Statistical Properties of Robust Satisficing

Abstract

Robust Satisficing is an emerging robust optimization method. However, it lacks the statistical guarantees of its widely-studied counterpart, DRO. This paper addresses this gap by deriving two-sided confidence intervals for optimal loss and finite-sample generalization error bounds under Wasserstein distance, extending also to distribution shifts. For f-divergence, we establish an asymptotic upper bound on generalization error. Numerical experiments show that RS outperforms ERM in small-sample regimes and under distribution shifts, and shows hyperparameter correspondence with DRO.

Keywords *Robust Satisficing, Distributionally Robust Optimization, Wasserstein distance, f-divergence, Generalization error*

1. Research Problem The Robust Satisficing (RS) model proposed by Long, Sim, and Zhou addresses the key drawbacks of classical robust method——Distributionally Robust Optimization (DRO), such as over-conservatism and the challenge of selecting an appropriate radius for the ambiguity set. The RS model shifts the focus from minimizing the worst-case loss to a satisficing strategy that balances performance and robustness more effectively.

The RS model is formulated as follows:

$$\begin{aligned} k_\tau = \min \quad & k \\ \text{s.t. } & \mathbb{E}_P[h(x,\xi)] - \tau \le k d(P, \hat{P}_N), \quad \forall P \in \mathbb{P} \\ & \boldsymbol{x} \in \mathcal{X}, \quad k \ge 0. \end{aligned} \tag{1}$$

In this formulation, $h(x,\xi)$ is the objective function with decision value x and random variable ξ, $\hat{P}_N$ is the empirical distribution with samples generated from true distribution P^*, and $d(\cdot,\cdot)$ is a measure that characterizes the discrepancy between distributions; here we consider the Wasserstein distance and f-divergence. RS uses a reference value τ as a hyperparameter, ensuring that any excess loss over this reference value is controlled by a multiple of the distance between distributions.

This approach avoids DRO's over-conservatism and potentially provides better generalization performance on the target distribution. However, the RS model lacks comprehensive statistical guarantees.

Our work delves into the statistical theory of the RS model, focusing on deriving and analyzing its statistical properties. We provide two-sided confidence intervals for the optimal loss and non-asymptotic upper bounds for the generalization error under the Wasserstein distance. We also extend our analysis to scenarios involving distribution shifts, where we present confidence intervals and generalization error bounds for the RS model optimizer. Additionally, we explore the case of f-divergence and provide an asymptotic upper bound on the generalization error. These results fill a crucial gap in the literature, demonstrating the RS model's robustness and practical advantages over DRO, particularly in small-sample regimes and under distribution shifts.

2. Key Methodology and Assumptions One of the key methodologies of this work is the systematic selection of the hyperparameter τ in the RS model. we choose $\tau_\epsilon := \inf_x \mathbb{E}_{\hat{P}_N}[h(x,\xi)] + \epsilon$, where ϵ is referred to as "tolerance value" that the RS model allows for excess empirical loss. We adopt this approach, focusing on characterizing the role of ϵ in the statistical guarantees provided by the RS model.

Another key methodology is the dual reformulation of the RS model. Under the Wasserstein distance, its dual reformulation is as follows:

$$\begin{aligned} \min \quad & k \geq 0 \\ \text{s.t.} \quad & \mathbb{E}_{\hat{P}_N}[\sup_{z\in\Xi} h(x,z) - kc(\xi,z)] \leq \tau. \\ & x \in \mathcal{X} \end{aligned} \tag{2}$$

Under f-divergence, its dual reformulation is as follows:

$$\begin{aligned} \min \quad & k \geq 0 \\ \text{s.t.} \quad & \min_{\mu}(\mu + kE_{\hat{P}_N} f^*(\frac{h(x,\xi)-\mu}{k})) \leq \tau. \\ & x \in \mathcal{X} \end{aligned} \tag{3}$$

Below are the assumptions required for our results.

Assumption 1 (Exponential tail decay in random variable)**.** *There exists an $a > 1$, such that $\mathbb{E}_{P^*}[\exp(||\xi||^a)] < \infty$.*

Assumption 2 (Lipschitz continuity of loss function)**.** *The loss $h(x,\xi)$ is Lipschitz with a uniform constant L in ξ.*

3. Main Results Denote $\hat{x}_N$ as the optimizer of RS model. Define optimal loss $J^* = \inf_{x\in\mathcal{X}} \mathbb{E}_{P^*}[h(x,\xi)]$ and generalization error $R(P^*, \hat{x}_N) = \mathbb{E}_{P^*}[h(\hat{x}_N,\xi)] - J^*$. Then under the Wasserstein distance, we have:

Theorem 1 (Confidence intervals of optimal loss)**.** *Suppose Assumptions 1 & 2 hold. For any N, let β_N be the confidence level. We have with probability at least $1-\beta_N$:*

$$-L \cdot r_N + (\tau_\epsilon - \epsilon) \leq J^* \leq \mathbb{E}_{P^*}[h(\hat{x}_N,\xi)] \leq L \cdot r_N + \tau_\epsilon, \tag{4}$$

where r_N, denoted as the "remainder", is solved from the below equation:

$$\beta_N = \begin{cases} c_1 \exp\big(-c_2 N r_N^{\max\{m,2\}}\big) & \text{if } r_N \leq 1, \\ c_1 \exp\big(-c_2 N r_N^a\big) & \text{if } r_N > 1, \end{cases}$$

with c_1, c_2 as positive constants that only depend on exponential decay rate a and the dimension m of ξ.

Theorem 2 (Finite sample generalization error bound)**.** *Suppose Assumptions 1 & 2 hold. With probability at least $1-\beta_N$, we have:*

$$R(P^*, \hat{x}_N) \leq \epsilon + 2L \cdot r_N, \tag{5}$$

where r_N is the reminder solved as in Theorem 1. Taking expectation with respect to data, we have:

$$\mathbb{E}_{P^*}[R(P^*, \hat{x}_N)] \leq \epsilon + O(L \cdot N^{-\min\{\frac{1}{m}, \frac{1}{2}\}}). \tag{6}$$

In distribution shift scenario, we evaluate the performance when applying $\hat{x}_N$ to another distribution $\tilde{P}$, which may shift from P^*, resulting in a certain degree of discrepancy. Define the optimal loss under the new distribution $\tilde{P}$ as $\tilde{J} = \inf_{x\in\mathcal{X}} \mathbb{E}_{\tilde{P}}[h(x,\xi)]$. Then we have:

Theorem 3 (Distribution Shift)**.** *Suppose Assumptions 1 & 2 hold. For any N, let β_N be some nominal confidence level. We have with probability at least $1-\beta_N$:*

$$-L\cdot r_N - L\cdot d_W(P^*,\tilde{P}) + (\tau_\epsilon - \epsilon) \le \tilde{J} \le \mathbb{E}_{\tilde{P}} h(\hat{x}_N,\xi) \le k_{\tau_\epsilon}\cdot r_N + k_\tau \cdot d_W(P^*,\tilde{P}) + \tau_\epsilon,$$

and

$$R(\tilde{P},\hat{x}_N) \le \epsilon + 2L\cdot d_W(P^*,\tilde{P}) + 2L\cdot r_N,$$

where the reminder r_N is solved as Theorem 1.

Taking the expectation on data, we have:

$$\mathbb{E}_{P^*}\left[R(\tilde{P},\hat{x}_N)\right] \le \epsilon + 2L\cdot d_W(P^*,\tilde{P}) + O\left(L\cdot N^{-\min\{\frac{1}{m},\frac{1}{2}\}}\right).$$

This result shows that results under distribution shifts merely require adding a multiple of the shift distance.

Finally, we extend the Robust Satisficing model to the commonly used f-divergences in machine learning.

Theorem 4 (Generalization error bound)**.** *Assuming that P^* is a discrete distribution with finite support Ξ. Choosing $\epsilon_N = C\sqrt{\frac{M|\Xi|\log N}{N}}$, For the Hellinger distance $d_H(P,Q) = \frac{1}{\sqrt{2}}\left(\int\left(\sqrt{dP}-\sqrt{dQ}\right)^2\right)^{\frac{1}{2}}$, the Le Cam distance $d_{LC}(P,Q) = \frac{1}{2}\int\frac{(dP-dQ)^2}{dP+dQ}$, and the Jensen-Shannon divergence $d_{JS}(P,Q) = d_{KL}(P,\frac{P+Q}{2}) + d_{KL}(Q,\frac{P+Q}{2})$, when $N\to\infty$, we have:*

$$\mathbb{E}_{P^*} R(P^*,\hat{x}_N) = O\Big(\sqrt{\frac{M|\Xi|\log N}{N}}\Big). \tag{7}$$

4. Numerical Experiments The experiment is conducted under the regression loss function $L(y-\langle x,\beta\rangle)$, where β is a parameter and the random variable is $\xi = (x,y)$. We conducted experiments on the RS performance both in small-sample regimes and under distribution shift ($\bar{\epsilon}$ represents ϵ scaled by the empirical loss).

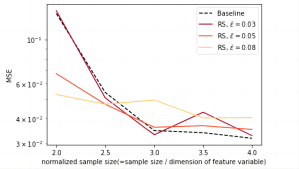

Figure 1: Performances across various sample sizes. RS outperforms the ERM baseline in small-sample regimes.

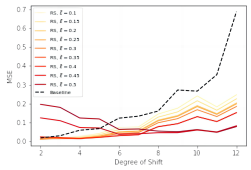

Figure 2: Performances across various degree of distribution shifts. RS outperforms the ERM baseline under distribution shifts.

Finally, we establish the correspondence of hyperparameters between RS and DRO under Lipschitz continuous loss functions. This determination is encapsulated by an optimization problem:

$$\sup_{r>0}\inf_x \frac{\frac{1}{N}\sum_{i=1}^N L(y_i-\langle x_i,\beta\rangle) + r\cdot||\beta||_2 - \tau}{r}. \tag{8}$$