A Consistent Estimator of the Expected Gradient Outerproduct

Shubhendu Trivedi* TTI-Chicago
Jialei Wang* University of Chicago
Samory Kpotufe TTI-Chicago
Gregory Shakhnarovich TTI-Chicago

Abstract

In high-dimensional classification or regression problems, the expected gradient outerproduct (EGOP) of the unknown regression function f, namely $\mathbb{E}_X \left(\nabla f(X) \cdot \nabla f(X)^\top \right)$, is known to recover those directions $v \in \mathbb{R}^d$ most relevant to predicting the output Y.

However, just as in gradient estimation, optimal estimators of the EGOP can be expensive in practice. We show that a simple rough estimator, much cheaper in practice, suffices to obtain significant improvements on real-world nonparametric classification and regression tasks. Furthermore, we prove that, despite its simplicity, this rough estimator remains statistically consistent under mild conditions.

1 INTRODUCTION

In high-dimensional nonparametric classification or regression problems, the output Y might not depend equally on all input variables in $X = (X^i)_{i=1}^d$. To be more precise, let $Y \approx f(X)$ for some unknown smooth f, it is often the case that f varies most along a few *relevant* coordinates, and varies little along most coordinates. This observation has given rise to many practical variable selection methods.

The usual assumption in variable selection is that $f(X) = g(PX)$, where $P \in \{0,1\}^{k \times d}$ projects X down to $k < d$ relevant coordinates. This assumption is generalized in *multi-index* regression (see e.g. [7, 9, 2, 13]) by letting $P \in \mathbb{R}^{k \times d}$ project X down to a k-dimensional subspace of $\mathbb{R}^d$. In other words, while f might vary significantly along all coordinates of X, it actually only depends on an unknown k-dimensional subspace.

Recovering this relevant subspace (sometimes called *effective dimension reduction* [7]) gives rise to the expected gradient outerproduct (EGOP):

$$\mathbb{E}_X G(X) \triangleq \mathbb{E}_X \left(\nabla f(X) \cdot \nabla f(X)^\top \right).$$

The EGOP recovers the average variation of f in all directions: the directional derivative at x along $v \in \mathbb{R}^d$ is given by $f'_v(x) = \nabla f(x)^\top v$, in other words $\mathbb{E}_X |f'_v(X)|^2 = \mathbb{E}_X \left(v^\top G(X) v \right) = v^\top \left(\mathbb{E}_X G(X) \right) v$.

It follows that, if f does not vary along v, v must be in the null-space of the EGOP matrix $\mathbb{E}_X G(X)$, since $\mathbb{E}_X |f'_v(X)|^2 = 0$. In fact, it is not hard to show that, under mild conditions (f continuously differentiable on a compact space $\mathcal{X}$), the column space of $\mathbb{E}_X G(X)$ is exactly the *relevant* subspace defined by P ([12]).

Interestingly, the EGOP is useful beyond the above multi-index motivation: even if there is no clearly relevant dimension-reduction P, as is likely in practice, one can expect that f does not vary equally in all directions. Instead of dimension-reduction, we might rather weight any direction $v \in \mathbb{R}^d$ according to its relevance as captured by the average variation of f along v (encoded in the EGOP). The weighting approach will be the main use of EGOP considered in this work.

The EGOP can be estimated in various sophisticated ways, which can however be prohibitively expensive. For instance an optimal way of estimating $\nabla f(x)$, and hence the EGOP, is to estimate the slope of a linear approximation to f locally at each $x = X_i$ in an n-sample $\{(X_i, Y_i)\}_{i=1}^n$. Local linear fits can however be prohibitively expensive since it involves multiplying and inverting large-dimensional matrices at all X_i. This can render the approach impractical although it is otherwise well motivated.

The main message of this work is that the EGOP need not be estimated optimally, but just well enough to use towards improving classification or regression, our practical end-goal.

The cheaper estimator considered here is as follows. Let f_n denote an initial estimate of f (we use a kernel estimate);

*Both authors contributed equally to this work

for the i-th coordinate of $\nabla f(x)$, we use the rough estimate

$$\Delta_{t,i} f_n(x) = (f_n(x + te_i) - f_n(x - te_i))/2t,\ t > 0.$$

Let $G_n(x)$ be the outer-product of the resulting gradient estimate $\hat{\nabla} f_n(x)$, the EGOP is estimated as $\mathbb{E}_n G_n(X)$, the empirical average of G_n. The exact procedure is given in Section 3.1.

We first show that this estimator is sound: despite being a rough approximation, it remains a statistically consistent estimate of the EGOP under very general distributional conditions, milder than the usual conditions on proper gradient estimation (see Section 3.2). The main consistency result and key difficulties (having to do with interdependencies in the estimate) are discussed in Section 4.

More importantly, we show through extensive experiments that preprocessing data with this cheaper EGOP estimate can significantly improve the performance of nonparametric classification and regression procedures in real-world applications. This described in Section 5.

In the next Section 2, we start with an overview of relevant work, followed by Section 3 describing the estimator and our theoretical setup.

2 SUMMARY OF RELEVANT WORK

The recent work of [6] considers estimating the coordinates f_i' of ∇f in a similar fashion as in the present work. However [6] is only concerned with a variable selection setting where each coordinate i of X is to be weighted by an estimate of $\mathbb{E}_X |f_i'(X)|$, which is their quantity of interest. This work addresses the more general approach of estimating the EGOP, its consistency and applicability.

Multiple methods have been developed for multi-index regression analysis, some using the so-called *inverse regression* approach (e.g. [7]), and many of them incorporating the estimation of derivate functionals of the unknown f. These approaches can already be found in early work such as [9], and typically estimate ∇f as the slope of local linear approximations of f.

Recent works of [12, 8] draws a clearer link between the various approaches to multi-index regression, and in particular relate the EGOP to the *covariance*-type matrices estimated in inverse regression. Furthermore, [8] proposes an alternative to estimating local linear slopes: their method estimates ∇f via a regularized least-squares objective over an RKHS. This is however still expensive since the least-square solution involves inverting an $n \times n$ feature matrix. In contrast our less sophisticated approach will take time in the order of n times the time to estimate f_n (f_n in practice could be a fast kernel regressor employing fast range-search methods).

The main use of the EGOP in the context of multi-index regression (as in the above cited work) is to recover the relevant subspace given by P in the model $f(x) = g(Px)$. The data can then be projected to the estimated subspace before projection.

While we do not argue for a particular way to use the EGOP to preprocess data, our experiments focus on the following use: let $VDV^\top$ be a spectral decomposition of the estimated EGOP, transform the input x as $D^{1/2}V^\top x$. Thus we do not rely on the multi-index model holding, but rather on a more general model where P might be a full-dimensional rotation (i.e. all directions are relevant), but g varies more in some coordinate than in others. The diagonal element $D_{i,i}$ recovers $\mathbb{E}_X(g_i'(X))^2$ where g_i' denotes coordinate i of ∇g, while $V^\top$ recovers P.

3 SETUP AND DEFINITIONS

We consider a regression or classification setting where the input X belongs to a space $\mathcal{X} \subset \mathbb{R}^d$, of bounded diameter 1. The output Y is real. We are interested in the unknown *regression function* $f(x) \triangleq \mathbb{E}[Y|X = x]$ (in the case of classification with $Y \in \{0, 1\}$, this is just the probability of 1 given x).

For a vector $x \in \mathbb{R}^d$, let $\|x\|$ denote the Euclidean norm, while for a matrix A, let $\|A\|_2$ denote the spectral norm, i.e. the largest singular value $\sigma_{\max}(A)$.

We use $A \circ B$ to denote the entry-wise product of matrices A and B.

3.1 ESTIMATING THE EGOP

We let μ denote the marginal of $P_{X,Y}$ on $\mathcal{X}$ and we let μ_n denote its empirical counterpart on a random sample $\mathbf{X} = \{X_i\}_{i=1}^n$. Given a labeled sample $(\mathbf{X}, \mathbf{Y}) = \{(X_i, Y_i)\}_1^n$ from $P_{X,Y}^n$, we estimate the EGOP as follows.

We consider a simple kernel estimator defined below, using a Kernel K satisfying the following admissibility conditions:

Definition 1 (Admissible Kernel)**.** *$K : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is non-increasing, $K > 0$ on $[0, 1)$, and $K(1) = 0$.*

Using such an admissible kernel K, and a bandwidth $h > 0$, we consider the regression estimate $f_{n,h}(x) = \sum_i \omega_i(x) Y_i$ where

$$\omega_i(x) = \frac{K(\|x - X_i\|/h)}{\sum_j K(\|x - X_j\|/h)} \text{ if } B(x, h) \cap \mathbf{X} \neq \emptyset,$$
$$\omega_i(x) = \frac{1}{n} \text{ otherwise.}$$

For any dimension $i \in [d]$, and $t > 0$, we first define

$$\Delta_{t,i} f_{n,h}(x) \triangleq \frac{f_{n,h}(x + te_i) - f_{n,h}(x - te_i)}{2t}.$$

This is a rough estimate of the line-derivative along coordinate i. However, for a robust estimate we also need to ensure that enough sample points contribute to the estimate. To this end, given a confidence parameter $0 < \delta < 1$ (this definiton for δ is assumed in the rest of this work), define $A_{n,i}(X)$ as the event that

$$\min_{s\in\{t,-t\}} \mu_n(B(X + se_i, h/2)) \geq \frac{2d \ln 2n + \ln(4/\delta)}{n}.$$

The gradient estimate is then given by the vector

$$\hat{\nabla} f_{n,h}(x) = \left(\Delta_{t,i} f_{n,h}(x) \cdot \mathbf{1}_{A_{n,i}(x)}\right)_{i\in[d]}.$$

Note that, in practice we can just replace $A_{n,i}(X)$ with the event that the balls $B(X + se_i, h), s \in \{-t, t\}$, contain samples.

Finally, define $G_n(x)$ as the outer-product of $\hat{\nabla} f_{n,h}(x)$, we estimate $\mathbb{E}_X G(X)$ as

$$\mathbb{E}_n G_n(X) \triangleq \frac{1}{n}\sum_{i=1}^{n} \hat{\nabla} f_{n,h}(X_i) \cdot \hat{\nabla} f_{n,h}(X_i)^\top.$$

3.2 DISTRIBUTIONAL QUANTITIES AND ASSUMPTIONS

For the analysis, our assumptions are quite general. In fact we could simply assume, as is common, that μ has lower-bounded density on a compact support $\mathcal{X}$, and that f is continuously differentiable; all the assumptions below will then hold. We list these more general detailed assumptions to better understand the minimal distributional requirements for consistency of our EGOP estimator.

A1 (Noise)**.** Let $\eta(X) \triangleq Y - f(X)$. We assume the following general noise model: $\forall\delta > 0$ there exists $c > 0$ such that $\sup_{x\in X} \mathbb{P}_{Y|X=x}(|\eta(x)| > c) \leq \delta$. We denote by $C_Y(\delta)$ the infimum over all such c. For instance, suppose $\eta(X)$ has exponentially decreasing tail, then $\forall\delta > 0$, $C_Y(\delta) \leq O(\ln 1/\delta)$.

Last the variance of $(Y|X = x)$ is upper-bounded by a constant σ_Y^2 uniformly over $x \in X$. The next assumption is standard for nonparametric regression/classification.

A2 (Bounded Gradient)**.** Define the *τ-envelope* of $\mathcal{X}$ as $\mathcal{X} + B(0,\tau) \triangleq \{z \in B(x,\tau), x \in \mathcal{X}\}$. We assume there exists τ such that f is continuously differentiable on the τ-envelope $\mathcal{X} + B(0,\tau)$. Furthermore, for all $x \in \mathcal{X} + B(0,\tau)$, we have $\|\nabla f(x)\| \leq R$ for some $R > 0$, and ∇f is uniformly continuous on $\mathcal{X} + B(0,\tau)$ (this is automatically the case if the support $\mathcal{X}$ is compact).

The next assumption generalizes common smoothness assumptions: it is typically required for gradient estimation that the gradient itself be Hölder continuous (or that f be second-order smooth). These usual assumptions imply the more general assumptions below.

A3 (Modulus of continuity of ∇f)**.** Let $\epsilon_{t,i} = \sup_{x\in\mathcal{X}, s\in[-t,t]} |f_i'(x) - f_i'(x + se_i)|$. We assume $\epsilon_{t,i} \xrightarrow{t\to 0} 0$ which is for instance the case when ∇f is uniformly continuous on an envelope $\mathcal{X} + B(0,\tau)$.

The next two assumptions capture some needed regularity conditions on the marginal μ. To enable local approximations of $\nabla f(x)$ over $\mathcal{X}$, the marginal μ should not concentrate on the boundary of $\mathcal{X}$. This is captured in the following assumption.

A4 (Boundary of $\mathcal{X}$)**.** Define the (t,i)**-boundary** of $\mathcal{X}$ as $\partial_{t,i}(\mathcal{X}) = \{x : \{x + te_i, x - te_i\} \not\subseteq \mathcal{X}\}$. Define the vector $\mu_{\partial_t} = (\mu(\delta_{t,i}(\mathcal{X})))_{i\in[d]}$. We assume that $\mu_{\partial_t} \xrightarrow{t\to 0} \mathbf{0}$. This is for instance the case if μ has a continuous density on $\mathcal{X}$.

Finally we assume that μ has mass everywhere, so that for samples X in dense regions, $X \pm te_i$ is also likely to be in a dense region.

A5 (Full-dimensionality of μ)**.** For all $x \in \mathcal{X}$ and $h > 0$, we have $\mu(B(x,h)) \geq C_\mu h^d$. This is for instance the case if μ has a lower-bounded density on $\mathcal{X}$.

4 CONSISTENCY OF THE ESTIMATOR $\mathbb{E}_n G_n(X)$ OF $\mathbb{E}_X G(X)$

We establish consistency by bounding $\|\mathbb{E}_n G_n(X) - \mathbb{E}_X G(X)\|_2$ for finite sample size n. The main technical difficulties in establishing the main result below have to do with the fact that each gradient approximation $\Delta_{t,h} f_{n,h}(X)$ at a sample point X depends on all other samples in $\mathbf{X}$. These inter-dependencies are circumvented by proceeding in steps which consider related quantities that are less sample-dependent.

Theorem 1 (Main)**.** *Assume A1, A2 and A5. Let $t < \tau$ and suppose $h \geq (\log^2(n/\delta)/n)^{1/d}$. There exist $C = C(\mu, K(\cdot))$ and $N = N(\mu)$ such that the following holds with probability at least $1 - 2\delta$. Define $A(n) = \sqrt{Cd \cdot \log(n/\delta)} \cdot C_Y^2(\delta/2n) \cdot \sigma_Y^2 / \log^2(n/\delta)$. Suppose $n \geq N$, we have:*

$$\|\mathbb{E}_n G_n(X)] - \mathbb{E}_X G(X)\|_2 \leq \frac{6R^2}{\sqrt{n}}\left(\sqrt{\ln d} + \sqrt{\ln\frac{1}{\delta}}\right) + \left(3R + \|\epsilon_t\| + \sqrt{d}\left(\frac{hR + C_Y(\delta/n)}{t}\right)\right) \cdot \left[\|\epsilon_t\| + \frac{\sqrt{d}}{t}\sqrt{\frac{A(n)}{nh^d} + 2h^2R^2} + R\left(\sqrt{\frac{d\ln\frac{d}{\delta}}{2n}} + \|\mu_{\partial_t}\|\right)\right]$$

Proof. Start with the decomposition

$$\|\mathbb{E}_n G_n(X) - \mathbb{E}_X G(X)\|_2 \leq \|\mathbb{E}_n G(X) - \mathbb{E}_X G(X)\|_2 + \|\mathbb{E}_n G_n(X) - \mathbb{E}_n G(X)\|_2. \quad (1)$$

The two terms of the r.h.s. are bounded separately in Lemma 2 and 12. □

Remark. Under the additional assumptions A3 and A4, the theorem implies consistency for $t \xrightarrow{n\to\infty} 0$, $h \xrightarrow{n\to\infty} 0$, $h/t^2 \xrightarrow{n\to\infty} 0$, and $(n/\log n)h^d t^4 \xrightarrow{n\to\infty} \infty$, this is satisfied for many settings, for example $t \propto h^{1/4}$, $h \propto (1/n)^{1/(2(d+1))}$.

The bound on the first term of (1) is a direct result of the below concentration bound for random matrices:

Lemma 1. *[10, 3]. Consider a random matrix $A \in \mathbb{R}^{d\times d}$ with bounded spectral norm $\|A\|_2 \leq M$. Let $A_1, A_2, ..., A_n$ be i.i.d. copies of A. With probability at least $1-\delta$, we have*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} A_i - \mathbb{E}A \right\|_2 \leq \frac{6M}{\sqrt{n}} \left(\sqrt{\ln d} + \sqrt{\ln \frac{1}{\delta}} \right).$$

We apply the above concentration to the i.i.d. matrices $G(X), X \in \mathbf{X}$, using the fact that $\|G(X)\|_2 = \|\nabla f(X)\|^2 \leq R^2$.

Lemma 2. *Assume A2. With probability at least $1-\delta$ over the i.i.d sample $\mathbf{X} \triangleq \{X_i\}_{i=1}^n$, we have*

$$\|\mathbb{E}_n G(X) - \mathbb{E}_X G(X)\|_2 \leq \frac{6R^2}{\sqrt{n}} \left(\sqrt{\ln d} + \sqrt{\ln \frac{1}{\delta}} \right).$$

The next Lemma provides an initial bound on the second term of (1).

Lemma 3. *Fix the sample $(\mathbf{X}, \mathbf{Y})$. We have:*

$$\|\mathbb{E}_n G_n(X) - \mathbb{E}_n G(X)\|_2 \leq \mathbb{E}_n \|\nabla f(X) - \hat{\nabla} f_{n,h}(X)\| \cdot \max_{x\in\mathbf{X}} \|\nabla f(x) + \hat{\nabla} f_{n,h}(x)\|. \quad (2)$$

Proof. We have by a triangle inequality $\|\mathbb{E}_n G_n(X) - \mathbb{E}_n G(X)\|_2$ is bounded by:

$$\mathbb{E}_n \left\| \left(\hat{\nabla} f_{n,h}(X) \cdot \hat{\nabla} f_{n,h}(X)^\top - \nabla f(X) \cdot \nabla f(X)^\top \right) \right\|_2.$$

To bound the r.h.s above, we use the fact that, for vectors a, b, we have

$$aa^\top - bb^\top = \frac{1}{2}(a-b)(b+a)^\top + \frac{1}{2}(b+a)(a-b)^\top,$$

implying that

$$\begin{aligned} \left\| aa^\top - bb^\top \right\|_2 &\leq \frac{1}{2} \left\| (a-b)(b+a)^\top \right\|_2 \\ &\quad + \frac{1}{2} \left\| (b+a)(a-b)^\top \right\|_2 \\ &= \left\| (b+a)(a-b)^\top \right\|_2 \end{aligned}$$

since the spectral norm is invariant under matrix transposition.

We therefore have that $\|\mathbb{E}_n G_n(X) - \mathbb{E}_n G(X)\|_2$ is at most

$$\begin{aligned} &\mathbb{E}_n \|(\nabla f(X) - \hat{\nabla} f_{n,h}(X)) \cdot (\nabla f(X) + \hat{\nabla} f_{n,h}(X))^\top\|_2 \\ &= \mathbb{E}_n \|\nabla f(X) - \hat{\nabla} f_{n,h}(X)\| \cdot \|\nabla f(X) + \hat{\nabla} f_{n,h}(X)\| \\ &\leq \mathbb{E}_n \|\nabla f(X) - \hat{\nabla} f_{n,h}(X)\| \cdot \max_{x\in\mathbf{X}} \|\nabla f(x) + \hat{\nabla} f_{n,h}(x)\|. \end{aligned}$$

□

Thus the matrix estimation problem is reduced to that of an average gradient estimation. The two terms of (2) are bounded in the following two subsections. These sections thus contain the bulk of the analysis. All omitted proofs are found in the supplementary.

4.1 BOUND ON $\mathbb{E}_n \|\nabla f(X) - \hat{\nabla} f_{n,h}(X)\|$

The analysis of this section relies on a series of approximations. In particular we relate the vector $\hat{\nabla} f_{n,h}(x)$ to the vector

$$\hat{\nabla} f(x) \triangleq \left(\Delta_{t,i} f(x) \cdot \mathbf{1}_{A_{n,i}(x)} \right)_{i\in[d]}.$$

In other words we start with the decomposition:

$$\mathbb{E}_n \|\nabla f(X) - \hat{\nabla} f_{n,h}(X)\| \leq \mathbb{E}_n \|\nabla f(X) - \hat{\nabla} f(X)\| + \mathbb{E}_n \|\hat{\nabla} f(X) - \hat{\nabla} f_{n,h}(X)\|. \quad (3)$$

We bound each term separately in the following subsections.

4.1.1 Bounding $\mathbb{E}_n \|\nabla f(X) - \hat{\nabla} f(X)\|$

We need to introduce vectors $\mathbf{I}_n(x) \triangleq \left(\mathbf{1}_{A_{n,i}(x)} \right)_{i\in[d]}$, and $\overline{\mathbf{I}_n(x)} \triangleq \left(\mathbf{1}_{\bar{A}_{n,d}(x)} \right)_{i\in[d]}$. We then have:

$$\mathbb{E}_n \|\nabla f(X) - \hat{\nabla} f(X)\| \leq \mathbb{E}_n \|\nabla f(X) \circ \overline{\mathbf{I}_n(X)}\| + \mathbb{E}_n \|\nabla f(X) \circ \mathbf{I}_n(X) - \hat{\nabla} f(X)\|. \quad (4)$$

The following lemma bounds the first term of (4).

Lemma 4. *Assume A2 and A5. Suppose $h \geq (\log^2(n/\delta)/n)^{1/d}$. With probability at least $1-\delta$ over the*

sample of $\mathbf{X}$*:*

$$\mathbb{E}_n \left\| \nabla f(X) \circ \overline{\mathbf{I}_n(X)} \right\| \leq R \cdot \left(\sqrt{\frac{d \ln \frac{d}{\delta}}{2n}} + \|\mu_{\partial_t}\| \right).$$

The second term of (4) is bounded in the next lemma.

Lemma 5. *Fix the sample* $\mathbf{X}$*. We have* $\max_{X \in \mathbf{X}} \|\nabla f(X) \circ \mathbf{I}_n(X) - \hat{\nabla} f(X)\| \leq \|\epsilon_t\|$.

The last two lemmas can then be combined using equation (4) into the final bound of this subsection.

Lemma 6. *Assume A2 and A5. Suppose* $h \geq (\log^2(n/\delta)/n)^{1/d}$*. With probability at least* $1-\delta$ *over the sample* $\mathbf{X}$*:*

$$\mathbb{E}_n \|\nabla f(X) - \hat{\nabla} f(X)\| \leq R \cdot \left(\sqrt{\frac{d \ln \frac{d}{\delta}}{2n}} + \|\mu_{\partial_t}\| \right) + \|\epsilon_t\|.$$

4.1.2 Bounding $\mathbb{E}_n \|\hat{\nabla} f(X) - \hat{\nabla} f_{n,h}(X)\|$

We need to consider bias and variance functionals of estimates $f_{n,h}(x)$. To this end we introduce the expected estimate $\tilde{f}_{n,h}(x) = \mathbb{E}_{\mathbf{Y}|\mathbf{X}} f_{n,h}(x) = \sum_{i=1}^n w_i(x) f(X_i)$.

The following lemma bounds the bias of estimates $f_{n,h}$. The proof relies on standard ideas.

Lemma 7 (Bias of $f_{n,h}$)**.** *Assume A2. Let* $t < \tau$*. We have for all* $X \in \mathbf{X}$*, all* $i \in [d]$*, and* $s \in \{-t, t\}$*:*

$$|\tilde{f}_{n,h}(X + se_i) - f(X + se_i)| \cdot \mathbf{1}_{A_{n,i}(x)} \leq hR.$$

The following lemma bounds the variance of estimates $f_{n,h}$ averaged over the sample $\mathbf{X}$. To obtain a high probability bound, we relie on results of Lemma 7 in [6]. However in [6], the variance of the estimator if evaluated at a point, therefore requiring local density assumptions. The present lemma has no such local density requirements given that we are interested in an average quantity over a collection of points.

Lemma 8 (Average Variance)**.** *Assume A1. There exist* $C = C(\mu, K(\cdot))$*, such that the following holds with probability at least* $1 - 2\delta$ *over the choice of the sample* $(\mathbf{X}, \mathbf{Y})$*. Define* $A(n) = \sqrt{Cd \cdot \ln(n/\delta)} \cdot C_Y^2(\delta/2n) \cdot \sigma_Y^2$*, for all* $i \in [d]$*, and all* $s \in \{-t, t\}$*:*

$$\mathbb{E}_n |\tilde{f}_{n,h}(X + se_i) - f_{n,h}(X + se_i)|^2 \cdot \mathbf{1}_{A_{n,i}(X)} \leq \frac{A(n)}{nh^d}$$

The main bound of this subsection is given in the next lemma which combines the above bias and variance results.

Lemma 9. *Assume A1 and A2. There exist* $C = C(\mu, K(\cdot))$*, such that the following holds with probability at least* $1 - 2\delta$ *over the choice of* $(\mathbf{X}, \mathbf{Y})$*. Define* $A(n) = \sqrt{Cd \cdot \ln(n/\delta)} \cdot C_Y^2(\delta/2n) \cdot \sigma_Y^2$*:*

$$\mathbb{E}_n \|\hat{\nabla} f(X) - \hat{\nabla} f_{n,h}(X)\| \leq \frac{\sqrt{d}}{t} \sqrt{\frac{A(n)}{nh^d} + 2R^2 h^2}.$$

Proof. In what follows, we first apply Jensen's inequality, and the fact that $(a+b)^2 \leq 2a^2 + 2b^2$. We have:

$$\begin{aligned}
&\mathbb{E}_n \|\hat{\nabla} f(X) - \hat{\nabla} f_{n,h}(X)\| \\
&= \mathbb{E}_n \left(\sum_{i \in [d]} |\Delta_{t,i} f_{n,h}(X) - \Delta_{t,i} f(X)|^2 \cdot \mathbf{1}_{A_{n,i}(X)} \right)^{1/2} \\
&\leq \left(\sum_{i \in [d]} \mathbb{E}_n |\Delta_{t,i} f_{n,h}(X) - \Delta_{t,i} f(X)|^2 \cdot \mathbf{1}_{A_{n,i}(X)} \right)^{1/2} \\
&\leq \frac{\sqrt{d}}{2t} \left(\max_{i \in [d], s \in \{-t,t\}} 4\mathbb{E}_n |f_{n,h}(\tilde{X}) - f(\tilde{X})|^2 \cdot \mathbf{1}_{A_{n,i}(X)} \right)^{1/2}
\end{aligned} \tag{5}$$

where $\tilde{X} = X + se_i$. Next, use the fact that for any $s \in \{-t, t\}$, we have the following decomposition into variance and bias terms

$$\begin{aligned}
&|f_{n,h}(X + se_i) - f(X + se_i)|^2 \\
&\leq 2|f_{n,h}(X + se_i) - \tilde{f}_{n,h}(X + se_i)|^2 \\
&+ 2|\tilde{f}_{n,h}(X + se_i) - f(X + se_i)|^2.
\end{aligned}$$

Combine this into (5) to get a bound in terms of the average bias and variance of estimates $f_{n,h}(X + se_i)$. Apply Lemma 7 and 8 and conclude. □

4.1.3 Main Result of this Section

The following theorem provides the final bound of this section on $\mathbb{E}_n \|\nabla f(X) - \hat{\nabla} f_{n,h}(X)\|$. It follows directly from the decomposition of equation 3 and Lemmas 6 and 9.

Lemma 10. *Assume A1, A2 and A5. Let* $t < \tau$ *and suppose* $h \geq (\log^2(n/\delta)/n)^{1/d}$*. With probability at least* $1 - 2\delta$ *over the choice of the sample* $(\mathbf{X}, \mathbf{Y})$*, we have*

$$\begin{aligned}
\mathbb{E}_n \|\nabla f(X) - \hat{\nabla} f_{n,h}(X)\| \leq \frac{\sqrt{d}}{t} \sqrt{\frac{A(n)}{nh^d} + 2R^2 h^2} \\
+ R \left(\sqrt{\frac{d \ln \frac{d}{\delta}}{2n}} + \|\mu_{\partial_t}\| \right) + \|\epsilon_t\|.
\end{aligned}$$

4.2 BOUNDING $\max_{X \in \mathbf{X}} \|\nabla f(X) + \hat{\nabla} f_{n,h}(X)\|$

Lemma 11. *Assume A1 and A2. With probability at least $1-\delta$, we have*

$$\|\nabla f(X) + \hat{\nabla} f_{n,h}(X)\| \leq 3R + \|\epsilon_t\| + \sqrt{d}\left(\frac{hR + C_Y(\delta/n)}{t}\right).$$

Proof. Fix $X \in \mathbf{X}$.We have

$$\begin{aligned}\|\nabla f(X) + \hat{\nabla} f_{n,h}(X)\| \leq & 2\|\nabla f(X)\| \\ & + \|\nabla f(x) - \hat{\nabla} f_{n,h}(X)\| \\ \leq & 2R + \|\nabla f(X) - \hat{\nabla} f(x)\| \\ & + \|\hat{\nabla} f(X) - \hat{\nabla} f_{n,h}(X)\|. \end{aligned} \tag{6}$$

We can bound the second term of (6) above as follows.

$$\begin{aligned}\|\nabla f(X) - \hat{\nabla} f(X)\| \leq & \|\nabla f(X) \circ \mathbf{I}_n(X) - \hat{\nabla} f(X)\| \\ & + \|\nabla f(X) \circ \overline{\mathbf{I}_n(X)}\| \\ \leq & \|\epsilon_t\| + R,\end{aligned}$$

where we just applied Lemma 5.

For the third term of (6), $\|\hat{\nabla} f(x) - \hat{\nabla} f_{n,h}(x)\|$ equals

$$\sqrt{\sum_{i \in [d]} (|\Delta_{t,i} f_{n,h}(x) - \Delta_{t,i} f(x)| \cdot \mathbf{1}_{A_{n,i}(x)})^2}.$$

As in the proof of Lemma 9, we decompose the above summand into bias and variance terms, that is:

$$\begin{aligned}& |\Delta_{t,i} f_{n,h}(x) - \Delta_{t,i} f(x)| \\ & \leq \frac{1}{t} \max_{s \in \{-t,t\}} |\tilde{f}_{n,h}(x + se_i) - f(x + se_i)| \\ & + \frac{1}{t} \max_{s \in \{-t,t\}} |\tilde{f}_{n,h}(x + se_i) - f_{n,h}(x + se_i)|.\end{aligned}$$

By Lemma 7, $|\tilde{f}_{n,h}(x + se_i) - f(x + se_i)| \leq Rh$ for any $s \in \{-t, t\}$.

Next, by definition of $C_Y(\delta/n)$, with probaility at least $1-\delta$, for each $j \in [n]$, Y_j has value within $C_Y(\delta)$ of $f(X_j)$. It follows that $|\tilde{f}_{n,h}(X + se_i) - f_{n,h}(X + se_i)| \leq C_Y(\delta/n)$ for $s \in \{-t, t\}$.

Thus, with probability at least $1-\delta$, we have

$$\|\hat{\nabla} f(X) - \hat{\nabla} f_{n,h}(X)\| \leq \sqrt{d}\left(\frac{hR + C_Y(\delta/n)}{t}\right).$$

Combine these bounds in (6) and conclude. □

4.3 FINAL BOUND ON $\|\mathbb{E}_n G_n(X) - \mathbb{E}_n G(X)\|_2$.

We can now combine the results of the last two subsections, namely Lemma 10 and 11, into the next lemma, using the bound of Lemma 3.

Lemma 12. *Assume A1, A2 and A5. Let $t < \tau$ and suppose $h \geq (\log^2(n/\delta)/n)^{1/d}$. With probability at least $1 - 2\delta$ over the choice of the sample $(\mathbf{X}, \mathbf{Y})$, we have that $\|\mathbb{E}_n G_n(X) - \mathbb{E}_n G(X)\|_2$ is at most*

$$\left(3R + \|\epsilon_t\| + \sqrt{d}\left(\frac{hR + C_Y(\delta/n)}{t}\right)\right) \cdot \left[\frac{\sqrt{d}}{t}\sqrt{\frac{A(n)}{nh^d} + 2h^2R^2} + R\left(\sqrt{\frac{d \ln \frac{d}{\delta}}{2n}} + \|\mu_{\partial_t}\|\right) + \|\epsilon_t\|\right].$$

5 EXPERIMENTAL EVALUATION

In this section we describe experiments aimed at evaluating the utility of EGOP as a metric estimation technique for regression or classification. We consider a family of non-parametric methods that rely on the notion of distance under a given Mahalanobis metric M, computed as $(x - x')^T M (x - x')$. In this setup, we consider three choices of M: (i) identity, i.e., Euclidean distance in the original space; (ii) the estimated gradient weights (GW) matrix as in [6], i.e., Euclidean distance weighted by the estimated $\Delta_{t,i} f_n$, and (iii) the estimated EGOP matrix $\mathbb{E}_n G_n(X)$. The latter corresponds to Euclidean distance in the original space under linear transform given by $[\mathbb{E}_n G_n(X)]^{1/2}$. Note that a major distinction between the metrics based on GW and EGOP is that the former only scales the Euclidean distance, whereas the latter introduces a rotation.

Each choice of M can define the set of neighbors of an input point x in two ways: (a) k nearest neighbors (kNN) of x for a fixed k, or (b) neighbors with distance $\leq h$ for a fixed h; we will refer to this as hNN. When the task is regression, the output values of the neighbors are simply averaged; for classification, the class label for x is decided by majority vote among neighbors. Note that hNN corresponds to kernel regression with the boxcar kernel.

Thus, we will consider six methods, based on combinations of the choice of metric M and the definition of neighbhors: kNN, kNN-GW, kNN-EGOP, hNN, hNN-GW, and hNN-EGOP.

5.1 SYNTHETIC DATA

We first discuss experiments on synthetic data, the goal of which is to examine the effect of varying the dependence of f on input dimensions on the quality of metric recovered with EGOP and alternative approaches. In these experiments, the output is generated i.i.d. as: $y = \sum_i sin(c_i x_i)$, where the sum is over the dimensions of $x \in \mathbb{R}^d$, and the profile of c determines the degree to which the value of

x_i affects the output. We used $d = 50$-dimensional input sampled over a bounded domain, and set $c[1] = 50$ and $c[i] = 0.6 * c[i-1]$ for $i = 2 : 50$. We consider two cases: (R) the input features are transformed by a random rotation in $\mathbb{R}^d$, *after* y has been generated; and (I) the input features are preserved. Under these conditions we evaluate the out of sample regression accuracy with original metric, GW and EGOP-based metrics, for different value of n; in each experiment, the values of h and t are tuned by cross-validation on the training set.

The first observation from results in Figures 1 is that adapting the metric by either GW or EGOP helps performance across the board. As can be expected, performance of EGOP, however, is not significantly affected by rotation. On the other hand, GW is able to recover a good metric in the no-rotation case, but much less so under rotation. Some insight into the nature of estimated metrics is obtained from the profile of the estimated feature relevance. For GW this consists of values on the diagonal of $\mathbf{M}$, and for EGOP of the (square roots) of the eigenvalues of $\mathbf{M}$. Plots in Figure 1 show these profiles (sorted in descending order). It is clear that EGOP is largely invariant to rotation of the features, and is consistently better at recovering the true relative relevance of features corresponding to the c described above.

5.2 REGRESSION EXPERIMENTS

In this section we present results on several real world datasets. The list of data sets with vital statistics (dimensionality and number of training/test points) is found in Table 1. For each data set, we report the results averaged over ten random training/test splits.

As a measure of performance we compute for each experiment the *normalized mean squared error* (nMSE): mean squared error over test set, divided by target variance over that set. This can be interpreted as fraction of variance in the target unexplained by the regressor.

In each experiment the input was normalized by the mean and standard deviation of the training set. For each method, the values of h or k as wel as t (the bandwidth used to estimate finite differences for GW and EGOP) were set by two fold cross-validation on the training set.

5.3 CLASSIFICATION EXPERIMENTS

The setup for classification data sets is very similar for regression, except that the task is binary classification, and the labels of the neighbors selected by each prediction method are aggregated by simple majority vote, rather than averaging as in regression. The performance measure of interest here is classification error. As in regression experiments, we normalized the data, tuned all relevant parameters by cross validation on training data, and repeated the entire experimental procedure ten times with random training/test splits.

In addition to the baselines listed above, in classification experiments we considered another competitor: the popular feature relevance determination method called ReliefF [4, 5]. A highly engineered method that includes heuristics honed over considerable time by practitioners, it has the same general form of assigning weights to features as do GW and EGOP.

5.4 RESULTS

The detailed results are reported in Tables 1 and 2. These correspond to a single value of training set size. Plots in Figures 2 and 3 show a few representative cases for regression and classification, respectively, of performance of different methods as a function of training set size; it is evident from these that while the performance of all methods tends to improve if additional training data are available, the gaps methods persist across the range of training set sizes.

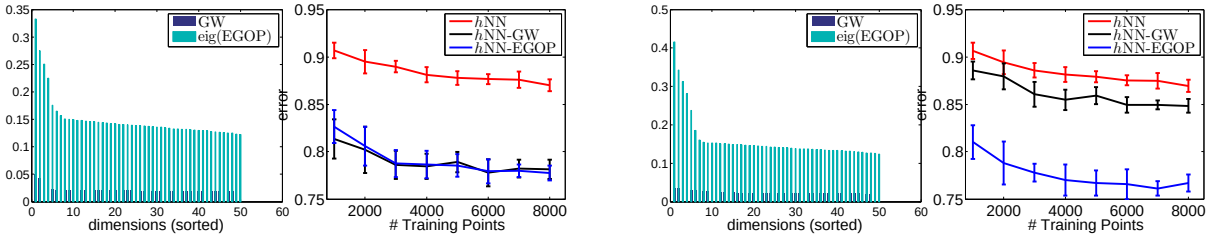
From the results in Tables 1 and 2, we can see that the -EGOP variants dominate the -GW ones, and that both produce gains relative to using the original metric. This is true both for kNN and for kernel regression (hNN) methods, suggesting general utility of EGOP-based metric, not tied to a particular non-parametric mechanism.

We also see that the metrics based on estimated EGOP are competitive with ReliefF, despite the latter benefiting from extensive engineering efforts over the years.

5.5 EXPERIMENTS WITH LOCAL LINEAR REGRESSION

As mentioned earlier in the paper, our estimator for EGOP is an alternative to an estimator based on computing the slope of locally linear regression (LLR) [1] over the training data. We have compared these two estimation methods on a number of data sets, and the results are plotted in Figure 4. In these experiments, the bandwidth of LLR was tuned by a 2-fold cross-validation on the training data.

We observe that despite its simplicity, the accuracy of predictors using EGOP-based metric estimated by our approach is competitive with or even better than the accuracy with EGOP estimated using LLR. As the sample size increases, accuracy of LLR improves. However, the computational expense of LLR-based estimator also grows with the size of data, and in our experiments it became dramatically slower than our estimator of EGOP for the larger data sizes. This confirms the intuition that our estimator is an appealing alternative to LLR-based estimator, offering a good tradeoff of speed and accuracy.



No rotation With rotation

Figure 1: Synthetic data, d=50, with and without rotation applied after generating y from $\mathbf{x}$. In each case we show error of hNN with different metrics (left) and the profile of derivatives recovered by GW and EGOP. The deterioration of the error performance of the Gradient Weights approach after the feature space is subject to a random rotation is noteworthy. See text for details.

Table 1: Regression results, with ten random runs per data set.

Dataset	d	train/test	hNN	hNN-GW	hNN-EGOP
Ailerons	5	3000/2000	0.3637 ± 0.0099	0.3381 ± 0.0087	**0.3264** ± 0.0095
Concrete	8	730/300	0.3625 ± 0.0564	0.2525 ± 0.0417	**0.2518** ± 0.0418
Housing	13	306/200	0.3033 ± 0.0681	**0.2628** ± 0.0652	0.2776 ± 0.0550
Wine	11	2500/2000	0.7107 ± 0.0157	0.7056 ± 0.0184	**0.6867** ± 0.0145
Barrett1	21	3000/2000	0.0914 ± 0.0106	**0.0740** ± 0.0209	0.0927 ± 0.0322
Barrett5	21	3000/2000	0.0906 ± 0.0044	**0.0823** ± 0.0171	0.0996 ± 0.0403
Sarcos1	21	3000/2000	0.1433 ± 0.0087	**0.0913** ± 0.0054	0.1064 ± 0.0101
Sarcos5	21	3000/2000	0.1101 ± 0.0033	0.0972 ± 0.0044	**0.0970** ± 0.0064
ParkinsonM	19	3000/2000	0.4234 ± 0.0386	0.3606 ± 0.0524	**0.3546** ± 0.0406
ParkinsonT	19	3000/2000	0.4965 ± 0.0606	**0.3980** ± **0.0738**	0.4168 ± 0.0941
TeleComm	48	3000/2000	0.1079 ± 0.0099	0.0858 ± 0.0089	**0.0380** ± 0.0059

Dataset	kNN	kNN-GW	kNN-EGOP
Ailerons	0.3364 ± 0.0087	0.3161 ± 0.0058	**0.3154** ± 0.0100
Concrete	0.2884 ± 0.0311	**0.2040** ± 0.0234	0.2204 ± 0.0292
Housing	0.2897 ± 0.0632	**0.2389** ± 0.0604	0.2546 ± 0.0550
Wine	0.6633 ± 0.0119	0.6615 ± 0.0134	**0.6574** ± 0.0171
Barrett1	0.1051 ± 0.0150	**0.0843** ± 0.0229	0.1136 ± 0.0510
Barrett5	0.1095 ± 0.0096	**0.0984** ± 0.0244	0.1120 ± 0.0315
Sarcos1	0.1222 ± 0.0074	**0.0769** ± 0.0037	0.0890 ± 0.0072
Sarcos5	0.0870 ± 0.0051	0.0779 ± 0.0026	**0.0752** ± 0.0051
ParkinsonM	0.3638 ± 0.0443	**0.3181** ± 0.0477	0.3211 ± 0.0479
ParkinsonT	0.4055 ± 0.0413	0.3587 ± 0.0657	**0.3528** ± 0.0742
TeleComm	0.0864 ± 0.0094	0.0688 ± 0.0074	**0.0289** ± 0.0031

Table 2: Classification results with 3000 training/2000 testing.

Dataset	d	hNN	hNN-GW	hNN-EGOP	hNN-ReliefF
Cover Type	10	0.2301 ± 0.0104	0.2176 ± 0.0105	0.2197 ± 0.0077	**0.1806** ± 0.0165
Gamma	10	0.1784 ± 0.0093	0.1721 ± 0.0082	**0.1658** ± 0.0076	0.1696 ± 0.0072
Page Blocks	10	0.0410 ± 0.0042	0.0387 ± 0.0085	**0.0383** ± 0.0047	0.0395 ± 0.0053
Shuttle	9	0.0821 ± 0.0095	0.0297 ± 0.0327	**0.0123** ± 0.0041	0.1435 ± 0.0458
Musk	166	0.0458 ± 0.0057	0.0477 ± 0.0069	**0.0360** ± 0.0037	0.0434 ± 0.0061
IJCNN	22	0.0523 ± 0.0043	0.0452 ± 0.0045	**0.0401** ± 0.0039	0.0510 ± 0.0067
RNA	8	0.1128 ± 0.0038	0.0710 ± 0.0048	**0.0664** ± 0.0064	0.1343 ± 0.0406

Dataset	kNN	kNN-GW	kNN-EGOP	kNN-ReliefF
Cover Type	0.2279 ± 0.0091	0.2135 ± 0.0064	0.2161 ± 0.0061	**0.1839** ± 0.0087
Gamma	0.1775 ± 0.0070	0.1680 ± 0.0075	0.1644 ± 0.0099	**0.1623** ± 0.0063
Page Blocks	0.0349 ± 0.0042	0.0361 ± 0.0048	**0.0329** ± 0.0033	0.0347 ± 0.0038
Shuttle	0.0037 ± 0.0025	0.0024 ± 0.0016	**0.0021** ± 0.0011	0.0028 ± 0.0021
Musk	0.2279 ± 0.0091	0.2135 ± 0.0064	0.2161 ± 0.0061	**0.1839** ± 0.0087
IJCNN	0.0540 ± 0.0061	0.0459 ± 0.0058	**0.0413** ± 0.0051	0.0535 ± 0.0080
RNA	0.1042 ± 0.0063	0.0673 ± 0.0062	**0.0627** ± 0.0057	0.0828 ± 0.0056

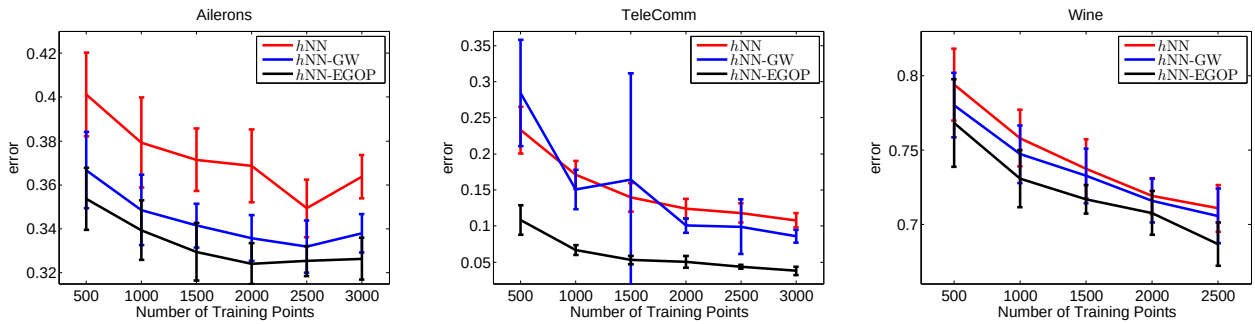


Figure 2: Regression error (nMSE) as a function of training set size for Ailerons, TeleComm, Wine data sets.

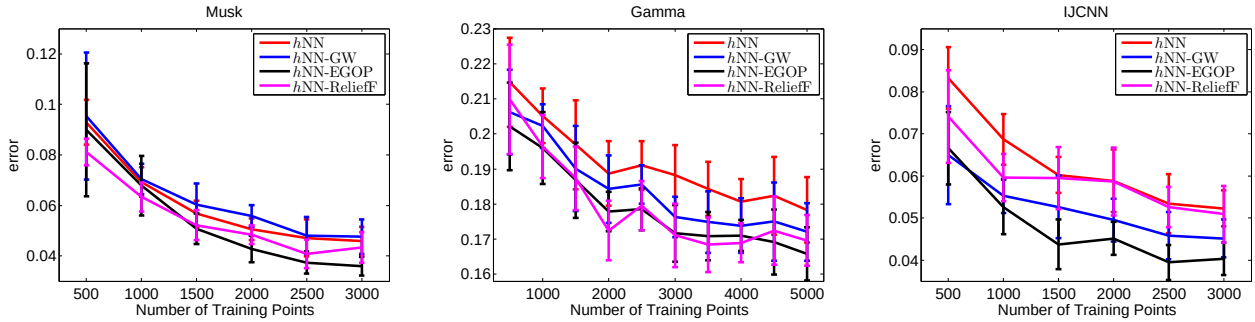


Figure 3: Classification error as a function of training set size for Musk, Gamma, IJCNN data sets.

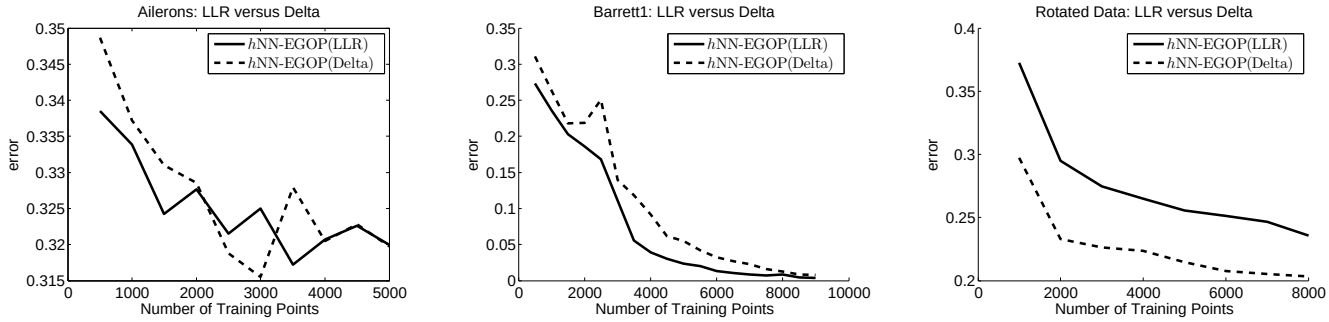


Figure 4: Comparison of EGOP estimated by our proposed method vs. locally linear regression, for Ailerons, Barrett1 and the synthetic (with rotation) data sets (this synthetic dataset is similar to the one used in section 5.1 but with $d = 12$ and $c = [5, 3, 1, .5, .2, .1, .08, .06, .05, .04, .03, .02]$). We also report the following running times (averaged over the ten random runs) for the same using our method and LLR respectively for the highest sample size used in the above real world datasets: Ailerons (128.13s for delta and 347.48s for LLR), Barrett (377.03s for delta and 1650.55s for LLR). Showing that our rough estimator is significantly faster than Local Linear Regression while giving competitive performance. These timings were recorded on an Intel i7 processor with CPU @ 2.40 GHz and 12 GB of RAM.

References

[1] W. S. Cleveland and S. J. Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

[2] W. Hardle, P. Hall, H. Ichimura, et al. Optimal smoothing in single-index models. *The annals of Statistics*, 21(1):157–178, 1993.

[3] S. Kakade. Lecture notes on multivariate analysis, dimensionality reduction, and spectral methods. *STAT 991, Spring*, 2010.

[4] K. Kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In W. R. Swartout, editor, *AAAI*, pages 129–134. AAAI Press / The MIT Press, 1992.

[5] I. Kononenko, E. Simec, and M. Robnik-Sikonja. Overcoming the myopia of inductive learning algo-

rithms with relieff. *Applied Intelligence*, 7:39–55, 1997.

[6] S. Kpotufe and A. Boularias. Gradient weights help nonparametric regressors. In *NIPS*, pages 2870–2878, 2012.

[7] K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[8] S. Mukherjee, Q. Wu, D.-X. Zhou, et al. Learning gradients on manifolds. *Bernoulli*, 16(1):181–207, 2010.

[9] J. L. Powell, J. H. Stock, and T. M. Stoker. Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430, 1989.

[10] J. A. Tropp. User-friendly tools for random matrices: An introduction. *Tutorial at NIPS*, 2012.

[11] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

[12] Q. Wu, J. Guinney, M. Maggioni, and S. Mukherjee. Learning gradients: predictive models that infer geometry and statistical dependence. *The Journal of Machine Learning Research*, 11:2175–2198, 2010.

[13] Y. Xia, H. Tong, W. Li, and L.-X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.

A Omitted proofs

Proof of Lemma 4. By assumption, $\|\nabla f(x)\| \leq R$, so we have

$$\mathbb{E}_n \left\| \nabla f(X) \circ \overline{\mathbf{I}_n(X)} \right\| \leq R \cdot \mathbb{E}_n \left\| \overline{\mathbf{I}_n(X)} \right\|. \tag{7}$$

We bound $\left\| \overline{\mathbf{I}_n(X)} \right\|$ as follows. For any $i \in [d]$, define the events $A_i(X) \equiv \min_{\{t,-t\}} \mu(B(X + se_i, h/2)) \geq 3 \cdot \frac{2d \ln 2n + \ln(4/\delta)}{n}$, and define the vector $\overline{\mathbf{I}(X)} \triangleq \left(\mathbf{1}_{\bar{A}_i(X)}\right)_{i \in [d]}$.

By relative VC bounds [11], let $\alpha_n = \frac{2d \ln 2n + \ln(4/\delta)}{n}$, then with probability at least $1 - \delta$ over the choice of $\mathbf{X}$, for all balls $B \in R^d$ we have $\mu(B) \leq \mu_n(B) + \sqrt{\mu_n(B)\alpha_n} + \alpha_n$. Therefore, with probability at least $1 - \delta$, $\forall i \in [d]$ and x in the sample $\mathbf{X}$, $\bar{A}_{n,i}(x) \Rightarrow \bar{A}_i(x)$.

Moreover, since $\|\overline{\mathbf{I}(X)}\| \leq \sqrt{d}$, by Hoeffding's inequality,

$$\mathbb{P}(\mathbb{E}_n \|\overline{\mathbf{I}(X)}\| - \mathbb{E}_X \|\overline{\mathbf{I}(X)}\| \geq \epsilon) \leq e^{-\frac{2n\epsilon^2}{d}}.$$

It follows that, with probability at least $1 - \delta$,

$$\begin{aligned} \mathbb{E}_n \|\overline{\mathbf{I}_n(X)}\| \leq & \mathbb{E}_n \|\overline{\mathbf{I}(X)}\| \\ \leq & \mathbb{E}_X \|\overline{\mathbf{I}(X)}\| + \sqrt{\frac{d \ln \frac{1}{\delta}}{2n}} \\ \leq & \sqrt{\mathbb{E}_X \|\overline{\mathbf{I}(X)}\|^2} + \sqrt{\frac{d \ln \frac{1}{\delta}}{2n}}, \end{aligned} \tag{8}$$

by Jensen's inequality. We bound each of the d terms of $\mathbb{E}_X \|\overline{\mathbf{I}(X)}\|^2 = \sum_{i \in [d]} \mathbb{E}_X \mathbf{1}_{\bar{A}_i(X)}$ as follows.

Fix any $i \in [d]$. We have $\mathbb{E}_X \mathbf{1}_{\bar{A}_i(X)} \leq \mathbb{E}_X[\mathbf{1}_{\bar{A}_i(X)} | X \in \mathcal{X} \backslash \partial_{t,i}(\mathcal{X})] + \mu(\partial_{t,i}(\mathcal{X}))$. Notice that $\mathbb{E}_X[\mathbf{1}_{\bar{A}_i(X)} | X \in \mathcal{X} \backslash \partial_{t,i}(\mathcal{X})] = 0$ since, by assumption, $\mu(B(x + se_i, h/2)) \geq C_\mu (h/2)^d \geq 3\alpha$ whenever $h \geq (\log^2(n/\delta)/n)^{1/d}$. Hence, we have

$$\sqrt{\mathbb{E}_X \|\overline{\mathbf{I}(X)}\|^2} \leq \sqrt{\sum_{i \in [d]} \mu^2(\partial_{t,i}(\mathcal{X}))}.$$

Combine this last inequality with (7) and (8) and conclude. □

Proof of Lemma 5. For a given coordinate $i \in [d]$, let f_i' denote the directional derivative $e_i^\top \nabla f$ along i. Pick any $x \in \mathcal{X}$. Since $f(x + te_i) - f(x - te_i) = \int_{-t}^{t} f_i'(x + se_i) ds$, we have

$$\begin{aligned} 2t(f_i'(x) - \epsilon_{t,i}) \leq & f(x + te_i) - f(x - te_i) \\ \leq & 2t(f_i'(x) + \epsilon_{t,i}) \end{aligned}$$

Thus $|\frac{1}{2t}(f(x+te_i)-f(x-te_i))-f_i'(x)| \le \epsilon_{t,i}$. We therefore have that $\|\nabla f(x)\circ \mathbf{I}_n(x) - \hat{\nabla} f(x)\|$ equals

$$\sqrt{\sum_{i=1}^{d}\left(f_i'(x)\cdot \mathbf{1}_{A_{n,i}(x)} - \Delta_{t,i}f(x)\cdot \mathbf{1}_{A_{n,i}(x)}\right)^2}$$
$$= \sqrt{\sum_{i=1}^{d}\left(\frac{1}{2t}(f(x+te_i)-f(x-te_i)) - f_i'(x)\right)^2}$$
$$\le \|\epsilon_t\|.$$

□

Proof of Lemma 7. Let $x = X + se_i$. Using a Taylor approximation on f to bound $|f(X_i)-f(x)|$, we have

$$\begin{aligned}|\tilde{f}_{n,h}(x) - f(x)| &\le \sum_{i\in[d]} w_i(x)|f(X_i)-f(x)| \\ &\le \sum_{i\in[d]} w_i(x)\|X_i - x\| \cdot \sup_{\mathcal{X}+B(0,\tau)}\|\nabla f\| \\ &\le hR.\end{aligned}$$

□

Proof of Lemma 8. Fix the sample $\mathbf{X}$ and consider only the randomness in $\mathbf{Y}$. The following result is implicit to the proof of Lemma 7 of [6]: with probability at least $1-2\delta$, for all $X \in \mathbf{X}$, $i \in [d]$, and $s \in \{-t, t\}$, we have (where, for simplicity, we write $x = X + se_i$) $|\tilde{f}_{n,h}(x) - f_{n,h}(x)|^2 \cdot \mathbf{1}_{A_{n,i}(X)}$ is at most

$$\frac{Cd\cdot\log(n/\delta)C_Y^2(\delta/2n)\cdot\sigma_Y^2}{n\mu_n((B(x,h/2))}.$$

Fix $i \in [d]$ and $s \in \{-t,t\}$. Taking empirical expectation, we get $\mathbb{E}_n|\tilde{f}_{n,h}(x) - f_{n,h}(x)|^2$ is at most

$$\frac{\sqrt{Cd\cdot\ln(n/\delta)}\cdot C_Y^2(\delta/2n)\cdot\sigma_Y^2}{n}\sum_{j\in[n]}\frac{1}{n(x_j,h/2)}$$

where $x_j = X_j + se_i$, and $n(x_i,h/2) = n\mu_n(B(x_i,h/2))$ is the number of samples in $B(x_i,h/2)$. Let $\mathcal{Z}\subset\mathbb{R}^d$ denote a minimal $h/4$-cover of $\{X_1,...,X_n\}$. Since $\mathcal{X}$ has bounded diameter, such a cover has size at most $C_{\mathcal{X}}(h/4)^d$ for some $C_{\mathcal{X}}$ depending on the support $\mathcal{X}$ of μ.

Assume every x_j is assigned to the closest $z \in \mathcal{Z}$, where ties can be broken any way, and write $x_j \to z$ to denote such an assignment. By definition of Z, x_j is contained in the ball $B(z,h/4)$, and we therefore have $B(z,h/4)\subset B(x_j,h/2)$.

Thus

$$\begin{aligned}\sum_{j\in[n]}\frac{1}{n(x_j,h/2)} &= \sum_{z\in\mathcal{Z}}\sum_{x_j\to z}\frac{1}{n(x_j,h/2)} \\ &\le \sum_{z\in\mathcal{Z}}\sum_{x_j\to z}\frac{1}{n(z,h/4)} \\ &\le \sum_{z\in\mathcal{Z}}\frac{n(z,h/4)}{n(z,h/4)} = |\mathcal{Z}| \le C_{\mathcal{X}}(h/4)^{-d}.\end{aligned}$$

Combining with the above analysis finishes the proof. □