Hardness Results for Structured Learning and Inference with Multiple Correct Outputs

Matthew Blaschko, Jiaqian Yu

Hardness Results for Structured Learning and Inference with Multiple Correct Outputs

Matthew B. Blaschko MATTHEW.BLASCHKO@INRIA.FR
Jiaqian Yu JIAQIAN.YU@CENTRALESUPELEC.FR
Inria & CentraleSupélec, Grande Voie des Vignes, 92295 Châtenay-Malabry, France

Abstract

In many domains of structured output prediction, multiple outputs can be considered correct. Several results exist showing that polynomial time computation both at training and test time is possible when a single correct output is present. In this work, we show that such guarantees do not hold when multiple outputs are correct. This is shown through three main results indicating that multiple correct outputs lead to NP-hard computation with existing convex surrogates for (i) learning with a supermodular loss function, (ii) learning with a submodular loss function, and (iii) test time inference with a diversity penalty term. These theoretical results highlight the importance of identifying sufficient conditions for tractable learning and inference with multiple correct outputs in practice.

1. Introduction

Many domains of structured prediction contain multiple correct outputs. In computer vision, multiple correct object detections may be present in an image (Blaschko, 2011). In text summarization, multiple paragraphs may be considered equally good summaries of a document (Sipos et al., 2012). In protein structure prediction, a molecule may have multiple possible configurations (Rohl et al., 2004). In this work, we show that the presence of multiple correct outputs leads to intractable computational problems in many common settings for which a single correct output leads to tractable problems.

We show three main hardness results for structured prediction with multiple correct outputs: (i) regularized risk minimization with a supermodular loss function is tractable with existing learning frameworks for a single correct output but NP-hard for multiple correct outputs (Proposition 3), (ii) regularized risk minimization with a submodular loss function is tractable with existing learning frameworks for a single correct output but NP-hard for multiple correct outputs (Proposition 4), and (iii) test time inference that is polynomial time solvable for a single correct output is NP-hard for multiple correct outputs when a diversity penalty is included (Proposition 5). These results suggest the use of alternative learning and approximate inference schemes when multiple correct outputs are present during training and/or testing.

Proceedings of the Constructive Machine Learning workshop @ ICML 2015.

2. Learning

In structured output learning, we assume that a task specific loss function is given that measures the disagreement between a prediction and a ground truth element. We denote a ground truth instance as $y^* \in \mathcal{Y}$ and the (incorrect) prediction $\tilde{y}$. We will distinguish between a loss function in which a single correct output is to be predicted, and a loss function in which p multiple correct outputs $Y^* \in \mathcal{Y}^p$ are possible:

$$\Delta_{\text{single}} : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R} \tag{1}$$

$$\Delta_{\text{multiple}} : \mathcal{Y}^p \times \mathcal{Y} \mapsto \mathbb{R}. \tag{2}$$

Specifically, we will assume that when there are multiple outputs we will take

$$\Delta_{\text{multiple}}(Y^*, \tilde{y}) := \min_{y^* \in Y^*} \Delta_{\text{single}}(y^*, \tilde{y}) \tag{3}$$

so that we require a prediction to be close to one of the ground truth outputs. We focus on two feasible families of loss functions for Δ_{single} for which convex surrogates have been developed: supermodular loss functions and submodular loss functions.

Submodular functions may be defined through several

equivalent properties. We use the following definition (Fujishige, 2005):

Definition 1. *A set function $l : \mathcal{P}(V) \mapsto \mathbb{R}$ is **submodular** if and only if for all subsets $A, B \subseteq V$, $l(A) + l(B) \geq l(A \cup B) + l(A \cap B)$.*

A function is *supermodular* iff its negative is submodular, and a function is modular iff it is both submodular and supermodular.

The computational implications of multiple correct ground truth instances for each of these families are explored in the following two subsections.

2.1. Supermodular Loss Functions and the Structured Output Support Vector Machine

The Structured Output Support Vector Machine (SOSVM) is one of the most popular frameworks for structured output prediction (Taskar et al., 2004; Tsochantaridis et al., 2005). Two variants bound a discrete loss function Δ with a convex surrogate: margin rescaling and slack rescaling. The margin-rescaling constraints and slack-rescaling constraints are:

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i, \quad \forall i, \forall \tilde{y} \in \mathcal{Y}: \tag{4}$$

$$\langle w, \phi(x_i, y_i)\rangle - \langle w, \phi(x_i, \tilde{y})\rangle \geq \Delta(y_i, \tilde{y}) - \xi_i \tag{5}$$

$$\text{or} \quad \langle w, \phi(x_i, y_i)\rangle - \langle w, \phi(x_i, \tilde{y})\rangle \geq 1 - \frac{\xi_i}{\Delta(y_i, \tilde{y})} \tag{6}$$

respectively.

Yu & Blaschko (2015) analyzed under which conditions margin and slack rescaling yield tight convex surrogates to the underlying loss function Δ. A central result is that the following notion of monotonicity is necessary:

Definition 2. *A set function $l : \mathcal{P}(V) \mapsto \mathbb{R}$ is **increasing** if and only if for all subsets $A \subset V$ and elements $x \in V \setminus A$, $l(A) \leq l(A \cup \{x\})$.*

Proposition 1 ((Yu & Blaschko, 2015))**.** *Slack rescaling yields an extension of a set function $\Delta(y, \cdot)$ iff $\Delta(y, \cdot)$ is an increasing function.*

It is a necessary, but not sufficient condition that $\Delta(y, \tilde{y})$ be increasing for margin rescaling to yield an extension. However, for all increasing $\Delta(y, \tilde{y})$ there exists a positive scaling $\gamma \in \mathbb{R}$ such that margin rescaling yields an extension.

Proposition 2 ((Yu & Blaschko, 2015))**.** *For all increasing set functions l such that $\exists y$ for which margin rescaling does not yeild an extension of $\Delta(y, \cdot)$, we can always find a positive scale factor γ specific to l such that margin rescaling yields an extension. We denote $\mathbf{M}\gamma\Delta$ and $\gamma\Delta$ as the rescaled functions.*

These results indicate that both slack and margin rescaling can be used to construct tight convex surrogates to increasing loss functions.

In the event that we have multiple correct outputs, we construct the loss imputed to the SOSVM by taking the minimum loss over all possible correct outputs. The resulting loss remains increasing, and therefore the resulting convex surrogate is tight, a positive result from a statistical perspective. However, multiple correct outputs may mean that computation of the subgradient of the convex surrogate is NP-hard. We show this making use of the following lemma:

Lemma 1. *Submodular functions are not closed over the* max *operation, i.e. if g and h are both submodular functions, $f = \max(g, h)$ is not necessarily submodular (see e.g. Section 1.2 (Krause & Golovin, 2014)).*

Proposition 3. *Computation of a subgradient of the slack and margin rescaling convex surrogates is NP-hard when there are multiple correct outputs.*

Proof. Slack and margin rescaling are computationally feasible only when Δ is supermodular (Yu & Blaschko, 2015). This is because subgradient computation requires solving

$$\arg\max_{\tilde{y}} \Delta(y, \tilde{y})(1 + \langle w, \phi(x, \tilde{y}) - \phi(x, y)\rangle) \tag{7}$$

and

$$\arg\max_{\tilde{y}} \Delta(y, \tilde{y}) + \langle w, \phi(x, \tilde{y})\rangle \tag{8}$$

for slack and margin rescaling, respectively. Optimization of Equations (7) and (8) is called loss augmented inference. The minimum over supermodular functions is the negative of the maximum over submodular functions, and submodular functions are not closed under maximization (Lemma 1). Finally we note that the arg max in Equations (7) and (8) will therefore be taken over non-supermodular functions, which is NP-hard in general. □

Consequently, even if we have a tractable loss augmented inference problem for a single correct output with a supermodular loss, the presence of multiple correct outputs will lead to NP-hard loss augmented inference problems.

2.2. Submodular Loss Functions and the Lovász Hinge

We now turn to submodular loss functions. These result in NP-hard subgradient computation for

SOSVMs, so Yu & Blaschko (2015) introduced an alternative convex surrogate based on the Lovász extension of a submodular set function, called the Lovász hinge.

Definition 3. *The Lovász hinge regularized risk is defined as*

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \tag{9}$$

$$s.t.\ \xi_i \geq \max_{\pi}\sum_{k=1}^{p} s_i^{\pi_k}\big(l(\{\pi_1,\cdots,\pi_k\}) - l(\{\pi_1,\cdots,\pi_{k-1}\})\big) \tag{10}$$

$$\forall i, \forall \tilde{y} \in \mathcal{Y}$$

where l is a submodular set function, $\pi = (\pi_1,\cdots,\pi_p)$ is a permutation of the index, $s_i^{\pi_k} = 1 - \langle w^{\pi_k}, x_i^{\pi_k}\rangle y_i^{\pi_k}$, and $l(\cdot)$ is derived from $\Delta(y^*,\cdot)$ by setting the argument to l to be the set of parts of $\tilde{y}$ not equal to y^*.

This convex surrogate generalizes hinge loss to multiple outputs when the loss function is submodular. Subgradient computation has a computational complexity of $\mathcal{O}(|y|\log|y|)$, where $|y|$ is the size of a single instance $y \in \mathcal{Y}$. We note that this convex surrogate (written as the maximum over linear constraints) is only tight when Δ remains submodular. Analogous to Proposition 3, we now show that multiple correct outputs results in a loss function Δ that is not guaranteed to be submodular, even if the loss function for a single correct output is submodular. We make use of the following result:

Lemma 2. *Submodular functions are not closed over the* min *operation, i.e. if g and h are both submodular functions, $f = \min(g,h)$ is not necessarily submodular (see e.g. Section 1.2 (Krause & Golovin, 2014)).*

Proposition 4. *Neither the Lovász hinge nor the structured output SVM provide a polynomial time tight convex surrogate to a submodular loss function when there are multiple correct outputs.*

Proof. Application of Equation (3) combined with Lemma 2 indicates that Δ_{multiple} is not submodular in general, even if Δ_{single} is. As Δ_{multiple} is neither submodular nor supermodular, neither the Lovász hinge nor the SOSVM yield polynomial time tight convex surrogates. □

3. Inference

In the previous section, we have shown that learning with both submodular and supermodular loss functions leads to NP-hard computation in order to compute subgradients of existing convex surrogate loss functions when there are multiple correct outputs. In this section, we further show a result due to Blaschko (2011) that test time inference becomes NP-hard in the presence of multiple correct outputs when a diversity penalty is included in the inference procedure.

Proposition 5. *Let $g(x,y)$ be a compatibility function for a structured prediction problem (e.g. $\langle w, \phi(x,y)\rangle$ from a SOSVM). The prediction of a set of $p \geq 2$ outputs with a diversity penalty is NP-hard in general:*

$$\arg\max_{Y\in\mathcal{Y}^p} \sum_{y\in Y} g(x,y) - \boldsymbol{\Omega}(Y) \tag{11}$$

where

$$\boldsymbol{\Omega}(\mathbf{Y}) = \sum_{i\neq j}\Omega(y_i,y_j) + \sum_{c\in\mathcal{C}}\Omega_c(y_c), \tag{12}$$

$\mathcal{C}$ is the set of higher order cliques in the penalty term (possibly $\mathcal{C} = \emptyset$), and Ω_c is supermodular for all $c \in \mathcal{C}$.

Proof. Section 3 of (Blaschko, 2011) shows that Equation (12) is supermodular for $\Omega \geq 0$. Consequently, the optimization in Equation (11) corresponds with a non-submodular minimization and is NP-hard. □

4. Conclusions

In this work, we have shown three main hardness results. The first two indicate that learning with supermodular (Proposition 3) and submodular (Proposition 4) loss functions are feasible when a single correct output is present, but NP-hard in general in the presence of multiple correct outputs. The third main result shows that test time inference is NP-hard when a diversity penalty is included (Proposition 5). In the absence of this penalty, test time inference with a single model will typically predict a set of highly overlapping predictions that are clustered around the single highest scoring output.

These results give substantial evidence that structured learning and inference with multiple correct outputs is fundamentally harder than when only a single output is considered correct. This points to two potentially productive directions of inquiry: (i) the exploration of tractable approximations to the NP-hard learning and inference problems, and (ii) the derivation of novel convex surrogates and sufficient conditions for polynomial time learning and inference that are applicable in practice to problems of interest.

Acknowledgements

This work is partially funded by ERC Grant 259112, and FP7-MC-CIG 334380. Jiaqian Yu is supported by a fellowship from the China Scholarship Council.

References

Blaschko, Matthew B. Branch and bound strategies for non-maximal suppression in object detection. In Boykov, Yuri, Kahl, Fredrik, Lempitsky, Victor, and Schmidt, Frank R. (eds.), *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 6819 of *Lecture Notes in Computer Science*, pp. 385–398. Springer, 2011.

Fujishige, Satoru. *Submodular functions and optimization*. Elsevier, 2005.

Krause, Andreas and Golovin, Daniel. Submodular function maximization. In Bordeaux, Lucas, Hamadi, Youssef, and Kohli, Pushmeet (eds.), *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, 2014.

Rohl, Carol A., Strauss, Charlie E. M., Misura, Kira M. S., and Baker, David. Protein structure prediction using Rosetta. In Brand, Ludwig and Johnson, Michael L. (eds.), *Numerical Computer Methods, Part D*, volume 383 of *Methods in Enzymology*, pp. 66–93. Academic Press, 2004.

Sipos, Ruben, Shivaswamy, Pannaga, and Joachims, Thorsten. Large-margin learning of submodular summarization models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 224–233, 2012.

Taskar, Ben, Guestrin, Carlos, and Koller, Daphne. Max-margin markov networks. In Thrun, Sebastian, Saul, Lawrence K., and Schölkopf, Bernhard (eds.), *Advances in Neural Information Processing Systems 16*, pp. 25–32. MIT Press, 2004.

Tsochantaridis, Ioannis, Joachims, Thorsten, Hofmann, Thomas, and Altun, Yasemin. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6 (9):1453–1484, 2005.

Yu, Jiaqian and Blaschko, Matthew B. Learning submodular losses with the Lovász hinge. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.