Functional Bundle Methods

Nathan Ratliff
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213

J. Andrew Bagnell
Robotics Institute and MLD
Carnegie Mellon University
Pittsburgh, PA 15213

1 Introduction

Recently, gradient descent based optimization procedures and their functional gradient based boosting generalizations have shown strong performance across a number of convex machine learning formulations. They are particularly alluring for structured prediction problems due to their low memory requirements [5], and recent theoretical work has show that they converge fast across a wide range of problems in terms of both optimization and generalization [5, 10, 11]. Importantly, they have effective and efficient nonlinear generalizations in the form of functional gradient boosting algorithms. These generalizations have seen success in a number of real-world problems [6, 3].

Unfortunately, these functional gradient boosting algorithms are often inefficient in terms of their representation: the algorithm adds a new nonlinear base learner to its hypothesis at each iteration, regardless of whether that new base learner already correlates strongly with a previous learner. Recent work in bundle methods for machine learning [12] has shown bundle optimization to be very efficient in terms of their representation, particularly for SVM learning problems [2]. In this paper, we expand on the idea of representation efficiency by generalizing bundle methods to function spaces using techniques from functional gradient boosting. In our derivation, we discuss generalizing bundle methods to function spaces [1]. In particular, we derive the functional bundle and use regularization path arguments of [9] to provide a straightforward and efficient method for optimizing it. This bundle optimization acts to more efficiently utilize each nonlinear learner. We demonstrate our approach on binary classification problems using the MNIST data set as well as on a structured prediction problem known as Maximum Margin Planning [4]. In all cases, we successfully learn the desired concept using only a small nonlinear function representation.

2 Derivation

Let our optimization problem take the following form:

$$\begin{aligned} \text{minimize } \mathcal{R}[f] &= \sum_{i=1}^{N} l_i(f(x_i)) \\ s.t.\ \|f\|_1 &\leq c, \end{aligned} \tag{1}$$

where f is in the linear span of a hypothesis space (base learner space) $\mathcal{H}$, and the norm measures the L_1 size of the coefficient expansion of f. Specifically, if $f = \sum_j \alpha_j h_j$ with $h_j \in \mathcal{H}$, then $\|f\|_1 = \sum_j |\alpha_j|$.

Following operations analogous to those of linear bundle methods [12], at each iteration, we find the functional gradient of the objective and use the set of functional gradients that have been encountered up to this point to lower bound the function. Denote the set of functional gradients computed at points f_t by $g_t = \sum_i \eta_t^i \delta_{x_i}$. Then the lower bound takes the following form:

$$\begin{aligned} \mathcal{R}[f] &\geq \max_t \Big\{ \mathcal{R}[f_t] + \langle g_t,\ f - f_t \rangle \Big\} \\ &= \max_t \left\{ \mathcal{R}[f_t] + \sum_{i=1}^{N} \eta_t^i \Big(f(x_i) - f_t(x_i) \Big) \right\}. \end{aligned} \tag{2}$$

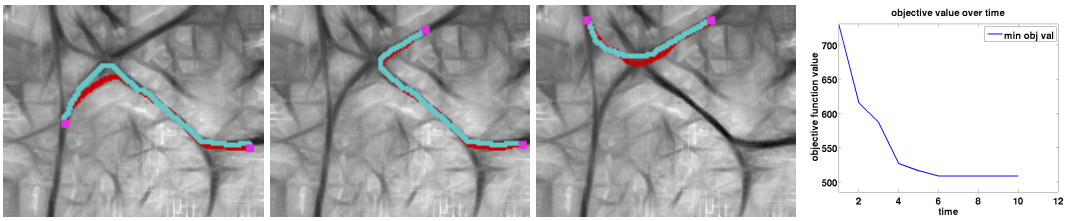

Figure 1: The first three images show the planned (learned) path (cyan) and the desired path (red) overlaying the learned cost across a hold-out region. The final image depicts the progression of objective values across iteration.

Using the regularization path arguments of [9], optimizing this lower bound subject to the constraint $\|f\|_1 \leq c$ is (approximately) equivalent to simply running functional gradient boosting for c/ϵ iterations, where $\epsilon > 0$ is a small step size. When we to do so, however, we'd find that each functional gradient we evaluate is a functional gradient we've already seen:

$$\nabla_f \max_t \left\{ \mathcal{R}[f_t] + \sum_{i=1}^{N} \eta_t^i \Big(f(x_i) - f_t(x_i) \Big) \right\} = \sum_{i=1}^{N} \eta_{t^*}^i \delta_{x_i} = g_{t^*},$$

where t^* is the maximizing t. We can therefore optimize the bundle efficiently without training additional base learners.

3 Preliminary experiments

Problem	4 vs 5	0 vs 1	3 vs 6
Accuracy	98.5	99.7	99.2
Size	7	7	7

We first demonstrate the performance on Regularized Least Squares Classification [8] using binary classification problems drawn from the MNIST data set. The table above shows that by choosing c appropriately, we can constrain the hypothesis representation to remain small while retaining good classification accuracy on these problems. In this case, we found high accuracy using only 7 neural network base learners, each consisting of only 4 hidden nodes.

We additionally implemented our algorithm for a imitation learning structured prediction problem known as Maximum Margin Planning [4]. Following the LEARCH algorithm [7], for this problem we optimize the functional bundle using exponentiated functional gradient descent to attain an exponentiated hypothesis of the form $\exp\{\sum_i \alpha_i h_i\}$. Figure 1 depicts the generalization performance under the functional bundle method. In this case, the algorithm successfully learned the concept using a final cost function consisting of only 4 neural network base learners, each consisting of 25 hidden nodes.

References

[1] S. Hassani. *Mathematical Physics*. Springer, 1998.

[2] T. Joachims. Training linear svms in linear time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

[3] L. Mason, J.Baxter, P. Bartlett, and M. Frean. Functional gradient techniques for combining hypotheses. In *Advances in Large Margin Classifiers*. MIT Press, 1999.

[4] N. Ratliff, J. A. Bagnell, and M. Zinkevich. Maximum margin planning. In *Twenty Second International Conference on Machine Learning (ICML06)*, 2006.

[5] N. Ratliff, J. A. Bagnell, and M. Zinkevich. (online) subgradient methods for structured prediction. In *Artificial Intelligence and Statistics (AIStats)*, 2007.

[6] N. Ratliff, D. Bradley, J. A. Bagnell, and J. Chestnutt. Boosting structured prediction for imitation learning. In *NIPS*, 2006.

[7] N. Ratliff, D. Silver, and J. A. Bagnell. Learning to search: Functional gradient techniques for imitation learning. *Submitted to Autonomous Robotics Special Issue on Robot Learning*, 2009.

[8] Y. Rifkin and Poggio. Regularized least squares classification. In e. a. Suykens, editor, *Advances in Learning Theory: Methods, Models and Applications*, volume 190. IOS Press, 2003.

[9] S. Rosset, J. Zhu, T. Hastie, and R. Schapire. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

[10] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *ICML*, 2007.

[11] S. Shalev-Shwartz and N. Srebro. Svm optimization: Inverse dependence on training set size. In *ICML*, 2008.

[12] A. Smola, S. Vishwanathan, and Q. Le. Bundle methods for machine learning. In *NIPS 20*, 2008.