Experimental Design for Linear Functionals in Reproducing Kernel Hilbert Spaces

Mojmír Mutný
ETH Zürich
mojmir.mutny@inf.ethz.ch

Andreas Krause
ETH Zürich
krausea.ethz.ch

Abstract

Optimal experimental design seeks to determine the most informative allocation of experiments to infer an unknown statistical quantity. In this work, we investigate the optimal design of experiments for *estimation of linear functionals in reproducing kernel Hilbert spaces (RKHSs)*. This problem has been extensively studied in the linear regression setting under an estimability condition, which allows estimating parameters without bias. We generalize this framework to RKHSs, and allow for the linear functional to be only approximately inferred, i.e., with a fixed bias. This scenario captures many important modern applications, such as estimation of gradient maps, integrals, and solutions to differential equations. We provide algorithms for constructing bias-aware designs for linear functionals. We derive non-asymptotic confidence sets for fixed and adaptive designs under sub-Gaussian noise, enabling us to certify estimation with bounded error with high probability.

1 Introduction

Optimal Experimental Design (OED) aims to determine data collection schemes – *designs* – to efficiently estimate unknown quantities of interest given limited resources (Chaloner and Verdinelli, 1995). As common, we model experiments via an *oracle* that yields a (noisy) response to a given input. A design is usually either an (adaptive) policy for querying the oracle, or a (nonadaptive) fixed allocation of query budget to different oracle inputs. OED has a rich history and close relations to the field of bandits (Szepesvari and Lattimore, 2020) and active learning (Settles, 2009).

We consider the regression setting, where observations at a fixed input $x \in \mathcal{X}$ can be obtained via the noisy oracle by:

$$y = \theta^\top \Phi(x) + \epsilon \text{ where } x \in \mathcal{X} \tag{1}$$

and $\cdot^\top \cdot$ denotes the inner product in a Hilbert space $\mathcal{H}_\kappa$, ϵ is independent, sub-Gaussian noise with *known* variance proxy σ^2, and θ is a bounded element from a separable reproducing kernel Hilbert space $\mathcal{H}_\kappa$ from kernel κ, with a bound $\theta^\top \mathcal{V}_0 \theta \leq \lambda^{-1}$, where $\mathcal{V}_0 : \mathcal{H}_k \to \mathcal{H}_k$ is a positive definite operator. [1] Depending on whether λ is known or unknown, we will propose different estimators. The norm constraint (under $\mathcal{V}_0$) can be interpreted as a bound on the total effect and its parameters if the Hilbert space is finite dimensional. For infinite dimensional spaces this assumption can model the additional constraint on regularity. For example, for kernels that induce Sobolev spaces, it explicitly bounds the total squared derivatives of functions – quantities known to govern complexity of approximation difficulty. The exact bound depends on the specific definition of κ, but if two spaces are isometrically isomorphic, the difference can be absorbed to $\mathcal{V}_0$ to induce the desired prior regularity, e.g., the regularity of the derivatives. Here we assume that kernel κ and the operator $\mathcal{V}_0$ are known due to prior analysis and/or first principles modeling of the system.

In contrast to the classical ED task of estimating θ (see Fedorov and Hackl, 1997), we are interested in estimating a *projection* of θ. Namely, let $\mathbf{C} : \mathcal{H}_\kappa \to \mathbb{R}^p$ be a *known linear operator*. The map $\mathbf{C}$ is

[1] In most cases, but not all, $\mathcal{V}_0$ is chosen to be the identity.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

such that $\mathbf{C}^\top \mathbf{C}$ is full rank p, where $\mathbf{C}^\top$ denotes the adjoint. Our goal is to identify an estimate of $\mathbf{C}\theta$ efficiently, i.e., with low query complexity or with a maximal reduction of uncertainty given a fixed query budget T.

The formalism above captures, or occurs as a subroutine in, numerous practical problems. For example, *evaluation* at specific target points, *integration* and *differentiation* are all linear operators, among many other useful linear functionals. Other examples include ordinary or partial differential equations operators, spectral transforms, and stability metrics from control engineering. In Section 7 we detail several example applications.

A naive approach would be to first obtain an estimate $\hat{\theta}$ of θ, and compute $\mathbf{C}\hat{\theta}$. However, the appeal of estimating linear functionals *directly* is that the number of unknowns of interest may be *much lower* than for the original overall unknown element θ (which might even be infinite-dimensional). Consequently, we would hope that the query complexity of reducing the variance of the estimate $\mathbf{C}\theta$ scales in the dimension of the range of $\mathbf{C}$, which is p. For example, when focusing on finite-dimensional RKHSs, the operator $\mathbf{C} : \mathbb{R}^m \to \mathbb{R}^p$, where m is the dimension of $\mathcal{H}_\kappa$, becomes a matrix. In this work, we study the cases where $p < m$, and show that the estimation error can indeed scale with p, and the geometry of the query set $\{\Phi(x) | x \in \mathcal{X}\}$. The estimation *bias* plays a central role in this work as for very large m estimating $\mathbf{C}\theta$ might not be possible up to any precision. The difficulty depends among other things the richness of the query set $\mathcal{X}$, as we will detail later in Sec. 3.

Sequential Experiment Design Apart from classical experiment design, where we first commit to what set of queries x we choose, referred to as a *fixed design*, we also consider sequential design, where the selected queries depend on past observations. Suppose for example, we are gathering data to test whether the null H_0: $\theta^\top \Phi(x) \geq 0$ for all $x \in \mathcal{X}$ or otherwise. We can incrementally gather evidence and check whether the null hypothesis has already been rejected. As our data depends on prior evaluation points it forms an *adaptive design*. In this work, we develop confidence sets for both fixed and *adaptive designs* – which are of paramount importance in the context of sequential experiment designs, e.g., to define stopping rules of adaptive hypothesis testing problems.

Contributions **A)** We consider objectives for experiment design for linear functionals in general RKHS spaces, which carefully take the bias of the estimator into account. **B)** We provide bounds on the query complexity required to reach ϵ accuracy to estimate linear functionals with high probability. **C)** We construct novel non-asymptotic confidence sets for linear estimators of linear functionals of RKHS elements, both for *fixed* and *adaptive* designs, where queries are *independent* of previous noise realizations, and where they are not, respectively. **D)** We demonstrate the improved inference error due to specially defined designs and new confidence sets on the problems of learning differential equations, linear bandits, gradient maps estimation, and stability verification of non-linear systems.

2 Background and Related Work

Linear Estimators Let $\mathcal{S} \subset \mathcal{X}$ be a finite set of selected evaluations s.t. $|\mathcal{S}| = n$, potentially repeated. We focus on linear estimators of the form $\mathbf{L} : \mathbb{R}^n \to \mathbb{R}^p$, where $\mathbf{L}y$ is estimating $\mathbf{C}\theta$. An *estimator* is understood here as the algorithm to find the random *estimate* $\mathbf{L}y$. Notice that given $\mathcal{S}$, $\mathbf{L}$ is not a random quantity; the randomness rather comes from the realizations y. To choose the estimator $\mathbf{L}$, one classically looks at the second moment of the residuals $\mathbf{C}\theta - \mathbf{L}y$, $\mathbf{E}(\mathbf{L}) = \mathbb{E}[(\mathbf{C}\theta - \mathbf{L}y)(\mathbf{C}\theta - \mathbf{L}y)^\top]$, where the argument signifies how the random variable y is transformed (noise is averaged)

$$\mathbf{E}(\mathbf{L}) = \sigma^2 \mathbf{L}\mathbf{L}^\top + (\mathbf{L}\mathbf{X} - \mathbf{C})\theta\theta^\top(\mathbf{L}\mathbf{X} - \mathbf{C})^\top, \quad \text{where} \quad \mathbf{E}(\mathbf{L}) \in \mathbb{R}^{p \times p}, \tag{2}$$

and the matrix $\mathbf{X}$ contains stacked evaluation functionals of the RKHS $\mathbf{X}_{i,x} = \Phi_i(x)$ for $x \in \mathcal{S} \subseteq \mathcal{X}$. We will then seek a way to transform y via estimator $\mathbf{L}$ such that the second moment of residuals is minimized in certain sense.

Importance of Bias and RKHS Classical experimental design studies estimation of $\mathbf{C} : \mathbb{R}^m \to \mathbb{R}^p$ where the RKHS is *finite* dimensional (Pukelsheim, 2006) and $\mathbf{X} \in \mathbb{R}^{n \times m}$. On top of that, they consider only estimators which are *unbiased* for *any* θ, in other words, $\mathbf{L}\mathbf{X} = \mathbf{C}$. While, with finite dimensions, this simplification might be reasonable, for infinite dimensionial RKHSs, bias is *inevitable* and must be controlled. Consider the case of *estimating the gradient* of a continuous function f. We can nearly never learn it up to arbitrary precision from noisy point queries. However, estimation up to a small error is always possible and sufficient. Estimating $\nabla_x f(x)$ from points close to $f(x)$ will incur small bias but larger variance as the change in f compared noise ϵ is small. Hence balancing these two sources of error is crucial for an informative design in RKHSs.

Experiment Design: Classical Perspective Consider an m dimensional version of the model in (1). Then, among all unbiased estimators, linear least-squares estimators minimize (2) under the Löwner order due to the famed Gauss-Markov theorem. Unbiasedness in this form is synonymous with *estimability*, which means that by repeating the evaluation in $\mathbf{X}$, arbitrary precision can be reached (Pukelsheim, 2006). The second moment of residuals then becomes $\mathbf{W}_\dagger^{-1} = \mathbf{C}\mathbf{V}^\dagger\mathbf{C}^\top$, where $\cdot^\dagger$ denotes a generalized pseudo-inverse and $\mathbf{V} = \mathbf{X}^\top\mathbf{X}$. The matrix $\mathbf{W}$ is often referred to as the *information matrix*. Gaffke (1987) and Krafft (1983) note that estimability implies that $\mathbf{W}_\dagger$ is non-singular. We relax this condition and allow the estimator to have a *bias*; this means that we cannot in general reduce the error arbitrarily by repeating the measurements. In fact, our extension uses the fact that $\mathbf{W}_\dagger$ as defined above will not be singular even if estimability condition is not satisfied. We will show that the matrix $\mathbf{W}_\dagger$ can still play the role of the information matrix.

Experiment Design: Modern Challenges Mutný et al. (2020) use experimental design to estimate the Hessian of an unknown RKHS function, while Kirschner et al. (2019) use it to estimate the gradient of it for use in Bayesian optimization. Perhaps most related, Shoham and Avron (2020) study over-parametrized experimental design for one-shot active deep learning and analyze the bias in connection to experiment design, similarly as in the seminal work of Bardow (2008). They do not consider bias arising from the limited design space nor do they treat linear functionals. More broadly, the uncertainty propagation is studied in the Bayesian framework in the field of *probabilistic numerics* for linear and nonlinear operators (see Cockayne et al., 2017; Owhadi and Scovel, 2016, and citations therein).

Confidence sets Unlike classical statistics, our focus is on non-asymptotic confidence sets on regression estimators. They can be found in, e.g., Draper and Smith (2014) and Abbasi-Yadkori et al. (2011), for fixed and adaptive designs, respectively. Our goal is to define the confidence sets in the appropriate norm such that their width scales with the dimension of the range of $\mathbf{C}$, i.e., p, and the geometry of $\{\Phi(x)|x \in \mathcal{X}\}$, but not directly $\dim(\mathcal{H}_\kappa)$ nor number of data T. Mutný et al. (2020) derive *non-asymptotic* confidence sets that scale in p, but grow with the number of points T as they do not use the appropriate norm (see Appendix B.4). Similarly, Khamaru et al. (2021) study estimators for adaptively collected data, and propose *asymptotic confidence* sets for their estimators. Without a specific condition, their confidence sets can grow with T, however, they consider more general noise distributions apart from sub-Gaussian as we do here.

3 Estimation and Bias

In this section, we motivate the linear estimators and identify *information matrices*. Information matrices are the inverses of second moment matrices of residuals $\mathbf{E}$ as in Eq. (2). They are an important object in the analysis of the error of the estimators and their confidence sets. Also, they depend on the evaluations that define the estimator $\mathbf{L}$, $\mathbf{X}$. *Maximizing the information matrices* as a function of the chosen observations and their proportions $\mathbf{X}$ gives rise to *optimal experimental designs.*

Further, we identify quantities that influence the bias of the estimators. We study two estimators: the least-norm estimator (interpolation), when the bound on $\|\theta\|_{\mathcal{V}_0}$ is unknown but finite, or the ridge regularized least squares estimator, where $\|\theta\|_{\mathcal{V}_0} \leq \lambda^{-1}$ is known. Both estimators are motivated as minimizing the error residuals $\mathbf{E}$ in trace norm under these two bound assumptions.

3.1 Estimators

Interpolation As apparent from Eq. (2), without the knowledge of an explicit bound on the norm θ, the worst case over θ causes the optimal estimator $\mathbf{L}$ to minimize *only* the trace of the bias. This leads to minmization of $(\mathbf{C} - \mathbf{L}\mathbf{X})\mathcal{V}_0^{-1/2}$ in Frobenius norm, leading to the familiar *interpolation estimator*,

$$\mathbf{C}\hat{\theta} := \mathbf{L}_\dagger y = \mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^\top\mathbf{K}^{-1}y, \quad \text{where} \quad \mathbf{K} = \mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^\top \tag{3}$$

The second moment of residuals can then be expressed as

$$\mathbf{E}(\mathbf{L}_\dagger) \preceq \underbrace{\sigma^2\mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^\top\mathbf{K}^{-2}\mathbf{X}\mathcal{V}_0^{-1}\mathbf{C}^\top}_{\text{variance}} + \underbrace{\frac{1}{\lambda}\mathbf{C}\mathcal{P}_\mathbf{X}\mathbf{C}^\top}_{\text{bias}},$$

where $\mathcal{P}_\mathbf{X} = \mathcal{V}_0^{-1/2}(\mathcal{V}_0^{-1/2}\mathbf{X}^\top\mathbf{K}^{-1}\mathbf{X}\mathcal{V}_0^{-1/2} - \mathcal{I})\mathcal{V}_0^{-1/2}$ is a scaled projection matrix. Unlike as in the classical ED treatment, the second moment has two terms: bias and variance. If the span of the scaled projection operator lies in the null space of $\mathbf{C}$, the bias (classically) vanishes. To control the error of estimation, we need to control both terms. For the special case of interpolation estimator,

we will control the second term separately using a *bias condition*, and the variance term will be controlled by the *information matrix* $\mathbf{W}_\dagger$ as we will see in Section 4

$$\mathbf{W}_\dagger(\mathbf{X}) = (\mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^\top\mathbf{K}^{-2}\mathbf{X}\mathcal{V}_0^{-1}\mathbf{C}^\top)^{-1}. \tag{4}$$

Regularized Regression The ridge regularized estimator is motivated by using $\theta^\top\mathcal{V}_0\theta \leq \lambda^{-1}$, and minimizing then the trace of this upper bound, leading to an estimator,

$$\mathbf{C}\hat{\theta}_\lambda := \mathbf{L}_\lambda y = \mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^\top(\lambda\sigma^2\mathbf{I} + \mathbf{K})^{-1}y. \tag{5}$$

Like above, we can give an upper bound on the second moment of the residuals. Conveniently, this estimator automatically balances the error due to variance and bias in one term: $\mathbf{E}(\mathbf{L}_\lambda) \preceq \mathbf{C}\theta\theta^\top\mathbf{C}^\top - \mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^\top(\mathbf{I}\sigma^2\lambda + \mathbf{K})^{-1}\mathbf{X}\mathcal{V}_0^{-1}\mathbf{C}^\top$. Using the matrix inversion lemma, we can express the above bound in a more concise form, and subsequently its inverse motivates the definition of the information matrix for the regularized estimator,

$$\mathbf{W}_\lambda(\mathbf{X}) = \sigma^{-2}(\mathbf{C}(\sigma^2\lambda\mathcal{V}_0 + \mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top)^{-1}. \tag{6}$$

3.2 Design objectives: Scalarization

The information matrices such as in (4) and (6) represent the inverse of estimation error, and the goal of optimal design is to maximize them (hence minimizing the error) with a proper choice of $\mathbf{X}$. As the Löwner order is not a total order, we need to resort to some scalarization of the information matrices, thus we solve $\max_{\mathbf{X}} f(\mathbf{W}(\mathbf{X}))$, where f is the scalarization $\mathbb{R}^p \to \mathbb{R}$. We focus on two common forms of scalarization and refer to them as E- and A-design (Pukelsheim, 2006).

- $f_E(\mathbf{W}) = \lambda_{\min}(\mathbf{W})$ – when constructing non-asymptotic high probability confidence sets;
- $f_A(\mathbf{W}) = 1/\operatorname{Tr}(\mathbf{W}^{-1})$ – when minimizing mean squared error of estimation.

Other popular criteria include D, V, and G-designs (Chaloner and Verdinelli, 1995), which we do not consider due to space considerations, but can be equivalently used should the experimenter have a reason for it. If $p = 1$, $\mathbf{C}$ is an element in $\mathcal{H}_k$ and the design problem degenerates into a special case known as c-optimality due to Elfving (1952) and, as $\mathbf{W} \in \mathbb{R}$, no scalarizations are needed.

Robust Designs The linear functionals are sometimes unknown, or parametrized by an *unknown* parameter γ as $\mathbf{C}_\gamma$, where γ belongs to a known set Γ. If we were to construct a design that has low estimation error for the worst case selection of γ, we can maximize the *information* in the following worst case metric $\max_{\mathbf{X}} \inf_{\gamma\in\Gamma} f\left(\left(\mathbf{C}_\gamma\mathcal{V}_0^{-1}\mathbf{X}^\top\mathbf{K}^{-2}\mathbf{X}\mathcal{V}_0^{-1}\mathbf{C}_\gamma^\top\right)^{-1}\right)$, for the interpolation estimator (and analogously for ridge regression). If the original function $f(\mathbf{W})$ is concave, then so is the function defined as the infimum over the compact index set Γ, which is true for the design criteria considered in this work (Boyd and Vandenberghe, 2004).

4 Fixed Designs and their Confidence Sets

In RKHS spaces, especially infinite dimensional ones, the *estimability* without bias is too restrictive. Given a finite evaluation budget T, we can only construct a discrete design $\mathbf{X} : \mathcal{H}_\kappa \to \mathbb{R}^T$, and there are many practical examples, where, given a finite query budget, $\mathbf{C}\theta$ cannot be learned to arbitrary precision, most prominently gradients and integrals among many others.

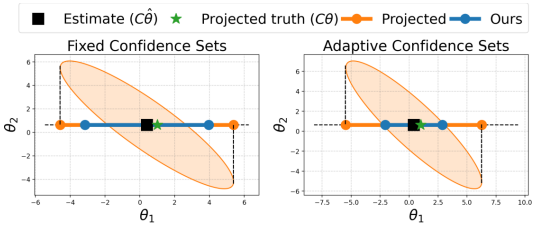


Figure 1: A two dim. example. Left: fixed design with our and projected confidence sets from two dimensions. Right: Adaptive confidence sets compared with projected ones due to Abbasi-Yadkori and Szepesvari (2012). In this example $\mathbf{C} = (1, 0)$ and $\theta = 0$.

Estimation with Bias and Interpolation Estimator Our goal is to establish a condition on the design space $\mathbf{X}$ such that the estimation with the interpolation estimator is possible up to a certain bias ϵ measured under the Euclidean norm. We measure the bias scaled by the magnitude of the Frobenius norm of the estimator and call this the relative ν-bias. This condition will allow us to balance the error due to the bias and noise.

Definition 1 (Relative ν-bias)**.** Let $\mathbf{C} : \mathcal{H}_\kappa \to \mathbb{R}^k$. The estimator $\mathbf{L}_\dagger$ on the design space $\mathbf{X}$ is said to have *relative ν-bias* if

$$\left\|(\mathbf{C} - \mathbf{L}_\dagger\mathbf{X})\mathcal{V}_0^{-1/2}\right\|_F^2 \leq \nu^2 \left\|\mathbf{L}_\dagger\right\|_F^2. \tag{7}$$

The $\|\cdot\|_F$ corresponds to the Frobenius norm of the maps $\mathcal{H}_\kappa \to \mathcal{H}_\kappa$. Due to the cyclic property of the trace, we can take the adjoint of the operator and calculate the quantity by taking a trace of $p \times p$ matrix instead. If $\nu = 0$ and $\dim(\mathcal{H}_\kappa) = m$, as we show in Lemma 1 in the Appendix, this is equivalent to the classical *estimability* condition due to Pukelsheim (2006). The left hand side corresponds to the *classical bias* $\text{bias}(\mathbf{L}_\dagger y) = \|\mathbb{E}[\mathbf{L}_\dagger y] - \mathbf{C}\theta\|_2$ of an estimator $\mathbf{L}_\dagger$ under the Frobenius norm. This is exactly the quantity that $\mathbf{L}_\dagger$ minimizes. The *classical bias* cannot be improved by repeated measurements or allocations thereof. Using a relation which we show formally in Proposition 5 in Appendix B, we can bound the $\text{bias}(\mathbf{L}_\dagger y) = \|\mathbb{E}[\mathbf{L}_\dagger y - \mathbf{C}\theta]\|_2 \leq \lambda_{\min}(\mathbf{W}_\dagger)^{-1/2}\nu/\sqrt{\lambda}$.

To check whether the condition (7) is satisfied given a design space $\mathbf{X}$, one needs to evaluate the trace. Finding the value of $p \times p$ matrix, however, depends strongly on the form of the operator $\mathbf{C}$, and a general recipe cannot be provided. Due to Riesz's representer theorem, and fact that the range of $\mathbf{C}$ is finite-dimensional, this is always possible. For example, for an integral operator $\int \Phi(x)^\top \cdot q(x)dx$, using shorthand $v = \mathbf{K}^{-1}y$, it holds that $\mathbf{L}_\dagger y = \sum_{i=1}^n \int \kappa(x, x_i)v_i dx$ where we used n points. Notice that we used only evaluations of kernel κ to do this calculation. Also notice that calculation of (7) reduces to the calculation of the maximum mean discrepancy (Gretton et al., 2005) between the empirical distribution (defined on $\mathbf{X}$), and $q(x)$ for this example. The general recipe is to (locally) project the Hilbert space on a truncated finite-dimensional basis $\Phi_m(x)$ with size m and, in that case, the check involves a solution to a linear system, but this is not always necessary.

4.1 Confidence sets

To construct confidence sets for both estimators considered, we split them into two categories: *fixed designs*, where the evaluation queries do not depend on the observed values y, and *adaptive designs*, where the queries $\Phi(x_i)$ *may depend* on prior evaluations $y_{i-1}, \dots y_1$. Proofs are in Appendix B.

Theorem 1 (Interpolation – Fixed Design)**.** *Under the regression model in Eq.* (1) *with T data point evaluations, let $\hat{\theta}$ be the estimate as in* (3)*. Let $\mathbf{X}$ satisfy the bias from Def. 1 with $\mathbf{W}_\dagger = (\mathbf{C}\mathcal{V}_0^{-1/2}\mathbf{V}^\dagger\mathcal{V}_0^{-1/2}\mathbf{C}^\top)^{-1}$. Then,*

$$\mathrm{P}\left(\left\|\mathbf{C}(\hat{\theta} - \theta)\right\|_{\mathbf{W}_\dagger} \geq \sigma\sqrt{\xi(\delta)} + \frac{\nu}{\sqrt{\lambda}}\right) \leq \delta, \tag{8}$$

where $\mathbf{V}^\dagger = \mathcal{V}_0^{-1/2}\mathbf{X}^\top(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^\top)^{-2}\mathbf{X}\mathcal{V}_0^{-1/2}$, and $\xi(\delta) = p + 2\sqrt{p\log(\frac{1}{\delta})}$.

Notice that in order to balance the source of the error due to bias and variance with high probability, we need to match $\sigma\sqrt{\xi(\delta)} \approx \frac{\nu}{\sqrt{\lambda}}$. Repeating the queries in $\mathbf{X}$ T times reduces σ by $1/\sqrt{T}$ but leaves the bias ν, as well as $\mathbf{W}_\dagger$, unchanged. This is because with the interpolation estimator the noisy repeated queries are averaged, which can be interpreted as a reduction in variance. Hence, by balancing $\nu/\sqrt{\lambda}$ with $\sigma\sqrt{\xi(\delta)}/\sqrt{T}$, we balance the bias and variance such that they are of the same magnitude. It does not make sense to repeat measurements more times if the bias dominates the error of estimation. A detailed example of estimating the gradient with fixed bias is given in Sec. 7.2.

We derive confidence sets for the regularized estimator of Eq. (5), albeit without the relative bias.

Proposition 2 (Regularized estimate – Fixed Design)**.** *Under the model in Eq.* (1) *with T data point evaluations, let $\hat{\theta}_\lambda$ be the regularized estimate as in* (5)*. Then $\mathrm{P}\left(\left\|\mathbf{C}(\hat{\theta}_\lambda - \theta)\right\|_{\mathbf{W}_\lambda} \geq \sqrt{\xi(\delta)} + 1\right) \leq \delta$, where $\mathbf{W}_\lambda$ in Eq.* (6)*, and $\xi(\delta) = p + 2\sqrt{p\log(\frac{1}{\delta})}$.*

Notice that, since the regularized estimator is designed to balance the bias and variance automatically, we do not need to specifically control the bias, which is contained within the information matrix $\mathbf{W}_\lambda$. Despite this elegant property, the regularized estimator involves a more challenging analysis. The main motivation to study the interpolation estimator is to understand the l_2 error as we show in Section 6.

5 Adaptive Design and Confidence Sets

To provide confidence sets for adaptively collected data, we need to project the data in $\mathbf{X}$ onto $\mathbf{C}$, where we will denote the projection by $\mathbf{Z}$ further on. With the data projected, we can reason about the reduction of the uncertainty of $\mathbf{C}\theta$ for each point separately, since to each $\Phi(x_i)$ we can associate a unique z_i in $\mathbb{R}^p$.

Definition 2 (Projected data)**.** Let $z(x) \in \mathbb{R}^p$, *projected data*, be a vector field s.t. $\Phi(x)\mathcal{V}_0^{-1/2} = z(x)\mathbf{C}\mathcal{V}_0^{-1/2} + j(x)$, where $x \in \mathcal{S} \subseteq \mathcal{X}$ and $j(x) \in \mathcal{H}_\kappa$, $|\mathcal{S}| = n$, such that $\mathbf{C}\mathcal{V}_0^{-1/2} j(x) = 0$.

Classically, the adaptive confidence sets are understood only for $\mathbf{C} = \mathbf{I}$, i.e., the identity (Abbasi-Yadkori et al., 2011). In fact, we can always derive confidence sets for $\mathbf{C}\theta$ from confidence sets for θ, as they only project the ellipsoid to a smaller dimensional space. However, their resulting size may be unnecessarily large, as the confidence parameter scales as $\mathcal{O}(\dim(\mathcal{H}_k))$ in general (see Figure 1 for a visual example).

The martingale analysis of Abbasi-Yadkori et al. (2011) and de la Peña et al. (2009) specifically assumes that information matrix $\mathbf{V}_t$, (where $\mathbf{C} = \mathbf{I}$) can be additively decomposed to information matrices due to a single evaluation $\mathbf{V}_t = \sum_{i=1}^t \Phi(x_i)\Phi(x_i)^\top$ each at a different time. With the matrix $\mathbf{W}_{\lambda,t}$, this additive decomposition is not always possible. Therefore, to utilize the martingale analysis, which requires this additive property, we consider a different information matrix, which upper bounds $\mathbf{W}_\lambda$. The information matrix we use, $\mathbf{\Omega}_\lambda$, is constructed from the projections $z(x_i)$, which have the necessary additive property. It gives rise to confidence sets, where, under the ellipsoidal norm, their size scales as $\Theta(p)$. The estimation error depends on $\mathbf{\Omega}_\lambda$ still, but under this norm, the confidence parameter and information matrix are decoupled in a similar way as for the fixed design.

Theorem 3 (Ridge estimate – Adaptive Design)**.** *Under the regression model in Eq.* (1) *with t adaptively collected data points, let $\hat{\theta}_{\lambda,t}$ be the regularized estimate as in* (5)*. Further, assume that $\mathbf{Z}$ is as in Def. 2 where $\mathbf{X}_t = \mathbf{Z}_t\mathbf{C} + \mathbf{J}_t\mathcal{V}_0^{1/2}$. Then for all $t \geq 0$,*

$$\left\|\mathbf{C}(\hat{\theta}_t - \theta)\right\|_{\mathbf{\Omega}_{\lambda,t}} \leq \sqrt{2\log\left(\frac{1}{\delta}\frac{\det(\mathbf{\Omega}_{\lambda,t})^{1/2}}{\det(\lambda\mathbf{S})^{1/2}}\right)} + 1 \tag{9}$$

with probability $1-\delta$, where $\mathbf{\Omega}_{\lambda,t} = \frac{1}{\sigma^2}\mathbf{Z}_t^\top\mathbf{Z}_t + \lambda\mathbf{S}$ and $\mathbf{S} = (\mathbf{C}\mathcal{V}_0^{-1}\mathbf{C}^\top)^{-1}$.

The matrix $\mathbf{Z}_t$ can be calculated by solving a least-squares problem (projection), whereupon $\mathbf{C}\mathcal{V}_0^{-1/2}\mathbf{J}_t = 0$ as needed by Definition 2. Notice that, on one hand, the above confidence parameter grows only when $\mathbf{Z}$ is large, but at the same time, the ellipsoid shrinks only in that case as well. Also, the ellipsoid above is necessarily smaller than the one with the information matrix $\mathbf{W}_\lambda$ as $\mathbf{W}_\lambda \preceq \mathbf{\Omega}_\lambda$ (Lemma 3 in Appendix). This means we can use the same confidence parameter to give a bound for $\|\cdot\|_{\mathbf{W}_{\lambda,t}}$. Notice that the estimator is the same as before, i.e., $\hat{\theta}_\lambda$, using $\mathbf{X}$ to define the regression, only the information matrix changes. We also present a visual comparison in Figure 1 on a simple example, showing that our fixed and adaptive sets are tighter than projected non-asymptotic sets. Using the shorthand $\beta_t(\delta)$ to define the confidence parameter in (9), we can in fact bound the error as $\|\mathbf{C}(\theta_{\lambda,t} - \theta)\|_2 \leq \lambda_{\min}(\mathbf{\Omega}_{\lambda,t})^{-1/2}\beta_t(\delta)$. It depends on $m = \dim(\mathcal{H}_\kappa)$ only via $\mathbf{\Omega}_{\lambda,t}$. The dependence of the confidence parameter can be at most $\Theta(\sqrt{\log(1+T)p})$, see Lemma 4 in Appendix B. The value of $\lambda_{\min}(\mathbf{\Omega}_{\lambda,t})^{-1}$ depends on the geometry of the set $\{\Phi(x)|x \in \mathcal{X}\}$ and the projection operator $\mathbf{C}$ as we show in Section 6.3. Note that the lower bound of (Szepesvari and Lattimore, 2020, Ex. 20.2.3) does not apply here, since it makes a statement only about the information matrix $\mathbf{W}_\lambda$.

6 Convex Relaxations, Geometry and Dimensions

Suppose we are given a candidate set of experiments, i.e., a unique subset of evaluations $\mathcal{S} \subset \mathcal{X}$, $|\mathcal{S}| = n$ and a budget T of total queries. We seek an allocation $\mathbf{X}$, where the rows of $\mathbf{X}$ contain potentially repeated evaluations from $\mathcal{S}$. How many times should we repeat each experiment in order to find $\max_{\mathbf{X}} f(\mathbf{W}(\mathbf{X}))$? To address this, experimental design literature relaxes this discrete optimization problem, and optimizes over fractional allocation $\eta \in \Delta^n$, where the number of repetitions for $\Phi(x_i)$ is recovered by rounding $\lceil \eta_i T \rceil$. With this interpretation, the objectives in Sec. 3.2 can be written as

$$\eta^* = \arg\max_{\eta \in \Delta^n} \left[f(\mathbf{W}_\dagger(\mathbf{D}(\eta)^{1/2}\mathbf{X}_\mathcal{S})) = f(\mathbf{C}(\mathbf{V}_0^{-1}\mathbf{X}_\mathcal{S}^\top\mathbf{D}(\eta)\mathbf{X}_\mathcal{S}\mathbf{V}_0^{-1})^\dagger\mathbf{C}^\top) \right], \tag{10}$$

where $\mathbf{D}(\eta)$ is the diagonalization operator that produces a diagonal matrix with vector η on the diagonal and $\mathbf{X}_\mathcal{S}$ contains non-repeated elements in $\mathcal{S}$. We have stated the problem above for the interpolation estimator and $\dim(\mathcal{H}_k) < \infty$, but it naturally generalizes to kernelized estimators, albeit in a less concise form (see Appendix C).

6.1 Optimizing allocations: experiment design algorithms

Given a subset $\mathcal{S}$, the problem (10) can be approximately solved using either convex optimization methods or a greedy algorithm. A comprehensive review of methods constructing designs η^* and rounding techniques to get $\mathbf{X}$ is *beyond the scope of this work*, and *not the core issue* addressed in this work. We briefly review two versatile approaches for completeness (more details in Appendix C).

Greedy selection Firstly, one can greedily maximize the scalarized information matrix with the update rule $\eta_{t+1} = \frac{t}{t+1}\eta_t + \frac{1}{1+t}\delta_t$, $\delta_t = \arg\max_{x\in\mathcal{X}} f\left(\mathbf{W}_\lambda\left(\frac{t}{t+1}\eta_t + \frac{1}{t+1}\delta_x\right)\right)$, where f refers to the scalarization and δ_x to the discrete measure corresponding to the feature map $\Phi(x)$. Due to the form of the update rule, $t\eta_t$ is always an integer.

6.2 Convex optimization

Alternatively, convex optimization can provably solve the problem to optimality. Specifically, we look for an allocation using convex optimization methods $\max_{\eta\in\Delta^n} f(\mathbf{W}_\circ(\mathbf{D}(\eta)^{1/2}\mathbf{X}))$ where $\circ \in \{\dagger, \lambda\}$. Care needs to be taken when selecting $\mathcal{S}$, as we discuss in Appendix C. The most common algorithms for ODE problems are the Frank-Wolfe algorithm (Todd, 2016) and mirror descent algorithm (Silvey et al., 1978). Additionally, semi-definite reformulation of the above optimization problems can be given as we shown in Appendix C.3. Optimal designs found via convex optimization need to be rounded in practice. State-of-art rounding techniques are discussed by Allen-Zhu et al. (2017) and Camilleri et al. (2021) for finite and infinite dimensional spaces, respectively. The exhaustive cover of rounding techniques is out-of-scope of the current work.

6.3 Estimation error and its dimension dependence

If we were to bound the squared error of estimation in high probability, the importance of the scalarization $\lambda_{\min}(\mathbf{W}_\dagger)$ becomes apparent. Using the Cauchy Schwarz inequality, we get

$$\left\|\mathbf{C}(\theta-\hat{\theta})\right\|_2 \leq \lambda_{\min}(\mathbf{W}_\dagger)^{-1/2}\left\|\mathbf{C}(\theta-\hat{\theta})\right\|_{\mathbf{W}_\dagger} \leq \sqrt{\frac{\lambda_{\min}(\mathbf{W}_\dagger(\eta^*))^{-1}}{T}}(\sigma\sqrt{\xi(\delta)}+\nu/\sqrt{\lambda}),$$

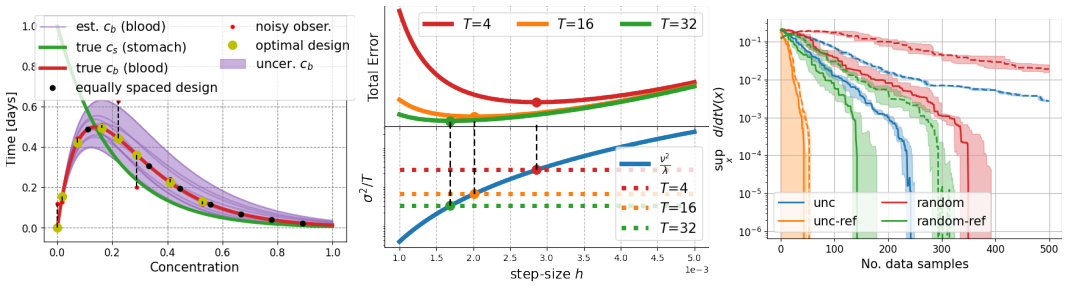
where the term due to Proposition 1 scales as $\mathcal{O}(p)$ when properly balanced. Using the optimal allocation η^*, the number of repetitive evaluations is equal to $\eta^* T$. Inverting the relation above yields a query complexity of the order $T \approx \mathcal{O}(\frac{p}{\epsilon^2}\lambda_{\min}(\mathbf{W}_\dagger(\eta^*)^{-1}))$. The optimal value $\lambda_{\min}(\mathbf{W}_\dagger(\eta^*))$ represents a problem dependent quantity that measures the *difficulty of estimation* and captures the geometry of the set $\{\Phi(x)|x\in\mathcal{S}\subset\mathcal{X}\}$. It cannot be bounded in general, but it has an elegant geometric interpretation. In particular, it corresponds to the square inverse of the *diameter of the largest inscribed ball in the convex hull of symmetrized* $\{\Phi(x)|x\in\mathcal{S}\}$ *in the range of* $\mathbf{C}$ (Pukelsheim and Studden, 1993).

At first glance, calculating this quantity might seem complicated, but with an example it is apparent. For example, if $\Phi(x) = x$ s.t. $\|x\|_2^2 \leq 1$ (unit l_2 ball in $\mathbb{R}^m$) and $\mathbf{C} = v$, where v is a unit vector ($p = 1$), the inverse diameter of largest ball we can inscribe in direction of v inside $\|x\|_2^2 \leq 1$ equates to 1, which is independent of d. This is not surprising, since it represents an "easy" design space, where for any direction v one can find an action x aligned with that coincides with the optimal design. This is true even if $\mathbf{C}$ has more rows. On the other hand, if we assume evaluation in the l_1 ball $\|x\|_1 \leq 1$ and $\mathbf{C}$ is a vector of ones (again $p = 1$). Then the diameter of the inscribed ball is proportional to the inverse square height of a simplex, namely, $\frac{1}{m}$, despite $p = 1$. Hence, despite the confidence parameter being $\mathcal{O}(1)$, the complexity depends primarily on the geometry of this set.

To give a more exotic example, if $\mathbf{C} = \nabla_x\Phi(x)$, and the design space are points with fixed length steps in all unit derection $\{x|\Phi(x\pm e_i h)\}$, where h is the stepsize and e_i are principal vectors in $\mathbb{R}^d$, we can show that $\lambda_{min}(\mathbf{W}_\dagger)^{-1} \leq dh + \mathcal{O}(h^2)$, leaving the overall complexity to learn a gradient to scale with d instead of the dimensionality of the RKHS which can be infinite. All formal proofs and references are provided in Appendix D.1.

7 Applications

We now discuss concrete applications that benefit from our contributions. Details of the experiments, and further applications, e.g., in statistical contamination, can be found in Appendices E and F. We provide information about optimization and rationale in choosing the κ, and $\mathcal{V}_0$ in C and F respectively.



(a) Pharmacokinetics (b) Gradient estimation (c) Controller stability

Figure 2: Experiments: **a)** *Pharmacokinetics*. We compare the equally spaced design (black) with the optimized design (yellow). The optimized design mimics the classical pharmacokinetics approach of spreading the initial measurements more densely after the initial dose (Gabrielsson and Weiner, 1995). For us, this rule elegantly emerges from first principles. In general, these trajectories follow a decaying pattern as the examples in light purple. The uncertainty in c_b due to the unknown dynamics γ is depicted in shaded region. **b)** *Gradient estimation*. The upper plot shows the total error of $\nabla f(x)$ we can certify with high probability as a function of the step-size h for a finite difference design. The minima of these errors exactly correspond to step-sizes derived from the bias-variance trade-off we proposed for the interpolation estimator. With increasing T, noise can be reduced more, and hence bias of the design needs to decrease accordingly, hence the decrease in the step size (e.g., red→green). **c)** *Stability:* We report the upper bound on the Lyapunov function in the whole operating domain, as a function of data points. The color coding (explained in the text) represents different data acquisition algorithms while dashed lines correspond to the confidence sets from prior work. Our confidence sets (solid) provide a tighter upper bound as a function of the number of data points, and allows faster termination, which happens when the upper bound is zero.

7.1 Linear bandits with finite number of arms

A special but important theoretical consequence of our adaptive confidence set is an improved regret bound for the upper confidence bound (UCB) algorithm (Auer, 2002) for linear bandits when the set of queries $\mathcal{X}$ is finite. The queries are known as actions in the bandit literature (Szepesvari and Lattimore, 2020). The UCB algorithm is a procedure which iteratively queries the action according to $x_t = \arg\max_{x\in\mathcal{X}} \max_{\vartheta\in C_{t-1}} \vartheta^\top \Phi(x)$, where C_t is confidence set for θ in round t. Different to this, in our analysis of the UCB algorithm, we use the confidence sets due to Theorem 3, which are confidence sets for the projections $\Phi(x)^\top\theta$ directly. We construct $|\mathcal{X}|$ confidence sets, one for each of the linear functionals $\mathbf{C}_x = \Phi(x)$, and denote them by $\mathrm{cf}_t(x)$. The UCB algorithm then just becomes $x_t = \arg\max_{x\in\mathcal{X}} \max_{y\in\mathrm{cf}_t(x)} y$. We now analyze its cumulative regret $R_T := \sum_{t=1}^T (\Phi(x^*) - \Phi(x_t))^\top\theta$, where $x^* = \arg\max_{x\in\mathcal{X}} \Phi(x)^\top\theta$, and θ is the true unknown *pay-off vector* of the bandit game.

Theorem 4. *Let $\theta \in \mathbb{R}^d$ be an unknown pay-off vector, and a set of actions $x \in \mathcal{X}$ such that $|\mathcal{X}| < \infty$. Then the cumulative regret of the UCB algorithm is bounded by*

$$R_T \leq \mathcal{O}(\sqrt{Td\log(T(|\mathcal{X}|+1)/\delta)}) \text{ with } 1-\delta \text{ probability}.$$

Thus, our adaptive confidence sets for linear functionals yield a regret bound that is a factor of $\sqrt{d}$ tighter than the standard bound. Note that this does not break the known lower bound $\mathcal{O}(d\sqrt{T})$ for $\mathcal{X}$ being the unit ball in $\mathbb{R}^d$ from Rusmevichientong and Tsitsiklis (2010), as we gain a logarithmic dependence on $|\mathcal{X}|$. For an ϵ covering of the extreme points of the unit ball in $\mathbb{R}^d$, one needs $|\mathcal{X}| \propto (1/\epsilon)^d$ points, leading to a regret scaling as $\mathcal{O}(d\sqrt{T}\log(1/\epsilon))$. Choosing $\epsilon = 1/T$ would, for a fixed T, recover the lower bound up to logarithmic terms. Using standard *doubling-trick* techniques, one can provide anytime result. However, for different query sets (such as the l_1 ball), whose extreme points (i.e., number of vertices) are finite (equal to d), these results can vastly improve the regret bounds. There are alternative algorithms based on arm-elimination, which achieve this regret bound, e.g., the algorithm of Szepesvari and Lattimore (2020), the SupLinRel algorithm of Auer (2002), and the algorithm of Valko et al. (2014). The proof of Theorem 4 is deferred to Appendix G, but it is a trivial application of previous results.

The above technique can be extended to infinite dimensional RKHS, where the role of dimension is played by, γ_T called maximum information gain (Srinivas et al., 2010), demonstrated in Appendix

G. The remaining challenge is to bound the ϵ-covering of extreme points of the set $\{\Phi(x)|x \in [-1,1]^d\} \subset \mathcal{H}_k$, where $[-1,1]^d$ is the continuous action set. If all evaluation operators have $\|\Phi(x)\|_k = 1$, for all $x \in [-1,1]$, such as for Matérn kernels (Mutný and Krause, 2018), we can use covering numbers of unit balls in these spaces as the upper bound on this number. These spaces are known to be isometrically isomorphic to Sobolev spaces (Wendland, 2004), and the bounds on covering numbers of unit balls are known to be of order $(\frac{1}{\epsilon})^{2d/s}$, where $s \in \mathbb{N}$, $s > d/2$, refers to regularity of the Sobolev space (Cucker and Smale, 2002). This yields a bound on the cumulative regret of the order $\mathcal{O}(\sqrt{C(d)\gamma_T T}\log T)$, where $C(d)$ can depend on dimension. The results in this paper, and the above technique provide an important tool for resolving the open problem about regret rates of kernelized bandit algorithms (Vakili et al., 2021), and suggest that, at least for Sobolev spaces, the bounds can be improved from the worst case $\mathcal{O}(\gamma_T\sqrt{T})$ to one scaling with $\mathcal{O}(\sqrt{C(d)\gamma_T T})$.

7.2 Gradient maps

Gradients of any order are linear operators. We can express the gradient at x as dimension-wise evaluation of the following operator $\nabla_x(\Phi(x)^\top\theta) = (\nabla_x\Phi(x))^\top\theta =: \mathbf{C}\theta$. Clearly, estimability is nearly always impossible, since any evaluation infinitesimally away from x will be insufficient to eliminate bias. Thus our estimates will invariably be biased for any function with infinite Taylor expansion. Yet, while estimating the gradient from evaluations very close to the original x leads to low bias, it at the same time increases the variance, since the difference in the functional value between the two point evaluations is very small compared to the noise magnitude. Given a finite budget or desired accuracy, the best we can do is to find the best design with optimal bias-variance trade-off given the kernel κ, budget T, and noise variance σ. We consider a class of parametrized finite difference designs $\{\Phi(x \pm he_i)\}$, where h is the stepsize and e_i are unit vectors. In Fig. 2b, we plot the total error with high probability with the budget T as a function of the step-size h. Notice that the lowest error occurs exactly when the variance of observations scaled with the confidence parameter is equal to the relative ν (two lines cross).

7.3 Learning linear ODE solutions and their parameters

A solution to a linear ordinary differential equation (ODE) u satisfies $\frac{d}{dt}u(t) = \mathbf{M}u(t) + s(t)$, where $\mathbf{M}$ is a linear operator and $s(t)$ is the non-homogeneous term. Assume that the solution to the equation $u(t) = \Phi(t)^\top u$ is a member of a Hilbert space $\mathcal{H}_\kappa$, then $\mathbf{T}u := \left(\frac{d}{dt} - \mathbf{M}(t)\right)\Phi(t)^\top u = s(t)$ for $t \in [t_0, t_1]$. Hence, the differential equation becomes a *linear constraint* for estimating u from samples. In fact, due to differential equations being fully specified by initial conditions, the only unknowns are due to initial conditions. To reveal the linear functional here, one needs to consider the solution to the differential equation, which can be written as $u = \mathbf{T}^\dagger s(t) + \mathbf{C}^\top v$, where $v \in \mathcal{H}_\kappa \backslash \mathrm{span}(\mathbf{T})$ belongs to the null space of $\mathbf{T}$. In this case, we span it with rows of $\mathbf{C}$. Consequently, the unknown element v can be found as $\mathbf{C}(u - \mathbf{T}^\dagger s)$. Thus, what needs to be estimated from samples, is the linear projection $\mathbf{C}u$, since $\mathbf{C}\mathbf{T}^\dagger s$ is known a priori. In Appendix F, we discuss implementation and calculation of the operator $\mathbf{C}$ on a discretized domain. Also note that, $\mathbf{Z}$ itself induces a Hilbert space $\mathcal{H}_z$ containing the solutions u (González et al., 2014). However, as we will see, for robust designs it is sometimes convenient to pick a larger Hilbert space $\mathcal{H}_\kappa$ s.t. $\mathcal{H}_z \subset \mathcal{H}_\kappa$ and then apply the differential equation constraint.

Robust Design As a specific example, consider an example of a pharmacokinetic model capturing the concentration of a medication in blood and stomach via differential equations: $(d/dt)c_s = -ac_s$ and $(d/dt)c_b = bc_s - dc_b$ for $t \in [t_0, t_1]$, where c_s and c_b are the concentration in stomach and blood respectively. The goal of this analysis is to infer $\gamma = (a, b, d)$ from the measurements of the blood concentration levels with fixed, but perhaps noisy, initial conditions (Gabrielsson and Weiner, 1995). To apply the above procedure, we need a fixed differential operator $\mathbf{T}_\gamma$ to define the operator $\mathbf{C}_\gamma$. However, γ is itself unknown. Instead, we can give a generally plausible set of $\gamma \in \Gamma$, initial conditions of concentration in blood $c_b(0) = 0$, and norm constraints on the initial stomach concentration $(c_s(0) - c_{\text{dose}})^2 \leq \lambda^{-1}$ (prior $\mathcal{V}_0$) in order to apply our framework. We use the squared exponential kernel to embed trajectories and consider robust A-optimal design (as in Sec. 3.2) by maximizing the worst case metric $\inf_{\gamma\in\Gamma} 1/\operatorname{Tr}(\mathbf{C}_\gamma \mathbf{V}_\lambda^{-1}\mathbf{C}_\gamma^\top)$ with the regularized estimator. This way, the design is appropriate for any γ. After estimating the trajectory, we can use the estimated trajectories (given γ) to optimize for γ via the maximum likelihood. Fig. 4a presents the concentrations c_b and c_s with their estimates and uncertainties due to the unknown γ. We show the equally spaced (black) and optimized designs (yellow). The more accurately we can infer the trajectory, the more accurately we can estimate γ, which we quantitatively show in Appendix E in Fig. 4b, where we decrease the MSE of estimating γ by a factor of 10 in comparison to equally spaced design.

7.4 Sequential Design: Certifying Lyapunov Stability

Consider a non-linear system with $x \in \mathbb{R}^d$ such that $\frac{d}{dt}x(t) = f(x(t),t) + u(x(t),t) = \mathbf{A}\Phi(x(t),t) + u(x(t),t)$, where the rows of $\mathbf{A}_i \in \mathcal{H}_\kappa$ for $i \in [d]$ model the system dynamics and $\Phi(x(t),t)$ are known evaluation functionals of $\mathcal{H}_\kappa$. We assume that the control laws $u(x(t),t)$ can be written as $u(x(t),t) = \mathbf{B}\Phi(x(t),t)$. We want to understand whether a given a control law $\mathbf{B}$ stabilizes the above system. A common approach is to create an estimate $\hat{\mathbf{A}}$ of $\mathbf{A}$ from data samples of trajectories, and use the controller $\mathbf{B} = \hat{\mathbf{A}} - \mathbf{P}$, where $\mathbf{P}$ is often a simple reverting law with a known gain that compensates the imprecise estimation of $\mathbf{A}$ with $\hat{\mathbf{A}}$, see below for an example. With this choice of u, the system can be written as $\frac{d}{dt}x(t) = (\mathbf{A} - \hat{\mathbf{A}} - \mathbf{P})\Phi(x(t),t) = (\tilde{\mathbf{A}} - \mathbf{P})\Phi(x(t),t)$, where the $\tilde{\mathbf{A}} = \mathbf{A} - \hat{\mathbf{A}}$ is the residual error in estimating $\mathbf{A}$. We can certify the stability of the resulting system using a *known* quadratic Lyapunov function as commonly done in control theory, in this case, using $V(x,t) = (x(t) - x_{\text{ref}}(t))^\top \mathbf{\Sigma}(x(t) - x_{\text{ref}}(t))$. Classical stability theory dictates that if the total time derivative of V,

$$dV/dt = (x(t) - x_{\text{ref}}(t))^\top \mathbf{\Sigma}(\tilde{\mathbf{A}} - \mathbf{P})\phi(x(t),t) + \partial V/\partial t, \tag{11}$$

is negative for all $x(t) \in O$ then we can guarantee stability in the operating region of $x \in O$ (Khalil, 2002). The above condition defines a linear operator $\mathbf{C}_x = (x - x_{\text{ref}})^\top \mathbf{\Sigma} \cdot \phi(x)$ operating on the unknown $\tilde{\mathbf{A}}$ for each $x \in O$. The linearity is best seen with vectorization as $\theta = \text{vec}(\mathbf{A})$, using the shorthand $z(t) = x(t) - x_{\text{ref}}(t)$, $z^\top \mathbf{\Sigma}\mathbf{A}\phi(x) = \text{vec}(\mathbf{\Sigma} z \phi(x)^\top)^\top \text{vec}(\mathbf{A}) = \mathbf{C}_x\theta$. The operator is parametrized by x, $\mathbf{C}_x$ for each $x \in O$. Even for continuous domains $x \in O$, $\mathbf{C}_x$ usually has low-rank structure, which can be calculated and depends on the size O and the size $\mathcal{H}_\kappa$. If the rank of the operator is small (e.g., the operating space is small) then we can certify negativity of Eq. (11) faster than learning the whole $\mathbf{A}$. In other words, we reduce uncertainty only where we need to as in the seminal work of Berkenkamp et al. (2016). We can sequentially query data points from x and check whether $\mathbf{C}_x\theta \leq 0$ for all x.

Consider a two dimensional nonlinear system from Lederer et al. (2021),

$$\frac{dx}{dt} = x + \frac{1}{1 + \exp(-2x_1)} \begin{pmatrix} 1 \\ -1 \end{pmatrix} + 0.5 \begin{pmatrix} \sin(\pi x_2) \\ \cos(\pi x_1) \end{pmatrix} + u.$$

The reverting controller is $\mathbf{P}\Phi(x(t),t) = -K(x - x_{\text{ref}}(t)) + (d/dt)x_{\text{ref}}$, where the reference trajectory corresponds to a circle $x_{\text{ref}}(t) = (\sin(t), \cos(t))$. The Lyapunov function is $V = (x(t) - x_{\text{ref}}(t))^\top (x(t) - x_{\text{ref}}(t))$. We assume that we can set the system to an initial condition $x(0)$ and observe a noisy observations of the state $y(0 + \Delta)$ at rapid sampling times Δ. From these, we create a derivative oracle, $(d/dt)x(t) \approx (x(t + \Delta) - x(t))/\Delta$, a common approach in nonlinear data-driven control (Umlauft et al., 2018). We use this example to showcase our adaptive confidence sets. We follow an adaptive stopping rule, where we query new data points if negativity cannot be certified. Due to this adaptive stopping rule, the data is adaptively collected.

We compare our confidence sets (solid) with the classical confidence sets of Abbasi-Yadkori et al. (2011) (dashed) and report upper bounds on the supremum over $x \in O$ of Eq. (11) in Fig. 2c. The operating region is a "tube" around the circular reference trajectory $x_{\text{ref}}(t)$. We see that our confidence sets shrink much faster than the classical confidence sets, since we can eliminate redundant information by projecting onto $\mathbf{C}$. Fig. 2c compares random sampling of datapoints from the whole domain (random) and the operating region (random-ref) and sampling according to uncertainty in the dynamics in the whole domain (unc) and within the tube around the reference trajectory (unc-ref). As expected, focused exploration methods work much better. However, and more importantly, the tightness of our confidence sets enable much quicker stability certification (termination). Note that the classical confidence sets may even *grow* in the cases where redundant information for estimation of $\mathbf{C}$ is inserted into the estimation, e.g., with random sampling.

8 CONCLUSION

We considered the problem of learning a linear function of an element in a reproducing kernel Hilbert space. We addressed the challenging case where linear estimators incur a non-negligible bias and provided confidence sets for the two most commonly used linear estimators. We demonstrated the generality of our approach and the tightness of our confidence sets on several challenging applications. We believe our results lay important foundations for principled and efficient experiment design in complex real-world settings.

Acknowledgement

This project has received funding Swiss National Science Foundation through NFP75 and this publication was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

Abbasi-Yadkori, Y. and Szepesvari, C. (2012). *Online learning for linearly parametrized control problems*. PhD thesis, University of Alberta.

Agrell, C. and Dahl, K. R. (2020). Sequential bayesian optimal experimental design for structural reliability analysis. *arXiv preprint arXiv:2007.00402*.

Allen-Zhu, Z., Li, Y., Singh, A., and Wang, Y. (2017). Near-optimal design of experiments via regret minimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 126–135, Sydney, Australia.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Bardow, A. (2008). Optimal experimental design of ill-posed problems: The meter approach. *Computers & Chemical Engineering*, 32(1-2):115–124.

Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175.

Berkenkamp, F., Schoellig, A. P., and Krause, A. (2016). Safe controller optimization for quadrotors with gaussian processes. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 491–496. IEEE.

Betke, U. and Henk, M. (1992). Estimating sizes of a convex body by successive diameters and widths. *Mathematika*, 39(2):247–257.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Cakmak, S., Astudillo Marban, R., Frazier, P., and Zhou, E. (2020). Bayesian optimization of risk measures. *Advances in Neural Information Processing Systems*, 33.

Camilleri, R., Jamieson, K., and Katz-Samuels, J. (2021). High-dimensional experimental design and kernel bandits. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1227–1237. PMLR.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statist. Sci.*, 10(3):273–304.

Cockayne, J., Oates, C., Sullivan, T., and Girolami, M. (2017). Bayesian Probabilistic Numerical Methods. *ArXiv e-prints*, stat.ME 1702.03673.

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49.

de la Peña, V. H., Klass, M. J., and Lai, T. L. (2009). Theory and applications of multivariate self-normalized processes. *Stochastic Processes and their Applications*, 119(12):4210 – 4227.

Draper, N. R. and Smith, H. (2014). *Applied Regression Analysis*. John Wiley and Sons.

Elfving, G. (1952). Optimum Allocation in Linear Regression Theory. *Annals of Mathematical Statistics*, 23(2):255–262.

Fedorov, V. V. and Hackl, P. (1997). *Model-Oriented Design of Experiments | Valerii V. Fedorov | Springer*. Springer-Verlag New York.

Gabrielsson, J. and Weiner, D. (1995). Pharmacokinetic and pharmacodynamic data analysis. *Trends in Pharmacological Sciences*, 16(4):143.

Gaffke, N. (1987). Further characterizations of design optimality and admissibility for partial parameter estimation in linear regression. *Ann. Statist.*, 15(3):942–957.

González, J., Vujačić, I., and Wit, E. (2014). Reproducing kernel hilbert space based estimation of systems of ordinary differential equations. *Pattern Recognition Letters*, 45:26–32.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Huszár, F. and Duvenaud, D. (2012). Optimally-weighted herding is bayesian quadrature. *arXiv preprint arXiv:1204.1664*.

Khalil, H. (2002). *Nonlinear Systems*. Patience Hall.

Khamaru, K., Deshpande, Y., Mackey, L., and Wainwright, M. J. (2021). Near-optimal inference in adaptive linear regression. *arXiv preprint arXiv:2107.02266*.

Kirschner, J., Mutný, M., Hiller, N., Ischebeck, R., and Krause, A. (2019). Adaptive and safe bayesian optimization in high dimensions via one-dimensional subspaces. *ICML 2019*.

Krafft, O. (1983). A matrix optimization problem. *Linear Algebra and its Applications*, 51:137 – 142.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338.

Lederer, A., Capone, A., Beckers, T., Umlauft, J., and Hirche, S. (2021). The impact of data on the stability of learning-based control. *Conference on Learning for Dynamics and Control*.

Leykekhman, D., Vexler, B., and Walter, D. (2020). Numerical analysis of sparse initial data identification for parabolic problems. *ESAIM: Mathematical Modelling and Numerical Analysis*.

Mutný, M., Johannes, K., and Krause, A. (2020). Experimental design for orthogonal projection pursuit regression. *AAAI2020*.

Mutný, M. and Krause, A. (2018). Efficient high dimensional bayesian optimization with additivity and quadrature fourier features. In *Neural and Information Processing Systems (NeurIPS)*.

Owhadi, H. and Scovel, C. (2016). Toward machine Wald. In *Springer Handbook of Uncertainty Quantification*, pages 1–35. Springer.

Pukelsheim, F. (2006). *Optimal Design of Experiments (Classics in Applied Mathematics) (Classics in Applied Mathematics, 50)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Pukelsheim, F. and Studden, W. J. (1993). E-optimal designs for polynomial regression. *The Annals of Statistics*, pages 402–415.

Rasmussen, C. and Williams, C. (2006). Gaussian processes for machine learning, vol. 1. *The MIT Press, Cambridge, doi*, 10:S0129065704001899.

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411.

Settles, B. (2009). *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.

Shoham, N. and Avron, H. (2020). Experimental design for overparameterized learning with application to single shot deep active learning. *arXiv preprint arXiv:2009.12820*.

Silvey, S., Titterington, D., and Torsney, B. (1978). An algorithm for optimal designs on a design space. *Communications in Statistics - Theory and Methods*, 7(14):1379–1389.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. *International Conference on Machine Learning*.

Szepesvari, C. and Lattimore, T. (2020). *Bandit Algorithms*. Cambridge University Press.

Todd, M. J. (2016). *Minimum-Volume Ellipsoids: Theory and Algorithms*. SIAM, mos-siam series on optimization edition.

Umlauft, J., Pöhler, L., and Hirche, S. (2018). An uncertainty-based control lyapunov approach for control-affine systems modeled by gaussian process. *IEEE Control Systems Letters*, 2(3):483–488.

Vakili, S., Scarlett, J., and Javidi, T. (2021). Open problem: Tight online confidence intervals for rkhs elements. In Belkin, M. and Kpotufe, S., editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4647–4652. PMLR.

Valko, M., Munos, R., Kveton, B., and Kocak, T. (2014). Spectral bandits for smooth graph functions. In *International Conference on Machine Learning*, pages 46–54.

Wendland, H. (2004). *Scattered data approximation*, volume 17. Cambridge university press.

Zhang, F. (2011). *Matrix Theory: Basic Results and Techniques*. Springer Science & Business Media.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] No. Theory paper.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] It's insignificant. I just used a personal machine.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary Material: Experimental Design for Linear Functionals in Reproducing Kernel Hilbert Spaces

A Estimability results

In the following section, we either describe proof of for implication or equivalences of certain conditions studied in this work. In A.1, we show consequence of Def. 1 which is used in the proofs of confidence sets. In the subsection following it, we establish relationship to classical ODE as in (Pukelsheim, 2006) showing that our bias condition generalizes notion stemming from there.

A.1 Equivalence of bias conditions

Proposition 5. *Let $\mathbf{L}_\dagger$ be the interpolation estimator and $\mathbf{W}_\dagger$ it associated information matrix then, then*

$$\left\|(\mathbf{C}-\mathbf{L}_\dagger\mathbf{X})\mathcal{V}_0^{-1/2}\right\|_F^2 \leq \nu^2 \left\|\mathbf{L}_\dagger\right\|_F^2 \implies \left\|\mathbf{W}_\dagger^{1/2}(\mathbf{C}-\mathbf{L}_\dagger\mathbf{X})\mathcal{V}_0^{-1/2}\right\|_2^2 \leq \nu^2 \tag{12}$$

Proof. Note that $\mathbf{V}^\dagger = \mathcal{V}_0^{-1/2}\mathbf{X}^\top\mathbf{K}^{-2}\mathbf{X}\mathcal{V}_0^{-1/2}$, where $\mathbf{K} = \mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^\top$.

$$\begin{aligned}
& \text{(LHS)} \\
= \;& \mathcal{V}_0^{-1/2}(\mathbf{C}-\mathbf{L}_\dagger\mathbf{X})^\top\mathbf{W}_\dagger(\mathbf{C}-\mathbf{L}_\dagger\mathbf{X})\mathcal{V}_0^{-1/2} \\
= \;& \mathcal{V}_0^{-1/2}(\mathcal{I}-\mathcal{V}_0^{-1}\mathbf{X}^\top\mathbf{K}^{-1}\mathbf{X})^\top\mathbf{C}^\top(\mathbf{C}\mathcal{V}_0^{-1/2}\mathbf{V}^\dagger\mathcal{V}_0^{-1/2}\mathbf{C}^\top)^{-1}\mathbf{C}(\mathcal{I}-\mathcal{V}_0^{-1}\mathbf{X}^\top\mathbf{K}^{-1}\mathbf{X})\mathcal{V}_0^{-1/2} \\
= \;& \mathcal{V}_0^{-1/2}(\mathcal{I}-\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^\top\mathcal{V}_0^{-1})\mathbf{C}^\top(\mathbf{C}\mathcal{V}_0^{-1/2}\mathbf{V}^\dagger\mathbf{C}^\top)^{-1}\mathcal{V}_0^{-1/2}\mathbf{C}(\mathcal{I}-\mathcal{V}_0^{-1}\mathbf{X}^\top\mathbf{K}^{-1}\mathbf{X})\mathcal{V}_0^{-1/2} \\
= \;& (\mathcal{I}-\mathcal{V}_0^{-1/2}\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^\top\mathcal{V}_0^{-1/2})\mathcal{V}_0^{-1/2}\mathbf{C}^\top(\mathbf{C}\mathcal{V}_0^{-1/2}\mathbf{V}^\dagger\mathcal{V}_0^{-1/2}\mathbf{C}^\top)^{-1}\mathbf{C}\mathcal{V}_0^{-1/2}(\mathcal{I}-\mathcal{V}_0^{-1/2}\mathbf{X}^\top\mathbf{K}^{-1}\mathbf{X}\mathcal{V}_0^{-1/2}) \\
= \;& (\mathcal{I}-\mathcal{P})\mathcal{V}_0^{-1/2}\mathbf{C}^\top(\mathbf{C}\mathcal{V}_0^{-1/2}\mathbf{V}^\dagger\mathcal{V}_0^{-1/2}\mathbf{C}^\top)^{-1}\mathbf{C}\mathcal{V}_0^{-1/2}(\mathcal{I}-\mathcal{P})
\end{aligned}$$

Now let $\tilde{\mathbf{X}} = \mathbf{X}\mathcal{V}_0^{-1/2}$ and $\tilde{\mathbf{C}} = \mathbf{C}\mathcal{V}_0^{-1/2}$. Also notice that, $(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^\dagger = \mathcal{V}_0^{-1/2}\mathbf{X}^\top\mathbf{K}^{-2}\mathbf{X}\mathcal{V}_0^{-1/2} = \mathbf{V}^\dagger$

We can apply Theorem 19, to get

$$\tilde{\mathbf{C}}^\top(\tilde{\mathbf{C}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^\dagger\tilde{\mathbf{C}}^\top)^{-1}\tilde{\mathbf{C}} = \mathcal{V}_0^{-1/2}\mathbf{C}^\top(\mathbf{C}\mathcal{V}_0^{-1/2}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^\dagger\mathcal{V}_0^{-1/2}\mathbf{C}^\top)^{-1}\mathbf{C}\mathcal{V}_0^{-1/2} \preceq c\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \mathcal{I}\nu^2$$

Using the above result, we show that

$$\begin{aligned}
\text{(LHS)} \;&\preceq\; (\mathcal{I}-\mathcal{P})(\mathcal{V}_0^{-1/2}\mathbf{X}^\top\mathbf{X}\mathcal{V}_0^{-1/2} + \mathcal{I}\nu^2)(\mathcal{I}-\mathcal{P}) \\
&=\; \nu^2(\mathcal{I}-\mathcal{P})^2 \preceq \mathcal{I}\nu^2
\end{aligned}$$

where we have used the fact that $\mathcal{I}-\mathcal{P}$ is projection matrix with orthogonal span to $\mathcal{V}_0^{-1/2}\mathbf{X}^\top\mathbf{X}\mathcal{V}_0^{-1/2}$.

□

Proposition 6. *Bias $\left\|(\mathbf{C}-\mathbf{L}_\dagger\mathbf{X})\mathcal{V}_0^{-1/2}\right\|$ depends only on its support not the allocations.*

Proof.

Let $\mathbf{L}_\eta$ be the estimator with design η, where

$$\mathbf{L}_\eta = \mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^\top\mathbf{D}(\eta)^{1/2}(\mathbf{D}(\eta)^{1/2}\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^\top\mathbf{D}(\eta)^{1/2})^{-1} \tag{13}$$

$$= \mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^\top)^{-1}\mathbf{D}(\eta)^{-1/2} = \mathbf{L}_\dagger\mathbf{D}(\eta)^{-1/2} \tag{14}$$

But at the same time $\mathbf{D}(\eta)^{1/2}\mathbf{X}$, hence these two cancel each other.

□

A.2 Basic equivalences and relations

To relate the estimability to terms used in the ED literature, we state the definition of the *feasibility cone* as in Pukelsheim (2006). We show that our condition in Def. 1 and Pukelsheims and estimability are equivalent under classical assumptions. We use a different formulation of the relative-bias condition due to Proposition 5.

Definition 3 ((Pukelsheim, 2006))**.** Let $\mathbf{C} : \mathcal{H}_\kappa \to \mathbb{R}^k$, denote $\mathcal{A}(\mathbf{C}) = \{\mathbf{A} \in \mathcal{P}(\mathcal{H}_\kappa) : \text{range}(\mathbf{C}^\top) \subseteq \text{range}(\mathbf{A})\}$, where $\mathcal{P}(\mathcal{H}_\kappa)$ is the space of positive-definite operators on $\mathcal{H}_\kappa$. We call $\mathcal{A}(\mathbf{C})$ the feasiblity cone of $\mathbf{K}$.

This definition is sometimes used as restatement of the estimability property.

Lemma 1 (Equivalence in $\nu = 0$)**.** *Let $\mathbf{C} \in \mathbb{R}^{k\times d}$ full rank k. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$, then if $\mathbf{M} = \mathbf{X}^\top\mathbf{X}$, the following are equivalent,*

1. ***Estimability:*** *There exists $\mathbf{L} \in \mathbb{R}^{k\times n}$ s.t. $\mathbf{C} = \mathbf{L}\mathbf{X}$.*
2. ***Feasibility cone:*** $\mathbf{M} \in \mathcal{A}(\mathbf{C})$

Proof.

- (2) $\implies$ (1): There exists $\mathbf{B}$ s.t. $\mathbf{C} = \mathbf{B}\mathbf{M}$, and $\mathbf{C} = \mathbf{B}\mathbf{X}^\top\mathbf{X}$, hence we can define $\mathbf{L} = \mathbf{B}\mathbf{X}$.
- (1) $\implies$ (2): As such $\mathbf{C} = \mathbf{L}\mathbf{X}$ implies that $\text{range}(\mathbf{C}) \subseteq \text{range}(\mathbf{X}^\top) = \text{range}(\mathbf{X}^\top\mathbf{X})$.

□

Definition 4 (Projected data)**.** Let $z(x) \in \mathbb{R}^p$ be a vector field s.t. $\Phi(x)\mathcal{V}_0^{-1/2} = z(x)\mathbf{C}\mathcal{V}_0^{-1/2} + j(x)$, where $x \in \mathcal{S} \subseteq \mathcal{X}$ and $j(x) \in \mathcal{H}_\kappa$, $|\mathcal{S}| = n$, such that $\mathbf{C}\mathcal{V}_0^{-1/2} j(x) = 0$. We call this vector field *projected data*. If in addition if $\{z(x) : x \in \mathcal{S}\}$ spans $\mathbb{R}^p$ it is said to be *approximately low-rank*.

Lemma 2. *The assumption in Definition 4 implies the assumption in Definition 1 with $\nu = \|\mathbf{J}\|_k \frac{\|\mathbf{Z}^\dagger\|_F}{\|\mathbf{L}_\dagger\|_F}$.*

Proof. Since $\mathbf{Z}$ spans whole $\mathbb{R}^k$, there exists a unique left pseudo-inverse. $\mathbf{Z}^\dagger\mathbf{X}\mathcal{V}_0^{-1/2} = \mathbf{C}\mathcal{V}_0^{-1/2} + \mathbf{Z}^\dagger\mathbf{J}$.

$$(\mathbf{C} - \mathbf{Z}^\dagger\mathbf{X})\mathcal{V}_0^{-1/2} = \mathbf{Z}^\dagger\mathbf{J}$$

Now taking the Frobenius norm,

$$\left\|(\mathbf{C} - \mathbf{Z}^\dagger\mathbf{X})\mathcal{V}_0^{-1/2}\right\|_F \leq \left\|\mathbf{Z}^\dagger\mathbf{J}\right\|_F = \|\mathbf{J}\|_F \left\|\mathbf{Z}^\dagger\right\|_F$$

So,

$$\frac{\left\|(\mathbf{C} - \mathbf{Z}^\dagger\mathbf{X})\mathcal{V}_0^{-1/2}\right\|_F}{\|\mathbf{Z}^\dagger\|_F} \leq \|\mathbf{J}\|_F$$

Now, since pseudo-inverse minimizes the LHS, we know that,

$$\left\|(\mathbf{C} - \mathbf{L}_\dagger\mathbf{X})\mathcal{V}_0^{-1/2}\right\|_F \leq \left\|(\mathbf{C} - \mathbf{Z}^\dagger\mathbf{X})\mathcal{V}_0^{-1/2}\right\|_F \leq \|\mathbf{J}\|_k \frac{\left\|\mathbf{Z}^\dagger\right\|_F}{\|\mathbf{L}_\dagger\|_F}\|\mathbf{L}_\dagger\|_F$$

□

B Confidence Sets: Proofs

This section includes proofs for the concentration results presented in the main text. In Section, B.4 we restate and prove results from Mutný et al. (2020) in the current notation for easier comparison.

B.1 Fixed Design with Interpolator

The following Theorem tries to qualify whether a design contains sufficient information to reduce confidence on a specific subspace $\mathbf{C}$ to the desired error.

Theorem 7 (Interpolation (Fixed Design))**.** *Under regression model in Eq. (1) with T data point evaluations, let $\hat{\theta}$ be the estimate as in (3). Further let $\mathbf{X}$ be s.t. satisfy the approximate estimability condition of Def. 1 with $T \geq k$. Then,*

$$\mathrm{P}\left(\left\|\mathbf{C}(\hat{\theta}-\theta)\right\|_{\mathbf{W}} \geq \sigma\sqrt{\xi(\delta)} + \frac{\nu}{\sqrt{\lambda}}\right) \leq \delta, \tag{15}$$

where $\mathbf{W} = (\mathbf{C}\mathcal{V}_0^{-1/2}\mathbf{V}^{\dagger}\mathcal{V}_0^{-1/2}\mathbf{C}^{\top})^{-1}$, $\mathbf{V}^{\dagger} = \mathcal{V}_0^{-1/2}\mathbf{X}^{\top}\mathbf{K}^{-2}\mathbf{X}\mathcal{V}_0^{-1/2}$, and $\xi(\delta) = p + 2\sqrt{p\log(\frac{1}{\delta})}$.

Proof. The pseudo inverse estimate of $\mathbf{C}\theta$ is $\mathbf{C}\hat{\theta} = \mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^{\top}(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^{\top})^{-1}y$, where $\mathbf{X} \in \mathbb{R}^{T\times m}$.

$$\begin{aligned}
\left\|\mathbf{C}(\hat{\theta}-\theta)\right\|_{\mathbf{W}}^2 &= \left\|\mathbf{C}(\mathcal{V}_0^{-1}\mathbf{X}^{\top}(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^{\top})^{-1}(\mathbf{X}\theta+\epsilon)-\theta)\right\|_{\mathbf{W}}^2 \\
&= \left\|\mathbf{C}(\mathcal{V}_0^{-1}\mathbf{X}^{\top}(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^{\top})^{-1}\mathbf{X}-\mathcal{I})\theta + \mathbf{C}\mathcal{V}_0^{-1}(\mathbf{X}^{\top}(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^{\top})^{-1}\epsilon)\right\|_{\mathbf{W}}^2 \\
&\leq \left\|\mathbf{C}(\mathcal{V}_0^{-1}\mathbf{X}^{\top}(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^{\top})^{-1}\mathbf{X}-\mathcal{I})\mathcal{V}_0^{-1/2}\mathcal{V}_0^{1/2}\theta\right\|_{\mathbf{W}}^2 \\
&\quad + \left\|\mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^{\top}(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^{\top})^{-1}\epsilon\right\|_{\mathbf{W}}^2 \\
&\leq \left\|\mathbf{W}^{1/2}\mathbf{C}(\mathcal{V}_0^{-1}\mathbf{X}^{\top}(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^{\top})^{-1}\mathbf{X}-\mathcal{I})\mathcal{V}_0^{-1/2}\right\|_2^2 \|\theta\|_{\mathcal{V}_0} \\
&\quad + \left\|\mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^{\top}(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^{\top})^{-1}\epsilon\right\|_{\mathbf{W}}^2 \\
&= \lambda^{-1}\nu^2 + \left\|\mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^{\top}(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^{\top})^{-1}\epsilon\right\|_{\mathbf{W}}^2
\end{aligned}$$

where the second to last line we used the relative-bias assumption, according to the Proposition 5, where we use the Def. 1 and $T \geq k$. The operator $\mathcal{I}$ is identify operator on $\mathcal{H}_k$.

If ϵ_G were Gaussially distributed then, $\mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^{\top}(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^{\top})^{-1}\epsilon_G$ is distributed as $\mathcal{N}(0, \mathbf{C}\mathbf{V}^{\dagger}\mathbf{C}^{\top})$. Likewise, $q_G = \mathbf{W}^{1/2}\mathbf{C}\mathcal{V}_0^{-1/2}\mathbf{X}^{\top}(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^{\top})^{-1}\epsilon_G$ is distributed as $\mathcal{N}(0, \mathbf{I}_k)$. Since ϵ is sub-Gaussian q needs to have tails of distribution which are below the tails of k-dimensional gaussian. In other words $P(|q_i| \leq t) \leq P(|q_{G_i}| \leq t)$. As such, $P(q_i^2 \leq t^2) \leq P(|q_{G_i}|^2 \leq t)$, where $q_{G_i}^2$ is chi-squared distributed. Using concentration for chi-squared random variables, $\mathrm{P}\big(\|z\|_2^2 \geq \sigma^2(p + 2\sqrt{px})\big) \leq \exp(-x)$ as in Laurent and Massart (2000). Rearranging the expression leads to the result with $\xi(\delta) = (p + 2\sqrt{p\log(\frac{1}{\delta})})$.

The last final inequality in the statement of the probability follows from taking a square root and triangle inequality, which finishes the proof. □

B.2 Fixed Design with Regularized Estimator

Proposition 8 (Fixed Design Ridge Regression)**.** *Under regression model in Eq. (1) with T data point evaluations, let $\hat{\theta}_\lambda$ be the regularized estimate as in (5). Then,*

$$\mathrm{P}\left(\left\|\mathbf{C}(\hat{\theta}_\lambda-\theta)\right\|_{\mathbf{W}_\lambda} \geq \sqrt{\beta(\delta)} = \sqrt{\xi(\delta)} + 1\right) \leq \delta,$$

where $\mathbf{W}_\lambda = \sigma^{-2}(\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{C}^{\top})^{-1}$, $\mathcal{V}_\lambda^{-1} = (\mathbf{X}^{\top}\mathbf{X} + \lambda\sigma^2\mathcal{V}_0)^{-1}$. In addition, $\xi(\delta) = p + 2\sqrt{p\log(\frac{1}{\delta})}$.

Proof. Notice that invertibility of $\mathbf{W}$ is guaranteed by full rank $\mathbf{C}$ and invertibility of $\mathbf{V}_\lambda$. Using shorthand $\mathcal{V} = \mathbf{X}^\top \mathbf{X}$,

$$\begin{aligned}
\left\|\mathbf{C}(\hat{\theta}_\lambda - \theta)\right\|_{\mathbf{W}_\lambda}^2 &= \left\|\mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^\top(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^\top + \sigma^2\lambda\mathbf{I})^{-1}(\mathbf{X}\theta + \epsilon) - \theta)\right\|_{\mathbf{W}_\lambda}^2 \\
&= \left\|\mathbf{C}\mathcal{V}_0^{-1/2}(\mathcal{V}_0^{-1/2}\mathbf{X}^\top\mathbf{X}\mathcal{V}_0^{-1/2} + \sigma^2\lambda\mathbf{I})^{-1}\mathcal{V}_0^{-1/2}\mathbf{X}^\top(\mathbf{X}\theta + \epsilon) - \theta)\right\|_{\mathbf{W}_\lambda}^2 \\
&= \left\|\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{X}^\top(\mathbf{X}\theta + \epsilon) - \theta)\right\|_{\mathbf{W}_\lambda}^2 \\
&= \left\|\mathbf{C}(\mathcal{V}_\lambda^{-1}\mathbf{X}^\top\epsilon + \mathcal{V}_\lambda^{-1}\mathcal{V}\theta - \theta)\right\|_{\mathbf{W}_\lambda}^2 \\
&\leq \left\|\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{X}^\top\epsilon\right\|_{\mathbf{W}_\lambda}^2 + \left\|\mathbf{C}(\mathcal{V}_\lambda^{-1}\mathbf{X}^\top\mathbf{X} - \mathcal{I})\theta\right\|_{\mathbf{W}_\lambda}^2
\end{aligned}$$

The second term,

$$\begin{aligned}
\left\|\mathbf{C}\mathcal{V}_\lambda^{-1}(\mathcal{V}_0\lambda\sigma^2)\theta)\right\|_{\mathbf{W}_\lambda}^2 &\overset{(64)}{\leq} \sigma^2\left\|\mathcal{V}_\lambda^{-1}(\mathcal{V}_0\lambda)\theta)\right\|_{\mathcal{V}_\lambda}^2 && (16) \\
&= \lambda\sigma^2\theta^\top\mathcal{V}_0\mathcal{V}_\lambda^{-1}(\mathcal{V}_0\lambda)\theta) \leq \lambda\theta^\top\mathcal{V}_0\theta \leq 1. && (17)
\end{aligned}$$

Let us analyze the first term,

$$\left\|\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{X}^\top\epsilon\right\|_{\mathbf{W}_\lambda}^2 = \epsilon^\top\mathbf{X}\mathcal{V}_\lambda^{-1}\mathbf{C}^\top\mathbf{W}_\lambda\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{X}^\top\epsilon \tag{18}$$

The distribution of $\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{X}^\top\epsilon$ is $\mathcal{N}(0, \sigma^2\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{V}\mathcal{V}_\lambda^{-1}\mathbf{C}^\top)$. Further, $(\sigma\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{C}^\top)^{-1/2}\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{X}^\top\epsilon$ is distributed as $z \sim \mathcal{N}(0, (\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{C}^\top)^{-1/2}\mathbf{C}\mathcal{V}_\lambda^{-1}\mathcal{V}\mathcal{V}_\lambda^{-1}\mathbf{C}^\top(\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{C}^\top)^{-1/2})$. Let us call the covariance matrix $\mathbf{\Sigma}$. It is easy to see that $\mathbf{\Sigma} \preceq \mathbf{P}$, an projection matrix with k unit eigenvalues.The random variable z can be generated as $z = q^\top\mathbf{\Sigma}q$, where $q \sim \mathcal{N}(0, \mathbf{I}_n)$. Our goal is to bound,$P(\|z\|_2^2 \geq t)$. Using eigenvalue decomposition of $\mathbf{\Sigma} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, we can show

$$\mathrm{P}(\|z\|_2^2 \geq t) = \mathrm{P}(q^\top\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top q^\top \geq t) = \mathrm{P}(q^\top\mathbf{\Lambda}q^\top \geq t) \leq \mathrm{P}(q^\top\mathbf{P}q^\top \geq t) = \mathrm{P}(\|\tilde{q}\|_2^2 \geq t)$$

where $\tilde{q}$ is p-dimensional standard normal. In order to bound the resulting chi-squared, we use concentration for chi-squared random variables, $\mathrm{P}\big(\|\tilde{q}\|_2^2 \geq (p + 2\sqrt{px})\big) \leq \exp(-x)$ as in Laurent and Massart (2000). Rearranging the expression leads to the result with $\xi(\delta) = (p + 2\sqrt{p\log(\frac{1}{\delta})})$. To deal with sub-Gaussianity we apply the same trick as in the previous Theorem.

□

B.3 Adaptive design and Regularized Estimator

Theorem 9 (Adaptive Design Regularized Regression)**.** *Under the regression model in Eq. (1) with t adaptively collected data points, let $\hat{\theta}_{\lambda,t}$ be the regularized estimate as in (5). Further, assume that $\mathbf{Z}$ is as in Def. 2 where $\mathbf{X}_t = \mathbf{Z}_t\mathbf{C} + \mathbf{J}_t\mathcal{V}_0^{1/2}$. Then for all $t \geq 0$,*

$$\left\|\mathbf{C}(\hat{\theta}_t - \theta)\right\|_{\mathbf{\Omega}_{\lambda,t}} \leq \sqrt{2\log\left(\frac{1}{\delta}\frac{\det(\mathbf{\Omega}_{\lambda,t})^{1/2}}{\det(\lambda\mathbf{S})^{1/2}}\right)} + 1 \tag{19}$$

with probability $1 - \delta$, where $\mathbf{\Omega}_{\lambda,t} = \frac{1}{\sigma^2}\mathbf{Z}_t^\top\mathbf{Z}_t + \lambda\mathbf{S}$ and $\mathbf{S} = (\mathbf{C}\mathcal{V}_0^{-1}\mathbf{C}^\top)^{-1}$.

Proof. Notice that the theorem is stated with $\mathbf{S}$ and $\mathbf{S}^{-1}$ having inverted roles in the proof, however the result above follows by swapping the two.

Note that $\mathbf{J}\mathcal{V}_0^{-1/2}\mathbf{C}^\top = 0$ as well as $\mathbf{C}\mathcal{V}_0^{-1/2}\mathbf{J}^\top = 0$ due to the assumption. Also, notice that $\mathbf{X} = \mathbf{Z}\mathbf{C} + \mathbf{J}\mathcal{V}_0^{1/2}$. We drop the t subscript for brevity.

$$
\begin{aligned}
\left\|\mathbf{C}\hat{\theta}-\mathbf{C}\theta\right\|^2_{\mathbf{\Omega}_\lambda} &= \left\|\mathbf{C}(\mathcal{V}_\lambda^{-1}\mathbf{X}^\top(\mathbf{X}\theta+\epsilon)-\theta)\right\|^2_{\mathbf{\Omega}_\lambda} = \left\|\mathbf{C}(\mathcal{V}_\lambda^{-1}\mathbf{X}^\top\epsilon+\mathcal{V}_\lambda^{-1}(\mathcal{V}-\mathcal{V}_\lambda)\theta)\right\|^2_{\mathbf{\Omega}_\lambda} \\
&\leq \left\|\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{X}^\top\epsilon\right\|^2_{\mathbf{\Omega}_\lambda} + \left\|\mathbf{C}\mathcal{V}_\lambda^{-1}(\mathcal{V}-\mathcal{V}_\lambda)\theta)\right\|^2_{\mathbf{\Omega}_\lambda}
\end{aligned}
$$

Now we analyze the two terms separately, The first term,

$$
\begin{aligned}
\left\|\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{X}^\top\epsilon\right\|^2_{\mathbf{\Omega}_\lambda} &= \left\|\mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^\top(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^\top+\sigma^2\lambda\mathbf{I})^{-1}\epsilon\right\|^2_{\mathbf{\Omega}_\lambda} \\
&= \left\|\mathbf{C}\mathcal{V}_0^{-1}(\mathbf{Z}\mathbf{C}+\mathbf{J}\mathcal{V}_0^{1/2})^\top(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^\top+\sigma^2\lambda\mathbf{I})^{-1}\epsilon\right\|^2_{\mathbf{\Omega}_\lambda} \\
&\leq \left\|\mathbf{C}\mathcal{V}_0^{-1}(\mathbf{Z}\mathbf{C})^\top(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^\top+\sigma^2\lambda\mathbf{I})^{-1}\epsilon\right\|^2_{\mathbf{\Omega}_\lambda} \\
&\quad + \left\|\underbrace{\mathbf{C}\mathcal{V}_0^{-1/2}\mathbf{J}^\top}_{=0}(\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^\top+\sigma^2\lambda\mathbf{I})^{-1}\epsilon\right\|^2_{\mathbf{\Omega}_\lambda} \\
&= \left\|\mathbf{C}\mathcal{V}_0^{-1}\mathbf{C}^\top\mathbf{Z}^\top((\mathbf{Z}\mathbf{C}+\mathbf{J}\mathcal{V}_0^{1/2})\mathcal{V}_0^{-1}(\mathbf{Z}\mathbf{C}+\mathbf{J}\mathcal{V}_0^{1/2})^\top+\sigma^2\lambda\mathbf{I})^{-1}\epsilon\right\|^2_{\mathbf{\Omega}_\lambda} \\
&= \left\|\mathbf{C}\mathcal{V}_0^{-1}\mathbf{C}^\top\mathbf{Z}^\top(\mathbf{Z}\mathbf{C}\mathcal{V}_0^{-1}\mathbf{C}^\top\mathbf{Z}+\underbrace{\mathbf{J}\mathcal{V}_0^{-1/2}\mathbf{C}}_{=0}\mathbf{Z}^\top+\mathbf{Z}\underbrace{\mathbf{C}\mathcal{V}_0^{-1/2}\mathbf{J}^\top}_{=0}+\mathbf{J}\mathbf{J}^\top\sigma^2\lambda\mathbf{I})^{-1}\epsilon\right\|^2_{\mathbf{\Omega}_\lambda} \\
&= \left\|\mathbf{C}\mathcal{V}_0^{-1}\mathbf{C}^\top\mathbf{Z}^\top(\mathbf{Z}\mathbf{C}\mathcal{V}_0^{-1}\mathbf{C}^\top\mathbf{Z}+\mathbf{J}\mathbf{J}^\top\sigma^2\lambda\mathbf{I})^{-1}\epsilon\right\|^2_{\mathbf{\Omega}_\lambda} \\
&\leq \left\|\mathbf{C}\mathcal{V}_0^{-1}\mathbf{C}^\top\mathbf{Z}^\top(\mathbf{Z}\mathbf{C}\mathcal{V}_0^{-1}\mathbf{C}^\top\mathbf{Z}+\sigma^2\lambda\mathbf{I})^{-1}\epsilon\right\|^2_{\mathbf{\Omega}_\lambda}
\end{aligned}
$$

Let us define shorthand $\mathbf{S}^{-1}=\mathbf{C}\mathcal{V}_0^{-1}\mathbf{C}^\top$ which is $\mathbb{R}^{k\times k}$ p.s.d. matrix.

$$
\begin{aligned}
\left\|\mathbf{C}\mathcal{V}_\lambda^{-1}\mathbf{Z}^\top\epsilon\right\|^2_{\mathbf{\Omega}_\lambda} &\leq \left\|\mathbf{S}^{-1/2}\mathbf{S}^{-1/2}\mathbf{Z}^\top(\mathbf{Z}\mathbf{S}^{-1/2}\mathbf{S}^{-1/2}\mathbf{Z}+\sigma^2\lambda\mathbf{I})^{-1}\epsilon\right\|^2_{\mathbf{\Omega}_\lambda} && (20) \\
&= \left\|\mathbf{S}^{-1/2}(\mathbf{S}^{-1/2}\mathbf{Z}^\top\mathbf{Z}\mathbf{S}^{-1/2}+\sigma^2\lambda\mathbf{I})^{-1}\mathbf{S}^{-1/2}\mathbf{Z}^\top\epsilon\right\|^2_{\mathbf{\Omega}_\lambda} && (21) \\
&= \left\|(\mathbf{Z}^\top\mathbf{Z}+\sigma^2\lambda\mathbf{S})^{-1}\mathbf{Z}^\top\epsilon\right\|^2_{\mathbf{\Omega}_\lambda} && (22) \\
&= \left\|\mathbf{Z}^\top\epsilon\right\|^2_{\mathbf{\Omega}_\lambda^{-1}} && (23)
\end{aligned}
$$

The term above is so called self-normalized noise, which can be handled by techniques of de la Peña et al. (2009) popularized by Abbasi-Yadkori et al. (2011). From now on the proof is generic. Let us define, the noise process $S_t=\sum_{i=1}^t z_i\frac{\epsilon_i}{\sigma^2}$ and variance $\mathbf{V}_t=\frac{z_i z_i^\top}{\sigma^2}$. Also, let $M_t(x)=\exp(S_t^\top x-\frac{1}{2}x^\top\mathbf{V}_t x)$ be a process with index t.

Due to sub-gaussianity of ϵ_i, $M_t(x)$ is a super-martingale under filtration that includes all $x_{t-1},\epsilon_{t-1},\dots x_1,\epsilon_1$. Now it is worth noting that this all has been introduces since $\exp(\left\|\mathbf{Z}^\top\epsilon\right\|^2_{\mathbf{\Omega}_\lambda^{-1}})=\sup_x M_t(x)$. This relation allow us to see that if we can upper bound $\sup_x M_t(x)$ we get what we want. We proceed by pseudo-maximization. We pick a fixed probability distribution $h(x)$ and define a process $\bar{M}_t=\int M_t(x)h(x)dx$. Since, $h(x)$ is fixed and normalized $\bar{M}_t$ is also a super-martingale. It turn out that $h(x)\sim\mathcal{N}(0,\lambda^{-1}\mathbf{S}^{-1})$ will achieve the desired result.

$$
\begin{aligned}
\bar{M}_t &= \int M_t(x)h(x)dx = \frac{1}{\sqrt{(2\pi)^k\det((\lambda)^{-1}\mathbf{S}^{-1})}}\int_{\mathbb{R}^k}\exp(x^\top S_t-\frac{1}{2}\left\|x\right\|_{\mathbf{V}_t}-\frac{\lambda}{2}\left\|x\right\|_{\mathbf{S}})dx \\
&= \left(\frac{\det(\mathbf{\Omega}_{\lambda,t})}{\det(\mathbf{S})}\right)^{1/2}\exp(\frac{1}{2}\left\|S_t\right\|^2_{\mathbf{\Omega}_{\lambda,t}})
\end{aligned}
$$

where we have applied standard rules for Gaussian integrals. Now, we use Ville's martingale inequality to bound,

$$
\begin{aligned}
P(\sup \bar{M}_t \geq \frac{1}{\alpha}) &\leq \mathbb{E}[\bar{M}_t]\alpha = \alpha && (24)\\
P(\log(\sup \bar{M}_t) \geq \log(\frac{1}{\alpha})) &\leq \alpha && (25)\\
P\left(\|S_t\|^2_{\mathbf{\Omega}^{-1}_{\lambda,t}} \geq 2\log\left(\frac{1}{\alpha}\right) + 2\log\left(\left(\frac{\det(\mathbf{S})}{\det(\mathbf{\Omega}_{\lambda,t})}\right)^{-1/2}\right)\right) &\leq \alpha && (26)\\
P(\|S_t\|^2_{\mathbf{\Omega}^{-1}_{\lambda,t}} \geq 2\log\left(\frac{1}{\alpha}\det\left(\frac{\det(\mathbf{S})}{\det(\mathbf{\Omega}_{\lambda,t})}\right)^{-1/2}\right) &\leq \alpha && (27)\\
& && (28)
\end{aligned}
$$

which finishes bounding the first term by $2\log\left(\frac{1}{\alpha}\left(\frac{\det(\mathbf{\Omega}_{\lambda,t})}{\det(\mathbf{S})}\right)^{1/2}\right)$

Now we turn to the second term,

$$
\begin{aligned}
\left\|\mathbf{C}\mathcal{V}_\lambda^{-1}(\mathcal{V}-\mathcal{V}_\lambda)\theta)\right\|^2_{\mathbf{\Omega}_\lambda} &= \left\|\mathbf{C}\mathcal{V}_\lambda^{-1}(\mathcal{V}_0\lambda\sigma^2)\theta)\right\|^2_{\mathbf{\Omega}_\lambda} && (29)\\
&\overset{(34)}{\leq} \left\|\mathbf{C}\mathcal{V}_\lambda^{-1}(\mathcal{V}_0\lambda\sigma^2)\theta)\right\|^2_{\mathbf{W}_\lambda} && (30)\\
&\overset{(64)}{\leq} \left\|\mathcal{V}_\lambda^{-1}(\mathcal{V}_0\lambda)\theta)\right\|^2_{\mathcal{V}_\lambda} && (31)\\
&= \lambda\theta^\top\mathcal{V}_0\mathcal{V}_\lambda^{-1}(\mathcal{V}_0\lambda)\theta) \leq \lambda^2\theta^\top\mathcal{V}_0\theta \leq 1 && (32)\\
& && (33)
\end{aligned}
$$

This proves the result. □

Lemma 3 (Ordering)**.**

$$\mathbf{W}_\lambda \preceq \mathbf{\Omega}_\lambda \tag{34}$$

Proof. Consider series of psd. manipulations,

$$
\begin{aligned}
\mathbf{C}^\top\mathbf{W}_\lambda\mathbf{C} &= \sigma^{-2}\mathbf{C}^\top(\mathbf{C}(\mathbf{X}^\top\mathbf{X}+\sigma^2\lambda\mathcal{V}_0)^{-1}\mathbf{C}^\top)^{-1}\mathbf{C}\\
&\overset{(64)}{\preceq} \sigma^{-2}\mathbf{X}^\top\mathbf{X}+\lambda\mathcal{V}_0\\
&= \sigma^{-2}(\mathbf{C}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{C}+\mathbf{C}^\top\mathbf{Z}^\top\mathbf{J}\mathcal{V}_0^{1/2}+\mathcal{V}_0^{1/2}\mathbf{J}^\top\mathbf{Z}\mathbf{C}+\mathcal{V}_0^{1/2}\mathbf{J}^\top\mathbf{J}\mathcal{V}_0^{1/2})\\
&\quad +\lambda\mathcal{V}_0\\
\mathbf{C}\mathcal{V}_0^{-1}\mathbf{C}^\top\mathbf{W}_\lambda\mathbf{C}\mathcal{V}_0^{-1}\mathbf{C}^\top &\preceq \mathbf{C}\mathcal{V}_0^{-1}(\sigma^{-2}\mathbf{C}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{C}+\sigma^{-2}\mathbf{C}^\top\mathbf{Z}^\top\mathbf{J}\mathcal{V}_0^{1/2}\\
&\quad +\sigma^{-2}\mathcal{V}_0^{1/2}\mathbf{J}^\top\mathbf{Z}\mathbf{C}+\sigma^{-2}\mathcal{V}_0^{1/2}\mathbf{J}^\top\mathbf{J}\mathcal{V}_0^{1/2}+\lambda\mathcal{V}_0)\mathcal{V}_0^{-1}\mathbf{C}^\top\\
&= \sigma^{-2}\mathbf{S}^{-1}\mathbf{Z}^\top\mathbf{Z}\mathbf{S}^{-1}+\lambda\mathbf{C}\mathcal{V}_0^{-1}\mathbf{C}^\top\\
&= \sigma^{-2}\mathbf{S}^{-1}\mathbf{Z}^\top\mathbf{Z}\mathbf{S}^{-1}+\lambda\mathbf{S}^{-1}\\
\mathbf{W}_\lambda &\preceq \frac{\mathbf{Z}^\top\mathbf{Z}}{\sigma^2}+\lambda\mathbf{S} = \mathbf{\Omega}_\lambda
\end{aligned}
$$

where we use the fact that $\mathbf{C}\mathcal{V}_0^{-1/2}\mathbf{J}^\top = 0$. □

Lemma 4 (confidence parameter size)**.** *Under assumptions of Thm. 3, where* $\|\Phi(x)\|_k^2 \leq L^2$, $\mathcal{V}_0 = \mathcal{I}$,

$$\sqrt{2\log\left(\frac{1}{\delta}\frac{\det(\mathbf{\Omega}_{\lambda,t})^{1/2}}{\det(\lambda\mathbf{S})^{1/2}}\right)} \leq \sqrt{p\log\left(\frac{tL^2}{p\lambda}+1\right)+2\log(1/\delta)} = \mathcal{O}(\sqrt{p\log(t/p)})$$

Further, if $\|\Phi(x)\|_k \approx (\sqrt{m})$ *grows with* m*, e.g.* $\Phi(x) = 1_m$ *(vector of ones) then*

$$\sqrt{2\log\left(\frac{1}{\delta}\frac{\det(\mathbf{\Omega}_{\lambda,t})^{1/2}}{\det(\lambda\mathbf{S})^{1/2}}\right)} = \mathcal{O}(\sqrt{p\log(tm/p\lambda+1)})$$

Proof. First let us determine the absolute bound on all $\|z_i\|_2^2$. Due to projected data,

$$\mathbf{S}^{-1/2}z_i = (\mathbf{C}\mathbf{C}^\top)^{1/2}(\mathbf{C}\mathbf{C}^\top)^{-1}\mathbf{C}\Phi(x_i) \implies$$

$$\left\|\mathbf{S}^{-1/2}z_i\right\|_2^2 \leq \left\|(\mathbf{C}\mathbf{C}^\top)^{1/2}(\mathbf{C}\mathbf{C}^\top)^{-1}\mathbf{C}\Phi(x_i)\right\|_2^2 \leq \left\|\mathbf{C}^\top(\mathbf{C}\mathbf{C}^\top)^{-1}\mathbf{C}\right\|_2^2 L^2 = L^2.$$

where the last step follows from properties of projection. Notice also that $\frac{1}{\sigma^2}\operatorname{Tr}(\mathbf{S}^{-1/2}\mathbf{Z}^\top\mathbf{Z}\mathbf{S}^{-1/2}) = \frac{1}{\sigma^2}\sum_{i=1}^t \operatorname{Tr}(\mathbf{S}^{-1/2}z_i z_i^\top\mathbf{S}^{-1/2}) \leq t\frac{L^2}{\sigma^2}$.

Using arithmetic-geometric mean inequality as in (Szepesvari and Lattimore, 2020) Lemma 19.4, we can show that

$$\begin{aligned}
\frac{\det(\mathbf{\Omega}_{\lambda,t})}{\det(\lambda\mathbf{S})} &= \frac{\det(\mathbf{S}^{-1/2}\mathbf{\Omega}_{\lambda,t}\mathbf{S}^{-1/2})}{\det(\lambda\mathbf{I})} \\
&\leq \frac{(\frac{1}{p}\operatorname{Tr}(\det(\mathbf{S}^{-1/2}\mathbf{\Omega}_{\lambda,t}\det(\mathbf{S}^{-1/2}))^p}{\lambda^p} = \frac{(\frac{1}{p}\operatorname{Tr}(\mathbf{S}^{-1/2}\mathbf{Z}\mathbf{Z}^\top\mathbf{S}^{-1/2}+\lambda\mathbf{I}))^p}{\lambda^p} \\
&\leq \frac{(\frac{1}{p}(tL^2+\lambda p))^p}{\lambda^p} \\
\log\left(\frac{\det(\mathbf{\Omega}_{\lambda,t})}{\det(\lambda\mathbf{S})}\right) &\leq p\log\left(\frac{tL^2}{p\lambda}+1\right)
\end{aligned}$$

The last line of the Lemma follows from noting that $L^2 = \mathcal{O}(m)$.

□

B.4 Adaptive design and Modified Regularized Estimator: Relation to prior work

With the adaptive estimator, we could alternatively directly define an estimator for $\mathbf{C}\theta$ as $\mathbf{C}\hat{\vartheta}$ using only the projected values $\mathbf{Z}$ as done by Mutný et al. (2020). In this case, however, additional bias growing in time t enters into the confidence parameter as (9) as $\frac{1}{\lambda}\sum_{i=1}^t \|j_i\|_{\mathcal{V}_0^{-1}}$, as we show in Theorem 10 below.

Theorem 10 (Adaptive Design Regularized Regression - biased)**.** *Let* $\mathbf{Z}_t$ *be s.t.* $\mathbf{X}_i = \mathbf{Z}_i\mathbf{C} + \mathbf{J}_i$*, where* $\mathbf{J}_i$ *is minimal as measured by squared norm, then the estimator*

$$\mathbf{C}\hat{\vartheta}_t = \arg\min_{\vartheta\in\mathbb{R}^k}\sum_{i=1}^n \frac{1}{\sigma^2}(y_i - \vartheta^\top z_i)^2 + \lambda\|\vartheta\|_{\mathbf{S}}^2 \tag{35}$$

has anytime confidence sets for all $t \geq 0$

$$\mathrm{P}\left(\left\|\mathbf{C}\theta - \mathbf{C}\hat{\vartheta}_t\right\|_{\mathbf{\Omega}_\lambda} \geq 1 + 2\log\left(\frac{1}{\delta}\frac{\det(\mathbf{\Omega}_{t,\lambda})}{\det(\tilde{\mathbf{V}}_0)}\right) + \sum_{i=1}^t \|\mathbf{J}_i\|_{\mathcal{V}_0^{-1}}\right) \leq \delta \tag{36}$$

where $\delta \in (0,1)$*, and* $\mathbf{\Omega}_{t,\lambda} = \frac{\mathbf{Z}_t^\top\mathbf{Z}_t}{\sigma^2} + \lambda\mathbf{S}$*, where* $\mathbf{S} = (\mathbf{C}\mathcal{V}_0\mathbf{C}^\top)^{-1}$

Notice that if $\mathbf{J}$ is small in the Frobenius norm, the confidence sets improve. In fact, they depend only on the norm.

Proof.

$$
\begin{aligned}
\left\|\mathbf{C}\theta - \mathbf{C}\hat{\vartheta}\right\|_{\mathbf{\Omega}_\lambda}^2 &= \left\|\mathbf{C}\theta - (\frac{1}{\sigma^2}\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{S})^{-1}\mathbf{Z}^\top(\mathbf{X}\theta + \epsilon)\right\|_{\mathbf{\Omega}_\lambda}^2 \\
&= \left\|\mathbf{C}\theta - \mathbf{\Omega}_\lambda^{-1}\mathbf{Z}^\top((\mathbf{Z}\mathbf{C} + \mathbf{J})\theta + \epsilon)\right\|_{\mathbf{\Omega}_\lambda}^2 \\
&= \left\|\mathbf{C}\theta - \mathbf{\Omega}_\lambda^{-1}\frac{1}{\sigma^2}\mathbf{Z}^\top\mathbf{Z}\mathbf{C}\theta + \frac{1}{\sigma^2}\mathbf{\Omega}_\lambda^{-1}\mathbf{Z}^\top\mathbf{J}\theta + \frac{1}{\sigma^2}\mathbf{\Omega}_\lambda^{-1}\mathbf{Z}^\top\epsilon\right\|_{\mathbf{\Omega}_\lambda}^2 \\
&= \left\|\mathbf{\Omega}_\lambda^{-1}(\mathbf{\Omega}_\lambda - \frac{1}{\sigma^2}\mathbf{Z}^\top\mathbf{Z})\mathbf{C}\theta + \frac{1}{\sigma^2}\mathbf{\Omega}_\lambda^{-1}\mathbf{Z}^\top\mathbf{J}\theta + \frac{1}{\sigma^2}\mathbf{\Omega}_\lambda^{-1}\mathbf{Z}^\top\epsilon\right\|_{\mathbf{\Omega}_\lambda}^2 \\
&= \left\|\mathbf{\Omega}_\lambda^{-1}\lambda\mathbf{S}\mathbf{C}\theta + \frac{1}{\sigma^2}\mathbf{\Omega}_\lambda^{-1}\mathbf{Z}^\top\mathbf{J}\theta + \frac{1}{\sigma^2}\mathbf{\Omega}_\lambda^{-1}\mathbf{Z}^\top\epsilon\right\|_{\mathbf{\Omega}_\lambda}^2 \\
&\leq \underbrace{\left\|\mathbf{\Omega}_\lambda^{-1}\lambda\mathbf{S}\mathbf{C}\theta\right\|_{\mathbf{\Omega}_\lambda}^2}_{\text{bias}} + \underbrace{\left\|\frac{1}{\sigma^2}\mathbf{\Omega}_\lambda^{-1}\mathbf{Z}^\top\mathbf{J}\theta\right\|_{\mathbf{\Omega}_\lambda}^2}_{\text{self-normalized bias}} + \underbrace{\left\|\frac{1}{\sigma^2}\mathbf{\Omega}_\lambda^{-1}\mathbf{Z}^\top\epsilon\right\|_{\mathbf{\Omega}_\lambda}^2}_{\text{self-normalized noise}}
\end{aligned}
$$

Let us now analyze each term separately. The self-normalized terms can be analyzed using a classical technique with a proper choice of mixture distribution as we show in the proof of the Theorem 3. In this case it is a normal distribution $\mathcal{N}(0, (\lambda\mathbf{S})^{-1})$.

The first term can be shown to be bounded by,

$$
\begin{aligned}
\left\|\lambda\mathbf{\Omega}_\lambda^{-1}\mathbf{S}\mathbf{C}\theta\right\|_{\mathbf{\Omega}_\lambda}^2 &= \left\|\lambda\mathbf{\Omega}_\lambda^{-1}(\mathbf{S}(\mathbf{C}\mathcal{V}_0\mathbf{C}^\top)\mathbf{S})\mathbf{C}\theta\right\|_{\mathbf{\Omega}_\lambda}^2 && (37)\\
&= \lambda^2\theta^\top\mathbf{C}^\top\mathbf{S}\mathbf{\Omega}_\lambda^{-1}\mathbf{S}\mathbf{C}\theta && (38)\\
&\leq \lambda^2\theta^\top\mathbf{C}^\top\mathbf{S}(\lambda\mathbf{S})^{-1}\mathbf{S}\mathbf{C}\theta && (39)\\
&\leq \lambda\theta^\top\mathbf{C}^\top(\mathbf{C}\mathcal{V}_0\mathbf{C}^\top)^{-1}\mathbf{C}\theta && (40)\\
&\leq \lambda\theta^\top\mathcal{V}_0\theta \leq 1 && (41)
\end{aligned}
$$

They can be analyzed in the same spirit as in Mutný et al. (2020) novel term

$$
\begin{aligned}
\left\|\frac{1}{\sigma^2}\mathbf{\Omega}_\lambda^{-1}\mathbf{Z}^\top\mathbf{J}\theta\right\|_{\mathbf{\Omega}_\lambda}^2 &= \frac{1}{\sigma^2}\theta\mathbf{J}^\top\mathbf{Z}\mathbf{\Omega}_\lambda^{-1}\mathbf{Z}^\top\mathbf{J}\theta && (42)\\
&= \theta\mathbf{J}^\top\frac{\mathbf{Z}}{\sigma}(\frac{\mathbf{Z}^\top\mathbf{Z}}{\sigma^2} + \tilde{\mathbf{V}}_0)^{-1}\frac{\mathbf{Z}^\top}{\sigma}\mathbf{J}\theta && (43)\\
&\leq \theta\mathbf{J}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^\dagger\mathbf{Z}^\top\mathbf{J}\theta && (44)\\
&\leq \theta\mathbf{J}^\top\mathbf{J}\theta = \sum_{i=1}^{t}(j_i^\top\theta)^2 \leq \sum_{i=1}^{t}\|j_i\|_{\mathcal{V}_0^{-1}}^2\frac{1}{\lambda} && (45)
\end{aligned}
$$

□

C Algorithms

Now we will briefly discuss algorithms that construct designs $\mathbf{X}$, and allocations over them η, which lead to low error either in expectation or with high probability - using A or E design, respectively. Depending on the aim and estimator, different algorithms might be preferable. For comprehensive reviews please refer to Todd (2016), Pukelsheim (2006),(Fedorov and Hackl, 1997) for optimization algorithms, and (Allen-Zhu et al., 2017), (Camilleri et al., 2021) for rounding techniques.

C.1 Greedy selection

A first idea is to greedily maximize the scalarized information matrix with the update rule $\eta_{t+1} = \frac{t}{t+1}\eta_t + \frac{1}{1+t}\delta_t$,

$$
\delta_t = \arg\max_{x\in\mathcal{X}} f\left(\mathbf{W}_\lambda\left(\frac{t}{t+1}\eta_t + \frac{1}{t+1}\delta_x\right)\right)
$$

where f refers to the scalarization and δ_x to the indicator of the discrete measure corresponding to the feature $\Phi(x)$. Notice that while stated in form of allocations, due to the form of the update rule $t\eta_t$ is always an integer.

Surprisingly, this algorithm can fail with the interpolation estimator, where adding a new row to $\mathbf{X}$, which does not lie in the kernel of $\mathbf{C}$ increases the variance. This is an artifact of the pseudo-inverse, but it demonstrates that the algorithm is not universal – hence we suggest using it with the regularized estimator only.

C.2 Convex optimization

Alternatively, one can first select a design space $\mathbf{X}$ supported on finitely many queries such that if the optimal design is supported on it, it leads to a desired low bias. After that, we optimize the allocation $\eta \in \Delta^n$ over it. This has the advantage that *a)* we can bound the query complexity to reach ϵ of learning as in Proposition 14, and *b)* we can provably achieve optimality with convex optimization, in contrast to the greedy heuristic.

We look for an allocation using convex optimization methods as in (10) or more generally, $\max_{\eta\in\Delta^n} f(\mathbf{W}_\circ(\mathbf{D}(\eta)^{1/2}\mathbf{X}))$ where $\circ \in \{\dagger, \lambda\}$. There are three possible ways to initialize the algorithm with $\mathbf{X}$: make an ansatz, greedily reduce bias first, or use a modification of the random projection initialization of Betke and Henk (1992) described below.

Mirror-descent and Regularity The objective above (or (10)) would be classically solved via Frank-Wolf algorithm (Todd, 2016), or a mirror descent algorithm (Beck and Teboulle, 2003) [2], which starts with the whole support, and reduces the weight of some of the queries η_i, as

$$\eta_{t+1} = \eta_t \frac{\exp(-s_t \nabla f(\mathbf{W}(\eta_t)))}{\sum_i \exp(-s_t \nabla_i f(\mathbf{W}(\eta_t)))}$$

where s_t is the stepsize; if $s_t \propto \sqrt{t}^{-1}$, the convergence is guaranteed. Both E and A-design objectives are concave as they are related by linear transform $\mathbf{C}$ from the classical objectives, which are known to be concave (Pukelsheim, 2006).

paragraphInitialization via random projections This algorithm is inspired by volume algorithm of Betke and Henk (1992). The algorithm proceeds by picking a random vector c_0 in the span of $\mathbf{C}^\top$ and picking two points $\bar{z}, \underline{z} = \arg\max_x c^\top x, \arg\min_x c^\top x$. Subsequently, we pick another c_1 which is orthogonal to $\underline{z} - \bar{z}$ and still in span of $\mathbf{C}^\top$. We repeat this procedure k times. Should this not generate a sufficiently accurate starting point with desired bias (as measured by Def. 1), we repeat this procedure until such design is obtained.

C.3 SDP reformulations of the objectives

We will now show that both A and E designs have an associated semi-definite reformulations which are possible if $\dim(\mathcal{H}_\kappa) = m$ [3]. One should bear in mind that these are possible only if the $\mathbf{C}\theta$ is estimable in the sense of Def. 3, Hence their might not be utilizable in many instances that this paper studies, and rather apply in cases classical experiment design. Nevertheless they do not appear in literature to the best of our knowledge.

On top of that these problems are more difficult to solve that lets say the alternative algorithm with mirror descent, however with the use of off-the shelf solvers for SDP problems such as cvxpy, they can be conveniently implemented. These refomrmulations are inspired by classical reformulations due to Boyd and Vandenberghe (2004), which show these for estimating θ directly. The key ingredient is to make use of generalized Shur-complement theorem of Zhang (2011).

A-design objective is $\max_{\eta\in\Delta_n} 1/\operatorname{Tr}(\mathbf{W}(\eta)^{-1})$ can be represented as semi-definite program:

$$\begin{aligned} &\min_{u\geq 0, \eta\in\Delta_n} && u_1 + u_2 + \cdots + u_k \\ &\text{subject to} && \begin{pmatrix} \mathbf{V}(\eta) & \mathbf{C}^\top e_i \\ e_i^\top \mathbf{C} & u_i \end{pmatrix} \succeq 0 \ \text{ for all } i \in [k] \end{aligned}$$

[2] Mirror descent is known as multiplicative algorithm in the experimental design literature (Silvey et al., 1978).

[3] We would like to thank Stephen Wright of University of Wisconsin for suggesting to revisit this idea.

E-design objective is $\max_{\eta\in\Delta_n} \lambda_{\min}(\mathbf{W}(\eta))$, can be again solved using semi-definite programming as:

$$\begin{aligned} &\min_{t\in\mathbb{R},\eta\in\Delta_n} \quad t \\ &\text{subject to} \quad \begin{pmatrix} \mathbf{V}(\eta) & \mathbf{C}^\top \\ \mathbf{C} & t\mathbf{I}_{k\times k} \end{pmatrix} \succeq 0 \end{aligned}$$

We will show how to apply the generalized for the first objective (A-design) only. The second reformulation follows analogously. Notice that the objective $\max_\eta 1/\operatorname{Tr}(\mathbf{W}^{-1})$ is equivalent to $\min_\eta \operatorname{Tr}(\mathbf{W}^{-1})$, and $\mathbf{W}^{-1} = \mathbf{C}\mathbf{V}(\eta)^\dagger\mathbf{C}^\top$. Generalized Shur-complement lemma of Zhang (2011) states that $\mathbf{C}\mathbf{V}(\eta)^\dagger\mathbf{C}^\top \preceq \mathbf{\Lambda}$ with the projection operator lying in the null space of $\mathbf{C}$, i.e. $(\mathbf{I} - \mathbf{V}(\eta)^\dagger\mathbf{V}(\eta))\mathbf{C}^\top = \mathbf{0}$, is equivalent to

$$\begin{pmatrix} \mathbf{V}(\eta) & \mathbf{C}^\top \\ \mathbf{C} & \mathbf{\Lambda} \end{pmatrix} \succeq 0.$$

The second condition is satisfied as long as the problem is estimable, in other words, there exists $\mathbf{L}$ s.t. $\mathbf{C} = \mathbf{L}\mathbf{X}$. Hence minimizing the trace of $\mathbf{\Lambda}$, we can minimize the trace of the information matrix. As the trace depends only on the diagonal elements of $\mathbf{\Lambda}$, we can choose it to be diagonal, i.e., $\mathbf{\Lambda} = \operatorname{Diag}(u)$.

C.4 Query complexity and Optimization

The objective in (10) was stated in the kernelized form which is only valid if the RKHS is finite-dimensional. We state here for completeness the version which operates in the kernelized setting where all the operators can be inverted (and calculated explicitly using only the kernel κ evaluation).

$$\eta^* = \arg\max_{\eta\in\Delta^n} \lambda_{\min}\left(\mathbf{C}\mathcal{V}_0^{-1}\mathbf{X}^\top(\mathbf{D}(\eta)\mathbf{X}\mathcal{V}_0^{-1}\mathbf{X}^\top\mathbf{D}(\eta))^{-2\dagger}\mathbf{X}\mathcal{V}_0^{-1}\mathbf{C}^\top\right)^{-1}. \tag{46}$$

D Geometry

We saw in Section 6 that properties of $\{\Phi(x) : x \in \mathcal{S}\}$ enter into consideration when we want to bound the total number of queries required to reach ϵ accuracy with high probability such that we can learn $\mathbf{C}\theta$. The dependence enters as the smallest eigenvalue $\lambda_{\min}$ of the information matrix $\mathbf{W}$. In this section, we relate the eigenvalue $\lambda_{\min}(\mathbf{W})$ directly to the geometry of the above set using seminal results about E-experimental design from Pukelsheim and Studden (1993).

D.1 Set Width

The geometrical property we are interested in *width of a set* that we define below. However, first, we want to make a general note that the formal results of our theorems hold for symmetrized sets $\{\Phi(x)|x \in \mathcal{S}\}$. This is without the loss of generality, since, an optimal design on symmetrized set and a non-symmetrized set is equivalent. This can be seen by considering $\mathbf{W}_1(\eta)$ defined on symmetrized set equates the one of non-symmetrized set $\mathbf{W}_2(\eta)$ due to the fact evaluations $\Phi(x)$ enter only as outer products $\mathbf{W}_1(\eta) = r(\sum_{i=1}^n \eta_i(\Phi(x_i)\Phi(x_i)^\top))$, where the sign clearly does not change anything in terms of the information matrix, and r correspond to further operations to define $\mathbf{W}$ properly (i.e. projection and inverse).

Definition 5 (Width)**.** Let $\mathcal{D}$ be a convex set in $\mathcal{H}$, then the width of it is defined as,

$$\operatorname{width}(\mathcal{D}) = \min_{\theta\in\mathcal{H}} \max_{x,y\in\mathcal{D}} \theta^\top(x-y) \tag{47}$$

where θ is the unit norm.

Width in the span of $\mathbf{C}$ is defined to be,

$$\operatorname{width}(\mathcal{D}, \mathbf{C}) = \min_{\theta=\mathbf{C}^\top\alpha \text{ s.t. } \alpha\in\mathbb{R}^k} \max_{x,y\in\mathcal{D}} \theta^\top(x-y) \tag{48}$$

where α is the unit norm.

Definition 6 (Gauge norm)**.** Let $\mathcal{D}$ be a set of points in $\mathcal{H}$, then the gauge norm of $\mathcal{D}$ is,

$$\rho_{\mathcal{S}}(x) = \inf_{r\geq 0}\{r : x \in r\operatorname{conv}(\mathcal{D})\} \tag{49}$$

Theorem 11 (Thm. 2.2 in (Pukelsheim and Studden, 1993))**.** *Let* $\mathbf{W}$ *be any the information matrix for* $\mathbf{C}\theta$*, and* $z \in \mathbb{R}^k$ *non-zero then*

$$\lambda_{\min}(\mathbf{W}) \leq \frac{\|z\|_2^2}{\rho(\mathbf{C}^\top z)^2}.$$

It holds with <u>*equality*</u> *if* $\mathbf{W}$ *has smallest eigenvalue of multiplicity one and* z *is the associated eigenvector.*

The treatment without multiplicity is somewhat more complicated and can be found in.

Proposition 12. *Under assumption of Thm. 11, no multiplicty,* z *being eigenvector of* $\mathbf{W}$*, let* $\mathbf{C} : \mathbb{R}^m \to \mathbb{R}^p$*,* $\mathcal{D}$ *be a symmetric convex set in* $\mathcal{H}$*, then*

$$\lambda_{\min}(\mathbf{W}) = \frac{\|z\|_2^2}{\rho(\mathbf{C}^\top z)^2} \geq \frac{1}{\mathrm{width}(\mathcal{S}, \mathbf{C})^2}$$

Proof. Due to homogenity of the gauge norm, $\rho(\mathbf{C}^\top z \frac{\|z\|_2}{\|z\|_2}) = \|z\|^2 \, (\rho \mathbf{C}^\top \hat{z})$, where $\hat{z}$ is a unit vector. Thus, the optimization can be restated as,

$$\frac{\|z\|_2^2}{\rho(\mathbf{C}^\top z)^2} = \frac{1.}{\rho(\mathbf{C}^\top \hat{z})^2} = \frac{1}{\mathrm{width}(\tilde{\mathcal{D}})^2} \geq \frac{1}{\mathrm{width}(\mathcal{D}, \mathbf{C}^\top \hat{z})^2}$$

where in the second to last step we use the fact that the set is symmetric and $\hat{z}$ is unit eigenvector. The last inequality follows by minimizing over all possible unit direction as int the definition of width. The set $\tilde{\mathcal{D}} \subset \mathcal{H}$ s.t. $\mathcal{D} = \{f \in \mathcal{H}_k | f = \mathbf{C}^\top \hat{z}\}$. Notice that this set lies in 1-dimension subspace, and is directly proportional to the gauge-norm. □

Given the above theorem, we can calculate the worst-case value of parameter $\frac{1}{\lambda_{\min}(\mathbf{W})}$ that bounds the complexity if the geometry of the set can be understood easily. Alternatively, we can resort to direct calculation. For example consider the gradient design problem, in which we know that for finite difference problem can be bounded by $dh + O(h)$, where h is the stepsize.

Proposition 13. *Under the assumption of the model (1), if* $\mathbf{C} = \nabla_x \Phi(x)$ *at* $x \in \mathbb{R}^d$*,* $\Phi(x)$ *is the evaluation functional, and the design space* $\mathbf{X}$ *contains rows* $\{\Phi(x \pm he_i\}_{i=1}^d$ *where* e_i *corresponds to the unit vectors in* $\mathbb{R}^d$ *and* h *is the step-size, then*

$$\frac{1}{\lambda_{\min}(\mathbf{W}(\eta)} \leq dh + O(h^2), \tag{50}$$

where η *corresponds to the design supported on all points with equal weight.*

Proof. The inverse of the information matrix is,

$$\begin{aligned}
\mathbf{W}_\dagger^{-1} &= \nabla_x \Phi(x) \left(\sum_{i=1}^d \eta_i (\Phi(x+he_i)\Phi(x+he_i)^\top + \Phi(x-he_i)\Phi(x-he_i)^\top) \right)^{-1} \nabla_x \Phi(x)^\top \\
&= \nabla_x \Phi(x) \left(\sum_{i=1}^d \eta_i (2\Phi(x)\Phi(x)^\top + h\nabla_x \Phi(x)^\top e_i e_i^\top \nabla_x \Phi(x) + O(h^2) \right)^{-1} \nabla_x \Phi(x)^\top \\
&\preceq \nabla_x \Phi(x) \left(h\nabla_x \Phi(x)^\top \mathbf{D}(\eta) \nabla_x \Phi(x) + O(h^2) \right)^{-1} \nabla_x \Phi(x)^\top \\
&= h\nabla_x \Phi(x) \left(\nabla_x \Phi(x)^\top \mathbf{D}(\eta) \nabla_x \Phi(x) + O(h^2) \right)^{-1} \nabla_x \Phi(x)^\top \\
&= dh\nabla_x \Phi(x) \left(\nabla_x \Phi(x)^\top \nabla_x \Phi(x) + O(h^2) \right)^{-1} \nabla_x \Phi(x)^\top \\
&\preceq dh\mathbf{I} + O(h^2)
\end{aligned}$$

In the first step we used Taylor's theorem and keep track of the order components. In the second we used that $\Phi(x)\Phi(x)^\top$ is psd, later we used the definition of the design and the properties of the projection matrix. Lastly,

$$\frac{1}{\lambda_{\min}(\mathbf{W}_\dagger)} = \lambda_{max}(\mathbf{W}_\dagger^{-1}) \leq dh + O(h^2),$$

which finishes the proof. □

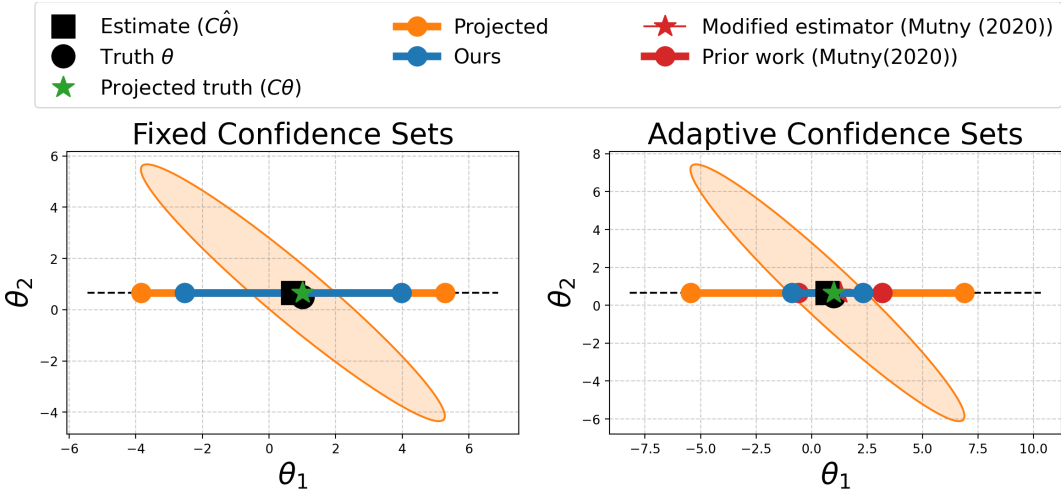


Figure 3: An example of confidence ellipsoids and intervals in two dimensions. You can see that in both regimes: fixed and adaptive. Our confidence sets (Ours) are the tightest for the estimation of the linear projection of $\mathbf{C}\theta$ value of the parameter $\theta = (1, 1)$, in this case $\mathbf{C} = (1, 0)$. On the left, we see fixed confidence sets while on the right are adaptive both with $T = 50$. We compare with (Mutný et al., 2020) and associated confidence sets, they are larger for this example (and grow with T); also the estimator is different as we delineate in Sec. B.4.

Corollary 14 (Query complexity)**.** *Under assumption of Prop. 1 where η^* is the optimum to Eq. 10, it holds that*

$$\left\|\mathbf{C}(\hat{\theta}_\dagger - \theta)\right\|_2 \leq \sqrt{\frac{\lambda_{\min}(\mathbf{W}_\dagger(\eta^*))^{-1}}{T}} \left(\sigma\sqrt{\xi(\delta)} + \frac{\nu}{\sqrt{\lambda}}\right).$$

with probability $1 - \delta$.

Proof.

$$\left\|\mathbf{C}(\hat{\theta}_\dagger - \theta)\right\|_2^2 \leq \frac{1}{\lambda_{\min}(\mathbf{W}_\dagger(\eta^*)} \left\|\mathbf{C}(\hat{\theta}_\dagger - \theta)\right\|_{\mathbf{W}_\dagger(\eta^*)} \tag{51}$$

$$= \frac{1}{\lambda_{\min}(\mathbf{W}_\dagger(\eta^*)}(\sigma^2\xi(\delta) + \frac{\nu^2}{\sqrt{\lambda}}) \tag{52}$$

Now the $\mathbf{W}_\dagger$ nor ν^2 do not depend on σ^2, however evaluating multiple times the same measurements Namely T times, reduces the variance σ^2 by $1/T$. After taking this into consideration and the square root it finishes the proof. □

E Extra Examples

In this section, we provide additional applications of our framework that deserve a mention, but before we do, we want to bring attention to Figure 3, where we show a comparison of our confidence sets with prior work and projected ones, similar to Figure 1. As a refresher, recall that

$$\mathbf{L} : \mathbb{R}^n \to \mathbb{R}^p$$
$$\mathbf{X} : \mathcal{H}_k \to \mathbb{R}^n$$
$$\mathbf{C} : \mathcal{H}_k \to \mathbb{R}^p.$$

E.1 Integral maps

Quadrature An integral is a special linear operator on the function space,

$$\mathbf{C} = \int q(x)\Phi(x)^\top \cdot dx = \int q(x)\Phi(x)dx^\top \cdot,$$

where $p = 1$. The approximate estimability condition can be written in a concise form

$$\|\mathbf{C} - \mathbf{L}\mathbf{X}\|_k = \left\| \int (q(x) - \sum_{y \in \mathcal{S}} l(y)\delta(x-y))\Phi(x)dx \right\|_k ,$$

where $l(y) = \mathbf{L}_y$, and $y \in \mathcal{S}$ which itself is equal to well-known integral distance $\mathrm{MMD}_{\mathcal{H}_\kappa}(q, \sum_{y \in \mathcal{D}} l(y)\delta(\cdot - y))$ referred to as Maximum Mean Discrepancy (MMD) multiplied by B. It measures the worst case integration error on the class of functions $\theta \in \mathcal{H}_\kappa$ s.t. $\|\theta\|_k \leq 1$ with nodes $x \in \mathcal{S}$ and weights $l(x)$. The method which minimizes this quantity is sometimes referred to as Bayesian Quadrature (Huszár and Duvenaud, 2012). Note that minimization of this quantity would correspond to the interpolation estimator.

Fourier Spectra: Low-pass Filters Suppose we are interested in approximately learning the Fourier spectrum of θ. This might be especially interesting if θ is composed of two signals one with specific low-frequency bands and the rest occurs only elsewhere and is of no interest. Let $\{\omega_i\}_{i=1}^p$ be the frequencies of interest, then the rows of $\mathbf{C}$ are $\mathbf{C}_j = \int_{\infty}^{\infty} \exp(i\omega_j x)\Phi(x)^\top \cdot dx$, which can be discretized or evaluated in closed form depending on the kernelized space.

Transductive Risk If the dimension of the response vector is very large much larger than the number of elements in the unlabeled set, which is a common appearance in deep learning or high dimensional statistics; there is still hope to have provable bounded error on selected important candidates. As opossed to study the full risk $\mathbb{E}_\epsilon[\mathbb{E}_{x\sim\rho}[(x^\top(\theta - \hat{\theta}))^2]]$ with data distribution ρ, we can optimize the risk on a set of candidates $\mathbb{E}_\epsilon[\sum_i^p[(q_i^\top(\theta - \hat{\theta}))^2]]$, which becomes $\mathrm{Tr}(\mathbf{Q}\mathbf{V}^\dagger\mathbf{Q})$, where $\mathbf{Q}$ contains the rows of k are the candidates.

E.2 Solution to Contamination (Low pass filter)

Often in estimation, the signal of interested is corrupted by another signal that we are not interested in. This is sometimes referred to as *contamination* (Fedorov and Hackl, 1997). In particular, consider the following model $\mathbb{E}[y] = (\alpha \quad \beta)^\top (\phi(x) \quad \psi(x))$, where we want to infer α only. Note that this is a special case of our framework, where the linear functional $\mathbf{C} = \mathbf{I}_S$ zeros out the variables β in $\theta = (\alpha, \beta)$. This problem is well-defined only if the spans of $\{\phi(x)\}_{x\in\mathcal{S}}$ and $\{\psi(x)\}_{x\in\mathcal{S}}$ are not contained in each other, which we neglect for our purposes here.

For illustration, consider a linear trend contaminated with function f, which has major frequency components that are *known*, i.e., $y = \alpha^\top x + f(x) + \epsilon$. We can stack these frequencies in an evaluation functional $\psi(x, \{\omega\}_{i=1}^m)$, and express the problem in the form above. In Fig. 5, we report the MSE for estimation of α with f which has frequency components $\{\pi l, \pi e l | l \in [16]\}$ weighted with $1/l^2$. The contamination-aware design tries to query points that eliminate the effect of f and leads to lower MSE for the same number of observations and performs much better in expectation than either random or full designs – an optimal design that learns both α and β. For this design, we chose to sample greedily first to reduce bias and then optimized the weights of each data point to achieve optimality.

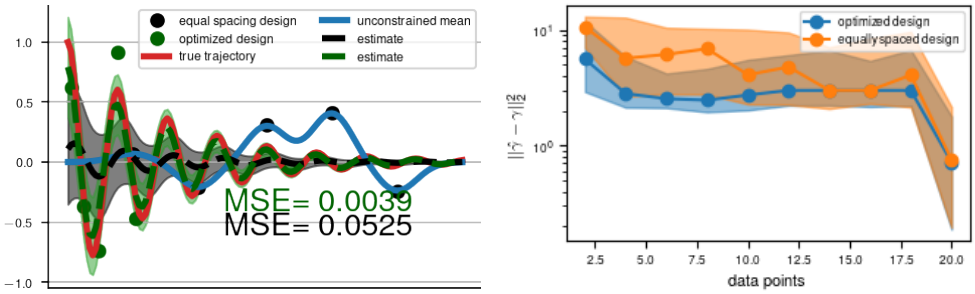
E.3 Learning ODE solutions

Consider an example of damped harmonic oscillator, $u(t)'' + \kappa u(t)' + u(t) = 0$ for $t \in [t_0, t_1]$. Adopting the above procedure and using the squared exponential kernel to embed trajectories [4] , we show in Figure 4a inference with the shape constraint using A-optimal design as well as equally space design, we see the resultant confidence bands and fit are much better for optimized design. We assume that $\kappa \in (\kappa_-, \kappa_+)$ and give a union of confidence sets as in the robust design case. The optimal design then corresponds to the unknown null subspace $\mathbf{C}$ and its variance part is equal to $\sup_{\kappa\in[\kappa_-,\kappa_+]} \mathrm{Tr}(\mathbf{C}_\kappa \mathbf{V}^\dagger \mathbf{C}_\kappa^\top)$.

E.4 Learning PDE solutions

In a similar fashion as the linear ODE feature, a linear constraint so do linear PDEs. We demonstrate this on a sensor placement problem, where consider heat equation in two dimensions with $u(t, x)$

[4] The nullspace of $\mathbf{Z}$ forms its own Hilbert space that spans the space of trajectories, however for estimation convenience it is sometimes easier to work with a larger space and then project.



(a) An estimate of the trajectory of a damped harmonic oscillator. The source of unknown here is due to the unknown initial conditions. Notice that with the optimized design we can achieve a much better fit. Also, the unconstrained fit with the squared exponential kernel is shown in blue. Below we also see the reduction in MSE for estimating this trajectory with the two designs.

(b) MLE error of estimating γ in the pharmacokinetic study. We see that with the optimized design we are consistently better in terms of MSE than equally spaced design. Error bars with 20 repetitions included. The y-axis denotes the number of data samples (draws of blood) required. Notice the logarithmic scale.

Figure 4: Further experiments showing application in ODE trajectory and parameter estimation.

with varying diffusion coefficient $c(x)$.

$$\partial_t u - c(x)\partial_{xx} u = 0 \quad \text{in} \quad (0,T)\times\Omega$$
$$u(0,x) = \langle \theta, \Phi(x)\rangle \quad \text{for} \quad x\in\Omega, z\in\partial\Omega.$$

We assume that the point value of initial conditions $u(0,x)$ can be measured by placing sensors along x. The goal is to place these sensors such that after elapsed time T, $u(T)$ can be inferred with the lowest error. This is the forward formulation of the inverse problem in Leykekhman et al. (2020). We assume that θ is a bounded member of RKHS due to the squared exponential kernel. In this case we are not interested in the full nullspace of $\mathbf{N}$ as in the previous example, instead we focus on the range space of $\mathbf{C} = \Phi(x,T)(\mathbf{N}^\top\mathbf{N})$ corresponding to information at $\Phi(x,T)$.

E.5 Sequential Design: Estimating CVar

Often the objective we are trying to estimate is inherently stochastic such as the following model:

$$\mathbb{E}_{z\sim W(x)}[\rho(\Phi(x,z)^\top\theta)]. \tag{53}$$

In particular one can be interested in x s.t. $\arg\sup_x \mathbb{E}_{z\sim W}[\Phi(x,z)^\top\theta]$. This is analyzed in sequential experiment design optimization with risk measures, where ρ is the risk measure (see (Cakmak et al., 2020) and (Agrell and Dahl, 2020)). It is often assumed that W cannot be evaluated explicitly, only sampled from with the evaluation oracle $y_i = \Phi(x_i,z_i)^\top + \epsilon_i$. We focus on a problem where we want to learn the risk of the actions $q\in Q, |Q| = k$, $\mathbb{E}_{z\sim W}[\rho(\Phi(q,z)^\top\theta)]$ instead of minimization and ρ is CVar risk measure. With a fixed value θ, this is a linear operator $\mathbf{C}_\theta$. Since θ is unknown, we adopt a sequential procedure, where we solve for the risk $C_{\hat\theta}$, where $\hat\theta$ is sampled uniformly from its confidence set.

F Experiments: Details

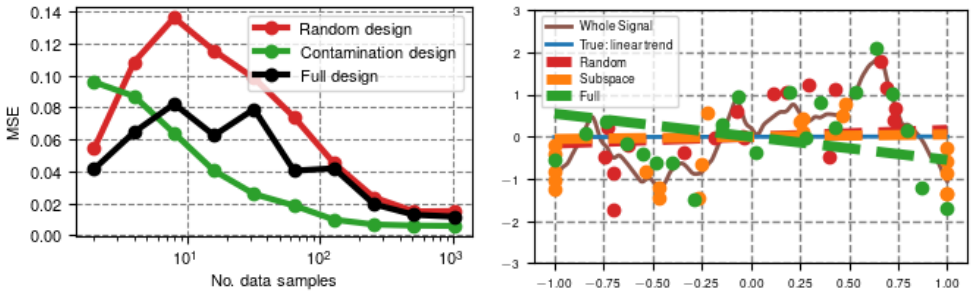
F.1 Gradient Estimation: Details

With gradient estimation we are interested in $\nabla_x\Phi(x)$ of an element $\theta\in\mathcal{H}_\kappa$. The example generated in Fig. 2b is in 2 dimensions, where we approximate $\Phi(x)$ of squared exponential kernel with lengthscale $l = 0.1$ with Quadrature Fourier Features of Mutný and Krause (2018) for the convenience of optimization. We consider a slightly off-set finite difference design, with the following elements,

$$\{\Phi(x), \Phi(x+he_1), \Phi(x+2he_1), \Phi(x-he_1), \Phi(x-he_2)\}.$$

We calculate the relative-bias ν due to the Def. 1 and balance it with the budget (as in the Figure). This way we can identify the optimal h^* before even solving for the optimal design. Given the identified h^*, we calculate the optimal design with mirror descent solving the objective (10), to get

$$\eta^* = (0.37, 0.09, 0.08, 0.09, 0.38),$$



(a) Mean squared error of the estimated linear trend. We see that contamination-aware design improves over full design as well as randomly selected design.

(b) Example of the signal in brown with design points as well as linear trend estimates in color-coding. Full denotes joint estimation of brown and blue signal. Random designates a random design.

Figure 5: Statistical contamination: a numerical experiment

which is not obvious without understanding the geometry of the problem. The value of σ was set to be 0.01. The total error in Fig. 2b is the combination of bias and variance which was calculated using the largest eigenvalue of $\mathbf{E}(\mathbf{L}_\dagger)$ as in (2). This is the error we can guarantee with high probability (not the actual error) which is somewhat lower.

F.2 The solution to Contamination (Statistics): Details

For this example, consider a linear trend contaminated with $f(x)$ which has known major frequency components, i.e., $y = \alpha^\top x + f(x) + \epsilon$. We can concatenate these frequencies, and construct an evaluation operator where $\Psi(x)_i = \cos(\omega_i x)$, and $\Psi(x)_{i+1} = \sin(\omega_i x)$ where ω_i is one of these frequencies. Then, we express the problem as $y = \theta^\top \Phi(x) = (\alpha, \beta)^\top (x, \Psi(x, \{\omega_i\}_{i=1}^m) + \epsilon$. Since we are interested only in α, the operator $\mathbf{C} = (1, \mathbf{0})$ selects it. Hence we look for a design (subset of points) which only minimizes the residuals due to estimation of α. In fact, such design might indirectly minimize the other residuals as well, but only if this is necessary to reduce the residuals due to α.

The example depicted in Fig. 5 is created using the frequencies $\{\pi l | l \in [16]\} \cup \{\pi e l | l \in [16]$ weighted with $1/l^2$. Such weighting of high frequencies creates contaminating signals that are similar to functions from Matérn kernel spaces (Rasmussen and Williams, 2006). The contamination-aware design performs much better.

The values of $\sigma = 0.5$ (relatively high). We compare with full design – one that minimizes the overall estimation error or in other words, where $\mathbf{C}$ is the identity. We first use the greedy algorithm to minimize the regularized estimator to obtain initial design space. Then we use convex optimization to find a proper allocation of a fixed budget (varying in the x-axis) to these queries. We again optimize with mirror descent.

F.3 Pharmacokinetics: Details

The pharmacokinetic two-compartment model models two organs. In this case the stomach and blood and the transfer of the medication between them. The model can be mathematically described as

$$\begin{aligned} \frac{d}{dt} c_s(t) &= -a c_s(t) \\ \frac{d}{dt} c_b(t) &= b c_s(t) - d c_b(t) \end{aligned}$$

where $c_s(t)$ represent concentration of the medication in the stomach, and $c_b(t)$ represents the concentration in the blood. The goal of pharmacokinetic studies is to identify (a, b, c) from draws of the blood of subjects to e.g. properly assign dosage.

In comparison to the example of the damped harmonic oscillator, the initial conditions are known – albeit imprecisely – what is the true unknown is $\gamma = (a, b, c)$. We make the following consideration, in order to infer γ precisely we need to have a precise estimation of the trajectory, hence we inject a slight perturbation to the initial condition and then search for the optimal design that is good for any

of the models γ: robust design. This way with any of the models from $\gamma \in \Gamma$ we will have a good estimation of trajectories and hence a good estimation of γ using MLE as we report in Fig. 4b.

The specific parameter used was $\sigma = 0.01$ which corrupted the observations of $c_b(t)$ at chosen times and the space which Embeds the trajectories was chosen to be that of the squared exponential kernel with lengthscale $l = 0.05$. The true parameters were set to be $(5, 10, 10)$ and the robust set of $\Gamma = ((4, 6), (9, 11), (9, 11))$ for each parameter. We use the greedy algorithm with regularized estimator and $\lambda = 1/2$ to optimize the robust A-optimal metric. The resultant set is visualized in Fig. 2a.

F.4 Control: Details

In the last experiment, we essentially model the same problem as in Lederer et al. (2021) as we explain. The only difference in our setup is the use of a gain $K = 200$ instead of $K = 15$. With a low gain of $K = 15$ one cannot provably stabilize with the given Lyapunov function. With $K = 200$ we were able to quickly stabilize the system once enough data points were obtained. Notice the reference trajectory and trajectory with "bad" and "good" controllers in Fig. 6. The driving non-linear dynamics are visualized likewise.

The experiment is designed to showcase the merit of the newly designed adaptive confidence sets with the stopping rule that stops once the controller has been verified to be stable. We repeated each experiment 10 times and reported the standard quantiles of the total derivative of the Lyapunov function in Fig. 2c. We always started with the initial set of 10 data points and then proceeded with exploration (adding specific data points) if the supermum of the total derivative was not negative. When it was we stopped. The stopping corresponds to the line going into negative values in the log plot in Fig. 2c. We considered these four strategies:

- *random* - randomly sampling data point from the whole domain [-1.5,1.5].
- *random-ref* - randomly sampling data point around the operating region.
- *unc* - sampling (greedily) the most uncertain query from the whole domain [-1.5,1.5].
- *unc-ref* - sampling (greedily) the most uncertain query from the operating region.

Berkenkamp et al. (2016) provides a better solution than uncertainty sampling, which is a special case of the linear functional studies in this work. In this example, this closely follows the *unc-ref* baseline and would create additional clutter. The inclusion of this algorithm would not further validate the benefit of the tighter confidence sets. The operating region is chosen to be a tube around the reference trajectory of width $h = 0.01$.

The dynamics is modeled using Nystöm features ($m = 400$) of the squared exponential kernel with lengthscale 0.25 and regularized estimator with λ estimated experimentally. We selected this with visual inspection since the example is known, but one can adopt the procedure in Umlauft et al. (2018). The noise std. $\sigma = 0.05$ and number of queries is $T = 500$.

G Improved regret of linear bandits: Proofs

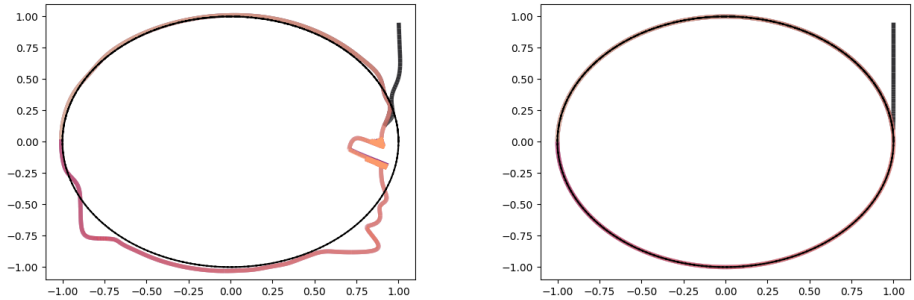
To prove the following theorem we require the famed potential lemma which we state for completeness,

$$\sum_{t=1}^{T} \|x_t\|_{\mathbf{V}_t^{-1}} \leq \mathcal{O}(\sqrt{dT \log(T/d)}).$$

A reference for this result can be found in (Szepesvari and Lattimore, 2020). Note that a bounded pay-off vector assumption has been used in the derivation. In general this result can be extended to kernelized bandits by noting that

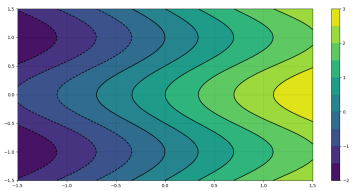
$$\sum_{t=1}^{T} \|x_t\|_{\mathbf{V}_t^{-1}} = \mathcal{O}(\sqrt{\gamma_T T}),$$

where $\gamma_T = \log\left(\frac{\det(\mathbf{V}_t)}{\det(\mathbf{I})}\right)$, due to matrix inversion lemma, this can be also expressed in kernelized form, as is usually referred to as information gain (Srinivas et al., 2010).

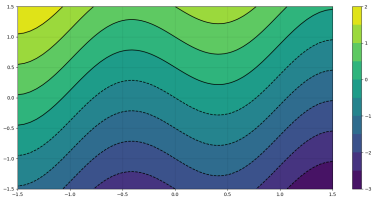


(a) An example of an unstable trajectory around the reference trajectory, started from (1,1) point

(b) An example of a stable trajectory around the reference trajectory, started from (1,1) point



(c) Non-linear dynamics function $f_1(x)$ which governs the first component x.



(d) Non-linear dynamics function $f_2(x)$ which governs the second component of x.

Figure 6: Control Experiment: Additional details.

Theorem 15. *Let $\theta \in \mathbb{R}^d$ be an unknown pay-off vector, and a set of actions $x \in \mathcal{X}$ such that $|\mathcal{X}|\infty$, then the regret of UCB algorithm is no more than*

$$R_T \leq \mathcal{O}(\sqrt{dT \log(T(|\mathcal{X}|+1)/\delta)}) \tag{54}$$

with probability $1-\delta$.

Proof of Theorem 15. We drop Φ and use x instead only. Note that due to Lemma 3, we have that

$$\left\| x_k^\top (\hat{\theta} - \theta) \right\|^2_{\mathbf{W}_{t,k}} \leq \left\| x_k^\top (\hat{\theta} - \theta) \right\|^2_{\mathbf{\Omega}_{t,k}} \leq \beta_t,$$

where β_t is defined for the information matrix $\mathbf{\Omega}_{t,k}$ in Theorem 3. The subscript k designates the dependence on x_k as we will need to consider all $x_k \in \mathcal{X}$, where we index $\mathcal{X}$ using $k \in \mathcal{N}$. Using this and $\delta_{k,t}$ indicator which is zero or one depending on whether action k was played at time t. Note

that,

$$
\begin{aligned}
R_T \quad & \leq \quad \sum_{t=1}^{T}(x)^\top\theta - x^\top\theta = \sum_{t=1}^{T}(x)^\top\theta - x_t^\top\tilde{\theta}_t + x_t^\top\tilde{\theta}_t - x_t^\top\theta \\
& \overset{\text{def. UCB}}{\leq} \quad \sum_{t=1}^{T} x_t^\top(\tilde{\theta}_t - \theta) = \sum_{t=1}^{T}\sum_{k=1}^{K}\delta_{k,t}x_t^\top(\tilde{\theta}_t - \theta) \leq \sum_{t=1}^{T}\sum_{k=1}^{K}\delta_{k,t}|x_t^\top(\tilde{\theta}_t - \theta)| \\
& = \quad \sum_{t=1}^{T}\sum_{k=1}^{K}\delta_{k,t}|x_t^\top(\tilde{\theta}_t - \theta)\mathbf{W}_{t,k}^{1/2}\mathbf{W}_{t,k}^{-1/2}| = \sum_{t=1}^{T}\sum_{k=1}^{K}\delta_{k,t}|x_t^\top(\tilde{\theta}_t - \theta)\mathbf{W}_{t,k}^{1/2}||\mathbf{W}_{t,k}^{-1/2}| \\
& = \quad \sum_{t=1}^{T}\sum_{k=1}^{K}\delta_{k,t}\underbrace{\sqrt{x_k^\top\mathbf{V}_t^{-1}x_k}}_{\mathbf{W}_{t,k}^{-1/2}}\left\|x_k^\top(\tilde{\theta}_t - \theta)\right\|_{\mathbf{W}_{t,k}} \leq \sum_{t=1}^{T}\sum_{k=1}^{K}\delta_{k,t}\sqrt{x_k^\top\mathbf{V}_t^{-1}x_k}\left\|x_k^\top(\tilde{\theta}_t - \theta)\right\|_{\mathbf{\Omega}_{t,k}} \\
& \overset{\text{Thm. 3}}{\leq} \quad \sum_{t=1}^{T}\sum_{k=1}^{K}\delta_{k,t}\sqrt{x_k^\top\mathbf{V}_t^{-1}x_k}\beta_{t,k} \\
& \overset{\text{Lemma 4}}{\leq} \quad \sum_{t=1}^{T}\sum_{k=1}^{K}\delta_{k,t}\left\|x_k\right\|_{\mathbf{V}_t^{-1}}\sqrt{\log(|\mathcal{X}|T/\delta)} \\
& \overset{\text{Potential Lemma}}{\leq} \quad \sqrt{\log(|\mathcal{X}|T/\delta)}\sum_{t=1}^{T}\left\|x_t\right\|_{\mathbf{V}_t^{-1}} \leq \mathcal{O}(\sqrt{dT\log((|\mathcal{X}|+1)T/\delta)})
\end{aligned}
$$

which finishes the proof. The $\tilde{\theta}_t$ is the vector that corresponds to the UCB action played by the algorithm. We took the union bound for the adaptive confidence bounds for each $x \in \mathcal{X}$, and hence the logarithmic dependence on $|\mathcal{X}|$. □

H Matrix Algebra Results

Definition 7 (Matrix slices and weighted slices)**.**

$$\mathbf{M}_{SS} \stackrel{\text{def}}{=} \mathbf{I}_{:S}^\top \mathbf{M} \mathbf{I}_{:S} \quad \text{and} \quad \mathbf{M}_{S\eta>0} = \mathbf{D}(\eta > 0)\mathbf{M}\mathbf{D}(\eta > 0) \tag{55}$$

$$\mathbf{M}_{S} \stackrel{\text{def}}{=} \mathbf{I}_{:S} \mathbf{M}_{SS} \mathbf{I}_{:S}^\top \quad \text{and} \quad \mathbf{M}_{S\eta} = \mathbf{D}(\eta)\mathbf{M}\mathbf{D}(\eta) \tag{56}$$

Lemma 5 (Zhang (2011))**.** *Let M be a positive definite matrix, and S be a subset of $[n]$, then*

$$(\mathbf{M}_S)^\dagger \preceq (\mathbf{M}^{-1})_S \tag{57}$$

Lemma 6 (Zhang (2011))**.** *If $\mathbf{A} \succeq \mathbf{B}$ where $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{l\times l}$ then for any $X \in \mathbb{R}^{d\times l}$, $\mathbf{X}\mathbf{A}\mathbf{X}^\top \succeq \mathbf{X}\mathbf{B}\mathbf{X}^\top$.*

Lemma 7. *Let $\mathbf{A} \in \mathbb{R}^{k\times m}$, then the matrix $\mathbf{P} = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}$ is projection matrix.*

Proof. Its easy to check $\mathbf{P}^\top = \mathbf{P}$ and $\mathbf{P}^2 = \mathbf{P}$. □

Lemma 8. *Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ full rank, and $\mathbf{C} \in \mathbb{R}^{k\times d}$ with rank k. Then $\mathbf{C}\mathbf{A}\mathbf{C}^\top$ has rank k.*

Proposition 16. *Let $\mathbf{X} \in \mathbb{R}^{m\times n}$ s.t. $\min(m, n) \geq |\eta|$, where $\mathbf{X}$ is full rank,*

$$\mathbf{V}(\eta)^\dagger = \mathbf{X}^\top \mathbf{D}(\eta)((\mathbf{X}\mathbf{X}^\top)_{S\eta})^\dagger \mathbf{D}(\eta)((\mathbf{X}\mathbf{X}^\top)_{S\eta})^\dagger \mathbf{D}(\eta)\mathbf{X}$$

Also,

$$\mathrm{Tr}(\mathbf{V}^\dagger) = \mathrm{Tr}(\mathbf{D}(\eta)^{1/2}((\mathbf{X}\mathbf{X}^\top)_{S\eta})^\dagger \mathbf{D}(\eta)^{1/2})$$

If $n = |\eta|$, then

$$\mathbf{V}(\eta)^\dagger = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{D}(\eta)^{-1}(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X} = \mathbf{X}^\dagger \mathbf{D}(\eta)^{-1}(\mathbf{X}^\top)^\dagger. \tag{58}$$

Proof. Tedious verification of four pseudo-inverse criteria. □

H.1 Auxiliary

Lemma 9. *Let $\mathbf{A} \in \mathbb{R}^{l\times l} \succ 0$, and $\mathbf{X} \in \mathbb{R}^{d\times l}$ diagonal where $d \leq l$, then*

$$\mathbf{A} \succeq \mathbf{X}^\top(\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\top)^{-1}\mathbf{X}. \tag{59}$$

Proof. Let S be a set that selects exactly the non-zero elements of matrix $\mathbf{X}$ and further suppose w.l.o.g. the matrix is in such permutations that this block is right-upper most.

By Lemma 5, we know that $(\mathbf{A}_{SS})^{-1} \preceq (\mathbf{A}^{-1})_{SS}$. Due to the monotonicity of the operation (Lemma 6) $\mathbf{X}_{SS}(\mathbf{A}_{SS})^{-1}\mathbf{X}_{SS}^\top \preceq \mathbf{X}_{SS}(\mathbf{A}^{-1})_{SS}\mathbf{X}_{SS}^\top$. Inverse operator reverse the above inequality, hence,

$$\begin{aligned}
(\mathbf{X}_{SS}(\mathbf{A}_{SS})^{-1}\mathbf{X}_{SS}^\top)^{-1} &\succeq (\mathbf{X}_{SS}(\mathbf{A}^{-1})_{SS}\mathbf{X}_{SS}^\top)^{-1} && (60)\\
(\mathbf{X}_{SS}^{-\top}(\mathbf{A}_{SS})\mathbf{X}_{SS}^{-1}) &\succeq (\mathbf{X}_{SS}(\mathbf{A}^{-1})_{SS}\mathbf{X}_{SS}^\top)^{-1} && (61)\\
(\mathbf{Y}_{SS}^\top(\mathbf{A}_{SS})\mathbf{Y}_{SS}) &\succeq (\mathbf{X}_{SS}(\mathbf{A}^{-1})_{SS}\mathbf{X}_{SS}^\top)^{-1} && (62)\\
(\mathbf{A}_{SS}) &\succeq \mathbf{X}_{SS}^\top(\mathbf{X}_{SS}(\mathbf{A}^{-1})_{SS}\mathbf{X}_{SS}^\top)^{-1}\mathbf{X}_{SS} && (63)
\end{aligned}$$

where definite following matrix $\mathbf{Y}$ s.t. $\mathbf{Y}_{SS} = (\mathbf{X}_{SS})^{-1}$ and zero-otherwise.

Lastly, we know that $\mathbf{A}_S$ has eigenvalues strictly smaller than $\mathbf{A}$, and also note that $\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\top = \mathbf{X}_{SS}(\mathbf{A}^{-1})\mathbf{X}_{SS}^\top$. Thus applying Lemma 6 we can lift the expression to $l \times l$ matrices,

$$\begin{aligned}
(\mathbf{A}) \succeq \mathbf{I}_{:S}(\mathbf{A})_{SS}\mathbf{I}_{S:} \succeq \mathbf{I}_{:S}\mathbf{X}_{SS}^\top(\mathbf{X}_{SS}(\mathbf{A}^{-1})_{SS}\mathbf{X}_{SS}^\top)^{-1}\mathbf{X}_{SS}\mathbf{I}_{S:} &\succeq \mathbf{X}^\top(\mathbf{X}_{SS}(\mathbf{A}^{-1})_{SS}\mathbf{X}_{SS}^\top)^{-1}\mathbf{X}\\
&= \mathbf{X}(\mathbf{X}\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}
\end{aligned}$$

□

Theorem 17. *Let* $\mathbf{A} \in \mathbb{R}^{l\times l} \succ 0$*, and* $\mathbf{M} \in \mathbb{R}^{d\times l}$*, where* $d \leq l$*, and* $\operatorname{rank}(\mathbf{M}) = d$ *then*

$$\mathbf{M}^\top(\mathbf{M}\mathbf{A}^{-1}\mathbf{M}^\top)^{-1}\mathbf{M} \preceq \mathbf{A} \tag{64}$$

Proof. Let $\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{d\times d}$ and $\mathbf{V} \in \mathbb{R}^{l\times l}$, be an SVD decomposition of $\mathbf{M}$ and $\mathbf{A} = \mathbf{R}^\top\mathbf{\Lambda}\mathbf{R}$, where $\mathbf{R} \in \mathbb{R}^{l\times l}$. be eigendecomposition of $\mathbf{A}$.

We perform the set of following equivalent operations to reduce the problem to diagonal case,

$$\begin{aligned}
\mathbf{R}^\top\mathbf{\Lambda}\mathbf{R} &\succeq (\mathbf{U}\mathbf{S}\mathbf{V}^\top)^\top(\mathbf{U}\mathbf{S}\mathbf{V}^\top\mathbf{R}^\top\mathbf{\Lambda}^{-1}\mathbf{R}(\mathbf{U}\mathbf{S}\mathbf{V}^\top)^\top)^{-1}\mathbf{U}\mathbf{S}\mathbf{V}^\top && (65)\\
\mathbf{R}^\top\mathbf{\Lambda}\mathbf{R} &\succeq \mathbf{V}\mathbf{S}^\top\mathbf{U}^\top(\mathbf{U}\mathbf{S}\mathbf{V}^\top\mathbf{R}^\top\mathbf{\Lambda}^{-1}\mathbf{R}\mathbf{V}\mathbf{S}^\top\mathbf{U})^{-1}\mathbf{U}\mathbf{S}\mathbf{V}^\top && (66)\\
\mathbf{R}^\top\mathbf{\Lambda}\mathbf{R} &\succeq \mathbf{V}\mathbf{S}^\top(\mathbf{S}\mathbf{V}^\top\mathbf{R}^\top\mathbf{\Lambda}^{-1}\mathbf{R}\mathbf{V}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{V}^\top && (67)\\
\mathbf{\Lambda} &\succeq \mathbf{G}\mathbf{S}^\top(\mathbf{S}\mathbf{G}^\top\mathbf{\Lambda}^{-1}\mathbf{G}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{G}^\top && (68)\\
\mathbf{B} &\succeq \mathbf{S}^\top(\mathbf{S}\mathbf{B}^{-1}\mathbf{S}^\top)^{-1}\mathbf{S} && (69)
\end{aligned}$$

where $\mathbf{V} = \mathbf{R}^\top\mathbf{G}$ is a new rotation matrix, and $\mathbf{B} = \mathbf{G}^\top\mathbf{\Lambda}\mathbf{G} = \mathbf{V}^\top\mathbf{R}^\top\mathbf{\Lambda}\mathbf{R}\mathbf{V} = \mathbf{V}^\top\mathbf{A}\mathbf{V}$. The rest follows from Lemma 9. □

Proposition 18. *Let* $\mathbf{X} \in \mathbb{R}^{n\times m}$*, and* $\mathbf{M} \in \mathbb{R}^{k\times m}$*, where* $k \leq m$*, and* $\operatorname{rank}(\mathbf{M}) = k$ *then if* $\exists\mathbf{A}$ *s.t.* $\mathbf{M} = \mathbf{A}\mathbf{X}$*, where* $\mathbf{A} \in k \times n$ *and* $k \leq n$*, the following holds,*

$$\mathbf{M}^\top(\mathbf{M}(\mathbf{X}^\top\mathbf{X})^+\mathbf{M}^\top)^{-1}\mathbf{M} \preceq \mathbf{X}^\top\mathbf{X}. \tag{70}$$

Proof. The pseudo-inverse of $(\mathbf{X}^\top\mathbf{X})^+ = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-2}\mathbf{X}$. Consequently,

$$\begin{aligned}
\mathbf{M}^\top(\mathbf{M}(\mathbf{X}^\top\mathbf{X})^+\mathbf{M}^\top)^{-1}\mathbf{M} &= \mathbf{X}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{X}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-2}\mathbf{X}\mathbf{X}^\top\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{X} && (71)\\
&= \mathbf{X}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{X} \preceq \mathbf{X}^\top\mathbf{X} && (72)
\end{aligned}$$

where the last line follows from the properties of a projection. □

H.2 Main Linear Algebra Theorem

The previous proposition has rather restrictive conditions. Namely, it requires the existence of a solution to k linear systems that are overdetermined. In particular m equations for n unknowns. A consequence of an approximate solution to such a system with a bounded error matrix is the content of the following theorem.

Theorem 19. *Let* $\mathbf{X} \in \mathbb{R}^{n\times m}$*, and* $\mathbf{C} \in \mathbb{R}^{k\times m}$*, where* $k \leq m$*, and* $\operatorname{rank}(\mathbf{C}) = k$ *then if the least squares solution* $\mathbf{A} \in \mathbb{R}^{k\times n}$ *satisfies* $\|\mathbf{C} - \mathbf{A}\mathbf{X}\|_F^2 \leq \delta^2 \|\mathbf{A}\|_F^2$ *and is full rank, the following holds,*

$$\mathbf{C}^\top(\mathbf{C}(\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{C}^\top)^{-1}\mathbf{C} \preceq (1 + \delta/\|\mathbf{X}\|_F)\mathbf{X}^\top\mathbf{X} + \delta^2\mathbf{I}. \tag{73}$$

Proof. Without loss of generality (as $n \geq k$), we assume that $\mathbf{C} - \mathbf{A}\mathbf{X} = \mathbf{A}\mathbf{\Sigma}$, where $\mathbf{\Sigma} \in \mathbb{R}^{n\times m}$. The condition in the statement is implied when $\|\mathbf{\Sigma}\|_F^2 \leq \delta^2$.

Due to $\mathbf{A}$, being OLS estimate, $\mathbf{A} = \mathbf{C}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}$. To select $\mathbf{\Sigma}$ we need to solve,

$$\mathbf{C} - \mathbf{C}(\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}) = \mathbf{C}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{\Sigma}$$

A specific least Frobenious norm choice for $\mathbf{\Sigma}$ leads to

$$\mathbf{\Sigma} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{C}^\top(\mathbf{C}\mathbf{V}^\dagger\mathbf{C}^\top)^{-1}\mathbf{C}(\mathbf{I}-\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}) = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{C}^\top(\mathbf{C}\mathbf{V}^\dagger\mathbf{C}^\top)^{-1}\mathbf{C}(\mathbf{I}-\mathbf{P}_\mathbf{X}),$$

for which $\mathbf{X}\mathbf{\Sigma}^\top = 0_{n\times n}$ and $\mathbf{\Sigma}\mathbf{X}^\top = 0_{n\times n}$. In the above, the pseudo-inverse of $\mathbf{V}^\dagger = (\mathbf{X}^\top\mathbf{X})^\dagger = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-2}\mathbf{X}$.

Consequently using $\mathbf{C} = \mathbf{A}(\mathbf{X} + \mathbf{\Sigma})$,

$$
\begin{aligned}
\mathbf{C}^\top(\mathbf{C}(\mathbf{X}^\top\mathbf{X})^+\mathbf{C}^\top)^{-1}\mathbf{C} &= (\mathbf{X}^\top+\mathbf{\Sigma}^\top)\mathbf{A}^\top(\mathbf{A}(\mathbf{X}+\mathbf{\Sigma})\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-2}\mathbf{X}(\mathbf{X}^\top+\mathbf{\Sigma}^\top)\mathbf{A}^\top)^{-1}\mathbf{A}(\mathbf{X}+\mathbf{\Sigma}) \\
&\overset{(64)}{\preceq} (\mathbf{X}^\top+\mathbf{\Sigma}^\top)[(\mathbf{X}+\mathbf{\Sigma})\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-2}\mathbf{X}(\mathbf{X}^\top+\mathbf{\Sigma}^\top)]^{-1}(\mathbf{X}+\mathbf{\Sigma}) \\
&= (\mathbf{X}^\top+\mathbf{\Sigma}^\top)[\mathbf{I}_{n\times n}+\mathbf{\Sigma}\mathbf{S}+\mathbf{S}^\top\mathbf{\Sigma}^\top+\mathbf{\Sigma}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-2}\mathbf{X}\mathbf{\Sigma}^\top]^{-1}(\mathbf{X}+\mathbf{\Sigma})
\end{aligned}
$$

where $\mathbf{S}=\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}$. Due to properties above, this simplifies to

$$
\begin{aligned}
\mathbf{C}^\top(\mathbf{C}(\mathbf{X}^\top\mathbf{X})^+\mathbf{C}^\top)^{-1}\mathbf{C} &\preceq (\mathbf{X}^\top+\mathbf{\Sigma}^\top)(\mathbf{X}+\mathbf{\Sigma}) \\
&= \mathbf{X}^\top\mathbf{X}+\mathbf{\Sigma}^\top\mathbf{X}+\mathbf{X}^\top\mathbf{\Sigma}+\mathbf{\Sigma}^\top\mathbf{\Sigma} \\
&\preceq \mathbf{X}^\top\mathbf{X}+\delta^2\mathbf{I}+\mathbf{\Sigma}^\top\mathbf{X}+\mathbf{X}^\top\mathbf{\Sigma} \\
&\preceq (1+\delta/\left\|\mathbf{X}\right\|_F)\mathbf{X}^\top\mathbf{X}+\delta^2\mathbf{I}
\end{aligned}
$$

The last step in the theorem follows by following claim. We look for a specific $\mathbf{\Sigma}$ s.t. norm constraint is satisfied as well as the expression is maximized in terms of Löwner ordering. Without loss of generality, there exists $\mathbf{B}\in\mathbb{R}^{n\times n}$ s.t. $\mathbf{\Sigma}=\mathbf{B}\mathbf{X}$, thus,

$$
\max_{\mathbf{B},\|\mathbf{B}\mathbf{X}\|_F^2\le\delta^2}\mathbf{X}^\top(\mathbf{B}+\mathbf{B}^\top)\mathbf{X}\preceq\mathbf{X}^\top\mathbf{X}\max_{\mathrm{Tr}(\mathbf{X}^\top\mathbf{B}^\top\mathbf{B}\mathbf{X})\le\delta^2}\lambda_{\max}(\mathbf{B}+\mathbf{B}^\top), \tag{74}
$$

where max is understood in Löwner ordering. We relax the constraint of the last objective,

$$
\begin{aligned}
\max_{\mathrm{Tr}(\mathbf{B}^\top\mathbf{B})\le\delta^2/\|\mathbf{X}\|_F^2}\lambda_{\max}(\mathbf{B}+\mathbf{B}^\top) &= \max_{\mathrm{Tr}(\mathbf{\Lambda}^\top\mathbf{\Lambda})\le\delta^2/\|\mathbf{X}\|_F^2,\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}}\lambda_{\max}(\mathbf{\Lambda}\mathbf{Q}+\mathbf{Q}^\top\mathbf{\Lambda}) \\
&= \max_{\mathrm{Tr}(\mathbf{\Lambda}^\top\mathbf{\Lambda})\le\delta^2/\|\mathbf{X}\|_F^2}\lambda_{\max}(2\mathbf{\Lambda}) \\
&= \delta/\left\|\mathbf{X}\right\|_F
\end{aligned}
$$

□