Performance-Aligned Learning Algorithms with Statistical Guarantees

Research Statement of Rizal Fathony

The goal of many prediction tasks in machine learning is to learn a prediction function that minimizes certain loss functions (e.g. zero-one, ordinal, and cost-sensitive loss) or maximizes certain performance metrics (e.g. accuracy, precision, recall, F1-score, and ROC curve) on the testing dataset. Unfortunately, optimizing those metrics directly via empirical risk minimization [6] is known to be intractable [7, 8]. In practice, convex surrogate losses over the desired metrics are needed in order to build efficient learning algorithms with the hope that optimizing the convex surrogates will indirectly optimize the original metrics given sufficient training data.

Among the most popular paradigms in building machine learning algorithms are the probabilistic approaches (e.g. logistic regression, conditional graphical models and many neural networks models) and large-margin approaches (e.g. support vector machine (SVM), and structured SVM) which differ in the way they construct convex surrogate losses. Probabilistic approaches construct prediction probability models and employ the logistic loss as the convex surrogate. Large-margin approaches aim to maximize the margin that separates correct predictions from the incorrect ones and use the hinge loss for the convex surrogates construction. Both approaches have their own strengths and weaknesses. In the case of multiclass classification, for example, the probabilistic approach (logistic regression) enjoys the statistical guarantee of Fisher consistency—meaning it optimizes the accuracy metric and produces Bayes optimal classifiers when they learn from any true distribution of data using a rich feature representation—, while the large-margin approach (SVM) enjoys computational efficiency via the kernel trick and dual parameter sparsity. However, the current formulations of large-margin approach suffer from Fisher consistency issues [9, 10, 11], while the probabilistic approach does not have dual parameter sparsity property.

When generalized to structured prediction, probabilistic methods such as conditional random field (CRF) [12] capture probabilistic structures in the model (which translates to Fisher consistency guarantees), with the downside that the computation of the normalization term may be intractable. The other weakness of probabilistic methods is that they do not have a mechanism to easily incorporate customized performance metrics and loss functions into their learning process, which is important in many structured prediction settings. Large-margin models like structured SVM (SSVM) [13, 14] have the flexibility to incorporate customized performance metrics and loss functions, but the Fisher consistency property is not guaranteed. This motivates the search for new approaches that overcome the weaknesses of the probabilistic and large-margin methods.

My research aims to address the challenges above by constructing new learning algorithms that simultaneously satisfy the desired properties of: (1) aligning with the learning objective by incorporating customized performance metrics or loss functions in the learning process; (2) providing the statistical guarantee of Fisher consistency; (3) enjoying computational efficiency; and (4) performing competitively in practice. My approach in constructing the learning algorithms is based on the robust adversarial formulation [15, 16, 17], i.e., *what predictor best maximizes the performance matrix (or minimizes the loss function) in the worst case given the statistical summaries of the empirical distributions?* In my previous works, I designed robust adversarial learning algorithms for multiclass classification, ordinal regression, bipartite matching in graphs, and graphical models with exact learning and inference. My future research directions focus on investigating the statistical properties of loss functions, such as stronger statistical guarantees, and the Fisher consistency of structured loss functions, as well as developing learning algorithms for various machine learning tasks, such as structured prediction, graphical models, multi-tasks learning, learning under specific constraints (e.g., learning with limited data, learning under budget, and learning with reject option), and combining the robust adversarial approach with deep learning.

Multiclass Classifications

In binary classification, logistic regression and SVM are among the most popular methods. Both techniques enjoy the statistical guarantee of Fisher consistency. SVMs provide the additional advantage of dual parameter sparsity so that when combined with kernel methods, extremely rich feature representations can be efficiently employed. Unfortunately, generalizing SVMs to multiclass classification with more than two labels is challenging and existing multiclass extensions of SVM tend to lose their consistency guarantees or produce low accuracy predictions in practice. The most popular multiclass extensions of SVMs are the WW model by Weston et al. [18], the CS model by Crammer and Singer [19] and LLW model by Lee at al. [20]. Of these, only the LLW model is Fisher consistent. However, as pointed by Dogan et. al. [11], the LLW model often performs poorly for datasets with low-dimensional feature spaces.

In our NIPS 2016 paper [1], we design a surrogate loss (we call it $\text{AL}^{0\text{-}1}$) for multiclass classification that overcome this long standing problem in the multiclass SVM formulation. The $\text{AL}^{0\text{-}1}$ is based on the robust adversarial formulation for multiclass zero-one loss, i.e. the robust predictor that minimize the zero-one loss in the worst case scenario given the summaries of the empirical distribution. It fills the long standing gap in multiclass classification by simultaneously: (1) guaranteeing Fisher consistency; (2) enabling computational efficiency via the kernel trick and dual parameter sparsity; and (3) providing competitive performance in practice.

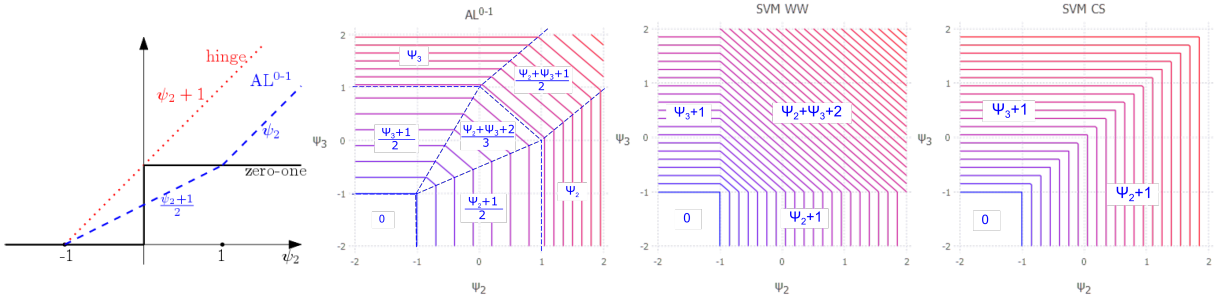

Figure 1: Plots of the $\text{AL}^{0\text{-}1}$ surrogate loss compared to the hinge loss for binary classification (leftmost), and $\text{AL}^{0\text{-}1}$ loss compared to the SVM-WW and the SVM-CS for 3 classes predictions (others).

Ordinal Classification/Regression

Many prevalent methods for ordinal classification/regression, where the discrete class labels have an inherent order, reduce the task to binary classification problems. Some view the task from the regression perspective and learn both a linear regression function and a set of thresholds that define class boundaries [21, 22, 23, 24, 25]. Other methods take a classification perspective and use tools from cost-sensitive classification [26, 27, 28]. However, since the ordinal regression loss is a non-convex and non-continuous, surrogate losses that approximate it have to be employed. Under both perspectives, surrogate losses for ordinal regression are constructed by transforming the surrogate losses for binary zero-one loss problems—such as the hinge loss, the logistic loss, and the exponential loss—to take into account the different penalties of the ordinal regression problem. Empirical evaluations have compared the appropriateness of different surrogate losses, but these still leave the possibility of undiscovered surrogates that align better with the ordinal regression loss.

Our NIPS 2017 paper [2] proposes a new way of constructing surrogate losses using the robust adversarial formulation which is trained to optimize the ordinal regression loss in the worst case scenario given statistical summaries of the training data. We shows that different types of statistical summaries of the empirical training data lead to thresholded regression-based predictions or classification-based predictions. In both cases, the surrogate loss is novel compared to existing surrogate losses. We also show the Fisher consistency of our surrogate losses as well as the empirical benefit of our formulation in several ordinal classification tasks.

Bipartite Matching in Graphs

Many important structured prediction problems, including recognizing correspondences in similar images, finding word alignments in text and providing ranked lists of items of information retrieval tasks can be formulated as weighted bipartite matching optimizations. Both probabilistic and large-margin approaches have significant drawbacks when applied under the constraints of perfect matchings. Probabilistic models such as CRF [29, 30] provide statistical guarantee of Fisher consistency but suffer computationally from the need to compute normalization terms that involve matrix permanent computations [29] (a #P-hard problem [31]). In contrast, large-margin approaches such as structured SVM [13, 14] provide computational efficiency but lack Fisher consistency guarantees.

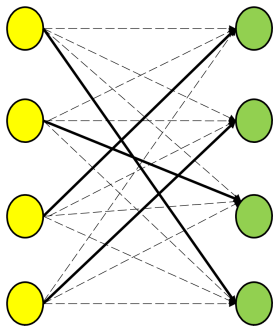

Figure 2: An example of bipartite matching task.

Our ICML 2018 paper [3] proposes the first learning algorithm for bipartite matching that enjoys both the statistical guarantee of Fisher consistency and computational efficiency. Our key algorithm design to avoid the need of searching for over $n!$ possible matchings is the reformulation of the adversarial robust formulation in terms of marginal probabilities. This reformulation reduces the number of variables to optimize from $n!$ to n^2 where the optimization is now over doubly-stochastic matrices. We derive an efficient algorithm to solve the marginal formulation and establish the convergence property of the algorithm. We show that the combination of efficient computation and Fisher consistency boosts the performance of the algorithm in video-tracking application to significantly outperform structured SVM. In many of these problems, due to the size of the problem, running CRF is impractical.

Conditional Graphical Models

Probabilistic graphical models such as CRF have the strength in modeling the structured probabilistic connections among the variables. Compared with the large-margin approach, CRF also provides the benefit of statistical consistency. The downside of CRF is, however, in contrast with structured SVM, it cannot easily incorporate custom performance metrics or loss functions into its learning process. This may limit the application of CRF for many structured prediction tasks where the goal of learning is to optimize customized performance metrics. On the other hand, although the structured SVM can easily incorporate custom metrics into its learning process, its statistical consistency is not guaranteed.

Our recent work [4] proposes the adversarial graphical models (AGMs) where the structured connections in the graph model the worst case conditional probability with respect to the optimized loss, marginalized over the pairwise variables. Our formulation can easily incorporate custom performance metrics or loss functions into its learning process while maintaining the statistical guarantee of Fisher consistency with respect to the metrics. We believe that the AGMs open new possible applications of graphical model optimization with customized performance metrics and loss functions in various application areas. In this work, we focus first on AGMs with graphical structures that are known to have exact inference and learning algorithms (including chain, tree, and low-treewidth graphs) with additively decomposable loss functions. We develop efficient learning algorithms for AGMs that have similar time complexity with the CRF and structured SVM and show the empirical benefit of our methods in several structured prediction tasks.

Future Directions

My future research directions focus on exploring statistical theory for loss functions, designing learning algorithms with statistical guarantees for many machine learning tasks, and applying the techniques to many application areas, including natural language processing, computer vision and bioinformatics.

Statistical Theory of Loss Functions

Is Fisher consistency sufficient for characterizing the desired theoretical properties of surrogate losses? Our previous research [1] shows that not all Fisher consistent surrogate losses are equal. While the

LLW approach [20] is Fisher consistent for multiclass classification, it performs poorly on datasets with low dimensional feature spaces [11]. In contrast, our Fisher consistent model performs competitively in both low and high dimensional spaces. My future works will investigate a stronger statistical guarantee for surrogate losses that aim to differentiate high performance Fisher consistent losses from the low performance ones. This statistical guarantees will need to consider how 'fast' a surrogate loss converge to the desired loss function based on the complexity of the hypothesis function used in the learning process.

Structured Performance Metrics or Loss Functions

For many standard classification problems, even though the problem itself is not a structured prediction problem, i.e., only predict single variable y for given $\mathbf{x}$, the performance metric in which the prediction is evaluated has some structure. Some examples of the desired structured performance metrics are precision, recall, F_β-score, Jaccard score, ROC-area, etc. I am interested in developing a plug-in classifier for these performance metrics that enjoys Fisher consistency based on the robust adversarial formulation. Previous works [32, 33] have investigated the problem for certain performance metrics, but the approaches require significant modifications (in some cases, rewriting the whole algorithm) when different metrics are used. I plan to investigate the common properties of different structured performance metrics and incorporate these properties in the algorithm design. My goal is to provide a Fisher consistent alternatives to the popular SVM^{perf} algorithm [34].

Structured Prediction and Graphical Models

I plan to continue my ongoing research in adversarial graphical models for structured prediction with more complex graphical structures and more general performance metrics or loss functions. In these cases, both exact and approximate learning and inference algorithms needs to be explored. This project aims to provide general graphical model frameworks that can be aligned with customized objective by incorporating customized performance metrics or loss functions as the input in the learning process and provide the statistical guarantee of Fisher consistency with respect to the metrics. I believe that this approach will benefit many applications in real word problems including natural language processing, computer vision, and bioinformatics.

Deep Learning

Deep neural networks models have been very successful in many machine learning applications and has become the state-of-the-art method for many prediction tasks. Deep learning techniques enable learning algorithms to simultaneously optimize the prediction accuracy and learn hierarchical representations of the dataset. In many cases, however, the prediction problem requires optimizing customized performance metrics where there are no easy ways to incorporate them into the learning pipeline of the deep learning models. In these cases, one needs to perform representation learning and prediction optimization separately. One solution to this problem, in my view, is to combine robust adversarial formulation with deep learning. Given a particular performance metric, robust adversarial formulations can be used in the last layer as the gradient source which can then be backpropagated to the neural network structure in the previous layers. This will enable end-to-end learning for many problems in which optimizing customized performance metrics is the goal of the learning process.

Multi-Task Learning

In many machine learning problems, leaning multiple tasks simultaneously improves the efficiency and prediction performance of the task-specific models when compared to training the models separately. Multi-task learning exploits the commonalities and differences across tasks to improve the efficiency and prediction performance. In each task, the performance metric used may be different, e.g., accuracy for the first task and F-score for the second task. I am interested in developing robust adversarial learning algorithms for multi-tasks learning that enjoy the statistical guarantee of Fisher consistency with respect to the combined performance metrics.

References (Publications)

[1] Rizal Fathony, Anqi Liu, Kaiser Asif, and Brian D. Ziebart. Adversarial multiclass classification: A risk minimization perspective. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 559–567, 2016.

[2] Rizal Fathony, Mohammad Ali Bashiri, and Brian D. Ziebart. Adversarial surrogate losses for ordinal regression. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 563–573, 2017.

[3] Rizal Fathony*, Sima Behpour*, Xinhua Zhang, and Brian D. Ziebart. Efficient and consistent adversarial bipartite matching. In *International Conference on Machine Learning (ICML)*, 2018.

[4] Rizal Fathony, Ashkan Rezaei, Mohammad Bashiri, and Brian D. Ziebart. Distributionally robust graphical models. *Submitted to NIPS 2018*, 2018.

[5] Anqi Liu, Rizal Fathony, and Brian D. Ziebart. Kernel robust bias-aware prediction under covariate shift. *arXiv preprint arXiv:1712.10050*, 2017.

Other References

[6] Vladimir Naumovich Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[7] Klaus-Uwe Hoffgen, Hans-Ulrich Simon, and Kevin S Vanhorn. Robust trainability of single neurons. *Journal of Computer and System Sciences*, 50(1):114–125, 1995.

[8] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[9] Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *The Journal of Machine Learning Research*, 8:1007–1025, 2007.

[10] Yufeng Liu. Fisher consistency of multicategory support vector machines. In *International Conference on Artificial Intelligence and Statistics*, pages 291–298, 2007.

[11] Ürün Doğan, Tobias Glasmachers, and Christian Igel. A unified view on multi-class support vector classification. *Journal of Machine Learning Research*, 17(45):1–32, 2016.

[12] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, volume 951, pages 282–289, 2001.

[13] Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on Machine learning*, pages 896–903. ACM, 2005.

[14] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. In *JMLR*, pages 1453–1484, 2005.

[15] Flemming Topsøe. Information theoretical optimization techniques. *Kybernetika*, 15(1):8–27, 1979.

[16] Peter D. Grünwald and A. Phillip Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32:1367–1433, 2004.

[17] Kaiser Asif, Wei Xing, Sima Behpour, and Brian D. Ziebart. Adversarial cost-sensitive classification. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2015.

[18] Jason Weston, Chris Watkins, et al. Support vector machines for multi-class pattern recognition. In *ESANN*, volume 99, pages 219–224, 1999.

[19] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, 2002.

[20] Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

[21] Amnon Shashua and Anat Levin. Ranking with large margin principle: Two approaches. In *Advances in Neural Information Processing Systems 15*, pages 961–968. MIT Press, 2003.

[22] Wei Chu and S Sathiya Keerthi. New approaches to support vector ordinal regression. In *Proceedings of the 22nd international conference on Machine learning*, pages 145–152. ACM, 2005.

[23] Hsuan-Tien Lin and Ling Li. Large-margin thresholded ensembles for ordinal regression: Theory and practice. In *International Conference on Algorithmic Learning Theory*, pages 319–333. Springer, 2006.

[24] Jason D. M. Rennie and Nathan Srebro. Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, pages 180–186, 2005.

[25] Ling Li and Hsuan-Tien Lin. Ordinal regression by extended binary classification. *Advances in neural information processing systems*, 19:865, 2007.

[26] Hsuan-Tien Lin. *From ordinal ranking to binary classification*. PhD thesis, California Institute of Technology, 2008.

[27] Hsuan-Tien Lin. Reduction from cost-sensitive multiclass classification to one-versus-one binary classification. In *Proceedings of the Sixth Asian Conference on Machine Learning*, pages 371–386, 2014.

[28] Han-Hsing Tu and Hsuan-Tien Lin. One-sided support vector regression for multiclass cost-sensitive classification. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1095–1102, 2010.

[29] James Petterson, Jin Yu, Julian J McAuley, and Tibério S Caetano. Exponential family graph matching and ranking. In *Advances in Neural Information Processing Systems*, pages 1455–1463, 2009.

[30] Maksims Volkovs and Richard S Zemel. Efficient sampling for bipartite matching problems. In *Advances in Neural Information Processing Systems*, pages 1313–1321, 2012.

[31] Leslie G Valiant. The complexity of computing the permanent. *Theoretical computer science*, 8(2):189–201, 1979.

[32] Hong Wang, Wei Xing, Kaiser Asif, and Brian Ziebart. Adversarial prediction games for multivariate losses. In *Advances in Neural Information Processing Systems*, pages 2710–2718, 2015.

[33] Zhan Shi, Xinhua Zhang, and Yaoliang Yu. Bregman divergence for stochastic variance reduction: Saddle-point and adversarial prediction. In *Advances in Neural Information Processing Systems*, pages 6033–6043, 2017.

[34] Thorsten Joachims. A support vector method for multivariate performance measures. In *Proceedings of the International Conference on Machine Learning*, pages 377–384, 2005.