regularization is summarized by an operator $\mathcal{R}_{k,P}:L^2([0,1])\to L^2([0,1])$ for which

$$\psi_k(P)=\mathcal{R}_{k,P}[T^*_{k,P}[r_{k,P}]],\ \forall P\in lin(\mathcal{M}\cup\mathcal{D}). \tag{4}$$

We assume that the regularization structure $(T_{k,P},r_{k,P},\mathcal{R}_{k,P})_{k\in\mathbb{N}}$ is such that: (1) $\lim_{k\to\infty}||\mathcal{R}_{k,P}[T^*_{k,P}[g]]-(T^*_PT_P)^{-1}T^*_P[g]||_{L^2([0,1])}=0$ pointwise over $g\in L^2([0,1])$; (2) $\lim_{k\to\infty}||\mathcal{R}_{k,P}[T^*_{k,P}[r_{k,P}-r_P]]||_{L^2([0,1])}=0$. We relegate a more thorough discussion and particular examples of the regularization to Appendix B.1. For now, it suffices to note that the first stage regularization encompasses commonly used regularizations such as the Kernel-based (e.g., Darolles et al. [2011], Hall and Horowitz [2005]) and the Series-Based (e.g., Ai and Chen [2003] and

[6]This restriction is mild and can be changed to accommodate discrete variables simply by requiring pdf's with respect to the counting measure.

[7]To restrict the support to $[0,1]$ is common in the literature (e.g. Hall and Horowitz [2005]). At this level of generality, one can always re-define h as $h\circ F_W^{-1}$ so that $p_W=U(0,1)$; of course this will affect the smoothness properties of h. The restriction $p_X=U(0,1)$ is really about p_X being known, since in that case, one can always take $F_X(X)$ as the instrument.

[8]Alternatively, we can define $T_P[h](X)=\int h(w)p(w|X)dw$ and $r_P(X)=\int yp(y|X)dy$. Depending on the type of the regularization one has at hand, it is more convenient to use one or the other.

Newey and Powell [2003]) regularizations, and the second stage regularization encompasses commonly used regularizations such as Tikhonov-/Penalization-based regularization (e.g., Darolles et al. [2011] and Hall and Horowitz [2005]) and Series-based regularization (e.g., Ai and Chen [2003] and Newey and Powell [2003]). For these combinations, conditions (1)-(2) haven been verified, under primitive conditions, in the literature; e.g. see Engl et al. [1996] Ch. 3-4.

It is easy to see that under conditions (1)-(2), the expression in 4 is in fact a regularization for $\psi(P)$ with $\mathbb{D}_{\psi} \supseteq \mathcal{M} \cup \mathcal{D}$ being a linear subspace specified in expression 19 in Appendix B.1. From this result, it also follows that $\{\gamma_k(P) \equiv \int \pi(w)\psi_k(P)(w)dw\}_{k\in\mathbb{N}}$ is a regularization for $\gamma(P)$ (in this case, $\Theta = \mathbb{R}$). $\triangle$

Example 3.3 (Regularized M-Estimators (cont.))**.** We impose the following assumptions over $(\mathcal{M}, \Theta, \phi)$: Θ is a subspace of L^q where $L^q \equiv L^q(\mathbb{Z}, \mu)$ for any $q \in [1, \infty)$ and some finite measure μ, and for $q = \infty$, $L^{\infty} = \mathbb{C}(\mathbb{Z}, \mathbb{R})$;[9] and $\theta \mapsto E_P[|\phi(Z,\theta)|]$ bounded and continuous, for all $P \in \mathcal{M}$.

The regularization is lifted from Pouzo [2015] and is defined using $\mathbb{K} = \mathbb{N}$ by: a sequence of nested linear subspaces of L^q, $(\Theta_k)_{k\in\mathbb{N}}$, such that $dim(\Theta_k) = k$ and the union is dense in Θ; a vanishing real-valued sequence $(\lambda_k)_{k\in\mathbb{N}}$ with $\lambda_k \in (0,1]$ and a lower-semi compact function $Pen : L^q \to \mathbb{R}_+$ such that, for each $k \in \mathbb{N}$[10]

$$\psi_k(P) \equiv \arg\min_{\theta\in\Theta_k} E_P[\phi(Z,\theta)] + \lambda_k Pen(\theta)$$

is a singleton for any $P \in \mathcal{M} \cup \mathcal{D}$.

It is clear that condition 1 in Definition 3.1 holds; we now show by contradiction that Condition 2 also holds. Suppose that there exists a $\epsilon > 0$ such that $||\psi_k(P) - \psi(P)||_{\Theta} \geq \epsilon$ for all k large. Let $\Pi_k\psi(P)$ be the projection of $\psi(P)$ onto Θ_k; for sufficiently large k, $||\Pi_k\psi(P) - \psi(P)||_{\Theta} \leq \epsilon$. Then, by optimality of $\psi_k(P)$ and some algebra, for large k, $\inf_{\theta\in\Theta:\,||\theta-\psi(P)||_{\Theta}\geq\epsilon} E_P[\phi(Z,\theta)] \leq E_P[\phi(Z,\psi(P))] - \{E_P[\phi(Z,\psi(P)) - \phi(Z,\Pi_k\psi(P))] + \lambda_k Pen(\Pi_k\psi(P))\}$. By continuity of $E_P[\phi(Z,\cdot)]$, $\lambda_k \downarrow 0$ and convergence of $\Pi_k\psi(P)$ to $\psi(P)$ the term in the curly bracket vanishes as k diverges, leading to the contradiction $\inf_{\theta\in\Theta:\,||\theta-\psi(P)||_{\Theta}\geq\epsilon} E_P[\phi(Z,\theta)] \leq E_P[\phi(Z,\psi(P))]$. $\triangle$

4 Continuous Regularizations

The results in this section extend the program started by Wolfowitz [1957] to regularized estimators with *continuous* regularizations. For such regularizations, it also presents a data

[9] The class $\mathbb{C}(\mathbb{Z}, \mathbb{R})$ is the class of continuous and uniformly bounded real-valued functions on $\mathbb{Z}$.
[10] A lower-semi compact function is one with compact lower contour sets.

driven method for choosing tuning parameters that yields consistent estimators as well as providing an explicit rate of convergence.

We say a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ is a modulus of continuity if f is continuous, non-decreasing and such that $f(t) = 0$ iff $t = 0$.

Definition 4.1 (Continuous Regularization). *A regularization $\boldsymbol{\psi}$ of ψ is continuous at $P \in \mathbb{D}_\psi$ with respect to d, if there exists a family of modulus of continuity $(\delta_k)_{k\in\mathbb{R}}$ such that for any $k \in \mathbb{K}$*

$$||\psi_k(P') - \psi_k(P)||_\Theta \leq \delta_k(d(P', P)) \tag{5}$$

for any $P' \in \mathbb{D}_\psi$.

The definition is equivalent to the standard "δ/ϵ"-definition of continuity because the modulus of continuity of ψ_k, δ_k, can converge to 0 arbitrary slowly. Moreover, the definition does not impose any *uniform* bounds on δ_k across different $k \in \mathbb{K}$. While such restriction would simplify the proofs considerably, it is too strong for many applications. Recall that the regularization is introduced precisely due to the poor behavior of ψ at P.

4.1 Consistency of Regularized Estimators

By imposing continuity with respect to a metric that ensures convergence of the empirical distribution P_n to P, the definition 4.1 readily implies consistency of $\psi_k(P_n)$ to $\psi_k(P)$ for any *fixed* $k \in \mathbb{N}$. However, in view of Condition 2 of that definition, unless there exists a $k \in \mathbb{K}$ such that $\psi_k(P) = \psi(P)$ this result is of limited interest. In order to guarantee consistency to $\psi(P)$ in general cases, k must be allowed to depend on n. To this end, the next lemma provides a uniform-in-k version of the aforementioned consistency results.

Lemma 4.1. *Suppose a regularization, $\boldsymbol{\psi}$, is continuous (at P) with respect to d and there exists a real-valued positive sequence $(r_n)_{n\in\mathbb{N}}$ such that $d(P_n, P) = O_P(r_n^{-1})$. Then, for any $\epsilon > 0$, there exists a $M > 0$ such that*

$$\limsup_{n\to\infty} \sup_{k\in\mathbb{K}} \mathbf{P}\left(||\psi_k(P_n) - \psi_k(P)||_\Theta \geq \delta_k(M r_n^{-1})\right) = 0$$

Proof. See Appendix C. □

The lemma implies that for any diverging sequence $(k_n)_{n\in\mathbb{N}}$ in $\mathbb{K}$, $||\psi_{k_n}(P_n) - \psi_{k_n}(P)||_\Theta$ is bounded by $\delta_{k_n}(M r_n^{-1})$ wpa1-P. This fact and the definition of regularization deliver consistency of $(\psi_{k_n}(P_n))_{n\in\mathbb{N}}$, provided that $\lim_{n\to\infty} \delta_{k_n}(M r_n^{-1}) = 0$. It turns out that under the assumption that if P_n consistently estimates P under d, then this claim follows by continuity of $t \mapsto \delta_k(t)$ for each $k \in \mathbb{K}$ and a simple diagonalization argument. The following theorem formalizes this idea.

Theorem 4.1 (Consistency of Regularized Estimators)**.** Suppose a regularization, $\boldsymbol{\psi}$, is continuous (at P) with respect to d such that $d(P_n, P) = o_P(1)$. Then there exists a $(k_n)_{n\in\mathbb{N}}$ in $\mathbb{K}$ such that

$$B_{k_n}(P) = o(1) \textit{ and } \delta_{k_n}(d(P_n, P)) = o_P(1),$$

and

$$||\psi_{k_n}(P_n) - \psi(P)||_\Theta = o_P(1).$$

Proof. See Appendix C. □

By the triangle inequality it is easy to see that the distance between the regularized estimator, $\psi_{k_n}(P_n)$, and the true parameter is bounded by the sum of two terms: the "sampling error", $\delta_{k_n}(d(P_n, P))$ and the "approximation error", $B_{k_n}(P)$, which generalizes the well-known "noise-bias" trade-off present in many applications.

4.2 Data-driven Choice of Tuning Parameter

Theorem 4.1 establish existence of a tuning parameter sequence that yields a consistent estimator, but it is silent about how to choose such sequence and what is the associated convergence rate. We now turn to these questions. In order to do this, let $\mathcal{G}_n$ be the (user-specified) set over which the tuning parameter is chosen; it is assumed to be a finite subset of $\mathbb{K}$.[11]

Given a rate of convergence of P_n to P under d, i.e., a positive real-valued sequence $(r_n)_{n\in\mathbb{N}}$ such that $d(P_n, P) = o_P(r_n^{-1})$ (observe the $o_P(.)$ as opposed to the $O_p(.)$; remark 4.1 below discusses the reason behind this choice). Theorem 4.1 suggests a criterion to construct tuning sequence $(k_n)_n$ that yields a consistent estimators: Choose a *diverging* sequence $(k_n)_n$ such that $\delta_{k_n}(r_n^{-1}) = o(1)$. For any $n \in \mathbb{N}$, an example of a choice that satisfies this criterion is what we call the *oracle choice over* $\mathcal{G}_n$,

$$\arg\min_{k\in\mathcal{G}_n}\{\delta_k(r_n^{-1}) + B_k(P)\},$$

which minimizes the trade-off between the approximation and the sampling errors. This choice represents commonly used heuristics and it is a good prescription to obtain approximately optimal rate of convergences (Birge and Massart [1998]). However, often times it is unfeasible, since it relies on knowledge of the approximation error, which is typically unknown because it depends on features of the unknown P.

[11] In Appendix D.3 we extend the main theorem of this section to the case where $\mathcal{G}_n$ is any closed set of $\mathbb{K}$, not necessarily finite.

It is thus desirable to construct a choice of tuning parameter that sidestep this issue while still providing approximately the same rates of convergence. We show that an adaptation of the Lepski method (e.g. Pereverzev and Schock [2006]) provides a data-driven choice that satisfies these properties. Due to the nature of the Lepski method, in order to establish the desired results we need monotonicity of the sampling and approximation errors as functions of the tuning parameter. Since these functions may not be monotonic, we replace them by monotonic majorants. Formally, let $k \mapsto \bar{B}_k(P)$ be a non-increasing function from $\mathbb{R}_+$ to itself such that $\bar{B}_k(P) \geq ||\psi_k(P) - \psi(P)||_\Theta$ for all $k \geq 0$, $\lim_{k\to\infty} \bar{B}_k(P) = 0$, and, for each $n \in \mathbb{N}$, let $k \mapsto \bar{\delta}_k(r_n^{-1})$ be a non-decreasing function from $\mathbb{R}_+$ to itself such that $\bar{\delta}_k(r_n^{-1}) \geq \delta_k(r_n^{-1})$.

In order to construct the data-driven choice for each $n \in \mathbb{N}$, we first define the following correspondence

$$(a_k)_k \mapsto \mathcal{L}_n((a_k)_k) \equiv \{k \in \mathcal{G}_n \colon k \in \mathcal{G}_n \colon ||\psi_k(P_n) - \psi_{k'}(P_n)||_\Theta \leq a_{k'},\ \forall k' \geq k \ in\ \mathcal{G}_n\} \quad (6)$$

where $(a_k)_k$ is a "test sequence" that defines the subset $\mathcal{L}_n((a_k)_k)$ of the grid $\mathcal{G}_n$ that can be chosen as our tuning parameter. For our results, the relevant sequence is given by $(4\bar{\delta}_k(r_n^{-1}))_k$ and the data-driven choice of tuning parameter is given by the minimal element of $\mathcal{L}_n \equiv \mathcal{L}_n((4\bar{\delta}_k(r_n^{-1}))_k)$, i.e., $\tilde{k}_n = \min\{k \colon k \in \mathcal{L}_n\}$ a.s.-$\mathbf{P}$.

Theorem 4.2. Suppose a regularization, $\boldsymbol{\psi}$, is continuous (at P) with respect to d and there exists a real-valued positive diverging sequence $(r_n)_{n\in\mathbb{N}}$ such that $d(P_n, P) = o_P(r_n^{-1})$. Then

$$||\psi_{\tilde{k}_n}(P_n) - \psi(P)||_\Theta = O_P\left(\inf_{k\in\mathcal{G}_n}\{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\}\right).$$

Proof. See Appendix D.1. □

Remark 4.1. The rate $(r_n)_n$ is defined as $d(P_n, P) = o_P(r_n^{-1})$, as opposed to $d(P_n, P) = O_P(r_n^{-1})$ as in Theorem 4.1. That is, r_n diverges (arbitrary) slower than the usual rates for P_n — which is typically given by $\sqrt{n}$ in our context. This type of lost is common when studying choice of tuning parameters (cf. Gine and Nickl [2008] and references therein). In our setup, it stems from the following fact: Take a rate $(s_n)_n$ such that $d(P_n, P) = O_P(s_n^{-1})$. For this rate, there are unknown constants (e.g. M in Lemma 4.1) which will render our data-driven choice infeasible. So to avoid them it suffices to replace s_n^{-1} by a (arbitrary) slower rate, e.g. $r_n^{-1} = \log(1+n)s_n^{-1}$ or $r_n^{-1} = \log(\log(1+n))s_n^{-1}$. △

Remark 4.2. The rate of convergence does not depend on the "complexity" of the set $\mathcal{G}_n$. This result stems from a certain "separability" property of the estimator: The probability statements stem from the behavior of $d(P_n, P)$ which does not depend on k nor in $\mathcal{G}_n$, the tuning parameter k only appear through the topological properties of the regularization. △

Remark 4.3 (Heuristics of the proof of Theorem 4.2)**.** Heuristically, for any $k \in \mathcal{G}_n$ that is larger or equal than $\tilde{k}_n$ it follows that $||\psi_{\tilde{k}_n}(P_n) - \psi(P)||_\Theta$ is bounded above (up to constants) by $\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)$ with probability approaching one. Lemma D.1 in Appendix D.1 formalizes this observation and shows that in order to establish the claim of the theorem it suffices to show existence of a tuning parameter in $\mathcal{G}_n$ that is larger or equal than $\tilde{k}_n$ (with probability approaching one) and minimizes (up to constants) $k \mapsto \{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\}$ over $\mathcal{G}_n$. Moreover, since $\tilde{k}_n$ is chosen as the minimal value in $\mathcal{L}_n$, to obtain the former condition it suffices to show that the tuning parameter belongs to $\mathcal{L}_n$ (with high probability).

By studying "projections" onto $\mathcal{G}_n$ of the tuning parameter that balances $k \mapsto \bar{\delta}_k(r_n^{-1})$ and $k \mapsto \bar{B}_k(P)$ we are able to explicitly construct a sequence of tuning parameters that satisfies these conditions; it is in this part that the monotonicity properties of these mappings are used. See Lemmas D.4 and D.3 in Appendix D.1. △

The following corollary is a direct consequence of Theorem 4.2 and its proof is omitted.

Corollary 4.1. Suppose $k \mapsto ||\psi_k(P) - \psi(P)||_\Theta$ and $k \mapsto \delta_k(r_n^{-1})$ are continuous, and non-increasing and non-decreasing resp.. Then under the conditions of Theorem 4.2, it follows

$$||\psi_{\tilde{k}_n(r_n)}(P_n) - \psi(P)||_\Theta = O_P\left(\inf_{k \in \mathcal{G}_n} \{\delta_k(r_n^{-1}) + ||\psi_k(P) - \psi(P)||_\Theta\}\right).$$

Theorem 4.2 and its corollary show that our data-driven choice of tuning parameter achieves the same rate as the one corresponding to the "oracle" choice, provided the monotonicity conditions hold.

The following proposition extends the result in Theorem 4.2 to an un-restricted one — where the infimum is not restricted to the set $\mathcal{G}_n$ but is taken over the whole $\mathbb{R}_+$. Unsurprisingly, in order to obtain this result, additional conditions are needed.

Proposition 4.1. Suppose all conditions in Theorem 4.2 hold, and that $k \mapsto \bar{\delta}_k(t)$ and $k \mapsto \bar{B}_k(P)$ are continuous and

$$\frac{\min_{k \in \mathcal{G}_n^+} \bar{\delta}_k(r_n^{-1})}{\max_{k \in \mathcal{G}_n^-} \bar{\delta}_k(r_n^{-1})} = O(1) \tag{7}$$

where $\mathcal{G}_n^+ \equiv \{k \in \mathcal{G}_n \colon \bar{\delta}_k(r_n^{-1}) \geq \bar{B}_k(P)\}$ and $\mathcal{G}_n^- \equiv \{k \in \mathcal{G}_n \colon \bar{\delta}_k(r_n^{-1}) \leq \bar{B}_k(P)\}$ are non-empty. Then

$$||\psi_{\tilde{k}_n}(P_n) - \psi(P)||_\Theta = O_P\left(\inf_{k \in \mathbb{R}_+} \{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\}\right).$$

Proof. See Appendix D.2. □

Remark 4.4 (On the conditions in the proposition)**.** The continuity condition is technical and it ensures that certain minimizers/maximizers are well-defined. Condition 7 imply the following two restrictions that are used in the proof:

(1) The fact that both $\mathcal{G}_n^+$ and $\mathcal{G}_n^-$ are non-empty ensures that the set $\mathcal{G}_n$ surrounds the choice of tuning parameter that balances the sampling error and the monotone envelope of the approximation error. If this condition fails, the minimal value of $k \mapsto \{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\}$ over $\mathcal{G}_n$ cannot be expected to be close to the value achieved when balancing both terms and thus close to the minimal value over $\mathbb{R}_+$. In Appendix D.2 we argue that $\mathcal{G}_n = \{1, ..., j(n)\}$ where $(j(n))_n$ is such that $\liminf_{n\to\infty} \bar{\delta}_{j(n)}(r_n^{-1}) > 0$ satisfies this assumption, at least for large n.

(2) The second role is more subtle. It essentially restricts — uniformly — the coarseness of the set $\mathcal{G}_n$ in terms of $k \mapsto \bar{\delta}_k(t)$. If $\bar{\delta}_k(t) = a(t) \times C_k$ and $\mathcal{G}_n = \mathbb{N}$, then the condition essentially imposes that $\limsup_{k\to\infty} C_{k+1}/C_k < \infty$; thus it allows for $C_k \asymp Poly(k)$ and $\log C_k \asymp k$ but not for $\log C_k \asymp k^2$. $\triangle$

4.3 Examples

Example 4.1 considers the case of bootstrapping the mean of a distribution when it is known to be non-negative. Andrews [2000] showed inconsistency of the bootstrap and proposed several consistent alternatives; we take one — the "k-out-of-n" bootstrap (Bickel and Freedman [1981]) — and illustrate how our methods can be used to derive the rate of convergence of this procedure and to choose the tuning parameter k that achieves this rate. To our knowledge this last result is novel.[12]

Example 4.2 provides primitive conditions for establishing continuity in M-estimation problems.

Example 4.1 (Bootstrap when the parameter is on the boundary)**.** Let $\mathcal{M}$ be the class of Borel probability measures over $\mathbb{R}$ with non-negative mean, unit variance and finite third moments; the non-negativity of the mean is a formalization that captures the issue of a parameter at the boundary. The object of interest is the law of an estimator of the mean, $\boldsymbol{z} \mapsto T_n(\boldsymbol{z}, P) = \sqrt{n}(\max\{n^{-1}\sum_{i=1}^n z_i, 0\} - \max\{E_P[Z], 0\})$. Thus, let, for each $k \in \mathbb{K} = \mathbb{N}$, $\psi_k : \mathcal{P}(\mathbb{R}) \to \mathcal{P}(\mathbb{R})$ be defined as

$$\psi_k(P)(A) \equiv \mathbf{P}\left(\{\boldsymbol{z} \colon T_k(\boldsymbol{z}, P) \in A\}\right),\ \forall A\ Borel.$$

In particular, for $P = P_n$, it follows that

$$\psi_k(P_n)(A) = \mathbf{P}_n\left(\sqrt{k}\left(\max\{k^{-1}\sum_{i=1}^k Z_i^*, 0\} - \max\{n^{-1}\sum_{i=1}^n Z_i, 0\}\right) \in A\right),\ \forall A\ Borel$$

[12] Bickel and Li [2006] and Bickel and Sakov [2008] perform a similar exercise but for a different case: estimation of largest order statistic.

where $(Z_i^*)_{i=1}^n$ is an IID sample drawn from P_n and $\mathbf{P}_n$ is the probability over $\mathbb{Z}^\infty$ induced by P_n. It is easy to see that $\psi_n(P_n)$ is the standard bootstrap estimator while $\psi_k(P_n)$ for $k < n$ is the k-out-of-n bootstrap estimator. Andrews [2000] showed that the "plug-in estimator", $\psi_n(P_n)$, while well-defined, fails to approximate the law of T_n, $\psi_n(P)$, even in the limit; but he showed that for certain sequences, $(k_n)_n$, $\psi_{k_n}(P_n) - \psi_n(P)$ converge to zero as n diverges. We now recast this result using the tools developed in this paper; by doing so we are able to provide a data-driven choice of the tuning parameter k_n.

To do this, we first show that the $(\psi_k)_{k\in\mathbb{N}}$ is continuous in the sense of Definition 4.1. Let $\Theta = \mathcal{P}(\mathbb{R})$ and let $||\cdot||_\Theta \equiv ||\cdot||_{LB}$, where recall LB is the class of real-valued Lipschitz with constant one function. This norm is one of the notions of distance typically used to establish validity of the Bootstrap. Also, let $\mathcal{W}(\cdot,\cdot)$ denote the Wassertein distance over $\mathcal{P}(\mathbb{Z})$, that is $\mathcal{W}(P,Q) \equiv \inf_{\zeta\in H(P,Q)} \int |z - z'|\zeta(dz, dz')$, where $H(P,Q)$ is the set of Borel probabilities over $\mathbb{Z}^2$ with marginals P and Q. The following proposition suggests the form of the modulus of continuity δ_k.

Proposition 4.2. For any $k \in \mathbb{N}$, $||\psi_k(P) - \psi_k(Q)||_\Theta \leq 2\sqrt{k}\mathcal{W}(P,Q)$ for any P and Q in $\mathcal{M} \cup \mathcal{D}$.

Proof. See Appendix D.4. □

The previous results suggests $\mathcal{W}$ as the natural distance over $\mathcal{P}(\mathbb{Z})$. In addition, the result also indicates that $\delta_k(t) = 2\sqrt{k}t$ for all $t \in \mathbb{R}_+$, which is increasing and continuous as a function of (t,k).

We now apply the results in Theorem 4.2 to choose the number of draws for the k-out-n bootstrap. Theorem 1 in Fournier and Guillin [2015] (their results are applied with $d = 1$, $p = 1$ and $q = 2$) shows that $\mathcal{W}(P_n, P) = O_P(n^{-1/2})$. Therefore, we take $r_n^{-1} = l_n n^{-1/2}$ where $(l_n)_n$ diverges arbitrary slowly. We also take $\mathcal{G}_n = \{1, ..., n\}$; it is clear that $k \mapsto \bar{B}_k(P) = B_k(P)$ and $k \mapsto \bar{\delta}_k(r_n^{-1}) = \delta_k(r_n^{-1})$. Given these choices, for each $n \in \mathbb{N}$, let $\tilde{k}_n$ be the choice of tuning parameter proposed above. Theorem 4.2 imply the following result.

Proposition 4.3. $||\psi_{\tilde{k}_n}(P_n) - \psi_n(P)||_{LB} = O_P\left(\inf_{k\in\{1,...,n\}}\{l_n\sqrt{k}n^{-1/2} + k^{-1/2}E_P[|Z|^3]\}\right)$.

Proof. See Appendix D.4. □

The RHS of the expression implies that the rate of convergence is given by $\sqrt{l_n}n^{-1/4}$. To our knowledge there is no data-driven method to choose the tuning parameter in this example. Bickel and Sakov [2008] propose a similar method to ours in a different example: Inference on the extrema of an IID sample. The authors obtain polynomial rates of convergence that are slower than ours but for a stronger norm. △

Example 4.2 (Regularized M-Estimators (cont.))**.** The following proposition shows that the regularization is continuous and more importantly it provides a "natural" choice of distance and illustrates the role of the regularization structure $\langle(\lambda_k, \Theta_k)_k, Pen\rangle$ and primitives (Θ, ϕ) for determining the rate of convergence of the regularized estimator. Henceforth, let $(\theta, P, k) \mapsto Q_k(P, \theta) \equiv E_P[\phi(Z, \theta)] + \lambda_k Pen(\theta)$.

Proposition 4.4. For each $k \in \mathbb{N}$ and $P \in \mathcal{M} \cup \mathcal{D}$,

$$||\psi_k(P) - \psi_k(P')||_{L^q} \leq \Gamma_k^{-1}(d(P, P')),\ \forall P' \in \mathcal{M} \cup \mathcal{D},$$

where for all $t > 0$

$$\Gamma_k(t) = \inf_{s \geq t} \left\{ \min_{\theta \in \Theta_k :\, ||\theta - \psi_k(P)||_{L^q} \geq s} \frac{Q_k(P, \theta) - Q_k(P, \psi_k(P))}{s} \right\}$$

and $d(P, P') \equiv \max_{k \in \mathbb{N}} ||P - P'||_{\mathcal{S}_k}$, where $\mathcal{S}_k \equiv \left\{ \frac{\phi(.,\theta) - \phi(.,\psi_k(P))}{||\theta - \psi_k(P)||_\Theta} : \theta \in \Theta_k \right\}$.[13]

Proof. See Appendix D.4 □

Following Shen and Wong [1994], the proof applies the standard arguments due to Wald — for establishing consistency of estimators — to "strips" of the sieve set Θ_k; by doing so, one improves the rates obtained from the standard Wald approach.

The proposition suggests the natural notion of distance over the space of probabilities, that is defined by the class of "test functions" given by $\left(\frac{\phi(z,\theta) - \phi(z,\psi_k(P))}{||\theta - \psi_k(P)||_\Theta} \right)_{\theta \in \Theta_k}$. By imposing additional conditions on ϕ and Θ_k one can embed the class $\mathcal{S}_k$ into well-known classes of functions for which one has a bound for the supremum of the empirical process $f \mapsto n^{-1} \sum_{i=1}^n f(Z_i) - E_P[f(Z)]$, and thus bounds for $d(P_n, P)$. For instance, if $\theta \mapsto \frac{d\phi(z,\theta)}{dz}$ is Lipschitz uniformly in z, then by using the mean value theorem and some algebra it follows that $\mathcal{S}_k \subseteq LB$ for every k, and thus $d(P_n, P) = O_P(n^{-1/2})$ (see van der Vaart and Wellner [1996]).

The modulus of continuity, Γ_k^{-1} is non-decreasing and is continuous over $t > 0$ (see the proof), and by definition $\Gamma_k(0) = 0$. Its behavior is determined by how well the criterion separates points in Θ_k relative to the norm $||.||_{L^q}$; the flatter $Q_k(P, \cdot)$ is around its minimizer, the larger Γ_k^{-1}. Importantly, even though $\Gamma_k(t) > 0$ for each k (recall that $\psi_k(P)$ is assumed to be unique), as k diverges, $\Gamma_k(t)$ may approach zero. This phenomena relates to the potential ill-posedness of the original problem, and will affect the rate of convergence of the estimator.

To shed some more light on the behavior of Γ_k and on the potential ill-posedness, consider the case where, $q = 2$, $Q(P, \cdot)$ is strictly concave and smooth, and $Pen(.) = ||.||^2_{L^2}$. Since

[13] We define $\Gamma_k(0) = 0$. The "$\inf_{s \geq t}$" ensures that Γ_k is non-decreasing; it can be omitted if such property is not needed. The "$max_{k \in \mathbb{N}}$" comes from the fact that d cannot depend on k in the definition of continuity.

$\psi_k(P)$ is a minimizer, $Q_k(P,\cdot)$ behaves locally as a quadratic function, in particular $\Gamma_k(t) \geq 0.5(C_k+\lambda_k)t$ for some non-negative constant C_k related to the Hessian of $Q(P,\cdot)$, and thus $\Gamma_k^{-1}(t) \precsim (C_k+\lambda_k)^{-1}t$. If $C_k \geq c > 0$ then $\Gamma_k^{-1}(t) \precsim t$; we deem this case to be well-posed as $||\psi_k(P') - \psi_k(P)||_{L^q} \precsim d(P',P)$.[14] On the other hand, if $\liminf_{k\to\infty} C_k = 0$ then, while the previous bound for the modulus of continuity is not possible, the following bound $\Gamma_k^{-1}(t) \precsim \lambda_k^{-1}t$ is. This case is deemed to be ill-posed and $||\psi_k(P')-\psi_k(P)||_{L^q} \precsim \lambda_k^{-1}d(P',P)$.

Finally, under the conditions discussed in the previous paragraph, in the ill-posed case, $k \mapsto \bar{\delta}_k(.) = \delta_k(.)$ if $k \mapsto \lambda_k$ is chosen to be non-increasing and continuous.[15] Thus Theorem 4.2 delivers a choice of tuning parameter that achieves consistency and a rate of $\min_{k\in\mathbb{N}}\{\lambda_k^{-1} \times r_n^{-1} + \inf_{l\geq k} ||\psi_l(P) - \psi(P)||_{L^q}\}$, where $(r_n)_n$ is such that $\max_{k\in\mathbb{N}} ||P_n - P||_{\mathcal{S}_k} = o_P(r_n^{-1})$. △

5 Differentiable Regularizations

In this section we derive asymptotic representations for the regularized estimator with *differentiable* regularizations. Before defining the concept of regularization, we define the notion of generalized asymptotic linearity (GAL). Throughout this section we assume $\Theta \subseteq \mathbb{R}$ to simplify the exposition; the results can be easily be extended to vector-valued parameters. In other cases where the parameter of interest is infinite-dimensional GAL is too weak and a stronger notion is needed, which we develop in Section 5.6.

5.1 Generalized Asymptotic Linearity

Let $\boldsymbol{\nu} \equiv (\nu_k)_{k\in\mathbb{K}}$ where, for all $k \in \mathbb{K}$, $\nu_k \in L_0^2(P) \equiv \{f \in L^2(P)\setminus\{0\} \colon E_P[f(Z)] = 0\}$.

Definition 5.1 (Generalized Asymptotic Linearity: GAL($\boldsymbol{k}$))**.** *A regularization $\boldsymbol{\psi}$ satisfies (weak) generalized asymptotic linearity for $\boldsymbol{k}\colon \mathbb{N} \to \mathbb{K}$ at $P \in \mathbb{D}_\psi$ with influence $\boldsymbol{\nu}$, if*

$$\left|\psi_{\boldsymbol{k}(n)}(P_n) - \psi_{\boldsymbol{k}(n)}(P) - n^{-1}\sum_{i=1}^n \nu_{\boldsymbol{k}(n)}(Z_i)\right| = o_P(n^{-1/2}||\nu_{\boldsymbol{k}(n)}||_{L^2(P)}). \quad (8)$$

If a regularization satisfies GAL($\boldsymbol{k}$) then, in order to study its asymptotic behavior, it suffices to study the behavior of $n^{-1/2}\sum_{i=1}^n \frac{\nu_{\boldsymbol{k}(n)}(Z_i)}{||\nu_{\boldsymbol{k}(n)}||_{L^2(P)}}$. Also, the reminder term is smaller than $n^{-1/2}||\nu_{\boldsymbol{k}(n)}||_{L^2(P)}$ as opposed to, say, $n^{-1/2}$, because the former is the proper order of the leading term, $n^{-1}\sum_{i=1}^n \nu_{\boldsymbol{k}(n)}(Z_i)$. That is, the natural scaling is given by $\sqrt{n}/||\nu_{\boldsymbol{k}(n)}||_{L^2(P)}$ as opposed to just $\sqrt{n}$; as the examples in Section 5.4 show, in many situations $\lim_{k\to\infty}||\nu_k||_{L^2(P)} = \infty$.

[14] This case relates to the so-called identifiable uniqueness condition (see White and Wooldridge [1991]).
[15] For the well-posed case the condition holds trivially.

The first result of this section is concerned with sufficient conditions ensuring existence of tuning sequences for which GAL holds. By a diagonalization argument (see Lemma C.1 in Appendix C), to obtain this result it suffices to show

$$\left|\psi_k(P_n) - \psi_k(P) - n^{-1}\sum_{i=1}^{n}\nu_k(Z_i)\right| = o_P(n^{-1/2})$$

for each $k \in \mathbb{K}$. That is, it suffices to show that for each $k \in \mathbb{K}$, ψ_k is asymptotic linear. As for the case of "plug-in" estimator, differentiability is the key property that allow us to achieve this result.

5.2 Definition of Differentiability

Let $\mathcal{T}_P \equiv \{a\mu\colon a \geq 0 \text{ and } \mu \in \mathcal{D} - \{P\}\}$. Throughout, let τ be a locally convex topology over $ca(\mathbb{Z})$ dominated by $||.||_{TV}$.[16]

Definition 5.2 (Differentiable Regularization: DIFF($P,\mathcal{C}$)). *A regularization* $\boldsymbol{\psi}$ *is differentiable at* $P \in \mathbb{D}_\psi$ *tangential to* $\mathcal{T}_P$ *under the class* $\mathcal{C} \subseteq 2^{\mathcal{T}_P}$*, if for any* $k \in \mathbb{K}$*, there exists a* $D\psi_k(P) : \mathcal{T}_P \to \Theta$ *τ-continuous and linear such that for any* $U \in \mathcal{C}$

$$\lim_{t\downarrow 0}\sup_{Q\in U}|\eta_k(tQ)|/t = 0, \text{ where } Q \mapsto \eta_k(Q) \equiv \psi_k(P+Q) - \psi_k(P) - D\psi_k(P)[Q]. \quad (9)$$

Remark 5.1. The functional $D\psi_k(P)$ acts as the gradient of ψ_k at P. The set $\mathcal{T}_P$ is the tangent set, i.e., the set that contains all the directions of the curves at P that we are considering; curves at P are of the form $t \mapsto P + tQ$ with $Q \in \mathcal{T}_P$. It turns out that it is enough to consider curves of the form $t \mapsto P + t\sqrt{n}(P_n - P)$ to obtain an asymptotic linear representation for the regularization. So, the choice of tangent set seems to be the most natural one. Of course, larger tangent sets will also deliver the desired results but establishing differentiability under them can be harder.

The definition does not impose any linear structure on $\mathcal{T}_P$ and t is restricted to be non-negative. This feature of the definition is analogous to the idea of directional derivative in Shapiro [1990] which has been shown to be sufficient for showing the validity of the Delta Method (see Shapiro [1990]), and turns out to be enough to also carry out our analysis. See also Fang and Santos [2014] and Cho and White [2017] for further references, examples and discussion. △

Remark 5.2. The class $\mathcal{C}$ determines the degree of uniformity of the limit and thus defines different notions of differentiability. It is known that common notions of differentiability

[16]Since we are working with measures, and not probabilities, it is convenient to allow for (non-metrizable) topologies. Locally convex topology means that it is constructed in terms of a family of semi-norms; dominated by $||.||_{TV}$ means that for any semi-norm, ρ, $\rho(Q) = O(||Q||_{TV})$ for all $Q \in ca(\mathbb{Z})$.

can be obtained from different choices of $\mathcal{C}$; see Dudley [2010] for a discussion. We now enumerate a few:

1. **τ-Gateaux Differentiability:** $\mathcal{C}$ is the class of finite subsets of $\mathcal{T}_P$; denoted by $\mathcal{J}_\tau$.
2. **τ-Hadamard Differentiability:** $\mathcal{C}$ is the class of τ-compact subsets of $\mathcal{T}_P$; denoted by $\mathcal{H}_\tau$.
3. **τ-Frechet Differentiability:** $\mathcal{C}$ is the class of τ-bounded subsets of $\mathcal{T}_P$; denoted by $\mathcal{E}_\tau$.

$\triangle$

5.3 Main Result

We now present the main result of this section.

Theorem 5.1. Suppose there exists a class $\mathcal{C} \subseteq 2^{\mathcal{T}}$ such that $\boldsymbol{\psi}$ is $DIFF(P,\mathcal{C})$ and [17]

> For any $\epsilon > 0$, there exists a $U \in \mathcal{C}$ and a N such that $\mathbf{P}\left(\sqrt{n}(P_n - P) \in U\right) \geq 1 - \epsilon$ for all $n \geq N$.

Then, there exists a $\boldsymbol{k}\colon \mathbb{N} \to \mathbb{K}$ for which $\boldsymbol{\psi}$ satisfies $GAL(\boldsymbol{k})$ and $\lim_{n\to\infty} \mathbf{k}(n) = \infty$.

Proof. See Appendix E. $\square$

It is easy to check that the influence of the regularization implied by the theorem is given by the sequence of $L_0^2(P)$ mappings, $(\varphi_k(P))_{k\in\mathbb{N}}$ where

$$z \mapsto \varphi_k(P)(z) \equiv D\psi_k(P)[\delta_z - P].$$

While the theorem shows existence of a sequence of tuning parameters for which generalized asymptotic linearity holds, it is silent about how to construct such sequence; we discuss this in Section 5.5.

Remark 5.3 (Heuristics of the Proof)**.** The proof is straightforward and is comprised of two steps. First, it is shown that $\boldsymbol{\psi}$ satisfies $GAL(\mathbf{k})$ for any fixed k, i.e., $\mathbf{k}(n) = k$. This result is analogous to the standard notion of asymptotic linearity applied but when ψ_k is used, thus,

[17] Implicit in the differentiability condition lies the assumption that for any $Q \in \mathcal{T}_P$, $t \mapsto P + tQ \in \mathbb{D}_\psi$. For this to hold, it is sufficient that P belongs to the algebraic interior of $\mathcal{M}$ relative to $\mathcal{T}_P$. However, by inspection of the proof of the Theorem, it can be seen that this assumption is not really needed since we only consider curves of the form $t \mapsto P + t_n a_n(P_n - P)$ where (t_n, a_n) are such that the curve equals P_n which is in $\mathbb{D}_\psi$.

it suffices to show that the reminder of the linear approximation is asymptotically negligible for each fixed k, i.e.,

$$\eta_k(P_n - P) = o_P(n^{-1/2}). \tag{10}$$

This is a standard condition for "plug-in" estimators (e.g. Van der Vaart [2000]), and the restriction over the class $\mathcal{C}$ and the definition of differentiability imply it. In some cases, however, it might be straightforward to verify condition (10) directly or by other means.

Second, a diagonalization argument is used to show existence of a diverging sequence. △

Remark 5.4. A common way of using Theorem 5.1 is by finding a class $\mathcal{S}$ that is P-Donsker, which implies that $(\sqrt{n}(P_n - P))_{n\in\mathbb{N}}$ is $||.||_{\mathcal{S}}$-compact (see Lemma E.1 in Appendix E), and ensuring $||.||_{\mathcal{S}}$-Hadamard differentiability; e.g. Van der Vaart [2000] Ch. 20. Dudley [2010] proposes an alternative way of using this result by showing that $\sqrt{n}(P_n - P)$ belongs, with high probability, to bounded p-variation sets, so the relevant notion of differentiability is Frechet differentiability (under the p-variation norm). △

5.4 Examples

The next example illustrates the implications of Theorem 5.1.

Example 5.1 (Integrated Square Density (cont.))**.** We now show that Definition 5.2 is satisfied by our class of regularizations and also establish a rate for the remainder term, $\eta_k(P_n - P)$ which is used to verify for which sequence of tuning parameter condition 10 holds.

Proposition 5.1. For any $P \in \mathcal{M}$, the regularization defined in expression (3) is $DIFF(P, \mathcal{E}_{||.||_{LB}})$. For each $k \in \mathbb{K}$,

$$Q \mapsto D\psi_k(P)[Q] = 2\int (\kappa_k \star P)(z)Q(dz),$$

and $Q \mapsto \eta_k(Q) = \int(\kappa_k \star Q)(z)Q(dz)$ is such that there exists a $L_k < \infty$ such that $|\eta_k(Q)| \leq L_k||Q||^2_{LB}$ for all $Q \in ca(\mathbb{Z})$.

Proof. See Appendix E.1. □

This proposition implies that for each $k \in \mathbb{K}$, ψ_k is $||.||_{LB}$-Frechet differentiable, and since LB is P-Donsker, the conditions in Theorem 5.1 are met. The influence is given by $z \mapsto \varphi_k(P)(z) \equiv 2\{(\kappa_k \star P)(z) - E_P[(\kappa_k \star P)(Z)]\}$, and since $\sup_k ||\varphi_k(P)||_{L^2(P)} \leq 2||p||_{L^\infty(\mathbb{R})}||\kappa||_{L^1(\mathbb{R})}$ (see Lemma E.2 in Appendix E.1), the natural scaling for GAL is $\sqrt{n}$. △

Next, we consider the NPIV example. It is not hard to see that the influence of $\boldsymbol{\gamma}$ will be given by $z \mapsto \int D\psi_k(P)^*[\pi](z) - E_P[D\psi_k(P)^*[\pi](Z)]$ provided $D\psi_k(P) : \mathcal{T}_P^* \to L^2([0,1])$ and its adjoint $D\psi_k(P)^* : L^2([0,1]) \to \mathcal{T}_P^*$ exists ($\mathcal{T}_P^*$ is the dual of $\mathcal{T}_P$). We show how our result can be used to characterize $D\psi_k(P)^*$ and thereby extend some results in the literature, for two widely used regularizations methods: Sieve-based and Penalization-based.

Example 5.2 (NPIV (cont.): The sieve-based Case)**.** We study the sieve-based regularization approach, which is constructed using two basis for $L^2([0,1])$, $(u_k, v_k)_{k\in\mathbb{N}}$, and two indices $k \mapsto (J(k), L(k))$ such that

$$
\begin{aligned}
&(g,x) \mapsto T_{k,P}[g](x) = (u^{J(k)}(x))^T Q_{uu}^{-1} E_P\left[u^{J(k)}(X) g(W)\right],\\
&x \mapsto r_{k,P}(x) = (u^{J(k)}(x))^T Q_{uu}^{-1} E_P[u^{J(k)}(X) Y],\\
&\mathcal{R}_{k,P} = (\Pi_k^* T_{k,P}^* T_{k,P} \Pi_k)^{-1}
\end{aligned}
$$

where $u^k(x) \equiv (u_1(x), ..., u_k(x))$, $v^k(w) \equiv (v_1(w), ..., v_k(w))$, $\Pi_k : L^2([0,1]) \to lin\{v^{L(k)}\} \subseteq L^2([0,1])$ is the projection operator, $g \mapsto \Pi_k[g] = (v^{L(k)})^T Q_{vv}^{-1} \int v^{L(k)}(w) g(w) dw$, and $Q_{uu} \equiv E_{Leb}[u^k(X)(u^k(X))^T]$, $Q_{uv} \equiv E_P[u^k(X)(v^k(W))^T]$ and $Q_{vv} \equiv E_{Leb}[v^k(W)(v^k(W))^T]$.

The next proposition proves differentiable of the regularization $\boldsymbol{\gamma}$ and provides the expression for the derivative.

Proposition 5.2. For any $P \in \mathcal{M}$, the sieve-based regularization $\boldsymbol{\gamma}$ is DIFF$(P, \mathcal{E}_{||.||_{LB}})$. For each $k \in \mathbb{N}$,

$$
Q \mapsto D\gamma_k(P)[Q] = \int D\psi_k(P)^*[\pi](z) Q(dz)
$$

where

$$
\begin{aligned}
D\psi_k(P)^*[\pi](y,w,x) =& (y - \psi_k(P)(w))(u^{J(k)}(x))^T Q_{uu}^{-1} Q_{uv} (Q_{uv}^T Q_{uu}^{-1} Q_{uv})^{-1} E_{Leb}[v^{L(k)}(W)\pi(W)]\\
&+ \left\{ E_P[(\psi(P)(W) - \psi_k(P)(W))(u^{J(k)}(X))^T] Q_{uu}^{-1} u^{J(k)}(x) \right.\\
&\left. \times (v^{L(k)}(w))^T (Q_{uv}^T Q_{uu}^{-1} Q_{uv})^{-1} E_{Leb}[v^{L(k)}(W)\pi(W)] \right\}. \qquad (11)
\end{aligned}
$$

And, for each $k \in \mathbb{N}$, the reminder of γ_k, η_k, is such that $|\eta_k(\zeta)| = o(||\zeta||_{LB})$ for any $\zeta \in \mathbb{D}_\psi$.[18]

Proof. See Appendix E.2. □

Even though expression for $D\psi_k(P)^*[\pi]$ may look cumbersome, it has an intuitive interpretation: It is identical to the influence function of the parameter $\int \theta^T v^{L(k)}(w)\pi(w)dw$ where θ is the estimand of a *misspecified linear GMM model* where the "endogenous variables" are $v^{L(k)}(W)$ and the "instrumental variables" are $u^{J(k)}(X)$; cf. Hall and Inoue [2003]. The first term in the RHS of expression 11 also has an intuitive interpretation: It is the influence function of the parameter $\int \theta^T v^{L(k)}(w)\pi(w)dw$ but in *well-specified linear GMM model.*

[18]The "o" function may depend on k.

The proposition implies that for the "fix-k" case, expression 11 is the proper influence function to be considered. However, one can ask whether as k diverges, the second term (the one in curly brackets) in RHS of expression 11 can be ignored. To shed light on this matter, it is convenient to use operator notation for expression 11:

$$\begin{aligned} D\psi_k^*(P)[\pi](y,w,x) =& T_{k,P}\mathcal{R}_{k,P}\Pi_k[\pi](x) \times (y - \psi_k(P)(w)) \\ &+ \mathcal{R}_{k,P}\Pi_k[\pi](w) \times T_{k,P}[\psi(P) - \psi_k(P)](x) \end{aligned} \tag{12}$$

(we derive this equality in expression 23 in Appendix E.2). The term $T_{k,P}[\psi(P) - \psi_k(P)]$ is multiplied by $\mathcal{R}_{k,P}\Pi_k[\pi]$, which is different to $T_{k,P}\mathcal{R}_{k,P}\Pi_k[\pi]$ — the factor multiplying $(y - \psi_k(P)(w))$. If $\pi \in Range(T_P)$ both multiplying factors converge to bounded quantities as k diverges.Thus, since $T_{k,P}[\psi(P) - \psi_k(P)]$ vanishes, the first summand in the RHS of expression 11 "asymptotically dominates" the second one. This is framework considered in Ackerberg et al. [2014]. However, if $\pi \notin Range(T_P)$ — and thus $\gamma(P)$ is not root-estimable (see Severini and Tripathi [2012]) — the situation is more subtle and without additional assumptions it is not clear which term in expression 11 dominates. The reason is that the aforementioned multiplying factors will no longer converge to a bounded quantity, and moreover, the rate of growth of $T_{k,P}\mathcal{R}_{k,P}\Pi_k[\pi]$ can can be dominated by the rate of $\mathcal{R}_{k,P}\Pi_k[\pi]$.

For this last case of $\pi \notin Range(T_P)$, the results closest to ours are those in Chen and Pouzo [2015] wherein the influence function for slower than root-n sieve estimators is derived. Their expression for the influence function is simpler than ours, but this arises from a different set of assumptions and, more importantly, a different approach that directly focus on expressions for "diverging k". $\triangle$

Example 5.3 (NPIV (cont.): The Penalization-based Case)**.** We study the penalization-based regularization case given by

$$\begin{aligned} &(x,g) \mapsto T_{k,P}[g](x) \equiv \int \kappa_k(x'-x) \int g(w)P(dw,dx') \\ &x \mapsto r_{k,P}(x) \equiv \int \kappa_k(x'-x) \int yP(dy,dx') \\ &\mathcal{R}_{k,P} = (T_{k,P}^* T_{k,P} + \lambda_k I)^{-1} \end{aligned}$$

where $\kappa_k(\cdot) = k\kappa(k\cdot)$ and κ is a smooth, symmetric around 0 pdf.

As opposed to the previous case, there is no obvious link to a "simpler" problem like GMM and thus it is not obvious a-priori what the influence function would be and what the proper scaling should be when $\gamma(P)$ is not root-n estimable. Theorem 5.1 suggests $D\psi_k^*(P)[\pi]$ and $\sqrt{n/Var_P(D\psi_k^*(P)[\pi])}$ as the influence function and scaling factor resp.; the next proposition characterizes it.

Proposition 5.3. For any $P \in \mathcal{M}$, the Penalization-based regularization $\boldsymbol{\gamma}$ is DIFF$(P, \mathcal{E}_{||.||_{LB}})$. For each $k \in \mathbb{N}$, $D\gamma_k(P)[\zeta] = \int D\psi_k(P)^*[\pi](z)\zeta(dz)$, where

$$D\psi_k^*(P)[\pi](y,w,x) = \mathcal{K}_k^2 T_P(T_P^*\mathcal{K}_k^2 T_P + \lambda_k I)^{-1}[\pi](x) \times (y - \psi_k(P)(w)) \\ + \lambda_k (T_P^*\mathcal{K}_k^2 T_P + \lambda_k I)^{-1}[\pi](w) \times \mathcal{K}_k^2 T_P(T_P^*\mathcal{K}_k^2 T_P + \lambda_k I)^{-1}[\psi_{id}(P)](x). \quad (13)$$

where $\mathcal{K}_k$ is the convolution operator $g \mapsto \mathcal{K}_k[g] = \kappa_k \star g$. And, for each $k \in \mathbb{N}$, the reminder of γ_k, η_k, is such that $|\eta_k(\zeta)| = o(||\zeta||_{LB})$ for any $\zeta \in \mathbb{D}_\psi$.[19]

Proof. See Appendix E.2. □

If $\pi \in Range(T_P)$, then the variance term converges to $||T_P[v^*](X)(Y - \psi(P)(W))||^2_{L^2(P)} = E_P[(T_P(T_P^*T_P)^{-1}[\pi](X))^2 E_P[(Y - \psi(P)(W))^2|X]]$ as k diverges, where $v^* \equiv (T_P^*T_P)^{-1}[\pi]$. The function $(y,w,x) \mapsto T_P[v^*](x)(y - \psi(P)(w))$ is the influence function one would obtained by employing the methods in Ai and Chen [2007a] (with identity weighting) and v^* is the Riesz representer of the functional $w \mapsto \int \pi(w)g(w)dw$ using their weak norm $||T_P[\cdot]||_{L^2(P)}$.

If $\pi \notin Range(T_P)$, the variance diverges, and, as in the sieve case, without additional assumptions it is not clear which term dominates the variance term $Var_P(D\psi_k^*(P)[\pi])$, as k diverges. This case illustrates how our results can be used to extend the results in Chen and Pouzo [2015] for irregular sieve-based estimators to more general regularization schemes. △

5.5 Data-driven Choice of Tuning Parameter and Undersmoothing

Theorem 5.1 implies existence of a $n \mapsto k(n)$ such that[20]

$$\frac{\sqrt{n}(\psi_{k(n)}(P_n) - \psi(P))}{||\varphi_{k(n)}(P)||_{L^2(P)}} - n^{-1/2}\sum_{i=1}^{n} \frac{\varphi_{k(n)}(P)(Z_i)}{||\varphi_{k(n)}(P)||_{L^2(P)}} = \frac{\sqrt{n}B_{k(n)}(P)}{||\varphi_{k(n)}(P)||_{L^2(P)}} + o_P(1). \quad (14)$$

I.e., the asymptotic behavior of the regularized estimator — once scaled and centered — is characterized by a term due to the approximation error and a stochastic term. Ideally, one would like to consider sequences $(k(n))_n$ satisfying Theorem 5.1 for which the approximation term in expression (14) vanishes, but, unfortunately it is known that this result is unattainable at this level of generality; e.g. Bickel and Ritov [1988] and Hall and Marron [1987].

[19]The "o" function may depend on k.

[20]The display hold provided $\liminf_{n\to\infty} ||\varphi_{k(n)}(P)||_{L^2(P)} > 0$ For the applications we have in mind, this restriction is natural and non-binding. Our results are not designed for cases where $\lim_{k\to\infty} ||\varphi_k(P)||_{L^2(P)} = 0$; this case can be handled separately — and rather easily — since both the approximation error and the rate of $k \mapsto \eta_k(P_n - P)$ decrease as k increases.

In view of this remark it is natural to seek choices of tuning parameter that make the terms in the RHS of expression (14) as "small as possible". Such choices will guarantee that GAL and the asymptotic negligibility of the approximation error both hold when possible, and otherwise, will at least yield good rates of convergence for $\psi_k(P_n) - \psi(P)$.

The result in this section shows that the data-driven way of choosing tuning parameters described in Section 4.1 satisfies this property. For each $n \in \mathbb{N}$, the data-driven choice of tuning parameter is of the form $\tilde{k}_n = \arg\min\{k \colon k \in \mathcal{L}_n((\Lambda_k)_k)\}$ for a suitable chosen sequence $(\Lambda_k)_k$. In section 4.1, the relevant sequence was $(4\bar{\delta}_k(r_n^{-1}))_k$; in this case, however, the structure of the problem is different. In particular, in addition to the reminder term $(\eta_k)_k$ implied by differentiability and the scaled approximation error, there is the additional term given by $n^{-1/2}\sum_{i=1}^{n} \frac{\varphi_{k(n)}(P)(Z_i)}{||\varphi_{k(n)}(P)||_{L^2(P)}}$. The following assumption introduces the quantities to construct $(\Lambda_k)_k$. For each n, let $\mathcal{G}_n$ be the grid defined as in Section 4.1.

Assumption 5.1. *There exists a* $(n,k) \mapsto \bar{\delta}_{j,k}(n)$ *for* $j \in \{1,2\}$ *non-decreasing and a* $N \in \mathbb{N}$ *such that*

(i) $\sup_{k \in \mathcal{G}_n} \frac{\sqrt{n}|\eta_k(P_n - P)|}{\bar{\delta}_{1,k}(n)} \leq 1$ *wpa1-P.*

(ii) $|\mathcal{G}_n| \sup_{k' \geq k \text{ in } \mathcal{G}_n} \frac{||\varphi_{k'}(P) - \varphi_k(P)||_{L^2(P)}}{\bar{\delta}_{2,k'}(n)} \leq 1$ *for all* $n \geq N$.

As the proof of Lemma E.7 in Appendix E.3 suggests, the sequence that defines our tuning parameter, for each $n \in \mathbb{N}$, is given by $k \mapsto \Lambda_k \equiv 4\frac{\bar{\delta}_{1,k}(n)+\bar{\delta}_{2,k}(n)}{\sqrt{n}}$.

Remark 5.5 (Discussion of Assumption 5.1)**.** Part (ii) implies that $(\bar{\delta}_{2,k}(n))_{n,k}$ acts as a growth rate for an object that, on the hand, involves the complexity of $\mathcal{G}_n$ — given by $|\mathcal{G}_n|$ — and on the other hand, involves the "length" of $\mathcal{G}_n$ — measured by $k \mapsto ||\varphi_k(P)||_{L^2(P)}$. In cases where $(||\varphi_k(P)||_{L^2(P)})_k$ is uniformly bounded, the "length" of $\mathcal{G}_n$ (measured by $k \mapsto ||\varphi_k(P)||_{L^2(P)}$) is also uniformly bounded and part (ii) boils down to $|\mathcal{G}_n| \leq \inf_{k \in \mathcal{G}_n} \delta_{2,k}(n)$). Part (i) implies that $(\bar{\delta}_{1,k}(n))_{n,k}$ also acts as the growth rate, but of a very different quantity: The reminder term of GAL (scaled by $\sqrt{n}$), uniformly on $k \in \mathcal{G}_n$.

To shed more light on Assumption 5.1, suppose there exists a norm $||.||_{\mathcal{S}}$ such that

C1: There exists, for each $k \in \mathbb{K}$ a modulus of continuity $\bar{\eta}_k : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\eta_k(Q) = \bar{\eta}_k(||Q||_{\mathcal{S}})$.

C2: There exists a real-valued positive diverging sequence $(r_n)_n$ such that $||P_n - P||_{\mathcal{S}} = o_P(r_n^{-1})$.

Condition C1 states that η_k is continuous with respect to some norm $||.||_{\mathcal{S}}$ and C2 ensure convergence of P_n to P under this norm. These conditions are analogous to the assumptions used to show Theorem 4.2.

Under these conditions it is easy to see that Part (i) follows by choosing $\bar{\delta}_{1,k}(n) = \sqrt{n}\bar{\eta}_k(r_n^{-1})$, which acts as $\delta_k(r_n^{-1})$ in Theorem 4.2, and, in particular, it does not depend on the grid. Part (ii), however, is not necessarily implied by this choice. If the growth rate of the reminder, $\sqrt{n}\bar{\eta}_k(r_n^{-1})$, is small compared to $|\mathcal{G}_n| \sup_{k' \geq k \text{ in } \mathcal{G}_n} ||\varphi_{k'}(P) - \varphi_k(P)||_{L^2(P)}$ then part (ii) requires that $\bar{\delta}_{2,k}(n)$ to be larger than the latter, i.e., $\bar{\delta}_{2,k}(n) \geq |\mathcal{G}_n| \sup_{k' \geq k \text{ in } \mathcal{G}_n} ||\varphi_{k'}(P) - \varphi_k(P)||_{L^2(P)}$.

Below we illustrate how to verify these assumptions in the context of Example 2.2.△

Proposition 5.4. Suppose all the conditions of Theorem 5.1 hold, and Assumption 5.1 holds. Then[21]

$$\left| \frac{\sqrt{n}(\psi_{\tilde{k}_n}(P_n) - \psi(P))}{||\varphi_{\tilde{k}_n}(P)||_{L^2(P)}} - \frac{n^{-1/2}\sum_{i=1}^n \varphi_{\tilde{k}_n}(P)(Z_i)}{||\varphi_{\tilde{k}_n}(P)||_{L^2(P)}} \right| = O_P\left(C_n^2 \inf_{k \in \mathcal{G}_n} \left\{ \frac{\bar{\delta}_{1,k}(n) + \bar{\delta}_{2,k}(n) + \sqrt{n}\bar{B}_k(P)}{||\varphi_k(P)||_{L^2(P)}} \right\} \right),$$

where $C_n \equiv \sup_{k',k \text{ in } \mathcal{G}_n} \frac{||\varphi_k(P)||_{L^2(P)}}{||\varphi_{k'}(P)||_{L^2(P)}}$.

Proof. See Appendix E.3. □

The rate in the proposition is — up to C_n^2 factor — the minimum value of the sum of two terms: $\frac{\bar{\delta}_{1,k}(n)+\bar{\delta}_{2,k}(n)}{||\varphi_k(P)||_{L^2(P)}}$, that controls the reminder term of GAL and another one, $\sqrt{n}\frac{B_k(P)}{||\varphi_k(P)||_{L^2(P)}}$, that controls the approximation term $\sqrt{n}\frac{|\psi_k(P)-\psi(P)|}{||\varphi_k(P)||_{L^2(P)}}$. Therefore, if there exists a choice of tuning parameter for which both these terms are asymptotically negligible, our result implies that

$$\sqrt{n}\frac{\psi_{\tilde{k}_n}(P_n) - \psi(P)}{||\varphi_{\tilde{k}_n}(P)||_{L^2(P)}} = n^{-1/2}\sum_{i=1}^n \frac{\varphi_{\tilde{k}_n}(P)(Z_i)}{||\varphi_{\tilde{k}_n}(P)||_{L^2(P)}} + o_P(1).$$

That is, the asymptotic distribution of $\sqrt{n}\frac{\psi_{\tilde{k}_n}(P_n)-\psi(P)}{||\varphi_{\tilde{k}_n}(P)||_{L^2(P)}}$ is given that of $n^{-1/2}\sum_{i=1}^n \frac{\varphi_{\tilde{k}_n}(P)(Z_i)}{||\varphi_{\tilde{k}_n}(P)||_{L^2(P)}}$. On the other hand, if no such sequence exists, the proposition readily implies a rate of convergence of the form $\left|\frac{\psi_{\tilde{k}_n}(P_n)-\psi(P)}{||\varphi_{\tilde{k}_n}(P)||_{L^2(P)}}\right| = O_P\left(n^{-1/2} + n^{-1/2}C_n^2 \inf_{k\in\mathcal{G}_n}\left\{\frac{\bar{\delta}_{1,k}(n)+\bar{\delta}_{2,k}(n)+\sqrt{n}\bar{B}_k(P)}{||\varphi_k(P)||_{L^2(P)}}\right\}\right)$.

The sequence $(C_n)_n$ quantifies the discrepancy of $k \mapsto ||\varphi_k(P)||_{L^2(P)}$ within the grid. In cases where Assumption 5.1(ii) holds for all k' and k in $\mathcal{G}_n$, it readily follows that $C_n = 1 + |\mathcal{G}_n|^{-1}\sup_{k\in\mathcal{G}_n} \bar{\delta}_{2,k}(n)/||\varphi_k(P)||_{L^2(P)}$.

In order to shed more light on these expressions and the assumptions, we applied our results to the estimation of the integrated square pdf (example 2.2). In this setting, Gine and Nickl [2008] provide a data-driven method to choose the bandwidth which is akin to ours. In fact, our method can be viewed as generalization of theirs to general regularizations.

[21] By arguments analogous to those in Proposition 4.1, this result can be extended to hold for $\inf_{k\in\mathbb{R}_+}$.

Example 5.4 (Integrated Square Density (cont.)). In this example the relevant tuning parameter is the bandwidth of the kernel, so we let $k \mapsto k^{-1}$, and as the grid, $\mathcal{G}_n$, we use the one proposed by Gine and Nickl (Gine and Nickl [2008]), i.e., $\mathcal{G}_n = \{k : k^{-1} \in \mathcal{H}_n\}$ where
$\mathcal{H}_n = \left\{ h \in \left[\frac{(\log n)^4}{n^2}, \frac{1}{n^{1-\delta}} \right] : \ h_0 = \frac{1}{n^{1-\delta}}, \ h_1 = \frac{\log n}{n}, \ h_2 = \frac{l_n^{-1}}{n}, \ h_{k+1} = h_k/a, \ \forall k = 2, 3, ... \right\}$
where $a > 1$, $(l_n)_n$ diverges to infinity slower than $\log n$ and $l_n^{-1} < \log n$ and $\delta > 0$ is arbitrary close to 1; in particular $\delta > 0$ is such that $2\varrho < \frac{1+\delta}{2(1-\delta)}$. Of importance to our analysis are the fact that $|\mathcal{G}_n| = O(\log n)$ and that for sufficiently large n, any two consecutive elements in $\mathcal{H}_n$ are such that $h_{k+1}/h_k \leq 1/a$.

The following lemma suggests an expression for the functions $(n, k) \mapsto \bar{\delta}_{i,k}(n)$ for $i \in \{1, 2\}$.

Lemma 5.1. *For any $M > 0$, there exists a N such that for all $n \geq N$,*

$$\sup_{h' \leq h \text{ in } \mathcal{H}_n} ||\varphi_{1/h}(P) - \varphi_{1/h'}(P)||_{L^2(P)} \leq 4||C||_{L^2(P)} h^{\varrho} E_{|\kappa|}[|U|^{m+\varrho}].$$

where the function C is the one in expression 2 in Example 3.1, and

$$\mathbf{P}\left(\sup_{h \in \mathcal{H}_n} \sqrt{n} |\eta_{1/h}(P_n - P)| \geq M \left(\frac{\kappa(0)}{\sqrt{n}h} + \frac{1}{\sqrt{nh}} \right) \right) \leq |\mathcal{H}_n| M^{-1}.$$

Proof. See Appendix E.3.2. □

Therefore, $\{(n, k) \mapsto \bar{\delta}_{i,k}(n)\}_{i=1,2}$ can be chosen as

$$(n, k) \mapsto \bar{\delta}_{1,k} \equiv (\log n)^3 \frac{k\kappa(0) + \sqrt{k}}{\sqrt{n}}, \ \text{and} \ (n, k) \mapsto \bar{\delta}_{2,k} \equiv (\log n)^3 k^{-(m+\varrho)}.$$

The lemma and this display illustrate the different nature of Assumptions 5.1(i)(ii). Part (i) bounds the reminder of the linear approximation and it increases with k and decreases with n; this is reflected in the term $\frac{(k\kappa(0)+\sqrt{k})}{\sqrt{n}}$ in the display. Part (ii) on the other hand essentially requires that the bandwidths in the grid $\mathcal{H}_n$ are not "too far apart". In particular, it depends on the size of the bandwidths in $\mathcal{H}_n$; this is reflected in the term $k^{-\varrho}$ in the display.

It follows that $\sup_{n \in \mathbb{N}} C_n < \infty$ because there exists a constant $C > 1$ such that $k \mapsto ||\varphi_k(P)||_{L^2(P)} \in [C^{-1}, C]$ and is continuous for all $k \geq 1$.

We verified that all assumptions of Proposition 5.4 hold. Moreover, Proposition 3.1 implies that $h \mapsto \bar{B}_{1/h}(P) = O(h^{2\varrho})$. Thus, the rate of Proposition 5.4 is given by
$\inf_{h \in \mathcal{H}_n} \{(\log n)^3 \left(\frac{\kappa(0)/h + 1/\sqrt{h}}{\sqrt{n}} + h^{\varrho} \right) + \sqrt{n} h^{2\varrho}\}$. In fact, given our choice of $\mathcal{H}_n$ and δ, some straightforward algebra shows that, at least for large n, the infimum over $\mathcal{G}_n$ and be replaced

by the infimum over $\mathbb{R}_+$. Therefore,

$$\left|\frac{\sqrt{n}(\psi_{\tilde{k}_n}(P_n)-\psi(P))}{||\varphi_{\tilde{k}_n}(P)||_{L^2(P)}}-\frac{n^{-1/2}\sum_{i=1}^n\varphi_{\tilde{k}_n}(P)(Z_i)}{||\varphi_{\tilde{k}_n}(P)||_{L^2(P)}}\right|=\begin{cases} O_P\left(\left(\frac{\log n}{n}\right)^{\frac{4\varrho}{1+4\varrho}-0.5}\right) & if\ \kappa(0)=0\\ O_P\left(\left(\frac{\log n}{n}\right)^{\frac{2\varrho}{1+2\varrho}-0.5}\right) & if\ \kappa(0)>0\end{cases}$$

For the case $\kappa(0)=0$, we replicate the results by Gine and Nickl [2008]: if $\varrho>0.25$, the reminder is negligible and root-n consistency follows, otherwise the optimal convergence rate is achieved. Our framework, however, also shows how the results in Gine and Nickl [2008] can be extended to other cases. △

5.6 Extension: Strong Generalized Asymptotic Linearity

We now establish an analogous result to Theorem 5.1 but for an stronger notion than GAL($\boldsymbol{k}$), one that is better suited when the parameter of interest is infinite dimensional. To do this, let Ξ be a subset of Θ^*, the dual of $(\Theta,||.||_\Theta)$.

Definition 5.3 (Strong Generalized Asymptotic Linearity: S-GAL($\Xi,\boldsymbol{k}$))**.** *A regularization $\boldsymbol{\psi}$ satisfies strong generalized asymptotic linearity for $\boldsymbol{k}\colon\mathbb{N}\to\mathbb{K}$ under Ξ at $P\in\mathbb{D}_\psi$ with influence $\boldsymbol{\nu}$, if for all $(n,\ell)\in\mathbb{N}\times\Xi$, $\ell[\nu_{\boldsymbol{k}(n)}]\in L^2_0(P)$ and*

$$\sup_{\ell\in\Xi}\left|\frac{\ell[\psi_{\boldsymbol{k}(n)}(P_n)-\psi_{\boldsymbol{k}(n)}(P)]}{||\ell[\nu_{\boldsymbol{k}(n)}]||_{L^2(P)}}-n^{-1}\sum_{i=1}^n\frac{\ell[\nu_{\boldsymbol{k}(n)}](Z_i)}{||\ell[\nu_{\boldsymbol{k}(n)}]||_{L^2(P)}}\right|=o_P(n^{-1/2}). \tag{15}$$

As, for each $\ell\in\Xi$, $\ell[\boldsymbol{\psi}]\equiv(\ell[\psi_k])_{k\in\mathbb{K}}$ is a real-valued sequence, Definition 5.2 can be applied to $\ell[\boldsymbol{\psi}]$. It immediately follows that $\eta_{k,\ell}(tQ)\equiv\ell[\eta_k(tQ)]=o(t)$ uniformly on $Q\in U\subseteq\mathcal{C}$ but *pointwise* on $\ell\in\Xi$. While this condition is enough when the parameter of interest is a vector-valued functional of $\psi(P)$; in other cases, where the parameter of interest is infinite-dimensional, this restriction is too weak, and the following strengthening of 9 is needed: For any $U\in\mathcal{C}$,

$$\lim_{t\downarrow 0}\sup_{U\in Q}\sup_{\ell\in\Xi}|\eta_{k,\ell}(tQ)/t|=0, \tag{16}$$

i.e., now the restriction in the reminder $\eta_{k,\ell}$ holds *uniformly* over $\ell\in\Xi$.

With a slight abuse of notation, we use $DIFF(P,\mathcal{C},\Xi)$ to denote the case when a regularization $\boldsymbol{\psi}$ is such that for each $\ell\in\Xi$, $\ell[\boldsymbol{\psi}]$ is $DIFF(P,\mathcal{C})$ (in the sense of Definition 5.2), and condition 16 holds.

Theorem 5.2. Suppose there exists a class $\mathcal{C}\subseteq 2^{\mathcal{T}_P}$ such that $\boldsymbol{\psi}$ is $DIFF(P,\mathcal{C},\Xi)$ (in the sense above), and

For any $\epsilon>0$, there exists a $U\in\mathcal{C}$ and a N such that $\mathbf{P}\left(\sqrt{n}(P_n-P)\in U\right)\geq 1-\epsilon$ for all $n\geq N$.

Then, there exists a $\boldsymbol{k}\colon \mathbb{N} \to \mathbb{N}$ for which $\boldsymbol{\psi}$ satisfies $S - GAL(\Xi, \boldsymbol{k})$ and $\lim_{n\to\infty} \mathbf{k}(n) = \infty$.

The proof is omitted since is completely analogous to the one of Theorem 5.1. As the following examples illustrate, the theorem can be used to construct inference for confidence bands. The first example simply illustrates how to apply the theorem to a known problem, the second example, however, provides novel results for general M-estimation problems.

Example 5.5 (Density Estimation (cont.))**.** Consider the setup in Example 2.1 but now the parameter of interest is $z \mapsto \psi(P)(z) = p(z)$ and the regularization is given by $z \mapsto \psi_k(P)(z) = (\kappa_k \star P)(z)$. The goal is to obtain an asymptotic linear representation *uniformly* over $z \in \mathbb{R}$, specifically,

$$\sup_{z\in\mathbb{R}} \left| \sqrt{n}\frac{(\kappa_{k_n} \star P_n)(z) - p(z)}{\sqrt{Var_P(\kappa_{k_n}(z-Z))}} - n^{-1/2}\sum_{i=1}^{n} \frac{\kappa_{k_n}(z-Z_i) - E_P[\kappa_{k_n}(z-Z)]}{\sqrt{Var_P(\kappa_{k_n}(z-Z))}} \right| = o_P(1),$$

for certain diverging sequences $(k_n)_{n\in\mathbb{N}}$. From this representation, by invoking known limit theorems results one can derive the asymptotic distribution of $\sup_{z\in\mathbb{R}} \left| \sqrt{n}\frac{(\kappa_{k_n}\star P_n)(z)-p(z)}{\sqrt{Var_P(\kappa_{k_n}(z-Z))}} \right|$; see Bickel and Rosenblatt [1973].

We now illustrate how Theorem 5.2 can be used to achieve the asymptotic representation in the previous display. Let $\Xi = \{\delta_x\colon x \in \mathbb{R}\}$. Suppose that Θ is the class of bounded continuous functions, then $\Xi \subseteq \Theta^*$ and for each $z \in \mathbb{R}$, there exist a corresponding $\ell = \delta_z \in \Xi$ such that $\ell[p] = p(z)$. Since ψ_k is linear, it also follows that $\ell[D\psi_k(P)[Q]] = \ell[\kappa_k \star Q] = (\kappa_k \star Q)(z_\ell)$ for any $Q \in \mathcal{T}_P = ca(\mathbb{R})$, and $\eta_{k,\ell}(.) = 0$. By Theorem 5.1, the regularization satisfies $GAL(\mathbf{k}, \Xi)$ for any $(k_n)_{n\in\mathbb{N}}$ with influence function given by $z \mapsto \ell[\varphi_k(P)](z) = \kappa_k(z_\ell - z) - E_P[\kappa_k(z_\ell - Z)]$. Hence, by Theorem 5.2 it follows that

$$\begin{aligned}&\sup_{z\in\mathbb{R}} \left| \sqrt{n}\frac{(\kappa_{k_n} \star P_n)(z) - p(z)}{\sqrt{Var_P(\kappa_{k_n}(z-Z))}} - n^{-1/2}\sum_{i=1}^{n} \frac{\kappa_{k_n}(z-Z_i) - E_P[\kappa_{k_n}(z-Z)]}{\sqrt{Var_P(\kappa_{k_n}(z-Z))}} \right| \\ &= O\left(\sup_{z\in\mathbb{R}} \left| \sqrt{n}\frac{(\kappa_{k_n} \star P)(z) - p(z)}{\sqrt{Var_P(\kappa_{k_n}(z-Z))}} \right| \right) + o_P(1)\end{aligned}$$

for any $(k_n)_{n\in\mathbb{N}}$.

Under our conditions over p, it follows that $\left| \sqrt{n}\frac{(\kappa_k\star P)(z)-p(z)}{\sqrt{Var_P(\kappa_k(z-Z))}} \right| \leq \frac{\sqrt{n}}{k^{1+\varrho}} E_{|\kappa|}[|U|^{1+\varrho}] \left| \frac{C(z)}{\sqrt{Var_P(\kappa_k(z-Z))}} \right|$ for any $z \in \mathbb{R}$. To give a more precise uniform bound for this term we need to estimate a lower bound for $\sqrt{Var_P(\kappa_k(z-Z))}$. If $C \asymp p$, it can be proved that $Var_P(\kappa_k(z-Z)) \asymp kp(z)||\kappa||^2_{L^2} + |p'(z)| \int \kappa(u)^2 u du$ at least for large k.[22] Hence, $\sup_{z\in\mathbb{R}} \left| \sqrt{n}\frac{(\kappa_{k_n}\star P)(z)-p(z)}{\sqrt{Var_P(\kappa_{k_n}(z-Z))}} \right| = O\left(\frac{\sqrt{n}}{k_n^{1.5+\varrho}} \right)$, and thus for any $(k_n)_{n\in\mathbb{N}}$ such that $n^{\frac{1}{3+2\varrho}} = o(k_n)$ the desired representation

[22] $C \asymp p$ is imposed to simplify the derivations in $Var_P(\kappa_k(z-Z))$.

follows. △

We now present sufficient conditions over $(\mathcal{M}, \Theta, \phi)$ and the regularization structure that guarantees differentiability of the regularization, and asymptotic confidence bands for the regularized M-estimator.

Example 5.6 (Regularized M-Estimators (cont.))**.** Let $\Theta \subseteq L^q \cap L^\infty$ for $q \in [1, \infty]$, thereby ensuring that the function-evaluation operation is well-defined; in Appendix F.1 we discuss the role of this assumption and offer an alternative procedure when it does not hold. In addition, we restrict our attention to linear sieves, i.e., for each $k \in \mathbb{N}$, Θ_k is the linear span of some basis functions $\kappa^k \equiv (\kappa_j)_{j=1}^k$. In view of the results in Example 4.2, we assume that there exists a positive vanishing sequence $(\epsilon_{k,n})_n$ for which, wpa1-P, $\psi_k(P_n) \in \Theta_k(\epsilon_{k,n}) \equiv \Theta_k \cap \Theta(\eta_{k,n})$ where $\Theta(\epsilon_{k,n}) \equiv \{\theta \in L^q \colon ||\theta - \psi_k(P)||_{L^q} \leq \epsilon_{k,n}\}$; henceforth, we can thus take $\Theta_k(\epsilon_{k,n})$ as the "relevant sieve" space. The next two assumptions impose smoothness restrictions on (ϕ, Pen).

Assumption 5.2. *(i) Pen is strictly convex and twice continuously differentiable; (ii) there exists a $C_0 < \infty$ and a $\varrho > 0$ such that for any $k \in \mathbb{N}$,*

$$\sup_{h \in \Theta_k(\epsilon_{k,n})} \sup_{(v_1,v_2) \in \Theta_k^2} \frac{\left| \frac{d^2 Pen(h)}{d\theta^2}[v_1, v_2] - \frac{d^2 Pen(\psi_k(P))}{d\theta^2}[v_1, v_2] \right|}{||v_1||_{L^q} ||v_2||_{L^q}} \leq C_0 ||h - \psi_k(P)||_{L^q}^{\varrho}.$$

For the following assumption, let $\mathcal{S}_{1,k}(\epsilon_{k,n}) = \left\{ \frac{d\phi(.,\theta)}{d\theta}[v/||v||_{L^q}] \colon \ (v, \theta) \in \Theta_k \times \Theta_k(\epsilon_{k,n}) \right\}$ and let $\mathcal{S}_{2,k}(\epsilon_{k,n})$ be defined analogously but using the second derivative; finally let $\Phi_{r,k,n}$ be the envelope of the class $\mathcal{S}_{r,k}(\epsilon_{k,n})$, i.e., $\Phi_{r,k,n}(.) = \sup_{f \in \mathcal{S}_{r,k}(\epsilon_{k,n})} |f(.)|$.

Assumption 5.3. *(i) $\theta \mapsto \phi(z, \theta)$ is convex and twice continuously differentiable; (ii) there exists a $\mathcal{S} \subseteq L^\infty(\mathbb{Z})$, such that for $r \in \{1, 2\}$ and all $(k, n) \in \mathbb{N}^2$, $\mathcal{S}_{r,k}(\epsilon_{k,n}) \subseteq \mathcal{S}$ and $\Phi_{2,k,n} \in \mathcal{S}$; (iii) for any $k \in \mathbb{N}$ and any $z \in \mathbb{Z}$,*

$$\sup_{h \in \Theta_k(\epsilon_{k,n})} \sup_{(v_1,v_2) \in \Theta_k^2} \frac{\left| \frac{d^2 \phi(z,h)}{d\theta^2}[v_1, v_2] - \frac{d^2 \phi(z,\psi_k(P))}{d\theta^2}[v_1, v_2] \right|}{||v_1||_{L^q} ||v_2||_{L^q}} \leq C_0 ||h - \psi_k(P)||_{L^q}^{\varrho}.$$

The class $\mathcal{S}$ imposes restrictions on the first and second derivative of ϕ as a function of z, at least it requires boundedness of these functions and it is the relevant class for constructing the norm over $ca(\mathbb{Z})$.

Proposition 5.5. Suppose Assumptions 5.2 and 5.3 hold. Then, for each $k \in \mathbb{N}$, ψ_k is $||.||_{\mathcal{S}}$-Frechet Differentiable with derivative given by,

$$D\psi_k(P)[Q] = (E_Q [\nabla_k(P)(Z)])^T \Delta_k(P)^{-1} \kappa^k, \ \forall Q \in \mathcal{T}_0,$$

where $\Delta_k(P) \equiv E_P\left[\frac{d^2\phi(Z,\psi_k(P))}{d\theta^2}[\kappa^k,\kappa^k]\right] + \lambda_k \frac{d^2 Pen(\psi_k(P))}{d\theta^2}[\kappa^k,\kappa^k] \in \mathbb{R}^{k\times k}$ and $z \mapsto \nabla_k(P)(z) \equiv \frac{d\phi(z,\psi_k(P))}{d\theta}[\kappa^k] \in \mathbb{R}^k$.

Proof. See Appendix F.1. □

This result implies that the regularization is $DIFF(P, \mathcal{E}_{||\cdot||_{\mathcal{S}}}, \Xi)$ for any bounded set Ξ in the dual of L^q, and that the influence is given by the sequence of functions $z \mapsto \varphi_k(P)(z) \equiv (\nabla_k(P)(z) - E_P[\nabla_k(P)(Z)])^T \Delta_k(P)^{-1}\kappa^k \in \Theta_k$. This result also implies a bound for $Q \mapsto \eta_{k,\ell}(Q)$ that is uniform over $||.||_{\mathcal{S}}$-bounded sets; see Appendix F.1.1 for an explicit bound. Thus, if the class $\mathcal{S}$ is P-Donsker — which means that $Q = \sqrt{n}(P_n - P)$ is a.s.-$\mathbf{P}$ a $||.||_{\mathcal{S}}$-bounded sequence (see Lemma E.1) — all the conditions in Theorem 5.2 hold.

For any, ℓ a linear functional over L^q,

$$||\ell[\varphi_k(P)]||^2_{L^2(P)} = (\ell[\kappa^k])^T \Delta_k(P)^{-1}\Sigma_k(P)\Delta_k(P)^{-1}(\ell[\kappa^k])$$

where $\Sigma_k(P) \in \mathbb{R}^{k\times k}$ is the covariance matrix of $\nabla_k(P)(Z)$.[23] Under our assumptions, for each k, $\Delta_k(P)$ is non-singular, so $||\ell[\varphi_k(P)]||^2_{L^2(P)} < \infty$. But, whether this holds uniformly over k depends on the limit behavior of the eigenvalues $\Delta_k(P)^{-1}$ and $\Sigma_k(P)$ which may diverge as k diverges.[24] By using $||\ell[\varphi_k(P)]||_{L^2(P)}$ as the scaling factor our approach adapts to either case, thus allowing the researcher to sidestep this discussion altogether.

We now provide asymptotic confidence bands for the regularized M-estimator. Since $\Theta \subseteq L^q \cap L^\infty$, the set $\Xi \equiv \{\delta_z : z \in \mathbb{Z}\}$ is a valid subset of the dual Θ^*. The following representation follows from Theorem 5.2.

Lemma 5.2. *Suppose Assumptions 5.2 and 5.3 hold and $\mathcal{S}$ is P-Donsker. Then, for any $q \in [1,\infty]$ and any $k \in \mathbb{N}$,*

$$\left\|\sqrt{n}\frac{\psi_k(P_n)-\psi_k(P)}{\sigma_k(P)} - (\kappa^k)^T\Delta_k(P)^{-1}n^{-1/2}\sum_{i=1}^{n}\frac{\nabla_k(P)(Z_i)}{\sigma_k(P)}\right\|_{L^q} = o_P(1),$$

where $\sigma_k(P) : \mathbb{Z} \to \mathbb{R}_+$ given by

$$z \mapsto \sigma_k^2(P)(z) = (\kappa^k(z))^T\Delta_k^{-1}(P)\Sigma_k(P)\Delta_k^{-1}(P)(\kappa^k(z)).$$

Proof. See Appendix F.1. □

Lemma 5.2 shows that in order to characterize the asymptotic distribution of $\left\|\sqrt{n}\frac{\psi_k(P_n)-\psi_k(P)}{\sigma_k(P)}\right\|_{L^q}$ it suffices to characterize the one of $\left\|(\kappa^k)^T\Delta_k(P)^{-1}n^{-1/2}\sum_{i=1}^n\frac{\nabla_k(P)(Z_i)}{\sigma_k(P)}\right\|_{L^q}$. The following

[23] The expression $\ell[\kappa^k]$ should be understood as ℓ applied to component-by-component to κ^k.

[24] For instance, if $E_P\left[\frac{d^2\phi(Z,\psi_k(P))}{d\theta^2}[\kappa^k,\kappa^k]\right]$ becomes singular or large k, the maximal eigenvalue of $\Delta_k(P)^{-1}$ diverges at rate λ_k^{-1}.

proposition accomplishes this by showing that the latter quantity can be approximated by a simple Gaussian process; the proof relies on coupling results (e.g. Pollard [2002]).

Proposition 5.6. Suppose Assumptions 5.2 and 5.3 hold. Then, for any $q \in [1, \infty]$ and any $k \in \mathbb{N}$,

$$\left\| \sqrt{n} \frac{\psi_k(P_n) - \psi_k(P)}{\sigma_k(P)} - \frac{(\kappa^k)^T \mathcal{Z}_k}{\sigma_k(P)} \right\|_{L^q} = O_{Pr}\left(\frac{\beta_k k}{\sqrt{n}} \left(1 + \frac{|\log\left(\frac{\sqrt{n}}{\beta_k k}\right)|}{k} \right) + r_{n,k}^{-1} \right),$$

where $\mathcal{Z}_k$ is such that $z \mapsto \frac{(\kappa^k)^T(z)\mathcal{Z}_k}{\sigma_k(P)} \sim N(0, I_k)$, $\beta_k \equiv E[||\Delta_k(P)^{-1}\nabla_k(P)(Z)||^3]$ and $r_{n,k}^{-1} \equiv n^{-1/2}\underline{e}_k(P)^{-1}C(l_n, P, k)$ where $\underline{e}_k(P) \equiv e_{min}\left(\Delta_k^{-1}(P)\Sigma_k(P)\Delta_k^{-1}(P)\right)$ and $(l_n)_n$ is any slowly diverging sequence.

Proof. See Appendix F.1. □

Proposition 5.6 provides the basis for constructing confidence bands under general L^q norms, and it also illustrates the type of restrictions on the sequence of tuning parameters that are needed to obtain these results. On the one hand, the sequence $(k_n)_n$ has to be such that $(\beta_{k_n} k_n + \underline{e}_{k_n}(P)^{-1}C(l_n, P, k_n))/\sqrt{n} = o(1)$; on the other hand, it has to be such that the approximation error is negligible, i.e., $||\sqrt{n}|\psi_{k_n}(P) - \psi(P)|/\sigma_{k_n}(P)||_{L^q} = o(1)$. If this last requirement does not hold, then Proposition 5.6 only provides a result for obtaining confidence bands of the "pseudo-true" parameter $\psi_{k_n}(P)$, which may or may not be of interest in certain applications.

Provided the aforementioned conditions hold and one has consistent estimators of $\sigma_{k_n}(P)$, $\Sigma_{k_n}(P)$ and $\Delta_{k_n}(P)$, the asymptotic distribution of $\left\|\sqrt{n}\frac{\psi_{k_n}(P_n)-\psi(P)}{\sigma_{k_n}(P)}\right\|_{L^q}$ can be approximated by the distribution of $\left\|\frac{(\kappa^k)^T\mathcal{Z}_k}{\sigma_k(P)}\right\|_{L^q}$. Belloni et al. [2015] obtained analogous results for the L^∞-norm in a linear regression model. Our methodology extends these results in two directions: general M-estimation problems and general L^q norms.[25] △

6 Conclusion

We propose an unifying framework to study the large sample properties of regularized estimators that extends the scope of the existing large sample theory for "plug-in" estimators to a large class containing regularized estimators. Our results suggest that the large sample theory for regularized estimators does not constitute a large departure from the existing large

[25] Chen and Christensen [2018] also derive confidence bands but in the NPIV model, which is beyond the scope of this example. We hope the results in the current paper could be used to extend their insights to general conditional moment models.

sample theory for "plug-in" estimators, in the sense that both are based on local properties of the mappings used for constructing the estimator. This last observation indicates that other large sample results developed for "plug-in" estimators can also be extended to the more general setting of regularized estimators; e.g., estimation of the asymptotic variance of the estimator and, more generally, inference procedure like the bootstrap. We view this as a potentially worthwhile avenue for future research.

References

D Ackerberg, X Chen, J Hahn, and Z Liao. Asymptotic efficiency of semiparametric two-step gmm. *The Review of Economic Studies*, 81(3):919–943, 2014.

C Ai and X Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

C. Ai and X. Chen. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141(1):5–43, 2007a.

C Ai and X Chen. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141(1):5–43, 2007b.

A. Ambrosetti and G. Prodi. *A primer of nonlinear analysis.* Number 34. Cambridge University Press, 1995.

D W K Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405, 2000.

A.D. Barbour and L.H.Y. Chen. *An Introduction to Stein's Method.* Lecture Notes Series - National University of Singapore, Institute for Mathematical Sciences. Singapore University Press, 2005.

A. Belloni, V. Chernozhukov, D. Chetverikov, and K. Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345 – 366, 2015.

P J Bickel and D A Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217, 1981.

P J Bickel and B. Li. Regularization in statistics. *Test*, 15(2):271–344, 2006.

P. J. Bickel and Y. Ritov. Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhya Series A*, 50(3):381–393, 1988.

P. J. Bickel and Y. Ritov. Achieving information bounds in non and semiparametric models. *The Annals of Statistics*, 18(2):925–938, 06 1990.

P. J. Bickel and M. Rosenblatt. On some global measures of the deviations of density function estimates. *The Annals of Statistics*, 1(6):1071–1095, 11 1973.

P J Bickel and A Sakov. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, 18(3):967–985, 2008.

P J Bickel, C AJ Klaassen, Y Ritov, and J A Wellner. *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag, 1998.

L. Birge and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 09 1998.

R Blundell, X Chen, and D Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

M D Cattaneo and M H Farrell. Optimal convergence rates, bahadur representation, and asymptotic normality of partitioning estimators. *Journal of Econometrics*, 174(2):127–143, 2013.

X. Chen. Large sample sieve estimation of semi-nonparametric models. volume 6 of *Handbook of Econometrics Chapter 76*, pages 5549 – 5632. 2007.

X. Chen and T. M. Christensen. Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation. *ArXiv pre-print 1508.03365*, 2015.

X Chen and T M Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84, 2018.

X. Chen and Z. Liao. Sieve m inference on irregular parameters. *Journal of Econometrics*, 182(1): 70–86, 2014.

X Chen and D Pouzo. Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–1079, 2015.

X Chen, Z Liao, and Y Sun. Sieve inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics*, 178:639–658, 2014.

J S Cho and H White. Directionally differentiable econometric models. *Econometric Theory*, pages 1–31, 2017.

S Darolles, Y Fan, J.-P. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

D. L. Donoho and R. C. Liu. Geometrizing rates of convergence, ii. *The Annals of Statistics*, 19 (2):633–667, 1991.

R M Dudley. Fréchet differentiability, p-variation and uniform donsker classes. In *Selected Works of RM Dudley*, pages 371–385. Springer, 2010.

P. P. B. Eggermont and V. N. LaRiccia. *Maximum penalized likelihood estimation*, volume I and II. Springer, 2001.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Z. Fang and A. Santos. Inference on directionally differentiable functions. *ArXiv pre-print 1404.3763*, 2014.

J.-P. Florens. *Inverse Problems and Structural Econometrics: The Example of Instrumental Variables*, volume 2 of *Econometric Society Monographs*, page 284311. Cambridge University Press, 2003.

N Fournier and A Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707, 2015.

C Gaillac and E Gautier. Adaptive estimation in the linear random coefficients model when regressors have limited variation. *arXiv preprint arXiv:1905.06584*, 2019.

E Gine and R Nickl. A simple adaptive estimator of the integrated square of a density. *Bernoulli*, 14(1):47–61, 02 2008.

E Giné and R Nickl. Uniform central limit theorems for kernel density estimators. *Probability Theory and Related Fields*, 141(3-4):333–387, 2008.

A R Hall and A Inoue. The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics*, 114(2):361–394, 2003.

P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929, 2005.

P. Hall and J.S. Marron. Estimation of integrated squared density derivatives. *Statistics and Probability Letters*, 6(2):109 – 115, 1987.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.

W. Härdle and O. Linton. Applied nonparametric methods. *Handbook of econometrics Chapter 38*, 4:2295–2339, 1994.

P. Lax. *Functional Analysis*. Wiley-Interscience, 2002.

O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, 12 1997.

W. K. Newey. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135, 1990.

W. K. Newey. Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, 10(2):233–253, 1994.

W K Newey and J L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

W. K. Newey, F. Hsieh, and J. M. Robins. Twicing kernels and a small bias property of semiparametric estimators. *Econometrica*, 72(3):947–962, 2004.

S. Pereverzev and E. Schock. On the adaptive selection of the parameter in regularization of ill-posed problems. *SIAM Journal on Numerical Analysis*, 43(5):2060–2076, 2006.

D Pollard. *A user's guide to measure theoretic probability*. Cambridge University Press, 2002.

D. Pouzo. On the Non-Asymptotic Properties of Regularized M-estimators. *ArXiv pre-print 1512.06290*, 2015.

J. L. Powell and T. M. Stoker. Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics*, 75(2):291 – 316, 1996.

J. L. Powell, J. H. Stock, and T. M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, pages 1403–1430, 1989.

T. A. Severini and G. Tripathi. Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. *Journal of Econometrics*, 170(2):491 – 498, 2012.

A Shapiro. On concepts of directional differentiability. *Journal of optimization theory and applications*, 66(3):477–487, 1990.

X Shen and W H Wong. Convergence rate of sieve estimates. *The Annals of Statistics*, 22(2): 580–615, 1994.

C. Stein. Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 1956.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 1996.

C Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

H. White and J. Wooldridge. Some results on sieve estimation with dependent observations. In W.A. Barnett, J. Powell, and G. Tauchen, editors, *Non-parametric and Semi-parametric Methods in Econometrics and Statistics*. 1991.

J. Wolfowitz. The minimum distance method. *The Annals of Statistics*, 28(1):75–88, 1957.

Table of Contents of the Appendix

Notation: Recall that $ca(X)$ for some set X is the Banach space of all Borel measures over X endowed with the the total variation norm, $||\mu||_{TV} = |\mu|(X)$ where $|.|$ is the total variation. For a real-valued sequence $(x_n)_n$, $x_n \uparrow a \in \mathbb{R} \cup \{\infty\}$ means that the sequence is non-decreasing and its limit is a; $x_n \downarrow a$ is defined analogously.

A Extensions of our Setup

In this Appendix we briefly discuss how to extend our theory to general stationary models (Section A.2), we also discuss how to extend our setup to capture some sample splitting procedure commonly used in the literature (Section A.1).

A.1 Sample-Splitting Procedures

Our regularized estimator — like the plug-in one — is defined in terms of P_n, and as such is permutation invariant. Thus, estimators that do not enjoy this property are not covered by our setup; perhaps the most notable class of estimators that falls in this category are estimators that rely on sample splitting procedures. We now argue that a slight generalization of our framework can encompass some splitting-sample procedures.

In order to illustrate the challenges and proposed solutions that arise from these procedures, we present the problem in a simple canonical example. Suppose the parameter of interest is comprised of two quantities: a vector, denoted as $h \in \mathcal{H}$ ($\mathcal{H}$ being some subset of a Euclidean space), and a real number, denoted as $\theta \in \mathbb{R}$. The former should be treated as a so-called "nuisance parameter" and the latter as the parameter of interest. Moreover the following "triangular structure" holds:

$$P \mapsto \psi(P) = (\theta(P, h(P)), h(P)), \tag{17}$$

where $P \mapsto h(P) \in \mathcal{H}$ is the mapping identifying the nuisance parameter and $(P, h) \mapsto \theta(P, h) \in \mathbb{R}$ is the mapping identifying the parameter of interest. The "triangular structure" means that h only depends on P whereas θ depends on both P and $h(P)$. An example of this structure is one where $\theta(P, h) = E_P[\phi(Z, h)]$ where ϕ is a known function that depends on the data Z but also the nuisance parameter.

Suppose the following estimator is considered. The data is divided in halves (for simplicity we assume the sample size to be even). An estimator, denoted as $\hat{\psi}_1$, is constructed by using the first half to construct an estimator of h and using this estimator and the second half of the sample to construct the estimator for θ. Another estimator, denoted as $\hat{\psi}_2$, is constructed by reversing the role of the first and second halves of the sample. The final estimator is simply $\hat{\psi} = 0.5\hat{\psi}_1 + 0.5\hat{\psi}_2$. To keep the setup as simple as possible, we assume, for now, that the plug-in estimator is used (within each sub-sample) for estimating both θ and h, i.e., there is no need to regularized the problem.

It is easy to see that $\hat{\psi}$ is not permutation invariant and thus does not fall in our framework. We now propose an alternative formulation of the original problem that, while seemingly redundant and even contrive at first glance, will allow us to extend our framework to this problem. This formulation entails thinking of ψ as a function of *two* probability distributions over $\mathbb{Z}$. Formally, $\bar{\psi} : \mathcal{M} \times \mathcal{M} \to \mathbb{R} \times \mathcal{H}$, where

$$\bar{\psi}(P_1, P_2) = (\theta(P_1, h(P_2)), h(P_2)). \tag{18}$$

At the population level this distinction is superfluous because, if the true probability is given by P, then $\psi(P) = \bar{\psi}(P, P)$. However, by taking $\bar{\psi}$ as the parameter mapping, the split-sample estimator can be formulated as follows. Let $P_n^{(1)}$ be the empirical distribution generated by the first half of the sample and $P_n^{(2)}$ be the empirical distribution generated by the second half of the sample. It follows that

$$\hat{\psi} = 0.5\bar{\psi}(P_n^{(1)}, P_n^{(2)}) + 0.5\bar{\psi}(P_n^{(2)}, P_n^{(1)}).$$

That is, the split-sample estimator can be seen as weighted average of two *plug in estimators using the parameter mapping* $\bar{\psi}$. Since the estimators $(P_n^{(1)}, P_n^{(2)})$ will converge to (P, P) under the same conditions that ensure convergence of P_n to P (except in the former case the relevant sample size is $n/2$ not n), then one can establish consistency and asymptotic linearity by using the typical results for plug-in estimators, but using $\bar{\psi}$ as the original parameter mapping, and not ψ.

The formulation using $\bar{\psi}$ allow us to tackle the case in which the estimation problem for h or θ needs to be regularized; e.g. if h is a function or a high-dimensional vector. We do this by proposing a regularization — in the sense of Definition 3.1 — *for* $\bar{\psi}$ *as opposed to* ψ, and construct the regularized estimator as

$$0.5\bar{\psi}_k(P_n^{(1)}, P_n^{(2)}) + 0.5\bar{\psi}_k(P_n^{(2)}, P_n^{(1)}), \ \forall k \in \mathbb{N}.$$

Thus our results can be applied to this case, by taking the regularization to be $(\bar{\psi}_k)_k$. For instance, to establish consistency, following Theorem 4.1, it suffices to verify continuity of $(\bar{\psi}_k)_k$.

The example given by expression (17) has 3 features that we believe are key in order to extend our general theory for regularized estimators to encompass sample-splitting procedures. We now extrapolate these feature from this simple canonical example to a more general setup

1. The number of splits in the sample is fixed, in the example was 2, in general it can be $s \in \mathbb{N}$ but s is assumed not to grow with n. Following the insight in expression 18, the new parameter is given by $\bar{\psi}(P_1, ..., P_s)$ where $P_1, ..., P_s$ belong to the model $\mathcal{M}$. Moreover, assuming, for simplicity, that $n = sm$ for some $m \in \mathbb{N}$, it also follows that one can construct a vector $P_n^{(1)}, ..., P_n^{(s)}$ of empirical probability distributions, one for each sub-sample.

2. The estimation procedure within each sub-sample admits a regularization as defined in our paper. That is, there exists a sequence $(\bar{\psi}_k)_k$ such that $\bar{\psi}_k(P_n^{(\pi_1)}, ..., P_n^{(\pi_s)})$ is well-defined for each permutation $\pi_1, ..., \pi_s$ of $\{1, ..., s\}$, and $\bar{\psi}_k(P, ..., P)$ converges to $\bar{\psi}(P, ..., P) = \psi(P)$ for each $P \in \mathcal{M}$.

3. The final estimator is a convex combination of the estimators $\bar{\psi}_k(P_n^{(\pi_1)}, ..., P_n^{(\pi_s)})$. For instance, if $s = 3$, then the final estimator is of the form $\sum_{i,j,k \in \{1,..,3\}} w_{i,j,k}\bar{\psi}_k(P_n^{(i)}, P_n^{(j)}, P_n^{(k)})$ where $(w_{i,j,k})_{i,j,k}$ are given weights. This last assumption is, in our opinion, less critical than the other two since we conjecture the convex combination can be replaces by a "smooth" operator.

We believe these features are general enough to encompass the sample-splitting procedures commonly used in applications They, however, do rule out cases where the sample splitting procedure demands number of splits that grow with the sample size.

A.2 Extension to General Stationary Models

We now briefly discuss how to extend our theory to general stationary models. In this case a **model** is a family of stationary probability distributions over $\mathbb{Z}^\infty$, i.e., a subset of $\mathcal{P}(\mathbb{Z}^\infty)$ (the set of *stationary* Borel probability distributions over $\mathbb{Z}^\infty$).

Let P denote the marginal distribution over Z_0 corresponding to $P \in \mathcal{P}(\mathbb{Z}^\infty)$ (by stationarity, the time dimension is irrelevant). For a given model $\mathcal{M}^\infty$, let $\mathcal{M}$ denote the set of marginal probability distribution over Z_0 corresponding to $\mathcal{M}^\infty$. A **parameter on model** $\mathcal{M}^\infty$ is a mapping from $\mathcal{M}$ to Θ. That is, we restrict attention to mappings that depend only on the marginal distribution. Our theory can also be extended to cases where ψ depends on the joint distribution of a *finite* sub-collections of $\mathbb{Z}^\infty$. Allowing for the mapping to depend on the entire $\mathbf{P}$ is mathematical possible, but such object is of little relevance since it cannot be estimated from the data.

A regularization of a parameter ψ is defined analogously and the (relevant) empirical distribution is given, for each $\boldsymbol{z} \in \mathbb{Z}^\infty$, by $P_n(A) \equiv n^{-1}\sum_{i=1}^n 1\{\boldsymbol{z} \colon Z_i(\boldsymbol{z}) \in A\}$ for any Borel set $A \subseteq \mathbb{Z}$.

Theorem 4.1 can be applied to this setup essentially without change, the difference with the i.i.d. setup lies on how to establish converges of P_n to P under d. Similarly, the notion of differentiability (Definition 5.2) can also be applied without change. The influence function will also be given by $z \mapsto D\psi_k(\mathbf{P})[\delta_z - P]$. The scaling, however, will be different, since

$$
\begin{aligned}
E_{\mathbf{P}}\left[\left(\sqrt{n} D\psi_k(\mathbf{P})[P_n - P]\right)^2\right] =& E_{\mathbf{P}}\left[\left(n^{-1/2}\sum_{i=1}^n D\psi_k(\mathbf{P})[\delta_{Z_i} - P]\right)^2\right] \\
=& ||\varphi_k(\mathbf{P})||^2_{L^2(P)} \\
& + 2n^{-1}\sum_{i<j} E_{\mathbf{P}}\left[\left(D\psi_k(\mathbf{P})[\delta_{Z_i} - P]\right)\left(D\psi_k(\mathbf{P})[\delta_{Z_j} - P]\right)\right] \\
=& ||\varphi_k(\mathbf{P})||^2_{L^2(P)} + 2n^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \gamma_{j-i,k}(\mathbf{P}) \\
\equiv& ||\varphi_k(\mathbf{P})||^2_{L^2(P)}(1 + 2\Phi_{n,k}(\mathbf{P}))
\end{aligned}
$$

where $\gamma_{j,k}(\mathbf{P}) \equiv E_{\mathbf{P}}\left[\left(D\psi_k(\mathbf{P})[\delta_{Z_0} - P]\right)\left(D\psi_k(\mathbf{P})[\delta_{Z_j} - P]\right)\right]$ and

$$
\Phi_{n,k}(\mathbf{P}) \equiv \sum_{i=1}^{n-1}\left(1 - \frac{i}{n}\right)\frac{\gamma_{i,k}(\mathbf{P})}{\gamma_{i,0}(\mathbf{P})}.
$$

Hence, the natural scaling is $||\varphi_k(\mathbf{P})||_{L^2(P)}\sqrt{(1 + 2\Phi_{n,k}(\mathbf{P}))}$ and not $||\varphi_k(\mathbf{P})||_{L^2(P)}$ as in the IID case. We note that our theory, a priori, does not require $\limsup_{n\to\infty} \Phi_{n,k}(\mathbf{P}) = \infty$.

In view of the previous discussion, the relevant restriction in Theorem 5.1 is

$$\sqrt{n}\frac{\eta_k(P_n - P)}{||\varphi_k(\mathbf{P})||_{L^2(P)}\sqrt{(1 + 2\Phi_{n,k}(\mathbf{P}))}} = o_{\mathbf{P}}(1).$$

An analogous amendment applies to Theorem 5.2.

B Appendix for Section 3

The next lemma formalize verifies Claims 1-3 and 1'-3' in the text. Throughout, let $\rho_k(\cdot) = k\rho(k\cdot)$ for any $k \in \mathbb{K}$.

Lemma B.1. *For all $h > 0$, $t \mapsto (\rho_{1/h} \star \rho_{1/h})(t) = (\rho \star \rho)_{1/h}(t)$.*

Proof. For all $t \in \mathbb{R}$,

$$\begin{aligned}(\rho_{1/h} \star \rho_{1/h})(t) = \int \rho_{1/h}(t-x)\rho_{1/h}(x)dx =& h^{-2}\int \rho((t-x)/h)\rho(x/h)dx = h^{-1}\int \rho(u)\rho(t/h-u)du \\ =& h^{-1}(\rho \star \rho)(t/h)\end{aligned}$$

where the last line follows from symmetry of ρ. □

Lemma B.2. *Claims 1-3 and 1'-3' in the text hold.*

Proof. For each case 1-3 and 1'-3', we show that the κ yields the associated estimators and that κ is a valid choice in each case.

(1) Follows directly from the fact that $\rho_{1/h} \star P_n = \hat{p}_{1/h}$.

(2) By Lemma B.1, $t \mapsto (\rho_{1/h} \star \rho_{1/h})(t) = h^{-1}(\rho \star \rho)(t/h)$. Hence, by taking $\kappa = \rho \star \rho$ it follows that $t \mapsto \kappa_{1/h}(t) = h^{-1}(\rho \star \rho)(t/h) = (\rho_{1/h} \star \rho_{1/h})(t)$. Moreover, κ is indeed a pdf, symmetric and continuously differentiable.

We now show the form of the implied estimator. We use the notation $\langle .,. \rangle$ to denote the dual inner product between $L^\infty(\mathbb{R})$ and $ca(\mathbb{R})$, so

$$\begin{aligned}\int (\kappa_{1/h} \star P)(x)P(dx) =& \langle \rho_{1/h} \star \rho_{1/h} \star P, P\rangle = \int\int \rho_{1/h}(x-y)(\rho_{1/h} \star P)(y)dyP(dx) \\ =& \int (\rho_{1/h} \star P)(y) \int \rho_h(y-x)P(dx)dy \\ =& \langle \rho_{1/h} \star P, \rho_{1/h} \star P\rangle_{L^2}\end{aligned}$$

where the second line follows by symmetry of ρ. Since $\rho_{1/h} \star P_n = \hat{p}_h$ the result follows.

(3) Take $\kappa(\cdot) \equiv (-\rho\star\rho(\cdot)+2\rho(\cdot))$. It follows that $\int \kappa(u)du = -\int \rho\star\rho(u)du+2\int \rho(u)du = 1$. Smoothness follows from smoothness of ρ. Finally, we note that one can write $\kappa(t)$ as $\{\rho \star \rho(t) + 2(\rho(t) - \rho \star \rho(t))\}$.

By Lemma B.1 $t \mapsto \kappa_{1/h}(t) = h^{-1}(\rho\star\rho)(t/h)+2h^{-1}(\rho(t/h)-\rho\star\rho(t/h)) = (\rho_{1/h}\star\rho_{1/h})(t)+ 2(\rho_{1/h}(t) - \rho_{1/h} \star \rho_{1/h}(t))$. So the expression of the estimator follows from simple algebra.

(1') Since P does not have atoms, $Z_i = Z_j$ iff $i = j$ a.s.-$\mathbf{P}$. It follows that the estimator is given by $n^{-2}\sum_{i,j} \kappa_{1/h}(Z_i - Z_j) = n^{-1}\kappa_{1/h}(0) + n^{-2}\sum_{i\neq j} \kappa_{1/h}(Z_i - Z_j)$ a.s.-$\mathbf{P}$ and the result follows since $\kappa(0) = 0$.

(2') The expression of the estimator follows from analogous calculations to those in 1'.

(3') By the calculations in (3)

$$\begin{aligned}\int (\hat{p}_h(z))^2 dz = \int (\rho_{1/h} \star \rho_{1/h} \star P_n)(x) P_n(dx) &= n^{-2}\sum_{i\neq j} \rho_{1/h} \star \rho_{1/h}(Z_i - Z_j) + n^{-1}\rho_{1/h} \star \rho_{1/h}(0) \\ &= n^{-2}\sum_{i\neq j} \rho_{1/h} \star \rho_{1/h}(Z_i - Z_j) + n^{-1}\int (\rho_{1/h}(z))^2 dz\end{aligned}$$

where the last line follows by symmetry. Hence

$$\begin{aligned}\int (\hat{p}_h(z))^2 dz - 2\int (\hat{p}_h(z))^2 dz + n^{-1}\int (\rho_{1/h}(z))^2 dz &= -n^{-2}\sum_{i\neq j} \rho_{1/h} \star \rho_{1/h}(Z_i - Z_j) \\ &= -n^{-2}\sum_{i,j} \rho_{1/h} \star \rho_{1/h}(Z_i - Z_j) \times 1\{Z_i - Z_j \neq 0\}\end{aligned}$$

where the last line follows because P does not have atoms, so $Z_i = Z_j$ iff $i = j$ a.s.-$\mathbf{P}$.

Similarly,

$$\begin{aligned}2n^{-1}\sum_{i=1}^{n} \hat{p}_h(Z_i) - 2\rho_{1/h}(0)/n &= 2\left(n^{-2}\sum_{i,j} \rho_{1/h}(Z_i - Z_j) - \rho_{1/h}(0)/n\right) \\ &= 2n^{-2}\sum_{i\neq j} \rho_{1/h}(Z_i - Z_j) \\ &= 2n^{-2}\sum_{i,j} \rho_{1/h}(Z_i - Z_j) \times 1\{Z_i - Z_j \neq 0\}.\end{aligned}$$

□

Proof of Proposition 3.1. Since $P \in \mathcal{M}$ it admits a smooth pdf, p, it follows that

$$\begin{aligned}\psi_k(P) - \psi(P) &= \int \left(\int \kappa_k(x-y)p(y)dy - p(x) \right) p(x)dx \\ &= \int \int \kappa(u)p(x-u/k)du - p(x)p(x)dx \\ &= \int (\bar{p} \star p(u/k) - \bar{p} \star p(0)) \, \kappa(u)du\end{aligned}$$

where $t \mapsto \bar{p}(t) \equiv p(-t)$. Henceforth, let $t \mapsto g(t) \equiv \bar{p} \star p(t)$.

Our condition (2) implies that p and $\bar{p}$ belong to the Besov space $\mathcal{B}^{\varrho}_{2,\infty}(\mathbb{R})$. Lemma 12 in Giné and Nickl [2008] implies that $g \in \mathcal{B}^{2\varrho}_{\infty,\infty}(\mathbb{R})$, in fact since $2\varrho \notin \mathbb{N}$, g is Hölder continuous with parameter 2ϱ. This implies and the previous display imply that $B_k(P) \leq Ck^{-2\varrho} \int |u|^{2\varrho} |\kappa(u)| du$ for some universal constant $C < \infty$.

□

Remark B.1 (Remarks about the Condition 2)**.** Gine and Nickl [2008] imposes $p \in H^{\varrho}_2(\mathbb{R})$, whereas our restriction essentially implies that $p \in \mathcal{B}^{\varrho}_{2,\infty}(\mathbb{R})$. In that paper and in ours the smoothness coefficient ϱ is less than 0.5, i.e., we have "low" degree of smoothness. Because of this, whether or not the kernel is a "twicing kernel" does not matter for the control of the approximation error. For larger levels of smoothness, e.g. $\varrho > 1$, we expect the "twicing kernel" — or higher order kernels in general — to yield different bounds for the approximation error. The goal of this example is to illustrate the scope of our methodology and thus we decided to stay as closed as possible to the existing literature and omit the case $\varrho > 0.5$. △

B.1 Some Remarks on the Regularization Structure in the NPIV Example.

The general regularization structure, $(\mathcal{R}_{k,P}, T_{k,P}, r_{k,P})_{k \in \mathbb{N}}$, and conditions 1-2 are taken from Engl et al. [1996] Ch. 3-4. It is clear from the problem that

$$\mathbb{D}_{\psi} = \{\mu \in ca(\mathbb{R} \times [0,1]^2) \colon E_{\mu}[|Y|^2] < \infty \; and \; E_{\mu}[|h(W)|^2] < \infty \; \forall h \in L^2([0,1])\}. \tag{19}$$

The next lemma presents useful properties of $\mathbb{D}_{\psi}$. The proof is straightforward and thus omitted.

Lemma B.3. *(1) $\mathbb{D}_{\psi} \supseteq \mathcal{M} \cup \mathcal{D}$; (2) $\mathbb{D}_{\psi}$ is a linear subspace.*

We now discuss canonical examples of regularizations methods for the first and second stage that we consider in this paper.

First Stage Regularization. For any $P \in \mathbb{D}_{\psi}$ and any $k \in \mathbb{N}$, we can generically write $r_{k,P}$ as

$$r_{k,P}(x) \equiv \int y \int U_k(x', x) P(dy, dx'), \ \forall x \in [0,1],$$

where $U_k \in L^{\infty}([0,1]^2)$ symmetric. For instance, if

$$(x', x) \mapsto U_k(x', x) = ku(k(x - x'))$$

where u is a symmetric around 0, smooth pdf, then $x \mapsto r_{k,P}(x) = \int y \int ku(k(x-x'))P(dy, dx')$, which is the the so-called kernel-based approach; e.g., for ill-posed inverse problems see Hall and Horowitz [2005] among others.

In the case one defined r_P using conditional probabilities, i.e., $r_P(x) = \int yp(y|x)dy$. The kernel approach becomes

$$x \mapsto r_{k,P}(x) = \int y \frac{\int ku(k(x - x'))P(dy, dx')}{\int ku(k(x - x'))P(dx')};$$

(e.g. Darolles et al. [2011]). Observe that r_P is only defined for probability measures for which the pdf exists.

Another approach is to directly set

$$(x', x) \mapsto U_k(x', x) = (u^k(x))^T Q_{uu}^{-1} u^k(x'),$$

where $(u_k)_{k\in\mathbb{N}}$ is some basis function in $L^2([0,1])$ and $Q_{uu} \equiv E_{Leb}[(u^k(X))(u^k(X))^T]$. In this case, $x \mapsto r_{k,P}(x) = (u^k(x))^T Q_{uu}^{-1} E_P[u^k(X)Y]$, which is the so-called series-based approach; e.g., for ill-posed inverse problems see Ai and Chen [2003], Newey and Powell [2003] among others.

Analogously, one can define $T_{k,P}$ as

$$g \mapsto T_{k,P}[g](x) \equiv \int g(w) \int U_k(x', x) P(dw, dx'), \ \forall x \in [0,1],$$

and the same observations above applied to this case.

The next lemma characterizes the adjoint for any $P \in \mathcal{M}$ (i.e., P as a pdf p). In this case, we can view the regularization as an operator acting on $T_P[g](x) = \int g(w)p(w,x)dw$, given by $\mathcal{U}_k : L^2([0,1]) \to L^2([0,1])$, where $\mathcal{U}_k T_P[g](x) \equiv \int U_k(x', x) \int g(w)p(w, x')dwdx'$.

Lemma B.4. *For any $k \in \mathbb{N}$ and any $P \in \mathcal{M}$ (in particular, it admits a pdf p), the adjoint of $T_{k,P}$ is $T^*_{k,P} : L^2([0,1]) \to L^2([0,1])$ and is given by*

$$f \mapsto T^*_{k,P}[f] = T^* \mathcal{U}_k[f].$$

Proof. For any $k \in \mathbb{N}$ and any $P \in \mathcal{M}$,

$$\begin{aligned}\langle T_{k,P}[g], f\rangle_{L^2([0,1])} =& \int (\mathcal{U}_k T_P[g](x)) f(x)dx \\ =& \int g(w) \int \int U_k(x',x) f(x) dx p(w,x') dx' dw \\ =& \langle g, T_P^* \mathcal{U}_k[f]\rangle_{L^2([0,1])}\end{aligned}$$

for any $g, f \in L^2([0,1])$. □

If $P \notin \mathcal{M}$, in particular if it does not have a pdf (with respect to Lebesgue), the adjoint operator is different; the reason being that T_P^* does not map onto a space of functions because P does not have a pdf. In this case, consider the operator $A_P : L^2([0,1]) \to ca([0,1])$ given by $f \mapsto A_P[f](B) = \int_{w\in B} \int f(x) P(dw, dx)$ for any $B \subseteq [0,1]$ Borel. Note that $|A_P[f](.)| \leq \int |f(x)| P(dx) < \infty$ provided that $f \in L^2(P)$, which is the case for any $P \in \mathbb{D}_\psi$. The next lemma characterizes the adjoint in this case.

Lemma B.5. *For any $k \in \mathbb{N}$ and any $P \in \mathbb{D}_\psi$, the adjoint of $T_{k,P}$ is given by*

$$f \mapsto T_{k,P}^*[f] = A_P \mathcal{U}_k[f].$$

Since $U_k \in L^2([0,1]^2)$, $T_{k,P}^*[f]([0,1]) \precsim ||f||_{L^2(P)} ||U_k||_{L^2([0,1]^2)}$ which is finite for $P \in \mathbb{D}_\psi$. So $T_{k,P}^*[f]$ in fact maps to $ca([0,1])$.

Proof. For any $k \in \mathbb{N}$ and any $P \in \mathbb{D}_\psi$,

$$\begin{aligned}\langle T_{k,P}[g], f\rangle_{L^2([0,1])} =& \int \int g(w) \int U_k(x',x) P(dw, dx') f(x) dx \\ =& \int g(w) \left(\int U_k(x',x) f(x) dx \right) P(dw, dx') \\ =& \int g(w) \int \mathcal{U}_k[f](x') P(dw, dx'),\end{aligned}$$

for any $g, f \in L^2([0,1])$. □

One possibility to avoid the aforementioned technical issue with the adjoint operator is to define a regularization given by

$$g \mapsto T_{k,P}[g](x) \equiv \int g(w) \left\{ \int U_k(x',x) V_k(w',w) P(dw',dx') \right\} dw, \ \forall x \in [0,1],$$

where $U_k \in L^\infty([0,1]^2)$ symmetric. For example, if $V_k(w',w) = h_k^{-1} v((w'-w)/h_k$ (and U_k is also given by the kernel-based approach), then in this case $(x,w) \mapsto \mathcal{W}_k[P](x,w) \equiv \int U_k(x',x) V_k(w',w) P(dw',dx')$ is a pdf over $[0,1]^2$ (regardless of whether P has a pdf or not), and thus

$$f \mapsto T^*_{k,P}[f](w) = \int f(x)\mathcal{W}_k[P](x,w)dx.$$

For instance, Hall and Horowitz [2005] considered a method akin to this.

In the case T_P is defined as a conditional operator, one can consider the sieve-based approach for U_k and $V_k(w',w) = (v^k(w))^T Q_{vv}^{-1} v^k(w')$ for some $(v_k)_{k\in\mathbb{N}}$ basis function in $L^2([0,1])$. Then, in this case,

$$\begin{aligned}
T_{k,P}[g](x) =& (u^k(x))^T Q_{uu}^{-1} E_P[u^k(X)(v^k(W))^T Q_{vv}^{-1} E_{Leb}[v^k(W)g(W)]] \\
=& (u^k(x))^T Q_{uu}^{-1} Q_{uv} Q_{vv}^{-1} E_{Leb}[v^k(W)g(W)] \\
=& \int g(w)\left\{\int (u^k(x))^T Q_{uu}^{-1} u^k(x')(v^k(w'))^T Q_{vv}^{-1} v^k(w) P(dw',dx')\right\} dw
\end{aligned}$$

where $Q_{vv} \equiv E_{Leb}[(v^k(W))(v^k(W))^T]$ and $Q_{uv} \equiv E_P[u^k(X)(v^k(W))^T]$, and

$$\begin{aligned}
f \mapsto T^*_{k,P}[f](w) =& (v^k(w))^T Q_{vv}^{-1} Q_{uv}^T Q_{uu}^{-1} E_{Leb}[u^k(X)f(X)] \\
=& \int f(x)\mathcal{W}_k[P](x,w)dx
\end{aligned}$$

where $\mathcal{W}_k[P](x,w) = \int (u^k(x))^T Q_{uu}^{-1} u^k(x')(v^k(w'))^T Q_{vv}^{-1} v^k(w) P(dw',dx')$.

Second Stage Regularization. For the second stage regularization, one widely used approach is the so-called Tikhonov- or Penalization-based approach, given by solving

$$\arg\min_{\theta\in\Theta}\{E_P\left[(r_{k,P}(X) - T_{k,P}[\theta](X))^2\right] + \lambda_k||\theta||^2_{L^2([0,1])}\}$$

which is non-empty and a singleton. This specification implies that

$$\mathcal{R}_{k,P} = (T^*_{k,P}T_{k,P} + \lambda_k I)^{-1},$$

which is well-known to be well-defined, i.e., 1-to-1 and bounded for any $\lambda_k > 0$.

Another widely used approach is the sieve-based approach that consists on setting up

$$\arg\min_{\theta\in\Theta_k} E_P\left[(r_{k,P}(X) - T_{k,P}[\theta](X))^2\right]$$

and specifie the $(\Theta_k)_k$ such that (1) $\cup_k\Theta_k$ is dense in Θ and Θ_k has dimension k, and (2) $\arg\min$ exists and is a singleton. For instance if Θ_k is convex, then a solution exists and is unique provided that $Kernel(T_{k,P}|\Theta_k) = \{0\}$. In this case

$$\mathcal{R}_{k,P} = (\Pi_k^* T^*_{k,P} T_{k,P}\Pi_k)^{-1},$$

where Π_k is the projection onto Θ_k.

Verification of Definition 3.1. The next Lemma shows that given Conditions 1-2 listed in Example 3.2, $(\psi_k(P))_{k\in\mathbb{N}}$ (and hence $(\gamma_k(P))_{k\in\mathbb{N}}$) is in fact a regularization.

Lemma B.6. *Suppose Conditions 1-2 listed in Example 3.2 hold. Then $(\psi_k(P))_{k\in\mathbb{N}}$ (and hence $(\gamma_k(P))_{k\in\mathbb{N}}$) is a regularization with $\mathbb{D}_\psi$ given in 19.*

Proof. Condition 1 in Definition 3.1 is satisfied by Lemma B.3. Regarding condition 2 in Definition 3.1, note that

$$||\psi_k(P) - \psi(P)||_{L^2([0,1])} \leq ||\mathcal{R}_{k,P}T^*_{k,P}[r_{k,P} - r_P]||_{L^2([0,1])} + ||(\mathcal{R}_{k,P}T^*_{k,P} - (T^*_PT_P)^{-1}T^*_P)[r_P]||_{L^2([0,1])},$$

which vanishes as k diverges by our conditions 1-2. □

C Appendix for Section 4.1

The next lemma provides an useful "diagonalization argument" that is used throughout the paper.

Lemma C.1. *Let $S = \{k_1, k_2, ...\}$ with $k_i < k_{i+1}$ for all $i \in \mathbb{N}$. Take a real-valued sequence $(x_{k,n})_{k\in S,n\in\mathbb{N}}$ such that, for each $k \in \mathbb{N}$, $\lim_{n\to\infty} |x_{k,n}| = 0$. Then, there exists a mapping $n \mapsto k(n) \in S$ such that (a) $\lim_{n\to\infty} |x_{k(n),n}| = 0$ and (b) $k(n) \uparrow \infty$.*

Proof. By pointwise convergence of the sequence $(x_{k,n})_n$, for any $l \in \mathbb{N}$, there exists a $n(l) \in \mathbb{N}$ such that $|x_{k_l,n}| \leq 1/2^{k_l}$ for all $n \geq n(l)$. WLOG we take $n(l+1) > n(l)$.

We now construct the mapping $n \mapsto k(n)$ as follows: For each $l \in \mathbb{N}$, let $k(n) \equiv k_l$ for all $n \in \{n(l)+1, ..., n(l+1)\}$; and $k(n) = 0$ for $n \in \{0, ..., n(0)\}$. Since the cutoffs $n(.)$ are increasing the set $\{n(l)+1, ..., n(l+1)\}$ is non-empty for each l. For integer $L > 0$, $k(n) > k_L$ for all $n \geq n(L)+1$; since $(k_l)_l$ diverges, (b) follows.

To show (a), for any $\epsilon > 0$ take l_ϵ such that $1/2^{k_{l_\epsilon}} \leq \epsilon$. Observe that for any $n \geq n(l_\epsilon)+1$, $|x_{k(n),n}| \leq 1/2^{l_\epsilon} \leq \epsilon$ by construction of $(n, k(n))$. Thus, (a) follows. □

Proof of Lemma 4.1. For any $k \in \mathbb{K}$ and any $n \in \mathbb{N}$, by continuity it follows $||\psi_k(P_n(\boldsymbol{Z})) - \psi_k(P)||_\Theta \leq \delta_k(d(P_n(\boldsymbol{Z}), P))$ a.s.-$\mathbf{P}$. In what follows, we omit the dependence on $\mathbf{Z}$.

So it suffices to show that there exists a diverging $(k_n)_n$ such that for any $\epsilon > 0$, there exists a $N \in \mathbb{N}$ and a $M > 0$ such that

$$\sup_{k\in\mathbb{K}} \mathbf{P}\left(\delta_k(d(P_n, P)) \geq \delta_k(Mr_n^{-1})\right) \leq \epsilon$$

for all $n \geq N$.

Since $t \mapsto \delta_k(t)$ is non-decreasing for all $k \in \mathbb{K}$, it follows that $\mathbf{P}\left(\delta_k(d(P_n, P)) \geq \delta_k(Mr_n^{-1})\right) \leq \mathbf{P}\left(d(P_n, P) \geq Mr_n^{-1}\right)$ for any $(n, k) \in \mathbb{N} \times \mathbb{K}$ and any $M > 0$. Hence,

$$\sup_{k\in\mathbb{K}} \mathbf{P}\left(\delta_k(d(P_n, P)) \geq \delta_k(Mr_n^{-1})\right) \leq \mathbf{P}\left(d(P_n, P) \geq Mr_n^{-1}\right)$$

for any $n \in \mathbb{N}$ and any $M > 0$.

By assumption, $r_n d(P_n, P) = O_P(1)$. This fact and the previous inequality imply the desired result. □

Proof of Theorem 4.1. Fix any $\epsilon > 0$. By the triangle inequality and laws of probability, for any $(k, n) \in \mathbb{N} \times \mathbb{K}$,

$$\mathbf{P}\left(||\psi_k(P_n) - \psi(P)||_\Theta \geq \epsilon\right) \leq \mathbf{P}\left(||\psi_k(P_n) - \psi_k(P)||_\Theta \geq 0.5\epsilon\right) + 1\{||\psi_k(P) - \psi(P)||_\Theta \geq 0.5\epsilon\}.$$

By assumption, there exists a $(r_n)_{n\in\mathbb{N}}$ such that $r_n^{-1} = o(1)$ and $d(P_n, P) = O_P(r_n^{-1})$. Thus, by Lemma 4.1, there exists a $N \in \mathbb{N}$ and a $M > 0$ such that for all $k \in \mathbb{K}$ and all $n \geq N$,

$$\mathbf{P}\left(||\psi_k(P_n) - \psi(P)||_\Theta \geq \epsilon\right) \leq \epsilon + 1\{\delta_k(Mr_n^{-1}) \geq 0.5\epsilon\} + 1\{||\psi_k(P) - \psi(P)||_\Theta \geq 0.5\epsilon\}.$$

Observe that for each k, $\delta_k(Mr_n^{-1}) = o(1)$. Since $\mathbb{K}$ is unbounded it contains a diverging increasing sequence, therefore, by Lemma C.1, there exists a diverging $(k_n)_{n\in\mathbb{N}}$ such that $\delta_{k_n}(Mr_n^{-1}) = o(1)$. This result, condition 2 in the definition of regularization and the previous display at $k = k_n$, imply that

$$\limsup_{n\to\infty} \mathbf{P}\left(||\psi_{k_n}(P_n) - \psi(P)||_\Theta \geq \epsilon\right) \leq \epsilon.$$

Finally, we show that $\delta_{k_n}(d(P_n, P)) = o_P(1)$. Since $t \mapsto \delta_k(t)$ is non-decreasing for all $k \in \mathbb{K}$ and $d(P_n, P) = O_P(r_n^{-1})$ it follows that $\mathbf{P}\left(\delta_{k_n}(d(P_n, P)) \geq \delta_{k_n}(Mr_n^{-1})\right) \leq \epsilon$ for all $n \geq N'$ (WLOG we take $N' = N$). Since $\delta_{k_n}(Mr_n^{-1}) = o(1)$ the result follows. □

D Appendix for Section 4.2

Observe that the set $\mathcal{L}_n$ is random. To stress this dependence, with some abuse of notation, we will sometimes use $\mathcal{L}_n(\mathbf{z})$ to denote the set.

D.1 Proof of Theorem 4.2

The next lemma provides two sufficient conditions that ensure the result in Theorem 4.2. To do this, for any $n \in \mathbb{N}$, let

$$D_n \equiv \{\boldsymbol{z} \in \mathbb{Z}^\infty : d(P_n(\boldsymbol{z}), P) \leq r_n^{-1}\}.$$

Lemma D.1. *Suppose there exists a sequence $(j_n)_{n\in\mathbb{N}}$ such that*

1. *For any $\epsilon > 0$, there exists a N such that $\mathbf{P}\left(\{\mathbf{z} \in \mathbb{Z}^\infty : j_n \in \mathcal{L}_n(\boldsymbol{z})\} \cap D_n\right) \geq 1 - \epsilon$ for all $n \geq N$.*
2. *There exists a constant $L < \infty$ such that $\bar{\delta}_{j_n}(r_n^{-1}) + \bar{B}_{j_n}(P) \leq L \inf_{k\in\mathcal{G}_n}\{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\}$.*

Then

$$||\psi_{\tilde{k}_n}(P_n) - \psi(P)||_\Theta = O_P\left(\inf_{k\in\mathcal{G}_n}\{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\}\right).$$

Proof. Let $A_n \equiv \{\boldsymbol{z} \in \mathbb{Z}^\infty : j_n \in \mathcal{L}_n(\boldsymbol{z})\}$.

For any $\boldsymbol{z} \in A_n \cap D_n$, it follows that

$$\begin{aligned}
||\psi_{\tilde{k}_n(\boldsymbol{z})}(P_n(\boldsymbol{z})) - \psi(P)||_\Theta \leq & ||\psi_{\tilde{k}_n(\boldsymbol{z})}(P_n(\boldsymbol{z})) - \psi_{j_n}(P_n(\boldsymbol{z}))||_\Theta + ||\psi_{j_n}(P_n(\boldsymbol{z})) - \psi(P)||_\Theta \\
\leq & ||\psi_{\tilde{k}_n(\boldsymbol{z})}(P_n(\boldsymbol{z})) - \psi_{j_n}(P_n(\boldsymbol{z}))||_\Theta + \bar{\delta}_{j_n}(r_n^{-1}) + \bar{B}_{j_n}(P) \\
\leq & 4\bar{\delta}_{j_n}(r_n^{-1}) + \bar{\delta}_{j_n}(r_n^{-1}) + \bar{B}_{j_n}(P),
\end{aligned}$$

where the first linear follows from triangle inequality; the second line follows from the fact that $\boldsymbol{z} \in D_n$ and $t \mapsto \bar{\delta}_k(t)$ is non-decreasing; the third line follows from the fact that $\boldsymbol{z} \in A_n$ and thus $j_n \geq \tilde{k}_n(\boldsymbol{z})$. Thus,

$$\begin{aligned}
A_n \cap D_n \subseteq & \left\{\boldsymbol{z} \in \mathbb{Z}^\infty : ||\psi_{\tilde{k}_n(\boldsymbol{z})}(P_n(\boldsymbol{z})) - \psi(P)||_\Theta \leq 5\left(\bar{\delta}_{j_n}(r_n^{-1}) + \bar{B}_{j_n}(P)\right)\right\} \\
\subseteq & \left\{\boldsymbol{z} \in \mathbb{Z}^\infty : ||\psi_{\tilde{k}_n(\boldsymbol{z})}(P_n(\boldsymbol{z})) - \psi(P)||_\Theta \leq 5L \inf_{k\in\mathcal{G}_n}\{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\}\right\}
\end{aligned}$$

where the last linear follows from the second condition. Since by condition 1, $A_n \cap D_n$ occurs with high probability, the result follows. □

We now construct a sequence $(h_n)_n$ that satisfies both conditions of the lemma. To do this, let for each $n \in \mathbb{N}$,

$$\begin{aligned}
\mathcal{G}_n^+ \equiv & \{k \in \mathcal{G}_n : \bar{\delta}_k(r_n^{-1}) \geq \bar{B}_k(P)\} \\
\mathcal{G}_n^- \equiv & \{k \in \mathcal{G}_n : \bar{\delta}_k(r_n^{-1}) \leq \bar{B}_k(P)\}.
\end{aligned}$$

Remark D.1. For any $n \in \mathbb{N}$, $\mathcal{G}_n^+$ or $\mathcal{G}_n^-$ are non-empty. △

For each $n \in \mathbb{N}$, let

$$T_n^+ = \bar{\delta}_{h_n^+}(r_n^{-1}) + \bar{B}_{h_n^+}(P)$$

if $\mathcal{G}_n^+$ is non-empty where

$$h_n^+ = \min\{k : k \in \mathcal{G}_n^+\};$$

and $T_n^+ = +\infty$, if $\mathcal{G}_n^+$ is empty. Similarly,

$$T_n^- = \bar{\delta}_{h_n^-}(r_n^{-1}) + \bar{B}_{h_n^-}(P)$$

if $\mathcal{G}_n^-$ is non-empty where

$$h_n^- = \max\{k\colon k \in \mathcal{G}_n^-\};$$

and $T_n^- = +\infty$, if $\mathcal{G}_n^-$ is empty.

Remark D.2. (1) Observe that when $\mathcal{G}_n^+$ (resp. $\mathcal{G}_n^-$) is non-empty, since it is discrete, h_n^+ (resp. h_n^-) is well-defined.

Intuitively, h_n^+ is the "round up" version within $\mathcal{G}_n$ of $k(n)$; and h_n^- is the "round down" version within $\mathcal{G}_n$ of $k(n)$.

(2) By our previous observation and the fact that either $\mathcal{G}_n^+$ or $\mathcal{G}_n^-$ is non-empty, it follows that either T_n^+ or T_n^- is finite. △

Finally, for each $n \in \mathbb{N}$, let $h_n \in \mathcal{G}_n$ be such that

$$h_n = h_n^+ 1\{T_n^+ \leq T_n^-\} + h_n^- 1\{T_n^+ > T_n^-\}.$$

Lemma D.2. *For each $n \in \mathbb{N}$, h_n exists and*

$$\bar{\delta}_{h_n}(r_n^{-1}) + \bar{B}_{h_n}(P) = \min\left\{T_n^-, T_n^+\right\}.$$

Proof. For each n, by our previous remark, either T_n^+ or T_n^- is finite.

If $T_n^+ = \infty$, then $T_n^- < \infty = T_n^+$ so h_n^- exists and $h_n = h_n^-$.

If $T_n^- = \infty$, then $T_n^+ < \infty = T_n^-$ so h_n^+ exists and $h_n = h_n^+$.

Finally, if both are finite, then both h_n^+ and h_n^- exist.

The fact that

$$\bar{\delta}_{h_n}(r_n^{-1}) + \bar{B}_{h_n}(P) = \min\left\{T_n^-, T_n^+\right\}$$

follows by construction. □

Lemma D.3. *For each $n \in \mathbb{N}$,*

$$\bar{\delta}_{h_n}(r_n^{-1}) + \bar{B}_{h_n}(P) \leq 2 \inf_{k \in \mathcal{G}_n} \{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\}.$$

Proof. Observe that

$$\inf_{k \in \mathcal{G}_n} \{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\} \geq \min\{\inf_{k \in \mathcal{G}_n^+} \{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\}, \inf_{k \in \mathcal{G}_n^-} \{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\}\}$$

where the infimum is defined as $+\infty$ if the corresponding set is empty.

Fix any $n \in \mathbb{N}$, if $\mathcal{G}_n^+ \neq \{\emptyset\}$,

$$\inf_{k \in \mathcal{G}_n^+} \{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\} \geq \inf_{k \in \mathcal{G}_n^+} \{\bar{\delta}_k(r_n^{-1})\} = \bar{\delta}_{h_n^+}(r_n^{-1}) \geq 0.5 \left(\bar{\delta}_{h_n^+}(r_n^{-1}) + \bar{B}_{h_n^+}(P)\right)$$

where the first inequality follows from the fact that $\bar{B}_k(P) \geq 0$; the second one (the equality) follows from the fact that $k \mapsto \bar{\delta}_k(r_n^{-1})$ is non-decreasing and that h_n^+ is minimal over $\mathcal{G}_n^+$; the third inequality follows from the fact that $\bar{\delta}_{h_n^+}(r_n^{-1}) \geq \bar{B}_{h_n^+}(P)$.

Similarly, if $\mathcal{G}_n^- \neq \{\emptyset\}$,

$$\inf_{k \in \mathcal{G}_n^-} \{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\} \geq \inf_{k \in \mathcal{G}_n^-} \{\bar{B}_k(P)\} = \bar{B}_{h_n^-}(P) \geq 0.5 \left(\bar{\delta}_{h_n^-}(r_n^{-1}) + \bar{B}_{h_n^-}(P)\right).$$

Observe that here we use monotonicity of $k \mapsto \bar{B}_k(P)$.

Thus,

$$\inf_{k \in \mathcal{G}_n} \{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\} \geq 0.5 \min\{T_n^-, T_n^+\},$$

and by Lemma D.2 the desired result follows.

□

Lemma D.4. *For any $n \in \mathbb{N}$, $\mathbf{P}(\{\mathbf{z} \in \mathbb{Z}^\infty : h_n \notin \mathcal{L}_n(\boldsymbol{z})\}) \leq \mathbf{P}(D_n^C)$.*

Proof. For any $n \in \mathbb{N}$,

$$\mathbf{P}(\{\mathbf{z} \in \mathbb{Z}^\infty : h_n \notin \mathcal{L}_n(\boldsymbol{z})\}) \leq \mathbf{P}(\{\mathbf{z} \in \mathbb{Z}^\infty : h_n \notin \mathcal{L}_n(\boldsymbol{z})\} \cap D_n) + \mathbf{P}(D_n^C).$$

By definition of $\mathcal{L}_n$ (omitting the dependence on $\mathbf{Z}$),

$$\{h_n \notin \mathcal{L}_n\} \subseteq \left\{\exists k \in \mathcal{G}_n : k > h_n \text{ and } ||\psi_k(P_n) - \psi_{h_n}(P_n)||_\Theta > 4\bar{\delta}_k(r_n^{-1})\right\}.$$

By triangle inequality and the fact that $t \mapsto \bar{\delta}_k(t)$ is non-decreasing,

$$\{h_n \notin \mathcal{L}_n\} \cap D_n \subseteq \left\{\exists k \in \mathcal{G}_n : k > h_n \text{ and } \bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P) + \bar{\delta}_{h_n}(r_n^{-1}) + \bar{B}_{h_n}(P) > 4\bar{\delta}_k(r_n^{-1})\right\}. \tag{20}$$

We now derive a series of useful claims.

Claim 1: If there exists $k \in \mathcal{G}_n$ such that $k > h_n$ and $h_n = h_n^-$, then $k \in \mathcal{G}_n^+$. **Proof:** If $h_n = h_n^-$, then h_n is the largest element of $\mathcal{G}_n^-$ and thus $k \notin \mathcal{G}_n^-$, which means that $k \in \mathcal{G}_n^+$. □

A corollary of this claim is that if there exists $k \in \mathcal{G}_n$ such that $k > h_n$ and $h_n = h_n^-$, then $\mathcal{G}_n^+$ is non-empty. From this claim, we derive the following two claims.

Claim 2: If there exists a $k > h_n$, then $\bar{\delta}_{h_n}(r_n^{-1}) + \bar{B}_{h_n}(P) \leq 2\bar{\delta}_{h_n^+}(r_n^{-1})$. **Proof:** If $h_n = h_n^+$, then $\bar{\delta}_{h_n}(r_n^{-1}) + \bar{B}_{h_n}(P) \leq \bar{\delta}_{h_n^+}(r_n^{-1}) + \bar{B}_{h_n^+}(P) \leq 2\bar{\delta}_{h_n^+}(r_n^{-1})$. If $h_n = h_n^-$, by the previous claim it follows that $\mathcal{G}_n^+$ is non-empty and thus h_n^+ is well-defined, thus $\bar{\delta}_{h_n}(r_n^{-1}) + \bar{B}_{h_n}(P) \leq \bar{\delta}_{h_n^+}(r_n^{-1}) + \bar{B}_{h_n^+}(P) \leq 2\bar{\delta}_{h_n^+}(r_n^{-1})$. □

Claim 3: For any $k > h_n$, $\bar{\delta}_k(r_n^{-1}) \geq \bar{B}_k(P)$. **Proof:** If $h_n = h_n^+$ then the claim follows because $k \mapsto \bar{\delta}_k(r_n^{-1}) - \bar{B}_k(P)$ is non-decreasing. If $h_n = h_n^-$, then $k \in \mathcal{G}_n^+$ by Claim 1 and thus $\bar{\delta}_k(r_n^{-1}) \geq \bar{B}_k(P)$. □

By Claims 2 and 3, it follows that if there exists $k \in \mathcal{G}_n$ such that $k \geq h_n$, then $\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P) + \bar{\delta}_{h_n}(r_n^{-1}) + \bar{B}_{h_n}(P) \leq 2\bar{\delta}_k(r_n^{-1}) + 2\bar{\delta}_{h_n^+}(r_n^{-1}) \leq 4\bar{\delta}_k(r_n^{-1})$ where the last inequality follows monotonicity of $k \mapsto \bar{\delta}_k(r_n^{-1})$ and the fact that $k > h_n^+$ because $k > h_n$ and so by Claim 1 $k \in \mathcal{G}_n^+$ and h_n^+ is minimal in this set. Applying this to expression 20, it follows that

$$\{h_n \notin \mathcal{L}_n\} \cap D_n \subseteq \left\{\exists k \in \mathcal{G}_n \colon k \geq h_n \text{ and } 4\bar{\delta}_k(r_n^{-1}) > 4\bar{\delta}_k(r_n^{-1})\right\}, \tag{21}$$

which is empty. Hence, $\mathbf{P}(\{\mathbf{z} \in \mathbb{Z}^\infty \colon h_n \notin \mathcal{L}_n(\boldsymbol{z})\}) \leq \mathbf{P}(D_n^C)$ as desired. □

Proof of Theorem 4.2. We verify that $(h_n)_{n\in\mathbb{N}}$ satisfies both conditions in Lemma D.1. By Lemma D.3 condition 2 in the Lemma D.1 holds with $L = 2$. To check condition 1 in the Lemma D.1, observe that

$$\mathbf{P}\left(\mathbb{Z}^\infty \setminus \{\{\mathbf{z} \in \mathbb{Z}^\infty \colon h_n \in \mathcal{L}_n(\boldsymbol{z})\} \cap D_n\}\right) \leq \mathbf{P}\left(\{\mathbf{z} \in \mathbb{Z}^\infty \colon h_n \notin \mathcal{L}_n(\boldsymbol{z})\}\right) + \mathbf{P}\left(D_n^C\right).$$

Thus, by Lemma D.4 and the fact $\lim_{n\to\infty} \mathbf{P}(D_n^C) = 0$, $(h_n)_{n\in\mathbb{N}}$ condition 2 is satisfied. □

D.2 Proof of Proposition 4.1

For any $n \in \mathbb{N}$, let

$$k(n) = \min\{k \in \mathbb{R}_+ \colon \bar{\delta}_k(r_n^{-1}) \geq \bar{B}_k(P)\}.$$

Lemma D.5. *For each $n \in \mathbb{N}$, $k(n)$ exists and solves*

$$\bar{\delta}_{k(n)}(r_n^{-1}) = \bar{B}_{k(n)}(P) = \min_{k\in\mathbb{R}_+} \max\{\bar{\delta}_k(r_n^{-1}), \bar{B}_k(P)\}.$$

Proof. For each n consider the set $\{k \in \mathbb{R}_+ \colon \bar{\delta}_k(r_n^{-1}) \geq \bar{B}_k(P)\}$. The set is closed since $k \mapsto \bar{B}_k(P)$ and $k \mapsto \bar{\delta}_k(r_n^{-1})$ are continuous. Since $\bar{\delta}_k(r_n^{-1}) > 0$ and $\bar{B}_k(P) = o(1)$, if follows that there exists a $K(n) < \infty$ such that $\bar{\delta}_k(r_n^{-1}) \geq \bar{B}_k(P)$ for any $k \geq K(n)$. Thus the set is non-empty and since we are minimizing the identity function, the minimizer exists and uniquely determined by

$$\bar{\delta}_{k(n)}(r_n^{-1}) = \bar{B}_{k(n)}(P).$$

The second equality is obvious. □

The next lemma shows that balancing the sampling and approximation error yields the same rate as the "optimal" choice.

Lemma D.6. *For any $n \in \mathbb{N}$,*

$$\bar{\delta}_{k(n)}(r_n^{-1}) \leq \inf_{k\in\mathbb{R}_+} \{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\} \leq 2\bar{\delta}_{k(n)}(r_n^{-1}).$$

Proof. Observe that for any $n \in \mathbb{N}$ and any $\epsilon > 0$, there exists $k^*(n)$ such that

$$\bar{\delta}_{k^*(n)}(r_n^{-1}) + \bar{B}_{k^*(n)}(P) - \epsilon \leq \inf_{k \in \mathbb{R}_+} \{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\} \leq 2\bar{\delta}_{k(n)}(r_n^{-1}).$$

The upper bound follows from the fact that $\inf_{k\in\mathbb{R}_+}\{\bar{\delta}_k(r_n^{-1})+\bar{B}_k(P)\} \leq \bar{\delta}_{k(n)}(r_n^{-1})+\bar{B}_{k(n)}(P)$ and definition of $k(n)$.

If $k^*(n) \geq k(n)$, then $\bar{\delta}_{k^*(n)}(r_n^{-1}) \geq \bar{\delta}_{k(n)}(r_n^{-1})$ since $k \mapsto \bar{\delta}_k(t)$ is non-decreasing for any $t \geq 0$ On the other hand, if $k^*(n) < k(n)$, then $\bar{B}_{k^*(n)}(P) \geq \bar{B}_{k(n)}(P) = \bar{\delta}_{k(n)}(r_n^{-1})$ where the last equality follows from Lemma D.5. Therefore, for any $n \in \mathbb{N}$ and any $\epsilon > 0$,

$$\bar{\delta}_{k(n)}(r_n^{-1}) - \epsilon \leq \inf_{k \in \mathbb{R}}\{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\} \leq 2\bar{\delta}_{k(n)}(r_n^{-1}).$$

Since $\epsilon > 0$ is arbitrary the result follows. □

Proof of Proposition 4.1. By inspection of the proof of Lemma D.1 it suffices to show existence of a sequence $(j_n)_n$ for which Condition 1 and the following strengthening of condition 2 holds:

Condition 2': There exists a constant $L < \infty$ such that $\bar{\delta}_{j_n}(r_n^{-1})+\bar{B}_{j_n}(P) \leq L \inf_{k\in\mathbb{R}_+}\{\bar{\delta}_k(r_n^{-1})+\bar{B}_k(P)\}$ for any $n \in \mathbb{N}$.

As for the proof of Theorem 4.2, we propose $j_n = h_n$ for all $n \in \mathbb{N}$. By Lemma D.4 condition 1 holds, so it only remains to show that Condition 2' holds.

Under the conditions in the proposition, h_n^+ and h_n^- are well-defined for all $n \in \mathbb{N}$. Moreover, they are either the same or consecutive elements in $\mathcal{G}_n$. Thus, under the conditions in the proposition, there exists a $C < \infty$ and a $N \in \mathbb{N}$ such that $\bar{\delta}_{h_n^+}(r_n^{-1}) \leq C\bar{\delta}_{h_n^-}(r_n^{-1})$ for all $n \geq N$. Therefore, for all $n \geq N$,

$$\begin{aligned}
\bar{\delta}_{h_n}(r_n^{-1}) + \bar{B}_{h_n}(P) =& \min\{\bar{\delta}_{h_n^-}(r_n^{-1}) + \bar{B}_{h_n^-}(P), \bar{\delta}_{h_n^+}(r_n^{-1}) + \bar{B}_{h_n^+}(P)\} \\
\leq& \min\{\bar{\delta}_{h_n^-}(r_n^{-1}) + \bar{B}_{h_n^-}(P), 2\bar{\delta}_{h_n^+}(r_n^{-1})\} \\
\leq& \min\{\bar{\delta}_{h_n^-}(r_n^{-1}) + \bar{B}_{h_n^-}(P), 2C\bar{\delta}_{h_n^-}(r_n^{-1})\} \\
\leq& 2C\bar{\delta}_{h_n^-}(r_n^{-1}).
\end{aligned}$$

Observe that $h_n^- \leq k(n)$ because, by definition $\bar{\delta}_{h_n^-}(r_n^{-1}) \leq \bar{B}_{h_n^-}(P)$ but $k(n)$ satisfies $\bar{\delta}_{k(n)}(r_n^{-1}) \geq \bar{B}_{k(n)}(P)$, so under the fact that $k \mapsto \bar{B}_k(P)$ is non-increasing it must follow that $h_n^- \leq k(n)$. Thus, $\bar{\delta}_{h_n^-}(r_n^{-1}) \leq \bar{\delta}_{k(n)}(r_n^{-1})$ and the result follows from Lemma D.6. □

Remark D.3 (Sufficient conditions for $\mathcal{G}_n^+$ and $\mathcal{G}_n^-$ to be non-empty in Proposition 4.1)**.** The condition that $\bar{\delta}_j(r_n^{-1}) < \bar{B}_j(P)$ for some $j \in \mathcal{G}_n$ is easy to satisfy as any fix j (e.g. $j = 1$) will satisfy this condition eventually. The other inequality is more delicate but the next lemma provides the basis for its verification.

Lemma D.7. $\limsup_{n\to\infty} \bar{\delta}_{k(n)}(r_n^{-1}) = 0.$

Proof. Suppose not. Then there exists a sub-sequence $(n(j))_j$ and a $c > 0$ such that $\bar{\delta}_{k(n(j))}(r_{n(j)}^{-1}) \geq c$ for all j. Clearly $(k(n(j)))_j$ must diverge, so $\bar{B}_{k(n(j))}(P) = o(1)$, but then $k(n(j))$ cannot be balancing both terms. □

Let $(j(n))_n$ be such that $\liminf_{n\to\infty} \bar{\delta}_{j(n)}(r_n^{-1}) > 0$. Since $\limsup_{n\to\infty} \bar{\delta}_{k(n)}(r_n^{-1}) = 0$, it follows that $j(n) > k(n)$ eventually and thus $\bar{\delta}_{j(n)}(r_n^{-1}) > \bar{B}_{j(n)}(P)$.

Thus, any set $\mathcal{G}_n$ such that $\mathcal{G}_n \ni \{1, j(n)\}$ will satisfy that $\mathcal{G}_n^+$ and $\mathcal{G}_n^-$ are non-empty, at least for sufficiently large n. △

D.3 An Extension of Theorem 4.2

In this section we show that Theorem 4.2 can be extended to whole when $\mathcal{G}_n$ is any *closed* set of $\mathbb{K}$. The extension is merely technical as one needs to ensure that some minimizers are attained over $\mathcal{G}_n$ when this set is not finite.

First, one needs to ensure that $\tilde{k}_n$ exists with probability approaching 1. Lemma D.4 shows that with probability approaching one, the set $\mathcal{L}_n$ is non-empty. Thus, it suffices to argue that $\mathcal{L}_n$ is closed. It is easy to see that the following assumption is sufficient for this.

Assumption D.1. *For each $P \in \mathcal{D}$ and each $t \geq 0$, the mapping $k \mapsto \psi_k(P)$ is continuous over $\mathcal{G}_n$.*

Observe that when $\mathcal{G}_n$ is finite this condition is trivially satisfied and that is why it is not imposed in the text.

The following theorem is an extension of Theorem 4.2 to the case where $\mathcal{G}_n$ is a closed set (not necessarily finite) of $\mathbb{K}$.

Theorem D.1. Suppose all assumptions in Theorem 4.2 hold. And suppose further that $\mathcal{G}_n$ is a closed set (not necessarily finite) of $\mathbb{K}$ and that Assumptions D.1 holds. Then

$$||\psi_{\tilde{k}_n}(P_n) - \psi(P)||_\Theta = O_P\left(\inf_{k\in\mathcal{G}_n}\{\bar{\delta}_k(r_n^{-1}) + \bar{B}_k(P)\}\right).$$

Proof. The proof is identical to the one of Theorem 4.2. Assumption D.1 and the fact that $\mathbb{K}$ is closed ensure that the quantities defined in the proof exist. □

D.4 Appendix for Section 4.3

For any probability P over a set A and any $k \in \mathbb{N}$, let $\bigotimes_{i=1}^k P$ be the product probability measure over $\prod_{i=1}^k A$ induced by P. Also, recal that the Wasserstein distance for $p \geq 1$ over $\mathcal{P}(\prod_{i=1}^k \mathbb{Z})$ for some $k \geq 1$ is defined as

$$\mathcal{W}_p(P,Q) \equiv \left(\inf_{\mu\in H(P,Q)} \int ||x-y||^p \mu(dx,dy)\right)^{1/p}$$

for any P, Q in $\mathcal{P}(\prod_{i=1}^{k} \mathbb{Z})$, where $H(P,Q)$ is the class of Borel probability measures over $\mathbb{Z}^{2k}$ with marginals P and Q. It is well-known that

$$(P,Q) \mapsto ||P-Q||_{Lip(\prod_{i=1}^{k} \mathbb{Z})} = \mathcal{W}_1(P,Q)$$

where for any set A, let $LB(A)$ to denote the class of bounded Lipschitz (with constant 1) real-valued functions on A; see Villani [2008] p. 60.

The following lemma is used in the proof

Lemma D.8. *For any P, Q in $\mathcal{P}(\mathbb{Z})$, any $k \in \mathbb{N}$ and any $\mu \in H(P,Q)$, $\bigotimes_{i=1}^{k} \mu \in H\left(\bigotimes_{i=1}^{k} P, \bigotimes_{i=1}^{k} Q\right)$.*

Proof of Lemma D.8. It is clear that the marginal of $\bigotimes_{i=1}^{k} \mu$ of a pair (x_i, y_i) is μ. Therefore, for any $A_1, ..., A_k$ Borel subsets on $\mathbb{Z}$,

$$\bigotimes_{i=1}^{k} \mu\left((A_1 \times \mathbb{Z}), ..., (A_k \times \mathbb{Z})\right) = \prod_{i=1}^{k} \mu(A_i \times \mathbb{Z}) =^{*} \prod_{i=1}^{k} P(A_i)$$

where $*$ follows because $\mu \in H(P,Q)$. Equivalently,

$$\int g(\vec{x}) \bigotimes_{i=1}^{k} \mu(d\vec{x}, d\vec{y}) = \int g(\vec{x}) \bigotimes_{i=1}^{k} P(d\vec{x})$$

for any g belonging to the class of "simple" functions on $\prod_{i=1}^{k} \mathbb{Z}$: The class of functions of the form $g(\vec{x}) = \sum_{i=1}^{k} 1_{A_i}(x_i)$ for any $A_1, ..., A_k$ Borel subsets on $\mathbb{Z}$. Since the class of "simple" functions is dense in $\mathbb{C}(\mathbb{Z}^k, \mathbb{R})$ (the class of continuous and bounded functions over $\mathbb{Z}$), by taking limits and using the previous display one can show that

$$\int f(\vec{x}) \bigotimes_{i=1}^{k} \mu(d\vec{x}, d\vec{y}) = \int f(\vec{x}) \bigotimes_{i=1}^{k} P(d\vec{x})$$

for any $f \in \mathbb{C}(\prod_{i=1}^{k} \mathbb{Z}, \mathbb{R})$. That is, the marginal probability of $\bigotimes_{i=1}^{k} \mu$ for the first k coordinates is $\bigotimes_{i=1}^{k} P$. A completely analogous argument shows that the marginal probability of $\bigotimes_{i=1}^{k} \mu$ for the last k coordinates is $\bigotimes_{i=1}^{k} Q$. $\square$

Proof of Proposition 4.2. First note that, for any $f \in LB$, $E_{\psi_n(P)}[f(Z)] = \int \psi_n(P)(f(Z) \geq t)dt = \int \mathbf{P}(f(T_n(\boldsymbol{z}, P)) \geq t)dt = E_{\mathbf{P}}\left[f\left(T_n(\boldsymbol{z}, P)\right)\right]$. Hence, it follows that

$$\begin{aligned}
||\psi_k(P) - \psi_k(Q)||_{\Theta} \leq & \sup_{f \in Lip} \left|E_{\mathbf{P}}\left[f\left(T_k(\boldsymbol{z}, P)\right)\right] - E_{Q^{\infty}}\left[f\left(T_k(\boldsymbol{z}, P)\right)\right]\right| \\
& + \sup_{f \in Lip} \left|E_{Q^{\infty}}\left[f\left(T_k(\boldsymbol{z}, P)\right) - f\left(T_k(\boldsymbol{z}, Q)\right)\right]\right| \\
\equiv & T_{1,k}(P,Q) + T_{2,k}(P,Q).
\end{aligned}$$

We now show that both terms, $T_{1,k}(P,Q)$ and $T_{2,k}(P,Q)$, are bounded by $\sqrt{k}\mathcal{W}_1(P,Q)$.

For any $k \in \mathbb{N}$, $f(T_k(\boldsymbol{z},P)) \equiv f_k\left(\sqrt{k}\max\{k^{-1}\sum_{i=1}^k Z_i(\boldsymbol{z}),0\}\right)$ where $f_k \equiv f(\cdot - \sqrt{k}\max\{E_P[Z],0\})$. It is easy to see that for any k, $f_k \in LB$ (given that $f \in LB$). Moreover, the mapping $t \mapsto g_k(t) \equiv f_k(\max\{t,0\})$ is also in Lip because

$$|g_k(t) - g_k(t')| \leq |\max\{t,0\} - \max\{t',0\}| \leq |t' - t|, \ \forall t, t'.$$

Finally, the mapping $t \mapsto g_k(\sqrt{k}t)/\sqrt{k}$ is also in LB since $g_k \in Lip$. Hence,

$$T_{1,k}(P,Q) \leq \sqrt{k} \sup_{f \in Lip} \left| E_{\bigotimes_{i=1}^k P}\left[f\left(k^{-1}\sum_{i=1}^k Z_i\right)\right] - E_{\bigotimes_{i=1}^k Q}\left[f\left(k^{-1}\sum_{i=1}^k Z_i\right)\right]\right|. \tag{22}$$

For any $g : \mathbb{R} \to \mathbb{R}$, let $\bar{g} : \mathbb{R}^k \to \mathbb{R}^k$ be defined as

$$\vec{z} \equiv (z_1, ..., z_k) \mapsto \bar{g}(\vec{z}) \equiv g\left(k^{-1}\sum_{i=1}^k z_i\right).$$

We now show that for any $g \in LB(\mathbb{R})$, $k\bar{g} \in LB(\mathbb{R}^k)$. This follows because

$$|\bar{g}(\vec{z}) - \bar{g}(\vec{z}')| \leq |k^{-1}\sum_{i=1}^k \{z_i - z_i'\}| \leq k^{-1}||\vec{z} - \vec{z}'||_1.$$

This result allow us to bound from above the LHS of the expression 22 so that

$$\begin{aligned}
&\sqrt{k} \sup_{f \in LB} \left| E_{\bigotimes_{i=1}^k P}\left[f\left(k^{-1}\sum_{i=1}^k Z_i\right)\right] - E_{\bigotimes_{i=1}^k Q}\left[f\left(k^{-1}\sum_{i=1}^k Z_i\right)\right]\right| \\
\leq & k^{-1/2} \sup_{f \in LB(\mathbb{R}^k)} \left| E_{\bigotimes_{i=1}^k P}\left[f(Z)\right] - E_{\bigotimes_{i=1}^k Q}\left[f(Z')\right]\right| \\
= & k^{-1/2}\mathcal{W}_1\left(\bigotimes_{i=1}^k P, \bigotimes_{i=1}^k Q\right).
\end{aligned}$$

For any $\mu \in H(P,Q)$, $\bigotimes_{i=1}^k \mu \in \mathcal{P}(\mathbb{Z}^k \times \mathbb{Z}^k)$ where $\mathbb{Z}^k \equiv \prod_{i=1}^k \mathbb{Z}$. Moreover, by Lemma D.8, $\bigotimes_{i=1}^k \mu \in H\left(\bigotimes_{i=1}^k P, \bigotimes_{i=1}^k Q\right)$.

For any $\eta > 0$, let $\mu^* \in H(P,Q)$ be the approximate minimizer of $\mathcal{W}_1(P,Q)$, i.e.,

$$\int |x - y|\mu^*(dx,dy) \leq \mathcal{W}_1(P,Q) + \eta,$$

as $\bigotimes_{i=1}^k \mu^* \in H\left(\bigotimes_{i=1}^k P, \bigotimes_{i=1}^k Q\right)$, it follows that

$$\begin{aligned}\mathcal{W}_1\left(\bigotimes_{i=1}^k P, \bigotimes_{i=1}^k Q\right) \leq & \int_{\mathbb{Z}^{2k}} ||\vec{x}-\vec{y}||_1 \bigotimes_{i=1}^k \mu^*(dx_i, dy_i) \\ =& \sum_{i=1}^k \int_{\mathbb{Z}^2} |x_i - y_i| \bigotimes_{i=1}^k \mu^*(dx_i, dy_i) \\ =& \sum_{i=1}^k \int_{\mathbb{Z}^2} |x_i - y_i| \mu^*(dx_i, dy_i) \\ =& k\mathcal{W}_1(P,Q) + k\eta.\end{aligned}$$

Since $\eta > 0$ is arbitrary, it follows that $\mathcal{W}_1(P^k, Q^k) \leq k\mathcal{W}_1(P,Q)$. Thus implying

$$T_{1,k}(P,Q) \leq \sqrt{k}\mathcal{W}_1(P,Q).$$

Regarding the term $T_{2,k}(P,Q)$, observe that

$$\begin{aligned}T_{2,k}(P,Q) \leq & |T_k(\boldsymbol{z}, P) - T_k(\boldsymbol{z}, Q)| \\ =& \sqrt{k}|\max\{E_P[Z], 0\} - \max\{E_Q[Z'], 0\}| \\ \leq& \sqrt{k}|E_P[Z] - E_Q[Z']|.\end{aligned}$$

Since $E_P[Z] = \int_{\mathbb{Z}^2} z\mu(dz, dz')$ for any $\mu \in H(P,Q)$.

$$T_{2,k}(P,Q) \leq \sqrt{k}\,|E_\mu[Z] - E_\mu[Z']| \leq \sqrt{k} \int |z - z'|\mu(dz, dz').$$

Choosing μ as the (approximate) minimizer of $\mathcal{W}_1(P,Q)$ it follows that

$$T_{2,k}(P,Q) \leq \sqrt{k}\mathcal{W}_1(P,Q).$$

□

To show the proposition 4.3, let $\psi(P) \in \mathcal{P}(\mathbb{R})$ be defined as the probability of $\max\{\zeta, 0\}$ if P is such that $E_P[Z] = 0$ and the probability of ζ otherwise, where $\zeta \sim N(0,1)$. The following lemma shows that $\psi(P)$ is the limit of $(\psi_k(P))_{k\in\mathbb{N}}$.

Lemma D.9. *For any $k \in \mathbb{N}$ and any $P \in \mathcal{M}$,*

$$||\psi_k(P) - \psi(P)||_{LB} \leq 6k^{-1/2}E_P[|Z|^3] + 1\{E_P[Z] > 0\}2\Phi(-\sqrt{k}E_P[Z])$$

Proof. Since $P \in \mathcal{M}$, $T_k(\boldsymbol{z}, P) = \max\{k^{-1/2}\sum_{i=1}^k (Z_i(\boldsymbol{z}) - E_P[Z]), -\sqrt{k}E_P[Z]\}$ for any $k \in \mathbb{N}$.

By triangle inequality and definition of $||.||_{LB}$,

$$\begin{aligned}||\psi_k(P)-\psi(P)||_{LB}\leq&\sup_{f\in LB}E\left[\left|f\left(T_k(\boldsymbol{z},P)\right)-f\left(\max\{\zeta,-\sqrt{k}E_P[Z]\}\right)\right|\right]\\&+\sup_{f\in LB}E\left[\left|f\left(\max\{\zeta,-\sqrt{k}E_P[Z]\}\right)-f\left(s_P(\zeta)\right)\right|\right]\\\equiv&Term_1(k)+Term_2(k),\end{aligned}$$

where $\zeta\sim N(0,1)$ and $t\mapsto s_P(t)=\max\{t,0\}\times 1\{E_P[Z]=0\}+t\times 1\{E_P[Z]>0\}$.

We now provide a bound for terms $Term_1(k)$. For any $f\in LB$ and any $k\in\mathbb{N}$, the mapping $t\mapsto f_k(t)\equiv f(\max\{t,-\sqrt{k}E_P[Z]\})$ satisfies, for any $t\leq t'$,

$$|f_k(t')-f_k(t)|\leq|\max\{t',-\sqrt{k}E_P[Z]\}-\max\{t,-\sqrt{k}E_P[Z]\}|$$

where the RHS is equal to 0 if $t\leq t'\leq -kE_P[Z]$, $t'-(-kE_P[Z])\leq t'-t$; if $t\leq -kE_P[Z]\leq t'$; and $t'-t$ if $-kE_P[Z]\leq t\leq t'$. Hence $|f_k(t')-f_k(t)|\leq|t'-t|$. The same inequality holds when $t'\leq t$, so f_k is in LB. Therefore,

$$Term_1(k)\leq\sup_{f\in LB}\left|E_{\mathbf{P}}\left[f\left(k^{-1/2}\sum_{i=1}^{k}(Z_i-E_P[Z])\right)-f\left(\zeta\right)\right]\right|\leq 6k^{-1/2}E_P[|Z|^3]$$

where the last line follows from Berry-Esseen Inequality for Lipschitz functions (see Barbour and Chen [2005] Thm. 3.2 in Ch. 1).

Regarding $Term_2(k)$, we note that if $E_P[Z]=0$, then $Term_2(k)=0$, because $s_P(\zeta)=\max\{\zeta,0\}$. So we only need a bound for $E_P[Z]>0$. Under this condition,

$$Term_2(k)\leq\sup_{f\in LB}E\left[1\{\zeta\leq-\sqrt{k}E_P[Z]\}\left|f\left(\max\{\zeta,-\sqrt{k}E_P[Z]\}\right)-f\left(\zeta\right)\right|\right].$$

Since $||f||_{L^\infty}\leq 1$, the inequality further implies that $Term_2(k)\leq 2E\left[1\{\zeta\leq-\sqrt{k}E_P[Z]\}\right]=2\Phi(-\sqrt{k}E_P[Z])$. □

Proof of Proposition 4.3. By the triangle inequality,

$$||\psi_{\tilde{k}_n}(P_n)-\psi_n(P)||_{LB}\leq||\psi_{\tilde{k}_n}(P_n)-\psi(P)||_{LB}+||\psi(P)-\psi_n(P)||_{LB}$$

By Lemma D.9, $||\psi_k(P)-\psi(P)||_{LB}\leq 6k^{-1/2}E_P[|Z|^3]+1\{E_P[Z]>0\}2\Phi(-\sqrt{k}E_P[Z])$. Thus, we can invoke Theorem 4.2 and its corollary to show that $||\psi_{\tilde{k}_n}(P_n)-\psi(P)||_{LB}=O_P\left(\inf_{k\in\{1,...,n\}}\{l_n\sqrt{k}n^{-1/2}+1\{E_P[Z]>0\}2\Phi(-\sqrt{k}E_P[Z])+k^{-1/2}E_P[|Z|^3]\}\right)$. It is clear that the choice k that achieves the infimum will diverge with n, so for this choice of k, $1\{E_P[Z]>0\}2\Phi(-\sqrt{k}E_P[Z])$ will eventually be dominated by $k^{-1/2}E_P[|Z|^3]$. Hence $||\psi_{\tilde{k}_n}(P_n)-\psi(P)||_{LB}=O_P\left(\inf_{k\in\{1,...,n\}}\{l_n\sqrt{k}n^{-1/2}+k^{-1/2}E_P[|Z|^3]\}\right)$

The desired result follows from the fact that for sufficiently large n, $||\psi(P)-\psi_n(P)||_{LB} \leq 8n^{-1/2}E_P[|Z|^3]$ and the fact that there exists a $C>0$ such that $\inf_{k\in\{1,\dots,n\}}\{l_n\sqrt{k}n^{-1/2}+k^{-1/2}E_P[|Z|^3]\}\geq Cn^{-1/2}$ for all n. $\square$

Proof of Proposition 4.4. Throughout, fix k and P,P' and let $||.||_\Theta \equiv ||.||_{L^q}$. Let $\Theta_k(M)\equiv\{\theta\in\Theta_k\colon ||\theta-\psi_k(P)||_\Theta\geq M\}$. And, let

$$t\mapsto U_k(t)\equiv \inf_{\theta\in\Theta_k(t)}\frac{Q_k(P,\theta)-Q_k(P,\psi_k(P))}{t}.$$

Towards the end of the proof we show that U_k is continuous. Let $t\mapsto\Gamma_k(t)\equiv\inf_{s\geq t}U_k(s)$; it follows that $\Gamma_k\leq U_k$, Γ_k is non-decreasing and by the Theorem of the Maximum Γ_k is continuous.

We show that $||\psi_k(P)-\psi_k(P')||_\Theta\geq M\equiv\Gamma_k^{-1}(d(P,P'))$ cannot occur.[26] To do this, we show that $1\{||\psi_k(P)-\psi_k(P')||_\Theta\geq M\}=0$. Observe that

$$\begin{aligned}1\{||\psi_k(P)-\psi_k(P')||_\Theta\geq M\} =&1\{\cup_{j\in\mathbb{N}}\{2^jM\geq||\psi_k(P)-\psi_k(P')||_\Theta\geq 2^{j-1}M\}\}\\ \leq&\max_{j\in\mathbb{N}}1\{2^jM\geq||\psi_k(P)-\psi_k(P')||_\Theta\geq 2^{j-1}M\}.\end{aligned}$$

For each $(j,k)\in\mathbb{N}^2$, let $S_{j,k}\equiv\{\theta\in\Theta_k\colon 2^jM\geq||\psi_k(P)-\theta||_\Theta\geq 2^{j-1}M\}$. It follows that, for any $j\in\mathbb{N}$,

$$1\{\psi_k(P')\in S_{j,k}\}\leq 1\left\{\inf_{\theta\in S_{j,k}}Q(P',\theta)\leq Q(P',\psi_k(P))\right\}$$

because $\psi_k(P)\in\Theta_k\setminus S_{j,k}$. Observe that, for any $\theta\in S_{j,k}\cup\{\psi_k(P)\}\subseteq\{\theta\in\Theta_k\colon||\theta-\psi_k(P)||_\Theta\leq 2^jM\}$

$$\begin{aligned}Q(P',\theta)-Q(P',\psi_k(P))\geq&Q(P,\theta)-Q(P,\psi_k(P))\\ &-|Q(P',\theta)-Q(P',\psi_k(P))-\{Q(P,\theta)-Q(P,\psi_k(P))\}|\\ \geq&Q(P,\theta)-Q(P,\psi_k(P))-2^jM\Delta_{j,k}(P,P')\end{aligned}$$

where

$$\Delta_{j,k}(P,P')\equiv\sup_{\theta\in\Theta_k\colon||\theta-\psi_k(P)||_\Theta\leq 2^jM}\frac{|Q(P',\theta)-Q(P',\psi_k(P))-\{Q(P,\theta)-Q(P,\psi_k(P))\}|}{||\theta-\psi_k(P)||_\Theta}.$$

[26] Note that $\bar{U}_k^{-1}(t)\equiv\inf\{s\colon\bar{U}_k(s)\geq t\}$.

Hence,

$$
\begin{aligned}
1\{\psi_k(P') \in S_{j,k}\} \leq & 1\left\{ \inf_{\theta\in\Theta_k\colon ||\theta-\psi_k(P)||_{\Theta}\geq 2^{j-1}M} \frac{Q(P,\theta)-Q(P,\psi_k(P))}{2^{j-1}M} \leq 0.5\mathit{\Delta}_{j,k}(P,P') \right\} \\
\leq & 1\left\{ \inf_{\theta\in\Theta_k\colon ||\theta-\psi_k(P)||_{\Theta}\geq 2^{j-1}M} \frac{Q(P,\theta)-Q(P,\psi_k(P))}{2^{j-1}M} \leq 0.5\max_{k\in\mathbb{N}} \mathit{\Delta}_{\infty,k}(P,P') \right\} \\
\leq & 1\left\{ \Gamma_k(2^{j-1}M) \leq 0.5\max_{k\in\mathbb{N}} \mathit{\Delta}_{\infty,k}(P,P') \right\}.
\end{aligned}
$$

Since $\bar{U}_k$ is non-decreasing, the previous display implies that

$$
1\{\psi_k(P') \in S_{j,k}\} \leq 1\left\{ 2^{j-1}M \leq \Gamma_k^{-1}(0.5\max_{k\in\mathbb{N}} \mathit{\Delta}_{\infty,k}(P,P')) \right\}
$$

which equals zero by the definition of M, the fact that Γ_k^{-1} is non-decreasing and $2^{j-1} \geq 1$.

We now show that $t \mapsto U_k(t)$ (and thus Γ_k) is continuous. Consider the problem $\inf_{\theta\in\Theta_k(M)} Q_k(P,\theta)$, and consider the set $L_k(M) \equiv \{\theta \in \Theta_k(M)\colon Pen(\theta) \leq \lambda_k^{-1}Q(P,\theta_k)\}$ for some (any) $\theta_k \in \Theta_k(M)$ which is non-empty and close. To solve the former minimization problem it suffices to solve $\inf_{\theta\in L_k(M)} Q_k(P,\theta)$, because the minimum value cannot be outside $L_k(M)$. Because Pen is lower-semi-compact, $L_k(M)$ is compact (a closed subset of a compact set) so this and lower-semi-continuity of $Q_k(P,\cdot)$ ensures that $\inf_{\theta\in L_k(M)} Q_k(P,\theta)$ is achieved by an element in $L_k(M)$ and the same is true for the original $V_k(M) \equiv \inf_{\theta\in\Theta_k(M)} Q_k(P,\theta)$. We just showed that the correspondence $M \mapsto L_k(M)$ is compact-valued, it is also continuous. By virtue of the Theorem of the Maximum, V_k is continuous; it is also non-decreasing. The function $t \mapsto U_k(t) = V_k(t)/t$ is also continuous in $t > 0$. □

E Appendix for Section 5

Proof of Theorem 5.1. We first show the desired result for a fixed k, i.e., $\mathbf{k}(n) = k$ for any $n \in \mathbb{N}$.

Let $(z,k) \mapsto \varphi_k(P) \equiv D\psi_k(P)[\delta_z - P]$ which is well-defined because $\delta_z - P \in \mathcal{T}_P$. We now show that $\varphi_k(P) \in L_0^2(P)$. The fact that has mean zero (provided it exists) is trivial, so we only show that $\int |\varphi_k(P)(z)|^2 P(dz) < \infty$. The topology is locally convex and thus generated by a family of semi-norms. Suppose there exists a $L < \infty$ such that $|D\psi_k(P)[\delta_z - P]| \leq \rho(\delta_z - P)$ where ρ is a member of the family. Because the topology τ is assumed to be dominated by $||.||_{TV}$ it follows that $\rho(\delta_z - P) \leq C||\delta_z - P||_{TV} \leq 2C$ for some finite C for any $z \in \mathbb{Z}$. And thus $\int |\varphi_k(P)(z)|^2 P(dz) \leq 2CL < \infty$ as desired.

We now show that there exists a member of the family of semi-norms, ρ, and a $L < \infty$ such that $|D\psi_k(P)[Q]| \leq L\rho(Q)$ for all $Q \in ca(\mathbb{Z})$. Suppose not, that is, for any $R > 0$ and any ρ, there exists a Q such that $\rho(Q) = 1$ and $|D\psi_k(P)[Q]| > R$. Since $D\psi_k(P)$ is continuous with respect to τ, there exists a member, ρ, of the family of semi-norms such that for any $\epsilon > 0$ there exists $\delta > 0$ such that if Q is such that $\rho(Q) \leq \delta$, then $|D\psi_k(P)[Q]| < \epsilon$.

Let $R = \epsilon/\delta$. There exists a Q such that $\rho(Q) = 1$ and $\delta|D\psi_k(P)[Q]| > \epsilon$. Let $\nu = \delta Q$, then $\rho(\nu) = \delta$ but $\delta|D\psi_k(P)[Q]| = |D\psi_k(P)[\delta Q]| = |D\psi_k(P)[\nu]| > \epsilon$ but this is a contradiction.

We now show that $\eta_k(P_n - P) = o_P(n^{-1/2})$ for each $k \in \mathbb{K}$. Let $n \mapsto \mathbb{G}_n \equiv \sqrt{n}(P_n - P)$. It follows that, a.s.-$\mathbf{P}$, $t \mapsto P + t\mathbb{G}_n$ is a valid curve because $\mathbb{G}_n \in \mathcal{T}_P$ a.s.-$\mathbf{P}$.

Fix any $\epsilon > 0$ and let $U \in \mathcal{C}$ be as in the condition of the statement of the theorem. Then, letting $D_n \equiv \{\boldsymbol{z} \in \mathbb{Z}^\infty \colon \mathbb{G}_n(\boldsymbol{z}) \in U\}$, it follows that

$$\mathbf{P}\left(\sqrt{n}|\eta_k(P_n - P)| \geq \epsilon\right) \leq \mathbf{P}\left(\sqrt{n}|\eta_k(P_n - P)| \geq \epsilon \mid D_n\right) + \mathbf{P}(D_n^C).$$

The second term in the RHS is less than ϵ for all $n \geq N$. Regarding the first term in the RHS it follows that, over D_n,

$$|\eta_k\left(P_n - P\right)|/t_n \leq \sup_{Q \in U} |\eta_k\left(t_n Q\right)|/t_n$$

where $t_n \equiv 1/\sqrt{n}$. Thus, by definition of differentiability, the first term in the RHS also vanishes as $n \to \infty$. So the desired result follows.

Therefore, for any $k \in \mathbb{K}$

$$\frac{1}{||\varphi_k(P)||_{L^2(P)}} \left| \sqrt{n}(\psi_k(P_n) - \psi_k(P)) - n^{-1}\sum_{i=1}^{n} D\psi_k(P)[\delta_{Z_i} - P] \right| = \frac{\sqrt{n}|\eta_k(P_n - P)|}{||\varphi_k(P)||_{L^2(P)}}$$

and $\frac{\sqrt{n}|\eta_k(P_n - P)|}{||\varphi_k(P)||_{L^2(P)}} = o_P(1)$, as desired.

We now shows existence of a diverging sequence by using the first part and the diagonalization lemma C.1.

For any $\epsilon > 0$, $k \in \mathbb{N}$ and $n \in \mathbb{N}$, let $T(\epsilon, k, n) \equiv \mathbf{P}\left(\sqrt{n}\frac{|\eta_k(P_n(\boldsymbol{z}) - P)|}{||\varphi_k(P)||_{L^2(P)}} \geq \epsilon\right)$. To show the desired result it suffices to show that there exists a non-decreasing diverging sequence $(j(n))_n$ such that for all $\epsilon > 0$, there exists a $\bar{N}$ such that

$$T(\epsilon, j(n), n) \leq \epsilon,$$

for all $n \geq \bar{N}$.

We shows that, for any $k \in \mathbb{K}$, $\lim_{n\to\infty} T(2^{-k}, k, n) = 0$. By Lemma C.1, there exists a non-decreasing diverging sequence $(j(n))_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} T(2^{-j(n)}, j(n), n) = 0$; i.e., for any $\epsilon > 0$, there exists a $N(\epsilon)$ such that $T(2^{-j(n)}, j(n), n) \leq \epsilon$ for all $n \geq N(\epsilon)$.

Since $j(.)$ diverges, there exists a N_ϵ such that $1/2^{j(n)} \leq \epsilon$ for all $n \geq N_\epsilon$. By these observations and the fact that $\epsilon \mapsto T(\epsilon, k, n)$ is non-increasing,

$$T(\epsilon, j(n), n) \leq T(2^{-j(n)}, j(n), n) \leq \epsilon$$

for all $n \geq \bar{N}_\epsilon \equiv \max\{N_\epsilon, N(\epsilon)\}$, and we thus showed the desired result.

□

The following result is a well-known representation result (see van der Vaart and Wellner [1996]) and is stated merely for convenience.

Lemma E.1. *Let* $\boldsymbol{z} \mapsto \mathbb{G}_n(\boldsymbol{z}) \equiv \sqrt{n}(P_n(\boldsymbol{z}) - P)$. *Suppose* $\mathcal{S}$ *is P-Donsker. Then there exists a tight Borel measurable* $\mathbb{G} \in L^\infty(\mathcal{S})$ *such that for any* $\epsilon > 0$, *there exists a Borel set* $A \subseteq \mathbb{Z}^\infty$ *such that* $\mathbf{P}(A) \geq 1 - \epsilon$ *and* $||\mathbb{G}_n(\boldsymbol{z}) - \mathbb{G}||_{\mathcal{S}} = o(1)$ *for all* $\boldsymbol{z} \in A$.

In the following proof, almost uniformly means that for any $\epsilon > 0$, there exists a Borel set $A \in \tilde{\mathbb{Z}^\infty}$ such that $\tilde{\boldsymbol{P}}(A) \geq 1 - \epsilon$ and $\sup_{\tau \in A} ||\tilde{\mathbb{G}}_n(\tau) - \tilde{\mathbb{G}}||_{L^\infty(\mathcal{S})} = o(1)$.

Proof of Lemma E.1. It is well-known that the following representation is also valid: $\mathbb{G}_n : \mathbb{Z}^\infty \to L^\infty(\mathcal{S})$. Since $\mathcal{S}$ is a Donsker Class, $\mathbb{G}_n$ converges weakly to some $\mathbb{G}$ tight Borel measurable element in $L^\infty(\mathcal{S})$ (e.g. see van der Vaart and Wellner [1996] Ch. 2.1). By Theorem 1.10.3 in van der Vaart and Wellner [1996] there exists a probability space $(\tilde{\mathbb{Z}^\infty}, \tilde{\boldsymbol{P}})$ and a sequence of maps $\tilde{\mathbb{G}}_n : \tilde{\mathbb{Z}^\infty} \to L^\infty(\mathcal{S})$ for all $n \in \mathbb{N}$ and $\tilde{\mathbb{G}} : \tilde{\mathbb{Z}^\infty} \to L^\infty(\mathcal{S})$ such that (i) $||\tilde{\mathbb{G}}_n - \tilde{\mathbb{G}}||_{L^\infty(\mathcal{S})} = o(1)$ almost uniformly; and (ii) $\tilde{\mathbb{G}}_n$ and $\tilde{\mathbb{G}}$ have the same law as $\mathbb{G}_n$ and $\mathbb{G}$ resp. □

E.1 Appendix for Example 5.1

Lemma E.2. *For all* $k \in \mathbb{N}$ *and* $P \in \mathcal{M}$,

$$||\varphi_k(P)||_{L^2(P)} \leq ||p||^2_{L^\infty(\mathbb{R})} ||\kappa||^2_{L^1(\mathbb{R})}.$$

Proof. It suffices to show that $E_P[|(\kappa_k \star P)(Z)|^2] \leq ||p||^2_{L^\infty} \left(\int |\kappa(u)|du\right)^2$. To do this, note that

$$\begin{aligned} E_P[|(\kappa_k \star P)(Z)|^2] =& \int \left(\int k\kappa((x-z)k)p(x)dx \right)^2 p(z)dx \\ =& \int \left(\int \kappa(u)p(z+u/k)du \right)^2 p(z)dz \\ \leq& ||p||^2_{L^\infty} \left(\int |\kappa(u)|du \right)^2 . \end{aligned}$$

□

Proof of Proposition 5.1. Consider the curve $t \mapsto P + tQ$ for any $Q \in ca(\mathbb{R})$. It is a valid curve because $\mathbb{D}_\psi = ca(\mathbb{R})$. Therefore

$$\begin{aligned} \psi_k(P+tQ) - \psi_k(P) =& t\left\{ \int (\kappa_k \star Q)(x)P(dx) + \int (\kappa_k \star P)(x)Q(dx) \right\} \\ &+ t^2 \int (\kappa_k \star Q)(x)Q(dx). \end{aligned}$$

Since κ is symmetric, $\int(\kappa_k \star P)(x)Q(dx) = \int(\kappa_k \star Q)(x)P(dx)$. From this display, $Q \mapsto \eta_k(Q) = \int(\kappa_k \star Q)(x)Q(dx)$ and $Q \mapsto D\psi_k(P)[Q] = 2\int(\kappa_k \star P)(x)Q(dx)$.

The mapping $Q \mapsto 2\int(\kappa_k \star P)(x)Q(dx)$ is clearly linear. Also, note that for any reals x and x'

$$\begin{aligned}|(\kappa_k \star P)(x) - (\kappa_k \star P)(x')| =& k\left|\int p(y)(\kappa(k(y-x))dy - \kappa(k(y-x')))dy\right| \\ \leq& Ck^2|x-x'|\end{aligned}$$

for some $C < \infty$. The last inequality follows from the fact that $x \mapsto \kappa(x)$ is smooth. Therefore $x \mapsto (k^2/C)(\kappa_k \star P)(x)$ is in LB. This implies that for any Q' and Q in $ca(\mathbb{Z})$,

$$\left|\int(\kappa_k \star P)(x)Q'(dx) - \int(\kappa_k \star P)(x)Q(dx)\right| \leq k^2C||Q'-Q||_{LB}$$

and thus $Q \mapsto 2\int(\kappa_k \star P)(x)Q(dx)$ is $||.||_{LB}$-continuous.

We now bound $Q \mapsto \eta_k(Q)$. For any $x' > x$ and any $Q \in \mathcal{T}_P$, it follows that

$$\begin{aligned}\kappa_k \star Q(x) - \kappa_k \star Q(x') =& \int k(\kappa(k(y-x)) - \kappa(k(y-x')))Q(dy) \\ =& \int k^2 \int_x^{x'} \kappa'(y-t)dtQ(dy) \\ =& k^2 \int_x^{x'} \int \kappa'(y-t)Q(dy)\end{aligned}$$

where the last line follows from the fact that $t \mapsto \kappa(t)$ is bounded. Since κ is smooth, $y \mapsto \kappa'(y-t)$ is Lipschitz (with some constant L) for any $t \in \mathbb{R}$. Hence

$$|\kappa_k \star Q(x) - \kappa_k \star Q(x')| \leq L|x'-x|||Q||_{LB}.$$

Thus, the mapping $x \mapsto (\kappa_k \star Q)(x)$ is bounded and Lipschitz with constant $L||Q||_{LB}$. Therefore,

$$|\eta_k(Q)| \leq L||Q||^2_{LB}.$$

□

E.2 Appendix for Examples 5.2 and 5.3

First, note that $\mathbb{D}_\psi \subseteq ca(\mathbb{R} \times [0,1]^2)$ (defined in Appendix B.1). For any $k \in \mathbb{N}$, let $F_k : L^2([0,1]) \times \mathbb{D}_\psi \to L^2([0,1])$ be such that

$$F_k(\theta, Q) \equiv \begin{cases} (T^*_{k,Q}T_{k,Q} + \lambda_k I)[\theta] - T^*_{k,Q}[r_{k,Q}] & for\ Penalization-Based \\ (\Pi^*_k T^*_{k,Q}T_{k,Q}\Pi_k)[\theta] - \Pi^*_k T^*_{k,Q}[r_{k,Q}] & for\ Sieve-Based\end{cases}$$

for any $(\theta, Q) \in L^2([0,1]) \times \mathbb{D}_\psi$. Note that for any $Q \in \mathbb{D}_\psi$ the integrals defining the operators are well-defined by assumptions and so is $T^*_{k,Q}$; see Appendix B.1 for a discussion.

Let $\varepsilon_P(Y,W) \equiv Y - \psi(P)(W)$ and $\varepsilon_{k,P}(Y,W) \equiv Y - \psi_k(P)(W)$. Also, throughout this section we use the notation introduced in Appendix B.1 to denote $T_{k,P}$ and other quantities.

Lemma E.3. *For any $P \in \mathbb{D}_\psi$ and any $k \in \mathbb{N}$, ψ_k is $||.||_{LB}$-Frechet differentiable tangential to $\mathbb{D}_\psi$ at P with derivative given by:*

(1) For the Penalization-Based:

$$\begin{aligned} D\psi_k(P)[Q] =&(T^*_{k,P}T_{k,P} + \lambda_k I)^{-1} T^*_{k,P}\mathbb{T}_{k,Q}[\varepsilon_{k,P}] \\ &- (T^*_{k,P}T_{k,P} + \lambda_k I)^{-1} T^*_{k,Q} T_{k,P}[\psi_k(P) - \psi(P)],\ \forall Q \in \mathbb{D}_\psi \end{aligned}$$

where $x \mapsto \mathbb{T}_{k,P}[g](x) \equiv \int \kappa_k(x' - \cdot) \int (\psi_k(P)(w) - y) Q(dy, dw, dx')$.

(2) For the Sieve-Based:

$$\begin{aligned} D\psi_k(P)[Q] =&(\Pi^*_k T^*_{k,P} T_{k,P} \Pi_k)^{-1} \Pi^*_k T^*_{k,P} \mathbb{T}_{k,Q}[\varepsilon_{k,P}] \\ &- (\Pi^*_k T^*_{k,P} T_{k,P} \Pi_k)^{-1} \Pi^*_k T^*_{k,Q} T_{k,P}[\psi_k(P) - \psi(P)],\ \forall Q \in \mathbb{D}_\psi \end{aligned}$$

where $x \mapsto \mathbb{T}_{k,P}[g](x) \equiv (u^{J(k)}(x))^T Q^{-1}_{uu} E_P[u^{J(k)}(X) g(Y,W)]$ *for any* $g \in L^2(P)$.

Proof. See the end of this Section. □

The following corollary trivially follows.

Corollary E.1. For the sieve-based and the penalization-based: For any $P \in \mathbb{D}_\psi$

(1) The regularization $\boldsymbol{\gamma}$ is $DIFF(P, \mathcal{E}_{||.||_{LB}})$.

(2) For each $k \in \mathbb{N}$, the reminder of γ_k, η_k, is such that $|\eta_k(\zeta)| = o(||\zeta||_{LB})$, for any $\zeta \in \mathbb{D}_\psi$.[27]

Proof. See the end of this Section. □

Proof of Proposition 5.2. The result follows from Corollary E.1. Lemma E.3(2) derives the expression for $D\psi_k(P)$; we now expand this expression in terms of the basis functions.

For any $g, f \in L^2([0,1])$,

$$\begin{aligned} T_{k,P}\Pi_k[g](x) =&T_{k,P}\left[(v^{L(k)}(.))^T Q^{-1}_{vv} E_{Leb}[v^{L(k)}(W) g(W)]\right](x) \\ =&(u^{J(k)}(x))^T Q^{-1}_{uu} Q_{uv} Q^{-1}_{vv} E_{Leb}[v^{L(k)}(W) g(W)], \end{aligned}$$

[27] The "o" function may depend on k.

and

$$\begin{aligned}\langle T_{k,P}\Pi_k[g], f\rangle_{L^2([0,1])} &= \int (u^{J(k)}(x))^T Q_{uu}^{-1} Q_{uv} Q_{vv}^{-1} E_{Leb}[v^{L(k)}(W)g(W)] f(x)dx \\ &= \int E_{Leb}[(u^{J(k)}(X))^T f(x)] Q_{uu}^{-1} Q_{uv} Q_{vv}^{-1} v^{L(k)}(w) g(w) dw\end{aligned}$$

so $\Pi_k^* T_{k,P}^* : L^2([0,1], Leb) \to L^2([0,1], Leb)$ and is given by

$$f \mapsto \Pi_k^* T_{k,P}^*[f](.) = E_{Leb}[(u^{J(k)}(X))^T f(X)] Q_{uu}^{-1} Q_{uv} Q_{vv}^{-1} v^{L(k)}(.).$$

Hence

$$\Pi_k^* T_{k,P}^* T_{k,P} \Pi_k[g](.) = (v^{L(k)}(.))^T Q_{vv}^{-1} Q_{uv}^T Q_{uu}^{-1} Q_{uv} Q_{vv}^{-1} E_{Leb}[v^{L(k)}(W) g(W)].$$

We now compute the inverse of this operator. Consider solving for $g(.) = (v^{L(k)}(.))^T Q_{vv}^{-1} b$ for some $b \in \mathbb{R}^k$ such that

$$\begin{aligned}&\Pi_k^* T_{k,P}^* T_{k,P} \Pi_k[g](.) = (v^{L(k)}(.))^T Q_{vv}^{-1} b \\ \iff& Q_{uv}^T Q_{uu}^{-1} Q_{uv} Q_{vv}^{-1} E_{Leb}[v^{L(k)}(W) g(W)] = b \\ \iff& E_{Leb}[v^{L(k)}(W) g(W)] = Q_{vv} (Q_{uv}^T Q_{uu}^{-1} Q_{uv})^{-1} b.\end{aligned}$$

Hence, $(\Pi_k^* T_{k,P}^* T_{k,P} \Pi_k)^{-1}[g](.) = (v^{L(k)}(.))^T (Q_{uv}^T Q_{uu}^{-1} Q_{uv})^{-1} b$. Therefore

$$\begin{aligned}(\Pi_k^* T_{k,P}^* T_{k,P} \Pi_k)^{-1} \Pi_k^* T_{k,P}^* \mathbb{T}_{k,Q}[\varepsilon_{k,P}] =& (v^{L(k)}(.))^T (Q_{uv}^T Q_{uu}^{-1} Q_{uv})^{-1} Q_{uv} Q_{uu}^{-1} E_{Leb}[u^{J(k)}(X) \mathbb{T}_{k,Q}[\varepsilon_{k,P}]] \\ =& (v^{L(k)}(.))^T (Q_{uv}^T Q_{uu}^{-1} Q_{uv})^{-1} Q_{uv} Q_{uu}^{-1} E_Q[u^{J(k)}(X) \varepsilon_{k,P}(Y,W)].\end{aligned}$$

And

$$\begin{aligned}&(\Pi_k^* T_{k,P}^* T_{k,P} \Pi_k)^{-1} \Pi_k^* T_{k,Q}^* T_{k,P}[\psi_k(P) - \psi_{id}(P)] \\ =& (v^{L(k)}(.))^T (Q_{uv}^T Q_{uu}^{-1} Q_{uv})^{-1} E_{Leb}[v^{L(k)}(W) \Pi_k^* T_{k,Q}^* T_{k,P}[\psi_k(P) - \psi_{id}(P)](W)].\end{aligned}$$

It is easy to see that for any Q, $D\gamma_k(P)[Q] = \int \pi(w) D\psi_k(P)(w)[Q]dw$, the goal is to cast this as $\int D\psi_k^*(P)[\pi](z)Q(dz)$. To this end, note that

$$\begin{aligned}&\int \pi(w) D\psi_k(P)(w)[Q]dw \\ =& \int \pi(w)(v^{L(k)}(w))^T (Q_{uv}^T Q_{uu}^{-1} Q_{uv})^{-1} Q_{uv} Q_{uu}^{-1} E_Q[u^{J(k)}(X) \varepsilon_{k,P}(Y,W)]dw \\ &- \int \pi(w)(v^{L(k)}(w))^T (Q_{uv}^T Q_{uu}^{-1} Q_{uv})^{-1} E_{Leb}[v^{L(k)}(W) \Pi_k^* T_{k,Q}^* T_{k,P}[\psi_k(P) - \psi_{id}(P)](W)]dw \\ \equiv& Term_{1,k} + Term_{2,k}.\end{aligned}$$

Regarding the first term, note that

$$Term_{1,k} = \int E_{Leb}[\pi(W)(v^{L(k)}(W))^T](Q_{uv}^T Q_{uu}^{-1} Q_{uv})^{-1} Q_{uv} Q_{uu}^{-1} u^{J(k)}(x)\varepsilon_{k,P}(y,w)Q(dy,dw,dx).$$

We can cast $Term_{2,k} = -\langle \Pi_k[\pi], (\Pi_k^* T_{k,P}^* T_{k,P}\Pi_k)^{-1}\left[\Pi_k^* T_{k,Q}^* T_{k,P}[\psi_k(P) - \psi_{id}(P)]\right]\rangle_{L^2}$, and thus

$$\begin{aligned} Term_{2,k} =& - \langle T_{k,Q}(\Pi_k^* T_{k,P}^* T_{k,P}\Pi_k)^{-1}\Pi_k[\pi], T_{k,P}[\psi_k(P) - \psi_{id}(P)]\rangle_{L^2} \\ =& - \int (u^{J(k)}(x))^T Q_{uu}^{-1} E_Q[u^{J(k)}(X)(\Pi_k^* T_{k,P}^* T_{k,P}\Pi_k)^{-1}\Pi_k[\pi](W)]T_{k,P}[\psi_k(P) - \psi_{id}(P)](x)dx \\ =& \int E_P[(\psi_{id}(P)(W) - \psi_k(P)(W))(u^{J(k)}(X))^T]Q_{uu}^{-1}u^{J(k)}(x)(\Pi_k^* T_{k,P}^* T_{k,P}\Pi_k)^{-1}\Pi_k[\pi](w) \\ & \times Q(dw,dx) \end{aligned}$$

where the second line follows from definition of $T_{k,P}$.

Therefore,

$$\begin{aligned} D\psi_k^*(P)[\pi](y,w,x) =& E_{Leb}[\pi(W)(v^{L(k)}(W))^T](Q_{uv}^T Q_{uu}^{-1} Q_{uv})^{-1} Q_{uv} Q_{uu}^{-1} u^{J(k)}(x)\varepsilon_{k,P}(y,w) \\ & + E_P[(\psi_{id}(P)(W) - \psi_k(P)(W))(u^{J(k)}(X))^T]Q_{uu}^{-1}u^{J(k)}(x) \\ & \times (v^{L(k)}(w))^T (Q_{uv}^T Q_{uu}^{-1} Q_{uv})^{-1} E_{Leb}[v^{L(k)}(W)\pi(W)]. \end{aligned}$$

In the operator notation this expression equals

$$\begin{aligned} D\psi_k^*(P)[\pi](y,w,x) =& T_{k,P}(\Pi_k^* T_{k,P}^* T_{k,P}\Pi_k)^{-1}\Pi_k[\pi](x)\varepsilon_{k,P}(y,w) \\ & + T_{k,P}[\psi_{id}(P) - \psi_k(P)](x) \times (\Pi_k^* T_{k,P}^* T_{k,P}\Pi_k)^{-1}\Pi_k[\pi](w). \end{aligned} \tag{23}$$

□

Proof of Proposition 5.3. The result follows from Corollary E.1. Lemma E.3(1) derives the expression for $D\psi_k(P)$; we now expand this expression in terms of the basis functions.

Note that $\psi_k(P) = (T_{k,P}^* T_{k,P} + \lambda_k I)^{-1} T_{k,P}^* r_{k,P}$ and $r_{k,P}(.) = \int \kappa_k(x' - \cdot)\int \psi_{id}(P)(w)p(w,x')dwdx'$ so that $\psi_k(P) = (T_{k,P}^* T_{k,P} + \lambda_k I)^{-1} T_{k,P}^* T_{k,P}[\psi_{id}(P)]$. Hence

$$\begin{aligned} & (T_{k,P}^* T_{k,P} + \lambda_k I)^{-1} T_{k,Q}^* T_{k,P}[\psi_k(P) - \psi_{id}(P)] \\ =& (T_{k,P}^* T_{k,P} + \lambda_k I)^{-1} T_{k,Q}^* T_{k,P}[((T_{k,P}^* T_{k,P} + \lambda_k I)^{-1} T_{k,P}^* T_{k,P} - I)\psi_{id}(P)] \\ =& - \lambda_k (T_{k,P}^* T_{k,P} + \lambda_k I)^{-1} T_{k,Q}^* T_{k,P} (T_{k,P}^* T_{k,P} + \lambda_k I)^{-1}[\psi_{id}(P)]. \end{aligned}$$

Thus

$$\begin{aligned}D\gamma_k(P)[Q] =&\langle \pi, \mathcal{R}_{k,P} T^*_{k,P} \mathbb{T}_{k,Q}[\varepsilon_{k,P}] \rangle_{L^2} \\ &+ \lambda_k \langle \pi, \mathcal{R}_{k,P} T^*_{k,Q} T_{k,P} \mathcal{R}_{k,P}[\psi_{id}(P)] \rangle_{L^2} \\ =&\langle T_{k,P}\mathcal{R}_{k,P}[\pi], \mathbb{T}_{k,Q}[\varepsilon_{k,P}] \rangle_{L^2} \\ &+ \lambda_k \langle T_{k,Q}\mathcal{R}_{k,P}[\pi], T_{k,P}\mathcal{R}_{k,P}[\psi_{id}(P)] \rangle_{L^2}.\end{aligned}$$

Note that

$$T_{k,P}[g](x) = \int \kappa_k(x'-x) \int g(w)p(w,x')dw, dx' = \int \kappa_k(x'-x) T_P[g](x')dx' = \mathcal{K}_k[T_P[g]](x)$$

and by symmetry of κ, $T^*_{k,P} = T_P[\mathcal{K}_k]$, where $\mathcal{K}_k$ is simply the convolution operator.

Therefore,

$$\begin{aligned}\langle T_{k,P}\mathcal{R}_{k,P}[\pi], \mathbb{T}_{k,Q}[\varepsilon_{k,P}] \rangle_{L^2} &= \int T_{k,P}\mathcal{R}_{k,P}[\pi](x) \int \kappa_k(x'-x) \int \varepsilon_{k,P}(y,w) Q(dy,dw,dx')dx \\ &= \int \left(\int \kappa_k(x-x') T_{k,P}\mathcal{R}_{k,P}[\pi](x)dx \right) \varepsilon_{k,P}(y,w) Q(dy,dw,dx') \\ &= \int \mathcal{K}^2_k T_P \mathcal{R}_{k,P}[\pi](x) \varepsilon_{k,P}(y,w) Q(dy,dw,dx).\end{aligned}$$

and

$$\begin{aligned}\langle T_{k,Q}\mathcal{R}_{k,P}[\pi], T_{k,P}\mathcal{R}_{k,P}[\psi_{id}(P)] \rangle_{L^2} &= \int\int \kappa_k(x'-x) \int \mathcal{R}_{k,P}[\pi](w) Q(dw,dx') T_{k,P}\mathcal{R}_{k,P}[\psi_{id}(P)](x)dx \\ &= \int \mathcal{R}_{k,P}[\pi](w) \left(\int \kappa_k(x'-x) T_{k,P}\mathcal{R}_{k,P}[\psi_{id}(P)](x)dx \right) Q(dw,dx') \\ &= \int \mathcal{R}_{k,P}[\pi](w) \mathcal{K}^2_k T_P \mathcal{R}_{k,P}[\psi_{id}(P)](x) Q(dw,dx').\end{aligned}$$

Therefore

$$D\psi^*_k(P)[\pi](y,w,x) = \mathcal{K}^2_k T_P \mathcal{R}_{k,P}[\pi](x)\varepsilon_{k,P}(y,w) + \lambda_k \mathcal{R}_{k,P}[\pi](w) \mathcal{K}^2_k T_P \mathcal{R}_{k,P}[\psi_{id}(P)](x).$$

□

E.2.1 Proofs of Supplementary Lemmas.

Proof of Lemma E.3. The proof follows by the Implicit Function Theorem in Ambrosetti and Prodi [1995] p. 38 with one minor modification.

First observe that F_k takes values in $L^2([0,1]) \times \mathbb{D}_\psi$ which is a subspace of $L^2([0,1]) \times \overline{\mathbb{D}}_\psi$ — the closure being taken with respect to $||.||_{LB}$. The space $L^2([0,1]) \times \overline{\mathbb{D}}_\psi$ is a Banach space under the norm $||.||_{L^2(P)} + ||.||_{LB}$.

We now check the rest of the assumptions of the theorem for each case separately.

(1) Observe that $\theta \mapsto F_k(\theta, Q)$ is linear, so $\frac{dF_k(\psi_k(P),P)}{d\theta} = (T^*_{k,P}T_{k,P} + \lambda_k I) : L^2([0,1]) \to L^2([0,1])$. By our conditions (1)-(2) stated in Example 3.2 $Kernel((T^*_{k,P}T_{k,P} + \lambda_k I)) = \{0\}$ and $(T^*_{k,P}T_{k,P} + \lambda_k I)$ has closed range — the range of an operator A is closed iff 0 is not an accumulation point of the spectrum of A^*A. Thus $\frac{dF_k(\psi_k(P),P)}{d\theta}$ is 1-to-1 and onto.

Thus, by the Implicit Function Theorem in Ambrosetti and Prodi [1995] p. 38, there exists a $||.||_{LB}$-open set U of P in $\overline{\mathbb{D}}_\psi$ such that $D\psi_k(P) = \left(\frac{dF_k(\psi_k(P),P)}{d\theta}\right)^{-1}\left[\frac{dF_k(\psi_k(P),P)}{dP}\right]$ for any $P \in U$.

We now characterize this expression. For any $k \in \mathbb{N}$ and any $Q \in \mathbb{D}_\psi$,

$$\begin{aligned}\frac{dF_k(\psi_k(P),P)}{dP}[Q] =&(T^*_{k,Q}T_{k,P} + T^*_{k,P}T_{k,Q})[\psi_k(P)] - \left(T^*_{k,Q}[r_{k,P}] + T^*_{k,P}[r_{k,Q}]\right)\\ =& - T^*_{k,Q}\left[r_{k,P} - T_{k,P}[\psi_k(P)]\right] + T^*_{k,P}\left[T_{k,Q}[\psi_k(P)] - r_{k,Q}\right]\\ \equiv& Term_1 + Term_2.\end{aligned}$$

Note that

$$\begin{aligned}r_{k,P} - T_{k,P}[\psi_k(P)](.) =& \int \kappa_k(x' - \cdot)(y - \psi_k(P)(w))p(y,w,x')dydwdx'\\ =& T_{k,P}[\psi(P) - \psi_k(P)](.)\end{aligned}$$

where the last equality follows because $\int(y - \psi(P)(w))p(y,w,X)dydw = 0$. Thus

$$Term_1 = - T^*_{k,Q}T_{k,P}[\psi(P) - \psi_{id}(P)].$$

Also

$$T_{k,Q}[\psi_k(P)](.) - r_{k,Q}(.) = \int \kappa_k(x' - \cdot)\int(\psi_k(P)(w) - y)Q(dy,dw,dx')$$

so

$$Term_3 = T_{k,P}\mathbb{T}_{k,Q}[\varepsilon_{k,P}]$$

Thus, the result follows.

(2) The proof is analogous to the one for part (1), so we only present an sketch. By assumption, $\Pi^*_k T^*_{k,P}T_{k,P}\Pi_k$ is 1-to-1 for each P.

Also, for any $k \in \mathbb{N}$ and any $Q \in \mathbb{D}_\psi$,

$$\begin{aligned}\frac{dF_k(\psi_k(P),P)}{dP}[Q] =&(\Pi^*_k T^*_{k,Q}T_{k,P}\Pi_k + \Pi^*_k T^*_{k,P}T_{k,Q}\Pi_k)[\psi_k(P)]\\ &- \Pi^*_k\left(T^*_{k,Q}[r_{k,P}] + T^*_{k,P}[r_{k,Q}]\right)\\ =&\Pi^*_k\left[(T^*_{k,Q}T_{k,P} + T^*_{k,P}T_{k,Q})[\psi_k(P)] - \left(T^*_{k,Q}[r_{k,P}] + T^*_{k,P}[r_{k,Q}]\right)\right]\end{aligned}$$

where the second line follows because $\Pi_k[\psi_k(P)] = \psi_k(P)$.

We note that

$$T_{k,Q}[\psi_k(P)] - r_{k,Q} = (u^{J(k)}(X))^T Q_{uu}^{-1} \int u^{J(k)}(x)(\psi_k(P)(w) - y)Q(dy, dw)$$
$$= - \mathbb{T}_{k,Q}[\varepsilon_{k,P}]$$

and

$$T_{k,P}[\psi_k(P)] - r_{k,P} = (u^{J(k)}(X))^T Q_{uu}^{-1} \int u^{J(k)}(x)(\psi_k(P)(w) - \psi_{id}(P)(w) - \varepsilon_P(y, w))P(dy, dw)$$
$$= T_{k,P}[\psi_k(P) - \psi_{id}(P)]$$

since $\int \varepsilon_P(y, w))P(dy, dw) = 0$. □

Proof of Corollary E.1. (1) By lemma E.3, for each $k \in \mathbb{N}$, ψ_k is $||.||_{LB}$-Frechet differentiable, i.e., for any $Q \in \mathbb{D}_\psi$,

$$\|\psi_k(Q) - \psi_k(P) - D\psi_k(P)[Q - P]\|_{L^2([0,1])} = o(||P - Q||_{LB}).$$

Since $\mathbb{D}_\psi$ is linear and $\mathbb{D}_\psi \supseteq lin(\mathcal{D} - \{P\})$ (see Lemma B.3), the curve $t \mapsto P + t\zeta$ with $\zeta \in \mathbb{D}_\psi$ maps into $\mathbb{D}_\psi$. Therefore, $\boldsymbol{\psi}$ is $DIFF(P, \mathcal{E}_{||.||_{LB}})$.

By duality

$$\sup_{\ell \in L^2([0,1]) : \, ||\ell||_{L^2([0,1])} = 1} |\ell[\psi_k(P + t\zeta) - \psi_k(P) - tD\psi_k(P)[\zeta]]| = to(||\zeta||_{LB})$$

(here we are abusing notation by using ℓ as both an element of $L^2([0, 1])$ and as the functional). Since γ_k is linear functional of ψ_k this display readily implies that γ_k is $||.||_{LB}$-Frechet differentiable. This in turn implies part (1).

(2) Part (1) implies that,

$$|\eta_k(\zeta)| = o(||\zeta||_{LB})$$

for any $\zeta \in \mathbb{D}_\psi$.

□

E.3 Appendix for Section 5.5

In this section, with a slight abuse of notation we will use $\mathcal{L}_n$ or $\mathcal{L}_n(\mathbf{z})$ to denote $\mathcal{L}_n(\Lambda_k)_k)$ for a given realization of the data $\mathbf{z}$.

E.3.1 Proof of Proposition 5.4

In analogy to the proof of Theorem 4.2, let $n \mapsto k(n)$ be defined as

$$k(n) \equiv \min \left\{ k \in \mathbb{R}_+ : \bar{\delta}_{1,k}(n) \geq \sqrt{n} \bar{B}_k(P) \right\}.$$

Also, for each $n \in \mathbb{N}$, let

$$A_n \equiv \left\{ \boldsymbol{z} \in \mathbb{Z}^\infty : \sup_{k \in \mathcal{G}_n} \frac{\sqrt{n} |\eta_k(P_n(\boldsymbol{z}) - P)|}{\bar{\delta}_{1,k}(n)} \leq 1 \right\},$$

and

$$B_n \equiv \left\{ \boldsymbol{z} \in \mathbb{Z}^\infty : \sup_{k' \geq k \ in \ \mathcal{G}_n} \frac{\sqrt{n} \, |D\psi_{k'}(P)[P_n(\boldsymbol{z}) - P] - D\psi_k(P)[P_n(\boldsymbol{z}) - P]|}{\bar{\delta}_{2,k'}(n)} \leq 1 \right\}.$$

and

$$k \mapsto \bar{\delta}_k(n) \equiv \bar{\delta}_{1,k}(n) + \bar{\delta}_{2,k}(n).$$

Lemma E.4. *Suppose there exists a sequence $(j_n)_n$ such that*

1. *For any $\epsilon > 0$, there exists a N such that $\mathbf{P}(\{\mathbf{z} \in \mathbb{Z}^\infty : j_n \in \mathcal{L}_n(\mathbf{z})\} \cap A_n \cap B_n) \geq 1 - \epsilon$ for all $n \geq N$.*
2. *There exists a $L < \infty$ such that $\frac{\bar{\delta}_{j_n}(n) + \bar{B}_{j_n}(P)}{||\varphi_{j_n}(P)||_{L^2(P)}} \leq L \inf_{k \in \mathcal{G}_n} \frac{\bar{\delta}_k(n) + \sqrt{n} \bar{B}_k(P)}{||\varphi_k(P)||_{L^2(P)}}$.*

Then for any $\epsilon > 0$, there exists a N such that

$$\mathbf{P}\left(\frac{\sqrt{n}}{||\varphi_{\tilde{k}(n)}(P)||_{L^2(P)}} \left| \psi_{\tilde{k}(n)}(P_n) - \psi(P) - D\psi_{\tilde{k}(n)}(P)[P_n - P] \right| \geq 2 C_n L \inf_{k \in \mathcal{G}_n} \frac{\bar{\delta}_k(n) + \sqrt{n} \bar{B}_k(P)}{||\varphi_k(P)||_{L^2(P)}} \right) \leq \epsilon$$

for all $n \geq N$, where C is as in Assumption 5.1(iv).

Proof. For any $n \in \mathbb{N}$ and any $\boldsymbol{z} \in \{\boldsymbol{z} \in \mathbb{Z}^\infty : j_n \in \mathcal{L}_n(\boldsymbol{z})\} \cap A_n \cap B_n$

$$\begin{aligned}
&\frac{\sqrt{n}}{||\varphi_{\tilde{k}_n(\boldsymbol{z})}(P)||_{L^2(P)}} \left| \psi_{\tilde{k}_n(\boldsymbol{z})}(P_n(\boldsymbol{z})) - \psi(P) - D\psi_{\tilde{k}_n(\boldsymbol{z})}(P)[P_n(\boldsymbol{z}) - P] \right| \\
\leq & \sqrt{n} \frac{|\eta_{j_n}(P_n(\boldsymbol{z}) - P)| + \bar{B}_{j_n}(P)}{||\varphi_{\tilde{k}_n(\boldsymbol{z})}(P)||_{L^2(P)}} \\
& + \frac{\sqrt{n}}{||\varphi_{\tilde{k}_n(\boldsymbol{z})}(P)||_{L^2(P)}} \left(\left| \psi_{\tilde{k}_n(\boldsymbol{z})}(P_n(\boldsymbol{z})) - \psi_{j_n}(P_n(\boldsymbol{z})) \right| + \left| D\psi_{\tilde{k}_n(\boldsymbol{z})}(P)[P_n(\boldsymbol{z}) - P] - D\psi_{j_n}(P)[P_n(\boldsymbol{z}) - P] \right| \right) \\
\leq & 2 \frac{\bar{\delta}_{1,j_n}(n) + \sqrt{n} \bar{B}_{j_n}(P)}{||\varphi_{\tilde{k}_n(\boldsymbol{z})}(P)||_{L^2(P)}} + \sqrt{n} \frac{\bar{\delta}_{2,j_n}(n)}{||\varphi_{\tilde{k}_n(\boldsymbol{z})}(P)||_{L^2(P)}}
\end{aligned}$$

where the first inequality follows from the definition of differentiability and simple algebra; the second inequality follows from the fact that $\boldsymbol{z} \in A_n \cap B_n$ and from the fact that $j_n \geq \tilde{k}_n(\boldsymbol{z})$ and the definition of $\mathcal{L}_n(\mathbf{z})$.

This result and the definition of $(C_n)_n$ in the Proposition 5.4 imply

$$\frac{\sqrt{n}}{||\varphi_{\tilde{k}_n(\boldsymbol{z})}(P)||_{L^2(P)}} \left| \psi_{\tilde{k}_n(\boldsymbol{z})}(P_n(\boldsymbol{z})) - \psi(P) - D\psi_{\tilde{k}_n(\boldsymbol{z})}(P)[P_n(\boldsymbol{z}) - P] \right| \leq 2C_n \frac{\bar{\delta}_{j_n}(n) + \sqrt{n}\bar{B}_{j_n}(P)}{||\varphi_{j_n}(P)||_{L^2(P)}}.$$

Thus, by condition 2,

$$\frac{\sqrt{n}}{||\varphi_{\tilde{k}_n(\boldsymbol{z})}(P)||_{L^2(P)}} \left| \psi_{\tilde{k}_n(\boldsymbol{z})}(P_n(\boldsymbol{z})) - \psi(P) - D\psi_{\tilde{k}_n(\boldsymbol{z})}(P)[P_n(\boldsymbol{z}) - P] \right| \leq 2LC_n \inf_{k \in \mathcal{G}_n} \frac{\bar{\delta}_k(n) + \sqrt{n}\bar{B}_k(P)}{||\varphi_k(P)||_{L^2(P)}}.$$

The result thus follows from condition 1. □

We now construct a sequence $(h_n)_n$ that satisfies both conditions of the lemma. The construction is completely analogous to the one in the proof of Theorem 4.2 but using $\bar{\delta}_k(n)/\sqrt{n}$ instead of $\delta_k(r_n^{-1})$.

Let, for each $n \in \mathbb{N}$,

$$\begin{aligned} \mathcal{G}_n^+ &\equiv \{k \in \mathcal{G}_n \colon \bar{\delta}_k(n)/\sqrt{n} \geq \bar{B}_k(P)\} \textit{ and} \\ \mathcal{G}_n^- &\equiv \{k \in \mathcal{G}_n \colon \bar{\delta}_k(n)/\sqrt{n} \leq \bar{B}_k(P)\}. \end{aligned}$$

For each $n \in \mathbb{N}$, let

$$T_n^+ = \frac{\bar{\delta}_{h_n^+}(n) + \sqrt{n}\bar{B}_{h_n^+}(P)}{||\varphi_{h_n^+}(P)||_{L^2(P)}}$$

if $\mathcal{G}_n^+$ is non-empty where

$$h_n^+ = \min\{k \colon k \in \mathcal{G}_n^+\};$$

and $T_n^+ = +\infty$, if $\mathcal{G}_n^+$ is empty. Similarly,

$$T_n^- = \frac{\bar{\delta}_{h_n^-}(n) + \sqrt{n}\bar{B}_{h_n^-}(P)}{||\varphi_{h_n^-}(P)||_{L^2(P)}}$$

if $\mathcal{G}_n^-$ is non-empty where

$$h_n^- = \max\{k \colon k \in \mathcal{G}_n^-\};$$

and $T_n^- = +\infty$, if $\mathcal{G}_n^-$ is empty.

Finally, for each $n \in \mathbb{N}$, let $h_n \in \mathcal{G}_n$ be such that

$$h_n = h_n^+ 1\{T_n^+ \leq T_n^-\} + h_n^- 1\{T_n^+ > T_n^-\}.$$

Lemma E.5. *For each $n \in \mathbb{N}$, h_n exists and*

$$\frac{\bar{\delta}_{h_n}(n) + \sqrt{n}\bar{B}_{h_n}(P)}{||\varphi_{h_n}(P)||_{L^2(P)}} = \min\left\{T_n^-, T_n^+\right\}.$$

Proof. The proof is identical to the one of Lemma D.2. □

Lemma E.6. *Suppose Assumption 5.1 holds. For each $n \in \mathbb{N}$,*

$$\frac{\bar{\delta}_{h_n}(n) + \sqrt{n}\bar{B}_{h_n}(P)}{||\varphi_{h_n}(P)||_{L^2(P)}} \leq 2C_n \inf_{k \in \mathcal{G}_n} \frac{\bar{\delta}_k(n) + \sqrt{n}\bar{B}_k(P)}{||\varphi_k(P)||_{L^2(P)}}$$

where $(C_n)_n$ is as in Proposition 5.4.

Proof. Observe that

$$\inf_{k \in \mathcal{G}_n} \frac{\bar{\delta}_k(n) + \sqrt{n}\bar{B}_k(P)}{||\varphi_k(P)||_{L^2(P)}} \geq \min\left\{\inf_{k \in \mathcal{G}_n^+} \frac{\bar{\delta}_k(n) + \sqrt{n}\bar{B}_k(P)}{||\varphi_k(P)||_{L^2(P)}}, \inf_{k \in \mathcal{G}_n^-} \frac{\bar{\delta}_k(n) + \sqrt{n}\bar{B}_k(P)}{||\varphi_k(P)||_{L^2(P)}}\right\}$$

where the infimum is defined as $+\infty$ if the corresponding set is empty.

Fix any $n \in \mathbb{N}$, if $\mathcal{G}_n^+ \neq \{\emptyset\}$,

$$\inf_{k \in \mathcal{G}_n^+} \frac{\bar{\delta}_k(n) + \sqrt{n}\bar{B}_k(P)}{||\varphi_k(P)||_{L^2(P)}} \geq \inf_{k \in \mathcal{G}_n^+} \frac{\bar{\delta}_k(n)}{||\varphi_k(P)||_{L^2(P)}} \geq C_n^{-1} \frac{\bar{\delta}_{h_n^+}(n)}{||\varphi_{h_n^+}(P)||_{L^2(P)}} \geq 0.5C_n^{-1} \left(\frac{\bar{\delta}_{h_n^+}(n) + \sqrt{n}\bar{B}_{h_n^+}(P)}{||\varphi_{h_n^+}(P)||_{L^2(P)}}\right)$$

where the first inequality follows from the fact that $\bar{B}_k(P) \geq 0$; the second one follows from the fact that h_n^+ is minimal over $\mathcal{G}_n^+$ and the fact that $\inf_{k \in \mathcal{G}_n^+} \frac{||\varphi_{h_n^+}(P)||_{L^2(P)}}{||\varphi_k(P)||_{L^2(P)}} = \left(\sup_{k \in \mathcal{G}_n^+} \frac{||\varphi_k(P)||_{L^2(P)}}{||\varphi_{h_n^+}(P)||_{L^2(P)}}\right)^{-1} \geq \left(\sup_{k \in \mathcal{G}_n : k \geq h_n^+} \frac{||\varphi_k(P)||_{L^2(P)}}{||\varphi_{h_n^+}(P)||_{L^2(P)}}\right)^{-1} \geq C_n^{-1}$.

Similarly, if $\mathcal{G}_n^- \neq \{\emptyset\}$,

$$\inf_{k \in \mathcal{G}_n^-} \frac{\bar{\delta}_k(n) + \sqrt{n}\bar{B}_k(P)}{||\varphi_k(P)||_{L^2(P)}} \geq \inf_{k \in \mathcal{G}_n^-} \frac{\sqrt{n}\bar{B}_k(P)}{||\varphi_k(P)||_{L^2(P)}} \geq C_n^{-1} \frac{\sqrt{n}\bar{B}_{h_n^-}(P)}{||\varphi_{h_n^-}(P)||_{L^2(P)}} \geq 0.5C_n^{-1} \left(\frac{\bar{\delta}_{h_n^-}(n) + \sqrt{n}\bar{B}_{h_n^-}(P)}{||\varphi_{h_n^-}(P)||_{L^2(P)}}\right).$$

Where here we use monotonicity of $k \mapsto \bar{B}_k(P)$ and the fact that $\inf_{k \in \mathcal{G}_n^-} \frac{||\varphi_{h_n^-}(P)||_{L^2(P)}}{||\varphi_k(P)||_{L^2(P)}} = \left(\sup_{k \in \mathcal{G}_n^-} \frac{||\varphi_k(P)||_{L^2(P)}}{||\varphi_{h_n^-}(P)||_{L^2(P)}}\right)^{-1} \geq \left(\sup_{k \in \mathcal{G}_n : k \leq h_n^-} \frac{||\varphi_k(P)||_{L^2(P)}}{||\varphi_{h_n^-}(P)||_{L^2(P)}}\right)^{-1} \geq C_n^{-1}$.

Thus,

$$\inf_{k \in \mathcal{G}_n} \{\delta_k(r_n^{-1}) + \bar{B}_k(P)\} \geq 0.5 \min\{T_n^-, T_n^+\},$$

and by Lemma E.5 the desired result follows.

□

Lemma E.7. *Suppose Assumption 5.1 holds. For any $n \in \mathbb{N}$, $\mathbf{P}(\{\boldsymbol{z} \in \mathbb{Z}^\infty : h_n \notin \mathcal{L}_n(\boldsymbol{z})\} \cap A_n) \leq \mathbf{P}(B_n^C)$.*

Proof. For any $n \in \mathbb{N}$,

$$\mathbf{P}(\{\boldsymbol{z} \in \mathbb{Z}^\infty : h_n \notin \mathcal{L}_n(\boldsymbol{z})\} \cap A_n) \leq \mathbf{P}(\{\boldsymbol{z} \in \mathbb{Z}^\infty : h_n \notin \mathcal{L}_n(\boldsymbol{z})\} \cap A_n \cap B_n) + \mathbf{P}(B_n^C).$$

By definition of $\mathcal{L}_n$,

$$\{\boldsymbol{z} \in \mathbb{Z}^\infty : h_n \notin \mathcal{L}_n(\boldsymbol{z})\} \subseteq C_n \equiv \left\{\boldsymbol{z} \in \mathbb{Z}^\infty : \exists k \in \mathcal{G}_n : k > h_n \textit{ and } |\psi_k(P_n(\boldsymbol{z})) - \psi_{h_n}(P_n(\boldsymbol{z}))| > 4\bar{\delta}_k(n)/\sqrt{n}\right\},$$

where $(n,k) \mapsto \bar{\delta}_k(n) \equiv \bar{\delta}_{1,k}(n) + \bar{\delta}_{2,k}(n)$.

For any $k \in \mathcal{G}_n$ such that $k \geq h_n$ and any $\boldsymbol{z} \in C_n \cap B_n \cap A_n$ (to ease the notational burden we omit $\boldsymbol{z}$ from the expressions below)

$$\begin{aligned}
|\psi_k(P_n) - \psi_{h_n}(P_n)| \leq & |\psi_k(P_n) - \psi(P) - D\psi_k(P)[P_n - P]| + |\psi_{h_n}(P_n) - \psi(P) - D\psi_{h_n}(P)[P_n - P]| \\
& + |D\psi_k(P)[P_n - P] - D\psi_{h_n}(P)[P_n - P]| \\
& + |\psi_k(P) - \psi(P)| + |\psi_{h_n}(P) - \psi(P)| \\
\leq & |\eta_k(P_n - P)| + |\eta_{h_n}(P_n - P)| + n^{-1/2}\bar{\delta}_{2,k}(n) + \bar{B}_k(P) + \bar{B}_{h_n}(P) \\
\leq & n^{-1/2}\bar{\delta}_{1,k}(n) + n^{-1/2}\bar{\delta}_{1,h_n}(n) + n^{-1/2}\bar{\delta}_{2,k}(n) + \bar{B}_k(P) + \bar{B}_{h_n}(P)
\end{aligned}$$

where the second inequality follows from the definition of η, the fact that $\boldsymbol{z} \in B_n$ and the fact that $k > h_n$; the third inequality follows from the fact that $\boldsymbol{z} \in A_n$. Thus,

$$\begin{aligned}
& \{h_n \notin \mathcal{L}_n\} \cap A_n \cap B_n \\
\subseteq & \left\{\exists k \in \mathcal{G}_n : k > h_n \textit{ and } \frac{\bar{\delta}_{1,k}(n)}{\sqrt{n}} + \frac{\bar{\delta}_{1,h_n}(n)}{\sqrt{n}} + \frac{\bar{\delta}_{2,k}(n)}{\sqrt{n}} + \bar{B}_k(P) + \bar{B}_{h_n}(P) > 4\frac{\bar{\delta}_k(n)}{\sqrt{n}}\right\}. \quad (24)
\end{aligned}$$

We now derive a series of useful claims.

Claim 1: If there exists $k \in \mathcal{G}_n$ such that $k > h_n$ and $h_n = h_n^-$, then $k \in \mathcal{G}_n^+$. **Proof:** If $h_n = h_n^-$, then h_n is the largest element of $\mathcal{G}_n^-$ and thus $k \notin \mathcal{G}_n^-$, which means that $k \in \mathcal{G}_n^+$. □

A corollary of this claim is that if there exists $k \in \mathcal{G}_n$ such that $k > h_n$ and $h_n = h_n^-$, then $\mathcal{G}_n^+$ is non-empty. From this claim, we derive the following two claims.

Claim 2: If there exists a $k > h_n$, then $\bar{\delta}_{1,h_n}(n) + \sqrt{n}\bar{B}_{h_n}(P) \leq 2\bar{\delta}_{h_n^+}(n)$. **Proof:** If $h_n = h_n^+$, then $\bar{\delta}_{h_n}(n) + \sqrt{n}\bar{B}_{h_n}(P) \leq \bar{\delta}_{h_n^+}(n) + \sqrt{n}\bar{B}_{h_n^+}(P) \leq 2\bar{\delta}_{h_n^+}(n)$. If $h_n = h_n^-$, by the previous claim it follows that $\mathcal{G}_n^+$ is non-empty and thus h_n^+ is well-defined, thus $\bar{\delta}_{h_n}(n) + \sqrt{n}\bar{B}_{h_n}(P) \leq \bar{\delta}_{h_n^+}(n) + \sqrt{n}\bar{B}_{h_n^+}(P) \leq 2\bar{\delta}_{h_n^+}(n)$. □

Claim 3: For any $k > h_n$, $\bar{\delta}_{1,k}(n) \geq \bar{B}_k(P)$. **Proof:** If $h_n = h_n^+$ then the claim follows because $k \mapsto \bar{\delta}_k(n) - \bar{B}_k(P)$ is non-decreasing under Assumption 5.1(i). If $h_n = h_n^-$, then $k \in \mathcal{G}_n^+$ by Claim 1 and thus $\bar{\delta}_k(n) \geq \bar{B}_k(P)$. □

By Claims 2 and 3, it follows that if there exists $k \in \mathcal{G}_n$ such that $k \geq h_n$, then $n^{-1/2}\bar{\delta}_k(n) + \bar{B}_k(P) + n^{-1/2}\bar{\delta}_{h_n}(n) + \bar{B}_{h_n}(P) \leq 2n^{-1/2}\bar{\delta}_k(n) + 2n^{-1/2}\bar{\delta}_{h_n}(n) \leq 4n^{-1/2}\bar{\delta}_k(n)$ where the last inequality follows from the fact that $k \mapsto \bar{\delta}_k(n)$ is non-decreasing by Assumption 5.1(i) and the fact that $k > h_n^+$ because $k > h_n$ and so by Claim 1 $k \in \mathcal{G}_n^+$ and h_n^+ is minimal in this set.

Hence applying this result to expression 24 and since $\bar{\delta}_k(n) = \bar{\delta}_{1,k}(n) + \bar{\delta}_{2,k}(n)$, it follows that

$$\{h_n \notin \mathcal{L}_n\} \cap A_n \cap B_n \subseteq \left\{\exists k \in \mathcal{G}_n \colon k \geq h_n \textit{ and } 4n^{-1/2}\bar{\delta}_k(n) > 4n^{-1/2}\bar{\delta}_k(n)\right\}, \tag{25}$$

which is empty. Hence, $\mathbf{P}(\{h_n \notin \mathcal{L}_n\} \cap A_n) \leq \mathbf{P}(B_n^C)$ as desired. □

Proof of Proposition 5.4. We verify that $(h_n)_{n\in\mathbb{N}}$ satisfies both conditions in Lemma E.4. By Lemma E.6 condition 2 in the Lemma E.4 holds with $L = 2C_n$. To check condition 1 in the Lemma E.4, observe that

$$\begin{aligned}\mathbf{P}\left(\mathbb{Z}^\infty \setminus \{\{\boldsymbol{z} \in \mathbb{Z}^\infty \colon h_n \in \mathcal{L}_n(\boldsymbol{z})\} \cap A_n \cap B_n\}\right) \leq & \mathbf{P}\left(\{\boldsymbol{z} \in \mathbb{Z}^\infty \colon h_n \notin \mathcal{L}_n(\boldsymbol{z})\}\right) + \mathbf{P}\left(A_n^C\right) + \mathbf{P}\left(B_n^C\right) \\ \leq & \mathbf{P}\left(\{\boldsymbol{z} \in \mathbb{Z}^\infty \colon h_n \notin \mathcal{L}_n(\boldsymbol{z})\} \cap A_n\right) + 2\mathbf{P}\left(A_n^C\right) + \mathbf{P}\left(B_n^C\right)\end{aligned}$$

Thus, by Lemma E.7

$$\mathbf{P}\left(\mathbb{Z}^\infty \setminus \{\{\boldsymbol{z} \in \mathbb{Z}^\infty \colon h_n \in \mathcal{L}_n(\boldsymbol{z})\} \cap A_n\}\right) \leq 2\mathbf{P}\left(A_n^C\right) + 2\mathbf{P}(B_n^C).$$

Under assumption 5.1(i) the first term in the RHS vanishes. Regarding the second term, by the union bound and the Markov inequality

$$\begin{aligned}\mathbf{P}(B_n^C) \leq & |\mathcal{G}_n|^2 \sup_{k' \geq k \textit{ in } \mathcal{G}_n} \bar{\delta}_{2,k'}^{-2}(n) E_{\mathbf{P}}\left[n\left(D\psi_{k'}(P)[P_n - P] - D\psi_k(P)[P_n - P]\right)^2\right] \\ = & |\mathcal{G}_n|^2 \sup_{k' \geq k \textit{ in } \mathcal{G}_n} \bar{\delta}_{2,k'}^{-2}(n) E_{\mathbf{P}}\left[\left(\varphi_{k'}(P)(Z) - \varphi_k(P)(Z)\right)^2\right] \\ = & o(1)\end{aligned}$$

where the last line follows from Assumption 5.1(ii).

□

E.3.2 Appendix for Example 5.4

Next, we provide an explicit characterization of $\eta_k(P_n - P)$.

Lemma E.8. *For any $P \in \mathcal{M}$ and any $k \in \mathbb{N}$,*

$$\eta_k(P_n - P) = O_P\left(\frac{\kappa_k(0)}{n} + \frac{2||\kappa||_{L^2}\sqrt{k}\sqrt{||p||_{L^\infty}}}{n} + \frac{||p||_{L^\infty}}{n} + \frac{2\sqrt{k}||\kappa||_{L^2}\sqrt{||p||_{L^\infty}}}{n^2}\right).$$

Proof. The proof is relegated to the end of this section. □

Proof of Lemma 5.1. Observe that $z \mapsto \varphi_{1/h}(P)(z) \equiv (\kappa_{1/h} \star P)(z) - E_P[(\kappa_{1/h} \star P)(Z)]$. So for any h and h',

$$\begin{aligned}
&E_P\left[\left(\varphi_{1/h}(P)(Z) - \varphi_{1/h'}(P)(Z)\right)^2\right] \\
=&E_P\left[\left(\left((\kappa_{1/h} \star P)(Z) - E_P[(\kappa_{1/h} \star P)(Z)] - \{(\kappa_{1/h'} \star P)(Z) - E_P[(\kappa_{1/h'} \star P)(Z)]\}\right)\right)^2\right] \\
=&E_P\left[\left((\kappa_{1/h} \star P)(Z) - (\kappa_{1/h'} \star P)(Z) - \{E_P[(\kappa_{1/h} \star P)(Z)] - E_P[(\kappa_{1/h'} \star P)(Z)]\}\right)^2\right] \\
\leq&4E_P\left[\left(\int \kappa(u)\{p(Z+hu) - p(Z+h'u)\}du\right)^2\right].
\end{aligned}$$

By expression (2), it follows that $|p(Z+hu) - p(Z+h'u)| \leq C(z)(|h|^{\varrho} + |h'|^{\varrho})|u|^{\varrho}$. Thus, for any z,

$$\begin{aligned}
&\int \kappa(u)\{p(z+hu) - p(z+h'u)\}du \\
\leq&\int \kappa(u)p'(z)(h+h')udu + C(z)(|h|^{m+\varrho} + |h'|^{\varrho})\int |\kappa(u)||u|^{\varrho}du.
\end{aligned}$$

By symmetry of κ, $\int \kappa(u)udu = 0$.

Therefore,

$$E_P\left[\left(\varphi_{1/h}(P)(Z) - \varphi_{1/h'}(P)(Z)\right)^2\right] \leq 4E_P\left[(C(Z))^2\right]\left(((h)^{\varrho} + (h')^{\varrho})\int |\kappa(u)||u|^{\varrho}du\right)^2.$$

By the proof of the Lemma E.8, for any $h > 0$ and any $M > 0$, there exists a N such that

$$\mathbf{P}\left(\frac{|\eta_{1/h}(P_n - P)|}{M\left(\frac{\kappa(0)}{hn} + \frac{1}{n\sqrt{h}}\right)} \geq 1\right) \leq M^{-1}$$

for all $n \geq N$. By the union bound

$$\mathbf{P}\left(\sup_{k \in \mathcal{G}_n} \frac{|\eta_k(P_n - P)|}{\frac{M}{n}\left(\kappa(0)k + \sqrt{k}\right)} \geq 1\right) \leq \sum_{k \in \mathcal{G}_n} \mathbf{P}\left(\frac{|\eta_k(P_n - P)|}{\frac{M}{n}\left(\kappa(0)k + \sqrt{k}\right)} \geq 1\right) \leq |\mathcal{G}_n| M^{-1}.$$

□

Proof of Lemma E.8. Consider the curve $t \mapsto P + tQ$. It is a valid curve because $\mathbb{D}_{\psi} = ca(\mathbb{R})$.

Therefore

$$\psi_k(P+tQ)-\psi_k(P) = t\left\{\int(\kappa_k\star Q)(x)P(dx)+\int(\kappa_k\star P)(x)Q(dx)\right\} + t^2\int(\kappa_k\star Q)(x)Q(dx).$$

Since κ is symmetric, $\int(\kappa_k\star P)(x)Q(dx)=\int(\kappa_k\star Q)(x)P(dx)$. From this display, $\eta_k(tQ)=t^2\int(\kappa_k\star Q)(x)Q(dx)$ and $D\psi_k(P)[Q]=2\int(\kappa_k\star P)(x)Q(dx)$.

The mapping $Q\mapsto 2\int(\kappa_k\star P)(x)Q(dx)$ is clearly linear. Also, note that $(\kappa_k\star P)(.)=\int\kappa(u)p(\cdot+ku)du$. Hence, for any reals x and x'

$$|(\kappa_k\star P)(x)-(\kappa_k\star P)(x')|=\int\kappa(u)\{p(x+u/k)-p(x'+u/k)\}du.$$

Thus, under the smoothness condition on p in expression (2), it follows that $x\mapsto(\kappa_k\star P)(x)$ is uniformly continuous and bounded, so the mapping $Q\mapsto 2\int(\kappa_k\star P)(x)Q(dx)$ is continuous with respect to the $||.||_{LB}$.

To establish the rate result, we use the Markov inequality. We also introduce the following notation $\int(\kappa_k\star Q)(x)Q(dx)=\langle\kappa_{h_k}\star Q,Q\rangle$ where $\langle .,.\rangle$ is the inner produce of the dual $(L^\infty(\mathbb{R}),ca(\mathbb{R}))$.

It follows that

$$\begin{aligned}\sqrt{E\left[(\eta_k(P_n-P))^2\right]} &= \sqrt{E\left[(\langle\kappa_k\star(P_n-P),P_n-P\rangle)^2\right]}\\ &= \sqrt{E\left[(\langle\kappa_k\star P_n,P_n\rangle-2\langle\kappa_k\star P,P_n\rangle+\langle\kappa_k\star P,P\rangle)^2\right]}\end{aligned}$$

where the second line follows from symmetry of κ which implies $\langle\kappa_k\star P_n,P\rangle=\langle P_n,\kappa_k\star P\rangle$.

We note that

$$\begin{aligned}\langle\kappa_k\star P_n,P_n\rangle &= \frac{\kappa_k(0)}{n}+\frac{1}{n^2}\sum_{i\neq j}\kappa_k(Z_i-Z_j)\\ &= \frac{\kappa_k(0)}{n}+\frac{1}{n^2}\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\kappa_k(Z_i-Z_j)+\sum_{i=2}^{n}\sum_{j=1}^{i-1}\kappa_k(Z_i-Z_j)\right),\end{aligned}$$

also

$$\begin{aligned}
\langle \kappa_k \star P, P_n \rangle =& \frac{1}{n}\sum_{i=1}^{n} (\kappa_k \star P)(Z_i) = \frac{1}{n}\sum_{i=1}^{n} E_P[\kappa_k(Z_i - Z)] \\
=& \frac{1}{n}\sum_{i=1}^{n} E_P[\kappa_k(Z_i - Z)]\frac{i}{n} + \frac{1}{n}\sum_{i=1}^{n} E_P[\kappa_k(Z_i - Z)]\frac{n-i}{n} \\
=& \frac{1}{n}\sum_{i=1}^{n} E_P[\kappa_k(Z_i - Z)]\frac{i}{n} + \frac{1}{n}\sum_{i=1}^{n-1} E_P[\kappa_k(Z_i - Z)]\frac{n-i}{n} \\
=& \frac{1}{n^2}\sum_{i=2}^{n}\sum_{j=1}^{i-1} E_P[\kappa_k(Z_i - Z_j)] + \frac{1}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E_P[\kappa_k(Z_i - Z_j)] \\
&+ \frac{1}{n^2} E_P[\kappa_k(Z_1 - Z)].
\end{aligned}$$

where the third line follows because $E_P[\kappa_k(Z_n - Z)]\frac{n-n}{n} = 0$, and the fourth one follows from the fact that by iid-ness, $E_P[\kappa_k(Z_i - Z_j)] = E_P[\kappa_k(Z_i - Z)]$ for all j.

Therefore,

$$\begin{aligned}
\langle \kappa_k \star P_n, P_n \rangle - 2\langle \kappa_k \star P, P_n \rangle =& \frac{\kappa_k(0)}{n} + \frac{1}{n^2}\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \kappa_k(Z_i - Z_j) + \sum_{i=2}^{n}\sum_{j=1}^{i-1} \kappa_k(Z_i - Z_j)\right) \\
&- \frac{2}{n^2}\sum_{i=2}^{n}\sum_{j=1}^{i-1} E_P[\kappa_k(Z_i - Z_j)] + \frac{1}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E_P[\kappa_k(Z_i - Z_j)] \\
&- \frac{2}{n^2} E_P[\kappa_k(Z_1 - Z)] \\
=& \frac{1}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \{\kappa_k(Z_i - Z_j) - 2E_P[\kappa_k(Z_i - Z_j)]\} \\
&+ \frac{1}{n^2}\sum_{i=2}^{n}\sum_{j=1}^{i-1} \{\kappa_k(Z_i - Z_j) - 2E_P[\kappa_k(Z_i - Z_j)]\} \\
&+ \frac{\kappa_k(0)}{n} - \frac{2}{n^2} E_P[\kappa_k(Z_1 - Z)] \\
=& \frac{2}{n^2}\sum_{i<j} \{\kappa_k(Z_i - Z_j) - 2E_P[\kappa_k(Z_i - Z_j)]\} \\
&+ \frac{\kappa_k(0)}{n} - \frac{2}{n^2} E_P[\kappa_k(Z_1 - Z)],
\end{aligned}$$

where the last line follows by symmetry of κ since $\kappa(Z_i - Z_j) = \kappa(Z_j - Z_i)$ for all i, j.

Since $\langle \kappa_k \star P, P\rangle = E_{P.P}[\kappa_k(Z-Z')] = \frac{1}{n^2}\sum_{i,j} E_{P.P}[\kappa_k(Z-Z')]$, it follows that

$$\begin{aligned}&\langle \kappa_k \star P_n, P_n\rangle - 2\langle \kappa_k \star P, P_n\rangle + \langle \kappa_k \star P, P\rangle \\ =&\frac{2}{n^2}\sum_{i<j}\bar{\kappa}_k(Z_i - Z_j) + \frac{\kappa_k(0)}{n} - \frac{2}{n^2}E_P[\kappa_k(Z_1 - Z)] + \frac{1}{n}E[\kappa_k(Z-Z')]\end{aligned}$$

where $(z, z,') \mapsto \bar{\kappa}_h(z-z') \equiv \kappa_h(z-z') - E_P[\kappa_h(z-Z)] - E_P[\kappa_h(z'-Z)] + E_{P.P}[\kappa_h(Z-Z')]$.

Therefore,

$$\begin{aligned}\sqrt{E\left[(\eta_k(P_n,P))^2\right]} \leq& 2\sqrt{E\left[\left(\frac{1}{n^2}\sum_{i<j}\bar{\kappa}_k(Z_i-Z_j)\right)^2\right]} + \frac{\kappa_k(0)}{n} + \frac{2}{n^2}\sqrt{E\left[(E[\kappa_k(Z_1-Z)])^2\right]} \\ &+ \frac{1}{n}E[\kappa_k(Z-Z')].\end{aligned}$$

We now bound each term on the RHS. First note that

$$\begin{aligned}\frac{1}{n}E[\kappa_h(Z-Z')] =& \frac{1}{n}\int k\kappa(k(z-z'))p(z)p(z')dzdz' \\ =& \frac{1}{n}\int \kappa(u)p(z'+u/k)p(z')dz'du \leq n^{-1}||p||_{L^\infty},\end{aligned}$$

and

$$\begin{aligned}\sqrt{E\left[(E[\kappa_k(Z_1-Z)])^2\right]} \leq& \sqrt{E\left[(\kappa_k(Z'-Z))^2\right]} \\ =& \sqrt{\int (k\kappa(k(z'-z)))^2 p(z)p(z')dzdz'} \\ =& \sqrt{k\int (\kappa(u))^2 p(z+u/k)p(u)dzdu} \leq k^{1/2}\sqrt{||p||_{L^\infty}}||\kappa||_{L^2}.\end{aligned}$$

where the first line follows by Jensen inequality. Finally, by Gine and Nickl [2008] Sec. 2

$$\sqrt{E\left[\left(\frac{1}{n^2}\sum_{i<j}\bar{\kappa}_k(Z_i-Z_j)\right)^2\right]} \leq \frac{2}{\sqrt{n^2}}\sqrt{E[(\bar{\kappa}_k(Z-Z'))^2]} \leq \frac{2||\kappa||_{L^2}\sqrt{||p||_{L^2}}}{n\sqrt{1/k}}.$$

□

F Appendix for Section 5.6

F.1 Appendix for Example 5.6

Let $\mathcal{T}_0 \equiv \cup_{k,n\in\mathbb{N}^2}\left\{\zeta \in ca(\mathbb{Z})\colon E_\zeta\left[\left|\frac{d\phi(Z,\theta)}{d\theta}[v]/||v||_{L^q}\right|^2\right] < \infty \text{ ,} \forall(v,\theta)\in\Theta_k\times\Theta_k(\epsilon_{k,n})\right\}$, which is a linear subspace in $ca(\mathbb{Z})$. The set $\mathcal{T}_0$ will act as the tangent space, Assumption 5.3(ii) — $\mathcal{S}_{1,k}(\epsilon_{k,n}) \subseteq \mathcal{S}$ for all k,n — ensures that elements of the form $a(P_n - P)$ for $a \geq 0$, belong to $\mathcal{T}_0$; i.e., $\mathcal{T}_P \subseteq \mathcal{T}_0$.

Proof of Proposition 5.5. Under our assumptions, $Q_k(\cdot, P)$ is strictly convex and smooth over Θ_k, so the following FOC holds:

$$\frac{dQ_k(\psi_k(P),P)}{d\theta}[\pi^T\kappa^k] = E_P\left[\frac{d\phi(Z,\psi_k(P))}{d\theta}[\pi^T\kappa^k]\right] + \lambda_k\frac{dPen(\psi_k(P))}{d\theta}[\pi^T\kappa^k] = 0$$

for any $\pi\in\mathbb{R}^k$. Let $H_k : \Theta_k(\epsilon_{k,n})\times\mathcal{T}_0 \subseteq \mathbb{R}^k\times ca(\mathbb{Z}) \to \mathbb{R}^k$ where $H_k(\pi,P) \equiv E_P\left[\frac{d\phi(Z,\pi^T\kappa^k)}{d\theta}[\kappa^k]\right] + \lambda_k\frac{dPen(\pi^T\kappa^k)}{d\theta}[\kappa^k]$. Letting $\pi_k(P)\in\mathbb{R}^k$ be the vector that $\psi_k(P) = \pi_k(P)^T\kappa^k$, it follows that $H_k(\pi_k(P),P) = 0$. Also, observe that

$$\frac{dH_k(\pi,P)}{dP}[Q] = E_Q\left[\frac{d\phi(Z,\pi^T\kappa^k)}{d\theta}[\kappa^k]\right] \in \mathbb{R}^k,$$

for any $Q\in\mathcal{T}_0$ and

$$\frac{dH_k(\pi,P)}{d\pi}[a] = \left(E_P\left[\frac{d^2\phi(Z,\pi^T\kappa^k)}{d\theta^2}[\kappa^k,\kappa^k]\right] + \lambda_k\frac{d^2Pen(\pi^T\kappa^k)}{d\theta^2}[\kappa^k,\kappa^k]\right)a$$

for any $a\in\mathbb{R}^k$; where $\Delta_k(P) \equiv E_P\left[\frac{d^2\phi(Z,\pi^T\kappa^k)}{d\theta^2}[\kappa^k,\kappa^k]\right] + \lambda_k\frac{d^2Pen(\pi^T\kappa^k)}{d\theta^2}[\kappa^k,\kappa^k] \in \mathbb{R}^{k\times k}$ and since Pen is strictly convex and ϕ convex under our assumptions, the matrix is positive-definite for all k.

We now show that the partial derivatives are $||.||+||.||_{\mathcal{S}}$-continuous and by Theorem 1.1.9 in Ambrosetti and Prodi [1995] this implies that the mapping H_k is Frechet differentiable at $\psi_k(P)$ under $||.||+||.||_{\mathcal{S}}$. To do this, consider a sequence (π_n, P_n) that converges to $(\pi \equiv \pi_k(P), P)$ under $||.||+||.||_{\mathcal{S}}$, moreover $\pi_m^T\kappa^k \in \Theta_k(\epsilon_{k,n})$ for all m. Then, letting $||A||_* \equiv \sup_{Q\in ca(\mathbb{Z})}\frac{||A[Q]||}{||Q||_{\mathcal{S}}}$,

$$\left\|\frac{dH_k(\pi_n,P_n)}{dP}[Q] - \frac{dH_k(\pi,P)}{dP}[Q]\right\|_* = \left\|E_Q\left[\frac{d\phi(Z,\pi_n^T\kappa^k)}{d\theta}[\kappa^k] - \frac{d\phi(Z,\pi^T\kappa^k)}{d\theta}[\kappa^k]\right]\right\|_*.$$

Observe that, for each $z \in \mathbb{Z}$,

$$\sup_{a \in \mathbb{R}^k} \frac{\left| \frac{d\phi(z,\pi_n^T \kappa^k)}{d\theta}[a\kappa^k] - \frac{d\phi(z,\pi^T \kappa^k)}{d\theta}[a\kappa^k] \right|}{||a||} \leq ||(\pi_n^T - \pi^T)\kappa^k||_{L^q} \sup_{a \in \mathbb{R}^k} \frac{||a^T \kappa^k||_{L^q}}{||a||} \times \sup_{h,v_1,v_2 \in \Theta_k(\epsilon_{k,n}) \times \Theta_k^2} \frac{\left| \frac{d^2\phi(z,h)}{d\theta^2}[v_1, v_2] \right|}{||v_1||_{L^q} ||v_2||_{L^q}}$$

where the first inequality follows from the assumption of twice differentiability and the Mean Value Theorem (Theorem 1.8 in Ambrosetti and Prodi [1995]). By Assumption 5.3, $\Phi_{2,k}(z) \equiv \sup_{h,v_1,v_2 \in \Theta_k(\epsilon_{k,n}) \times \Theta_k^2} \frac{\left| \frac{d^2\phi(z,h)}{d\theta^2}[v_1,v_2] \right|}{||v_1||_{L^q} ||v_2||_{L^q}}$ and $\Phi_{2,k} \in \mathcal{S}$. Hence, given that $||x|| = \sup_{a \in \mathbb{R}^k} |ax|/||a||$,

$$\left\| \frac{dH_k(\pi_n, P_n)}{dP}[Q] - \frac{dH_k(\pi, P)}{dP}[Q] \right\|_* \leq ||(\pi_n^T - \pi^T)\kappa^k||_{L^q} \sup_{a \in \mathbb{R}^k} \frac{||a^T \kappa^k||_{L^q}}{||a||} \frac{|E_Q[\Phi_{2,k}(Z)]|}{||Q||_{\mathcal{S}}}.$$

Since $\frac{|E_Q[\Phi_{2,k}(Z)]|}{||Q||_{\mathcal{S}}} \leq 1$ and $\sup_{a \in \mathbb{R}^k} \frac{||a^T \kappa^k||_{L^q}}{||a||} < \infty$, this display implies that $(a, P) \mapsto \frac{dH_k(a,P)}{dP}$ is continuous.

Also,

$$\begin{aligned}\left\| \frac{dH_k(\pi_n, P_n)}{d\pi}[a] - \frac{dH_k(\pi, P)}{d\pi}[a] \right\| \leq & \left\| \left(E_{P_n}\left[\frac{d^2\phi(Z, \pi_n^T \kappa^k)}{d\theta^2}[\kappa^k, \kappa^k] \right] - E_P\left[\frac{d^2\phi(Z, \pi^T \kappa^k)}{d\theta^2}[\kappa^k, \kappa^k] \right] \right) a \right\| \\ & + \lambda_k \left\| \left(\frac{d^2 Pen(\pi_n^T \kappa^k)}{d\theta^2}[\kappa^k, \kappa^k] - \frac{d^2 Pen(\pi^T \kappa^k)}{d\theta^2}[\kappa^k, \kappa^k] \right) a \right\| \\ \equiv & Term_{1,n} + Term_{2,n}.\end{aligned}$$

The first term in the RHS is bounded by two terms: $Term_{1,1,n} \equiv \left\| E_{P_n - P}\left[\frac{d^2\phi(Z,\pi^T\kappa^k)}{d\theta^2}[\kappa^k, \kappa^k] \right] \right\| ||a||$ and $Term_{1,2,n} \equiv \sup_{\{b:\ ||b||=1\}} \left| \left(E_{P_n}\left[\frac{d^2\phi(Z,\pi_n^T\kappa^k)}{d\theta^2}[b^T\kappa^k, a^T\kappa^k] - \frac{d^2\phi(Z,\pi^T\kappa^k)}{d\theta^2}[a^T\kappa^k, b^T\kappa^k] \right] \right) \right|$. By assumption 5.3, for each j, l, $z \mapsto \frac{d^2\phi(z,\psi_k(P))}{d\theta^2}[\kappa_j, \kappa_l] \in \mathcal{S}$.

Hence, each component of $E_{P_n - P}\left[\frac{d^2\phi(Z,\pi^T\kappa^k)}{d\theta^2}[\kappa^k, \kappa^k] \right]$ vanishes as $||P_n - P||_{\mathcal{S}}$ does, to $Term_{1,1,n}$ does too. By Assumption 5.3, $h \mapsto \frac{d^2\phi(z,h)}{d\theta^2}$ is continuous at $\psi_k(P)$ uniformly on $z \in \mathbb{Z}$, so $Term_{1,2,n}$ vanishes as π_n converges to π.

By Assumption 5.2, $a \mapsto \frac{d^2 Pen(a\kappa^k)}{d\theta^2}$ is continuous and thus $Term_{2,n}$ vanishes as π_n converges to π. Therefore, $(a, P) \mapsto \frac{dH_k(a,P)}{d\pi}$ is continuous.

By continuity of both derivatives of H_k, H_k is Frechet differentiable (Theorem 1.1.9 in Ambrosetti and Prodi [1995]). Hence, this and the fact that $\Delta_k(P)$ is non-singular imply by the implicit function theorem (see Theorem 2.3 in Ambrosetti and Prodi [1995]) that, in a neighborhood of P, π_k is Frechet differentiable under $||.||_{\mathcal{S}}$ with the derivative given by

$$Q \mapsto \left(\Delta_k(P)^{-1} \frac{dH_k(\pi, P)}{dP}[Q] \right)^T.$$

Since $\psi_k(P) = \pi_k(P)^T \kappa^k$, this implies the result. □

Proof of Lemma 5.2. Since μ is a finite measure, it suffices to show the result for $q = \infty$. Since $\Theta \subseteq L^\infty$, evaluation functionals are linear and bounded and thus $\Xi \subseteq \Theta^*$.

Note that for any $\ell \in \Xi$ denoted by $\ell = \delta_z$, $t \mapsto \ell[\varphi_k(P)](t) = (\kappa^k(z))^T \Delta_k(P)^{-1} \nabla_k(P)(t)$ and $||\ell[\varphi_k(P)]||^2_{L^2(P)} = \sigma_k^2(P)(z)$. Also,

$$\sup_{\ell \in \Xi} \sqrt{n} \frac{\eta_{k,\ell}(\zeta)}{||\ell[\varphi_k(P)]||_{L^2(P)}} = \sup_{z \in \mathbb{Z}} \sqrt{n} \frac{\eta_{k,\delta_z}(\zeta)}{\sigma_k(P)(z)}, \tag{26}$$

for any $\zeta \in \mathcal{T}_0$.

By proposition F.1 and the fact that $\sigma_k(P)(z) \geq \underline{e}||\kappa^k(z)||$ where $\underline{e}_k(P) \equiv e_{min}\left(\Delta_k^{-1}(P)\Sigma_k(P)\Delta_k^{-1}(P)\right)$ (where e_{min} is the minimal Eigenvalue), it follows that for any $B_0 < \infty$ and any $k \in \mathbb{N}$,

$$\sup_{z \in \mathbb{Z}} \sqrt{n} \frac{\eta_{k,\delta_z}(\zeta)}{\sigma_k(P)(z)} \leq \underline{e}_k(P)^{-1} \max\{||\zeta||^3_{\mathcal{S}}, ||\zeta||^{1+\varrho}_{\mathcal{S}}\} C(B_0, P, k),$$

which is bounded for any ζ such that $||\zeta||_{\mathcal{S}}$. Thus, in Theorem 5.2 we can take $\mathcal{T}_0 \supseteq \mathcal{T}_P$ and $\mathcal{C}$ as the class of $||.||_{\mathcal{S}}$-bounded sets. It only remains to show the condition in the theorem. To do this, observe that $\mathcal{S}$ is P-Donsker. Hence $\mathbb{G}_n = \sqrt{n}(P_n - P)$ is $||.||_{\mathcal{S}}$-bounded a.s.-$\mathbf{P}$, and thus the condition in the theorem holds.

Therefore, by Theorem 5.2,

$$\sup_{z \in \mathbb{Z}} \left| \sqrt{n} \frac{\psi_k(P_n)(z) - \psi_k(P)(z)}{\sigma_k(P)(z)} - (\kappa^k(z))^T \Delta_k(P)^{-1} n^{-1/2} \sum_{i=1}^n \frac{\nabla_k(P)(Z_i)}{\sigma_k(P)(z)} \right| = o_P(1),$$

as desired.

□

The following Lemma is used in the proof of Proposition 5.6.

Lemma F.1. *For any $(n,k) \in \mathbb{N}^2$ and any $q \in [1, \infty]$,*

$$\left\| \frac{(\kappa^k)^T \Delta_k(P)^{-1}}{\sigma_k(P)} \left(n^{-1/2} \sum_{i=1}^n \nabla_k(P)(Z_i) - \Delta_k(P)\mathcal{Z}_k \right) \right\|_{L^q} = O_{Pr}\left(\frac{\beta_k k}{\sqrt{n}} \left(1 + \frac{|\log\left(\frac{\sqrt{n}}{\beta_k k}\right)|}{k} \right) \right)$$

where $\beta_k \equiv E_P[||\Sigma_k(P)^{-1/2} \nabla_k(P)(Z)||^3]$, and the Pr is the product measure between $\mathbf{P}$ and standard Gaussian, and the "O" does not depend on n or k.

Proof. By Lemma 5.2 and the fact that μ is a finite measure, it is sufficient to show that

$$\left\| \frac{(\kappa^k)^T \Delta_k(P)^{-1}}{\sigma_k(P)} \left(n^{-1/2} \sum_{i=1}^n \nabla_k(P)(Z_i) - \Delta_k(P)\mathcal{Z}_k \right) \right\|_{L^\infty} = O_{Pr}\left(\frac{\beta_k k}{\sqrt{n}} \left(1 + \frac{|\log\left(\frac{\sqrt{n}}{\beta_k k}\right)|}{k} \right) \right),$$

where Pr is the product measure between $\mathbf{P}$ and the measure of $\mathcal{Z}_k$. By letting $\mathcal{T}_k \sim N(0, I_k)$ such that $\Delta_k(P)\mathcal{Z}_k = \Sigma_k(P)^{1/2}\mathcal{T}_k$, the LHS equals

$$\sup_{z\in\mathbb{Z}} \left| (\kappa^k(z))^T \frac{\Delta_k(P)^{-1}\Sigma_k(P)^{1/2}}{\sigma_k(P)(z)} \left(n^{-1/2} \sum_{i=1}^n \Psi_k(P)(Z_i) - \mathcal{T}_k \right) \right|,$$

where $\Psi_k(P) \equiv \Sigma_k(P)^{-1/2}\nabla_k(P)$. By Cauchy-Swarchz inequality,

$$\begin{aligned} &\left\| \frac{(\kappa^k)^T \Delta_k(P)^{-1}}{\sigma_k(P)} \left(n^{-1/2} \sum_{i=1}^n \nabla_k(P)(Z_i) - \Delta_k(P)\mathcal{Z}_k \right) \right\|_{L^\infty} \\ \leq &\sup_{z\in\mathbb{Z}} \left\| (\kappa^k(z))^T \frac{\Delta_k(P)^{-1}\Sigma_k(P)^{1/2}}{\sigma_k(P)(z)} \right\| \left\| n^{-1/2} \sum_{i=1}^n \Psi_k(P)(Z_i) - \mathcal{T}_k \right\| \\ \leq &\left\| n^{-1/2} \sum_{i=1}^n \Psi_k(P)(Z_i) - \mathcal{T}_k \right\| \end{aligned}$$

where the last line follows from the fact that

$$\left\| (\kappa^k(z))^T \frac{\Delta_k(P)^{-1}\Sigma_k(P)^{1/2}}{\sigma_k(P)(z)} \right\| = \sqrt{\frac{(\kappa^k(z))^T \Delta_k(P)^{-1}\Sigma_k(P)\Delta_k(P)^{-1}(\kappa^k(z))}{\sigma_k^2(P)(z)}} \leq 1.$$

By Pollard [2002] Thm. 10, for any $\delta > 0$,

$$\Pr\left(\left\| n^{-1/2} \sum_{i=1}^n \Psi_k(P)(Z_i) - \mathcal{T}_k \right\| \geq 3\delta \right) \leq C_0 \frac{\beta k}{n^{1/2}} \delta^{-3} \left(1 + \frac{|\log(n^{1/2}\delta^3/(\beta k))|}{k} \right)$$

where $\beta \equiv E_P[||\Psi_k(P)(Z)||^3]$ and C_0 is some universal constant. □

Proof of Proposition 5.6. By the triangle inequality

$$\begin{aligned} \left\| \sqrt{n} \frac{\psi_k(P_n) - \psi_k(P)}{\sigma_k(P)} - \frac{(\kappa^k)^T \mathcal{Z}_k}{\sigma_k(P)} \right\|_{L^q} \leq &\left\| \frac{(\kappa^k)^T \Delta_k(P)^{-1}}{\sigma_k(P)} \left(n^{-1/2} \sum_{i=1}^n \nabla_k(P)(Z_i) - \Delta_k(P)\mathcal{Z}_k \right) \right\|_{L^q} \\ &+ \left\| \sqrt{n} \frac{\psi_k(P_n) - \psi_k(P)}{\sigma_k(P)} - \frac{(\kappa^k)^T \Delta_k(P)^{-1}}{\sigma_k(P)} n^{-1/2} \sum_{i=1}^n \nabla_k(P)(Z_i) \right\|_{L^q}. \end{aligned}$$

Lemma F.1 bounds the first term in the RHS. The second term in the RHS is bounded by

$\sup_{z\in\mathbb{Z}}\sqrt{n}\frac{\eta_{k,\delta_z}(P_n-P)}{\sigma_k(P)(z)}$, so it remains to derive the rate of convergence of this term. That is, we want to find a $(r_{n,k})_{n,k}$ such that any $\delta>0$ there exists a C_δ and a N_δ such that

$$\mathbf{P}\left(\sup_{z\in\mathbb{Z}} r_{n,k}\sqrt{n}\frac{\eta_{k,\delta_z}(P_n-P)}{\sigma_k(P)(z)}\geq C_\delta\right)\leq\delta,$$

for all $n\geq N_\delta$.

By the proof of Lemma 5.2, for any $\epsilon>0$, there exists a $M_\epsilon\geq 1$ and a N_ϵ such that for all $n\geq N_\epsilon$ and all $k\in\mathbb{N}$,

$$\begin{aligned}
&\mathbf{P}\left(\sup_{z\in\mathbb{Z}} r_{n,k}\sqrt{n}\frac{\eta_{k,\delta_z}(P_n-P)}{\sigma_k(P)(z)}\geq \epsilon\right)\\
\leq&\mathbf{P}\left(\sup_{z\in\mathbb{Z}} r_{n,k}\sqrt{n}\frac{\eta_{k,\delta_z}(P_n-P)}{\sigma_k(P)(z)}\geq \epsilon\cap A_\epsilon\right)+\epsilon\\
\leq&\mathbf{P}\left(r_{n,k}n^{-1/2}\sup_{z\in\mathbb{Z}}\frac{\max\{||\mathbb{G}_n||^3_{\mathcal{S}},||\mathbb{G}_n||^{1+\varrho}_{\mathcal{S}}\}||\kappa^k(z)||\times C(M_\epsilon,P,k)}{\sigma_k(P)(z)}\geq\epsilon\cap A_\epsilon\right)+\epsilon\\
\leq&1\left\{r_{n,k}M_\epsilon^3\times n^{-1/2}\underline{e}_k(P)^{-1}C(M_\epsilon,P,k)\geq\epsilon\right\}+\epsilon,
\end{aligned}$$

for all $n\geq N_\epsilon$.

Therefore, setting $\delta=\epsilon$, $C_\delta=2M_\delta^3$ and $r_{n,k}^{-1}\equiv n^{-1/2}\underline{e}_k(P)^{-1}C(l_n,P,k)$, it follows that

$$\begin{aligned}
&\mathbf{P}\left(\sup_{z\in\mathbb{Z}} r_{n,k}\sqrt{n}\frac{\eta_{k,\delta_z}(P_n-P)}{\sigma_k(P)(z)}\geq C_\delta\right)\\
\leq&\mathbf{P}\left(\sup_{z\in\mathbb{Z}} r_{n,k}\sqrt{n}\frac{\eta_{k,\delta_z}(P_n-P)}{\sigma_k(P)(z)}\geq C_\delta\cap A_\delta\right)+\delta\\
\leq&\mathbf{P}\left(r_{n,k}n^{-1/2}\sup_{z\in\mathbb{Z}}\frac{\max\{||\mathbb{G}_n||^3_{\mathcal{S}},||\mathbb{G}_n||^{1+\varrho}_{\mathcal{S}}\}||\kappa^k(z)||\times C(M_\delta,P,k)}{\sigma_k(P)(z)}\geq C_\delta\cap A_\delta\right)+\delta\\
\leq&1\left\{r_{n,k}M_\delta^3\times n^{-1/2}\underline{e}_k(P)^{-1}C(M_\delta,P,k)\geq M_\delta\right\}+\delta\\
\leq&1\left\{M_\delta^3\geq C_\delta\right\}+\delta=\delta
\end{aligned}$$

where the last line follows because $C(M,P,k)/C(l_n,P,k)<1$ for any constant M and n sufficiently large. □

F.1.1 Explicit bound for $\eta_{k,\ell}$

Frechet differentiability of ψ_k under $||.||_{\mathcal{S}}$ implies a bound for $Q\mapsto\eta_{k,\ell}(Q)$ that is uniform over $||.||_{\mathcal{S}}$-bounded sets. Unfortunately, this general result is silent about how this bound depends on (k,ℓ). By using the Mean Value Theorem on the first derivative, the following proposition presents an explicit bound for $\eta_{k,\ell}$.

Proposition F.1. For any $P \in \mathcal{M}$, any $B_0 < \infty$ and any $k \in \mathbb{N}$:[28]

$$\eta_{k,\ell}(tQ) \leq t||Q||_{\mathcal{S}} \sup_{t\in[0,1]} \|\ell[D\psi_k(P+tQ) - D\psi_k(P)]\|_* ,$$

for any $\ell \in \Theta^*$, any $Q \in \mathcal{Q} \equiv \{\zeta \in ca(\mathbb{Z})\colon ||\zeta||_{\mathcal{S}} \leq B_0\}$ and any $t \geq 0$. Moreover, there exists a $T \in (0,1]$ such that[29]

$$\|\ell[D\psi_k(P+tQ) - D\psi_k(P)]\|_* \leq t \max\{||Q||^2_{\mathcal{S}}, ||Q||^{\varrho}_{\mathcal{S}}\}||\ell[\kappa^k]|| \times \mathbb{C}(B_0, P, k)$$

for any $\ell \in \Theta^*$, any $Q \in \mathcal{Q}$ and any $t \leq T$, where $\mathbb{C}(B_0, P, k)$ is given in expression 27 in Appendix F.1.

Proof of Proposition F.1. By the calculations in p. 14 in Ambrosetti and Prodi [1995], it follows that

$$||\psi_k(P+tQ) - \psi_k(P) - tD\psi_k(P)[Q]||_\Theta \leq t||Q||_{\mathcal{S}} \sup_{t\in[0,1]} \|D\psi_k(P+tQ) - D\psi_k(P)\|_* ,$$

and, for any $\ell \in \Theta^*$,

$$||\ell[\psi_k(P+tQ) - \psi_k(P) - tD\psi_k(P)[Q]]||_\Theta \leq t||Q||_{\mathcal{S}} \sup_{t\in[0,1]} \|\ell[D\psi_k(P+tQ) - D\psi_k(P)]\|_* .$$

Thus, we can take $\eta_{k,\ell}(P+tQ, P) \leq t||Q||_{\mathcal{S}} \sup_{t\in[0,1]} \|\ell[D\psi_k(P+tQ) - D\psi_k(P)]\|_*$. Observe that

$$\begin{aligned}
\ell[D\psi_k(P+tQ)[Q] - D\psi_k(P)[Q]] =& \left(E_Q[\nabla_k(P+tQ)(Z)]\right)^T \left(\Delta_k(P+tQ)^{-1} - \Delta_k(P)^{-1}\right) \ell[\kappa^k] \\
&+ \left(E_Q[\nabla_k(P+tQ)(Z) - \nabla_k(P)(Z)]\right)^T \left(\Delta_k(P)^{-1}\right) \ell[\kappa^k] \\
=& \left(E_Q[\nabla_k(P+tQ)(Z) - \nabla_k(P)(Z)]\right)^T \left(\Delta_k(P+tQ)^{-1} - \Delta_k(P)^{-1}\right) \\
&\times \ell[\kappa^k] \\
&+ \left(E_Q[\nabla_k(P)(Z)]\right)^T \left(\Delta_k(P+tQ)^{-1} - \Delta_k(P)^{-1}\right) \ell[\kappa^k] \\
&+ \left(E_Q[\nabla_k(P+tQ)(Z) - \nabla_k(P)(Z)]\right)^T \Delta_k(P)^{-1} \ell[\kappa^k] \\
\equiv& Term_{1,k,n} + Term_{2,k,n} + Term_{3,k,n}.
\end{aligned}$$

[28]Recall that the norm $||.||_*$ is the operator norm associated, in this case, to $||.||_{L^q}$.

[29]T may depend on B_0, P and k, but it does not depend on ℓ or Q.

In Step 1 below we show that

$$\begin{aligned}
||Term_{1,k,n}|| \leq & t^{1+\varrho} \max\{||Q||_{\mathcal{S}}^{2+\varrho}, ||Q||_{\mathcal{S}}^{3}\} \sup_{t\in[0,1]} ||D\psi_k(P+tQ)||_* \sup_{a\in\mathbb{R}^k} \frac{||a^T\kappa^k||_{L^q}}{||a||} \\
& \times \left((\lambda_k C_0 + C_0) \sup_{t\in[0,1]} ||D\psi_k(P+tQ)||_*^{\varrho} + 1 \right) \sup_{(a,b)\in\mathbb{R}^{2k}} \frac{||a^T\kappa^k||_{L^q}||b^T\kappa^k||_{L^q}}{||a||||b||} \\
& \times ||\ell[\kappa^k]|| \\
||Term_{2,k,n}|| \leq & t^{\varrho} \max\{||Q||_{\mathcal{S}}^{1+\varrho}, ||Q||_{\mathcal{S}}^{2}\}||\Delta_k(P)^{-1}|| \\
& \times \left((\lambda_k C_0 + C_0) \sup_{t\in[0,1]} ||D\psi_k(P+tQ)||_*^{\varrho} + 1 \right) \sup_{(a,b)\in\mathbb{R}^{2k}} \frac{||a^T\kappa^k||_{L^q}||b^T\kappa^k||_{L^q}}{||a||||b||} \\
& ||\Delta_k(P+tQ)^{-1}|| \times ||\ell[\kappa^k]|| \\
||Term_{3,k,n}|| \leq & t||Q||_{\mathcal{S}}^{2} \sup_{t\in[0,1]} \|D\psi_k(P+tQ)\|_* \sup_{a\in\mathbb{R}^k} \frac{||a^T\kappa^k||_{L^q}}{||a||} \left\|\Delta_k(P)^{-1}\right\| \times ||\ell[\kappa^k]||.
\end{aligned}$$

Observe that $||a^T\kappa^k||_{L^q} = \left(\int |a^T\kappa^k(z)|^q\mu(dz)\right)^{1/q} \leq ||a|| \left(\int ||\kappa^k(z)||^q\mu(dz)\right)^{1/q} = ||a|| \times ||||\kappa^k||||_{L^q}$. Thus

$$\begin{aligned}
||Term_{1,k,n}|| \leq & t^{1+\varrho} \max\{||Q||_{\mathcal{S}}^{2+\varrho}, ||Q||_{\mathcal{S}}^{3}\} \sup_{t\in[0,1]} ||D\psi_k(P+tQ)||_* \times ||||\kappa^k||||_{L^q}^3 \\
& \times \left((\lambda_k C_0 + C_0) \sup_{t\in[0,1]} ||D\psi_k(P+tQ)||_*^{\varrho} + 1 \right) \\
& \times ||\ell[\kappa^k]|| \\
||Term_{2,k,n}|| \leq & t^{\varrho} \max\{||Q||_{\mathcal{S}}^{1+\varrho}, ||Q||_{\mathcal{S}}^{2}\}||\Delta_k(P)^{-1}|| \\
& \times \left((\lambda_k C_0 + C_0) \sup_{t\in[0,1]} ||D\psi_k(P+tQ)||_*^{\varrho} + 1 \right) ||||\kappa^k||||_{L^q}^2 \\
& ||\Delta_k(P+tQ)^{-1}|| \times ||\ell[\kappa^k]|| \\
||Term_{3,k,n}|| \leq & t||Q||_{\mathcal{S}}^{2} \sup_{t\in[0,1]} \|D\psi_k(P+tQ)\|_* ||||\kappa^k||||_{L^q} \left\|\Delta_k(P)^{-1}\right\| \times ||\ell[\kappa^k]||.
\end{aligned}$$

By step 2, $||\Delta_k(P+tQ)^{-1}|| \leq 2||\Delta_k(P)^{-1}||$ for all t less or equal than some $T_{B_0,P,k}$ (specified in Step 2). Therefore

$$\begin{aligned}
||Term_{2,k,n}|| \leq & 2t^{\varrho} \max\{||Q||_{\mathcal{S}}^{1+\varrho}, ||Q||_{\mathcal{S}}^{2}\}||\Delta_k(P)^{-1}||^2 \\
& \times \left((\lambda_k C_0 + C_0) \sup_{t\in[0,1]} ||D\psi_k(P+tQ)||_*^{\varrho} + 1 \right) ||||\kappa^k||||_{L^q}^2 \\
& \times ||\ell[\kappa^k]||.
\end{aligned}$$

So, after some simple algebra and the fact that $\lambda_k \leq 1$,

$$\begin{aligned}&||\ell[D\psi_k(P+tQ)[Q] - D\psi_k(P)[Q]]||\\ \leq& 3t\max\{||Q||^3_{\mathcal{S}}, ||Q||^{1+\varrho}_{\mathcal{S}}\}||\ell[\kappa^k]||\\ &\times \max\{1, ||||\kappa^k||||^3_{L^q}\} \times \left(\max\left\{1, \sup_{t\in[0,1],Q\in\mathcal{Q}} ||D\psi_k(P+tQ)||_*, 2||\Delta_k^{-1}(P)||\right\}\right)^2\\ &\times \left(2C_0 \sup_{t\in[0,1],Q\in\mathcal{Q}} ||D\psi_k(P+tQ)||^{\varrho}_* + 1\right).\end{aligned}$$

By letting

$$\begin{aligned}\mathbb{C}(B_0,k,P) \equiv& 3\max\{1, ||||\kappa^k||||^3_{L^q}\} \times \left(\max\left\{1, \sup_{t\in[0,1],Q\in\mathcal{Q}} ||D\psi_k(P+tQ)||_*, 2||\Delta_k^{-1}(P)||\right\}\right)^2\\ &\times \left(2C_0 \sup_{t\in[0,1],Q\in\mathcal{Q}} ||D\psi_k(P+tQ)||^{\varrho}_* + 1\right) \qquad (27)\end{aligned}$$

the result follows.

STEP 1. By the calculations in the proof of Proposition 5.5 and the Mean Value Theorem,

$$\begin{aligned}\|E_Q[\nabla_k(P+tQ) - \nabla_k(P)]\| \leq& ||\psi_k(P+tQ) - \psi_k(P)||_{L^q} \sup_{a\in\mathbb{R}^k} \frac{||a^T\kappa^k||_{L^q}}{||a||}||Q||_{\mathcal{S}}\\ \leq& t||Q||^2_{\mathcal{S}} \sup_{t\in[0,1]} \|D\psi_k(P+tQ)\|_* \sup_{a\in\mathbb{R}^k} \frac{||a^T\kappa^k||_{L^q}}{||a||}.\end{aligned}$$

Hence

$$||Term_{3,k,n}|| \leq t||Q||^2_{\mathcal{S}} \sup_{t\in[0,1]} \|D\psi_k(P+tQ)\|_* \sup_{a\in\mathbb{R}^k} \frac{||a^T\kappa^k||_{L^q}}{||a||} \left\|\Delta_k(P)^{-1}\right\| \times ||\ell[\kappa^k]||.$$

Regarding $Term_{2,k,n}$, it follows that

$$\begin{aligned}||Term_{2,k,n}|| =& ||E_Q[\nabla_k(P)(Z)]\Delta_k(P)^{-1}(\Delta_k(P) - \Delta_k(P+tQ))\Delta_k(P+tQ)^{-1}\ell[\kappa^k]||\\ \leq& ||E_Q[\nabla_k(P)(Z)]\Delta_k(P)^{-1}|| \times ||\Delta_k(P) - \Delta_k(P+tQ)|| \times ||\Delta_k(P+tQ)^{-1}\ell[\kappa^k]||.\end{aligned}$$

By Assumption 5.3(ii), $\frac{d\phi(.,\psi_k(P))}{d\theta}[a^T\kappa/||a^T\kappa||_{L^q}] \in \mathcal{S}$ for any $a \in \mathbb{R}^k$, so

$$||E_Q[\nabla_k(P)(Z)]\Delta_k(P)^{-1}|| \leq ||E_Q[\nabla_k(P)(Z)]|| \times ||\Delta_k(P)^{-1}|| \leq ||Q||_{\mathcal{S}} \sup_{a\in\mathbb{R}^k} \frac{||a^T\kappa||_{L^q}}{||a||}||\Delta_k(P)^{-1}||.$$

Also

$$||\Delta_k(P) - \Delta_k(P+tQ)|| \leq \lambda_k \left\| \frac{d^2 Pen(\psi_k(P))}{d\theta^2}[\kappa^k, \kappa^k] - \frac{d^2 Pen(\psi_k(P+tQ))}{d\theta^2}[\kappa^k, \kappa^k] \right\| + \left\| E_P\left[\frac{d^2\phi(Z,\psi_k(P))}{d\theta^2}[\kappa^k, \kappa^k] \right] - E_{P+tQ}\left[\frac{d^2\phi(Z,\psi_k(P+tQ))}{d\theta^2}[\kappa^k, \kappa^k] \right] \right\| \equiv Term_{4,k,n} + Term_{5,k,n},$$

and in turn

$$Term_{5,k,n} \leq \left\| E_P\left[\frac{d^2\phi(Z,\psi_k(P))}{d\theta^2}[\kappa^k, \kappa^k] - \frac{d^2\phi(Z,\psi_k(P+tQ))}{d\theta^2}[\kappa^k, \kappa^k] \right] \right\| + t\left\| E_Q\left[\frac{d^2\phi(Z,\psi_k(P+tQ))}{d\theta^2}[\kappa^k, \kappa^k] \right] \right\| \equiv Term_{6,k,n} + Term_{7,k,n}.$$

Observe that

$$Term_{4,k,n} = \lambda_k \sup_{(a,b)\in\mathbb{R}^{2k}} \frac{\left| \frac{d^2 Pen(\psi_k(P))}{d\theta^2}[a^T\kappa^k, b^T\kappa^k] - \frac{d^2 Pen(\psi_k(P+tQ))}{d\theta^2}[a^T\kappa^k, b^T\kappa^k] \right|}{||a|| \times ||b||}.$$

By assumption 5.2,

$$\left| \frac{d^2 Pen(\psi_k(P))}{d\theta^2}[a^T\kappa^k, b^T\kappa^k] - \frac{d^2 Pen(\psi_k(P+tQ))}{d\theta^2}[a^T\kappa^k, b^T\kappa^k] \right| \leq C_0 ||\psi_k(P) - \psi_k(P+tQ)||^{\varrho}_{L^q} ||a^T\kappa^k||_{L^q} ||b^T\kappa^k||_{L^q},$$

and thus

$$Term_{4,k,n} \leq \lambda_k ||\psi_k(P) - \psi_k(P+tQ)||^{\varrho}_{\Theta} \sup_{(a,b)\in\mathbb{R}^{2k}} \frac{||a^T\kappa^k||_{L^q}||b^T\kappa^k||_{L^q}}{||a|| \times ||b||} \leq \lambda_k C_0 ||tQ||^{\varrho}_{\mathcal{S}} \sup_{t\in[0,1]} ||D\psi_k(P+tQ)||^{\varrho}_{*} \sup_{(a,b)\in\mathbb{R}^{2k}} \frac{||a^T\kappa^k||_{L^q}||b^T\kappa^k||_{L^q}}{||a|| \times ||b||}.$$

By Assumption 5.3,

$$\left| \frac{d^2\phi(z,\psi_k(P))}{d\theta^2}[a^T\kappa^k, b^T\kappa^k] - \frac{d^2\phi(z,\psi_k(P+tQ))}{d\theta^2}[a^T\kappa^k, b^T\kappa^k] \right| \leq C_0 ||\psi_k(P) - \psi_k(P+tQ)||^{\varrho}_{L^q} ||a^T\kappa^k||_{L^q} ||b^T\kappa^k||_{L^q},$$

so

$$Term_{6,k,n} \leq C_0||tQ||_{\mathcal{S}}^{\varrho} \sup_{t\in[0,1]} ||D\psi_k(P+tQ)||_*^{\varrho} \sup_{(a,b)\in\mathbb{R}^{2k}} \frac{||a^T\kappa^k||_{L^q}||b^T\kappa^k||_{L^q}}{||a||\times||b||}.$$

Also, by Assumption 5.3(ii)

$$Term_{7,k,n} \leq t||Q||_{\mathcal{S}} \sup_{t\in[0,1]} \sup_{(a,b)\in\mathbb{R}^{2k}} \frac{||a^T\kappa^k||_{L^q}||b^T\kappa^k||_{L^q}}{||a||\times||b||}.$$

Hence

$$\begin{aligned}||Term_{2,k,n}|| \leq& ||Q||_{\mathcal{S}}||\Delta_k(P)^{-1}|| \\ &\times\left((\lambda_k C_0 + C_0)t^{\varrho}||Q||_{\mathcal{S}}^{\varrho} \sup_{t\in[0,1]} ||D\psi_k(P+tQ)||_*^{\varrho} + t||Q||_{\mathcal{S}}\right) \frac{||a^T\kappa^k||_{L^q}||b^T\kappa^k||_{L^q}}{||a||||b||} \\ &||\Delta_k(P+tQ)^{-1}||||\ell[\kappa^k]||.\end{aligned}$$

Finally, by the previous calculations,

$$\begin{aligned}||Term_{1,k,n}|| \leq& t||Q||_{\mathcal{S}}^2 \sup_{t\in[0,1]} ||D\psi_k(P+tQ)||_* \frac{||a^T\kappa^k||_{L^q}}{||a||} \\ &\times\left((\lambda_k C_0 + C_0)t^{\varrho}||Q||_{\mathcal{S}}^{\varrho} \sup_{t\in[0,1]} ||D\psi_k(P+tQ)||_*^{\varrho} + t||Q||_{\mathcal{S}}\right) \frac{||a^T\kappa^k||_{L^q}||b^T\kappa^k||_{L^q}}{||a||||b||} \\ &\times ||\ell[\kappa^k]||.\end{aligned}$$

Step 2. We now show that there exists a $T_{B_0,P,k}$ such that for all $t \leq T_{B_0,P,k}$, $||\Delta_k(P+tQ)^{-1}|| \leq 2||\Delta_k(P)^{-1}||$. To do this, note that

$$||\Delta_k(P+tQ)^{-1}|| = \frac{1}{e_{min}(\Delta_k(P+tQ))}.$$

If $e_{min}(\Delta_k(P+tQ)) \geq e_{min}(\Delta_k(P))$ then $||\Delta_k(P+tQ)^{-1}|| \leq ||\Delta_k(P)^{-1}||$. If $e_{min}(\Delta_k(P+tQ)) \leq e_{min}(\Delta_k(P))$, then, by Weyl inequality, $e_{min}(\Delta_k(P+tQ)) \geq e_{min}(\Delta_k(P)) - ||\Delta_k(P+$

$tQ) - \Delta_k(P)||$. By our previous calculations for terms 4 and 5,

$$
\begin{aligned}
||\Delta_k(P+tQ) - \Delta_k(P+tQ)|| \leq & \left((\lambda_k C_0 + C_0) t^{\varrho} ||Q||^{\varrho}_{\mathcal{S}} \sup_{t \in [0,1]} ||D\psi_k(P+tQ)||^{\varrho}_{*} + t||Q||_{\mathcal{S}} \right) \\
& \times \sup_{a,b \in \mathbb{R}^{2k}} \frac{||a^T \kappa^k||_{L^q} ||b^T \kappa^k||_{L^q}}{||a|| ||b||} \\
\leq & 2B_0 \left(C_0 t^{\varrho} \sup_{t \in [0,1], Q \in \mathcal{Q}} ||D\psi_k(P+tQ)||^{\varrho}_{*} + t \right) \\
& \times \sup_{a,b \in \mathbb{R}^{2k}} \frac{||a^T \kappa^k||_{L^q} ||b^T \kappa^k||_{L^q}}{||a|| ||b||}
\end{aligned}
$$

because $||Q||_{\mathcal{S}} \leq B_0$ and $\lambda_k \leq 1$. Hence, there exists a $T = T_{B_0,P,k}$ such that for all $t \leq T$ $e_{min}(\Delta_k(P+tQ)) \geq 0.5 e_{min}(\Delta_k(P))$ and thus $||\Delta_k(P+tQ)^{-1}|| \leq 2||\Delta_k(P)^{-1}||$ for any $t \leq T_{B_0,P,k}$. □

Admittedly the bound presented here might be loose, but it does hint at how k, ℓ, t and $||Q||_{\mathcal{S}}$ affect the reminder.

F.1.2 The Role of the Choice of Θ

In the text we assumed that $\Theta \subseteq L^q \cap L^{\infty}$, this assumption ensured that the evaluation functional is well-defined for elements of Θ, formally, it ensure that the evaluation functional belongs to Θ^*. If this is not the case, i.e., $\Theta \subseteq L^q$ but not necessarily $\Theta \subseteq L^{\infty}$, then $z \mapsto \sqrt{n} \frac{\psi_k(P)_n(z) - \psi_k(P)(z)}{\sigma_k(P)(z)}$ may not even be well-defined and the approach developed in the text is not valid. However, we now show that by exploiting results for duals of L^q (see Lax [2002]), Theorem 5.2 still can be used as the basis of inferential results for L^q-confidence bands; what changes is the scaling.

For all $q < \infty$, let $\Xi_q = \{f \in L^{q^*} : ||f||_{q^*} \leq 1\}$ with $q^* = \frac{q}{q-1}$ and $\Xi_{\infty} = \{\delta_z : z \in \mathbb{Z}\}$. We note that by defining Ξ_q in this way, we are abusing notation, because the mapping ℓ is a linear bounded functional operating over L^q, the fact that we view this mapping as an element of Ξ_q is due to dual representation results; see Lax [2002].

Also, let

$$
\bar{\sigma}^2_{k,q}(P) = \sup_{\ell \in \Xi_q} (\ell[\kappa^k])^T \Delta_k^{-1}(P) \Sigma_k(P) \Delta_k^{-1}(P) (\ell[\kappa^k]).
$$

Observe that for $q = \infty$, $\bar{\sigma}^2_{k,\infty}(P) = \sup_z \sigma^2_k(P)(z)$. But this relationship only holds in $q = \infty$, for $q < \infty$, $\sigma_k(P)(z)$ may not be even well-defined.

For these choices, it follows that

Lemma F.2. *Suppose Assumptions 5.2 and 5.3 hold and $\mathcal{S}$ is P-Donsker. Then, for any*

$q \in [1,\infty]$ *and any* $k \in \mathbb{N}$,

$$\left\| \sqrt{n}\frac{\psi_k(P_n)-\psi_k(P)}{\bar{\sigma}_{k,q}(P)} - (\kappa^k)^T \Delta_k(P)^{-1} n^{-1/2} \sum_{i=1}^{n} \frac{\nabla_k(P)(Z_i)}{\bar{\sigma}_{k,q}(P)} \right\|_{L^q} = o_P(1).$$

Proof. To ease the notational burden, we use $||.||_q$ to denote $||.||_{L^q}$. Also, recall that $\Xi_q = \{f \in L^{q^*} \colon ||f||_{q^*} \leq 1\}$ and $\Xi_\infty = \{\delta_z \colon z \in \mathbb{Z}\}$.

Let, for any $q \in [1,\infty)$, $(f,\eta) \mapsto \langle f,\eta \rangle_q \equiv \int f(z')\eta(z)\mu(dz)$ and $(f,\eta) \mapsto \langle f,\eta\rangle_\infty \equiv \int f(z)\eta(dz)$. By duality $||.||_q = \sup_{\ell \in \Xi_q} |\langle ., \ell \rangle_q|$ (see Lax [2002]), and

$$||f||_\infty = \sup_{z\in\mathbb{Z}} |f(z)| = \sup_{z\in\mathbb{Z}} |\int f(z')\delta_z(dz')| = \sup_{\ell\in\Xi_\infty} |\int f(z')\ell(dz')|,$$

so that $||.||_q = \sup_{\ell\in\Xi_q} |\langle ., \ell\rangle_q|$.

By duality and straightforward algebra,

$$\begin{aligned}
&\left\| \sqrt{n}\frac{\psi_k(P_n)-\psi_k(P)}{\bar{\sigma}_{k,q}(P)} - (\kappa^k)^T \Delta_k(P)^{-1} n^{-1/2} \sum_{i=1}^{n} \frac{\nabla_k(P)(Z_i)}{\bar{\sigma}_{k,q}(P)} \right\|_q \\
=& \sup_{\ell\in\Xi_q} \left| \frac{\ell[\sqrt{n}(\psi_k(P_n)-\psi_k(P)) - (\kappa^k)^T\Delta_k(P)^{-1}n^{-1/2}\sum_{i=1}^n \nabla_k(P)(Z_i)]}{\bar{\sigma}_{k,q}(P)} \right| \\
\leq& \sup_{\ell\in\Xi_q} \left| \frac{\ell[\sqrt{n}(\psi_k(P_n)-\psi_k(P))] - (\ell[\kappa^k])^T\Delta_k(P)^{-1}n^{-1/2}\sum_{i=1}^n \nabla_k(P)(Z_i)}{\sqrt{(\ell[\kappa^k])^T\Delta_k^{-1}(P)\Sigma_k(P)\Delta_k^{-1}(P)(\ell[\kappa^k])}} \right| \\
&\times \sup_{\ell\in\Xi_q} \left| \frac{\sqrt{(\ell[\kappa^k])^T\Delta_k^{-1}(P)\Sigma_k(P)\Delta_k^{-1}(P)(\ell[\kappa^k])}}{\bar{\sigma}_{k,q}(P)} \right|.
\end{aligned}$$

The second term equals one by definition of $\bar{\sigma}_{k,q}(P)$. The first term can be bounded by analogous arguments to those employed on the proof of Lemma 5.2 and they will be omitted. □

Lemma 5.2 shows that in order to characterize the asymptotic distribution of $\left\|\sqrt{n}\frac{\psi_k(P_n)-\psi_k(P)}{\bar{\sigma}_{k,q}(P)}\right\|_{L^q}$ it suffices to characterize the one of $\left\|n^{-1/2}\sum_{i=1}^n \frac{\nabla_k(P)(Z_i)}{\bar{\sigma}_{k,q}(P)}\right\|_{L^q}$. In the following proposition we accomplish by employing coupling results (e.g. Pollard [2002]).

Proposition F.2. Suppose Assumptions 5.2 and 5.3 hold and $\mathcal{S}$ is P-Donsker. Then, for any $q \in [1,\infty]$ and any $k \in \mathbb{N}$,

$$\left\| \sqrt{n}\frac{\psi_k(P_n)-\psi_k(P)}{\bar{\sigma}_{k,q}(P)} - \frac{(\kappa^k)^T\mathcal{Z}_k}{\bar{\sigma}_{k,q}(P)} \right\|_{L^q} = O_{Pr}\left(\frac{\beta_k k}{\sqrt{n}} \left(1 + \frac{|\log\left(\frac{\sqrt{n}}{\beta_k k}\right)|}{k} \right) + r_{k,n}^{-1} \right),$$

where $\mathcal{Z}_k \sim N(0, \Delta_k(P)^{-1}\Sigma_k(P)\Delta_k(P)^{-1})$ and $\beta_k \equiv E[||\Delta_k(P)^{-1}\nabla_k(P)(Z)||^3]$.

Proof. By noting that

$$\left\| \sqrt{n}\frac{\psi_k(P_n) - \psi_k(P)}{\bar{\sigma}_{k,q}(P)} - \frac{(\kappa^k)^T \mathcal{Z}_k}{\bar{\sigma}_{k,q}(P)} \right\|_{L^q} \leq \left\| \sqrt{n}\frac{\psi_k(P_n) - \psi_k(P) - (\kappa^k)^T \mathcal{Z}_k}{\sqrt{(\ell[\kappa^k])^T \Delta_k^{-1}(P)\Sigma_k(P)\Delta_k^{-1}(P)(\ell[\kappa^k])}} \right\|_{L^q},$$

the proof is analogous to that of Proposition 5.6 and thus omitted. □