Generalization Bounds in the Predict-then-Optimize Framework

Othman El Balghiti (Rayens Capital), Adam N. Elmachtoub (Columbia University), Paul Grigas (University of California, Berkeley) and Ambuj Tewari (University of Michigan)

NeurIPS 2019

Outline of Topics

- Predict-then-optimize framework and preliminaries
- Combinatorial dimension based generalization bounds
- Margin-based generalization bounds under strong convexity
- Conclusions and future directions

Motivation

- Large-scale optimization problems arising in practice almost always involve unknown parameters
- Often there is a relationship between the unknown parameters and some contextual/auxiliary data
- Given historical data, one approach is to build a predictive statistical/machine learning model from data (e.g. using linear regression)
  - First predict the unknown parameters, then optimize given the predictions
  - Predict phase and optimize phase are naively decoupled
  - There is an opportunity for the prediction model to be informed by the downstream optimization task

Contextual Stochastic Linear Optimization

We consider stochastic optimization problems of the form:

$$\begin{aligned} \min_{w} \quad & \mathbb{E}_{c\sim\mathcal{D}_x}\left[c^T w\right] \\ \text{s.t.} \quad & w \in S \end{aligned}$$

Notation:

- S is a given convex and compact set
- c is an unknown cost vector of the linear objective function
- $\mathcal{D}_x$ is the conditional distribution of c given an auxiliary feature/context vector $x \in \mathbb{R}^p$

Various approaches for dealing with the above problem in the literature: often without constraints, with very simple constraints, or without directly accounting for the optimization structure

Contextual Stochastic Linear Optimization, cont.

$$\begin{aligned} \min_{w} \quad & \mathbb{E}_{c\sim\mathcal{D}_x}\left[c^T w\right] \\ \text{s.t.} \quad & w \in S \end{aligned}$$

Notice that the linearity assumption implies that

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

Hence, it is sufficient to focus on estimating/predicting the vector $\mathbb{E}_{c\sim\mathcal{D}_x}[c|x]$

Predict-then-optimize (PO) Paradigm

We define $P(\hat{c})$ to be the optimization task with predicted cost vector $\hat{c}$

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

$w^*(\hat{c})$ denotes an arbitrary optimal solution of $P(\hat{c})$

Predict-then-Optimize (PO) Paradigm

- Given a new feature vector x, predict $\hat{c}$ based on x
- Make decision $w^*(\hat{c})$
- Incur cost $c^T w^*(\hat{c})$ with respect to the actual ("true") realized c

Predict-then-Optimize (PO) Loss Function

Within the predict-then-optimize paradigm, we can naturally define a loss function referred to as the "Smart predict-then-optimize" (SPO) loss function [Elmachtoub and G 2017]:

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

Given historical training data $(x_1, c_1), \ldots, (x_n, c_n)$ and a hypothesis class $\mathcal{H}$ of cost vector prediction models (i.e., $f : \mathbb{R}^p \to \mathbb{R}^d$ for $f \in \mathcal{H}$), the ERM principle yields:

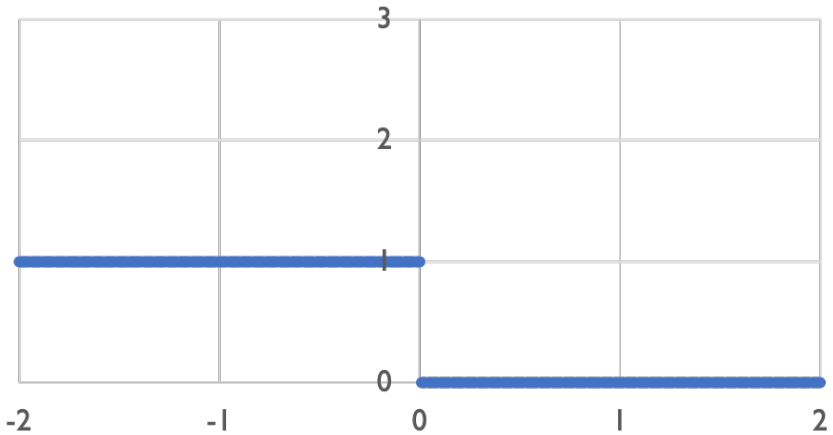
Empirical Risk Minimization with the SPO Loss

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

Binary and Multiclass Classification as a Special Case

It turns out that the SPO loss is a special case of the classical 0-1 loss in binary classification

- This equivalence happens with $S = [-1/2, +1/2]$ and $c \in \{-1, +1\}$
- This example can also be generalized to multiclass classification where S is now the unit simplex



Empirical Risk Minimization with the SPO Loss

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

It turns out that the SPO loss is nonconvex, and in fact may be discontinuous depending on the structure of S

Thus, the above optimization problem is challenging even for simple hypothesis classes such as linear functions $\mathcal{H} = \{x \mapsto Bx : B \in \mathbb{R}^{d \times p}\}$

There are several approaches for addressing this problem computationally

- An appealing idea is based on a surrogate loss function approach (see, e.g., [Elmachtoub and G 2017], [Ho-Nguyen and Kilinc-Karzan 2019])

Generalization Bounds for the SPO Loss

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

The focus of this work is not on optimization for the above problem, but on generalization

- Generalization bounds verify that trying to solve the above problem (based on training data) is at all reasonable

Let us define the empirical and expected SPO loss as:

$$\hat{R}_{\mathrm{SPO}}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathrm{SPO}}(f(x_i), c_i)\,, \text{ and } R_{\mathrm{SPO}}(f) := \mathbb{E}_{(x,c)\sim\mathcal{D}}\left[\ell_{\mathrm{SPO}}(f(x), c)\right]$$

Generalization Bounds for the SPO Loss

$$\hat{R}_{\mathrm{SPO}}(f) := \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{SPO}}(f(x_i), c_i) \ , R_{\mathrm{SPO}}(f) := \mathbb{E}_{(x,c)\sim\mathcal{D}}\left[\ell_{\mathrm{SPO}}(f(x), c)\right]$$

A generalization bound relates the above two quantities and verifies that minimizing the empirical loss also (approximately) minimizes the expected loss

Importantly the bound should hold uniformly over $f \in \mathcal{H}$ and with high probability over $(x_i, c_i) \sim \mathcal{D}^n$

A generalization bound implies an "on average" (over x) guarantee for the problem of interest:

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

Rademacher Complexity and Generalization

We follow a standard approach to establishing generalization bounds based on Rademacher compelxity

Given the observed data $(x_1, c_1), \ldots, (x_n, c_n)$, define the empirical Rademacher complexity of $\mathcal{H}$ w.r.t. to the SPO loss as:

$$\hat{\mathfrak{R}}^n_{\mathrm{SPO}}(\mathcal{H}) := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell_{\mathrm{SPO}}(f(x_i), c_i) \right] ,$$

where σ_i are i.i.d. Rademacher random variables uniformly distributed on $\{-1, +1\}$

Let us also assume that $\ell_{\mathrm{SPO}} \in [0, \omega]$ for some $\omega > 0$, which follows from the boundedness of S and the distribution of c

Rademacher Complexity and Generalization, cont.

The following is a celebrated result yielding a generalization bound based on Rademacher complexity

Theorem [Bartlett and Mendelson 2002]

Let $\mathcal{H}$ be a family of functions mapping from $\mathbb{R}^p$ to $\mathbb{R}^d$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample drawn from the distribution $\mathcal{D}$, the following holds for all $f \in \mathcal{H}$

$$R_{\mathrm{SPO}}(f) \leq \hat{R}_{\mathrm{SPO}}(f) + 2\hat{\mathfrak{R}}^n_{\mathrm{SPO}}(\mathcal{H}) + 3\omega\sqrt{\frac{\log(2/\delta)}{2n}} \ .$$

The remaining challenge is to bound $\hat{\mathfrak{R}}^n_{\mathrm{SPO}}(\mathcal{H})$, which is difficult due to the nonconvex and discontinuous nature of the SPO loss

Bounds Based on Combinatorial Dimension

Let us first consider the case where:

- S is a polytope with set of extreme points $\mathfrak{S}$
- $\mathcal{H} = \mathcal{H}_{\text{lin}} := \{x \mapsto Bx : B \in \mathbb{R}^{d \times p}\}$ is the set of linear predictors

Theorem

Under the above two conditions, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample drawn from the distribution $\mathcal{D}$, the following holds for all $f \in \mathcal{H}_{\text{lin}}$

$$R_{\text{SPO}}(f) \leq \hat{R}_{\text{SPO}}(f) + 2\omega\sqrt{\frac{2dp\log(n|\mathfrak{S}|^2)}{n}} + \omega\sqrt{\frac{\log(1/\delta)}{2n}}$$

Bounds Based on Combinatorial Dimension, cont.

Proof of the previous theorem is based on "reducing" the problem to a multiclass classification problem where the classes correspond to the extreme points of S

- This is not a complete reduction, since the SPO loss function is more complicated
- We can then leverage the notion of Natarajan dimension [Natarajan 1989], which is an extension of VC-dimension to the multiclass case
- Key result is relating the SPO Rademacher complexity to the Natarajan dimension
- Related techniques appeared recently in [Gupta and Kallus 2019]

Extension to Convex Sets

Using a discretization argument, we can extend the previous result to any bounded convex set S

- We presume that $\|w\|_2 \leq \rho_w$ for all $w \in S$

Theorem

In the case of linear predictors and general compact and convex S, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample drawn from the distribution $\mathcal{D}$, the following holds for all $f \in \mathcal{H}_{\text{lin}}$

$$R_{\text{SPO}}(f) \leq \hat{R}_{\text{SPO}}(f) + 4d\omega\sqrt{\frac{2p\log(2n\rho_w d)}{n}} + 3\omega\sqrt{\frac{\log(2/\delta)}{2n}} + O\left(\frac{1}{n}\right)$$

Question: Can we improve the dependence on the dimensions d and p and replace them with more "natural" quantities?

Extension to Convex Sets

Using a discretization argument, we can extend the previous result to any bounded convex set S

- We presume that $\|w\|_2 \leq \rho_w$ for all $w \in S$

Theorem

In the case of linear predictors and general compact and convex S, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample drawn from the distribution $\mathcal{D}$, the following holds for all $f \in \mathcal{H}_{\text{lin}}$

$$R_{\text{SPO}}(f) \leq \hat{R}_{\text{SPO}}(f) + 4d\omega\sqrt{\frac{2p\log(2n\rho_w d)}{n}} + 3\omega\sqrt{\frac{\log(2/\delta)}{2n}} + O\left(\frac{1}{n}\right)$$

Question: Can we improve the dependence on the dimensions d and p and replace them with more "natural" quantities?

Strongly Convex Sets

We now make the additional assumption that S is μ-strongly convex with respect to the ℓ_2-norm:

- Namely, for any $w_1, w_2 \in S$ and $\lambda \in [0,1]$, it holds that a ball centered at $\lambda w_1 + (1-\lambda)w_2$ of radius $\left(\frac{\mu}{2}\right)\lambda(1-\lambda)\|w_1 - w_2\|^2$ is contained in S
- Examples include certain norm balls and Schatten norm balls, as well as level sets of smooth and strongly convex functions

Intuitively, strong convexity of S implies that linear optimization over S is "poorly behaved" only when c is near zero

Formally, we can prove that the linear optimization oracle $w^*(\cdot)$ must satisfy the following "Lipschitz-like" property:

$$\|w^*(\hat{c}_1) - w^*(\hat{c}_2)\|_2 \leq \frac{1}{\mu \cdot \min\{\|\hat{c}_1\|_2, \|\hat{c}_2\|_2\}}\|\hat{c}_1 - \hat{c}_2\|_2 \text{ for any } \hat{c}_1, \hat{c}_2 \in \mathbb{R}^d$$

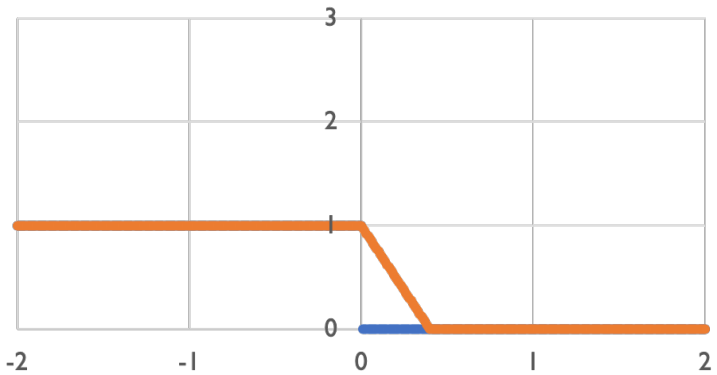
Correcting the SPO Loss Near Zero

Recall the definition of the SPO loss $\ell_{\text{SPO}}(\hat{c}, c) := c^T(w^*(\hat{c}) - w^*(c))$ and that $\ell_{\text{SPO}} \in [0, \omega]$

We are motivated to "correct" the poor behavior of this loss function near zero by considering the "γ-margin SPO loss" defined by:

$$\ell^{\gamma}_{\text{SPO}}(\hat{c}, c) := \begin{cases} \ell_{\text{SPO}}(\hat{c}, c) & \text{if } \|\hat{c}\|_2 > \gamma \\ \left(\frac{\|\hat{c}\|_2}{\gamma}\right) \ell_{\text{SPO}}(\hat{c}, c) + \left(1 - \frac{\|\hat{c}\|_2}{\gamma}\right) \omega & \text{if } \|\hat{c}\|_2 \leq \gamma \end{cases}$$

Analogue in binary classification is the ramp loss:



Correcting the SPO Loss Near Zero

The γ-margin SPO loss is defined by:

$$\ell^{\gamma}_{\mathrm{SPO}}(\hat{c}, c) := \begin{cases} \ell_{\mathrm{SPO}}(\hat{c}, c) & \text{if } \|\hat{c}\|_2 > \gamma \\ \left(\frac{\|\hat{c}\|_2}{\gamma}\right) \ell_{\mathrm{SPO}}(\hat{c}, c) + \left(1 - \frac{\|\hat{c}\|_2}{\gamma}\right) \omega & \text{if } \|\hat{c}\|_2 \leq \gamma \end{cases}$$

Based on the "Lipschitz-like" property of the optimization oracle we can prove that $\ell^{\gamma}_{\mathrm{SPO}}(\cdot, c)$ is a Lipschitz function:

Theorem

For any fixed $c \in \mathbb{R}^d$ and $\gamma > 0$, it holds that:

$$|\ell^{\gamma}_{\mathrm{SPO}}(\hat{c}_1, c) - \ell^{\gamma}_{\mathrm{SPO}}(\hat{c}_2, c)| \leq \frac{5\rho_c}{\gamma\mu}\|\hat{c}_1 - \hat{c}_2\|_2 \quad \text{for all } \hat{c}_1, \hat{c}_2 \in \mathbb{R}^d ,$$

where ρ_c is such that $\|c\|_2 \leq \rho_c$ with probability 1

Improved Generalization Bound

Following [Bertsimas and Kallus 2014] and [Maurer 2016], we define the multivariate Rademacher complexity of $\mathcal{H}$ as:

$$\hat{\mathfrak{R}}^n(\mathcal{H}) := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \sigma_{ij} f_j(x_i) \right]$$

where σ_{ij} are i.i.d. Rademacher random variables

Theorem

Suppose that S is μ-strongly convex and let $\gamma > 0$ be fixed. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample drawn from the distribution $\mathcal{D}$, the following holds for all $f \in \mathcal{H}$

$$R_{\mathrm{SPO}}(f) \leq \hat{R}^{\gamma}_{\mathrm{SPO}}(f) + \frac{10\sqrt{2}\rho_c \hat{\mathfrak{R}}^n(\mathcal{H})}{\gamma\mu} + 3\omega\sqrt{\frac{\log(2/\delta)}{2n}}$$

Notice that $\hat{R}^{\gamma}_{\mathrm{SPO}}(f)$ is the empirical γ-margin SPO loss

Improved Generalization Bound

Following [Bertsimas and Kallus 2014] and [Maurer 2016], we define the multivariate Rademacher complexity of $\mathcal{H}$ as:

$$\hat{\mathfrak{R}}^n(\mathcal{H}) := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \sigma_{ij} f_j(x_i) \right]$$

where σ_{ij} are i.i.d. Rademacher random variables

Theorem

Suppose that S is μ-strongly convex and let $\gamma > 0$ be fixed. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample drawn from the distribution $\mathcal{D}$, the following holds for all $f \in \mathcal{H}$

$$R_{\mathrm{SPO}}(f) \leq \hat{R}^{\gamma}_{\mathrm{SPO}}(f) + \frac{10\sqrt{2}\rho_c \hat{\mathfrak{R}}^n(\mathcal{H})}{\gamma\mu} + 3\omega\sqrt{\frac{\log(2/\delta)}{2n}}$$

Notice that $\hat{R}^{\gamma}_{\mathrm{SPO}}(f)$ is the empirical γ-margin SPO loss

Improved Generalization Bound, cont.

Margin-based Generalization Bound

$$R_{\mathrm{SPO}}(f) \leq \hat{R}^{\gamma}_{\mathrm{SPO}}(f) + \frac{10\sqrt{2}\rho_c \hat{\mathfrak{R}}^n(\mathcal{H})}{\gamma\mu} + 3\omega\sqrt{\frac{\log(2/\delta)}{2n}}$$

Some comments:

- The proof of the above bound is based on a vector contraction inequality for Lipschitz functions [Maurer 2016]
- Notice that there is no direct dependence on the dimensions, instead the bound involves the more natural/analytic quantity $\hat{\mathfrak{R}}^n(\mathcal{H})$
- In many situations, we can further bound $\hat{\mathfrak{R}}^n(\mathcal{H})$ using well-established techniques (see, e.g., [Kakade, Sridharan, Tewari 2009])
- It is straightforward to also extend/modify this result for situations where the strong convexity is embedded in the cost function

"Margin-based" Generalization Bound?

Margin-based Generalization Bound

$$R_{\mathrm{SPO}}(f) \leq \hat{R}^{\gamma}_{\mathrm{SPO}}(f) + \frac{10\sqrt{2}\rho_c \hat{\mathfrak{R}}^n(\mathcal{H})}{\gamma\mu} + 3\omega\sqrt{\frac{\log(2/\delta)}{2n}}$$

Some comments:

- The parameter γ is analogous to the margin in binary classification, whereby we require a classifier to not only be correct but to correct by a certain margin
- The above bound is not meaningful for every distribution, instead it is effective when the distribution $\mathcal{D}$ has a "favorable margin"
- In fact, the above is an exact extension of a well-known result for binary classification [Koltchinskii, Panchenko, et al. 2002]

Extension to a Uniform Result Over γ

Margin-based Generalization Bound

$$R_{\mathrm{SPO}}(f) \leq \hat{R}^{\gamma}_{\mathrm{SPO}}(f) + \frac{10\sqrt{2}\rho_c \hat{\mathfrak{R}}^n(\mathcal{H})}{\gamma\mu} + 3\omega\sqrt{\frac{\log(2/\delta)}{2n}}$$

Notice that the above bound involves a tradeoff with respect to γ:

- Smaller value of γ yields $\hat{R}^{\gamma}_{\mathrm{SPO}}(f) \approx \hat{R}_{\mathrm{SPO}}(f)$
- Large value of γ yields a better Lipschitz constant, hence the $O(1/\gamma)$ term is smaller

We can actually extend this bound to a result that holds uniformly over $\gamma \in (0, \bar{\gamma}]$ with only an additional logarithmic factor

- Given a dataset and a predictor f computed on that dataset, one can do a line search over γ to obtain the best bound
- A uniform result over $\gamma \in (0, \bar{\gamma}]$ makes this line search procedure statistically valid

Conclusions and Future Work

Conclusions:

- Reviewed the predict-then-optimize framework to predict cost vectors for the purpose of solving a downstream optimization task
- Provided combinatorial generalization bounds in the polyhedral and general convex settings
- Provided improved margin-based generalization bounds under strong convexity

Many exciting future directions:

- Extending the margin theory to polyhedral and general convex sets
- Developing improved bounds in other situations, e.g., perhaps based on local Rademacher complexity
- Minimax lower bounds
- Furthering the theory of surrogate losses, nonlinear cost functions, etc.