The Non-linear F-Design and Applications to Interactive Learning

Alekh Agarwal [1] **Jian Qian** [2] **Alexander Rakhlin** [2] **Tong Zhang** [3]

Abstract

We propose a generalization of the classical G-optimal design concept to non-linear function classes. The criterion, termed *F-design*, coincides with G-design in the linear case. We compute the value of the optimal design, termed the *F-condition number*, for several non-linear function classes. We further provide algorithms to construct designs with a bounded F-condition number. Finally, we employ the F-design in a variety of *interactive machine learning* tasks, where the design is naturally useful for data collection or exploration. We show that in four diverse settings of *confidence band construction, contextual bandits, model-free reinforcement learning, and active learning*, F-design can be combined with existing approaches in a black-box manner to yield state-of-the-art results in known problem settings as well as to generalize to novel ones.

1. Introduction

Developed by pioneers like Fisher and Wald, experimental design is a foundational area of study in statistics and machine learning (Cochran & Cox, 1948; Montgomery, 2017; Casella et al., 2008; Pronzato & Pázman, 2013). In contrast to vanilla statistical learning, where the data are assumed to be sampled from a fixed and unknown distribution, experimental design allows the learner to choose a distribution over the underlying input space to their advantage. Good experimental designs form the bedrock of efficient data collection and learning in sample-constrained scenarios, with applications in social and natural sciences, healthcare settings, and beyond (Jackson & Cox, 2013; Keskin Gündoğdu et al., 2016). More recently, the increasing popularity of interactive paradigms such as active learning, bandits, and reinforcement learning have also made experimental design a key component of machine learning algorithms (Lattimore & Szepesvári, 2020; Wagenmaker et al., 2021; Foster et al., 2021; Agarwal & Zhang, 2022; Mhammedi et al., 2023). The classical theory of experimental design is most well-developed when the experimental data is subsequently modeled using (generalized) linear functions. In such cases, the target is to obtain closed-form expressions or nearly-sharp bounds for a number of definitions of an optimal design, along with efficient algorithms for computing them.

The core intuition behind experimental design, when we subsequently fit data using functions from a particular class, is the following. We want a small subset of the inputs, such that querying the values of the desired outputs at these instances allows us to (approximately) extrapolate to the rest of the domain, for any function in our class. That is, the set of query instances serves as a basis tailored to the function class. When the function class has an underlying linear structure, the Euclidian basis is always a sufficient (and without assumptions on the domain, an optimal) design.

When the functions under consideration are non-linear, finding such a minimal set is quite non-trivial, as different functions can have varying degrees of sensitivity in different parts of the instance space. Consequently, much of the literature on non-linear experimental design has focused on *local* or *asymptotic* design criteria. In such approaches, we assume that the functions under consideration have already been localized to a “small” neighborhood of the optimal solution, and we seek a good design for this local region of the parameter space, often using Taylor or other expansions to reduce the local design problem to a linear one. However, the availability of such a local region is a stringent assumption in practice, where no prior knowledge might be available until we carry out data collection. This is a particularly pronounced concern in interactive ML scenarios, where we often start with no initial samples and seed good models using a limited amount of data.

One notable exception for the non-linear case is the eluder dimension, which was originally proposed in Russo & Van Roy (2013) and further refined in several works (Jin et al., 2021; Foster et al., 2020). Eluder dimension is a complexity measure of a function class with a similar perspective of quantifying how many samples suffice to reliably evaluate some measure of the fitness of functions from a class over the entire input space, and we discuss it in more

[1]Google [2]MIT [3]UIUC. Correspondence to: Alekh Agarwal <alekhagarwal@google.com>, Jian Qian <jianqian@mit.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

	Simultaneous Confidence Bands	Contextual Bandit (realizable + infinite actions)	Model-free RL (realizable + Bellman rank + infinite actions)	Active Learning
Objective	$\lvert\widehat{f}(x) - f^*(x)\rvert$	PAC Regret	PAC Regret	Query complexity
Bounds	$\mathcal{O}(\sqrt{\mathcal{V}(\rho^*, x)\lVert\widehat{f} - f^*\rVert^2_{\rho^*}})$	$\mathcal{O}(\sqrt{\mathcal{V}^* \log\lvert\mathcal{F}\rvert/T})$	$\mathcal{O}(\log(\lvert\mathcal{F}\rvert T)\sqrt{\mathsf{br}\cdot\mathcal{V}^* H/T})$	$\mathcal{O}(\mathcal{V}^* \log\lvert\mathcal{F}\rvert)$

Table 1. Illustration of the applications enabled by F-design. $\mathcal{V}^*$ denotes the optimal F-condition number for some underlying function class $\mathcal{F}$ that is used in the application, ρ^* is the optimal F-design and $\mathcal{V}(\rho^*, x)$ denotes the F-condition number of ρ^* on x. For example, $\mathcal{V}^* = d$ when $\mathcal{F}$ is linear. In simultaneous confidence bands, we have regression data sampled from ρ^* and labeled according to an unknown $f^* \in \mathcal{F}$ and want error bounds at all x, around some estimator $\widehat{f}$ obtained by squared regression. In contextual bandits, we consider the realizable setting where expected rewards are equal to $f^\star \in \mathcal{F}$, and control the regret of the learned policy after T rounds of exploration, in terms of optimal F-condition number for potentially infinite action spaces. In RL, the class $\mathcal{F}$ models Bellman errors for an MDP with Bellman rank br, and we bound the PAC regret of the learner's policy after observing T episodes, to the optimal policy, again in infinite action settings. In active learning, we use $\mathcal{F}$ to model the value of the binary label in a noiseless setting and bound the query complexity to find a consistent classifier. All our bounds match the best results for the linear class using $\mathcal{V}^* = d$, but the extension to the general non-linear case is novel. New results for RL are discussed in the text.

detail in later technical development. A more detailed discussion of related works that tailor design to several of the specific applications that we consider, as well as the broader classical literature on design can be found in Appendix A.

Motivated by the questions above, we take a fresh look at experimental design in its full generality in this work. We provide a novel definition of *F-design* that generalizes the classical idea of G-design to arbitrary, non-linear function classes. We also demonstrate the benefits of this generalization with applications in diverse learning scenarios. Concretely, our paper makes the following contributions:

- We develop the **F-condition number** to measure the quality of a design for non-linear function classes. We show that the classical G-optimal design optimizes F-condition number when the function class is linear, and the criterion admits natural bounds for non-linear classes such as Lipschitz or smooth functions. We further show that the F-condition number can always be upper bounded in terms of the eluder dimension up to log terms, but can be *exponentially smaller* than it for general function classes.
- We give **computationally efficient algorithms**, under a natural argmax oracle, that find a design with an upper bounded F-condition number.
- We demonstrate **applications** of our techniques in a variety of learning settings. Concretely, we show that the optimal F-condition number can be used to obtain state-of-the-art guarantees for *contextual bandits, reinforcement learning, and active learning*, where we use an optimal F-design to guide the exploration process. We also show applications to *estimating simultaneous confidence bands* in non-linear regression settings. Given that these applications are some of the key beneficiaries of classical experimental design, these results highlight the ability of F-design to serve as a drop-in replacement in the non-linear setting. Table 1 summarizes the results we get for these applications.
- Our applications are investigated in most detail for the case of **model-free RL**, where we match state-of-the art results in most prior settings, and also provide new results for sample-efficient RL in some previously unknown settings that are summarized below, with details in Section 5.3:
  - Bellman rank with linearly embedded backups (includes finite actions, known from (Agarwal & Zhang, 2022)).
  - Bellman rank with $Q_\star$ smooth over actions (**new**).
  - Value-based Linear Quadratic Regulators (also implied by Agarwal & Zhang (2022)).

2. Preliminaries

We start with a discussion of the experimental design paradigm and set up some basic notation. In experimental design, the learner is given the set of instances $\mathcal{X}$, which is also referred to as the *design space*, where each instance is viewed as an elementary experiment that can reveal some information about the environment. Querying (or, "experimenting with") an instance $x \in \mathcal{X}$ generates a stochastic observation $y = f^*(x) + \epsilon$ for some unknown f^* in a given function class $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ and a zero-mean ϵ. The goal is to select a distribution (or, *a design*) $\rho \in \mathcal{P}(\mathcal{X})$ on the design space $\mathcal{X}$ to optimize a criterion, to be specified shortly. Here $\mathcal{P}(\mathcal{X})$ denotes the set of probability measures on $\mathcal{X}$ with an appropriately defined σ-algebra. For simplicity, we focus on the problem with a finite function class and a finite design space for the most part, though our ideas easily extend to more general settings with covering numbers, as some of our examples illustrate. Various design criteria are considered in the literature, of which the most relevant is the classical G-optimal design, proposed in Smith (1918).

G-optimal design. For any bounded $\mathcal{X} \subset \mathbb{R}^d$, the G-design value for any distribution ρ on $\mathcal{X}$ is defined as

$$\sup_{x\in\mathcal{X}} x^\top \left(\mathbb{E}_{z\sim\rho} zz^\top\right)^\dagger x. \tag{1}$$

where $(\cdot)\dagger$ denotes the pseudo inverse (Pronzato & Pázman, 2013). The G-design value corresponds to the maximum asymptotic variance of the least square predictor over the design space $\mathcal{X}$. The optimal design is the distribution that minimizes the design value. The following properties are well-known for the G-design. The optimal design value is smaller than or equal to the dimension d (with equality obtained when linear combinations of $\mathcal{X}$ span $\mathbb{R}^d$) (Kiefer & Wolfowitz, 1960). There exists an optimal design that is supported on at most $\widetilde{\mathcal{O}}(d)$ instances (Todd, 2016). There is a Frank-Wolfe style algorithm that can obtain a design with a value matching the optimal design value up to constant within $\widetilde{\mathcal{O}}(d)$ number of steps with access to a maximization oracle over $x \in \mathcal{X}$ for (1)(Todd, 2016).

For the general non-linear case, the notion of G-optimal design is also studied (Pronzato & Pázman, 2013) with an emphasis on local asymptotic optimality. However, the properties of the linear G-optimal design aforementioned are not known for the general non-linear extension.

Eluder dimension. In the spirit of controlling worst-case prediction errors through design, another notable complexity measure is the eluder dimension, defined below.

Definition 2.1 (Eluder dimension (Russo & Van Roy, 2013)). For any function class $\mathcal{F}$ and $\epsilon_0 > 0$, let the eluder dimension $\mathrm{dimE}(\mathcal{F};\epsilon_0)$ be the length of the longest sequence of tuples $(x_1, f_1, f_1'), ..., (x_m, f_m, f_m')$ such that there exists $\epsilon \geq \epsilon_0$ and functions $f_i, f_i' \in \mathcal{F}$ for $i = 1, ..., m$ that

$$|f_i(x_i) - f_i'(x_i)| > \epsilon, \text{ and } \sum_{j<i}(f_i(x_j) - f_i'(x_j))^2 \leq \epsilon^2.$$

Suppose $(x_1, f_1, f_1'), ..., (x_m, f_m, f_m')$ are the longest sequence as in the above definition. The design $\rho_E = \frac{1}{m}\sum_{i=1}^m \mathbb{1}(x_i)$, where $\mathbb{1}(x_i)$ is the point distribution supported on x_i, has the property that for any $f, f' \in \mathcal{F}$, if $\mathbb{E}_{x\sim\rho_E}(f(x) - f'(x))^2 \leq \epsilon_0^2/m$, then $\sup_x |f(x) - f'(x)| \leq \epsilon_0$. This implication of controlling the infinity norm through 2-norm by the eluder dimension captures the number of "directions" for the linear case. But the ratio between the two norms can be *exponentially larger* than the minimal design, as we illustrate in the next section.

Other complexity measures. Other complexity measures of function classes like the VC dimension, Radermacher complexity, Decision-Estimation Coefficient and Decoupling Coefficient (Foster et al., 2021; Zhang, 2022; Zhong et al., 2022) are tailored to specific learning tasks and not comparable in the guarantees obtained. We contrast with these criteria for suitable applications later in the paper. See more detailed discussion in Section 6.

Notation We denote by $\mathbb{R}_{\geq 0} = [0, \infty)$. For any integer $T \geq 1$, we denote by $[T]$ the set $\{1, 2, ..., T\}$. For any positive semi-definite matrix $A \in \mathbb{R}^{d\times d}$, the norm $\|\cdot\|_A$ on $\mathbb{R}^d$ is defined as $\|x\|_A^2 := x^\top Ax$. We use $\mathbb{1}(\mathcal{E})$ to denote the indicator function for event $\mathcal{E}$. For any set $\mathcal{X}$, we abuse the notation $\mathbb{1}(x)$ to also mean the point distribution supported on $x \in \mathcal{X}$. For any set $\mathcal{X}$ and function $f : \mathcal{X} \to \mathbb{R}$, we denote by $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$ the infinity norm of f. For any distribution $\rho \in \mathcal{P}(\mathcal{X})$ and function f, we denote by $\|f\|_{\rho,2}^2 = \mathbb{E}_{x\sim\rho}(f(x))^2$ the 2-norm of f with respect to ρ. We define $\mathcal{O}(\cdot)$, $\Omega(\cdot)$, $o(\cdot)$, $\Theta(\cdot)$, $\widetilde{\mathcal{O}}(\cdot)$, $\widetilde{\Omega}(\cdot)$, $\widetilde{\Theta}(\cdot)$ following standard non-asymptotic big-oh notation.

3. General Formulation of Non-linear Experiment Design

We now introduce the F-condition number (Section 3.1) and show that it coincides with the G-optimal design criterion in the linear case. We also give some examples where the optimal F-condition number can be computed up to constants (Section 3.2), and compare the optimal F-condition number with the eluder dimension (Section 3.3).

3.1. The F-condition Number and F-design

Let $\mathcal{DF} = \mathcal{F} - \mathcal{F} := \{f - f' : f, f' \in \mathcal{F}\}$. Let $\nu_0 : \mathcal{DF} \to \mathbb{R}_{\geq 0}$ be a base functional which is a map from the class $\mathcal{DF}$ to non-negative values. For any function class $\mathcal{F}$ with design space $\mathcal{X}$, any base functional ν_0, function $h \in \mathcal{DF}$, design $\rho \in \mathcal{P}(\mathcal{X})$ and instance x, the F-condition number of ρ on x with respect to h is defined as

$$\mathcal{V}(\mathcal{F}, \mathcal{X}, \rho, x, h; \nu_0) := \frac{h^2(x)}{\nu_0(h) + \mathbb{E}_{x'\sim\rho} h^2(x')}.$$

(F-condition number)

With $h = f - f'$ for some $f, f' \in \mathcal{F}$, the numerator in this definition captures the squared error at x, when $f^* = f'$ and we instead use an estimate f. The second term in the denominator is the same squared error but under the design ρ. Hence, the ratio relates how well controlling the squared error under ρ bounds the squared error on some other query point x, and captures the ability of ρ to cover the space $\mathcal{X}$ effectively from the perspective of the class $\mathcal{F}$. The term $\nu_0(h)$ is added for regularity to ensure that the ratio stays bounded away from zero, and is often set to a small positive constant independent of h in our applications.

The F-condition number of ρ and the optimal F-condition number $\mathcal{V}^*$ are defined by taking supremum over all instances $x \in \mathcal{X}$ and $h \in \mathcal{DF}$:

$$\mathcal{V}(\mathcal{F}, \mathcal{X}, \rho; \nu_0) := \sup_{x\in\mathcal{X}, h\in\mathcal{DF}} \mathcal{V}(\mathcal{F}, \mathcal{X}, \rho, x, h; \nu_0) \quad \text{and}$$

$$\mathcal{V}^*(\mathcal{F}, \mathcal{X}; \nu_0) := \min_{\rho\in\mathcal{P}(\mathcal{X})} \mathcal{V}(\mathcal{F}, \mathcal{X}, \rho; \nu_0),$$

We omit the dependence on $\mathcal{X}$ for simplicity when it is apparent from the context. Our design criteria is to minimize $\mathcal{V}(\mathcal{F}, \rho; \nu_0)$, and the optimal F-design is the minimizer

$$\rho_{\mathcal{V}}(\mathcal{F}; \nu_0) := \operatorname*{argmin}_{\rho \in \mathcal{P}(\mathcal{X})} \mathcal{V}(\mathcal{F}, \rho; \nu_0). \tag{2}$$

One powerful implication of (2) is that for any $f, f' \in \mathcal{F}$:

$$\|f - f'\|_\infty \le \mathcal{V}^*(\mathcal{F}, \mathcal{X}; \nu_0)\, \mathbb{E}_{x \sim \rho_{\mathcal{V}}} \left[(f(x) - f'(x))^2 \right],$$

which gives us a "condition number" to convert ℓ_2 to ℓ_∞ error bounds. We note that one might expect a similar conclusion from the eluder design ρ_E using $\dim\mathrm{E}(\mathcal{F}; \epsilon_0)$ instead, but the specifics of Definition 2.1 do not allow this. In Section 5.1, we utilize a sharper version of this bound to obtain simultaneous confidence bands for regression. As a final remark on notation, we use $\rho_{\mathcal{V}}(\mathcal{F}; \epsilon_0)$ to denote the case where $\nu_0(h) \equiv \epsilon_0$ for all $h \in \mathcal{DF}$.

3.2. Examples of F-design

We start by showing that the optimal F-design as defined in Equation 2 is equivalent to the classical G-optimal design for the linear function class.

Example 1 (Linear class). Let $\mathcal{F}^{\mathsf{lin}} = \{f_w : f_w(x) = w^\top x, w \in \mathbb{R}^d, \|w\|_2 \le 1\}$, $\mathcal{X} = \{x : \|x\|_2 \le 1\|\}$, and consider the base functional ν_{lin} given by $\nu_{\mathsf{lin}}(h) = \lambda \mathbb{E}_{x \sim \rho_0} h^2(x)$, with ρ_0 being the uniform distribution on the canonical basis vectors $\{e_i\}_{i=1}^d$, for some $\lambda > 0$ and $h \in \mathcal{DF}$. We have $\mathcal{V}^*(\mathcal{F}^{\mathsf{lin}}; \nu_{\mathsf{lin}}) \le d$.

Lemma 3.1. *The optimal F-design $\rho_{\mathcal{V}}(\mathcal{F}^{\mathsf{lin}}; \nu_{\mathsf{lin}})$ also minimizes the G-design objective* (1) *with a regularization λ/d:*

$$\rho_{\mathcal{V}}(\mathcal{F}^{\mathsf{lin}}; \nu_{\mathsf{lin}}) = \operatorname*{argmin}_{\rho \in \mathcal{P}(\mathcal{X})} \sup_{x \in \mathcal{X}} \|x\|^2_{(\Sigma_\rho(\lambda/d))^{-1}},$$

where $\Sigma_\rho(a) = \mathbb{E}_{x \sim \rho} x x^\top + a I_d$ and I_d is the identity matrix.

Combining Lemma 3.1 with the classical results of Kiefer & Wolfowitz (1960), we know that the optimal F-condition number is bounded by the dimension d in the linear case.

Example 2 (Hölder classes). Let U be any bounded set in $\mathbb{R}^d$ with positive volume and let β be a positive number. For any index $\mathbf{k} = (k_1, ..., k_d) \in \mathbb{Z}^d_{\ge 0}$, the partial derivative operator $D^{\mathbf{k}} := \partial^k / \partial^{k_1}_{x_1} \cdots \partial^{k_d}_{x_d}$ is termed as of order k The Hölder class $\mathcal{F}^{\mathsf{H}}_{\beta,d}$ is defined as the set of $k = \lfloor \beta \rfloor$ times differentiable functions f, whose partial derivatives $D^{\mathbf{k}} f$ of order k satisfy $|D^{\mathbf{k}} f(x) - D^{\mathbf{k}} f(y)| \le \|x - y\|^{\beta - k}$ for all $x, y \in U$ and all partial derivatives of order less than k are bounded by 1. For any $\epsilon_0 \ge 0$ as a constant function, we have $\mathcal{V}^*(\mathcal{F}^{\mathsf{H}}_{\beta,d}; \epsilon_0) = \Theta(\epsilon_0^{-d/(2\beta+d)})$.

The near-optimal design involves a two-stage construction where we have a uniform design over a covering of the design space and a local G-optimal design via linearization for each local region. This approach is pretty representative, as shown in the next example.

Example 3 (Fractional related class). Let $B > 2$ be a positive number. Consider the design space $\mathcal{X} = [-B, B] \subset \mathbb{R}$ and function class $\mathcal{P}_{k,B} := \{f(x) = p(x)/(1 + x^2) \mid p(x)$ is a polynomial with degree at most $k.\}$. We have $\mathcal{V}^*(\mathcal{P}_{k,B}; \epsilon_0) = \mathcal{O}(k \log B)$.

For this function class, we use the G-optimal design to build an F-design following a similar two-stage construction as mentioned in the previous example.

Theorem 3.2 (General function class). *For any bounded function class $\mathcal{F}$ on any design space $\mathcal{X}$ and $\epsilon_0 > 0$, let $\epsilon > 0$ be such that $\epsilon_0 > 4\epsilon^2 / \mathcal{N}(\mathcal{DF}, \|\cdot\|_\infty, \epsilon)$, where $\mathcal{N}(\mathcal{DF}, \|\cdot\|_\infty, \epsilon)$ is the minimum covering number of the function class $\mathcal{F}$ in infinity norm. Then we have $\mathcal{V}^*(\mathcal{F}; \epsilon_0) = \mathcal{O}(\mathcal{N}(\mathcal{DF}, \|\cdot\|_\infty, \epsilon))$.*

While the log covering number commonly serves as a standard for measuring complexity, aligning with the F-condition number in the linear case and the Hölder classes, the F-condition number scales with the covering number in the worst case. This is unavoidable, as seen by choosing the function class $\mathcal{F}_{\mathcal{X}} := \{f_x(y) = \mathbb{1}(y = x) \mid x \in \mathcal{X}\} \subset (\mathcal{X} \to \mathbb{R})$ for any set $\mathcal{X}$.

Collectively, these examples highlight that the optimal F-condition number has an expected scaling with the complexity of the underlying function class, and is easily tailored to different underlying structures. We proceed in the next section to compare the optimal F-condition number with the eluder dimension.

3.3. Comparison with the Eluder Dimension

As discussed earlier, for (generalized) linear functions, there is only a logarithmic gap between the optimal F-condition number and the eluder dimension, as both scale with the ambient dimension. In the non-linear case, the optimal F-condition number is upper bounded by the eluder dimension up to log terms, but can be exponentially smaller.

Lemma 3.3. *For any function class $\mathcal{F} : \mathcal{X} \to [-1, 1]$ and $\epsilon_0 \in (0, 1/4)$, let $\dim\mathrm{E}(\mathcal{F}; \sqrt{\epsilon_0}) = d$. We have*

$$\mathcal{V}^*(\mathcal{F}; \epsilon_0) = \mathcal{O}\left(d \log(1/\epsilon_0) \cdot (\log d + \log\log(1/\epsilon_0))\right).$$

This is a corollary of Theorem 4.2. We get $\dim\mathrm{E}(\mathcal{F}; \sqrt{\epsilon_0})$ due to a difference between measuring $h^2(x)$ versus $h(x)$ at the worst-case x between F-condition number and eluder dimension. On the other hand, we show an exponential gap between the optimal F-condition number and the eluder dimension in the following cheating code example. For an upper bound of the F-condition number by the related disagreement coefficient (Foster et al., 2020), we refer to Proposition D.3 in Appendix D.4.

Example 4 (Cheating code (Jun & Zhang, 2020))**.** For any integer $k \geq 1$, suppose the design space $\mathcal{X} = \mathcal{A}^k \cup \mathcal{B}^k$ consists of two sets $\mathcal{A}^k = \{a_0, ..., a_{2^k-1}\}$ and $\mathcal{B}^k = \{b_0, ...b_{k-1}\}$, where the first has exponentially more arms than the second. Consider the function class given by $\mathcal{F}^k = \{f^i | i \in \{0, ..., 2^k - 1\}\}$ with

$$\begin{cases} f^i(a_j) = \mathbb{1}(i = j) & \text{for all } j \in \{0, ..., 2^k - 1\}, \\ f^i(b_l) = \frac{1}{2} \cdot (\text{the } l\text{-th bit of } i) & \text{for all } l \in \{0, \dots, k-1\}. \end{cases}$$

Lemma 3.4. *For the Example 4, let $\nu_0 > 0$ be any positive base functional and $\epsilon_0 \in (0, 1)$. We then have $\mathcal{V}^*(\mathcal{F}^k; \nu_0) \leq 4k$ and $\mathrm{dimE}(\mathcal{F}^k; \epsilon_0) \geq 2^k - 1$.*

We note that this exponential gap also exists for the disagreement coefficient and the star number (Foster et al., 2020). Interested readers can refer to Appendix D.4 for details.

Having established these favorable properties of F-design, we next investigate algorithms to construct designs with a good F-condition number.

4. Computational Properties

So far, we have motivated that the F-condition number provides a meaningful complexity measure tailored to the underlying function class. However, leveraging its properties in practice requires the ability to approximate a near-optimal design computationally. In this section, we focus on these computational aspects. Our computational results largely resemble those for the linear G-optimal design and we address the connections during our discussion. While Definition (2) considers all possible distributions ρ over the design space $\mathcal{X}$, representing and sampling from an arbitrarily dense distribution can be intractable when $|\mathcal{X}|$ is large. To this end, we begin with a basic sparsification argument that shows that any design's F-condition number can be well approximated with another design, that is only supported on a sparse subset of $\mathcal{X}$. We then proceed to give algorithms for constructing sparse designs with a bounded F-condition number under a natural computational oracle.

4.1. Existence of Sparse Designs

One of the most favorable properties of the linear G-optimal design is that the optimal design can be approximated in the design value by designs with sparse support (of size $\widetilde{\mathcal{O}}(d)$). We now show that any design can always be approximated by a sparse one, even in the non-linear case, through a probabilistic sparsification argument.

Lemma 4.1. *For any bounded function class $\mathcal{F}$, suppose that the base functional ν_0 is such that $\nu_0(h) > 0$ for all $h \in \mathcal{DF}$. Let $\delta \in (0, 1)$. Then for any design ρ if we sample $x_1, ..., x_n$ from ρ, with $n = \mathcal{O}(\mathcal{V}(\mathcal{F}, \rho; \nu_0) \log(|\mathcal{F}|/\delta))$, then with probability at least $1 - \delta$, we have $\mathcal{V}(\mathcal{F}, \hat{\rho}; \nu_0) \leq 4\mathcal{V}(\mathcal{F}, \rho; \nu_0)$, where $\hat{\rho} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(x_i)$.*

Algorithm 1 Greedy optimization of F-condition number

```
Require: Design space $\mathcal{X}$, function class $\mathcal{F}$, time horizon $T$, parameter $\epsilon_0 > 0$.
for $t = 1, \dots, T$ do
    Find $x_t \in \mathcal{X}$ such that
        $$x_t = \arg\max_{x \in \mathcal{X}} \sup_{h \in \mathcal{DF}} \frac{h^2(x)}{T\epsilon_0 + \sum_{s=1}^{t-1} h^2(x_s)}.$$
end for
return $\rho_T = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(x_t)$.
```

This argument allows us to restrict attention to finitely supported sparse designs computationally. We now describe such an algorithm and analyze the design it constructs.

4.2. Computation with Argmax Oracle

Another favorable property of the linear G-optimal design is that the optimal design can be approximated in design value up to constants efficiently with access to an argmax oracle. The argmax oracle takes an input design ρ and outputs the worst instance x under that design. For classical G-design, this is the x at which the norm in Equation 1 is maximized. We give the analogous definition for the general case below.

The argmax oracle The oracle computes the hardest instance x for any given design $\rho \in \mathcal{P}(\mathcal{X})$, i.e., it computes

$$x = \underset{x'}{\mathrm{argmax}} \sup_{h \in \mathcal{DF}} \mathcal{V}(\mathcal{F}, \rho, x', h; \nu_0).$$

This oracle is natural since it resembles the maximization oracle required by the Frank-Wolfe algorithm in the linear G-design case (Todd, 2016). Approximate oracles are sufficient for Algorithm 1 and implementation of such an oracle in the linear case can be found in Walsh (2022).

Using this oracle, Algorithm 1 gives a greedy approach to construct a design. The algorithm begins with an initially empty design, and at each iteration, finds the point x_t which is returned by the argmax oracle for the design ρ_{t-1}. The design ρ_t after t iterations is simply the uniform distribution over the points $\{x_i : i = 1, \dots, t\}$. This algorithm enjoys the following guarantee on the F-condition number of ρ_T.

Theorem 4.2. *For any function class $\mathcal{F} : \mathcal{X} \to [-1, 1]$, $\epsilon_0 \in (0, 1/4)$, let $\mathrm{dimE}(\mathcal{F}; \sqrt{\epsilon_0}) = d$ and let $T = Cd \log(1/\epsilon_0)(\log d + \log\log(1/\epsilon_0))$ for C large enough.*

Then Algorithm 1 returns a distribution ρ_T such that

$$\mathcal{V}(\mathcal{F}, \rho_T; \epsilon_0) = \mathcal{O}\left(d \log(1/\epsilon_0) \cdot (\log d + \log\log(1/\epsilon_0))\right).$$

In words, the F-condition number of ρ_T is bounded in terms of the eluder dimension of $\mathcal{F}$, but not $\mathcal{V}^*(\mathcal{F}; \epsilon_0)$ [1]. This guarantee is reasonable for classes like the (generalized) linear classes since there is no essential gap between $\mathrm{dimE}(\mathcal{F}; \epsilon_0)$ and $\mathcal{V}^*(\mathcal{F}; \epsilon_0)$, both of which scale with the dimension. On the other hand, it can be highly suboptimal in cases such as Example 4. As we show in Appendix E.3, the bound in Theorem 4.2 is *unimprovable* in the worst case for a class of algorithms which construct designs only supported on the convex hull of the points x_t that are returned from all previous queries to the argmax oracle. Nevertheless, the algorithm would run into trouble only when the x's returned by the argmax oracle are not the ones with the most information. Concretely, if one alters the cheating code example by setting the value for cheating arms to be $f^i(b_l) = 2 \cdot$ (the l-th bit of i), then the argmax oracle would return the cheating arms and obtain the F-condition number which is exponentially smaller than the eluder dimension in the altered cheating code.

Remark 4.3 (**Approximate optimality** via subgradient descent)**.** The F-condition number, as a supremum of convex functions with respect to the design, is convex in the design. Hence, with a slightly stronger argmax oracle (in that the supremum achieving functions are also required) defined in Appendix E.4, we can use projected subgradient descent to guarantee convergence to the optimal F-condition number $\mathcal{V}^*(\mathcal{F}; \epsilon_0)$. This approach requires $\Omega(|\mathcal{X}|)$ computation and hence is only suitable for small design spaces $\mathcal{X}$.

5. Applications

In this section, we study several applications of F-design.

5.1. Simultaneous Confidence Bands for Regression

We can use the F-design approach to compute the Simultaneous Confidence Bands (SCBs) for least squares regression. This is natural as the definition of F-design encodes transferring an average squared loss under the design ρ to a worst-case (ℓ_∞) error bound over the design space $\mathcal{X}$.

In the SCBs problem, the learner is given the function class $\mathcal{F}$ with the instance space $\mathcal{X}$. The learner can choose a design $\rho \in \mathcal{P}(\mathcal{X})$, and sample n samples $x_1, \ldots, x_n$ i.i.d. according to the design. The corresponding measurements y_i that the learner observes for each sample x_i, satisfy $\mathbb{E}[y_i | x_i] = f^\star(x_i)$ for some (unknown) $f^\star \in \mathcal{F}$. We further assume that $|f(x)| \le B$ for all $f \in \mathcal{F}$. The goal of the learner is to find a design ρ, such that for any estimator $\widehat{f}$ we can upper bound the worst-case prediction error $|\widehat{f}(x) - f^*(x)|$ over any query point $x \in \mathcal{X}$. The SCBs are first constructed in Bickel & Rosenblatt (1973) and are extensively studied in the literature (see e.g. (Johnston, 1982; Härdle, 1989; Xia, 1998; Wang & Yang, 2009; Cai et al., 2014; 2019) and references therein).

For this problem, we can choose the design as the optimal F-design $\rho_{\mathcal{V}} := \rho_{\mathcal{V}}(\mathcal{F}, \nu_0)$ according to (2) for some base functional ν_0. We further define the F-condition number of any design ρ on x to be $\mathcal{V}(\mathcal{F}, \rho, x; \nu_0) := \sup_{h \in \mathcal{D}\mathcal{F}} \mathcal{V}(\mathcal{F}, \rho, x, h; \nu_0)$ for the following result where we omit further the dependence on $\mathcal{F}$ and ν_0 for brevity.

Theorem 5.1. *Suppose $\sup_{\Delta f \in \mathcal{D}\mathcal{F}} \nu_0(\Delta f) \le \epsilon_0$. For the optimal F-design $\rho_{\mathcal{V}}$, we have for all $\widehat{f} \in \mathcal{F}$ and $x \in \mathcal{X}$,*

$$|\widehat{f}(x) - f^*(x)| = \mathcal{O}\left(\sqrt{\mathcal{V}(\rho_{\mathcal{V}}, x)(\epsilon_0 + \|\widehat{f} - f^*\|^2_{\rho_{\mathcal{V}},2})}\right).$$

For instance, if $\widehat{f} = \widehat{f}_n$ is the least squares estimator on n samples drawn i.i.d. according to $\rho_{\mathcal{V}}$, then we have $\|\widehat{f}_n - f^*\|^2_{\rho_{\mathcal{V}},2} = O(B \log(|\mathcal{F}|/\delta)/n)$, with probability $1 - \delta$. When applied to the optimal F-design for the class $\mathcal{F}^{\mathsf{lin}}$, this yields confidence intervals which scale as $\widetilde{\mathcal{O}}(d/\sqrt{n})$, since both $\mathcal{V}^*(\mathcal{F}^{\mathsf{lin}}, \mathcal{X})$ and $\mathcal{N}(\mathcal{F}^{\mathsf{lin}}, 1/n)$ (the ℓ_∞ covering number of $\mathcal{F}^{\mathsf{lin}}$ to accuracy $1/n$ in the ℓ_∞ norm) scale as $\mathcal{O}(d)$. This is inferior to confidence bounds for i.i.d. problems by a factor of $\sqrt{d}$ (Wainwright, 2019). Understanding the optimality of the SCBs implied by F-design is an interesting future direction.

5.2. Pure Exploration for Contextual Bandits

A contextual bandit (CB) problem is a class of reinforcement learning problems where the learning agent repeatedly interacts with an environment, but there are no long-term consequences of its actions. We focus on the stochastic version of the problem in the pure exploration setting (Audibert & Bubeck, 2010; Jamieson & Nowak, 2014), although adversarial settings and regret minimization are also considered in the literature (Bubeck et al., 2012b).

In pure exploration, the learner interacts with the environment for T rounds. At round t, the learner observes some context $z_t \sim D$,[2] chooses an action $a_t \in \mathcal{A}$ in response and receives a reward $r_t \in [0, 1]$, and we assume that $r_t \sim D(\cdot | z_t, a_t)$ for some unknown distribution D for all rounds t. With a slight abuse of notation, we also use D to refer to the joint distribution over $z, \boldsymbol{r}$, where $\boldsymbol{r} = (r(a))_{a \in \mathcal{A}}$ is a function specifying rewards for all actions. We make

[1] The parameter ϵ_0 serves as a small regularization. In the linear case, it can be chosen to recover the G-optimal design to be 0. For all the downstream tasks mentioned in Section 5, ϵ_0 is usually chosen to be less than $1/T$.

[2] We denote contexts by z rather than the more standard choice of x to avoid confusion with the elements of the design space.

the following assumption, often called realizability, in this section.

Assumption 5.2 (CB Realizability)**.** There exists a function $f^\star \in \mathcal{F}$ such that $\mathbb{E}[r|z,a] = f^\star(z,a)$ for all $z \in \mathcal{Z}, a \in \mathcal{A}$.

Since the rewards are in $[0,1]$, we also assume that $f(z,a) \in [0,1]$ for all $f \in \mathcal{F}$, $z \in \mathcal{Z}$ and $a \in \mathcal{A}$. The goal of the learner is to use the samples (z_t, a_t, r_t) collected over the T rounds and find a policy $\pi_T : \mathcal{Z} \to \mathcal{A}$ to minimize

$$\mathsf{Reg}(\pi_T) := \mathbb{E}_{z\sim D}\left[\max_a f^\star(z,a) - f^\star(z,\pi_T(z))\right]. \quad (3)$$

With the optimal F-design, we can achieve this goal through the following process. Upon observing z_t, we let $\rho_{\mathcal{V}}(\cdot|z_t)$ denote the optimal design as per our objective (2) for some base functional ν_0, with the design space $\mathcal{X} = \mathcal{A}$ and the function class $\mathcal{F}_{z_t} = \{f_{z_t}(a) = f(z_t,a) \ : \ f \in \mathcal{F}\}$ as the projection of $\mathcal{F}$ onto the observed context. Then for the dataset from T rounds, suppose the estimator $\widehat{f}_T$ satisfies, with probability at least $1-\delta$, an upper bound $\mathbf{Est}_{\mathsf{Off}}(T,\delta)$:

$$\mathbb{E}_{z\sim D, a\sim\rho_{\mathcal{V}}(\cdot|z)}[(\widehat{f}_T(z,a) - f^\star(z,a))^2] \le \mathbf{Est}_{\mathsf{Off}}(T,\delta). \quad (4)$$

The greedy policy $\pi_T(z) := \mathrm{argmax}_a \, \widehat{f}_T(z,a)$ satisfies:

Theorem 5.3. *Suppose ν_0 is such that $\sup_{h\in\mathcal{D}\mathcal{F}_z} \nu_0(h) \le \epsilon_0$ and $\mathcal{V}^*(\mathcal{F}_z, \mathcal{A}; \nu_0) \le d$ for all $z \in \mathcal{Z}$. Let π_T be the greedy policy of the function $\widehat{f}_T$ that satisfies* (4)*. Then for any $\delta \in (0,1)$, we have with probability at least $1-\delta$:*

$$\mathsf{Reg}(\pi_T) = \mathcal{O}\left(\sqrt{d \cdot (\mathbf{Est}_{\mathsf{Off}}(T,\delta) + \epsilon_0)}\right).$$

For example, using $\widehat{f}_T = \mathrm{argmin}_{f\in\mathcal{F}} \sum_{t=1}^T (f(z_t,a_t) - r_t)^2$, we have with probability at least $1-\delta$, $\mathbf{Est}_{\mathsf{Off}}(T,\delta) = \mathcal{O}(\log(|\mathcal{F}|/\delta)/T)$, which is the bound we use for Table 1. For details, please see Lemma F.2 in the appendix.

The result reduces to known optimal results (Agarwal et al., 2012; Dani et al., 2008) in the finite-action and (generalized) linear cases. For a detailed discussion of related approaches using the PAC DEC criterion in the general case, we refer the reader to Appendix F.2.1. Our results can be applied to Lipschitz or smooth function classes as shown in Section 3.2, and there is extensive literature on regret minimization in such non-parametric bandit settings (Kleinberg et al., 2019; Bubeck et al., 2011).

Lipschitz bandit If we directly apply Theorem 5.3 with covering numbers, the PAC regret guarantee scales with the rate $T^{(2d+1)/(2d+2)}$ where the optimal is $T^{(d+1)/(d+2)}$ for non-contextual Lipschitz bandits with dimension d. The gap is due to the fact that in the non-contextual case, we can get direct guarantees on the L^∞ norm of $\hat{f} - f^\star$. Because on each of the discretized points of a Lipschitz bandit as an action, we can count the number of pulls and obtain an L^∞ norm bound between $\hat{f}$ and $f^\star$ without going through the L^2 norm. The detour from the L^2 norm to the L^∞ norm causes the degradation in the rate. Meanwhile, we mention in passing that a slight improvement can be done by replacing the covering number bound by the chaining bound for bounding the estimation error $\mathbf{Est}_{\mathsf{Off}}$. This obtains a matching regret bound when $d <= 4$ and a suboptimal bound of $T^{1-2/(3d)}$ when $d > 4$. In a nutshell, our algorithm is designed for contextual cases where the L^∞ norm bound is hard to obtain in general, which leads to our approach of going through the L^2 norm.

5.3. Model-free Reinforcement Learning

In this section, we study the more general setting of long-horizon reinforcement learning (RL), with the goal of finding a near optimal policy in a Markov Decision Process (MDP) (Puterman, 2014). Concretely, we consider an episodic, finite horizon MDP with a horizon H, which is a stochastic process parameterized by the tuple $(D, \{\mathcal{Z}^h\}_{h\in[H]}, \{\mathcal{A}^h\}_{h\in[H]}, \{P^h\}_{h\in[H]}\{R^h\}_{h\in[H]})$. In this setting, the agent interacts with its environment over episodes $t = 1, 2, \ldots$. The episode t begins with an initial state $z_t^1 \sim D$. At each step $h \in [H]$, the agent chooses an action a_t^h, observes a reward r_t^h and the next state $z_t^{h+1} \sim P^h(\cdot|z_t^h, a_t^h)$ according the unknown transition dynamics P of the MDP. Here $z^h \in \mathcal{Z}^h$, $a^h \in \mathcal{A}^h$ and $\mathbb{E}[r^h|z^h,a^h] = R^h(z^h,a^h)$. The goal is to find a policy which maximizes the H-step return, that is:

$$\pi_\star = \underset{(\pi^1,\ldots,\pi^h)}{\mathrm{argmax}} \, \mathbb{E}\big[\sum_{h=1}^H r^h | a^h \sim \pi^h(\cdot|z^h)\big],$$

with $z^1 \sim D$ and $z^{h+1} \sim P^h(\cdot|z^h,a^h)$. Given a policy π, which refers to an H tuple $(\pi^1,\ldots,\pi^H)$, we define the Q-value function:

$$Q_\pi^h(z,a) = \mathbb{E}\big[\sum_{h'=h}^H r^{h'} | a^{h'} \sim \pi^{h'}, z^h = z, a^h = a\big].$$

Define $V_\star^h(z^h) = Q_\star^h(z^h, \pi_\star^h(z^h))$, where $Q_\star^h := Q_{\pi_\star}^h$. We try to approximate $Q_\star$ using functions from $\mathcal{F} = (\mathcal{F}^1,\ldots,\mathcal{F}^H)$, and make the following assumption.

Assumption 5.4 ($Q_\star$ Realizability)**.** $\forall\, h \in [H]$, $Q_\star^h \in \mathcal{F}^h$.

Given a function $f = (f^1,\ldots,f^H) \in \mathcal{F}$, a common approach in model-free RL is to evaluate the Bellman consistency of f as a surrogate for how well it approximates $Q_\star$. That is, we define the *Bellman error* for any h, z, a as:

$$\mathcal{E}^h(f,z,a) = f^h(z,a) - \mathbb{E}[r^h + \max_{a'\in\mathcal{A}^{h+1}} f^{h+1}(z',a')|z,a]. \quad (5)$$

In particular, it is well know that $\mathcal{E}^h(Q_\star^h, z, a) = 0$ for all $z \in \mathcal{Z}^h$, $a \in \mathcal{A}^h$ and $h \in [H]$ (Jiang et al., 2017).

For the model-free RL setting studied here, a large number of algorithms search for functions f with a small Bellman error under samples z^h, a^h collected by some exploratory policies at each $h \in [H]$. Examples include Q-learning (Dayan & Watkins, 1992), where the policy is simply ϵ-greedy with respect to the current f, as well as a large number of algorithms with bounded sample complexity under various structural conditions (Jiang et al., 2017; Du et al., 2021; Jin et al., 2021; Agarwal & Zhang, 2022; Foster et al., 2023b), all of which use carefully construct their exploration policies. These algorithms have sample complexity guarantees under some structural conditions over the MDP, and we build upon the *Bellman rank* introduced in Jiang et al. (2017).

Assumption 5.5 (Bellman factorization and Bellman rank)**.** We assume that for all $f, f' \in \mathcal{F}$, and $h \in [H]$, there exist (unknown) functions $u^h, \psi^{h-1} : \mathcal{F} \to \mathbb{R}^{\mathsf{br}^h}$ and an inner product $\langle \cdot, \cdot \rangle$ such that for any starting state $z^1 \in \mathcal{Z}$

$$\mathbb{E}_{z^h \sim \pi_{f'}|z^1} \mathcal{E}(f, z^h, \pi_f(z^h)) = \left\langle \psi^{h-1}(f', z^1), u^h(f, z^1) \right\rangle .$$

We assume that $\sup_{f \in \mathcal{F}, z^1 \in \mathcal{Z}} \|u^h(f, z^1)\|_2 \leq B_1$. The Bellman rank is defined as $\mathsf{br} := \sum_{h=1}^H \mathsf{br}^{h-1}$.

Prior work gives sample-efficient RL algorithms under Assumptions 5.4 and 5.5, for finite action spaces (Jiang et al., 2017), as well as linearly embedded Bellman errors in infinite action spaces (Agarwal & Zhang, 2022).

The last result is particularly relevant to our development, as it essentially combines the Bellman rank machinery with G-optimal design to handle infinite action spaces in certain linear settings. Here, we extend this to general non-linear scenarios. To proceed further, note that both the OLIVE algorithm of Jiang et al. (2017) and the TS^2-D algorithm of Agarwal & Zhang (2022) use an exploration policy ρ^h for the first $h-1$ steps, to effectively explore over the states at step h. They subsequently invoke a policy π^h_{exp} at step h for one-step exploration in the action space, which is uniform over actions in OLIVE and a G-optimal design in TS^2-D.

To apply F-design to this problem, we first choose an appropriate function class. Like the contextual bandit example, given some state z^h at step h, we define $\pi^h_{\text{exp}} = \rho_{\mathcal{V}}(\mathcal{F}_{z^h}; \nu_0)$, where $\mathcal{F}_{z^h} = \{f(z^h, \cdot) \ : \ f \in \mathcal{F}\}$ and ν_0a base functional. To analyze this approach, we need the following additional assumption, which is quite common in the literature.

Assumption 5.6 (Bellman completeness)**.** $\forall h \in [H]$ and $f^{h+1} \in \mathcal{F}^{h+1}$, $\exists$ a function $g^h_f \in \mathcal{F}^h$ such that $\forall\ z \in \mathcal{Z}^h$ and $a \in \mathcal{A}^h$, we have $g^h(z, a) = \mathbb{E}[r^h + f^{h+1}(z')|z, a]$.

Informally, the assumption says that the class $\mathcal{F}$ can express one step Bellman backups of all functions $f \in \mathcal{F}$. Importantly, under this assumption, the class $\mathcal{DF}$ used in our design objective is a superset of all the Bellman errors, which is critical for our analysis.

With these concepts, we give Algorithm 6 in Appendix F.3, which is a modification of Algorithm 2 in Agarwal & Zhang (2022) with F-design instead of G-design. Replacing Lemma 32 of their paper with the more general analog, Lemma F.7, in their proof immediately gives the result.

Theorem 5.7. *Under Assumptions 5.4, 5.5 and 5.6, suppose further that for any state $z \in \mathcal{Z}^h$ and all function class $\mathcal{F}^h_z$, we have $\mathcal{V}^*(\mathcal{F}^h_z; 0) \leq d$. Then there exists an algorithm that interacts with the environment for T rounds starting at state z^1_t and outputs one function f_t (and greedy policy $\pi_t = \pi_{f_t}$) at each time step $t \in [T]$ that achieves*

$$\mathbb{E}\left[\sum_{t=1}^T V^1_\star(z^1_t) - V^1_{\pi_t}(z^1_t)\right] = \mathcal{O}\left(\log(|\mathcal{F}|T)\sqrt{\mathsf{br} \cdot dHT}\right),$$

The dependence on br, d, and H matches the best previously known results in the special cases studied in prior works.

On Bellman completeness. Note that while the Bellman completeness assumption is not needed in many prior works, this is because both finite action and linearly embedded Bellman error settings allow a good design for π^h_{exp} without this assumption. Eliminating Assumption 5.6 in the general non-linear setting is an interesting question for future study.

For concrete examples of interest, we consider MDPs with the following low-rank transition structure.

Definition 5.8 (Low-rank MDP ((Jiang et al., 2017; Jin et al., 2020)))**.** An MDP has a low-rank transition structure if for any $h \in [H]$, there exist d and an unknown feature map $\phi^h : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}^{\mathsf{br}^h}$ and (signed) measures $\mu^{h+1} = (\mu^{h+1,1}, ..., \mu^{h+1,\mathsf{br}^h})$ over $\mathcal{Z}^{h+1}$, such that for any $(z^h, a^h) \in \mathcal{Z}^h \times \mathcal{A}^h$, we have

$$P^h(z^{h+1}|z^h, a^h) = \left\langle \phi^h(z^h, a^h), \mu^{h+1}(z^{h+1}) \right\rangle ,$$

we assume $\max_{h, z^{h+1}} \|\mu^{h+1}(z^{h+1})\| \leq B_1$.

The MDPs with the low-rank transitions have a Bellman rank at most $\mathsf{br} = \sum_h \dim(\phi^h)$ (Jiang et al., 2017). We consider $\phi^h \equiv \phi$ for all $h \in [H]$ for simplicity here.

Example 5 (Low-rank MDP + realizable + Bellman complete with Hölder function class)**.** Consider any low-rank MDP class with $\mathcal{A} \subset \mathbb{R}^d$ a bounded set. Furthermore, suppose $\mathcal{F}$ is realizable and Bellman complete with $\mathcal{F}^h_z \subset \mathcal{F}^{\mathsf{H}}_{\beta,d}$ for all $z \in \mathcal{Z}$.

Corollary 5.9. *For the example above, there exists an algorithm that achieves a PAC regret upper bounded by $\mathcal{O}\left((H\mathsf{br})^{1/2} T^{(\beta^2+4\beta d+d^2)/((\beta+d)(2\beta+d))}\right)$ which is sublinear whenever $\beta > d$.*

Example 6 (LQR)**.** Our results give a value-based method to learning in Linear Quadratic Regulators, since the optimal value functions are quadratic and admit efficient design, and LQRs have a bounded Bellman rank (Jiang et al., 2017).

In this section, we mainly consider the application of the F-design in model-free reinforcement learning. For model-based reinforcement learning where the F-design is applied to obtain online regret minimization guarantees, we refer to Appendix B.1.

5.4. Pool-based Active Learning

In this section, we consider active learning in the pool-based model, where an unlabeled pool of data is made available to the algorithm, and the goal is to query labels on a subset of the data to achieve the same statistical performance as if training were carried out on the entire pool.

We consider the binary realizable classification problem. We denote by $\mathcal{X}$ the instance space, by $\mathcal{Y} = \{\pm 1\}$ the output space, and by D an unknown distribution over $\mathcal{X} \times \mathcal{Y}$. The corresponding random variables are denoted by x and y. We also denote by $D_{\mathcal{X}}$ the marginal distribution of D over $\mathcal{X}$. Given a hypothesis f mapping $\mathcal{X}$ to $\mathcal{Y}$, the population loss (often referred to as risk) of f is denoted by $L(f)$, and defined as $L(f) = \mathbb{E}_{(x,y)\sim D}[\ell(f(x), y)]$, where $\ell(y, y') = \mathbb{1}(y = y')$ is the binary loss function.

Assumption 5.10 (Realizability for binary classification)**.** There exists a hypothesis $f^* \in \mathcal{F}$ such that almost surely for any $(x, y) \sim D$, we have $y = f^*(x)$.

In this setting, we apply F-design to the function class $\mathcal{F}$, with the design space $\mathcal{U}$ as the pool of unlabeled instances $x_1, \ldots, x_T$ that are given to us. We subsequently query labels y_i at n points sampled according to the optimal design $\rho_{\mathcal{V}}$ for this class[3], and define $\widehat{f}$ to be any function in $\mathcal{F}$ that achieves a misclassification error of 0 on these samples. Combining the definition of F-condition number with standard arguments, we get the following result.

Theorem 5.11. *To achieve with probability at least* $1 - \delta$ *that:* $L(\widehat{f}) \leq \mathcal{O}\left((\mathrm{VCdim}(\mathcal{F}) \log T + \log(1/\delta))/T\right)$ *under Assumption 5.10, we need a sample size at most*

$$n = \mathcal{O}(\mathcal{V}^*(\mathcal{F}, \mathcal{U}; 0) \log(|\mathcal{F}|/\delta)).$$

Here $\mathrm{VCdim}(\mathcal{F})$ *is the VC dimension of* $\mathcal{F}$*.*

Application of G-optimal design in active learning has been considered by (Hazan & Karnin, 2014) for linear cases with a hard margin. Our result extends to the non-linear cases. Similar settings of active learning have been considered in the literature (Balcan et al., 2007; Balcan & Long, 2013; Zhang & Li, 2021; Gentile et al., 2022). For the most recent result of Gentile et al. (2022), the number of queries they require for the non-linear case scales with the eluder dimension. Thus they may require an exponentially larger number of samples than the optimal F-condition number.

6. Discussion

Our work introduces the non-linear F-design and applies it to several learning tasks. We close with two general directions for future research.

Computing the optimal design efficiently. In this paper, we can compute a F-design with the F-condition number scaling with the eluder dimension in an oracle-efficient way. However, it is not known how to compute/approximate the optimal F-design while also scaling sublinearly (ideally as log) in $|\mathcal{X}|$. In Appendix E.3, we show that a broader class of algorithms depending on the argmax oracle cannot achieve this, suggesting we need a different approach.

Other optimal design criteria In this paper, we mainly study non-linear G-optimal design and its application to various learning tasks. It would be interesting to study potential extensions of other classical design criteria in future research, as they might have complementary benefits.

Online regret minimization and relation to the DEC For online regret minimization, there are naive cases where the F-design is suboptimal. For instance, in the case where every function agrees with the optimal action, the regret minimization problem is trivial, but the F-design aims to control the L^∞ divergence, which is non-trivial. Nevertheless, the F-design can be used for online regret minimization by bounding the DEC as shown by Theorem B.4. At the current stage, applying the F-design for regret minimization remains largely unexplored, and we hope to reveal deeper links in our future work.

Acknowledgements

Jian Qian acknowledges support from ARO through award W911NF-21-1-0328 and from the Simons Foundation and NSF through award DMS-2031883.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Agarwal, A. and Zhang, T. Non-linear reinforcement learning in large action spaces: Structural conditions and sample-efficiency of posterior sampling. *arXiv preprint arXiv:2203.08248*, 2022.

Agarwal, A., Dudík, M., Kale, S., Langford, J., and Schapire, R. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pp. 19–

[3] We sample with replacement. If an instance x_i is sampled more than once, the label y_i would remain the same each time x_i is sampled. This does not hurt the guarantee we obtain.

26. PMLR, 2012.

Audibert, J.-Y. and Bubeck, S. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pp. 13–p, 2010.

Awerbuch, B. and Kleinberg, R. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

Balcan, M.-F. and Long, P. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pp. 288–316, 2013.

Balcan, M.-F., Broder, A., and Zhang, T. Margin based active learning. In *International Conference on Computational Learning Theory*, pp. 35–50. Springer, 2007.

Ball, K. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31:1–58, 1997.

Bickel, P. J. and Rosenblatt, M. On some global measures of the deviations of density function estimates. *The Annals of Statistics*, pp. 1071–1095, 1973.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011.

Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, pp. 41–1. JMLR Workshop and Conference Proceedings, 2012a.

Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012b.

Bubeck, S., Dekel, O., Koren, T., and Peres, Y. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In *Conference on Learning Theory*, pp. 266–278, 2015.

Cai, L., Liu, R., Wang, S., and Yang, L. Simultaneous confidence bands for mean and variance functions based on deterministic design. *Statistica Sinica*, 29(1):505–525, 2019.

Cai, T. T., Low, M., and Ma, Z. Adaptive confidence bands for nonparametric regression functions. *Journal of the American Statistical Association*, 109(507):1054–1070, 2014.

Casella, G., Fienberg, S., and Olkin, I. *Statistical design*. Springer, 2008.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Chen, F., Mei, S., and Bai, Y. Unified algorithms for rl with decision-estimation coefficients: No-regret, pac, and reward-free learning. *arXiv preprint arXiv:2209.11745*, 2022.

Cochran, W. G. and Cox, G. M. Experimental designs. Technical report, North Carolina State University. Dept. of Statistics, 1948.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory (COLT)*, 2008.

Dayan, P. and Watkins, C. Q-learning. *Machine learning*, 8 (3):279–292, 1992.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pp. 2826–2836. PMLR, 2021.

Foster, D. J. and Rakhlin, A. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *International Conference on Machine Learning (ICML)*, 2020.

Foster, D. J., Rakhlin, A., Simchi-Levi, D., and Xu, Y. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Foster, D. J., Golowich, N., and Han, Y. Tight guarantees for interactive decision making with the decision-estimation coefficient. *arXiv preprint arXiv:2301.08215*, 2023a.

Foster, D. J., Golowich, N., Qian, J., Rakhlin, A., and Sekhari, A. Model-free reinforcement learning with the decision-estimation coefficient. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Gentile, C., Wang, Z., and Zhang, T. Fast rates in pool-based batch active learning. *arXiv preprint arXiv:2202.05448*, 2022.

Härdle, W. Asymptotic maximal deviation of m-smoothers. *Journal of Multivariate Analysis*, 29(2):163–179, 1989.

Hazan, E. and Karnin, Z. Hard-margin active linear regression. In *International Conference on Machine Learning*, pp. 883–891. PMLR, 2014.

Hazan, E. and Karnin, Z. Volumetric spanners: an efficient exploration basis for learning. *The Journal of Machine Learning Research*, 17(1):4062–4095, 2016.

Jackson, M. and Cox, D. R. The principles of experimental design and their application in sociology. *Annual Review of Sociology*, 39:27–49, 2013.

Jamieson, K. and Nowak, R. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6. IEEE, 2014.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143, 2020.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

John, F. Extremum problems with inequalities as subsidiary conditions. *Traces and emergence of nonlinear programming*, pp. 197–215, 2014.

Johnston, G. J. Probabilities of maximal deviations for nonparametric regression function estimates. *Journal of Multivariate Analysis*, 12(3):402–414, 1982.

Jun, K.-S. and Zhang, C. Crush optimism with pessimism: Structured bandits beyond asymptotic optimality. *Advances in Neural Information Processing Systems*, 33: 6366–6376, 2020.

Katz-Samuels, J., Zhang, J., Jain, L., and Jamieson, K. Improved algorithms for agnostic pool-based active classification. In *International Conference on Machine Learning*, pp. 5334–5344. PMLR, 2021.

Keskin Gündoğdu, T., Deniz, I., Çalışkan, G., Şahin, E. S., and Azbar, N. Experimental design methods for bioengineering applications. *Critical reviews in biotechnology*, 36(2):368–388, 2016.

Kiefer, J. and Wolfowitz, J. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12: 363–366, 1960.

Kleinberg, R., Slivkins, A., and Upfal, E. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66 (4):1–77, 2019.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Mhammedi, Z., Block, A., Foster, D. J., and Rakhlin, A. Efficient model-free exploration in low-rank mdps. *arXiv preprint arXiv:2307.03997*, 2023.

Modi, A., Jiang, N., Tewari, A., and Singh, S. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 2010–2020. PMLR, 2020.

Montgomery, D. C. *Design and analysis of experiments*. John wiley & sons, 2017.

Pronzato, L. and Pázman, A. Design of experiments in nonlinear models. *Lecture notes in statistics*, 212:1, 2013.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Shor, N. Z. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 2012.

Smith, K. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12 (1/2):1–85, 1918.

Tikhomirov, V. M. ε-entropy and ε-capacity of sets in functional spaces. *Selected Works of AN Kolmogorov: Volume III: Information Theory and the Theory of Algorithms*, pp. 86–170, 1993.

Todd, M. J. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

Vaart, A. v. d. and Wellner, J. A. Empirical processes. In *Weak Convergence and Empirical Processes: With Applications to Statistics*, pp. 127–384. Springer, 2023.

Vapnik, V. N. *Estimation of dependences based on empirical data*, volume 40. Springer-Verlag New York, 1982.

Wagenmaker, A., Katz-Samuels, J., and Jamieson, K. Experimental design for regret minimization in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3088–3096. PMLR, 2021.

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

Walsh, S. J. Overview of optimal experimental design and a survey of its expanse in application to agricultural studies. 2022.

Wang, J. and Yang, L. Polynomial spline confidence bands for regression curves. *Statistica Sinica*, pp. 325–342, 2009.

Xia, Y. Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):797–811, 1998.

Zhang, C. and Li, Y. Improved algorithms for efficient active learning halfspaces with massart and tsybakov noise. In *Conference on Learning Theory*, pp. 4526–4527. PMLR, 2021.

Zhang, T. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

Zhong, H., Xiong, W., Zheng, S., Wang, L., Wang, Z., Yang, Z., and Zhang, T. Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond. *arXiv preprint arXiv:2211.01962*, 2022.

Zhu, Y., Foster, D. J., Langford, J., and Mineiro, P. Contextual bandits with large action spaces: Made practical. In *International Conference on Machine Learning*, pp. 27428–27453. PMLR, 2022.

A. Related Works

Design-based methodology and applications are extensively considered in the literature mostly with applications in linear cases. In the following, we describe several related works in this regard.

Decomposition and Spanners John's decomposition (John, 2014; Ball, 1997) shows that for any bounded convex set in the d-dimensional space, there exists a set of bases with size at most $\mathcal{O}(d^2)$ that can span the convex set with coefficients in certain forms. Bubeck et al. (2012a) leverage this decomposition for exploration in linear bandits and obtains the minimax optimal rate. The idea of John's decomposition is extended by the barycentric spanners and the volumetric spanners which are proposed with similar motivations for finding a set of bases in the linear class for any bounded convex set (Awerbuch & Kleinberg, 2008; Hazan & Karnin, 2016). The barycentric spanner tries to find a set of bases where all vectors in the set can be written as linear combinations of the spanner with each coefficient bounded by $[-1, 1]$. Despite the original motivation of online linear optimization, it is applied to problems that include contextual bandits and RL (Foster et al., 2021; Zhu et al., 2022). The volumetric spanner tries to find a set of bases where all vectors in the set can be written as linear combinations of the spanner with coefficients that lie in a unit ball. The volumetric spanner has applications in experimental design and linear bandits (Hazan & Karnin, 2016). Both of these spanners are developed with a focus on linear applications. We refer the reader to the texts (Bubeck et al., 2015; Lattimore & Szepesvári, 2020) for a more in-depth discussion on the uses of design in the bandit literature.

Applications of Linear G-Optimal Design Applications of linear G-optimal design to linear bandits and low-rank MDPs are also considered (Foster et al., 2021; Wagenmaker et al., 2021). In such cases, the G-optimal design offers a strategy for exploration that is sufficiently balanced to cover all the actions. More related to our RL results, Agarwal & Zhang (2022) apply G-optimal design within an optimistic posterior sampling algorithm for a subclass of the problems that we address here. These applications are again restricted to the linear case.

Design-Based Approach for Active Learning Katz-Samuels et al. (2021) consider the agnostic pool-based PAC active classification task where the learner iteratively queries samples from the environment for obtaining a classifier with a given performance guarantee. For this task, they design querying strategies based on a design value which is the ratio between the number of disagreements expected to be observed and the difference in error. Gentile et al. (2022) consider the pool-based batch active-learning as we describe in Section 5.4. They propose a design principle in the spirit of the eluder dimension, i.e., iteratively finding the worst sample in terms of the ratio between square loss and the sum of all the square losses with respect to all the samples chosen in history. As we show in our Section 5.4 such a method can be improved in a pool-dependent fashion and possibly exponentially due to the exponential gap between the F-condition number and the eluder dimension.

Non-linear Experimental Design Many classical design criteria, such as A,c,D,G,L-design , are studied in the statistics literature (Pronzato & Pázman, 2013). These design criteria originate from controlling different aspects of the (asymptotic) variance of the linear least squares predictor over the design space $\mathcal{X}$. The natural extensions of these criteria to the non-linear cases are thus local and asymptotic (Pronzato & Pázman, 2013). Global and non-asymptotic minimax criteria are also considered in the most general form (Pronzato & Pázman, 2013, Section 8.4) with difficulties in terms of existence.

B. Additional Results

B.1. Model-based Reinforcement Learning

In this section, we continue the study of reinforcement learning following the model-based approach where a class $\mathcal{M}$ of MDPs are given, where the MDP models in $\mathcal{M}$ share the same (known) $(D, \{\mathcal{Z}^h\}_{h\in[H]}, \{\mathcal{A}^h\}_{h\in[H]})$ while paramentrized by $M = (\{P^h\}_{h\in[H]}\{R^h\}_{h\in[H]})$. The optimal policy for each model $M \in \mathcal{M}$ is denoted by π_M. The Q-value function is denoted by $Q^h_{M,\pi}$. The optimal value function is defined as $V^h_M(z^h) = Q^h_M(z^h, \pi_M(z^h))$ where $Q^h_M(z^h, a^h) = Q^h_{M,\pi_M}(z^h, a^h)$. We also denote $J_M(\pi) = \mathbb{E}_{z^1\sim D}[Q^1_{M,\pi}(z^1, \pi(z^1))]$ and $\mathrm{Reg}_M(\pi) = J_M(\pi_M) - J_M(\pi)$ for all $M \in \mathcal{M}$ and $\pi \in \Pi$. In the model based approach, we try to estimate the true model $M_\star$ with the following realizability assumption.

Assumption B.1 ($M_\star$ Realizability)**.** $M_\star \in \mathcal{M}$.

For any two models $M, M' \in \mathcal{M}$, policy π and $h \in [H]$, we abuse the notation and define similarly the *model-based*

Bellman error as

$$\mathcal{E}^h(M, M', \pi) := \mathbb{E}^{M,\pi}\big[Q^h_{M'}(z^h, a^h) - r^h - V^h_{M'}(z^{h+1})\big],$$

where $\mathbb{E}^{M,\pi}$ denote the expectation running policy π in model M.

For the model-based setting, algorithms have been developed to search for M' such that $\mathcal{E}^h(M_\star, M', \pi)$ is small for all π (Du et al., 2021; Foster et al., 2021). For such a setup, we define the following adaptation of F-condition number.

Definition B.2 (Regret F-condition number)**.** For any RL model class $\mathcal{M}$, $\overline{M} \in \mathcal{M}$, and any parameter $\eta \geq 0$, consider the function class $\mathcal{F}^h_\eta(\overline{M}) = \left\{ \frac{\mathcal{E}^h(\overline{M},M,\cdot)}{\sqrt{1+\eta\cdot\mathrm{Reg}_{\overline{M}}(\cdot)}} \mid M \in \mathcal{M} \right\} \subset (\Pi \to \mathbb{R})$. For any base measure $\nu_0 : \cup_{h\in[H]}\mathcal{F}^h_\eta \to \mathbb{R}_{\geq 0}$, the regret F-condition number is defined as

$$\mathcal{V}^\star_\eta(\mathcal{M}; \overline{M}, \nu_0) := \sup_{h\in[H]} \inf_{\rho\in\Delta(\Pi)} \sup_{f\in\mathcal{F}^h_\eta(\overline{M}), M\in\mathcal{M}} \frac{f^2(\pi_M)}{\nu_0(f) + \mathbb{E}_{\pi\sim\rho}[f^2(\pi)]}.$$

The regret F-condition number differs from the F-condition number in that the supreme on nominator is only taken over the possible optimal policies $\Pi_{\mathcal{M}} = \{\pi_M | M \in \mathcal{M}\}$, whereas the design can be chosen on the whole policy class. The regret F-condition number is bounded by the dimension d for the cases where

$$\mathcal{E}^h(\overline{M}, M, \pi_{M'}) = \big\langle X_h(M'; \overline{M}), W_h(M; \overline{M}) \big\rangle\,,$$

for some functions $X_h(\cdot; \overline{M}), W_h(\cdot; \overline{M}) : \mathcal{M} \to \mathbb{R}^d$ following Example 1. These classes include MDPs with linear $Q_\star/V_\star$ (Du et al., 2021), MDPs with low occupancy measure (Du et al., 2021), and linear mixture MDPs (Modi et al., 2020). For all such models, we have the following theorem.

Theorem B.3. *Suppose* $\sup_f \nu_0(f) \leq 1/T$ *and* $\mathcal{V}^\star_0(\mathcal{M}; \overline{M}, \nu_0) \leq d$*. There exists an algorithm that interacts with the true model for* T *rounds and returns a policy* π_T *that achieves*

$$\mathbb{E}\big[V^1_\star(z^1) - V^1_{\pi_T}(z^1)\big] = \mathcal{O}\left(\sqrt{\frac{H^2 d \log|\mathcal{M}|}{T}}\right).$$

This shows that the F-condition number captures a wide range of structural properties. Varying the parameter η in the defintion of regret F-condition number, we can also obtain guarantees for online regret minimization. For this we have the following theorem.

Theorem B.4. *Suppose* $\sup_f \nu_0(f) \leq 1/T$*,* $\mathcal{V}^\star_\eta(\mathcal{M}; \overline{M}, \nu_0) \leq d$ *for some* $\eta = \Theta(H\sqrt{dT}/\log|\mathcal{M}|)$*. There exists an algorithm that interacts with the true model for* T *rounds with policy* π_t *at round* t *for* $t \in [T]$ *that achieves*

$$\sum_{t=1}^{T} \mathbb{E}\big[V^1_\star(z^1) - V^1_{\pi_t}(z^1)\big] = \mathcal{O}\left(\sqrt{H^2 d \log|\mathcal{M}| \cdot T}\right).$$

C. Technical Tools

C.1. Subgradient Descent

Proposition C.1 ((Shor, 2012))**.** *Let* $d > 0$ *be any interger. For any* L*-Lipschitz convex function* l *defined on a closed bounded convex set* $\mathcal{C} \subset \mathbb{R}^d$ *and any initial point* x_1*, the iteration*

$$x_{t+1} = P_{\mathcal{C}}(x_t - g_t/\sqrt{t}),$$

where $P_{\mathcal{C}}(x) = \operatorname{argmin}_{y\in\mathcal{C}} \|x - y\|$ *is the projection map and* $g_t \in \partial l(x_t)$ *satisfies*

$$\min_{s\leq t} l(x_s) - \min_{x\in\mathcal{C}} l(x) \leq \frac{\|x_1 - x^*\|^2 + L^2 \log t}{\sqrt{t}},$$

where $x^* = \operatorname{argmin}_{x\in\mathcal{C}} l(x)$*.*

C.2. ERM Guarantee

Lemma C.2. *For any distribution $\rho \in \Delta(\mathcal{X})$, if we sample $x_1, ..., x_n$ i.i.d. from ρ, then we have that with probability at least $1-\delta$,*

$$\mathbb{E}_{x\sim\rho}(\widehat{f}_n(x) - f^\star(x))^2 = \mathcal{O}\left(\frac{B\log(|\mathcal{F}|/\delta)}{n}\right),$$

for any $\delta \in (0, 1/2)$, where $\widehat{f}_n$ is the ERM estimator.

Proof of Lemma C.2. The proof is standard. We include this for completeness.

$$\begin{aligned}\sum_{i=1}^n (f^*(x_i) - \widehat{f}_n(x_i))^2 &= \sum_{i=1}^n \Big((\widehat{f}_n(x_i) - y_i)^2 - (f^*(x_i) - y_i)^2 + 2(\widehat{f}_n(x_i) - f^*(x_i))(y_i - f^*(x_i))\Big)\\ &\leq 2\sum_{i=1}^n (\widehat{f}_n(x_i) - f^*(x_i))\epsilon_i,\end{aligned}$$

where the last inequality is by the definition of $\widehat{f}_n$ and $\epsilon_i := y_i - f^*(x_i)$ is mean-zero and bounded in $[-B, B]$. Then by Azuma's inequality, we have with probability at least $1 - \delta/2$,

$$\sum_{i=1}^n (\widehat{f}_n(x_i) - f^*(x_i))\epsilon_i \leq 4B\sqrt{\sum_{i=1}^n (\widehat{f}_n(x_i) - f^*(x_i))^2 \log(2|\mathcal{F}|/\delta)}.$$

Thus,

$$\sum_{i=1}^n (\widehat{f}_n(x_i) - f^*(x_i))^2 \leq 16B^2\log(2|\mathcal{F}|/\delta).$$

Then by Bernstein inequality, we have with probability at least $1 - \delta/2$ for any $f, f' \in \mathcal{F}$,

$$\begin{aligned}\mathbb{E}_{x\sim\rho}(f(x) - f'(x))^2 &\leq \frac{1}{n}\sum_{i=1}^n (f(x_i) - f'(x_i))^2 + \sqrt{\frac{2\mathrm{Var}\left((f(x_1) - f'(x_1))^2\right)\log(2|\mathcal{F}|/\delta)}{n}} + \frac{B\log(2|\mathcal{F}|/\delta)}{3n}\\ &\leq \frac{1}{n}\sum_{i=1}^n (f(x_i) - f'(x_i))^2 + \sqrt{\frac{2B^2\mathbb{E}_{x\sim\rho}(f(x) - f'(x))^2\log(2|\mathcal{F}|/\delta)}{n}} + \frac{B\log(2|\mathcal{F}|/\delta)}{3n}\\ &\leq \frac{1}{n}\sum_{i=1}^n (f(x_i) - f'(x_i))^2 + \frac{1}{2}\mathbb{E}_{x\sim\rho}(f(x) - f'(x))^2 + \frac{2B\log(2|\mathcal{F}|/\delta)}{n},\end{aligned}$$

where the second inequality is by $\mathrm{Var}\left((f(x_1) - f'(x_1))^2\right) \leq \mathbb{E}_{x\sim\rho}(f(x) - f'(x))^4 \leq B^2\mathbb{E}_{x\sim\rho}(f(x) - f'(x))^2$ and the third inequality is by Cauchy-Schwarz inequality. In all, we have with probability at least $1 - \delta/2$ that

$$\mathbb{E}_{x\sim\rho}(\widehat{f}_n(x) - f^*(x))^2 \leq \frac{2}{n}\sum_{i=1}^n (\widehat{f}_n(x_i) - f^*(x_i))^2 + \frac{4B\log(2|\mathcal{F}|/\delta)}{n}.$$

Finally, by the union bound, we have with probability at least δ,

$$\mathbb{E}_{x\sim\rho}(\widehat{f}_n(x) - f^*(x))^2 \leq \frac{36B\log(2|\mathcal{F}|/\delta)}{n}.$$

□

D. Proofs from Section 3

D.1. Equivalence with G-optimal Design in the Linear Case

Proof of Lemma 3.1. Let $\mathbb{B}_d$ refer to the Euclidian unit ball in $\mathbb{R}^d$. By the definition of $\mathcal{F}^{\mathsf{lin}}$, we have that for any fixed ρ,

$$\mathcal{V}(\mathcal{F}^{\mathsf{lin}}, \rho; \nu_0) = \sup_{x\in\mathcal{X}} \sup_{w,w'\in\mathbb{B}_d} \frac{\left((w-w')^\top x\right)^2}{(w-w')^\top \Sigma_\rho(\lambda/d)(w-w')}.$$

Now we note that the objective is invariant to the norm of $w - w'$, so the objective is equivalent to

$$\begin{aligned}\mathcal{V}(\mathcal{F}^{\mathsf{lin}}, \rho; \nu_0) &= \sup_{x\in\mathcal{X}} \sup_{u\neq 0} \frac{\left(u^\top x\right)^2}{u^\top \Sigma_\rho(\lambda/d) u} \\ &= \sup_{x\in\mathcal{X}} \sup_{v\neq 0} \frac{(v^\top \Sigma_\rho(\lambda/d)^{-1/2} x)^2}{\|v\|_2^2} \\ &= \sup_{x\in\mathcal{X}} \|x\|^2_{\Sigma_\rho(\lambda/d)^{-1}},\end{aligned}$$

where the final step uses the closed form expression for the maximizing $v = \Sigma_\rho(\lambda/d)^{-1/2}x$. This last term is exactly the residual variance on the point x, given data from ρ, and forms the objective for classical G-design. This completes the proof. □

D.2. Optimal F-condition Number for Example Function Classes

Proof of Example 2. Without loss of generality, assume $U \subset [0,1]^d$. Let $h \in (0,1)$ be some width to be determined and let $1/h$ be an integer. Divide U into $N = (1/h)^d$ cubes each with width h. Consider the design space of

$$\mathcal{X}_h = \left\{ \left(x_1^{k_1} \cdots x_d^{k_d}\right)_{\substack{(k_1,\dots,k_d)\in\mathbb{Z}^d_{\geq 0} \\ k_1+\cdots+k_d\leq\lfloor\beta\rfloor}} \mid (x_1, ..., x_d) \in [0,h]^d \right\} \subset \mathbb{R}^{(\lfloor\beta\rfloor+1)^d}.$$

For this design space in $\mathbb{R}^{(\lfloor\beta\rfloor+1)^d}$, we can consider the corresponding G-optimal design π_h which corresponds naturally to a distribution on $[0,h]$. Let $T^j(x) = (x_1 + j_1 h, x_2 + j_2 h, ..., x_d + j_d h)$ for $x = (x_1, ..., x_d) \in [0,1]^d$ be the translation map for any $j = (j_1, ..., j_d) \in \{0, 1, ..., 1/h - 1\}^d$. Then we consider the distribution

$$\rho = \frac{1}{N} \sum_{j\in\{0,1,\dots,1/h-1\}^d} T^j_{\#}\pi_h,$$

where $T^j_{\#}$ denote the push-forward map induced by T^j. For this ρ, we have

$$\mathcal{V}(\mathcal{F}^{\mathsf{H}}_{\beta,d}, \rho; \epsilon_0) = \sup_{g\in\mathcal{D}\mathcal{F}^{\mathsf{H}}_{\beta,d}} \frac{\|g\|_\infty^2}{\epsilon_0 + \int g^2 \mathrm{d}\rho}.$$

For any $g \in \mathcal{D}\mathcal{F}^{\mathsf{H}}_{\beta,d}$, let x^g attain the maximum in the infinity norm. Furthermore, for any $j = (j_1, ..., j_d) \in \{0, 1, ..., 1/h - 1\}^d$, let $x^j = (j_1 h, j_2 h, ..., j_d h)$. Suppose $m \in \{0, 1, ..., 1/h - 1\}^d$ satisfies for $x^g = (x_1^g, ..., x_d^g)$ that $x_l^g \in [x_l^m, x_l^m + h)$ for each coordinate $l \in [d]$. Then we let the polynomial approximation of g up to degree $\lfloor\beta\rfloor$ be

$$p_g(x) = \sum_{\substack{\mathbf{k}=(k_1,\dots,k_d)\in\mathbb{Z}^d_{\geq 0} \\ k_1+\cdots+k_d\leq\lfloor\beta\rfloor}} \frac{D^{\mathbf{k}} g(x^m)}{k_1! \cdots k_d!} (x_1 - x_1^m)^{k_1} \cdots (x_d - x_d^m)^{k_d}.$$

Then, since g is Holder, we have

$$|g(x^g) - p_g(x^g)| \leq C_{\beta,d} h^\beta.$$

Furthermore, since π_h is G-optimal and p_g is linear in the $\mathcal{X}_h$, we have

$$p_g^2(x^g) \lesssim (\beta+1)^d \int p_g^2 \mathrm{d}T^m_{\#}\pi_h.$$

Combine with the inequality that $a^2 \geq b^2/2 - 3(a-b)^2$ and we will choose h small comparing with ϵ_0 to ensure the positivity of the denominator, we have

$$\begin{aligned}
\frac{\|g\|_\infty^2}{\epsilon_0 + \int g^2 \mathrm{d}\rho} &\lesssim \frac{(g(x^g) - p_g(x^g))^2 + p_g^2(x^g)}{\epsilon_0 + (\int p_g^2/2 - 3|g-p_g|^2 \mathrm{d}T^m_{\#}\pi_h)/N} \\
&\lesssim \frac{C^2_{\beta,L,d} h^{2\beta} + \int p_g^2 \mathrm{d}T^m_{\#}\pi_h}{\epsilon_0 - 3C^2_{\beta,L,d} h^{2\beta}/N + \int p_g^2 \mathrm{d}T^m_{\#}\pi_h/(2N)} \\
&\lesssim N,
\end{aligned}$$

where we choose $\epsilon_0 = C \cdot h^{2\beta}/N$ for C large enough, we have $N = (1/h)^d = \mathcal{O}(\epsilon_0^{-d/(2\beta+d)})$.

For the lower bound side, we adopt the lower bound example from Tikhomirov (1993). We assume L is large enough. Otherwise, we only need to scale accordingly. Consider

$$\phi(x) = \begin{cases} a\prod_{i=1}^d (1+x_i)^\beta (1-x_i)^\beta, & \text{if } |x_i| \leq 1, i \in [d], \\ 0 & \text{otherwise,} \end{cases}$$

where $a \geq 0$ is a scaling factor. This function belongs to $\mathcal{F}^{\mathsf{H}}_{\beta,d}$ (Tikhomirov, 1993) for a small enough. For any ϵ_0 sufficiently small, let $\epsilon = \epsilon_0^{\beta/(2\beta+d)}$. For any set U of non-empty interior, choose a set of points $x^1, ..., x^N$ that satisfies $|x^i_k - x^j_k| > 2\Delta$ for all $i, j \in [N]$, $k \in [d]$, and $\Delta = (\epsilon/a)^{1/\beta}$. By volumetric argument, we can find such a set with $N = \Omega(\epsilon^{-d/\beta})$. Consider functions

$$\mathcal{H} := \left\{ h^r(x) := \Delta^\beta \phi\left(\frac{x - x^r}{\Delta}\right) \mid r \in [N] \right\}.$$

The function set $\mathcal{H} \subset \mathcal{D}\mathcal{F}^{\mathsf{H}}_{\beta,d}$ for L large enough. Consider the cubes $C^r = \{x \mid \sup_{k\in[d]} |x_k - x^r_k| \leq \Delta\}$ for $r \in [N]$. Since these cubes are disjoint, then for any design ρ, there exists $r \in [N]$ such that $\rho(C^r) \leq 1/N$. Then we have,

$$\begin{aligned}
\mathcal{V}^\star(\mathcal{F}^{\mathsf{H}}_{\beta,d}; \epsilon_0) &\geq \frac{\|h^r\|_\infty^2}{\epsilon_0 + \int (h^r)^2 d\rho} \\
&\geq \frac{\epsilon^2/2^{4d\beta}}{\epsilon_0 + \epsilon^2/(2^{4d\beta} \cdot N)} \\
&\gtrsim \frac{\epsilon_0^{\beta/(2\beta+d)}}{\epsilon_0 + \epsilon_0^{\beta/(2\beta+d)} \cdot \epsilon_0^{d/(2\beta+d)}} \\
&= \Omega(\epsilon_0^{-d/(2\beta+d)}),
\end{aligned}$$

where the second inequality is by the fact that $\|\phi\|_\infty = a/2^{2\beta d}$ and the definition of Δ and the third inequality is by the definition of ϵ and N.

□

Proof of Example 3. We first separate the interval $[-B, B]$ into $\mathcal{O}(\log B)$ number of subintervals. Concretely, let $I_l = [2^l, 2^{l+1}]$, $J_l = [-2^{l+1}, -2^l]$ for $l \geq 1$, and $U = [-2, 2]$. For each interval in $I \in \mathcal{I} = \{I_l\}_{l\in[\log B]} \cup \{J_l\}_{l\in[\log B]} \cup \{U\}$. $\max_{x\in I}(1+x^2) \leq 5\min_{x\in I}(1+x^2)$. And for the polynomial function class $\mathcal{P}_k$ with degree bounded by k on interval I, there exists a design $\rho_{k,I}$ such that $\mathcal{V}(\mathcal{P}_k, \rho_{k,I}; 0) \leq k$. Finally, we consider $\rho = \frac{1}{|\mathcal{I}|}\sum_{I\in\mathcal{I}} \rho_{k,I}$ and obtain an upper bound of $\mathcal{V}^*(\mathcal{P}_{k,B}; \epsilon_0) = \mathcal{O}(k \log B)$. □

Proof of Theorem 3.2. For every $h \in \mathcal{DF}$, find any x_h such that $2|h(x_h)| \geq \|h\|_\infty$. For any ϵ that satisfies the assumption in the statement. Consider the minimum covering $\mathcal{DF}_\epsilon$ of $\mathcal{DF}$ in $\|\cdot\|_\infty$ and the uniform design ρ on the set $\mathcal{X}_\epsilon := \{x_h | h \in \mathcal{DF}_\epsilon\}$. Let $N = \mathcal{N}(\mathcal{DF}, \|\cdot\|_\infty, \epsilon) = |\mathcal{DF}_\epsilon|$. Then for any $h \in \mathcal{DF}$, suppose $\hat{h} \in \mathcal{DF}_\epsilon$ is such that $\|h - \hat{h}\|_\infty \leq \epsilon$. We have

$$\begin{aligned}
\frac{\|h\|_\infty^2}{\epsilon_0 + \int h^2 d\rho} &\leq \frac{2\|h-\hat{h}\|_\infty^2 + 2\|\hat{h}\|_\infty^2}{\epsilon_0 + h^2(x_{\hat{h}})/N} \\
&\leq \frac{2\|h-\hat{h}\|_\infty^2 + 2\|\hat{h}\|_\infty^2}{\epsilon_0 - 3(h-\hat{h})^2(x_{\hat{h}})/N + \|\hat{h}\|_\infty^2/(4N)} \\
&\leq \frac{2\epsilon^2 + 2\|\hat{h}\|_\infty^2}{\epsilon_0 - 3\epsilon^2/N + \|\hat{h}\|_\infty^2/(4N)} \\
&\leq \frac{2\epsilon^2 + 2\|\hat{h}\|_\infty^2}{\epsilon^2/N + \|\hat{h}\|_\infty^2/(4N)} \\
&\leq 8N.
\end{aligned}$$

where the second inequality is by $a^2 \geq b^2/2 - 3(a-b)^2$ and $|\hat{h}(x_{\hat{h}})| \geq \|\hat{h}\|_\infty/2$, the third inequality is by the definition of $\hat{h}$, and the fourth inequality is by the assumption that $\epsilon_0 > 4\epsilon^2/N$. □

D.3. Exponential Gap with Eluder Dimmension

Proof of Lemma 3.4. Let $\rho^k = \mathrm{Unif}(\mathcal{B}^k)$, then we have

$$\begin{aligned}
\mathcal{V}^*(\mathcal{F}^k; \nu_0) &\leq \mathcal{V}^*(\rho_k; \nu_0) \\
&= \sup_{x,h \in \mathcal{X} \times \mathcal{DF}} \frac{h^2(x)}{\nu_0(h) + \mathbb{E}_{x' \sim \rho^k}[h(x')^2]}.
\end{aligned}$$

For any $0 \neq h \in \mathcal{DF}$, there exists l such that $|h(b^l)| = \frac{1}{2}$. Together with the fact that $\sup_{x,h} h(x)^2 \leq 1$, we have

$$\mathcal{V}^*(\mathcal{F}^k; \nu_0) \leq \frac{1}{\frac{1}{k} h(b^l)^2} = 4k.$$

On the other hand, it is easy to verify that the sequence of (a_t, f^0, f^t) for $t \in [2^k - 1]$ with $\epsilon = 3/4$ satisfies the condition in the definition of eluder dimension (Definition 2.1). Thus we have

$$\mathrm{dimE}(\mathcal{F}^k; \epsilon_0) \geq 2^k - 1.$$

□

D.4. Exponential Gap for Disagreement Coefficient

Definition D.1 (Disagreement coefficient (Foster et al., 2020))**.** The disagreement coefficient for function class $\mathcal{F}$ and positive numbers $\Delta_0, \epsilon_0 \in (0, 1)$ is defined as

$$\boldsymbol{\theta}(\mathcal{F}, \Delta_0, \epsilon_0) = \sup_{\rho \in \mathcal{P}(\mathcal{X})} \sup_{\Delta \geq \Delta_0, \epsilon \geq \epsilon_0} \left\{ \frac{\Delta^2}{\epsilon^2} \cdot \mathbb{P}_{x \sim \rho} \left(\exists f \in \mathcal{F} : |f(x)| > \Delta, \mathbb{E}_{x \sim \rho} f^2(x) \leq \epsilon \right) \right\} \vee 1.$$

Proposition D.2. *For any integer $k \geq 1$, $\epsilon_0 = 1/2^k$, and $\Delta_0 \in (\epsilon_0, 1)$, we have for the cheating code function class $\mathcal{F}^k$,*

$$\boldsymbol{\theta}(\mathcal{F}^k, \Delta_0, \epsilon_0) \geq 2^{2k}.$$

Proof of Proposition D.2. Take $\rho = \mathrm{Unif}(\mathcal{A}_k)$, thus for any $a_i \in \mathcal{A}_k$, f^i satisfies $|f^i(a_i)| > \Delta_0$, $\mathbb{E}_{x \sim \rho} f^2(x) \leq \epsilon_0$. Thus we have

$$\boldsymbol{\theta}(\mathcal{F}^k, \Delta_0, \epsilon_0) \geq \frac{1}{\epsilon_0^2} = 2^{2k}$$

□

Proposition D.3. *For any* $\Delta_0, \epsilon_0 \in (0,1)$ *and function class* $\mathcal{F}$ *bounded in* $[-1,1]$, *we have*

$$\mathcal{V}^*(\mathcal{F};\epsilon_0) \le \mathcal{V}^*(\mathcal{F};\epsilon_0) \le 2\theta(\mathcal{F},\Delta_0,\sqrt{\epsilon_0})\log(1/\epsilon_0)\log(1/\Delta_0) + \frac{\Delta_0^2}{\epsilon_0} + 1.$$

Proof of Proposition D.3. We follow the proof of Lemma E.2 of Foster et al. (2021).

$$\begin{aligned}
\mathcal{V}^*(\mathcal{F};\epsilon_0) &= \inf_\rho \sup_{h,x} \frac{h^2(x)}{\epsilon_0 + \mathbb{E}_{x'\sim\rho} h^2(x')} \\
&= \inf_\rho \sup_\mu \mathbb{E}_{x\sim\mu}\left[\sup_{h\in\mathcal{DF}} \frac{h^2(x)}{\epsilon_0 + \mathbb{E}_{x'\sim\rho} h^2(x')}\right] \\
&\le \sup_\mu \mathbb{E}_{x\sim\mu}\left[\sup_{h\in\mathcal{DF}} \frac{h^2(x)}{\epsilon_0 + \mathbb{E}_{x'\sim\mu} h^2(x')}\right] \\
&\le \frac{\Delta_0^2}{\epsilon_0} + \sup_\mu \mathbb{E}_{x\sim\mu}\left[\sup_{h\in\mathcal{DF}} \frac{h^2(x) - \Delta_0^2}{\epsilon_0 + \mathbb{E}_{x'\sim\mu} h^2(x')}\right].
\end{aligned}$$

We fix any design μ. Then since for any $\epsilon_0 \le X \le 1$, we have

$$\frac{1}{X^2} = 2\int_X^1 \frac{1}{t^3} dt + 1 = 2\int_\epsilon^1 \frac{1}{t^3}\mathbb{1}(t \ge X) dt + 1.$$

Since $|f| \le 1$, we have

$$\mathbb{E}_{x\sim\mu}\left[\sup_{h\in\mathcal{DF}} \frac{h^2(x) - \Delta_0^2}{\epsilon_0 + \mathbb{E}_{x'\sim\mu} h^2(x')}\right] \le 1 + 2\mathbb{E}_{x\sim\mu}\left[\sup_{h\in\mathcal{DF}} \int_{\sqrt{\epsilon_0}}^1 \frac{h^2(x) - \Delta_0^2}{\epsilon^3}\mathbb{1}\left(\mathbb{E}_{x'\sim\mu} h^2(x') \le \epsilon^2\right) d\epsilon\right].$$

Similarly, we have for any $\Delta_0 \le X \le 1$

$$X^2 - \Delta_0^2 = \int_{\Delta_0^2}^1 \mathbb{1}(X^2 > t) dt = 2\int_{\Delta_0}^1 \mathbb{1}(X > t) t dt.$$

This leads to the upper bound

$$\begin{aligned}
&2\mathbb{E}_{x\sim\mu}\left[\sup_{h\in\mathcal{DF}} \int_{\sqrt{\epsilon_0}}^1 \frac{h^2(x) - \Delta_0^2}{\epsilon^3}\mathbb{1}\left(\mathbb{E}_{x'\sim\mu} h^2(x') \le \epsilon^2\right) d\epsilon\right] \\
&\le 4\mathbb{E}_{x\sim\mu}\left[\sup_{h\in\mathcal{DF}} \int_{\sqrt{\epsilon_0}}^1 \int_{\Delta_0}^1 \frac{\delta}{\epsilon^3}\mathbb{1}\left(|h(x)| > \delta \wedge \mathbb{E}_{x'\sim\mu} h^2(x') \le \epsilon^2\right) d\delta d\epsilon\right] \\
&\le 4\mathbb{E}_{x\sim\mu}\left[\int_{\sqrt{\epsilon_0}}^1 \int_{\Delta_0}^1 \frac{\delta}{\epsilon^3}\mathbb{1}\left(\exists h : |h(x)| > \delta \wedge \mathbb{E}_{x'\sim\mu} h^2(x') \le \epsilon^2\right) d\delta d\epsilon\right] \\
&\le 4\int_{\sqrt{\epsilon_0}}^1 \int_{\Delta_0}^1 \frac{\delta}{\epsilon^3}\mathbb{P}_{x\sim\mu}\left(\exists h : |h(x)| > \delta \wedge \mathbb{E}_{x'\sim\mu} h^2(x') \le \epsilon^2\right) d\delta d\epsilon \\
&\le 4\theta(\mathcal{F},\Delta_0,\sqrt{\epsilon_0}) \int_{\epsilon_0}^1 \int_{\Delta_0}^1 \frac{1}{\delta\epsilon} d\delta d\epsilon \\
&\le 2\theta(\mathcal{F},\Delta_0,\sqrt{\epsilon_0})\log(1/\epsilon_0)\log(1/\Delta_0).
\end{aligned}$$

In all, we have

$$\mathcal{V}^*(\mathcal{F};\epsilon_0) \le 2\theta(\mathcal{F},\Delta_0,\sqrt{\epsilon_0})\log(1/\epsilon_0)\log(1/\Delta_0) + \frac{\Delta_0^2}{\epsilon_0} + 1.$$

□

Definition D.4 (Star number (Foster et al., 2020))**.** For any function $f^* \in \mathcal{F}$, the value function star number $\mathcal{S}_{f^\star}(\mathcal{F}, \epsilon_0)$ is the length of the longest sequence of pairs $x_1, \ldots, x_m$ such that there exists $\epsilon > \epsilon_0$, where for all i, there exists $f_i \in \mathcal{F}$ such that

$$\left|f_i\left(x_i\right) - f^\star\left(x_i\right)\right| > \epsilon, \quad \text{ and } \quad \sum_{j \neq i}\left(f_i\left(x_j\right) - f^\star\left(x_j\right)\right)^2 \leq \epsilon^2.$$

Proposition D.5. *For any integer $k \geq 1$ and $\epsilon_0 \in (0,1)$, we have for the cheating code function class $\mathcal{F}^k$, for any $f^* \in \mathcal{F}^k$ such that*

$$\mathcal{S}_{f^*}(\mathcal{F}^k, \epsilon_0) \geq 2^k - 1.$$

Proof of Proposition D.5. Without loss of generality, take $f^* = f^0$, take $x_i = a_i$ for $i \in [2^k - 1]$ and $f_i = f^i$ in the definition of star number. We have $\mathcal{S}_{f^*}(\mathcal{F}^k, \epsilon_0) \geq 2^k - 1$. □

E. Proofs from Section 4

E.1. Proof of Design Sparsification

Proof of Lemma 4.1. For any $h \in \mathcal{DF}$, by Bernstein's inequality, we have with probability at least $1 - \delta'$,

$$\mathbb{E}_{x\sim\rho}(h(x))^2 - \frac{1}{n}\sum_{i=1}^{n}(h(x_i))^2 \leq \sqrt{\frac{2\mathrm{Var}_\rho(h^2)\ln(1/\delta)}{n}} + \frac{\ln(1/\delta')}{3n} \cdot \sup_x (h(x))^2.$$

Since $\mathrm{Var}_\rho(h^2) \leq \mathbb{E}_{x\sim\rho}(h(x))^2 \cdot \sup_x (h(x))^2$, we have

$$\begin{aligned}\sqrt{\frac{2\mathrm{Var}_\rho(h^2)\ln(1/\delta)}{n}} &\leq \sqrt{\frac{2\mathbb{E}_{x\sim\rho}(h(x))^2 \cdot \sup_x (h(x))^2 \ln(1/\delta')}{n}} \\ &\leq \frac{1}{2}\mathbb{E}_{x\sim\rho}(h(x))^2 + \frac{\ln(1/\delta')}{n} \cdot \sup_x (h(x))^2,\end{aligned}$$

where the second inequality is by AM-GM inequality. Together we have

$$\mathbb{E}_{x\sim\rho}(h(x))^2 - \frac{1}{n}\sum_{i=1}^{n}(h(x_i))^2 \leq \frac{1}{2}\mathbb{E}_{x\sim\rho}(h(x))^2 + \frac{2\ln(1/\delta')}{n} \cdot \sup_x (h(x))^2.$$

Reorganizing, we have

$$\mathbb{E}_{x\sim\rho}(h(x))^2 \leq 2 \cdot \left(\frac{1}{n}\sum_{i=1}^{n}(h(x_i))^2\right) + \frac{4\ln(1/\delta')}{n} \cdot \sup_x (h(x))^2.$$

Furthermore, by the choice of $n = 16\mathcal{V}(\mathcal{F}, \rho; \nu_0)\log(|\mathcal{DF}|/\delta) \leq 32\mathcal{V}(\mathcal{F}, \rho; \nu_0)\log|\mathcal{F}|$, $\delta' = \frac{\delta}{|\mathcal{DF}|}$ and the definition of $\mathcal{V}(\rho; \nu_0)$, we have

$$\begin{aligned}\mathbb{E}_{x\sim\rho}(h(x))^2 &\leq 2 \cdot \left(\frac{1}{n}\sum_{i=1}^{n}(h(x_i))^2\right) + \frac{4\ln(1/\delta')}{n} \cdot \sup_x (h(x))^2 \\ &\leq 2 \cdot \left(\frac{1}{n}\sum_{i=1}^{n}(h(x_i))^2\right) + \frac{\sup_x (h(x))^2}{2\mathcal{V}(\mathcal{F}, \rho; \nu_0)} \\ &\leq 2 \cdot \left(\frac{1}{n}\sum_{i=1}^{n}(h(x_i))^2\right) + \frac{1}{2}\left(\nu_0(h) + \mathbb{E}_{x\sim\rho}(h(x))^2\right).\end{aligned}$$

Reorganizing, we have with probability at least $1 - \delta/(2|\mathcal{DF}|)$,

$$\mathbb{E}_{x\sim\rho}(h(x))^2 + \nu_0(h) \leq 4 \cdot \left(\frac{1}{n}\sum_{i=1}^{n}(h(x_i))^2 + \nu_0(h)\right).$$

Finally, by the union bound we conclude that with probability at least $1-\delta$ that the empirical distribution $\hat{\rho} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(x_i)$ satisfies

$$\mathcal{V}(\mathcal{F}, \hat{\rho}; \nu_0) = \sup_{h,x} \frac{(h(x))^2}{\nu_0(h) + \mathbb{E}_{x' \sim \hat{\rho}}(h(x'))^2} \leq 4\mathcal{V}(\mathcal{F}, \rho; \nu_0).$$

□

E.2. Proofs of Theorem 4.2 and Lemma 3.3

Definition E.1 (Smoothed eluder dimension)**.** For the function class $\mathcal{F}$, $T > 0$ and $\epsilon_0 > 0$, let $\{(x_t, h_t) : t = 1, \ldots, T\}$ be an arbitrary sequence in $\mathcal{X} \times \mathcal{DF}$. We can define

$$\widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0) = \sup_{\{(x_t,h_t)\}_{t=1}^T} \sum_{t=1}^T \frac{h_t(x_t)^2}{T\epsilon_0 + \sum_{i=1}^t h_t(x_i)^2}.$$

Lemma E.2. *For any function class $\mathcal{F}$, $\epsilon_0 > 0$, and $\theta > 0$, suppose the sequence $\{(x_t, h_t) : t = 1, \ldots, T\}$ satisfies $h_t(x_t) \in (\theta, 2\theta]$ for all $t \in [T]$. Let $\mathrm{dimE}(\mathcal{F}; \theta) = d$ We have*

$$\sum_{t=1}^T \frac{h_t(x_t)^2}{T\epsilon_0 + \sum_{i=1}^t h_t(x_i)^2} \leq 8d\log(T/d).$$

Proof of Lemma E.2. Consider the following bucketing process as in Algorithm 2:

Algorithm 2 Bucketing

```
Initialize a list of buckets $\mathcal{L}_1 = \emptyset$
for $t = 1, \ldots, T$ do
    for $l = 1, \ldots, |\mathcal{L}_t|$ do
        if the $l$-th bucket $B_{l,t} \in \mathcal{L}_t$ satisfies $\sum_{x \in B_{l,t}} h_t^2(x) < \theta^2$ then
            Let $l_t = l$
            Update $B_{l,t+1} = B_{l,t} \cup \{x_t\}$
            break
        end if
    end for
    If $x_t$ is not added not any of the bucket in $\mathcal{L}_t$, then create a new bucket in the list $\mathcal{L}_{t+1} = \mathcal{L}_t \cup \{B_{|\mathcal{L}_t|+1,t+1} = \{x_t\}\}$
    and let $l_t = |\mathcal{L}_t| + 1$.
    For $l \neq l_t$, maintain the bucket $B_{l,t+1} = B_{l,t}$.
end for
```

For any $l \in [|\mathcal{L}_T|]$ and element x_t that lies in the bucket $B_{l,T}$ in the end, we have

$$\frac{h_t(x_t)^2}{T\epsilon_0 + \sum_{i=1}^t h_t(x_i)^2} \leq \frac{4\theta^2}{T\epsilon_0 + l\theta^2}.$$

Furthermore, by the definition of $\mathrm{dimE}(\mathcal{F}, \theta)$, each bucket can not contain more than $\mathrm{dimE}(\mathcal{F}, \theta)$ elements. Finally, by monotonicity of the upper bound $\frac{4\theta^2}{T\epsilon_0 + l\theta^2}$ with respect to the bucket number l, thus we have

$$\sum_{t=1}^T \frac{h_t(x_t)^2}{T\epsilon_0 + \sum_{i=1}^t h_t(x_i)^2} \leq 4d \sum_{l=1}^{T/d} \frac{\theta}{T\epsilon_0 + l\theta} \leq 8d\log T.$$

□

Lemma E.3. *For any function class $\mathcal{F}$ bounded by 1 and $\epsilon_0 \in (0,1)$, we have*

$$\widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0) \le 9\log^2(1/\epsilon_0) \cdot \mathrm{dimE}(\mathcal{F}; \sqrt{\epsilon_0}).$$

Proof of Lemma E.3. Let $K = \log(1/\epsilon_0)$ then we have $2^K\sqrt{\epsilon_0} \ge 1$. For any sequence $\{(x_t, h_t) : t = 1, \ldots, T\}$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{h_t(x_t)^2}{T\epsilon_0 + \sum_{i=1}^{t} h_t(x_i)^2} \\
&= \sum_{t=1}^{T} \frac{h_t(x_t)^2}{T\epsilon_0 + \sum_{i=1}^{t} h_t(x_i)^2} \mathbb{1}(h_t(x_t) \le \sqrt{\epsilon_0}) + \sum_{k=1}^{K}\sum_{t=1}^{T} \frac{h_t(x_t)^2}{T\epsilon_0 + \sum_{i=1}^{t} h_t(x_i)^2} \mathbb{1}(2^{k-1}\sqrt{\epsilon_0} < h_t(x_t) \le 2^k\sqrt{\epsilon_0}) \\
&\le 1 + \sum_{k=1}^{K}\sum_{t=1}^{T} \frac{h_t(x_t)^2}{T\epsilon_0 + \sum_{i=1}^{t} h_t(x_i)^2} \mathbb{1}(2^{k-1}\sqrt{\epsilon_0} < h_t(x_t) \le 2^k\sqrt{\epsilon_0}).
\end{aligned}$$

For any fixed $k \in [K]$, we can apply Lemma E.2. Thus we have

$$\begin{aligned}
\sum_{t=1}^{T} \frac{h_t(x_t)^2}{T\epsilon_0 + \sum_{i=1}^{t} h_t(x_i)^2} &\le 1 + 8\log T \sum_{k=1}^{K} \mathrm{dimE}(\mathcal{F}; 2^{k-1}\sqrt{\epsilon_0})) \\
&\le 9K \log T \cdot \mathrm{dimE}(\mathcal{F}; \sqrt{\epsilon_0})),
\end{aligned}$$

where the last inequality is by noting that the eluder dimension is monotonic in its second argument. Taking supremum on the sequence yields the desired result. □

Proof of Theorem 4.2. We first show that the maximization in line 2 of Algorithm 1 is equivalent to the related maximization problem:

$$x_t = \arg\max_{x \in \mathcal{X}} \sup_{h \in \mathcal{DF}} \frac{h^2(x)}{T\epsilon_0 + h^2(x) + \sum\limits_{s=1}^{t-1} h^2(x_s)} \tag{6}$$

The equivalence happens because the functions z/a and $z/(a+z)$ are maximized at the same argument z for any $a > 0$ by monotonicity. For the rest of the proof, we consider the form of x_t from Equation 6, as it is more amenable to our analysis.

We note that $\widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0) < T$ by definition. This is because

$$\begin{aligned}
\widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0) &= \sup_{\{(x_t,h_t)\}_{t=1}^{T}} \sum_{t=1}^{T} \frac{h_t(x_t)^2}{T\epsilon_0 + \sum_{i=1}^{t} h_t(x_i)^2} \\
&\le \sup_{\{(x_t,h_t)\}_{t=1}^{T}} \sum_{t=1}^{T} \frac{h_t(x_t)^2}{T\epsilon_0 + h_t(x_t)^2} < T.
\end{aligned}$$

For any $t \in [T]$ and arbitrary x_{T+1}, we have

$$\begin{aligned}
\sup_{h \in \mathcal{DF}} \frac{h(x_{t+1})^2}{T\epsilon_0 + \sum_{s=1}^{t+1} h(x_s)^2} &= \sup_{h \in \mathcal{DF}} \frac{h(x_{t+1})^2}{T\epsilon_0 + h(x_{t+1})^2 + \sum_{s=1}^{t} h(x_s)^2} \\
&\le \sup_{h \in \mathcal{DF}} \frac{h(x_{t+1})^2}{T\epsilon_0 + h(x_{t+1})^2 + \sum_{s=1}^{t-1} h(x_s)^2} \\
&\le \sup_{x} \sup_{h \in \mathcal{DF}} \frac{h(x)^2}{T\epsilon_0 + h(x)^2 + \sum_{s=1}^{t-1} h(x_s)^2} \\
&= \sup_{h \in \mathcal{DF}} \frac{h(x_t)^2}{T\epsilon_0 + h(x_t)^2 + \sum_{s=1}^{t-1} h(x_s)^2} \\
&= \sup_{h \in \mathcal{DF}} \frac{h(x_t)^2}{T\epsilon_0 + \sum_{s=1}^{t} h(x_s)^2},
\end{aligned}$$

where the first inequality is by the fact that $\sum_{s=1}^{t} h(x_s)^2 > \sum_{s=1}^{t-1} h(x_s)^2$, the second inequality is by viewing x_{t+1} as a variable then replace it with its supreme and the second equality is by the definition of x_t. With these inductive inequalities, we further have for any $t \in [T]$,

$$\sup_{h \in \mathcal{DF}} \frac{h(x_{T+1})^2}{T\epsilon_0 + \sum_{s=1}^{T+1} h(x_s)^2} \le \sup_{h \in \mathcal{DF}} \frac{h(x_t)^2}{T\epsilon_0 + \sum_{s=1}^{t} h(x_s)^2}.$$

Summing up both sides for $t \in [T]$, we have

$$\begin{aligned} T \sup_{h \in \mathcal{DF}} & \frac{h(x_{T+1})^2}{T\epsilon_0 + \sum_{s=1}^{T+1} h(x_s)^2} \\ & \le \sum_{t=1}^{T} \sup_{h \in \mathcal{DF}} \frac{h(x_t)^2}{T\epsilon_0 + \sum_{s=1}^{t} h(x_s)^2} \\ & = \sup_{(h_t)_{t=1}^T \in (\mathcal{DF})^{\otimes T}} \sum_{t=1}^{T} \frac{h_t(x_t)^2}{T\epsilon_0 + \sum_{s=1}^{t} h_t(x_s)^2} \\ & \le \sup_{\{(x_t, h_t)\}_{t=1}^T} \sum_{t=1}^{T} \frac{h_t(x_t)^2}{T\epsilon_0 + \sum_{i=1}^{t} h_t(x_i)^2} = \widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0), \end{aligned}$$

where the first equality is because that h_t are independently maximizing each summand and the the second inequality is by viewing x_t as variables and then replacing them by the supreme with abuse of notation. Recall that x_{T+1} can be arbitrary. This implies that for any h and x,

$$T \frac{h(x)^2}{T\epsilon_0 + h(x)^2 + \sum_{s=1}^{T} h(x_s)^2} \le \widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0).$$

Changing the formulation we have

$$(T - \widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0))h(x)^2 \le T\left(\epsilon_0 + \frac{1}{T}\sum_{s=1}^{T} h(x_s)^2\right)\widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0) = T(\epsilon_0 + \mathbb{E}_{x' \sim \rho_T} h(x')^2)\widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0).$$

Moving terms around and taking supreme over x, h, we can obtain

$$\mathcal{V}(\mathcal{F}, \rho_T; \nu_0) \le \frac{T}{T - \widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0)} \cdot \widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0).$$

Combine the above result with Lemma E.3 and the assumption that $18\log(1/\epsilon_0)\log T \cdot \mathrm{dimE}(\mathcal{F}; \sqrt{\epsilon_0}) \le T$, we have

$$\begin{aligned} \mathcal{V}(\mathcal{F}, \rho_T; \nu_0) & \le \frac{T}{T - \widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0)} \cdot \widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0) \\ & \le 2\widetilde{\mathrm{dimE}}(\mathcal{F}, \epsilon_0) \\ & \le 18\log(1/\epsilon_0)\log T \cdot \mathrm{dimE}(\mathcal{F}; \sqrt{\epsilon_0}). \end{aligned}$$

Finally, $18\log(1/\epsilon_0)\log T \cdot \mathrm{dimE}(\mathcal{F}; \sqrt{\epsilon_0}) \le T$ is achieved when $T = Cd\log(1/\epsilon_0)(\log d + \log\log(1/\epsilon_0))$ for C large enough. Thus plug in this T on the right-hand side, and we have

$$\mathcal{V}(\mathcal{F}, \rho_T; \epsilon_0) = \mathcal{O}\left(d\log(1/\epsilon_0) \cdot (\log d + \log\log(1/\epsilon_0))\right).$$

□

Algorithm 3 FW algorithm for F-design

```
Require: Design space $\mathcal{X}$, function class $\mathcal{F}$, time horizon $T$, base measure $\nu_0 > 0$, stepsizes $\eta_1, \dots, \eta_T \in [0, 1]$
$\rho_0 = \mathrm{Unif}(\mathcal{X})$
for $t = 1, \dots, T$ do
    Find $x_t \in \mathcal{F}$ so that
        $x_t = \arg\max_{x \in \mathcal{X}} \sup_{h \in \mathcal{DF}} \frac{h^2(x)}{\nu_0(h) + \mathbb{E}_{x' \sim \rho_{t-1}} h^2(x')}.$
    Update $\rho_t = (1 - \eta_t)\rho_{t-1} + \eta_t \mathbb{1}(x_t)$.
end for
Return $\rho_T$.
```

E.3. Hardness with Argmax Oracle

We consider the generalization of FW algorithm for linear G-optimal design (Todd, 2016) which is motivated by the Frank-Wolfe algorithm to a more general form of Algorithm 1 as shown in Algorithm 3. This algorithm reduces to Algorithm 1 if the learning rates $\eta_t = \frac{1}{t}$ for all $t \in [T]$.

A more favorbale property of Algorithm 3 is that by choosing $\{\eta_t\}_{t=1}^T$ more adaptively, the algorithm can converge in $\mathcal{O}(d(\log\log|\mathcal{X}| + \log d))$ (Todd, 2016) steps in the linear case. However, in this section, we show that it is not possible to achieve $\mathcal{V}(\mathcal{F}, \rho_T; \nu_0) \leq \mathcal{O}(\mathcal{V}^*(\mathcal{F}; \nu_0))$ within $\widetilde{\mathcal{O}}(\mathcal{V}^*(\mathcal{F}; \nu_0))$ number of steps. In fact, for the cheating code example (Example 4), the algorithm will not converge to $\mathcal{O}(\mathcal{V}^*(\mathcal{F}; \nu_0))$ at all for any $T \geq 1$.

Theorem E.4. *For any integer $k > 1$, let $\mathcal{F}^k$ be the cheating code class as in Example 4. Then for any choice of $\{\eta_t\}_{t=1}^T$ (possibly adaptively) and $t \in [T]$, the ρ_t in Algorithm 3 satisfies*

$$\mathcal{V}(\mathcal{F}^k, \rho_t; \nu_0) \geq \min\left\{\frac{2^k}{k+2}, \frac{1}{\sup_{h \in \mathcal{DF}^k} \nu_0(h)}\right\}.$$

This result suggests that we need to consider a different class of algorithms for obtaining near-optimal designs efficiently.

Proof of Theorem E.4. For any fixed k and design ρ,

$$\hat{x} = \operatorname*{argmax}_{x \in \mathcal{X}} \sup_{f, f' \in \mathcal{F}^k} \frac{(f(x) - f'(x))^2}{\nu_0(f - f') + \mathbb{E}_{x' \sim \rho_{t-1}}(f(x') - f'(x'))^2} = \operatorname*{argmax}_{x \in \mathcal{X}} (f^i(x) - f^j(x))^2,$$

where f^i, f^j are the two functions that achieve the inner supreme. It is clear that $f^i \neq f^j$, thus $\hat{x} = a_i$ or a_j. Thus we have $x_t \in \mathcal{A}^k$ for any $t \in [T]$. Therefore, for any chosen sequence of $\{\eta_t\}_{t=1}^T$, we know that the mass on $\mathcal{A}^k$ will only increase, i.e.,

$$\rho_T(\mathcal{A}^k) \geq \rho_{T-1}(\mathcal{A}^k) \geq \cdots \geq \rho_0(\mathcal{A}^k) = \frac{2^k}{2^k + k}.$$

Thus there exists i, j such that $\rho_T(a_i) + \rho_T(a_j) \leq \frac{1}{2^{k-1}}$ by pigeohole, and also we have $\rho_T(\mathcal{B}^k) \leq \frac{k}{2^k + k}$. Thus

$$\begin{aligned}
\mathcal{V}(\mathcal{F}^k, \rho_T; \nu_0) &\geq \frac{(f^i(a_i) - f^j(a_i))^2}{\nu_0(f^i - f^j) + \mathbb{E}_{x' \sim \rho_T}(f^i(x') - f^j(x'))^2} \\
&\geq \frac{1}{\sup_{h \in \mathcal{DF}} \nu_0(h) + \frac{1}{2^{k-1}} + \frac{k}{2^k + k}} \\
&\geq \min\left\{\frac{2^k}{k+2}, \frac{1}{\sup_{h \in \mathcal{DF}} \nu_0(h)}\right\}.
\end{aligned}$$

□

Algorithm 4 Subgradient descent for non-linear G-optimal design

```
Require: Function class $\mathcal{F}$, $\mathcal{C} = \mathcal{P}(\mathcal{X})$, time horizon $T$
Set $\rho_0 = \mathrm{Unif}(\mathcal{X})$.
for $t = 1, 2, \ldots, T$ do
    Find $x_t \in \mathcal{X}, h_t \in \mathcal{H}_{\mathcal{F}}$ so that
        $x_t, h_t = \arg\max_{x,h} \frac{h^2(x)}{\nu_0(h) + \mathbb{E}_{x' \sim \rho_{t-1}} h^2(x')}$.
    Let $g_t = \frac{h_t^2(x_t)}{(\nu_0(h_t) + \mathbb{E}_{x' \sim \rho_{t-1}} h_t^2(x'))^2} \cdot h_t^2$.
    Let $\rho_t = P_{\mathcal{C}}\left(\rho_{t-1} + g_t/\sqrt{t}\right)$.
end for
return $\hat{\rho}_T = \mathrm{argmin}_{\rho \in \{\rho_s\}_{s=1}^T} \mathcal{V}(\mathcal{F}, \rho; \nu_0)$.
```

E.4. Projected Subgradient Descent with Argmax Oracle

Theorem E.5. *Suppose function class $\mathcal{F}$ is bounded by 1 and the base measure ν_0 is lower bounded $\nu_0(h) \geq \epsilon > 0$ for all $h \in \mathcal{DF}$. Then Algorithm 4 ensures that*

$$\mathcal{V}(\mathcal{F}, \hat{\rho}_T; \nu_0) - \mathcal{V}^*(\mathcal{F}; \nu_0) \leq \frac{2\epsilon^4 + 256|\mathcal{X}| \log t}{\epsilon^4 \sqrt{t}}.$$

Proof of Theorem E.5. To apply Proposition C.1, we first verify that $\mathcal{V}(\mathcal{F}, \rho; \nu_0)$ is a Lipschitz convex function in ρ. For convexity, we have since

$$\mathcal{V}(\mathcal{F}, \rho; \nu_0) = \sup_{h \in \mathcal{D}f, x \in \mathcal{X}} \frac{h(x)^2}{\nu_0(h) + \mathbb{E}_{x' \sim \rho} h(x')^2}$$

is a supreme of convex functions. For the Lipschitzness, we have that for any fixed $h, x \in \mathcal{DF} \times \mathcal{X}$, the gradient is bounded by

$$\begin{aligned} \left\| \partial_\rho \frac{h(x)^2}{\nu_0(h) + \mathbb{E}_{x' \sim \rho} h^2(x')} \right\| &= \left\| \frac{h^2(x) h^2(\cdot)}{\left(\nu_0(h) + \mathbb{E}_{x' \sim \rho} h^2(x')\right)^2} \right\| \\ &\leq \frac{16}{\epsilon^2} \|e\| \leq \frac{16}{\epsilon^2} \sqrt{|\mathcal{X}|}, \end{aligned}$$

where $e = (1, \ldots, 1) \in \mathbb{R}^{\mathcal{X}}$. Thus as $\mathcal{V}(\mathcal{F}, \rho; \nu_0)$ as a supreme of such functions also has the same Lipschitz constant. Finally, since $\mathcal{V}(\mathcal{F}, \rho; \nu_0)$ is a supreme of convex functions, by Danskin's theorem, we verify that

$$-\frac{h_t^2(x_t) h_t^2(\cdot)}{(\nu_0(h_t) + \mathbb{E}_{x' \sim \rho_{t-1}} h_t^2(x'))^2} = \partial_\rho \frac{h_t^2(x_t)}{\nu_0(h_t) + \mathbb{E}_{x' \sim \rho_{t-1}} h_t^2(x')}$$

is indeed a subgradient of the function $\mathcal{V}(\mathcal{F}, \rho_{t-1}; \nu_0)$. Combine above and invoke Proposition C.1 we have

$$\mathcal{V}(\mathcal{F}, \hat{\rho}_T; \nu_0) - \mathcal{V}^*(\mathcal{F}; \nu_0) \leq \frac{2 + 256|\mathcal{X}| \log t / \epsilon^4}{\sqrt{t}} = \frac{2\epsilon^4 + 256|\mathcal{X}| \log t}{\epsilon^4 \sqrt{t}}.$$

□

The above result shows that with an argmax oracle, we can approximate the non-linear G-optimal design in principle. Nevertheless, the computation is not ideal since each time we have to update the policy ρ_t in a very complicated way which in the worst case has a computation complexity that scales with the number of possible instances $|\mathcal{X}|$.

F. Proofs from Section 5

F.1. Proofs for Simultaneous Confidence Bands

Proof of Theorem 5.1. By the definition of $\rho_{\mathcal{V}}$, we have for all $x \in \mathcal{X}$,

$$\begin{aligned} |\widehat{f}_n(x) - f^\star(x)| \leq& \sqrt{\nu_0(\widehat{f}_n - f^\star) + \mathbb{E}_{x\sim\rho_{\mathcal{V}}}(\widehat{f}(x) - f^*(x))^2} \cdot \sqrt{\mathcal{V}(\mathcal{F}, \rho^*, x; \nu_0)} \\ \leq& \mathcal{O}\left(\sqrt{\epsilon_0 + \|\widehat{f} - f^*\|^2_{\rho_{\mathcal{V}},2}} \cdot \sqrt{\mathcal{V}(\mathcal{F}, \rho_{\mathcal{V}}, x; \nu_0)}\right). \end{aligned}$$

□

F.2. Proofs for Contextual Bandits

Lemma F.1. *Let f_T be as defined as satisfying* (4) *and π_T be its greedy policy. Then we have*

$$\begin{aligned} \mathbb{E}_{z\sim D}\left[f^\star(z, a^\star(z)) - f^\star(z, \pi_T(z))\right] &\leq 2\sqrt{\mathbb{E}_{z\sim D}\left[\max_{a\in\mathcal{A}}\left(f^\star(z,a) - \widehat{f}_T(z,a)\right)^2\right]} \\ &\leq 2\sqrt{\mathbb{E}_{z\sim D}\left[\sum_{a\in\mathcal{A}}\left(f^\star(z,a) - \widehat{f}_T(z,a)\right)^2\right]}. \end{aligned}$$

Proof. We adapt the proof from Agarwal et al. (2012). Let $a^\star(z) = \max_a f^\star(z, a)$. By Jensen's inequality, we have

$$\begin{aligned} &\mathbb{E}_{z\sim D}\left[f^\star(z, a^\star(z)) - f^\star(z, \pi_T(z))\right] \leq \sqrt{\mathbb{E}_{z\sim D}\left[\left(\max_a f^\star(z,a) - f^\star(z,\pi_T(z))\right)^2\right]} \\ =&\sqrt{\mathbb{E}_{z\sim D}\left[\left(f^\star(z,a^\star(z)) \pm \widehat{f}_T(z,\pi_T(z)) - f^\star(z,\pi_T(z))\right)^2\right]} \\ \leq&\sqrt{\mathbb{E}_{z\sim D}\left[\left(f^\star(z,a^\star(z)) - \widehat{f}_T(z,a^\star(z)) + \widehat{f}_T(z,\pi_T(z)) - f^\star(z,\pi_T(z))\right)^2\right]}, \end{aligned}$$

where the last inequality follows since $f^\star(z, a^\star(z)) - f^\star(z, \pi_T(z)) \geq 0$ and $f_T(z, \pi_T(z)) \geq f_T(z, a^\star(z))$. Proceeding further, we have by Cauchy-Schwarz inequality

$$\begin{aligned} &\mathbb{E}_{z\sim D}\left[f^\star(z, a^\star(z)) - f^\star(z, \pi_T(z))\right] \\ \leq&\sqrt{2\mathbb{E}_{z\sim D}\left[\left(f^\star(z,a^\star(z)) - \widehat{f}_T(z,a^\star(z))\right)^2 + \left(\widehat{f}_T(z,\pi_T(z)) - f^\star(z,\pi_T(z))\right)^2\right]} \\ \leq&\sqrt{2\mathbb{E}_{z\sim D}\left[2\max_{a\in\mathcal{A}}\left(f^\star(z,a) - \widehat{f}_T(z,a)\right)^2\right]}. \end{aligned}$$

□

Proof of Theorem 5.3. The proof is relatively simple given assumption (4) and Lemma F.1. By definition of optimal design, we have for any $z \in \mathcal{Z}$ and $a \in \mathcal{A}$:

$$\begin{aligned} (\widehat{f}_T(z,a) - f^\star(z,a))^2 \leq& d\left(\nu_0(f - f^*) + \mathbb{E}_{a'\sim\rho_{\mathcal{V}}(\cdot|z)}[(\widehat{f}_T(z,a') - f^\star(z,a'))^2]\right) \\ \leq& d\left(\epsilon_0 + \mathbb{E}_{a'\sim\rho_{\mathcal{V}}(\cdot|z)}[(\widehat{f}_T(z,a') - f^\star(z,a'))^2]\right). \end{aligned}$$

Take expectation over $z \sim D$, and combine with assumption (4), we have with probability at least $1-\delta$,

$$\begin{aligned}
\mathsf{Reg}(\pi_T) &= \mathbb{E}_{z\sim D}\left[f^\star(z, a^\star(z)) - f^\star(z, \pi_T(z))\right] \\
&\le \sqrt{2\mathbb{E}_{z\sim D}\left[2\max_{a\in\mathcal{A}}\left(f^\star(z,a) - \widehat{f}_T(z,a)\right)^2\right]} \\
&\le 4\sqrt{d\left(\epsilon_0 + \mathbb{E}_{z\sim D, a\sim\rho_{\mathcal{V}}(\cdot|z)}[(\widehat{f}_T(z,a) - f^\star(z,a))^2]\right)} \\
&= \mathcal{O}\left(\sqrt{d\cdot(\mathbf{Est}_{\mathsf{Off}}(T,\delta)+\epsilon_0)}\right).
\end{aligned}$$

□

Lemma F.2. *Let $(z_t, a_t, r_t)_{t=1}^T$ be samples where $z_t \overset{i.i.d.}{\sim} D$, $a_t \sim \rho_{\mathcal{V}}(\cdot|z_t)$, and $r_t \sim D(\cdot|z_t, a_t)$. Then for the least-squares regressor $\widehat{f}_T$ defined as:*

$$\widehat{f}_T = \arg\min_{f\in\mathcal{F}} \sum_{t=1}^T (f(z_t, a_t) - r_t)^2,$$

we have with probability at least $1-\delta$:

$$\mathbb{E}_{z\sim D, a\sim\rho_{\mathcal{V}}(\cdot|z)}[(f_T(z,a) - f^\star(z,a))^2] = \mathcal{O}\left(\frac{\log(|\mathcal{F}|/\delta)}{T}\right).$$

Proof of Lemma F.2. The distribution of the joint input of (z_t, a_t) are i.i.d., thus we can apply Lemma C.2. □

F.2.1. Connections to Contextual PAC DEC

In this section, we show that the offset version of the Probably Approximately Correct (PAC) version of the Decision Estimation Coefficient (DEC) (Chen et al., 2022; Foster et al., 2023a) for contextual bandits can be upper bounded by the F-condition number. However, the minimax PAC regret upper bound obtained from Theorem 5.3 and that obtained from the PAC DEC are not comparable. We start by introducing the definition of PAC DEC as follows.

Definition F.3 (Offset PAC DEC for contextual bandit)**.** For any function class $\mathcal{F}$ and parameter $\gamma > 0$, the offset version of PAC DEC for context bandit with square loss is defined as

$$\mathsf{p\text{-}dec}_\gamma(\mathcal{F}) \triangleq \sup_{\bar{f}, z} \inf_{p,q} \sup_{f\in\mathcal{F}_z} f(z, \pi_f(z)) - \mathbb{E}_{a\sim p(z)} f(z,a) - \gamma\mathbb{E}_{a\sim q(z)}(f(z,a) - \bar{f}(z,a))^2,$$

where $\pi_f(z) := \operatorname{argmax}_a f(z,a)$, p, q are any map from $\mathcal{Z}$ to $\mathcal{P}(\mathcal{A})$ and $\bar{f}$ is any function from $\mathcal{Z}\times\mathcal{A}$ to $[0,1]$.

We can bound the offset PAC DEC via the F-condition number.

Theorem F.4. *For any function class $\mathcal{F}$, let $\sup_z \mathcal{V}^*(\mathcal{F}_z, \mathcal{A}; \nu_0) \le d$ and $\sup_{h\in\mathcal{D}\mathcal{F}_z} \nu_0(h) \le \epsilon_0$, then we have*

$$\mathsf{p\text{-}dec}_\gamma(\mathcal{F}) \le \frac{d}{\gamma} + 2\sqrt{d\epsilon_0}.$$

Proof of Theorem F.4. For any context z, $f, \bar{f} \in \mathcal{F}_z$, we have

$$\begin{aligned}
f(z,\pi_f(z)) - f(z,\pi_{\bar{f}}(z)) &= f(z,\pi_f(z)) - \bar{f}(z,\pi_f(z)) + \bar{f}(z,\pi_f(z)) - \bar{f}(z,\pi_{\bar{f}}(z)) + \bar{f}(z,\pi_{\bar{f}}(z)) - f(z,\pi_{\bar{f}}(z)) \\
&\le 2\sup_a |f(z,a) - \bar{f}(z,a)|,
\end{aligned}$$

where $\pi_f(z) := \operatorname{argmax}_a f(z,a)$ and $\pi_{\bar{f}}(z) := \operatorname{argmax}_a \bar{f}(z,a)$ are defined as the greedy policy with respect to f and f' and the inequality is by $\bar{f}(z,\pi_f(z)) \le \bar{f}(z,\pi_{\bar{f}}(z))$.

Thus for any $\bar{f}$ and z, take $p = \pi_{\bar{f}}$ and $q = \rho_{\mathcal{V}}(\mathcal{F}, \mathcal{X}; \nu_0)$, we have

$$\begin{aligned}
\textsf{p-dec}_\gamma(\mathcal{F}) &\triangleq \sup_{\bar{f},z} \inf_{p,q} \sup_{f\in\mathcal{F}_z} f(z,\pi_f(z)) - \mathbb{E}_{a\sim p(z)} f(z,a) - \gamma\mathbb{E}_{a\sim q(z)}(f(z,a)-\bar{f}(z,a))^2 \\
&\le \sup_{\bar{f},z} \sup_{f\in\mathcal{F}_z} \sup_{a'} 2|f(z,a')-\bar{f}(z,a')| - \gamma\mathbb{E}_{a\sim q(z)}(f(z,a)-\bar{f}(z,a))^2 \\
&\le \sup_{\bar{f},z} \sup_{f\in\mathcal{F}_z} 2\sqrt{d\left(\nu_0(f-\bar{f}) + \mathbb{E}_{a\sim q(z)}(f(z,a)-\bar{f}(z,a))^2\right)} - \gamma\mathbb{E}_{a\sim q(z)}(f(z,a)-\bar{f}(z,a))^2 \\
&\le 2\sqrt{d\epsilon_0} + \frac{d}{\gamma},
\end{aligned}$$

where the second inequality is by the definition of q and the final inequality is by Cauchy-Schwarz inequality.

□

To bound the minimax PAC regret through the PAC DEC, the results in Foster et al. (2023a); Chen et al. (2022) require, as a sub-procedure, an online estimation oracle $\mathbf{Alg}_{\mathbf{Est}}$. Suppose at step t, the learner chooses policy $q_t : \mathcal{Z} \to \mathcal{A}$. This oracle, at each step t, takes the input of $z_1, a_1, r_1, ..., z_{t-1}, a_{t-1}, r_{t-1}$, then outputs a predictor $\widehat{f}_t$ such that with probability at least $1-\delta$,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{z\sim D, a\sim q_t(\cdot|z)}[(\widehat{f}_t(z,a) - f^\star(z,a))^2] \le \mathbf{Est}_{\textsf{On}}(T,\delta), \tag{7}$$

where $\mathbf{Est}_{\textsf{On}}(T,\delta)$ is an known upper bound.

Proposition F.5 ((Chen et al., 2022)). *For any parameter $\gamma > 0$, with the online oracle $\mathbf{Alg}_{\mathbf{Est}}$, there is an algorithm that achieves with probability at least $1-\delta$,*

$$\textsf{Reg}(\pi_T) = \mathcal{O}\left(\textsf{p-dec}_\gamma(\mathcal{F}) + \gamma \cdot \mathbf{Est}_{\textsf{On}}(T,\delta)\right)$$

for $\delta \in (0,1)$.

Algorithm 5 Contextual Explorative E2D

```
Require: Exploration parameter $\gamma > 0$, online estimation oracle $\mathbf{Alg}_{\mathbf{Est}}$.
for $t = 1, 2, \cdots, T$ do
    Receive context $z_t$.
    Obtain the estimation from the estimation oracle, $\widehat{f}_t = \mathbf{Alg}_{\mathbf{Est}}\left(\{(z_i, a_i, r_i)\}_{i=1}^{t-1}\right)$.
    Define
        $p_t, q_t := \operatorname*{argmin}_{p,q:\mathcal{Z}\to\mathcal{P}(\mathcal{A})} \sup_{f\in\mathcal{F}_z} f(z,\pi_f(z)) - \mathbb{E}_{a\sim p(z)} f(z,a) - \gamma\mathbb{E}_{a\sim q(z)}(f(z,a) - \widehat{f}_t(z,a))^2.$
    Sample decision $a_t \sim q_t(\cdot|z_t)$ and update estimation oracle with $(z_t, a_t, r_t)$.
end for
return $\pi_T = \frac{1}{T}\sum_{t=1}^{T} p_t$
```

Proof of Proposition F.5. We go through the contextualization in section 8 of Foster et al. (2021) for Theorem 10 of Chen et al. (2022). By the definition of the PAC DEC, we first have

$$\begin{aligned}
\max_a f^\star(z,a) - f^\star(z,\pi_T(z)) &= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{a\sim p_t(z)}[f^\star(z,\pi^*(z)) - f^\star(z,a)] \\
&\le \frac{\gamma}{T}\sum_{t=1}^{T} \mathbb{E}_{a\sim q_t(z)}\left[\left(f^\star(z,a) - \widehat{f}_t(z,a)\right)^2\right] + \textsf{p-dec}_\gamma(\mathcal{F}).
\end{aligned}$$

Thus we have

$$\begin{aligned}\mathsf{Reg}(\pi_T) &= \mathbb{E}_{z\sim D}\left[\max_a f^\star(z,a) - f^\star(z,\pi_T(z))\right]\\ &\le \textsf{p-dec}_\gamma(\mathcal{F}) + \frac{\gamma}{T}\sum_{t=1}^T \mathbb{E}_{z\sim D, a\sim q_t(z)}\left[\left(f^\star(z,a) - \widehat{f}_t(z,a)\right)^2\right]\\ &= \mathcal{O}\left(\textsf{p-dec}_\gamma(\mathcal{F}) + \gamma\cdot \mathbf{Est}_{\mathsf{On}}(T,\delta)\right),\end{aligned}$$

with probability at least $1-\delta$ for $\delta\in(0,1)$. □

As an example, the exponentially-weighted aggregating forecaster (Cesa-Bianchi & Lugosi, 2006) using the square loss achieves with probability at least $1-\delta$, a known upper bound of $\mathbf{Est}_{\mathsf{On}}(T,\delta) = \mathcal{O}(\log(|\mathcal{F}|/\delta)/T)$.

Lemma F.6. *The exponentially weighted aggregating forecaster of*

$$\widehat{f}_t = \sum_{f\in\mathcal{F}} \mu_t(f) f, \quad \textit{where} \quad \mu_t(f) \propto \exp\left(-\sum_{s=1}^{t-1}(f(z_s,a_s) - r_s)^2\right),$$

achieves with probability at least $1-\delta$

$$\sum_{t=1}^T \mathbb{E}_{z\sim D, a\sim q_t(z)}\left[\left(f^\star(z,a) - \widehat{f}_t(z,a)\right)^2 \mid \mathfrak{F}_{t-1}\right] \le 2\log|\mathcal{F}| + 16\log(1/\delta),$$

for $\delta\in(0,1)$.

Proof of Lemma F.6. The proof is standard and we follow the development of Foster & Rakhlin (2020) with only the difference in the choice of the filtration. The exponential weight aggregation with square loss (Proposition 3.2 of Cesa-Bianchi & Lugosi (2006)) satisfies

$$\sum_{t=1}^T (\widehat{f}_t(z_t,a_t) - r_t)^2 - \sum_{t=1}^T (f^*(z_t,a_t) - r_t)^2 \le \log|\mathcal{F}|.$$

Consider the filtration:

$$\mathfrak{F}_{t-1} = \sigma((z_1,a_1,r_1),...,(z_{t-1},a_{t-1},r_{t-1})).$$

Define $M_t = (\widehat{f}_t(z_t,a_t) - r_t)^2 - (f^*(z_t,a_t) - r_t)^2$ and $Z_t = \mathbb{E}[M_t|\mathcal{F}_{t-1}] - M_t$. Then we have $|Z_t|\le 1$ and

$$\begin{aligned}\mathbb{E}[Z_t^2 \mid \mathfrak{F}_{t-1}] &\le \mathbb{E}[M_t^2 \mid \mathfrak{F}_{t-1}]\\ &= \mathbb{E}[(\widehat{f}_t(z_t,a_t) - f^*(z_t,a_t))^2(\widehat{f}_t(z_t,a_t) + f^*(z_t,a_t) - 2r_t)^2 \mid \mathfrak{F}_{t-1}]\\ &\le 4\mathbb{E}[(\widehat{f}_t(z_t,a_t) - f^*(z_t,a_t))^2 \mid \mathfrak{F}_{t-1}]\\ &= 4\mathbb{E}[M_t \mid \mathfrak{F}_{t-1}].\end{aligned}$$

Then by Lemma 1 of (Foster & Rakhlin, 2020), take $\eta = 1/8$, we have

$$\sum_{t=1}^T \mathbb{E}_{z\sim D, a\sim q_t(z)}\left[\left(f^\star(z,a) - \widehat{f}_t(z,a)\right)^2 \mid \mathfrak{F}_{t-1}\right] \le 2\log|\mathcal{F}| + 16\log(1/\delta).$$

□

Although the PAC DEC can be upper bounded via the optimal F-condition number, the bounds obtained from Theorem 5.3 and from Proposition F.5 are not comparable as $\mathbf{Est}_{\mathsf{Off}}(T,\delta)$ and $\mathbf{Est}_{\mathsf{On}}(T,\delta)$ might differ.

F.3. Proofs for Model-free RL

An upshot of Assumption 5.6 is that it gives us a function class to express the possible Bellman errors, since for any h and $f \in \mathcal{F}$, we can write $\mathcal{E}^h(f, z, a) = f^h(z, a) - g^h(z, a)$ for some $g^h \in \mathcal{F}^h$. This allows us to obtain a one-step exploration policy π^h_{exp} by choosing the optimal F-design (2) with sample space $\mathcal{Z} = \mathcal{A}$ and function class $\mathcal{F}^h_z = \{f^h(z, a) \,:\, f^h \in \mathcal{F}^h\}$, for each $z \in \mathcal{Z}$. For this approach, we have the following lemma.

Lemma F.7. *Let* $\sup_{h \in \mathcal{D}\mathcal{F}^h_z} \nu_0(h) \leq \epsilon_0$ *for some* ϵ_0 *and* $\pi^h_z = \rho_{\mathcal{V}}(\mathcal{F}^h_z; \nu^h_0)$ *be an optimal* F*-design. Suppose that* $\mathcal{V}(\mathcal{F}^h_z, \pi^h_z; \nu_0) \leq d^h$ *for all* $z \in \mathcal{Z}$*. Then we have for all* $f \in \mathcal{F}$

$$\max_{a \in \mathcal{A}} \mathcal{E}^h(f, z, a)^2 \leq d^h \mathbb{E}_{a \sim \pi^h_z} \mathcal{E}^h(f, z, a)^2 + d^h \epsilon_0.$$

Proof of Lemma F.7. By the definitions (2), we know that under the conditions of the lemma, for any $z \in \mathcal{Z}$ and $f, g \in \mathcal{F}^h_z$, we have

$$\max_{a \in \mathcal{A}} (f(z, a) - g(z, a))^2 \leq d^h \left(\nu^h_0(f - g) + \mathbb{E}_{a \sim \pi^h_z} (f(z, a) - g(z, a))^2 \right).$$

In particular, choosing $g(z, a) = \mathbb{E}[r + f^{h+1}(z')|z, a]$ and take supreme over the space of $f - g \in \mathcal{D}\mathcal{F}^h_z$ completes the proof. □

Definition F.8. For any $f \in \mathcal{F}$, we define the set $\mathcal{F}(\epsilon, f) = \left\{ g \in \mathcal{F} : \sup_{z,a,h} |\mathcal{E}^h(g, f; z, a)| \leq \epsilon \right\}$ of functions that have a small Bellman error with f for all z, a. Further assume that $\mathcal{F}$ has an L_∞ cover $f_1, \ldots, f_N \in \mathcal{F}$ for $N = N(\epsilon)$, so that $\forall f \in \mathcal{F}$, $\min_i \sup_{z,a} |f(z, a) - f_i(z, a)| \leq \epsilon$. Then we define $\kappa(\epsilon) = \sup_{f \in \mathcal{F}} - \ln p_0(\mathcal{F}(\epsilon, f))$ and $\kappa'(\epsilon) = \ln N(\epsilon)$.

Proposition F.9 (Theorem 9 of Agarwal & Zhang (2022))**.** *Under Assumptions 5.2, 5.5 and 5.6, suppose further that for any state* $z \in \mathcal{Z}^h$ *and all function class* $\mathcal{F}^h_z$*, we have* $\mathcal{V}^*(\mathcal{F}^h_z; \epsilon_0) \leq d(\epsilon_0)$ *for* $\epsilon_0 > 0$*. Suppose we run TS*^{2}*-ND (Alg. 6) with parameters* $\gamma = 0.1$ *and* $\eta \leq c/(\kappa(\epsilon) + \kappa'(\epsilon) + \log T)$ *for a universal constant c. Then for any* $\epsilon > 0$*, and* $\lambda \leq \eta$ *we have*

$$\mathbb{E}\left[\sum_{t=1}^{T} V^1_\star(z^1_t) - V^1_{\pi_t}(z^1_t)\right] = \mathcal{O}\left(\frac{1}{\lambda}\left(\epsilon T + \kappa(\epsilon) + \kappa'(\epsilon)\right) + T \cdot \widetilde{\epsilon}(\lambda/\eta)\right),$$

where $\widetilde{\epsilon}(\lambda/\eta) := \frac{\lambda \mathsf{br} H d(\epsilon_0)}{\eta} + \frac{\epsilon_0 \eta}{\lambda}$ *for all* $\epsilon, \epsilon_0 > 0$*.*

Proof of Proposition F.9. We introduce some notations from Agarwal & Zhang (2022). We define the Bellman residual of $f \in \mathcal{F}$ using another $g \in \mathcal{F}$ as:

$$\mathcal{E}^h(g, f, z^h, a^h) = g(z^h, a^h) - \mathcal{T}^h f(z^h, a^h),$$

where $\mathcal{T}^h f(z^h, a^h) := \mathbb{E}[r^h + f(z^{h+1}, \pi_f(z^{h+1})) \mid z^h, a^h]$. At any round t of the algorithm, using the observed tuple $(x^h_t, a^h_t, r^h_t, x^{h+1}_t)$, we define

$$\hat{\Delta}^h_t(g, f) = g(x^h_t, a^h_t) - r^h_t - f(x^{h+1}_t), \tag{8}$$

With these definitions, we consider Algorithm 6. For Algorithm 6, which is a modification of Algorithm 2 in Agarwal & Zhang (2022). Following their analysis, but replacing Lemma 32 of their paper with the more general analog in Lemma F.7 immediately gives this proposition. □

Proof of Theorem 5.7. Apply Proposition F.9 with p_0 being the uniform distribution on $\mathcal{F}$ and $\epsilon = 0$ in Definition F.8 . This gives $\kappa(0) = \kappa'(0) = \log |\mathcal{F}|$. Let the base functional $\nu_0 = 0$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} V^1_\star(z^1_t) - V^1_{\pi_t}(z^1_t)\right] = \mathcal{O}\left(\frac{\log |\mathcal{F}|}{\lambda} + T \cdot \frac{\lambda \mathsf{br} H d(\epsilon_0)}{\eta}\right)$$

Choose $\eta = c/(\kappa(\epsilon) + \kappa'(\epsilon) + \log T) = 1/\log(|\mathcal{F}|T)$ and balancing λ, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} V^1_\star(z^1_t) - V^1_{\pi_t}(z^1_t)\right] = \mathcal{O}\left(\log(|\mathcal{F}|T)\sqrt{\mathsf{br} \cdot dHT}\right).$$

□

Algorithm 6 Two timeScale Thompson Sampling with Non-linear Design (TS^2-ND)

```
Require: Function class $\mathcal{F}$, prior $p_0 \in \mathcal{P}(\mathcal{F})$, learning rates $\eta, \gamma$ and optimism coefficient $\lambda$.
Set $S_0 = \emptyset$.
for $t = 1, \dots, T$ do
    Observe $z_t^1 \sim \mathcal{D}$ and draw $h_t \sim \{1, \dots, H\}$ uniformly at random.
    Define $q_t(g) = p(g|f, S_{t-1}) \propto p_0(g) \exp(-\gamma \sum_{s=1}^{t-1} \hat{\Delta}_s^{h_s}(g, f)^2)$.    ▷ Inner loop TS update
    Define $L_t^h(f) = \eta \hat{\Delta}_t^h(f, f)^2 + \frac{\eta}{\gamma} \ln \mathbb{E}_{g \sim q_t} \left[\exp(-\gamma \hat{\Delta}_t^h(g, f)^2)\right]$.    ▷ Likelihood function
    Define $p_t(f) = p(f|S_{t-1}) \propto p_0(f) \exp(\sum_{s=1}^{t-1} (\lambda f(z_s^1) - L_s^{h_s}(f)))$ as the posterior.    ▷ Outer loop Optimistic TS update
    Draw $f_t \sim p_t$ from the posterior. Let $\pi_t = \pi_{f_t}$ and execute $a_t^h = \pi_t(z_t^h)$ for $h = 1, \dots, h_t - 1$ to observe $z_t^{h_t}$.
    Let $\rho_t = \rho_{\mathcal{V}}(\mathcal{F}^{h_t}(z_t^{h_t}), \mathcal{A}; 0) \in \Delta(\mathcal{A})$ be the non-linear $G$-optimal design for $\mathcal{F}^{h_t}(z_t^{h_t})$ Draw $a_{h_t}^t \sim \rho_t$ and observe $r_t^{h_t}$ and $z_t^{h_t+1}$.    ▷ $G$-optimal design
    Update $S_t = S_{t-1} \cup \{z_t^h, a_t^h, r_t^h, z_t^{h+1}\}$ for $h = h_t$.
end for
Return $(\pi_1, \dots, \pi_T)$.
```

Proof of Corollary 5.9. By the standard covering result (Vaart & Wellner, 2023, Theorem 2.7,1), we have $\kappa(\epsilon) = \kappa'(\epsilon) \leq \mathcal{O}(\epsilon^{-d/\beta})$. Moreover by Example 2, we have $d(\epsilon_0) = \mathcal{O}(\epsilon_0^{-d/(2\beta+d)})$. Thus applying Proposition F.9, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} V_\star^1(z_t^1) - V_{\pi_t}^1(z_t^1)\right] &= \mathcal{O}\left(\frac{1}{\lambda}\left(\epsilon T + \kappa(\epsilon) + \kappa'(\epsilon)\right) + T\left(\frac{\lambda \mathsf{br} H d(\epsilon_0)}{\eta} + \frac{\epsilon_0 \eta}{\lambda}\right)\right)\\ &= \mathcal{O}\left(\frac{1}{\lambda}\left(\epsilon T + \frac{1}{\epsilon^{d/\beta}}\right) + T\left(\frac{\lambda \mathsf{br} H}{\eta \epsilon_0^{d/(2\beta+d)}} + \frac{\epsilon_0 \eta}{\lambda}\right)\right)\\ &= \mathcal{O}\left((H \mathsf{br})^{1/2} T^{(\beta^2 + 4\beta d + d^2)/((\beta+d)(2\beta+d))}\right)\end{aligned}$$

with the choice of $\eta = cT^{-d/(\beta+d)}$, $\lambda = (H\mathsf{br})^{-(\beta(2d+\beta))/((\beta+d)(2\beta+d))}$, $\epsilon = T^{-\beta/(\beta+d)}$ and $\epsilon_0 = T^{(d-\beta)}/(\beta+d)$. This bound is sublinear in T whenever $\beta > d$. □

F.4. Proofs for Active Learning

Proof of Theorem 5.11. By the definition of non-linear G-optimal design $\rho_{\mathcal{V}} := \rho_{\mathcal{V}}(\mathcal{F}, \mathcal{P}; 0)$, we have for any $f, f' \in \mathcal{F}$ and $x \in \mathcal{U}$,

$$(f(x) - f'(x))^2 \leq d \cdot \mathbb{E}_{x' \sim \rho_{\mathcal{V}}}(f(x') - f'(x'))^2.$$

By Lemma 4.1, then there exists a subset $\mathcal{U}_n = \{z_1, ..., z_n\} \subset \mathcal{P}$ where $n = \mathcal{O}(d \log |\mathcal{F}|)$ such that for all $x \in \mathcal{U}$,

$$(f(x) - f'(x))^2 \leq 4d \cdot \frac{1}{n}\sum_{i=1}^{n}(f(z_i) - f'(z_i))^2.$$

Then by querying the labels for instances of $z_1, ..., z_n$, we can consider any $\widehat{f} \in \mathcal{F}$ in the version space that satisfies $\widehat{f}(z_i) = f^*(z_i)$ for all $i \in [n]$. This would imply that for all $x \in \mathcal{P}$,

$$(\widehat{f}(x) - f^*(x))^2 \leq 4d \cdot \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(z_i) - f^*(z_i))^2 = 0.$$

This implies that the estimator $\widehat{f}$ will recover all the correct labels for instances $x_1, \dots, x_T$, then by the classical result of Vapnik (1982), we have there exists an algorithm that achieves with probability at least $1 - \delta$,

$$L(\widehat{f}) \leq \mathcal{O}\left(\frac{\mathrm{VCdim}(\mathcal{F}) \log T + \log(1/\delta)}{T}\right).$$

□

F.5. Proofs for Model-based RL

To prove Theorem B.3, we prove the relationship between the offset PAC DEC with the regret F-condition number, which implies that Explorative E2D achieves the desired PAC-regret bound. For this we first introduce the offset PAC DEC for general model class from Chen et al. (2022).

Definition F.10 (Offset PAC DEC for general model class)**.** For any model class $\mathcal{M}$ with the policy class Π, we define

$$\mathsf{p\text{-}dec}_\gamma(\mathcal{M}, \overline{M}) \triangleq \inf_{p,q\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}[\mathrm{Reg}_M(\pi)] - \gamma\mathbb{E}_{\pi\sim q}\left[D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right)\right].$$

Theorem F.11. *For any positive value $\gamma > 0$, the offset PAC DEC is upper bounded by the regret F-condition number with $\eta = 0$, i.e.,*

$$\mathsf{p\text{-}dec}_\gamma(\mathcal{M}, \overline{M}) = \mathcal{O}\left(\frac{H^2 \cdot \mathcal{V}_0^\star(\mathcal{M}; \overline{M}, \nu_0)}{\gamma} + \gamma \cdot \sup_f \nu_0(f)\right).$$

Proof of Theorem F.11. Consider the set of designs $\rho_1, \rho_2, \ldots, \rho_H$ such that for all $h \in [H]$,

$$\sup_{f\in\mathcal{F}_0^h(\overline{M}), M\in\mathcal{M}} \frac{f^2(\pi_M)}{\nu_0(f) + \mathbb{E}_{\pi\sim\rho_h}[f^2(\pi)]} \le \mathcal{V}_0^\star(\mathcal{M}; \overline{M}, \nu_0).$$

Then consider the exploration policy of $q = \frac{1}{H}\sum_{h=1}^{H} \rho_h$ with the exploitation policy of $p = \pi_{\overline{M}}$, we bound the offset PAC DEC by

$$\mathsf{p\text{-}dec}_\gamma(\mathcal{M}, \overline{M}) \le \sup_M \left(J_M(\pi_M) - J_M(\pi_{\overline{M}}) - \gamma \cdot \mathbb{E}_{\pi\sim q}\left[D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right)\right]\right).$$

For any $M \in \mathcal{M}$, we have

$$\begin{aligned} J_M(\pi_M) - J_M(\pi_{\overline{M}}) &= (J_M(\pi_M) - J_{\overline{M}}(\pi_M)) + (J_{\overline{M}}(\pi_M) - J_{\overline{M}}(\pi_{\overline{M}})) + (J_{\overline{M}}(\pi_{\overline{M}}) - J_{\overline{M}}(\pi_{\overline{M}})) \\ &\le 2\sup_{M'} |J_M(\pi_{M'}) - J_{\overline{M}}(\pi_{M'})|. \end{aligned}$$

Then by simulation lemma and AM-GM, we have

$$\begin{aligned} |J_M(\pi_{M'}) - J_{\overline{M}}(\pi_{M'})| &= \left|\sum_{h=1}^{H} \mathbb{E}^{\overline{M},\pi_{M'}}\left[Q_M^h(z^h, a^h) - r^h - V_M^h(z^{h+1})\right]\right| \\ &\le \sum_{h=1}^{H} \left|\mathcal{E}^h(\overline{M}, M, \pi_{M'})\right| \\ &\le \sum_{h=1}^{H}\left(\frac{1}{\xi H} + \xi H\left(\mathcal{E}^h(\overline{M}, M, \pi_{M'})\right)^2\right). \end{aligned}$$

Now by the definition of the design, we have

$$\left(\mathcal{E}^h(\overline{M}, M, \pi_M)\right)^2 \le \mathcal{V}_0^\star(\mathcal{M}; \overline{M}, \nu_0) \cdot \left(\nu_0(\mathcal{E}^h(\overline{M}, M, \cdot)) + \mathbb{E}_{\pi\sim\rho_h}\left[\left(\mathcal{E}^h(\overline{M}, M, \pi)\right)^2\right]\right).$$

We note further that $\mathbb{E}^{M,\pi}\left[Q_M^h(z^h, a^h) - r^h - V_M^h(z^{h+1})\right] = 0$, thus

$$\begin{aligned} \left(\mathbb{E}^{\overline{M},\pi}\left[Q_M^h(z^h, a^h) - r^h - V_M^h(z^{h+1})\right]\right)^2 &= \left((\mathbb{E}^{\overline{M},\pi} - \mathbb{E}^{M,\pi})\left[Q_M^h(z^h, a^h) - r^h - V_M^h(z^{h+1})\right]\right)^2 \\ &\le \left(D_{\mathrm{TV}}\left(M(\pi) \,\|\, \overline{M}(\pi)\right)\right)^2 \\ &\le D_{\mathrm{H}}^2\left(M(\pi), \overline{M}(\pi)\right). \end{aligned}$$

Combining above, we have

$$\begin{aligned}
J_M(\pi_M) - J_M(\pi_{\overline{M}}) &\le 2\sup_{M'} |J_M(\pi_{M'}) - J_{\overline{M}}(\pi_{M'})| \\
&\le \frac{2}{\xi} + 2\xi H \sum_{h=1}^{H} \left(\mathcal{E}^h(\overline{M}, M, \pi_M)\right)^2 \\
&\le \frac{2}{\xi} + 2\xi H \cdot \mathcal{V}_0^\star(\mathcal{M}; \overline{M}, \nu_0) \cdot \sum_{h=1}^{H} \left(\nu_0(\mathcal{E}^h(\overline{M}, M, \cdot)) + \mathbb{E}_{\pi\sim\rho_h}\left[\left(\mathcal{E}^h(\overline{M}, M, \pi)\right)^2\right]\right) \\
&\le \frac{2}{\xi} + 2\xi H \cdot \mathcal{V}_0^\star(\mathcal{M}; \overline{M}, \nu_0) \cdot \sum_{h=1}^{H} \left(\nu_0(\mathcal{E}^h(\overline{M}, M, \cdot)) + \mathbb{E}_{M'\sim\rho_h}\left[D^2_{\mathrm{H}}\left(M(\pi_{M'}), \overline{M}(\pi_{M'})\right)\right]\right)
\end{aligned}$$

Choosing $\xi = \frac{\gamma}{2H^2 \cdot \mathcal{V}_0^\star(\mathcal{M}; \overline{M}, \nu_0)}$, we have for any $M \in \mathcal{M}$,

$$J_M(\pi_M) - J_M(\pi_{\overline{M}}) - \gamma \cdot \mathbb{E}_{\pi\sim q}\left[D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right)\right] = \mathcal{O}\left(\frac{H^2 \cdot \mathcal{V}_0^\star(\mathcal{M}; \overline{M}, \nu_0)}{\gamma} + \gamma \cdot \sup_f \nu_0(f)\right).$$

□

Proof of Theorem B.3. This theorem is a corollary from Theorem F.11 combined with Chen et al. (2022, Theorem 10). □

To prove Theorem B.4, we further consider the relationship of regret F-condition number with offset DEC introduced by (Foster et al., 2021).

Definition F.12 (Offset DEC for general model class)**.** For any model class $\mathcal{M}$ with the policy class Π, we define

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \triangleq \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}[\mathrm{Reg}_M(\pi)] - \gamma \mathbb{E}_{\pi\sim p}\left[D^2_{\mathrm{H}}\left(M(\pi), \overline{M}(\pi)\right)\right].$$

Theorem F.13. *For any positive value $\gamma > 0$, the offset DEC is upper bounded by the regret F-condition number, i.e.,*

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) = \mathcal{O}\left(\frac{H^2 \mathcal{V}_\eta^\star(\mathcal{M}; \overline{M}, \nu_0)}{\gamma} + \gamma \cdot \sup_f \nu_0(f)\right),$$

whenever $\eta = \Theta\left(\frac{\gamma}{H^2 \mathcal{V}_\eta^\star(\mathcal{M}; \overline{M}, \nu_0)}\right) < \frac{\gamma}{8H^2 \mathcal{V}_\eta^\star(\mathcal{M}; \overline{M}, \nu_0)}$.

This theorem shows that our F-design is useful as a black-box replacement whenever linear G-optimal design is used.

Proof of Theorem F.13. Consider the set of designs $\rho_1, \rho_2, \ldots, \rho_H$ such that for all $h \in [H]$,

$$\sup_{f\in\mathcal{F}_\eta^h(\overline{M}), M\in\mathcal{M}} \frac{f^2(\pi_M)}{\nu_0(f) + \mathbb{E}_{\pi\sim\rho_h}[f^2(\pi)]} \le \mathcal{V}_\eta^\star(\mathcal{M}; \overline{M}, \nu_0).$$

Then consider the policy of $q = \frac{1}{2H} \sum_{h=1}^{H} \rho_h + \frac{1}{2}\mathbb{1}(\cdot = \pi_{\overline{M}})$. And the weighted inverse gap weighting policy $p \in \Delta(\Pi)$ of

$$p(\pi) = \frac{q(\pi)}{\lambda + \eta \cdot \mathrm{Reg}_{\overline{M}}(\pi)},$$

for the $\lambda \in [1/2, 1]$ such that $\sum_\pi p(\pi) = 1$. We first show that such a λ always exists. Consider the function $K(\lambda) = \sum_\pi \frac{q(\pi)}{\lambda + \eta \cdot \mathrm{Reg}_{\overline{M}}(\pi)}$. It is clear that $K(1/2) \ge \frac{q(\pi_{\overline{M}})}{1/2} = 1$. On the other hand $K(1) \le \sum_\pi q(\pi) \le 1$. Furthermore, since $K(\lambda)$ is monotonically decreasing, there exists a $\lambda \in [1/2, 1]$ such that $K(\lambda) = 1$.

To bound the offset DEC, we have

$$\mathsf{dec}_\gamma(\mathcal{M},\overline{M}) \le \sup_M \mathbb{E}_{\pi\sim p}\big[J_M(\pi_M) - J_M(\pi) - \gamma\cdot D^2_{\mathrm{H}}\left(M(\pi),\overline{M}(\pi)\right)\big].$$

For any $M\in\mathcal{M}$, we have

$$\begin{aligned}\mathbb{E}_{\pi\sim p}[J_M(\pi_M) - J_M(\pi)] = {} & (J_M(\pi_M) - J_{\overline{M}}(\pi_M)) + (J_{\overline{M}}(\pi_M) - J_{\overline{M}}(\pi_{\overline{M}})) + \mathbb{E}_{\pi\sim p}[J_{\overline{M}}(\pi_{\overline{M}}) - J_{\overline{M}}(\pi)]\\ & + \mathbb{E}_{\pi\sim p}[J_{\overline{M}}(\pi) - J_M(\pi)].\end{aligned}$$

We bound the four parts separately. We first deal with the last two terms which are easy. We note

$$\mathbb{E}_{\pi\sim p}[J_{\overline{M}}(\pi_{\overline{M}}) - J_{\overline{M}}(\pi)] = \sum_\pi \frac{q(\pi)\mathrm{Reg}_{\overline{M}}(\pi)}{\lambda+\eta\cdot\mathrm{Reg}_{\overline{M}}(\pi)} \le \frac{1}{\eta}$$

and

$$\mathbb{E}_{\pi\sim p}[J_{\overline{M}}(\pi) - J_M(\pi)] \le \frac{\gamma}{2}\cdot\mathbb{E}_{\pi\sim p}\big[D^2_{\mathrm{H}}\left(M(\pi),\overline{M}(\pi)\right)\big] + \frac{1}{2\gamma}.$$

For the first term, by simulation lemma and AM-GM, we have

$$\begin{aligned}J_M(\pi_M) - J_{\overline{M}}(\pi_M) &= \sum_{h=1}^H \mathbb{E}^{\overline{M},\pi_M}\big[Q^h_M(z^h,a^h) - r^h - V^h_M(z^{h+1})\big]\\ &= \sum_{h=1}^H\left(\sqrt{1+\eta\cdot\mathrm{Reg}_{\overline{M}}(\pi_M)}\cdot\frac{\mathcal{E}^h(\overline{M},M,\pi_M)}{\sqrt{1+\eta\cdot\mathrm{Reg}_{\overline{M}}(\pi_M)}}\right)\\ &\le \sum_{h=1}^H\left(\frac{1+\eta\cdot\mathrm{Reg}_{\overline{M}}(\pi_M)}{\eta H} + \eta H\cdot\frac{\left(\mathcal{E}^h(\overline{M},M,\pi_M)\right)^2}{1+\eta\cdot\mathrm{Reg}_{\overline{M}}(\pi_M)}\right).\end{aligned}$$

Now by the definition of the design, we have

$$\frac{\left(\mathcal{E}^h(\overline{M},M,\pi_M)\right)^2}{1+\eta\cdot\mathrm{Reg}_{\overline{M}}(\pi_M)} \le \mathcal{V}^\star_\eta(\mathcal{M};\overline{M},\nu_0)\cdot\left(\sup_f \nu_0(f) + \mathbb{E}_{\pi\sim\rho_h}\left[\frac{\left(\mathcal{E}^h(\overline{M},M,\pi)\right)^2}{1+\eta\cdot\mathrm{Reg}_{\overline{M}}(\pi)}\right]\right).$$

We note further that $\mathbb{E}^{M,\pi}\big[Q^h_M(z^h,a^h) - r^h - V^h_M(z^{h+1})\big] = 0$, thus

$$\begin{aligned}\left(\mathbb{E}^{\overline{M},\pi}\big[Q^h_M(z^h,a^h) - r^h - V^h_M(z^{h+1})\big]\right)^2 &\le \left((\mathbb{E}^{\overline{M},\pi} - \mathbb{E}^{M,\pi})\big[Q^h_M(z^h,a^h) - r^h - V^h_M(z^{h+1})\big]\right)^2\\ &\le \left(D_{\mathrm{TV}}\left(M(\pi)\,\|\,\overline{M}(\pi)\right)\right)^2\\ &\le D^2_{\mathrm{H}}\left(M(\pi),\overline{M}(\pi)\right).\end{aligned}$$

Altogether, we have

$$\begin{aligned}&\mathbb{E}_{\pi\sim p}[J_M(\pi_M) - J_M(\pi)]\\ &\quad\le \sum_{h=1}^H\left(\frac{1+\eta\cdot\mathrm{Reg}_{\overline{M}}(\pi_M)}{\eta H} + \eta H\mathcal{V}^\star_\eta(\mathcal{M};\nu_0)\cdot\left(\sup_f\nu_0(f) + \mathbb{E}_{\pi\sim\rho_h}\left[\frac{D^2_{\mathrm{H}}\left(M(\pi),\overline{M}(\pi)\right)}{1+\eta\cdot\mathrm{Reg}_{\overline{M}}(\pi)}\right]\right)\right)\\ &\qquad - \mathrm{Reg}_{\overline{M}}(\pi_M) + \frac{\gamma}{2}\cdot\mathbb{E}_{\pi\sim p}\big[D^2_{\mathrm{H}}\left(M(\pi),\overline{M}(\pi)\right)\big] + \frac{1}{\gamma} + \frac{1}{\eta}\\ &\quad\le \frac{1}{\eta} + \eta H\mathcal{V}^\star_\eta(\mathcal{M};\overline{M},\nu_0)\sum_{h=1}^H\sup_f\nu_0(f) + \left(\eta H^2\mathcal{V}^\star_\eta(\mathcal{M};\nu_0) + \frac{\gamma}{2}\right)\cdot\mathbb{E}_{\pi\sim p}\big[D^2_{\mathrm{H}}\left(M(\pi),\overline{M}(\pi)\right)\big] + \frac{1}{\gamma} + \frac{1}{\eta}.\end{aligned}$$

Then by the assumption of $\eta = \Theta\left(\frac{\gamma}{H^2\mathcal{V}^\star_\eta(\mathcal{M};\overline{M},\nu_0)}\right) < \frac{\gamma}{8H^2\mathcal{V}^\star_\eta(\mathcal{M};\overline{M},\nu_0)}$, we obtain

$$\mathbb{E}_{\pi\sim p}[J_M(\pi_M) - J_M(\pi)] - \gamma\cdot\mathbb{E}_{\pi\sim p}\big[D^2_{\mathrm{H}}\left(M(\pi),\overline{M}(\pi)\right)\big] = \mathcal{O}\left(\frac{H^2\mathcal{V}^\star_\eta(\mathcal{M};\nu_0)}{\gamma} + \gamma\cdot\sup_f\nu_0(f)\right).$$

□

Proof of Theorem B.4. This theorem is a corollary from Theorem F.13 combing with Foster et al. (2021, Theorem 4.1). □