Wasserstein Distributionally Robust Gaussian Process Regression and Linear Inverse Problems

Xuhui Zhang\textsuperscript{1}, Jose Blanchet\textsuperscript{1}, Youssef Marzouk\textsuperscript{2}, Viet Anh Nguyen\textsuperscript{3}, Sven Wang\textsuperscript{2}

\textsuperscript{1}Stanford University \hspace{1cm} \textsuperscript{2}MIT \hspace{1cm} \textsuperscript{3}VinAI Research

Abstract

We study a distributionally robust optimization formulation (i.e., a min-max game) for problems of nonparametric estimation: Gaussian process regression and, more generally, linear inverse problems. We choose the best mean-squared error predictor on an infinite-dimensional space against an adversary who chooses the worst-case model in a Wasserstein ball around an infinite-dimensional Gaussian model. The Wasserstein cost function 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 the game has a well-defined value (i.e., strong duality holds in the sense that max-min equals min-max) and show existence of 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 properties of the Nash equilibrium and the effects of hyperparameters through a set of numerical experiments, demonstrating the versatility of our modeling framework.

1 Introduction

Nonparametric estimation of a conditional expectation is a ubiquitous task in engineering and science, and other areas of statistical application. In general, this task involves solving an infinite dimensional optimization problem. The problem is also ill-posed because the underlying distribution of the covariates or predictors and the quantity to predict is unknown. However, under the assumption that the prediction quantity and the predictors jointly follow a Gaussian model, the prediction task is greatly simplified. In this case the best mean-square prediction is an affine function of the predictors. In addition to gaining tractability, the Gaussian-model enables the construction of confidence bands around the regression function, which is very convenient in terms of evaluating the prediction error and the level of confidence attached to such error estimate. Of course, all of this relies heavily on Gaussianity. There is a rich literature in non-parametric Bayesian statistics
that provides some comfort in the sense of asymptotic consistency as the number of observations increases to infinity under suitable assumptions [15, 30, 32]. However, our focus in this paper is on the case in which the number of observations is fixed and we are interested in taking advantage of the convenient properties of the Gaussian model (e.g. tractability and error quantification) while also incorporating that reality is most likely (if not with overwhelming certainty) not Gaussian.

More generally, we focus on solving ill-posed linear inverse problems of which regression estimation is just a particular case that arises when the associated linear operator is simply the identity map. Ultimately, this paper delivers an optimal approach for adding a layer of robustness in order to hedge against the impact in mean-squared error prediction implied by assuming the incorrect model. The approach results in using a modified Gaussian model which turns out to be distributionally robust in a non-parametric sense to be discussed in the sequel. Finally, the approach is practical since the distributionally robust model turns out to be also Gaussian and therefore the best-mean square prediction is affine in the observations. In simple words, our results imply that Gaussian models can be used to account not only to quantify prediction errors (assuming the Gaussian model is correct) but also to quantify actual model errors within a broad non-parametric class of models. In this paper, we address the issue of quantifying model misspecification using ideas borrowed from the distributionally robust optimization literature [5, 13, 24, 20, 10]. A distributionally robust optimization (DRO) formulation is a min-max game between the decision-maker and an adversary which is introduced to assess the impact of model misspecification in the decision-maker’s chosen criterion. In the nonparametric Gaussian regression setting, for instance, the decision-maker wishes to choose the best (in mean-square error) predictor against an adversary that chooses a model which departs (in a nonparametric way, subject to a budget constraint) from the baseline Gaussian model assumed by the decision maker. In order to allow nonparametric (and therefore non-Gaussian) adversarial specifications, we allow the adversary to select a model within a suitable $\delta$-Wasserstein ball around the nominal model. Because of the nature of the Wasserstein distance (to be specified later), the adversary is allowed to select models which depart significantly not only from the Gaussian assumption but from implicit smoothness regularity properties implied by the choice of the decision-maker’s prior. Consequently, our min-max formulation allows to efficiently explore and quantify the impact of model misspecification.

As mentioned earlier, our focus is not only on Gaussian process regression but we also consider more general tasks, including a class of linear inverse problems. Precisely, we consider equations which are linear in the sense that the solution of the equation, denoted as $u$ in the sequel, can be written in the form $u = T(b)$, where $b$ represents “input conditions” and $T$ represents a linear operator. This formulation includes, for example, the Laplace equation and the heat equation as particular cases. In these settings, for instance, $b$ may represent the boundary condition (e.g., Laplace’s equation) or the initial value (e.g., heat equation). We are interested in the case where $u$ and $b$ take values in suitable infinite-dimensional (Banach or Hilbert) spaces. Despite the infinite-dimensional nature of the
model, in reality one can only access finitely many noisy observations of the function $u$. The noise, which is often also assumed to be Gaussian, can be used to represent a degree of uncertainty in the operator $T$. Given the partial observations, we are interested in two tasks, either recovering the $u$ at other locations (corresponding to a regression problem) or in recovering the input $b$ (an inverse problem). The standard Gaussian regression case corresponds in this setting to letting $T$ being the identity map.

In the context of inverse problems, the Bayesian approach is to assume a prior on the unknown input $b$ (typically Gaussian) and computing the relevant posterior distribution given the observations by applying Bayes’ rule; see, for example, [27, 9] for a summary of this perspective in inverse problems. The success of this method critically relies on the accurate specification of the operator $T$, typically a differential or an integral equation, representing the underlying physical law and the correct elicitation of the prior and noise distributions, all of which are not guaranteed in practical complex systems. Moreover, the Gaussian prior specification can significantly impact the inverse problem task. Our DRO (min-max) formulation applies to the inverse problem setting quantifies not only the impact of specifying an incorrect probabilistic model but also, thorough the the noise in the observations, a degree of misspecification in the operator $T$ as well.

In more precise terms, the standard (either Gaussian regression or Bayesian inverse) problem that is considered in the community is the following. Under some nominal model, $P_0$, we have that $u : D \to \mathbb{R}$ is ultimately a continuous Gaussian random field taking values on a compact set $D$. The sample paths of $u$ are guaranteed to live on a Hilbert space $\mathcal{H}_w$ with norm $\| \cdot \|_{\mathcal{H}_w}$ which controls the sample-path smoothness of $u$ parametrically via a sequence of weights, $w$. We can access $Y_1, \ldots, Y_m$ (noisy) measurements of $u$. The noise and $u$ are jointly Gaussian under $P_0$. To estimate the ‘full path’ of $u$ from the noisy measurements, the decision maker chooses a mapping $\phi : \mathbb{R}^m \to \mathcal{H}_w$, subject to an optimality criterion that minimizes the $L_2(P_0 \times D)$ error (i.e., the $P_0$ expectation of the integrated squared error between the outcome and the estimated outcome at each location over $D$). To this standard framework, we add an adversarial layer. We introduce an adversary that can change the model $P_0$ so that $u$ and the noise may no longer be Gaussian and the resulting prediction function may have different smoothness properties than those implicitly assumed by the prior (by changing the weighting parameter $w$ in the Wasserstein distance specification). Moreover, the adversary is allowed to observe the decision maker’s map $\phi$. Precisely, we are interested in studying the following min-max game,

$$\min_{\phi} \max_{P: W(P,P_0) \leq \delta} \mathbb{E}_P \left[ \|u - \phi(Y_1, \ldots, Y_m)\|_{L^2(D)}^2 \right].$$ (1)

Here, $W$ is a suitable Wasserstein distance in which the underlying transportation cost is given by $\| \cdot \|_{\mathcal{H}_w}$. Typical Gaussian regression and Bayesian inverse problem specifications are recovered when the size of the uncertainty (i.e., the adversarial budget), $\delta$, is set equal to zero. An important novelty element from the standpoint of distributionally robust
optimization is to propose a Wasserstein distance \( W \) as a convenient vehicle to control features such as the amount of roughness that the adversary is allowed to inject in the sample path. As mentioned earlier, the hierarchy of norms \( \| \cdot \|_{H_w} \) are defined in terms of a sequence of weights which, combined with suitable basis functions, control the degree of smoothness in the elements of the underlying space. Our Wasserstein distance is designed to explore the impact of distributions with potentially rougher (and, more importantly, non-Gaussian) sample paths in the prediction problems.

Wasserstein-type distances defined in the space of stochastic processes were studied by [3, 1, 4]. Their focus is on processes indexed by a (one-dimensional) parameter representing time. Examples of these settings include price processes in finance [3], and the focus is on defining a Wasserstein distance that respects the casual structure (i.e., filtrations). In contrast, we consider uncertainty of roughness in a multi-parameter field setting.

The contributions of this paper include a theoretical analysis of the min-max game \( (1) \) and associated algorithmic developments. Note that in the distributionally robust formulation \( (1) \) we have an infinite dimensional action space for the outer player, an infinite dimensional action space for the inner player and, moreover, the actions of the inner player are actually probability measures on infinite dimensional spaces themselves. These features differentiate our analysis from prior work, such as [5, 20, 24]. In most (if not all, except for the current work) situation in the distributionally robust optimization literature, as far as we know, either the action set of the decision maker is finite dimensional or the probability measures (the action set of the inner player) are supported in finite dimensional spaces.

Our contributions are as follows (under reasonable assumptions to be discussed in our next sections).

- We show that strong duality holds for problem \( (1) \) in the sense that min-max value equals to the max-min value.
- We show that there exists \( \delta_0 > 0 \), such that if \( 0 < \delta < \delta_0 \) problem \( (1) \) also admits a unique Nash equilibrium pair. Moreover, the worst case distribution involves a modified Gaussian process with potentially rougher paths than the prior. Consequently, the robustified decision remains affine in the observations.
- Our analysis technique proceeds by approximating problem \( (1) \) by a sequence of finite-dimensional versions and therefore we obtain an algorithm for computing the associated Nash equilibrium. Our algorithm is an adaptation of the (finite-dimensional) Frank-Wolfe algorithm in [24].

One way to interpret our result is that Gaussian process regression (or inverse recovery) can be made strongly robust in a nonparametric sense. The worst-case covariance function is essential in understanding the error (since it is basically all is needed for computing a bound for the mean-squared error) and, in this sense, it provides the means for computing an upper bound for the quality of the robust regression function. The structural
properties of the worst-case covariance function and the associated distributionally robust regression function are explored in our numerical section. Further, numerical experiments are performed to explore the structure of the Nash equilibrium. We find that the worst-case distributions (prior and posterior) are perturbed so as to induce greater uncertainties in regions where information is limited, which intuitively guarantees greater robustness of the predictions. We observe that in cases where there is: a smoother prior; less penalty on higher modes that induce roughness; a larger $\delta$; or smaller noise, we see more easily that the behavior of the worst-case that induces sharper contrasts between the observed and unobserved locations in both prior and posterior sample paths.

In the rest of the paper, we introduce our notations and problem setup in Section 2. We present our main theoretical results in Section 3, illustrate the applicability of our general framework through a number of examples in Section 4, and present the numerical experiments in Section 5.

2 Notations and Problem Setup

Let $\mathcal{D} \subset \mathbb{R}^d$, $d \geq 1$ be a compact set. We write $C(\mathcal{D})$ to denote the space of continuous functions on $\mathcal{D}$, which is naturally endowed with the sup-norm $\| \cdot \|_{C(\mathcal{D})}$. We denote by $L^2(\mathcal{D})$ the space of 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, $b^0$, follows a $C(\mathcal{D})$-valued centered Gaussian random field. 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, $k(x, x') = E(b^0(x)b^0(x'))$ and thus an associated reproducing kernel Hilbert space (RKHS) which is obtained as the closure of functions of the form $f(x) = E[(a_1b^0(x_1) + \ldots + a_nb^0(x_n))b^0(x)]$. The closure can be taken relative to the norms $\| \cdot \|_{C(\mathcal{D})}$ and $\| \cdot \|_{L^2(\mathcal{D})}$ - both limiting procedures coincide (see, [31, Lemma 8.1]). By the spectral decomposition of the covariance operator of $b^0$ (i.e., $k(\cdot)$, [16, 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 occurs in $C(\mathcal{D})$ (and also in $L^2(\mathcal{D})$) almost surely. We impose a full-rank assumption on the prior $b^0$ in the following sense.

---

1For theories of Banach space-valued Gaussian random variables, we refer to [16, Chapter 2] and [31].
Assumption 2.1 (Full rank). The closure of the RKHS generated by $b^0$ (as explained above) under the norm $\| \cdot \|_{C(D)}$ is equal to the space $C(D)$. Equivalently, $\kappa_n \neq 0$ for all $n \geq 1$.

We impose Assumption 2.1 so that the support of $b^0$ is not contained in a proper subspace of $C(D)$. As it turns out, this assumption will be important in order to ensure that the worst-case distribution in our analysis is unique (see the proof of Theorems 3.1 and 3.2 below).

As an example, a general class of Gaussian smoothness priors can be constructed via the Laplace operator on $D$. Specifically, we will use the following class of Matérn processes in Sections 4 and 5 which provides natural prior distributions for $\alpha$-regular functions vanishing at $\partial D$.

Example 2.1 (Matérn prior [19, Equation (2)]). Suppose $D$ has a smooth boundary. The prior with a Matérn covariance function with parameter $\kappa \geq 0$ and $\alpha > \frac{d}{2}$ (controlling the smoothness) can be expressed as

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

where the eigenvalues $\lambda_n$ and eigenfunctions $e_n$ correspond to the Dirichlet-Laplacian operator on $D$. See [28, Corollary 5.1.5]. The eigenvalue $\lambda_n$ satisfy Weyl’s law $\lambda_n = \Theta(n^{2/d})$ [29, Corollary 8.3.5] and the eigenfunctions $e_n \in C^\infty(\bar{D})$, $n \geq 1$, where $\bar{D}$ denotes the closure of $D$ and $C^\infty(\bar{D})$ is the space of smooth functions on $\bar{D}$.

We consider perturbations to the nominal prior $b^0$ by borrowing ideas from the field of distributionally robust optimization. We assume the perturbations are supported in a space of continuous functions $H_w$ that is also a RKHS (and in particular a Polish space - this is important when invoking key duality results to study the maximization in our DRO formulation). The important feature with RKHS that we make use of is that the point evaluation functionals are well-defined and continuous with respect to the Hilbert space norm. In particular, we define the space (where we abbreviate the inner product $\langle \cdot, \cdot \rangle_{L^2(D)}$ as $\langle \cdot, \cdot \rangle$)

$$H_w = \left\{ f \in L^2(D) : \sum_{n=1}^{\infty} w_n (f, e_n)^2 < \infty \right\}$$

parametrized by a positive sequence $w = (w_n)_{n \geq 1}$. Typically, the Hilbert norm on $H_w$ is stronger than the usual $L^2(D)$ norm. More precisely, we impose the following assumption on $w$ and the basis $\{e_n\}_{n=1}^{\infty}$.

Assumption 2.2 (RKHS). Assume that $\lim_{n \to \infty} w_n = \infty$ and $H_w$ is endowed with the inner product

$$\langle f, \tilde{f} \rangle_{H_w} = \sum_{n=1}^{\infty} w_n (f, e_n) \langle \tilde{f}, e_n \rangle \quad \forall f, \tilde{f} \in H_w.$$
We assume that the space $\mathcal{H}_w$ is a RKHS with the Hilbert space norm $\| \cdot \|_{\mathcal{H}_w}$ in $C(\mathcal{D})$, identifying $f \in \mathcal{H}_w$ with its continuous version in $L^2(\mathcal{D})$.

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 be used to define our adversarial perturbations. The next example, which is a continuation of Example 2.1, provides intuition about the interpretation of $w$ in terms of roughness.

Example 2.2. Let the eigenvalues $\lambda_n$ and eigenfunctions $e_n$ correspond to the Dirichlet-Laplacian operator on $\mathcal{D}$. Consider $w_n = \Theta(\lambda^n \beta_n)$, then $\beta$ 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 $\mathcal{D}$. Thus for any $\beta > \frac{d}{2}$, by the Sobolev embedding theorem [28, 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 $C(\mathcal{D})$.

We assume the ability to sample noisy observations of $u$ at $m$ design points in $\mathcal{D}$. In particular, let $x_i \in \mathcal{D}, i = 1, \ldots, m$ be the design points, and, also under $P_0$, let $\epsilon^0 = (\epsilon^0_1, \ldots, \epsilon^0_m)$ be a vector of independent $\mathcal{N}(0, \sigma^2)$ errors, we observe the data

$$Y_i = T(b^0)(x_i) + \epsilon^0_i = u^0(x_i) + \epsilon^0_i, \quad i = 1, \ldots, m,$$

where $T(b^0)(x_i) = u^0(x_i), i = 1, \ldots, m$ are point evaluations of a single sample path. We further assume that both $b^0$ and $\epsilon^0$ are independent under $P_0$. Consequently, under $P_0$, the pair $(b^0, \epsilon^0)$ constitutes a Gaussian random variable on the product space $C(\mathcal{D}) \times \mathbb{R}^m$ (which is Banach with the product norm). We make the following assumption on the forward map $T$.

**Assumption 2.3 (Operator).** We assume the following.

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

(ii) There exists a positive sequence $\tilde{w} = \{\tilde{w}_n\}_{n=1}^{\infty}$ and a corresponding space $\mathcal{H}_{\tilde{w}}$ as in (2), with its Hilbert space norm $\| \cdot \|_{\mathcal{H}_{\tilde{w}}}$, which constitutes a RKHS in $C(\mathcal{D})$ such that

$$\|T(f)\|_{\mathcal{H}_{\tilde{w}}} \leq C_{\tilde{w}} \|f\|_{\mathcal{H}_{\tilde{w}}}, \quad \forall f \in \mathcal{H}_{\tilde{w}}.$$

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

\footnote{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 $c_0 \in (0, \infty)$.}
We will see examples in Section 4 where Assumption 2.3 is satisfied. Assumption 2.3 entails certain compatibility of the forward map 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 \[17,2\]. We note that other assumptions in the literature exist, e.g., the “band-limited” assumption in \[23\]. In our framework, we will consider Wasserstein perturbations to the prior, which encompasses a much richer collection of prior family than typically considered in the literature.

2.1 The Nominal Estimation Problem

Let \( M \) denote the space of measurable maps from \( \mathbb{R}^m \) (data space) to \( C(\mathcal{D}) \) (parameter space) and let \( P \) 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

\[
\min_{\phi \in M} E_{P_0} \left[ \|u^0 - \phi(Y_1, \ldots, Y_m)\|_{L^2(\mathcal{D})}^2 \right],
\]

where \( \phi(Y_1, \ldots, Y_m) \) is the predictor for \( u^0 \) 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 conditional mean, i.e.,

\[
\phi_0(Y_1, \ldots, Y_m)(x) = E_{P_0}[u^0(x)|Y_1, \ldots, Y_m],
\]

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

\[
E_{P_0}[u^0(x)|Y_1, \ldots, Y_m] = (k(x, x_1), \ldots, k(x, x_m)) \cdot (K)^{-1} \cdot (Y_1, \ldots, Y_m)^\top,
\]

where \( k(x, x_i) = E_{P_0}[u^0(x)Y_i] \), and \( K = (E_{P_0}[Y_iY_j])_{ij} \in S_{++}^m \), where we denote \( S_{++}^m \) as the set of strictly positive-definite matrices.

2.2 The Distributionally Robust Optimization Formulation

Instead of considering the nominal measure \( P_0 \) on the prior and noise distributions, we postulate a min-max game where an adversarial nature chooses a measure \( P \) in opposition of the decision-maker’s decision. In particular, we have the observation system

\[
\begin{cases}
u = T(b), \\ (b, \epsilon_1, \ldots, \epsilon_m) \sim P, \\ Y_1 = u(x_1) + \epsilon_1, \ldots, Y_m = u(x_m) + \epsilon_m,
\end{cases}
\]

(3)
where \(u(x_1), \ldots, u(x_m)\) are point evaluations of a single sample path. Instead of the nominal Bayes risk, we consider a ‘worst-case Bayes risk’ with respect to all possible mis specification on both \(b\) and \(\varepsilon\), i.e., the whole data-generating process. Our goal is to minimize the worst-case Bayes risk by solving

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

where nature’s admissible choice of \(P\) is constrained by the Wasserstein distance \(W(P, P_0)\) relative to the nominal measure \(P_0\). The Wasserstein distance \(W\) is defined through optimal transport as follows.

**Definition 2.1 (Optimal transport cost).** The optimal transport cost between two probability measures on \(C(D) \times \mathbb{R}^m\) is defined as

\[
D_c(P, P_0) = \inf_{\pi} \{ E_{\pi} [c((b, \varepsilon), (b^0, \varepsilon^0))] : \pi_{(b, \varepsilon)} = P, \pi_{(b^0, \varepsilon^0)} = P_0 \},
\]

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

The existence of an optimal coupling (i.e., an optimal solution to problem (5)) is guaranteed whenever \(c\) is non-negative and lower semi-continuous with respect to the product norm on the Polish space \((C(D) \times \mathbb{R}^m)^2\), see e.g. [33, Theorem 4.1]. In this paper, we use the ground cost function \(c\) defined as

\[
c((b, \varepsilon), (b^0, \varepsilon^0)) = \|\varepsilon - \varepsilon^0\|^2_2 + \|b - b^0\|^2_{H_w} = \sum_{i=1}^m (\varepsilon_i - \varepsilon^0_i)^2 + \sum_{n \geq 1} \left( \langle b, e_n \rangle - \langle b^0, e_n \rangle \right)^2 w_n.
\]

To see that \(c\) is lower semi-continuous, note that if \(\|b_k - b_\infty\|_{C(D)} \to 0\) and \(\|\tilde{b}_k - \tilde{b}_\infty\|_{C(D)} \to 0\) as \(k \to \infty\), then \(\lim \inf_{k \to \infty} \sum_{n \geq 1} \left( \langle b_k, e_n \rangle - \langle \tilde{b}_k, e_n \rangle \right)^2 w_n \geq \sum_{n \geq 1} \left( \langle b_\infty, e_n \rangle - \langle \tilde{b}_\infty, e_n \rangle \right)^2 w_n\) for each \(n\) since \(e_n \in C(D)\), thus

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

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 \(W = \sqrt{D_c}\) is a Wasserstein-type distance on its domain of finiteness \(\{P \in \mathcal{P} : D_c(P, P_0) < \infty\}\) [33, Definition 6.1].

We use the Wasserstein distance \(W\) as 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 $u$). It is important to stress that $W$ depends on the specification of the function class $\mathcal{H}_w$, or equivalently on the Hilbert space norm $\| \cdot \|_{\mathcal{H}_w}$. 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 $\infty$ slowly, then the adversary is under-penalized for moving mass corresponding to the higher “modes”, resulting in rougher sample paths of $b$. The modeling of the behavior of the adversary 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 2.4 (Operator and adversarial cost).** Suppose that we can select $\tilde{w}$ in Assumption 2.3 such that $\tilde{w}_n = o(w_n)$ as $n \to \infty$.

Intuitively, Assumption 2.4 simply says that the operator is bounded even if the adversarial perturbations are made to be 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 in the standard Gaussian regression case, namely, when $T$ equals the identity map.

**Remark 2.1.** Having introduced the setup of our framework, a few comments are in order.

(i) It is natural to consider random elements on a general Banach space $B$ other than $C(D)$, e.g., by embedding the Banach space $B$ in its second dual $B^{**}$ and identify a Borel measurable random element $b$ in $B$ with the stochastic process $(b^*(b) : b^* \in B^*)$, but at the expense of technicality, see a discussion in [31, Section 2.3]. We choose $C(D)$ to mainly illustrate our conceptual contribution of a distributionally robust formulation of nonparametric regression and inverse problems.

(ii) The $L^2(D)$ norm in the objective of the formulation (4) can potentially 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 [28, Section 4.1], and we leave the extension to a future work.

(iii) It is tempting to replace the Hilbert space norm with the $L^2(D)$ norm in the definition of the ground cost function $c$. However, point evaluations are not continuous under the $L^2(D)$ norm. Our proofs for the main results rely crucially on the continuity of point evaluations, 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 regression problem (4).

**Theorem 3.1 (Strong duality for the regression problem).** Suppose that Assumptions 2.1-2.4 hold. For any $\delta > 0$, the strong duality holds:

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

(6)

The idea of the proof to Theorem 3.1 is to first show a strong duality for a sequence of finite-dimensional approximations. In particular, define $\text{span} \left\{ e_n : 1 \leq n \leq N \right\}$ as the (closed) linear subspace of $C(D)$ spanned by the basis vectors $(e_n : 1 \leq n \leq N)$. We consider truncating $P$ (resp. $P_0$) into the space $\text{span} \left\{ e_n : 1 \leq n \leq N \right\} \times \mathbb{R}^m$, and denote the induced measure as $Q^{(N)}$ (resp. $Q_0^{(N)}$). The truncation is through the coordinate projections after expanding functions in the $L^2(D)$ basis $(e_n : \infty \leq n = 1)$. Since the coordinate projections are bounded linear mappings, $Q_0^{(N)}$ is centered Gaussian. Notice that the space $\text{span} \left\{ e_n : 1 \leq n \leq N \right\} \times \mathbb{R}^m$ is isomorphic to $\mathbb{R}^{N+m}$, thus we view the truncated measures $Q^{(N)}$ and $Q_0^{(N)}$ as finite-dimensional measures on $\mathbb{R}^{N+m}$.

For convenience, denote

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

and abusing notation slightly we use $Obj(\phi, Q^{(N)})$ for the truncated measure. In the proof we construct the finite-dimensional approximations and the related strong duality reads

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

(7)

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

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

where $\pi_r$ and $\pi_s$ are projections onto the first and second component of the coupling $\pi$. In the definition of $W_N^2$, $c_N$ is the induced cost function on $\mathbb{R}^{N+m}$ with

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

for any $r, s \in \mathbb{R}^{N+m}$.
We note that in (7), the minimizer in the $\phi$ variable and the maximizer in the $Q^{(N)}$ variable exist, which justifies the minimization and the maximization operators. Denote by $\phi^*_N$ (resp. $Q^*_N$) the (unique) solution to the outer optimization problem in the left (resp. right) hand side of (7). Then $(\phi^*_N, Q^*_N)$ is the (unique) pair of Nash equilibrium for problem (7) in the sense that

$$\text{Obj}(\phi_N^*, Q_N^*) = \min_{\phi \in \mathcal{M}} \text{Obj}(\phi, Q_N^*) = \max_{Q^{(N)}: W_N(Q^{(N)}, Q_0^{(N)}) \leq \delta} \text{Obj}(\phi_N^*, Q^{(N)}).$$

The structural properties of the optimal solutions reveal that $Q^*_N$ is a centered Gaussian distribution, and $\phi^*_N$ is a linear prediction rule, namely, for any $x \in \mathcal{D}$

$$\phi^*_N(Y_1, \ldots, Y_m) = \mathbb{E}_{Q^*_N}[u(x)|Y_1, \ldots, Y_m] = \left(k^{(N)}_e(x, x_1), \ldots, k^{(N)}_e(x, x_m)\right) \cdot (K^{(N)}_e)^{-1} \cdot (Y_1, \ldots, Y_m)^\top,$$

where $k_e^{(N)}(x, x_i) = \mathbb{E}_{Q^*_N}[u(x)|Y_i]$, and $K^{(N)}_e = \left(\mathbb{E}_{Q^*_N}[Y_i Y_j]\right)_{ij} \in \mathbb{S}^{m}$ is invertible. Similar finite-dimensional duality results have been established in [22, 24], but with a slight difference in the definition of the Wasserstein distance and the ambiguity set. The rest of the proof then argues that the error of approximations are negligible as the dimension grows to infinity. In particular, one intermediate result we rely on using in Section 5 is the following

**Proposition 3.1** (Approximation of objective values). Let $\phi^*_N$ be the (unique) solution to the min-max problem in (7), and $Q^*_N$ be the (unique) solution to the max-min problem in (7). We have

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

asymptotically as the number of finite-dimensional modes $N \to \infty$.

### 3.2 Existence, Uniqueness and Construction of the Nash Equilibrium

In this part we show that (6) admits a unique pair of Nash equilibrium under certain conditions. Recall the Nash equilibrium $(\phi^*_N, Q^*_N)$ corresponding to the finite dimensional approximation as recalled in the last section. We now add the tail of $P_0$ to $Q_N^*$, namely, we denote $P_N^* \in \mathcal{P}$ as the measure such that $(\{b_n\}_{n=1}^N, \epsilon)$ and $(\{b_n, e_n\}_{n>N} = d \setminus \{b_0, e_0\}_{n>N}$ for $(b, \epsilon) \sim P_N^*$ and $(b_0, e_0) \sim P_0$, where $d$ means “equality in distribution”. We can extract a subsequence from $P_N^*, N \geq 1$ by a compactness argument.

**Proposition 3.2** (Compactness of the ambiguity set). For every sequence $P_N \in \mathcal{P}, N \geq 1$ that satisfies $W(P_N, P_0) \leq \delta$, there is a weakly convergent subsequence $P_{N_l}, l \geq 1$ with $P_{N_l} \to P_{\infty}$, such that the limit $P_{\infty} \in \mathcal{P}$ also satisfies $W(P_{\infty}, P_0) \leq \delta$. 

12
By Proposition 3.2, we find a weakly convergent subsequence of $P_N^*$, denoted as $P_{N_l}^*$, $l \geq 1$, with a limit $P_\infty^*$ that is feasible. The sequence $P_{N_l}^*$ is centered Gaussian, thus the limit $P_\infty^*$ is also centered Gaussian. To see this, consider any bounded linear functional $F : C(D) \times \mathbb{R}^m \to \mathbb{R}$, and construct the Skorohod representations $Z_{N_l} \sim P_{N_l}^*$ and $Z_\infty \sim P_\infty^*$, where $Z_{N_l} \to Z_\infty$ almost surely. It then follows that $F(Z_{N_l}) \to F(Z_\infty)$ almost surely, and we note it is an elementary fact that the limit of centered univariate Gaussians must be centered univariate Gaussian. 

Denote $K_\varepsilon = (\mathbb{E}_{P_N^*}[Y_iY_j])_{ij}$. Under the condition that $K_\varepsilon$ is invertible, the solution $\phi_\infty^*$ to $\min_\phi \text{Obj}(\phi, P_\infty^*)$ is well-defined

$$\phi_\infty^*(Y_1, \ldots, Y_m)(x) = \mathbb{E}_{P_\infty^*}[u(x)|Y_1, \ldots, Y_m]$$

$$= (k(x,x_1), \ldots, k(x,x_m)) \cdot (K_\varepsilon)^{-1} \cdot (Y_1, \ldots, Y_m)^\top,$$

where $k(x,x_i) = \mathbb{E}_{P_N^*}[u(x)Y_i]$. The main result of this section is the following theorem.

**Theorem 3.2** (Nash equilibrium). Under the condition that $K_\varepsilon$ is invertible for the limit $P_\infty^*$ of some weakly convergent subsequence of $P_N^*$, we have that $(\phi_\infty^*, P_\infty^*)$ consists of a pair of Nash equilibrium to problem \( \mathcal{P} \), i.e.,

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

Therefore, $(\phi_\infty^*, P_\infty^*)$ represents a pair of equilibrium strategies where no player (i.e., the decision-maker or the adversary) has anything to gain by changing only their own strategy. Moreover, the pair $(\phi_\infty^*, P_\infty^*)$ is unique, $\lim_{N \to \infty} \phi_N^*(Y_1, \ldots, Y_m)(x) = \phi_\infty^*(Y_1, \ldots, Y_m)(x)$ for any data $Y_1, \ldots, Y_m$ and any $x \in D$, and $P_N^* \Rightarrow P_\infty^*$ weakly 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. 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 $K_\varepsilon$ is invertible for the limit of some weakly convergent subsequence of $P_N^*$. Fortunately, due to the following simple proposition, in practice it is not hard to check whether this condition holds.

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

1. the matrix $K_\varepsilon$ is invertible for the limit $P_\infty^*$ of some weakly convergent subsequence of $P_N^*$, or

2. we have $\det(K_\varepsilon^{(N)}) \to 0$ 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 3.3.** There exists a constant $\delta_0 = \delta_0(T, m, (x_i)_i, \mathcal{H}_w, \mathcal{H}_{\tilde{w}}, \sigma^2) > 0$, such that if $\delta < \delta_0$, then for all $P$ such that $\mathcal{W}(P,P_0) \leq \delta$, the matrix $(\mathbb{E}_P[Y_iY_j])_{ij}$ is invertible.
3.3 Strong duality for the inverse problem

Alternatively, we propose a distributionally robust formulation for the inverse problem, where our primary interest lies in recovering the unknown input $b$. Under the observation system (3), the goal of the decision-maker is to seek for a nonparametric predictor $\phi_b \in \mathcal{M}$ that minimizes the worst-case objective

$$\inf_{\phi_b \in \mathcal{M}} \sup_{P \in \mathcal{P}, W(P, P_0) \leq \delta} E_P \left[ \|b - \phi_b(Y_1, \ldots, Y_m)\|_{L^2(D)}^2 \right], \quad (8)$$

where nature’s admissible choice of $P$ is constrained by the Wasserstein distance $W(P, P_0)$ constructed in (5). We state the strong duality of (8).

**Theorem 3.4** (Strong duality for the inverse problem). Suppose that Assumptions 2.1-2.4 hold. For any $\delta > 0$,

$$\inf_{\phi_b \in \mathcal{M}} \sup_{P \in \mathcal{P}, W(P, P_0) \leq \delta} E_P \left[ \|b - \phi_b(Y_1, \ldots, Y_m)\|_{L^2(D)}^2 \right] = \sup_{P \in \mathcal{P}, W(P, P_0) \leq \delta} \inf_{\phi_b \in \mathcal{M}} E_P \left[ \|b - \phi_b(Y_1, \ldots, Y_m)\|_{L^2(D)}^2 \right]. \quad (9)$$

The proof of Theorem 3.4 works verbatim as the proof of Theorem 3.1. Though strong duality holds for both the regression and the inverse problems under our formulation, 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 $u$ [11, 7, 8, 18].

As to the Nash equilibrium associated with (9), 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 to Matérn process priors and the space of perturbation given by Examples 2.1 and 2.2. 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 the Gaussian process regression. Assumptions 2.3-2.4 are satisfied for $\tilde{w}_n = \lambda_n^\beta$ and any $\frac{d}{2} < \beta < \beta$. If $D$ 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

\[ \begin{align*}
\Delta u(x) &= b(x) \quad \forall x \in D, \\
u(x) &= 0 \quad \forall x \in \partial D.
\end{align*} \]

We have that the forward map \( T \) is the inverse-Laplacian operator, thus

\[ T(f) = \sum_{n \geq 1} -\lambda_n^{-1}(f, e_n)e_n. \]

It is straightforward to see that Assumptions \[2.3\] \[2.4\] are satisfied for \( \tilde{w}_n = \lambda_n^\cdot \) and any \( d \frac{1}{2} < \beta < \beta. \)

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

\[ \begin{align*}
u_t &= \nu_{xx} \quad 0 < x < 1, \\
u(x, 0) &= b(x) \quad 0 < x < 1, \\
u(0, t) &= \nu(1, t) = 0 \quad t \geq 0,
\end{align*} \]

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

\[ \nu(x, t) = \sum_{n \geq 1} e^{-n^2 \pi^2 t} (b, e_n)e_n. \]

Thus we have the (time-dependent) forward map \( T \) satisfies, for any \( t \geq 0 \),

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

Assumptions \[2.3\] \[2.4\] are satisfied for \( \tilde{w}_n = \lambda_n^\cdot \) and any \( d \frac{1}{2} < \beta < \beta. \)

Example 4.4 (Radon transform in the plane). The Radon transform of a function \( f \) is the function

\[ T(f)(s, \omega) = \int_{-\infty}^{\infty} f(s\omega + t\omega^\perp) dt, \quad s \in \mathbb{R}, \omega \in S^1, \]

where \( S^1 \) is the unit circle, and \( \omega^\perp \) is the vector in \( S^1 \) obtained by rotating \( \omega \) counterclockwise by 90°. Recall we consider \( f \) to be supported in a compact domain \( D \subset \mathbb{R}^2 \), and thus \( T(f) \) vanishes outside a compact subset of \( \mathbb{R} \times S^1 \). It is straightforward to see that \( T : C(D) \rightarrow C(\mathbb{R} \times S^1) \) 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 \[21\] Theorems II.5.1 and II.5.2, the Radon transform has the Sobolev estimate:

\[ \|T(f)\|_{H^\beta(\mathbb{R} \times S^1)} \leq C_{\tilde{w}} \|f\|_{\mathcal{H}_w} \quad \forall f \in \mathcal{H}_w. \]
for \( \tilde{w}_n = \lambda_n^{\tilde{\beta}} \) and any \( \frac{d}{2} < \tilde{\beta} < \beta \), where \( H^{\tilde{\beta}}(\mathbb{R} \times S^1) \) is the usual order-\( \tilde{\beta} \) Sobolev space. Identifying \( S^1 \) with \([0, 2\pi)\), and by the Sobolev embedding theorem [28, Proposition 4.1.3], we see that point evaluations in \( H^{\tilde{\beta}}(\mathbb{R} \times S^1) \) is continuous. Our theory in Section 3 applies with this slight modification of Assumptions 2.3-2.4 after inspecting the proofs.

5 Numerical Experiments

Our focus in this paper is the framework and theoretical properties for a min-max formulation of regression and inverse problems in an infinite-dimensional setting. To gain further insights into our min-max formulation, we now present a series of experiments to explore (both qualitatively and quantitatively) properties of the Nash equilibrium. In particular, we compute the Nash equilibrium of the finite-dimensional approximation \((\phi_N^*, Q_N^*)\) (see definitions in Section 3) with \( N = 200 \), by an adaptation of the Frank-Wolfe algorithm in [24]. Under the condition 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 value of the game is (asymptotically) consistent. Our focus is the one-dimensional Gaussian process regression as described in Example 4.1. Throughout this section we set \( \kappa_n = \lambda_n^{-\alpha/2} \) and \( w_n = \lambda_n^{\beta} \).

First we fix a set of baseline parameters where \( \alpha = 2, \beta = 0.51, \delta^2 = 0.1 \) and \( \sigma = 0.1 \). We choose 10 designs equi-spaced on either \((0, 1)\) or \((0, 0.5)\) interval, excluding the end points. Since both the nominal and the worst-case measures are Gaussian, we have that the posterior covariances do not depend on the realizations of the data. We first visualize the correlation functions on \([0,1]^2\) as in Figures 1 and 2. Comparing to the nominal measures, we observe that in the worst-case measures there are ripples corresponding to reductions of correlations between the sampled locations. We next plot the 95% confidence bands of the nominal and the worst-case sample paths from 500 Monte Carlo simulations in Figures 3 and 4, where we sample the data according to the prior measure in plotting the posterior sample paths. In particular, we use the vector of observations

\[
(-0.17, -0.09, 0.02, 0.04, 0.12, 0.05, -0.03, 0.03, -0.28, -0.15)
\]

for 10 designs equi-spaced in \((0,1)\) and

\[
(0.03, -0.05, 0.08, -0.08, 0.15, 0.12, -0.25, -0.24, 0.16, 0.02)
\]

for 10 designs equi-spaced in \((0,0.5)\). Comparing to the nominal prior measures, we observe that the worst-case prior measures have roughly the same overall variances, and a sharper contrast between the observed and unobserved locations, especially in the first case of designs. More impressively, comparing to the nominal posterior measures, we observe that in the worst-case posterior measures there are significant increase of variance in regions away from the sampled locations, while the variance increase in regions surrounding the sampled locations is only moderate. Hence the worst-cases are perturbed so as to induce
greater uncertainties in regions where information is limited, which intuitively guarantees
greater robustness of the predictions.

Figure 1: Correlation functions on $[0,1]^2$ with 10 designs equi-spaced in $(0,1)$.

We next vary the baseline parameters to see the effect of the worst-case perturbations
on the (prior and posterior) sample paths comparing to the nominal measures. We focus
on the first case of designs. In each parameter setting we sample the paths for 5 times
and gauge their qualitative behavior, while we compute the Förstner metric [12] of the
(posterior covariance, prior covariance) pencils in the Laplacian eigenspace to quantify the
contrast between posterior and prior covariances, induced by the nominal and worst-case
measures respectively. To plot the posterior sample paths, we use the same vector of
observations as before. In particular, with the remaining parameters fixed to be the base
case, we

1. choose $\alpha \in \{0.51, 2, 4\}$, where we summarize the results in Figures 5-7 (note that we
Figure 2: Correlation functions on $[0, 1]^2$ with 10 designs equi-spaced in $(0, 0.5)$.

include the baseline $\alpha = 2$ for comparison);

2. choose $\beta \in \{0.7, 1\}$, where we summarize the results in Figure 8 (note that the nominal prior and posterior are the same as the baseline setting);

3. choose $\delta^2 \in \{0.01, 1\}$, where we summarize the results in Figure 9 (note that the nominal prior and posterior are the same as the baseline setting);

4. choose $\sigma \in \{0.01, 1\}$, where we summarize the results in Figures 10 and 11.

Notably in the choice of parameters we do not check the condition put forward by Lemma 3.3 as we believe that the condition is sufficient but not necessary. We summarize results on the Förstner metric in Table 1. Combining the qualitative and quantitative results, we observe that in cases where there is: a larger $\alpha$ (i.e. a smoother prior); a smaller $\beta$ (i.e. less
penalty on higher modes that induce roughness); a larger $\delta$ (i.e. more admissible perturbations); or a smaller $\sigma$ (i.e. smaller noise), we can more easily see the behavior of the worst-case that induces sharper contrasts between the observed and unobserved locations in both prior and posterior sample paths.

![Figure 3: 95% confidence band of sample paths with 10 designs equi-spaced in (0, 1).](image)

| | 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: Förstner metric [12] of the (posterior covariance, prior covariance) pencils in the Laplacian eigenspace with 10 designs equi-spaced in (0, 1).
Figure 4: 95% confidence band of sample paths with 10 designs equi-spaced in (0, 0.5).

Acknowledgements: Blanchet gratefully acknowledges support from the NSF via grant DMS-EPSRC 2118199 as well as NSF-DMS 1915967, and from the AFOSR via grant MURI FA9550-20-1-0397.
(a) Nominal Prior  (b) Worst-case Prior  (c) Nominal Posterior  (d) Worst-case Posterior

Figure 5: Sample paths with $\alpha = 0.51$ and 10 designs equi-spaced in $(0, 1)$.

(a) Nominal Prior  (b) Worst-case Prior  (c) Nominal Posterior  (d) Worst-case Posterior

Figure 6: Sample paths with $\alpha = 2$ and 10 designs equi-spaced in $(0, 1)$.

References

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

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

[3] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder, Adapted Wasserstein distances and stability in mathematical finance, Finance and Stochastics, 24 (2020), pp. 601–632.

[4] J. Bion-Nadal and D. Talay, On a Wasserstein-type distance between solutions to stochastic differential equations, The Annals of Applied Probability, 29 (2019), pp. 1609 – 1639.

[5] J. Blanchet, Y. Kang, and K. Murthy, Robust wasserstein profile inference and applications to machine learning, Journal of Applied Probability, 56 (2019), p. 830–857.

[6] J. Blanchet and K. Murthy, Quantifying distributional model risk via optimal transport, Mathematics of Operations Research, 44 (2019), pp. 565–600.
Figure 7: Sample paths with $\alpha = 4$ and 10 designs equi-spaced in $(0, 1)$.

Figure 8: Worst-case sample paths with varying $\beta$ and 10 designs equi-spaced in $(0, 1)$.

[7] L. Cavalier, *Inverse problems in statistics*, in Inverse problems and high-dimensional estimation, vol. 203 of Lect. Notes Stat. Proc., Springer, Heidelberg, 2011, pp. 3–96.

[8] L. Cavalier and A. Tsybakov, *Sharp adaptation for inverse problems with random noise*, Probability Theory and Related Fields, 123 (2002), pp. 323–354.

[9] M. Dashti and A. M. Stuart, *The Bayesian Approach to Inverse Problems*, Springer International Publishing, 2017, pp. 311–428.

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

[11] D. L. Donoho, *Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition*, Appl. Comput. Harmon. Anal., 2 (1995), pp. 101–126.

[12] W. Förstner and B. Moonen, *A Metric for Covariance Matrices*, Springer Berlin Heidelberg, 2003, pp. 299–309.

[13] R. Gao, X. Chen, and A. J. Kleywegt, *Wasserstein distributional robustness and regularization in statistical learning*, arXiv preprint arXiv:1712.06050, (2017).

[14] R. Gao and A. J. Kleywegt, *Distributionally robust stochastic optimization with Wasserstein distance*, arXiv preprint arXiv:1604.02199, (2016).
Figure 9: Worst-case sample paths with varying $\delta$ and 10 designs equi-spaced in $(0, 1)$.

Figure 10: Sample paths with $\sigma = 0.01$ and 10 designs equi-spaced in $(0, 1)$.

[15] S. Ghosal and A. Roy, *Posterior consistency of Gaussian process prior for non-parametric binary regression*, The Annals of Statistics, 34 (2006), pp. 2413–2429.

[16] E. Giné and R. Nickl, *Mathematical Foundations of Infinite-Dimensional Statistical Models*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2015.

[17] S. Gugushvili, A. van der Vaart, and D. Yan, *Bayesian linear inverse problems in regularity scales*, Annales de l’Institut Henri Poincaré, Probabilités et Statistiques, 56 (2020), pp. 2081 – 2107.

[18] B. Knapik, A. W. van der Vaart, and J. H. van Zanten, *Bayesian inverse problems with Gaussian priors*, Ann. Statist., 39 (2011), pp. 2626–2657.

[19] F. Lindgren, H. Rue, and J. Lindström, *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 (2011), pp. 423–498.

[20] P. Mohajerin Esfahani and D. Kuhn, *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations*, Mathematical Programming, 171 (2018), pp. 115–166.
Figure 11: Sample paths with $\sigma = 1$ and 10 designs equi-spaced in $(0, 1)$.

[21] F. Natterer, *The Mathematics of Computerized Tomography*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 2001.

[22] V. A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani, *Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization*, Mathematics of Operations Research, (forthcoming).

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

[24] S. Shafieezadeh-Abadeh, V. Nguyen, D. Kuhn, and P. Mohajerin Esfahani, *Wasserstein distributionally robust Kalman filtering*, in Advances in Neural Information Processing Systems 31, 2018, pp. 8483–8492.

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

[26] A. V. Skorokhod, *Limit theorems for stochastic processes*, Theory of Probability & Its Applications, 1 (1956), pp. 261–290.

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

[28] M. Taylor, *Partial Differential Equations I: Basic Theory*, Applied mathematical sciences, Springer, 1996.

[29] ———, *Partial Differential Equations: Qualitative studies of linear equations. II*, Applied mathematical sciences, Springer New York, 1996.

[30] A. W. van der Vaart and J. H. van Zanten, *Rates of contraction of posterior distributions based on Gaussian process priors*, The Annals of Statistics, 36 (2008), pp. 1435–1463.
[31] A. W. van der Vaart and J. H. van Zanten, *Reproducing kernel Hilbert spaces of Gaussian priors*, Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh, (2008), p. 200–222.

[32] A. W. van der Vaart and J. H. van Zanten, *Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth*, The Annals of Statistics, 37 (2009), pp. 2655 – 2675.

[33] C. Villani, *Optimal Transport: Old and New*, Springer Science & Business Media, 2008.
6 Appendix

6.1 Proof to Theorem 3.1

Proof of the strong duality. We start with a few definitions used in the proof.

**Definition 6.1 (Affine predictor).** For \( \phi(Y_1, \ldots, Y_m) \) a mapping from \( \mathbb{R}^m \) (data space) to \( C(\mathcal{D}) \) (parameter space), we define that \( \phi \) is affine if it can be written as

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

where \( \alpha_j \in C(\mathcal{D}) \) \( \forall 0 \leq j \leq m \).

**Definition 6.2 (Finite-dimensional truncation).** For the orthonormal system \( \{e_n\}_{n=1}^{\infty} \) introduced in Assumption 2.1, we denote \( \text{span}\{e_n\}_{n=1}^{N} \) as the (closed) linear subspace of \( C(\mathcal{D}) \) spanned by \( \{e_n\}_{1 \leq n \leq N} \). Similarly we define \( \text{span}\{T(e_n)\}_{n=1}^{N} \).

For affine \( \phi \), we say \( \text{coef}(\phi) \in \text{span}\{T(e_n)\}_{n=1}^{N} \) if

\[
\alpha_j \in \text{span}\{T(e_n)\}_{n=1}^{N} \forall j.
\]

For convenience, denote the shorthand

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

Note that for all \( N \geq 1 \), we have the chain of inequalities

\[
\inf_{\phi, \text{affine}, \text{coef}(\phi) \in \text{span}\{T(e_n)\}_{n=1}^{N}} \sup_{P: \mathbb{W}(P, P_0) \leq \delta} \text{Obj}(\phi, P) \geq \inf_{\phi, \text{affine}} \sup_{P: \mathbb{W}(P, P_0) \leq \delta} \text{Obj}(\phi, P) \geq \inf_{\phi} \sup_{P: \mathbb{W}(P, P_0) \leq \delta} \text{Obj}(\phi, P) \geq \sup_{P: \mathbb{W}(P, P_0) \leq \delta} \inf_{\phi} \text{Obj}(\phi, P) \geq \sup_{P: \mathbb{W}(P, P_0) \leq \delta, (b, \epsilon_n) \overset{d}{=} (b', \epsilon_n) \forall n > N} \phi \text{Obj}(\phi, P).
\]

where \( \overset{d}{=} \) means equality in distribution. We define additionally

**Definition 6.3 (Truncated measure).** For any \( N \geq 1 \), denote by \( P^{(N)} \) the measure induced by truncating \( P \in \mathcal{P} \) into the space \( \text{span}\{e_n\}_{n=1}^{N} \times \mathbb{R}^m \). Namely, we have \( (b, \epsilon_n) = 0 \) a.s. for \( (b, \epsilon) \sim P^{(N)} \) and \( n > N \).
Our proof consists of showing the following three claims:

**Claim 1**: The finite-dimensional version of the strong duality holds, i.e.,
\[
\inf_{\phi} \sup_{P: W(P,P_0) \leq \delta, \langle b,e_n \rangle d = \langle b_0,e_n \rangle} \text{Obj}(\phi, P^{(N)}) \leq \delta, \langle b,e_n \rangle d = \langle b_0,e_n \rangle \forall n > N
\]

\[
\text{(12)}
\]

**Claim 2**: The truncation of \( P \) to \( P^{(N)} \) in the last term of (11) preserves the chain of inequalities, i.e.,
\[
\sup_{P: W(P,P_0) \leq \delta, \langle b,e_n \rangle d = \langle b_0,e_n \rangle} \inf_{\phi} \text{Obj}(\phi, P) \geq \sup_{P: W(P,P_0) \leq \delta, \langle b,e_n \rangle d = \langle b_0,e_n \rangle} \inf_{\phi} \text{Obj}(\phi, P^{(N)}).
\]

\[
\text{(13)}
\]

**Claim 3**: The truncation of \( P \) to \( P^{(N)} \) in the first term of (11) has an error asymptotically negligible, i.e.,
\[
\inf_{\phi, \text{affine}, \text{coef}(\phi) \in \text{span}\{T(e_n)\}} \sup_{P: W(P,P_0) \leq \delta} \text{Obj}(\phi, P) - \inf_{\phi, \text{affine}, \text{coef}(\phi) \in \text{span}\{T(e_n)\}} \sup_{P: W(P,P_0) \leq \delta} \text{Obj}(\phi, P^{(N)}) \leq o(1)
\]

\[
\text{(14)}
\]

as \( N \to \infty \). Combining the above three claims and the chain of inequalities (11), we conclude that
\[
\inf_{\phi} \sup_{P: W(P,P_0) \leq \delta} \text{Obj}(\phi, P) = \sup_{P: W(P,P_0) \leq \delta} \inf_{\phi} \text{Obj}(\phi, P)
\]

\[
\text{(15)}
\]

by letting \( N \to \infty \). For ease of notation, we make it implicit that \( \inf_{\phi} \) is understood as minimizing over the space of measurable maps from \( \mathbb{R}^m \) (data space) to \( C(D) \) (parameter space), and we make it implicit that \( P \) always takes value in \( \mathcal{P} \), the space of (Borel) probability measures on \( C(D) \times \mathbb{R}^m \).

We now provide the proof of the three claims.

**Proof of Claim 1.** We first define

**Definition 6.4** (Marginal measure). Denote the marginal measure, under \( P \in \mathcal{P} \), of the coordinates
\[
(\langle b,e_1 \rangle, \cdots, \langle b,e_N \rangle, \epsilon)
\]

on \( \mathbb{R}^{N+m} \) as \( Q^{(N)} \).
Note that the marginal measure is finite-dimensional, while the truncated measure in Definition 6.3 belongs to $\mathcal{P}$. We have that $Q_0^{(N)}$, as the marginal measure of $P_0$ per Definition 6.3, is a multivariate Gaussian. Note that

$$\sup_{P:W(P,P_0)\leq \delta, \pi \in \pi} \mathbb{E}_{Q^{(N)}}[g(R)],$$

where $R$ is a random variable distributed according to $Q^{(N)}$, and $W_N$ is defined as

$$W_N^2(Q^{(N)}, Q_0^{(N)}) = \min_{\pi} \{ \mathbb{E}_\pi [c_N(r, s)]: \pi_r = Q^{(N)}, \pi_s = Q_0^{(N)} \},$$

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

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

If $\phi$ is affine and $\text{coef}(\phi) \in \text{span}\{ T(e_n) \}_{n=1}^N$, then $g(r)$ is a convex (non-constant) quadratic function of $r$.

Suppose first that $\phi$ is affine and $\text{coef}(\phi) \in \text{span}\{ T(e_n) \}_{n=1}^N$. By Theorem 1 in [6], we have

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

and that there exists an optimal dual optimizer $\gamma^*$. Since the distributions in the set

$$\{ Q^{(N)}: W_N(Q^{(N)}, Q_0^{(N)}) \leq \delta \}$$

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

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

Since $\sup_{s \in \mathbb{R}^{N+m}} g(s) = \infty$, we have necessarily that $\gamma^* > 0$. Observe that $g(r)$ can be written in the form $g(r) = r^\top G r + c^\top r + \|\alpha_0\|_{L^2(D)}^2$ for some $G \in \mathbb{S}_+^{N+m}$ and $c \in \mathbb{R}^{N+m}$. Denote

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

Then almost surely for $R$ distributed under $Q_0^{(N)}$, we have

$$g(s) - \gamma^* c_N(R, s) = s^\top (G - \gamma^* W_N) s + (2\gamma^* R^\top W_N + c^\top) s + \|\alpha_0\|_{L^2(D)}^2 + \gamma^* R^\top W_N R.$$
Since $R$ follows a non-degenerate multivariate Gaussian distribution, we have that necessarily
\[
\gamma^* W_N - G \in \mathbb{S}_{++}^{N+m},
\]
otherwise \(\sup_{s \in \mathbb{R}^{N+m}} g(s) - \gamma^* c_N(R, s) = \infty\) almost surely. The solution to the left side of equation (16), if it exist, must concentrates on the graph of the affine push-forward map
\[
s^*(R) = -\frac{1}{2}(G - \gamma^* W_N)^{-1}(2\gamma^* W_N R + c).
\]
(17)
The existence of a solution $Q_N^{(N)}$ to the left hand side of (16) as well as of an optimal coupling between $Q_0^{(N)}$ and $Q_N^{(N)}$, which we denote by $\pi^*$, can be verified by [14, Corollary 1(i)]. Indeed, this follows from $\kappa < \gamma^*$ as $\gamma^* W_N - G \in \mathbb{S}_{++}^{N+m}$, where the growth-rate $\kappa$ defined in [14] is
\[
\kappa = \lim \sup_{r \to \infty} \frac{r^\top Gr}{r^\top W_N r}.
\]
It follows that $Q_N^{(N)}$ is in fact also Gaussian, whence we may write (16) as
\[
\sup_{Q_N^{(N)}: W_N(Q_N^{(N)}, Q_0^{(N)}) \leq \delta} \mathbb{E}_{Q_N^{(N)}}[g(R)] = \sup_{Q_N^{(N)}: W_N(Q_N^{(N)}, Q_0^{(N)}) \leq \delta, Q_N^{(N)} \text{ normal}} \mathbb{E}_{Q_N^{(N)}}[g(R)]
\]
\[
= \sup_{(\mu, \Sigma) \in \mathcal{S}_N} \langle G, \Sigma + \mu \mu^\top \rangle + c^\top \mu + \|\alpha_0\|_{L^2(D)}^2,
\]
where the last inequality is because $c^\top \mu \geq 0$ after a possible sign change in $\mu$. Therefore
\[
\inf_{\phi(\cdot), \text{affine}, \text{coef}(\phi) \in \text{span}\{T(e_n)\}_{n=1}^N} \sup_{Q_N^{(N)}: W_N(Q_N^{(N)}, Q_0^{(N)}) \leq \delta} \mathbb{E}_{Q_N^{(N)}}[g(R)]
\]
\[
= \inf_{\phi(\cdot), \text{affine}, \text{coef}(\phi) \in \text{span}\{T(e_n)\}_{n=1}^N, \alpha_0 = 0} \sup_{(\mu, \Sigma) \in \mathcal{S}_N} \langle G, \Sigma + \mu \mu^\top \rangle
\]
\[
= \inf_{\phi(\cdot), \text{affine}, \text{coef}(\phi) \in \text{span}\{T(e_n)\}_{n=1}^N, \alpha_0 = 0} \langle G, \Sigma \rangle,
\]
where the last equality is because $s^*(R)$ is zero-mean whenever $c = 0$, which is a consequence of $\alpha_0 = 0$. $\mathcal{S}_N$ is a compact and convex set coming from a modified Gelbrich distance
\[
\mathcal{S}_N = \left\{ (\mu, \Sigma) : \|\mu\|^2 + tr(W_N \Sigma) + tr(W_N \Sigma_0) - 2tr \left[ \left( \sqrt{\Sigma_0 W_N \Sigma W_N \sqrt{\Sigma_0}} \right) \right] \leq \delta^2 \right\},
\]
29
where $\Sigma_0 \in S_{++}^{N+m}$ is the covariance matrix of $Q_0^{(N)}$. We show in passing how the modified Gelbrich distance arises. Note that the usual Euclidean cost function is $\|r-s\|_2^2$, which gives rise to the usual Gelbrich distance [24, Proposition 2.2]. Our new cost is $(r-s)\top W_N(r-s)$. Thus the optimal coupling $\pi^*$ for the new cost solves

$$
\min_{\pi(r,s)} \left(-\int r\top W_N s d\pi(r,s)\right).
$$

Using substitution of variables $\tilde{r} = \sqrt{W_N}r$ and $\tilde{s} = \sqrt{W_N}s$, there is a one-to-one correspondence between $\pi(r,s)$ and $\tilde{\pi}(\tilde{r},\tilde{s})$, and $\tilde{\pi}^*$ (corresponding to $\pi^*$) solves

$$
\min_{\tilde{\pi}(\tilde{r},\tilde{s})} \left(-\int \tilde{r}\top \tilde{s} d\tilde{\pi}(\tilde{r},\tilde{s})\right).
$$

By Sion’s minimax theorem [25], we have

$$
\inf_{\phi(\cdot),\text{affine},\text{coef}(\phi)\in \text{span}\{T(e_n)\}_{n=1}^N} \sup_{(G,\Sigma)\in S_N} \langle G, \Sigma \rangle = \sup_{(0,\Sigma)\in S_N} \inf_{\phi(\cdot),\text{affine},\text{coef}(\phi)\in \text{span}\{T(e_n)\}_{n=1}^N} \langle G, \Sigma \rangle = \sup_{Q^{(N)}:W_N(Q^{(N)},Q_0^{(N)})\leq \delta} \inf_{Q^{(N)}:W_N(Q^{(N)},Q_0^{(N)})\leq \delta} \mathbb{E}_{Q^{(N)}}[g(R)].
$$

The last “full-rank” assertion comes from the fact that $s^*(R)$ is a non-degenerate linear transform of $R$. Thus

$$
\inf_{\phi(\cdot),\text{affine},\text{coef}(\phi)\in \text{span}\{T(e_n)\}_{n=1}^N} \sup_{P:W(P,P_0)\leq \delta, (b,\epsilon_n)\in (R,\epsilon_n)\forall n>N} \text{Obj}(\phi, P^{(N)}) = \inf_{\phi(\cdot),\text{affine},\text{coef}(\phi)\in \text{span}\{T(e_n)\}_{n=1}^N} \sup_{Q^{(N)}:W_N(Q^{(N)},Q_0^{(N)})\leq \delta} \mathbb{E}_{Q^{(N)}}[g(R)] = \inf_{\phi(\cdot),\text{affine},\text{coef}(\phi)\in \text{span}\{T(e_n)\}_{n=1}^N, \alpha_0=0} \text{Obj}(\phi, P^{(N)}).
$$

For $W(P,P_0) \leq \delta$, note that $u$ is a process with continuous sample paths, whence we can interchange the integration

$$
\mathbb{E}_P \left[\|u - \phi(Y_1,\ldots,Y_m)\|_{L^2(D)}^2\right] = \mathbb{E}_P \left[\int_D |u(x) - \phi(Y_1,\ldots,Y_m)(x)|^2 dx\right]
\leq \int_D \mathbb{E}_P \left[|u(x) - \phi(Y_1,\ldots,Y_m)(x)|^2\right] dx.
$$

30
Thus the optimal solution to

$$\inf_{\phi \in \mathcal{M}} \mathbb{E}_P \left[ \| u - \phi(Y_1, \ldots, Y_m) \|_{L^2(D)}^2 \right]$$  \hspace{1cm} (21)$$

is given by the conditional mean

$$\phi(Y_1, \ldots, Y_m)(x) = \mathbb{E}_P [u(x)|Y_1, \ldots, Y_m].$$  \hspace{1cm} (22)$$

For any $P$ a centered Gaussian random variable, it is easy to see that for $x \in D$,

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

is jointly Gaussian and

$$\mathbb{E}_P [u(x)|Y_1, \ldots, Y_m] = (k_\epsilon(x, x_1), \ldots, k_\epsilon(x, x_m)) \cdot (K_\epsilon)^{-1} \cdot (Y_1, \ldots, Y_m)^\top$$  \hspace{1cm} (23)$$

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

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

provided the matrix is invertible. Moreover, the optimal objective value of (21) is given by

$$\int_D \mathbb{E}_P [\text{Var}_P[u(x)|Y_1, \ldots, Y_m]] dx$$

$$= \int_D k(x, x) - (k_\epsilon(x, x_1), \ldots, k_\epsilon(x, x_m)) \cdot (K_\epsilon)^{-1} \cdot (k_\epsilon(x, x_1), \ldots, k_\epsilon(x, x_m))^\top dx$$

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

$$\sup_{P:W(P, P_0) \leq \delta, (b, e_n) \sim (b_0, e_n) \forall n, Q^{(N)}} \inf_{\phi(-), \text{affine}, \text{coef}(\phi) \in \text{span}\{T(e_n)\}} \text{Obj} \left( \phi, P^{(N)} \right)$$

$$= \sup_{P:W(P, P_0) \leq \delta, (b, e_n) \sim (b_0, e_n) \forall n, Q^{(N)}} \inf_{\phi(-)} \text{Obj}(\phi, P^{(N)})$$

$$\leq \sup_{P:W(P, P_0) \leq \delta, (b, e_n) \sim (b_0, e_n) \forall n, Q^{(N)}} \inf_{\phi(-)} \text{Obj}(\phi, P^{(N)})$$

On the other hand,

$$\sup_{P:W(P, P_0) \leq \delta, (b, e_n) \sim (b_0, e_n) \forall n, Q^{(N)}} \inf_{\phi(-)} \text{Obj}(\phi, P^{(N)})$$

$$\leq \inf_{\phi(-)} \sup_{P:W(P, P_0) \leq \delta, (b, e_n) \sim (b_0, e_n) \forall n, Q^{(N)}} \text{Obj}(\phi, P^{(N)})$$

$$\leq \inf_{\phi(-), \text{affine}, \text{coef}(\phi) \in \text{span}\{T(e_n)\}} \sup_{P:W(P, P_0) \leq \delta, (b, e_n) \sim (b_0, e_n) \forall n, Q^{(N)}} \text{Obj}(\phi, P^{(N)}).$$
Thus combining the above chains of inequalities and \((20)\), we have

\[
\inf_{\phi(\cdot), \text{affine}, \text{coef}(\phi) \in \text{span}\{T(e_n)\}_{n=1}^{N}} \sup_{P : \mathcal{W}(P,P_0) \leq \delta, \langle b, e_n \rangle \overset{d}{=} \langle b^0, e_n \rangle \forall n > N} \text{Obj}(\phi, P^{(N)})
\]

\[
= \inf_{\phi(\cdot), \text{affine}, \text{coef}(\phi) \in \text{span}\{T(e_n)\}_{n=1}^{N}, \alpha_0 = 0} \sup_{P : \mathcal{W}(P,P_0) \leq \delta, \langle b, e_n \rangle \overset{d}{=} \langle b^0, e_n \rangle \forall n > N} \text{Obj}(\phi, P^{(N)})
\]

\[
= \inf_{\phi((\cdot))} \sup_{P : \mathcal{W}(P,P_0) \leq \delta, \langle b, e_n \rangle \overset{d}{=} \langle b^0, e_n \rangle \forall n > N} \text{Obj}(\phi, P^{(N)})
\]

\[
\inf_{\phi((\cdot))} \sup_{P : \mathcal{W}(P,P_0) \leq \delta, \langle b, e_n \rangle \overset{d}{=} \langle b^0, e_n \rangle \forall n > N} \text{Obj}(\phi, P^{(N)})
\]

\[
\sup_{P : \mathcal{W}(P,P_0) \leq \delta, \langle b, e_n \rangle \overset{d}{=} \langle b^0, e_n \rangle \forall n > N, Q^{(N)} \text{ is centered full-rank normal } \phi((\cdot))
\]

Thus we have established \((12)\). \(\square\)

**Proof of Claim 2.** Denote \((\phi^*_N, P^*_N)\) as the pair of Nash equilibrium of \((12)\). In particular, we can choose \(\phi^*_N\) is affine, \(Q^{(N)}_N\) (as the marginal measure of \(P^*_N\) per Definition 6.4 which we denote by \(Q^{(N)}_N\) for ease of notation in the sequel) is centered full-rank normal, \((\phi^*_N, Q^{(N)}_N)\) is the Nash equilibrium of \((18)\) and \((19)\), and

\[
(\{\langle b, e_n \rangle\}_{n=1}^{N}), \langle b, e_{N+1} \rangle, \ldots
\]

are mutually-independent and \(\langle b, e_n \rangle \overset{d}{=} \langle b^0, e_n \rangle \forall n > N\) for \((b, \epsilon) \sim P^*_N\) and \((b^0, \epsilon^0) \sim P_0\). Note that \(P^*_N\) is a Borel-measurable Gaussian random variable on \(C(D)\). We now show our claim \((13)\). 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) if of zero-mean since it has the same distribution under \(P_0\). Thus for any affine \(\phi\),

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

Let \(\{\tilde{e}_n\}_{n=1}^{N}\) be the Hilbert-Schmidt normalization of \(\{T(e_n)\}_{n=1}^{N}\), with \(\tilde{N} = \tilde{N}(N) \to \infty\) as \(N \to \infty\), where if \(\{T(e_n)\}_{n=1}^{\infty}\) is finite-dimensional, then we formally add \(\tilde{e}_n\) as the zero
function if necessary to avoid notation burdens. Thus we can write

$$\sup_{P: W(P,P_0) \leq \delta, \langle b, e_n \rangle \equiv \langle b^0, e_n \rangle} \inf_{\phi} \text{Obj}(\phi, P)$$

$$\geq \inf_{\phi} \text{Obj}(\phi, P_N^*)$$

$$= \inf_{\phi, \text{affine}} E_{P_N^*} \left[ \| u - \phi(Y_1, \ldots, Y_m) \|_{L^2(D)}^2 \right]$$

Due to Gaussianity

$$\geq \inf_{\phi, \text{affine}} E_{P_N^*} \left[ \| u \|_{\text{span}\{T(e_n)\}_{n=1}^N} - \phi(Y_1, \ldots, Y_m) \|_{\text{span}\{T(e_n)\}_{n=1}^N} \|_{L^2(D)}^2 \right]$$

$$= \inf_{\phi, \text{affine}} E_{P_N^*} \left[ \left( u - \sum_{j=1}^m \alpha_j \left( \sum_{k=1}^N \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right), \tilde{e}_n \right)^2 \right]$$

$$+ \sum_{n=1}^N \tilde{N} E_{P_N^*} \left[ \sum_{j=1}^m \alpha_j R_j,N, \tilde{e}_n \right]$$

by independence and zero-mean of $R_j,N$

$$\geq \inf_{\phi, \text{affine}} E_{P_N^*} \left[ \left( u - \sum_{j=1}^m \alpha_j \left( \sum_{k=1}^N \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right), \tilde{e}_n \right)^2 \right]$$

$$= \sup_{P: W(P,P_0) \leq \delta, \langle b, e_n \rangle \equiv \langle b^0, e_n \rangle} \inf_{\phi} \text{Obj}(\phi, P(N)),$$

where $f|_{\text{span}\{T(e_n)\}_{n=1}^N}$ denote the orthogonal projection of function $f \in C(D)$ to the subspace $\text{span}\{T(e_n)\}_{n=1}^N$, thus we have established Claim 2.

**Proof of Claim 3.** Note that $\phi_N^*$ has the form $\phi_N^*(Y_1, \ldots, Y_m) = \sum_{j=1}^m \alpha_j^{(N)} Y_j$, where $\alpha_j^{(N)} \in \text{span}\{T(e_n)\}_{n=1}^N$. Also note that we have

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

where $C$ is independent of $N$, since the objective in (18) is lower bounded by $C' \sum_{j=1}^m \| \alpha_j \|_{L^2(D)}^2$ for some constant $C'$ independent of $N$, by choosing $(0, \Sigma) \in \mathcal{S}_N$ in the inner constraint of (18) to be the nominal measure and note that the noise is independent of the prior. For

33
any $N$, we have

$$\sup_{P: W(P, P_0) \leq \delta} \text{Obj}(\phi_N^*, P) - \sup_{P: W(P, P_0) \leq \delta, \langle b, e_n \rangle \equiv (b^0, e_n) \forall n > N} \text{Obj}(\phi_N^*, P^{(N)})$$

$$= \sup_{P: W(P, P_0) \leq \delta} \left( \sum_{n=1}^{\tilde{N}} \mathbb{E}_P(\langle u - \phi_N^*, \tilde{e}_n \rangle)^2 + \sum_{n=N+1}^{\infty} \mathbb{E}_P(\langle u, \tilde{e}_n \rangle)^2 \right)$$

$$- \sup_{P: W(P, P_0) \leq \delta, \langle b, e_n \rangle \equiv (b^0, e_n) \forall n > N} \sum_{n=1}^{\tilde{N}} \mathbb{E}_{P^{(N)}}(\langle u - \phi_N^*, \tilde{e}_n \rangle)^2$$

$$\leq \sup_{P: W(P, P_0) \leq \delta} \sum_{n=1}^{\tilde{N}} \mathbb{E}_P(\langle u - \phi_N^*, \tilde{e}_n \rangle)^2$$

$$- \sup_{P: W(P, P_0) \leq \delta, \langle b, e_n \rangle \equiv (b^0, e_n) \forall n > N} \sum_{n=1}^{\tilde{N}} \mathbb{E}_{P^{(N)}}(\langle u - \phi_N^*, \tilde{e}_n \rangle)^2$$

$$+ \sup_{P: W(P, P_0) \leq \delta} \sum_{n=\tilde{N}+1}^{\infty} \mathbb{E}_P(\langle u, \tilde{e}_n \rangle)^2.$$  \hspace{1cm} (25)

For the last term in (25), note that

$$\sup_{P: W(P, P_0) \leq \delta} \mathbb{E}_P \left[ \sum_{n=\tilde{N}+1}^{\infty} (\langle u, \tilde{e}_n \rangle)^2 \right]$$

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

$$\leq o(1) + O(1) \sup_{P: W(P, P_0) \leq \delta} \mathbb{E}_\pi \left[ \|u - u^0 - \sum_{n=1}^{N} \langle b - b^0, e_n \rangle T(e_n) \|_{L^2(\mathcal{P})}^2 \right].$$
where $\pi$ is the optimal coupling between $P_0$ and $P$, and

\[
\sup_{P:W(P, P_0) \leq \delta} \mathbb{E}_\pi \left[ \right] \sum_{n=1}^{N} \langle \bar{b} - b^0 , e_n \rangle \bar{e}_n \right] \right]^{2} \right. \\
\leq O(1) \sup_{P:W(P, P_0) \leq \delta} \mathbb{E}_\pi \left[ \right] \sum_{n=1}^{N} \langle \bar{b} - b^0 , e_n \rangle \bar{e}_n \right] \right]^{2} \right. \\
\leq O(1) \sup_{P:W(P, P_0) \leq \delta} \mathbb{E}_\pi \left[ \right] \sum_{n=1}^{N} \langle \bar{b} - b^0 , e_n \rangle \bar{e}_n \right] \right]^{2} \right. \\
\leq o(1) \sup_{P:W(P, P_0) \leq \delta} \mathbb{E}_\pi \left[ \right] \sum_{n=1}^{N} \langle \bar{b} - b^0 , e_n \rangle \bar{e}_n \right] \right]^{2} \right. \\
\to 0 \text{ as } N \to \infty,
\]

where the penultimate inequality comes from that $\bar{w}_n = o(w_n)$ and $b - b^0 - \sum_{n=1}^{N} (b - b^0) e_n$ is in the space spanned by $\{e_n\}_{n=N+1}$. For the first two terms in (25), we write

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

Thus for any feasible $P$ in $\{ P : W(P, P_0) \leq \delta \}$, we have

\[
\mathbb{E}_P(\langle u - \phi_N^*, \bar{e}_n \rangle)^2 \\
= \mathbb{E}_P \left( \right) \left( \sum_{j=1}^{m} \alpha_j^{(N)} (x) \left( \sum_{k=1}^{N} \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right) \right)^2 \right) \\
- 2 \mathbb{E}_P \left( \right) \left( \sum_{j=1}^{m} \alpha_j^{(N)} (x) \left( \sum_{k=1}^{N} \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right) \right) \left( \sum_{j=1}^{m} \alpha_j^{(N)} (x) R_{j,N} , e_n \right) \\
+ \mathbb{E}_P \left( \right) \left( \sum_{j=1}^{m} \alpha_j^{(N)} (x) R_{j,N} , e_n \right)^2 \\
\right)
\]

Note that

\[
\sum_{n=1}^{N} \mathbb{E}_P \left( \right) \left( \sum_{j=1}^{m} \alpha_j^{(N)} (x) \left( \sum_{k=1}^{N} \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_j \right) \right)^2 \right) \\
\leq \sup_{P:W(P, P_0) \leq \delta, b, e_n \leq \langle b^0, e_n \rangle \forall n > N \geq 1} \sum_{n=1}^{N} \mathbb{E}_{P^{(N)}} \left( \langle u - \phi_N^*, \bar{e}_n \rangle \right)^2.
\]

35
Also, by Cauchy-Schwarz

\[
\hat{N} \sum_{n=1} E_P \left( \left< u - \sum_{j=1}^m \alpha_j^{(N)}(x) \left( \sum_{k=1}^N \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_i \right), \hat{e}_n \right> \right) \leq \sqrt{E_P \left( \hat{N} \sum_{n=1} \left< u - \sum_{j=1}^m \alpha_j^{(N)}(x) \left( \sum_{k=1}^N \langle b, e_k \rangle T(e_k)(x_j) + \epsilon_i \right), \hat{e}_n \right>^2 \right)}.
\]

\[
\leq \sup_{P : W(P, P_0) \leq \delta} \sum_{n=1}^{\hat{N}} E_{P(n)} \left( \left< u - \phi^*_N, \hat{e}_n \right> \right)^2 \leq \sup_{P : W(P, P_0) \leq \delta} \sum_{n=1}^{\hat{N}} E_{P(n)} \left( \left< u - \phi^*_N, \hat{e}_n \right> \right)^2 \leq 2 \sup_{P : W(P, P_0) \leq \delta} \left< u(x_j) - \sum_{k=1}^N \langle b^0, e_k \rangle T(e_k)(x_j), \hat{e}_n \right>^2 \leq 2 \sup_{P : W(P, P_0) \leq \delta} \left< u(x_j) - \sum_{k=1}^N \langle b^0, e_k \rangle T(e_k)(x_j), \hat{e}_n \right>^2.
\]

Therefore, denote \( \pi \) as the optimal coupling between \( P \) and \( P_0 \), we have

\[
\sup_{P : W(P, P_0) \leq \delta} \mathbb{E}_P (R_{j,N})^2
\leq 2 \mathbb{E}_{P_0} \left( u^0(x_j) - \sum_{k=1}^N \langle b^0, e_k \rangle T(e_k)(x_j) \right)^2 + 2 \sup_{P : W(P, P_0) \leq \delta} \mathbb{E}_\pi \left( u(x_j) - u^0(x_j) - \sum_{k=1}^N \langle b - b^0, e_k \rangle T(e_k)(x_j) \right)^2 \leq o(1) + 2 \sup_{P : W(P, P_0) \leq \delta} \mathbb{E}_\pi \left( u(x_j) - u^0(x_j) - \sum_{k=1}^N \langle b - b^0, e_k \rangle T(e_k)(x_j) \right)^2.
\]
where

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

$$= \sup_{P: W(P, P_0) \leq \delta} \mathbb{E}_\pi \left[ \left( T(b - b^0 - \sum_{k=1}^N \langle b - b^0, e_k \rangle e_k)(x_j) \right)^2 \right]$$

$$\leq O(1) \sup_{P: W(P, P_0) \leq \delta} \mathbb{E}_\pi \left[ \| T(b - b^0 - \sum_{k=1}^N \langle b - b^0, e_k \rangle e_k) \|_{\mathcal{H}_w}^2 \right]$$

$$\leq O(1) \sup_{P: W(P, P_0) \leq \delta} \mathbb{E}_\pi \left[ \| b - b^0 - \sum_{k=1}^N \langle b - b^0, e_k \rangle e_k \|_{\mathcal{H}_w}^2 \right]$$

$$\leq o(1) \sup_{P: W(P, P_0) \leq \delta} \mathbb{E}_\pi \left[ \| b - b^0 \|_{\mathcal{H}_w}^2 \right]$$

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

where the first inequality follows because $\mathcal{H}_w$ is an RKHS so that the point evaluation at $x_j$ is bounded. The second inequality follows from Assumption 2.3 while the third equality follows from Assumption 2.4. Finally, the last inequality follows from $\{e_n\}$ being an orthogonal system. Therefore, we have

$$\sup_{P: W(P, P_0) \leq \delta} \mathbb{E}_P (R_{j,N})^2 \rightarrow 0 \text{ as } N \rightarrow \infty.$$

Therefore, we have

$$\sup_{P: W(P, P_0) \leq \delta} \text{Obj}(\phi_N^*, P) - \sup_{P: W(P, P_0) \leq \delta, (b,e_n)=\langle b^0,e_n \rangle \forall n \geq N} \text{Obj}(\phi_N^*, P^{(N)}) \leq o(1) \text{ as } N \rightarrow \infty,$$

thereby establishing our Claim 3.

6.2 Proof to Proposition 3.1

*Proof to Proposition 3.1.* We note that this is an immediate consequence of inequalities (11) and Claims 1-3 in the proof to Theorem 3.1.

6.3 Proof to Proposition 3.2

*Proof of Proposition 3.2.* Take an arbitrary positive sequence $a_n$ such that $\sum_{n \geq 1} a_n^2 w_n < \infty$, the sets

$$\mathcal{C}_a = \{(\Delta b, \epsilon) \in \mathcal{H}_w \times \mathbb{R}^m : |\langle \Delta b, e_n \rangle| \leq a \cdot a_n \ \forall n \geq 1, |\epsilon_j| \leq a \ \forall j = 1, \ldots, m\},$$
for any $a > 0$ is a compact subset of $\mathcal{H}_w \times \mathbb{R}^m$. For $W(P_N, P_0) \leq \delta$ and $(b, \epsilon) \sim P_N$ optimally coupled to $(b^0, \epsilon^0) \sim P_0$, and denote $Q_N$ as the measure of the difference $(b - b^0, \epsilon - \epsilon^0)$, we see that for any $\eta$, there exists a sufficiently large, such that $Q_N(C_\alpha) \geq 1 - \eta$ for all $N$. Otherwise we find $\eta$, and a subsequence $N_k$, such that $Q_{N_k}(C_k) \leq \eta$ for all $k \geq 1$, so that $W(P_{N_k}, P_0) \to \infty$, a contradiction!

Thus the sequence $(Q_N, P_0)$ is tight, by Prokhorov’s theorem there exists a weakly convergent subsequence (by abuse of notation assume the subsequence is the original sequence) with a limit denoted as $(Q_\infty, P_0)$. Recovering the measure $P_\infty$ on $(b, \epsilon)$ from the coupling $(b - b^0, \epsilon - \epsilon^0) \sim Q_\infty$ and $(b^0, \epsilon^0) \sim P_0$, we have

$$W(P_\infty, P_0) \leq \delta.$$  

This completes the proof.$\square$

### 6.4 Proof to Theorem 3.2

We start with a simple but useful result, which is well-known.

**Lemma 6.1** (Strong duality and existence of Nash equilibrium). Let $\phi^*$ be the optimal solution to the outer infimum of the problem

$$\inf_{\phi \in \mathcal{M}} \sup_{P \in \mathcal{P} : W(P, P_0) \leq \delta} \text{Obj}(\phi, P),$$

and let $P^*$ be the optimal solution to the outer supremum of the problem

$$\sup_{P \in \mathcal{P} : W(P, P_0) \leq \delta} \inf_{\phi \in \mathcal{M}} \text{Obj}(\phi, P).$$

Then $(\phi^*, P^*)$ consists of a pair of Nash equilibrium of problem (15).

**Proof of Lemma 6.1**. We have

$$\sup_{P \in \mathcal{P} : W(P, P_0) \leq \delta} \inf_{\phi} \text{Obj}(\phi, P) = \inf_{\phi} \text{Obj}(\phi, P^*) \leq \text{Obj}(\phi^*, P^*) \leq \sup_{P \in \mathcal{P} : W(P, P_0) \leq \delta} \text{Obj}(\phi^*, P) = \inf_{\phi} \sup_{P \in \mathcal{P} : W(P, P_0) \leq \delta} \text{Obj}(\phi, P).$$

By the strong duality, all inequalities become equalities, thus we have

$$\text{Obj}(\phi^*, P^*) = \min_{\phi} \text{Obj}(\phi, P^*) = \max_{P : W(P, P_0) \leq \delta} \text{Obj}(\phi^*, P).$$

The proof is complete.$\square$
Proof of Theorem 3.2. It is easy to compute the objective value \( \text{Obj}(\phi_N^*, Q_N^*) \) as

\[
\text{Obj}(\phi_N^*, Q_N^*) = \int_D \mathbb{E}_{Q_N^*} [\text{Var}_{Q_N^*} [u(x)|Y_1, \ldots, Y_m]] dx
\]

\[
= \int_D k^{(N)}(x, x)
- \left( K^{(N)}(x, x_1), \ldots, k^{(N)}(x, x_m) \right) (K^{(N)}_\epsilon)^{-1} \left( k^{(N)}_\epsilon(x, x_1), \ldots, k^{(N)}_\epsilon(x, x_m) \right)^\top dx,
\]

where \( k^{(N)}(x, x) = \mathbb{E}_{Q_N^*}[u(x)^2] \) and the objective value \( \text{Obj}(\phi_\infty^*, P_\infty^*) \) is

\[
\text{Obj}(\phi_\infty^*, P_\infty^*) = \int_D \mathbb{E}_{P_\infty^*} [\text{Var}_{P_\infty^*} [u(x)|Y_1, \ldots, Y_m]] dx
\]

\[
= \int_D k(x, x) - (k_\epsilon(x, x_1), \ldots, k_\epsilon(x, x_m)) \cdot (K_\epsilon)^{-1} \cdot (k_\epsilon(x, x_1), \ldots, k_\epsilon(x, x_m))^\top dx,
\]

where \( k(x, x) = \mathbb{E}_{P_\infty^*}[u(x)^2] \). Note that \( P_{N_l}^* \) is a Gaussian convergent sequence, \( T \) is bounded linear operator, and note that the tail-difference in \( Q_{N_l}^* \), and \( P_{N_l}^* \) is negligible. Thus by [16, Exercise 2.1.4], for a Gaussian sequence convergence in distribution implies convergence of the second moment. Thus uniformly for \( x \in D \), \( k^{(N_l)}(x, x) \rightarrow k(x, x) \) and \( k^{(N_l)}_\epsilon(x, x_i) \rightarrow k_\epsilon(x, x_i) \). Also \( K^{(N_l)}_\epsilon \rightarrow K_\epsilon \). Thus

\[
\text{Obj}(\phi_{N_l}^*, Q_{N_l}^*) = \min_\phi \text{Obj}(\phi, Q_{N_l}^*) \rightarrow \text{Obj}(\phi_\infty^*, P_\infty^*) = \min_\phi \text{Obj}(\phi, P_\infty^*)
\]

Thus \( P_\infty^* \) solves the RHS of (13).

We next show that \( \phi_\infty^* \) solves the LHS of (15). We first note that due to convergence of moments of \( P_{N_l}^* \), that

\[
\phi_{N_l}^*(\cdot, \ldots, \cdot)(\cdot) \rightarrow \phi_\infty^*(\cdot, \ldots, \cdot)(\cdot) \text{ everywhere point-wise as } l \rightarrow \infty.
\]

Next, by Fatou’s lemma, we have for any \( P \)

\[
\mathbb{E}_P \left[ \| u - \phi_\infty^*(Y_1, \ldots, Y_m) \|_{L^2(D)}^2 \right] \leq \liminf_{l \rightarrow \infty} \mathbb{E}_P \left[ \| u(\cdot) - \phi_{N_l}^*(Y_1, \ldots, Y_m) \|_{L^2(D)}^2 \right].
\]

Therefore,

\[
\sup_{P: W(P, P_0) \leq \delta} \mathbb{E}_P \left[ \| u - \phi_\infty^*(Y_1, \ldots, Y_m) \|_{L^2(D)}^2 \right]
\]

\[
\leq \liminf_{l \rightarrow \infty} \sup_{P: W(P, P_0) \leq \delta} \mathbb{E}_P \left[ \| u - \phi_{N_l}^*(Y_1, \ldots, Y_m) \|_{L^2(D)}^2 \right].
\]

39
Note that due to the finite-dimensional strong duality (12), \( \phi_N^* \) also solves
\[
\min_{\phi, \text{affine}, \text{coef}(\phi) \in \text{span}(T(\epsilon_n))} \max_{P: W(P, P_0) \leq \delta, (b, \epsilon_n) \in \mathbb{R}^{2N}} \text{Obj}(\phi, P^{(N)}).
\]

By (14) and the strong duality
\[
\liminf_{N \to \infty} \sup_{P: W(P, P_0) \leq \delta} \mathbb{E}_P \left[ \left\| u(\cdot) - \phi_N^*(Y_1, \ldots, Y_m) \right\|_{L^2(D)}^2 \right] = \min_{\phi, P: W(P, P_0) \leq \delta} \text{Obj}(\phi, P)
\]

Thus \( \phi_N^* \) solves the min-max problem.

Finally, we note that \( \phi^* \) is unique, since any \( \phi^* \) has to solve \( \min_{\phi} \text{Obj}(\phi, P_N^*) \) by Lemma 6.1. Thus we have for the entire sequence
\[
\phi_N^*(\cdot, \ldots, \cdot)(\cdot) \to \phi_N^*(\cdot, \ldots, \cdot)(\cdot) \text{ everywhere point-wise as } N \to \infty.
\]

Moreover, we claim that \( P_N^* \) is also unique, from which the entire sequence \( P_N^* \) converges to \( P_\infty^* \) in the weak topology. By Lemma 6.1, any \( P^* \) has to solve \( \max_{P: W(P, P_0) \leq \delta} \text{Obj}(\phi_N^*, P) \).

By Theorem 1 in [6], the problem admits the reformulation
\[
\max_{P: W(P, P_0) \leq \delta} \text{Obj}(\phi_N^*, P) = \inf_{\gamma \geq 0} \left( \gamma \delta^2 + \mathbb{E}_P \left[ \sup_{(b, \epsilon) \in C(D) \times \mathbb{R}^m} \left( \left\| T(b) - \phi_N^*((T(b)(x_i) + \epsilon_i)i) \right\|_{L^2(D)}^2 - \gamma \left( \left\| b - b^0 \right\|_{H_w}^2 + \left\| \epsilon - \epsilon^0 \right\|_2^2 \right) \right] \right) \right)
\]

We claim that the optimal dual \( \gamma^* \) is sufficiently large so that
\[
\left\| T(b) - \phi_N^*((T(b)(x_i) + \epsilon_i)i) \right\|_{L^2(D)}^2 - \gamma^* \left( \left\| b - b^0 \right\|_{H_w}^2 + \left\| \epsilon - \epsilon^0 \right\|_2^2 \right) < 0 \quad \forall (b, \epsilon) \neq 0 \in H_w \times \mathbb{R}^m. \quad (26)
\]

It is easy to see that \( \gamma^* > 0 \). Suppose (26) does not hold, then
\[
\left\| T(b^*) - \phi_N^*((T(b^*)(x_i) + \epsilon^*_i)i) \right\|_{L^2(D)}^2 - \gamma^* \left( \left\| b^* \right\|_{H_w}^2 + \left\| \epsilon^* \right\|_2^2 \right) \geq 0,
\]
for some \( (b^*, \epsilon^*) \neq 0 \). Then for all \( (b^0, \epsilon^0) \in C(D) \times \mathbb{R}^m \) satisfying
\[
\langle T(b^*) - \phi_N^*((T(b^*)(x_i) + \epsilon^*_i)i), T(b^0) - \phi_N^*((T(b^0)(x_i) + \epsilon^0_i)i) \rangle \neq 0,
\]

Moreover, observe that the constraint

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

\[ \geq \sup_t \left( \| t (T(b^*) - \phi_\infty^*((T(b^*)(x_i) + \epsilon^*_i)_i)) + (T(b^0) - \phi_\infty^*((T(b^0)(x_i) + \epsilon^0_i)_i)) \|_{L^2(D)}^2 \right) \]

\[ - t^2 \gamma^*(\| b^* \|_{\mathcal{H}_w}^2 + \| \epsilon^* \|_2^2) \]

\[ \geq \sup_t t^2 \left( \| T(b^*) - \phi_\infty^*((T(b^*)(x_i) + \epsilon^*_i)_i) \|_{L^2(D)}^2 - 2 \gamma^*(\| b^* \|_{\mathcal{H}_w}^2 + \| \epsilon^* \|_2^2) \right) \]

\[ + 2t \langle T(b^*) - \phi_\infty^*((T(b^*)(x_i) + \epsilon^*_i)_i), T(b^0) - \phi_\infty^*((T(b^0)(x_i) + \epsilon^0_i)_i) \rangle \]

\[ + \| T(b^0) - \phi_\infty^*((T(b^0)(x_i) + \epsilon^0_i)_i) \|_{L^2(D)}^2 \]

\[ = \infty. \]

Moreover, observe that the constraint

\[ \langle T(b^*) - \phi_\infty^*((T(b^*)(x_i) + \epsilon^*_i)_i), T(b^0) - \phi_\infty^*((T(b^0)(x_i) + \epsilon^0_i)_i) \rangle = 0 \quad (27) \]

is a linear constraint on \((b^0, \epsilon^0)\), and \((b^0, \epsilon^0)\) cannot take the value \((b^*, \epsilon^*)\). Thus the collection of \((b^0, \epsilon^0)\) satisfying (27) is subset of a proper linear subspace of \(C(D) \times \mathbb{R}^m\). The collection is closed since \(T : C(D) \to C(D)\) is bounded. Thus event (27) has probability strictly less than one. Thus concluding claim (26).

Denote

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

By calculus of variation, any maximizer of \( J \) satisfies

\[ \frac{\partial}{\partial t} J(b + t \Delta b, \epsilon + t \Delta \epsilon) \bigg|_{t=0} = 0, \quad \forall (\Delta b, \Delta \epsilon) \in C(D) \times \mathbb{R}^m. \]

It is easy to compute that

\[ \frac{1}{2} \frac{\partial}{\partial t} J(b + t \Delta b, \epsilon + t \Delta \epsilon) \bigg|_{t=0} \]

\[ = \langle T(b) - \phi_\infty^*((T(b)(x_i) + \epsilon_i)_i), T(\Delta b) - \phi_\infty^*((T(\Delta b)(x_i) + \Delta \epsilon_i)_i) \rangle \]

\[ - \gamma^* ((b - b^0, \Delta b)_{\mathcal{H}_w} + (\epsilon - \epsilon^0, \Delta \epsilon)). \]

Suppose \((b, \epsilon)\) and \((\bar{b}, \bar{\epsilon})\) are two maximizers of \( J \), choosing \((\Delta b, \Delta \epsilon) = (b - \bar{b}, \epsilon - \bar{\epsilon})\), we have

\[ \| T(\Delta b) - \phi_\infty^*((T(\Delta b)(x_i) + \Delta \epsilon_i)_i) \|_{L^2(D)}^2 - 2 \gamma^*(\| \Delta b \|_{\mathcal{H}_w}^2 + \| \Delta \epsilon \|_2^2) = 0. \]

From our earlier claim (26), we conclude that \((b, \epsilon) = (\tilde{b}, \tilde{\epsilon})\). By Theorem 1 in [6], this shows that the primal optimal transport plan is unique, thus \( P^* \) is unique. \(\square\)
6.5 Proof to Proposition 3.3

Proof to Proposition 3.3. First, suppose that $K_\epsilon$ is invertible for the limit $P^\star_N$ of some weakly convergent subsequence of $P^\star_N$. Then by Theorem 3.2, the entire sequence $P^\star_N \Rightarrow P^\star_\infty$. Since the sequence of $P^\star_N$ is a Gaussian convergent sequence, we have $\det(K_\epsilon(N)) \rightarrow \det(K_\epsilon) \neq 0$ as $N \rightarrow \infty$.

Next, suppose $K_\epsilon$ is not invertible (equivalently, $\det(K_\epsilon) = 0$) for the limit $P^\star_\infty$ of every weakly convergent subsequence of $P^\star_N$, and suppose that $\det(K_\epsilon(N))$ does not converge to 0 as $N \rightarrow \infty$. Thus there exists a subsequence of $P^\star_N$, denoted by $P^\star_{N_l}$, $l \geq 1$, such that $\det(K_\epsilon(N_l))$ is uniformly bounded away from 0 and $P^\star_{N_l}$ has a weak limit (upon passing into a further subsequence, e.g., by Proposition 3.2). It is clear that $K_\epsilon$ for the limit $P^\star_\infty$ of $P^\star_{N_l}$ is invertible. A contradiction!

6.6 Proof of Lemma 3.3

Proof of Lemma 3.3. For any $P$ such that $W(P, P_0) \leq \delta$, to show that the matrix in question is strictly positive definite, it suffices to show that

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

Note that we can always find a coupling $\pi$ such that

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

where $(b, \epsilon) \sim P$ and $(b^0, \epsilon^0) \sim P_0$. 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 $H_\tilde{w}$)

$$|f(x_i)| \leq C_m \|f\|_{H_\tilde{w}}, \quad \forall f \in H_\tilde{w}, i = 1, \ldots, m.$$
By Fatou's lemma

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

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

\[
\leq C_m^2 \mathbb{E}_\pi \left[ \| T \left( \sum_{k=1}^{N} \langle b - b^0, e_k \rangle e_k \right) \|_{H_w}^2 \right]
\]

\[
\leq C_m^2 C_w^2 \mathbb{E}_\pi \left[ \| \sum_{k=1}^{N} \langle b - b^0, e_k \rangle e_k \|_{H_w}^2 \right]
\]

\[
\leq C_m^2 C_w^2 O(1) \mathbb{E}_\pi \left[ \| \sum_{k=1}^{N} \langle b - b^0, e_k \rangle e_k \|_{H_w}^2 \right]
\]

\[
\leq O(1) \mathbb{E}_\pi \left[ \| b - b^0 \|_{H_w}^2 \right].
\]

For any \( \xi \in \mathbb{R}^d \) and \( \| \xi \|_2 = 1 \),

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

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

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

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

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

\[
\leq 2 \mathbb{E}_P \left[ \left( \sum_{i=1}^{m} \xi_i (u(x_i) + \epsilon_i) \right)^2 \right] + O(1) \delta^2.
\]
Note that \((\mathbb{E}_{P_0}[Y_iY_j])_{ij} \geq \sigma^2 I_{m \times m}\), since \(\epsilon_i^0\) are independent \(\mathcal{N}(0, \sigma^2)\) noise under \(P_0\). Thus

\[
\inf_{\xi \in \mathbb{R}^d, \|\xi\|_2 = 1} \mathbb{E}_{P_0} \left[ \left( \sum_{i=1}^{m} \xi_i (u(x_i) + \epsilon_i) \right)^2 \right] \geq \frac{1}{2} \sigma^2 - O(1) \delta^2 > 0,
\]

for \(\delta < \delta_0\) where \(\delta_0\) depends on \(T, m, (x_i)_i, H_w, H_{\tilde{w}}, \sigma^2\). 

\[\square\]

### 6.7 Proof to Theorem 3.4

The proof for the inverse problem works verbatim as the proof to Theorem 3.1 with minor modifications. For example, instead of considering \(\text{coef}(\phi) \in \text{span}\{T(e_n)\}_{n=1}^N\), we consider \(\text{coef}(\phi) \in \text{span}\{e_n\}_{n=1}^N\). We suppress details to avoid repetitions.