Tight Generalization Bounds for Noiseless Inverse Optimization

[1,2] **Pouria Fatemi** [3,4] **Hoomaan Maskan** [1,2] **Suvrit Sra** [5] **Peyman Mohajerin Esfahani**

[1] Technical University of Munich, School of CIT, Dept. of Mathematics
[2] Munich Center for Machine Learning
[3] Uppsala University, Dept. of Information Technology
[4] Uppsala University, Science for Life Laboratory
[5] University of Toronto

Abstract

Inverse optimization (IO) seeks to infer the parameters of a decision-maker's objective from observed context–action data. We study *noiseless IO*, where demonstrations are generated by a ground-truth objective. We provide a high-probability $\mathcal{O}(d/T)$ generalization bound for the induced action set, where d is the number of unknown parameters and T is the size of the training dataset. We strengthen this guarantee under additional conditions that ensure uniqueness of the chosen action, bringing our IO guarantees in line with best-arm identification results in the bandit literature. We further show that the $\mathcal{O}(d/T)$ rate is tight over all consistent estimators considered here, and extend the result to both instantaneous and cumulative regret. Notably, the resulting regret lower bound matches the corresponding upper bounds in the adversarial setting, indicating that the stochastic IO setting is effectively adversarial for the class of estimators studied here. Finally, we propose a parameter-free algorithm with lower per-iteration complexity than generic solvers. Experiments validate the predicted rates and illustrate the tightness of our bounds.

1 Introduction

We consider the contextual decision-making problem

$$a_\theta(s) \in \mathcal{A}_\theta(s) := \operatorname*{argmax}_{a \in \mathbb{A}(s)} F_\theta(s, a), \tag{1}$$

where $s \in \mathcal{S}$ is a random context (state) drawn from a distribution $\mathbb{P}_\mathcal{S}$ over $\mathcal{S}$, $\mathbb{A}(s)$ is the set of feasible actions given s, and $F_\theta(s, a) = \langle \theta, \psi(s, a) \rangle$ is a linear score parameterized by $\theta \in \mathbb{R}^d$. We define $\mathcal{A}_\theta(s)$ as the set of optimal actions for a context s under parameter θ and assume that it is non-empty. We also define $\psi : \mathcal{S} \times \mathbb{A} \to \mathbb{R}^d$ as a generic feature map where $\mathbb{A} := \cup_{s \in \mathcal{S}} \mathbb{A}(s)$.

Suppose there is an expert who observes a context s and then selects an action $a_{\theta^\star}(s)$ according to (1) using a parameter $\theta^\star$ that is unknown to a learner. But the learner gets to observe T demonstrations $D_T := \{(s_t, a_t^\star)\}_{t=1}^T$, $a_t^\star := a_{\theta^\star}(s_t)$, with $(s_t)_{t=1}^T$ drawn i.i.d. from $\mathbb{P}_\mathcal{S}$, and wishes to use them to mimic the expert's actions.[1]

This learning paradigm is commonly referred to as *inverse optimization* (IO) (Ahuja & Orlin, 2001). The goal of IO is to infer a parameter θ such that the corresponding actions generalize to unseen contexts drawn from $\mathbb{P}_\mathcal{S}$. Using the demonstrations, the learner outputs an estimate $\hat{\theta}$, inducing a

[1] Throughout, we assume that the expert's optimal action is unique, i.e., $\mathcal{A}_{\theta^\star}(s) = \{a_{\theta^\star}(s)\}$ for all s. If $\mathcal{A}_{\theta^\star}(s)$ is not a singleton, all of our generalization bounds can be written in the form of $\mathbb{P}_\mathcal{S}\big(\mathcal{A}_{\theta^\star}(s) \cap \mathcal{A}_{\hat{\theta}}(s) = \emptyset\big)$.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

set-valued greedy map $\mathcal{A}_{\hat{\theta}}(s)$. Natural measures to judge a learner's generalization performance are

$$\text{Set-level mismatch:} \qquad \mathbb{P}_{\mathcal{S}}\big(a_{\theta^\star}(s) \notin \mathcal{A}_{\hat{\theta}}(s)\big), \tag{2a}$$

$$\text{Action-level mismatch:} \qquad \mathbb{P}_{\mathcal{S}}(a_{\theta^\star}(s) \neq a_{\hat{\theta}}(s)), \tag{2b}$$

where (2a) measures how often the learned model excludes the expert's demonstrated action on a fresh state $s \sim \mathbb{P}_{\mathcal{S}}$, while (2b) is a stronger notion, which is also known as *best-arm identification* in the bandit literature (Lattimore & Szepesvári, 2020).

IO has been extensively studied from algorithmic, modeling, and optimization perspectives (Chan et al., 2025), and has witnessed a recent surge of interest, e.g., in learning theory (Mohajerin Esfahani et al., 2018; Aswani et al., 2018; Ren et al., 2025), in reinforcement learning (Zattoni Scroccaro et al., 2025b), for practical problems in control (Akhtar et al., 2021), robotics (Dimanidis et al., 2025), routing (Zattoni Scroccaro et al., 2025b), and finance (Li, 2021).

Despite this interest, IO's statistical foundations are far less developed. Since IO can be viewed as a supervised learning problem, it is natural to seek generalization guarantees (2) that characterize its out-of-sample performance. Through the lens of (2a), our problem can be seen as a binary classification problem, for which standard learning-theoretic tools could provide generalization guarantees (see Appendix A.3). Similarly, the action-level counterpart (2b) can be viewed as multiclass or structured prediction, whose PAC complexity is studied via the Natarajan dimension in finite-label settings and characterized by the DS dimension in general multiclass learning (Natarajan, 1989; Ben-David et al., 1995; Shalev-Shwartz & Ben-David, 2014; Brukhim et al., 2022; Pabbaraju, 2026). Nonetheless, this viewpoint is limited, as these complexity notions can be hard to quantify and may yield conservative bounds that are often not tight. For instance, existing upper and lower bounds either exhibit gaps in their dependence on the key variables (d, T) or scale with the number of actions, making them uninformative when the action space is infinite (Shalev-Shwartz & Ben-David, 2014, Theorem 29.3).

To the best of our knowledge, this is the first work that addresses *tight* generalization guarantees in IO. With this background, let us summarize our key contributions.

(i) **Generalization upper bounds.** We develop a *scenario program* view of noiseless IO, where each demonstration induces a random convex feasibility constraint on the parameter θ (Campi & Garatti, 2008). This yields a high-probability $\mathcal{O}(d/T)$ bound on set-level mismatch (Proposition 2.1). We further upgrade this guarantee to action-level mismatch using an incenter estimator that promotes uniqueness of the greedy action (Theorem 2.2). The action-level bound implies instantaneous regret guarantees and, in a stochastic online protocol, logarithmic cumulative regret (Proposition 2.4).

(ii) **Generalization lower bounds.** We show that the $\mathcal{O}(d/T)$ action-level mismatch rate is minimax optimal (Proposition 3.1) and construct a single IO instance whose mismatch performance (2) attains the upper bound regardless of the choice of consistent estimator in the considered class (Theorem 3.2, Remark 3.4). We further extend these results to instantaneous and expected cumulative regret, as the same IO instance applies across datasets of varying sample sizes (Proposition 3.5); Table 1 summarizes these results within the literature. Notably, the resulting $\Omega(d \log(T/d))$ lower bound on expected cumulative regret matches the cited adversarial upper bounds, even though in our case contexts are sampled i.i.d. For the considered estimator class, our construction exhibits this behavior across a broad class of context distributions. The corresponding lower bound holds with high probability at each fixed round, showing that the per-round difficulty is not driven solely by rare training histories.

(iii) **A parameter-free algorithm.** We propose a parameter-free algorithm that empirically converges to generalizable solutions after only T iterations. In particular, the proposed method has lower per-iteration computational cost than off-the-shelf solvers (Remark 4.1).

Our numerical experiments empirically validate our theoretical predictions and illustrate the tightness of the proposed generalization bounds.

1.1 Related work

Generalization in IO. Classical IO infers objectives and/or constraints that rationalize observed decisions in-sample; guarantees concern exact/approximate optimality on the training instances. A recent out-of-sample analysis by Besbes et al. (2025) proves a geometrical upperbound on the minimum possible instantaneous regret a policy could achieve in the adversarial offline setting. Their

Table 1: Comparison of generalization bounds in IO under adversarial and stochastic settings.

	Protocol	Metric	Upper bound	Lower bound
Adversarial	Offline	Instantaneous regret (9)	Geometric characterization; no T-dependent rate (Besbes et al., 2025)	Matching geometric lower bound (Besbes et al., 2025)
	Online	Cumulative regret (9)	$\mathcal{O}(d^4 \log T)$ (Besbes et al., 2025); $\mathcal{O}(d \log T)$ (Gollapudi et al., 2021; Sakaue et al., 2025)	$\Omega(d)$ (Sakaue et al., 2025)
Stochastic	Offline	Set/Action-level mismatch (2)	**This work**: $\mathcal{O}(d/T)$ (Theorem 2.2)	**This work**: $\Omega(d/T)$ (Proposition 3.1‡ and Theorem 3.2†‡)
	Online	Cumulative regret (9)	**This work**: $\mathcal{O}(d \log(T/d))$ (Proposition 2.4)	**This work**: $\Omega(d \log(T/d))$ (Proposition 3.5†)

† The lower bounds are restricted to the algorithm class specified in the corresponding result.
‡ The lower bounds differ in their order of quantification (min–max versus max–min); see Remark 3.3.

result depends on the uncertainty angle of the information set but does not provide an explicit rate in the sample size. Our generalization result relies on the set-level and action-level mismatch for a stochastic environment.

Supervised learning. From a supervised learning perspective, the IO problem objective in (1) can be considered as a hypothesis class for learning the mapping between the context s and the actions a. This essentially requires minimizing a loss function to find the model parameter θ. In IO, the choice of the loss function plays a crucial role. Examples of different loss functions include the KKT loss (Keshavarz et al., 2011), first order loss (Bertsimas et al., 2015), predictability loss (Aswani et al., 2018), suboptimality loss (Mohajerin Esfahani et al., 2018), predictive loss, and constrained suboptimality loss (Ren et al., 2025). It is important to note that directly training the objective $F_{\theta^\star}$ is not possible, since we do not have direct measurements of the ground-truth objective values in IO.

Smart predict, then optimize. In a related study to IO, Elmachtoub & Grigas (2022) propose the *Smart predict, then optimize* (SPO) framework, where access to the optimal cost value of $F_{\theta^\star}$ is assumed and the SPO loss function is proposed. Given feature-cost pairs $\{s_t, c_t\}_{t=1}^T$, a predictor is trained such that it minimizes a loss function that penalizes predictions leading to bad decisions. SPO assumes access to costs/predictions $c_t := \langle \theta, s_t \rangle$ and actions/decisions $a_t^\star$, while we only observe states-action pairs $\{s_t, a_t^\star\}$. In a recent work, El Balghiti et al. (2023) propose a generalization result of $\mathcal{O}(1/\sqrt{T})$ when assuming a joint distribution over (s_t, c_t). With fewer distributional assumptions but access to more data, they achieve a worse generalization result compared to ours.

Online learning. Online learning studies the problem of updating a model sequentially as data arrives, with the aim of achieving low cumulative loss or regret (Zinkevich, 2003; Hazan et al., 2016). When performed in an online manner, IO can be viewed as a model-learning task, making it conceptually closer to online learning. The major difference between these frameworks is that the available data in online learning is the loss function value, while in IO, it is the optimal decision. Recently, the online version of IO has been studied. Besbes et al. (2025) show that in the adversarial case, a naive application of a *circumcenter policy* fails. They also propose an algorithm achieving a regret bound of $\mathcal{O}(d^4 \log T)$. For a similar setting, Gollapudi et al. (2021); Sakaue et al. (2025) improve this rate to $\mathcal{O}(d \log T)$ while Sakaue et al. (2025) achieve this rate with improved per-iteration complexity using an online Newton step. In comparison, we propose a high-probability regret bound of $\mathcal{O}(d \log T)$ when nature is stochastic using a simple policy. This result is interesting since it shows that in the stochastic environment, simple policies can work, while this is not necessarily the case in the adversarial setting.

1.2 Problem terminology

For each pair $(s_t, a_t^\star)$ and every $a \in \mathbb{A}(s_t)$, optimality of the expert under $\theta^\star$ implies $\langle \theta^\star, \psi(s_t, a) \rangle \leq \langle \theta^\star, \psi(s_t, a_t^\star) \rangle$. For a generic parameter $\theta \in \mathbb{R}^d$, we say that θ is *consistent* with a demonstration $(s_t, a_t^\star)$ if the demonstrated expert action is also greedy under θ, i.e.,

$$\left\langle \theta,\ \psi(s_t, a) - \psi(s_t, a_t^\star) \right\rangle \leq 0,\ \forall a \in \mathbb{A}(s_t),\ t \in [T]. \tag{3}$$

To express condition (3) compactly, define the suboptimality gap (Mohajerin Esfahani et al., 2018):

$$\ell_\theta^{\mathrm{sub}}(s,a) := \max_{a' \in \mathbb{A}(s)} \langle \theta, \psi(s,a') - \psi(s,a) \rangle. \tag{4}$$

By definition, $\ell_\theta^{\mathrm{sub}}(s,a) \geq 0$ for all $s \in S$ and $a \in \mathbb{A}(s)$. Meanwhile, for satisfying condition (3) we need θ such that $\ell_\theta^{\mathrm{sub}}(s_t, a_t^\star) \leq 0$. Combining these two is equivalent to $\ell_\theta^{\mathrm{sub}}(s_t, a_t^\star) = 0$. Therefore, we define the consistency set:

$$\mathcal{C}_T := \Big\{ \theta \in \mathbb{R}^d : \ \ell_\theta^{\mathrm{sub}}(s_t, a_t^\star) = 0, \ \forall t \in [T] \Big\}. \tag{5}$$

By construction, $\theta \in \mathcal{C}_T$ if and only if $a_t^\star \in \mathcal{A}_\theta(s_t)$ for all $t \in [T]$, i.e., the demonstrated actions remain greedy under θ on the training states. When $\mathcal{A}_\theta(s_t)$ contains multiple actions, the constraint only enforces that the demonstrated action $a_t^\star$ belongs to that set. If ties occur, $\mathcal{A}_\theta(s)$ may contain multiple actions.

Remark 1.1 (Parameter set regularity)**.** For any $\theta \in \mathcal{C}_T$, the scaled parameter $\alpha\theta$ is also in $\mathcal{C}_T$ for every $\alpha > 0$. As a result, the consistency set $\mathcal{C}_T$ forms a cone. This is expected, since the greedy action is invariant under positive rescaling of θ. To avoid this scale ambiguity, we define the closed convex parameter domain $\Theta \subset \mathbb{R}^d$ as a normalized set containing at most one representative from each positive ray: if $\theta \in \Theta$, then $\alpha\theta \notin \Theta$ for all $\alpha \neq 1$. In particular, this normalization implies $0 \notin \Theta$. Since positive rescaling does not change the induced actions, we select the normalized representative of $\theta^\star$ and assume, without loss of generality, that $\theta^\star \in \Theta$.

The set $\mathcal{C}_T$ provides an exact characterization of all parameters that rationalize the observed data. However, this constraint-based description does not by itself yield out-of-sample guarantees. Each demonstration induces a family of inequalities, and generalization is governed by the probability that these random feasibility constraints are violated on unseen states. In the next section, we build on this viewpoint and derive generalization bounds for IO.

2 Generalization Guarantees

In this section, we propose high-probability generalization upperbounds for set-level mismatch. Later, we strengthen this guarantee to action-level mismatch as the core result of this work. Through our action-level analysis, we propose regret bounds for a stochastic online variant of our IO framework.

2.1 Set-level mismatch upperbound

Since any IO estimator should belong to $\mathcal{C}_T$, a natural way to select such an estimator is

$$\hat{\theta}_T^{\mathrm{sub}} \ := \ \arg\min_{\theta \in \Theta} \ J(\theta) \quad \text{s.t.} \quad \ell_\theta^{\mathrm{sub}}(s_t, a_t^\star) \leq 0, \quad \forall t \in [T], \tag{6}$$

for a convex function $J : \mathbb{R}^d \to \mathbb{R}$ such that the program (6) admits a *unique* optimizer $\hat{\theta}_T^{\mathrm{sub}}$ almost surely. Existence and uniqueness are guaranteed, for example, when J is coercive and strictly convex.

Problem (6) reveals a direct connection between IO and the well-studied paradigm of *scenario optimization*. Each demonstration induces an i.i.d. feasibility constraint in parameter space. This connection is significant because it recasts inverse optimization generalization as a constraint-violation problem, focusing on the probability that the learned parameter violates the population feasibility constraint on a fresh state. To the best of our knowledge, this scenario-program viewpoint for IO estimation is novel and equips us with new tools.

We use the result from (Campi & Garatti, 2008) to prove the generalizability of $\hat{\theta}_T^{\mathrm{sub}}$. In particular, the following proposition (see Appendix B.3 for the proof) states that if T is large enough, the solution $\hat{\theta}_T^{\mathrm{sub}}$ has a low set-level mismatch with high probability.

Proposition 2.1 (Set-level mismatch)**.** *Fix $\beta \in (0,1)$ and choose $\varepsilon \in (0,1]$ such that*

$$T \geq N(\varepsilon, \beta) := \min\Big\{ n \in \mathbb{N} \ \Big| \ \sum_{i=0}^{d-1} \binom{n}{i} \varepsilon^i (1-\varepsilon)^{n-i} \leq \beta \Big\}.$$

Let $\hat{\theta}_T^{\mathrm{sub}}$ be the unique optimizer of (6). Then, with probability at least $1-\beta$ over the draw of D_T,

$$\mathbb{P}_{\mathcal{S}}\big(a_{\theta^\star}(s) \notin \mathcal{A}_{\hat{\theta}_T^{\mathrm{sub}}}(s)\big) \ \leq \ \varepsilon. \tag{7}$$

In particular, if $T \geq 2\big(d + \log(1/\beta)\big)$, then, with probability at least $1-\beta$ over the draw of D_T,

$$\mathbb{P}_{\mathcal{S}}\big(a_{\theta^\star}(s) \notin \mathcal{A}_{\hat{\theta}_T^{\mathrm{sub}}}(s)\big) \ \leq \frac{2}{T}\big(d + \log(1/\beta)\big). \tag{8}$$

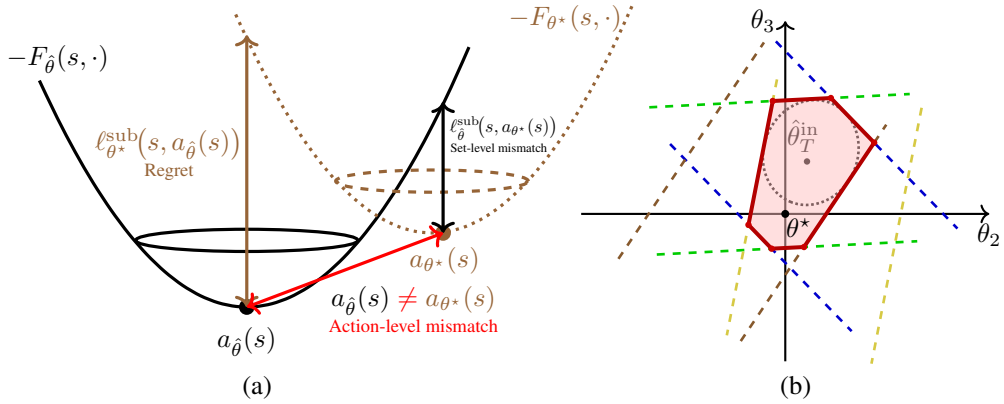


Figure 1: (a) Relation between the set-level mismatch, the action-level mismatch, and the true suboptimality loss (regret). Since additive shifts do not affect the optimal actions, we shift the objective curves vertically so that their optimal values coincide. (b) Geometric intuition for tightness in the two-dimensional case $\theta = (1, \theta_2, \theta_3)$ for Theorem 3.2. Each demonstration induces an asymmetric strip constraint. Intersecting these constraints yields the consistency polytope (shaded) that contains $\theta^\star$. The incenter estimator $\hat{\theta}^{\mathrm{in}}_T$ has the maximum distance from the facets of $\mathcal{C}_T$. The dotted ellipsoid represents the intersection of the circular cone around $\hat{\theta}^{\mathrm{in}}_T$ with Θ.

The guarantee in Proposition 2.1 is set-valued, i.e., it only controls whether the expert action is contained in the argmax set $\mathcal{A}_{\hat{\theta}^{\mathrm{sub}}_T}(s)$. When these argmax sets are large, the bound can be uninformative about the actual action taken by a greedy policy (which depends on tie-breaking). We illustrate this issue with the following example.

Example 1 (Issues with set-level guarantees)**.** Let $\mathcal{S} = \{0, 1\}$ with $\mathbb{P}_{\mathcal{S}}(0) = \mathbb{P}_{\mathcal{S}}(1) = \frac{1}{2}$, and let $\mathbb{A} = [-1, 1]^2$. Define $\psi\big(s, (a_1, a_2)\big) := (a_1, a_2, s\, a_2) \in \mathbb{R}^3$. For each s, the map $a \mapsto \psi(s, a)$ is injective since the first two coordinates recover (a_1, a_2). Let $\theta^\star = (1, -1, 2)$ and define the expert selection $a_{\theta^\star}(s) \in \mathcal{A}_{\theta^\star}(s)$. Since $\langle \theta^\star, \psi(s, a)\rangle = a_1 + (-1 + 2s)a_2$, the expert argmax sets are singletons and the (unique) maximizers are $a_{\theta^\star}(0) = (1, -1)$ and $a_{\theta^\star}(1) = (1, 1)$. Now, take $\hat{\theta}^{\mathrm{sub}}_T = (1, 0, 0)$. This choice of $\hat{\theta}^{\mathrm{sub}}_T$ is in $\mathcal{C}_T$ since for any given dataset $D_T := \{(s_t, a^\star_t)\}_{t=1}^T$, any t, and any $a = (a_1, a_2) \in [-1, 1]^2$, the expert always chooses $a^\star_t = (1, \pm 1)$. This means that $(a^\star_t)_1 = 1$, and therefore

$$\langle \hat{\theta}^{\mathrm{sub}}_T, \psi(s_t, a) - \psi(s_t, a^\star_t)\rangle = a_1 - 1 \leq 0,$$

where the last inequality is due to $a_1 \leq 1$. Therefore, $\hat{\theta}^{\mathrm{sub}}_T = (1, 0, 0) \in \mathcal{C}_T$. For this $\hat{\theta}^{\mathrm{sub}}_T$, we have $\langle \hat{\theta}^{\mathrm{sub}}_T, \psi(s, a)\rangle = a_1$ and hence

$$\mathcal{A}_{\hat{\theta}^{\mathrm{sub}}_T}(s) = \{(1, a_2) : a_2 \in [-1, 1]\},$$

for all $s \in \mathcal{S}$. This set is a full line segment for every state. In particular, $a_{\theta^\star}(s) \in A_{\hat{\theta}^{\mathrm{sub}}_T}(s)$ for both $s = 0, 1$, and therefore $\mathbb{P}_{\mathcal{S}}\big(a_{\theta^\star}(s) \notin \mathcal{A}_{\hat{\theta}^{\mathrm{sub}}_T}(s)\big) = 0$.

To address the issue explained in Example 1, we need to shift toward action-level mismatch and specify how ties in $\mathcal{A}_\theta(s)$ are resolved at test time. To this end, we seek an estimator of $\theta^\star$ that guarantees recovery of the unique action $a_{\theta^\star}(s)$. Moreover, controlling the action-level mismatch relates the set-level guarantees to regret bounds. We define instantaneous regret r_t and cumulative regret R_T as

$$r_t := \big\langle \theta^\star, \psi(s_t, a^\star_t) - \psi(s_t, a_t)\big\rangle, \quad R_T := \sum_{t=1}^{T} r_t, \tag{9}$$

where $a_t := a_{\hat{\theta}_{t-1}}(s_t)$ is the learner's greedy action. Figure 1 (a) illustrates how set-level mismatch, action-level mismatch, and the notion of regret, relate to each other.

2.2 Action-level mismatch upperbound

As illustrated in Example 1, set-level guarantees can be uninformative when the learned action set $\mathcal{A}_{\hat{\theta}^{\mathrm{sub}}_T}(s)$ is large. It is possible to avoid such large action sets, but this requires care in choosing the

estimator. In Example 1, the estimate $\hat{\theta}$ lies on the boundary of the consistency set $\mathcal{C}_T$. With this estimator, the expert's action remains optimal but ties with alternatives at the observed states. Such ties can persist on unseen states. However, this does not rule out suitable solutions in the interior of $\mathcal{C}_T$. Rather, it motivates selecting an estimator that sits away from the tie hyperplanes.

In this section, we use an *incenter* estimator inspired by Zattoni Scroccaro et al. (2025a), which selects $\hat{\theta}_T^{\text{in}}$ in the interior of $\mathcal{C}_T$. We will show that with high probability, the corresponding action set to the incenter estimator is singleton. Based on this, we improve our generalization bound from set-level to action-level. Consider the incenter loss function:

$$\ell_\theta^{\text{in}}(s, a_{\theta^\star}(s)) := \max_{a \in \mathbb{A}(s)} \Big(\langle \theta, \delta(s,a) \rangle + \|\delta(s,a)\| \Big), \tag{10}$$

for $\delta(s,a) := \psi(s,a) - \psi\big(s, a_{\theta^\star}(s)\big)$. For fixed s, $\theta \mapsto \ell_\theta^{\text{in}}(s, a_{\theta^\star}(s))$ is convex as a pointwise supremum of convex functions. Using $\ell_\theta^{\text{in}}(s_t, a_t^\star)$ instead of the sub-optimality loss in (6) would result in the following unique incenter estimator:

$$\hat{\theta}_T^{\text{in}} := \arg\min_{\theta \in \mathbb{R}^d} J(\theta) \quad \text{s.t.} \quad \ell_\theta^{\text{in}}(s_t, a_t^\star) \le 0, \quad \forall t \in [T]. \tag{11}$$

If $\text{int}(\mathcal{C}_T) \neq \emptyset$, then (11) is feasible over $\mathbb{R}^d$ (Zattoni Scroccaro et al., 2025a). Moreover, the constraint $\ell_\theta^{\text{in}}(s_t, a_t^\star) \le 0$ is equivalent to

$$\langle \theta, \delta(s_t, a) \rangle \le -\|\delta(s_t, a)\|, \qquad \forall a \in \mathbb{A}(s_t). \tag{12}$$

Under our standing assumption that the expert's optimal action is unique, every alternative $a \neq a_t^\star$ satisfies $\delta(s_t, a) \neq 0$. Thus, (12) enforces a strictly negative margin separating the demonstrated action from every alternative. In contrast, the consistency constraints only require $\langle \theta, \delta(s_t, a) \rangle \le 0$, permitting boundary solutions that can induce ties. Thus, the incenter loss keeps the estimator away from the tie hyperplanes $\langle \theta, \delta(s_t, a) \rangle = 0$. Next, we establish the generalization guarantee for $\hat{\theta}_T^{\text{in}}$. The proof is given in Appendix B.4.

Theorem 2.2 (Action-level guarantees via incenter)**.** *Assume* $\text{int}(\mathcal{C}_T) \neq \emptyset$ *almost surely. Fix* $\beta \in (0,1)$ *and let* $\varepsilon \in (0,1]$ *satisfy* $T \ge N(\varepsilon, \beta)$*. Then, with probability at least* $1-\beta$ *over the draw of* D_T*,* $\hat{\theta}_T^{\text{in}}$ *satisfies*

$$\mathbb{P}_{\mathcal{S}}\Big(\mathcal{A}_{\hat{\theta}_T^{\text{in}}}(s) = \{a_{\theta^\star}(s)\} \Big) \ge 1 - \varepsilon. \tag{13}$$

Theorem 2.2 is our main generalization result. It strengthens the set-level guarantee by ensuring that the learned objective has a unique optimal action matching the expert, with the same sample complexity $N(\varepsilon, \beta)$. The only additional structural requirement is $\text{int}(\mathcal{C}_T) \neq \emptyset$ almost surely. Remark 2.3 proposes a sufficient condition for this requirement.

Remark 2.3 (Finite action sets & PAC learnability)**.** For finite feasible action sets $\mathbb{A}(s_t)$, the condition $\text{int}(\mathcal{C}_T) \neq \emptyset$ holds automatically under our assumption that the expert's optimal action is unique. Indeed, at every observed context, each alternative $a \in \mathbb{A}(s_t) \setminus \{a_t^\star\}$ satisfies $\langle \theta^\star, \delta(s_t, a) \rangle < 0$. Since there are only finitely many such strict linear inequalities, they remain strict throughout a sufficiently small neighborhood of $\theta^\star$. This neighborhood is therefore contained in $\mathcal{C}_T$, proving $\theta^\star \in \text{int}(\mathcal{C}_T)$. This observation connects our action-level guarantee to PAC learnability for linear structured prediction with finite action sets (Daniely & Shalev-Shwartz, 2014, Theorem 5). In this setting, our sample-complexity bound in Theorem 2.2 improves their upper bound by removing the additional $\log(1/\varepsilon)$ factor multiplying d. Furthermore, our result extends their finite-action guarantee to infinite feasible action sets $\mathbb{A}(s_t)$ under the assumption that $\text{int}(\mathcal{C}_T) \neq \emptyset$ almost surely.

Revisiting Example 1 for incenter solution. Under the incenter formulation (11), the parameter $\hat{\theta}_T^{\text{sub}} = (1,0,0)$ is no longer feasible. Indeed, for $s = 0$ and the competitor $a' = (1,1)$, we have $a_{\theta^\star}(0) = (1,-1)$ and $\delta(0,a') = (0,2,0)$, and therefore $\langle (1,0,0), \delta(0,a') \rangle + \|\delta(0,a')\| = 2 > 0$, violating the incenter constraint. A feasible incenter solution is, for instance, $\hat{\theta}_T^{\text{in}} = (2,-2,6)$, which satisfies all incenter constraints and yields the greedy score

$$\langle \hat{\theta}_T^{\text{in}}, \psi(s,a) \rangle = 2a_1 + (-2+6s)a_2.$$

As a result, the induced argmax sets are singletons: $\mathcal{A}_{\hat{\theta}_T^{\text{in}}}(0) = \{(1,-1)\}, \mathcal{A}_{\hat{\theta}_T^{\text{in}}}(1) = \{(1,1)\}$. Hence, the incenter estimator yields a unique action set.

2.3 Regret upperbound

We extend our IO framework to a stochastic online protocol. At each round $t = 1, 2, \ldots, T$, a fresh context $s_t \sim \mathbb{P}_{\mathcal{S}}$ arrives; the learner (i) computes an IO estimate from the demonstrations observed so far, (ii) selects a greedy action under this estimate for the new context, and (iii) observes the expert action under the same context. For the regret analysis, we impose the following boundedness assumption on the feature map.

Assumption 1 (Bounded features)**.** There exists $C > 0$ such that $\forall (s, a) \in \mathcal{S} \times \mathbb{A}, \|\psi(s, a)\|_2 \leq C$.

This assumption ensures that each action mismatch incurs bounded regret, allowing us to translate action-level mismatch guarantees into regret bounds. Without such control, even a small mismatch probability can lead to arbitrarily large regret. Recall that $\delta(s, a) = \psi(s, a) - \psi\big(s, a_{\theta^\star}(s)\big)$. Then, by Assumption 1, $\|\delta(s, a)\|_2 \leq B := 2C$. We evaluate performance via regret with respect to the expert score under $\theta^\star$ (see definitions of instantaneous regret r_t and cumulative regret R_T in (9)).

Proposition 2.4 (Regret upper bound)**.** *Suppose Assumption 1 holds and* $\mathrm{int}(\mathcal{C}_t) \neq \emptyset$ *almost surely for every* $t \in \{1, \ldots, T-1\}$*. Consider the stochastic online protocol based on the incenter estimator, where* $a_t = a_{\hat{\theta}^{\mathrm{in}}_{t-1}}(s_t)$*. Fix* $\beta \in (0, 1)$ *and let* $\varepsilon \in (0, 1]$ *satisfy* $t - 1 \geq N(\varepsilon, \beta)$*. Then, with probability at least* $1 - \beta$ *over the draw of* D_{t-1}*,*

$$\mathbb{E}_{\mathcal{S}}[r_t \mid D_{t-1}] \leq \|\theta^\star\|_2 \, B \, \varepsilon. \tag{14}$$

Moreover, for every $\delta \in (0, 1)$ *and* $T \geq d + 1$*, with probability at least* $1 - \delta$ *over* D_T*,*

$$R_T = \mathcal{O}\left(\|\theta^\star\|_2 \, B \left(d + \log \frac{T}{\delta} \right) \log \frac{T}{d} \right). \tag{15}$$

A more complete statement and proof of Proposition 2.4 are given in Appendix B.5. Our resulting $\mathcal{O}(d \log T)$-type regret matches the guarantees obtained by cutting-plane approaches (Gollapudi et al., 2021) and second-order online updates (Sakaue et al., 2025). A key conceptual distinction of this work with (Besbes et al., 2025) is the role of the environment. In particular, their analysis shows that a naive greedy circumcenter policy can suffer linear regret under adversarial instances, whereas in our stochastic setting the greedy incenter protocol enjoys logarithmic regret.

3 Tightness of the Guarantees

We establish two complementary tightness results. First, a minimax lower bound shows that the action-level generalization rate of $\hat{\theta}^{\mathrm{in}}_T$ is optimal up to constants among all learning algorithms. Second, we construct a single IO instance that attains the set-level bound in Proposition 2.1 exactly for every estimator in a restricted class and every $T \geq d$. This common-instance construction also establishes instantaneous and expected cumulative regret lower bounds for that class.

3.1 Minimax lower bound

We first show that the action-level rate in Theorem 2.2 cannot be improved in general. This result follows by connecting realizable binary classification to our IO framework and applying the lower bound established by Ehrenfeucht et al. (1989).

Proposition 3.1 (Minimax lower bound)**.** *Let* $\mathcal{A}_T$ *be any algorithm mapping* D_T *to a measurable policy* $\widehat{a}_T = \mathcal{A}_T(D_T)$*. There exist constants* $c, \varepsilon_0, \delta_0 > 0$ *such that, for every* $\varepsilon \in (0, \varepsilon_0]$ *and* $\delta \in (0, \delta_0]$*, if* $T \leq c\frac{d + \log(1/\delta)}{\varepsilon}$*, then*

$$\inf_{\mathcal{A}_T} \sup_{\theta^\star, P_S} \mathbb{P}_{D_T, \mathcal{A}_T} \left(P_S \left(\widehat{a}_T(s) \neq a_{\theta^\star}(s) \right) > \varepsilon \right) \geq \delta.$$

The proof of Proposition 3.1 is given in Appendix B.6. The lower bound applies to every learning algorithm, in particular to the incenter estimator. Consequently, the action-level rate in Theorem 2.2 is minimax optimal up to universal constants.

Proposition 3.1 holds for every algorithm, but the hard instance $(\theta^\star, P_S)$ may depend on the algorithm and the sample size. The following subsections ask whether a single IO instance can be hard for every estimator in a specified class and every $T \geq d$. We establish a uniform-instance lower bound for scenario-program estimators induced by the following restricted objective class in (6):

$$\mathcal{J} := \{\theta \mapsto f(\langle c, \theta \rangle + b) \mid f : \mathbb{R} \to \mathbb{R} \text{ is strictly monotone}, \; b \in \mathbb{R}, \; \langle c, \theta \rangle \text{ is non-constant on } \Theta\}.$$

3.2 Set-level and action-level mismatch lowerbounds

The bound from (Campi & Garatti, 2008, Theorem 1) provides a tightness guarantee when the scenario program satisfies a fully-supported condition (see Definition A.5 for the definition of a fully-supported problem). We build an IO instance corresponding to a fully-supported scenario program, which shows that the bound from Proposition 2.1 is not improvable in general.

Theorem 3.2 (Tightness of set-level mismatch)**.** *For every $T \geq d$ and $\varepsilon \in (0,1]$, $\hat{\theta}_T^{\mathrm{sub}}$ satisfies*

$$\sup_{\theta^\star, \mathbb{P}_{\mathcal{S}}} \inf_{J \in \mathcal{J}} \mathbb{P}_D\Big(\mathbb{P}_{\mathcal{S}}\big(a_{\theta^\star}(s) \notin \mathcal{A}_{\hat{\theta}_T^{\mathrm{sub}}}(s)\big) > \varepsilon\Big) \geq \beta := \sum_{i=0}^{d-1} \binom{T}{i} \varepsilon^i (1-\varepsilon)^{T-i}. \tag{16}$$

In particular, if ε is chosen so that the right-hand side equals β, then, up to constants, $\varepsilon \asymp \frac{d+\log(1/\beta)}{T}$.

Proof. Let $\mathcal{S} = \mathbb{S}^{d-1}$, and let $s \sim \mathbb{P}_{\mathcal{S}}$ have a density on the sphere, i.e., $\mathbb{P}_{\mathcal{S}}$ is absolutely continuous with respect to the uniform surface-area measure on $\mathbb{S}^{d-1}$. Let $\mathbb{A} = [-1, 1]$. Parametrize $\theta \in \mathbb{R}^{d+1}$ as $\theta = (\theta_1, \theta_{-1})$ and restrict to the affine slice $\Theta := \{(1, \theta_{-1}) : \theta_{-1} \in \mathbb{R}^d\}$. Define ψ by

$$\psi(s, a) := \begin{pmatrix} -2|a| + a \\ a\, s \end{pmatrix}.$$

Then $F_\theta(s, a) = -2|a| + a + a\, s^\top \theta_{-1}$. Let $\theta^\star = (1, \mathbf{0})$. Since $F_{\theta^\star}(s, a) = -2|a| + a \leq 0$, with equality only at $a = 0$, we have $a_{\theta^\star}(s) = 0$ for all $s \in \mathcal{S}$. For the action 0, we have $F_\theta(s, 0) = 0$, and hence $\ell_\theta^{\mathrm{sub}}(s, 0) = \max_{a \in [-1,1]} F_\theta(s, a)$. Splitting over $a \in [0, 1]$ and $a \in [-1, 0]$ gives

$$\ell_\theta^{\mathrm{sub}}(s, 0) = \max\{0,\ s^\top \theta_{-1} - 1,\ -s^\top \theta_{-1} - 3\}.$$

Thus $\ell_\theta^{\mathrm{sub}}(s, 0) \leq 0$ is equivalent to the asymmetric strip constraints $s^\top \theta_{-1} \leq 1$ and $-s^\top \theta_{-1} \leq 3$. For any objective $J(\theta) = f(\langle c, \theta \rangle + b) \in \mathcal{J}$, strict monotonicity of f and non-constancy of $\langle c, \theta \rangle$ on Θ imply that minimizing J over the feasible polytope is equivalent to either minimizing or maximizing a non-constant linear functional in θ_{-1}. Hence the scenario program (6) reduces to an LP of the form

$$\hat{\theta}_{-1} \in \arg\min_{\theta_{-1} \in \mathbb{R}^d} c_{-1}^\top \theta_{-1} \quad \text{s.t.} \quad s_t^\top \theta_{-1} \leq 1, \quad -s_t^\top \theta_{-1} \leq 3, \quad t \in [T].$$

Under the absolute continuity of $\mathbb{P}_{\mathcal{S}}$, for $T \geq d$ the directions $(s_t)_{t=1}^T$ span $\mathbb{R}^d$ almost surely, so the feasible set is a bounded polytope. Moreover, the linear objective yields a unique vertex of the polytope almost surely. At such an optimum, exactly d constraints are active almost surely. Since the two inequalities generated by the same sample cannot be simultaneously active, these active constraints correspond to d distinct samples. As a result, the scenario program is fully supported with support size d almost surely. Hence, by (Campi & Garatti, 2008, Theorem 1), the violation probability of the scenario optimizer obeys the exact tail identity. Since this violation probability is precisely $\mathbb{P}_{\mathcal{S}}\big(a_{\theta^\star}(s) \notin \mathcal{A}_{\hat{\theta}_T^{\mathrm{sub}}}(s)\big)$, taking the infimum over $J \in \mathcal{J}$ and then the supremum over $(\theta^\star, \mathbb{P}_{\mathcal{S}})$ gives the stated lower bound.

In the simplest nontrivial case $\theta_{-1} \in \mathbb{R}^2$, each sample induces a random strip constraint, so the consistency set becomes a random polytope around $\theta^\star$. Minimizing J selects a vertex determined by d active constraints; see Figure 1 (b). □

Theorem 3.2 shows that the dependence on (d, T, β) in Proposition 2.1 cannot be improved in a distribution-free sense and without assuming additional structure.

Remark 3.3 (Min–max vs. max–min lower bounds)**.** Let $L(\mathcal{A}, I)$ denote the mismatch probability of algorithm $\mathcal{A}$ on instance I. For any fixed algorithm class,

$$\sup_I \inf_{\mathcal{A}} L(\mathcal{A}, I) \leq \inf_{\mathcal{A}} \sup_I L(\mathcal{A}, I).$$

Thus, a max–min lower bound is stronger than the corresponding minimax lower bound for the same class: it requires a common hard instance rather than allowing the instance to depend on the algorithm. Proposition 3.1 establishes a minimax lower bound over all learners, whereas Theorem 3.2 gives this stronger conclusion within the restricted class $\mathcal{J}$, constructing a single instance that is hard for every estimator induced by $\mathcal{J}$. Specifying this class is essential: for each fixed instance, an inner infimum over arbitrary convex objectives could select $J(\theta) = \|\theta - \theta^\star\|_2^2$, and the resulting estimator uniquely recovers $\theta^\star$ and has zero mismatch. Restricting the objective to $\mathcal{J}$ rules out this degeneracy while retaining the structured class of boundary-selecting estimators considered in Theorem 3.2.

Remark 3.4 (Tightness of action-level mismatch)**.** For any tie-breaking rule $a_{\hat{\theta}_T^{\mathrm{sub}}}(s) \in \mathcal{A}_{\hat{\theta}_T^{\mathrm{sub}}}(s)$, if $a_{\theta^\star}(s) \notin \mathcal{A}_{\hat{\theta}_T^{\mathrm{sub}}}(s)$ then we have $a_{\theta^\star}(s) \neq a_{\hat{\theta}_T^{\mathrm{sub}}}(s)$. Hence, the lower bound on set-level mismatch in Theorem 3.2 immediately transfers to action-level mismatch.

3.3 Regret lower bound

We next convert the action-level lower bound in Remark 3.4 into a regret lower bound. In the proof of Theorem 3.2, every action-level error incurs constant positive regret. Thus, the instantaneous regret inherits the same lower tail, and summing over time yields a logarithmic cumulative regret lower bound. The proof of the following result is given in Appendix B.7.

Proposition 3.5 (Tightness of regret)**.** *Consider a protocol that at each round t selects $a_t \in \mathcal{A}_{\hat{\theta}_{t-1}^{\mathrm{sub}}}(s_t)$ according to any measurable tie-breaking rule. Then, for every $t \geq d+1$ and $\varepsilon \in (0,1]$,*

$$\sup_{\theta^\star, \mathbb{P}_{\mathcal{S}}} \inf_{J \in \mathcal{J}} \mathbb{P}_{D_{t-1}}(\mathbb{E}_{\mathcal{S}}[r_t \mid D_{t-1}] \geq \varepsilon) \geq \sum_{i=0}^{d-1} \binom{t-1}{i} \varepsilon^i (1-\varepsilon)^{t-1-i}. \tag{17}$$

In particular, for every $T \geq d+1$, we have

$$\sup_{\theta^\star, \mathbb{P}_{\mathcal{S}}} \inf_{J \in \mathcal{J}} \mathbb{E}[R_T] \geq d \log \frac{T+1}{d+1}. \tag{18}$$

As shown in Table 1, this $\Omega(d \log T)$ lower bound in Proposition 3.5 matches the known adversarial upper bounds. Thus, within the considered estimator class, the stochastic IO setting is effectively not easier than the adversarial one.

4 A Parameter-free Algorithm

Our analysis has mainly focused on statistical guarantees for exact solutions returned by either $\hat{\theta}_T^{\mathrm{sub}}$ in (6) or $\hat{\theta}_T^{\mathrm{in}}$ in (11). These programs can be solved using standard convex solvers. Recall that the auxiliary objective $J(\theta)$ is used to ensure uniqueness of the optimizer. Since $J(\theta)$ is generic, we expect that any optimization method that produces an approximate solution in $\mathcal{C}_T$ exhibits similar generalization behavior in practice. Define

$$f_T(\theta) := \max_{1 \leq t \leq T} \ell_\theta^{\mathrm{sub}}(s_t, a_t^\star) = \max_{1 \leq t \leq T} \max_{a \in \mathbb{A}(s_t)} \langle \theta, \delta(s_t, a) \rangle.$$

By definition of the consistency set $\mathcal{C}_T$, finding a point in $\mathcal{C}_T$ is equivalent to minimizing $f_T(\theta)$, whose optimal value is zero. Since f_T is a non-smooth convex function, we can minimize it with a subgradient method using the Polyak step-size, yielding a parameter-free procedure. Let θ_i^{Pol} denote the iterate at iteration i. Choose the most violated constraint t_i and its corresponding action a_i at θ_i^{Pol} as the maximizer of $\langle \theta_i^{\mathrm{Pol}}, \delta(s_t, a) \rangle$ for $1 \leq t \leq T$ and $a \in \mathbb{A}(s_t)$. Then $\delta(s_{t_i}, a_i)$ is a subgradient of f_T at θ_i^{Pol}, and the Polyak update is

$$\theta_{i+1}^{\mathrm{Pol}} \leftarrow \theta_i^{\mathrm{Pol}} - \frac{f_T(\theta_i^{\mathrm{Pol}})}{\|\delta(s_{t_i}, a_i)\|_2^2} \delta(s_{t_i}, a_i). \tag{19}$$

Standard results for Polyak step-sizes on non-smooth convex objectives yield a convergence rate of $\mathcal{O}(1/\sqrt{N})$ after N iterations of (19). In particular, for a finite number of iterations we cannot guarantee that $\theta_N^{\mathrm{Pol}} \in \mathcal{C}_T$, and therefore the generalization bound of Proposition 2.1 does not directly apply to θ_N^{Pol}. Nevertheless, as we show empirically in Section 5, θ_N^{Pol} exhibits generalization behavior similar to $\hat{\theta}_T^{\mathrm{sub}}$ when $N = T$, i.e., when we perform T Polyak updates. This observation suggests that extending our theory to approximate solutions is an interesting direction for future work. Finally, to avoid the trivial $\theta = 0$ solution, one should initialize (19) away from the origin.

Remark 4.1 (Complexity reduction)**.** The algorithm (19) has lower per-iteration cost than generic solvers (e.g., interior-point methods) for (6). When $\mathbb{A}$ is finite, each update (19) identifies the most violated pair (t_i, a_i) via a scan over $t \in [T]$ and $a \in \mathbb{A}$, costing $\mathcal{O}(T|\mathbb{A}|)$. In contrast, an interior-point method must handle $T|\mathbb{A}|$ inequality constraints, and its worst-case per-iteration cost is $\mathcal{O}(T^3|\mathbb{A}|^3)$.

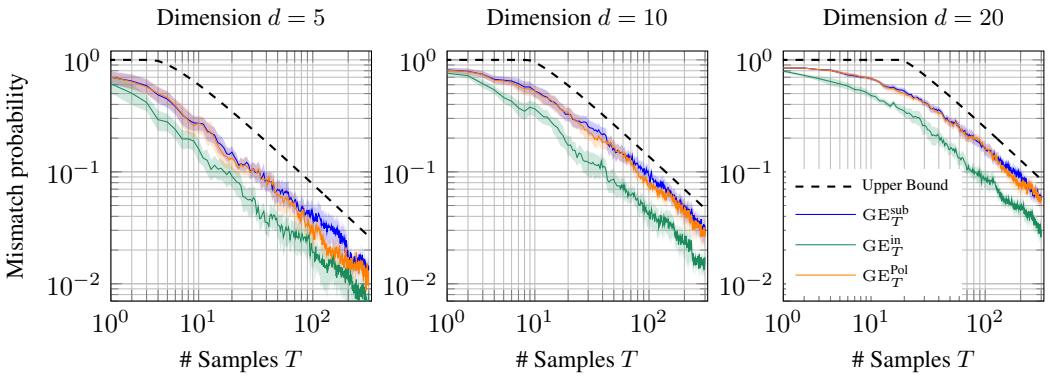


Figure 2: Generalization mismatch probability GE_T on fresh contexts versus T for $d \in \{5, 10, 20\}$. The depicted results are averaged over 10 runs with 90% confidence bands. The dashed line shows the theoretical upper bound from Proposition 2.1 with $\beta = 0.1$. The experiments confirm our theoretical findings.

5 Numerical Experiments

In this section, we verify the generalization guarantees in Proposition 2.1, Theorem 2.2, and the parameter-free method of Section 4 on synthetic data. The details of our experimental setup are given in Appendix A.4. We compare three estimators $\hat{\theta}_T^{\mathrm{sub}}$, $\hat{\theta}_T^{\mathrm{in}}$, and $\hat{\theta}_T^{\mathrm{Pol}}$ derived from (6), (11), and the parameter-free approach in Section 4, respectively. Figure 2 shows the generalization mismatch probability on fresh contexts for $d \in \{5, 10, 20\}$. In all three panels, the mismatch drops as we collect more data, and the curves follow an $\mathcal{O}(1/T)$ decay, matching the predictions of Proposition 2.1 and Theorem 2.2. The incenter estimator $\hat{\theta}_T^{\mathrm{in}}$ performs best across all dimensions, suggesting that enforcing a strict margin reduces ties in the greedy rule and improves action prediction, consistent with Theorem 2.2. The parameter-free Polyak method $\hat{\theta}_T^{\mathrm{Pol}}$ tracks $\hat{\theta}_T^{\mathrm{sub}}$ closely, providing a solver-free alternative with similar statistical performance. Comparing the three panels, the mismatch is larger for higher dimensions, which agrees with the linear dependence on d suggested by our theory.

6 Limitations and Concluding Remarks

We established tight generalization guarantees for IO in the noiseless regime by viewing empirical loss as a scenario program for the first time. Our work comes with its own limitations that naturally suggest future research directions.

Faster rate under further structure. We derived a tight set-level mismatch bound of order $\mathcal{O}(d/T)$, later extended to an action-level result using the incenter estimator. It remains an open question whether faster rates are achievable under additional assumptions, e.g., strong convexity/smoothness of the feature map or distributional assumptions on the context.

Active learning. Our analysis assumes the contexts $\{s_t\}_{t=1}^T$ are drawn i.i.d. from a fixed distribution $\mathbb{P}_{\mathcal{S}}$. However, in many interactive settings, the learner can adaptively choose which contexts to query next, suggesting an active variant of IO. Designing s_{t+1} based on past observations to shrink the consistency set $\mathcal{C}_T$ more rapidly is a promising direction.

Noisy dataset. Although our IO problem has a stochastic relation with the environment, it does not take any additive noise into account. Since additive noise bridges IO and online learning in more realistic scenarios, extending the analysis to noisy observations is an important next step.

Computational complexity vs. generalization. We proposed a parameter-free algorithm with lower per-iteration cost than generic solvers for computing $\hat{\theta}_T^{\mathrm{sub}}$. We empirically observed that this method exhibits comparable generalization behavior, but we have not yet provided a corresponding theory. Establishing such a formal tradeoff between computational complexity and the resulting generalization guarantees is also an important future direction.

Acknowledgments

Pouria Fatemi and Suvrit Sra acknowledge generous support from the Alexander von Humboldt Foundation. Hoomaan Maskan acknowledges support from the Science for Life Laboratory (SciLifeLab), and the Wallenberg AI, Autonomous Systems, and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation. Peyman Mohajerin Esfahani acknowledges the support by the ERC Starting Grant TRUST-949796. The paper is based primarily on the work done while Hoomaan Maskan was at Umeå University. The authors declare no competing interests.

References

Ahuja, R. K. and Orlin, J. B. Inverse optimization. *Operations Research*, 49(5):771–783, 2001.

Akhtar, S. A., Kolarijani, A. S., and Mohajerin Esfahani, P. Learning for control: An inverse optimization approach. *IEEE Control Systems Letters*, 6:187–192, 2021.

Aswani, A., Shen, Z.-J., and Siddiq, A. Inverse optimization with noisy data. *Operations Research*, 66(3):870–892, 2018.

Ben-David, S., Cesa-Bianchi, N., and Long, P. M. Characterizations of learnability for classes of {0,..., n}-valued functions. *Journal of computer and system sciences (Print)*, 50(1):74–86, 1995.

Bertsimas, D., Gupta, V., and Paschalidis, I. C. Data-driven estimation in equilibrium using inverse optimization. *Mathematical Programming*, 153(2):595–633, 2015.

Besbes, O., Fonseca, Y., and Lobel, I. Contextual inverse optimization: Offline and online learning. *Operations Research*, 73(1):424–443, 2025.

Brukhim, N., Carmon, D., Dinur, I., Moran, S., and Yehudayoff, A. A characterization of multiclass learnability. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 943–955. IEEE, 2022.

Campi, M. C. and Garatti, S. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008.

Chan, T. C., Mahmood, R., and Zhu, I. Y. Inverse optimization: Theory and applications. *Operations Research*, 73(2):1046–1074, 2025.

Daniely, A. and Shalev-Shwartz, S. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pp. 287–316. PMLR, 2014.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Dimanidis, I., Ok, T., and Mohajerin Esfahani, P. Offline reinforcement learning via inverse optimization. *arXiv preprint arXiv:2502.20030*, 2025.

Dzhaparidze, K. and Van Zanten, J. On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117, 2001.

Ehrenfeucht, A., Haussler, D., Kearns, M., and Valiant, L. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

El Balghiti, O., Elmachtoub, A. N., Grigas, P., and Tewari, A. Generalization bounds in the predict-then-optimize framework. *Mathematics of Operations Research*, 48(4):2043–2065, 2023.

Elmachtoub, A. N. and Grigas, P. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

Freedman, D. A. On tail probabilities for martingales. *The Annals of Probability*, pp. 100–118, 1975.

Gollapudi, S., Guruganesh, G., Kollias, K., Manurangsi, P., Leme, R., and Schneider, J. Contextual recommendations and low-regret cutting-plane algorithms. *Advances in Neural Information Processing Systems*, 34:22498–22508, 2021.

Hazan, E. et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Keshavarz, A., Wang, Y., and Boyd, S. Imputing a convex objective function. In *IEEE International Symposium on Intelligent Control*, pp. 613–619, 2011.

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

Li, J. Y.-M. Inverse optimization of convex risk functions. *Management Science*, 67(11):7113–7141, 2021.

Mohajerin Esfahani, P., Shafieezadeh-Abadeh, S., Hanasusanto, G. A., and Kuhn, D. Data-driven inverse optimization with imperfect information. *Mathematical Programming*, 167(1):191–234, 2018.

Natarajan, B. K. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

Pabbaraju, C. The optimal sample complexity of multiclass and list learning. *arXiv preprint arXiv:2604.24749*, 2026.

Ren, K., Mohajerin Esfahani, P., and Georghiou, A. Inverse optimization via learning feasible regions. *arXiv preprint arXiv:2505.15025*, 2025.

Sakaue, S., Tsuchiya, T., Bao, H., and Oki, T. Online inverse linear optimization: Efficient logarithmic-regret algorithm, robustness to suboptimality, and lower bound. *Advances in Neural Information Processing Systems*, 34, 2025.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Zattoni Scroccaro, P., Atasoy, B., and Mohajerin Esfahani, P. Learning in inverse optimization: Incenter cost, augmented suboptimality loss, and algorithms. *Operations Research*, 73(5):2661–2679, 2025a.

Zattoni Scroccaro, P., van Beek, P., Mohajerin Esfahani, P., and Atasoy, B. Inverse optimization for routing problems. *Transportation Science*, 59(2):301–321, 2025b.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the International Conference on Machine Learning*, pp. 928–936, 2003.

A Background and Helpful Lemmas

In this section, we present some of the important lemmas and background used in this work.

A.1 Chance constraints and scenario programs

We now recall finite-sample scenario program and present its corresponding chance-constrained optimization. Let $\Theta \subset \mathbb{R}^d$ be a convex set and let $J : \Theta \to \mathbb{R}$ be a convex function. Consider a measurable function $g : \Theta \times \mathcal{S} \to \mathbb{R}$ that is convex in θ for every $s \in \mathcal{S}$. The associated chance-constrained program is

$$\begin{cases} \min\limits_{\theta} & J(\theta) \\ \text{s.t.} & \mathbb{P}_{\mathcal{S}}\big[g(\theta, s) \le 0\big] \ge 1 - \varepsilon, \\ & \theta \in \Theta, \end{cases} \tag{CCP$_\varepsilon$}$$

where $\varepsilon \in [0, 1]$ is the violation level and $\mathbb{P}_{\mathcal{S}}$ denotes the distribution of states. Let $(s_i)_{i=1}^T$ be T i.i.d. samples from $\mathbb{P}_{\mathcal{S}}$. The corresponding *scenario program* is

$$\begin{cases} \min\limits_{\theta} & J(\theta) \\ \text{s.t.} & g(\theta, s_i) \le 0, \quad i = 1, \dots, T, \\ & \theta \in \Theta. \end{cases} \tag{SCP}$$

Since (SCP) depends on the random samples $(s_i)_{i=1}^T$, its optimizer is a random variable. We assume that (SCP) admits a unique solution with probability 1. The following theorem states that if T is large enough, the unique solution of (SCP) is a feasible solution of (CCP$_\varepsilon$) with high probability (Campi & Garatti, 2008).

Theorem A.1 ((Campi & Garatti, 2008))**.** *Let $\beta \in [0, 1]$ and $T \ge N(\varepsilon, \beta)$, where*

$$N(\varepsilon, \beta) := \min \left\{ n \in \mathbb{N} \;\middle|\; \sum_{i=0}^{d-1} \binom{n}{i} \varepsilon^i (1 - \varepsilon)^{n-i} \le \beta \right\}.$$

Then, with probability at least $1 - \beta$ over the random draw of $(s_i)_{i=1}^T$, the optimizer of (SCP) is a feasible solution of (CCP$_\varepsilon$).

Remark A.2 (Low-data regime guarantees)**.** The condition $T \ge N(\varepsilon, \beta)$ is meaningful only when $T \ge d$. If $T < d$, then for every $\varepsilon \in (0, 1)$,

$$\sum_{i=0}^{d-1} \binom{T}{i} \varepsilon^i \, (1 - \varepsilon)^{T-i} = \sum_{i=0}^{T} \binom{T}{i} \varepsilon^i \, (1 - \varepsilon)^{T-i} = 1,$$

so the defining inequality for $N(\varepsilon, \beta)$ cannot hold for any $\beta < 1$. Equivalently, when $T < d$ the theorem can only be stated with $\varepsilon = 1$, making the guarantee non-informative.

The following lemma reformulates the result from Theorem A.1 for any $T = \Omega(d + \log(1/\beta))$.

Lemma A.3. *Fix $\beta \in (0, 1)$. For any $T \ge 2\big(d + \log(1/\beta)\big)$, the choice $\varepsilon = \dfrac{2}{T}\big(d + \log(1/\beta)\big)$ satisfies $T \ge N(\varepsilon, \beta)$.*

The proof of this lemma is given in Appendix B.1. Next, we present the definition of support constraint.

Definition A.4 (Support constraint)**.** A constraint $g(\theta, s_i) \le 0$, for $i = 1, \dots, T$, is called a support constraint of (SCP) if its removal changes the solution of (SCP).

We can now present the notion of a fully supported problem.

Definition A.5 (Fully supported problem)**.** The scenario program (SCP) with $T \ge d$ is called fully supported if the number of its support constraints is exactly d. The chance constraint program (CCP$_\varepsilon$) is fully supported if for $T \ge d$, its corresponding (SCP) is fully supported with probability 1.

A.2 Online protocol

We make the dependence on the history explicit for the proofs. Let $\mathcal{F}_t := \sigma\big((s_1, a_1^\star), \dots, (s_t, a_t^\star)\big)$ be the sigma-algebra generated by the expert demonstrations up to round t (and set $\mathcal{F}_0$ to be the trivial σ-algebra). At round $t \ge 1$, the learner computes $\hat{\theta}_{t-1}^{\text{in}}$ from $D_{t-1} := \{(s_i, a_i^\star)\}_{i=1}^{t-1}$, plays $a_t := a_{\hat{\theta}_{t-1}^{\text{in}}}(s_t)$, and observes $a_t^\star = a_{\theta^\star}(s_t)$. We have $s_t \sim \mathbb{P}_{\mathcal{S}}$ and $s_t \perp\!\!\!\perp \mathcal{F}_{t-1}$.

Under Assumption 1, define $0 \le r_t \le \|\theta^\star\|_2 \, B$ for all t.

For $N \ge 1$, let $\hat{\theta}_N^{\mathrm{in}}$ be the incenter estimator from N i.i.d. demonstrations, and define

$$X_N := \mathbb{P}_{\mathcal{S}}\big(a_{\hat{\theta}_N^{\mathrm{in}}}(s) \neq a_{\theta^\star}(s)\big) \in [0,1].$$

In particular, in the online protocol we write $X_{t-1} := \mathbb{P}_{\mathcal{S}}(a_{\hat{\theta}_{t-1}^{\mathrm{in}}}(s) \neq a_{\theta^\star}(s))$. To bound the expected cumulative regret, we need to bound the expectation of X_N using our generalization result. The proof of this result given below can be found in Appendix B.2.

Lemma A.6 (Expected mismatch probability for the incenter estimator)**.** *Assume* $\mathrm{int}(\mathcal{C}_N) \neq \emptyset$ *almost surely. Let* X_N *be as above. Then for all* $N \ge d$*,*

$$\mathbb{E}_D[X_N] \;\le\; \frac{d}{N+1}, \tag{20}$$

where $\mathbb{E}_D$ *denotes expectation over the* N *i.i.d. demonstrations used to construct* $\hat{\theta}_N^{\mathrm{in}}$*.*

To achieve a high probability bound on cumulative regret, we use Freedman's bound from the martingale concentration inequality literature.

Lemma A.7 (Freedman's inequality (Freedman, 1975; Dzhaparidze & Van Zanten, 2001))**.** *Let* $(M_t)_{t\ge 0}$ *be a martingale with* $M_0 = 0$ *and differences* $Z_t := M_t - M_{t-1}$ *satisfying* $|Z_t| \le b$ *a.s. Let the predictable quadratic variation be*

$$V_t := \sum_{i=1}^{t} \mathbb{E}[Z_i^2 \mid \mathcal{F}_{i-1}].$$

Then, for any fixed $t \ge 1$*, any deterministic* $v \ge 0$*, and any* $\delta \in (0,1)$*,*

$$\mathbb{P}\left(M_t > \sqrt{2v \log\frac{1}{\delta}} + \frac{2}{3} b \log\frac{1}{\delta} \;\text{ and }\; V_t \le v \right) \le \delta.$$

A.3 A classical VC-dimension view of IO

Noiseless inverse optimization admits an exact realizable binary classification interpretation for the set-level mismatch. The dataset $D_T = \{(s_t, a_t^\star)\}_{t=1}^T$ may be viewed as an i.i.d. sample of supervised examples drawn from the distribution of the random pair $(s, a_{\theta^\star}(s))$ with $s \sim \mathbb{P}_{\mathcal{S}}$. For each $\theta \in \mathbb{R}^d$, define the induced binary classifier

$$h_\theta(s,a) := \mathbf{1}\{a \in \mathcal{A}_\theta(s)\} = \mathbf{1}\{\ell_\theta^{\mathrm{sub}}(s,a) = 0\},$$

and let

$$\mathcal{H} := \{h_\theta : \theta \in \mathbb{R}^d\}.$$

Thus, $h_\theta(s,a) = 1$ if and only if action a is greedy under parameter θ in state s. Under this identification, the population risk is

$$R(\theta) := \mathbb{P}_{\mathcal{S}}\big(h_\theta(s, a_{\theta^\star}(s)) = 0\big) = \mathbb{P}_{\mathcal{S}}\big(a_{\theta^\star}(s) \notin \mathcal{A}_\theta(s)\big) = \mathbb{P}_{\mathcal{S}}\big(\ell_\theta^{\mathrm{sub}}(s, a_{\theta^\star}(s)) > 0\big)\,,$$

which is exactly the set-level mismatch probability studied in this paper. Since $R(\theta^\star) = 0$, the induced classification problem is realizable.

Likewise, the empirical classification error is

$$\widehat{R}_T(\theta) := \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{a_t^\star \notin \mathcal{A}_\theta(s_t)\},$$

and

$$\widehat{R}_T(\theta) = 0 \quad\Longleftrightarrow\quad \ell_\theta^{\mathrm{sub}}(s_t, a_t^\star) = 0 \;\; \forall t \in [T] \quad\Longleftrightarrow\quad \theta \in C_T.$$

Hence, any learning rule $D_T \mapsto \widehat{\theta}_T(D_T)$ satisfying $\widehat{\theta}_T(D_T) \in C_T$ is sample-consistent for $\mathcal{H}$, in the sense that $\widehat{R}_T(\widehat{\theta}_T) = 0$. Consequently, whenever $d_{\mathrm{VC}} := \mathrm{VCdim}(\mathcal{H}) < \infty$, standard realizable VC theory implies that, for any confidence level $\delta \in (0,1)$, any such sample-consistent rule satisfies,

$$R(\widehat{\theta}_T) = \mathcal{O}\left(\frac{d_{\mathrm{VC}} \log(T/d_{\mathrm{VC}}) + \log(1/\delta)}{T} \right),$$

with probability at least $1-\delta$, for $T \geq d_{\mathrm{VC}}$ (Shalev-Shwartz & Ben-David, 2014).

This provides a clean conceptual PAC-learning interpretation of noiseless inverse optimization. At the same time, it also highlights the limitation of a purely VC-based viewpoint. Indeed, the observed supervision is one-sided: each demonstration only certifies that the expert action must remain in the argmax set $\mathcal{A}_\theta(s)$ under a candidate parameter θ, and therefore each sample acts by eliminating inconsistent parameters from the parameter space, thereby forming the consistency set C_T. As a result, the relevant complexity is not naturally expressed in terms of arbitrary labelings of state-action pairs. In particular, the quantity d_{VC} is determined by the geometry of the induced hypothesis class $\mathcal{H}$, rather than directly by the parameter dimension d. Therefore, while the VC reduction is conceptually useful, it does not by itself explain the dimension-dependent $O(d/T)$ guarantee established in this paper. This observation motivates the scenario-based view, where each demonstration is treated directly as a random feasibility constraint in parameter space.

A.4 Numerical experiment setup

We consider a contextual linear decision model with $K = |\mathbb{A}| = 15$ discrete actions and d-dimensional features. The expert follows a greedy policy for a fixed but unknown parameter $\theta^\star \in \mathbb{R}^d$ satisfying $\sum_{i=1}^d \theta_i^\star = 1$. For each $k \in [K]$, we generate a fixed action vector $a_k \sim \mathcal{N}(0, I_d)$ once at the beginning of the experiment and normalize it, and set $\mathbb{A} = \{a_1, \ldots, a_K\}$. At each round, we draw a fresh random linear map $s_t \in \mathbb{R}^{d\times d}$, $(s_t)_{ij} \sim \mathrm{Unif}[-3, 3]$, and define the context features as $\psi(s_t, a_k) = s_t a_k \in \mathbb{R}^d$. The learner observes $\{(s_t, a_t^\star)\}_{t=1}^T$ for $T \leq T_{\max} = 300$ and uses a greedy plug-in policy based on an estimate $\hat{\theta}_T$ to select $a_{\hat{\theta}_T}(s)$. Ties are broken by choosing the smallest-index action. Throughout, we set $J(\theta) = \|\theta\|_2^2$ and $\Theta = \{\theta \in \mathbb{R}^d \mid \sum_{i=1}^d \theta_i = 1\}$. We compare three estimators $\hat{\theta}_T^{\mathrm{sub}}$, $\hat{\theta}_T^{\mathrm{in}}$, and $\hat{\theta}_T^{\mathrm{Pol}}$ derived from (6), (11), and the parameter-free approach in Section 4, respectively. To directly test generalization, we evaluate the current estimate $\hat{\theta}_T$ on $N_{\mathrm{test}} = 1000$ fresh contexts and record the empirical mismatch rate

$$\mathrm{GE}_T := \frac{1}{N_{\mathrm{test}}} \sum_{j=1}^{N_{\mathrm{test}}} \mathbb{I}\Big\{ a_{\hat{\theta}_T}(s_j) \neq a_{\theta^\star}(s_j) \Big\}.$$

We repeat the full $T_{\max}$-round experiment for 10 independent Monte-Carlo runs. For each curve, we report the mean across runs together with a two-sided 90% normal confidence band. The convex programs (6) and (11) are solved using CVXPY (Diamond & Boyd, 2016). All experiments were run on a personal computer and required no specialized hardware.

B Proofs

B.1 Proof of Lemma A.3

Recall that

$$\sum_{i=0}^{d-1} \binom{T}{i} \varepsilon^i (1-\varepsilon)^{T-i} = \mathbb{P}[X \le d-1], \qquad X \sim \mathrm{Bin}(T, \varepsilon),$$

with mean $\mu := \mathbb{E}[X] = T\varepsilon$. For our choice of ε we have $\mu = T\varepsilon = 2\big(d + \log(1/\beta)\big) > d-1$, hence $\delta := 1 - \frac{d-1}{\mu} \in (0,1)$ and $(1-\delta)\mu = d-1$. By the Chernoff lower-tail bound,

$$\mathbb{P}[X \le d-1] = \mathbb{P}\big[X \le (1-\delta)\mu\big] \le \exp\left(-\frac{\delta^2 \mu}{2}\right) = \exp\left(-\frac{(\mu - d + 1)^2}{2\mu}\right).$$

Now, if $T\varepsilon \ge 2\big(\log(1/\beta) + d\big)$, then in particular $T\varepsilon \ge 2\big(\log(1/\beta) + d - 1\big)$ and therefore

$$\frac{(T\varepsilon - d + 1)^2}{2T\varepsilon} = \frac{(T\varepsilon - (d-1))^2}{2T\varepsilon} \ge \frac{T\varepsilon}{2} - (d-1) \ge \log(1/\beta).$$

Thus $\mathbb{P}[X \le d-1] \le \exp(-\log(1/\beta)) = \beta$, i.e.,

$$\sum_{i=0}^{d-1} \binom{T}{i} \varepsilon^i (1-\varepsilon)^{T-i} \le \beta,$$

which means $T \ge N(\varepsilon, \beta)$ by definition of $N(\varepsilon, \beta)$.

B.2 Proof of Lemma A.6

To compute $\mathbb{E}_D[X_N]$, use the tail-integral formula:

$$\mathbb{E}_D[X_N] = \int_0^1 \mathbb{P}_D(X_N > \varepsilon)\, d\varepsilon.$$

By Theorem 2.2,

$$\mathbb{P}_D(X_N > \varepsilon) \le \sum_{i=0}^{d-1} \binom{N}{i} \varepsilon^i (1-\varepsilon)^{N-i}.$$

Hence

$$\mathbb{E}_D[X_N] \le \sum_{i=0}^{d-1} \binom{N}{i} \int_0^1 \varepsilon^i (1-\varepsilon)^{N-i}\, d\varepsilon.$$

The integral is a Beta integral:

$$\int_0^1 \varepsilon^i (1-\varepsilon)^{N-i}\, d\varepsilon = \frac{i!(N-i)!}{(N+1)!}.$$

Therefore

$$\binom{N}{i} \cdot \frac{i!(N-i)!}{(N+1)!} = \frac{N!}{i!(N-i)!} \cdot \frac{i!(N-i)!}{(N+1)!} = \frac{1}{N+1}.$$

Summing over $i = 0, \dots, d-1$, we obtain

$$\mathbb{E}_D[X_N] \le \sum_{i=0}^{d-1} \frac{1}{N+1} = \frac{d}{N+1},$$

which implies (20).

B.3 Proof of Proposition 2.1

Consider the chance-constrained program

$$\min_{\theta\in\Theta} J(\theta) \quad \text{s.t.} \quad \mathbb{P}_{\mathcal{S}}\big(\ell_\theta^{\text{sub}}(s, a_{\theta^\star}(s)) \le 0\big) \;\ge\; 1-\varepsilon. \tag{21}$$

For a small ε, any feasible solution of (21) is a generalizable result. Since $a_t^\star = a_{\theta^\star}(s_t)$, the constraints in (6) are exactly the scenario constraints obtained by sampling $s_t \sim \mathbb{P}_{\mathcal{S}}$ and enforcing $\ell_\theta^{\text{sub}}(s_t, a_{\theta^\star}(s_t)) \le 0$ for all $t \in [T]$. Therefore, we use the result from (Campi & Garatti, 2008, Theorem 1). This theorem states that if $T \ge N(\varepsilon, \beta)$, then with probability at least $1-\beta$ over the draw of $(s_t)_{t=1}^T$, the unique optimizer $\hat{\theta}_T^{\text{sub}}$ of the scenario program (6) is feasible for the chance constraint problem (21), i.e.,

$$\mathbb{P}_{\mathcal{S}}\big(\ell_{\hat{\theta}_T^{\text{sub}}}^{\text{sub}}(s, a_{\theta^\star}(s)) \le 0\big) \;\ge\; 1-\varepsilon. \tag{22}$$

By definition, $\ell_\theta^{\text{sub}}(s, a_{\theta^\star}(s)) \ge 0$ for all (θ, s). Therefore, the event $\{\ell_\theta^{\text{sub}}(s, a_{\theta^\star}(s)) \le 0\}$ is equivalent to $\{\ell_\theta^{\text{sub}}(s, a_{\theta^\star}(s)) = 0\}$. Moreover, by (4), $\ell_\theta^{\text{sub}}(s, a_{\theta^\star}(s)) = 0$ is equivalent to $a_{\theta^\star}(s) \in \mathcal{A}_\theta(s)$. Applying this equivalence to (22) yields:

$$\mathbb{P}_{\mathcal{S}}\big(a_{\theta^\star}(s) \in \mathcal{A}_{\hat{\theta}_T^{\text{sub}}}(s)\big) \;\ge\; 1-\varepsilon,$$

which is equivalent to (7). It is also possible to make the choice of ε explicit as a function of T. In particular, for any $T \ge 2\big(d + \log(1/\beta)\big)$, the choice $\varepsilon = \frac{2}{T}\big(d + \log(1/\beta)\big)$ satisfies $T \ge N(\varepsilon, \beta)$. See Lemma A.3 for the proof.

B.4 Proof of Theorem 2.2

Consider the chance-constrained problem

$$\min_{\theta\in\mathbb{R}^d} J(\theta) \quad \text{s.t.} \quad \mathbb{P}_{\mathcal{S}}\big(\ell_\theta^{\text{in}}(s, a_{\theta^\star}(s)) \le 0\big) \ge 1-\varepsilon,$$

whose associated scenario program is (11). For brevity, write $\hat{\theta} := \hat{\theta}_T^{\text{in}}$ and $\ell(s) := \ell_{\hat{\theta}}^{\text{in}}\big(s, a_{\theta^\star}(s)\big)$. By (Campi & Garatti, 2008, Theorem 1), if $T \ge N(\varepsilon, \beta)$, with probability at least $1-\beta$ over the draw of the training sample D_T,

$$\mathbb{P}_{\mathcal{S}}(\ell(s) \le 0) \;\ge\; 1-\varepsilon. \tag{23}$$

We now show that, for every $s \in \mathcal{S}$,

$$\ell(s) \le 0 \implies \Big(\mathcal{A}_{\hat{\theta}_T^{\text{in}}}(s) = \{a_{\theta^\star}(s)\}\Big). \tag{24}$$

Fix any s with $\ell(s) \le 0$. By the definition of ℓ in (10), for every $a \in \mathbb{A}(s)$,

$$\langle \hat{\theta}, \delta(s,a)\rangle \le -\|\delta(s,a)\|. \tag{25}$$

For $a = a_{\theta^\star}(s)$ we have $\delta(s,a) = 0$, hence (25) holds with equality.

If $a \ne a_{\theta^\star}(s)$, uniqueness of the expert's optimum implies $\langle \theta^\star, \delta(s,a)\rangle < 0$, and hence $\delta(s,a) \ne 0$, so $\|\delta(s,a)\| > 0$ and (25) implies: $\langle \hat{\theta}, \delta(s,a)\rangle < 0$. Equivalently,

$$\langle \hat{\theta}, \psi(s,a)\rangle < \langle \hat{\theta}, \psi(s, a_{\theta^\star}(s))\rangle, \qquad \forall a \ne a_{\theta^\star}(s).$$

Thus $a_{\theta^\star}(s)$ is the unique maximizer of $a \mapsto \langle \hat{\theta}, \psi(s,a)\rangle$, which proves (24). Combining (23) and (24) yields, on the same event of probability at least $1-\beta$ over D_T,

$$\mathbb{P}_{\mathcal{S}}\Big(\mathcal{A}_{\hat{\theta}_T^{\text{in}}}(s) = \{a_{\theta^\star}(s)\}\Big) \ge \mathbb{P}_{\mathcal{S}}(\ell(s) \le 0) \ge 1-\varepsilon.$$

This is the claimed guarantee.

B.5 Proof of Proposition 2.4

Proposition B.1 (Regret upper bound (complete version))**.** *Suppose Assumption 1 holds and* $\mathrm{int}(\mathcal{C}_t) \neq \emptyset$ *almost surely for every* $t \in \{1, \dots, T-1\}$*. Consider the stochastic online protocol based on the incenter estimator, where* $a_t = a_{\hat{\theta}^{\mathrm{in}}_{t-1}}(s_t)$*. Fix* $\beta \in (0,1)$ *and let* $\varepsilon \in (0,1]$ *satisfy* $t-1 \geq N(\varepsilon, \beta)$*. Then, with probability at least* $1-\beta$ *over the draw of* D_{t-1}*,*

$$\mathbb{E}[r_t \mid D_{t-1}] \leq \|\theta^\star\|_2 \, B \, \varepsilon. \tag{26}$$

In particular, for every $t \geq d+1$*,*

$$\mathbb{E}[r_t] \leq \|\theta^\star\|_2 \, B \, \frac{d}{t}. \tag{27}$$

Consequently, for every $T \geq d+1$*,*

$$\mathbb{E}[R_T] = \mathcal{O}\left(\|\theta^\star\|_2 \, B \, d \left(1 + \log \frac{T}{d}\right)\right). \tag{28}$$

Moreover, for every $\delta \in (0,1)$ *and* $T \geq d+1$*, with probability at least* $1-\delta$ *over the online trajectory,*

$$R_T = \mathcal{O}\left(\|\theta^\star\|_2 \, B \, \left(d + \log \frac{T}{\delta}\right) \left(1 + \log \frac{T}{d}\right)\right). \tag{29}$$

Proof. By Assumption 1, we have $0 \leq r_t \leq \|\theta^\star\|_2 \, B$ for all t. Moreover, $r_t = 0$ whenever $a_t = a_t^\star$. Conditioning on D_{t-1}, $\hat{\theta}^{\mathrm{in}}_{t-1}$ is fixed, and $s_t \sim \mathbb{P}_{\mathcal{S}}$ independently of D_{t-1}. Then,

$$\begin{aligned} \mathbb{E}[r_t \mid D_{t-1}] &= \mathbb{E}[r_t \, \mathbf{1}\{a_t \neq a_t^\star\} \mid D_{t-1}] \\ &\leq \|\theta^\star\|_2 \, B \, \mathbb{P}_{\mathcal{S}}\Big(a_{\hat{\theta}^{\mathrm{in}}_{t-1}}(s) \neq a_{\theta^\star}(s)\Big) = \|\theta^\star\|_2 \, B \, X_{t-1}. \end{aligned} \tag{30}$$

By Theorem 2.2, if $t-1 \geq N(\varepsilon, \beta)$, then $X_{t-1} \leq \varepsilon$ with probability at least $1-\beta$ over D_{t-1}. This proves (26).

Taking expectations in (30) yields

$$\mathbb{E}[r_t] \leq \|\theta^\star\|_2 \, B \, \mathbb{E}_D[X_{t-1}].$$

For $t \geq d+1$, Lemma A.6 gives

$$\mathbb{E}_D[X_{t-1}] \leq \frac{d}{t},$$

and hence

$$\mathbb{E}[r_t] \leq \|\theta^\star\|_2 \, B \, \frac{d}{t}.$$

This proves (27).

For cumulative regret, split the sum at d:

$$\mathbb{E}[R_T] = \sum_{t=1}^{\min\{d,T\}} \mathbb{E}[r_t] + \sum_{t=d+1}^{T} \mathbb{E}[r_t].$$

For $t \leq d$, $\mathbb{E}[r_t] \leq \|\theta^\star\|_2 \, B$, and therefore

$$\sum_{t=1}^{\min\{d,T\}} \mathbb{E}[r_t] \leq d \, \|\theta^\star\|_2 \, B.$$

For $t \geq d+1$, using (27),

$$\sum_{t=d+1}^{T} \mathbb{E}[r_t] \leq \|\theta^\star\|_2 \, B \, d \sum_{t=d+1}^{T} \frac{1}{t} \leq \|\theta^\star\|_2 \, B \, d \log \frac{T}{d}.$$

Thus,

$$\mathbb{E}[R_T] = \mathcal{O}\left(\|\theta^\star\|_2 \, B \, d \left(1 + \log \frac{T}{d}\right)\right),$$

which proves (28).

It remains to prove the high-probability cumulative regret bound. Define $\mu_t := \mathbb{E}[r_t \mid \mathcal{F}_{t-1}]$ and the martingale

$$M_t := \sum_{i=1}^{t} (r_i - \mu_i), \qquad M_0 = 0.$$

Then

$$R_T = \sum_{t=1}^{T} \mu_t + M_T. \tag{31}$$

From (30), we have $\mu_t \le \|\theta^\star\|_2 \, B \, X_{t-1}$.

Fix $\delta \in (0,1)$ and set

$$\bar{\beta} := \frac{\delta}{2T}, \qquad t_0 := \min\Big\{T+1,\ \big\lceil 2\big(d + \log \tfrac{2T}{\delta}\big)\big\rceil + 1\Big\}.$$

For any $t \in \{t_0, \dots, T\}$, if this index set is nonempty, we have $t - 1 \ge 2\big(d + \log(1/\bar{\beta})\big)$. Applying Theorem 2.2 with sample size $t-1$ and confidence $\bar{\beta}$, together with Lemma A.3, gives

$$\mathbb{P}\Big(X_{t-1} \le 2\,\frac{d + \log(1/\bar{\beta})}{t-1}\Big) \ge 1 - \bar{\beta}, \qquad t = t_0, \dots, T.$$

A union bound gives an event E with $\mathbb{P}(E) \ge 1 - \delta/2$ on which

$$\forall\, t \in \{t_0, \dots, T\} :\ X_{t-1} \le 2\,\frac{d + \log \frac{2T}{\delta}}{t-1}. \tag{32}$$

For $t < t_0$ we use the trivial bound $X_{t-1} \le 1$. Therefore, on E,

$$\sum_{t=1}^{T} \mu_t \le \|\theta^\star\|_2 \, B \left[(t_0 - 1) + 2\Big(d + \log \frac{2T}{\delta}\Big) \sum_{t=t_0}^{T} \frac{1}{t-1} \right], \tag{33}$$

where the sum is interpreted as zero if $t_0 = T + 1$.

Let $Z_t := M_t - M_{t-1} = r_t - \mu_t$. Then $|Z_t| \le \|\theta^\star\|_2 \, B$ almost surely. Moreover,

$$\mathbb{E}[Z_t^2 \mid \mathcal{F}_{t-1}] = \mathrm{Var}(r_t \mid \mathcal{F}_{t-1}) \le \mathbb{E}[r_t^2 \mid \mathcal{F}_{t-1}] \le (\|\theta^\star\|_2 \, B)^2 X_{t-1}.$$

Thus, on E, the predictable quadratic variation satisfies

$$V_T := \sum_{t=1}^{T} \mathbb{E}[Z_t^2 \mid \mathcal{F}_{t-1}] \le (\|\theta^\star\|_2 \, B)^2 \left[(t_0 - 1) + 2\Big(d + \log \frac{2T}{\delta}\Big) \sum_{t=t_0}^{T} \frac{1}{t-1} \right]. \tag{34}$$

Set

$$H_T := (t_0 - 1) + 2\Big(d + \log \frac{2T}{\delta}\Big) \sum_{t=t_0}^{T} \frac{1}{t-1}.$$

Applying Lemma A.7 with confidence $\delta/2$, $b = \|\theta^\star\|_2 B$, and $v = (\|\theta^\star\|_2 B)^2 H_T$, gives

$$\mathbb{P}\left(M_T > \|\theta^\star\|_2 B \sqrt{2 H_T \log \frac{2}{\delta}} + \frac{2}{3} \|\theta^\star\|_2 B \log \frac{2}{\delta} \ \text{ and } \ V_T \le (\|\theta^\star\|_2 B)^2 H_T \right) \le \frac{\delta}{2}.$$

Since, by (34), the event E implies $V_T \le (\|\theta^\star\|_2 B)^2 H_T$, we have

$$\mathbb{P}\left(E \cap \left\{ M_T > \|\theta^\star\|_2 B \sqrt{2 H_T \log \frac{2}{\delta}} + \frac{2}{3} \|\theta^\star\|_2 B \log \frac{2}{\delta} \right\} \right) \le \frac{\delta}{2}.$$

Since also $\mathbb{P}(E^c) \le \delta/2$, a union bound gives that, with probability at least $1 - \delta$, the event E holds and

$$M_T \le \|\theta^\star\|_2 B \sqrt{2 H_T \log \frac{2}{\delta}} + \frac{2}{3} \|\theta^\star\|_2 B \log \frac{2}{\delta}.$$

On this event, combining (31), (33), and (34) gives

$$R_T \le \|\theta^\star\|_2 B \left[(t_0 - 1) + 2\Big(d + \log\frac{2T}{\delta}\Big) \sum_{t=t_0}^{T} \frac{1}{t-1} \right] + \|\theta^\star\|_2 B \sqrt{2\left[(t_0 - 1) + 2\Big(d + \log\frac{2T}{\delta}\Big) \sum_{t=t_0}^{T} \frac{1}{t-1} \right] \log\frac{2}{\delta}} + \frac{2}{3} \|\theta^\star\|_2 B \, \log\frac{2}{\delta}. \tag{35}$$

Finally, since

$$t_0 - 1 = \mathcal{O}\Big(d + \log\frac{T}{\delta}\Big) \qquad \text{and} \qquad \sum_{t=t_0}^{T} \frac{1}{t-1} \le 1 + \log\frac{T}{d},$$

we obtain

$$R_T = \mathcal{O}\left(\|\theta^\star\|_2 \, B \, \left(d + \log\frac{T}{\delta}\right) \left(1 + \log\frac{T}{d}\right) \right),$$

which proves (29). □

B.6 Proof of Proposition 3.1

The proof is a direct reduction to realizable binary classification. Consider

$$A(s) = \{0, 1\}, \qquad \psi(s, 0) = 0, \qquad \psi(s, 1) = s.$$

Under a fixed tie-breaking rule,

$$a_\theta(s) = \mathbf{1}\{\langle \theta, s \rangle \ge 0\}.$$

The resulting expert-policy class is the class of homogeneous linear classifiers in $\mathbb{R}^d$, whose VC dimension is d. The standard realizable PAC lower bound applies to arbitrary-output, possibly randomized learners and gives sample complexity

$$\Omega\left(\frac{d + \log(1/\delta)}{\varepsilon}\right)$$

see Ehrenfeucht et al. (1989).

B.7 Proof of Proposition 3.5

Consider the IO instance constructed in Theorem 3.2. In this construction, $\theta^\star = (1, \mathbf{0})$ and $a_{\theta^\star}(s) = 0$ for every $s \in \mathcal{S}$. Moreover,

$$F_{\theta^\star}(s, a) = -2|a| + a,$$

so

$$F_{\theta^\star}(s, 0) - F_{\theta^\star}(s, 1) = 1, \qquad F_{\theta^\star}(s, 0) - F_{\theta^\star}(s, -1) = 3.$$

Fix a round $t \ge d + 1$ and condition on D_{t-1}. Then $\hat{\theta}^{\mathrm{sub}}_{t-1}$ is fixed. Write $\hat{\theta} := \hat{\theta}^{\mathrm{sub}}_{t-1} = (1, \hat{\theta}_{-1})$ and for a fresh state s, let $x = s^\top \hat{\theta}_{-1}$. From the construction in Theorem 3.2, the learned score is

$$F_{\hat{\theta}}(s, a) = -2|a| + a + ax.$$

Hence the unique greedy action under $\hat{\theta}$ is 1 when $x > 1$, is -1 when $x < -3$, and is 0 when $-3 < x < 1$. The boundary events $\{x = 1\}$ and $\{x = -3\}$ have $\mathbb{P}_{\mathcal{S}}$-measure zero, since $\mathbb{P}_{\mathcal{S}}$ is absolutely continuous on the sphere. Therefore, $\mathbb{P}_{\mathcal{S}}$-almost surely, the action-level mismatch event $\{a_{\hat{\theta}}(s) \ne a_{\theta^\star}(s)\}$ implies $a_{\hat{\theta}}(s) \in \{1, -1\}$. On this event, the regret is either 1 or 3, and hence is at least 1. Consequently,

$$r_t(s) \ge \mathbf{1}\Big\{ a_{\hat{\theta}^{\mathrm{sub}}_{t-1}}(s) \ne a_{\theta^\star}(s) \Big\} \qquad \mathbb{P}_{\mathcal{S}}\text{-a.s.}$$

Since $s_t \sim \mathbb{P}_{\mathcal{S}}$ independently of D_{t-1}, it follows that

$$\mathbb{E}[r_t \mid D_{t-1}] \ge \mathbb{P}_{\mathcal{S}}\Big(a_{\hat{\theta}^{\mathrm{sub}}_{t-1}}(s) \ne a_{\theta^\star}(s) \Big) = X_{t-1}. \tag{36}$$

Now apply Remark 3.4 with sample size $t-1$. For the constructed IO instance, and for every $J \in \mathcal{J}$,

$$\mathbb{P}_{D_{t-1}}(X_{t-1} > \varepsilon) \geq \sum_{i=0}^{d-1} \binom{t-1}{i} \varepsilon^i (1-\varepsilon)^{t-1-i}.$$

Combining this with (36) gives

$$\mathbb{P}_{D_{t-1}}(\mathbb{E}[r_t \mid D_{t-1}] \geq \varepsilon) \geq \sum_{i=0}^{d-1} \binom{t-1}{i} \varepsilon^i (1-\varepsilon)^{t-1-i}.$$

Taking the infimum over $J \in \mathcal{J}$ and then the supremum over $(\theta^\star, \mathbb{P}_{\mathcal{S}})$ proves the first claim.

We now lower bound the expected instantaneous regret. By (36),

$$\mathbb{E}[r_t] \geq \mathbb{E}_{D_{t-1}}[X_{t-1}].$$

Using the tail-integral formula and the action-level lower bound above,

$$\mathbb{E}_{D_{t-1}}[X_{t-1}] = \int_0^1 \mathbb{P}_{D_{t-1}}(X_{t-1} > \varepsilon)\, d\varepsilon \geq \sum_{i=0}^{d-1} \binom{t-1}{i} \int_0^1 \varepsilon^i (1-\varepsilon)^{t-1-i}\, d\varepsilon.$$

The integral is a Beta integral:

$$\int_0^1 \varepsilon^i (1-\varepsilon)^{t-1-i}\, d\varepsilon = \frac{i!(t-1-i)!}{t!}.$$

Therefore,

$$\binom{t-1}{i} \frac{i!(t-1-i)!}{t!} = \frac{1}{t}.$$

Summing over $i = 0, \ldots, d-1$, we obtain

$$\mathbb{E}_{D_{t-1}}[X_{t-1}] \geq \frac{d}{t}.$$

Hence, for every $t \geq d+1$,

$$\mathbb{E}[r_t] \geq \frac{d}{t}.$$

Finally, since regret is nonnegative,

$$\mathbb{E}[R_T] = \sum_{t=1}^{T} \mathbb{E}[r_t] \geq \sum_{t=d+1}^{T} \mathbb{E}[r_t] \geq d \sum_{t=d+1}^{T} \frac{1}{t}.$$

Moreover,

$$\sum_{t=d+1}^{T} \frac{1}{t} \geq \int_{d+1}^{T+1} \frac{dx}{x} = \log \frac{T+1}{d+1}.$$

Therefore,

$$\mathbb{E}[R_T] \geq d \log \frac{T+1}{d+1}.$$

Since the constructed IO instance satisfies the above lower bounds for every $J \in \mathcal{J}$, taking the infimum over $J \in \mathcal{J}$ and then the supremum over $(\theta^\star, \mathbb{P}_{\mathcal{S}})$ gives the claimed result.