

A PREDICT-THEN-OPTIMIZE APPROACH FOR PORTFOLIO SELECTION

PHƯƠNG PHÁP DỰ BÁO SAU ĐÓ TỐI ƯU CHO VIỆC LỰA CHỌN DANH MỤC ĐẦU TƯ

Le Hoang Son, Phan Nguyen Ky Phuc

Department of Industrial and System Engineering, International University (VNU-HCM)

ABSTRACT

Many real-world financial analytic problems involve two challenged and complexed tasks: predict and optimization, which leads to the standard paradigm is predict-then-optimize. Typically, these two tasks run separately which means that the machine learning models try to minimize the prediction error and do not consider how such predictions will be used for the downstream optimization problem. To tackle it, in this paper we introduce the Smart Predict then Optimize"(SPO) framework which leverages the optimization structure for designing a better prediction model. Moreover, for the optimization structure, we apply Distributionally robust chance constrained optimization over Wasserstein ambiguity sets to define the feasible solution on a set of conic constraints. We train our model under subgradient methods. Finally, we evaluate our approach with least square loss to compare the performance between them.

Keywords: Smart Predict then optimize; Portfolio selection; Subgradient.

TÓM TẮT

Nhiều vấn đề phân tích tài chính trong thế giới thực liên quan đến hai nhiệm vụ đầy thách thức và phức tạp: dự đoán và tối ưu hóa, dẫn đến mô hình tiêu chuẩn là dự đoán-sau đó-tối ưu hóa. Thông thường, hai tác vụ này chạy riêng biệt, có nghĩa là các mô hình học máy cố gắng giảm thiểu lỗi dự đoán và không xem xét các dự đoán đó sẽ được sử dụng như thế nào cho bài toán tối ưu hóa hạ lưu. Để giải quyết vấn đề này, trong bài báo này, chúng tôi giới thiệu khuôn khổ Dự đoán sau đó tối ưu hóa"(SPO) thông minh thúc đẩy cấu trúc tối ưu hóa để thiết kế mô hình dự đoán tốt hơn. Hơn nữa, đối với cấu trúc tối ưu hóa, chúng tôi áp dụng tối ưu hóa hạn chế cơ hội mạnh mẽ phân tán trên các tập mờ hồ Wasserstein để xác định giải pháp khả thi trên tập hợp các ràng buộc conic. Chúng tôi đào tạo mô hình của mình theo các phương pháp dưới cấp. Cuối cùng, chúng tôi đánh giá cách tiếp cận của mình với tổn thất bình phương ít nhất để so sánh hiệu suất giữa chúng.

Từ khóa: Dự đoán sau đó tối ưu; Lựa chọn danh mục đầu tư; Subgradient.



1. INTRODUCTION

In many real-world applications of operations research, or in particular finance usually involve two important tasks: prediction and optimization. Typically, these challenges are separated into two distinguished complex processes, which is so-predict then optimize paradigm. Machine learning oracles will acquire the prediction tasks which intend to predict the unknown parameters of optimization problems and minimize the prediction error, but do not take into account how the predictions affect the downstream optimization. For example, consider a portfolio selection problem that may be computationally challenging. At first, a trained prediction will provide predictions for the impact factors of portfolio cost on current transaction day. Then, an optimization solver will provide near-optimal/feasible decisions of portfolio investment using the predicted factor as input. In this paper, together with the DRO approach, we combine a new and leverage framework with aiming to improve the decisions, namely Smart "Predict, then Optimize" (SPO), which directly incorporate the machine learning models to the optimization problems structure for coming up superior decisions. A main component of this framework is the SPO loss function which measures the error of decisions induced by a prediction.

Integrating prediction and optimization or Predict-then-optimize arise in many real-world applications. Using and exploiting the KKT conditions around the optimal solutions, researchers has proposed many frameworks on predict-then-optimize problems. For example, [1] propose a framework which computes the gradient of the solution of the QP with respect to the predictions by applying the implicit function theorem to differentiate through the

KKT conditions around the optimal point, [2] uses the same approach for linear programs with small modifications by adding the quadratic term to convert it to concave QP programs. [3] uses the trained machine learning models to directly predict the optimal solution of a newsvendor problem from data with effectiveness and tractability properties, but it is unclear that this approach can work well or not when there are constraints. [4] estimates the un-known optimization objective using ML models accurately by using a weighted combination of training samples such nearest neighbors and decision trees. They predict the objective of an instance by applying the same weights initiated by the ML model to the associated objective functions of those samples. This approach also does not rely on the specification of optimization problems. [5] proposes are regret minimization in context of portfolio optimization with historical data without features. Another related work is data-driven inverse optimization paradigm where feasible or optimal solutions of an optimization problems are solved and objective function needs to be learned. Typically, this approach handles a single unknown objective, which does not provide any prior samples of objective function, see [6]. Our work in this paper is based on Smart "Predict then Optimize (SPO) framework proposed by [7] which is a utilized framework for combining prediction and optimization tasks. Commonly, two-stage learning method is widely used in industry which is first predicting the unknown parameters, then these parameters will be input to optimization problems. Exclude a few limited options, this approach is used for predictions without considering how these predictions affect the downstream optimization problems. On the other hand, [7] introduces the Smart "Predict then Optimize" which utilizes

the structural information of optimization problems for machine learning models that predict unknown parameters. Fundamentally, SPO framework is an integrated prediction and optimization framework that aims to minimize the decision errors, not the prediction errors. Given a training dataset $D: \{(x_1, c_1), \dots, (x_n, c_n)\}$ which is the historical feature data x_i corresponding to observed parameters data c_i , we predict the unobserved parameters $(\hat{c}_1, \dots, \hat{c}_n)$ under a trained model. To evaluate the quality of prediction, we can use the least square loss, $\|c_i - \hat{c}_i\|_2^2$ to measure errors. However, this measurement totally ignores the decisions included to the predictions. The key advantage of SPO loss over the least square loss is that it includes the decision in to prediction by taking the true cost created by $\hat{c}_i$, minus the optimal cost induced from true parameters c_i . Hence, the SPO loss focuses on maintaining the sequentially predicting then optimizing without ignoring the structure of explicit optimization problems. Moreover, SPO loss predicts the unknown parameters of the contextual optimization problems in which the parameters appear in the linear objective function. Moreover, even if the prediction task is challenging due to model misspecification, the SPO framework can still yield suboptimal decisions. This is, in fact, because the structure of SPO framework require that the prediction is directly attached in to the downstream optimization problem. One more beneficial is that the downstream optimization problem is typically more computationally tractable and more attractive to practitioners, which means that SPO framework can integrate the optimization and prediction task together easily, rather than making a more complex procedure. Overall, SPO framework provides a clear foundation for creating operation-driven machine learning tools that can be leveraged in real-world optimization settings.

2. METHODOLOGY

2.1. SPO loss and its surrogate function

SPO framework proposed by [6] interests in the nominal optimization problem with a linear objective, where the decision variable $\mathbf{w} \in \mathbb{R}^K$ and a feasible region well-defined with certainty $\mathcal{S} \subseteq \mathbb{R}^K$. Nevertheless, the cost vector of objective, $\mathbf{c} \in \mathbb{R}^K$, is not known at the time the decision are made; instead of knowing the available associated feature vector $\mathbf{x} \in \mathbb{R}^P$. Since $\mathbf{x}$ is correlated to $\mathbf{c}$, we can predict $\mathbf{c}$ using the conditional information from $\mathbf{x}$. Our goal is to solve the contextual stochastic optimization problem.

$$\min_{\mathbf{w} \in \mathcal{S}} \mathbb{E}_{\mathbf{c} \sim \mathcal{D}_x} [\mathbf{c}^\top \mathbf{w} \mid \mathbf{x}] = \min_{\mathbf{w} \in \mathcal{S}} \mathbb{E}_{\mathbf{c} \sim \mathcal{D}_x} [\mathbf{c} \mid \mathbf{x}]^\top \mathbf{w}$$

where $\mathcal{D}_x$ is the conditional distribution of $\mathbf{c}$ given $\mathbf{x}$. SPO framework bases on predicting $\mathbf{c}$ given $\mathbf{x}$ which is $\mathbb{E}_{\mathbf{c} \sim \mathcal{D}_x} [\mathbf{c} \mid \mathbf{x}]$. We denote this by $\hat{\mathbf{c}}$, and thus solve the deterministic version of the optimization model relied on by $\hat{\mathbf{c}}$. Moreover, we might not have the structure of conditional distribution $\mathcal{D}_x$, but we could have the training data sample which contains both of them.

Next, we start with the predict-then-optimize paradigm in which we measure the surplus cost provoked when making a suboptimal decision because of an inexact cost vector prediction. Given a cost vector prediction $\hat{\mathbf{c}}$, a decision $\mathbf{w}^*(\hat{\mathbf{c}})$ is made with respect to the cost $\hat{\mathbf{c}}^\top \mathbf{w}^*(\hat{\mathbf{c}})$. After the decision $\mathbf{w}^*(\hat{\mathbf{c}})$ is carried out, the excess cost incurred with respect to the true cost vector $\mathbf{c}$ is measured by $\hat{\mathbf{c}}^\top \mathbf{w}^*(\hat{\mathbf{c}}) - \mathbf{c}^\top \mathbf{w}^*(\hat{\mathbf{c}})$, which is so-called the SPO loss. Note that, the $\mathbf{w}^*(\hat{\mathbf{c}})$ may be suboptimal with respect to $\mathbf{c}$. Here, we denote the decision $\mathbf{w}^*(\mathbf{c}) := \arg \min_{\mathbf{w} \in \mathcal{S}} \{\mathbf{c}^\top \mathbf{w}\}$ where $\mathcal{S} \subseteq \mathbb{R}^K$ is non-empty, compact and convex set .

that represent the feasible region, the optimal objective value is $\mathbf{z}^*(\mathbf{c}) := \min_{\mathbf{w} \in \mathcal{S}} \mathbf{c}^\top \mathbf{w}$.

Given a cost vector prediction $\hat{\mathbf{c}}$ and a realized cost vector $\mathbf{c}$, the SPO loss, $\ell_{SPO}(\hat{\mathbf{c}}, \mathbf{c})$ with respect to the optimization oracle $\mathbf{w}^*(\cdot)$ is defined as $\ell_{SPO}(\hat{\mathbf{c}}, \mathbf{c}) := \mathbf{c}^\top \mathbf{w}^*(\hat{\mathbf{c}}) - \mathbf{c}^\top \mathbf{w}^*(\mathbf{c})$. Even the SPO loss is a convex function, $\ell_{SPO}(\hat{\mathbf{c}}, \mathbf{c})$ could not be continuous in $\hat{\mathbf{c}}$ because $\mathbf{w}^*(\hat{\mathbf{c}})$ may also be discontinuous in $\hat{\mathbf{c}}$ and exponential in size. As a surrogate lost function of SPO loss function, SPO+ loss approximates $\ell_{SPO+}(\hat{\mathbf{c}}, \mathbf{c})$. When finding the prediction model that minimizes the empirical risk using the SPO+ loss, this prediction model also estimates the empirical risk using SPO loss. We recall the definition of SPO+ loss as following form 1_ $\ell_{SPO+}(\hat{\mathbf{c}}, \mathbf{c}) := \max_{\mathbf{w} \in \mathcal{S}} \{\mathbf{c}^\top \mathbf{w} - 2\hat{\mathbf{c}}^\top \mathbf{w}\} + 2\hat{\mathbf{c}}^\top \mathbf{w}^*(\mathbf{c}) - \mathbf{z}^*(\mathbf{c})$.

As a surrogate loss function of SPO loss, the SPO+ loss is $\ell_{SPO+}(\hat{\mathbf{c}}, \mathbf{c}) \geq \ell_{SPO}(\hat{\mathbf{c}}, \mathbf{c})$. Moreover, the SPO+ loss is a convex function of the cost vector prediction $\hat{\mathbf{c}}$ since $\max_{\mathbf{w} \in \mathcal{S}} \{\mathbf{c}^\top \mathbf{w} - 2\hat{\mathbf{c}}^\top \mathbf{w}\}$ is a maximum of linear functions which is a convex one. Unfortunately, $\ell_{SPO+}(\hat{\mathbf{c}}, \mathbf{c})$ is a non-smooth function because it is not differentiable since $\max_{\mathbf{w} \in \mathcal{S}} \{\mathbf{c}^\top \mathbf{w} - 2\hat{\mathbf{c}}^\top \mathbf{w}\}$

is not generally differentiable. This implies that training the $\ell_{SPO+}(\hat{\mathbf{c}}, \mathbf{c})$ using gradient method is inapplicable. To train this function, [6] shows that $\mathbf{w}^*(\mathbf{c}) - \mathbf{w}^*(2\hat{\mathbf{c}} - \mathbf{c})$ is a subgradient of $\ell_{SPO+}(\hat{\mathbf{c}}, \mathbf{c})$, which is $\mathbf{g} = \mathbf{w}^*(\mathbf{c}) - \mathbf{w}^*(2\hat{\mathbf{c}} - \mathbf{c}), \mathbf{g} \in \partial \ell_{SPO+}(\hat{\mathbf{c}}, \mathbf{c})$.

Finally, [6] shows that minimizing risk SPO+ loss is also minimizes the SPO loss. Thus, it is reasonable to train the SPO+ loss rather than SPO loss. We aim to solve the empirical risk minimization of SPO in form of

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell_{SPO+}(f(x_i), c_i).$$

Given the definition of subgradient, the empirical risk SPO+ loss problem could be optimized using the Stochastic Subgradient Descent method.

2.2. The downstream portfolio selection models

In this paper, we choose the chance-constrained portfolio selection problem with Wasserstein distance adapted from [8] for the downstream optimization $\mathbf{w}^*(\hat{\mathbf{c}})$. The problem is formulated as:

$$\begin{aligned} \min_{\mathbf{q}, \mathbf{r}, t, \mathbf{w}} \quad & \mathbf{c}^\top \mathbf{w} \\ \text{s.t.} \quad & \mathbf{q} \in \{0, 1\}^N, t \geq 0, \mathbf{r} \geq \mathbf{0}, \mathbf{w} \geq \mathbf{0}, \\ & \epsilon t \geq \theta \|\mathbf{w}\|_* + \frac{1}{N} \sum_{i \in [N]} r_i \\ & M_i(1 - q_i) \geq t - r_i, i \in [N], \\ & \mathbf{w}^\top \xi_i - \omega + M_i q_i \geq t - r_i, i \in [N], \\ & \sum_{i \in [N]} q_i \leq \lfloor \epsilon N \rfloor \\ & \|\mathbf{w}\|_0 \leq \nu \end{aligned}$$

The objective is to find a minimum cost of investment $\mathbf{w}$ into K assets with random return $\xi = (\xi_1, \dots, \xi_K)^\top$ while reaching a authorized target return with probability at least $1 - \epsilon$. To be more precise, the vector $\mathbf{w}$ represent the percentage of a total wealth invested of one dollar in each K available assets, and ξ is the vector of random return of these assets. In other words, the vector of cost $\mathbf{c}$ can be viewed as negative of the expected investment returns of K assets, i.e. $\mathbf{c} = -E(\xi)$. Thus, the decision makers or agents would like to minimize the investment cost subject to having the target

return greater or equal to a desired level ω with probability at least $1 - \epsilon$, which is equivalent to maximize the expected return of K assets. We refer readers to the paper [8] for more detail on this problem. The random return $\xi > 0$ for all $\xi \in \mathbb{R}^K$ captures the random yields of financial assets encoding the ratio of end price and the initial price of a financial asset, which is profitable if the ratio is greater than 1 and otherwise. Here, c is the cost vector and ω is the authorized target return. Here, $w \geq 0$ means that there is no-short selling decision. Constraint $\|w\|_0 \leq \nu$ is the budget constraint that controls the total investment is lower than.

3. NUMERICAL EXPERIMENT

3.1. Experiment setup

In this section, we conduct a computational experiment in order to investigate the effect of using loss function taking into account the downstream optimization problem information, which is fusion of prediction and optimization. To make our experiment comparative, we examine the SPO+ loss $\ell_{SPO+}(\hat{c}, c)$ and least square loss $\ell_{LS}(\hat{c}, c)$. Recall that the least square loss does not consider any information about the optimization problem. On the other hands, the SPO+ loss function incorporates information of the optimization problem. Hence, the empirical risk minimization problems of least square loss

$$\text{is } \ell_{LS}(\hat{c}, c) := \min_{V \in \mathbb{R}^{K \times p}} \frac{1}{2n} \sum_{i=1}^n \|Vx_i - c_i\|^2,$$

and SPO+ loss is

$$\ell_{SPO+}(\hat{c}, c) := \min_{V \in \mathbb{R}^{K \times p}} \frac{1}{n} \sum_{i=1}^n \ell_{SPO+}(Vx_i, c_i).$$

In all problem classes, we compare the linear predictor between SPO+ and LS loss

where V is obtained by solving the empirical risk minimization problem in (1.) and (2.) on same his historical data $\{(x_i, c_i)\}_{i \in [n]}$. We estimate V via optimizing the least squares loss and the SPO+ loss. Next, we first take the feature x , then use the V to compute the decision $w^*(Vx)$, then compute the median of optimality gap between prediction and actual cost of each loss $L = c^T w^*(Vx) - c^T w^*(c)$. To evaluate the effectiveness of two losses in terms of incorporating information of optimization, we use "Improvement" to denote the percentage of difference between the value of these two methods.

$$\text{Improvement}(\%) = \frac{L_{LS} - L_{SPO+}}{L_{LS}} \times 100$$

We conduct all of experiment on an self-laptop with Intel Core i7 8950HK 2.8GHz processor with 6 core 32GB memory. All experiment are developed in Julia, and is run on single-node mode. We use CPLEX as an external MIP solver for downstream optimization problems with JuMP as a wrapper. For solving the empirical risk minimization problems for unregularized least square and SPO+ loss, we utilize the is the Stochastic Subgradient Descent. To be fairness in both methods, we do not use the early stopping technique of SGD. For the inner optimization problem which is probabilistic portfolio, we set number of assets $K \in \{20, 30, 50\}$ and $\omega = 1$. On the set of values for the Wasserstein radius and risk tolerance ϵ , we choose $\theta \in \{0.05, 0.08, 0.01\}$ and $\epsilon \in \{0.05, 0.1\}$. We also set the N scenarios in as, $N \in \{100, 200, 300\}$. As we use l_2 -norm to describe the Wassertein distance, the downstream portfolio selection becomes the mixed integer second-order cone programs. For historical training data $\{(x_i, c_i)\}_{i \in [n]}$, we follow the generation scheme of [6].



3.2. Computational result

For comparison between SPO+ and LS loss, we generally observe that SPO+ provides superior result to LS, which emphasizes the roles of incorporating optimization information to prediction tasks. As long as the number of assets becomes greater, the SPO+ method performs significantly compare to LS loss. As for the θ values and N scenarios grow up which elevate the complexity of downstream optimization, we observe that SPO+ still show improvement over LS loss. Especially for the case of 50 assets, SPO+ method tends to offer improvement over LS method in all sectors. However, one cannot claim that SPO+ outperforms LS due to the fact that the improvement of SPO+ over LS is not considerable large, whereby most of them are lower than 5%. Finally, SPO+, which combine the information of downstream optimization problems into prediction task, is better approach than LS in purpose of integrating prediction and optimization task. The detailed result is shown as:

Table 1. Computation result

	θ	K		
		~20	~30	~50
$N = 100$	0.05	3.055	-3.353	1.472
	0.08	1.721	7.571	2.876
	0.1	-2.578	-2.862	7.612
$N = 200$	0.05	3.005	-2.826	3.578
	0.08	0.742	-4.203	2.164
	0.1	3.665	2.937	-2.843
$N = 300$	0.05	1.918	-0.972	1.534
	0.08	-2.065	-2.605	1.917
	0.1	2.876	-4.273	3.455

4. CONCLUSIONS

In this paper, we provide the formulation

for chance constrained portfolio selection when the uncertainty affects to the decision variables using Wasserstein ambiguity sets, which implies that we handle the unknown parameters in constraints. We also provide the SPO and SPO+ frameworks for developing prediction model under predict-then-optimize paradigm. By using SPO and SPO+, we could include the contextual information to optimization, which usually ignore in practice. However, there is still some drawbacks and possible new research opportunities in our approach. ♦

References:

- [1]. P. Donti, B. Amos, and J. Z. Kolter; *Task-based end-to-end model learn-ing in stochastic optimization*. In, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/3fc2c60b5782f641f76bcefc39fb2392-Paper.pdf>.
- [2]. B. Wilder, B. Dilkina, and M. Tambe; *Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization*. *Proceeding sof the AAAI Conference on Artificial Intelligence*, 33(01):1658–1665, Jul. 2019. doi: 10.1609/aaai.v33i01.33011658. URL <https://ojs.aaai.org/index.php/AAAI/article/view/3982>.
- [3]. G.-Y. Ban and C. Rudin; *The big data newsvendor: Practical insights from machine learning*. *Operations Research*, 67(1):90–108, 2019.
- [4]. D. Bertsimas and N. Kallus; *From predictive to prescriptive analytics*. *Management Science*, 66(3):1025–1044, 20.
- [5]. E. Lim, J. G. Shanthikumar, and G.-Y. Vahn; *Robust portfolio choice with learning in the framework of regret: Single-period case*. *Management Science*, 58(9):1732–1746, 2012.
- [6]. N. Elmachtoub and P. Grigas; *Smart “predict, then optimize”*. *Management Science*, 2021.
- [7]. P. M. Esfahani, S. Shafieezadeh-Abadeh, G. A. Hanasusanto, and D. Kuhn; *Data-driven inverse optimization with imperfect information*. *Mathematical Programming*, 167(1):191–234, 2018.
- [8]. Ho-Nguyen, F. Kılınç-Karzan, S. Küçükyavuz, and D. Lee; *Strong formulations for distributionally robust chance-constrained programs with left-hand side uncertainty under wasserstein ambiguity*. *arXiv preprint arXiv:2007.06750*, 2020.