Distributionally Robust Regularization of Sparse Integer Programming Trained Learning Models

Sanjeeb Dash Soumyadip Ghosh João Gonçalves Mark S. Squillante
Mathematics of Computation, IBM Research, Yorktown Heights, USA
sanjeebd,ghoshs,jpgoncal,mss@us.ibm.com

Abstract

Building explainable machine learning models is crucial for human users to be able to interpret the proposed statistical relationship obtained from the training data. Mixed integer optimization formulations are often used to train such models with explicit sparsity constraints, aiming to hit the right trade off between sparsity and prediction accuracy. Existing methods to find the right choice of sparsity – e.g., via cross-validation – are computationally expensive. For convex model training formulations, recent advances in distributionally robust optimization (DRO) provide strong generalization while sidestepping this computational burden. We describe an extension of such regularization via DRO to mixed integer sparse programs, providing statistical guarantees as a function of an associated sparsity parameter of the formulation. We illustrate the use of this approach in the case of building explainable binary classification models using sets of feature value rules.

1 Introduction

The need for machine learning (ML) models to be easily understandable to humans is critically important in many application domains, whereas many popular ML methods are hard to interpret and analyze. In applications having societal impact, such as criminal justice, medicine and pharmacology, ML models that are not interpretable may not be satisfactory [24, 25]. These trends have led in recent years to a considerable emphasis on explainability and interpretability in ML models, including methods such as LIME [23] that explain the predictions of a classifier by approximating it locally with an interpretable model. Such trends have also led to revisiting rule sets [7, 26], rule lists [16], and decision trees [14] for interpretable classification, as well as constructing interpretable ensembles of models hand-picked for explainability [9].

The need for interpretable ML models has resulted in many sparse problem formulations in the literature; refer to, e.g., [1, 27, 13]. For an integer $n \in \mathbb{N}$, define $[n] := \{1, \ldots, n\}$. In this paper we consider stochastic problems of the form:

$$L_\dagger^* := \min_{v,w} L(v, w) = \mathbb{E}_{\xi \sim P^\dagger} l(v, w, \xi) \;\text{ s.t. }\; g_j(v, w) \le 0,\ j \in [J];\ \ h_m(w) \le 0,\ m \in [M], \quad (1)$$

where $v \in \mathbb{R}^d$ and $w \in \mathbb{Z}^p$ are real and integer-valued variables, respectively, and $d, p, J, M \in \mathbb{N}$. The constraints h_m may include constraints enforcing sparsity. We assume $l(\cdot, w, \xi)$ and $g_j(\cdot, w)$ are convex for fixed w and ξ.

The only information provided to the user on the true distribution $P^\dagger$ of ξ (which models, for example, a data generation process) is a dataset of observations $\mathcal{T} := \{\xi_n,\ n \in [N]\}$. A natural counterpart to (1) is then to replace the expectation with its empirical average taken over $\mathcal{T}$. One of the foundational concerns of statistical learning is to ensure that the optimal model parameters obtained by solving the empirical counterpart does not overfit the dataset of observations $\mathcal{T}$. For the

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

version of (1) with only continuous variables, methods such as LASSO regularization add a weighed regularization term to the objective in order to penalize solutions with many components of v that take non-zero values. Sparse formulations explicitly encode this using the deterministic constraints h_m on w. For instance, setting $d = p$ and imposing $w \in \{0,1\}^p$, one can ensure that each v_i is non-zero only if w_i is 1, and then sparsity is controlled via a constraint $\sum_i w_i \leq C$. While sparsity constraints provide more direct control on regularization, the issue of hyperparameter tuning is not alleviated since, as in the continuous case with the weight of the penalty term, the parameter C needs to be well tuned. Following the standard procedure of cross-validation can be expensive with the need to repeatedly solve the integer program (IP) (1).

Our Contributions: We present a general framework to regularize the mixed-integer convex program (MICP) (1) based on a distributionally robust optimization (DRO) reformulation. We give statistical guarantees in Section 2 and show that solutions to the DRO program provide adequate coverage of the optimal performance under the true (unknown) distribution in (1), where we relate the key parameters of the DRO formulation to the corresponding parameters of the sparsity formulation in (1). Of equal importance is the computationally challenge that arises in finding good candidate solutions to the DRO formulation of the typically expensive to solve original MICP (1). Unlike DRO formulations of convex programs that have generic meta-heuristic prescriptions [12], the DRO-MICP version requires case-dependent attention. In Section 3, we describe an MICP formulation for sparse explainable binary classification, the implications of our results on the statistical guarantees for DRO formulations of MICPs, and our algorithm to solve its DRO reformulation. We conclude with a small set of results illustrating the computational savings of using the DRO regularization.

2 Distributionally Robust Regularization

As an alternative to address the costs of cross-validation, techniques from DRO have become popular for improved generalization in efficiently selecting models with continuous-valued variables. There is a rich literature on parameterizing the DRO formulation of the convex model selection program appropriately to provide statistical guarantees without needing expensive hyperparameter tuning [2, 18, 20]. The out-of-sample performance of stochastic optimization formulations has been analyzed [3, 10, 17, 18] utilizing the empirical likelihood methodology [21]. Moreover, [12] prescribe a general procedure that solves these concave-convex formulations efficiently to produce models of similar or higher generalization quality on a significantly lower computational budget. Denote by P^0 the equal-weight distribution over $\mathcal{T}$ and let e be the vector of all ones. The *robust* loss $R(v, w)$ is:

$$R(v,w) := \max_{P \in \mathcal{P}(\rho)} \mathbb{E}_P \, l(v,w,\xi), \text{ where } \mathcal{P}(\rho) := \left\{P \mid D(P,P^0) \leq \rho, \; e^\top P = 1, \; P \geq 0\right\}. \quad (2)$$

If ρ is clear from the context, we denote $\mathcal{P}(\rho)$ by $\mathcal{P}$. The set $\mathcal{P}$ is centered at the empirical probability mass function (pmf) P^0 and includes all pmfs P that are within a distance $D(P, P^0)$ of ρ from P^0, where the distance is measured by the Kullback-Leibler (KL) divergence: $D(P, P^0) = -\frac{1}{N}\sum_n \log(NP_n)$. The DRO procedure selects the model that solves for the minimizer $R^* := \min_{v,w} R(v, w)$ subject to all the constraints present in (1). In the case we only have continuous-valued variables v in (1) and $l(\cdot, \xi)$ is convex for fixed ξ, it is broadly true [2, 18, 20] that if ρ is well-chosen the robust formulation provides a good statistical proxy for the performance under the true distribution $P^\dagger$. In particular, the optimal robust loss value R^* is an asymptotically (in N) valid upper bound on the optimal loss $L^*_\dagger$ under the true (unknown) distribution. Moreover, typically we need to set $\rho = O(\mathfrak{C})/N$, where $\mathfrak{C}$ is a measure of the explanatory power of the model class.

Our main result similarly relates the radius ρ of the probability ball $\mathcal{P}$ to an asymptotic statistical coverage guarantee on the performance at the true distribution $P^\dagger$ under two assumptions on the MICP formulation (1). First, the residual program upon fixing integral w is convex in v and has $J(w)$ active constraints $g_j(\cdot, w)$. Second, the set $\mathcal{H}$ of unique values for w admitted by the constraints in (1) is finite. Let $|\mathcal{H}|$ stand for the cardinality of $\mathcal{H}$.

Theorem 1 *Let $\bar{J} = \max_{w \in \mathcal{H}} J(w)$. Choose an $\alpha \in (0, 1)$ and let $\tilde{\alpha} = \alpha\,|\mathcal{H}|$. Set $\rho = \chi^2_{\bar{J},1-\tilde{\alpha}}/N$, the $(1-\tilde{\alpha})$-th quantile of the χ^2 distribution with $\bar{J}$ degrees of freedom. Then, we have that $\rho \leq \frac{1}{N}\max\{\bar{J}\,, \log|\mathcal{H}|\}$ and $\lim_{N\to\infty} \mathbb{P}\big(L^*_\dagger \leq R^*\big) = (1-\alpha)$.*

This result indicates, broadly, that results of DRO for convex programs can be extended to MICPs of the form (1) that satisfy our additional assumptions. Our proof employs the tool of likelihood maximization given estimating equations [21, Chapter 3.4], which traces back to [22]. The main proof is set up as a union bound over the finite collection $\mathcal{H}$ of all possible configurations of integral w. For each fixed configuration w, our assumptions yield a convex program with $J(w)$ constraints, and the DRO regularization of this constrained convex program is analyzed following [18]. The values of $\tilde{\alpha}$ and the χ^2 quantile set as ρ to regulate each convex program are carefully chosen to yield the result above following upper bounds on the tail of χ^2 distributions from [15].

Thm 1 shows that the optimal solution of the DRO regularized version of a MICP in form (1) asymptotically provides a good upper bound to the best performance achievable under the unknown true distribution. This is inline with the DRO theory for convex programs. The power of Thm 1 lies in relating the ρ necessary to achieve this to key formulation parameters $\bar{J}$ and $|\mathcal{H}|$. The $\bar{J} \leq J$ clearly, but often $\bar{J} \ll J$. In the next section, we explain how a sparse formulation for binary classification is DRO regularized with the aid of Thm 1, yielding Corollary 1 below. Importantly, it provides an outline of the specific heuristic used to generate good candidate robust solutions.

3 Sparse Ruleset Classification

Binary classification is the problem of finding a model that correctly classifies a data point x as having label $y \in \{0, 1\}$. We assume that data x is as an ordered set of ℓ binary-valued features, i.e., $x \in \{0, 1\}^{\ell}$. Data with integral or categorical features can be transformed into this setting by using, for instance, one-hot encoding. Real-valued features can be encoded by binning the range of all observed values. The true distribution $P^{\dagger}$ that generates the data pair $\xi := (x, y)$ is seldom known, and model fitting is done based on a (finite) training dataset $\mathcal{T} = \{\xi_1, \ldots, \xi_N\}$ of size N.

We consider the model class of convex ensembles of rulesets for binary classification. A *simple rule* takes the form $r(x) := \mathbb{I}(x^j = 0)$ or $r(x) := \mathbb{I}(x^j = 1)$ for data x and $j \in [\ell]$. A *rule* is a conjunction $t(x)$ of simple rules and has the form $t(x) = \prod_{v=1}^{C} r_v(x)$. The conjunction $t(x)$ checks if a subset of components of x have certain desired values. For a fixed C, there are $\binom{2\ell}{C}$ such rules. A *ruleset*, or disjunctive normal form (DNF) classifier, takes the form $h(x) := t_1(x) \vee t_2(x) \vee \ldots \vee t_M(x)$. It assigns x a label of 1 if any rule t_m is satisfied. This form allows the classifier to capture non-linear relationships between the features x and the label y. For example, the loan default risk classification problem from [11] applies label '*high risk*' to a disjunction of two conjunction terms (*#Loans* ≥ 7) $\vee$ ((*#Loans* ≤ 5) $\wedge$ (*Total Amount* $\geq \$10,000$)). A *convex ensemble* of rule sets takes the form $F(x) = \sum_k v_k h_k(x)$ where the non-negative v_k satisfy $\sum_k v_k = 1$. The convex ensemble predicts labels by applying a threshold, setting $\hat{y} = \mathbb{I}(F(x) \geq 1/2)$. The quality of the convex ensemble F is determined by the misclassification loss at data ξ, and takes values in $[0, 1]$ with a piecewise-linear convex form $l(v, \xi) = \max\{0,\ (1 - 2y)(2\sum_k v_k h_k(x) - 1)\}$.

To score the explanatory power of a rule set, we define its complexity based on the number of rules and the number of terms in each rule. We associate with a conjunction $t(x)$ of C simple rules a cost $c(t) = C + 1$. The cost of a DNF term h is then given by $c(h) = \sum_{m=1}^{M} c(t_m) = M + \sum_{m=1}^{M} C_m$. The ensemble of rulesets similarly bear a cost of the sum of the individual costs $c(F) = \sum_k c(h_k)$. We follow [7] and construct *sparse* rule sets with a constraint on their complexity, This is modeled by introducing $\{0, 1\}$-integer variables w, one for each disjunction h_k in F, and solving:

$$L^*_{\dagger} \ := \ \min_{v,w} \mathbb{E}_{\xi \sim P^{\dagger}} l(v, \xi) \ \text{ s.t. } \sum_k v_k = 1\,;\ v_k \leq w_k,\ \forall h_k\,; \text{ and } \sum_k w_k c(h_k) \leq \mathfrak{C}\,. \tag{3}$$

For each fixed configuration w, the complexity constraint implies at the most $\mathfrak{C}$ constraints of form $v_k \leq 1$. Let $\mathcal{H}$ denote the space of all ruleset ensembles satisfying the constraints in (3).

Corollary 1 *For the sparse ruleset MICP* (3)*, we have that* $\log |\mathcal{H}| = O(\mathfrak{C})$*. The DRO formulation studied in Thm 1 sets radius* $\rho \leq \frac{O(\mathfrak{C})}{N}$ *to obtain asymptotically valid upper bounds* R^* *on* $L^*_{\dagger}$*.*

Solving the DRO-MICP: The set of feasible rulesets h_k that obey $c(h_k) \leq \mathfrak{C}$ is of size $O(\exp(\mathfrak{C}))$ and thus a straightforward DRO reformulation of the MICP (1) over all such rulesets is cumbersome to solve. We follow the column generation approach studied in [7] for iteratively selecting rulesets to include in the ensemble. This is interleaved with the inner maximization in the robust objective (2) that optimizes for the worst pmf P assigned to the training dataset $\mathcal{T}$. The t-th iteration of the procedure consists of: (a) using column generation to identify a ruleset h_k of stricter cost bound

$\mathfrak{C}' < \mathfrak{C}$ that minimizes the misclassification loss at the data pmf P^t; (b) constructing the optimal ensemble from all rulesets identified up until iteration t that minimizes the loss at P^t without imposing the max-complexity $\mathfrak{C}$ bound; and (c) updating the data pmf to P^{t+1} via maximization (2). The procedure stops when no significant progress is being made in the robust loss objective in (c). At the end, a sparse ensemble is selected from all the generated rulesets that minimizes training loss over P^0 while satisfying the max complexity bound $\mathfrak{C}$.

Following the numerical experiments of Dash et al. [7], we conduct numerous tests on classification datasets from the UCI repository [19] and the dataset from the FICO Explainable Machine Learning Challenge [11]. Each input dataset is split uniformly at random into two datasets – the training subset that contains 90% of the data and the testing subset that contains the other 10%. A total of 10 such permutations are considered to produce the reported means and confidence intervals (CIs). In addition to our DRO-based rule set induction approach (DR) and the sparse rule set algorithm (CG) of Dash et al. [7], we consider the CART [5] method as an alternative to creating rules-based explainable models whose complexity can be calculated using our cost model. The random forest [4] (RF) method is included as representative of the state-of-the-art in producing models that generalize well, while at the same time its dense ensembles of decision trees are not appropriate for explainability. A representative sample of our empirical results is presented in Table 1 where we set $\mathfrak{C}' = 5$ since this low value grants more generalization flexibility. We refer the reader to [6] for additional details.

Table 1: *Classification Performance, Complexity of Models, and Time Ratio Between CG and DR over Unseen Test Data of Models Created by Competing Methods for Seven UCI Repository Datasets and the FICO Dataset. Mean Rank of Competing Methods given in Last Row. All Results shown as Mean* (95% *CI Width).*

Dataset	Test Performance (%)				Model Complexity			Time ratio
Name	DR	CG	CART	RF	DR	CG	CART	CG/DR
heart	78.6(4.3)	78.9(4.7)	***81.6***(4.7)	82.5(1.4)	15.8(1.9)	**11.3**(3.5)	32.0(15.9)	3.6
ILPD	**71.5**(1.5)	69.6(2.4)	67.4(3.1)	69.8(1.0)	*14.5*(1.4)	**10.9**(5.3)	56.5(21.4)	6.4
FICO	**72.1**(0.8)	71.7(1.0)	70.9(0.6)	73.1(0.2)	*14.9*(2.6)	**13.3**(8.0)	155.0(53.9)	1.9
ionosphere	***92.6***(4.3)	90.0(3.5)	87.2(3.5)	93.6(1.4)	*15.0*(1.5)	**12.3**(5.9)	46.1(8.2)	1.5
liver	*59.1*(4.3)	***59.7***(4.7)	55.9(2.7)	60.0(1.6)	12.8(1.0)	**5.2**(2.4)	60.2(30.6)	4.3
pima	**76.2**(3.7)	*74.1*(3.7)	72.1(2.5)	*76.1*(1.6)	14.5(1.5)	**4.5**(2.5)	34.7(11.4)	4.2
transfusion	*77.5*(1.7)	*77.9*(2.7)	**78.7**(2.2)	*77.3*(0.6)	10.7(1.3)	**5.6**(2.4)	14.3(4.5)	1.1
WDBC	**94.9**(0.8)	94.0(2.4)	93.3(1.8)	97.2(0.4)	*16.2*(2.0)	**13.9**(4.7)	*15.6*(4.3)	3.6
Mean Rank	2.000	2.714	3.571	**1.714**	2.143	**1.000**	2.857	

For both the test accuracy and model complexity results in Table 1, the method with the best average outcomes among the interpretable models is highlighted in **bold** and the best average outcome overall is underlined. Any method with average outcomes within the CI of the overall best outcome is highlighted in *italics*. We also report the mean rank of each method following [8].

The leftmost set of columns in Table 1 shows that DR provides the best overall generalization performance among all methods in two of the eight cases (ILPD, pima) and among all interpretable methods in all but three of the eight cases (heart, liver, transfusion), as well as either performing best or statistically identical to the best method in all but three of the eight cases (heart, FICO, WDBC). This helps to explains why the solution quality of DR yields the second best mean rank among all methods across all datasets, relatively close to the best mean rank of the non-interpretable RF method and significantly better than the mean rank of the competing explainable methods of CG and CART.

The middle set of columns in Table 1 shows that the complexity of the estimated models from DR is within a reasonable range of CG, even statistically identical to CG in half of the cases. As reflected in the mean rank, CG has the lowest model complexity with DR consistently providing better complexity than CART in all but one case (WDBC). The DR method is able to produce sparse convex mixtures of rulesets of complexity of only $\mathfrak{C}' = 5$ that maintain a low generalization error with a slightly higher complexity. In contrast, the individual rule sets produced by CG use higher complexity than 5 to obtain a reasonable balance between reduced training error and lower test efficacy, but overall do not perform well on test data. Increasing the model complexity in CG may improve training power but leads to a higher generalization gap, and it is not readily clear that the gap in test performance can be closed for any value of $\mathfrak{C}$ by the CG method. The last column of Table 1 presents the computational savings of DR over CG as a multiplicative factor ranging from 1.1 to 6.4 with an average factor of 3.3. This is caused by the repeated calls to solve the IPs needed in the cross-validation procedure to tune the complexity bound $\mathfrak{C}$ in CG. The DR method, in stark contrast, significantly economizes on the calls to column generation to produce additional rulesets to add to the ensemble, thus efficiently generalizing better and achieving the main goal in introducing the DRO-based regularization procedure.

References

[1] D. Bertsimas, A. King, and R. Mazumdar. Best subset selection via a modern optimization lens. *Annals of Statistics*, 44(2):813–852, 2016.

[2] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, Sep 2019.

[3] J. H. Blanchet, Y. Kang, F. Zhang, and Z. Hu. A distributionally robust boosting algorithm. *2019 Winter Simulation Conference (WSC)*, pages 3728–3739, 2019.

[4] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[5] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and regression trees. *Biometrics*, 40:874, 1984.

[6] S. Dash, S. Ghosh, J. Goncalves, and M. S. Squillante. Obtaining Explainable Classification Models using Distributionally Robust Optimization. *ArXiv e-prints*, 2025.

[7] S. Dash, O. Gunluk, and D. Wei. Boolean decision rules via column generation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[8] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[9] A. Dhurandhar, K. Shanmugam, and R. Luss. Enhancing simple models by exploiting what they already know. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[10] J. Duchi, P. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

[11] FICO. FICO explainable machine learning challenge. https://community.fico.com/community/xml, 2018. Last accessed 2018-05-16.

[12] S. Ghosh, M. S. Squillante, and E. Wollega. Efficient generalization with distributionally robust learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28310–28322. Curran Associates, Inc., 2021.

[13] S. T. Goh, L. Semenova, and C. Rudin. Sparse density trees and lists: An interpretable alternative to high-dimensional histograms. *INFORMS Journal on Data Science*, 2024.

[14] X. Hu, C. Rudin, and M. Seltzer. Optimal sparse decision trees. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[15] T. Inglot. Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30:339–351, 2010.

[16] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1675–1684, 2016.

[17] H. Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

[18] H. Lam and E. Zhou. The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters*, 45(4):301–307, 2017.

[19] M. Lichman. UCI machine learning repository, 2013.

[20] H. Namkoong and J. C. Duchi. Variance-based regularization with convex objectives. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2971–2980. Curran Associates, Inc., 2017.

[21] A. Owen. *Empirical Likelihood*. Chapman and Hall/CRC, 2001.

[22] J. Qin and J. Lawless. Empirical Likelihood and General Estimating Equations. *The Annals of Statistics*, 22(1):300 – 325, 1994.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, page 1135–1144, 2016.

[24] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.*, 1(5):206—215, 2019.

[25] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statist. Surv.*, 16:1–85, 2022.

[26] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille. A bayesian framework for learning rule sets for interpretable classification. *Journal of Machine Learning Research*, pages 1–37, 2017.

[27] R. Zhang, R. Xin, M. Seltzer, and C. Rudin. Optimal sparse regression trees. In *Proceedings of AAAI*, 2023.