DISTRIBUTIONALLY ROBUST GAUSSIAN PROCESS REGRESSION AND BAYESIAN INVERSE PROBLEMS

BY XUHUI ZHANG[1,a], JOSE BLANCHET[1,b], YOUSSEF MARZOUK[2,c], VIET ANH NGUYEN[3,e], AND SVEN WANG[2,d]

[1]*Stanford University,* [a]*xuhui.zhang@stanford.edu;* [b]*jose.blanchet@stanford.edu*

[2]*Massachusetts Institute of Technology,* [c]*ymarz@mit.edu;* [d]*svenwang@mit.edu*

[3]*Chinese University of Hong Kong,* [e]*nguyen@se.cuhk.edu.hk*

We study a distributionally robust optimization formulation (i.e., a min-max game) for two representative problems in Bayesian nonparametric estimation: Gaussian process regression and, more generally, linear inverse problems. Our formulation seeks the best mean-squared error predictor in an infinite-dimensional space against an adversary who chooses the worst-case model in a Wasserstein ball around a nominal infinite-dimensional Bayesian model. The transport cost is chosen to control features such as the degree of roughness of the sample paths that the adversary is allowed to inject. We show that strong duality holds in the sense that max-min equals min-max, and that there exists a unique Nash equilibrium that can be computed by a sequence of finite-dimensional approximations. Crucially, the worst-case distribution is itself Gaussian. We explore the properties of the Nash equilibrium and the effects of hyperparameters through numerical experiments, demonstrating the versatility of our modeling framework.

1. Introduction. Bayesian nonparametric estimation [19] is used ubiquitously in science, engineering, and other statistical application areas for both 'direct' nonparametric regression and solving inverse problems. The computation of the posterior (or conditional) mean estimators, which are the most commonly used Bayesian point estimators, involves the solution of an infinite-dimensional optimization problem whose specification requires knowledge of the distributions at hand. In general, this problem has no closed-form solution. If the observations and the parameter of interest are jointly Gaussian, however, then the problem immediately becomes tractable [20, 41]. The conditional expectation (the best mean-square estimator of the parameter) is an affine function of the observations. Conditional covariances also can be evaluated in closed forms, enabling quantification of prediction uncertainty.

Gaussianity is a strong assumption that is easily violated in reality. More generally, the joint probabilistic model for the parameter of interest and the observations is often—perhaps inevitably—*misspecified*. This misspecification can take myriad forms, including incorrect prior assumptions on the smoothness or dependence structure of the unknown parameter and incorrect assumptions on the nature of the data-generating process—which in turn involves assumptions on both the observational noise and, in the inverse problem setting, on the "forward" operator relating the parameters to the observations. In these situations, it is desirable to ensure some form of *robustness*, e.g., to construct a nonparametric estimator that hedges against the impact of model misspecification on mean-square error. One also would like to represent possible modeling errors nonparametrically and in a way that fully reflects the infinite-dimensional nature of the regression and inverse problem settings.

This paper addresses model misspecification in nonparametric settings by adapting and extending ideas from distributionally robust optimization (DRO) [13, 32]. DRO formulations create a min-max game between a decision-maker and a fictitious adversary: the decision-maker finds a decision or an estimator that minimizes the worst-case expected loss over all possible distributions the adversary can choose. The set containing the adversary's feasible choice is the *ambiguity set*. In Wasserstein DRO, this ambiguity set is prescribed as a Wasserstein neighborhood around a nominal distribution [24, 9, 7]. By choosing appropriately the nominal distribution and the Wasserstein distance, the Wasserstein DRO solutions can exhibit many desirable properties, e.g. in the setting of (group) LASSO [6], norm regularization [35], shrinkage [30], ridge regression [27], etc.

The first line of work on Wasserstein DRO set the nominal distribution as the empirical distribution of the available data [15, 17]. Subsequently, finite-dimensional parametric information under the form of Gaussianity is injected into the nominal distribution for the estimation task [30, 36]. To capture more structural information about the problem, independence has also added to the ambiguity set to model the independence between the finite-dimensional signal and the finite-dimensional error in the minimum mean square error estimation problem [31]. In Wasserstein DRO when the nominal distribution is a generic stochastic process, the papers [18, 8] derived some general duality results quantifying the worst-case expected loss over the Wasserstein ambiguity set (i.e., the inner maximum of a DRO formulation), while the full min-max game was not studied in this setting.

This paper significantly extends the existing literature by introducing a DRO formulation for *in*finite-dimensional Gaussian models. In our formulation, the decision-maker seeks an estimator of the unknown parameter that minimizes a certain Bayes risk, and the adversary chooses a probabilistic model that departs from the baseline/nominal Gaussian model assumed by the decision maker. Specifically, we will allow the adversary to select a model in a nonparametric way within a δ-Wasserstein ball around the nominal model, where δ is the ambiguity size. The particular Wasserstein geometry that we impose allows the adversary to select models that depart significantly not only *from the Gaussian assumption* but also *from the smoothness properties* dictated by the decision-makers' choice of prior. Consequently, our min-max formulation allows us to explore and assess the impact of model misspecification effectively.

We briefly contrast the nominal and robust inverse problems to make these ideas concrete; a precise and technical presentation is deferred to Section 2. Let b^0 represent the parameter of interest, modeled as a real-valued random process with continuous sample paths on some compact domain $\mathcal{D} \subseteq \mathbb{R}^d$. Suppose we have a finite number of real-valued observations $(Y_1, \ldots, Y_m)$, specified as

$$Y_i = T(b^0)(x_i) + \epsilon_i^0,$$

where T is a bounded linear operator, $(x_i)_{i=1}^m \in \mathcal{D}$ are collocation or design points, and $(\epsilon_i^0)_{i=1}^m$ are real-valued random variables with independent entries representing observational noise. In the nominal case, b^0 and ϵ_i^0 are endowed with a Gaussian prior measure P_0 on $L^2(\mathcal{D}) \times \mathbb{R}^m$, and under P_0, ϵ_i^0 are independent of b^0. We seek an estimator $\phi : \mathbb{R}^m \to L^2(\mathcal{D})$ that minimizes the Bayes risk

$$(1) \qquad \min_\phi \ \mathbb{E}_{P_0} \left[\left\| b^0 - \phi(Y_1, \ldots, Y_m) \right\|_{L^2(\mathcal{D})}^2 \right].$$

It is well known that this estimator is given by the conditional expectation $\phi(Y_1, \ldots, Y_m) = \mathbb{E}[b^0 \,|\, Y_1, \ldots, Y_m]$ of the posterior distribution. We build the Wasserstein distributionally robust counterpart of this problem by introducing an ambiguity set of probability measures, $\{P : \mathbb{W}(P, P_0) \le \delta\}$, where $\mathbb{W}$ is a Wasserstein distance on the space of (Borel) probability

measures on $C(\mathcal{D}) \times \mathbb{R}^m$. We then seek an estimator ϕ that minimizes the worst-case Bayes risk over this ambiguity set:

$$(2) \qquad \inf_{\phi} \sup_{P : \mathrm{W}(P, P_0) \leq \delta} \mathbb{E}_P \left[\| b - \phi(Y_1, \ldots, Y_m) \|_{L^2(\mathcal{D})}^2 \right],$$

where now $Y_i = T(b)(x_i) + \epsilon_i$, and b and $\epsilon_1, \ldots, \epsilon_m$ are jointly distributed according to P. The robust formulation (2) thus adds a fictitious adversary to the nominal formulation (1), and the adversary is the inner player solving the supremum subproblem.

By choosing among distributions in the ambiguity set, the adversary may inject additional roughness to sample paths of b, beyond what is specified by the nominal distribution P_0. The ability to add roughness is directly tied to our specification of $\mathbb{W}$: we consider perturbations to b that are elements of an RKHS $\mathcal{H}_w$ whose norm is parameterized by a sequence of weights $(w_n)_{n \geq 1}$. Using these weights, we obtain a flexible class of ground cost defining our Wasserstein distance which can adjust the penalty for transportation along different "modes" of the sequence decomposition of b in the Karhunen-Loeve expansion. In particular, the choice of weights $(w_n)_{n \geq 1}$ induces a hierarchy of the Hilbert space norms so that rougher perturbations correspond to spaces whose weights $(w_n)_{n \geq 1}$ grow to infinity in a slower rate. This infinite-dimensional construction is an important novelty from the DRO perspective. To guide intuition, consider the Matérn process (see Example 2.1 in the main body). The smoothness of the Matérn kernel is controlled by a parameter α, which determines the regularity of the sample paths. When α is large, the paths are smooth, while a smaller α results in rougher paths. In our framework, optimal transport perturbation of b with a slower-growing weights $(w_n)_{n \geq 1}$ is analogous to reducing the value of α, which introduce more roughness into the sample paths of b. The adversary also can modify the distribution of the additive noise ϵ_i, which can be understood in part as compensating for the misspecification of the nominal forward operator T [22]. Moreover, the adversary can replace P_0 (in any of its marginals or jointly) with a non-Gaussian distribution.

Our assumptions on the operator T in the formulation above will encompass many linear inverse problems [41, 12], e.g., learning the initial or boundary conditions of a heat equation, or canonical problems in computerized tomography [29]. By setting $T = \mathrm{Id}$, however, we recover the important case of Gaussian process regression, to which our main results immediately apply. For non-identity T, our results apply both to recovering b (e.g., solving the inverse problem) and to estimating u (e.g., PDE-constrained regression).

Wasserstein-type distances defined on the space of stochastic processes were recently studied by [4, 1, 5]. The focus therein is on processes indexed by a one-dimensional parameter (representing, for instance, time). A typical setting involves price processes in finance [4, 1], where one needs to define a Wasserstein distance that respects causal structure (i.e., filtrations). In contrast, we consider in this work a Wasserstein distance on multi-dimensional fields.

Note also that in our formulation (2), we have an infinite-dimensional action space for the outer player and an infinite-dimensional action space for the inner player. Moreover, the actions of the inner player are themselves probability measures on infinite-dimensional spaces. These features differentiate our analysis from prior work, such as [6, 15, 36]. To our knowledge, previous work in the DRO literature assumes either the action set of the decision maker to be finite-dimensional or the probability measures (the action set of the fictitious adversary) to be supported on finite-dimensional spaces.

Alternative to the Wasserstein distance, one can also naturally formulate problem (2) with an information divergence. Distributionally robust conditional mean estimation with a relative entropy ambiguity set centered on a multivariate Gaussian prior has been studied in [25, 26, 47]. As shown in [25, Theorem 1] and [26, Theorem 1], however, the resulting robust

estimator coincides exactly with its non-robust counterpart, and only the posterior covariance is inflated. The same conclusion holds for ambiguity sets constructed from the τ-divergence family [48, 49]. In strong contrast to these results, the robust estimator of our formulation (2) typically differs from its non-robust counterpart.

While the results just discussed were derived in the finite-dimensional setting, it is reasonable to expect that similar properties would be preserved under reasonable assumptions in the infinite-dimensional setting, which is our concern here. Distributionally robust formulations of non-causal filtering that employ a relative entropy or τ-divergence ambiguity set centered on a stationary Gaussian process prior have been studied in [26, 49], where it was shown that the nominal non-causal Wiener filter remains optimal. Moreover, an infinite-dimensional DRO formulation based on, e.g., relative entropy or any other criterion that requires the existence of a likelihood ratio will typically restrict the adversary to preserve sample path properties, such as the degree of sample path smoothness under the baseline model (i.e., the prior P_0). Choosing instead a Wasserstein ambiguity set, as we will explain, permits adversarial distributions with rougher sample paths than the prior, so that *smoothness misspecification* is naturally addressed by our robust estimation problem.

We now summarize our main contributions. Under reasonable assumptions to be made precise later:

- We analyze the Wasserstein distributionally robust inverse problem (2) and show that strong duality holds, in the sense that the infimum and the supremum operators can be switched without any loss of optimality.
- We show that there exists an upper bound $\delta_0 > 0$ such that if $0 < \delta < \delta_0$, problem (2) also admits a unique Nash equilibrium pair $(\phi_\infty^\star, P_\infty^\star)$. Moreover, the worst-case distribution $P_\infty^\star$ involves a modified Gaussian process with potentially rougher paths than the prior. Consequently, the robustified conditional mean remains affine in the observations, and the optimization problem is tractable.
- We numerically approximate problem (2) by a sequence of finite-dimensional counterparts and, therefore, obtain a procedure to compute the associated Nash equilibrium. Our numerical algorithm adapts the finite-dimensional Frank-Wolfe algorithm in [36].

One way to interpret our results is that the solution of linear inverse problems and Gaussian process regression can be made robust in a nonparametric sense. The interpretation of the worst-case covariance function (i.e., the covariance of the Gaussian process $P_\infty^\star$) enables the computation of the upper bound on the worst-case mean square error and thus an upper bound on the quality of the robust solution. In our numerical examples, we explore the structural properties of this worst-case covariance function and the associated robust estimator. In particular, we explore the prior and posterior covariances of b under P_0 and $P_\infty^\star$ (i.e., $\mathbb{C}\mathrm{ov}_P[b]$ and $\mathbb{C}\mathrm{ov}_P[b \,|\, y_1, \ldots, y_m]$ for $P = P_0$ and $P = P_\infty^\star$) and find that the worst-case distributions (both prior and posterior) have greater uncertainty in regions of $\mathcal{D}$ where information is limited, which intuitively guarantees greater robustness of the predictions. Moreover, we observe that in cases where there is (i) a smoother nominal prior, (ii) a smaller transport penalty in basis directions that induce roughness, (iii) a larger δ, or (iv) a smaller nominal observational noise, the worst-case distributions induce sharper contrasts between the observed and unobserved locations along both the prior and posterior sample paths. One criticism of our results is that the distributionally robust solution may be pessimistic. While our framework provides the flexibility of tuning hyperparameters such as the weights in defining the Wasserstein distance $\mathbb{W}$, we leave the precise specification of these hyperparameters for future work.

The remainder of this paper is organized as follows. We introduce our problem setup in Section 2. We present our main theoretical results in Section 3, illustrate the applicability of our general framework by highlighting several examples in Section 4, and present simple

numerical experiments in Section 5. All proofs are deferred to Appendix A. For convenience, we provide a glossary of notations in Appendix B (list of important notations, their brief description, and page of reference).

2. Problem Statement. Let $\mathcal{D} \subset \mathbb{R}^d$, $d \geq 1$ be a compact set. We write $C(\mathcal{D})$ to denote the space of real-valued continuous functions on D, which is naturally endowed with the sup-norm $\| \cdot \|_{C(\mathcal{D})}$. We denote by $L^2(\mathcal{D})$ the space of real-valued square-integrable functions on $\mathcal{D}$. Since $\mathcal{D}$ is compact, we have $C(\mathcal{D}) \subseteq L^2(\mathcal{D})$.

We introduce a probability measure P_0 under which the so-called prior input process b^0 is a $C(\mathcal{D})$-valued centered Gaussian random field.[1] Further, under the inclusion $C(\mathcal{D}) \hookrightarrow L^2(\mathcal{D})$ where $\hookrightarrow$ denotes the inclusion map, the random field b^0 can be viewed as $L^2(\mathcal{D})$-valued. This random field generates a positive definite kernel, namely, $\mathcal{K}(x, x') = \mathbb{E}_{P_0}[b^0(x)b^0(x')]$ and thus an associated reproducing kernel Hilbert space (RKHS) which is obtained as the closure of functions of the form $f(x) = \mathbb{E}_{P_0}[(a_1 b^0(x_1) + \cdots + a_n b^0(x_n))b^0(x)]$. The closure can be taken relative to the norms $\| \cdot \|_{C(\mathcal{D})}$ and $\| \cdot \|_{L^2(\mathcal{D})}$; both limiting procedures coincide [44, Lemma 8.1]. By the spectral decomposition of the covariance operator of b^0 (i.e., $\mathcal{K}(\cdot, \cdot)$, [20, Example 2.6.15]), there exists a complete orthonormal system $\{e_n\}_{n=1}^\infty$ of $L^2(\mathcal{D})$ where $e_n \in C(\mathcal{D})$), an i.i.d. sequence of standard univariate normal random variables $\{g_n\}_{n=1}^\infty$, and a non-negative sequence of "eigenvalues" $\{\kappa_n^2\}_{n=1}^\infty$ satisfying $\sum_{n=1}^\infty \kappa_n^2 < \infty$, such that under P_0,

$$b^0 = \sum_{n \geq 1} \kappa_n g_n e_n,$$

where the convergence of the above infinite sum occurs in $C(\mathcal{D})$, and thus also in $L^2(\mathcal{D})$, almost surely. We impose a full-rank assumption on the prior b^0 in the following sense, which is necessary to ensure that the worst-case distribution is unique in the proof of Theorems 3.1 and 3.2 below.

ASSUMPTION 2.1 (Full rank). *We assume that $\kappa_n \neq 0$ for all $n \geq 1$.*

A general class of Gaussian smoothness priors can be constructed via the Laplace operator on $\mathcal{D}$. Specifically, we will use the following class of Matérn processes in Sections 4 and 5, which provides natural prior distributions for α-regular functions vanishing at the boundary $\partial\mathcal{D}$[2].

EXAMPLE 2.1 (Matérn prior [28, Equation (2)]). *Suppose $\mathcal{D}$ has a smooth boundary. The prior with a Matérn covariance function with parameters $\kappa \geq 0$ and $\alpha > \frac{d}{2}$ controlling the smoothness can be expressed as*

$$b^0 = \sum_{n \geq 1} \left(\kappa^2 + \lambda_n\right)^{-\frac{\alpha}{2}} g_n e_n,$$

where the eigenvalues λ_n and eigenfunctions e_n correspond to the Dirichlet-Laplace operator on $\mathcal{D}$ [42, Corollary 5.1.5]. The eigenvalues λ_n satisfy Weyl's law $\lambda_n = \Theta(n^{2/d})$ [43, Corollary 8.3.5] and the eigenfunctions $e_n \in C^\infty(\bar{\mathcal{D}})$, $n \geq 1$, where $\bar{\mathcal{D}}$ denotes the closure of $\mathcal{D}$ and $C^\infty(\bar{\mathcal{D}})$ is the space of infinitely smooth functions on $\bar{\mathcal{D}}$.

---

[1] For the theory of Banach space-valued Gaussian random variables, we refer to [20, Chapter 2] and [44].

[2] α-regular functions refer to functions that are Hölder (and Sobolev)-continuous of order $\alpha - d/2 - q$ for any $q > 0$.

We consider perturbations to the nominal prior b^0 by borrowing ideas from the field of distributionally robust optimization (DRO). We assume that the perturbations are supported in a space of continuous functions $\mathcal{H}_w$ that is also an RKHS; in particular $\mathcal{H}_w$ is also a Polish space, which is important to invoke key duality results to study the maximization in our DRO formulation. The useful feature of RKHS is that point evaluation functionals are well-defined and continuous with respect to the Hilbert space norm [44, Example 2.2]. Specifically, we define the space

$$(3) \qquad \mathcal{H}_w = \left\{ f \in L^2(\mathcal{D}) : \sum_{n \geq 1} w_n \langle f, e_n \rangle^2 < \infty \right\},$$

which is parameterized by a positive sequence $w = (w_n)_{n \geq 1}$. Notice that from this point, we abbreviate the inner product $\langle \cdot, \cdot \rangle_{L^2(\mathcal{D})}$ as $\langle \cdot, \cdot \rangle$. Typically, the Hilbert norm on $\mathcal{H}_w$ is stronger than the usual $L^2(\mathcal{D})$ norm. More precisely, we impose the following assumption on w and the basis $\{e_n\}_{n=1}^{\infty}$:

ASSUMPTION 2.2 (RKHS conditions). *Assume that $\lim_{n \to \infty} w_n = \infty$ and $\mathcal{H}_w$ is endowed with the inner product*

$$\langle f, \tilde{f} \rangle_{\mathcal{H}_w} = \sum_{n \geq 1} w_n \langle f, e_n \rangle \langle \tilde{f}, e_n \rangle \qquad \forall f, \tilde{f} \in \mathcal{H}_w.$$

Further, the space $\mathcal{H}_w$ equipped with the Hilbert space norm $\| \cdot \|_{\mathcal{H}_w}$ is an RKHS compactly embedded in $C(\mathcal{D})$.

Under Assumption 2.2, $\mathcal{H}_w$ is a separable Hilbert space such that point evaluation functionals are well-defined and continuous. Note that the sequence w controls the roughness (or equivalently the smoothness) of functions in $\mathcal{H}_w$. Throughout the rest of the paper, we will fix a given sequence w satisfying Assumption 2.2. The norm $\| \cdot \|_{\mathcal{H}_w}$ will define our adversarial perturbations. The next example, a continuation of Example 2.1, provides intuition about the interpretation of w in terms of roughness.

EXAMPLE 2.2 (RKHS space). *Let the eigenvalues λ_n and eigenfunctions e_n correspond to the Dirichlet-Laplacian operator on $\mathcal{D}$. Consider $w_n = \Theta(\lambda_n^{\beta})$.[3] Then β controls the roughness of functions in $\mathcal{H}_w$. Note that the "spectrally defined" spaces $\mathcal{H}_w$ are subspaces of the classical Sobolev spaces on $\mathcal{D}$. For any $\beta > \frac{d}{2}$, by the Sobolev embedding theorem [42, Proposition 4.1.3], we can identify $f \in \mathcal{H}_w$ with its continuous version. Under this identification $\mathcal{H}_w$ is a RKHS in $C(\mathcal{D})$.*

Now, suppose that we have a linear forward operator T that maps sample paths of the prior input process b^0 to sample paths of another process u, which also takes values in $C(\mathcal{D})$. We make the following assumptions on this operator.

ASSUMPTION 2.3 (Operator). *We assume the following:*

(i) *The forward map $T : C(\mathcal{D}) \to C(\mathcal{D})$ is linear and bounded (with operator norm $C_T > 0$).*

---

[3]For positive sequences $\{a_n\}, \{b_n\}$, the notation $a_n = \Theta(b_n)$ means that $1/c_0 \leq a_n/b_n \leq c_0$, for some $c_0 \in (0, \infty)$.

(ii) *There exists a positive sequence $\tilde{w} = \{\tilde{w}_n\}_{n=1}^{\infty}$ and a corresponding space $\mathcal{H}_{\tilde{w}}$ as in* (3), *with Hilbert space norm $\| \cdot \|_{\mathcal{H}_{\tilde{w}}}$, which constitutes a RKHS continuously embedded in $C(\mathcal{D})$ such that for some positive constant $C_{\tilde{w}}$,*

$$\|T(f)\|_{\mathcal{H}_{\tilde{w}}} \le C_{\tilde{w}} \|f\|_{\mathcal{H}_{\tilde{w}}} \qquad \forall f \in \mathcal{H}_{\tilde{w}}.$$

In other words, T is bounded when restricted to $\mathcal{H}_{\tilde{w}}$.

The forward operator T defines our data-generating process. In particular, let $x_i \in \mathcal{D}$, $i = 1, \ldots, m$ be the design points. Also under P_0, let $\epsilon^0 = (\epsilon_1^0, \ldots, \epsilon_m^0)$ be a vector of errors. Then we observe a single, noisy path of u at the design points, i.e.,

$$Y_i = u^0(x_i) + \epsilon_i^0, \quad i = 1, \ldots, m,$$

where $u^0(x_i) = T(b^0)(x_i)$, $i = 1, \ldots, m$ are point evaluations of a single sample path. We are now ready to state the core assumption about the nominal distribution P_0.

ASSUMPTION 2.4 (Independence).   *We assume that under the nominal distribution P_0, ϵ^0 is a vector of independent $\mathcal{N}(0, \sigma^2)$ errors, and that b^0 and ϵ^0 are independent. Consequently, under P_0, the pair (b^0, ϵ^0) constitutes a Gaussian random variable on the product space $C(\mathcal{D}) \times \mathbb{R}^m$, which is Banach with the product norm.*

We will see examples in Section 4 where Assumption 2.3 on the forward operator is satisfied; these include Gaussian process regression (where $T = \mathrm{Id}$) and several canonical linear inverse problems. Assumption 2.3 entails certain compatibility of the forward operator with the prior basis $\{e_n\}_{n=1}^{\infty}$, which is similar to the assumption of "norm equivalence on regularity scales" in the literature on *linear* Bayesian inverse problems [21, 2]. We note that other assumptions in the literature exist, e.g., the "band-limited" assumption in [34]. In our framework, we will consider Wasserstein perturbations to the prior P_0, which encompass a much richer collection of prior families than typically considered in the literature.

2.1. *The Nominal Inverse Problem.*   Let $\mathcal{M}$ denote the space of measurable maps from the data space $\mathbb{R}^m$ to the signal space $C(\mathcal{D})$ and let $\mathcal{P}$ denote the space of (Borel) probability measures on $C(\mathcal{D}) \times \mathbb{R}^m$. The classical Bayes risk minimization for $L^2(\mathcal{D})$ loss is

$$(4) \qquad \min_{\phi \in \mathcal{M}} \mathbb{E}_{P_0} \left[\|b^0 - \phi(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2 \right],$$

where $\phi(Y_1, \ldots, Y_m)$ is the estimator for b^0 given observations $(Y_1, \ldots, Y_m)$. We regard the problem as the nominal estimation problem since the Bayes risk is evaluated under the nominal measure P_0. The solution of the nominal problem is the posterior mean (or conditional expectation), i.e.,

$$\phi_0(Y_1, \ldots, Y_m)(x) = \mathbb{E}_{P_0} \left[b^0(x) | Y_1, \ldots, Y_m \right].$$

This follows by noting that the $L^2(\mathcal{D})$-norm is a Lebesgue integral and by using Fubini's theorem. Since we assume that the nominal distribution P_0 is Gaussian, this estimator corresponds to the linear prediction rule,

$$\mathbb{E}_{P_0} \left[b^0(x) | Y_1, \ldots, Y_m \right] = \left(k^0(x, x_1), \ldots, k^0(x, x_m) \right) \cdot (K^0)^{-1} \cdot (Y_1, \ldots, Y_m)^{\top},$$

where $k^0(x, x_i) = \mathbb{E}_{P_0}[b^0(x) Y_i]$, $K_{ij}^0 = (\mathbb{E}_{P_0}[Y_i Y_j])$, and $K^0 \in \mathbb{S}_{++}^m$, where $\mathbb{S}_{++}^m$ denotes the set of strictly positive definite matrices. The invertibility of K^0 can be directly deduced from Assumption 2.4.

2.2. *The Distributionally Robust Optimization Inverse Problem.* Instead of considering the nominal measure P_0 on the prior and noise distributions, we postulate a min-max game where an adversary chooses a measure P in opposition to the decision-maker's choice of estimator. In particular, for some forward map T describing the relationship between the parameter b and the function $u = T(b)$, we assume that the observations $Y_1, \ldots, Y_m$ arise as

$$
(5) \qquad \begin{cases} (b, \epsilon_1, \ldots, \epsilon_m) \sim P, \\ Y_1 = u(x_1) + \epsilon_1, \ldots, Y_m = u(x_m) + \epsilon_m, \end{cases}
$$

where $u(x_1), \ldots, u(x_m)$ are point evaluations of a single sample path. Note that we consider a fixed-design setting where $x_1, \ldots, x_m$ is fixed throughout. Instead of the nominal Bayes risk, we consider a 'worst-case Bayes risk' with respect to all possible mis-specification on both b and ϵ, i.e., the whole data-generating process. Our goal is to minimize the worst-case Bayes risk by solving

$$
(6) \qquad \inf_{\phi \in \mathcal{M}} \sup_{P \in \mathcal{P}, \mathbb{W}(P, P_0) \leq \delta} \mathbb{E}_P \left[\| b - \phi(Y_1, \ldots, Y_m) \|^2_{L^2(\mathcal{D})} \right],
$$

where the adversary's admissible choice of P is constrained by the Wasserstein distance $\mathbb{W}(P, P_0)$ relative to the nominal measure P_0.

The Wasserstein distance $\mathbb{W}$ is constructed based on the optimal transport cost between P and P_0, which is defined as follows.

DEFINITION 2.1 (Optimal transport cost). *The optimal transport cost between the probability measures P and P_0 on $C(\mathcal{D}) \times \mathbb{R}^m$ is defined as*

$$
(7) \qquad D_c(P, P_0) = \inf_\pi \left\{ \mathbb{E}_\pi [c((b, \epsilon), (b^0, \epsilon^0))] \ : \ \pi_{(b, \epsilon)} = P, \ \pi_{(b^0, \epsilon^0)} = P_0 \right\},
$$

where the infimum is taken over all couplings π between (b, ϵ) and (b^0, ϵ^0) with marginals P and P_0, and c is some ground cost on $C(\mathcal{D}) \times \mathbb{R}^m$.

Here, our definition of optimal transport cost follows the general literature, see [45], with specification that the probability measure consists of a process component (namely b) and a finite-dimensional vector component (namely ϵ). The existence of an optimal coupling (i.e., an optimal solution to problem (7)) is guaranteed whenever c is non-negative and lower semi-continuous with respect to the product norm on the Polish space $(C(\mathcal{D}) \times \mathbb{R}^m)^2$; see e.g., [45, Theorem 4.1]. The particular choice of c is typically motivated by the geometric properties of the problem at hand. For instance, the geometry of finite-dimensional spaces differs from that of infinite-dimensional spaces. In this paper, we use the ground cost function c defined as

$$
c((b, \epsilon), (b^0, \epsilon^0)) = \| \epsilon - \epsilon^0 \|^2_2 + \| b - b^0 \|^2_{\mathcal{H}_w}
$$

$$
= \sum_{i=1}^m (\epsilon_i - \epsilon_i^0)^2 + \sum_{n \geq 1} \left(\langle b, e_n \rangle - \langle b^0, e_n \rangle \right)^2 w_n.
$$

To see that c is lower semi-continuous, note that if $\| b_k - b_\infty \|_{C(\mathcal{D})} \to 0$ and $\| \tilde{b}_k - \tilde{b}_\infty \|_{C(\mathcal{D})} \to 0$ as $k \to \infty$, then $\liminf_{k \to \infty} \left(\langle b_k, e_n \rangle - \langle \tilde{b}_k, e_n \rangle \right)^2 w_n \geq \left(\langle b_\infty, e_n \rangle - \langle \tilde{b}_\infty, e_n \rangle \right)^2 w_n$ for each n since $e_n \in C(\mathcal{D})$. Thus, we have

$$
\liminf_{k \to \infty} \sum_{n \geq 1} \left(\langle b_k, e_n \rangle - \langle \tilde{b}_k, e_n \rangle \right)^2 w_n \geq \sum_{n \geq 1} \left(\langle b_\infty, e_n \rangle - \langle \tilde{b}_\infty, e_n \rangle \right)^2 w_n.
$$

With this choice of the ground cost c, the optimal coupling between P and P_0 exists for any P and P_0. Note that D_c is not a distance because it does not satisfy the triangle inequality, but its square root $\sqrt{D_c}$ is a Wasserstein-type distance on its domain of finiteness $\{P \in \mathcal{P} : D_c(P, P_0) < \infty\}$ [45, Definition 6.1]. Throughout this paper, we will use

$$\mathbb{W} := \sqrt{D_c}.$$

Our choice of the ground cost function c penalizes discrepancies in both the observation error component (with a usual l_2 norm in finite-dimensional Euclidean space) and the stochastic process component (with a Hilbert space norm), which to our knowledge, is novel in the literature on DRO. Allowing the adversary to modify the distribution of the additive noise ϵ_i can also be understood in part as compensating for the misspecification of the nominal forward operator T [22]. One technical reason that we chose optimal transport cost instead of e.g. relative entropy or any other criterion that requires the existence of a likelihood ratio is that P and P_0 can have different support, which is needed when dealing with non-parametric models and permits adversarial distributions with different smoothness properties of the sample paths.

It is important to stress that our definition of the Wasserstein distance $\mathbb{W}$ depends on the specification of the function class $\mathcal{H}_w$, or equivalently on the Hilbert space norm $\|\cdot\|_{\mathcal{H}_w}$. This dependence provides a convenient tool to control features such as the amount of roughness or smoothness that the adversary is allowed to inject in the sample path of the process b (and hence of u). Intuitively speaking, the sequence w puts different penalties on the mass transportation of different "modes" of the spectral decomposition of the sample path of b. For example, if the sequence w increases to infinity slowly, the adversary is under-penalized for moving mass corresponding to the higher "modes," resulting in rougher sample paths of b. The modeling of the adversary's behavior thus conveniently reduces to the specification of the Hilbert norm $\|\cdot\|_{\mathcal{H}_w}$.

We impose a final compatibility assumption between the operator T and the adversarial cost introduced.

ASSUMPTION 2.5 (Operator and adversarial cost).   *We select $\tilde{w}$ in Assumption 2.3 such that $\tilde{w}_n = o(w_n)$ as $n \to \infty$.*

Intuitively, Assumption 2.5 says that the operator is bounded even if the adversarial perturbations are slightly rougher than the adversarial choice. This assumption, we believe, is purely technical. The natural condition to impose is that the operator is bounded only on the chosen adversarial space. Our results hold under this more natural (and weaker) assumption in the case of standard Gaussian process regression, namely when T equals the identity map.

REMARK 2.1.   *Several comments are in order regarding our framework:*

(i) *It is natural to consider random elements on a general Banach space $\mathbb{B}$ other than $C(\mathcal{D})$, e.g., by embedding the Banach space $\mathbb{B}$ in its second dual $\mathbb{B}^{\star\star}$ and identifying a Borel measurable random element b in $\mathbb{B}$ with the stochastic process $(b^\star(b) : b^\star \in \mathbb{B}^\star)$, but at the expense of additional technicality, see a discussion in [44, Section 2.3]. We choose $C(\mathcal{D})$ to illustrate our main conceptual contribution of a distributionally robust formulation of nonparametric regression and inverse problems.*
(ii) *The $L^2(\mathcal{D})$ norm in the objective function of the formulation (6) can be replaced by another member in the hierarchy of the Hilbert space norms. However, the latter norm lacks the Lebesgue integral representation, especially for those with fractional power [42, Section 4.1], and we leave the extension to future work.*

(iii) *It is tempting to replace the Hilbert space norm with the $L^2(\mathcal{D})$ norm in the definition of the ground cost function c. However, point evaluations are not continuous under the $L^2(\mathcal{D})$ norm. Our proofs for the main results rely crucially on the continuity of point evaluations, and thus, we resort to the RKHS in Assumption 2.3.*

3. Main Results.

3.1. *Strong Duality.* Our first main result is a minimax theorem, which states that one may interchange the infimum and supremum operators in the inverse problem (6).

THEOREM 3.1 (Strong duality for the inverse problem). *Suppose that Assumptions 2.1–2.5 hold. For any $\delta > 0$, strong duality holds, that is,*

$$
\inf_{\phi \in \mathcal{M}} \sup_{P \in \mathcal{P}, \mathbb{W}(P, P_0) \leq \delta} \mathbb{E}_P \left[\|b - \phi(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2 \right]
$$

$$
(8) \qquad = \sup_{P \in \mathcal{P}, \mathbb{W}(P, P_0) \leq \delta} \inf_{\phi \in \mathcal{M}} \mathbb{E}_P \left[\|b - \phi(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2 \right].
$$

The full proof of Theorem 3.1 is deferred to Appendix A. To prove Theorem 3.1, our technical contribution consists of a novel asymptotic approximation technique. First, the left hand side of (8) can be trivially upper bounded by the min-max problem constraining ϕ as *affine estimators* (see Definition A.1), namely

$$
\inf_{\phi \in \mathcal{M} \text{ and affine}} \sup_{P \in \mathcal{P}, \mathbb{W}(P, P_0) \leq \delta} \mathbb{E}_P \left[\|b - \phi(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2 \right]
$$

$$
(9) \qquad \geq \inf_{\phi \in \mathcal{M}} \sup_{P \in \mathcal{P}, \mathbb{W}(P, P_0) \leq \delta} \mathbb{E}_P \left[\|b - \phi(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2 \right].
$$

Second, define $\text{span}\{e_n\}_{n=1}^N$ as the (closed) linear subspace of $C(\mathcal{D})$ spanned by the basis vectors $(e_n : 1 \leq n \leq N)$, the right hand side of (8) can be lower bounded by constraining the ambiguity set, namely

$$
\sup_{P \in \mathcal{P}, \mathbb{W}(P, P_0) \leq \delta} \inf_{\phi \in \mathcal{M}} \mathbb{E}_P \left[\|b - \phi(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2 \right]
$$

$$
(10) \qquad \geq \sup_{P \in \mathcal{W}_N(\delta)} \inf_{\phi \in \mathcal{M}} \mathbb{E}_P \left[\|b - \phi(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2 \right],
$$

where $\mathcal{W}_N(\delta)$ denotes the Wasserstein balls around P_0 arising from perturbations constrained in the first N coordinates (see equation (16) in the appendix). In the end, we argue that the left hand side of the relaxation (9) and the asympototic limit of the right hand side of the relaxation (10) coincides, and since they bound problem (8) the strong duality follows.

To proceed, we first show a strong duality result for a sequence of finite-dimensional approximations. In particular, we consider truncating P and P_0 into the space $\text{span}\{e_n\}_{n=1}^N \times \mathbb{R}^m$, and denote the induced measure as $Q^{(N)}$ and $Q_0^{(N)}$, respectively. The truncation is through the coordinate projections after expanding functions in the basis $\{e_n\}_{n=1}^\infty$. Since the coordinate projections are bounded linear mappings, $Q_0^{(N)}$ is centered Gaussian. Notice that the space $\text{span}\{e_n\}_{n=1}^N \times \mathbb{R}^m$ is isomorphic to $\mathbb{R}^{N+m}$, thus we view the truncated measures $Q^{(N)}$ and $Q_0^{(N)}$ as finite-dimensional measures on $\mathbb{R}^{N+m}$. This reduction to finite-dimensional setting enables the use of existing results (e.g., Sion's minimax theorem [37]).

For convenience, denote

$$\mathrm{Obj}(\phi, P) = \mathbb{E}_P\left[\|b - \phi(Y_1, \ldots, Y_m)\|^2_{L^2(\mathcal{D})}\right],$$

and by a slight abuse of notation we denote $\mathrm{Obj}(\phi, Q^{(N)})$ for the truncated measure. The related strong duality in finite-dimensions reads

(11)
$$\min_{\phi \in \mathcal{M}} \max_{Q^{(N)}:\mathrm{W}_N(Q^{(N)}, Q_0^{(N)}) \leq \delta} \mathrm{Obj}(\phi, Q^{(N)}) = \max_{Q^{(N)}:\mathrm{W}_N(Q^{(N)}, Q_0^{(N)}) \leq \delta} \min_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, Q^{(N)}),$$

where W_N^2 is the induced optimal transport cost on $\mathbb{R}^{N+m}$

$$\mathrm{W}_N^2(Q^{(N)}, Q_0^{(N)}) = \min_{\pi}\left\{\mathbb{E}_\pi[c_N(r,s)] : \pi_r = Q^{(N)}, \pi_s = Q_0^{(N)}\right\}.$$

Here π_r and π_s are projections onto the first and second components of the coupling π. In the definition of W_N^2, c_N is the induced cost function on $\mathbb{R}^{N+m}$ with

$$c_N(r,s) = \sum_{n=1}^N (r_n - s_n)^2 w_n + \sum_{j=1}^m (r_{N+j} - s_{N+j})^2 \quad \text{for any } r,\ s \in \mathbb{R}^{N+m},$$

where recall that w_n are the parameters of the Hilbert space norm $\|\cdot\|_{\mathcal{H}_w}$. We note that in (11), the minimizer in the ϕ variable and the maximizer in the $Q^{(N)}$ variable exist, which justifies the min and the max operators. Denote by $\phi_N^\star$ and $Q_N^\star$ the (unique) solution to the outer optimization problem in the left-hand and the right-hand side of (11), respectively. Then $(\phi_N^\star, Q_N^\star)$ is the (unique) pair of Nash equilibrium for problem (11) in the sense that

$$\mathrm{Obj}(\phi_N^\star, Q_N^\star) = \min_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, Q_N^\star) = \max_{Q^{(N)}:\mathrm{W}_N(Q^{(N)}, Q_0^{(N)}) \leq \delta} \mathrm{Obj}(\phi_N^\star, Q^{(N)}).$$

The structural properties of the optimal solutions reveal that $Q_N^\star$ is a centered Gaussian distribution, and $\phi_N^\star$ is a linear prediction rule (see details in the full proof). Namely, for any $x \in \mathcal{D}$,

$$\phi_N^\star(Y_1, \ldots, Y_m)(x) = \mathbb{E}_{Q_N^\star}\left[b(x)\big|Y_1, \ldots, Y_m\right]$$
$$= \left(k_\epsilon^{(N)}(x, x_1), \ldots, k_\epsilon^{(N)}(x, x_m)\right) \cdot (K_\epsilon^{(N)})^{-1} \cdot (Y_1, \ldots, Y_m)^\top,$$

where $k_\epsilon^{(N)}(x, x_i) = \mathbb{E}_{Q_N^\star}[b(x)Y_i]$, and $K_\epsilon^{(N)} = \left(\mathbb{E}_{Q_N^\star}[Y_iY_j]\right)_{ij} \in \mathbb{S}_{++}^m$ is invertible. Similar finite-dimensional duality results have been established in [31, 36], but with a different definition of the Wasserstein distance and the ambiguity set.

Having established the strong duality in finite-dimensions, the rest of the proof then argues that the left hand side of the relaxation (9) can be upper bounded by the value of the finite-dimensional min-max problem with an error of $o(1)$, and the right hand side of the relaxation (10) can be lower bounded by the value of the finite-dimensional max-min problem. Therefore, as the dimension N grows to infinity, the approximation error becomes negligible. This limiting argument is a key technical novelty, which is achieved by a careful analysis of the tail modes of the spectral decomposition. We state one intermediate result that guarantees the approximation of objective values.

PROPOSITION 3.1 (Approximation of objective values). *Let $\phi_N^\star$ be the (unique) solution to the min-max problem in (11), and $Q_N^\star$ be the (unique) solution to the max-min problem in (11). We have*

$$\mathrm{Obj}(\phi_N^\star, Q_N^\star) = \sup_{P \in \mathcal{P}, \mathrm{W}(P, P_0) \leq \delta} \mathrm{Obj}(\phi_N^\star, P) + o(1) = \inf_{\phi \in \mathcal{M}} \sup_{P \in \mathcal{P}, \mathrm{W}(P, P_0) \leq \delta} \mathrm{Obj}(\phi, P) + o(1),$$

asymptotically as the number of basis vectors in the approximation tends to infinity, i.e., as $N \to \infty$.

3.2. *Existence, Uniqueness, and Construction of the Nash Equilibrium.* In this part, we show that (8) admits a unique pair of Nash equilibrium under certain conditions. Recall the Nash equilibrium $(\phi_N^\star, Q_N^\star)$ corresponding to the finite-dimensional approximation in the last section. We now add the tail of P_0 to $Q_N^\star$, namely, we denote $P_N^\star \in \mathcal{P}$ as the measure under which the two random elements

(12)
$$(\{\langle b, e_n \rangle\}_{n=1}^N, \epsilon) \quad \text{and} \quad \{\langle b, e_n \rangle\}_{n=N+1}^\infty$$

are independent. Moreover, we have that

$$\mathcal{L}^{P_N^\star}\big(\{\langle b, e_n \rangle\}_{n=N+1}^\infty\big) = \mathcal{L}^{P_0}\big(\{\langle b^0, e_n \rangle\}_{n=N+1}^\infty\big),$$

where $\mathcal{L}^P$ denotes the law under P. We can extract a subsequence from $P_N^\star$, $N \geq 1$ by a compactness argument.

PROPOSITION 3.2 (Compactness of the ambiguity set). *Under the conditions of Theorem 3.1, for every sequence $P_N \in \mathcal{P}, N \geq 1$ that satisfies $\mathbb{W}(P_N, P_0) \leq \delta$, there exists a weakly convergent subsequence $P_{N_l}, l \geq 1$ with $P_{N_l} \Rightarrow P_\infty$, such that the limit $P_\infty \in \mathcal{P}$ also satisfies $\mathbb{W}(P_\infty, P_0) \leq \delta$.*

By Proposition 3.2, we find a weakly convergent subsequence of $P_N^\star$, denoted as $P_{N_l}^\star$, $l \geq 1$, with a limit $P_\infty^\star$ that belongs to the feasible set. The subsequence $P_{N_l}^\star$ is centered Gaussian, thus the limit $P_\infty^\star$ is also centered Gaussian. To see this, consider any bounded linear functional $F : C(\mathcal{D}) \times \mathbb{R}^m \to \mathbb{R}$, and construct the Skorohod representations $Z_{N_l} \sim P_{N_l}^\star$ and $Z_\infty \sim P_\infty^\star$, where Z_{N_l} converges to Z_∞ almost surely [38] on the common probability space. It follows that $F(Z_{N_l})$ converges to $F(Z_\infty)$ almost surely, and we note that the limit of centered univariate Gaussians must also be centered univariate Gaussian.

Define the matrix $K_{\epsilon, P_\infty^\star} = \big(\mathbb{E}_{P_\infty^\star}[Y_i Y_j]\big)_{ij} \in \mathbb{S}_+^m$. Under the condition that $K_{\epsilon, P_\infty^\star}$ is invertible, the solution $\phi_\infty^\star$ to $\min_\phi \mathrm{Obj}(\phi, P_\infty^\star)$ is well-defined as

(13)
$$\phi_\infty^\star(Y_1, \ldots, Y_m)(x) = \mathbb{E}_{P_\infty^\star}[b(x)|Y_1, \ldots, Y_m]$$
$$= \big(k_{\epsilon, P_\infty^\star}(x, x_1), \ldots, k_{\epsilon, P_\infty^\star}(x, x_m)\big) \cdot (K_{\epsilon, P_\infty^\star})^{-1} \cdot (Y_1, \ldots, Y_m)^\top,$$

where $k_{\epsilon, P_\infty^\star}(x, x_i) = \mathbb{E}_{P_\infty^\star}[b(x)Y_i]$. The main result of this section is the following theorem.

THEOREM 3.2 (Nash equilibrium). *Suppose that the conditions of Theorem 3.1 hold. Let $P_\infty^\star$ denote the limit of a weakly convergent subsequence of $P_N^\star, N \geq 1$, and assume that $K_{\epsilon, P_\infty^\star}$ is invertible under $P_\infty^\star$. Then, for $\phi_\infty^\star$ given by (13), the pair $(\phi_\infty^\star, P_\infty^\star)$ forms a Nash equilibrium to problem (8), i.e.,*

$$\mathrm{Obj}(\phi_\infty^\star, P_\infty^\star) = \min_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P_\infty^\star) = \max_{P \in \mathcal{P}: \mathbb{W}(P, P_0) \leq \delta} \mathrm{Obj}(\phi_\infty^\star, P).$$

Thus, $(\phi_\infty^\star, P_\infty^\star)$ represents a pair of equilibrium strategies where neither the decision-maker nor the adversary, the two players of the zero-sum game, benefits from changing their strategy. Moreover, the pair $(\phi_\infty^\star, P_\infty^\star)$ is unique, with components respectively given by the pointwise limit

$$\lim_{N \to \infty} \phi_N^\star(Y_1, \ldots, Y_m)(x) = \phi_\infty^\star(Y_1, \ldots, Y_m)(x) \quad \forall Y_1, \ldots, Y_m \in \mathbb{R}, \quad x \in \mathcal{D},$$

and the weak limit $P_N^\star \Rightarrow P_\infty^\star$ as $N \to \infty$.

One important consequence of Theorem 3.2 is that the worst-case distribution involves a modified Gaussian process with potentially rougher paths than the prior. Hence, the robustified decision, as the conditional mean of the worst-case distribution, remains affine in the observations and, therefore, is tractable. We formally state the following theorem.

THEOREM 3.3 (Nash equilibrium property).   *Under the notation of Theorem 3.2, the robustified decision $\phi_\infty^\star$ is an affine function in the observations, and the worst-case distribution $P_\infty^\star$ is a Gaussian process.*

The convergence statement in Theorem 3.2 readily gives rise to an algorithm for computing the associated Nash equilibrium.

The result of Theorem 3.2 requires that $K_{\epsilon,P_\infty^\star}$ is invertible for the limit of some weakly convergent subsequence of $P_N^\star$. Fortunately, due to the following proposition, it is not difficult to check whether this condition holds in practice.

PROPOSITION 3.3.   *Either one (and only one) of the following cases occurs.*

1. *there exists a weakly convergent subsequence of $P_N^\star$, with the limit denoted by $P_\infty^\star$, such that the matrix $K_{\epsilon,P_\infty^\star}$ is invertible under $P_\infty^\star$, or*
2. *the sequence of determinants $det(K_\epsilon^{(N)}) \to 0$ as $N \to \infty$.*

To conclude this section, we provide a sufficient condition to ensure that the first case of Proposition 3.3 occurs.

LEMMA 3.4 (Invertibility).   *There exists a strictly positive constant δ_0 that depends on $(T, m, (x_i)_i, \mathcal{H}_w, \mathcal{H}_{\tilde{w}}, \sigma^2)$ such that for any $\delta < \delta_0$ and P satisfying $\mathbb{W}(P, P_0) \leq \delta$, the matrix $(\mathbb{E}_P[Y_i Y_j])_{ij}$ is invertible.*

3.3. *Strong Duality for the Regression Problem.*   Problem (4) is a linear inverse problem where our primary interest lies in recovering the unknown input b. Alternatively, we mention in passing a distributionally robust formulation for regressing $u = T(b)$ given noisy pointwise observations (i.e., a regression problem). Under the observation system (5), the goal of the decision-maker is to seek a nonparametric predictor $\phi_u \in \mathcal{M}$ that minimizes the worst-case objective

$$(14) \qquad \inf_{\phi_u \in \mathcal{M}} \sup_{P \in \mathcal{P}, \mathbb{W}(P, P_0) \leq \delta} \mathbb{E}_P\left[\|u - \phi_u(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2\right],$$

where nature's admissible choice of P is constrained by the Wasserstein distance $\mathbb{W}(P, P_0)$ constructed from (7). We state the strong duality of (14).

THEOREM 3.5 (Strong duality for the regression problem).   *Suppose that Assumptions 2.1–2.5 hold. For any $\delta > 0$,*

$$\inf_{\phi_u \in \mathcal{M}} \sup_{P \in \mathcal{P}, \mathbb{W}(P, P_0) \leq \delta} \mathbb{E}_P\left[\|u - \phi_u(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2\right]$$

$$(15) \qquad = \sup_{P \in \mathcal{P}, \mathbb{W}(P, P_0) \leq \delta} \inf_{\phi_u \in \mathcal{M}} \mathbb{E}_P\left[\|u - \phi_u(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2\right].$$

The proof of Theorem 3.5 works verbatim as that of Theorem 3.1. Though strong duality holds for both the regression and the inverse problems under our formulations, we note that

for ill-posed inverse problems in the Bayesian nonparametrics literature, the minimax rate for estimating b is slower than the minimax rate for estimating u [14, 10, 11, 23].

Regarding the Nash equilibrium associated with (15), it is not hard to see, after examining the proof of Theorem 3.2, that we can develop the same theory verbatim to that of Section 3.2. For ease of exposition, we suppress the details here.

4. Some Examples. In this section, we give several examples that illustrate the applicability of our general framework. We restrict our attention to Matérn process priors and the Sobolev-type space of perturbations given by Examples 2.1 and 2.2, respectively. We assume the relation $\kappa_n = \lambda_n^{-\alpha/2}$ (so that $\kappa = 0$ in Example 2.1) and $w_n = \lambda_n^{\beta}$, where $\alpha > \frac{d}{2}$ and $\beta > \frac{d}{2}$.

EXAMPLE 4.1 (Gaussian process regression). *By choosing T as the identity operator, we recover Gaussian process regression. Assumptions 2.3–2.5 are satisfied for $\tilde{w}_n = \lambda_n^{\tilde{\beta}}$ and any $\frac{d}{2} < \tilde{\beta} < \beta$. If $\mathcal{D}$ is the one-dimensional interval $[0,1]$, the eigenvalues are $\lambda_n = n^2\pi^2$, and the eigenfunctions are*

$$e_n(x) = \sqrt{2}\sin(n\pi x) \quad \forall x \in [0,1].$$

EXAMPLE 4.2 (Laplace equation). *The Laplace equation with a homogeneous Dirichlet boundary condition is*

$$\begin{cases} \triangle u(x) = b(x) & \forall x \in \mathcal{D}^o, \\ u(x) = 0 & \forall x \in \partial\mathcal{D}, \end{cases}$$

where $\mathcal{D}^o$ is the interior of $\mathcal{D}$, while $\partial\mathcal{D}$ denotes its boundary. We have that the forward map T is the inverse-Laplacian operator, thus,

$$T(f) = \sum_{n \geq 1} -\lambda_n^{-1}\langle f, e_n\rangle e_n.$$

It is straightforward to see that Assumptions 2.3–2.5 are satisfied for $\tilde{w}_n = \lambda_n^{\tilde{\beta}}$ and any $\frac{d}{2} < \tilde{\beta} < \beta$. Similarly, one could treat other elliptic inverse problems.

EXAMPLE 4.3 (Heat equation). *The one-dimensional homogeneous heat equation without source is*

$$\begin{cases} u_t = u_{xx} & 0 < x < 1, \\ u(x,0) = b(x) & 0 < x < 1, \\ u(0,t) = u(1,t) = 0 & t \geq 0, \end{cases}$$

where $u(\cdot, t)$ is the temperature profile at time t, and b is the initial condition. By separation of variables, the solution to the heat equation is

$$u(x,t) = \sum_{n \geq 1} e^{-n^2\pi^2 t}\langle b, e_n\rangle e_n.$$

Therefore, the (time-dependent) forward map T satisfies, for any $t \geq 0$,

$$T(f) = \sum_{n \geq 1} e^{-n^2\pi^2 t}\langle f, e_n\rangle e_n.$$

Assumptions 2.3–2.5 are satisfied for $\tilde{w}_n = \lambda_n^{\tilde{\beta}}$ and any $\frac{d}{2} < \tilde{\beta} < \beta$.

EXAMPLE 4.4 (Radon transform in the plane).   *The Radon transform of a function f is the function*

$$T(f)(s,\omega) = \int_{-\infty}^{\infty} f(s\omega + t\omega^{\perp})\mathrm{d}t, \quad s \in \mathbb{R}, \ \omega \in \mathcal{S}^1,$$

where $\mathcal{S}^1$ is the unit circle and $\omega^{\perp}$ is the vector in $\mathcal{S}^1$ obtained by rotating ω counterclockwise by $90°$. Recall we consider f to be supported in a compact domain $\mathcal{D} \subset \mathbb{R}^2$, and thus $T(f)$ vanishes outside a compact subset of $\mathbb{R} \times \mathcal{S}^1$. It is straightforward to see that $T : C(\mathcal{D}) \to C(\mathbb{R} \times \mathcal{S}^1)$ is linear and bounded, where we allow a slight modification of our framework since T takes value in a different space from its domain of definition. By [29, Theorems II.5.1 and II.5.2], the Radon transform satisfies the following Sobolev estimate:

$$\|T(f)\|_{H^{\tilde{\beta}}(\mathbb{R}\times\mathcal{S}^1)} \leq C_{\tilde{w}}\|f\|_{\mathcal{H}_{\tilde{w}}} \quad \forall f \in \mathcal{H}_{\tilde{w}},$$

for $\tilde{w}_n = \lambda_n^{\tilde{\beta}}$ and any $1 = \frac{d}{2} < \tilde{\beta} < \beta$, where $H^{\tilde{\beta}}(\mathbb{R} \times \mathcal{S}^1)$ is the usual order-$\tilde{\beta}$ Sobolev space. Identifying $\mathcal{S}^1$ with $[0, 2\pi)$, and by the Sobolev embedding theorem [42, Proposition 4.1.3], we see that point evaluations in $H^{\tilde{\beta}}(\mathbb{R} \times \mathcal{S}^1)$ are continuous. Our theory in Section 3 applies with this slight modification of Assumptions 2.3–2.5, after inspecting the proofs.

5. Numerical Experiments.   Our focus in this paper is formulating a min-max framework for regression and linear inverse problems in an infinite-dimensional setting and elucidating key theoretical properties of this formulation. We now present some numerical experiments that offer further insights into the properties of the Nash equilibrium. In particular, we compute the Nash equilibrium $(\phi_N^\star, Q_N^\star)$ of a finite-dimensional approximation of the robust estimation problem (see definitions in Section 3) with $N = 200$, by an adaptation of the Frank-Wolfe algorithm in [36]. The concrete algorithm as developed in [36] and the details of our adaptation is discussed in Appendix A.9. Under the conditions of Theorem 3.2, the use of the finite-dimensional Nash equilibrium is justified. When the condition fails, Proposition 3.1 still guarantees that the objective values of the finite-dimensional games converge to those of the *in*finite-dimensional games. We illustrate our results for Gaussian process regression on the unit interval $\mathcal{D} = [0, 1]$, as described in Example 4.1. Throughout this section we set $\kappa_n = \lambda_n^{-\alpha/2}$ and $w_n = \lambda_n^{\beta}$. Recall that $(\kappa_n)_{n\geq 1}$, and hence α, control the smoothness of the nominal prior, while $(w_n)_{n\geq 1}$, and hence β, control the roughness of the adversarial perturbations to the prior via the $\|\cdot\|_{\mathcal{H}_w}$ component of the transport cost.

First we fix a set of baseline parameters: $\alpha = 2$ and $\beta = 0.51$; $\delta^2 = 0.1$ (size of the Wasserstein ball); and $\sigma = 0.1$ (observational noise magnitude). We choose $m = 10$ design points $(x_i)_{i=1}^{10}$ equispaced on either the $(0, 1)$ or $(0, 0.5)$ intervals, excluding the endpoints. More specifically, we choose $x_i = i/11$ and $x_i = i/22$ respectively. Below we visualize the prior and posterior covariances of b under the nominal and worst-case measures $Q_0^{(N)}$ and $Q_N^\star$; i.e., $\mathbb{C}\mathrm{ov}_P[b]$ and $\mathbb{C}\mathrm{ov}_P[b \,|\, y_1, \ldots, y_m]$ for $P \in \{Q_0^{(N)}, Q_N^\star\}$. Since both the nominal and the worst-case measures are Gaussian, the posterior covariances do not depend on the realization of the data.

We first visualize the four correlation functions corresponding to these covariances for our two designs in Figures 1 and 2. The geometry of each design is evident in the worst-case measures. Compared to the nominal measures, we observe "ripples" in the worst-case measures corresponding to reductions of correlation between the observed locations. Next, in Figures 3 and 4, we plot marginal intervals (at each x) containing 95% of the nominal and the worst-case sample paths. Data to obtain the two posteriors shown here were drawn from the nominal prior measure. In particular, we used the vector of observations
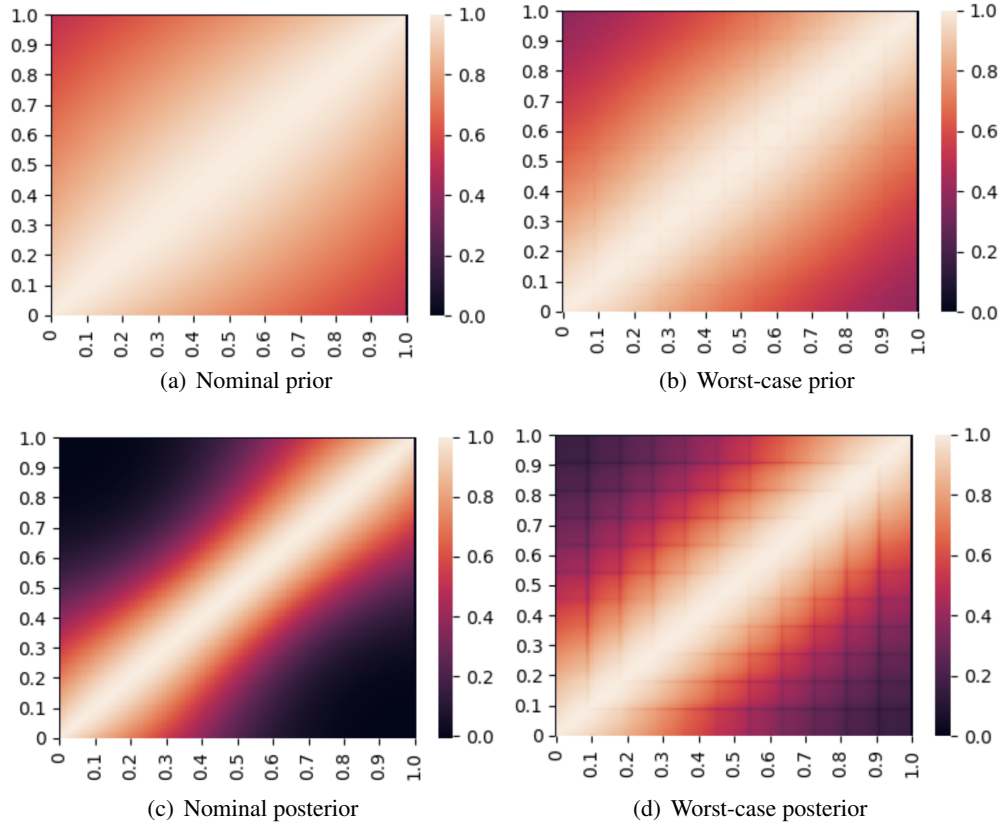
(a) Nominal prior   (b) Worst-case prior

(c) Nominal posterior   (d) Worst-case posterior

FIG 1. *Correlation functions on $[0,1]^2$ with 10 design points equispaced on $(0,1)$.*

$$(-0.17, -0.09, 0.02, 0.04, 0.12, 0.05, -0.03, 0.03, -0.28, -0.15)$$

for the $(0,1)$ design and

$$(0.03, -0.05, 0.08, -0.08, 0.15, 0.12, -0.25, -0.24, 0.16, 0.02)$$

for the $(0, 0.5)$ design. Comparing to the nominal prior measures, we observe that the worst-case prior measures have roughly the same overall variance magnitude, but a sharper contrast between the observed and unobserved locations, especially for the $(0, 1)$-equispaced designs. On the other hand, comparing to the nominal posterior measures, we observe that the worst-case posterior measures have significantly higher marginal variances in regions away from the observed locations, while the variance increase moderately in regions surrounding the observations. The worst cases are thus perturbed to induce greater uncertainty in regions where information is limited; intuitively, this guarantees greater robustness of the estimates.

Next, we vary the baseline parameters to see the effect of the worst-case perturbations on (prior and posterior) sample paths of b compared to sample paths of b under the nominal measures. We focus on the $(0, 1)$-equispaced design. For each parameter setting, we draw and visualize five independent sample paths to gauge their qualitative behavior. The posterior sample paths are conditioned on the same observation values as before. As another way of quantifying the impact of the worst-case perturbation, we compute the distance between the prior and posterior covariance matrices on $\mathbb{S}_{++}^N$, where these matrices are induced by either the nominal or the worst-case measure. The distance we employ is the natural geodesic distance on the manifold of symmetric positive-definite matrices, also known as the Förstner distance [16] and (up to a constant) Rao's distance [3, 33]. This distance is invariant under
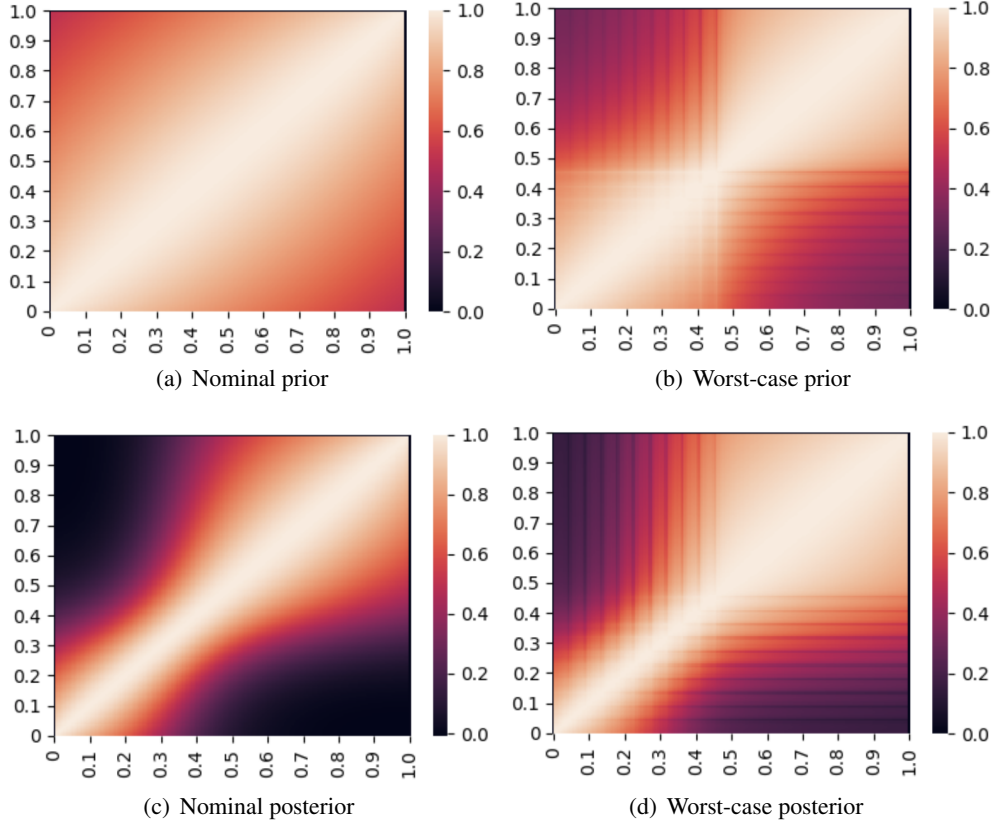
(a) Nominal prior

(b) Worst-case prior

(c) Nominal posterior

(d) Worst-case posterior

FIG 2. *Correlation functions on* $[0,1]^2$ *with 10 design points equispaced on* $(0, 0.5)$.

affine transformations and under inversion and has been used extensively to compare covariance matrices in previous work [16, 40, 39].

With the remaining parameters fixed to the base case, we explore the following parameter variations.

1. Prior smoothness: choose $\alpha \in \{0.51, 2, 4\}$. Results are shown in Figures 5–7; note that we include the baseline value $\alpha = 2$ for comparison.
2. Adversarial perturbation smoothness: choose $\beta \in \{0.7, 1\}$. Results are shown in Figure 8. Note that the nominal prior and posterior are the same as in the baseline setting.
3. Size of the Wasserstein ambiguity set: choose $\delta^2 \in \{0.01, 1\}$. Results are shown in Figure 9. Note that the nominal prior and posterior are the same as in the baseline setting.
4. Magnitude of the observation noise: choose $\sigma \in \{0.01, 1\}$. Results are shown in Figures 10 and 11.

The corresponding nominal and worst-case prior-to-posterior distances are reported Table 1. Combining these qualitative and quantitative results, we observe that in cases where there is: a larger α (i.e., a smoother prior); a smaller β (i.e., smaller penalty on modes that induce roughness); a larger δ (i.e., wider range of admissible perturbations); or a smaller σ (i.e., smaller observation noise), the worst-case distributions induce sharper contrasts between the observed and unobserved locations in both the prior and posterior sample paths.

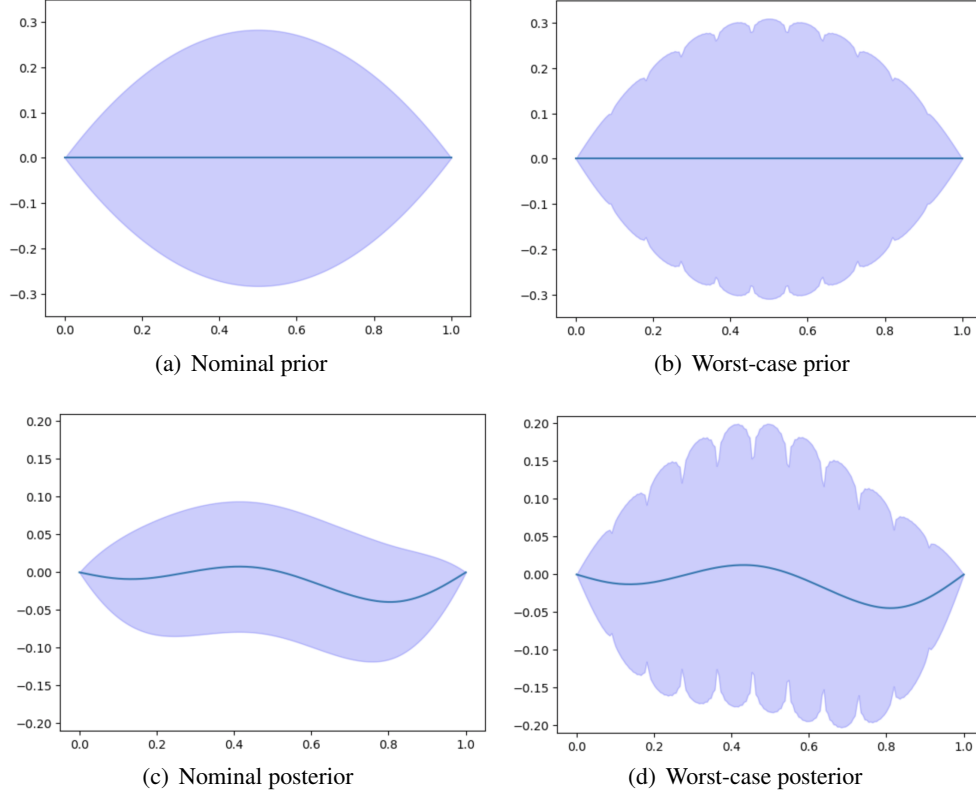## APPENDIX A: PROOFS OF THEOREMS AND ADDITIONAL DETAILS

A.1. Proof of Theorem 3.1.

(a) Nominal prior

(b) Worst-case prior

(c) Nominal posterior

(d) Worst-case posterior

FIG 3. *95% intervals of sample paths with 10 design points equispaced on* $(0,1)$.

	baseline	$\alpha = 0.51$	$\alpha = 4$	$\beta = 0.7$	$\beta = 1$	$\delta^2 = 0.01$	$\delta^2 = 1$	$\sigma = 0.01$	$\sigma = 1$
Nominal	2.56	16.24	0.11	2.56	2.56	2.56	2.56	8.92	0.11
Worst-case	12.82	16.15	5.33	8.97	2.13	9.88	10.90	19.74	3.64

TABLE 1

Natural distance between the prior and posterior covariance matrices under different problem settings, for 10 design points equispaced on $(0,1)$.

PROOF OF THE STRONG DUALITY. We start with a few definitions used in the proof. Recall that $\mathcal{M} = \{\phi : \mathbb{R}^m \to C(\mathcal{D}),\ \phi \text{ is measurable}\}$ is the set of estimators for b.

DEFINITION A.1 (Affine predictor). *For* $\phi(Y_1, \ldots, Y_m)$ *a mapping from* $\mathbb{R}^m$ *(data space) to* $C(\mathcal{D})$ *(parameter space), we say that* ϕ *is affine if it is of the form*

$$\phi(Y_1, \ldots, Y_m) = \alpha_0 + \sum_{j=1}^m \alpha_j Y_j,$$

for some functions $\alpha_j \in C(\mathcal{D})$, $0 \le j \le m$. *We write* $\mathcal{M}_{\text{aff}} = \{\phi : \mathbb{R}^m \to C(\mathcal{D}),\ \phi \text{ is affine}\} \subset \mathcal{M}$.

Let $\text{span}\{e_n\}_{n=1}^N$ and $\text{span}\{T(e_n)\}_{n=1}^N$ denote the closed subspaces of $C(\mathcal{D})$ spanned by the first N basis functions from Assumption 2.1, and by the action of T on these basis, respectively. For affine ϕ, we say that $\text{coef}(\phi) \in \text{span}\{e_n\}_{n=1}^N$ if for all $0 \le j \le m$, $\alpha_j \in \text{span}\{e_n\}_{n=1}^N$, and we denote the set of such estimators by

$$\mathcal{M}_{\text{aff},N} = \{\phi \in \mathcal{M}_{\text{aff}} : \text{coef}(\phi) \in \text{span}\{e_n\}_{n=1}^N\}.$$
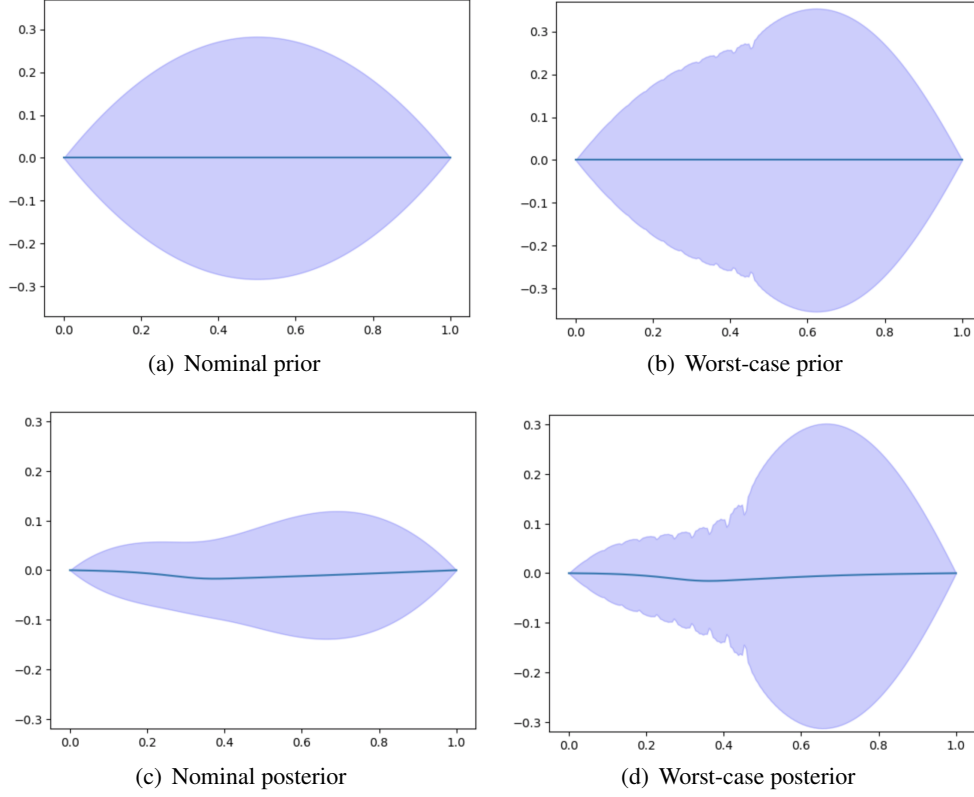
(a) Nominal prior

(b) Worst-case prior

(c) Nominal posterior

(d) Worst-case posterior

FIG 4. *95% intervals of sample paths with 10 design points equispaced on* $(0, 0.5)$.

For convenience, we write

$$\mathrm{Obj}(\phi, P) = \mathbb{E}_P \left[\|b - \phi(Y_1, \ldots, Y_m)\|^2_{L^2(\mathcal{D})} \right]$$

for the Bayes risk of an estimator $\phi \in \mathcal{M}$ under the distribution P. We denote the Wasserstein ball around P_0 as

$$\mathcal{W}(\delta) = \{P \in \mathcal{P} : \mathbb{W}(P, P_0) \leq \delta\}.$$

Then, denoting the Wasserstein balls around P_0 arising from perturbations in the first N coordinates by

$$(16) \qquad \mathcal{W}_N(\delta) = \{P \in \mathcal{W}(\delta) : \mathcal{L}^P\big(\{\langle b, e_n \rangle\}_{n=N+1}^\infty\big) = \mathcal{L}^{P_0}\big(\{\langle b, e_n \rangle\}_{n=N+1}^\infty\big)\},$$

where $\mathcal{L}^P$ denotes the law under P. We have for any $N \geq 1$ that

$$\inf_{\phi \in \mathcal{M}_{\mathrm{aff}, N}} \sup_{P \in \mathcal{W}(\delta)} \mathrm{Obj}(\phi, P) \geq \inf_{\phi \in \mathcal{M}_{\mathrm{aff}}} \sup_{P \in \mathcal{W}(\delta)} \mathrm{Obj}(\phi, P)$$

$$\geq \inf_{\phi \in \mathcal{M}} \sup_{P \in \mathcal{W}(\delta)} \mathrm{Obj}(\phi, P)$$

$$\geq \sup_{P \in \mathcal{W}(\delta)} \inf_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P)$$

$$(17) \qquad \geq \sup_{P \in \mathcal{W}_N(\delta)} \inf_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P).$$

(a) Nominal prior

(b) Worst-case prior

(c) Nominal posterior
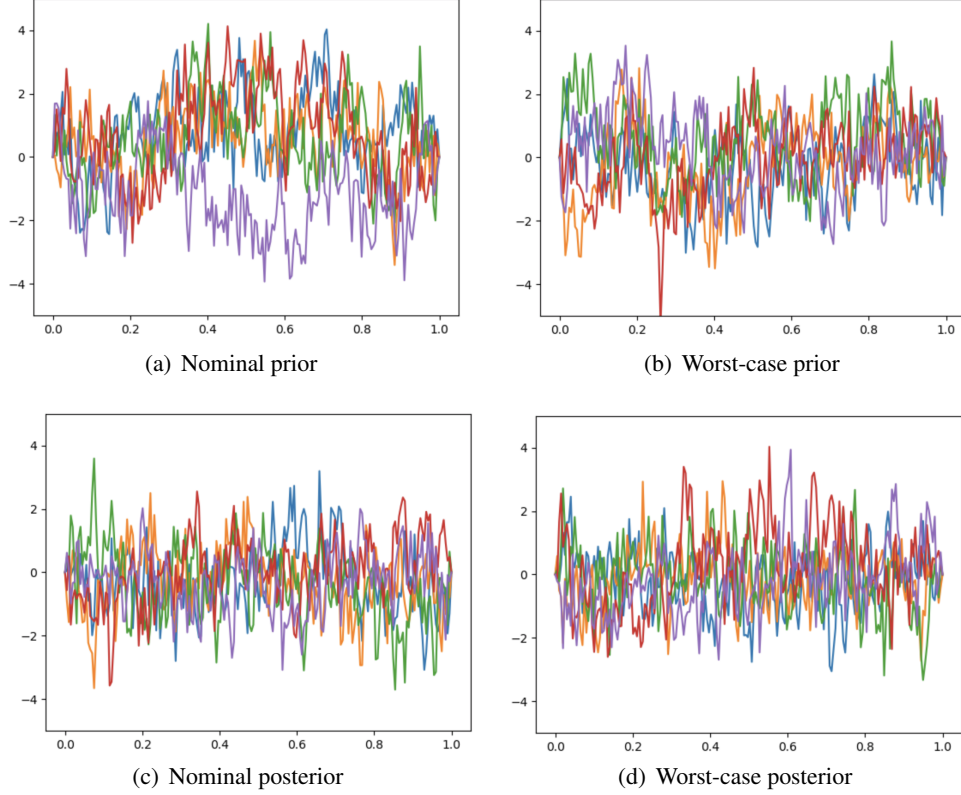
(d) Worst-case posterior

FIG 5. *Sample paths with* $\alpha = 0.51$ *and 10 designs equispaced in* $(0, 1)$.

Above, the first two inequalities follow from the inclusion $\mathcal{M}_{\mathrm{aff},N} \subset \mathcal{M}_{\mathrm{aff}} \subset \mathcal{M}$, the third inequality follows from weak duality, and the last equality follows from the inclusion $\mathcal{W}_N(\delta) \subset \mathcal{W}(\delta)$. Finally, for any $P \in \mathcal{P}$ and $N \geq 1$, we denote the joint measure induced by projecting b onto $\mathrm{span}\{e_n\}_{n=1}^N$ whilst keeping ϵ intact by

$$(18) \qquad P^{(N)} := \mathcal{L}^P\left(\left(\sum_{1 \leq n \leq N} e_n \langle b, e_n \rangle, \epsilon\right)\right).$$

Our proof consists of showing the following three claims:

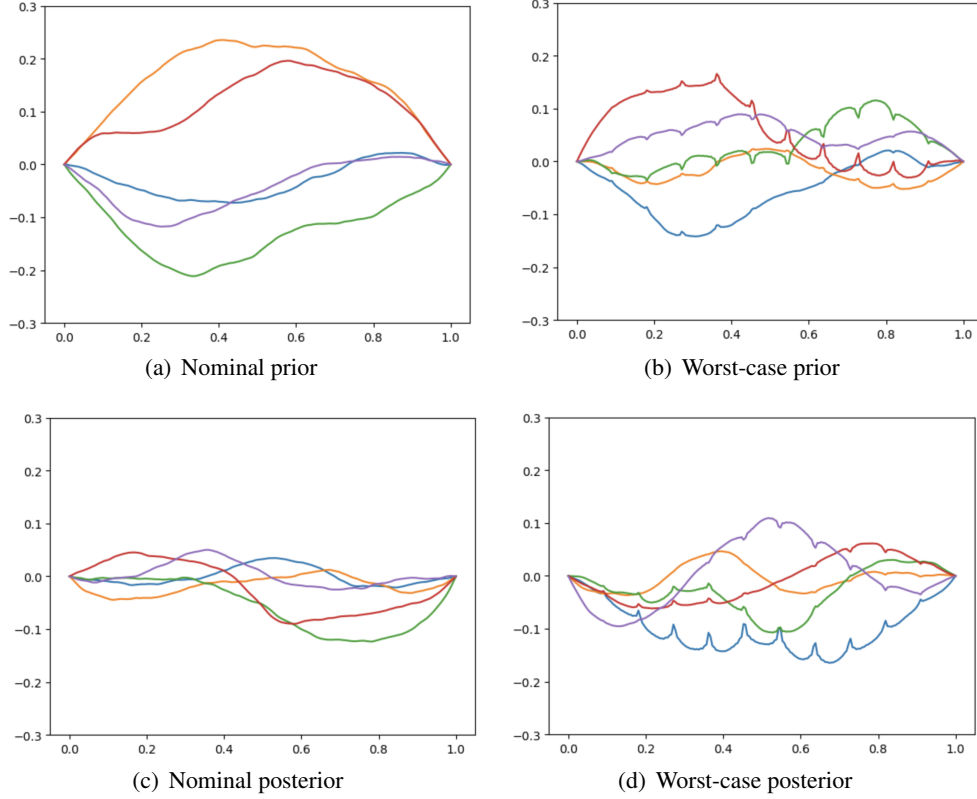**Claim 1**: The finite-dimensional version of strong duality holds, i.e.,

$$\inf_{\phi \in \mathcal{M}} \sup_{P \in \mathcal{W}_N(\delta)} \mathrm{Obj}(\phi, P^{(N)}) = \inf_{\phi \in \mathcal{M}_{\mathrm{aff},N}} \sup_{P \in \mathcal{W}_N(\delta)} \mathrm{Obj}(\phi, P^{(N)})$$

$$(19) \qquad\qquad\qquad\qquad = \sup_{P \in \mathcal{W}_N(\delta)} \inf_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P^{(N)}).$$

Claim 2: The truncation of P to $P^{(N)}$ in the last term of (17) preserves the chain of inequalities, i.e.,

$$(20) \qquad \sup_{P \in \mathcal{W}_N(\delta)} \inf_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P) \geq \sup_{P \in \mathcal{W}_N(\delta)} \inf_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P^{(N)}).$$

Claim 3: The truncation of P to $P^{(N)}$ in the first term of (17) has an error which is asymptotically negligible, i.e., as $N \to \infty$

$$(21) \qquad \inf_{\phi \in \mathcal{M}_{\mathrm{aff},N}} \sup_{P \in \mathcal{W}(\delta)} \mathrm{Obj}(\phi, P) - \inf_{\phi \in \mathcal{M}_{\mathrm{aff},N}} \sup_{P \in \mathcal{W}_N(\delta)} \mathrm{Obj}(\phi, P^{(N)}) = o(1).$$

(a) Nominal prior

(b) Worst-case prior

(c) Nominal posterior

(d) Worst-case posterior

FIG 6. *Sample paths with $\alpha = 2$ and 10 designs equispaced in $(0,1)$.*

Combining the above three claims and the chain of inequalities (17), we conclude that

$$
(22) \qquad \inf_{\phi \in \mathcal{M}} \sup_{P \in \mathcal{W}(\delta)} \mathrm{Obj}(\phi, P) = \sup_{P \in \mathcal{W}(\delta)} \inf_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P)
$$

by letting $N \to \infty$. $\square$

We now provide the proof of the three claims.

PROOF OF CLAIM 1.   For $P \in \mathcal{P}$, we denote the marginal distribution of the first N basis coefficients $\langle b, e_n \rangle$ and ϵ by

$$
(23) \qquad Q^{(N)} := \mathcal{L}^P \big((\langle b, e_1 \rangle, \cdots, \langle b, e_N \rangle, \epsilon) \big).
$$

Note that $Q^{(N)}$ is a probability measure on $\mathbb{R}^{N+m}$, while the distribution $P^{(N)}$ is supported on $C(\mathcal{D}) \times \mathbb{R}^m$. In particular, we have $Q_0^{(N)} := \mathcal{L}^{P_0} \big((\langle b, e_1 \rangle, \cdots, \langle b, e_N \rangle, \epsilon) \big)$ is a multivariate Gaussian. Now, for the weighted cost function

$$
c_N(r, s) = \sum_{n=1}^{N} (r_n - s_n)^2 w_n + \sum_{j=1}^{m} (r_{N+j} - s_{N+j})^2 \quad \forall r, s \in \mathbb{R}^{N+m}
$$

(where the variable order of r (or s) is the same as the variable order in equation (23)) on $\mathbb{R}^{N+m}$, we denote the corresponding optimal transport cost between probability measures τ and ν on $\mathbb{R}^{N+m}$ by

$$
\mathrm{W}_N^2(\tau, \nu) = \min_{\pi: \pi_r = \tau, \pi_s = \nu} \mathbb{E}_{(R,S) \sim \pi} [c_N(R, S)],
$$

(a) Nominal prior

(b) Worst-case prior

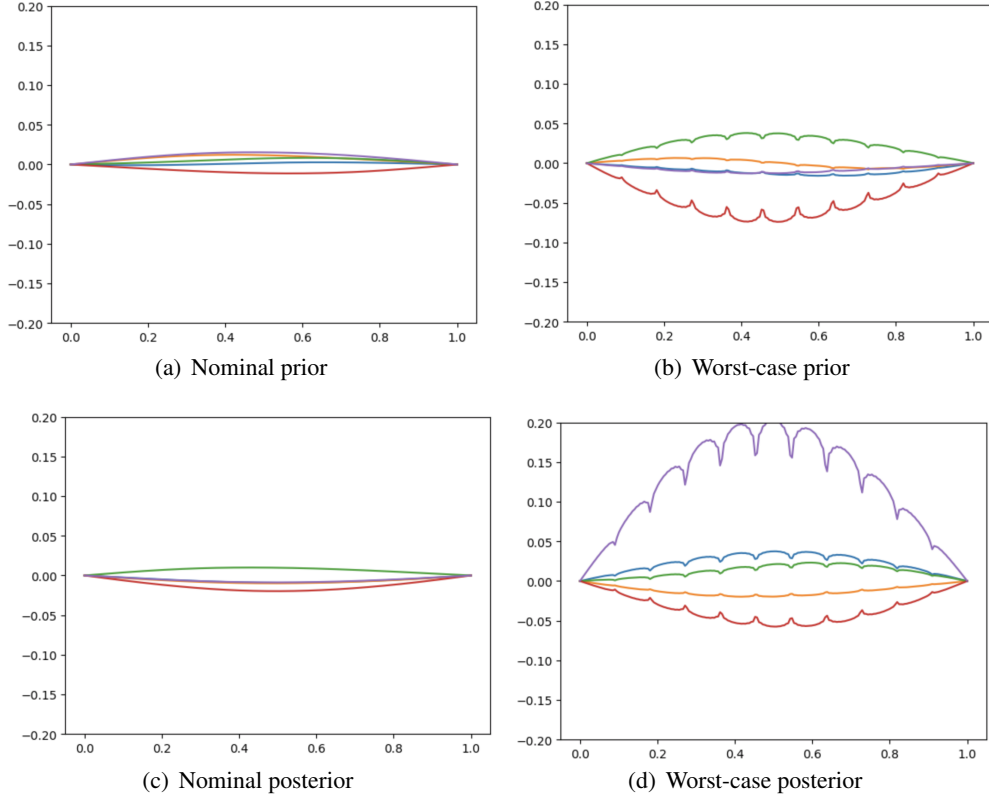(c) Nominal posterior

(d) Worst-case posterior

FIG 7. *Sample paths with $\alpha = 4$ and 10 designs equispaced in $(0,1)$.*

where we implicitly imposed π as a probability measure to avoid cluttered notations. Thus, by writing the shorthand

$$g(r) := \left\| \sum_{n=1}^{N} r_n e_n - \phi\left(\left(\sum_{n=1}^{N} r_n T(e_n)(x_j) + r_{N+j}\right)_{j=1}^{m}\right) \right\|_{L^2(\mathcal{D})}^2,$$

we obtain that

$$\sup_{P \in \mathcal{W}_N(\delta)} \mathrm{Obj}(\phi, P^{(N)}) = \sup_{Q^{(N)}: \mathbb{W}_N(Q^{(N)}, Q_0^{(N)}) \leq \delta} \mathbb{E}_{R \sim Q^{(N)}}[g(R)].$$

Assume momentarily that $\phi \in \mathcal{M}_{\mathrm{aff}, N}$. In this case, g is a convex (non-constant) quadratic function and [8, Theorem 1] implies the dual formulation
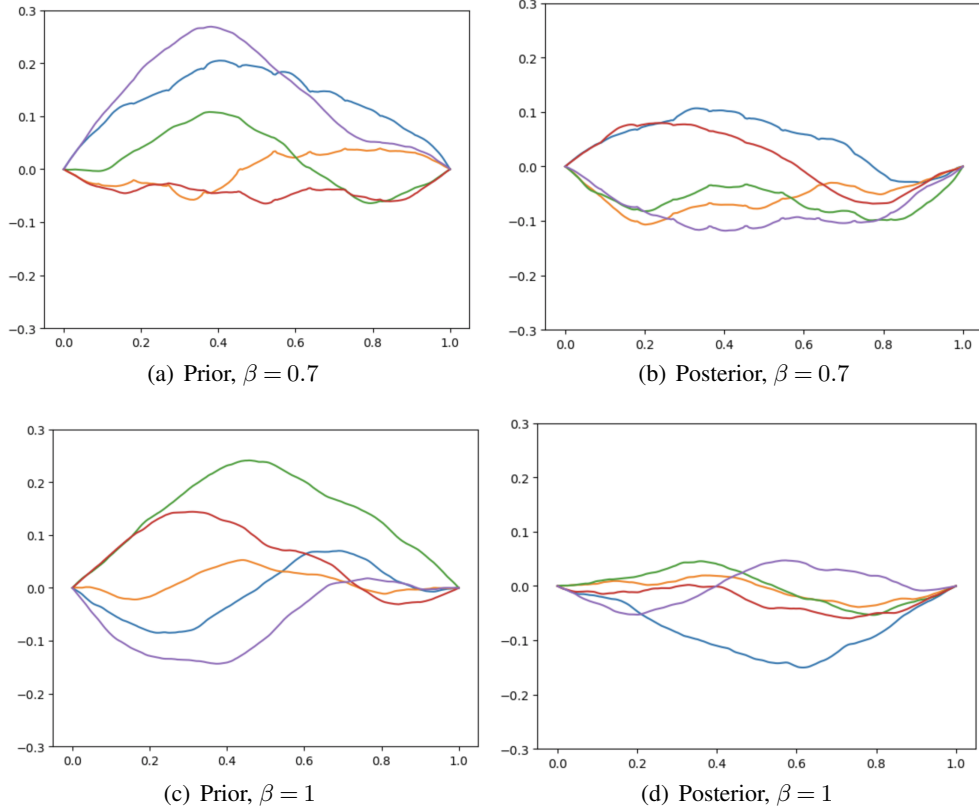
(24)

$$\sup_{Q^{(N)}: \mathbb{W}_N(Q^{(N)}, Q_0^{(N)}) \leq \delta} \mathbb{E}_{R \sim Q^{(N)}}[g(R)] = \inf_{\gamma \geq 0} \left(\gamma \delta^2 + \mathbb{E}_{R \sim Q_0^{(N)}} \left[\sup_{s \in \mathbb{R}^{N+m}} \{g(s) - \gamma c_N(R, s)\} \right] \right).$$

The same theorem moreover implies that there exists a dual optimizer $\gamma^\star$ to the right-hand side minimization problem of (24). Since the distributions in the set

$$\{Q^{(N)} : \mathbb{W}_N(Q^{(N)}, Q_0^{(N)}) \leq \delta\}$$

are uniformly bounded in the second moment and g is quadratic in r, we have that

$$\sup_{Q^{(N)}: \mathbb{W}_N(Q^{(N)}, Q_0^{(N)}) \leq \delta} \mathbb{E}_{R \sim Q^{(N)}}[g(R)] < \infty.$$

(a) Prior, $\beta = 0.7$
(b) Posterior, $\beta = 0.7$
(c) Prior, $\beta = 1$
(d) Posterior, $\beta = 1$

FIG 8. *Worst-case sample paths with varying β and 10 designs equispaced in $(0,1)$.*

Since also $\sup_{s \in \mathbb{R}^{N+m}} g(s) = \infty$, we have necessarily that $\gamma^\star > 0$.

Observe that $g(r)$ can be written in the form $g(r) = r^\top G r + c^\top r + \|\alpha_0\|_{L^2(\mathcal{D})}^2$ for some $G \in \mathbb{S}_+^{N+m}$ and $c \in \mathbb{R}^{N+m}$, where we denote $\mathbb{S}_+^{N+m}$ as the set of positive semi-definite matrices with a dimension of $N + m$. Denoting the matrix (note that not to be confused with W_N, which denotes optimal transport distance)

$$W_N = \mathrm{diag}(w_1, \ldots, w_N, 1, \ldots, 1),$$

then, almost surely for $R \sim Q_0^{(N)}$, we have

$$g(s) - \gamma^\star c_N(R, s) = s^\top (G - \gamma^\star W_N)s + (2\gamma^\star R^\top W_N + c^\top)s + \|\alpha_0\|_{L^2(\mathcal{D})}^2 + \gamma^\star R^\top W_N R.$$

Since R follows a non-degenerate multivariate Gaussian distribution, we have that necessarily $\gamma^\star W_N - G \in \mathbb{S}_{++}^{N+m}$, otherwise $\sup_{s \in \mathbb{R}^{N+m}} g(s) - \gamma^\star c_N(R, s) = \infty$ almost surely. Note that [8, Theorem 1] also implies that the optimal coupling between $Q_0^{(N)}$ and the maximizer to the left side of (24), if it exists, must be given as the law of $(R, s^\star(R))$, for the affine push-forward map

$$(25) \qquad s^\star(R) = -\frac{1}{2}(G - \gamma^\star W_N)^{-1}(2\gamma^\star W_N R + c).$$

The existence of a solution $Q_\star^{(N)}$ to the left hand side of (24) as well as of an optimal coupling between $Q_0^{(N)}$ and $Q_\star^{(N)}$, which we denote by $\pi^\star$, can be verified by [18, Corollary 1(i)].
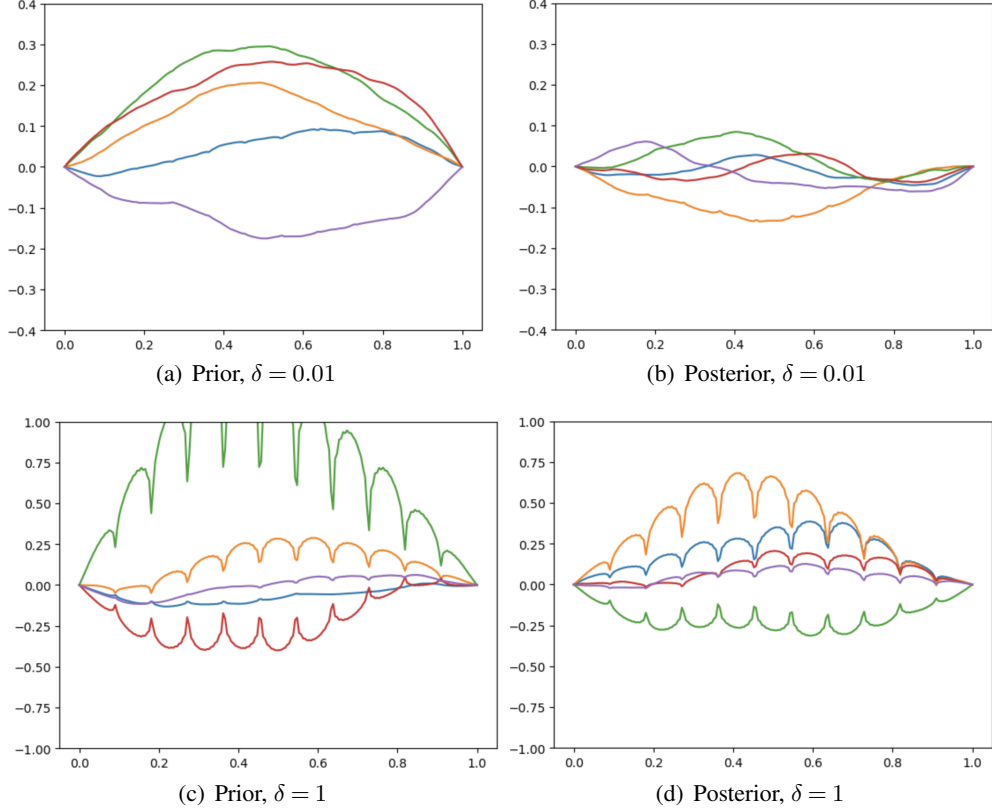
FIG 9. *Worst-case sample paths with varying δ and 10 designs equispaced in $(0,1)$.*

Indeed, since $\gamma^\star W_N - G \in \mathbb{S}_{++}^{N+m}$, for the growth rate κ defined to be

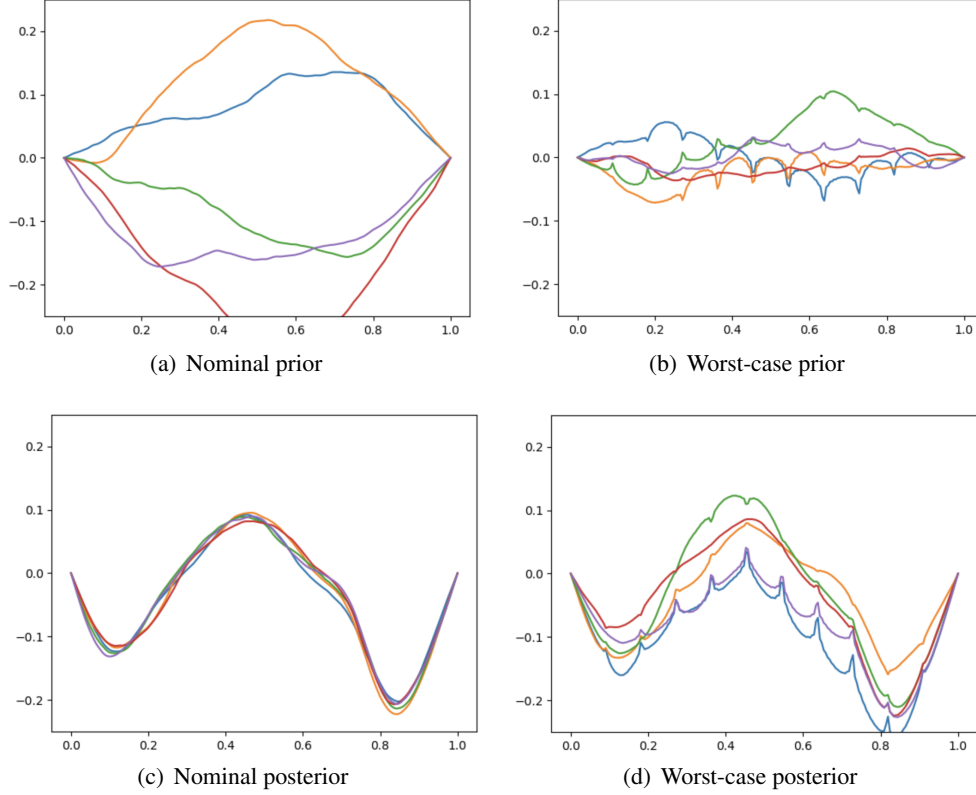$$\kappa = \limsup_{r \to \infty} \frac{r^\top G r}{r^\top W_N r},$$

we have that $\kappa < \gamma^\star$ so that Condition (a) in [18, Corollary 1(i)] holds. By the affine push-forward map (25), $Q_\star^{(N)}$ is also Gaussian, whence we may write (24) as

$$\sup_{Q^{(N)}:\mathrm{W}_N(Q^{(N)},Q_0^{(N)})\leq\delta} \mathbb{E}_{R\sim Q^{(N)}}[g(R)] = \sup_{Q^{(N)}:\mathrm{W}_N(Q^{(N)},Q_0^{(N)})\leq\delta, Q^{(N)}\text{ normal}} \mathbb{E}_{R\sim Q^{(N)}}[g(R)]$$

$$= \sup_{(\mu,\Sigma)\in\mathcal{S}_N} \langle G, \Sigma + \mu\mu^\top \rangle + c^\top\mu + \|\alpha_0\|_{L^2(\mathcal{D})}^2$$

$$\geq \sup_{(\mu,\Sigma)\in\mathcal{S}_N} \langle G, \Sigma + \mu\mu^\top \rangle,$$

where the last inequality is because $c^\top\mu \geq 0$ after a possible sign change in μ. Here $\mathcal{S}_N$ is a compact and convex set

$$\mathcal{S}_N = \left\{(\mu,\Sigma) : \|\mu\|_2^2 + tr(W_N\Sigma) + tr(W_N\Sigma_0) - 2tr\left[\left(\sqrt{\Sigma_0}W_N\Sigma W_N\sqrt{\Sigma_0}\right)^{1/2}\right] \leq \delta^2\right\},$$

where $\Sigma_0 \in \mathbb{S}_{++}^{N+m}$ is the covariance matrix of $Q_0^{(N)}$. To see how the set $\mathcal{S}_N$ arises, note that the usual squared-Euclidean cost function is $\|r - s\|_2^2$ between r and $s \in \mathbb{R}^{N+m}$, which gives

(a) Nominal prior

(b) Worst-case prior

(c) Nominal posterior

(d) Worst-case posterior

FIG 10. *Sample paths with $\sigma = 0.01$ and 10 designs equispaced in $(0, 1)$.*

rise to the usual Gelbrich distance (see [36, Proposition 2.2])

$$\sqrt{\|\mu\|_2^2 + tr(\Sigma) + tr(\Sigma_0) - 2tr\left[\left(\sqrt{\Sigma_0}\Sigma\sqrt{\Sigma_0}\right)^{1/2}\right]}$$

that coincides with the type-2 Wasserstein distance between multivariate Gaussians. Herein, our new cost is $(r - s)^\top W_N (r - s)$, thus the optimal coupling $\pi^\star$ for the new cost solves

$$\min_{\pi:\pi_r=\tau,\pi_s=\nu}\left(-\int r^\top W_N s \, \mathrm{d}\pi(r, s)\right).$$

Using substitution of variables $\tilde{r} = \sqrt{W_N}r$ and $\tilde{s} = \sqrt{W_N}s$, we find that the above problem reduces to the optimal transport problem with a squared-Euclidean cost.

In the sequel we denote the Wasserstein balls around P_0 arising from perturbations in the first N coordinates and restricting to the family of Gaussian distributions as

(26) $$\mathcal{W}_{\mathrm{nor},N}(\delta) = \{P \in \mathcal{W}_N(\delta) : Q^{(N)} \text{ is centered full-rank normal}\}.$$

Therefore, we obtain (recall α_0 in Definition A.1)

$$\inf_{\phi\in\mathcal{M}_{\mathrm{aff},N},\alpha_0=0}\ \sup_{Q^{(N)}:\mathrm{W}_N(Q^{(N)},Q_0^{(N)})\leq\delta}\ \mathbb{E}_{R\sim Q^{(N)}}[g(R)]$$

$$\geq \inf_{\phi\in\mathcal{M}_{\mathrm{aff},N}}\ \sup_{Q^{(N)}:\mathrm{W}_N(Q^{(N)},Q_0^{(N)})\leq\delta}\ \mathbb{E}_{R\sim Q^{(N)}}[g(R)]$$

$$\geq \inf_{\phi\in\mathcal{M}_{\mathrm{aff},N}}\ \sup_{(\mu,\Sigma)\in\mathcal{S}_N}\ \langle G, \Sigma + \mu\mu^\top\rangle$$

(a) Nominal prior
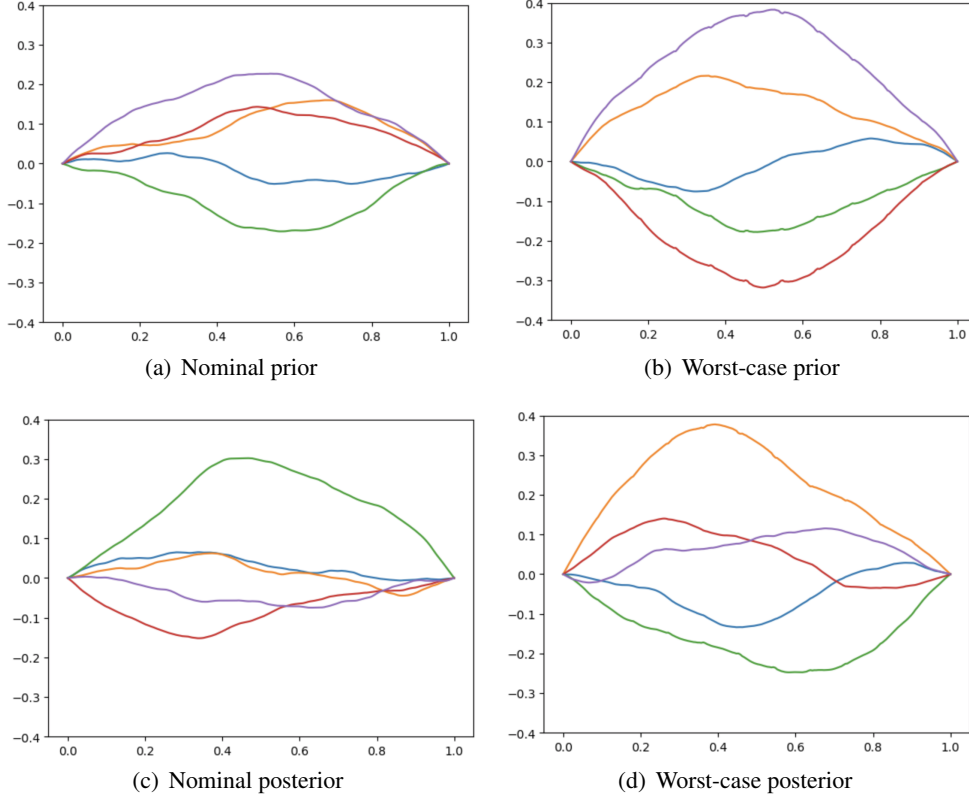
(b) Worst-case prior

(c) Nominal posterior

(d) Worst-case posterior

FIG 11. *Sample paths with $\sigma = 1$ and 10 designs equispaced in $(0,1)$.*

$$\geq \inf_{\phi \in \mathcal{M}_{\text{aff},N}} \sup_{(\mu,\Sigma) \in \mathcal{S}_N} \langle G, \Sigma \rangle$$

$$= \inf_{\phi \in \mathcal{M}_{\text{aff},N}, \alpha_0 = 0} \sup_{(0,\Sigma) \in \mathcal{S}_N} \langle G, \Sigma \rangle$$

$$= \inf_{\phi \in \mathcal{M}_{\text{aff},N}, \alpha_0 = 0} \sup_{Q^{(N)}: W_N(Q^{(N)}, Q_0^{(N)}) \leq \delta} \mathbb{E}_{R \sim Q^{(N)}}[g(R)],$$

where the last equality is because $s^\star(R)$ is zero-mean whenever $c = 0$, which in turn follows from $\alpha_0 = 0$. Hence all inequalities become equalities, and we conclude that

$$\inf_{\phi \in \mathcal{M}_{\text{aff},N}} \sup_{Q^{(N)}: W_N(Q^{(N)}, Q_0^{(N)}) \leq \delta} \mathbb{E}_{R \sim Q^{(N)}}[g(R)] = \inf_{\phi \in \mathcal{M}_{\text{aff},N}, \alpha_0 = 0} \sup_{(0,\Sigma) \in \mathcal{S}_N} \langle G, \Sigma \rangle.$$

Remind that the matrix G is dependent on the coefficients of the affine estimator ϕ. By Sion's minimax theorem [37], we have

$$(27) \qquad \inf_{\phi \in \mathcal{M}_{\text{aff},N}, \alpha_0 = 0} \sup_{(0,\Sigma) \in \mathcal{S}_N} \langle G, \Sigma \rangle$$

$$(28) \qquad = \sup_{(0,\Sigma) \in \mathcal{S}_N} \inf_{\phi \in \mathcal{M}_{\text{aff},N}, \alpha_0 = 0} \langle G, \Sigma \rangle$$

$$= \sup_{Q^{(N)}: W_N(Q^{(N)}, Q_0^{(N)}) \leq \delta, Q^{(N)} \text{ is centered full-rank normal}} \inf_{\phi \in \mathcal{M}_{\text{aff},N}, \alpha_0 = 0} \mathbb{E}_{R \sim Q^{(N)}}[g(R)].$$

The last "full-rank" assertion comes from the fact that $s^\star(R)$ is a non-degenerate linear transformation of R, and that a linear transformation of a multivariate Gaussian is also Gaussian.

Thus

$$\inf_{\phi\in\mathcal{M}_{\text{aff},N}}\sup_{P\in\mathcal{W}_N(\delta)}\text{Obj}(\phi,P^{(N)})=\inf_{\phi\in\mathcal{M}_{\text{aff},N}}\sup_{Q^{(N)}:\text{W}_N(Q^{(N)},Q_0^{(N)})\leq\delta}\mathbb{E}_{R\sim Q^{(N)}}[g(R)]$$

$$=\inf_{\phi\in\mathcal{M}_{\text{aff},N},\alpha_0=0}\sup_{Q^{(N)}:\text{W}_N(Q^{(N)},Q_0^{(N)})\leq\delta}\mathbb{E}_{R\sim Q^{(N)}}[g(R)]$$

$$(29)\qquad\qquad=\sup_{P\in\mathcal{W}_{\text{nor},N}(\delta)}\inf_{\phi\in\mathcal{M}_{\text{aff},N},\alpha_0=0}\text{Obj}(\phi,P^{(N)}),$$

where recall that $\mathcal{W}_{\text{nor},N}(\delta)$ was defined in equation (26). For $\text{W}(P,P_0)\leq\delta$, note that b is a process with continuous sample paths, whence we can interchange the integration

$$\mathbb{E}_P\left[\|b-\phi(Y_1,\ldots,Y_m)\|_{L^2(\mathcal{D})}^2\right]=\mathbb{E}_P\left[\int_{\mathcal{D}}|b(x)-\phi(Y_1,\ldots,Y_m)(x)|^2\mathrm{d}x\right]$$

$$=\int_{\mathcal{D}}\mathbb{E}_P\left[|b(x)-\phi(Y_1,\ldots,Y_m)(x)|^2\right]\mathrm{d}x.$$

Thus the optimal solution to the estimation problem

$$(30)\qquad\qquad\inf_{\phi\in\mathcal{M}}\mathbb{E}_P\left[\|b-\phi(Y_1,\ldots,Y_m)\|_{L^2(\mathcal{D})}^2\right]$$

is given by the conditional expectation function

$$(31)\qquad\qquad\phi(Y_1,\ldots,Y_m)(x)=\mathbb{E}_P\left[b(x)|Y_1,\ldots,Y_m\right].$$

For any distribution P that is a centered Gaussian random variable, it is easy to see that for $x\in\mathcal{D}$, the random vector

$$(b(x),Y_1,\ldots,Y_m)=(b(x),b(x_1)+\epsilon_1,\ldots,b(x_m)+\epsilon_m)$$

is jointly Gaussian and the conditional expectation satisfies

$$(32)\quad\mathbb{E}_P\left[b(x)|Y_1,\ldots,Y_m\right]=(k_{\epsilon,P}(x,x_1),\ldots,k_{\epsilon,P}(x,x_m))\cdot(K_{\epsilon,P})^{-1}\cdot(Y_1,\ldots,Y_m)^\top$$

where $k_{\epsilon,P}(x,x_j)=\mathbb{E}_P[b(x)(b(x_j)+\epsilon_j)]$, and

$$K_{\epsilon,P}=(\mathbb{E}_P[(b(x_i)+\epsilon_i)(b(x_j)+\epsilon_j)])_{ij}\in\mathbb{S}_{++}^m$$

provided the matrix $K_{\epsilon,P}$ is invertible. Moreover, the optimal value of the Bayes risk (30) is given by

$$\int_{\mathcal{D}}\mathbb{E}_P[\mathbb{V}\mathrm{ar}_P[b(x)|Y_1,\ldots,Y_m]]\mathrm{d}x$$

$$=\int_{\mathcal{D}}k_P(x,x)-(k_{\epsilon,P}(x,x_1),\ldots,k_{\epsilon,P}(x,x_m))\cdot(K_{\epsilon,P})^{-1}\cdot(k_{\epsilon,P}(x,x_1),\ldots,k_{\epsilon,P}(x,x_m))^\top\mathrm{d}x$$

where $k_P(x,x)=\mathbb{E}_P[b(x)b(x)]$. Thus

$$\sup_{P\in\mathcal{W}_{\text{nor},N}(\delta)}\inf_{\phi\in\mathcal{M}_{\text{aff},N},\alpha_0=0}\text{Obj}(\phi,P^{(N)})=\sup_{P\in\mathcal{W}_{\text{nor},N}(\delta)}\inf_{\phi\in\mathcal{M}}\text{Obj}(\phi,P^{(N)})$$

$$\leq\sup_{P\in\mathcal{W}_N(\delta)}\inf_{\phi\in\mathcal{M}}\text{Obj}(\phi,P^{(N)}),$$

On the other hand,

$$\sup_{P\in\mathcal{W}_N(\delta)}\inf_{\phi\in\mathcal{M}}\text{Obj}(\phi,P^{(N)})\leq\inf_{\phi\in\mathcal{M}}\sup_{P\in\mathcal{W}_N(\delta)}\text{Obj}(\phi,P^{(N)})\leq\inf_{\phi\in\mathcal{M}_{\text{aff},N}}\sup_{P\in\mathcal{W}_N(\delta)}\text{Obj}(\phi,P^{(N)}).$$

Thus combining the above chains of inequalities and the equality (29), we have

$$\inf_{\phi \in \mathcal{M}_{\mathrm{aff},N}} \sup_{P \in \mathcal{W}_N(\delta)} \mathrm{Obj}(\phi, P^{(N)}) = \inf_{\phi \in \mathcal{M}_{\mathrm{aff},N}, \alpha_0 = 0} \sup_{P \in \mathcal{W}_N(\delta)} \mathrm{Obj}(\phi, P^{(N)})$$

$$= \inf_{\phi \in \mathcal{M}} \sup_{P \in \mathcal{W}_N(\delta)} \mathrm{Obj}(\phi, P^{(N)})$$

(33)
$$= \sup_{P \in \mathcal{W}_N(\delta)} \inf_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P^{(N)})$$

$$= \sup_{P \in \mathcal{W}_{\mathrm{nor},N}(\delta)} \inf_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P^{(N)}).$$

Thus we have established (19). $\qquad\square$

PROOF OF CLAIM 2. By the proof of Claim 1, there exists a Nash equilibrium of (19), which we denote by $(\phi_N^\star, P_N^\star)$. In particular, by the previous Claim 1, we can choose $\phi_N^\star$ as affine and $Q_N^\star := \mathcal{L}^{P_N^\star}(\langle b, e_1\rangle, \cdots, \langle b, e_N\rangle, \epsilon)$ as some centered full-rank normal distribution. Also, the pair $(\phi_N^\star, Q_N^\star)$ is the unique Nash equilibrium corresponding to the game (27) and (28). Moreover, under the law $P_N^\star$, the random elements

(34)
$$(\{\langle b, e_n\rangle\}_{n=1}^N, \epsilon) \quad \text{and} \quad \{\langle b, e_n\rangle\}_{n=N+1}^\infty$$

are independent. Finally, we have that

$$\mathcal{L}^{P_N^\star}\left(\{\langle b, e_n\rangle\}_{n=N+1}^\infty\right) = \mathcal{L}^{P_0}\left(\{\langle b, e_n\rangle\}_{n=N+1}^\infty\right).$$

Note that $P_N^\star$ is a Gaussian measure supported on $C(\mathcal{D})$. We now show our claim (20). Note that

$$Y_j = u(x_j) + \epsilon_j = \sum_{k=1}^N \langle b, e_k\rangle T(e_k)(x_j) + \epsilon_j + \left(u(x_j) - \sum_{k=1}^N \langle b, e_k\rangle T(e_k)(x_j)\right),$$

where the last term (denoted by $R_{j,N}$ as a shorthand) is of zero-mean under $P_N^\star$. Thus any affine estimator ϕ can be written as (recall Definition A.1)

$$\phi(Y_1, \ldots, Y_m) = \sum_{j=1}^m \alpha_j \left(\sum_{k=1}^N \langle b, e_k\rangle T(e_k)(x_j) + \epsilon_j\right) + \sum_{j=1}^m \alpha_j R_{j,N},$$

where w.l.o.g. $\alpha_0 = 0$. Let $\{\tilde{e}_k\}_{k=1}^\infty = \{e_k\}_{k=1}^\infty$. (In the proof of Theorem 3.5 we modify $\{\tilde{e}_k\}_{k=1}^\infty$ to be the Gram-Schmidt orthonormalization of $\{T(e_k)\}_{k=1}^\infty$. See discussion in Appendix A.8). We have

$$\sup_{P \in \mathcal{W}_N(\delta)} \inf_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P)$$

$$\geq \inf_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P_N^\star)$$

$$= \inf_{\phi \in \mathcal{M}} \mathbb{E}_{P_N^\star}[\|b - \phi(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2]$$

$$= \inf_{\phi \in \mathcal{M}_{\mathrm{aff}}} \mathbb{E}_{P_N^\star}[\|b - \phi(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2] \quad \text{due to Gaussianity}$$

$$= \inf_{\phi \in \mathcal{M}_{\mathrm{aff}}} \mathbb{E}_{P_N^\star}\left[\sum_{n=1}^\infty (\langle b - \phi(Y_1, \ldots, Y_m), \tilde{e}_n\rangle)^2\right]$$

$$
\geq \inf_{\phi \in \mathcal{M}_{\text{aff}}} \mathbb{E}_{P_N^\star} \left[\sum_{n=1}^N \left(\langle b - \phi(Y_1, \ldots, Y_m), \tilde{e}_n \rangle \right)^2 \right]
$$

$$
= \inf_{\phi \in \mathcal{M}_{\text{aff}}} \sum_{n=1}^N \mathbb{E}_{P_N^\star} \left(\left\langle b - \sum_{j=1}^m \alpha_j \left(\sum_{k=1}^N \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right), \tilde{e}_n \right\rangle^2 \right)
$$

$$
+ 2 \sum_{n=1}^N \mathbb{E}_{P_N^\star} \left(\left\langle b - \sum_{j=1}^m \alpha_j \left(\sum_{k=1}^N \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right), \tilde{e}_n \right\rangle \left\langle \sum_{j=1}^m \alpha_j R_{j,N}, \tilde{e}_n \right\rangle \right)
$$

$$
+ \sum_{n=1}^N \mathbb{E}_{P_N^\star} \left(\left\langle \sum_{j=1}^m \alpha_j R_{j,N}, \tilde{e}_n \right\rangle^2 \right)
$$

$$
= \inf_{\phi \in \mathcal{M}_{\text{aff}}} \sum_{n=1}^N \mathbb{E}_{P_N^\star} \left(\left\langle b - \sum_{j=1}^m \alpha_j \left(\sum_{k=1}^N \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right), \tilde{e}_n \right\rangle^2 \right)
$$

$$
+ \sum_{n=1}^N \mathbb{E}_{P_N^\star} \left(\left\langle \sum_{j=1}^m \alpha_j R_{j,N}, \tilde{e}_n \right\rangle^2 \right)
$$

$$
\geq \inf_{\phi \in \mathcal{M}_{\text{aff}}} \sum_{n=1}^N \mathbb{E}_{P_N^\star} \left(\left\langle b - \sum_{j=1}^m \alpha_j \left(\sum_{k=1}^N \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right), \tilde{e}_n \right\rangle^2 \right)
$$

$$
= \sup_{P \in \mathcal{W}_N(\delta)} \inf_{\phi \in \mathcal{M}} \text{Obj}(\phi, P^{(N)}).
$$

The penultimate equality is because that

$$
\mathbb{E}_{P_N^\star} \left(\left\langle b - \sum_{j=1}^m \alpha_j \left(\sum_{k=1}^N \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right), \tilde{e}_n \right\rangle \left\langle \sum_{j=1}^m \alpha_j R_{j,N}, \tilde{e}_n \right\rangle \right)
$$

$$
= \mathbb{E}_{P_N^\star} \left(\left\langle b - \sum_{j=1}^m \alpha_j \left(\sum_{k=1}^N \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right), \tilde{e}_n \right\rangle \right) \mathbb{E}_{P_N^\star} \left(\left\langle \sum_{j=1}^m \alpha_j R_{j,N}, \tilde{e}_n \right\rangle \right)
$$

$$
= 0,
$$

where we have used the independence in (34) and the fact that $R_{j,N}$ have zero mean. Therefore, we have established Claim 2. $\qquad\square$

PROOF OF CLAIM 3. Since $\phi_N^\star$ is affine, we can write $\phi_N^\star$ in the form $\phi_N^\star(Y_1, \ldots, Y_m) = \sum_{j=1}^m \alpha_j^\star Y_j$, where $\alpha_j^\star \in \text{span}\{e_n\}_{n=1}^N$ and w.l.o.g. $\alpha_0^\star = 0$. Here, we have suppressed the dependence of $\alpha_j^\star$ on N to simplify the notations. Also note that there exists a constant C independent of N, such that for any $1 \leq j \leq m$,

$$
\|\alpha_j^\star\|_{L^2(\mathcal{D})}^2 = \sum_{n=1}^N |\langle \alpha_j^\star, \tilde{e}_n \rangle|^2 \leq C < \infty.
$$

This is because we can choose $(0, \Sigma) \in \mathcal{S}_N$ in the inner constraint of (27) to be the nominal measure, and thereby conclude that the optimal objective value in (27) is lower bounded by

$C' \sum_{j=1}^{m} \|\alpha_j\|_{L^2(\mathcal{D})}^2$, for some constant C' independent of N. For any N, we have

$$\sup_{P \in \mathcal{W}(\delta)} \mathrm{Obj}(\phi_N^\star, P) - \sup_{P \in \mathcal{W}_N(\delta)} \mathrm{Obj}(\phi_N^\star, P^{(N)})$$

$$= \sup_{P \in \mathcal{W}(\delta)} \left(\sum_{n=1}^{N} \mathbb{E}_P(\langle b - \phi_N^\star, \tilde{e}_n \rangle)^2 + \sum_{n=N+1}^{\infty} \mathbb{E}_P(\langle b, \tilde{e}_n \rangle)^2 \right)$$

$$- \sup_{P \in \mathcal{W}_N(\delta)} \sum_{n=1}^{N} \mathbb{E}_{P^{(N)}}(\langle b - \phi_N^\star, \tilde{e}_n \rangle)^2$$

$$\leq \sup_{P \in \mathcal{W}(\delta)} \sum_{n=1}^{N} \mathbb{E}_P(\langle b - \phi_N^\star, \tilde{e}_n \rangle)^2 - \sup_{P \in \mathcal{W}_N(\delta)} \sum_{n=1}^{N} \mathbb{E}_{P^{(N)}}(\langle b - \phi_N^\star, \tilde{e}_n \rangle)^2$$

$$(35) \qquad + \sup_{P \in \mathcal{W}(\delta)} \sum_{n=N+1}^{\infty} \mathbb{E}_P(\langle b, \tilde{e}_n \rangle)^2.$$

For the last term in (35), note that

$$\sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_P \left[\sum_{n=N+1}^{\infty} (\langle b, \tilde{e}_n \rangle)^2 \right]$$

$$\leq 2 \mathbb{E}_{P_0} \left[\sum_{n=N+1}^{\infty} (\langle b^0, \tilde{e}_n \rangle)^2 \right] + 2 \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_\pi \left[\sum_{n=N+1}^{\infty} (\langle b - b^0, \tilde{e}_n \rangle)^2 \right]$$

$$\leq o(1) + O(1) \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_\pi \left[\|b - b^0 - \sum_{n=1}^{N} \langle b - b^0, e_n \rangle e_n\|_{L^2(\mathcal{D})}^2 \right],$$

where π is the optimal coupling between the marginals P_0 and P, and

$$\sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_\pi \left[\|b - b^0 - \sum_{n=1}^{N} \langle b - b^0, e_n \rangle e_n\|_{L^2(\mathcal{D})}^2 \right]$$

$$\leq O(1) \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_\pi \left[\|b - b^0 - \sum_{n=1}^{N} \langle b - b^0, e_n \rangle e_n\|_{\mathcal{H}_{\tilde{w}}}^2 \right]$$

$$\leq o(1) \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_\pi \left[\|b - b^0 - \sum_{n=1}^{N} \langle b - b^0, e_n \rangle e_n\|_{\mathcal{H}_w}^2 \right]$$

$$\to 0 \text{ as } N \to \infty,$$

where the second inequality comes from Assumption 2.5 and the fact that $b - b^0 - \sum_{n=1}^{N} \langle b - b^0, e_n \rangle e_n$ is in the space spanned by $\{e_n\}_{n=N+1}^{\infty}$. For the first two terms in (35), we write

$$\phi_N^\star(Y_1, \ldots, Y_m) = \sum_{j=1}^{m} \alpha_j^\star \left(\sum_{k=1}^{N} \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right) + \sum_{j=1}^{m} \alpha_j^\star R_{j,N}.$$

Thus for any feasible P in $\mathcal{W}(\delta)$, we have

$$\mathbb{E}_P(\langle b - \phi_N^\star, \tilde{e}_n \rangle)^2$$

$$= \mathbb{E}_P \left(\left\langle b - \sum_{j=1}^{m} \alpha_j^{\star}(x) \left(\sum_{k=1}^{N} \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right), \tilde{e}_n \right\rangle^2 \right)$$

$$- 2\mathbb{E}_P \left(\left\langle b - \sum_{j=1}^{m} \alpha_j^{\star}(x) \left(\sum_{k=1}^{N} \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right), \tilde{e}_n \right\rangle \left\langle \sum_{j=1}^{m} \alpha_j^{\star}(x) R_{j,N}, \tilde{e}_n \right\rangle \right)$$

$$+ \mathbb{E}_P \left(\left\langle \sum_{j=1}^{m} \alpha_j^{\star}(x) R_{j,N}, \tilde{e}_n \right\rangle^2 \right).$$

Note that

$$\sum_{n=1}^{N} \mathbb{E}_P \left(\left\langle b - \sum_{j=1}^{m} \alpha_j^{\star}(x) \left(\sum_{k=1}^{N} \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right), \tilde{e}_n \right\rangle^2 \right)$$

$$\leq \sup_{P \in \mathcal{W}_N(\delta)} \sum_{n=1}^{N} \mathbb{E}_{P^{(N)}} (\langle b - \phi_N^{\star}, \tilde{e}_n \rangle)^2.$$

Also, by Cauchy-Schwarz, we obtain

$$\sum_{n=1}^{N} \mathbb{E}_P \left(\left\langle b - \sum_{j=1}^{m} \alpha_j^{\star}(x) \left(\sum_{k=1}^{N} \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_i \right), \tilde{e}_n \right\rangle \left\langle \sum_{j=1}^{m} \alpha_j^{\star}(x) R_{j,N}, \tilde{e}_n \right\rangle \right)$$

$$\leq \sqrt{\mathbb{E}_P \left(\sum_{n=1}^{N} \left\langle b - \sum_{j=1}^{m} \alpha_j^{\star}(x) \left(\sum_{k=1}^{N} \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_i \right), \tilde{e}_n \right\rangle^2 \right)} \cdot$$

$$\sqrt{\mathbb{E}_P \left(\sum_{n=1}^{N} \left\langle \sum_{j=1}^{m} \alpha_j^{\star}(x) R_{j,N}, \tilde{e}_n \right\rangle^2 \right)}$$

$$\leq \sqrt{\sup_{P \in \mathcal{W}_N(\delta)} \sum_{n=1}^{N} \mathbb{E}_{P^{(N)}} (\langle b - \phi_N^{\star}, \tilde{e}_n \rangle)^2} \sqrt{mC \left(\sum_{j=1}^{m} \mathbb{E}_P \left[R_{j,N}^2 \right] \right)}.$$

Therefore, denoting π as the optimal coupling between P and P_0, we have

$$\sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_P \left(R_{j,N} \right)^2$$

$$= \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_P \left(u(x_j) - \sum_{k=1}^{N} \langle b, e_k \rangle T(e_k)(x_j) \right)^2$$

$$\leq 2\mathbb{E}_{P_0} \left(u^0(x_j) - \sum_{k=1}^{N} \langle b^0, e_k \rangle T(e_k)(x_j) \right)^2$$

$$+ 2 \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_\pi \left(u(x_j) - u^0(x_j) - \sum_{k=1}^{N} \langle b - b^0, e_k \rangle T(e_k)(x_j) \right)^2$$

$$\leq o(1) + 2 \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_\pi \left(u(x_j) - u^0(x_j) - \sum_{k=1}^{N} \langle b - b^0, e_k \rangle T(e_k)(x_j) \right)^2.$$

Moreover, we also find

$$\sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_\pi \left(u(x_j) - u^0(x_j) - \sum_{k=1}^{N} \langle b - b^0, e_k \rangle T(e_k)(x_j) \right)^2$$

$$= \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_\pi \left[\left(T(b - b^0 - \sum_{k=1}^{N} \langle b - b^0, e_k \rangle e_k)(x_j) \right)^2 \right]$$

$$\leq O(1) \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_\pi \left[\| T(b - b^0 - \sum_{k=1}^{N} \langle b - b^0, e_k \rangle e_k) \|_{\mathcal{H}_{\tilde{w}}}^2 \right]$$

$$\leq O(1) \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_\pi \left[\| b - b^0 - \sum_{k=1}^{N} \langle b - b^0, e_k \rangle e_k \|_{\mathcal{H}_{\tilde{w}}}^2 \right]$$

$$\leq o(1) \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_\pi \left[\| b - b^0 - \sum_{k=1}^{N} \langle b - b^0, e_k \rangle e_k \|_{\mathcal{H}_w}^2 \right]$$

$$\leq o(1) \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_\pi \left[\| b - b^0 \|_{\mathcal{H}_w}^2 \right]$$

$$\to 0 \text{ as } N \to \infty,$$

where the first inequality follows because $\mathcal{H}_{\tilde{w}}$ is an RKHS consisting of continuous functions, so that point evaluations at the x_j's are bounded, according to Assumption 2.3(ii). The second inequality also follows from Assumption 2.3(ii), while the third equality follows from Assumption 2.5. Finally, the last inequality follows because $\{e_n\}_{n=1}^\infty$ is an orthogonal system. Therefore, we have that

$$\sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_P \left(R_{j,N} \right)^2 \to 0 \text{ as } N \to \infty.$$

It then follows that

$$\sup_{P \in \mathcal{W}(\delta)} \text{Obj}(\phi_N^\star, P) - \sup_{P \in \mathcal{W}_N(\delta)} \text{Obj}(\phi_N^\star, P^{(N)}) = o(1) \text{ as } N \to \infty,$$

establishing our Claim 3. $\qquad\square$

A.2. Proof of Proposition 3.1.

PROOF OF PROPOSITION 3.1. This is an immediate consequence of the inequality (17) and Claims 1-3 in the proof of Theorem 3.1. $\qquad\square$

A.3. Proof of Proposition 3.2.

PROOF OF PROPOSITION 3.2. Since the space $C(\mathcal{D}) \times \mathbb{R}^m$ is Polish, if suffices to show that the sequence $P_N, N = 1, \ldots, \infty$ is tight. First note that P_0 is a finite Radon measure on $C(\mathcal{D}) \times \mathbb{R}^m$, and hence P_0 is tight. Therefore, for any $0 < \eta < 1$, there exists a compact set $\mathcal{C}_0$ in $C(\mathcal{D}) \times \mathbb{R}^m$, such that $P_0(\mathcal{C}_0) \geq 1 - \frac{\eta}{2}$. Now consider the set

$$\mathcal{C}_1 = \{(b, \epsilon) \in C(\mathcal{D}) \times \mathbb{R}^m : \exists\, (b^0, \epsilon^0) \in \mathcal{C}_0,\ \| b - b^0 \|_{\mathcal{H}_w}^2 + \| \epsilon - \epsilon_0 \|_2^2 \leq 2\delta^2/\eta \},$$

which is compact since the space $\mathcal{H}_w$ is compactly embedded in $C(\mathcal{D})$ by Assumption 2.2. Denote Q_N as the law of the difference $(b - b^0, \epsilon - \epsilon^0)$ under the optimal coupling between $(b, \epsilon) \sim P_N$ and $(b^0, \epsilon^0) \sim P_0$. Then,

$$1 - P_N(\mathcal{C}_1) \leq 1 - P_0(\mathcal{C}_0) + Q_N(\{(b, \epsilon) \in \mathcal{H}_w \times \mathbb{R}^m : \|b\|_{\mathcal{H}_w}^2 + \|\epsilon\|_2^2 > 2\delta^2/\eta\})$$

$$\leq \frac{\eta}{2} + \frac{\mathbb{W}^2(P_N, P_0)}{2\delta^2/\eta} \leq \frac{\eta}{2} + \frac{\eta}{2} = \leq \eta,$$

where in the second inequality we used the Cauchy-Schwarz inequality. Therefore, we conclude that

$$P_N(\mathcal{C}_1) \geq 1 - \eta \quad \forall N,$$

and hence that the sequence P_N is tight. It follows that there exists a weakly convergent subsequence $P_{N_l}, l \geq 1$ with $P_{N_l} \Rightarrow P_\infty$. Since the Wasserstein distance is lower semi-continuous, the limit P_∞ necessarily satisfies $\mathbb{W}(P_\infty, P_0) \leq \delta$.

$\square$

A.4. Proof of Theorem 3.2. We start with a simple but useful result, which is well-known [46].

LEMMA A.1 (Minimax theorem and Nash equilibrium). *Let $\phi^\star$ be an optimal solution to the outer infimum of the problem*

$$\inf_{\phi \in \mathcal{M}} \sup_{P \in \mathcal{W}(\delta)} \mathrm{Obj}(\phi, P),$$

and let $P^\star$ be an optimal solution to the outer supremum of the problem

$$\sup_{P \in \mathcal{W}(\delta)} \inf_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P).$$

Then $(\phi^\star, P^\star)$ is a Nash equilibrium of (22).

PROOF OF LEMMA A.1. We have

$$\sup_{P \in \mathcal{W}(\delta)} \inf_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P) = \inf_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P^\star) \leq \mathrm{Obj}(\phi^\star, P^\star)$$

$$\leq \sup_{P \in \mathcal{W}(\delta)} \mathrm{Obj}(\phi^\star, P) = \inf_{\phi \in \mathcal{M}} \sup_{P \in \mathcal{W}(\delta)} \mathrm{Obj}(\phi, P).$$

By strong duality, all inequalities become equalities. Thus, we have

$$\mathrm{Obj}(\phi^\star, P^\star) = \min_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P^\star) = \max_{P \in \mathcal{W}(\delta)} \mathrm{Obj}(\phi^\star, P).$$

This completes the proof.

$\square$

PROOF OF THEOREM 3.2. Recall that we denote

$$Q_N^\star := \mathcal{L}^{P_N^\star}\big(\langle b, e_1 \rangle, \cdots, \langle b, e_N \rangle, \epsilon\big)$$

in the proof to Claim 2 of Theorem 3.1. We also denote additionally

$$P_\star^{(N)} := \mathcal{L}^{P_N^\star}\Big(\big(\sum_{1 \leq n \leq N} e_n \langle b, e_n \rangle, \epsilon\big)\Big).$$

Letting $k^{(N)}(x,x) = \mathbb{E}_{P_\star^{(N)}}[b(x)^2]$, the objective value $\mathrm{Obj}(\phi_N^\star, P_\star^{(N)})$ can be written as

$$\mathrm{Obj}(\phi_N^\star, P_\star^{(N)})$$

$$= \int_{\mathcal{D}} \mathbb{E}_{P^{(N)}}[\mathbb{V}\mathrm{ar}_{P^{(N)}}[b(x)|Y_1, \ldots, Y_m]]\mathrm{d}x$$

$$= \int_{\mathcal{D}} k^{(N)}(x,x)$$

$$- \left(k_\epsilon^{(N)}(x,x_1), \ldots, k_\epsilon^{(N)}(x,x_m) \right) (K_\epsilon^{(N)})^{-1} \left(k_\epsilon^{(N)}(x,x_1), \ldots, k_\epsilon^{(N)}(x,x_m) \right)^\top \mathrm{d}x.$$

While the limiting objective value $\mathrm{Obj}(\phi_\infty^\star, P_\infty^\star)$ is

$$\mathrm{Obj}(\phi_\infty^\star, P_\infty^\star)$$

$$= \int_{\mathcal{D}} \mathbb{E}_{P_\infty^\star}[\mathbb{V}\mathrm{ar}_{P_\infty^\star}[b(x)|Y_1, \ldots, Y_m]]\mathrm{d}x$$

$$= \int_{\mathcal{D}} k_{P_\infty^\star}(x,x) - \left(k_{\epsilon,P_\infty^\star}(x,x_1), \ldots, k_{\epsilon,P_\infty^\star}(x,x_m) \right) \cdot (K_{\epsilon,P_\infty^\star})^{-1} \cdot \left(k_{\epsilon,P_\infty^\star}(x,x_1), \ldots, k_{\epsilon,P_\infty^\star}(x,x_m) \right)^\top \mathrm{d}x,$$

where $k_{P_\infty^\star}(x,x) = \mathbb{E}_{P_\infty^\star}[b(x)^2]$. Recall that $P_{N_l}^\star$ is a (Gaussian) subsequence of $P_N^\star$, which converges weakly to $P_\infty^\star$. Note that T is a bounded linear operator and that the tail-difference in $P_\star^{(N_l)}$ and $P_{N_l}^\star$ is negligible. By [20, Exercise 2.1.4], convergence in distribution implies convergence of the second moment, thus uniformly for $x \in \mathcal{D}$, $k^{(N_l)}(x,x) \to k_{P_\infty^\star}(x,x)$ and $k_\epsilon^{(N_l)}(x,x_i) \to k_{\epsilon,P_\infty^\star}(x,x_i)$. Also, it holds that $K_\epsilon^{(N_l)} \to K_{\epsilon,P_\infty^\star}$. Thus, we have

$$\mathrm{Obj}(\phi_{N_l}^\star, P_\star^{(N_l)}) = \min_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P_\star^{(N_l)}) \to \mathrm{Obj}(\phi_\infty^\star, P_\infty^\star) = \min_{\phi \in \mathcal{M}} \mathrm{Obj}(\phi, P_\infty^\star),$$

and hence that $P_\infty^\star$ solves the right-hand side of (22).

We next show that $\phi_\infty^\star$ solves the left-hand side of (22). We first note that due to the convergence of moments of $P_{N_l}^\star$, it holds that

$$\lim_{l \to \infty} \phi_{N_l}^\star(Y_1, \ldots, Y_m)(x) = \phi_\infty^\star(Y_1, \ldots, Y_m)(x), \quad \forall Y_1, \ldots, Y_m \in \mathbb{R}, \quad x \in \mathcal{D},$$

Next, by Fatou's lemma, we have that for any $P \in \mathcal{W}(\delta)$,

$$\mathbb{E}_P \left[\|b - \phi_\infty^\star(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2 \right] \leq \liminf_{l \to \infty} \mathbb{E}_P \left[\|b - \phi_{N_l}^\star(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2 \right].$$

Therefore,

$$\sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_P \left[\|b - \phi_\infty^\star(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2 \right] \leq \liminf_{l \to \infty} \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_P \left[\|b - \phi_{N_l}^\star(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2 \right].$$

Note that due to the finite-dimensional strong duality (19), $\phi_N^\star$ also solves

$$\min_{\phi \in \mathcal{M}_{\mathrm{aff},N}} \max_{P \in \mathcal{W}_N(\delta)} \mathrm{Obj}(\phi, P^{(N)}).$$

By (21) and strong duality, we have

$$\liminf_{N \to \infty} \sup_{P \in \mathcal{W}(\delta)} \mathbb{E}_P \left[\|b - \phi_N^\star(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2 \right] = \min_{\phi \in \mathcal{M}} \sup_{P \in \mathcal{W}(\delta)} \mathrm{Obj}(\phi, P)$$

Thus the estimator $\phi_\infty^\star$ solves the min-max problem on the left-hand side of (22).

By Lemma A.1, since $\phi_\infty^\star$ has to solve $\min_{\phi\in\mathcal{M}} \mathrm{Obj}(\phi, P_\infty^\star)$, we have that $\phi_\infty^\star$ is uniquely determined (independent of the choice of the subsequence $\phi_{N_l}^\star$), thus

$$\lim_{N\to\infty} \phi_N^\star(Y_1,\ldots,Y_m)(x) = \phi_\infty^\star(Y_1,\ldots,Y_m)(x) \quad \forall\, Y_1,\ldots,Y_m \in \mathbb{R}, \quad x\in\mathcal{D}.$$

Moreover, by Lemma A.1, $P_\infty^\star$ has to solve $\max_{P\in\mathcal{W}(\delta)} \mathrm{Obj}(\phi_\infty^\star, P)$. We show in the sequel that $P_\infty^\star$ is uniquely determined (independent of the choice of the subsequence $P_{N_l}^\star$), from which the sequence $P_N^\star$ converges to $P_\infty^\star$ in the weak topology as a whole. By Theorem 1 in [8], we have the reformulation

$$\max_{P\in\mathcal{W}(\delta)} \mathrm{Obj}(\phi_\infty^\star, P)$$

$$= \inf_{\gamma\geq 0} \Big(\gamma\delta^2 +$$

$$\mathbb{E}_{P_0}\Big[\sup_{(b,\epsilon)\in C(\mathcal{D})\times\mathbb{R}^m} \Big(\|b - \phi_\infty^\star((T(b)(x_i)+\epsilon_i)_i)\|_{L^2(\mathcal{D})}^2 - \gamma(\|b-b^0\|_{\mathcal{H}_w}^2 + \|\epsilon-\epsilon^0\|_2^2) \Big) \Big] \Big).$$

We claim that the optimal dual $\gamma^\star$ is sufficiently large so that for all $(b,\epsilon)\in\mathcal{H}_w\times\mathbb{R}^m$ such that $(b,\epsilon)\neq 0$:

$$(36) \qquad \|b - \phi_\infty^\star((T(b)(x_i)+\epsilon_i)_i)\|_{L^2(\mathcal{D})}^2 - \gamma^\star(\|b\|_{\mathcal{H}_w}^2 + \|\epsilon\|_2^2) < 0.$$

It is easy to see that $\gamma^\star > 0$. Suppose that (36) does not hold, then

$$(37) \qquad \|b^\star - \phi_\infty^\star((T(b^\star)(x_i)+\epsilon_i^\star)_i)\|_{L^2(\mathcal{D})}^2 - \gamma^\star(\|b^\star\|_{\mathcal{H}_w}^2 + \|\epsilon^\star\|_2^2) \geq 0,$$

for some $(b^\star,\epsilon^\star)\in\mathcal{H}_w\times\mathbb{R}^m$ and $(b^\star,\epsilon^\star)\neq 0$. Then for all $(b^0,\epsilon^0)\in C(\mathcal{D})\times\mathbb{R}^m$ satisfying

$$\langle b^\star - \phi_\infty^\star((T(b^\star)(x_i)+\epsilon_i^\star)_i), b^0 - \phi_\infty^\star((T(b^0)(x_i)+\epsilon_i^0)_i)\rangle \neq 0,$$

we have that

$$\sup_{(b,\epsilon)\in C(\mathcal{D})\times\mathbb{R}^m} \Big(\|b - \phi_\infty^\star((T(b)(x_i)+\epsilon_i)_i)\|_{L^2(\mathcal{D})}^2 - \gamma^\star(\|b-b^0\|_{\mathcal{H}_w}^2 + \|\epsilon-\epsilon^0\|_2^2) \Big)$$

$$\geq \sup_t \Big(\|t\,(b^\star - \phi_\infty^\star((T(b^\star)(x_i)+\epsilon_i^\star)_i)) + (b^0 - \phi_\infty^\star((T(b^0)(x_i)+\epsilon_i^0)_i)) \|_{L^2(\mathcal{D})}^2 \Big)$$

$$\quad - t^2\gamma^\star(\|b^\star\|_{\mathcal{H}_w}^2 + \|\epsilon^\star\|_2^2)$$

$$\geq \sup_t t^2 \Big(\|b^\star - \phi_\infty^\star((T(b^\star)(x_i)+\epsilon_i^\star)_i)\|_{L^2(\mathcal{D})}^2 - \gamma^\star(\|b^\star\|_{\mathcal{H}_w}^2 + \|\epsilon^\star\|_2^2) \Big)$$

$$\quad + 2t\langle b^\star - \phi_\infty^\star((T(b^\star)(x_i)+\epsilon_i^\star)_i), b^0 - \phi_\infty^\star((T(b^0)(x_i)+\epsilon_i^0)_i)\rangle$$

$$\quad + \|b^0 - \phi_\infty^\star((T(b^0)(x_i)+\epsilon_i^0)_i)\|_{L^2(\mathcal{D})}^2$$

$$= \infty,$$

where the infinite optimal value is due to the nonnegativity of the coefficient associated with the quadratic term t^2. Moreover, observe that the constraint

$$(38) \qquad \langle b^\star - \phi_\infty^\star((T(b^\star)(x_i)+\epsilon_i^\star)_i), b^0 - \phi_\infty^\star((T(b^0)(x_i)+\epsilon_i^0)_i)\rangle = 0$$

is a linear constraint in (b^0,ϵ^0), thus the collection of (b^0,ϵ^0) satisfying (38), denoted by $\mathcal{C}_{(b^\star,\epsilon^\star)}$, is a linear subspace of $C(\mathcal{D})\times\mathbb{R}^m$. This collection is closed (with respect to the product norm $\|\cdot\|_{C(\mathcal{D})}\times\|\cdot\|_{\ell_2}$) because $T: C(\mathcal{D})\to C(\mathcal{D})$ is bounded. Note that the support of P_0, denoted by $\mathrm{supp}(P_0)$, contains basis vectors of the form

$$\bar{B} = \{(e_k, 0,\ldots,0)\}_{k=1}^\infty \cup \{(\mathbf{0}, \bar{e}_j)\}_{j=1}^m$$

where $\{e_k\}_{k=1}^{\infty}$ is the orthonormal basis of $C(\mathcal{D})$, $\mathbf{0} \in C(\mathcal{D})$ is the zero-function, and $\bar{e}_j \in \mathbb{R}^m$ denotes the j-th standard basis on $\mathbb{R}^m$. Suppose that $\bar{B} \subseteq \mathcal{C}_{(b^\star, \epsilon^\star)}$. Since $b^\star \in \mathcal{H}_w \subseteq \mathcal{H}_{\tilde{w}}$, by expanding $(b^\star, \epsilon^\star)$ in the basis vectors from $\bar{B}$ (which converges in the product norm $\|\cdot\|_{\mathcal{H}_{\tilde{w}}} \times \|\cdot\|_{\ell_2}$), and considering that $T : \mathcal{H}_{\tilde{w}} \to \mathcal{H}_{\tilde{w}}$ is bounded from Assumption 2.3 (ii) as well as the fact that $\|\cdot\|_{\mathcal{H}_{\tilde{w}}}$ is a stronger norm than the $L^2(\mathcal{D})$-norm, we conclude that $(b^\star, \epsilon^\star) \in \mathcal{C}_{(b^\star, \epsilon^\star)}$, which is a contradiction to (37). Hence there is at least one basis vector of $\bar{B}$ not in $\mathcal{C}_{(b^\star, \epsilon^\star)}$. Since $\text{supp}(P_0)$ is the smallest closed subset of $C(\mathcal{D}) \times \mathbb{R}^m$ on which the probability measure evaluates to one, we have that $P_0(\mathcal{C}_{(b^\star, \epsilon^\star)}) < 1$. This concludes the proof of claim (36).

Denoting the shorthand

$$J(b, \epsilon) = \|b - \phi_\infty^\star((T(b)(x_i) + \epsilon_i)_i)\|_{L^2(\mathcal{D})}^2 - \gamma^\star(\|b - b^0\|_{\mathcal{H}_w}^2 + \|\epsilon - \epsilon^0\|_2^2),$$

any maximizer of J has to satisfy the first-order optimality condition

$$\left.\frac{\partial}{\partial t} J(b + th_b, \epsilon + th_\epsilon)\right|_{t=0} = 0 \quad \forall (h_b, h_\epsilon) \in C(\mathcal{D}) \times \mathbb{R}^m.$$

It is easy to compute that

$$\left.\frac{1}{2}\frac{\partial}{\partial t} J(b + th_b, \epsilon + th_\epsilon)\right|_{t=0}$$
$$= \langle b - \phi_\infty^\star((T(b)(x_i) + \epsilon_i)_i), h_b - \phi_\infty^\star((T(h_b)(x_i) + (h_\epsilon)_i)_i)\rangle$$
$$- \gamma^\star\left(\langle b - b^0, h_b\rangle_{\mathcal{H}_w} + \langle \epsilon - \epsilon^0, h_\epsilon\rangle\right).$$

Suppose that (b, ϵ) and $(\tilde{b}, \tilde{\epsilon})$ are two maximizers of J, by choosing $(h_b, h_\epsilon) = (b - \tilde{b}, \epsilon - \tilde{\epsilon})$, we have

$$\|h_b - \phi_\infty^\star((T(h_b)(x_i) + (h_\epsilon)_i)_i)\|_{L^2(\mathcal{D})}^2 - \gamma^\star(\|h_b\|_{H_w}^2 + \|h_\epsilon\|_2^2) = 0.$$

From our earlier claim (36), we conclude that $(b, \epsilon) = (\tilde{b}, \tilde{\epsilon})$. By Theorem 1 in [8], this shows that the solution to $\max_{P \in \mathcal{W}(\delta)} \text{Obj}(\phi_\infty^\star, P)$ is unique. $\square$

A.5. Proof of Theorem 3.3.

PROOF OF THEOREM 3.3. Since $P_N^\star \Rightarrow P_\infty^\star$ as $N \to \infty$, and each $P_N^\star$ is Gaussian, it follows directly that $P_\infty^\star$ is also a Gaussian process. Since $\phi_\infty^\star$ has to solve $\text{Obj}(\phi_\infty^\star, P_\infty^\star) = \min_{\phi \in \mathcal{M}} \text{Obj}(\phi, P_\infty^\star)$, $\phi_\infty^\star$ is the conditional mean of the worst-case distribution $P_\infty^\star$, which is affine in the observations.

$\square$

A.6. Proof of Proposition 3.3.

PROOF OF PROPOSITION 3.3. First, suppose that the following statement holds true:
(39)
"$K_{\epsilon, P_\infty^\star}$ is invertible for the limit $P_\infty^\star$ of some weakly convergent subsequence of $P_N^\star$".

Then by Theorem 3.2, the sequence $P_N^\star$ weakly converges to $P_\infty^\star$ as a whole. Since the sequence $P_N^\star$ is Gaussian, we have $\det(K_\epsilon^{(N)}) \to \det(K_{\epsilon, P_\infty^\star}) \neq 0$ as $N \to \infty$.

Next, suppose that condition (39) does not hold, and suppose that $\det(K_\epsilon^{(N)})$ does not converge to 0 as $N \to \infty$. Then there exists a subsequence of $P_N^\star$, denoted by $P_{N_l}^\star, l \geq 1$,

such that the sequence of determinants $\det(K_\epsilon^{(N_l)})$ is uniformly bounded away from 0. Furthermore, $P_{N_l}^\star$ has a weak limit (upon passing into a further subsequence, e.g., by Proposition 3.2). It is clear that condition (39) applies to this subsequence $P_{N_l}^\star$, which is a contradiction! $\hfill\square$

A.7. Proof of Lemma 3.4.

PROOF OF LEMMA 3.4. Let P be such that $\mathbb{W}(P, P_0) \leq \delta$. To show that the matrix in question is strictly positive definite, it suffices to show that

$$\inf_{\xi \in \mathbb{R}^d : \|\xi\|_2 = 1} \xi^\top \left(\mathbb{E}_P[Y_i Y_j]\right)_{ij} \xi = \inf_{\xi \in \mathbb{R}^d : \|\xi\|_2 = 1} \mathbb{E}_P \left[\left(\sum_{i=1}^m \xi_i(u(x_i) + \epsilon_i)\right)^2\right] > 0.$$

To this end, let π be a coupling such that

$$\mathbb{E}_\pi \left[c((b, \epsilon), (b^0, \epsilon^0))\right] = \mathbb{E}_\pi \left[\|\epsilon - \epsilon^0\|_2^2 + \|b - b^0\|_{\mathcal{H}_w}^2\right] \leq \delta^2,$$

where the marginal distribution of (b, ϵ) and (b^0, ϵ^0) are P and P_0, respectively. We denote $u = T(b)$ and $u^0 = T(b^0)$. Let C_m be the constant such that (due to the RKHS property of $\mathcal{H}_{\tilde{w}}$)

$$|f(x_i)| \leq C_m \|f\|_{\mathcal{H}_{\tilde{w}}} \qquad \forall f \in \mathcal{H}_{\tilde{w}}, \ i = 1, \ldots, m.$$

By Fatou's lemma

$$\mathbb{E}_\pi \left[(u(x_i) - u^0(x_i))^2\right] \leq \liminf_{N \to \infty} \mathbb{E}_\pi \left[\left(\sum_{k=1}^N \langle b - b^0, e_k \rangle T(e_k)(x_i)\right)^2\right]$$

$$= \liminf_{N \to \infty} \mathbb{E}_\pi \left[\left(T\left(\sum_{k=1}^N \langle b - b^0, e_k \rangle e_k\right)(x_i)\right)^2\right]$$

$$\leq C_m^2 \liminf_{N \to \infty} \mathbb{E}_\pi \left[\left\|T(\sum_{k=1}^N \langle b - b^0, e_k \rangle e_k)\right\|_{\mathcal{H}_{\tilde{w}}}^2\right]$$

$$\leq C_m^2 C_{\tilde{w}}^2 \liminf_{N \to \infty} \mathbb{E}_\pi \left[\left\|\sum_{k=1}^N \langle b - b^0, e_k \rangle e_k\right\|_{\mathcal{H}_{\tilde{w}}}^2\right]$$

$$\leq C_m^2 C_{\tilde{w}}^2 \tilde{C} \liminf_{N \to \infty} \mathbb{E}_\pi \left[\left\|\sum_{k=1}^N \langle b - b^0, e_k \rangle e_k\right\|_{\mathcal{H}_w}^2\right]$$

$$(40) \qquad\qquad \leq C_m^2 C_{\tilde{w}}^2 \tilde{C} \mathbb{E}_\pi \left[\|b - b^0\|_{\mathcal{H}_w}^2\right],$$

where the constant $C_{\tilde{w}}$ is from Assumption 2.3(ii), and the constant $\tilde{C}$ is due to Assumption 2.5.

For any $\xi \in \mathbb{R}^d$ with $\|\xi\|_2 = 1$,

$$\mathbb{E}_\pi \left[\left(\sum_{i=1}^m \xi_i(u^0(x_i) + \epsilon_i^0)\right)^2\right]$$

$$= \mathbb{E}_\pi \left[\left(\sum_{i=1}^m \xi_i(u(x_i) + \epsilon_i) + \sum_{i=1}^m \xi_i(u^0(x_i) - u(x_i) + \epsilon_i^0 - \epsilon_i) \right)^2 \right]$$

$$\leq 2\mathbb{E}_\pi \left[\left(\sum_{i=1}^m \xi_i(u(x_i) + \epsilon_i) \right)^2 \right] + 4\mathbb{E}_\pi \left[\left(\sum_{i=1}^m \xi_i(u^0(x_i) - u(x_i)) \right)^2 \right]$$

$$+ 4\mathbb{E}_\pi \left[\left(\sum_{i=1}^m \xi_i(\epsilon_i^0 - \epsilon_i) \right)^2 \right]$$

$$\leq 2\mathbb{E}_P \left[\left(\sum_{i=1}^m \xi_i(u(x_i) + \epsilon_i) \right)^2 \right] + 4\mathbb{E}_\pi \left[\sum_{i=1}^m (u^0(x_i) - u(x_i))^2 \right] + 4\mathbb{E}_\pi[\|\epsilon^0 - \epsilon\|_2^2]$$

$$\leq 2\mathbb{E}_P \left[\left(\sum_{i=1}^m \xi_i(u(x_i) + \epsilon_i) \right)^2 \right] + O(1)\delta^2,$$

where we used the previous inequality (40) to estimate $\mathbb{E}_\pi \left[\sum_{i=1}^m (u^0(x_i) - u(x_i))^2 \right]$. Note that $(\mathbb{E}_{P_0}[Y_i Y_j])_{ij} \succeq \sigma^2 I_{m \times m}$, since ϵ_i^0 are independent $\mathcal{N}(0, \sigma^2)$ noise under P_0. Therefore,

$$\inf_{\xi \in \mathbb{R}^d, \|\xi\|_2 = 1} \mathbb{E}_P \left[\left(\sum_{i=1}^m \xi_i(u(x_i) + \epsilon_i) \right)^2 \right] \geq \frac{1}{2}\sigma^2 - O(1)\delta^2 > 0,$$

for all $\delta < \delta_0$, where δ_0 is a positive constant that depends on $T, m, (x_i)_i, \mathcal{H}_w, \mathcal{H}_{\tilde{w}}$ and σ^2. This completes the proof. $\square$

A.8. Proof of Theorem 3.5. The proof for the inverse problem works verbatim as the proof to Theorem 3.1 with minor modifications. Instead of considering $\text{coef}(\phi) \in \text{span}\{e_n\}_{n=1}^N$, we consider $\text{coef}(\phi) \in \text{span}\{T(e_n)\}_{n=1}^N$. Additionally, in the proof to Theorem 3.1 Claim 2, we modify $\{\tilde{e}_k\}_{k=1}^\infty$ to be the Gram-Schmidt orthonormalization of $\{T(e_k)\}_{k=1}^\infty$. In the process of orthonormalizing $\{T(e_k)\}_{k=1}^\infty$, for any n such that the function $T(e_n)$ is already included in the span of $\{T(e_k)\}_{k=1}^{n-1}$, we define $\tilde{e}_n$ as the zero function.

A.9. Details of Frank-Wolfe Algorithm. To numerically compute the Nash equilibrium of the min-max problem in the finite-dimensional approximation, we borrow the computational technique developed in [36]. In particular, we note that the finite-dimensional problem (11) has the same structure as [36, Equation (2)]. To see this, denoting

$$Y = (Y_1, \ldots, Y_m),$$

we can rewrite problem (11) as

(41a)
$$\min_{\phi \in \mathcal{M}} \max_{Q^{(N)} : W_N(Q^{(N)}, Q_0^{(N)}) \leq \delta} \text{Obj}(\phi, Q^{(N)})$$

(41b)
$$= \min_{\psi \in \mathcal{M}_N} \max_{Q^{(N)} : W_N(Q^{(N)}, Q_0^{(N)}) \leq \delta} \mathbb{E}_{Q^{(N)}} \left[\sum_{n=1}^N (\langle b, e_n \rangle - \psi_n(Y))^2 \right],$$

where $\mathcal{M}_N$ denotes the family of all measurable functions from $\mathbb{R}^m$ to $\mathbb{R}^N$, and $\psi = (\psi_1, \ldots, \psi_N)$. We note that the vector $(\{\langle b, e_n \rangle\}_{n=1}^N, Y)$ can be obtained by the matrix transform

$$(\{\langle b, e_n \rangle\}_{n=1}^N, Y) = A \cdot (\{\langle b, e_n \rangle\}_{n=1}^N, \epsilon),$$

where $\cdot$ denotes matrix multiplication, and

$$A = \begin{pmatrix} I_N & 0 \\ A_{21} & I_m \end{pmatrix}, \quad (A_{21})_{jn} = T(e_n)(x_j) \;\forall 1 \le n \le N, 1 \le j \le m.$$

Hence the covariance of $(\{\langle b, e_n \rangle\}_{n=1}^N, Y)$ under the nominal measure is given by

$$A\Sigma_0 A^\top,$$

where

$$\Sigma_0 = \mathbb{Cov}_{Q_0^{(N)}} \left((\{\langle b, e_n \rangle\}_{n=1}^N, \epsilon) \right).$$

By [36, Theorem 2.5] and its proof, the finite-dimensional problem (41b) is thus equivalent to the finite convex program

$$\sup \quad \mathrm{tr}\left[S_{11} - S_{12} S_{22}^{-1} S_{21} \right]$$

$$\text{s.t.} \quad S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \in \mathbb{S}_{++}^{N+m}, \quad S_{11} \in \mathbb{S}_{++}^N, \quad S_{22} \in \mathbb{S}_{++}^m, \quad S_{12} = S_{21}^\top \in \mathbb{R}^{N \times m},$$

$$(42) \qquad \mathrm{tr}\left[BSB^\top + \tilde{\Sigma} - 2\left(\tilde{\Sigma}^{1/2} BSB^\top \tilde{\Sigma}^{1/2} \right)^{1/2} \right] \le \delta^2, \quad S \succeq \overline{\sigma} I_{N+m}.$$

where $B = \sqrt{W_N} A^{-1}, \tilde{\Sigma} = \sqrt{W_N} \Sigma_0 \sqrt{W_N}$, and $\overline{\sigma}$ is the minimum eigenvalue of $A\Sigma_0 A^\top$. Assuming the finite convex program has optimal solutions $S^\star, S_{11}^\star, S_{22}^\star$ and $S_{12}^\star$. Then the optimal solution $\psi^\star$ to equation (41b) is $\psi^\star(Y) = S_{12}^\star (S_{22}^\star)^{-1} Y$, and hence the optimal solution $\phi^\star$ to equation (41a) is $\phi^\star = \sum_{n=1}^N \psi_n^*(Y) e_n$. The worst-case prior on $(\{\langle b, e_n \rangle\}_{n=1}^N)$ is given by a finite-dimensional Gaussian with zero-mean and covariance matrix S_{11}^* and the worst-case posterior on $(\{\langle b, e_n \rangle\}_{n=1}^N)$ (conditional on Y) is given by a finite-dimensional Gaussian with zero-mean and covariance matrix $S_{11}^\star - S_{12}^\star (S_{22}^\star)^{-1} S_{21}^\star$.

The finite convex problem (42) is a nonlinear semi-definite program (SDP) and to solve it tractably [36] proposed a tailored first-order method (Frank-Wolfe algorithm). The crucial observation in [36, Definition 3.1] is that the function $f(S) = \mathrm{tr}\left[S_{11} - S_{12} S_{22}^{-1} S_{21} \right]$ satisfies the relation

$$f(S) = \langle S, \nabla f(S) \rangle,$$

because

$$\langle S, \nabla f(S) \rangle = \left\langle \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}, \begin{pmatrix} I_N & -S_{12} S_{22}^{-1} \\ -S_{22}^{-1} S_{21} & S_{22}^{-1} S_{21} S_{12} S_{22}^{-1} \end{pmatrix} \right\rangle = f(S).$$

Hence, the objective in problem (42) can be replaced with a linear approximation. In particular, following the iterative algorithm in [36, Equations 7(a) and 7(b)] we consider

$$(43) \qquad\qquad S^{(k+1)} = a_k F(S^k) + (1 - a_k) S^k, \forall k \ge 0,$$

with $S^{(0)} = B^{-1} \tilde{\Sigma} B^{-\top} = A\Sigma_0 A^\top$ is the starting solution, a_k represents a judiciously chosen step-size, and

$$(44) \quad F(S) = \begin{cases} \arg\max \langle L, \nabla f(S) \rangle, \\ \qquad \text{s.t.} \; L \succeq \bar{\sigma} I_{N+m}, \; \mathrm{tr}\left[BLB^\top + \tilde{\Sigma} - 2\left(\tilde{\Sigma}^{1/2} BLB^\top \tilde{\Sigma}^{1/2} \right)^{1/2} \right] \le \delta^2. \end{cases}$$

Problem (44) can be further rewritten by considering the transformation $\tilde{L} \leftarrow BLB^\top$:

$$(45) \qquad F(S) = \begin{cases} \arg\max \langle \tilde{L}, B^{-\top} \nabla f(S) B^{-1} \rangle, \\ \qquad \text{s.t.} \; \tilde{L} \succeq \tilde{\sigma} I_{N+m}, \; \mathrm{tr}\left[\tilde{L} + \tilde{\Sigma} - 2\left(\tilde{\Sigma}^{1/2} \tilde{L} \tilde{\Sigma}^{1/2} \right)^{1/2} \right] \le \delta^2, \end{cases}$$

which is exactly [36, Equations 7(b)].

In conclusion, the finite-dimensional min-max problem (41a) can be solved using the iterations (43) with the oracle (45), which is exactly [36, Equations 7(a) and 7(b)]. Therefore, the Frank-Wolfe algorithm in [36] can be applied to numerically compute the distributionally robust predictor and the worst-case prior. We refer interested readers to [36, Algorithms 1 and 2] for the concrete algorithm and the analysis thereof, which is beyond the scope of this work.

APPENDIX: B LIST OF NOTATIONS

$(\phi_\infty^\star, P_\infty^\star)$ pair of Nash equilibrium in infinite-dimensions 13

$(\phi_N^\star, Q_N^\star)$ pair of Nash equilibrium in finite-dimensional approximation 11

α smoothness parameter in Example 2.1 5

α_j coefficient functions in affine predictor 18

β roughness parameter in Example 2.2 6

δ radius of the Wasserstein Ball 8

ϵ denotes a generic choice of observation errors by nature 8

ϵ^0 nominal vector of errors 7

κ_n eigen-values in Karhunen Loeve expansion 5

λ_n Dirichlet-Laplacian eigenvalues in Example 2.1 5

$\mathbb{S}_+^m$ positive semi-definite matrix of dimension m-by-m 23

$\mathcal{M}$ space of measurable maps 7

$\mathcal{M}_{\text{aff}}$ affine subspace of $\mathcal{M}$ 18

$\mathcal{P}$ space of (Borel) probability measures 7

$\mathcal{W}_N(\delta)$ Wasserstein balls around P_0 arising from perturbations in the first N coordinates 19

$\mathcal{M}_N$ measurable functions from $\mathbb{R}^m$ to $\mathbb{R}^N$ 38

$\mathcal{D}$ compact set 5

$\mathcal{H}_w$ RKHS for perturbations 6

$\mathcal{L}^P$ the law under P 12

$\mathcal{M}_{\text{aff},N}$ affine subspace of $\mathcal{M}$ with $\text{coef}(\phi) \in \text{span}\{e_n\}_{n=1}^N$ 19

$\mathcal{S}_N$ modified Gelbrich distance 24

$\mathcal{W}(\delta)$ Wassertein ball around P_0 of radius δ 19

$\mathcal{W}_{\text{nor},N}(\delta)$ Wasserstein balls around P_0 arising from perturbations in the first N coordinates and restricting to the family of Gaussian distributions 25

$\mathbb{W}$ Wasserstein distance 8

$\mathbb{S}_{++}^m$ positive-definite matrix of dimension m-by-m 11

ϕ non-parametric predictor for b 7

ϕ_u non-parametric predicator for u 13

π coupling of probability measures 8

Σ_0 covariance matrix of $Q_0^{(N)}$ 24

$\text{span}\{e_n\}_{n=1}^N$ linear subspace of spanned by the first N basis vectors 10

$\mathbb{W}_N$ induced optimal transport cost on $\mathbb{R}^{N+m}$ 11

$\{e_n\}_{n=1}^\infty$ orthonormal system of $L^2(\mathcal{D})$ 5

b denotes a generic choice of input process by nature 8

b^0 prior input process 5

c ground cost function 8

c_N induced cost function on $\mathbb{R}^{N+m}$ 11

$K_\epsilon^{(N)}$ covariance matrix of observations Y under $Q_N^\star$ 11

$k_\epsilon^{(N)}(x, x_i)$ covariance function of $b(x)$ and Y_i under $Q_N^\star$ 11

$k_P(x,x)$ variance function of $b(x)$ under P 27
$K_{\epsilon,P}$ covariance matrix of observations Y under P 12
$k_{\epsilon,P}(x,x_i)$ covariance function of $b(x)$ and Y_i under P 12
m number of design points 7
N number of basis vectors in finite-dimensional approximation 10
P denotes a generic choice of probability measure by nature 8
$P^{(N)}$ Projection of P onto $\mathrm{span}\{e_n\}_{n=1}^{N}$ whilst keeping ϵ intact 20
P_0 nominal probability measure 5
$P_\star^{(N)}$ truncation of $P_N^\star$ without the tails 34
$P_N^\star$ the probability measure composing of $Q_N^\star$ and the tail of P_0 12
$Q^{(N)}$ induced measure of P in $\mathbb{R}^{N+m}$ 10
$Q_0^{(N)}$ induced measure of P_0 in $\mathbb{R}^{N+m}$ 10
T linear forward operator 6
u observation process $T(b)$ 6
W_N matrix with diagonals w_n 23
w_n Sequence of weights parametrizing the Hilbert space norm 6
x_i design points 7
Y_i observations at design point x_i 7

REFERENCES

[1] ACCIAIO, B., VERAGUAS, J. B. and ZALASHKO, A. (2017). Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization. *arXiv preprint arXiv:1611.02610*.

[2] AGAPIOU, S., LARSSON, S. and STUART, A. M. (2013). Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Processes and their Applications* **123** 3828–3860.

[3] ATKINSON, C. and MITCHELL, A. F. (1981). Rao's distance measure. *Sankhyā: The Indian Journal of Statistics, Series A* 345–365.

[4] BACKHOFF-VERAGUAS, J., BARTL, D., BEIGLBÖCK, M. and EDER, M. (2020). Adapted Wasserstein distances and stability in mathematical finance. *Finance and Stochastics* **24** 601–632.

[5] BION-NADAL, J. and TALAY, D. (2019). On a Wasserstein-type distance between solutions to stochastic differential equations. *The Annals of Applied Probability* **29** 1609 – 1639.

[6] BLANCHET, J., KANG, Y. and MURTHY, K. (2019). Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability* **56** 830–857.

[7] BLANCHET, J., LI, J., LIN, S. and ZHANG, X. (2024). Distributionally Robust Optimization and Robust Statistics. *arXiv preprint arXiv:2401.14655*.

[8] BLANCHET, J. and MURTHY, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research* **44** 565–600.

[9] BLANCHET, J., MURTHY, K. and NGUYEN, V. A. (2021). Statistical analysis of Wasserstein distributionally robust estimators. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications* 227–254. INFORMS.

[10] CAVALIER, L. (2011). Inverse problems in statistics. In *Inverse problems and high-dimensional estimation. Lect. Notes Stat. Proc.* **203** 3–96. Springer, Heidelberg. MR2868199

[11] CAVALIER, L. and TSYBAKOV, A. (2002). Sharp adaptation for inverse problems with random noise. *Probability Theory and Related Fields* **123** 323–354. https://doi.org/10.1007/s004400100169 MR1918537

[12] DASHTI, M. and STUART, A. M. (2017). *The Bayesian Approach to Inverse Problems* In *Handbook of Uncertainty Quantification* 311–428. Springer International Publishing.

[13] DELAGE, E. and YE, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* **58** 595–612.

[14] DONOHO, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied and Computational Harmonic Analysis* **2** 101–126. https://doi.org/10.1006/acha.1995.1008 MR1325535

[15] MOHAJERIN ESFAHANI, P. and KUHN, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* **171** 115–166.

[16] FÖRSTNER, W. and MOONEN, B. (2003). *A Metric for Covariance Matrices* In *Geodesy-The Challenge of the 3rd Millennium* 299–309. Springer Berlin Heidelberg.

[17] GAO, R., CHEN, X. and KLEYWEGT, A. J. (2017). Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*.

[18] GAO, R. and KLEYWEGT, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*.

[19] GHOSH, J. K. and RAMAMOORTHI, R. V. (2003). *Bayesian Nonparametrics*. *Springer Series in Statistics*. Springer.

[20] GINÉ, E. and NICKL, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

[21] GUGUSHVILI, S., VAN DER VAART, A. and YAN, D. (2020). Bayesian linear inverse problems in regularity scales. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **56** 2081 – 2107.

[22] KENNEDY, M. C. and O'HAGAN, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63** 425–464.

[23] KNAPIK, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Annals of Statistics* **39** 2626–2657. https://doi.org/10.1214/11-AOS920 MR2906881

[24] KUHN, D., ESFAHANI, P. M., NGUYEN, V. A. and SHAFIEEZADEH-ABADEH, S. (2019). Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning. *INFORMS TutORials in Operations Research* 130-166.

[25] LEVY, B. C. and NIKOUKHAH, R. (2004). Robust least-squares estimation with a relative entropy constraint. *IEEE Transactions on Information Theory* **50** 89-104.

[26] LEVY, B. C. and NIKOUKHAH, R. (2013). Robust State Space Filtering Under Incremental Model Perturbations Subject to a Relative Entropy Tolerance. *IEEE Transactions on Automatic Control* **58** 682-695.

[27] LI, J., LIN, S., BLANCHET, J. and NGUYEN, V. A. (2022). Tikhonov regularization is optimal transport robust under martingale constraints. *Advances in Neural Information Processing Systems* **35** 17677–17689.

[28] LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach [with Discussion]. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **73** 423–498.

[29] NATTERER, F. (2001). *The Mathematics of Computerized Tomography. Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics.

[30] NGUYEN, V. A., KUHN, D. and MOHAJERIN ESFAHANI, P. (2018). Distributionally Robust Inverse Covariance Estimation: The Wasserstein Shrinkage Estimator. *arXiv preprint arXiv:1805.07194*.

[31] NGUYEN, V. A., SHAFIEEZADEH-ABADEH, S., KUHN, D. and MOHAJERIN ESFAHANI, P. (forthcoming). Bridging Bayesian and Minimax Mean Square Error Estimation via Wasserstein Distributionally Robust Optimization. *Mathematics of Operations Research*.

[32] RAHIMIAN, H. and MEHROTRA, S. (2019). Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*.

[33] RAO, C. R. (1987). Differential metrics in probability spaces. *Differential Geometry in Statistical Inference* **10** 217–240.

[34] RAY, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electronic Journal of Statistics* **7** 2516 – 2549.

[35] SHAFIEEZADEH-ABADEH, S., KUHN, D. and ESFAHANI, P. M. (2019). Regularization via mass transportation. *Journal of Machine Learning Research* **20** 1–68.

[36] SHAFIEEZADEH-ABADEH, S., NGUYEN, V. A., KUHN, D. and MOHAJERIN ESFAHANI, P. (2018). Wasserstein distributionally robust Kalman filtering. In *Advances in Neural Information Processing Systems 31* 8483-8492.

[37] SION, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics* **8** 171–176.

[38] SKOROKHOD, A. V. (1956). Limit Theorems for Stochastic Processes. *Theory of Probability & Its Applications* **1** 261-290.

[39] SPANTINI, A., CUI, T., WILLCOX, K., TENORIO, L. and MARZOUK, Y. (2017). Goal-oriented optimal approximations of Bayesian linear inverse problems. *SIAM Journal on Scientific Computing* **39** S167–S196.

[40] SPANTINI, A., SOLONEN, A., CUI, T., MARTIN, J., TENORIO, L. and MARZOUK, Y. (2015). Optimal low-rank approximations of Bayesian linear inverse problems. *SIAM Journal on Scientific Computing* **37** A2451–A2487.

[41] STUART, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numerica* **19** 451–559.

[42] TAYLOR, M. E. (1996). *Partial Differential Equations I: Basic Theory. Applied mathematical sciences*. Springer.

[43] TAYLOR, M. E. (1996). *Partial Differential Equations II: Qualitative studies of linear equations. Applied mathematical sciences*. Springer.

[44] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh* 200–222.

[45] VILLANI, C. (2008). *Optimal Transport: Old and New*. Springer Science & Business Media.

[46] VON NEUMANN, J. and MORGENSTERN, O. (2004). *Theory of Games and Economic Behavior. Princeton classic editions*. Princeton University Press.

[47] YI, S. and ZORZI, M. (2022). Robust Kalman Filtering Under Model Uncertainty: The Case of Degenerate Densities. *IEEE Transactions on Automatic Control* **67** 3458-3471.

[48] ZORZI, M. (2017). Robust Kalman Filtering Under Model Perturbations. *IEEE Transactions on Automatic Control* **62** 2902-2907.

[49] ZORZI, M. (2017). On the robustness of the Bayes and Wiener estimators under model uncertainty. *Automatica* **83** 133-140.