Active Learning by Prediction Uncertainty: Applications in Contextual Stochastic Linear Optimization

SUBMISSION 1653

We design a new active learning algorithm for regression problems, accommodating a wide range of loss functions beyond the conventional squared loss. In this setting, unlabeled samples arrive sequentially and we need to decide whether to inquire about the label of each sample. The goal of active learning is to use a minimal number of labels to build a prediction model that achieves low losses according to a broadly defined loss function. Due to the potential complicated structures of the loss function, calculating the optimal label acquisition policy usually suffers from computational complexity. To address this challenge, we propose a new active learning algorithm that utilizes the prediction uncertainty within a certain confidence set of candidate predictors. Compared to previous active learning algorithms, our algorithm is computationally tractable and efficient by leveraging the local prediction uncertainty. We derive an upper bound for the label complexity of our algorithm, demonstrating its advantage over traditional supervised learning algorithms in the setting of contextual stochastic linear optimization (CSLO). Our numerical experiments show the practical value of our algorithm in the personalized treatment design problem.

Contents

Manuscript submitted for review to the 25th ACM Conference on Economics & Computation (EC'24).

1 INTRODUCTION

When building a general prediction model that predicts the label based on the feature of samples, the accuracy of the prediction is closely related to the size of the training set. The more observations we have in the training set, the more accurate the prediction model should be. However, in practice, data collection can be a time-consuming and expensive process. In some cases, although feature vectors are usually cheap to acquire, acquiring the label of one sample could take some unignorable cost or time. For example, in the inventory problem, in order to observe the demand for a new product, we need to spend some days observing sales and keep inventory levels at a high enough level to avoid losing sales. In the example of clinical trials, there exist time, ethical, and economic costs to recruit volunteers to test the effect of drugs. Thus, given a large number of unlabeled samples, the cost of acquiring the labels of all features could be prohibitively large. As a consequence, how to efficiently collect the labels of features in order to build a good prediction model is a critical question.

Algorithms that sequentially acquire the labels of samples for training a prediction model are in the area called active learning. In this literature, the number of required labeled samples when attaining the prediction model at a desired performance level is called *label complexity*. To reduce the label complexity, active learning algorithms focus on identifying the "informative" features. Intuitively, features with higher "informational value" will be labeled at a higher probability. By adapting the label probability to the "informational value" of each sample, the label complexity of the active learning algorithms is expected to be smaller than the sample complexity of the supervised learning.

However, the design of the active label acquisition policy depends on the loss function of interest. The "informational value" of each sample varies when the loss function changes. Most existing active learning literature for regression problems focuses on the squared loss. However, in practice, depending on the use of the prediction model, some other loss functions may be more interesting. For example, if the prediction model is used to predict the unknown parameters for some decision-making problem, then the decision loss from the decision-making problem is more interesting than the simple squared loss function. We provide a detailed example in Section 1.1.

When the loss function exhibits complicated structures, designing optimal label acquisition algorithms becomes computationally challenging. This complexity arises from the difficulty of integrating the structure of the loss function into the evaluation of each feature's informational value, rendering the process computationally intractable. For instance, in scenarios where a prediction model is used to estimate unknown parameters in decision-making problems, the decision loss can be nonconvex, making it difficult to incorporate into the label acquisition policy.

In designing active label acquisition policies, most existing literature on active learning for regression problems employs prediction uncertainty with the goal of minimizing squared loss. This prediction uncertainty typically represents the maximum ℓ_2 norm of the prediction difference for the predictors within a confidence set. Therefore, when dealing with general loss functions, an intriguing question arises: Can we directly use prediction uncertainty to estimate the 'information value' of a feature for general loss functions?

If the answer to the above question is yes, we can reduce the computational complexity involved in calculating the importance of one sample and avoid analyzing the complicated structure of the loss function.

In our paper, we propose a stream-based active learning algorithm for regression problems under the general loss function, where the sampling probability of each sample is based on the prediction uncertainty. This algorithm is particularly interesting for contextual stochastic linear optimization (CSLO), where the output of the prediction model is used as the parameters for a stochastic linear optimization problem. Compared to previous active learning methods, which require analyzing the structure of the linear programming, our algorithm has a smaller computational complexity by utilizing the local prediction uncertainty. Our contributions are summarized as follows.

- We propose an importance-weighted active learning algorithm for general loss functions. Our algorithm achieves computational tractability by assigning labeling probabilities based on prediction uncertainty.
- We derive the non-asymptotic risk bounds and label complexities for our proposed active label acquisition policy. Particularly, under the general loss function, we derive the bounds for the risk and label complexity regarding the general loss function.
- When applying our proposed active learning algorithm in the contextual stochastic linear optimization problem, we further derive the following theoretical guarantees for the decision loss:
  - When CSLO has general feasible regions, we derive an upper bound for the expected loss of the prediction model. Under certain conditions, we further derive the sublinear label complexity which demonstrates the advantage of our algorithm over supervised learning.
  - When the feasible region of CSLO is a strongly-convex set, we derive a sharper upper bound for the expected loss of the prediction model under some natural noise conditions.
  - When considering the low-noise conditions of the noise distribution in the CSLO, we derive a further smaller bound for the label complexity.
- Numerical experiments demonstrate that our active learning algorithm can reduce the size of the training set when achieving the same level of risk, compared to supervised learning. Besides, compared to the existing active learning algorithms, our active learning algorithm based on prediction uncertainty has a better performance.

1.1 Examples of application: Personalized Treatment Design

In this section, we provide an example of the application of the active learning algorithm in stochastic contextual programming. Stochastic contextual programming involves some unknown parameters. How to deal with these unknown parameters is the crux of obtaining a good decision and has been the focus of stochastic optimization studies over the decades. To address these unknown parameters, one common paradigm is *predict-then-optimize*, where we use the collected data to fit a model that predicts the unknown parameters in the downstream optimization problem. Next, we solve the downstream optimization problem supposing that the prediction is accurate. This predict-then-optimize framework allows us to provide personalized decisions for each sample based on its feature information.

Specifically, we focus on its application in personalized treatment planning for patients, where the careful design of drug doses is crucial, taking into account individual patient conditions. For instance, post-surgery recovery often involves administering multiple medications, such as antibiotics, analgesics, and anti-inflammatories. The appropriate dosages of these drugs depend on factors like the surgery's extent, the patient's age, gender, past medical history, and more.

The personalized treatment design problem can be formulated as follows:

$$\min_{w \in \mathbb{B}^{n,3}} : \mathbb{E}[c_N^T w - c_P^T w | x] \tag{1}$$

$$\text{subject to } Aw \leq B \tag{2}$$

$$w_{i,\text{low}} + w_{i,\text{median}} + w_{i,\text{high}} = 1, \forall i = 1, ..., n \tag{3}$$

The decision vector w represents three levels of dosage for each drug. In other words, for each drug i, the possible dosage are $w_{i,\text{low}}$, $w_{i,\text{median}}$, and $w_{i,\text{high}}$. The random vector $c_P \in \mathbb{R}^{n,3}$ represents the positive effect of each drug at each dosage level. Similarly, the random vector $c_N \in \mathbb{R}^{n,3}$ represents the negative effect of each drug at each level. For example, $c_{N,1,\text{high}}$ represents the side-effect on this patient for drug 1 with high volume. These two vectors c_P and c_N depend on the feature vector x of the patient. The objective when designing the therapy is to minimize the expected net cost, which is the expectation of the negative effect minus the positive effect.

The constraints $Aw \leq B$ represent some constraints between different drugs. For example, if Drug 1 and 2 cannot be set as high volume at the same time, then we have $w_{1,\text{high}} + w_{2,\text{high}} = 1$. If Drug 2 is high volume, then Drug 3 has to be high volume as well. This constraint can be written as $w_{3,\text{high}} \geq w_{2,\text{high}}$.

To solve the above problem and determine the personalized treatment, suppose that we have a prediction model that outputs effect vector c_N^T and c_P^T based on the feature x of customers. However, the prediction of c_N^T and c_P^T can be different from the actual effect. To enhance and refine the existing prediction model, we can ask doctors to collect patient feedback during the recovery process. Since obtaining feedback from all patients is impractical, the goal is to identify informational and representative patients.

Thus, we aim to design an active learning algorithm that helps doctors to identify representative patients and guides their decisions for tracking and adjusting the treatment accordingly.

It is worth noting that to observe the full vector of c_N and c_P, doctors do not have to test every level of doses. It is because, for one type of drug, the effects of different dosages on the same patient usually have a known relation. For example, for the antibiotics, we have that $c_{P,1,\text{high}} = 5\mathfrak{C}$, $c_{P,1,\text{median}} = 4\mathfrak{C}$, and $c_{P,1,\text{low}} = \mathfrak{C}$, for some $\mathfrak{C} > 0$. Thus, as long as doctors observe the effectiveness of one level, doctors are able to gauge the effectiveness of the other levels.

1.2 Related work

The active learning algorithm for regression problems has been studied in Castro et al. [2005], Sugiyama and Nakajima [2009], Cai et al. [2016], and Beygelzimer et al. [2009]. Beygelzimer et al. [2009] propose an importance-weighted algorithm (IWAL) that extends the disagreement-based methods in the classification problem to the regression problem. It is the first work that considers the general regression problems. In this work, they calculate the maximum difference between the loss of all possible labels as the “informational value” of one feature. However, when the loss function is not the squared ℓ_2 norm of the prediction errors, this algorithm becomes computationally challenging when calculating the importance weight for one feature.

As illustrated in Section 1.1, one of the motivations for considering the general loss function in active learning arises from the contextual stochastic linear optimization. In CSLO, when evaluating prediction models, instead of considering the prediction error, it is more natural to directly consider the cost of the decisions induced by the predicted parameters. This loss is called *Smart Predict-then-Optimize (SPO)* loss in [Elmachtoub and Grigas, 2022]. In this case, ideally, we aim to evaluate the uncertainty of one feature regarding the SPO loss, which is intricate due to its non-convexity and discontinuity.

To address the intractability of the SPO loss in active learning, [Liu et al., 2023] proposes a margin-based approach that is based on the concept of *distance to degeneracy*. They demonstrate the advantage of considering SPO loss in the active learning framework both theoretically and empirically. However, their approach relies on the accessibility of the margin information, which may be computationally challenging when the feasible region is complicated. Furthermore, their method disregards the difference between the informational values of the near-degeneracy samples.

Different from the approach in [Liu et al., 2023], our general active learning algorithm does not depend on the margin structure. This empowers our algorithm with computational advantages when the feasible region is non-polyhedron or complicated. Our theoretical and numerical results further demonstrate the advantage of our active learning algorithm.

2 IMPORTANCE-WEIGHTED ACTIVE LEARNING FOR VECTOR-VALUED REGRESSION

In this section, we introduce the importance-weighted active learning algorithm in the general setting of vector-valued regression. We first introduce the notation and preliminaries of vector-valued regression and then present our algorithm and its analysis.

2.1 Preliminary knowledge about active learning

Let $x \in \mathcal{X}$ denote a generic feature vector, where $\mathcal{X} \subseteq \mathbb{R}^p$ is the feature space. The label, or response, vector is denoted by $c \in C \subseteq \mathbb{R}^d$.[1] We assume there is a fixed but unknown distribution $\mathcal{D}$ over pairs (x, c) living in $\mathcal{X} \times C$.

Our goal is to learn a prediction function $h : \mathcal{X} \to \mathbb{R}^d$ that predicts the associated label based on a given feature vector. We assume that there is a hypothesis class $\mathcal{H}$ of predictor functions h and it is well-specified, whereby $\mathbb{E}_c[c|x] \in \mathcal{H}$. For simplicity of exposition, we assume that $\mathcal{H}$ is a finite class, and the cardinality of $\mathcal{H}$ is denoted by $|\mathcal{H}|$, but our analysis can be extended to classes with finite pseudo-dimension by standard covering arguments (See [Cortes et al., 2010, Wainwright, 2019] for examples). Throughout this section, we assume that the learning algorithm is based on specifying a loss function $\ell(\cdot, \cdot) : \mathbb{R}^d \times C \to \mathbb{R}_+$ for regression, where $\mathbb{R}_+$ denotes the nonnegative real space. For example, a common choice is the squared ℓ_2-norm, namely $\ell(\hat{c}, c) = \frac{1}{2}\|\hat{c} - c\|_2^2$. Given the choice of loss function, the (expected) risk of a predictor $h \in \mathcal{H}$ is defined by $R_\ell(h) := \mathbb{E}[\ell(h(x), c)]$, and the corresponding minimum risk value is $R_\ell^* := \min_{h \in \mathcal{H}} R_\ell(h)$. Given the predictor class $\mathcal{H}$, we use $\hat{C}$ to denote the space of the predicted labels, namely, $\hat{C} := \{c \in \mathbb{R}^d : c = h(x), \text{ for some } x \in \mathcal{X} \text{ and } h \in \mathcal{H}\}$. For loss function ℓ, we define the maximum loss by $\omega_\ell(\hat{C}, C) := \sup_{\hat{c}, c \in C} \ell(\hat{c}, c)$. We further define the bound of the label space by $\rho(C) := \max_{c \in C} \|c\|$.

The broad goal of active learning is to find a "good" predictor from the hypothesis class with a small number of labeled samples. Compared to the standard supervised learning that acquires the labels of all the samples before training the model, active learning algorithms reduce the label cost by choosing which samples to label sequentially and interactively. An active learner aims to use a small number of labeled samples to achieve a small prediction loss. In other words, for a given $\epsilon > 0$, the goal of an active learning method is to find a predictor $\hat{h}$ trained on the data with the minimal number of labeled samples, such that $R_\ell(\hat{h}) \leq R_\ell^* + \epsilon$ with high probability. The number of labels acquired is referred to as the label complexity.

2.2 Importance-Weighted Active Learning Algorithm Based on Prediction Uncertainty

Our importance-weighted active learning algorithm based on prediction uncertainty (IWAL-PU) is stated in Algorithm 1. The algorithm operates in a sequential learning environment where, at the beginning of each iteration t, we observe a feature vector x_t that follows the distribution $\mathcal{D}_\mathcal{X}$. The algorithm maintains a confidence set of predictors $H_t \subseteq \mathcal{H}$. Initially, the confidence set of the predictor is set as the entire label space $\mathcal{H}$. To obtain the labeling probability p_t of this feature x_t, we calculate the maximum ℓ_2 norm of prediction difference for the predictors within the current confidence set H_t. This difference is characterized by the ℓ_2 norm, which is denoted by $\|\cdot\|$. In this step, calculating the labeling probability is computationally easy, especially when the confidence set H_t contains a finite number of predictors. To decide whether to acquire the label of x_t, we

[1] Note that we frequently use the label terminology to be consistent with the active learning literature, and we use c to denote the label (instead of the more standard y) to be consistent with our application to the contextual stochastic linear optimization setting.

Algorithm 1 Importance-Weighted Active Learning Based on Prediction Uncertainty (IWAL-PU)

```
**Input:** A sequence of slackness $r_t$.
Set $W_0 \leftarrow \emptyset$, $n_0 \leftarrow 0$, $H_0 \leftarrow \mathcal{H}$.
**for** $t$ from 1,2,...,$T$ **do**
    Receive $x_t$.
    $p_t \leftarrow \frac{1}{\rho(C)} \max_{h_1,h_2 \in H_t} \{\|h_1(x_t) - h_2(x_t)\|\}$.
    Flip a coin $Q_t \in \{0,1\}$ with $\mathbb{P}(Q_t = 1) = p_t$.
    **if** $Q_t = 0$ **then**
        Do not request a label associated with $x_t$.
        Set $W_t \leftarrow W_{t-1}$, $n_t \leftarrow n_{t-1}$.
    **else**
        Request a label $c_t$ associated with $x_t$.
        Update $W_t \leftarrow W_{t-1} \cup \{(x_t, c_t, \frac{1}{p_t})\}$, $n_t \leftarrow n_{t-1} + 1$.
    **end if**
    Let $\hat{\ell}^t(h) \leftarrow \frac{1}{t} \sum_{(x_i, c_i, \frac{1}{p_i}) \in W_t} \frac{1}{p_i} \ell(h(x_i), c_i)$.
    Update $h_t \leftarrow \arg\min_{h \in H_{t-1}} \hat{\ell}^t(h)$ and $\hat{\ell}^{t,*} \leftarrow \min_{h \in H_{t-1}} \hat{\ell}^t(h)$.
    Update the confidence set of the predictor $H_t$ by $H_t \leftarrow \{h \in H_{t-1} : \hat{\ell}^t(h) \leq \hat{\ell}^{t,*} + r_t\}$.
**end for**
**Return** $h_T$.
```

flip a coin with probability of heads given by p_t. If the coin is head-up, then we acquire a label c_t drawn from the conditional distribution of c given x_t. Then, we add the sample (x_t, c_t) and its corresponding weight $\frac{1}{p_t}$ into the existing training set W_{t-1}. If the coin lands tails up, then we do not acquire a label for x_t. After each iteration, we update the predictor by minimizing the empirical re-weighted loss, which is defined as

$$\ell^{\mathsf{rew}}(h; (x_t, c_t, \tfrac{1}{p_t})) := \begin{cases} \frac{1}{p_t} \ell(h(x_t), c_t), & \text{if } x_t \text{ is labeled,} \\ 0, & \text{otherwise.} \end{cases}$$

In Algorithm 1, the random variables at iteration t are (x_t, c_t, p_t, Q_t), where the random variable $Q_t \in \{0,1\}$ represents the outcome of the coin flip that determines if we acquire the label of this sample or not. For simplicity, we use random variable $z_t \in \mathcal{Z} := \mathcal{X} \times C \times (0,1] \times \{0,1\}$ to denote the tuple of random variables $z_t := (x_t, c_t, p_t, Q_t)$. Thus, z_t depends on $z_1, \ldots, z_{t-1}$ and the classical convergence results for i.i.d. samples do not apply to our importance-weighted sampling algorithm. We define $\mathcal{F}_{t-1}$ as the σ-field of all random variables until the end of iteration $t-1$, i.e., $z_1, \ldots, z_{t-1}$. With slight abuse of notation, in Algorithm 1, the re-weighted loss function at iteration t can be rewritten as

$$\ell^{\mathsf{rew}}(h; z_t) = \frac{Q_t}{p_t} \ell(h(x_t), c_t).$$

Given the sampling probability q_t, Q_t is a Bernoulli random variable that is independent of other randomness and $\mathbb{E}[Q_t] = p_t$, the expectation of the re-weighted loss is

$$\mathbb{E}[\ell^{\mathsf{rew}}(h; z_t)] = \frac{\mathbb{E}[Q_t]}{p_t} \mathbb{E}[\ell(h(x_t), c_t)] = \mathbb{E}[\ell(h(x_t), c_t)] = \mathbb{E}[\ell(h(x), c)] = R_\ell(h).$$

This implies that the expectation of the re-weighted loss is an unbiased estimator for the risk of h. Thus, under appropriate regularity conditions that are guaranteed by our assumptions, the empirical re-weighted loss is expected to converge to the risk, and its minimizer h_t over the confidence set H_{t-1} is expected to converge to the true h^* that minimizes the risk.

In Algorithm 1, the confidence set of the predictor H_t is constructed by allowing a slackness r_t for the current re-weighted loss in line 16. By shrinking the value of r_t, we construct a sequence of nested sets $H_t \subseteq H_{t-1} \subseteq \ldots \subseteq H_0 = \mathcal{H}$ that get smaller as iterations go on. These smaller confidence sets imply that the maximum prediction error, p_t calculated in line 5, gets smaller as well. Since p_t is also the labeling probability, our algorithm becomes more selective as iterations go on.

When deciding the value of r_t, we need to consider the convergence rate of the loss function. Intuitively, on the one hand, we would like r_t to decrease quickly enough to reduce the label complexity by becoming more selective. On the other hand, r_t cannot shrink so quickly that the true predictor h^* fails to be included in the confidence set. In the next section, we provide an explicit form for r_t and analyze the label complexity of our algorithm.

2.3 Convergence and label complexity analysis

To guarantee the convergence of IWAL-PU (Algorithm 1), we make the following assumptions concerning the loss function and the joint distribution $\mathcal{D}$.

Assumption 1. *The loss function $\ell(\cdot,\cdot)$ and distribution $\mathcal{D}$ satisfy:*

*(1) **(Lipschitz property)** $\ell(\cdot, c)$ is L-Lipschitz for all $c \in C$, i.e., $|\ell(c_1, c) - \ell(c_2, c)| \le L\|c_1 - c_2\|$ for all $c \in C$ and $c_1, c_2 \in \hat{C}$.*

*(2) **(Local error bounds)** There exists a function $\phi(\cdot,\cdot) : \mathbb{R}_+ \times \mathcal{X} \to \mathbb{R}_+$, with $\phi(0, x) = 0$ and $\phi(\cdot, x)$ non-decreasing for all $x \in \mathcal{X}$, such that for any $x \in \mathcal{X}$ and $h \in \mathcal{H}$,*

$$R_\ell(h) - R_\ell^* \le \epsilon \Rightarrow \|h(x) - h^*(x)\| \le \phi(\epsilon, x).$$

Assumption 1.(1) is the Lipschitz property of the loss function. Assumption 1.(2) assumes that there exists a function $\phi(\epsilon, x)$ such that when the excess risk of h is small, the prediction error on x is small. We further use $\bar{\phi}$ to denote the uniform upper bound for $\phi(\epsilon, x)$ for all $x \in \mathcal{X}$, i.e., $\bar{\phi}(\epsilon) := \sup_{x\in\mathcal{X}} \phi(\epsilon, x)$. This function $\bar{\phi}$ was considered in [Liu et al., 2023]. Our function $\phi(\epsilon, x)$ in Assumption 1.(2) is expected to be smaller than $\bar{\phi}$ defined in [Liu et al., 2023]. Both functions $\bar{\phi}$ and ϕ can be used in the following analysis, while ϕ incorporates the dependence on the feature x, and is able to provide more insights on the label acquisition, which will be shown in Section 4.

Under Assumption 1, Theorem 1 specifies the slackness r_t and provides the excess risk for our IWAL-PU algorithm. It also provides an upper bound for the expected number of acquired labels after T iterations.

Theorem 1 (Label Complexity Guarantees for Vector-Valued Regression). *Suppose that Assumption 1 holds. Let $\delta \in (0, 1]$ be a given parameter, and set $r_t \leftarrow 2L\sqrt{\frac{\ln(2t|\mathcal{H}|/\delta)}{t}}$ for $t \ge 1$ and $r_0 \leftarrow 2\omega_\ell(\hat{C}, C)$. Then, the following guarantees hold simultaneously with probability at least $1-\delta$ for all $T \ge 1$:*

- *(a) The excess risk satisfies $R_\ell(h_T) - R_\ell^* \le 2r_T$,*
- *(b) The expectation of the number of labels acquired, $\mathbb{E}[n_T]$, deterministically satisfies $\mathbb{E}[n_T] \le \sum_{t=1}^{T} \sup_{x\in\mathcal{X}}\{\phi(2r_t, x)\} + \delta T$.*

The proof of Theorem 1 is provided in the next section. In Theorem 1, at iteration t, the slackness $r_t \le \tilde{O}(t^{-1/2})$, so part *(a)* indicates that the excess risk converges to zero at rate $\tilde{O}(T^{-1/2})$. The order of the upper bound for the number of expected labels, $\mathbb{E}[n_T]$ depends on the function $\phi(\epsilon, x)$, which further depends on the property of the loss function in Assumption 1.(2). In Example 1, we provide one simple illustrative example of $\phi(\epsilon, x)$ in the mean estimation problem when we use the squared ℓ_2 norm as the loss function.

Example 1 (Example of Single Dimension Mean estimation). *In this mean estimation problem, we decide whether to observe the outcome of a random variable $y \in \mathbb{R}$. Our goal is to predict the mean of y, which is denoted by $\bar{y}$. The loss function ℓ is simply the squared loss, i.e., $\ell(y_1, y_2) = (y_1 - y_2)^2$. We denote our estimation result of y by $\hat{y}$. We use the importance-weighting algorithm, Algorithm 1, to reduce the number of acquired labels when achieving a small prediction error $\ell(\hat{y}, \bar{y})$. At each iteration, if we decide to acquire the outcome of y, we observe a random outcome $y + \varepsilon$, where $\varepsilon \in \mathbb{R}$ is a noise term with zero mean. The label complexity in Theorem 1 depends on the function ϕ. We consider the forms of ϕ under two different noise conditions.*

In the first case, there is no additional assumption for ε. The excess risk for our prediction $\hat{y}$ is $|\hat{y} - \bar{y}|^2$, so the form of ϕ is simply $\phi(\epsilon, x) = \sqrt{\epsilon}$. By Theorem 1, setting $\delta \longleftarrow 1/T^2$, the order of the expected number of acquired labels $\mathbb{E}[n_T]$ is at most $\sum_{t=1}^{T} \tilde{O}(\sqrt{t^{-1/2}}) = \sum_{t=1}^{T} \tilde{O}(t^{-1/4}) \leq \tilde{O}(T^{3/4})$. Combining this order with the result in argument (a) in Theorem 1, we have that when we inquire $\bar{n}$ labels on average, the excess squared loss is at most $\tilde{O}(\bar{n}^{-\frac{1}{2}\cdot\frac{4}{3}}) = \tilde{O}(\bar{n}^{-\frac{2}{3}})$. This order is slower than $\tilde{O}(\bar{n}^{-1})$, the typical learning rate of the mean estimation, and the order of minimax lower bound under the squared loss. Next, we consider a special noise condition to demonstrate a smaller label complexity.

In the second case, we assume that the noise term ε are some random integers from $\{0, \pm 1, \pm 2, ..., \pm \Lambda\}$, where Λ is a positive integer. In this case, we can reduce label complexity by reducing the search space. For example, we can reduce the search space to the discrete grid of $y_1 + z$, where z is an integer. Thus, when $\hat{y} \neq \bar{y}$, we have that $|\hat{y} - \bar{y}| \geq 1$. Therefore, we have that $|\hat{y} - \bar{y}| \leq |\hat{y} - \bar{y}|^2$. Therefore, by the definition of ϕ, we have that $\phi(\epsilon, x) \leq \epsilon$. By Theorem 1, setting $\delta \longleftarrow 1/T^2$, the order of the expected number of acquired labels $\mathbb{E}[n_T]$ is at most $\sum_{t=1}^{T} \tilde{O}(t^{-1/2}) \leq \tilde{O}(T^{1/2})$. Combining this order with the result in argument (a) in Theorem 1, we have that when we inquire $\bar{n}$ labels on average, the excess squared loss is at most $\tilde{O}(1/\bar{n})$. This order is faster than the label complexity in the first case. It implies that the theoretical performance of our active learning algorithm depends on the noise distribution. Under certain noise conditions, our active learning algorithm can achieve a small label complexity. □

Example 1 shows that the theoretical performance of our algorithm depends on the function ϕ. When we have additional knowledge about noise distribution, the hypothesis class, and the structure of the loss function, we can derive a smaller function ϕ, which will lead to a smaller label complexity. Particularly, in Section 3, we demonstrate that under certain conditions, in the predict-then-optimize framework, the label complexity regarding the SPO risk of our active learning algorithm is smaller than the best known sample complexity for the predict-then-optimize framework.

2.3.1 *Proof of Theorem 1.*

In this section, we provide the proof of Theorem 1.

Recall that for all $t \geq 1$,

$$\mathbb{E}[\ell^{\mathsf{rew}}(h; z_t)|\mathcal{F}_{t-1}] = \mathbb{E}[\ell(h(x_t), c_t)] = R_\ell(h).$$

Consider the above applied to both $h \in \mathcal{H}$ and h^* and averaged over $i \in \{1, \dots, t\}$ to yield:

$$R_\ell(h) - R_\ell(h^*) \;=\; \frac{1}{t} \sum_{i=1}^{t} \left(\mathbb{E}[\ell^{\mathsf{rew}}(h; z_i)|\mathcal{F}_{i-1}] - \mathbb{E}[\ell^{\mathsf{rew}}(h^*; z_i)|\mathcal{F}_{i-1}]\right). \tag{4}$$

For any given $h \in H_{T-1}$, we denote the discrepancy between the conditional expectation and the realized excess re-weighted loss of predictor h at time t by Z_h^{t}, i.e., $Z_h^{\mathsf{t}} := \mathbb{E}[\ell^{\mathsf{rew}}(h; z_t) - \ell^{\mathsf{rew}}(h^*; z_t)|\mathcal{F}_{t-1}] - (\ell^{\mathsf{rew}}(h; z_t) - \ell^{\mathsf{rew}}(h^*; z_t))$. Since $Z_{h_T}^{\mathsf{t}}$ is bounded and $\mathbb{E}[Z_{h_T}^{\mathsf{t}}|\mathcal{F}_{t-1}] = 0$, we have that $\sum_{t=1}^{T} Z_{h_T}^{\mathsf{t}}$ is a martingale.

Thus, for any given $h \in H_{T-1}$, (4) is equivalently written as:

$$R_\ell(h) - R_\ell(h^*) \ = \ \frac{1}{t}\sum_{i=1}^{t} Z_h^{\mathtt{i}} + \frac{1}{t}\sum_{i=1}^{t}\left(\ell^{\mathtt{rew}}(h;z_i) - \ell^{\mathtt{rew}}(h^*;z_i)\right). \tag{5}$$

Before providing the proof of Theorem 1, we first show that the confidence set H_{T-1} contains the true optimal predictor h^* at each iteration if r_t satisfies some conditions in Lemma 1.

Lemma 1. *Given $T \geq 1$, if r_t satisfies that $\sup_{h\in H_{t-1}} \left|\frac{1}{t}\sum_{i=1}^{t} Z_h^{\mathtt{i}}\right| \leq r_t$, for any $t \leq T-1$, then we have $h^* \in H_{T-1}$.*

Proof of Lemma 1. Since $H_{T-1} \subseteq H_{T-2} \subseteq \ldots \subseteq H_0$, we prove Lemma 1 by induction. Obviously, we have that $h^* \in H_0 = \mathcal{H}$. Assume that $h^* \in H_t$ for all $t \leq T-2$. Next, we will show that $h^* \in H_{T-1}$.

Since $H_{T-1} = \{h \in H_{T-2} : \hat{\ell}^{T-1}(h) \leq \hat{\ell}^{T-1,*} + r_{T-1}\}$, to show $h^* \in H_{T-1}$, it suffices to show that $\frac{1}{T-1}\sum_{i=1}^{T-1} \ell^{\mathtt{rew}}(h^*;z_i) \leq \frac{1}{T-1}\sum_{i=1}^{T-1} \ell^{\mathtt{rew}}(h_{T-1};z_i) + r_{T-1}$.

Since $R_\ell(h_{T-1}) - R_\ell(h^*) \geq 0$, by (5), we have that

$$R_\ell(h_{T-1}) - R_\ell(h^*) = \frac{1}{T-1}\sum_{i=1}^{T-1} Z_{h_{T-1}}^{\mathtt{i}} + \frac{1}{T-1}\sum_{i=1}^{T-1}\left(\ell^{\mathtt{rew}}(h_{T-1};z_i) - \ell^{\mathtt{rew}}(h^*;z_i)\right) \geq 0.$$

Since $h_{T-1} \in H_{T-2}$, by the condition in Lemma 1, we have $\frac{1}{T-1}\sum_{i=1}^{T-1} Z_{h_{T-1}}^{\mathtt{i}} \leq r_{T-1}$. Therefore, we obtain that

$$\frac{1}{T-1}\sum_{i=1}^{T-1} \ell^{\mathtt{rew}}(h^*;z_i) \leq \frac{1}{T-1}\sum_{i=1}^{T-1} \ell^{\mathtt{rew}}(h_{T-1};z_i) + r_{T-1}.$$

Thus, we obtain $h^* \in H_{T-1}$. □

We are ready to provide the proof of Theorem 1.

Proof of Theorem 1. We provide the proof for each part separately.

Part *(a)*. Let us first prove part *(a)*. Indeed, we prove a stronger version of part *(a)*: with probability $1-\delta$, for all $T \geq 1$, we have

$$(A)\text{: For any } h \in H_{T-1}\text{, we have } R_\ell(h) - R_\ell^* \leq 2r_T \text{ and } \sup_{h\in H_{T-1}} \left|\frac{1}{T}\sum_{t=1}^{T} Z_h^{\mathtt{t}}\right| \leq r_T.$$

We prove Claim (A) by the chain rule of probability. When $T = 0$, part *(a)* holds by the definition of $r_0 \geq 2\omega_\ell(\hat{C}, C)$. Next, we assume that Claim (A) holds for all $t \leq T-1$ and our goal is to show that Claim *(A)* holds for T with probability at least $1 - \frac{\delta}{2T^2}$. In other words, for any $h \in H_{T-1}$, we have $R_\ell(h) - R_\ell^* \leq 2r_T$ and $\sup_{h\in H_{T-1}} \left|\frac{1}{T}\sum_{t=1}^{T} Z_h^{\mathtt{t}}\right| \leq r_T$ with probability at least $1 - \frac{\delta}{2T^2}$. If it is true, by the chain rule of probability and taking the union bound over $T \in \{1, 2, \ldots, \infty\}$, we will have that Claim *(A)* holds for all $T \geq 1$ with probability at least

$$1 - \sum_{T=1}^{\infty} \frac{\delta}{2T^2} \geq 1 - \frac{\delta\pi^2}{12} \geq 1 - \delta.$$

Thus, we will obtain Claim (A).

The rest of the proof of part *(a)* is to show that Claim *(A)* holds for T with probability at least $1 - \frac{\delta}{2T^2}$.

Given any $h \in H_{T-1} \subseteq H_{T-2}$, since Claim *(A)* holds for $t \leq T-1$, we have that $R_\ell(h) - R_\ell(h^*) \leq 2r_{T-1}$. Thus, by the definition of ϕ, we have that for any $x \in \mathcal{X}$, $\|h(x) - h^*(x)\| \leq \phi(2r_{T-1}, x)$. By the Lipschitz property of $\ell(\cdot, c)$ and the decreasing property of r_t, we have that for any $i \leq T$,

$$
\begin{aligned}
|\ell^{\mathsf{rew}}(h; z_i) - \ell^{\mathsf{rew}}(h^*; z_i)| &= \left| \frac{q_i(\ell(h(x_i), c_i) - \ell(h^*(x_i), c_i))}{\phi(2r_{i-1}, x_i)} \right| \\
&\leq \left| \frac{q_i(\ell(h(x_i), c_i) - \ell(h^*(x_i), c_i))}{\phi(2r_{T-1}, x_i)} \right| \\
&\leq \frac{L\|h(x_i) - h^*(x_i)\|}{\phi(2r_{T-1}, x_i)} \leq L.
\end{aligned}
$$

It implies that the loss function $\ell^{\mathsf{rew}}(h; z_t) - \ell^{\mathsf{rew}}(h^*; z_t)$ is upper bounded by L, so we can apply Azuma's inequality to the sequence $\sum_{t=1}^T Z_h^{\mathrm{t}}$. By taking the average of Z_h^{t}, we have that $\left|\frac{1}{T}\sum_{t=1}^T Z_h^{\mathrm{t}}\right| \leq \epsilon$, with probability at least $1 - 2e^{-\frac{\epsilon^2 T}{2L^2}}$.

By setting the probability $1 - 2e^{-\frac{\epsilon^2 T}{2L^2}} = 1 - \frac{\delta}{2T^2|\mathcal{H}|^2}$, we obtain that $\epsilon \leq 2L\sqrt{\frac{\ln(2T|\mathcal{H}|/\delta)}{T}} = r_T$. By applying the union bound over all $h \in H_{T-1} \subseteq \mathcal{H}$, and all $h^* \in \mathcal{H}$, we have that

$$
\sup_{h \in H_{T-1}} \left| \frac{1}{T} \sum_{t=1}^{T} Z_h^{\mathrm{t}} \right| \leq \epsilon = r_T,
$$

with probability at least

$$
1 - |\mathcal{H}|^2 \cdot \frac{\delta}{2T^2|\mathcal{H}|^2} \geq 1 - \frac{\delta}{2T^2}.
$$

Next, conditioning on the occurrence of this event, we will prove that $R_\ell(h) - R_\ell(h^*) \leq 2r_T$. Since Claim (A) holds for all $t \leq T-1$, it implies that for any $t \leq T-1$, $\sup_{h \in H_{t-1}} \left|\frac{1}{t}\sum_{i=1}^T Z_h^{\mathrm{i}}\right| \leq r_t$. Thus, the condition in Lemma 1 holds, and by Lemma 1, we have that $h^* \in H_{T-1}$.

Since $h \in H_{T-1}$ and $h^* \in H_{T-1}$, we have that $\ell^{\mathsf{rew}}(h_T; z_t) - \ell^{\mathsf{rew}}(h^*; z_t) \leq r_T$ in (5) and we obtain that $R_\ell(h) - R_\ell(h^*) \leq 2r_T$ if $\left|\frac{1}{T}\sum_{t=1}^T Z_h^{\mathrm{t}}\right| \leq r_T$.

Therefore, we obtain that Claim *(A)* holds for T with probability at least $1 - \frac{\delta}{2T^2}$.

Part *(b)*. Lastly, we prove part *(b)*. Condition on the event in part *(a)*, the label probability at iteration t is at most

$$
\sup_{x \in \mathcal{X}} \phi(2r_t, x).
$$

Since part *(a)* happens with probability at least $1 - \delta$, the label probability at iteration t is at most $\sup_{x \in \mathcal{X}}\{\phi(2r_t, x)\} + \delta$. Thus, the expectation of the total number of acquired labels is at most $\sum_{t=1}^T [\sup_{x \in \mathcal{X}}\{\phi(2r_t, x)\} + \delta] = \sum_{t=1}^T \sup_{x \in \mathcal{X}}\{\phi(2r_t, x)\} + \delta T$. □

3 SMALL LABEL COMPLEXITY FOR CONTEXTUAL STOCHASTIC LINEAR OPTIMIZATION

In this section, we consider our importance-weighted active learning method for the contextual stochastic linear optimization problem, which is also referred to as the predict-then-optimize framework. We first present the motivation of applying this algorithm in the predict-then-optimize framework in Section 3.2. Next, we analyze the convergence rate and label complexity of IWAL-PU in Section 3.3. Finally, we demonstrate how to improve these bounds when the feasible region is a strongly convex set or a polyhedron in Sections 3.4 and 3.5. These results demonstrate a smaller label complexity than the naive supervised learning approach that acquires labels of all samples.

3.1 Preliminaries

Here, we introduce the preliminary knowledge about the predict-then-optimize framework. In particular, we introduce the SPO loss function and the surrogate loss for the SPO loss.

In this setting, label vectors to be predicted are the parameters in the downstream optimization problem. Particularly, the downstream optimization problem has a linear objective, and the label vector is the cost vector of the objective. Given the prediction of the cost vector, the deterministic linear optimization is solved to make a decision. Thus, acquiring a “label" corresponds to collecting the cost vector data c that corresponds to a given feature vector x.

Let $w \in S$ denote the decision variable of the downstream optimization problem, where the feasible region $S \subseteq \mathbb{R}^d$ is a convex and compact set that is assumed to be fully known to the decision-maker. We denote the diameter of the set $S \subset \mathbb{R}^d$ by $D_S := \sup_{w,w' \in S} \|w - w'\|$. Given an observed feature vector x, the ultimate goal is to solve the contextual stochastic optimization problem:

$$\min_{w \in S} \mathbb{E}_c[c^T w|x] = \min_{w \in S} \mathbb{E}_c[c|x]^T w. \tag{6}$$

(6) indicates that the downstream optimization problem in the predict-then-optimize framework relies on a prediction of the conditional expectation $\mathbb{E}_c[c|x]$. Given such a prediction $\hat{c}$, a decision is made by then solving the deterministic version of the downstream optimization problem:

$$P(\hat{c}) : \quad \min_{w \in S} \hat{c}^T w. \tag{7}$$

For simplicity, we assume $w^*(\cdot) : \mathbb{R}^d \to S$ is an oracle for solving (7), whereby $w^*(\hat{c})$ is an optimal solution of $P(\hat{c})$.

In the predict-then-optimize framework, given the prediction $h(x)$ for any newly observed feature vector x, our decision is $w^*(h(x))$. The ultimate goal of our active learning algorithms is to select $h \in \mathcal{H}$ that can lead to optimal decisions. In the predict-then-optimize setting, the loss of prediction in terms of the downstream decision making is the SPO loss introduced by Elmachtoub and Grigas [2022], which characterizes the regret in decision error due to an incorrect prediction, and is formally defined as

$$\ell_{\mathrm{SPO}}(\hat{c}, c) := c^T w^*(\hat{c}) - c^T w^*(c),$$

for any cost vector prediction $\hat{c}$ and realized cost vector c. We further define the risk of a prediction function $h(\cdot)$ as $R_{\mathrm{SPO}}(h) := \mathbb{E}_{(x,c)\sim\mathcal{D}}[\ell_{\mathrm{SPO}}(h(x), c)]$, and the excess risk of $h(\cdot)$ as $R_{\mathrm{SPO}}(h) - \inf_{h' \in \mathcal{H}} R_{\mathrm{SPO}}(h')$.

Since the SPO loss is non-convex and even non-continuous, instead of minimizing the SPO loss directly, a common approach is to consider surrogate loss functions ℓ that have better computational properties and are still (ideally) aligned with the original SPO loss, for example, the SPO+ loss introduced in [Elmachtoub and Grigas, 2022], which is defined by

$$\ell_{\mathrm{SPO+}}(\hat{c}, c) := \max_{w \in S} \left\{(c - 2\hat{c})^T w\right\} + 2\hat{c}^T w^*(c) - c^T w^*(c).$$

It is an upper bound on the SPO loss, i.e., $\ell_{\mathrm{SPO}}(\hat{c}, c) \leq \ell_{\mathrm{SPO+}}(\hat{c}, c)$ for any $\hat{c} \in \hat{C}$ and $c \in C$. Elmachtoub and Grigas [2022] demonstrate the computational tractability of the SPO+ surrogate loss, conditions for Fisher consistency of the SPO+ risk with respect to the true SPO risk, as well as strong numerical evidence of its good performance with respect to the downstream optimization task. Liu and Grigas [2021] further demonstrate sufficient conditions that imply that when the excess surrogate SPO+ risk of a prediction function h is small, the excess true SPO risk of a prediction function h is also small.

3.2 Motivation and algorithm

The motivation for considering the importance weighted active learning algorithm under the SPO loss is to reduce the label probability of samples that hold less 'importance' compared to others. Intuitively, when observing a particular feature x, we can estimate the potential SPO risk associated with the decision for that sample. If this SPO risk is small, we can strategically reduce the total label cost by assigning a lower probability to acquire the label of its sample. Lemma 2 below provides an upper bound for the excess SPO risk on a feature x.

Lemma 2 (Bounds for the Excess SPO Risk on a Single Point). *Suppose c is a random vector, and $\mathbb{E}[c] = \bar{c}$. For any $\hat{c} \in C$, we have*

$$\mathbb{E}[\ell_{\text{SPO}}(\hat{c}, c) - \ell_{\text{SPO}}(\bar{c}, c)] \leq \|\hat{c} - \bar{c}\| \|w^*(\hat{c}) - w^*(\bar{c})\|.$$

Proof of Lemma 2. By the definition of the SPO loss, we have that

$$\begin{aligned}\mathbb{E}[\ell_{\text{SPO}}(\hat{c}, c) - \ell_{\text{SPO}}(\bar{c}, c)] &= \mathbb{E}\left[c^T(w^*(\hat{c}) - w^*(c)) - c^T(w^*(\bar{c}) - w^*(c))\right] \\ &= \mathbb{E}\left[c^T(w^*(\hat{c}) - w^*(\bar{c}))\right] = \bar{c}^T(w^*(\hat{c}) - w^*(\bar{c})).\end{aligned}$$

Since $\hat{c}^T(w^*(\bar{c}) - w^*(\hat{c})) \geq 0$, we have that

$$\begin{aligned}\bar{c}^T(w^*(\hat{c}) - w^*(\bar{c})) &\leq \bar{c}^T(w^*(\hat{c}) - w^*(\bar{c})) + \hat{c}^T(w^*(\bar{c}) - w^*(\hat{c})) \\ &= (\bar{c} - \hat{c})^T(w^*(\hat{c}) - w^*(\bar{c})) \leq \|\hat{c} - \bar{c}\| \|w^*(\hat{c}) - w^*(\bar{c})\|.\end{aligned}$$

Thus, we obtain Lemma 2. □

Lemma 2 indicates that given a feature vector x, the excess SPO risk on this feature is bounded by the product of $\|h^*(x) - \hat{h}(x)\|$ and $\|w^*(h^*(x)) - w^*(\hat{h}(x))\|$. Obviously, when $w^*(h^*(x)) = w^*(\hat{h}(x))$, the decisions are the same, so the excess SPO risk is zero and we do not have to acquire the label of x. More interestingly, when the decisions are not the same, Lemma 2 indicates that the SPO risk can be upper bounded by the norm of the prediction error. This motivates us to use a weighted sampling approach based on the prediction error of $\hat{h}(x)$. Intuitively, when this prediction error is small, although we are not confident about picking the optimal decision, we do not have to acquire that label with a large probability, i.e., the label probability of that sample can still be small. Thus, considering the prediction error could help us save the label complexity, especially when the current prediction error is not small enough to distinguish the optimal decision from the suboptimal.

When applying Algorithm 1 in the predict-then-optimize framework, using the SPO loss directly is computationally challenging due to its non-convexity. Thus, we use a surrogate loss of SPO loss, which is continuous and convex. Assumption 2 requires the consistency of the surrogate loss, which is a basic requirement for choosing the surrogate loss function.

Assumption 2. ***(Consistency)*** *The unique minimizer of the SPO risk h^* is also the unique minimizer of the surrogate risk, and $h^*(x) = \mathbb{E}[c|x]$ for all $x \in \mathcal{X}$.*

Among all possible convex surrogate loss functions, the SPO+ loss enjoys some merits. First, the SPO+ loss satisfies Assumption 2 under certain noise conditions, which is studied in [Elmachtoub and Grigas, 2022]. Second, the function ϕ can have a nice form when the distribution $\mathcal{D}$ satisfies some conditions in Section 4. Thirdly, compared to the squared loss, the SPO+ loss incorporates information from the downstream decision-making problem. In the following sections, we analyze the label complexity regarding the SPO risk when the surrogate loss function satisfies Assumptions 1 and 2.

3.3 Small label complexity for SPO risk

In this section, we analyze the label complexity of IWAL-PU regarding the SPO risk. In particular, based on the risk guarantees in Theorem 1, when the surrogate loss further satisfies the consistency condition, then Theorem 2 provides the bound for the SPO risk after T iterations. To analyze the bound for the SPO risk, we assume that the marginal distribution density of x, $\mathcal{D}_{\mathcal{X}}$ is known, which is denoted by $\mu(x)$.

Theorem 2 (SPO risk bounds for IWAL-PU). *Suppose that Assumptions 1 and 2 hold. Set* $r_t \leftarrow 2L\sqrt{\frac{\ln(2t|\mathcal{H}|/\delta)}{t}}$ *for* $t \geq 1$ *and* $r_0 \leftarrow 2\omega_\ell(\hat{C}, C)$. *Then, with probability at least* $1 - \delta$, *for all* $T \geq 1$, *the events in Theorem 1 holds and the excess SPO risk satisfies* $R_{\text{SPO}}(h_T) - R^*_{\text{SPO}} \leq D_S \int_{x \in \mathcal{X}} \mu(x)\phi(2r_T, x)dx$.

In Theorem 2, the order of SPO risk depends on function ϕ. As will be shown later in Section 4, the function $\phi(\epsilon, x)$ is a square root function of ϵ when the distribution $\mathcal{D}$ satisfies some natural conditions and the surrogate loss is SPO+. Thus, next in Proposition 1, by substituting $\phi(\epsilon, x)$ with a square root function, we provide the order of the label complexity and the risk bounds.

Proof of Theorem 2. By Lemma 2 and the definition of D_S, we have that

$\mathbb{E}[\ell_{\text{SPO}}(h_T(x), c) - \ell_{\text{SPO}}(h^*(x), c)] \leq \|h_T(x) - h^*(x)\| \|w^*(h_T(x)) - w^*(h^*(x))\| \leq \|h_T(x) - h^*(x)\| D_S.$

By part *(a)*, we have $\|h_T(x) - h^*(x)\| \leq \phi(2r_T, x)$ for any $x \in \mathcal{X}$. Thus, by taking the expectation of x over density function $\mu(x)$, we have

$$R_{\text{SPO}}(h_T) - R^*_{\text{SPO}} \leq D_S \int_{x \in \mathcal{X}} \mu(x)\phi(2r_T, x)dx,$$

which is the result of Theorem 2. □

Proposition 1. *Under the same assumptions of Theorem 2, suppose that there exists a constant* $C' > 0$, *such that* $\phi(\epsilon, x) \leq C'\sqrt{\epsilon}$. *Then, the following guarantees hold simultaneously with probability at least* $1 - \delta$ *for all* $T \geq 1$:

- *(a) The excess surrogate risk satisfies* $R_{\text{SPO+}}(h_T) - R^*_{\text{SPO+}} \leq \tilde{O}(T^{-1/2})$,
- *(b) The excess SPO risk satisfies* $R_{\text{SPO}}(h_T) - R^*_{\text{SPO}} \leq \tilde{O}(T^{-1/4})$,
- *(c) The expectation of the number of labels acquired,* $\mathbb{E}[n_T]$, *deterministically satisfies* $\mathbb{E}[n_T] \leq \tilde{O}(T^{3/4})$, *when* $\delta \leq O(T^{-2})$.

Proof of Proposition 1. Since $r_T \leq \tilde{O}(T^{-1/2})$, the excess surrogate risk is at most $2r_T \leq \tilde{O}(T^{-1/2})$, which is the result of part *(a)*. Since $\phi(\epsilon, x) \leq C'\sqrt{\epsilon}$, the excess SPO risk is at most $D_S \int_{x \in \mathcal{X}} \mu(x)\phi(2r_T, x)dx \leq D_S C' \sqrt{2r_T} \leq \tilde{O}(T^{-1/4})$, which is the result of part *(b)*. When $\delta \leq O(T^{-2})$, the last term δT in part *(c)* in Theorem 2 is less than $\tilde{O}(1)$, so it suffices to focus on the first term. Since $\sum_{t=1}^T \sup_{x \in \mathcal{X}}\{\phi(2r_t, x)\} \leq C' \sum_{t=1}^T \sqrt{2r_t} \leq \sum_{t=1}^T \tilde{O}(t^{-1/4}) \leq \tilde{O}(T^{3/4})$. □

Proposition 1 provides the bounds for the surrogate risk, SPO risk, and the number of acquired labels. These results imply that when the expected number of acquired labels is $\bar{n}$, the excess surrogate risk is at most $\tilde{O}(\bar{n}^{-2/3})$, which is faster than the typical supervised learning whose risk is at most $\tilde{O}(\bar{n}^{-1/2})$. Besides, the excess SPO risk is at most $\tilde{O}(\bar{n}^{-1/3})$, which is also faster than the supervised learning bound of $\tilde{O}(\bar{n}^{-1/4})$ for the polyhedral case in [Liu and Grigas, 2021]. Note that the above results do not depend on any margin structure of the problem or the low-noise conditions.

Note that when using the SPO+ loss, the condition in Proposition 1, $\phi(\epsilon, x) \leq C'\sqrt{\epsilon}$, imposes additional conditions on the noise distribution. Examples of these noise conditions are provided in Section 4 later. Note that these additional conditions from the existence of ϕ make the label complexity of our algorithm smaller than the minimax lower bound of the sample complexity of the SPO risk provided in [Hu et al., 2022].

In Sections 3.4 and 3.5, we show that we can further improve the bound for the SPO risk when the feasible region S is a strongly-convex set or considering some low-noise conditions.

3.4 Refined bounds in strongly convex feasible regions

In this section, we provide refined bounds for the SPO risk in Theorem 2 where the feasible region is a strongly convex set. These results further reduce the label complexity in terms of the SPO risk in Proposition 1.

Definition 1 provides the definition of the strongly convex feasible region.

Definition 1 (Strongly convex feasible region). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a μ_S-strongly convex and L_S-smooth function for some $L_S \geq \mu_S > 0$. Suppose that the feasible region S is defined by $S = \{w \in \mathbb{R}^d : f(w) \leq r\}$ for some constant $r > f_{\min} := \min_w f(w)$.*

Proposition 2 (Refined bounds for the strongly convex feasible regions). *Let S be a strongly convex set with parameters L_S, μ_S and r. Suppose that the distribution of c satisfies that $\|c\| \geq c_{\min}$ almost surely for some $c_{\min} > 0$ and $\|h(x)\| \geq c_{\min}$ for all $x \in \mathcal{X}$. Then, under the same setting of Theorem 2 and Proposition 1, we have that for any $T \geq 1$, the excess SPO risk satisfies $R_{\text{SPO}}(h_T) - R^*_{\text{SPO}} \leq \tilde{O}(T^{-1/2})$.*

Together with Proposition 1, Proposition 2 implies that the error bound for the SPO risk in terms of the expected number of acquired labels $\bar{n}$ is $\tilde{O}(\bar{n}^{-2/3})$, which is faster than the convergence rate of the supervised learning, $\tilde{O}(\bar{n}^{-1/2})$, in [Liu and Grigas, 2021]. Thus, it demonstrates a further smaller label complexity than supervised learning.

3.5 Refined bounds by low-noise conditions

In this section, we provide a further refined bound for SPO risk under some low-noise conditions, which are also studied in [Liu et al., 2023]. To define the low-noise condition, we first review the definition of distance to degeneracy defined in [El Balghiti et al., 2022].

Definition 2. *(Distance to Degeneracy, [El Balghiti et al., 2022]). The set of degenerate cost vector predictions is $C^o := \{\hat{c} \in \mathbb{R}^d : P(\hat{c})$ has multiple optimal solutions$\}$. Given a norm $\|\cdot\|$ on $\mathbb{R}^d$, the distance to degeneracy of the prediction $\hat{c}$ is $\nu_S(\hat{c}) := \inf_{c \in C^o}\{\|c - \hat{c}\|\}$.* □

The distance to degeneracy can be easily computed in the case of a polyhedral feasible region with known extreme point representations. [El Balghiti et al., 2022] provide the exact formulas of the distance to degeneracy function in this case.

Next, to describe how the density of the distribution $\nu_S(h^*(x))$ is allocated near the points of degeneracy, we review the definition of the near-degeneracy function Ψ defined in [Liu et al., 2023] in Definition 3.

Definition 3 (Near-degeneracy function). *The near-degeneracy function $\Psi : \mathbb{R}_+ \to [0, 1]$ with respect to the distribution of $x \sim \mathcal{D}_\mathcal{X}$ is defined as:*

$$\Psi(b) := \mathbb{P}\left(\inf_{h^* \in \mathcal{H}^*}\{\nu_S(h^*(x))\} \leq b\right).$$

□

The near-degeneracy function Ψ measures the probability that the distance to degeneracy of $h^*(x)$ is smaller than b, when x follows the marginal distribution of x in $\mathcal{D}_\mathcal{X}$. Based on the near-degeneracy function, we can obtain refined guarantees for IWAL-PU.

Theorem 3 (Refined Guarantees for IWAL-PU with near-degneracy function). *Suppose that Assumption 1 holds. Under the same setting of Theorem 2, the following guarantees hold simultaneously with probability at least $1-\delta$ for all $T \geq 1$:*

- *(a) The excess surrogate risk satisfies $R_\ell(h_T) - R_\ell^* \leq 2r_T$,*
- *(b) The excess SPO risk satisfies*

$$R_{\text{SPO}}(h_T) - R_{\text{SPO}}^* \leq D_S \Psi\left(\sup_{x\in\mathcal{X}}\{\phi(2r_t, x)\}\right)\int_{x\in\mathcal{X}} \mu(x)\phi(2r_T, x)dx,$$

- *(c) The expectation of the number of labels acquired, $\mathbb{E}[n_T]$, deterministically satisfies $\mathbb{E}[n_T] \leq \sum_{t=1}^{T}\sup_{x\in\mathcal{X}}\{\phi(2r_t, x)\} + \delta T$.*

Compared to the results in Theorem 2, the SPO risk bound in Theorem 3 has an additional multiplier $\Psi\left(\sup_{x\in\mathcal{X}}\{\phi(2r_t, x)\}\right)$. This demonstrates the influence of the near-degeneacy function on the SPO risk. Since $\Psi\left(\sup_{x\in\mathcal{X}}\{\phi(2r_t, x)\}\right)$ can be very small or even zero when r_t is small, this demonstrate an improvement over Theorem 2.

Next, we consider one example of the low-noise condition, which is characterized by the near-degeneracy condition, defined in Assumption 3.

Assumption 3 (Near-degeneracy condition). *There exist constants $b_0, \kappa > 0$ such that*

$$\Psi(b) = \mathbb{P}\left(\inf_{h^*\in\mathcal{H}^*}\{\nu_S(h^*(x))\} \leq b\right) \leq (b/b_0)^\kappa.$$

Assumption 3 controls the rate at which $\Psi(b)$ approaches 0 as b approaches 0. In other words, for small enough b so that $\frac{b}{b_0} < 1$, when the parameter κ is larger the probability near the degeneracy is smaller at a faster rate. Proposition 3 below further demonstrates the order of the SPO risk under the low noise condition.

Proposition 3 (Refined SPO risk bounds under low-noise conditions). *Suppose that Assumption 3 holds. Under the same setting of Theorem 2 and Proposition 1, we have that for any $T \geq 1$, the excess SPO risk satisfies $R_{\text{SPO}}(h_T) - R_{\text{SPO}}^* \leq \tilde{O}(T^{-\frac{\kappa+1}{4}})$.*

Proposition 3 indicates that the SPO risk bound by iteration t is at most $\tilde{O}(T^{-\frac{\kappa+1}{4}})$. In comparison, this SPO risk bound is smaller than the results of the margin-based approach in [Liu et al., 2023], where the rate is $\tilde{O}(T^{-\frac{\kappa}{4}})$. Their results do not require the well-specification of the hypothesis class, while our results need this assumption. Together with part *(c)* in Proposition 1, Proposition 3 indicates that the SPO risk in terms of the expectation of the acquired labels $\bar{n}$ is at most $\tilde{O}(\bar{n}^{-\frac{\kappa+1}{3}})$.

4 EXAMPLES OF $\phi(\epsilon, x)$

In the preceding sections, we assume the function ϕ is a squared root function. For example, in Propositions 1, 2, and 3, we assume $\phi(\epsilon, x) \leq C'\sqrt{\epsilon}$ for some $C' > 0$. In this section, we provide some examples of function $\phi(\epsilon, x)$.

Lemma 3 (General predictors). *Suppose that $\mathcal{X}$ contains finite support. If the distribution of $\mathcal{D}$ satisfies the following condition: there exists a constant $C_1 > 0$, such that for any $h \in \mathcal{H}$, $\mathbb{E}_x[\|h(x) - h^*(x)\|^2] \leq C_1 \cdot (R_\ell(h) - R_\ell(h^*))$. Then, we have that for any $x \in \mathcal{X}$ and any $h \in \mathcal{H}$,*

$$\phi(\epsilon, x) \leq \sqrt{\frac{C_1 \cdot \epsilon}{\mu(x)}}.$$

The condition of Lemma 3 holds for the common squared loss obviously with $C_1 = 1$. It also holds for the SPO+ loss under certain noise conditions. For example, in Appendix A in [Liu et al., 2023], various natural noise conditions for the existence of C_1 in Lemma 3 are studied when using SPO+ as the surrogate loss. Thus, under these same conditions, Lemma 3 provides an example of function $\phi(\epsilon, x)$, which implies that $\phi(\epsilon, x)$ is a square root function of ϵ.

Lemma 3 indicates that the function $\phi(\epsilon, x)$ becomes larger when the marginal distribution $\mu(x)$ gets larger. Next, when the true prediction model is linear, we can further utilize the directional information of x to derive function ϕ.

In this example, suppose that the true model is $c = \Theta^* x + \epsilon$, where the matrix $\Theta^* \mathbb{R}^{p\times d}$ is the true parameter in the linear model and $\epsilon \in \mathbb{R}^p$ is the noise vector with zero mean. Define the expectation $\Xi := \mathbb{E}[xx^T]$. Without loss of generality, we assume that Ξ is invertible such that all columns in feature vectors are independent of each other. Given a positive definite matrix M, we denote the weighted norm of a vector x by $\|x\|_M := \sqrt{x^T M x}$.

Lemma 4 (Linear predictors). *Suppose that the hypothesis class $\mathcal{H}$ is set of matrix $\Theta \in \mathbb{R}^{p\times d}$ and Ξ is invertible. If the distribution of $\mathcal{D}$ satisfies the following condition: there exists a constant $C_1 > 0$, such that for any $h \in \mathcal{H}$, $\mathbb{E}_x[\|h(x) - h^*(x)\|^2] \leq C_1 \cdot (R_\ell(h) - R_\ell(h^*))$. Then, we have that*

$$\phi(\epsilon, x) \leq \sqrt{C_1 \epsilon \cdot \|x\|_{\Xi^{-1}}}.$$

Lemma 4 indicates that the function ϕ is upper bounded by a squared root function of $\|x\|_{\Xi^{-1}}$. When the feature vector x is along with the direction of the eigenvector with the largest eigenvalue of Ξ^{-1}, $\|x\|_{\Xi^{-1}}$ is maximized. On the other hand, if the feature vector x is along with the direction of the eigenvector with the smallest eigenvalue of Ξ^{-1}, then the informational value $\phi(\epsilon, x)$ gets smaller. It implies that when x is in a similar direction to most feature vectors in the underlying distribution $\mathcal{D}$, the informational value $\phi(\epsilon, x)$ is small. These directional information are the typical factors to be considered in the fixed design problem of linear regression.

5 NUMERICAL EXPERIMENTS

In this section, we examine the empirical performance of our IWAL-PU algorithm using synthetic data. Particularly, we consider a personalized treatment design problem. In this setting, we have three different drugs and each drug has three possible dosage levels: low, medium, and high. To facilitate doctors to determine the best personalized treatment for each patient, we aim to build a prediction model that predicts the effect of three drugs on one patient based on the patient's feature information. Based on our predicted effect of three drugs at different levels, doctors can determine the best treatment to minimize the expected negative effect or maximize the positive effect.

We suppose that the following two rules must be followed when determining the best treatment: (1) The dosage of Drug 2 must be higher than the dosage of Drug 1. (2) At most one of these three drugs can be given at a high level. We assume that each patient has 5 different independent covariates, so the patients' feature vector is in the dimension of six, which includes the interception term. The cost vector to be predicted is the net effects of all drugs at all possible levels, so the dimension of the cost vector is 9.

The data generation process is as follows. We assume that the entries in feature vectors are integers between $[-2, 2]$. The distribution of the feature vectors $\mu(x)$ follows some mixed binomial distribution, where the center of each binomial distribution is generated randomly. We randomly generate a binary matrix $\Theta \in \mathbb{B}^{5\times 9}$ as the true model. To generate cost vector for each patient, we assume that $c = x\Theta + \epsilon$, where $\epsilon \in \mathbb{B}^9$ is a noise vector. We assume each entry in the noise vector ϵ is randomly drawn from $-1, 0$, and 1. The hypothesis class $\mathcal{H}$ contains all integer matrices in $\mathbb{R}^{12\times 5}$ whose entries are between $[-10, 10]$.

We use the SPO+ loss as the surrogate loss. To implement our IWAL-PU algorithm, we need to calculate the maximum prediction difference of $h(x)$ for all predictors h within the confidence class H_t. This exact calculation is challenging because our objective is to maximize the ℓ_2 norm under $t-1$ constraints. Thus, we adopt the following approximation method. We first relax the first $t-2$ constraints of the confidence set of H_t, i.e., H_t is the set of $\{h \in \mathcal{H} : \hat{\ell}^t(h) \leq \hat{\ell}^{t,*} + r_t\}$. Then, given the current predictor h_t, we use the local search idea. We randomly generate a noise coefficient matrix, where each entry is randomly drawn from -1, 0 and 1. Then, we add this noise matrix to the current predictor h_t to obtain a random predictor h'. If this predictor h' is within H_t, then we record the prediction $h'(x_t)$ in a possible prediction set. We repeat this random local exploration 100 times, and calculate the maximum distance between the vectors in the final possible prediction set.

Besides the above local approximation method, we can also utilize the insights from Section 4. Suppose we know the marginal distribution of feature $\mu(x)$, but we do not know the outcome in the test set. Then, we can use $c_1\sqrt{1/\mu(x)}$ and $c_2\sqrt{\|x\|_{\Xi^{-1}}}$ as the approximation of the maximum prediction error, where c_1 and c_2 are some parameters to be tuned. Using $\sqrt{1/\mu(x)}$ allows us to assign a larger sampling probability for the features with small probability to occur, while using $\sqrt{\|x\|_{\Xi^{-1}}}$ allows us to assign a smaller sampling probability for the features in the less stretched direction. After collecting the data set, to find the best predictor within the hypothesis class, we minimize the empirical SPO+ loss by the projected gradient descent method. After each update of the predictors, we round the value of each entry to the closest integer.

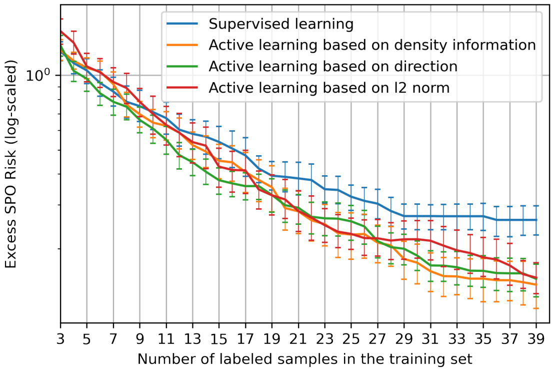

Fig. 1. Risk on the test set during the training process for various sample selection criteria.

The results of 25 independent trials are shown in Figure 1 with 90% confidence intervals. The x-axis is the number of acquired labels, while the y-axis is the excess SPO risk in the test set. Our IWAL-PU is the red curve in Figure 1. Compared to the supervised learning result in the blue curve, it shows that when acquiring the same number of labels in the training set, our IWAL-PU achieves a lower SPO risk in the test set than the supervised learning algorithm, which always acquires the labels of samples. We also test two variations of the importance weights by using the distribution information. The green curve utilizes the direction information of $\sqrt{\|x\|_{\Xi^{-1}}}$, while the orange curve uses the density information of $\sqrt{1/\mu(x)}$. These two curves demonstrate that these heuristic sampling probabilities can achieve a similar performance as the IWAL-PU approach.

Next, we compare the performance of SPO+ loss and the simple squared ℓ_2 norm loss. The simple squared ℓ_2 norm does not consider the downstream decision-making problem when training the model, while SPO+ loss incorporates the downstream decision-making problem. Figure 2 shows the distribution of the excess SPO risk on the test set when the number of acquired labels is 35, 36, and 37. It shows that although using the squared loss can have a smaller average risk, but the variance of the SPO risk is much higher the SPO+ loss. It demonstrates that using the SPO+ loss can reduce the variance of the risk for the downstream decision-making problem.

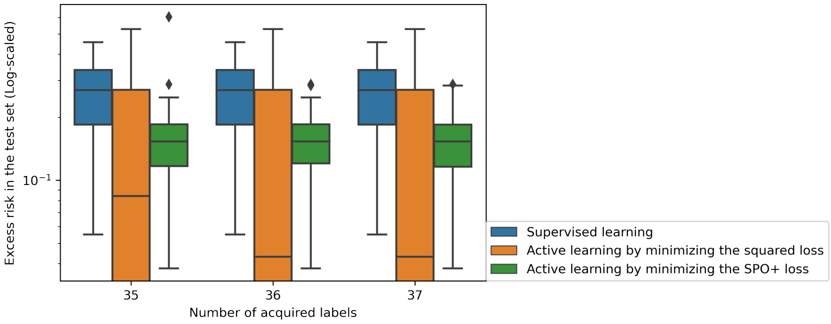

Fig. 2. Risk on the test set for different loss functions.

Finally, we compare our IWAL-PU algorithm with the traditional margin-based active learning approach. Figure 3 shows that our IWAL-PU has a smaller average risk than the margin-based approach.

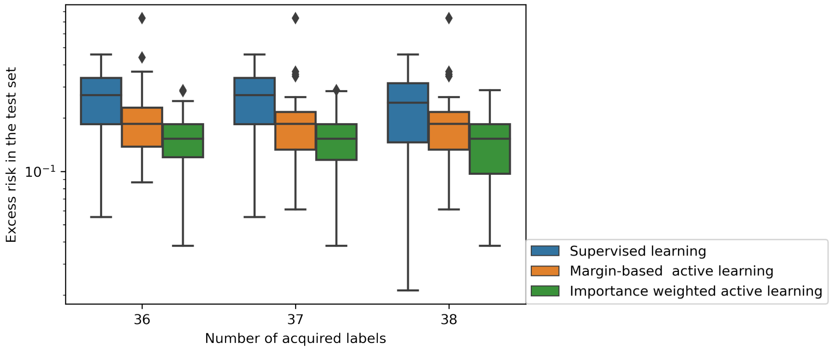

Fig. 3. Risk on the test set for margin-based approach.

The reason why IWAL-PU has a smaller risk than the margin-based approach is because of the structure of distribution. In our numerical experiments, when generating the data, the cost vectors are not far away from the degenerate cost vectors. Thus, the variation of distance to degeneracy provides less information about the potential SPO risk than the prediction error. As a consequence, using the distance to degeneracy as the sample selection approach has a slightly higher risk than our IWAL-PU approach.

6 CONCLUSIONS AND FUTURE DIRECTIONS

Our work develops a prediction uncertainty-weighted active learning algorithm for the regression problem under the general surrogate loss function. When applying this algorithm in the CSLO, we provide non-asymptotic bounds for the surrogate risk, SPO risk, and label complexity of our algorithm under various conditions. These bounds show that our IWAL-PU achieves a smaller label complexity under the supervised learning that acquires the labels of all samples. Our numerical results demonstrate that our IWAL-PU can reduce the size of the training set when achieving the same level of SPO risk, compared to the supervised learning and existing active learning algorithm.

There are several interesting future research directions: Our work assumes that the feature space $\mathcal{X}$ and the hypothesis class both have finite supports. In the future, it will be interesting to relax these assumptions and consider a more general setting with continuous feature space and infinite predictors in the hypothesis class. It is also worth studying some new active learning algorithms by integrating our prediction uncertainty-based approach into the existing margin-based approach.

REFERENCES

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. 2009. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*. 49–56.

Wenbin Cai, Muhan Zhang, and Ya Zhang. 2016. Batch mode active learning for regression with expected model change. *IEEE transactions on neural networks and learning systems* 28, 7 (2016), 1668–1681.

Rui Castro, Rebecca Willett, and Robert Nowak. 2005. Faster rates in regression via active learning. In *NIPS*, Vol. 18. 179–186.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. 2010. Learning bounds for importance weighting. *Advances in neural information processing systems* 23 (2010).

Othman El Balghiti, Adam N Elmachtoub, Paul Grigas, and Ambuj Tewari. 2022. Generalization bounds in the predict-then-optimize framework. *Mathematics of Operations Research* (2022).

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

Yichun Hu, Nathan Kallus, and Xiaojie Mao. 2022. Fast rates for contextual linear optimization. *Management Science* (2022).

Heyuan Liu and Paul Grigas. 2021. Risk bounds and calibration for a smart predict-then-optimize method. *Advances in Neural Information Processing Systems* 34 (2021).

Mo Liu, Paul Grigas, Heyuan Liu, and Zuo-Jun Max Shen. 2023. Active Learning in the Predict-then-Optimize Framework: A Margin-Based Approach. *arXiv preprint arXiv:2305.06584* (2023).

Masashi Sugiyama and Shinichi Nakajima. 2009. Pool-based active learning in approximate linear regression. *Machine Learning* 75, 3 (2009), 249–274.

Martin J Wainwright. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press.

A COMPLEMENTARY PROOFS

Proof of Proposition 2. Recall that in the proof of Theorem 2, we have that

$$\mathbb{E}[\ell_{\text{SPO}}(h_T(x), c) - \ell_{\text{SPO}}(h^*(x), c)] \leq \|h_T(x) - h^*(x)\| \|w^*(h_T(x)) - w^*(h^*(x))\|.$$

By Lemma 2 in [Liu and Grigas, 2021], we have that

$$\|w^*(h_T(x)) - w^*(h^*(x))\| \leq \frac{\sqrt{2L_s(r - f_{\min})}}{\mu_S} \left\| \frac{h_T(x)}{\|h_T(x)\|} - \frac{h^*(x)}{\|h^*(x)\|} \right\|.$$

Since $\|h\| \geq c_{\min}$, we further have that

$$\left\| \frac{h_T(x)}{\|h_T(x)\|} - \frac{h^*(x)}{\|h^*(x)\|} \right\| \leq \frac{1}{c_{\min}} \|h_T(x) - h^*(x)\|.$$

Thus, combining the above results, we have

$$\mathbb{E}[\ell_{\text{SPO}}(h_T(x), c) - \ell_{\text{SPO}}(h^*(x), c)] \leq \frac{\sqrt{2L_s(r - f_{\min})}}{\mu_S c_{\min}} \|h_T(x) - h^*(x)\|^2.$$

Since $\|h_T(x) - h^*(x)\| \le C'\sqrt{r_T} \le \tilde{O}(T^{-1/4})$, we have that $\mathbb{E}[\ell_{\text{SPO}}(h_T(x), c) - \ell_{\text{SPO}}(h^*(x), c)] \le \tilde{O}(T^{-1/2})$ for all $x \in \mathcal{X}$. Thus, taking the expectation over $x \sim \mu(x)$, we obtain that $R_{\text{SPO}}(h_T) - R^*_{\text{SPO}} \le \tilde{O}(T^{-1/2})$. □

Proof of Theorem 3. Parts *(a)* and *(c)* are the same results as Theorem 2, so the focus of the proof is to show part *(b)*. Recall that Lemma 2 indicates that

$$\mathbb{E}[\ell_{\text{SPO}}(h_T(x), c) - \ell_{\text{SPO}}(h^*(x), c)] \le \|h_T(x) - h^*(x)\| \|w^*(h_T(x)) - w^*(h^*(x))\|.$$

Thus, when $\|w^*(h_t(x)) - w^*(h^*(x))\| = 0$, the excess SPO risk at feature x is zero. Next, we show that when $\nu_S(h^*(x)) \ge 2\phi(2r_t, x_t)$, we have $w^*(h_t(x)) = w^*(h^*(x))$ and thus the SPO risk at x is zero.

Since ν_S is a 1-Lipschitz distance function, we have that $|\nu_S(h^*(x)) - \nu_S(h_t(x))| \le \|h_t(x) - h^*(x)\|$, which implies that

$$\nu_S(h_t(x_t)) \ge \nu_S(h^*(x_t)) - \|h_t(x_t) - h^*(x_t)\| \ge \nu_S(h^*(x_t)) - \phi(2r_t, x_t).$$

Thus, when $\nu_S(h^*(x)) \ge 2\phi(2r_t, x_t)$, we have that

$$\nu_S(h_t(x_t)) \ge \phi(2r_t, x_t) \ge \|h_t(x) - h^*(x)\|.$$

Thus, by Lemma 1 in [Liu et al., 2023], the condition that $\|h_t(x) - h^*(x)\| \le \max\{\nu_S(h_t(x)), \nu_S(h^*(x))\}$ is satisfied, and we have that $w^*(h_t(x)) = w^*(h^*(x))$. By the definition of near-degenearcy function, the probability that $\nu_S(h^*(x)) \le 2\phi(2r_t, x_t)$ is at most $\Psi(2\phi(2r_t, x_t))$. Thus, the risk at feature x is at most

$$\Psi(2\phi(2r_t, x_t))\|h_t(x) - h^*(x)\| + (1 - \Psi(2\phi(2r_t, x_t))) \cdot 0 = \Psi(2\phi(2r_t, x_t))\|h_t(x) - h^*(x)\|.$$

Since $\|h_t(x) - h^*(x)\| \le \phi(2r_t, x_t)$, by taking the expectation of x over probability density $\mu(x)$, we have

$$\begin{aligned} R_{\text{SPO}}(h_T) - R^*_{\text{SPO}} &\le D_S \int_{x \in \mathcal{X}} \mu(x) \Psi\left(\phi(2r_t, x)\right) \phi(2r_T, x) dx \\ &\le D_S \Psi\left(\sup_{x \in \mathcal{X}} \{\phi(2r_t, x)\}\right) \int_{x \in \mathcal{X}} \mu(x) \phi(2r_T, x) dx, \end{aligned}$$

which is the result in part *(b)* in Theorem 3. □

Proof of Proposition 3. By Theorem 3, the excess SPO risk is at most

$$R_{\text{SPO}}(h_T) - R^*_{\text{SPO}} \le D_S \Psi\left(\sup_{x \in \mathcal{X}} \{\phi(2r_t, x)\}\right) \int_{x \in \mathcal{X}} \mu(x) \phi(2r_T, x) dx.$$

Since $\sup_{x \in \mathcal{X}} \{\phi(2r_t, x)\} \le C'\sqrt{2r_t}$, $r_t \le \tilde{O}(t^{-1/2})$, we have that $\sup_{x \in \mathcal{X}} \{\phi(2r_t, x)\} \le \tilde{O}(t^{-1/4})$. Since $\Psi(\epsilon) \le \epsilon^\kappa$, we have that $\Psi\left(\sup_{x \in \mathcal{X}} \{\phi(2r_t, x)\}\right) \le \tilde{O}(t^{-\kappa/4})$.

Thus, we have that $R_{\text{SPO}}(h_T) - R^*_{\text{SPO}} \le \tilde{O}(t^{-\kappa/4}) \cdot \tilde{O}(t^{-1/4}) \le \tilde{O}(t^{-\frac{\kappa+1}{4}})$. □

Proof of Lemma 3. Since $\mathcal{X}$ contains finite supports and $\|h(x) - h^*(x)\|^2 \ge 0$ for all $x' \in \mathcal{X}$, we have that $\mathbb{E}[\|h(x) - h^*(x)\|^2] = \sum_{x' \in \mathcal{X}} \mu(x') \|h(x') - h^*(x')\|^2 \ge \mu(x) \|h(x) - h^*(x)\|^2$. Then, combining it with the condition in Lemma 3 will yield the result:

$$\|h(x) - h^*(x)\| \le \sqrt{\frac{C_1}{\mu(x)} \cdot (R_{\text{SPO+}}(h) - R_{\text{SPO+}}(h^*))}.$$

□

Proof of Lemma 4. We use β_i to denote the ith column in matrix Θ. By the definition of Ξ, we have that

$$\mathbb{E}[\|h(x) - h^*(x)\|^2] = \mathbb{E}\left[\left\|(\hat{\Theta} - \Theta^*)x\right\|^2\right] = \sum_{i=1}^{p} \|\hat{\beta}_i - \beta_i^*\|_{\Xi}.$$

By the Cauchy Schwarz inequality, we further have that

$$\|h(x) - h^*(x)\| = \sum_{i=1}^{p} \|(\hat{\beta}_i - \beta_i^*)x\| \leq \sum_{i=1}^{p} \sqrt{(\hat{\beta}_i - \beta_i^*)^T(\hat{\beta}_i - \beta_i^*) \cdot x^T x}.$$

Since $(\hat{\beta}_i - \beta_i^*)^T(\hat{\beta}_i - \beta_i^*) \cdot x^T x \leq \|\hat{\beta}_i - \beta_i^*\|_{\Xi} \cdot \|x\|_{\Xi^{-1}}$, we have that

$$\|h(x) - h^*(x)\| \leq \|x\|_{\Xi^{-1}}(\sum_{i=1}^{p} \|\hat{\beta}_i - \beta_i^*\|_{\Xi}) = \|x\|_{\Xi^{-1}} \cdot (\mathbb{E}[\|h(x) - h^*(x)\|^2]).$$

Combining the above results with $\mathbb{E}_x[\|h(x) - h^*(x)\|^2] \leq C_1 \cdot (R_\ell(h) - R_\ell(h^*))$, we can obtain the form of ϕ in Lemma 4. □