Scalable Online Optimization algorithms
&
Modern Approaches to Structured Prediction

William St-Arnaud, Elyes Lamouchi, Frédéric Boileau

Introduction

Structured prediction is concerned with predicting labels from a given set of features where the labels have some inherent structure which should be considered when training the model. The structures considered are usually combinatorial in nature which means comparing the scores of two assignments can be a challenge with respect to tractability. We review some of the ways this is tackled with a large margin approach in the modern literature. We thus focus on the structural SVM framework which has the computational advantage over MLE approaches by dispensing with the need to compute the partition function which is #P-complete in general; e.g. for matchings and mincuts.(Lacoste-Julien et al. 2012).

Using convex analysis it is not hard to see that since the primal optimization problem of struct SVM has an exponential number of constraints in the length of the input the dual will have an exponential number of variables. One approach is thus to solve the dual while imposing some notion of sparsity on the dual variables. Intuitively the success of this approach can be explained by the fact that we only expect a small number of constraints to be active or nearly active at a certain point.

The Frank-Wolfe algorithm (Frank and Wolfe 1956) is a popular method for constrained convex optimization consisting in taking a first-order approximation of the objective function and doing a linear search over the set of constraints to update the current point; leveraging sparsity . Over the years, many improvements and alternatives were developed to address the particularities of various frameworks. Jaggi (Jaggi 2013) presents a variety of stronger convergence results for Frank-Wolfe-type algorithms found in the literature (e.g. F-W with approximate linear subproblems, away steps, etc.). Lacoste-Julien et al. (Lacoste-Julien et al. 2012) propose a randomized block-coordinate variant of the F-W algorithm. The paper goes on to demonstrate its use in the case of SVM's and the advantages it yields over stochastic subgradient methods and other available SVM solvers. S. Lacoste-Julien and M.Jaggi (Lacoste-Julien and Jaggi 2015) also go on to show that the "Away-steps", "Pairwise FW", "fully-corrective FW", and "Minimum norm point" variants of the F-W algorithm all achieve linear convergence under a weaker condition than strong convexity.

Another approach developed in (Lacoste-Julien et al. 2012) is to to leverage the min-max structure of the loss-augmented decoding problem. The latter arises naturally in struct SVM as we are minimizing the loss over the labels which maximize the score. In many cases of interest the saddle-point problem has a linear objective function which allows us to use efficient convex relaxation techniques. Moreover we can exploit the structure of the problem in conjuction with ideas from combinatorial optimization to guarantee tractability in some well known cases.

Project

The type of project we wish to present is at the crossing of two lines of work: analysis and practical evaluation. Our project will, therefore, be composed of two main parts:

- The first part is a literature review on the subject. We intend to go deeper into some of the papers cited above, studying the assumptions they make and identifying the critical points in the proofs that forbid using softer assumptions. We will also compare the papers, showing how they relate to each other and how they overcome previous difficulties related to structured predictions. Finally, this part would serve as an overview of all previous results that will allow us to identify the insights and limitations of this class of results and potentially some holes in the results which could be used as a starting point for future research.
- The second part consists of a practical evaluation of the method proposed in and of other methods depending on the time constraints. The idea is to see first-handedly the effect of the various variants of FW on real data. This way, we can better compare/justify the usage of these methods in real-world scenarios. We intend to run some evaluation of the model on both toy/synthetic datasets and real datasets with a few numbers of network configurations. datasets, we aim to construct a practical example in which this method helps the model to converge to an interesting global minimum. On real datasets, the objective would be to determine if the gain from using this method is significant or not in real case uses when compared to more conventional methods.

Our project will take the form of an article detailing the analysis of the papers and our experiments on real data, as well as a repository including the code used for the experiments.

References

Frank, Marguerite and Philip Wolfe (1956). "An algorithm for quadratic programming". In: *Naval Research Logistics Quarterly* 3.1-2, pp. 95–110. URL: https://EconPapers.repec.org/RePEc:wly:navlog:v:3:y:1956:i:1-2:p:95-110.

Jaggi, Martin (2013). "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization." In: *Proceedings of the 30th International Conference on Machine Learning*, pp. 427–435. URL: http://infoscience.epfl.ch/record/229246.

Lacoste-Julien, Simon and Martin Jaggi (2015). "On the Global Linear Convergence of Frank-Wolfe Optimization Variants". In: *arXiv e-prints*, arXiv:1511.05932, arXiv:1511.05932. arXiv: 1511.05932 [math.OC].

Lacoste-Julien, Simon et al. (2012). "Stochastic Block-Coordinate Frank-Wolfe Optimization for Structural SVMs". In: *CoRR* abs/1207.4747. arXiv: 1207.4747. URL: http://arxiv.org/abs/1207.4747.

Taskar, Benjamin, Simon L. Julien, and Michael I. Jordan (2006). "Structured Prediction, Dual Extragradient and Bregman Projections". In: *Journal of Machine Learning Research* 7, pp. 1627–1653. URL: http://dblp.uni-trier.de/rec/bibtex/journals/jmlr/TaskarLJ06.