Numerical Optimization to AI, and Back

Sathya N. Ravi
University of Wisconsin, ravi5@wisc.edu

The impact of numerical optimization on modern data analysis has been quite significant – today, these methods lie at the heart of most statistical machine learning applications in domains spanning genomics, finance and medicine. The expanding scope of these applications (and the complexity of the associated data) has continued to raise the expectations of various criteria associated with the underlying algorithms. Broadly speaking, my research work can be classified into two AI categories: Optimization in ML (Opt-ML) and Optimization in CV (Opt-CV). The expanding scope of these applications (and the complexity of the associated data) has continued to raise the expectations of various criteria associated with the underlying algorithms. It is well known that problems in these areas are not only mathematically interesting but also motivated by practical considerations that arise in the analysis of real world datasets. My research contributes to this endeavor by focusing on the algorithmic and learning issues involved in ML and CV by borrowing ideas from Statistics and Probability theory. I will describe two projects in detail for each of the above two categories (one ongoing and one published for each category), and a brief description of a slightly more theoretical project that I have contributed to in nontrivial ways, during the course of my PhD at UW Madison. A checkmark (✓) indicates that the project has been peer reviewed and published, whereas a 🔍 indicates that the paper is under review.

✓Experimental Design (ED) is a problem with deep foundations dating back at least to the early 1900s in Opt-ML. Here, given the features/covariates x_i's, an experimenter must conduct an experiment in order to obtain the value of the dependent (or response) variables y_i's. The focus of much of the classical work on this topic is to maximize the amount of information that the full experiment yields for a given (or least) amount of work. Although the literature is very mature, not many strategies are available when these design problems appear in the context of sparse linear models commonly encountered in high dimensional machine learning. In this work, we study this budget constrained design where the underlying regression model involves a ℓ_1-regularized linear function. More practically, this makes the model interpretable in the sense that it identifies the important features that are used for predictions. While the obvious strategy involves solving a combinatorial nonlinear optimization problem, we provided two tractable formulations: the first is motivated geometrically whereas the second is algebraic in nature. We show how both these formulations can be solved efficiently for large datasets. In particular, our algorithms require one largest eigenvector computation or a rank one update in addition to the first order oracle (gradient, function value). As practical application of our algorithms, with the help of my adviser, I was able to test the formulations on a large neuroimaging dataset and show that cost savings in longitudinal studies aimed at clinical trials are possible (Ravi et al. 2016). Our paper was presented in ICML, 2016.

✓Fast Filter Flow is a framework that can be used to model various problems in CV. We proposed an algorithm to solve the *Filter Flow* problem (Ravi et al. 2017). To state the problem, let us suppose that relationships between pairs of images are modeled using a linear transformation on a (very) high dimensional space. Then, various constraints are specifically imposed based on the problem one wishes to solve, say Optical Flow, Stereo Matching etc.. The goal is to learn an appropriate relationship that minimizes reconstruction error while satisfying the catalog of constraints. But the computational time involved in solving Filter Flow makes it infeasible for practical purposes, for instance, it takes 10 hours to compute optical flow using standard optimization solvers. I showed that even if one wishes to impose these set of constraints, the problem is amenable to powerful convex optimization algorithms that can exploit multiple processors in a computer. In this paper, I explored different optimization formulations to make it efficient and at the same time preserved the theoretical properties that Filter Flow offers. While the formulation was proposed earlier in 2013, the authors used generic all-purpose optimization software which can be extremely slow on real world examples. Using techniques from convex optimization, we developed a lock-free algorithm that can be used to efficiently solve the filter flow problem. Empirically, we achieved a 20x speed up for optimization and also provided convergence analysis for our proposed algorithm. Our paper was presented in CVPR, 2017.

🔍How to use constraints in the era of Deep Learning? A number of results have recently demonstrated the benefits of incorporating various constraints when training deep architectures in vision and machine learning. The advantages range from guarantees for statistical generalization to better accuracy to compression. But support for general constraints within widely used libraries remains scarce and their broader deployment within many applications that can benefit from them remains under-explored. Part of the reason is that Stochastic gradient descent (SGD), the workhorse for training deep neural networks, *does not natively deal with constraints with global scope very well.* In this paper, we revisit a classical first order scheme from numerical optimization, Conditional Gradients (CG), that has, thus far had limited applicability in training deep models. We show via rigorous analysis how various constraints can be naturally handled by modifications of this algorithm. We designed an algorithm that can explicitly constrain the most successful measure of complexity of neural networks called as the Path norm. We improved upon the performance of Path norm regularized networks with no additional effort, using Path norm constrained networks. We provide convergence guarantees and show a suite of *immediate benefits* that are possible — from training ResNets with fewer layers but better accuracy simply by substituting in our version of CG to faster training of GANs with 50% fewer epochs in image inpainting applications to provably better generalization guarantees using efficiently implementable forms of recently proposed regularizers. Our paper has been will be presented at AAAI 2019.

🔍Robustness in CV. The practicality of problems in CV has attracted many researchers thereby accelerating the pace of research. While this pace is shown to be fruitful in terms of increasing the performance of the end-to-end framework, recent flurry of papers has shown the fragility of the algorithms, specifically with respect to random perturbations. We revisited the Blind Deconvolution problem with a focus on understanding its robustness and convergence properties. Provable robustness to noise and input perturbation is receiving recent interest in vision, from obtaining immunity to adversarial attacks to assessing and describing failure modes of algorithms in mission critical applications. Further, many blind deconvolution methods based on deep architectures internally make use of or optimize the basic formulation. Hence, a clearer understanding of how this submodule behaves, when it can be solved, and what noise injection it can tolerate is a first order requirement. We derived new insights into the theoretical underpinnings of blind deconvolution. The algorithm that emerges has nice convergence guarantees and is provably robust in a sense we formalize in the paper. Interestingly, these technical results play out very well in practice, where on standard datasets our algorithm yields results competitive with or superior to the state of the art. This is ongoing work, which we are planning to submit to a peer reviewed conference/journal soon. A preliminary version of our paper can be found in Arxiv (Ravi, Mehta, and Singh 2018).

🔍Algorithmic summarization using Coresets. Smoothness has played a huge role in the success of ML systems in predictive analysis especially in large scale settings. In fact, there is a common belief that if an algorithm A_1 that outperforms A_2 in the smooth setting is expected to behave similarly for simple ML problems even in the nonsmooth setting. The Frank-Wolfe (FW) Method is a classical optimization algorithm that has been extensively applied for *smooth* problems in Machine Learning. But many problems in these areas are naturally expressed as a non-smooth (often convex) optimization model, for example, Hinge loss SVM, Multiway Graph cuts to name a few. In this work, the goal was to evaluate whether FW type methods can be derived for such non-smooth problems in Vision and Machine Learning. I showed that by bringing together both the classical $\epsilon-$subdifferential and approximate subdifferential idea introduced in a 1990s work by (White 1993), we can define FW type algorithm for such problems. Using this construction, we were also able to analyze the sparsity of the solution at a given iteration which turns out to be intricately related to a concept that is well studied for over a decade in the Computational Geometry literature called as a "coreset". Intuitively, a coreset is a subset of the original dataset which behaves like the entire dataset, that is, any statistic computed using a coreset will be provably close to the quantity if computed using the entire dataset. Hence in some cases, the algorithm provides solutions to these problems in time complexity bounds that are "independent" of the size of the input problem. We then provide analysis to various problems in ML to demonstrate the applications of the proposed algorithm, see our paper (Ravi, Collins, and Singh 2017). We are currently revising the paper for publication in Informs Journal of Optimization.

Future Plans: I believe that the optimization techniques that I have used over the years for developing both practical and theoretical understanding of ML and CV problems can be applied to many more open problems. Designing more efficient algorithms for many problems is an active line of research especially in the large scale setting. Having said that, recent set of results in the ML community also suggest that the term "efficiency" requires some rethinking in mission critical real world applications. In particular, other properties such as stability, fairness, privacy etc. are becoming increasingly important like never before. I am really excited about these aspects of research in ML/CV and their interplay with practical constraints.

References

Ravi, S. N.; Ithapu, V.; Johnson, S.; and Singh, V. 2016. Experimental design on a budget for sparse linear models and applications. In *ICML*.

Ravi, S. N.; Xiong, Y.; Mukherjee, L.; and Singh, V. 2017. Filter flow made practical: Massively parallel and lock-free. In *CVPR*.

Ravi, S. N.; Collins, M. D.; and Singh, V. 2017. A deterministic nonsmooth frank wolfe algorithm with coreset guarantees. *arXiv:1708.06714. Under Review*.

Ravi, S. N.; Mehta, R.; and Singh, V. 2018. Robust blind deconvolution via mirror descent. *arXiv:1803.08137*.

White, D. 1993. Extension of the frank-wolfe algorithm to concave nondifferentiable objective functions. *Journal of optimization theory and applications*.