Active Learning in the Predict-then-Optimize Framework: A Margin-Based Approach

Mo Liu

INFORMS Workshop on Data Science

IEOR, University of California, Berkeley

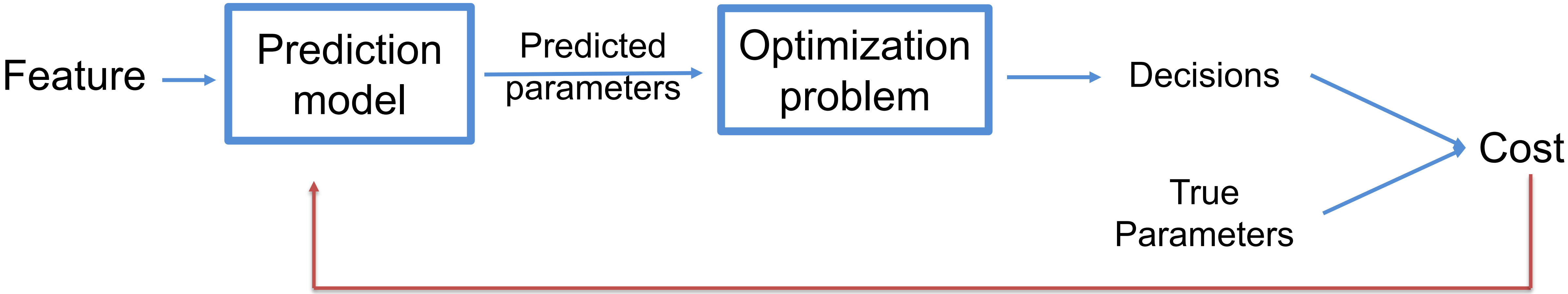
Joint work with Paul Grigas, Heyuan Liu and Zuo-Jun Max Shen

Build a prediction model to predict unknown parameters

Prediction	→	Decision
Customers' preference	→	Product recommendation
Demand	→	Inventory level
Price elasticity	→	Optimal prices

Predict-then-optimize framework

- Consider a stochastic optimization problem with unknown parameters:



Higher prediction accuracy ? ⇒ Lower cost

Data collection in predict-then-optimize framework

Realization of one sample: feature + parameters (labels of the samples)

😟 Acquiring the labels for one sample could be very expensive.

In a personalized pricing problem, to realize the purchase probability under all prices:

- Customer investigation
- Price trials

- How can we minimize the number of labels acquired while learning an effective prediction model?

- Select representative samples to acquire their labels

➢ Active learning + predict-then-optimize framework

Agenda

- Information gathering process for predict-then-optimize framework

➢ **Smart predict-then-optimize loss (SPO) and preliminaries**

- Theoretical motivation for margin-based algorithm
- Algorithm
- Analysis
- Numerical Experiments

Predict-then-optimize framework

- Optimization problem:

$$\min_{w \in S} \mathbb{E}_{c \sim D_X}[c^T w | x] = \min_{w \in S} \mathbb{E}_{c \sim D_X}[c^T | x] w$$

- Suppose we have access to the optimization oracle:

$$w^*(c) = \arg\min_{w \in S} c^T w$$

- SPO (Smart Predict-then-optimize) loss function:

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

- Predictor $h \in \mathcal{H}$
- SPO Risk:

$$R_{SPO}(h) = \mathbb{E}_{x,c \sim D}[\ell_{SPO}(h(x), c)]$$

- Best predictor:

$$h^* := \arg\min_{h \in \mathcal{H}} R_{SPO}(h)$$

Constructing the training set

- During the data collection process:
  - Feature x_i is readily available
  - Cost vector c_i is expensive
    - Large label cost
    - Time-consuming label process

- How can we minimize the number of labels acquired while achieving a small SPO risk?

Active learning

From iteration 1 to T, at each iteration t:

- Given one unlabeled sample with feature x_t from some unknown distribution
- Decide whether to acquire the label c_t of this sample x_t

– Goal: Use a small number of inquiries to achieve good performance

min: Number of acquired labels after T iterations

subject to: The final prediction model after T iterations has a good performance

- Label complexity: The minimum number of inquiries we need to make to attain performance at a given level

Goal of active learning

- Traditionally, active learning focuses on minimizing the prediction error

- In the predict-then-optimize framework:

Can we select samples to minimize the SPO loss directly, instead of the prediction error?
After T iterations, the prediction model h_T is obtained by using the selected training set

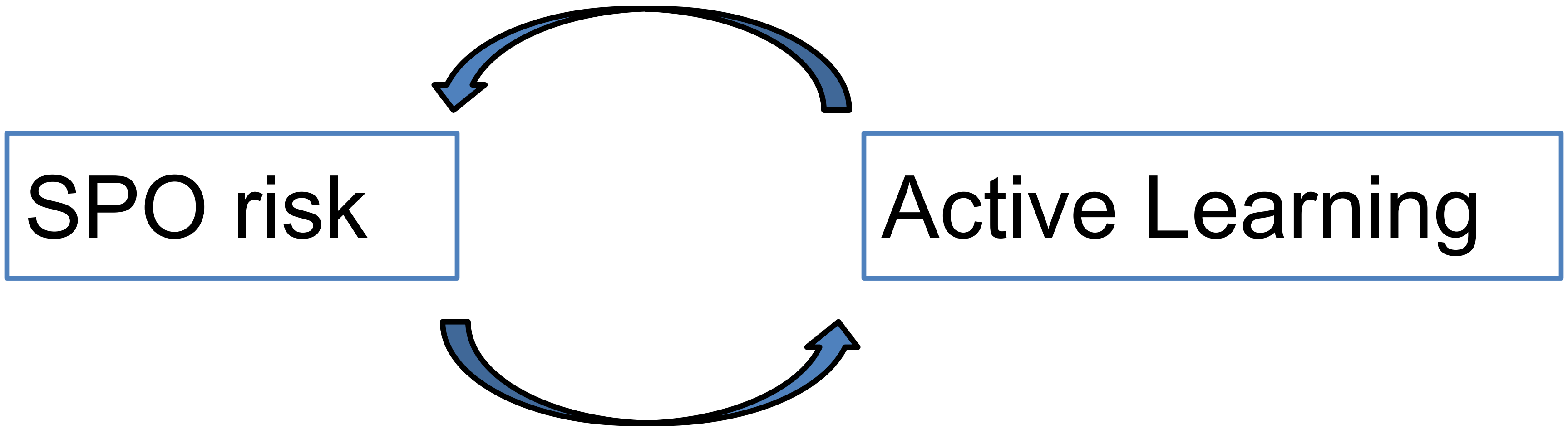
$$\begin{aligned} \min: &\quad \text{Number of acquired labels after } T \text{ iterations} \\ \text{subject to:} &\quad R_{SPO}(h_T) - R_{SPO}(h^*) \leq \epsilon \end{aligned}$$

Agenda

- Information gathering process for predict-then-optimize framework
- Smart predict-then-optimize loss (SPO) and preliminaries
- ➢ **Theoretical motivation for margin-based algorithm**
- Algorithm
- Analysis
- Numerical Experiments

Motivation in theory

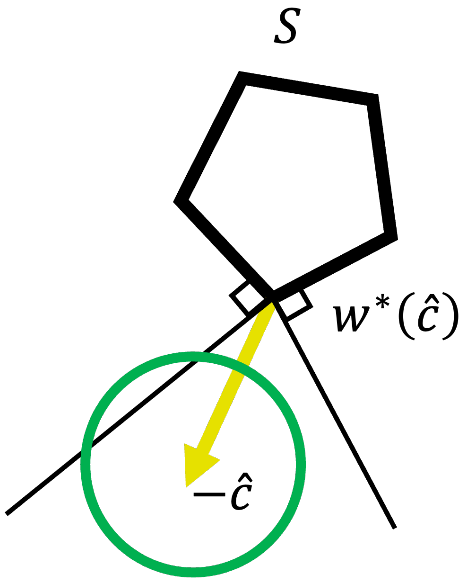
Active learning can help to minimize SPO loss



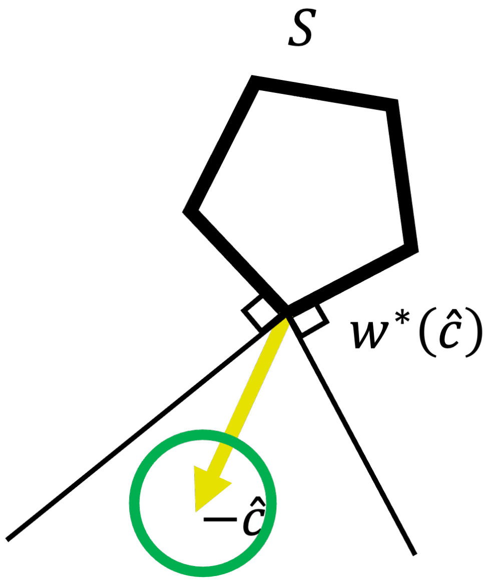
SPO loss can help to select samples for active learning

Motivation

- Vector $\hat{c}$ is the prediction from $\hat{h}(x)$
- Green circle is the "confidence region" of $\hat{c}$



Size of training set: 5

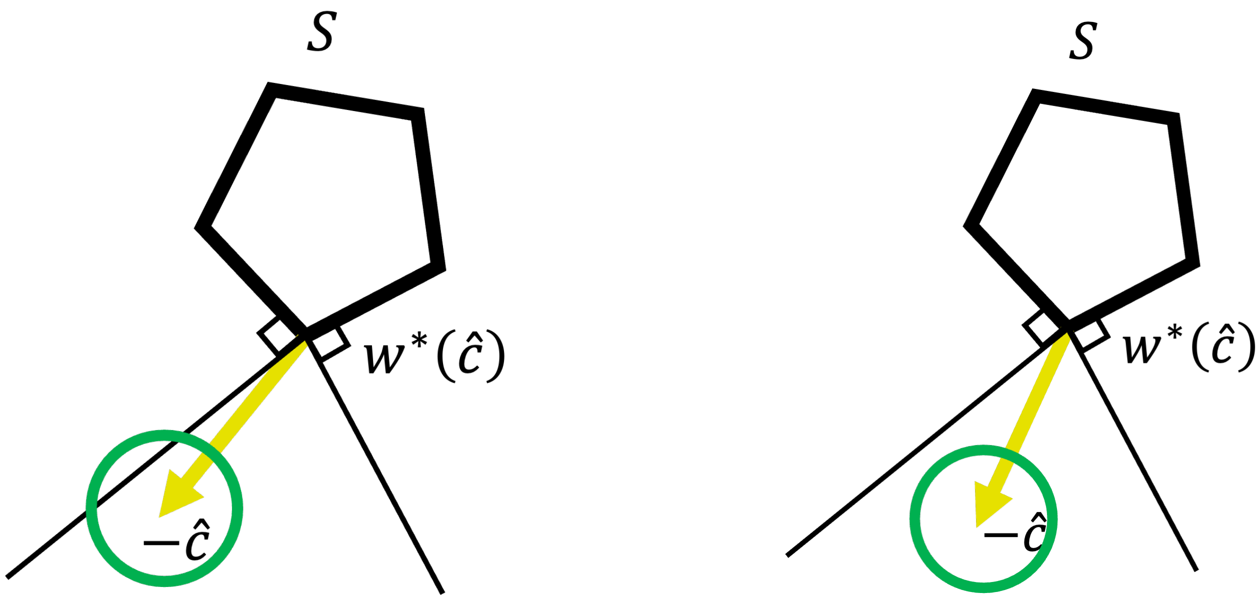


Size of training set: 5,000

- Active learning help identify critical samples to minimize SPO

Motivation

- Green circle has the same radius but different locations



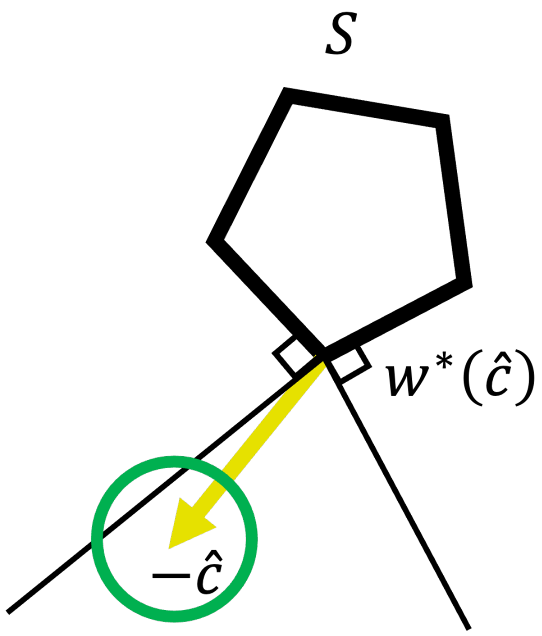
- SPO can make active learning more selective

Agenda

- Information gathering process for predict-then-optimize framework
- Smart predict-then-optimize loss (SPO) and preliminaries
- Theoretical motivation for margin-based algorithm
- **Algorithm**
- Analysis
- Numerical Experiments

Margin-based algorithm

- Idea: If the green circle (confidence region) intersects the boundary of the cone, then we acquire the label of that sample



Margin-based algorithm

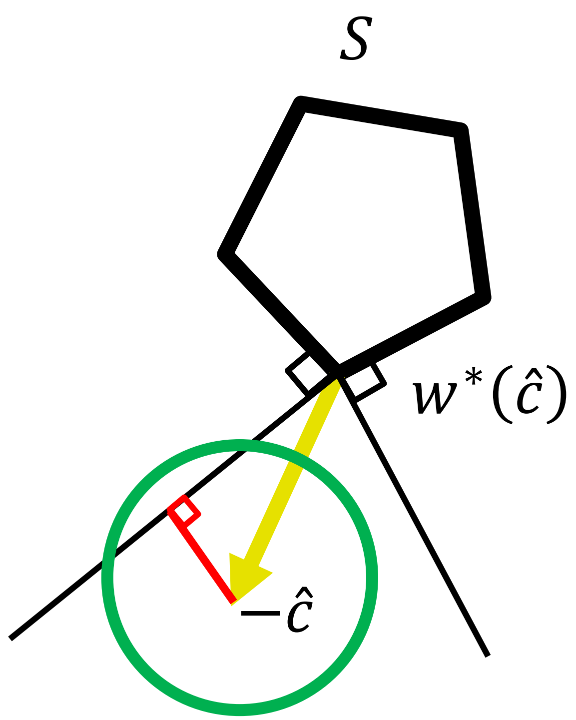
- Suppose $\mathcal{C}^0$ is the set of cost vectors that have multiple optimal decisions

- Distance to degeneracy:

$$\nu_S(\hat{c}) := \inf_{c \in \mathcal{C}^0} \{\|c - \hat{c}\|\}$$

- If $\nu_S\big(h_{t-1}(x_t)\big) < b_{t-1}$:

  Acquire the label of this sample x_t



Model training in the prediction-then-optimize framework

- After constructing a training set, how to obtain the predictor h_T?

- Minimize empirical loss in the selected training set
  - Squared loss

  - SPO+ loss
    - A specialized training loss that considering the downstream optimization:
    - Proposed in *Elmachtoub and Grigas (2022)*:

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

- There is some benefit when using the SPO+ loss

Agenda

- Information gathering process for predict-then-optimize framework
- Smart predict-then-optimize loss (SPO) and preliminaries
- Theoretical motivation for margin-based algorithm
- Algorithm
- ➢ **Analysis**
- Numerical Experiments

Theoretical guarantees for MBAL-SPO

- Label complexity vs. Sample complexity

Without any assumptions about the noise distribution:

- The label complexity is the same as the sample complexity in the supervised learning *(Kääriäinen M (2006))*

- We need some additional noise conditions

Label complexity from the low noise conditions

- Near-degeneracy function Ψ: the CDF of the distance to degeneracy
  - Difficulty in distinguishing the optimal decisions from the sub-optimal decisions

Low-noise condition: $\Psi(b) \leq b_0 \cdot b^{\kappa}$, $b < 1$, for some $\kappa > 0$

- κ gets larger $\rightarrow$ $\Psi(b)$ gets smaller $\rightarrow$ easier to find the optimal decisions
- Low-noise condition is closely related to Hu et al. 2022 and Tsybakov's noise condition

Assumption:

When the excess surrogate risk is at most Δ $\rightarrow$ The prediction error for $\hat{h}(x)$ is at most $\mathcal{O}(\sqrt{\Delta})$

Overview of the results

- After T iterations:

	Active learning	Supervised learning
Excess surrogate risk	$\mathcal{O}\left(T^{-1/2}\right)$	$\mathcal{O}\left(T^{-1/2}\right)$
Excess SPO risk	$\Psi\left(\mathcal{O}\left(T^{-\frac{\kappa}{4}}\right)\right)$	$\Psi\left(\mathcal{O}\left(T^{-\frac{\kappa}{4}}\right)\right)$
Number of labels	$\sum_{t=1}^{T}\Psi\left(\mathcal{O}\left(t^{-\frac{\kappa}{4}}\right)\right)$	T

Overview of the results

- After T iterations with low-noise conditions:

	Active learning	Supervised learning
Excess surrogate risk	$\mathcal{O}(T^{-1/2})$	$\mathcal{O}\left(T^{-1/2}\right)$
Excess SPO risk	$\mathcal{O}(T^{-\kappa/4})$	$\mathcal{O}(T^{-\kappa/4})$
Number of labels	$\mathcal{O}(T^{1-\kappa/4})$	T

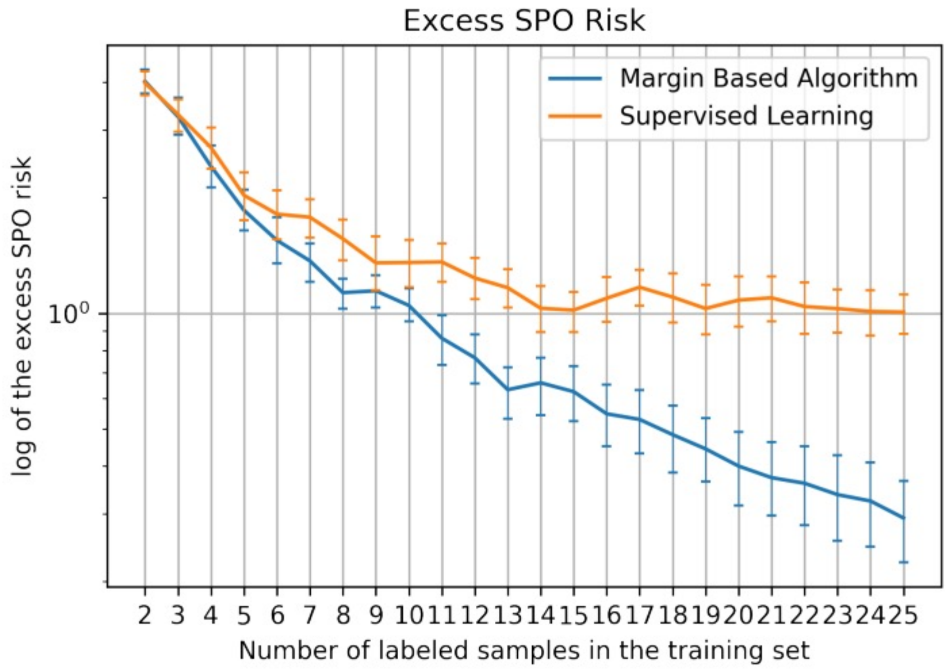
Agenda

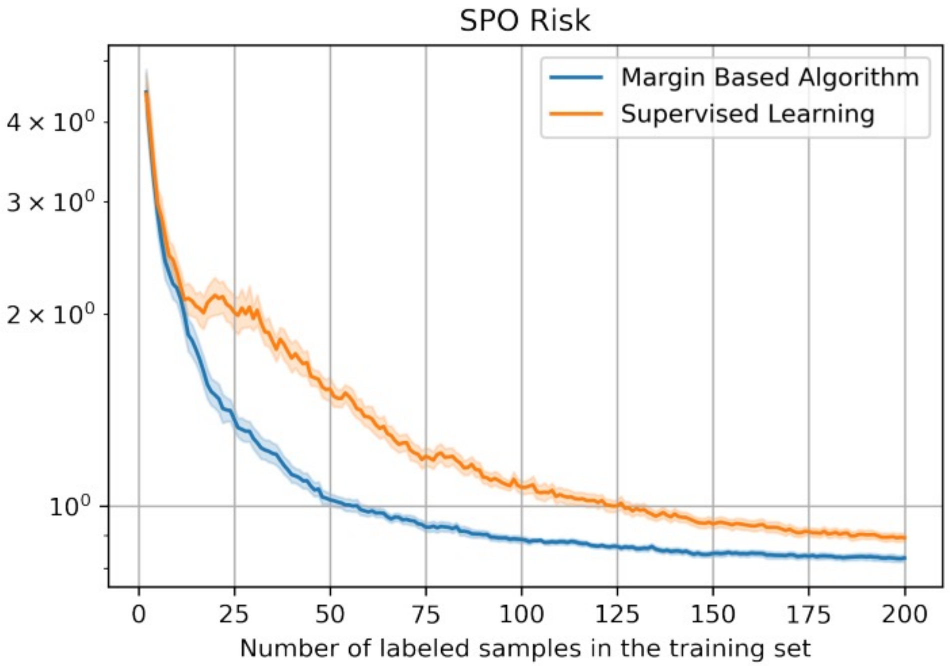
- Information gathering process for predict-then-optimize framework
- Smart predict-then-optimize loss (SPO) and preliminaries
- Theoretical motivation for margin-based algorithm
- Algorithm
- Analysis

➢ **Numerical Experiments**

Numerical experiments: shortest path problem

- Shortest path problem on $3{\times}3$ and $5{\times}5$ grid networks
- Predict the traveling time of each edge based on some features
- Using the SPO+ as the surrogate training loss





Numerical experiments: personalized pricing problem

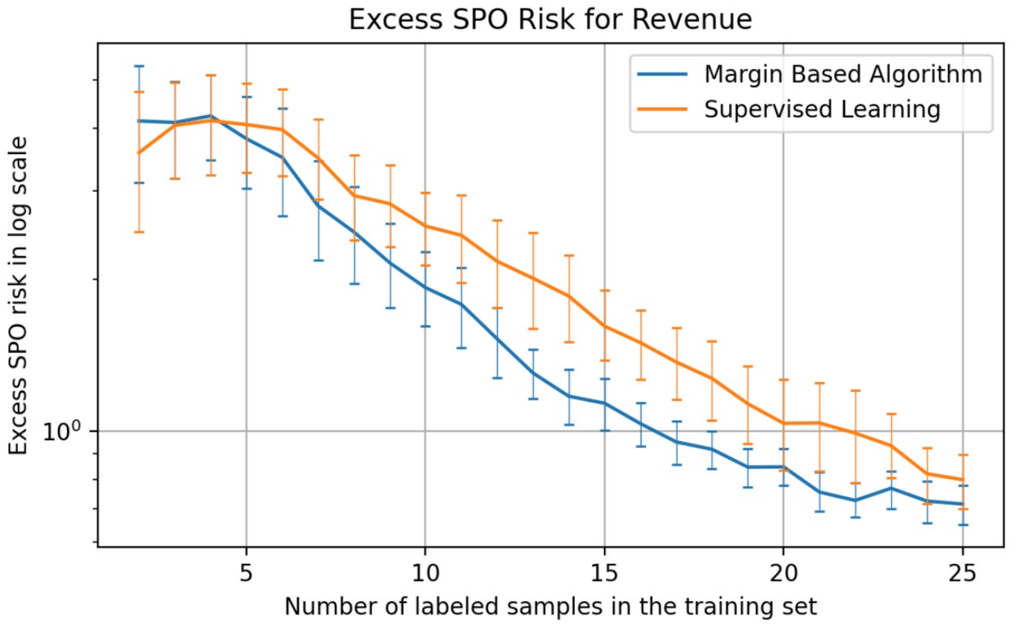
- Unknown function: $d_j(p^i) \in [0,1]$
  - Purchase probability of product j under price p^i
  - $d_j(p^i)$ depends on customer feature x
- Decisions: $w_{i,j} \in \{0,1\}$, whether the price of product j is set as p^i

$$
\begin{aligned}
\max_{\mathbf{w}}: \quad & \mathbb{E}[\sum_{j=1}^{\mathfrak{J}} \sum_{i=1}^{\mathcal{I}} d_j(p_i) p_i w_{i,j} | x] \\
& \sum_{i=1}^{\mathcal{I}} w_{i,j} = 1, \forall j = 1, 2, ..., \mathfrak{J} \\
& \mathbf{A}\mathbf{w} \leq b \\
& w_{i,j} \in \{0,1\}, \quad i = 1, 2, ..., \mathcal{I}, j = 1, 2, ..., \mathfrak{J}.
\end{aligned}
$$

- SPO loss: Revenue loss of the personalized prices based on the prediction for $d_j(p^i)$

Numerical experiments: personalized pricing problem

- Personalized pricing problem for three products
- The hypothesis class is mis-specified (The true model is exponential while the hypothesis class is linear.)



Thank you

https://arxiv.org/pdf/2305.06584.pdf

Variations of the algorithm

- If the noise does not satisfy the separable conditions or we use general surrogate loss:

- We have two variations:

- Variation 1:
  - Construct a confidence set of the optimal predictor at each iteration
  - $h_t \in H_t \subset H_{t-1} \subset \cdots \subset H_0$
  - Minimize the training loss within the confidence set
- Variation 2:
  - At each iteration, when $\nu_S\left(h_{t-1}(x_t)\right) > b_{t-1}$, reject samples with some probability smaller than 1

Overview of the results

- After T iterations under separability condition:

	Active learning	Supervised learning
Excess surrogate risk	$\mathcal{O}(T^{-1/2})$	$\mathcal{O}\left(T^{-1/2}\right)$
Excess SPO risk	$\mathcal{O}(\min\{T^{-\kappa/4}, T^{-1/2}\})$	$\mathcal{O}(T^{-\kappa/4})$
Number of labels	$\mathcal{O}(1)$	T

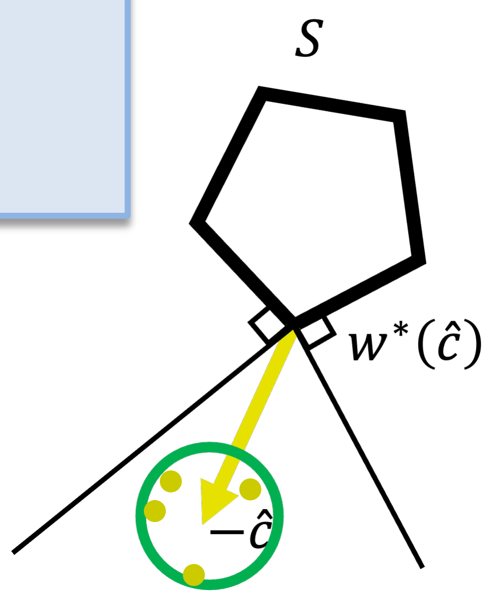
Property of SPO+ loss

- SPO+ loss in *Elmachtoub and Grigas (2022)*:

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

Separability condition: for some $\varrho \in (0,1)$:

$$\|h^*(x) - c\| \leq \varrho \nu_S\big(h^*(x)\big)$$

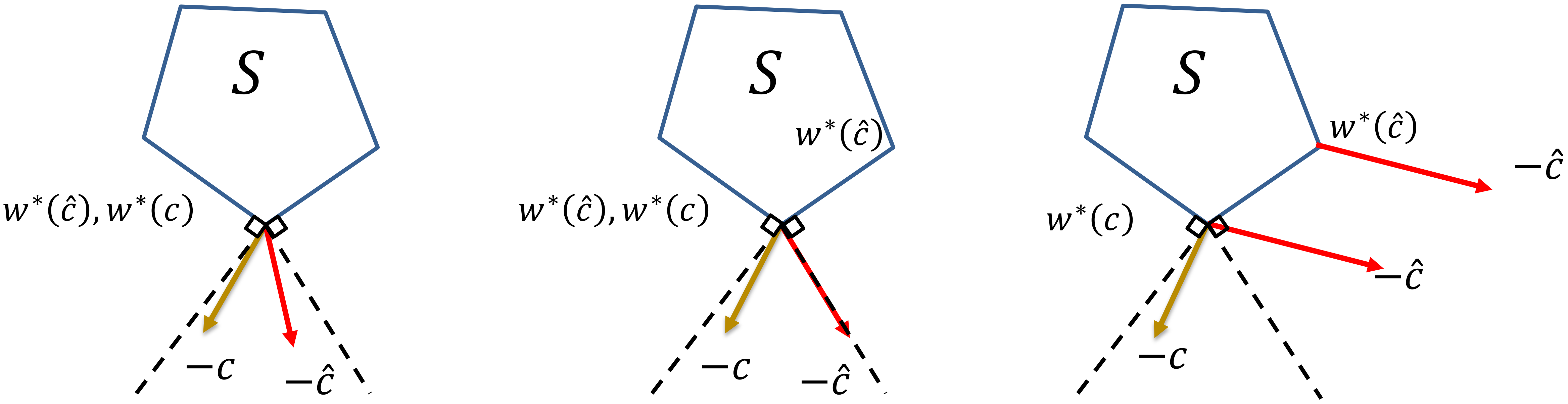


Separability condition implies:

– The minimum SPO+ risk and SPO risk are both zero.

Geometric Interpretation

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



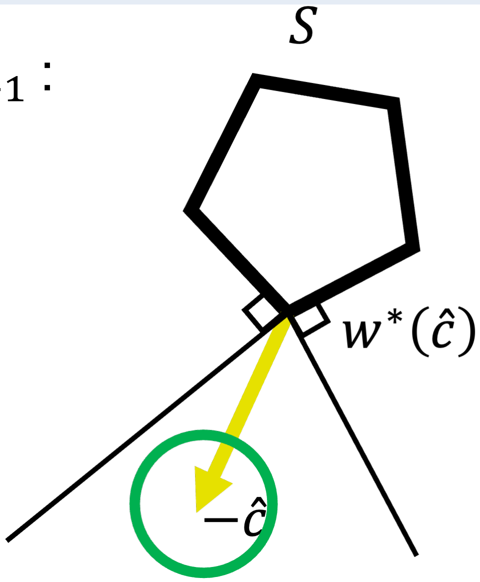
- SPO loss is discontinuous and nonconvex.
- Convex surrogate loss function: SPO+, squared loss, …

Summary: Three versions of MBAL-SPO

- Given a sequence b_t, and $\tilde{p}$ (If $\tilde{p} > 0$: soft-rejection; Otherwise, hard-rejection.)
- At each iteration t:
- Observe x_t
- If $v_S\left(h_{t-1}(x_t)\right) \geq b_{t-1}$:
  - Flip a coin with heads-up probability $\tilde{p}$
  - If the coin gets heads-up:
    - Acquire its label and update the training set
  - Else:
    - Reject x_t.
- Else:
  - Acquire its label and update the training set
- Update the predictor h_t by minimizing the empirical risk within the confidence set H_t
- Update a confidence set of predictor H_t if using general surrogate loss under hard-rejection.

Margin-Based Algorithm

- What if $v_S\left(h_{t-1}(x_t)\right) \geq b_{t-1}$:



  - Reject it directly?
  - The SPO loss of this sample is zero, so rejecting it does not change the total SPO loss.

$$\underbrace{\boxed{\ell_{SPO}(h_{t-1}(x_t), c_t)}}_{0} + \sum_{i=1}^{t-1} \ell_{SPO}(h_{t-1}(x_i), c_i)$$

😟 We are minimizing the surrogate loss, instead of SPO loss.

😟 Although SPO loss is zero, the surrogate loss is possibly nonzero.

Margin-Based Algorithm

- When rejecting the sample directly:

😟 Empirical surrogate loss $\nrightarrow$ Surrogate risk.

Soft rejection	Hard rejection with SPO+ surrogate	Hard rejection with general surrogate function

Soft rejection Algorithm

- If $v_S\big(h_{t-1}(x_t)\big) \geq b_{t-1}$(Green circle does not intersect with the boundary):
- ➢ Acquire the label with probability $\tilde{p} > 0$. (soft-rejection)

- If this sample eventually gets labeled, the weight of this sample is $\frac{1}{\tilde{p}}$
  - The expectation of re-weighted empirical surrogate loss equals the surrogate risk.

😟 The expected number of labels up to time T is at least $O(\tilde{p}T)$

Proof Sketch

1. Convergence for the surrogate risk:

$$R_\ell(h) - R_\ell(h^*) = \text{average excess risk for hard rejected samples}$$
$$+ \text{ uniform convergence rate for the reweighted risk}$$
$$+ \text{ average empirical excess risk of } h$$

Proof Sketch

1. Convergence for the surrogate risk:

$R_\ell(h) - R_\ell(h^*) =$ average excess risk for hard rejected samples (Holder's property)

+ uniform convergence rate for the reweighted risk (Sequential complexity)

+ average empirical excess risk of h ($h^*$is within H_t)

Proof Sketch

1. Convergence for the surrogate risk:

$$R_\ell(h) - R_\ell(h^*) = \text{average excess risk for hard rejected samples (Holder's property)}$$
$$+ \text{ uniform convergence rate for the reweighted risk (Sequential complexity)}$$
$$+ \text{ average empirical excess risk of } h \ (h^* \text{is within } H_t)$$

2. From surrogate risk to the SPO risk: near-degeneracy function.

3. Label complexity: Bound for the label probability at each step

4. For the soft-rejection, optimize $\tilde{p}$ as a function of T to achieve a small label complexity.

Examples of operations management problems

Unknown parameters	Decision
Customers' preference	Product recommendation
Demand	Inventory level
Price elasticity	Optimal prices