OPTIMAL LINEAR AGGREGATION

BY SPENCER GIBSON[1,a],

[1]*Carnegie Mellon University,* [a]*sjgibson@andrew.cmu.edu*

Let $f_1,...,f_n : \mathbb{R}^m \to \mathbb{R}$ be estimators of the true regression function $\mu(x) = \mathbb{E}[y|x]$ for a distribution $(x,y) \sim P$. We give conditions for the existence of an optimal linear aggregation $f(x) = \sum_{i=1}^n \alpha_i f_i(x)$ that minimizes the mean-squared error over the distribution. The existence depends on the invertibility of a Gram matrix, which measures the pairwise similarity between the base functions over the distribution. We also define a vector which measures the similarity between estimators' outputs and true values over the distribution. We then show that consistent estimators of the Gram matrix and the similarity vector yield a consistent estimator of the optimal weighting. We prove this holds even when the Gram matrix estimator is singular. We finish the paper with a plug-in estimator and reduce it to the least squares solution with transformed features.

1. Introduction. A common goal in machine learning and statistics is to learn the relationship between a set of features x, and a response variable, y, governed by a distribution $(x,y) \sim P$. To do so, we often train estimators $f_1,...,f_n$ on a dataset $D_T = \{(x,y)\} \sim P$. Leveraging base learners to create a more accurate model is the crux of ensembling. This was first notably addressed by Freund and Schapire ([2]) who introduced the concept of boosting. Many new boosting and ensembling techniques have been published since including gradient boosting machines ([3]), random forest ([1]), and stacking ([10]).

A similar problem to ensembling is defining an aggregate estimator from a set of base learners $f_1,...,f_n$. Intuitively, an aggregate estimator is a function with the best behavior among the base estimators. Tsybakov ([9]) discusses three well-known problems in the field. The first, linear aggregation (different from our problem), aims to construct an estimator $\hat{f}$ that is as good as the best linear combination of the base learners up to a small remainder term that does not depend on the learners. The second, convex aggregation, considers only convex linear combinations. He lastly considers an aggregate estimator that performs at least as well as the best of the base learners up to a remainder term independent of the learners.

In this paper, we work with linear aggregation. Given a set of base learners $f_1,...,f_n$, a linear aggregate is a linear combination of them, i.e. $f = \sum_{i=1}^n \alpha_i f_i$. We aim to minimize the mean-squared error of f over the joint distribution of feature-value tuples. It is clear from the problem definition that linear aggregation is a specific instance of ensembling as well as a linear aggregate estimator in the terminology of Tsybakov.

The problem setting might appear niche but in fact it is adjacent to many fields, including statistical learning, ensemble methods, and nonparametric regression. Linear aggregation is closely related to boosting ([2]). The general goal of both is to leverage a set of learners to obtain a function with lower mean-squared error, variance, bias, etc. Linear aggregation is a subset of linear regression with basis functions ([6]). Both are similar to kernel regression ([8]). Generalized additive models ([4]) are also closely related.

Why choose linear aggregation among all ensemble methods? Linear aggregation is interpretable because of its simplicity. Indeed, we show in Estimator Implementation (section

MSC2020 subject classifications: Primary 62G08, 62G20; secondary 62C05.

Keywords and phrases: Linear aggregation, consistent estimator, linear regression with basis functions, ensemble methods, convergence in probability, law of large numbers.

4), that the plug-in estimator for the optimal weighting reduces to the estimator for simple linear regression with transformed features. Interpretability is important in practice when engineers want an ensemble with clear meaning or scientists want to draw conclusions from an experiment. Linear aggregation also lends itself to a myriad of computational approaches. We show in Theory and Proofs (section 3) that consistent estimators of the optimal weighting can be derived from consistent estimators of two relevant matrices. These can be calculated by any relevant statistical method like empirical mean, Monte Carlo approximation ([7]) and its derivatives, variational methods ([5]), and deep learning techniques.

2. Paper Structure. The rest of the paper is split in two sections - Theory and Proofs, and Estimator Implementation.

In the Theory and Proofs section, we formalize the optimal solution to linear aggregation in terms of the mean-squared error over the joint distribution under the condition that the base learners are not too redundant. We show the optimal weighting in terms of the Gram matrix of the base learners. Lastly, we show the estimator of the optimal weighting is consistent even when the estimators of the Gram matrix are not necessarily invertible.

In the Estimator Implementation section, we show the plug-in estimator of the optimal weighting is consistent and reduces to the typical finite sample least squares solution with transformed features.

3. Theory and Proofs. We begin with functions $M_1, \dots, M_n$, with $M_i : \mathbb{R}^m \to \mathbb{R}$ for each $i \in [n]$. In the machine learning context, these would be models trained on a dataset $D_T = \{(X_i, Y_i)_{i=1}^N\}$ with $(X_i, Y_i) \sim P$. The range of the functions is $\mathbb{R}$ but the results can be extended simply to higher dimensions.

A simple ensemble of these models is a linear aggregation $M(x) = \sum_{i=1}^n \alpha_i M_i(x)$. We prove that consistent estimators of the optimal weighting exist as long as the base learners are not linearly dependent almost everywhere.

ASSUMPTION 3.1. The functions $\{M_i\}_{i=1}^n$ are not linearly dependent almost everywhere. Specifically, for any $z \in \mathbb{R}^m$ we have $P(\{x \in \mathbb{R}^m \mid (\sum_{i=1}^n z_i M_i(x)) \neq 0\}) > 0$.

ASSUMPTION 3.2. $\mathbb{E}_{x \sim P(x)}[M_i(x)^2] < \infty$ for $i \in \{1, \dots, n\}$, and $\mathbb{E}_{y \sim P(Y)}[y^2] < \infty$.

DEFINITION 3.3. Define the Gram matrix C whose (i, j)-th entry is given by $(C)_{ij} = \mathbb{E}_{x \sim P(x)}[M_i(x) M_j(x)]$. The vector b is defined to have i-th entry $b_i = \mathbb{E}_{(x,y) \sim P}[M_i(x) y]$.

COROLLARY 3.4. *The matrix C is well-defined.*

PROOF. Follows immediately from Cauchy-Schwarz and assumption 3.2. □

DEFINITION 3.5 (Optimal Aggregation Minimizes MSE). For a fixed value of α, define

$$MSE(M) = \mathbb{E}_{(x,y) \sim P}[(y - M(x))^2]$$

We define the optimal weighting to be the one that minimizes the mean-squared error of the model:

$$\alpha^\star = \arg\min_{\alpha} MSE(M)$$

DEFINITION 3.6. The Frobenius norm is used throughout the paper. For a vector v, the Frobenius norm of v is

$$||v||_F = (\sum_{i=1}^{n} v_i^2)^{\frac{1}{2}}$$

For a matrix X, the Frobenius norm of X is given by

$$||X||_F = ||vec(X)||_F = (\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}^2)^{\frac{1}{2}}$$

DEFINITION 3.7. A consistent estimator is one that converges to the true value in probability, e.g. $\hat{x}_n$ is an estimator of x and satisfies the following property: For any $\epsilon > 0$, $\lim_{n\to\infty} Pr(||\hat{x}_n - x||_F > \epsilon) = 0$.

REMARK 3.8. We use the continuous mapping theorem throughout the paper. The theorem is stated below. Its proof can be found in any classic probability text.

THEOREM 3.9 (Continuous Mapping Theorem). *Let $\{X_n\}$ and X be random variables taking values in a metric space S. Suppose that $g : S \to S'$ is a function between metric spaces (S, d_S) and $(S', d_{S'})$, where S' is another metric space. Let D_g denote the set of discontinuity points of g, and assume that Pr$[X \in D_g] = 0$. If $X_n \xrightarrow{\mathbb{P}} X$ as $n \to \infty$, then:*

$$g(X_n) \xrightarrow{\mathbb{P}} g(X) \quad \text{as } n \to \infty.$$

LEMMA 3.10. *C is positive semi-definite.*

PROOF. $C_{ij} = \mathbb{E}_{x\sim P(x)}[M_i(x)M_j(x)] = \mathbb{E}_{x\sim P(x)}[M_j(x)M_i(x)] = C_{ji}$ so C is symmetric. Take any $z \in \mathbb{R}^n$. We have

$$(Cz)_l = \sum_{j=1}^{n} \mathbb{E}[M_l(x)M_j(x)]z_j = \mathbb{E}[\sum_{j=1}^{n} M_l(x)M_j(x)z_j]$$

And

$$\begin{aligned} z^T Cz &= \sum_{i=1}^{n} z_i(Cz)_i \\ &= \sum_{i=1}^{n} z_i \mathbb{E}\left[\sum_{j=1}^{n} M_i(x)M_j(x)z_j\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} z_i M_i(x)M_j(x)z_j\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n} z_i M_i(x)\right)^2\right] \geq 0. \end{aligned}$$

□

REMARK 3.11. In practice, one uses an ensemble when they believe their base models are not redundant. The next lemma shows that unless the functions in the linear aggregate are redundant almost everywhere then C is invertible and a clear theoretically optimal solution exists.

LEMMA 3.12. *If the functions $\{M_i\}_{i=1}^n$ are not linearly dependent almost everywhere then C is positive definite and hence invertible.*

PROOF. Under the assumption that the functions $\{M_i\}_{i=1}^n$ are not linearly dependent almost everywhere, this implies $(\sum_{i=1}^n z_i M_i(x))^2 > 0$ for a region of positive probability measure, and thus,

$$\forall z \in \mathbb{R}^n, z^T C z = \mathbb{E}[(\sum_{i=1}^n z_i M_i(x))^2] > 0$$

This shows C is positive definite. Invertibility follows from positive definiteness. □

COROLLARY 3.13. *C is positive definite and invertible*

PROOF. Follows immediately from assumption 3.1 and Lemma 3.12. □

LEMMA 3.14. *The optimal weighting is given by $\alpha^\star = C^{-1}b$.*

PROOF. The ensemble aims to minimize the mean-squared error over the joint distribution where $MSE(M) = \mathbb{E}_{(x,y)\sim P}[(y - M(x))^2]$. Recall that $\alpha^\star = \arg\min_\alpha MSE(M)$.

From linearity of expectation and finiteness from assumption 3.2, we can write $MSE(M)$ as:

$$E_{(x,y)\sim P}[y^2 - 2y\sum_{i=1}^n \alpha_i M_i(x) + (\sum_{i=1}^n \alpha_i M_i(x))^2] = \alpha^T C\alpha - 2\alpha^T b + E_{(x,y)\sim P}[y^2]$$

The mean-squared error is convex with respect to α since $\frac{\partial^2 MSE(M)}{\partial \alpha^2} = 2C \succeq 0$ from Lemma 3.10.

The local minimizer, and hence global minimizer, is given by $\frac{\partial MSE(M)}{\partial \alpha} = 2C\alpha - 2b = 0 \implies \alpha^\star = C^{-1}b$. Thus, $\alpha^\star = C^{-1}b$ is the optimal weighting.

□

THEOREM 3.15. *For consistent estimators $\hat{C}_n \in \mathrm{GL}(n, \mathbb{R})$ of C, and $\hat{b}_n$ of b, we get a consistent estimator $\hat{\alpha}_n = (\hat{C}_n)^{-1}\hat{b}_n$ of $\alpha^\star$.*

PROOF. The function $f : \mathrm{GL}(n, \mathbb{R}) \to \mathrm{GL}(n, \mathbb{R})$ given by $f(X) = X^{-1}$ is continuous. Since $\hat{C}_n \xrightarrow{\mathbb{P}} C$, we get that $(\hat{C}_n)^{-1} = f(\hat{C}_n) \xrightarrow{\mathbb{P}} f(C) = C^{-1}$. Thus, $\hat{\alpha}_n = (\hat{C}_n)^{-1}\hat{b}_n \xrightarrow{\mathbb{P}} C^{-1}b = \alpha^\star$ and it is consistent. □

REMARK 3.16. In practice, estimators of C and b may not be invertible for finite sample sizes. The theorem below shows the estimator $\hat{\alpha}_n$ using pseudoinverses is still consistent.

THEOREM 3.17. *For consistent estimators $\hat{C}_n$ of C, and $\hat{b}_n$ of b, we get a consistent estimator $\hat{\alpha}_n = (\hat{C}_n)^{+}\hat{b}_n$ of $\alpha^\star$.*

PROOF. We first show that $C_n^+ \xrightarrow{\mathbb{P}} C^{-1}$. Fix an $\epsilon > 0$. We can write

$$Pr(||C_n^+ - C^{-1}||_F > \epsilon) \leq Pr(||C_n^+ - C^{-1}||_F > \epsilon \mid \det(C_n) \neq 0) Pr(\det(C_n) \neq 0) + Pr(\det(C_n) = 0).$$

Since det is continuous and $C_n \xrightarrow{\mathbb{P}} C$ then $det(C_n) \xrightarrow{\mathbb{P}} det(C)$. C is invertible and so $det(C) \neq 0$ so $Pr(det(C_n) \neq 0) \to 1$ as $n \to \infty$. Conversely, $Pr(det(C_n) = 0) \to 0$ as $n \to \infty$. Lastly, $Pr(||C_n^+ - C^{-1}||_F > \epsilon \mid det(C_n) \neq 0) = Pr(||C_n^{-1} - C^{-1}||_F > \epsilon)$ for a sequence $C_n \xrightarrow{\mathbb{P}} C$ where $\forall n, C_n \in \mathrm{GL}(n, \mathbb{R})$. We showed in the previous lemma that this probability converges to 0 as $n \to \infty$. Combining the three limits gives that $C_n^+ \xrightarrow{\mathbb{P}} C^{-1}$.

Since $b_n \xrightarrow{\mathbb{P}} b$, we get that $\hat{\alpha}_n = (\hat{C}_n)^+ \hat{b}_n \xrightarrow{\mathbb{P}} C^{-1}b = \alpha^\star$. □

4. Estimator Implementation.

We give the plug-in estimator assuming that C is invertible and show the estimator is consistent.

COROLLARY 4.1. *Assuming that* $Var(M_i(x)M_j(x)) < \infty$ *and* $Var(M_i(x)Y) < \infty$ *for all* i, j, *using the plug-in (sample mean) estimators from a dataset* $D_E \sim_{iid} P$ *for each term in* C *and* b, *and assuming* $C \in \mathrm{GL}(n, \mathbb{R})$, *then* $\hat{\alpha}_n = (\hat{C}_n)^+ \hat{b}_n$ *is a consistent estimator of* $\alpha^\star = C^{-1}b$.

PROOF. From our variance assumptions, the law of large numbers gives that each component of $(\hat{C}_n)_{ij} \xrightarrow{\mathbb{P}} C_{ij}$, and from basic continuity arguments, $\hat{C}_n \xrightarrow{\mathbb{P}} C$. The same logic shows $\hat{b}_n \xrightarrow{\mathbb{P}} b$. Applying Theorem 3.17 gives the result. □

LEMMA 4.2. *Define* $(\Phi)_{ij} = M_i(x_j)$ *and assume* Φ *is invertible. The usual least squares estimator for* $\alpha^\star$ *is given by* $\hat{\alpha}_{LS(n)} = (\Phi^T\Phi)^{-1}\Phi^T y$. *The plug-in estimator simplifies to the usual least squares solution.*

PROOF. Note $\hat{C}_n = \frac{1}{n}\Phi^T\Phi$ and $(\hat{C}_n)^{-1} = n(\Phi^T\Phi)^{-1}$. Note $\hat{b}_n = \frac{1}{n}\Phi^T y$. Combining yields $\hat{\alpha}_n = (\hat{C}_n)^+ \hat{b}_n = (\Phi^T\Phi)^{-1}\Phi^T y = \hat{\alpha}_{LS(n)}$ as desired. □

REFERENCES

[1] BREIMAN, L. (2001). Random forests. *Machine learning* **45** 5–32.

[2] FREUND, Y., SCHAPIRE, R. E. et al. (1996). Experiments with a new boosting algorithm. In *icml* **96** 148–156. Citeseer.

[3] FRIEDMAN, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232.

[4] HASTIE, T. J. (2017). Generalized additive models. In *Statistical models in S* 249–307. Routledge.

[5] JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning* **37** 183–233.

[6] KOHN, R., SMITH, M. and CHAN, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing* **11** 313–322.

[7] METROPOLIS, N. and ULAM, S. (1949). The monte carlo method. *Journal of the American statistical association* **44** 335–341.

[8] NADARAYA, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications* **9** 141–142.

[9] TSYBAKOV, A. B. (2003). Optimal rates of aggregation. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings* 303–313. Springer.

[10] WOLPERT, D. H. (1992). Stacked generalization. *Neural networks* **5** 241–259.