Single Index Convex Experts: Efficient Estimation via Adapted Bregman Losses

Pradeep Ravikumar[†], Martin Wainwright[†‡], Bin Yu[†‡]
Department of Statistics[†] and Department of EECS[‡]
University of California, Berkeley

1 Introduction

Semiparametric models, consisting of a (finite-dimensional) parametric component, and a non-parametric functional component, are a powerful tool for high-dimensional data analysis. In this work, we study efficient methods for estimating a particular class of semiparametric models, known as *single index regression* models. This model is parameterized by a vector $\beta \in \mathbb{R}^p$ and a function $g : \mathbb{R} \to \mathbb{R}$. The response variables $Y_i \in \mathbb{R}$ are linked to the covariates $X_i \in \mathbb{R}^p$ by the semiparametric regression model

$$Y_i = g(\beta^\top X_i) + \epsilon_i \tag{1}$$

where ϵ_i is an additive zero mean noise, independent of X_i. The function g is assumed to belong to some given class of functions $\mathcal{G}$, for instance differentiable functions, or monotonically increasing functions. Given n i.i.d. samples $S = \{(X_i, Y_i), i = 1, \ldots, n\}$ from the model (1), the task is to estimate both the parametric component β and the unknown function g.

In this model (1), the nonparametric component g takes as its argument a unidimensional summary of the covariates—called the index–via the linear projection $\beta^T X$. For this reason, estimation in this model does not suffer from the curse of dimensionality intrinsic to general non-parametric estimation. Single index models have thus been used for density estimation in many machine learning applications, for instance as product of a collection of single index functions when g is specified [2, 5, 4], as well as sums of single index models; for instance the projection pursuit regression procedure [1] estimates single index models in a stepwise manner, obtaining an additive combination of single index functions as an estimate of the regression function. The methods above, with or without the estimation of g, however all entail non-convex estimation problems. The solutions thus obtained are thus not only suboptimal estimates with respect to their models, but can also be unstable.

In this work, we develop a novel two-stage estimation procedure, in which the loss function applied to β is adapted as a function of the current estimate of g. For the case of monotonic functions g, by using appropriate classes of Bregman divergences, we obtain an overall procedure that involves only tractable convex optimization steps, and is provably Fisher consistent. We start off by noting that estimating a single index model using the least squares loss function is a non-convex estimation task. Consider the population least squares functional, namely $\min_{g \in \mathcal{G}, \beta \in \mathbb{R}^p} \mathbb{E}(Y - g(\beta^\top X))^2$. By computing the Hessian with respect to β, it is straightforward to see that this function is not convex in terms of β for general functions g. (It is convex, for instance, for linear g.) Given this non-convexity, we are motivated to consider a larger class of loss functions, in particular the class of Bregman divergences. For any Bregman function F (roughly, a strictly convex differentiable function), the Bregman divergence $D_F(a, b)$ is defined as,

$$D_F(a \,\|\, b) := F(a) - F(b) - \nabla F(b)^\top (a - b) \tag{2}$$

The Bregman divergence induced by a univariate Bregman function F, between Y and $g(\beta^\top X)$ is then given by,

$$D_F\left(Y, \| \, g(\beta^\top x)\right) = F(Y) - F(g(\beta^\top X)) - f(g(\beta^\top X))(Y - g(\beta^\top X)) \tag{3}$$

where $f = F'$. The least squares loss function is a special case, obtained by setting $F(z) = \frac{1}{2}z^2$. Of interest to us are alternative choices of Bregman distances; in particular, the following result shows that for any monotonic g, there is always a Bregman divergence for which estimation of β reduces to a convex problem:

Proposition 1. *Consider the single index model* (1) *when g belongs to the class $\mathcal{G}$ of monotonically increasing functions. Then for any $g \in \mathcal{G}$, there exists a Bregman divergence $D_{F(g)}$ for which the estimation of β is a convex problem. In particular, define $G(v) = \int_{-\infty}^{v} g(t)dt$, and define the function*

$$F(u) \quad = \quad \sup_{v \in \mathbb{R}} \left\{ v^T u - G(v) \right\} \tag{4}$$

The Bregman divergence $D_{F(g)}$ induced by this choice of F, when applied to the pair y and $g(\beta^\top x)$, takes the form

$$D_{F(g)}\left(y \,\|\, g(\beta^\top x)\right) = G(\beta^\top x) - \beta^\top x y + F(y), \tag{5}$$

which is a convex function of β whenever g is monotonic.

Note that the function (4) is the Fenchel conjugate [3] of the function G. Overall, this result motivates the following practical scheme. Since G is convex for monotonic g, optimizing the "surrogate" function (5) for β is a convex program. On the other hand, for fixed β, estimation of the function g in the single index model (1) is a standard problem in isotonic regression. Thus, we have the following two-stage procedure for estimating a single index model:

(a) Given a convex estimate $\widehat{G}$, alternatively, a monotone estimate $\widehat{g}$, minimize the associated "surrogate" loss function (5) with respect to β.

(b) given an estimate $\widehat{\beta}$, obtain an estimate $\widehat{g}$ by performing a monotone or isotonic regression of the response Y on the covariates $\widehat{\beta}^\top X$.

Note that the key aspect of this method is that the Bregman loss in step (a) is adapted, depending on the current estimate $\widehat{g}$ of the semiparametric component. We have implemented this scheme, and found excellent practical performance on various models. It can be shown that minimization of the surrogate loss (5) is always Fisher consistent for β. In particular, given g^* from the true model, consider the surrogate loss function (5) defined by $G^* = \int g^*$. This method is Fisher consistent, in the sense that the population minimizer is always equal to the true β^*. We are currently exploring conditions under which it can be guaranteed that the sample minimizers over both β and g converge to this population optimum.

References

[1] J.H. Friedman and J.W. Tukey. A projection pursuit algorithm for exploratory data analysis. pages 881–889, 1974.

[2] G. Hinton. Products of experts. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN99)*, 1999.

[3] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*, volume 1. Springer-Verlag, New York, 1993.

[4] S. Roth and M.J. Black. Fields of experts: a framework for learning image priors. In *Computer Vision and Pattern Recognition, CVPR*, 2005.

[5] M. Welling, G.E. Hinton, and S. Osindero. Learning sparse topographic representations with products of student-t distributions. In *Advances in Neural Information Processing Systems, volume 15*, 2002.