Scalable Quasi-Bayesian Inference for Instrumental Variable Regression

Ziyu Wang[1,*], **Yuhao Zhou**[1,*], **Tongzheng Ren**[2], **Jun Zhu**[1,‡]
[1] Dept. of Comp. Sci. and Tech., BNRist Center, State Key Lab for Intell. Tech. & Sys.,
Institute for AI, Tsinghua-Bosch Joint Center for ML, Tsinghua University
[2] Department of Computer Science, UT Austin
{wzy196,yuhaoz.cs}@gmail.com, tongzheng@utexas.edu, dcszj@tsinghua.edu.cn

Abstract

Recent years have witnessed an upsurge of interest in employing flexible machine learning models for instrumental variable (IV) regression, but the development of uncertainty quantification methodology is still lacking. In this work we present a scalable quasi-Bayesian procedure for IV regression, building upon the recently developed kernelized IV models. Contrary to Bayesian modeling for IV, our approach does not require additional assumptions on the data generating process, and leads to a scalable approximate inference algorithm with time cost comparable to the corresponding point estimation methods. Our algorithm can be further extended to work with neural network models. We analyze the theoretical properties of the proposed quasi-posterior, and demonstrate through empirical evaluation the competitive performance of our method.

1 Introduction

Instrumental variable (IV) regression is a standard approach for estimating causal effect from confounded observational data. In the presence of confounding, any regression method estimating $\mathbb{E}(\mathbf{y} \mid \mathbf{x})$ cannot recover the causal relation $f^\dagger$ between the outcome $\mathbf{y}$ and the treatment $\mathbf{x}$, since the residual $\mathbf{u} = \mathbf{y} - f^\dagger(\mathbf{x})$ is correlated with $\mathbf{x}$ due to the unobserved confounders. IV regression enables identification of the causal effect through the introduction of *instruments*, variables $\mathbf{z}$ that are known to influence $\mathbf{y}$ only through $\mathbf{x}$.

IV regression is widely used in economics [1], epidemiology [2] and clinical research [3], but modeling nonlinear effect in IV regression can be challenging. Recent years have seen great development in adopting modern machine learning models for IV regression [4–8]. However, there is still a lack of uncertainty quantification measures for these flexible IV models. Uncertainty quantification is especially important for IV analysis, since unlike in standard supervised learning scenarios, we do not have (unconfounded) validation data, from which we could deduce the error pattern of the estimated model. Moreover, the instrument of choice may be weak, meaning that it only provides limited information for $\mathbf{x}$; in such cases point estimators suffer from high variance [9]. This problem is exacerbated in the nonparametric setting, where IV estimation is usually an ill-posed inverse problem, where instruments provide vanishing information for the higher-order nonlinear effects [10].

The IV setting brings unique challenges for uncertainty quantification. For example, while it is natural to consider a Bayesian approach, specification of the likelihood requires knowledge of the entire data generating process, which is typically unavailable in IV regression. Consequently, most, if not all, existing work on Bayesian IV [11–15] assumes the following data generating process:

[*] ZW and YZ contribute equally. [‡] JZ is the corresponding author. An extended version is available at https://arxiv.org/abs/2106.08750v2.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

$\mathbf{x} = g(\mathbf{z}) + \mathbf{u}_1,\ \mathbf{y} = f(\mathbf{x}) + \mathbf{u}_2$, where $(\mathbf{u}_1, \mathbf{u}_2)$ are correlated and independent of $\mathbf{z}$. These methods then conduct posterior inference on g, f as well as the *generative model* for the unobserved confounders $(\mathbf{u}_1, \mathbf{u}_2)$. However, the additive error in $\mathbf{x}$ is an unnecessary assumption for most point estimation procedures, and is difficult to check in high dimensions. Furthermore, the need to model the generating process of $(\mathbf{u}_1, \mathbf{u}_2)$ introduces an extra risk of model misspecification, and Bayesian inference on the generative model is computationally expensive, especially on complex high-dimensional datasets. None of these issues present if only point estimation is needed.

For the above reasons, it is appealing to turn to an alternative *quasi-Bayesian* approach [16–18]. Quasi-Bayesian analysis views IV estimation as a generalized method-of-moments (GMM) procedure, and defines the quasi-posterior as a Gibbs distribution constructed from a chosen prior and violation of the moment constraints. It does not require full knowledge of the data generating process, and thus does not suffer from the aforementioned drawbacks of Bayesian inference. However, computation of the quasi-posterior is non-trivial, as its density contains a conditional expectation term $\mathbb{E}(\mathbf{y} - f(\mathbf{x}) \mid \mathbf{z})$, which itself needs to be estimated from data. This makes both approximate inference and theoretical analysis difficult. So far, quasi-Bayesian analysis for nonparametric IV is only developed on classical models such as wavelet basis [17, 18] or Nadaraya-Watson smoothing [19], while adoption of flexible machine learning models such as kernel machines or neural networks remains an open challenge. Moreover, numerical study has been limited in previous work, so little is known about the empirical performance of the quasi-Bayesian approach.

In this work, we present a novel quasi-Bayesian procedure for IV regression, building upon the recent development in kernelized IV models [7, 20]. We employ a Gaussian process (GP) prior and construct a quasi-likelihood using a kernel conditional expectation estimator. We establish theoretical properties of the resultant quasi-posterior, proving its consistency and showing that it may quantify instrument strength. Furthermore, inspired from the minimax formulation of IV estimation [5, 8, 20, 21], we design a principled approximate inference algorithm using random feature expansion and a novel adaptation of the "randomized prior trick" [22]. The algorithm has the form of stochastic gradient descent-ascent and is thus scalable and easy to implement. It can also be adapted to work with flexible neural network models, in which case its behavior can be formally justified by analyzing the neural networks in the kernel regime. Empirical evaluation shows that the proposed method produces informative uncertainty estimates, scales to high-dimensional and complex nonlinear problems, and is especially advantageous when the instrument is weak.

The rest of the paper is organized as follows: In Section 2 we set up the problem. We then derive the quasi-posterior and analyze its theoretical properties in Section 3, and present the approximate inference algorithm in Section 4. Section 5 reviews related work, and Section 6 presents numerical studies. Finally, we discuss conclusion and future work in Section 7.

2 Notations and Setup

Notations We use boldface $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ to represent random variables on the space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, regular font (x, y, z) to denote deterministic values. $[n]$ denotes the set $\{1, 2, \cdots, n\}$. $\{(x_i, y_i, z_i) : i \in [n]\}$ indicates the training data. We use the notations $X := (x_1, \ldots, x_n) \in \mathcal{X}^n$, $f(X) := (f(x_1), \ldots, f(x_n))$; likewise for Y, Z. For finite-dimensional vectors $\theta, \theta' \in \mathbb{R}^m$, we use $\|\theta\|_2, \langle \theta, \theta' \rangle_2$ to denote the Euclidean norm and inner product, respectively. For any operator $A : H_1 \to H_2$ between Hilbert spaces H_1 and H_2, we denote its adjoint by $A^* : H_2 \to H_1$. When $H_1 = H_2$ and $\lambda \in \mathbb{R}$, we use the notation $A_\lambda := A + \lambda I$ for simplicity.

IV regression Denote the treatment and response variables as $\mathbf{x}, \mathbf{y}$, the instrument as $\mathbf{z}$, and the true *structural function* of interest as $f^\dagger$. Consider the data generating process $\mathbf{y} = f^\dagger(\mathbf{x}) + \mathbf{u}$, where the unobserved $\mathbf{u}$ satisfies $\mathbb{E}(\mathbf{u} \mid \mathbf{z}) = 0$, but may correlate with $\mathbf{x}$. Then $f^\dagger$ satisfies

$$\mathbb{E}(\mathbf{y} - f^\dagger(\mathbf{x}) \mid \mathbf{z}) = 0 \ \text{ a.s. } [P(dz)], \tag{1}$$

where P denotes the data distribution. This *conditional moment restriction* (CMR) formulation is the standard definition in literature [e.g., 23, 18], and is used in the recent work on machine learning models for IV. It connects to GMM as (1) can be viewed as a continuum of generalized moment constraints. Note that (1) does not place any structural constraint on the conditional distribution $p(\mathbf{x} \mid \mathbf{z})$, such as additive noise; hence, it does not require full knowledge of the data generating

process. Also, as discussed in, e.g., Hartford et al. [4], the setup can also be extended to incorporate observed confounders $\mathbf{v}$, by including $\mathbf{v}$ in both $\mathbf{x}$ and $\mathbf{z}$.

Kernelized IV and a dual view Let $\mathcal{H}, \mathcal{I}$ be suitably chosen function spaces on $\mathcal{X}, \mathcal{Z}$, respectively. The generalized moment constraint (1) motivates the use of the following objective

$$\min_{f \in \mathcal{H}} \quad \mathcal{L}(f) := d_n^2(\hat{E}_n f - \hat{b}) + \bar{\lambda}\Omega(f), \tag{2}$$

where $\hat{E}_n : \mathcal{H} \to \mathcal{I}$ is an empirical approximation to the conditional expectation operator $E : f \mapsto \mathbb{E}(f(\mathbf{x}) \mid \mathbf{z} = \cdot)$, $\hat{b}$ is an estimator of $\mathbb{E}(\mathbf{y} \mid \mathbf{z} = \cdot)$, $\{d_n\}$ is a sequence of suitable (semi-)norm on $\mathcal{I}$, and $\Omega : \mathcal{H} \to \mathbb{R}$ is a regularization term.

When $\mathcal{H}, \mathcal{I}$ are reproducing kernel Hilbert spaces (RKHS) with corresponding kernels k_x, k_z, it is natural to set $\Omega(f) = \frac{1}{2}\|f\|_{\mathcal{H}}^2$, and define $\hat{b}$ as the result of kernel ridge regression on $\mathbf{y}$ with respect to $\mathbf{z}$. In this case $\hat{E}_n$ can be defined with the empirical *kernel conditional expectation* operator: let $C_{zx} = \mathbb{E}(k(\mathbf{x}, \cdot) \otimes k(\mathbf{z}, \cdot))$, $C_{zz} = \mathbb{E}(k(\mathbf{z}, \cdot) \otimes k(\mathbf{z}, \cdot))$. Assuming E maps all $f \in \mathcal{H}$ to $Ef \in \mathcal{I}$, we have [24]

$$C_{zz} E f = C_{zx} f, \ \ \forall f \in \mathcal{H}.$$

This motivates the use of $\hat{E}_n := \hat{C}_{zz,\bar{\nu}}^{-1} \hat{C}_{zx}$ where $\hat{C}_{zz} := \frac{1}{n}\sum_{i=1}^n k(z_i, \cdot) \otimes k(z_i, \cdot)$, $\hat{C}_{zx}$ is defined similarly, and $\bar{\nu}$ is a regularization hyperparameter.[2] The choice of d_n is flexible; the dual IV formulation uses $d_n^2(g) := \frac{1}{2n}\sum_{j=1}^n g(z_j)^2 + \frac{\bar{\nu}}{2}\|g\|_{\mathcal{I}}^2 = \frac{1}{2}\langle g, \hat{C}_{zz,\bar{\nu}} g\rangle_{\mathcal{I}}$. Introducing the evaluation operator $S_z : \mathcal{I} \to \mathbb{R}^n, S_z g := (g(z_1), \ldots, g(z_n))$, we have $\hat{b} = \hat{C}_{zz,\bar{\nu}}^{-1} \frac{S_z^* Y}{n}$, and

$$\mathcal{L}(f) = \frac{1}{2}\left\| \hat{C}_{zz,\bar{\nu}}^{-1/2} \left(\hat{C}_{zx} f - \frac{S_z^* Y}{n} \right) \right\|_{\mathcal{I}}^2 + \frac{\bar{\lambda}}{2}\|f\|_{\mathcal{H}}^2 \tag{3}$$

$$= \max_{g \in \mathcal{I}} \frac{1}{n}\sum_{j=1}^n \left((f(x_j) - y_j) g(z_j) - \frac{g(z_j)^2}{2} \right) - \frac{\bar{\nu}}{2}\|g\|_{\mathcal{I}}^2 + \frac{\bar{\lambda}}{2}\|f\|_{\mathcal{H}}^2, \tag{4}$$

where (4) holds because of the Fenchel duality $\frac{1}{2}u^2 = \sup_{v \in \mathbb{R}} \left(uv - \frac{1}{2}v^2\right)$ and the equality $\mathbb{E}(u(\mathbf{x}, \mathbf{y}) g(\mathbf{z})) = \mathbb{E}(\mathbb{E}(u(\mathbf{x}, \mathbf{y}) \mid \mathbf{z}) g(\mathbf{z}))$; see [25, 8, 21, 26]. The dual formulation (4) circumvents the need to compute $\hat{E}_n$ directly, an operator between the typically infinite-dimensional spaces $\mathcal{H}$ and $\mathcal{I}$, and leads to a scalable estimation procedure based on stochastic gradient descent-ascent (SGDA). It can also be generalized to work with deep neural networks (DNNs) instead of kernel machines, by replacing the RKHS regularizer with a suitable regularizer for DNNs, although theoretical analysis for the resulted algorithm requires separate effort [8, 26]. As we shall see, (4) will also enable the construction of a scalable approximate inference algorithm.

3 Quasi-Bayesian Analysis of Dual IV

Introduce the notations $S_x : \mathcal{H} \to \mathbb{R}^n, S_x f := (f(x_1), \ldots, f(x_n))$, so that $\hat{C}_{zx} = \frac{1}{n} S_z^* S_x, \hat{C}_{zz} = \frac{1}{n} S_z^* S_z$. Define $\lambda := n\bar{\lambda}, \nu = n\bar{\nu}$. Now we can re-express (3) in an equivalent form as:

$$\bar{\mathcal{L}}(f) := \frac{n}{\lambda}\mathcal{L}(f) = \frac{1}{2}(f(X) - Y)^\top (\lambda^{-1} L)(f(X) - Y) + \frac{1}{2}\|f\|_{\mathcal{H}}^2, \tag{5}$$

where $L := \frac{1}{n} S_z \hat{C}_{zz,\bar{\nu}}^{-1} S_z^*$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n$, and thus can be identified with an $n \times n$ matrix. Since the first term above is equivalent to the log density of the multivariate normal distribution $\mathcal{N}(Y \mid f(X), \lambda L^{-1})$, we can view (5) as the objective of a kernel ridge regression problem, which has a data-dependent noise covariance λL^{-1}.[3] The connection between kernel ridge regression and Gaussian process regression [28] thus motivates the use of the *quasi-posterior*

[2] With an abuse of notation, k refers to the reproducing kernel of the corresponding RKHS (k_x for $\mathcal{H}$ or k_z for $\mathcal{I}$) whenever the denotation is clear.

[3] Here we assume the invertibility of L for brevity. Alternatively, observe that (5) defines a linear inverse problem with the finite-dimensional observation operator $f \mapsto \sqrt{L} f(X)$ and noise variance $\lambda^{-1} I$, and we can follow [27, Chapter 6] to derive the same quasi-posterior.

$\Pi(df \mid \mathcal{D}^{(n)})$, defined through the following Radon-Nikodym derivative w.r.t. the standard GP prior $\Pi = \mathcal{GP}(0, k_x)$:

$$\frac{d\Pi(\cdot \mid \mathcal{D}^{(n)})}{d\Pi}(f) \propto \exp\Big(-\frac{1}{2}(f(X)-Y)^\top(\lambda^{-1}L)(f(X)-Y)\Big) = \exp\Big(-\frac{n}{\lambda}d_n^2(\hat{E}_n f - \hat{b})\Big). \quad (6)$$

Note that contrary to standard Bayesian modeling, we <u>*do not*</u> assume $Y \sim \mathcal{N}(f(X), \lambda L^{-1})$ is part of the true data generating process, nor does the theoretical analysis below rely on it. Instead, the quasi-posterior (6) should be interpreted as a Gibbs distribution which trades off between the properly scaled *evidence* $\lambda^{-1}nd_n^2(\hat{E}_n f - \hat{b})$, which characterizes the estimated violation of the GMM constraint (1), and our *prior belief* $\Pi(df)$. This trade-off is most clear from the well-known variational characterization of the Gibbs distribution [29],

$$\Pi(\cdot \mid \mathcal{D}^{(n)}) = \arg\min_{\Psi} \mathbb{E}_{f\sim\Psi}[\lambda^{-1}nd_n^2(\hat{E}_n f - \hat{b})] + \mathrm{KL}(\Psi \,\|\, \Pi). \quad (7)$$

Nonetheless, the fictitious data generating process $f \sim \mathcal{GP}(0, k), Y \sim \mathcal{N}(f(X), \lambda L^{-1})$ is useful for deriving the quasi-posterior, since its conditional distribution $p_{\mathrm{fic}}(df \mid Y)$ coincides with (6) [27, Chapter 6]. In the probability space of this fictitious data generating process, for any finite set of test inputs x_*, we have

$$p_{\mathrm{fic}}(f(x_*), Y) \sim \mathcal{N}\left(0, \begin{bmatrix} K_{**} & K_{*x} \\ K_{x*} & K_{xx} + \lambda L^{-1} \end{bmatrix}\right),$$

where $K_{(\cdot)}$ denote the corresponding Gram matrices with subscript $*$ denoting x_* and x denoting X (so, e.g., $K_{*x} := k(x_*, X)$). Thus by the Gaussian conditioning formula, we have

$$\Pi(f(x_*) \mid \mathcal{D}^{(n)}) = p_{\mathrm{fic}}(f(x_*) \mid Y) = \mathcal{N}(m, S), \quad (8)$$

$$\text{where} \quad m := K_{*x}(\lambda I + LK_{xx})^{-1}LY, \quad (9)$$

$$S := K_{**} - K_{*x}L(\lambda I + K_{xx}L)^{-1}K_{x*}, \quad (10)$$

$$L = K_{zz}(K_{zz} + \nu I)^{-1}. \quad (11)$$

In the above $K_{zz} := k(Z, Z)$ denotes the Gram matrix, and (11) follows from the definition $L = n^{-1}S_z(n^{-1}S_z^*S_z + \bar{\nu}I)^{-1}S_z^*$, the Woodbury identity, and the observation that $S_zS_z^* = K_{zz}$.

Theoretical Analysis For the quasi-posterior $\Pi(\cdot \mid \mathcal{D}^{(n)})$ to be a useful measure of uncertainty, it needs to satisfy the following informal criteria: as $n \to \infty$, we expect

(C1) $\Pi(\cdot \mid \mathcal{D}^{(n)})$ will exclude incorrect solutions;

(C2) In cases of non-identification, $\Pi(\cdot \mid \mathcal{D}^{(n)})$ will not exclude any valid solution in the model.

In the following, we formalize these criteria, and demonstrate that with appropriately chosen hyper-parameters, the quasi-posterior satisfies both criteria. We will work with the following assumptions:

Assumption 3.1. *The restriction of the conditional expectation operator $f \mapsto \mathbb{E}(f(\mathbf{x}) \mid \mathbf{z} = \cdot)$ on $\mathcal{H}$, denoted as E, has its image contained in $\mathcal{I}$. $E : \mathcal{H} \to \mathcal{I}$ is bounded.*

Assumption 3.1 is intuitive, and is also slightly more general than some previous work [20, 26] that require the conditional expectation operator to be bounded on the entire hypothesis space, which typically corresponds to the "sample space" of the GP prior in our setting, and is much larger than $\mathcal{H}$ (see Appendix A). Nonetheless, we note that Hypothesis 4 in Singh et al. [7] may be more general, although they impose extra smoothness assumptions on E.

Assumption 3.2. *The true structural function $f^\dagger(x)$ is such that, there exists a sequence $\{\tau_m : m \in \mathbb{N}\}$ satisfying $\tau_m \to 0$, and*

$$m\tau_m^2 \geq \inf_{f_m^\dagger \in \mathcal{H}: \|f^\dagger - f_m^\dagger\| \leq \tau_m} \|f_m^\dagger\|_{\mathcal{H}}^2 - \log \Pi(\{f : \|f\| \leq \tau_m\}), \quad (12)$$

where $\|\cdot\|$ denotes the sup norm.

Assumption 3.2 requires that $f^\dagger$ can be well approximated in $\mathcal{H}$; this is more general than previous work [e.g., 20, 7] that require $f^\dagger \in \mathcal{H}$, but is typical in the GP literature [30, 31]. The sequence $\{\tau_m\}$ is determined by the complexity of $\mathcal{H}$ and its ability for approximating $f^\dagger$, and is usually the optimal posterior contraction rate for Gaussian process regression on the unconfounded dataset $\{(x_i, f^\dagger(x_i) + \epsilon_i) : i \in [m]\}$ [31].

Assumption 3.3. *$\mathcal{X}$ and $\mathcal{Z}$ are Polish spaces. The kernels k_x, k_z are continuous, $\sup_{x\in\mathcal{X}} k(x,x) + \sup_{z\in\mathcal{Z}} k(z,z) \le \kappa^2$, and Mercer's representations [32] of k_x and k_z exist. The random variable $\mathbf{y} - f^\dagger(\mathbf{x})$ is sub-exponential.*

Assumption 3.3 imposes technical conditions frequently assumed in literature [e.g., 33]. The requirements on the kernels can be satisfied by e.g. continuous kernels on compact subsets of $\mathbb{R}^d$.

Given the assumptions above, we characterize (C1) by showing that as $n \to \infty$, the posterior places vanishing mass on the region of functions that violates the GMM constraints (1). Concretely,

Theorem 3.1 (Proof in Appendix B)**.** *There exist a constant $M > 0$ depending on the data distribution and the kernels of choice, and a sufficiently slowly growing sequence $\{\gamma_n\} \to \infty$ (e.g., we can always have $\gamma_n \le \log\log n$) such that when taking $\bar\lambda = n^{-1/2}, \bar\nu = \min\{\tau^2_{\lceil\sqrt{n}\rceil}, n^{-1/2}\gamma_n\}$, it has*

$$\mathbb{E}_{\mathcal{D}^{(n)}}\Pi(\{f : \mathbb{E}^2(f(\mathbf{x}) - \mathbf{y} \mid \mathbf{z}) > M\tau^2_{\lceil\sqrt{n}\rceil}\} \mid \mathcal{D}^{(n)}) \to 0. \tag{13}$$

Remark 3.1. The above result should primarily be interpreted as a consistency result, although it also provides a posterior contraction rate [34] in the order of $\tau_{\lceil\sqrt{n}\rceil}$ in the semi-norm $f \mapsto \|Ef\|_{L_2(P(dz))}$, where $\{\tau_m\}$ is defined in Assumption 3.2. As an example, suppose the regularity of $f^\dagger$ is similar to the Matérn-$1/2$ RKHS; then for suitable kernels we have $\tau_{\lceil\sqrt{n}\rceil} = O(n^{-1/8})$, see Remark A.3.

The rate is suboptimal, and can be immediately improved if we impose further regularity assumption on k_z: if we assume the critical radius of the local Rademacher complexity [35] of the unit-norm ball $\mathcal{I}_1$ is $\delta_n \asymp \tau_n \asymp n^{-\frac{b}{2(b+1)}}$, a rate of $O(n^{-\frac{b-1}{2(b+1)}})$ for the semi-norm can be established following similar arguments. This is still worse than [26], which provides the rate $O(\max\{\tau_n, \delta_n\})$ for what corresponds to our posterior mean estimator. Nonetheless, their result is also generally suboptimal, because the choice of $\delta_n \asymp \tau_n$ is actually suboptimal when the IV regression problem is ill-posed. To date, minimax optimal rates have only been established under additional assumptions on the relations between the conditional expectation operator and the model for f. Our extended version establishes such results, for both this semi-norm and more intuitive norms such as $L_2(P(dx))$.

Remark 3.2. The use of an increasing $\lambda = n\bar\lambda$ is a technical artifact due to our minimal assumptions on k_z. It is common to impose extra regularization in nonparametric IV, due to the need to estimate E from data [19], and, in the quasi-Bayesian setting, also due to the additional technical challenges [18]. However, this is not necessary for our method if additional assumptions are imposed, as we show in the extended version. In practice, our hyperparameter selection procedure does not produce λ with significant growth.

The following proposition characterizes (C2):

Proposition 3.1 (Proof in Appendix B)**.** *The scaled log quasi-likelihood estimate $\ell_n(f) = d_n^2(\hat{E}_n f - \hat{b})$ satisfies*

$$\Pi\left(\left\{\forall f : \forall\delta > 0, \lim_{n\to\infty} \mathbb{P}_{\mathcal{D}^{(n)}}\left(\left|\ell_n(f) - \mathbb{E}_{\mathbf{z}\sim P(z)}(\mathbb{E}(f(\mathbf{x}) - \mathbf{y} \mid \mathbf{z})^2)\right| > \delta\right) = 0\right\}\right) = 1.$$

In words, for Π-almost every f, the log quasi-likelihood estimate scaled by n^{-1} converges to $\mathbb{E}_{\mathbf{z}\sim P(z)}(\mathbb{E}(f(\mathbf{x}) - \mathbf{y} \mid \mathbf{z})^2)$, which characterizes the violation of (1). While the restriction to the probability-1 subset can be concerning in the nonparametric setting [34], the proof only depends on f satisfying similar approximability conditions to $f^\dagger$; see (45).

Remark 3.3. To understand why the proposition characterizes (C2), suppose there are multiple f that satisfy (1). Then all of them will eventually have significantly higher log likelihood than functions violating (1): the difference is $\Theta(\lambda^{-1}n)$. If these functions have a similar level of regularity in the sense that their *concentration functions*, which is the right-hand side of (12), have the same asymptotics as $\epsilon \to 0$, Borell's inequality [e.g., 34, Proposition 11.17] implies that the posterior mass of small $\|\cdot\|$-norm balls around them will also have the same asymptotics.

Finally, we provide the following (over-)simplified example, which compares the behavior of the quasi-posterior and bootstrap in the context of (C2):

Example 3.1. Suppose $\mathbf{z}$ is completely non-informative so that $\mathbb{E}(f(\mathbf{x}) \mid \mathbf{z}) \equiv \mathbb{E}f(\mathbf{x})$ for all f; and suppose the estimated conditional expectation $\hat{E}_n$ is sufficiently accurate, so that we replace it

with E.[4] In this case bootstrap on (3) will always return the point estimator $f \equiv 0$ due to the non-zero regularization on $\|f\|_{\mathcal{H}}$, while the quasi-posterior behaves like the prior, correctly reflecting the complete lack of evidence in data.

While this example is oversimplified, and in practice the estimation error of E plays an important role, it is known that bootstrap uncertainty estimates can be problematic given weak IVs [36–38]. We also observe similar failures for bootstrap in the experiments (see Section 6.1).

4 Scalable Approximate Inference via a Randomized Prior Trick

We now turn to approximate inference with parametric models such as random feature expansion or wide NNs. Scalable inference for the IV quasi-posterior appears difficult, since for any f, computing the quasi-likelihood involves computing $\hat{E}_n f$, which in turn requires either inverting an $n \times n$ Gram matrix, judging from (5) and (11), or solving an optimization problem specific to f from (4). Nonetheless, we show that it is possible, by extending the "randomized prior" trick for Gaussian process regression [22] to work with (quasi-)likelihoods with an optimization formulation as in (4).

Our algorithm works with random feature models. A random feature model for k_z approximates $k_z(z, z') \approx \tilde{k}_{z,m}(z, z') := \frac{1}{m}\phi_{z,m}(z)^\top \phi_{z,m}(z')$, where $\phi_{z,m}$ takes value in $\mathbb{R}^m$. Then the map $\varphi \mapsto \frac{1}{\sqrt{m}}\varphi^\top \phi_{z,m}(\cdot) =: g(\cdot; \varphi)$ parameterizes an approximate RKHS $\tilde{\mathcal{H}}$; and for all $c > 0$, the random function $g(\cdot; \varphi)$, where $\varphi \sim \mathcal{N}(0, cI)$, is distributed as $\mathcal{GP}(0, c\tilde{k}_{z,m})$. The notations related to k_x are similar and thus omitted. Now we can state the objective function:

Proposition 4.1 (Proof in Appendix C.1)**.** *Let $\phi_0 \sim \mathcal{N}(0, \lambda\nu^{-1}I), \theta_0 \sim \mathcal{N}(0, I)$, $\tilde{y}_i \sim \mathcal{N}(y_i, \lambda)$. Then the optima θ^* of*

$$\min_{\theta\in\mathbb{R}^m} \max_{\phi\in\mathbb{R}^m} \sum_{i=1}^n \left((f(x_i;\theta) - \tilde{y}_i) g(z_i;\phi) - \frac{g(z_i;\phi)^2}{2} \right) - \frac{\nu}{2}\|\phi - \phi_0\|_2^2 + \frac{\lambda}{2}\|\theta - \theta_0\|_2^2 \tag{14}$$

parameterizes a random function which follows the quasi-posterior distribution (6)*, where the kernels are replaced by the random feature approximations.*

Given the above proposition, we can sample from the random feature-approximated quasi-posterior by solving (14) with stochastic gradient descent-ascent; the approximation errors will be analyzed in the following. The objective (14) is closely related to (4); as we show in Appendix C.1.1, it is equivalent to

$$\min_{f\in\tilde{\mathcal{H}}} \max_{g\in\tilde{\mathcal{I}}} \sum_{i=1}^n \left((f(x_i) - \tilde{y}_i) g(z_i) - \frac{g(z_i)^2}{2} \right) - \frac{\nu}{2}\|g - g_0\|_{\tilde{\mathcal{I}}}^2 + \frac{\lambda}{2}\|f - f_0\|_{\tilde{\mathcal{H}}}^2, \tag{15}$$

which differs from (4) only in the regularizers: instead of regularizing the norm of f and g, (15) encourages the functions to stay close to randomly sampled *anchors* [39]. Alternatively, we can view (15) as *perturbing* the point estimator (4), so that it has a covariance matching that of the quasi-posterior. A similar relation is also observed in the original randomized prior trick, which transforms GP regression to the optimization problem $\min_{f\in\tilde{\mathcal{H}}} \sum_{i=1}^n (f(x_i) - \tilde{y}_i)^2 + \lambda\|f - f_0\|_2^2$. In both cases, the resultant algorithm for approximate inference has the same time complexity as ensemble training for point estimation.

While the algorithm can be directly applied to neural network models as in [22], we follow [40] and modify the objective, to account for the difference between the neural tangent kernel (NTK) [41] of a wide neural network architecture and the NNGP kernel of the corresponding infinite-width Bayesian neural network [42–44]. Concretely, we modify (14) as

$$\min_{\theta} \max_{\phi} \sum_{i=1}^n \left((\tilde{f}_\theta(x_i) - \tilde{y}_i)\tilde{g}_\phi(z_i) - \frac{\tilde{g}_\phi(z_i)^2}{2} \right) - \frac{\nu}{2}\|\phi - \phi_0\|_2^2 + \frac{\lambda}{2}\|\theta - \theta_0\|_2^2, \tag{16}$$

$$\text{where } \tilde{g}_\phi(z) := g(z;\phi) - g(z;\phi_0) + \tilde{g}_0(z), \;\; \tilde{g}_0(z) := \sqrt{\frac{\lambda}{\nu}}\Big\langle \bar{\phi}_0, \; \frac{\partial g}{\partial \phi}\Big|_{\phi=\phi_0}(z) \Big\rangle,$$

[4] This setting could be realistic in the sample splitting setup considered in [7]. Note that as long as we have finite samples for the estimation of f (the "second stage"), we still need to have $\bar{\lambda} > 0$.

and ϕ_0 denotes the initial value of ϕ, and $\bar{\phi}_0 \sim \mathcal{N}(0, I)$ is a set of randomly initialized NN parameters independent of ϕ_0; and $\tilde{f}_\theta$ is defined similarly.

We only give a formal justification for the modification, *under the assumption*[5] that the NNs remain in the kernel regime throughout training, so that $g(z;\phi) - g(z;\phi_0) = \langle \phi - \phi_0, \frac{\partial g(z)}{\partial \phi}|_{\phi_0} \rangle_2$ [47]. Thus for the purpose of analyzing $g(\cdot;\phi) - g(\cdot;\phi_0)$, we can view g as a random feature model with the parameterization $\phi \mapsto \langle \phi, \frac{\partial g(z)}{\partial \phi}|_{\phi_0} \rangle_2$. Thus by the argument in Appendix C.1.1, we can show that the weight regularizer $\|\phi - \phi_0\|_2$ is equivalent to $\|g(\cdot;\phi) - g(\cdot;\phi_0)\|_{\tilde{\mathcal{I}}} = \|\tilde{g}_\phi - \tilde{g}_0\|_{\tilde{\mathcal{I}}}$, where $\tilde{\mathcal{I}}$ is determined by the NTK $k_{g,ntk}(z,z') := \langle \frac{\partial g(z)}{\partial \phi}|_{\phi_0}, \frac{\partial g(z')}{\partial \phi}|_{\phi_0} \rangle_2$. Similar arguments can be made for $\tilde{f}_\theta$ and $\tilde{f}_0$. Consequently, (16) is equivalent to an instance of (15) with $\tilde{\mathcal{H}}, \tilde{\mathcal{I}}$ defined by the NTKs.

Implementation details for the algorithm, including hyperparameter selection, are discussed in Appendix D.

Convergence analysis We now complete the analysis of the inference algorithm, by showing that for any fixed set of test points x_*, SGDA can approximate the marginal distribution $\Pi(f(x_*) \mid \mathcal{D}^{(n)})$ arbitrarily well given a sufficient computational budget. This implies that the approximate posterior is good enough for prediction purposes.

We place several mild assumptions on the random feature model, listed in Appendix C.2; they are satisfied by common approximations such as the random Fourier features [48]. The SGDA algorithm is described in detail in Appendix C.4. Under this setup, we have

Proposition 4.2 (Proof in Appendix C.5)**.** *Fix $\mathcal{D}^{(n)}$ and $\lambda, \nu > 0$. Then there exist a sequence of choices of m and SGDA step-size schemes, such that for any $l \in \mathbb{N}$, we have*

$$\sup_{x^* \in \mathcal{X}^l} \max(\|\hat{\mu}_m - \mu_m\|_2, \|\hat{S}_m - S_m\|_F) \xrightarrow{p} 0.$$

In the above, $\hat{\mu}_m, \hat{S}_m$ denote the mean and covariance of the approximate marginal posterior for $f(x_*)$, μ, S correspond to the true posterior, $\|\cdot\|_F$ denotes the Frobenius norm, and the convergence in probability is defined with respect to the sampling of random feature basis.

5 Related Work

Quasi-Bayesian analysis for GMM estimation problems was first developed in [49, 50, 16], which provided theoretical analysis for parametric models. The use of the quasi-posterior is motivated from the maximum entropy principle, based on which similar ideas have been developed in the machine learning literature [51–53]. Alternatively, it can also be obtained as a non-informative limit of certain Bayesian posteriors [54]. [17, 18] first analyzed the use of nonparametric models for quasi-Bayesian IV; no numerical study was presented. Closer to our work is [19] which constructed a quasi-posterior using Nadaraya-Watson smoothing as $\hat{E}_n$. While similar in form to our method, the smoothing approach relies on stronger assumptions, especially for high-dimensional data; see [7, Appendix A.2.1] which compared the corresponding point estimators. Their approach is also harder to scale to large datasets.

Our quasi-Bayesian procedure builds upon the kernelized IV models [7, 20] and the dual formulation of IV regression [6, 20, 21, 26]. The formulation (3)-(4) is from [20, 26]. [7] proposes a similar method. Although it was motivated differently as a kernelized two-stage least squares (2SLS) estimator, its objective is asymptotically equivalent to (3); however, the slight difference in regularization prevents the use of estimation procedures similar to (4), as we discuss in Appendix C.1.4. An alternative kernel-based estimator is proposed in [55, 56]; in particular, [55] also includes a Gaussian process construction, but for the different purpose of computing a leave-one-out validation statistics. Other recent work on ML models for IV include [4, 57, 58]. It remains interesting future work to develop scalable quasi-Bayesian procedures for these methods, although the mean estimator derived from our quasi-posterior implementation has competitive performance.

[59] studies semi-parametric IV estimation using an exponentially tilted empirical likelihood (ETEL) estimator. ETEL also connect to Bayesian methods. However, similar to the quasi-Bayesian

[5] The same linearization assumption has been employed in [40, 45]. We refer readers to [46, 21] for analysis of the linearization error in various settings similar to ours.

approach, the need to estimate the conditional moment restrictions complicates the understanding of its frequentist behavior, as well as the design of scalable inference algorithms; in their case, it is especially unclear if the empirical likelihood can be computed in sublinear time. Other approaches for uncertainty quantification include Bayesian inference and bootstrap. We have discussed previous works on Bayesian IV and their limitations in Section 1. Bootstrap is typically justified in the asymptotic regime, which does not cover many scenarios where uncertainty is most needed; this is different from the quasi-Bayesian approach which can always be justified through (7). Moreover, standard bootstrap inference on 2SLS is known to be unreliable when the instrument strength is weak [36–38]; while remedies exist (e.g., [38]), they heavily rely on the linearity and additive noise assumptions, and are thus difficult to generalize to the nonlinear setting we are interested in. As the kernelized IV methods generalize 2SLS, we expect similar issues to exist in our setting.

6 Experiments

Code to reproduce the experiments is available at `https://github.com/meta-inf/qbdiv`.

6.1 1D Simulation

We first experiment on a variety of 1D synthetic datasets, constructed by modifying the setup in [5] to incorporate a nonlinear first stage, in a way similar to [7, 60]:

$$z := \text{sigmoid}(w), \quad x := \text{sigmoid}\left(\frac{\alpha w + (1-\alpha)u'}{\sqrt{\alpha^2 + (1-\alpha)^2}}\right), \quad y_i \sim \mathcal{N}(f_0(2x-1) + 2u, 0.1),$$

where (u, u') are normal random variables with unit variance and a correlation of 0.5, $w \sim \mathcal{N}(0, 1)$ is independent of (u, u'), α is a parameter controlling the instrument strength, and f_0 is constructed from the `sine`, `step`, `abs` or a `linear` function. We choose $N \in \{200, 1000\}$ and $\alpha \in \{0.05, 0.5\}$.

Our baselines include BayesIV [15], a state-of-the-art Bayesian model based on B-splines and Dirichlet process mixture; we also include bootstrap on 2SLS with ridge regularization, either applied directly to the input features (Linear), on their polynomial expansion (Poly), or on the same kernelized models (KIV) as ours.[6] Hyperparameter for the kernelized IV methods are selected by cross validation based on the observable first-stage and second-stage losses as in previous work [7, 20]; see Appendix D.1. For kernels we choose the RBF and Matérn kernels, although results for Matérn kernels are deferred to appendix for brevity. See Appendix D.3 for the detailed setup.

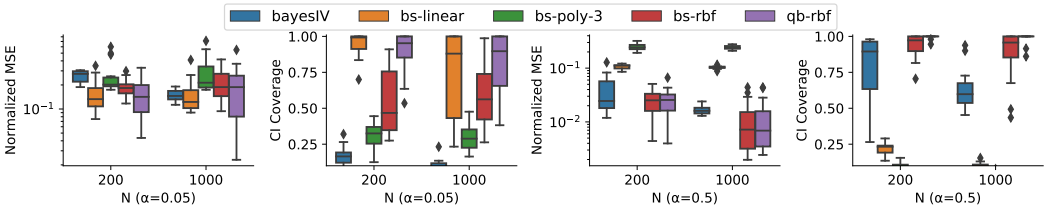


Figure 1: Test MSE and CI coverage on the `sine` dataset. The left two plots correspond to $\alpha = 0.05$, while the right two correspond to $\alpha = 0.5$. `bs` denotes bootstrap, `qb` denotes quasi-Bayesian.

Normalized MSE and coverage rate of 95% credible intervals (CI) on the `sine` datasets are plotted in Figure 1. We report results on 20 independently generated datasets. As we can see, quasi-Bayesian inference provides the most reliable uncertainty estimates, and is especially advantageous in the weak IV setting ($\alpha = 0.05$). While its CI can be conservative, we note that it is still informative, and properly reflects the sample size and instrument strength; see Appendix D.4 for visualizations. These results connect to previous work showing (in a different setting) that the radius of credible set produced by a GP posterior can have the correct order of magnitude [61].

We report the average run time in Table 1. We can see that in addition to the improved flexibility, our method is also more scalable compared with the standard Bayesian baseline, as it avoids the costly modeling of the joint residual distribution.

[6] We do not compare with [19] since their source code is unavailable. Note that their smoothing-based method does not scale to large datasets, and there is no numerical study on the credible interval in [19].

Table 1: Average run-time (in seconds) in the 1D simulation, for a single set of hyperparameters. N/A: does not converge after 20 minutes. For the approximate inference algorithm, we report the average time cost for parallel runs on a single accelerator; see Appendix D.3 for details.

N	1000	5000	20000
BayesIV (CPU)	655	N/A	N/A
QB (closed-form, CPU)	2.37	49.1	1150
QB (approx. inf., GPU)	~ 10	~ 40	~ 140

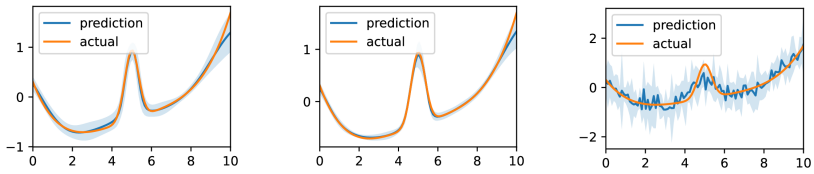
Full results on all datasets, visualizations and additional experiments are deferred to Appendix D.4. As a summary, (i) on the `abs` and `linear` datasets where all modeling assumptions hold, the results are qualitatively similar to the sin dataset. Moreover, the over-smoothed RBF kernel appears to have similar coverage comparing with the optimal kernel, and follow a similar contraction rate, as the previous work on GP regression [62] suggests. (ii) On the `step` dataset which violates Assumption 3.2, the quasi-posterior still provides more coverage than the baselines. (iii) Uncertainty estimates produced by the approximate inference algorithm are similar to that from the exact quasi-posterior.

6.2 Airline Demand

We now turn to the more challenging demand simulation first proposed by [4]. The dataset simulates a scenario where we need to predict the demand of airline tickets y, as a function of the price x, and two observed confounders: customer type s, and time of year t. The data generating process is

$$x := (z+3)\psi(t) + 25 + u', \quad y := f_0(x,t,s) + u, \quad f_0(x,t,s) := 100 + (10+x)\cdot s \cdot \psi(t) - 2x$$

where (u, u') are standard normal variables with correlation ρ, $z \sim \mathcal{N}(0,1)$ is independent of (u, u'), and ψ is a nonlinear function whose shape is given in Figure 2. The variable s either varies across $\{0, \dots, 6\}$ (the lower-dimensional setting), or is observed as an MNIST image representing the corresponding digit; the latter case represents the real-world scenario where only high-dimensional surrogates of the true confounder is observed. We only report results for $\rho = 0.5$, noting that results using other choices of ρ have been similar. We use $n \in \{1000, 10000\}$ for the lower-dimensional setting, and use $n = 50000$ for the image-based setting.



(a) lower-dim, $N = 10^3$ (b) lower-dim, $N = 10^4$ (c) image setup

Figure 2: Approximate quasi-posteriors in the demand simulation. We plot a cross-section by fixing s, x to their mean values and varying t.

We compare our method with bootstrap on the same model, BayesIV, and bootstrap on linear or polynomial models. Performance of other point estimators on this dataset has been reported in [7, 20, 58], compared with which our method is generally competitive. We implement the dual IV model using both an RBF kernel and DNN models. See Appendix D.5 for details.

We report the test MSE and CI coverage for $N = 1000$ in Table 2, and visualize all approximate quasi-posteriors for the NN models in Figure 2. As we can see, when implemented with DNNs, our method produces uncertainty estimates with excellent coverage, which also correctly reflects the information available in the dataset: the CI is wider when N is smaller, or in the high-dimensional experiment where estimation is harder. Bootstrap has a noticeably worse performance when $N = 1000$. Still, it performs well in the (arguably less interesting) large-sample setting, with a CI coverage similar to our method; see Appendix D.6. This is because on this dataset, the total instrument strength is stronger due to the presence of observed confounders, and the NN model is a good fit. Consequently, the asymptotic behavior of bootstrap can be observed when N is large.

Both methods have poorer performances when we switch to the RBF kernel, although the quasi-posterior is still more reliable. We conjecture that both $\mathcal{H}$ and $\mathcal{I}$ are misspecified in this case, and

thus the credible intervals can only reflect the uncertainty within the prior model. These results show that NN models can be advantageous, which our inference algorithm supports.

The other baselines perform poorly due to their inflexibility; in particular, note that BayesIV uses additive models for both stages (e.g., $f(x,t,s) = f_1(x) + f_2(t) + f_3(s)$) which do not approximate this data generating process well. Full results and visualizations are deferred to Appendix D.6.

Table 2: Test normalized MSE, average CI coverage and CI width on the demand dataset with $N = 1000$. Results averaged over 20 trials.

Method	BS-Linear	BS-Poly	BayesIV	BS-RBF	QB-RBF	BS-NN	QB-NN
NMSE	$.37 \pm .01$	$.31 \pm .06$	$.28 \pm .04$	$.17 \pm .01$	$.17 \pm .01$	$.06 \pm .03$	$.04 \pm .00$
CI Cvg.	$.09 \pm .01$	$.15 \pm .03$	$.27 \pm .06$	$.45 \pm .02$	$.77 \pm .02$	$.86 \pm .02$	$.94 \pm .01$
CI Wid.	$.09 \pm .01$	$.16 \pm .04$	$.08 \pm .06$	$.18 \pm .02$	$.37 \pm .01$	$.14 \pm .01$	$.26 \pm .04$

7 Conclusion

In this work we propose a scalable quasi-Bayesian procedure for IV regression. We analyze the theoretical properties of the proposed quasi-posterior, and derive a scalable algorithm for approximate inference. Empirical evaluations show that the proposed method scales to large and high-dimensional datasets, and can be particularly advantageous when the instrument strength is weak.

Beyond IV regression, formulations like (1) also appear in various other problems in causal inference and statistics, as discussed in, e.g., [21]; our method can be readily applied to these problems. Future work includes extension to more general conditional moment restriction problems with nonlinear constraints [63, 17].

Acknowledgement

We thank the anonymous reviewers for their helpful feedback and references. Z.W., Y.Z. and J.Z. were supported by NSFC Projects (Nos. 61620106010, 62061136001, 61621136008, 62076147, U19B2034, U19A2081, U1811461), Beijing NSF Project (No. JQ19016), Beijing Academy of Artificial Intelligence (BAAI), Tsinghua-Huawei Joint Research Program, a grant from Tsinghua Institute for Guo Qiang, Tiangong Institute for Intelligent Computing, and the NVIDIA NVAIL Program with GPU/DGX Acceleration.

Broader Impact

IV analysis is widely used in social science and clinical research, where it can be highly undesirable to have a systematic error or increased prediction variance on subpopulations which are underrepresented in the training dataset, or on which the instrument is locally weak. Uncertainty quantification is an important first step in detecting such issues, and the quasi-Bayesian approach we take can be preferable since it has a clear interpretation in a non-asymptotic setting, and when the instrument strength is weak. However, epistemic uncertainty estimates are only as good as the model allows, and it should not create a false sense of complete security. Caution must be taken when employing complex models in scenarios with potential fairness implications.

References

[1] J. D. Angrist and J.-S. Pischke, *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

[2] S. Greenland, "An introduction to instrumental variables for epidemiologists," *International Journal of Epidemiology*, vol. 29, no. 4, pp. 722–729, 2000.

[3] J. Cuzick, R. Edwards, and N. Segnan, "Adjusting for non-compliance and contamination in randomized clinical trials," *Statistics in Medicine*, vol. 16, no. 9, pp. 1017–1029, 1997.

[4] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy, "Deep IV: A flexible approach for counterfactual prediction," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1414–1423.

[5] G. Lewis and V. Syrgkanis, "Adversarial generalized method of moments," *arXiv preprint arXiv:1803.07164*, 2018.

[6] A. Bennett, N. Kallus, and T. Schnabel, "Deep generalized method of moments for instrumental variable analysis," *arXiv preprint arXiv:1905.12495*, 2019.

[7] R. Singh, M. Sahani, and A. Gretton, "Kernel Instrumental Variable Regression," *arXiv:1906.00232 [cs, econ, math, stat]*, Jul. 2020, arXiv: 1906.00232.

[8] A. Bennett and N. Kallus, "The Variational Method of Moments," *arXiv:2012.09422 [cs, econ, math, stat]*, Dec. 2020, arXiv: 2012.09422.

[9] J. H. Stock, J. H. Wright, and M. Yogo, "A survey of weak instruments and weak identification in generalized method of moments," *Journal of Business & Economic Statistics*, vol. 20, no. 4, pp. 518–529, 2002.

[10] W. K. Newey, "Nonparametric instrumental variables estimation," *American Economic Review*, vol. 103, no. 3, pp. 550–56, 2013.

[11] F. Kleibergen and H. K. Van Dijk, "Bayesian simultaneous equations analysis using reduced rank structures," *Econometric theory*, pp. 701–743, 1998.

[12] F. Kleibergen and E. Zivot, "Bayesian and classical approaches to instrumental variable regression," *Journal of Econometrics*, vol. 114, no. 1, pp. 29–72, 2003.

[13] T. G. Conley, C. B. Hansen, R. E. McCulloch, and P. E. Rossi, "A semi-parametric Bayesian approach to the instrumental variable problem," *Journal of Econometrics*, vol. 144, no. 1, pp. 276–305, 2008.

[14] H. F. Lopes and N. G. Polson, "Bayesian instrumental variables: priors and likelihoods," *Econometric Reviews*, vol. 33, no. 1-4, pp. 100–121, 2014.

[15] M. Wiesenfarth, C. M. Hisgen, T. Kneib, and C. Cadarso-Suarez, "Bayesian nonparametric instrumental variables regression based on penalized splines and dirichlet process mixtures," *Journal of Business & Economic Statistics*, vol. 32, no. 3, pp. 468–482, 2014.

[16] V. Chernozhukov and H. Hong, "An MCMC approach to classical estimation," *Journal of Econometrics*, vol. 115, no. 2, pp. 293–346, Aug. 2003.

[17] Y. Liao and W. Jiang, "Posterior consistency of nonparametric conditional moment restricted models," *The Annals of Statistics*, vol. 39, no. 6, pp. 3003–3031, Dec. 2011.

[18] K. Kato, "Quasi-Bayesian analysis of nonparametric instrumental variables models," *The Annals of Statistics*, vol. 41, no. 5, pp. 2359–2390, Oct. 2013.

[19] J.-P. Florens and A. Simoni, "Nonparametric estimation of an instrumental regression: A quasi-Bayesian approach based on regularized posterior," *Journal of Econometrics*, vol. 170, no. 2, pp. 458–475, Oct. 2012.

[20] K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj, "Dual Instrumental Variable Regression," *arXiv:1910.12358 [cs, econ, stat]*, Oct. 2020, arXiv: 1910.12358.

[21] L. Liao, Y.-L. Chen, Z. Yang, B. Dai, Z. Wang, and M. Kolar, "Provably efficient neural estimation of structural equation model: An adversarial approach," *arXiv preprint arXiv:2007.01290*, 2020.

[22] I. Osband, J. Aslanides, and A. Cassirer, "Randomized prior functions for deep reinforcement learning," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 8626–8638.

[23] W. K. Newey and J. L. Powell, "Instrumental variable estimation of nonparametric models," *Econometrica*, vol. 71, no. 5, pp. 1565–1578, 2003.

[24] K. Fukumizu, F. R. Bach, and A. Gretton, "Statistical Consistency of Kernel Canonical Correlation Analysis," *Journal of Machine Learning Research*, p. 23, 2004.

[25] B. Dai, N. He, Y. Pan, B. Boots, and L. Song, "Learning from Conditional Distributions via Dual Embeddings," *arXiv:1607.04579 [cs, math, stat]*, Dec. 2016, arXiv: 1607.04579.

[26] N. Dikkala, G. Lewis, L. Mackey, and V. Syrgkanis, "Minimax estimation of conditional moment models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 248–12 262.

[27] A. M. Stuart, "Inverse problems: a Bayesian perspective," *Acta numerica*, vol. 19, pp. 451–559, 2010.

[28] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur, "Gaussian processes and kernel methods: A review on connections and equivalences," *arXiv preprint arXiv:1807.02582*, 2018.

[29] T. Zhang *et al.*, "From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation," *The Annals of Statistics*, vol. 34, no. 5, pp. 2180–2210, 2006.

[30] A. W. van der Vaart and J. H. van Zanten, "Rates of contraction of posterior distributions based on Gaussian process priors," *The Annals of Statistics*, vol. 36, no. 3, Jun. 2008.

[31] ——, "Information Rates of Nonparametric Gaussian Process Regression," *Journal of Machine Learning Research*, vol. 12, pp. 2095–2119, 2011.

[32] I. Steinwart and C. Scovel, "Mercer's Theorem on General Domains: On the Interaction between Measures, Kernels, and RKHSs," *Constructive Approximation*, vol. 35, no. 3, pp. 363–417, Jun. 2012.

[33] L. H. Dicker, D. P. Foster, and D. Hsu, "Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators," *Electronic Journal of Statistics*, vol. 11, no. 1, Jan. 2017.

[34] S. Ghosal and A. Van der Vaart, *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, 2017, vol. 44.

[35] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.

[36] M. Moreira, J. R. Porter, and G. A. Suarez, "Bootstrap and higher-order expansion validity when instruments may be weak," 2004.

[37] A. Flores-Lagunes, "Finite sample evidence of IV estimators under weak instruments," *Journal of Applied Econometrics*, vol. 22, no. 3, pp. 677–694, Apr. 2007.

[38] R. Davidson and J. G. Mackinnon, "Wild Bootstrap Tests for IV Regression," *Journal of Business & Economic Statistics*, vol. 28, no. 1, pp. 128–144, 2010, publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[39] T. Pearce, F. Leibfried, and A. Brintrup, "Uncertainty in neural networks: Approximately Bayesian ensembling," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. PMLR, 26–28 Aug 2020, pp. 234–244.

[40] B. He, B. Lakshminarayanan, and Y. W. Teh, "Bayesian deep ensembles via the neural tangent kernel," *arXiv preprint arXiv:2007.05864*, 2020.

[41] A. Jacot, F. Gabriel, and C. Hongler, "Neural Tangent Kernel: Convergence and Generalization in Neural Networks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.

[42] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 2012, vol. 118.

[43] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, "Deep neural networks as Gaussian processes," *arXiv preprint arXiv:1711.00165*, 2017.

[44] A. G. d. G. Matthews, M. Rowland, J. Hron, R. E. Turner, and Z. Ghahramani, "Gaussian process behaviour in wide deep neural networks," *arXiv preprint arXiv:1804.11271*, 2018.

[45] W. Hu, Z. Li, and D. Yu, "Simple and effective regularization methods for training on noisily labeled data with generalization guarantee," in *International Conference on Learning Representations*, 2020.

[46] T. Hu, W. Wang, C. Lin, and G. Cheng, "Regularization matters: A nonparametric perspective on overparametrized neural network," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 13–15 Apr 2021, pp. 829–837.

[47] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[48] A. Rahimi, B. Recht *et al.*, "Random features for large-scale kernel machines." in *NIPS*, vol. 3, no. 4. Citeseer, 2007, p. 5.

[49] A. Zellner, "Bayesian Method of Moments (BMOM) Analysis of Mean and Regression Models," *arXiv:bayes-an/9511001*, Dec. 1995, arXiv: bayes-an/9511001.

[50] J.-Y. Kim, "Limited information likelihood and Bayesian analysis," *Journal of Econometrics*, p. 19, 2002.

[51] T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," in *Proceedings of the 12th International Conference on Neural Information Processing Systems*, 1999, pp. 470–476.

[52] M. Dudík, S. J. Phillips, and R. E. Schapire, "Maximum entropy density estimation with generalized regularization and an application to species distribution modeling," *Journal of Machine Learning Research*, 2007.

[53] J. Zhu and E. P. Xing, "Maximum entropy discrimination Markov networks." *Journal of Machine Learning Research*, vol. 10, no. 11, 2009.

[54] J.-P. Florens and A. Simoni, "Gaussian processes and bayesian moment estimation," *Journal of Business & Economic Statistics*, vol. 39, no. 2, pp. 482–492, 2021.

[55] R. Zhang, M. Imaizumi, B. Schölkopf, and K. Muandet, "Maximum Moment Restriction for Instrumental Variable Regression," *arXiv:2010.07684 [cs]*, Oct. 2020, arXiv: 2010.07684.

[56] A. Mastouri, Y. Zhu, L. Gultchin, A. Korba, R. Silva, M. J. Kusner, A. Gretton, and K. Muandet, "Proximal causal learning with kernels: Two-stage estimation and moment restriction," *arXiv preprint arXiv:2105.04544*, 2021.

[57] A. Puli and R. Ranganath, "General control functions for causal effect estimation from instrumental variables," *Advances in Neural Information Processing Systems*, vol. 33, p. 8440, 2020.

[58] L. Xu, Y. Chen, S. Srinivasan, N. de Freitas, A. Doucet, and A. Gretton, "Learning Deep Features in Instrumental Variable Regression," *arXiv:2010.07154 [cs, stat]*, Oct. 2020, arXiv: 2010.07154.

[59] S. Chib, M. Shin, and A. Simoni, "Bayesian estimation and comparison of conditional moment models," 2019.

[60] X. Chen and T. M. Christensen, "Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression," *Quantitative Economics*, vol. 9, no. 1, pp. 39–84, 2018.

[61] B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten, "Bayesian inverse problems with Gaussian priors," *The Annals of Statistics*, vol. 39, no. 5, pp. 2626–2657, Oct. 2011, arXiv: 1103.2692.

[62] A. Van der Vaart, H. Van Zanten *et al.*, "Bayesian inference with rescaled Gaussian process priors," *Electronic Journal of Statistics*, vol. 1, pp. 433–448, 2007.

[63] C. Ai and X. Chen, "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, vol. 71, no. 6, pp. 1795–1843, 2003.

[64] S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart, "Convergence rates of posterior distributions," *Annals of Statistics*, pp. 500–531, 2000.

[65] A. Caponnetto and E. De Vito, "Optimal rates for the regularized least-squares algorithm," *Foundations of Computational Mathematics*, vol. 7, no. 3, pp. 331–368, 2007.

[66] K. Fukumizu, "Nonparametric Bayesian inference with kernel mean embedding," in *Modern Methodology and Applications in Spatial-Temporal Modeling*. Springer, 2015, pp. 1–24.

[67] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, F. Odone, and P. Bartlett, "Learning from examples as an inverse problem." *Journal of Machine Learning Research*, vol. 6, no. 5, 2005.

[68] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018, vol. 47.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See, e.g., the discussion around the assumptions and theoretical results in Section 3.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We warn against potential misuse of this work in the broader impact section.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.1.
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs are in Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix D.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] We did not use any existing asset.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Our research does not involve human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Background on Gaussian Process Regression

We review some standard results on Gaussian process regression. They will be needed in our proof in the following, and provide more context to the results in the main text. For a thorough overview of this subject, see, for example, [34].

Notations The appendix uses the following additional notations: $\lesssim, \gtrsim$ represent inequality up to a universal constant, $\asymp$ denotes equivalent up to constants. $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert-Schmidt norm.

For infinite-dimensional Gaussian process models, the prior draws almost surely fall out of the corresponding RKHS. Therefore, our posterior analysis will rely on the following result, showing that the GP prior support can be approximated with increasing accuracy using balls in the RKHS with increasing norm, in terms of a weaker norm that can be defined on the entire prior support (e.g., the sup norm).

Theorem A.1 ([30], Theorem 2.1)**.** *Let W be a Borel measurable, zero-mean Gaussian random element in a separable Banach space $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$, and let w_0 be contained in the closure of $\mathcal{H}$ in $\mathbb{B}$. Let $\tau_n^2 > 0$ be a number such that*

$$\phi_{w_0}(\tau_n) \leq n\tau_n^2, \tag{17}$$

where

$$\phi_{w_0}(\tau) = \inf_{h \in \mathcal{H}: \|h - w_0\| < \tau} \|h\|_{\mathcal{H}}^2 - \log P(\|W\| < \tau). \tag{18}$$

Then, for any $C_\Theta > 1$ with $e^{-C_\Theta n\tau_n^2} < 1/2$, the set

$$\Theta_n = \tau_n \mathbb{B}_1 + \underline{J}_n \mathcal{H}_1 \tag{19}$$

is measurable and satisfy

$$\log N(3\tau_n, \Theta_n, \|\cdot\|) \leq 6C_\Theta n\tau_n^2, \tag{20}$$

$$\mathbb{P}(W \notin \Theta_n) \leq e^{-C_\Theta n\tau_n^2}, \tag{21}$$

$$\mathbb{P}(\|W - w_0\| < 2\tau_n) \geq e^{-n\tau_n^2}. \tag{22}$$

In the above $\mathbb{B}_1, \mathcal{H}_1$ are the unit norm balls in the corresponding spaces, and $J_n = -2\Phi^{-1}(e^{-C_\Theta n\tau_n^2})$ where Φ^{-1} is the inverse CDF of the standard normal distribution.

Our analysis will make use of the following:

Corollary A.1. *Fix any $w_0 \in \mathbb{B}$. Then for any $n \in \mathbb{N}$,*

(i). $\underline{J}_n \leq 2\sqrt{2C_\Theta n\tau_n^2} =: J_n$.

(ii). there exists $w_n^\dagger \in \mathcal{H}$ such that $\|w_n^\dagger - w_0\| \leq \tau_n$, and

$$\mathbb{P}(\|W - w_n^\dagger\| \leq 2\tau_n) \geq e^{-3n\tau_n^2}. \tag{23}$$

Proof. (i) holds because $\Phi(t) \geq 1 - e^{-t^2/2}$. for (ii), from (17)-(18) we can see that such $w_n^\dagger \in \mathcal{H}$ exists, and we can find $w_n^\dagger$ so that

$$\|w_n^\dagger\|_{\mathcal{H}} \leq 2\phi_{w_0}(\tau_n) \leq 2n\tau_n^2.$$

(23) follows from thet inequality

$$-\log P(\|W - w_n^\dagger\| \leq 2\tau_n) \overset{(a)}{\leq} \phi_{w_n^\dagger}(\tau_n) \leq \|w_n^\dagger\|_{\mathcal{H}}^2 - \log P(\|W\| < \tau_n) \overset{(b)}{\leq} 3n\tau_n^2.$$

where (a) can be found in Lemma I.28, [34]; and (b) from (17). □

Remark A.1. For Gaussian processes the space $\mathbb{B}$ is a function space. In the analysis of our algorithms, we require that the norm $\|\cdot\|$ in the space $\mathbb{B}$ is at least equivalent to the sup norm $\|f\|_\infty := \sup_x |f(x)|$, i.e., $\|f\|_\infty \lesssim \|f\|$. This requirement is natural for most examples. For example, the space $\mathbb{B}$ is generally chosen to be the continuous function space equipped with the sup norm.

Choices of τ_n Choices of τ_n will affect the posterior contraction rate. In general, τ_n is determined by the ability of RKHS $\mathcal{H}$ to approximate the target function w_0, and the *small-ball probability* $\log P(\|W\| \le \tau)$ which is usually determined by the metric entropy $\log N(\tau, \mathcal{H}_1, \|\cdot\|)$ [34, Lemma I.30]. For the standard Matérn and RBF kernels and the sup norm as $\|\cdot\|$, valid choices for τ_n are provided in [31], which we review below.

Lemma A.1 (Matérn kernel. Lemma 3-4, [31])**.** *If $\mathcal{H}$ is the RKHS corresponding to the Matérn-α kernel in $[0,1]^d$, $w_0 \in C^\beta([0,1]^d) \cap H^\beta([0,1]^d)$,[7] the condition* (17) *will be satisfied with*

$$\tau_n^2 \asymp n^{-\frac{2\min(\alpha,\beta)}{2\alpha+d}},$$

where the constant hidden in $\asymp$ may depend on w_0.

Remark A.2. τ_n above usually determines the posterior contraction rate of GP regression using a normal likelihood with fixed variance [31]. For any fixed $\beta > 0$, it is minimized when we set $\alpha = \beta$. When $\alpha > d/2$, samples from the the corresponding GP belong to the space $C^{\underline{\alpha}}[0,1]^d \cap H^{\underline{\alpha}}[0,1]^d$ with probability 1, for all $\underline{\alpha} < \alpha$: see [31, pp. 2104], and [28, pp. 37-38]. Therefore, when $\alpha = \beta > \frac{d}{2}$, the above lemma applies to w_0 in a space that is very slightly smaller than the "sample space" of the prior. And in this case, τ_n matches the minimax rate for regression in $H^\beta([0,1]^d)$.

The practice of choosing kernel so that the GP sample space (approximately) matches the regularity of the target function is different from in kernel ridge regression, where the kernel is chosen so that the corresponding RKHS, a much smaller space than the GP sample space, matches the target regularity. Still, in all cases we can always invoke the above lemma when w_0 has less regularity. Although the resulted τ_n^2 may be worse, it is known that using an "over-smoothed" prior does not lead to worse rates if we allow the noise variance parameter to vary with n [62].

Remark A.3. When $w_0 = f^\dagger \in C^\beta[0,1]^d \cap H^\beta[0,1]^d$ with $\beta = \frac{d+1}{2}$, we can invoke the above theorem with $\alpha = \frac{d+1}{2}$ and obtain $\tau_n^2 \asymp n^{-1/2}$. The RKHS of the Matérn-$1/2$ kernel is norm equivalent to H^β [28], and C^β is often referred to as qualitatively having the same degree of regularity as H^β (see, e.g., [31]). This is a very basic assumption for regularity, since the eigendecay of the Matérn-$1/2$ kernel is $\lambda_j \asymp j^{-\frac{d+1}{d}}$; if we further slow down the decaying rate below j^{-1}, $\mathcal{C}_x$ will no longer be trace-class; equivalently, k_x will no longer be bounded, contradicting our Assumption 3.3.

The following lemma applies when RKHS $\mathcal{H}$ corresponding to the standard RBF kernel $k(x,x') := \exp(-\|x-x'\|^2/2)$, and $f \in A^{\gamma,r}$ which is a function space requiring exponential decrease of the Fourier transform.[8]

Lemma A.2 (RBF kernel. Lemma 6, 9, [31])**.** *Let f_0 be the restriction to $[0,1]^d$ an element of $A^{\gamma,r}(\mathbb{R}^d)$. Then*

(i). For $r > 2$ or $r = 2, \gamma \ge 4$, f_0 is in $\mathcal{H}$.

(ii). For $r \in (0,2)$, we have

$$\inf_{h\in\mathcal{H}:\|h-f_0\|<\tau} \|h\|_{\mathcal{H}}^2 - \log P(\|W\| \le \tau) \le C_1 \exp\left(\frac{\left(\log \tau^{-1}\right)^{2/r}}{4\gamma^{2/r}}\right) + C_2\left(\log \tau^{-1}\right)^{1+d}.$$

where C_1, C_2 only depends on d and f_0. Consequently, for any $r \ge 1$ and $w_0 \in A^{\gamma,r}(\mathbb{R}^d)$, we have

$$\tau_n^2 \asymp \frac{(\log n)^{2/r}}{n}.$$

Remark A.4. The Gaussian process using RBF kernel takes value in the space of real analytic functions, which corresponds to $A^{\gamma,r}$ with $r = 1$ [31]. Therefore, the above lemma applies to all functions in the "sample space" of the GP prior.

Remark A.5. Finally, note that in the sequel we will always assume that

$$n\tau_n^2 \to \infty.$$

As τ_n upper bounds the posterior contraction rate in Gaussian process regression, the above will always holds for infinite-dimensional models of interest; in general, as $\liminf n\tau_n^2 > 0$ must hold by (17), we can increase τ_n by, e.g., a logarithm factor, although for finite-dimensional models the analysis can be simplified considerably.

[7] C^β denotes the Hölder space of order β, and H^β denotes the Sobolev space of order β.

[8] The specific form is irrelevant for our purposes; see van der Vaart and van Zanten [31].

B Analysis of the Quasi-Posterior

In this section we prove Theorem 3.1 and Proposition 3.1. Our proof is similar to the adaptation of the posterior contraction framework [64] in [18], and involves bounding the log quasi-likelihood on certain events. However, it is different since in our case, $\hat{E}_n$ and the GP prior are not constructed with orthonormal basis in $L_2(p_{data})$. Moreover, we directly provide guarantee on the true GMM conditions (1), and do not make additional assumption on the data distribution; whereas [18] analyzed the estimated GMM conditions constructed from $\hat{E}_n$ and the empirical data distribution, and then moved to analyze f under various assumptions about the joint distribution $p_{data}(dz \times dx)$, including identifiability.

We introduce the following event

$$B_n(r, L) := A_n(r) \cap \left\{ \|\hat{C}_{zx} - C_{zx}\| \leq \frac{L}{\sqrt{n}} \right\} \cap \left\{ \left\| \frac{S_z^*}{n}(f^\dagger(X) - Y) \right\|_{\mathcal{I}} \leq \frac{L}{\sqrt{n}} \right\}, \tag{24}$$

where the event $A_n(r) := \left\{ \left\| C_{zz,\bar{\nu}}^{-1/2}(\hat{C}_{zz} - C_{zz})C_{zz,\bar{\nu}}^{-1/2} \right\| \leq r \right\}$. We will then bound the (scaled) log quasi-likelihood

$$\ell_n(f) := -\frac{2\lambda}{n} \log \frac{d\Pi(\cdot \mid \mathcal{D}^{(n)})}{d\Pi} = \left\| \hat{C}_{zz,\bar{\nu}}^{-1/2} \frac{S_z^*(f(X) - Y)}{n} \right\|_{\mathcal{I}}^2 \tag{25}$$

in both directions on the event $B_n(r, L)$.

B.1 Bounds on the Quasi-likelihood

Lemma B.1. *Conditioned on the event $B_n(r, L)$ for $r \in (0, 1/2)$, we have for all $f \in \mathcal{H}$*

$$-r_n^{(0)} + \sqrt{\frac{1-2r}{1-r}} \, \|C_{zz}^{-\frac{1}{2}} C_{zx} f\|_{\mathcal{I}} \leq \|\hat{C}_{zz,\bar{\nu}}^{-\frac{1}{2}} \hat{C}_{zx} f\|_{\mathcal{I}} \leq r_n^{(0)} + \sqrt{\frac{1}{1-r}} \, \|C_{zz}^{-\frac{1}{2}} C_{zx} f\|_{\mathcal{I}}, \tag{26}$$

where

$$r_n^{(0)} \lesssim \left(\frac{L}{\sqrt{\bar{\nu} n}} + \sqrt{\bar{\nu}} \right) \|f\|_{\mathcal{H}}. \tag{27}$$

Proof. On the event $A_n(r)$, we have

$$\begin{aligned}
|\|\hat{C}_{zz,\bar{\nu}}^{-1/2} C_{zx} f\|_{\mathcal{I}}^2 - \|C_{zz,\bar{\nu}}^{-1/2} C_{zx} f\|_{\mathcal{I}}^2| &= |\langle C_{zx} f, (C_{zz,\bar{\nu}}^{-1} - \hat{C}_{zz,\bar{\nu}}^{-1}) C_{zx} f \rangle_{\mathcal{I}}| \\
&\leq \|C_{zz,\bar{\nu}}^{1/2}(C_{zz,\bar{\nu}}^{-1} - \hat{C}_{zz,\bar{\nu}}^{-1}) C_{zz,\bar{\nu}}^{1/2}\| \cdot \|C_{zz,\bar{\nu}}^{-1/2} C_{zx} f\|_{\mathcal{I}}^2 \\
&= \|I - C_{zz,\bar{\nu}}^{1/2} \hat{C}_{zz,\bar{\nu}}^{-1} C_{zz,\bar{\nu}}^{1/2}\| \cdot \|C_{zz,\bar{\nu}}^{-1/2} C_{zx} f\|_{\mathcal{I}}^2 \\
&\leq \frac{r}{1-r} \|C_{zz,\bar{\nu}}^{-1/2} C_{zx} f\|_{\mathcal{I}}^2,
\end{aligned}$$

where the last inequality above uses (51) in Lemma B.6. Thus

$$\sqrt{\frac{1-2r}{1-r}} \|C_{zz,\bar{\nu}}^{-1/2} C_{zx} f\|_{\mathcal{I}} \leq \|\hat{C}_{zz,\bar{\nu}}^{-1/2} C_{zx} f\|_{\mathcal{I}} \leq \sqrt{\frac{1}{1-r}} \|C_{zz,\bar{\nu}}^{-1/2} C_{zx} f\|_{\mathcal{I}}.$$

Since $\|C_{zz,\bar{\nu}}^{-1/2} C_{zz}^{1/2}\| \leq 1$, the right hand side above is $\leq \sqrt{1/(1-r)} \|C_{zz}^{-1/2} C_{zx} f\|_{\mathcal{I}}$; for the left hand side, observe that

$$\|C_{zz}^{-1/2} C_{zx} f\|_{\mathcal{I}} - \|C_{zz,\bar{\nu}}^{-1/2} C_{zx} f\|_{\mathcal{I}} \leq \|C_{zz}^{1/2} - C_{zz,\bar{\nu}}^{-1/2} C_{zz}\| \|Ef\|_{\mathcal{I}}$$

where we recall $E = C_{zz}^{-1} C_{zx}$ is bounded by Assumption 3.1. To bound $\|C_{zz}^{1/2} - C_{zz,\bar{\nu}}^{-1/2} C_{zz}\|$, denote by $\{\lambda_i\}$ the eigenvalues of C_{zz}, then the i-th eigenvalue of $C_{zz}^{1/2} - C_{zz,\bar{\nu}}^{-1/2} C_{zz}$ is

$$\lambda_i^{1/2} - \frac{\lambda_i}{\sqrt{\lambda_i + \bar{\nu}}} = \frac{\sqrt{\lambda_i^2 + \lambda_i \bar{\nu}} - \sqrt{\lambda_i^2}}{\sqrt{\lambda_i + \bar{\nu}}} \overset{(a)}{\leq} \frac{\bar{\nu}/2}{\sqrt{\lambda_i + \bar{\nu}}} \leq \sqrt{\bar{\nu}}/2,$$

where (a) follows from the concavity of the square root function. Thus

$$\|C_{zz}^{-1/2}C_{zx}f\|_{\mathcal{I}} - \|C_{zz,\bar{\nu}}^{-1/2}C_{zx}f\|_{\mathcal{I}} \le \sqrt{\bar{\nu}}\|Ef\|_{\mathcal{I}}/2 \le \sqrt{\bar{\nu}}\|E\|\|f\|_{\mathcal{H}}/2,$$

and

$$\sqrt{\frac{1-2r}{1-r}}\left(\|C_{zz}^{-1/2}C_{zx}f\|_{\mathcal{I}} - \frac{\sqrt{\bar{\nu}}}{2}\|E\|\|f\|_{\mathcal{H}}\right) \le \|\hat{C}_{zz,\bar{\nu}}^{-1/2}C_{zx}f\|_{\mathcal{I}} \le \sqrt{\frac{1}{1-r}}\|C_{zz}^{-1/2}C_{zx}f\|_{\mathcal{I}}. \tag{28}$$

Note that on the event $B_n(r, L)$, we have

$$|\|\hat{C}_{zz,\bar{\nu}}^{-1/2}\hat{C}_{zx}f\|_{\mathcal{I}} - \|\hat{C}_{zz,\bar{\nu}}^{-1/2}C_{zx}f\|_{\mathcal{I}}| \le \|\hat{C}_{zz,\bar{\nu}}^{-1/2}\|\|\hat{C}_{zx} - C_{zx}\|\|f\|_{\mathcal{H}} \le \frac{L}{\sqrt{n\bar{\nu}}}\|f\|_{\mathcal{H}}. \tag{29}$$

Combining (28) and (29) completes the proof. □

Lemma B.2. *Conditioned on the event $B_n(r, L)$ for $r \in (0, 1/2)$, we have for all $f \in \mathbb{B}$ that can be written as $f = f_h + f_e$ where $f_h \in \mathcal{H}$, $\|f_e\| \le 2\tau_m$ and for arbitrary $m \in \mathbb{N}$,*

$$-r_{n,m}^{(1)} + \sqrt{\frac{1-2r}{1-r}}\|\mathbb{E}(f - \mathbf{y} \mid \mathbf{z})\|_p \le \sqrt{\ell_n(f)} \le r_{n,m}^{(1)} + \sqrt{\frac{1}{1-r}}\|\mathbb{E}(f - \mathbf{y} \mid \mathbf{z})\|_p, \tag{30}$$

where $\ell_n(f)$ is defined in (25)*, and $f_m^\dagger \in \mathcal{H}$ is an approximation of $f^\dagger$ in $\mathcal{H}$ such that $\|f^\dagger - f_m^\dagger\| \le \tau_m$, and*

$$r_{n,m}^{(1)} \lesssim \left(\frac{L}{\sqrt{\bar{\nu}n}} + \sqrt{\bar{\nu}}\right)(\|f_h - f_m^\dagger\|_{\mathcal{H}} + 1) + \tau_m. \tag{31}$$

Proof. Define the random vectors

$$R := Y - f^\dagger(X), \;\; E := f^\dagger(X) - f_m^\dagger(X) - f_e(X),$$

so that $\mathbb{E}(R \mid \mathbf{Z}) = 0$, $\|E\|_\infty \le 2\tau_m$. Consider the decomposition

$$\begin{aligned}
\sqrt{\ell_n(f)} &= \left\|\hat{C}_{zz,\bar{\nu}}^{-1/2}\left(-\frac{S_z^*(R+E)}{n} + \hat{C}_{zx}(f_h - f_m^\dagger)\right)\right\|_{\mathcal{I}} \\
&\le \left\|\hat{C}_{zz,\bar{\nu}}^{-1/2}\frac{S_z^*(R+E)}{n}\right\|_{\mathcal{I}} + \left\|\hat{C}_{zz,\bar{\nu}}^{-1/2}\hat{C}_{zx}(f_h - f_m^\dagger)\right\|_{\mathcal{I}}, \qquad (32) \\
\sqrt{\ell_n(f)} &\ge -\left\|\hat{C}_{zz,\bar{\nu}}^{-1/2}\frac{S_z^*(R+E)}{n}\right\|_{\mathcal{I}} + \left\|\hat{C}_{zz,\bar{\nu}}^{-1/2}\hat{C}_{zx}(f_h - f_m^\dagger)\right\|_{\mathcal{I}}. \qquad (33)
\end{aligned}$$

On the event $B_n(r, L)$, we have

$$\left\|\hat{C}_{zz,\bar{\nu}}^{-1/2}\frac{S_z^*R}{n}\right\|_{\mathcal{I}} \le \bar{\nu}^{-1/2}\left\|\frac{S_z^*R}{n}\right\|_{\mathcal{I}} \le L(n\bar{\nu})^{-1/2}.$$

And since

$$\begin{aligned}
\left\|\hat{C}_{zz,\bar{\nu}}^{-1/2}\frac{S_z^*E}{n}\right\|_{\mathcal{I}}^2 &= \frac{1}{n^2}\left\langle S_z C_{zz,\bar{\nu}}^{-1}S_z^*E, E\right\rangle \le \frac{1}{n^2}\|S_z C_{zz,\bar{\nu}}^{-1}S_z^*\|\|E\|_2^2 \\
&= \underbrace{\|K_{zz}(K_{zz} + \bar{\nu}nI)^{-1}\|}_{\le 1} \cdot \frac{1}{n}\|E\|_2^2 \le 9\tau_m^2,
\end{aligned}$$

where the last inequality follows from the fact that $\|E\|_\infty \le \|E\| \le 3\tau_m$, we have

$$\left\|\hat{C}_{zz,\bar{\nu}}^{-1/2}\frac{S_z^*(R+E)}{n}\right\|_{\mathcal{I}} \le L(n\bar{\nu})^{-1/2} + 3\tau_m. \tag{34}$$

For the second term in (32) and (33), recall that by Lemma B.1 we have

$$\begin{aligned}
-r_n^{(0)} + \sqrt{\frac{1-2r}{1-r}}\|C_{zz}^{-1/2}C_{zx}(f_h - f_m^\dagger)\|_{\mathcal{I}} &\le \|\hat{C}_{zz,\bar{\nu}}^{-1/2}\hat{C}_{zx}(f_h - f_m^\dagger)\|_{\mathcal{I}} \\
&\le r_n^{(0)} + \sqrt{\frac{1}{1-r}}\|C_{zz}^{-1/2}C_{zx}(f_h - f_m^\dagger)\|_{\mathcal{I}},
\end{aligned}$$

where

$$r_n^{(0)} \lesssim \left(L(\bar{\nu}n)^{-1/2} + \bar{\nu}^{1/2}\right) \|f_h - f_m^\dagger\|_{\mathcal{H}}.$$

From the triangle inequality $||a| - |b|| \le |a - b|$, we have

$$\begin{aligned}
&|\|\mathbb{E}(f - \mathbf{y} \mid \mathbf{z})\|_p - \|C_{zz}^{-1/2} C_{zx}(f_h - f_m^\dagger)\|_{\mathcal{I}}| \\
&= |\|\mathbb{E}(f - f^\dagger \mid \mathbf{z})\|_p - \|\mathbb{E}(f_h - f_m^\dagger \mid \mathbf{z})\|_p| \\
&\le \|\mathbb{E}(f_e + f_m^\dagger - f^\dagger \mid \mathbf{z})\|_p \\
&\le \|f_e + f_m^\dagger - f^\dagger\|_\infty \le 3\tau_m.
\end{aligned}$$

Since $r \in (0, 1/2)$, we know $\sqrt{\frac{1}{1-r}} \le 2$ and $\sqrt{\frac{1-2r}{1-r}} \le 1$. Thus, we have

$$\sqrt{\frac{1-2r}{1-r}} \|\mathbb{E}(f - \mathbf{y} \mid \mathbf{z})\|_p - r_n^{(0)} - 4\tau_m \le \|\hat{C}_{zz,\bar{\nu}}^{-1/2} \hat{C}_{zx}(f_h - f_m^\dagger)\|_{\mathcal{I}} \tag{35}$$

$$\le \sqrt{\frac{1}{1-r}} \|\mathbb{E}(f - \mathbf{y} \mid \mathbf{z})\|_p + r_n^{(0)} + 6\tau_m. \tag{36}$$

Plugging (34), (35) and (36) to (32) and (33) completes the proof. □

B.2 Proof of Theorem 3.1

Let $\{m_n : n \in \mathbb{N}\}$ be an increasing sequence to be determined later. We drop the subscript n below for brevity. Let $\{\Theta_m : m \in \mathbb{N}\}$ be defined as in Theorem A.1 with $w_0 = f^\dagger$; recall that C_Θ can be set arbitrarily large. In the event $B_n(r, L)$ we fix $r = 1/3$ and determine L later; both parameters r, L will be dropped for brevity.

We define the unnormalized quasi-posterior measure as follows:

$$\tilde{\Pi}(A \mid \mathcal{D}^{(n)}) := \int_A \exp\left(-\frac{n}{2\lambda}\ell_n(f)\right) \Pi(df). \tag{37}$$

Consider the decomposition

$$\begin{aligned}
\mathbb{E}(\Pi(\mathrm{err}_{n,f} \mid \mathcal{D}^{(n)})) &\le \mathbb{E}(\Pi(\mathrm{err}_{n,f} \mid \mathcal{D}^{(n)}) \mid B_n) + (1 - \mathbb{P}(B_n)) \\
&\le \mathbb{E}\left(\frac{\tilde{\Pi}(\Theta_m^c \mid \mathcal{D}^{(n)}) + \tilde{\Pi}(\mathrm{err}_{n,f} \cap \Theta_m \mid \mathcal{D}^{(n)})}{\tilde{\Pi}(\Theta \mid \mathcal{D}^{(n)})} \mid B_n\right) + (1 - \mathbb{P}(B_n)) \\
&\le \mathbb{E}\left(\frac{\Pi(\Theta_m^c) + \tilde{\Pi}(\mathrm{err}_{n,f} \cap \Theta_m \mid \mathcal{D}^{(n)})}{\tilde{\Pi}(\Theta \mid \mathcal{D}^{(n)})} \mid B_n\right) + (1 - \mathbb{P}(B_n)) \\
&=: (\mathrm{I}) + (\mathrm{II}),
\end{aligned}$$

where the last inequality follows from $-\frac{n}{2\lambda}\ell_n(f) \le 0$ and the definition of $\tilde{\Pi}$ in (37).

By Assumption 3.2, we can find $f_m^\dagger \in \mathcal{H}$ such that $\|f^\dagger - f_m^\dagger\| \le \tau_m$ and

$$\|f_m^\dagger\|_{\mathcal{H}}^2 \le \inf_{h \in \mathcal{H}: \|h - f^\dagger\| \le \tau_m} \|h\|_{\mathcal{H}}^2 + 1 \le m\tau_m^2 + 1. \tag{38}$$

We first consider the denominator $\tilde{\Pi}(\Theta \mid \mathcal{D}^{(n)})$ in (I). For any $f \in \mathbb{B}$ with $\|f - f_m^\dagger\| \le 2\tau_m$, using Lemma B.2 with $f_h = f_m^\dagger$ and $f_e = f - f_m^\dagger$, we have the following on the event B_n:

$$\sqrt{\ell_n(f)} \le r_{n,m}^{(1)} + \sqrt{\frac{3}{2}} \|\mathbb{E}(f - \mathbf{y} \mid \mathbf{z})\|_p \le r_{n,m}^{(1)} + 4\tau_m,$$

where $r_{n,m}^{(1)}$ is defined in (31) and the last inequality follows from that

$$|\mathbb{E}(f - \mathbf{y} \mid \mathbf{z})| \le \mathbb{E}(|f - f_m^\dagger| \mid \mathbf{z}) + \mathbb{E}(|f_m^\dagger - f^\dagger| \mid \mathbf{z}) \le \|f - f_m^\dagger\| + \|f_m^\dagger - f^\dagger\| \le 3\tau_m.$$

Plugging $f_h - f_m^\dagger = 0$ into the definition of $r_{n,m}^{(1)}$ yields

$$\ell_n(f) \lesssim \frac{L^2}{\bar{\nu} n} + \bar{\nu} + \tau_m^2, \quad \text{if } B_n \text{ and } \|f - f_m^\dagger\| \le 2\tau_m \text{ hold.} \tag{39}$$

Thus, on the event B_n, we have for some fixed constant $C_1 > 0$,

$$\begin{aligned}
\tilde{\Pi}(\Theta \mid \mathcal{D}^{(n)}) &\ge \int_{\{f \in \Theta : \|f - f_m^\dagger\| \le 2\tau_m\}} \exp\left(-\frac{n}{2\lambda}\ell_n(f)\right) \Pi(df) \\
&\ge \Pi(\{\|f - f_m^\dagger\| \le 2\tau_m\}) \cdot \exp\left(-\frac{C_1 n}{\lambda}\left(\frac{L^2}{\bar{\nu} n} + \bar{\nu} + \tau_m^2\right)\right) \\
&\overset{(23)}{\ge} \exp\left(-3m\tau_m^2 - \frac{C_1 n}{\lambda}\left(\frac{L^2}{\bar{\nu} n} + \bar{\nu} + \tau_m^2\right)\right).
\end{aligned} \tag{40}$$

Now we consider the numerators in (I). First by Theorem A.1 we have $\Pi(\Theta_m^c) \le \exp(-C_\Theta m\tau_m^2)$, where C_Θ is any constant such that $e^{-C_\Theta m\tau_m^2} \le 1/2$, to be determined later. Thus,

$$\frac{\Pi(\Theta_m^c)}{\tilde{\Pi}_n(\Theta \mid \mathcal{D}^{(n)})} \le \exp\left(-\left(C_\Theta - 3 - \frac{C_1 n}{m\lambda}\right) m\tau_m^2 + \frac{C_1 n}{2\lambda}\left(\frac{L^2}{\bar{\nu} n} + \bar{\nu}\right)\right). \tag{41}$$

We now turn to the $\tilde{\Pi}(\mathrm{err}_{n,f} \cap \Theta_n)$ term in the numerators of (I). Noting that for non-negative numbers, $a \ge b - c$ implies $2a^2 \ge b^2 - 2c^2$, and by Lemma B.2, on the event B_n, for any $f \in \mathrm{err}_{n,f} \cap \Theta_m$ we have

$$\ell_n(f) \ge \frac{1}{4}\|\mathbb{E}(f - \mathbf{y} \mid \mathbf{z})\|_p^2 - \left(r_{n,m}^{(1)}\right)^2 \ge \frac{M\epsilon_n^2}{4} - \left(r_{n,m}^{(1)}\right)^2. \tag{42}$$

Recalling that when $f \in \Theta_m$, we can write $f = f_h + f_e$, where $f_h \in \underline{J}_m \mathcal{H}_1$ and $\|f_e\| \le \tau_n$. In view of (17), (18) and (38), we find $\|f_m^\dagger\|_{\mathcal{H}}^2 \le m\tau_m^2 + 1$. From Corollary A.1 and (31), we know

$$\begin{aligned}
\left(r_{n,m}^{(1)}\right)^2 &\lesssim \left(\frac{L^2}{\bar{\nu} n} + \bar{\nu}\right) \cdot (\|f_h\|_{\mathcal{H}}^2 + \|f_m^\dagger\|_{\mathcal{H}}^2 + 1) + \tau_m^2 \\
&\lesssim \left(\frac{L^2}{\bar{\nu} n} + \bar{\nu}\right) \cdot \left((C_\Theta + 1)m\tau_m^2 + 1\right) + \tau_m^2.
\end{aligned} \tag{43}$$

Combining (40), (42) and (43), we know there is a fixed constant $C_2 > C_1$ such that the following holds on the event B_n,

$$\frac{\tilde{\Pi}(\Theta_m \cap \mathrm{err}_{n,f} \mid \mathcal{D}^{(n)})}{\tilde{\Pi}(\Theta \mid \mathcal{D}^{(n)})} \le \exp\left(-\frac{Mn\epsilon_n^2}{8\lambda} + \Gamma_1 m\tau_m^2 + \Gamma_2\left(\frac{L^2}{\bar{\nu} n} + \bar{\nu}\right)\right), \tag{44}$$

where

$$\begin{aligned}
\Gamma_1 &:= 3 + \frac{C_2 n}{m\lambda}, \\
\Gamma_2 &:= 1 + (C_\Theta + 1)m\tau_m^2 + \frac{C_2 n}{2\lambda}.
\end{aligned}$$

Setting $C_\Theta = 4 + 2C_1$, $\epsilon_n = \tau_m$, $m = \lambda = \sqrt{n}$, $\bar{\nu} = L/\sqrt{n}$, $L = \min\{m\tau_m^2, \gamma_n\}$ where $\gamma_n \to \infty$ is a sequence with arbitrarily slow growth, we can verify that there exists an $M > 0$ such that both (41) and (44) converges to zero by noting that as $n \to \infty$, $m\tau_m^2 \to \infty, \tau_m^2 \to 0$. Hence, the term (I) converges to zero.

Next, we shall show that (II) tends to zero as $n \to \infty$. This is equivalent to verify that the right hand sides of (47), (48) and (49) tend to zero. Since $L \to \infty$, we know (48) and (49) will vanish. The following inequality shows that (47) also vanishes.

$$\bar{\nu} n - \log N(\bar{\nu}) = L\sqrt{n} - \log N(\bar{\nu}) \ge \sqrt{n} - \log O\left(\frac{\sqrt{n}}{L}\right) \to \infty,$$

where the inequality follows from the fact $N(\bar{\nu}) = O(\bar{\nu}^{-1})$ (see [65, Proposition 3]).

B.3 Proof of Proposition 3.1

We follow the choice of parameters (except for r, which will be set to $1/\max\{3, L\}$) as in Theorem 3.1 to show that

$$\Pi\left(\left\{f : \lim_{n\to\infty} \mathbb{P}_{\mathcal{D}^{(n)}}\left(\left|\sqrt{\ell_n(f)} - \|\mathbb{E}(f - \mathbf{y} \mid \mathbf{z})\|_p\right| > \delta\right) = 0,\ \forall \delta > 0\right\}\right) = 1.$$

From (17) and (18) we can see that for any τ_m that satisfies the condition of Theorem A.1, $\tilde{\tau}_m \geq \tau_m$ will also satisfy it. Thus we choose $\tilde{\tau}_m := \max\{\tau_m, \sqrt{2(C_\Theta m)^{-1}\log m}\}$ and define $\tilde{\Theta}_m$ accordingly. Then by Theorem 3.1 and the Borel-Cantelli Lemma, the set

$$S := \{f \in \mathbb{B} : \text{there exists } M_f > 0 \text{ such that } f \in \tilde{\Theta}_m \text{ for every } m > M_f\} \tag{45}$$

has prior probability 1 since $\sum_{m\geq 1} e^{-C_\Theta m \tilde{\tau}_m^2} \leq \sum_{m\geq 1} m^{-2} < \infty$.

For $f \in S$ and $m > M_f$, by Lemma B.2, we know the following holds on the event $B_n(r, L)$:

$$-r_{n,m}^{(1)} + \sqrt{\frac{1-2r}{1-r}}\|\mathbb{E}(f - \mathbf{y} \mid \mathbf{z})\|_p \leq \sqrt{\ell_n(f)} \leq r_{n,m}^{(1)} + \sqrt{\frac{1}{1-r}}\|\mathbb{E}(f - \mathbf{y} \mid \mathbf{z})\|_p,$$

with $m = \sqrt{n}, \bar{\nu} = L/\sqrt{n}$ as in Theorem A.1, the above becomes

$$r_{n,m}^{(1)} \lesssim \left(\frac{L}{\sqrt{\bar{\nu}n}} + \sqrt{\bar{\nu}}\right)(\|f_h - f_m^\dagger\|_{\mathcal{H}} + 1) + \tilde{\tau}_m \lesssim \sqrt{\frac{L}{n}}(\sqrt{m\tilde{\tau}_m^2} + 1) + \tilde{\tau}_m \lesssim \sqrt{L}\tilde{\tau}_m.$$

Since the growth of L can be arbitrarily slow, and $\tilde{\tau}_m \to 0$, we have $\lim_{n\to\infty} r_{n,m}^{(1)} = 0$. Note that $r := 1/\max\{3, L\} \to 0$, from (47), (48) and (49), it can be verified that $\lim_{n\to\infty} \mathbb{P}(B_n(r, L)) = 1$. Combining with the above inequality, we know that

$$\lim_{n\to\infty} \mathbb{P}_{\mathcal{D}^{(n)}}\left(\left|\sqrt{\ell_n(f)} - \|\mathbb{E}(f - \mathbf{y} \mid \mathbf{z})\|_p\right| > \delta\right) = 0, \quad \forall f \in S, \delta > 0.$$

B.4 Auxiliary Results

In this section, we collect several auxiliary results used in our proofs.

Lemma B.3. *For $r \in (0, 1)$, define*

$$A_n(r) := \{\|C_{zz,\bar{\nu}}^{-1/2}(\hat{C}_{zz} - C_{zz})C_{zz,\bar{\nu}}^{-1/2}\| \leq r\}. \tag{46}$$

Then when $\bar{\nu} \leq \sup_z k(z, z) =: \kappa^2$, and $r \geq \sqrt{\kappa^2/(\bar{\nu}n)} + \kappa^2/(3\bar{\nu}n)$, we have

$$1 - \mathbb{P}(A_n(r)) \leq 4N(\bar{\nu}) \exp\left(-\frac{\bar{\nu}nr^2}{2\kappa^2(1 + r/3)}\right), \tag{47}$$

where $N(\bar{\nu}) := \mathrm{Tr}(C_{zz}C_{zz,\bar{\nu}}^{-1})$ is the effective dimension of C_{zz}.

Proof. This is Lemma 1 in [33], with $\delta = 0$ (in their notation). □

The following lemma is a standard concentration result on the operator C_{zx}. See, e.g., Caponnetto and De Vito [65], Fukumizu [66], De Vito et al. [67]. We will give its proof for completeness.

Lemma B.4. *If $\sup_x k(x, x) \leq \kappa^2$ and $\sup_z k(z, z) \leq \kappa^2$, then for any $\delta \in (0, 1)$, we have for any constant $C > 0$,*

$$1 - \mathbb{P}\left(\|\hat{C}_{zx} - C_{zx}\| \leq \frac{C}{\sqrt{n}}\right) \leq 2\exp\left(-\frac{C}{4\kappa^2}\right). \tag{48}$$

Proof. Define the random variable $\xi := k(x, \cdot) \otimes k(z, \cdot)$. It is easy to verify that ξ is a Hilbert-Schmidt operator from $\mathcal{H}$ to $\mathcal{I}$, and $\mathbb{E}_{x,z}\xi = C_{zx}$. Note that $\|\xi\|_{\mathrm{HS}} = \sqrt{k(x, x)k(z, z)} \leq \kappa^2$ and $\mathbb{E}\|\xi\|_{\mathrm{HS}}^2 \leq \kappa^4$. From Proposition 2 in [65], we conclude that for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(\|\hat{C}_{zx} - C_{zx}\|_{\mathrm{HS}} \leq \frac{4\kappa^2}{\sqrt{n}} \log\frac{2}{\delta}\right) \geq 1 - \delta.$$

Finally, this lemma can be proved by a simple algebra and the fact that $\|\cdot\| \leq \|\cdot\|_{\mathrm{HS}}$. □

Lemma B.5. *Assume that $f^\dagger(x)-y$ is a Λ-subexponential random variable and $\sup_z k(z,z) \le \kappa^2$, then there exists a universal constant c_1 such that for all $C > 0$,*

$$1 - \mathbb{P}\left(\left\|\frac{S_z^*}{n}(f^\dagger(X) - Y)\right\|_{\mathcal{I}} \le \frac{C}{\sqrt{n}}\right) \le 2\exp\left(-\frac{C}{c_1\Lambda\kappa}\right). \tag{49}$$

Proof. Define the random variable $\xi := k(z,\cdot)(f^\dagger(x) - y)$. Since $f^\dagger(x) - y$ is Λ-subexponential, we know $\left(\mathbb{E}|f^\dagger(x) - y|^p\right)^{1/p} \le c_0\Lambda p$ for all $p \ge 1$ for some universal constant c_0 (See, e.g., Proposition 2.7.1 in Vershynin [68]). Recall that the Stirling's formula $\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n} \le n!$, we know $\mathbb{E}\|\xi\|_{\mathcal{I}}^n = \mathbb{E}k(z,z)^{\frac{n}{2}}|f^\dagger(x) - y|^n \le cn!(c\Lambda\kappa)^n$ for some universal constant c. Thus, from the fact that $\mathbb{E}\xi = \mathbb{E}(k(z,\cdot)\mathbb{E}(f^\dagger(x) - y \mid z)) = 0$ and Proposition 2 in [65], it has

$$\mathbb{P}\left(\left\|\frac{S_z^*}{n}(f^\dagger(X) - Y)\right\|_{\mathcal{I}} \le \frac{4c\kappa\Lambda}{\sqrt{n}}\log\frac{2}{\delta}\right) \ge 1 - \delta.$$

The final conclusion follows by a simple algebra. □

Lemma B.6. *On the event $A_n(r)$, we have*

$$\|C_{zz}^{1/2}\hat{C}_{zz,\bar{\nu}}^{-1}C_{zz}^{1/2}\| \le \|C_{zz,\bar{\nu}}^{1/2}\hat{C}_{zz,\bar{\nu}}^{-1}C_{zz,\bar{\nu}}^{1/2}\| \le \frac{1}{1-r}, \tag{50}$$

$$\|C_{zz,\bar{\nu}}^{1/2}(C_{zz,\bar{\nu}}^{-1} - \hat{C}_{zz,\bar{\nu}}^{-1})C_{zz,\bar{\nu}}^{1/2}\| \le \frac{r}{1-r}. \tag{51}$$

Proof. (50) is Eq. (19) in [33], with (in their notation) $z = 1$. For (51), note that

$$\begin{aligned}\|C_{zz,\bar{\nu}}^{1/2}(C_{zz,\bar{\nu}}^{-1} - \hat{C}_{zz,\bar{\nu}}^{-1})C_{zz,\bar{\nu}}^{1/2}\| &= \|I - C_{zz,\bar{\nu}}^{1/2}(C_{zz,\bar{\nu}} - (C_{zz} - \hat{C}_{zz}))^{-1}C_{zz,\bar{\nu}}^{1/2}\| \\ &= \|I - (I - C_{zz,\bar{\nu}}^{-1/2}(C_{zz} - \hat{C}_{zz})C_{zz,\bar{\nu}}^{-1/2})^{-1}\|.\end{aligned}$$

Define $D := C_{zz,\bar{\nu}}^{-1/2}(C_{zz} - \hat{C}_{zz})C_{zz,\bar{\nu}}^{-1/2}$. Then on the event $A_n(r)$, the right hand side above is

$$\|(I - D)^{-1} \cdot (-D)\| \le \|(I - D)^{-1}\|\|D\| \le \frac{1}{1-r} \cdot r,$$

where the last inequality uses the fact that $\|D\| \le r$ on $A(r)$, and that $\|(I - D)^{-1}\| \le (1 - \|D\|)^{-1}$. □

C Analysis of the Approximate Inference Algorithm

C.1 Proof of the Double Randomized Prior Trick

C.1.1 A Function-Space Equivalent to Proposition 4.1

We first claim that Proposition 4.1 is equivalent to the following function-space version, the proof of which is deferred to Section C.1.3:

Proposition C.1. *Let $\tilde{\mathcal{H}}, \tilde{\mathcal{I}}$ be finite-dimensional RKHSes with kernels k_x, k_z, respectively,*

$$g_0 \sim \mathcal{GP}(0, \lambda\nu^{-1}\tilde{k}_z),\ f_0 \sim \mathcal{GP}(0, \tilde{k}_x),\ \tilde{y}_i \sim \mathcal{N}(y_i, \lambda).$$

Then the optima f^* of

$$\min_{f\in\tilde{\mathcal{H}}} \max_{g\in\tilde{\mathcal{I}}} \mathcal{L}(f,g) := \sum_{i=1}^{n} \left((f(x_i) - \tilde{y}_i)g(z_i) - \frac{g(z_i)^2}{2} \right) - \frac{\nu}{2}\|g - g_0\|_{\tilde{\mathcal{I}}}^2 + \frac{\lambda}{2}\|f - f_0\|_{\tilde{\mathcal{H}}}^2 \tag{52}$$

follows the posterior distribution (6)*, with the kernels $\tilde{k}_x, \tilde{k}_z$.*

Proof of the equivalence. Observe that (52) is exactly the same as (14) when the random feature parameterization $\phi \mapsto g(z;\phi)$ is injective,[9] in which case we have $\|\phi\|_2 = \|g(\cdot;\phi)\|_{\tilde{\mathcal{I}}}$. Otherwise, observe that on the subspace

$$\Phi_s := \operatorname{span}\{\phi_{z,m}(z') : z' \in \mathcal{Z}\},$$

$\|\phi\|_2 = \|g(\cdot;\phi)\|_{\tilde{\mathcal{I}}}$ always holds: this follows by definition of $\tilde{k}_z$ when ϕ is a finite linear combination of the ϕ's, and the general case follows by continuity (note that $\tilde{\mathcal{I}}$ is already defined by $\tilde{k}_z$). Clearly any $g - g_0 \in \tilde{\mathcal{I}}$ can be parameterized with some ϕ in this subspace, so the optima of (52) is a valid candidate solution for (14). On the other hand, for any $\phi - \phi_0$ outside the aforementioned subspace, we have $\|\phi - \phi_0\|_2 > \|g(\cdot;\phi) - g(\cdot;\phi_0)\|_{\tilde{\mathcal{I}}}$. Therefore, the optimal ϕ of (14) must satisfy $\|\phi - \phi_0\|_2 = \|g(\cdot;\phi) - g(\cdot;\phi_0)\|_{\tilde{\mathcal{I}}}$, and thus solves (52). As a similar result also holds for f, we conclude that the two objectives are equivalent. □

Remark C.1. The non-injective setting above justifies the formal analysis of (16) in the main text. We also remark that any parameter θ, ϕ visited by the SGDA algorithm on (14) or (16) (starting from θ_0, ϕ_0) satisfies

$$\theta - \theta_0 \in \Theta_s,\ \ \phi - \phi_0 \in \Phi_s.$$

Thus $\|\phi - \phi_0\|_2 = \|g(\cdot;\phi) - g(\cdot;\phi_0\|_{\tilde{\mathcal{I}}}$ (and similarly for θ), and from the perspective of the SGDA algorithm, the objectives (52) and (14) are *always* the same. This can be proved by induction. Take ϕ for example; clearly $\phi = \phi_0$ satisfies the above. For ϕ_ℓ obtained at the ℓ-th step of SGDA, we have

$$\phi_\ell - \phi_0 = (1-\nu)(\phi_{\ell-1} - \phi_0) + V_\ell^\top \phi_{z,m}(Z),$$

where $V_\ell \in \mathbb{R}^n$ is independent of ϕ_ℓ. Thus $\phi_\ell - \phi_0 \in \Phi_s$ by definition of Φ_s and the inductive hypothesis.

C.1.2 Matrix Identities

We list two identities here that will be used in the derivations.

Lemma C.1. *Let U, C, V, S be operators between appropriate Banach spaces, $\lambda \in \mathbb{R} \setminus \{0\}$, then*

$$(\lambda I + UCV)^{-1} = \lambda^{-1}(I - U(\lambda C^{-1} + VU)^{-1}V), \tag{53}$$

$$S(S^*S + \lambda I)^{-1} = (SS^* + \lambda I)^{-1}S. \tag{54}$$

Proof. Recall the Woodbury identity:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

[9] Most random feature models, such as the random Fourier feature model, satisfies this property almost surely.

Then, we have

$$\begin{aligned}(\lambda I + UCV)^{-1} &= \lambda^{-1}I - \lambda^{-2}U(C^{-1} + \lambda^{-1}VU)^{-1}V \\ &= \lambda^{-1}(I - U(\lambda C^{-1} + VU)^{-1}V).\end{aligned}$$

And,

$$\begin{aligned}S(S^*S + \lambda I)^{-1} &= S(\lambda^{-1}I - \lambda^{-2}S^*(\lambda^{-1}SS^* + I)^{-1}S) \\ &= \lambda^{-1}(S - SS^*(SS^* + \lambda I)^{-1}S) \\ &= (SS^* + \lambda I)^{-1}S.\end{aligned}$$

□

C.1.3 Proof of Proposition C.1

Define $Y = (y_1, \ldots, y_n)$, $\tilde{Y} = (\tilde{y}_1, \ldots, \tilde{y}_n)$. We rewrite the objective as

$$\begin{aligned}\mathcal{L}(f, g) &= \left(\langle n\hat{C}_{zx}f - S_z^*\tilde{Y}, g\rangle_{\tilde{\mathcal{I}}} - \frac{1}{2}\langle n\hat{C}_{zz}g, g\rangle_{\tilde{\mathcal{I}}} - \frac{\nu}{2}\|g - g_0\|_{\tilde{\mathcal{I}}}^2\right) + \frac{\lambda}{2}\|f - f_0\|_{\tilde{\mathcal{H}}}^2 \\ &= n\left(\langle \hat{C}_{zx}f - n^{-1}S_z^*\tilde{Y}, g\rangle_{\tilde{\mathcal{I}}} - \frac{1}{2}\langle \hat{C}_{zz,\bar{\nu}}\, g, g\rangle_{\tilde{\mathcal{I}}} + \bar{\nu}\langle g, g_0\rangle_{\tilde{\mathcal{I}}} - \frac{\bar{\nu}}{2}\|g_0\|_{\tilde{\mathcal{I}}}^2\right) + \frac{\lambda}{2}\|f - f_0\|_{\tilde{\mathcal{H}}}^2,\end{aligned}$$

where $S_z, \hat{C}_{zx}, \hat{C}_{zz}$ are now defined w.r.t. the approximate kernels. The optimal g^* for fixed f is

$$g^*(f) = \hat{C}_{zz,\bar{\nu}}^{-1}(\hat{C}_{zx}f - n^{-1}S_z^*\tilde{Y} + \bar{\nu}g_0). \tag{55}$$

Plugging g^* back to the objective, we have

$$\begin{aligned}\mathcal{L}(f, g^*(f)) &= \frac{n}{2}\langle g^*, \hat{C}_{zz,\bar{\nu}}g^*\rangle_{\tilde{\mathcal{I}}} + \frac{\lambda}{2}\|f - f_0\|_{\tilde{\mathcal{H}}}^2 - \frac{n\bar{\nu}}{2}\|g_0\|_{\tilde{\mathcal{I}}}^2, \\ \partial_f\mathcal{L} &= n\hat{C}_{xz}\hat{C}_{zz,\bar{\nu}}^{-1}\hat{C}_{zz,\bar{\nu}}g^* + \lambda(f - f_0) \\ &= n\hat{C}_{xz}\hat{C}_{zz,\bar{\nu}}^{-1}(\hat{C}_{zx}f - n^{-1}S_z^*\tilde{Y} + \bar{\nu}g_0) + \lambda(f - f_0).\end{aligned}$$

Setting $\partial_f\mathcal{L}$ to zero, we obtain

$$f^* = (n\hat{C}_{xz}\hat{C}_{zz,\bar{\nu}}^{-1}\hat{C}_{zx} + \lambda I)^{-1}(n\hat{C}_{xz}\hat{C}_{zz,\bar{\nu}}^{-1}(n^{-1}S_z^*\tilde{Y} - \bar{\nu}g_0) + \lambda f_0). \tag{56}$$

Since

$$\begin{aligned}(n\hat{C}_{xz}\hat{C}_{zz,\bar{\nu}}^{-1}\hat{C}_{zx} + \lambda I)^{-1} &= (n^{-1}S_x^*S_z\hat{C}_{zz,\bar{\nu}}^{-1}S_z^*S_x + \lambda I)^{-1} \\ &= (S_x^*LS_x + \lambda I)^{-1} \\ &\overset{(53)}{=} \lambda^{-1}(\underbrace{I - S_x^*(\lambda L^{-1} + S_xS_x^*)^{-1}S_x}_{\text{defined as } \mathcal{C}}),\end{aligned} \tag{57}$$

we can rewrite f^* as

$$f^* = \lambda^{-1}\mathcal{C}(\hat{C}_{xz}\hat{C}_{zz,\bar{\nu}}^{-1}(S_z^*\tilde{Y} - \nu g_0) + \lambda f_0).$$

Clearly, f^* is a Gaussian process. Suppose $f^*(x_*) \sim \mathcal{N}(S_*\mu', S_*\mathcal{C}'S_*^*)$, then

$$\begin{aligned}\mu' = \lambda^{-1}\mathcal{C}\, n\hat{C}_{xz}\hat{C}_{zz,\bar{\nu}}^{-1}(n^{-1}S_z^*Y) &= \lambda^{-1}(I - S_x^*(\lambda L^{-1} + S_xS_x^*)^{-1}S_x)S_x^*LY \\ &= \lambda^{-1}S_x^*(I - (\lambda L^{-1} + S_xS_x^*)^{-1}S_xS_x^*)LY \\ &= S_x^*(\lambda L^{-1} + S_xS_x^*)^{-1}Y.\end{aligned}$$

The RHS above matches the posterior mean (9) (with k_x, k_z replaced by their random feature approximations) since $S_xS_x^* = K_{xx}$ and

$$S_*\mu' = S_*S_x^*(\lambda L^{-1} + S_xS_x^*)^{-1}Y = K_{*x}(\lambda L^{-1} + K_{xx})^{-1}Y = K_{*x}(\lambda + LK_{xx})^{-1}LY.$$

As $\tilde{Y} - Y, g_0$ and f_0 are independent, the covariance operator of f^* is

$$\begin{aligned}\mathcal{C}' &= \lambda^{-1}\mathcal{C}(\hat{C}_{xz}\hat{C}_{zz,\bar{\nu}}^{-1}(n\lambda\hat{C}_{zz} + \lambda\nu I)\hat{C}_{zz,\bar{\nu}}^{-1}\hat{C}_{zx} + \lambda^2 I)\lambda^{-1}\mathcal{C} \\ &= \lambda^{-1}\mathcal{C}(\lambda n\hat{C}_{xz}\hat{C}_{zz,\bar{\nu}}^{-1}\hat{C}_{zx} + \lambda^2 I)\lambda^{-1}\mathcal{C} \overset{(57)}{=} \mathcal{C}.\end{aligned}$$

In view of (57), we know

$$\begin{aligned}S_*\mathcal{C}'S_*^* &= S_*S_*^* - S_*S_x^*(\lambda L^{-1} + S_xS_x^*)^{-1}S_xS_*^* \\ &= K_{**} - K_{*x}(\lambda L^{-1} + K_{xx})^{-1}K_{x*},\end{aligned}$$

which matches the posterior covariance matrix (10) with replaced kernels.

C.1.4 Discussion of KernelIV [7]

The KernelIV estimator [7] is motivated as a kernelized generalization for 2SLS. Its *first stage* consists of estimating the conditional expectation operator E, restricted on $\mathcal{H}$; we can see from Theorem 1 therein that their estimator E^n_λ coincides with our choice of $\hat{E}_n = \hat{C}^{-1}_{zz,\bar{\nu}}\hat{C}_{zx}$. Thus when the domain of the response variable $\mathcal{Y} = \mathbb{R}$, their second-stage objective reduces to

$$\begin{aligned}\hat{\mathcal{E}}_n(f) &:= \frac{1}{n}\sum_{i=1}^n (\tilde{y}_i - \langle f, \hat{E}^*_n k(\tilde{z}_i, \cdot)\rangle_{\mathcal{H}})^2 + \bar{\lambda}\|f\|^2_{\mathcal{H}} \\ &\equiv \langle f, (\hat{C}_{xz}\hat{C}^{-1}_{zz,\bar{\nu}}\hat{C}_{zz}\hat{C}^{-1}_{zz,\bar{\nu}}\hat{C}_{zx} + \bar{\lambda}I)f\rangle_{\mathcal{H}} + \left\langle f, \hat{C}_{xz}\hat{C}^{-1}_{zz,\bar{\nu}}\frac{S^*_{\tilde{z}}\tilde{Y}}{n}\right\rangle_{\mathcal{H}} + \bar{\lambda}\|f\|^2_{\mathcal{H}} \end{aligned} \tag{58}$$

where in the last equality we have dropped the quadratic term about $\tilde{Y}$ as it is independent of f. Comparing with the kernelized DualIV objective (3), (58) is only different in their use of separate samples $(\tilde{z}_i, \tilde{y}_i)$,[10] and the replacement of $\hat{C}^{-1}_{zz,\bar{\nu}}$ in (3) with the asymptotically equivalent $\hat{C}^{-1}_{zz,\bar{\nu}}\hat{C}_{zz}\hat{C}^{-1}_{zz,\bar{\nu}}$. The similarity between the two objectives is also supported by previous report that empirically, the resulted estimators perform similarly [20].

(58) has an optimization-based equivalent form, similar to (4) to (3). Indeed, using a similar argument to Appendix C.1.3, we can see that

$$\langle f, \hat{C}_{xz}\hat{C}^{-1}_{zz,\bar{\nu}}\hat{C}_{zz}\hat{C}^{-1}_{zz,\bar{\nu}}\hat{C}_{zx}f\rangle_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^n (2g(\tilde{z}_i)f(\tilde{x}_i) - g(\tilde{z}_i)^2) - 2\bar{\nu}\|g\|^2_{\mathcal{I}},$$

where $g = \hat{C}^{-1}_{zz,\bar{\nu}}\hat{C}_{zx}f$ solves

$$\max_{g\in\mathcal{I}} \frac{1}{n}\sum_{i=1}^n (2g(\tilde{z}_i)f(\tilde{x}_i) - g(\tilde{z}_i)^2) - \bar{\nu}\|g\|^2_{\mathcal{I}} \tag{59}$$

which is equivalent to the KRR objective. Following this we can see that

$$\hat{\mathcal{E}}_n(f) \equiv \frac{1}{n}\sum_{i=1}^n (2g(\tilde{z}_i)f(\tilde{x}_i) - g(\tilde{z}_i)^2 + f(\tilde{x}_i)h(\tilde{z}_i)) - 2\bar{\nu}\|g\|^2_{\mathcal{I}} + \bar{\lambda}\|f\|^2_{\mathcal{H}}, \tag{60}$$

where $h = \hat{C}^{-1}_{zz,\bar{\nu}}\frac{S^*_{\tilde{z}}\tilde{Y}}{n}$ represents $\hat{b}_n$ in (2). However, note the different regularizers on g in (60) and (59) above, which is due to the replacement of $\hat{C}^{-1}_{zz,\bar{\nu}}$ with $\hat{C}^{-1}_{zz,\bar{\nu}}\hat{C}_{zz}\hat{C}^{-1}_{zz,\bar{\nu}}$ in (58); consequently, the objective $\hat{\mathcal{E}}_n$ no longer has a minimax formulation, and it is less clear whether a GDA-like algorithm will converge to the expected optima.

Finally, we note that Mastouri et al. [56] provides additional discussions on the difference between the kernelIV estimator and the kernelized dualIV estimator.

C.2 Assumptions used in Proposition 4.2

The analysis in the subsequent subsections relies on the following assumptions on the random feature expansion. We only state them for x for conciseness; the requirements for z are similar.

The following assumption holds for, e.g., random Fourier features [48].

Assumption C.1.

$$\sup_{x,x'\in\mathcal{X}} \left|k_x(x,x') - \tilde{k}_{x,m}(x,x')\right| \xrightarrow{p} 0, \quad \textit{as } m\to\infty, \tag{61}$$

The following assumption may be relaxed to require $\sup_x \tilde{k}_{x,m}(x,x)$ to have finite higher-order moments; we use this for simplicity.

Assumption C.2. *There exists a constant $\tilde{\kappa} > 0$ such that* $\max_{m\in\mathbb{N}} \sup_{x\in\mathcal{X}} \tilde{k}_{x,m}(x,x) \le \tilde{\kappa}$.

[10] Note that $\tilde{y}_i$ here refers to the separate batch of unperturbed samples (see [7]), as opposed to the perturbed samples in the main text; we also assume that the two set of samples have the same sample size for simplicity.

C.3 Analysis of Random Feature Approximation

We recall the following facts: for $A, B \in \mathbb{R}^{n\times n}$,

$$\|A\| \le \|A\|_F \le \sqrt{n}\|A\|, \;\; A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}.$$

Lemma C.2. *For all $m \in \mathbb{N}$, let $k_{x,m}$ be a random feature approximation to k_x such that* (61) *holds, and let $\tilde{k}_{z,m}$ be an approximation to k_z satisfying a similar requirement as above. Then the random feature-approximated posterior $\Pi_m(f(x_*) \mid \mathcal{D}^{(n)}) = \mathcal{N}(\tilde{\mu}, \tilde{S})$ satisfies*

$$\lim_{m\to\infty} \sup_{x^*\in\mathcal{X}^l} \|\mu - \tilde{\mu}\|_2 = 0, \quad \lim_{m\to\infty} \sup_{x^*\in\mathcal{X}^l} \|\tilde{S} - S\|_F = 0,$$

for any fixed training data (X, Y, Z), $l \in \mathbb{N}$, and $\lambda, \nu > 0$. In the above, $\tilde{\mu}$ and $\tilde{S}$ are defined as

$$\begin{aligned}
\tilde{\mu} &= \tilde{K}_{*x}(\lambda I + \tilde{L}\tilde{K}_{xx})^{-1}\tilde{L}Y, \\
\tilde{S} &= \tilde{K}_{**} - \tilde{K}_{*x}\tilde{L}(\lambda I + \tilde{K}_{xx}\tilde{L})^{-1}\tilde{K}_{x*}, \\
\tilde{L} &= \tilde{K}_{zz}(\tilde{K}_{zz} + \nu I)^{-1},
\end{aligned}$$

and the Gram matrices are defined using $\tilde{k}_{x,m}$ and $\tilde{k}_{z,m}$.

Proof. Define

$$\epsilon_m = \max\left(\sup_{x,x'\in\mathcal{X}} \left| k(x,x') - \tilde{k}_{x,m}(x,x') \right|, \sup_{z,z'\in\mathcal{Z}} \left| k(z,z') - \tilde{k}_{z,m}(z,z') \right| \right).$$

By assumption $\epsilon_m \xrightarrow{p} 0$. For $\tilde{S}$ we consider the decomposition

$$\begin{aligned}
\|\tilde{S} - S\| &\le \|\tilde{K}_{**} - K_{**}\| \\
&\quad + \|\tilde{K}_{*x} - K_{*x}\|\|\tilde{L}\|\|(\lambda I + \tilde{K}_{xx}\tilde{L})^{-1}\tilde{K}_{x*}\| \\
&\quad + \|K_{*x}\|\|\tilde{L} - L\|\|(\lambda I + \tilde{K}_{xx}\tilde{L})^{-1}\tilde{K}_{x*}\| \\
&\quad + \|K_{*x}L\|\|(\lambda I + \tilde{K}_{xx}\tilde{L})^{-1} - (\lambda I + K_{xx}L)^{-1}\|\|\tilde{K}_{x*}\| \\
&\quad + \|K_{*x}L(\lambda I + K_{xx}L)^{-1}\|\|\tilde{K}_{x*} - K_{x*}\| \\
&=: \text{(I)} + \text{(II)} + \text{(III)} + \text{(IV)} + \text{(V)}.
\end{aligned}$$

In the following, we use $O(\cdot)$ and $O_p(\cdot)$ to represent the asymptotic behaviour when $m \to \infty$. Since n and l are fixed, the operator norms of the matrices K_{*x}, L, K_{xx} are $O(1)$. Observe that $\|K_{zz} - \tilde{K}_{zz}\| \le \sqrt{n}\epsilon_m$. By the triangle inequality, the inequality $\|\cdot\| \le \|\cdot\|_F$ and the boundedness of $\tilde{k}_{x,m}$ and $\tilde{k}_{z,m}$, we have $\|\tilde{K}_{*x}\| = O(1)$. Both $O(\cdot)$ terms above are independent of x^*. Finally, recall that $\|L\| = \|K_{zz}(K_{zz} + \nu I)^{-1}\| \le 1$ and similarly $\|\tilde{L}\| \le 1$. Using these facts, we have

$$\begin{aligned}
\text{(I)} &\le \|\tilde{K}_{**} - K_{**}\|_F \le l\epsilon_m \to 0. \\
\text{(II)} &\le \sqrt{ln}\epsilon_m \cdot 1 \cdot \lambda^{-1} \cdot O(1) \to 0. \\
\|\tilde{L} - L\| &= \|K_{zz}(K_{zz} + \nu I)^{-1} - \tilde{K}_{zz}(\tilde{K}_{zz} + \nu I)^{-1}\| \to 0. \\
&\le \|K_{zz} - \tilde{K}_{zz}\| \cdot \nu^{-1} + \|\tilde{K}_{zz}(\tilde{K}_{zz} + \nu I)^{-1}\|\|(K_{zz} - \tilde{K}_{zz})(K_{zz} + \nu I)^{-1}\| \\
&\le 2\sqrt{n}\epsilon_m \cdot \nu^{-1} \to 0. \\
\text{(III)} &\le O(1) \cdot \|\tilde{L} - L\| \cdot \lambda^{-1}O(1) \to 0 \\
\text{(IV)} &= O(1) \cdot \|(\lambda I + \tilde{K}_{xx}\tilde{L})^{-1}\|\|\tilde{K}_{xx}\tilde{L} - K_{xx}L\|\|(\lambda I + K_{xx}L)^{-1}\| \\
&\le O(1) \cdot \lambda^{-2} \cdot (\|\tilde{K}_{xx} - K_{xx}\|\|\tilde{L}\| + \|K_{xx}\|\|\tilde{L} - L\|) \to 0. \\
\text{(V)} &= O(1) \cdot \sqrt{ln}\epsilon_m \to 0.
\end{aligned}$$

Moreover, the converges above are all independent of the choice of x^*. Thus we have

$$\sup_{x^*\in\mathcal{X}^l} \|\tilde{S} - S\|_F \le l \sup_{x^*\in\mathcal{X}^l} \|\tilde{S} - S\| \to 0.$$

Using a similar argument we have

$$\sup_{x^*\in\mathcal{X}^l} \|\tilde{\mu} - \mu\|_2 \to 0.$$

□

C.4 Analysis of the Optimization Algorithm

Algorithm 1: Modified randomized prior algorithm for approximate inference.

Input: Hyperparameters $\nu, \lambda \in \mathbb{R}$. Random feature models $\theta \mapsto f(\cdot;\theta)$, $\varphi \mapsto g(\cdot;\varphi)$.
Result: A single sample from the approximate posterior
Initialize: draw $\theta_0 \sim \mathcal{N}(0, I), \varphi_0 \sim \mathcal{N}(0, \lambda\nu^{-1}I), \tilde{Y} \sim \mathcal{N}(Y, \lambda I)$;
for $\ell \leftarrow 1, \ldots, L-1$ **do**
  $\hat{\theta}_\ell \leftarrow \theta_{\ell-1} - \eta_\ell \hat{\nabla}_\theta \mathcal{L}_{\mathrm{rf}}(\theta_{\ell-1}, \varphi_{\ell-1}, \theta_0, \varphi_0)$;
  $\hat{\varphi}_\ell \leftarrow \varphi_{\ell-1} + \eta_\ell \hat{\nabla}_\varphi \mathcal{L}_{\mathrm{rf}}(\theta_{\ell-1}, \varphi_{\ell-1}, \theta_0, \varphi_0)$;
  $\theta_{\ell+1} \leftarrow \mathrm{Proj}_{B_f}(\hat{\theta}_\ell)$;
  $\varphi_{\ell+1} \leftarrow \mathrm{Proj}_{B_g}(\hat{\varphi}_\ell)$;
end
return $f(\cdot;\theta_L)$

For the purpose of the analysis we consider the standard SGDA algorithm as outlined in Algorithm 1. In the algorithm $\mathcal{L}_{\mathrm{rf}}$ denotes the objective in (14), and Proj_B denotes the projection into the ℓ_2-norm ball with radius B, and $\hat{\nabla}\mathcal{L}_{\mathrm{rf}}$ represents a stochastic (unbiased) approximation of the gradient $\nabla\mathcal{L}_{\mathrm{rf}}$. In the following, we will suppress the dependency of $\mathcal{L}_{\mathrm{rf}}$ on θ_0, φ_0 for simplicity.

Concretely, we introduce the notations

$$\Phi_f := \frac{1}{\sqrt{m}} \begin{bmatrix} \phi_{x,m}(x_1)^\top \\ \vdots \\ \phi_{x,m}(x_n)^\top \end{bmatrix} \in \mathbb{R}^{n\times m}, \quad \Phi_g := \frac{1}{\sqrt{m}} \begin{bmatrix} \phi_{z,m}(z_1)^\top \\ \vdots \\ \phi_{z,m}(z_n)^\top \end{bmatrix} \in \mathbb{R}^{n\times m},$$

where we recall $X := (x_1, \ldots, x_n)$ and $Z := (z_1, \ldots, z_n)$ are the training data.

Observe that $\Phi_f\theta = f(X;\theta), \Phi_g\varphi = g(Z;\varphi)$, we can rewrite the objective (14) as

$$\mathcal{L}_{\mathrm{rf}}(\theta,\varphi) = \theta^\top\Phi_f^\top\Phi_g\varphi - \tilde{Y}^\top\Phi_g\varphi - \frac{1}{2}\varphi^\top\Phi_g^\top\Phi_g\varphi - \frac{\nu}{2}\|\varphi-\varphi_0\|_2^2 + \frac{\lambda}{2}\|\theta-\theta_0\|_2^2. \tag{62}$$

We additionally define

$$\mathcal{L}_i(\theta,\varphi) = n\left(\theta^\top\Phi_f^\top E_i\Phi_g\varphi - \tilde{Y}^\top E_i\Phi_g\varphi - \frac{1}{2}\varphi^\top\Phi_g^\top E_i\Phi_g\varphi\right) - \frac{\nu}{2}\|\varphi-\varphi_0\|_2^2 + \frac{\lambda}{2}\|\theta-\theta_0\|_2^2,$$

where $E_i := e_i e_i^\top$ and $\{e_i\}_{i\in[n]}$ is the standard orthogonal basis of $\mathbb{R}^n$. We can see that

$$\mathcal{L}_{\mathrm{rf}}(\theta,\varphi) = \frac{1}{n}\sum_{i\in[n]}\mathcal{L}_i(\theta,\varphi).$$

Therefore, the stochastic gradient in Algorithm 1 can be defined as

$$\hat{\nabla}\mathcal{L}_{\mathrm{rf}}(\theta,\varphi) := \nabla\mathcal{L}_{\mathcal{I}}(\theta,\varphi) = \sum_{i\in[n]}\nabla\mathcal{L}_i(\theta,\varphi)\mathbf{1}_{i=\mathcal{I}}, \tag{63}$$

where $\mathcal{I}$ is a random variable sampled from the uniform distribution of the set $[n]$.

In practice we run the algorithm concurrently on J sets of parameters, starting from independent draws of initial conditions $\{\theta_0^{(j)}, \phi_0^{(j)}\}$; moreover, the projection is not implemented, and there are various other modifications to further improve stability, as described in Appendix D.2.

The following lemma is a convergence theorem of Algorithm 1 under the choice of stochastic gradient defined in (63).

Lemma C.3. *Fix an $m \in \mathbb{N}$. Denote by θ^* the optima of* (14) *and take $\eta_\ell := \frac{1}{\mu(\ell+1)}$ with $\mu = \min\{\lambda,\nu\}$. Then for any $\epsilon, B_1, B_2, B_3 > 0$, there exist $B_f, B_g > 0$ such that when $L = \Omega(\delta^{-1}\epsilon^{-2})$, the approximate optima θ_L returned by Algorithm 1 satisfies*

$$\mathbb{P}\left(\{\|\theta_L - \theta^*\|_2 > \epsilon\} \cap E_n\right) \le \delta,$$

where

$$E_n := \left\{ \|\theta_0\|_2 + \|\varphi_0\|_2 \le B_1, \|\tilde{Y}\|_2 \le B_2, \sup_{z\in\mathcal{Z}} \tilde{k}_{z,m}(z,z) + \sup_{x\in\mathcal{X}} \tilde{k}_{x,m}(x,x) \le B_3 \right\},$$

and $\tilde{k}_{\cdot,m}$ denotes the random feature-approximated kernel. The randomness in the statement above is from the sampling of the initial values θ_0, φ_0, the gradient noise.

Proof. Recall from (56) that θ^* is a sum of bounded linear transforms of θ_0, φ_0 and $\tilde{Y}_0$. Thus on the event E_n the norm of the optima $\|\theta^*\|_2$ is bounded. Similarly, $\|\varphi^*\|_2$ is also bounded on E_n by (55). We choose B_f and B_g to be their maximum values on the event E_n.

Notice that $\mathcal{L}_{\mathrm{rf}}$ is strongly-convex in θ, and strongly-concave in φ, so it has the unique stationary point (θ^*, φ^*). We will then bound $\|\theta_\ell - \theta^*\|_2^2 + \|\varphi_\ell - \varphi^*\|_2^2$. Let σ_f, σ_g be the minimal constants such that $\|\nabla_\theta \mathcal{L}_i(\theta,\varphi)\|_2^2 \le \sigma_f^2$, $\|\nabla_\varphi \mathcal{L}_i(\theta,\varphi)\|_2^2 \le \sigma_g^2$ for all $i \in [n]$, $\|\theta\|_2 \le B_f$ and $\|\varphi\|_2 \le B_g$. Introducing the notation $B := \max\{B_f, B_g\}$, so we have $\|\theta\|_2, \|\varphi\|_2 \le B$. Define

$$r_\ell = \mathbb{E}\left[\|\theta_\ell - \theta^*\|_2^2 + \|\varphi_\ell - \varphi^*\|_2^2\right].$$

We want to know how r_ℓ contracts. We first make a stochastic gradient step on θ_ℓ with step size η_ℓ, i.e., $\hat{\theta}_{\ell+1} := \theta_\ell - \eta_\ell \hat{\nabla}_\theta \mathcal{L}_{\mathrm{rf}}(\theta_\ell, \varphi_\ell)$ with $\hat{\nabla}\mathcal{L}_{\mathrm{rf}}$ defined in (63). Then,

$$\mathbb{E}[\|\hat{\theta}_{\ell+1} - \theta^*\|_2^2 \mid \theta_\ell, \varphi_\ell] \le \|\theta_\ell - \theta^*\|_2^2 - 2\eta_\ell \langle \theta_\ell - \theta^*, \nabla_\theta \mathcal{L}(\theta_\ell, \varphi_\ell)\rangle + \eta_\ell^2 \sigma_f^2,$$

where the expectation is taken with respect to the randomness of the gradient. For the above inner product term, we have that

$$\begin{aligned}\langle \theta_\ell - \theta^*, \nabla_\theta \mathcal{L}_{\mathrm{rf}}(\theta_\ell, \varphi_\ell)\rangle &= \langle \theta_\ell - \theta^*, \nabla_\theta \mathcal{L}_{\mathrm{rf}}(\theta_\ell, \varphi_\ell) - \nabla_\theta \mathcal{L}_{\mathrm{rf}}(\theta^*, \varphi^*)\rangle \\ &= \lambda \|\theta_\ell - \theta^*\|_2^2 + \langle \theta_\ell - \theta^*, \Phi_f^\top \Phi_g (\varphi_\ell - \varphi^*)\rangle.\end{aligned}$$

Next, we consider the gradient step on φ_ℓ with step size η_ℓ, i.e., $\hat{\varphi}_{\ell+1} := \varphi_\ell + \eta_\ell \hat{\nabla}_\varphi \mathcal{L}_{\mathrm{rf}}(\theta_\ell, \varphi_\ell)$. Then, we have that

$$\mathbb{E}[\|\hat{\varphi}_{\ell+1} - \varphi^*\|_2^2 \mid \theta_\ell, \varphi_\ell] \le \|\varphi_\ell - \varphi^*\|_2^2 + 2\eta_\ell \langle \varphi_\ell - \varphi^*, \nabla_\varphi \mathcal{L}_{\mathrm{rf}}(\theta_\ell, \varphi_\ell)\rangle + \eta_\ell^2 \sigma_g^2.$$

We similarly deal with the inner product term:

$$\begin{aligned}\langle \varphi_\ell - \varphi^*, \nabla_\varphi \mathcal{L}_{\mathrm{rf}}(\theta_\ell, \varphi_\ell)\rangle &= \langle \varphi_\ell - \varphi^*, \nabla_\varphi \mathcal{L}_{\mathrm{rf}}(\theta_\ell, \varphi_\ell) - \nabla_\varphi \mathcal{L}_{\mathrm{rf}}(\theta^*, \varphi^*)\rangle \\ &= -\langle \varphi_\ell - \varphi^*, (\Phi_g^\top \Phi_g + \nu I)(\varphi_\ell - \varphi^*)\rangle + \langle \varphi_\ell - \varphi^*, \Phi_g^\top \Phi_f (\theta_\ell - \theta^*)\rangle \\ &\le -\nu \|\varphi_\ell - \varphi^*\|_2^2 + \langle \varphi_\ell - \varphi^*, \Phi_g^\top \Phi_f (\theta_\ell - \theta^*)\rangle,\end{aligned}$$

Combining the above results, we have

$$r_{\ell+1} \le \mathbb{E}[\|\hat{\theta}_{\ell+1} - \theta^*\|_2^2 + \|\hat{\varphi}_{\ell+1} - \varphi^*\|_2^2 \mid \theta_\ell, \varphi_\ell] \le (1 - 2\mu\eta_\ell) r_\ell + \eta_\ell^2(\sigma_f^2 + \sigma_g^2),$$

where we have set $\mu := \min\{\nu, \lambda\}$, and the first inequality follows from the fact that the projection onto a convex set is a contraction map, i.e., $\|\mathrm{Proj}_B(x) - \mathrm{Proj}_B(y)\| \le \|x - y\|$.

Let $\sigma^2 = \sigma_f^2 + \sigma_g^2$ and $\eta_\ell = \frac{\xi}{\ell+1}$ for some $\xi > \frac{1}{2\mu}$, by induction we have

$$r_\ell \le \frac{c_\xi}{\ell + 1}, \quad \text{where } c_\xi = \max\left\{ r_0, \frac{2\xi^2\sigma^2}{2\mu\xi - 1} \right\}.$$

Specifically, taking $\xi = \mu^{-1}$, we have

$$r_\ell \le \frac{1}{\ell+1} \max\left\{ r_0, \frac{2\sigma^2}{\mu^2} \right\}. \tag{64}$$

We now track the constants we have used in (64). Note that on the event E_n,

$$r_0 \le 2\left(\|\theta_0\|_2^2 + \|\theta^*\|_2^2 + \|\varphi_0\|_2^2 + \|\varphi^*\|_2^2\right) \le 4(B_1^2 + B^2).$$

Recall that the definition of σ^2 is

$$\sigma^2 = \max_{i\in[n], \|\theta\|_2, \|\varphi\|_2 \le B} \|\nabla_\theta \mathcal{L}_i(\theta,\varphi)\|_2^2 + \max_{i\in[n], \|\theta\|_2, \|\varphi\|_2 \le B} \|\nabla_\varphi \mathcal{L}_i(\theta,\varphi)\|_2^2 =: \mathrm{(I)} + \mathrm{(II)}.$$

For the first term, we have

$$\begin{aligned}(\mathrm{I}) &= \max_{i\in[n],\|\theta\|_2,\|\varphi\|_2\le B} \|\lambda(\theta-\theta_0) + n\Phi_f^\top E_i \Phi_g \varphi\|_2^2 \\ &\le \max_{i\in[n],\|\theta\|_2,\|\varphi\|_2\le B} \left(2\lambda^2\|\theta-\theta_0\|_2^2 + 2n^2\|\Phi_f^\top E_i\Phi_g\varphi\|_2^2\right) \\ &\le 4\lambda^2(B^2+B_1^2) + 2n^2B_3^2B^2.\end{aligned}$$

Similarly, for the second term, we have

$$\begin{aligned}(\mathrm{II}) &= \max_{i\in[n],\|\theta\|_2,\|\varphi\|_2\le B} \|n(\theta^\top\Phi_f^\top E_i\Phi_g - \tilde{Y}^\top E_i\Phi_g - \Phi_g^\top E_i\Phi_g\varphi) - \nu(\varphi-\varphi_0)\|_2^2 \\ &\le 4n^2B_3^2B^2 + 2n^2B_2^2B^2 + 4\nu^2(B^2+B_1^2).\end{aligned}$$

Thus, we know that

$$\sigma^2 \le 8(\lambda^2+\nu^2)(B^2+B_1^2) + 6n^2B_3^2B^2 + 2n^2B_2^2B^2 =: \tilde{C}.$$

Taking $L_\delta = \delta^{-1}\epsilon^{-2}\max\{4B_1^2+4B^2, \tilde{C}\mu^{-1}\}$ and $\eta_\ell = \frac{1}{\mu(\ell+1)}$, by (64), we know that

$$\mathbb{P}(\|\theta_L-\theta^*\|_2 > \epsilon) \le \epsilon^{-2}\mathbb{E}\|\theta_L-\theta^*\|_2^2 \le \epsilon^{-2}r_\ell \le \delta.$$

□

C.5 Proof of Proposition 4.2

By Lemma C.2, for any $\epsilon_1 > 0$ we have

$$\lim_{m\to\infty}\mathbb{P}\left(\left\{\sup_{x^*\in\mathcal{X}^l}\|\tilde{\mu}-\mu\|_2 > \epsilon_1\right\}\cup\left\{\sup_{x^*\in\mathcal{X}^l}\|\tilde{S}-S\|_F > \epsilon_1\right\}\right) = 0, \tag{65}$$

where the randomness is from the sampling of random feature bases.

Fix an arbitrary set of $\epsilon_1 > 0, \delta_0 > 0$. Then we can find $m \in \mathbb{N}$ such that the event in (65) has probability smaller than δ_0. Combining Assumption C.2 with the fact that $\theta_0, \phi_0, \tilde{Y}_0$ are now Gaussian random variables with fixed dimensionality, for any $\delta_1 > 0$, we can choose B_1, B_2, B_3 such that the event E_n defined in Lemma C.3 has probability $1-\delta_1$. Thus for any $\epsilon_2 > 0$, when the number of iteration steps exceeds $\Omega(\delta_1^{-1}\epsilon_2^{-2})$, we have

$$\mathbb{P}(\|\hat{\theta}_m-\theta_m^*\|_2 > \epsilon_2) \le \mathbb{P}(\{\|\hat{\theta}_m-\theta_m^*\|_2 > \epsilon_2\}\cap E_n) + \mathbb{P}(E_n^c) \le 2\delta_1, \tag{66}$$

where $\hat{\theta}_m$ denotes the approximate optima returned by Algorithm 1 after $\Omega(\delta_1^{-1}\epsilon_2^{-2})$ iterations, θ_m^* denotes the exact optima of the minimax objective, and the randomness is from the gradient noise as well as the perturbations $f_0, g_0, \tilde{Y}$. Thus we have

$$\mathbb{E}\|\hat{\theta}_m-\theta_m^*\|_2 \le \epsilon_2 + 2\delta_1(\mathbb{E}\|\hat{\theta}_m\|_2 + \mathbb{E}\|\theta_m^*\|_2) \le \epsilon_2 + 4\delta_1 B.$$

From the choice of B in Lemma C.3, we can see that $\delta_1 B \le \mathbb{E}(\|\theta_m^*\|\cdot(1-\mathbf{1}_{E_n}))$, and thus converges to 0 as $\delta_1 \to 0$. Therefore, $\mathbb{E}\|\hat{\theta}_m-\theta_m^*\|_2$ converges to 0, and for any $x^* \in \mathcal{X}^l$,

$$\begin{aligned}\mathbb{E}\sup_{x^*\in\mathcal{X}^l}\|f(x^*;\hat{\theta}_m) - f(x^*;\theta_m^*)\|_2 &= \mathbb{E}\sup_{x^*\in\mathcal{X}^l}\|\phi_{x,m}(x^*)^\top(\hat{\theta}_m-\theta_m^*)\|_2 \\ &\le l\sqrt{\tilde{\kappa}}\cdot\mathbb{E}\|\hat{\theta}_m-\theta_m^*\|_2 \to 0,\end{aligned}$$

where the expectation is taken with respect to the gradient noise, perturbations, and random feature draws. Hence, the mean and covariance of $f(x^*;\hat{\theta}_m)$ converges to that of $f(x^*;\theta_m^*)$ as intended, and we know that the following holds with probability at least $1-\delta_0$

$$\sup_{x^*\in\mathcal{X}^l}\max\left\{\|\mathbb{E}(f(x^*;\hat{\theta}_m)) - \mathbb{E}(f(x^*;\theta_m))\|_2, \|\mathrm{Cov}(f(x^*;\hat{\theta}_m)) - \mathrm{Cov}(f(x^*;\theta_m))\|_F\right\} \le \epsilon_1$$

Combining this with (65) completes the proof.

D Implementation Details, Experiment Setup and Additional Results

D.1 Hyperparameter Selection

We follow the strategy in previous work [e.g., 7, 20] and select hyperparameters by minimizing the *observable* first or second stage loss, depending on which part they directly correspond to.

For the first stage, the loss is

$$\mathcal{L}_{v1} = \mathrm{Tr}(K_{xx} - 2K_{x\tilde{x}}L + K_{\tilde{x}\tilde{x}}L^\top L) = \mathbb{E}_{f\sim\mathcal{GP}(0,k)}\|f(X) - Lf(\tilde{X})\|_2^2$$

where $L := K_{z\tilde{z}}(K_{\tilde{z}\tilde{z}} + \nu I)^{-1}$, and tilde indicates the held-out data. From the above equality we can see that a Monte-Carlo estimator for L_1 can be constructed with the following procedure:

(i). Draw $f \sim \mathcal{GP}(0, k_x)$.

(ii). Perform kernel ridge regression on the dataset $\{(\tilde{z}_i, f(\tilde{x}_i))\}$.

(iii). Return the mean squared error on the dataset (X, Z).

This procedure can also be implemented for the NN-based models.

For the second stage, the loss $\sum_{i=1}^n \hat{d}_n(\hat{E}_n f, \hat{b})$ can be computed directly, for both the closed-form quasi-posterior and the random feature approximation. For the approximate inference algorithm, as we can see from (15) that the dual functions $\{g(\cdot;\varphi^{(k)})\}$ are samples from Gaussian process posteriors centered at the needed point estimates $\hat{E}_n f(\cdot;\theta^{(k)})$, instead of the point estimates themselves, we train separate validator models to approximate the latter. The validator models have the architecture to the dual functions used for training, and follow the same learning rate schedule. The validator models are trained before estimating the validation statistics, and we run SGD until convergence to ensure an accurate estimate.

D.2 Details in the Approximate Inference Algorithm

To draw multiple samples from the quasi-posterior efficiently, our algorithm runs J SGDA chains in parallel, with different perturbations $\{(\tilde{Y}^{(j)}, f_0^{(j)}, g_0^{(j)}) : j \in [J]\}$. While the convergence analysis works with the extremely simple Algorithm 1, in practice we extend it to improve stability and accelerate convergence:

(i). we employ early stopping based on the validation statistics;

(ii). before the main optimization loop we initialize the dual parameters at the approximate optima $\arg\min_\varphi L_{\mathrm{rf}}(f^{(j)}, g(\cdot;\varphi))$, by running SGD until convergence;

(iii). in each SGDA iteration, we use $K_1 > 1$ GD steps on g and one GA step for f;

(iv). after every K_2 epochs, we fix $\theta^{(j)}$ and train the dual parameters $\varphi^{(j)}$ for one epoch.

All the above choices are shown to improve the observable validation statistics. We fix $K_1 = 3, K_2 = 2$ which are determined on the 1D datasets using the validation statistics.

D.3 1D Simulation: Experiment Setup Details

In constructing the datasets, let $\tilde{f}_0$ denote the sine, step, abs or linear ($\tilde{f}_0(x) = x$) function; we then set $f_0 = \tilde{f}_0(4 \cdot (2x-1))$ if $\tilde{f}_0$ is sine, abs or linear, $\mathbf{1}_{\{2x-1<0\}} + 2.5 \cdot \mathbf{1}_{\{2x-1\geq 0\}}$ otherwise. These choices are made to maintain similarity with previous work [5, 6], which used the same transformed step function and defined $\mathbf{x}$ so that it has a range of approximately $[-4, 4]$.

For 2SLS and the kernelized IV methods, we determine λ and ν following D.1. To improve stability, we repeat the procedures on 50 random partitions of the combined training and validation set, and choose the hyperparameters that minimize the average loss. The hyperparameters are chosen from a log-linearly scaled grid consisting of 10 values in the range of $[0.1, 30]$. We note that the occasional instability of hyperparameter selection is also reported in [20]. For BayesIV, we run the MCMC sampler for 25000 iterations, discard the first 5000 iterations for burn in, and take one sample out of every 80 consecutive iterations to construct the approximate posterior. For bootstrap we use 20

samples. In both cases we verify that further increasing the computational budget does not improve the final performance.

We normalize the dataset to have zero mean and unit variance. For all kernel methods we set the kernel bandwidth using the median trick.

Finally, we provide details in the run time measurement in Table 1: BayesIV is evaluated on a machine with an i9-9900k processor (8 cores, 16 threads; 5.0 GHz) and 64GB RAM; the closed-form quasi-posterior is evaluated on a machine with two Xeon E5-2620v4 processors (total 16 cores, 32 threads; 3.0 GHz) and 220GB RAM; the approximate inference method is evaluated on the same machine, with a GeForce GTX 1080Ti GPU. The use of different machines is because BayesIV requires a Windows environment; it should not put BayesIV into disadvantage, as it only makes efficient use of 4 CPU cores. For all methods, the reported runtime excludes non-computational tasks such as data preparation and JIT compilation. For the approximate inference algorithm, we report the runtime for the optimal hyperparameters; runtime for suboptimal hyperparameters is typically lower due to early stopping. As a single run of the algorithm does not fully utilize the GPU, we run 6 experiments in parallel and report the elapsed time divided by 6. This is a realistic evaluation setting, since in practice all methods require multiple runs for hyperparameter selection, and will benefit from parallelization whenever possible.

D.4 1D Simulation: Full Results and Visualizations

Full results are reported in Table 3-4. As we can see, the gap in CI coverage between bootstrap and the quasi-posterior consistently appears across all datasets, and is most evident in the small-sample setting or when Matérn kernels are used instead of the RBF kernel.

We provide the following visualizations:

(i). We visualize the quasi-posterior and the bootstrap predictive distribution on all datasets, using the nonparametric kernel that best matches the smoothness of the target function. This amounts to Matérn-$3/2$ for abs and step, and RBF for sin and linear.[11] Results for $\alpha = 0.5$ are plotted in Figure 3, and $\alpha = 0.05$ in Figure 4. We can see that

- The credible intervals produced by our method shrink when N or α increases, correctly reflecting the increased amount of available information in training data. Their width also has the same order of magnitude as bootstrap, when $\alpha = 0.5$ (i.e., when bootstrap is more reliable).
- When the instrument strength is weak ($\alpha = 0.05$), our method is significantly more robust than bootstrap, especially when the sample size is smaller.
- On the step dataset where Assumption 3.2 is violated, our method still provides good coverage.

(ii). We plot the quasi-posterior using over-smoothed kernels on the abs dataset, which include the RBF kernel and the Matérn-$5/2$ kernel, in Figure 5 (b-c).

- We can see that both kernels produce CIs with good coverage, and the CIs have similar (albeit slightly smaller) width comparing with the Matérn-$3/2$ kernel. This is consistent with previous results on GP regression using oversmoothed priors [62]; the slight shrink in CI width could be attributed to the fact that the abs function is smoother than C^0 in most regions.

(iii). We plot the approximate quasi-posterior using the approximate inference algorithm in Figure 5 (d).[12] Comparing Figure 5 (c) and (d), we can see that the approximate and exact quasi-posterior are visually similar.

[11] None of the kernels match the discontinuous step function, so we use the least smooth one; for the linear function, we skip the linear kernel, since numerical study of quasi-posteriors using low-dimensional parametric models exists in literature [16].

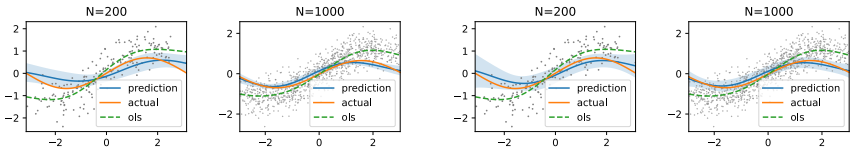
[12] We use 400 random Fourier feature basis to approximate the RBF kernel. Regularization hyperparameters are determined using the closed-form validation statistics, and optimization hyperparamaters are determined by grid search following the setting of the lower-dimensional demand experiment below.

Method	bayesIV	bs-lin	qb-lin	bs-poly	qb-poly	bs-ma3	qb-ma3	bs-ma5	qb-ma5	bs-rbf	qb-rbf
$f_0 = \text{sin}, N = 200, \alpha = 0.5$											
MSE	.024 (.038)	.111 (.011)	.109 (.010)	.243 (.037)	.243 (.034)	.023 (.010)	.025 (.014)	.022 (.011)	.021 (.015)	.025 (.013)	.026 (.015)
CI Cvg.	.895 (.252)	.232 (.039)	.110 (.019)	.077 (.032)	.045 (.017)	.965 (.065)	1.00 (.000)	.972 (.065)	1.00 (.000)	.972 (.079)	1.00 (.013)
CI Wid.	.188 (.035)	.143 (.028)	.072 (.004)	.078 (.025)	.041 (.006)	.283 (.031)	.661 (.066)	.288 (.032)	.569 (.067)	.293 (.040)	.408 (.063)
$f_0 = \text{sin}, N = 1000, \alpha = 0.5$											
MSE	.016 (.003)	.103 (.006)	.103 (.006)	.237 (.020)	.239 (.020)	.009 (.011)	.008 (.014)	.008 (.012)	.007 (.015)	.007 (.012)	.007 (.013)
CI Cvg.	.598 (.155)	.097 (.020)	.049 (.006)	.036 (.012)	.017 (.004)	.962 (.113)	1.00 (.000)	.954 (.111)	1.00 (.000)	.957 (.168)	1.00 (.036)
CI Wid.	.085 (.006)	.061 (.011)	.032 (.001)	.038 (.009)	.019 (.002)	.186 (.032)	.602 (.037)	.173 (.029)	.509 (.037)	.164 (.029)	.326 (.041)
$f_0 = \text{abs}, N = 200, \alpha = 0.5$											
MSE	.042 (.038)	.454 (.052)	.456 (.053)	.478 (.085)	.477 (.089)	.033 (.026)	.035 (.025)	.032 (.024)	.031 (.025)	.031 (.019)	.031 (.021)
CI Cvg.	.863 (.184)	.190 (.039)	.085 (.198)	.055 (.027)	.020 (.072)	.945 (.110)	1.00 (.004)	.942 (.118)	1.00 (.004)	.920 (.125)	1.00 (.030)
CI Wid.	.207 (.027)	.214 (.035)	.077 (.220)	.110 (.038)	.043 (.091)	.316 (.037)	.676 (.072)	.317 (.034)	.599 (.079)	.277 (.037)	.462 (.082)
$f_0 = \text{abs}, N = 1000, \alpha = 0.5$											
MSE	.024 (.011)	.448 (.016)	.449 (.016)	.468 (.025)	.469 (.026)	.020 (.006)	.019 (.005)	.019 (.006)	.018 (.006)	.017 (.007)	.016 (.006)
CI Cvg.	.507 (.196)	.083 (.018)	.111 (.092)	.028 (.007)	.009 (.003)	.857 (.075)	1.00 (.000)	.823 (.079)	1.00 (.000)	.829 (.102)	1.00 (.016)
CI Wid.	.092 (.005)	.098 (.019)	.127 (.103)	.061 (.016)	.019 (.002)	.181 (.024)	.646 (.050)	.174 (.026)	.530 (.056)	.168 (.024)	.383 (.061)
$f_0 = \text{step}, N = 200, \alpha = 0.5$											
MSE	.045 (.026)	.075 (.010)	.075 (.010)	.179 (.025)	.180 (.023)	.041 (.013)	.047 (.017)	.043 (.012)	.046 (.015)	.046 (.011)	.048 (.013)
CI Cvg.	.845 (.176)	.347 (.069)	.220 (.045)	.110 (.048)	.067 (.022)	.797 (.101)	1.00 (.022)	.787 (.085)	.998 (.038)	.710 (.085)	.917 (.051)
CI Wid.	.194 (.018)	.139 (.023)	.072 (.004)	.068 (.022)	.041 (.006)	.300 (.047)	.665 (.083)	.285 (.041)	.593 (.082)	.252 (.035)	.453 (.061)
$f_0 = \text{step}, N = 1000, \alpha = 0.5$											
MSE	.023 (.009)	.070 (.004)	.069 (.004)	.178 (.012)	.176 (.011)	.035 (.012)	.038 (.015)	.038 (.011)	.040 (.014)	.039 (.012)	.040 (.014)
CI Cvg.	.616 (.116)	.185 (.042)	.098 (.014)	.053 (.021)	.030 (.005)	.784 (.153)	1.00 (.016)	.739 (.164)	.976 (.028)	.661 (.134)	.839 (.077)
CI Wid.	.086 (.004)	.060 (.011)	.032 (.001)	.032 (.010)	.019 (.002)	.206 (.062)	.565 (.053)	.196 (.073)	.483 (.065)	.168 (.031)	.312 (.054)
$f_0 = \text{linear}, N = 200, \alpha = 0.5$											
MSE	.009 (.011)	.002 (.002)	.001 (.002)	.128 (.019)	.128 (.017)	.012 (.008)	.017 (.013)	.011 (.009)	.014 (.014)	.011 (.011)	.013 (.016)
CI Cvg.	.948 (.091)	1.00 (.153)	1.00 (.130)	.087 (.026)	.060 (.018)	.995 (.044)	1.00 (.001)	.995 (.050)	1.00 (.003)	.990 (.060)	1.00 (.043)
CI Wid.	.129 (.025)	.088 (.012)	.072 (.004)	.055 (.017)	.041 (.006)	.269 (.031)	.520 (.049)	.261 (.031)	.438 (.058)	.239 (.034)	.298 (.041)
$f_0 = \text{linear}, N = 1000, \alpha = 0.5$											
MSE	.005 (.002)	.000 (.001)	.000 (.001)	.121 (.012)	.121 (.012)	.006 (.004)	.007 (.005)	.006 (.003)	.007 (.005)	.006 (.003)	.005 (.004)
CI Cvg.	.626 (.130)	1.00 (.174)	1.00 (.230)	.034 (.008)	.026 (.004)	.992 (.094)	1.00 (.000)	.990 (.095)	1.00 (.000)	.975 (.103)	1.00 (.000)
CI Wid.	.051 (.002)	.039 (.006)	.032 (.001)	.026 (.006)	.019 (.002)	.171 (.031)	.508 (.043)	.156 (.032)	.418 (.044)	.135 (.024)	.242 (.043)

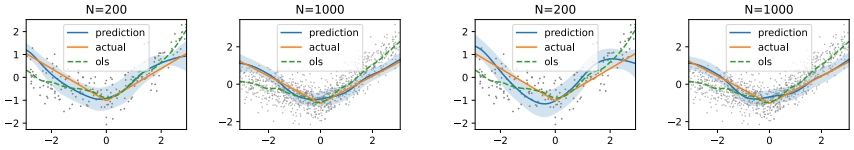
Table 3: Full results in the 1D simulation, for $\alpha = 0.5$

Method	bayesIV	bs-lin	qb-lin	bs-poly	qb-poly	bs-ma3	qb-ma3	bs-ma5	qb-ma5	bs-rbf	qb-rbf
$f_0 = \text{sin}, N = 200, \alpha = 0.05$											
MSE	.275 (.045)	.133 (.070)	.193 (.125)	.202 (.121)	.215 (.161)	.231 (.037)	.190 (.068)	.209 (.037)	.163 (.073)	.183 (.047)	.142 (.081)
CI Cvg.	.165 (.077)	.992 (.080)	1.00 (.134)	.325 (.092)	.425 (.080)	.270 (.082)	.960 (.084)	.332 (.121)	.955 (.086)	.468 (.219)	.952 (.134)
CI Wid.	.192 (.030)	.589 (.166)	.971 (.481)	.394 (.171)	.771 (.265)	.192 (.036)	.712 (.045)	.233 (.054)	.694 (.065)	.297 (.085)	.638 (.105)
$f_0 = \text{sin}, N = 1000, \alpha = 0.05$											
MSE	.146 (.025)	.123 (.077)	.169 (.315)	.213 (.145)	.228 (.192)	.246 (.071)	.216 (.113)	.238 (.085)	.214 (.131)	.188 (.096)	.188 (.131)
CI Cvg.	.082 (.060)	.880 (.296)	.888 (.346)	.289 (.086)	.344 (.135)	.373 (.111)	.819 (.131)	.436 (.145)	.840 (.185)	.562 (.218)	.897 (.212)
CI Wid.	.095 (.018)	.552 (.367)	.852 (.568)	.405 (.294)	.588 (.400)	.254 (.036)	.605 (.049)	.298 (.042)	.568 (.050)	.373 (.086)	.536 (.055)
$f_0 = \text{abs}, N = 200, \alpha = 0.05$											
MSE	.806 (.478)	.487 (.259)	.500 (.505)	.489 (.233)	.472 (.453)	.392 (.064)	.349 (.156)	.368 (.074)	.350 (.180)	.336 (.094)	.393 (.221)
CI Cvg.	.122 (.197)	.435 (.167)	.775 (.201)	.247 (.119)	.545 (.122)	.217 (.102)	.795 (.137)	.287 (.130)	.832 (.163)	.352 (.243)	.805 (.250)
CI Wid.	.239 (.049)	.526 (.311)	1.30 (.473)	.303 (.152)	.895 (.279)	.294 (.047)	.712 (.045)	.351 (.065)	.694 (.065)	.424 (.103)	.638 (.105)
$f_0 = \text{abs}, N = 1000, \alpha = 0.05$											
MSE	1.45 (.250)	.479 (.105)	.500 (.083)	.472 (.159)	.472 (.111)	.376 (.090)	.390 (.144)	.374 (.109)	.367 (.181)	.304 (.134)	.265 (.214)
CI Cvg.	.019 (.014)	.464 (.179)	.742 (.249)	.332 (.165)	.562 (.191)	.306 (.109)	.665 (.207)	.374 (.165)	.667 (.248)	.505 (.269)	.625 (.308)
CI Wid.	.139 (.017)	.460 (.328)	1.24 (.576)	.340 (.238)	.923 (.384)	.339 (.044)	.605 (.049)	.367 (.056)	.561 (.049)	.504 (.114)	.536 (.061)
$f_0 = \text{step}, N = 200, \alpha = 0.05$											
MSE	.226 (.127)	.105 (.069)	.148 (.199)	.157 (.092)	.192 (.148)	.214 (.036)	.183 (.070)	.194 (.038)	.176 (.077)	.160 (.056)	.147 (.090)
CI Cvg.	.193 (.090)	.777 (.158)	.787 (.197)	.382 (.079)	.438 (.066)	.262 (.103)	.952 (.068)	.300 (.158)	.920 (.089)	.432 (.282)	.890 (.149)
CI Wid.	.184 (.032)	.621 (.297)	.767 (.452)	.456 (.179)	.652 (.292)	.262 (.051)	.712 (.045)	.310 (.070)	.694 (.065)	.365 (.103)	.638 (.112)
$f_0 = \text{step}, N = 1000, \alpha = 0.05$											
MSE	.079 (.444)	.083 (.044)	.131 (.207)	.152 (.105)	.236 (.120)	.231 (.059)	.214 (.107)	.224 (.070)	.208 (.127)	.192 (.082)	.170 (.130)
CI Cvg.	.149 (.231)	.715 (.171)	.696 (.201)	.366 (.087)	.390 (.120)	.290 (.075)	.841 (.165)	.353 (.144)	.853 (.212)	.515 (.229)	.800 (.189)
CI Wid.	.125 (.021)	.544 (.295)	.815 (.564)	.408 (.187)	.539 (.423)	.298 (.037)	.605 (.049)	.330 (.052)	.561 (.049)	.420 (.095)	.543 (.065)
$f_0 = \text{linear}, N = 200, \alpha = 0.05$											
MSE	.083 (.023)	.013 (.032)	.041 (.202)	.103 (.065)	.151 (.149)	.052 (.015)	.046 (.024)	.046 (.016)	.045 (.023)	.046 (.020)	.053 (.023)
CI Cvg.	.238 (.133)	1.00 (.284)	1.00 (.172)	.372 (.099)	.412 (.109)	.490 (.194)	1.00 (.000)	.750 (.219)	1.00 (.000)	.895 (.188)	1.00 (.026)
CI Wid.	.121 (.030)	.559 (.213)	.595 (.406)	.495 (.181)	.543 (.327)	.220 (.028)	.712 (.045)	.270 (.041)	.694 (.065)	.323 (.074)	.638 (.105)
$f_0 = \text{linear}, N = 1000, \alpha = 0.05$											
MSE	.051 (.007)	.018 (.034)	.022 (.165)	.125 (.092)	.188 (.411)	.070 (.023)	.069 (.041)	.070 (.028)	.070 (.047)	.057 (.031)	.066 (.043)
CI Cvg.	.094 (.041)	1.00 (.215)	1.00 (.200)	.338 (.113)	.298 (.122)	.455 (.107)	1.00 (.028)	.528 (.192)	1.00 (.066)	.684 (.237)	1.00 (.096)
CI Wid.	.058 (.007)	.535 (.289)	.423 (.470)	.435 (.309)	.406 (.365)	.205 (.020)	.605 (.049)	.230 (.033)	.561 (.049)	.297 (.089)	.530 (.061)

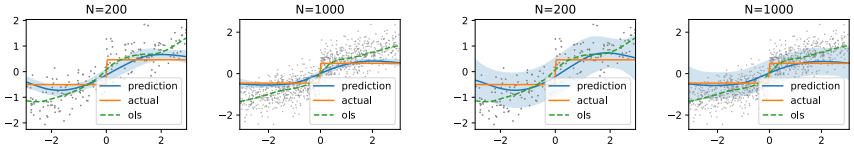
Table 4: Full results in the 1D simulation, for $\alpha = 0.05$



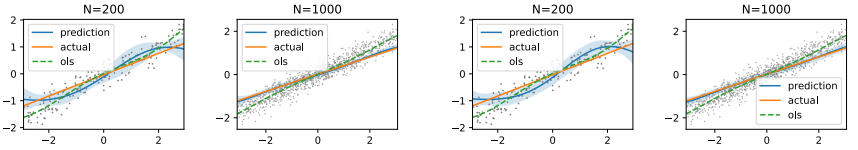
(a) sin, bootstrap (b) sin, quasi-Bayes



(c) abs, bootstrap (d) abs, quasi-Bayes

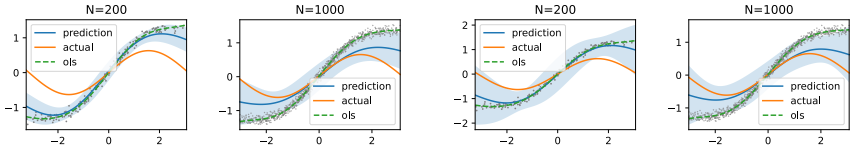


(e) step, bootstrap (f) step, quasi-Bayes

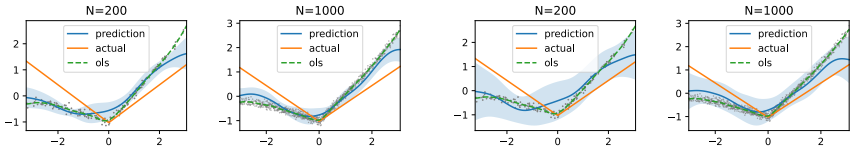


(g) linear, bootstrap (h) linear, quasi-Bayes

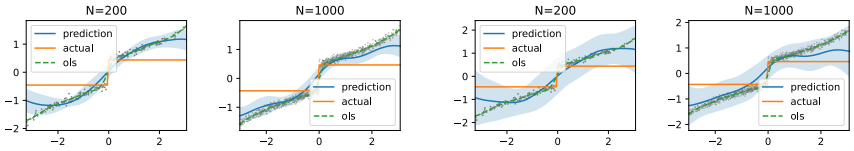
Figure 3: 1D datasets: visualizations of predictive distribution with $\alpha = 0.5$. Dot indicates the training data, and "ols" indicates biased regression predictions using KRR. Shade indicates 95% CI.



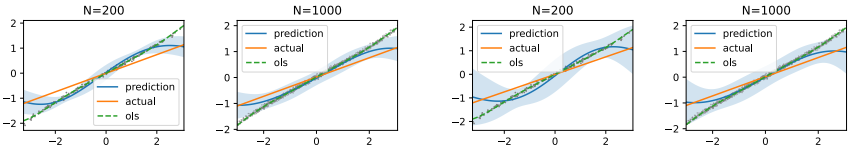
(a) sin, bootstrap (b) sin, quasi-Bayes



(c) abs, bootstrap (d) abs, quasi-Bayes

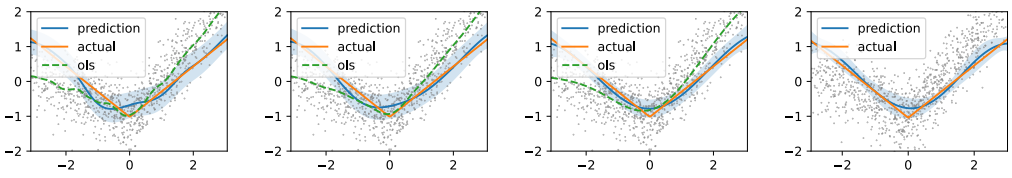


(e) step, bootstrap (f) step, quasi-Bayes



(g) linear, bootstrap (h) linear, quasi-Bayes

Figure 4: 1D datasets: visualizations of predictive distribution with $\alpha = 0.05$. Best viewed when zoomed. Due to the hyperparameter selection procedure, the CIs do not always shrink as N increases.



(a) Matérn-3/2 kernel (b) Matérn-5/2 kernel (c) RBF, closed form (d) RBF, RF approx

Figure 5: 1D datasets: visualization of the quasi-posterior on the abs dataset using various models. We fix $N = 1000, \alpha = 0.5$.

D.5 Demand Simulation: Experiment Setup Details

All variables in the dataset are normalized to have zero mean and unit variance. For BayesIV, we run the MCMC sampler for 50000 iterations, discard the first 10000 samples as burn in, and take every 80th sample for inference. For the kernelized methods, hyperparamater selection follows the 1D experiments. For the NN-based methods, implementation details are discussed in Appendix D.2; for both our method and bootstrap, we draw $J = 10$ samples from the predictive distribution.

We select hyperparameters by applying the procedure in Appendix D.1 to a fixed train / validation split, since on this dataset we observe little variation in its results. Hyperparameters include λ, ν, and the learning rate schedule (initial learning rate η_0 and period of learning rate decay τ). The learning rate is adjusted by multiplying it by a factor of 0.8 every τ iterations. We fix the optimizer to Adam, and train until validation statistics no longer improves.

For the lower-dimensional setup, we select λ and ν from a log-linearly scaled grid of 10 values, with the range of $[5 \times 10^{-3}, 5]$ and $[0.05, 1]$, respectively. The ranges are chosen based on preliminary experiments using the range of $[0.1, 30]$. We determine η from $\{5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 5 \times 10^{-2}\}$, and τ from $\{80, 160, 320, 640\}$. We fix the batch size at 256. The NN architecture consists of two fully-connected layers, with 50 hidden units and the tanh activation. We also experimented with NNs with 3 hidden layers or with ReLU activation, and made this choice based on the validation statistics.

For the image-based setup, the range of λ and η follows the above. For ν, τ we consider $\nu \in [1, 100], \tau \in \{640, 1280, 2560, 5120\}$, based on preliminary experiments. We fix the batch size at 80. The network architecture is adapted from [4], and consists of two 3×3 convolutional layers with 64 filters, followed by max pooling, dropout, and three fully-connected layers with $64, 32$ and 1 units.

Following the setup in all previous work, we use a uniform grid on $[5, 30] \times [0, 10] \times \{0, \dots, 6\}$ as the test set.

Computational cost We report the typical training time for a single set of hyperparameters, excluding JIT compilation time, on a GeForce GTX 1080Ti GPU. In the lower-dimensional experiments, training takes around 25 minutes for a single set of hyperparameters when $N = 10^3$, or around 30 minutes when $N = 10^4$; in both cases 6 experiments can be carried out in parallel on a single GPU. In the image experiment, training takes around 7.5 hours.

The time cost reported above is for the optimal hyperparameter configuration; experiments using suboptimal hyperparameters usually take a shorter period of time due to early stopping. It can also be improved by switching to low-precision numerical operations, or with various heuristics in the hyperparameter search (e.g., using a smaller J in an initial search).

D.6 Demand Simulation: Full Results and Visualizations

Results in the large-sample settings are presented in Table 5. We only include 2SLS for comparison, since the time complexity of the other baselines is too high. The results are consistent with the discussion in the main text.

We plot the predictive distributions for all methods in Figure 7, on the same cross-section as in the main text, for $N = 1000$. (We omit the plot for $N = 10^4$ and the image experiment, since in

those settings bootstrap and the quasi-posterior have similar behaviors.) As we can see, all non-NN baselines except BayesIV produce overly smooth predictions, presumably due to the lack of flexibility in these models. Note that the visualizations only correspond to an intersection of the true function $f(x_0, t_0, s)$, with x_0, t_0 fixed; the complete function has the form of $x \cdot s \cdot \psi(t)$, ignoring the less significant terms, and thus may incur a large norm penalty in the less flexible RKHSes. The issue is further exacerbated by the discrepancy between the training and test distributions: the former is non-uniform due to confounding. As we can see from Figure 6, in the region where t is close to 5, the data is scarce for most values of x, which may explain the reason that the RBF-based methods fail to provide good coverage around $t = 5$ (and $s = 3, x = 17.5$, as used in the visualizations), and the reason that both NN-based methods assign higher uncertainty around this location.

BayesIV has a different failure mode: as it employs additive regression models for both stages $p(\mathbf{x} \mid \mathbf{z}), p(\mathbf{y} \mid \mathbf{x})$, it approximates this cross-section relatively well. However, as the true structural function does not have an additive decomposition, its prediction in other regions can be grossly inaccurate; we plot one such cross-section in Figure 8(a).

When implemented with the NN model, bootstrap CIs are more optimistic in regions with more training data, although the difference is often insignificant. The difference in out-of-distribution regions is more significant, where bootstrap is often less robust. An example is provided in Figure 8.

Table 5: Deferred results on the demand design. Results are averaged over 20 trials for the low-dimensional experiment, and 10 trials for the image experiment.

Setting	Low-dimensional, $N = 10^4$			Image, $N = 5 \times 10^4$		
Method	BS-2SLS	BS-NN	QB-NN	BS-2SLS	BS-NN	QB-NN
NMSE	$.371 \pm .003$	$.014 \pm .003$	$.020 \pm .002$	$.559 \pm .008$	$.168 \pm .027$	$.138 \pm .037$
CI Cvg.	$.024 \pm .005$	$.944 \pm .009$	$.957 \pm .008$	$.112 \pm .005$	$.892 \pm .022$	$.909 \pm .017$
CI Wid.	$.014 \pm .002$	$.136 \pm .015$	$.203 \pm .013$	$.132 \pm .039$	$.636 \pm .027$	$.597 \pm .024$

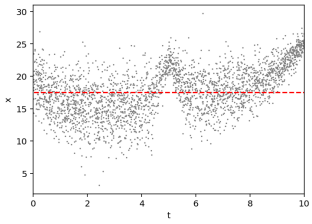
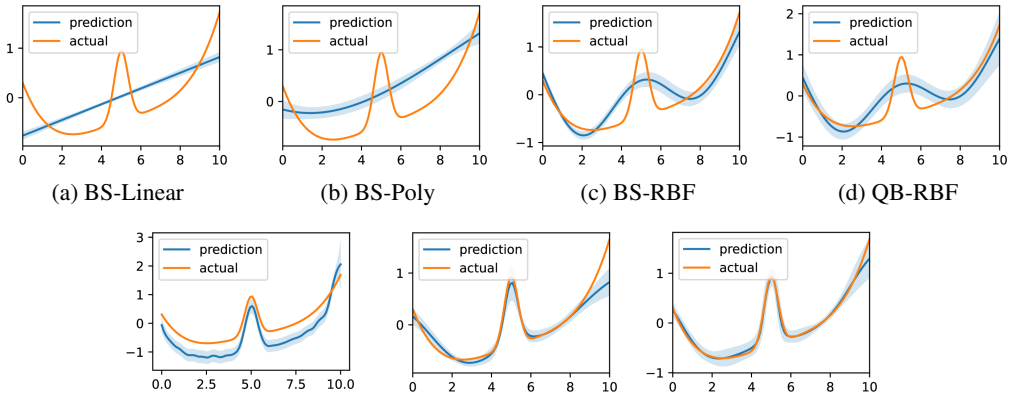


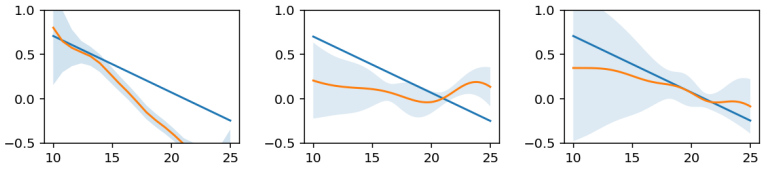
Figure 6: Demand experiment: scatter plot of 10^4 samples from the training data distribution $p(x, t \mid s = 4)$. The dashed line indicates the cross-section used in Figure 2.



(a) BS-Linear (b) BS-Poly (c) BS-RBF (d) QB-RBF

(e) BayesIV (f) BS-NN (g) QB-NN

Figure 7: Demand experiment: visualizations of the predictive distributions for $N = 1000$, on the same cross-section as in Figure 2.



(a) BayesIV (b) BS-NN (c) QB-NN

Figure 8: Demand experiment: visualizations of the predictive distributions for $N = 1000$ on a out-of-distribution cross-section, obtained by fixing $t = 9, s = 6$ and varying x.