Smart “Predict, then Optimize”

Adam N. Elmachtoub, Paul Grigas .

Management Science,2022

解读者：屠雪永

2022.02.21

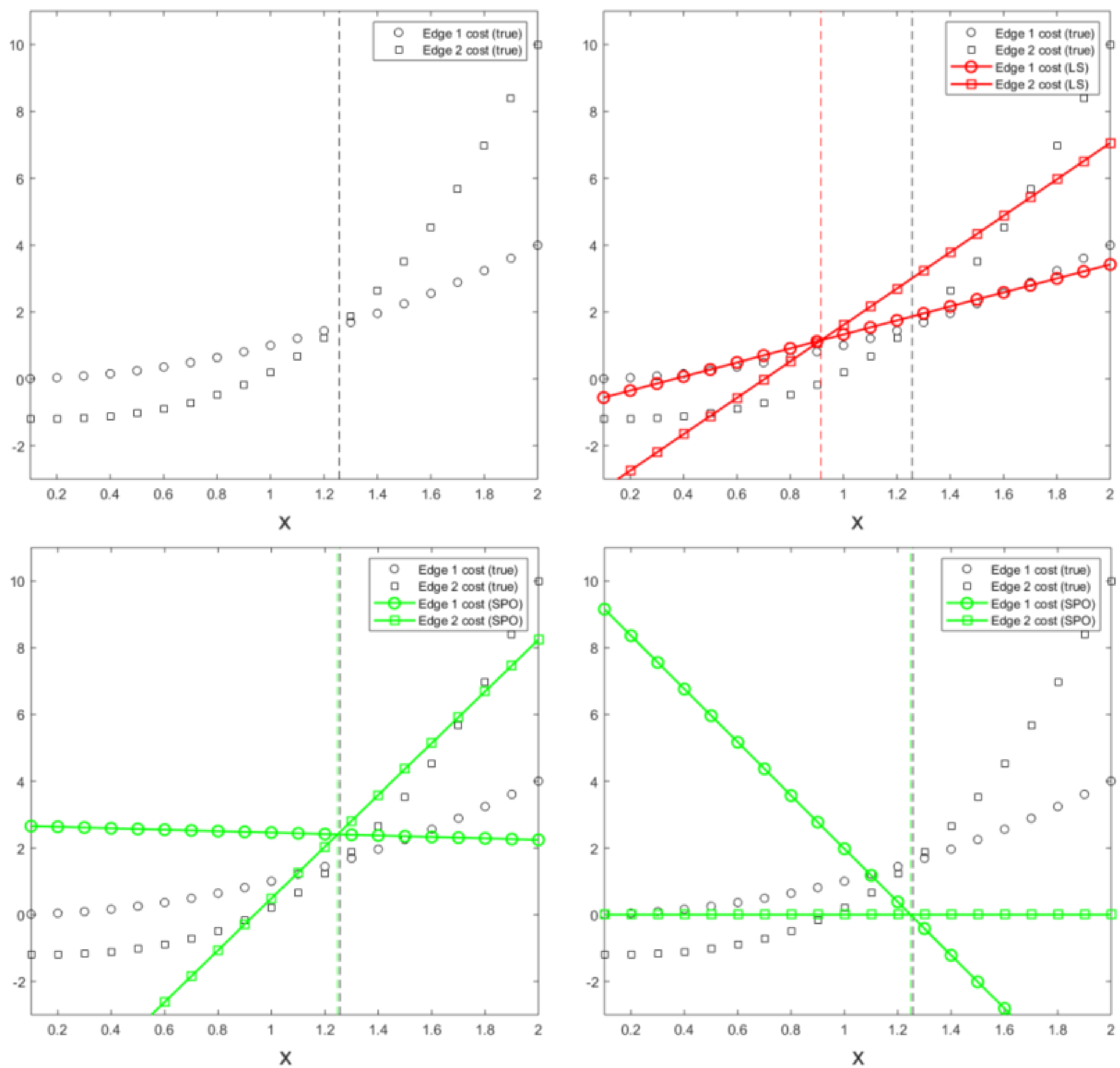
Outline

- Introduction
- “Predict, then Optimize” Framework
- SPO Loss Functions
- Consistency of the SPO+ Loss Function
- Computational Approaches and Experiments
- Conclusion

1. I

•

Figure 3 Illustrative Example.



1. Introduction-- Motivation

- Two significant challenges: prediction and optimization.
- The standard paradigm is predict-then-optimize
  - do not account for how the predictions will be used in the optimization problem
- Smart “Predict, then Optimize” (SPO)
  - directly leverages the optimization problem structure for designing better prediction models.

1. Introduction-- Applications

- Vehicle Routing
  - the cost of each edge
- Inventory Management
  - demand predictions
- Portfolio Optimization
  - the returns

1. Introduction-- Related Literature

- Nominal optimization problem has **no constraints** and **bypass issues of non-uniqueness** of solutions
  - Kao et al. (2009), . Donti et al. (2017), Ban and Rudin (2019)
- ML models has **no constraints and nonlinear** in the parameter.
  - Ban and Rudin (2019), Bertsimas and Kallus (2020)
- **Not the true parameters and without features**
  - Tulabandhula and Rudin (2013), Gupta and Rusmevichientong (2017)

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

- Data-driven inverse optimization, **no previous samples** of the objective
  - Bertsimas et al. (2015), Esfahani et al. (2018)

x→c→w

- The general setting of structured prediction ,SSVM (x→w)
  - Osokin et al. (2017)

1. Introduction-- Framework

"Predict, then Optimize" Framework

$$P(c): \quad z^*(c) := \min_w c^T w \quad \text{s.t. } w \in S,$$

Consider decision error

The SPO Loss Functions

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

Change nonconvex to convex

the SPO+ Loss Function

Prove the Consistency

Computational Approaches and Experiment

1. Introduction-- Contribution

- 1. We first formally define a **new loss function**, which we call the SPO loss.
- 2. Given the intractability of the SPO loss function, we develop a surrogate loss function which we call the **SPO+ loss**.
- 3. We prove a key **consistency** result of the SPO+ loss function.
- 4. we validate our framework through **numerical experiments** on the shortest path and portfolio optimization problem.

2. “Predict, then Optimize” Framework

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

- 1.Nominal (downstream) optimization problem

$$\begin{aligned} P(c):\quad z^*(c) := \ & \min_{w} c^T w \\ & \text{s.t. } w \in S\ , \end{aligned}$$

where w are the decision variables, c is the problem data describing the linear objective function, and $S \subseteq \mathbb{R}^d$ is a nonempty, compact and convex

- 2. **Training data** of the form $(x_1, c_1), (x_2, c_2), \ldots, (x_n, c_n)$, where x_i is a feature vector

- 3. A hypothesis class $\mathcal{H}$ of cost vector prediction models $f : X \to \mathbb{R}^d$, where $\hat{c} := f(x) = Bx$

2. “Predict, then Optimize” Framework

- 4. A loss function $\ell(\cdot,\cdot): \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_+$, whereby $\ell(\hat{c}, c)$ quantifies the error in making prediction $\hat{c}$ when the realized (true) cost vector is actually c.

- The optimization problem

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

$$\ell(\hat{c}, c) = \frac{1}{2} \|\hat{c} - c\|_2^2$$

$$\begin{aligned} &\min \mathrm{c}^T w \\ &s.t.\ w^T \Sigma \mathrm{w} \leq \Upsilon \\ &e^T w \ <= \ 1 \\ &w \ >= \ 0 \end{aligned}$$

$$\mathrm{x} \xrightarrow{\hat{c} := \ f(x) = Bx} \mathrm{c} \xrightarrow{\qquad\qquad} \mathrm{w}$$

3. Smart "Predict, then Optimize" Framework

- **Definition 1 (SPO Loss).** Given a cost vector prediction $\hat{c}$ and a realized cost vector c, the true SPO loss $l_{SPO}^{w^*}(\hat{c}, c)$ w.r.t. optimization oracle $w^*(\cdot)$ is defined as $l_{SPO}^{w^*}(\hat{c}, c) := c^T w^*(\hat{c}) - z^*(c)$.

$$P(c): \quad z^*(c) := \min_{w} c^T w \quad \text{s.t. } w \in S ,$$

- Deficiency: $w^*(\hat{c})$ may not be unique

$$\min_{w \in W^*(\hat{c})} c^T w - z^*(c)$$

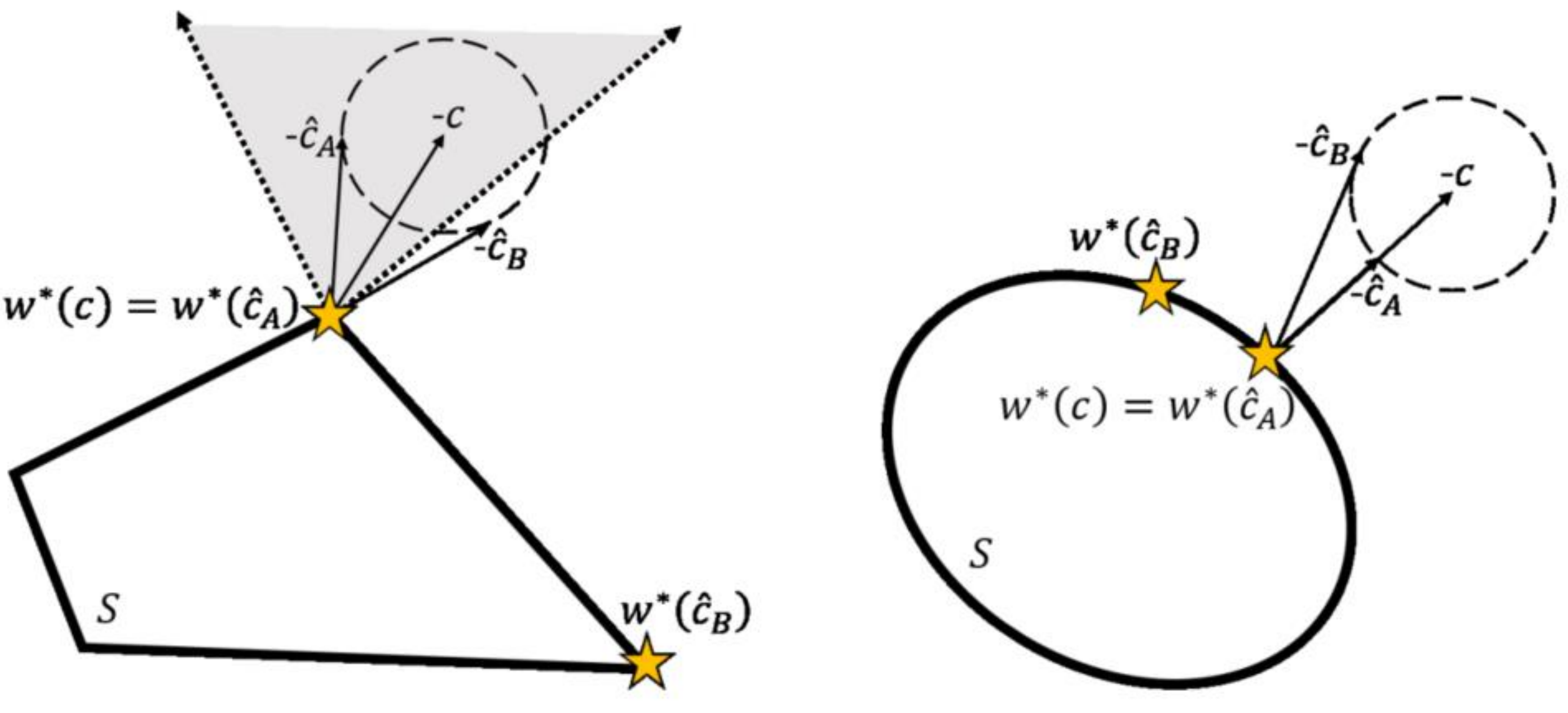
→degenerate prediction $\hat{c}$ = 0 since $W^*(0)$ = S

- **Definition 2 (Unambiguous SPO Loss).** Given a cost vector prediction $\hat{c}$ and a realized cost vector c, the true SPO loss $l_{SPO}(\hat{c}, c)$ is defined as $l_{SPO}(\hat{c}, c) := max_{w \in \mathrm{W}^*(\hat{c})} c^T w - z^*(c)$.

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

3. Smart “Predict, then Optimize” Framework

- An Illustrative Example

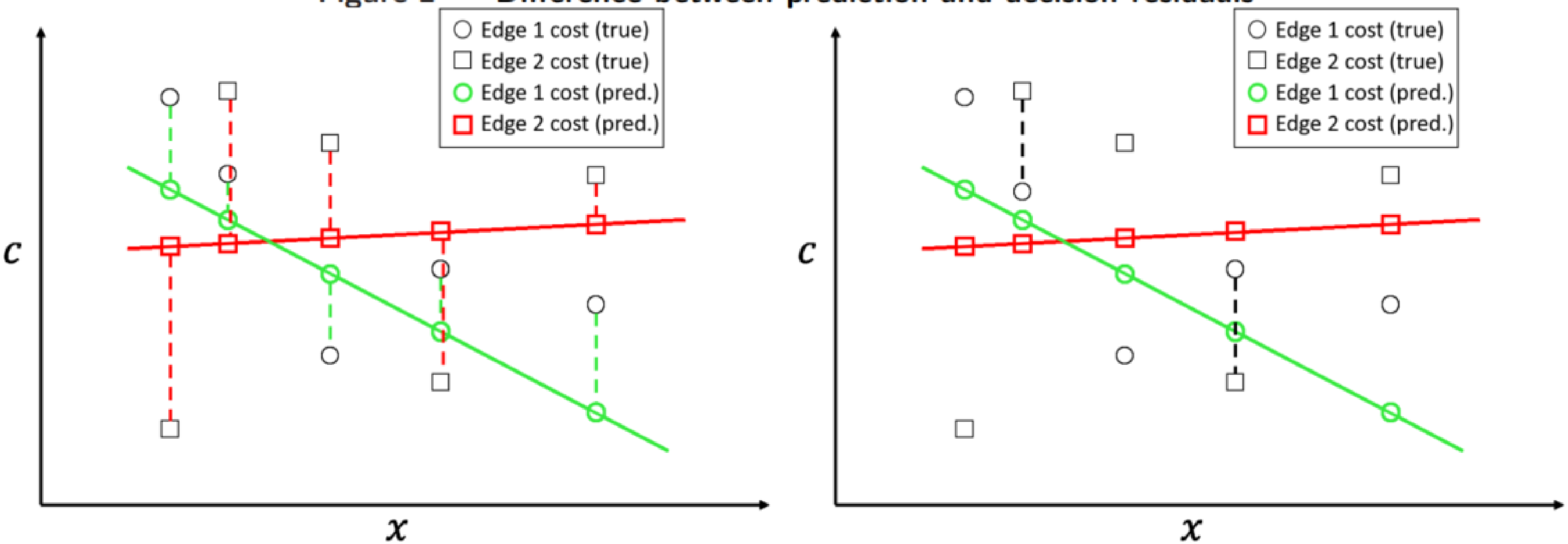


(a) Polyhedral feasible region (b) Elliptic feasible region

3. Smart “Predict, then Optimize” Framework

- An Illustrative Example

Figure 2 Difference between prediction and decision residuals



(a) Prediction residuals (b) Decision residuals

3. Smart "Predict, then Optimize" Framework

- The SPO+ Loss Function
- $l_{SPO}(\hat{c}, c) \coloneqq max_{w \in W^*(\hat{c})}\, c^T w - z^*(c)$.

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

$$z^*(\hat{c}) = \hat{c}^T w \text{ for all } w \in W^*(\hat{c})$$

$$\ell_{\text{SPO}}(\hat{c}, c) = \max_{w \in W^*(\hat{c})} \{c^T w - \alpha \hat{c}^T w\} + \alpha z^*(\hat{c}) - z^*(c)$$

$$\ell_{\text{SPO}}(\hat{c}, c) \;\le\; \inf_{\alpha} \left\{ \max_{w \in S} \{c^T w - \alpha \hat{c}^T w\} + \alpha z^*(\hat{c}) \right\} - z^*(c) \;.$$

PROPOSITION 2 **(Dual Representation of SPO Loss)**. *For any cost vector prediction $\hat{c} \in \mathbb{R}^d$ and realized cost vector $c \in \mathbb{R}^d$, the function $\alpha \mapsto \max_{w \in S} \{c^T w - \alpha \hat{c}^T w\} + \alpha z^*(\hat{c})$ is monotone decreasing on $\mathbb{R}$, and the true SPO loss function may be expressed as*

$$\ell_{\text{SPO}}(\hat{c}, c) \;=\; \lim_{\alpha \to \infty} \left\{ \max_{w \in S} \{c^T w - \alpha \hat{c}^T w\} + \alpha z^*(\hat{c}) \right\} - z^*(c) \;. \tag{7}$$

3. Smart "Predict, then Optimize" Framework

- The SPO+ Loss Function

$$\ell_{\mathrm{SPO}}(\hat{c}, c) \;=\; \lim_{\alpha\to\infty}\left\{\max_{w\in S}\left\{c^T w - \alpha\hat{c}^T w\right\} + \alpha z^*(\hat{c})\right\} - z^*(c)\;.$$

- The SPO ERM problem

$$\min_{f\in\mathcal{H}}\;\frac{1}{n}\sum_{i=1}^{n}\lim_{\alpha_i\to\infty}\left\{\max_{w\in S}\left\{c_i^T w - \alpha_i f(x_i)^T w\right\} + \alpha_i z^*(f(x_i))\right\} - z^*(c_i)$$

$$\leq\;\min_{f\in\mathcal{H}}\;\frac{1}{n}\sum_{i=1}^{n}\max_{w\in S}\left\{c_i^T w - 2f(x_i)^T w\right\} + 2f(x_i)^T w^*(c_i) - z^*(c_i)\;.$$

DEFINITION 3 (SPO+ LOSS). Given a cost vector prediction $\hat{c}$ and a realized cost vector c, the *SPO+ loss* is defined as $\ell_{\mathrm{SPO+}}(\hat{c}, c) := \max_{w\in S}\left\{c^T w - 2\hat{c}^T w\right\} + 2\hat{c}^T w^*(c) - z^*(c)$.

3. Smart "Predict, then Optimize" Framework

DEFINITION 3 (SPO+ LOSS). Given a cost vector prediction $\hat{c}$ and a realized cost vector c, the *SPO+ loss* is defined as $\ell_{\mathrm{SPO+}}(\hat{c}, c) := \max_{w \in S} \{c^T w - 2\hat{c}^T w\} + 2\hat{c}^T w^*(c) - z^*(c)$.

PROPOSITION 3 (**SPO+ Loss Properties**). *Given a fixed realized cost vector c, it holds that:*

1. $\ell_{\mathrm{SPO}}(\hat{c}, c) \leq \ell_{\mathrm{SPO+}}(\hat{c}, c)$ *for all* $\hat{c} \in \mathbb{R}^d$,
2. $\ell_{\mathrm{SPO+}}(\hat{c}, c)$ *is a convex function of the cost vector prediction* $\hat{c}$, *and*
3. *For any given* $\hat{c}$, $2(w^*(c) - w^*(2\hat{c} - c))$ *is a subgradient of* $\ell_{\mathrm{SPO+}}(\cdot)$ *at* $\hat{c}$, *i.e.*, $2(w^*(c) - w^*(2\hat{c} - c)) \in \partial \ell_{\mathrm{SPO+}}(\hat{c}, c)$.

- Consistency of the SPO+ Loss Function
  - When minimizing the SPO+ loss is equivalent to minimizing the SPO loss

4. Computational Approaches

The SPO+ ERM:

$$\min_{B\in\mathbb{R}^{d\times p}} \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{SPO+}}(Bx_i, c_i) \ + \lambda\Omega(B)$$

$$\mathcal{H} \ = \ \{f \, : f(x) \ = \ Bx \ \text{ for some } B \ \in \mathbb{R}^{d\times p}, \Omega(B) \ \le \ \rho\}$$

$$\Omega(B) = \tfrac{1}{2}\|B\|_F^2$$

$$c^T w = c^T w^*(\hat{c}) = c^T w^*(\hat{B}x)$$

Solve: CPLEX and Gurobi →medium sized problem

Stochastic gradient methods→large scale instances

4. Computational Approaches

$$2(w^*(c) - w^*(2\hat{c} - c)) \in \partial \ell_{\mathrm{SPO+}}(\hat{c}, c)$$

Algorithm 1 Stochastic Subgradient Descent with Mini-Batching for Problem (13)

Initialize $B_0 \in \mathbb{R}^{d \times p}$ (typically $B_0 \leftarrow 0$), $t \leftarrow 0$. Set batch size parameter $N \geq 1$.

At iteration $t \geq 0$:

1. For $j = 1, \ldots, N$:

   Sample i uniformly at random from the set $\{1, \ldots, n\}$.

   Compute $\tilde{w}_t^j \leftarrow w^*(2B_t x_i - c_i)$.

   Set $\tilde{G}_t^j \leftarrow (w^*(c_i) - \tilde{w}_t^j) x_i^T$.

2. Select $\gamma_t > 0$ and compute:

   $\Psi_t \in \partial \Omega(B_t)$

   $G_t \leftarrow \frac{1}{N} \sum_{j=1}^{N} \tilde{G}_t^j \ + \ \lambda \Psi_t$

   $B_{t+1} \leftarrow B_t - \gamma_t G_t$

   $\bar{B}_t \leftarrow \frac{1}{\sum_{s=0}^{t} \gamma_s} \sum_{s=0}^{t} \gamma_s B_s$.

5. Computational Experiments--Portfolio Optimization

1. the previously described SPO+ method

2. $l(\hat{c}, c) = \frac{1}{2}\|\hat{c} - c\|_2^2$

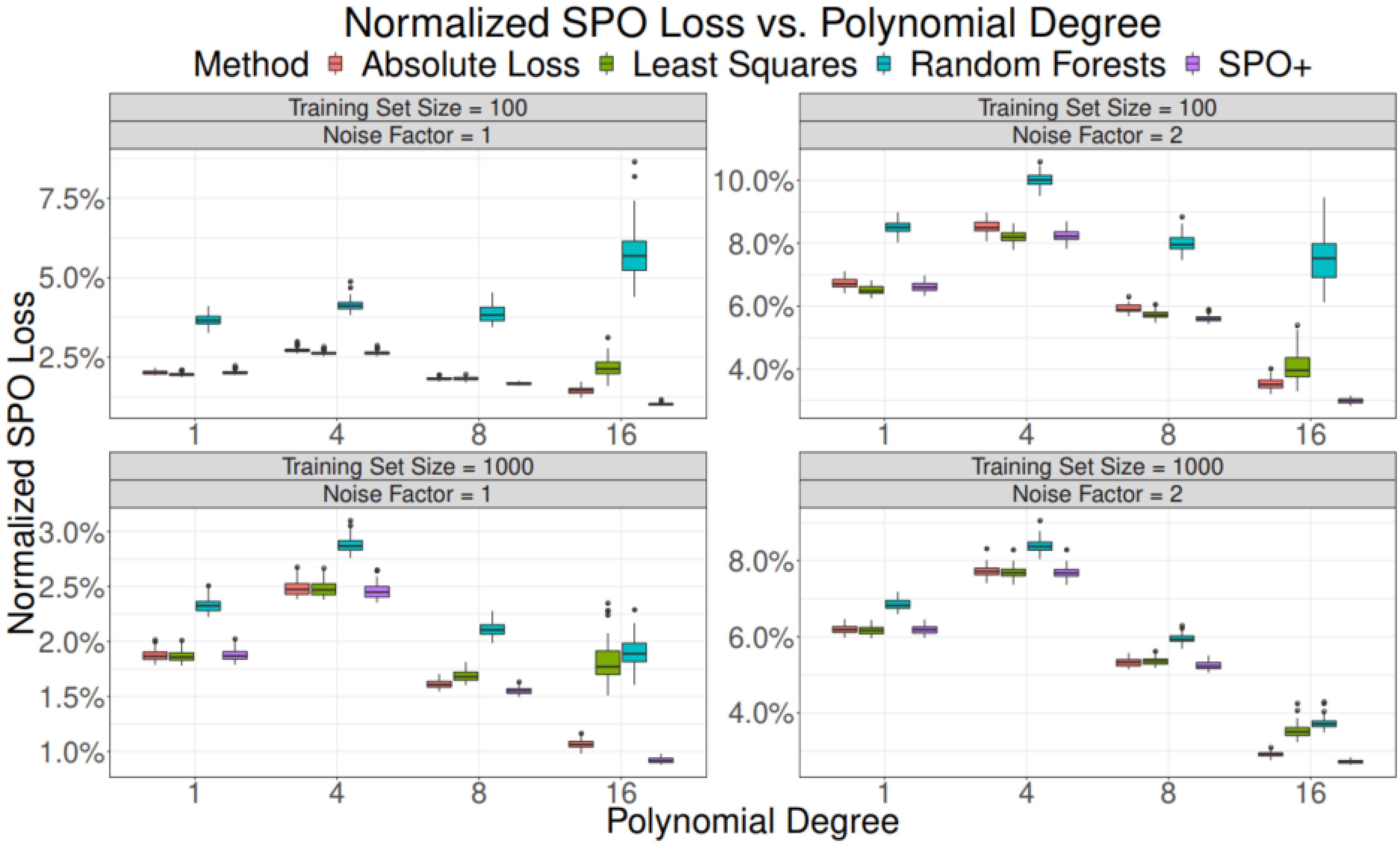
3. $l(\hat{c}, c) = \frac{1}{2}\|\hat{c} - c\|_1$

4. A random forests approach

$$\min_{B \in \mathbb{R}^{d \times p}} \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathrm{SPO+}}(Bx_i, c_i) + \lambda \Omega(B)$$

Data generate:

1. The conditional mean $\bar{r}_{ij}$ of the j^{th} asset return is set equal to $\bar{r}_{ij} := \left(\frac{0.05}{\sqrt{p}}(B^* x_i)_j + (0.1)^{1/\deg}\right)^{\deg}$, where deg is a fixed positive integer parameter.
2. The observed return vector $\tilde{r}_i$ is set to $\tilde{r}_i := \bar{r}_i + Lf + 0.01\tau\varepsilon$, where $f \sim N(0, I_4)$ and $\varepsilon \sim N(0, I_{50})$. The cost vector c_i is set to $c_i := -\tilde{r}_i$.



$$\mathrm{NormSPOTest}(\hat{f}) := \frac{\sum_{i=1}^{n_{\mathrm{test}}} \ell_{\mathrm{SPO}}(\hat{f}(\tilde{x}_i), \tilde{c}_i)}{\sum_{i=1}^{n_{\mathrm{test}}} z^*(\tilde{c}_i)}$$

V. Conclusion

- We provide a new SPO framework for developing prediction models under the predict-then-optimize paradigm.
- We also derived the convex SPO+ loss function.
- Empirically, SPO+ strongly outperforms all approaches when there is model misspecification.