The Exponentiated Functional Gradient Algorithm for Structured Prediction Problems

J. Andrew Bagnell

with John Langford, Nathan Ratliff, and David Silver

Maximum margin structured prediction has recently gained prominence within the learning community as a powerful method for predicting outputs that are naturally interconstrained, including, for instance route prediction, parsing, and sequence labeling. Until recently, this approach was unfortunately limited in terms of scalability, convergence, and memory requirements. [Taskar06] investigated saddle-point gradient methods for optimization that are well suited to a class of optimization problems where the prediction problem can be written as a linear program. [Ratliff06, Ratliff07b] developed an alternative approach using a convex regularized risk formulation of the maximum margin structured prediction problem. This objective is then optimized by a direct generalization of gradient descent, popular in convex optimization, called the subgradient method. Importantly, the implementation of this learning algorithm is simple and intuitive. The central computation during training involves iteratively executing the *same* inference algorithm that will also be run during the prediction (test) phase of the algorithm.

The primal version of the maximum margin problem depends on a linear representation in terms of pre-specified features. In recent work [Ratliff07a], the range of applicability of the structured maximum margin formulation was extended by a method (inspired by the AnyBoost algorithm of [Mason99]) for “boosting” in new features that make the structured prediction problem easier to solve. These methods can be understood as “functional gradient” techniques: they try to approximate, using a regression or classification algorithm, the direction in the space of functions that would best improve the structured prediction algorithm’s performance. The resulting approximations are added together, as gradient steps, in a linear model to produce the final score function for the structured prediction problem. We argue, both theoretically and empirically, that for a large class of these prediction problems it is more natural to generalize *exponentiated gradient descent* to the space of functions that it is to generalize vanilla gradient descent. We present two novel exponentiated functional gradient algorithms: one that builds an additive model; and a second with the flavor of a reduction algorithm [Beygelzimer05] that iteratively trains a single regressor to help solve the overall structured prediction problem.

We demonstrate that these exponentiated functional gradient descent algorithms yield state-of-the-art results on a series of tasks including imitation learning in outdoor robotics, sequence labeling, and extraction of road networks from satellite imagery. Empirically the exponentiated functional gradient approaches significantly outperform the existing linear or structured boosting approaches.

[Beygelzimer05] Beygelzimer, A., Dani, V., Hayes, T., Langford, J., & Zadrozny, B. Error limiting reductions between classification tasks. ICML 2005.

[Mason99] Mason, L., J.Baxter, Bartlett, P., & Frean, M. Functional gradient techniques for combining hypotheses. Advances in Large Margin Classifiers. MIT Press.

[Ratliff06] Ratliff, N., Bagnell, J. A., & Zinkevich, M. (2006). Maximum margin planning. ICML 2006.

[Ratliff07a] Ratliff, N., Bradley, D., Bagnell, J. A., Chestnutt, J. Boosting Structured Prediction for Imitation Learning, NIPS 2007.

[Ratliff07b] Ratliff, N., Bagnell, J.A., & Zinkevich, M. (Approximate) Sub-gradient methods for Structured Prediction. AISTAT 2007.

[Taskar05]Taskar, B., Chatalbashev, V., Guestrin, C., & Koller, D. Learning structured prediction models: A large margin approach. ICML 2005.

[Taskar06] Taskar, B. Lacost-Julien, S. & Jordan, M. (2006) Structured prediction via the extra-gradient method. NIPS 2006.