The EGOP Flow: Local features for Continuous Index Learning

Alex Kokot, Anand Hemmady, Vydhourie Thiyageswaran, Marina Meila

May 23, 2025

Abstract

We introduce the setting of *continuous index learning*, where a function of many variables varies only along a small number of directions at each point. For efficient estimation, it is beneficial for a learning algorithm to *adapt* to the subspace that captures the local variability of the function f. We pose this task as kernel adaptation along a manifold with noise, and present the *Average Gradient Outerproduct (AGOP) Descent* feature learning algorithm, and its continuous counterpart the *Expected Gradient Outer Product (EGOP) flow.* We prove that the EGOP flow adapts to the regularity of the function of interest, showing that under a *supervised noisy manifold* hypothesis, intrinsic dimensional learning rates are achieved for arbitrarily high dimensional noise. On synthetic data, we show that AGOP descent mirrors the feature learning capabilities of deep learning, while two-layer neural networks fail to do so efficiently.

1 Introduction

Kernel methods have recently risen in popularity in the study of machine learning algorithms. Many algorithms have leveraged the corresponding RKHS structure for efficient estimation of sufficiently regular functions [Wainwright, 2019], and functionals [Rao, 2014]. Further, many popular learning algorithms, such as certain neural network architectures and random forests, have been shown to asymptotically correspond to carefully chosen kernels [Jacot et al., 2018, Scornet, 2016]. Thus, the question of kernel engineering [Belkin et al., 2018], the selection of kernels tailored to problems of interest, is of central importance. This not only allows for potential efficiency gains, but also closely emulates the feature learning properties of deep neural networks.

A modern incarnation of kernel engineering is *multi-index learning*, particularly in the case of neural networks ([Boix-Adsera et al., 2023, Damian et al., 2023, Mousavi-Hosseini et al., 2022], etc.). This literature aims to show that the desirable properties of kernel engineering, such as data adaptivity and dimension reduction, are captured implicitly in certain machine learning models. Much work has been done in the *single-index* case, where the outcome of interest depends on the features x solely through their evaluation in a fixed direction v, $x^T v$. These works have shown that two-layer neural networks not only learn this dependence, but also do so efficiently, leading to rapid increases in prediction quality ([Abbe et al., 2024, Bietti et al., 2022, Lee et al., 2024], etc.). In this paper, we consider a setting we call *continuous index learning*, a regression task where the response only depends locally on directions v_x that change smoothly with the features x.

Of course, kernel engineering is a rich field in its own right. Of central interest is the design of specialized kernels suited to particular data structures ([Barla et al., 2002, Chapelle et al., 1999,

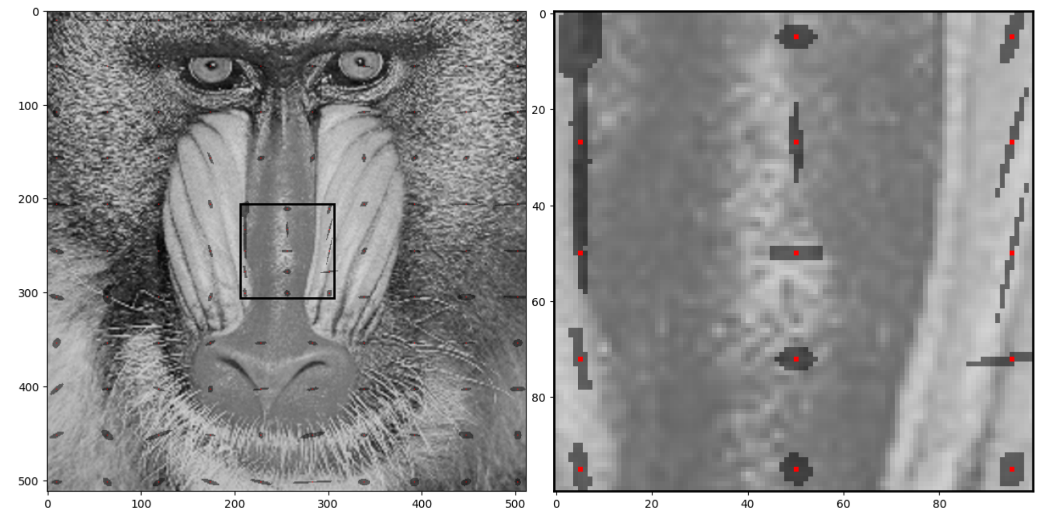

Figure 1: Localizations from AGOP Descent (Algorithm 1) centered at each of the highlighted points (in red) trained on a grayscale image of a mandrill Bush [2021]. Here X is the location and Y the grayscale value of the image. For visualization purposes the flow was stopped early to enforce neighborhoods of 75 pixels at each point. On the right the image is magnified to the highlighted region boxed-off on the left.

Joachims, 1998, Kondor and Jebara, 2003, Odone et al., 2005, Vishwanathan et al., 2010], etc.), statistical principles ([Genton, 2001, Osborne, 2010, Schölkopf et al., 1997], etc.), and problems of interest (Gong et al. [2024], Kokot and Luedtke [2025], etc.). In regression settings, the principle of local feature learning, in which kernels are augmented by differential information at points of interest ([Lowe, 1999, Schmid and Mohr, 1997, Wallraven et al., 2003]), emerged. Earlier nonparametric methods developed a similar framework, with datasets being recursively partitioned to improve the quality of local fits [Breiman and Meisel, 1976, Friedman, 1979, Heise, 1971]. Our method bears particular resemblance to "kernel steering" developed in the image processing literature [Takeda et al., 2007] (see also the follow-up paper [Takeda et al., 2008]).

The goal of this method is to allow the kernel size and shape to change in a data-dependent way, adapting not only to sample location and density, but also to local features in the data. A special case of these data-adaptive kernels is the popular bilateral filter in computer vision [Tomasi and Manduchi, 1998], [Elad, 2002]. In Takeda et al. [2007], they propose *kernel steering*, an iterative procedure that estimates gradients about a point of interest. These are then used to "steer" the kernel locally, adopting the empirical covariance of the gradients $\hat{\Diamond}$ as a Mahalanobis distance for subsequent estimation. Applying this method to the Gaussian kernel gives the steering kernel

$$K_{h,\hat{\Diamond}}(x_i - x^*) = \frac{\sqrt{\det(\hat{\Diamond})}}{2\pi^2 h^2 \mu^2} \exp\left\{-\frac{(x_i - x^*)^T \hat{\Diamond}(x_i - x^*)}{2h^2\mu^2}\right\},$$

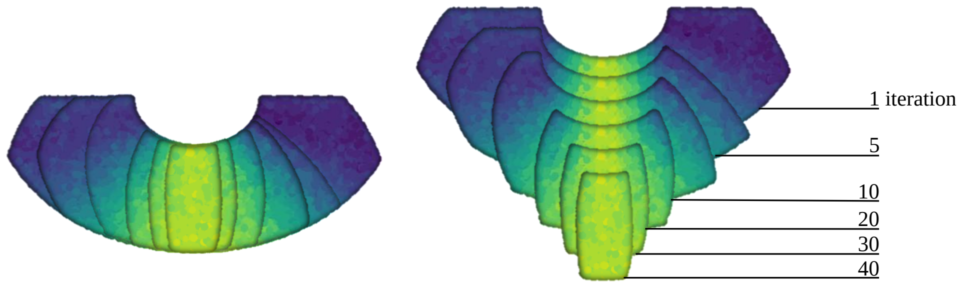

Figure 2: AGOP Descent performed on data from a noisy circle in dimension $D = 2$. The overlaid heatmap represents the values of the function of interest f that varies with the angle, up to additive Gaussian noise. Left: domain of X and function values $f(X)$. Right: neighborhood localization by AGOP Descent iteration with convergence toward the central point.

which can then be used to estimate the function again, and further refine the estimate of $\hat{\Diamond}$. This directional adaptation enables more effective denoising and image recovery, and is closely related to the expected gradient outer product (EGOP) $\Diamond := \mathbb{E}[\nabla f(X) \nabla f(X)^T]$ [Hristache et al., 2001, Samarov, 1993, Trivedi et al., 2014, Xia et al., 2002, Yuan et al., 2023].

Recently, a similar procedure has been proposed in Radhakrishnan et al. [2022, 2025], which further emphasizes the importance of the empirical covariance matrix of the estimated gradients $\hat{\Diamond}$, also called the Average Gradient Outerproduct (AGOP). In that paper, both empirical and theoretical results closely relate this object to the performance of simple neural networks, and it was shown that their method can adapt to the global regularity of the function of interest [Radhakrishnan et al., 2025].

However, $\Diamond$ may generically be full rank, making procedures such as the above subject to the curse of dimensionality. The present work focuses on exactly such a setting. We say that tuples (X, Y) satisfy the *supervised noisy manifold hypothesis* if they are such that $X \sim M + E_M$ where M is sampled from a d-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$, E_M is orthogonal to $\mathcal{M}$ at M, and $f(X) = f(M)$ for $f(x) := \mathbb{E}[Y|X = x]$. That is, the regression target does not depend on the noise, $f(x) = f \circ \pi(x)$, for π the projection onto the manifold. For this to be well-defined, we additionally assume that E_M lies within the reach of $\mathcal{M}$ Federer [1959] almost surely. Thus, in this example, $\Diamond$ is locally of low-dimension, however globally it may not degenerate. The structure of the features X is typical in manifold learning (Aamari and Levrard [2018], Genovese et al. [2012], etc.).

When learning such a label of interest $f(x^*)$, one would like to pool data orthogonally to the manifold, leading to a vast reduction in variance compared to typical isotropic kernel methods. In this ellipsoidal region, $\Diamond$ is approximately of low rank with principal eigencomponents tangential to $\mathcal{M}$. In this work, we bridge local [Takeda et al., 2007] and global [Radhakrishnan et al., 2022] iterative AGOP schemes by developing a method that automatically forms such a localization,

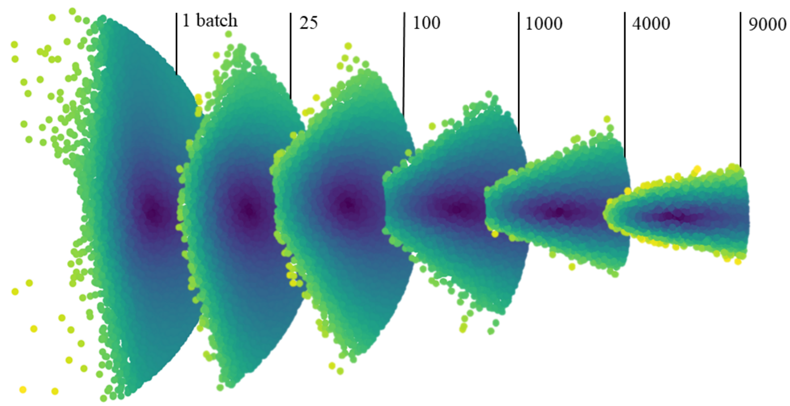

Figure 3: Localizations produced by a deep neural network while training on 10^5 points from the noisy 1-sphere. Each localization is generated at a different phase of training indicated by the batch number. Distances relative to the central point are computed in the learned embedding space, and we sample 10^5 points with replacement using weights $w_i \propto \exp(-\|x - x_i\|^2_{\text{Embed}}/8)$. Overlaid is a heatmap of the distances in the embedding space.

without a priori knowledge of the manifold or the underlying regularity of the function. The resulting estimator achieves adaptability reminiscent of deep neural networks [Cloninger and Klock, 2021], going beyond the multi-index setting.

As a first indicator of our method, observe that

$$v^T \lozenge v = \mathbb{E}[(\nabla f(X)^T v)^2] = \mathbb{E}[\partial_v f(X)^2]$$

where $\partial_v f$ denotes the directional derivative of f along v. Hence, the quadratic form generated by the EGOP is directly related to the directions of maximal variation of our function. This indicates that it is desirable to shift the features X proportionately to this operator, which we make precise via the theory of Wasserstein flows as described in the following section. This leads us to a discretized algorithm we call AGOP Descent, which we study by its continuum counterpart, the "EGOP flow". We prove that under the supervised noisy manifold hypothesis, the resulting regressor has intrinsic dimensional learning rates, regardless of the ambient dimension.

We support this theoretical result with the following numerical examples. First, expanding on the noisy manifold setting, we show how our regression error rates remain invariant to the injection of high-dimensional noise. Then, we compare our method to the performance of a deep neural network on toy data. In particular, we show how the feature embeddings produced by transformers trained on a simple example is qualitatively similar to the localizations generated by our procedure. In contrast, in Section 5.3 we demonstrate that two-layer neural networks are not able to efficiently learn in the noisy manifold setting, with a sharp decrease in performance compared to AGOP descent. Finally, we apply this procedure to estimate backbone angles in Molecular Dynamics (MD) data, where we leverage the noisy manifold structured data to improve prediction quality.

2 An Isotropic Vignette

As a starting point, we reframe the classic Nadaraya-Watson [Nadaraya, 1964, Watson, 1964] kernel regression algorithm through the viewpoint of Wasserstein flows. A rigorous treatment of Wasserstein flows can be found in Ambrosio et al. [2006]; in this section, we provide some intuition. An illuminating viewpoint is that of a particle system evolving overtime. Let $x_t(x_0)$ denote the location at time t of a particle with initial position x_0. The particle system follows the velocity field v_t if the instantaneous velocity $\dot{x}_t(x_0)$ of a particle with initial position x_0 equals $v_t(x_t(x_0))$ for all x_0. The law μ_t of the distribution of the particles at time t is exactly the pushforward[1] $\mu_t = (x_t)_\sharp \mu_0$ of the law μ_0 of initial positions of the particles. It is known (see Ambrosio et al. [2006]) that μ_t satisfies the continuity equation $\partial_t \mu_t = -\nabla(v_t \mu_t)$, with the gradient being defined in the weak sense

$$\partial_t \int \phi \, d\mu_t = \int \langle \nabla \phi, v_t \rangle \, d\mu_t.$$

For a functional on 2-Wasserstein space, $\mathcal{F} : \mathcal{P}_2(\mathcal{X}) \to \mathbb{R}$, the analogous formula

$$\partial_t \mathcal{F}(\mu_t) = \int \langle \nabla_W \mathcal{F}(\mu_t), v_t \rangle \, d\mu_t,$$

holds, where $\nabla_W \mathcal{F}(\mu_t)$ can be explicitly derived as the gradient of the first-variation (closely connected to other first order notions such as Fréchet and Hadamard derivatives) of $\mathcal{F}$ at μ_t.

We now relate Nadaraya-Watson to the isotropic flow with velocity field $v_t(x) = -(x - x^*)$. For convenience, we will translate our function so that $x^* = 0$ without loss of generality. The ODE $\dot{x}_t = -x$ has explicit solution $x_t = \exp(-t)x_0$, which greatly simplifies our analysis. The first object of interest that we will highlight is the "covariance" matrix $\Sigma_t = \int xx^T d\mu_t$. This operator provides a simple encoding of the extent to which μ_t is warped in each direction. Further, its determinant is related to the extent to which we have localized or contracted the measure. Explicit computation reveals $\det(\Sigma_t) = \exp(-2Dt)\det(\Sigma_0)$, but one can alternately derive, with the Wasserstein formalism, the ODE $\partial_t \det(\Sigma_t) = -2\,\mathrm{tr}(I)\det(\Sigma_t)$, or equivalently $\partial_t \log\det(\Sigma_t) = -2D$, which allows for the same explicit solution.

We now introduce the smoothing operators

$$\mathcal{P}_{\Sigma_t} f(x) = C(x)^{-1} \int k(\Sigma_t^{-1/2}[x-y]) f(y) \, d\mu,$$

$$\hat{\mathcal{P}}_{\Sigma_t} f(x) = \hat{C}(x)^{-1} \frac{1}{n} \sum_{i=1}^{n} k(\Sigma_t^{-1/2}[x - X_i]) f(Y_i)$$

where $C(x), \hat{C}(x)$ are the normalizing constants given so that the above convolutions are constant on constant functions, and k is a kernel function (see [Tsybakov, 2009] for specific properties k satisfies). This formula appears strange at first, but for this simple flow, one can directly compute

$$\mathcal{P}_{\Sigma_t} f = \int k(\Sigma_0^{-1/2}[x-y]/\exp(-t)) f(y) \, d\mu = \int k([x-y]/h) f(y) \, d\mu$$

setting initial covariance to I and identifying $\exp(-2t)$ with the bandwidth parameter h. Hence, this is nothing more than Nadaraya-Watson. We can interpret the quadratic form $v^T(\Sigma_t)^{-1}v$ as a

[1] To preserve the flow of the presentation, some definitions are omitted; the reader is invited to consult Ambrosio et al. [2006] for details.

natural normalization of the vector v by the relative proportion of variation the distribution exhibits in its direction, observing $\int x^T \Sigma_t^{-1} x d\mu_t = \mathrm{tr}(\Sigma_t \Sigma_t^{-1}) = D$. This metric has appeared in diverse contexts from local features [Schmid and Mohr, 1997] to VAEs [Chadebec and Allassonnière, 2022]. To study the quality of this smoothing estimator, we use the usual bias-variance trade-off to get

$$\mathbb{E}\left[\int (f - \hat{\mathcal{P}}_{\Sigma_t} f)^2 \, d\mu_t\right] \leq \int (f - \mathcal{P}_{\Sigma_t} f)^2 \, d\mu_t + \mathbb{E}\left[\int (\mathcal{P}_{\Sigma_t} f - \hat{\mathcal{P}}_{\Sigma_t} f)^2 \, d\mu_t\right].$$

The first term can be bounded above by a Poincaré inequality. Indeed, we can locally interpret the above convolution as a diffusion operator on μ_t under the metric Σ_t^{-1}. Below we define the EGOP functional.

Definition 1 (EGOP functional)**.** *We define the EGOP functional to be*

$$W(\mu_t) = \int \nabla f^T \Sigma_t \nabla f \, d\mu_t$$

Then, we have the following control on the "squared bias".

Proposition 1 (Bias)**.** $\int (f - \mathcal{P}_{\Sigma_t} f)^2 \, d\mu_t = O\left(W(\mu_t)\right).$

The EGOP functional, $W(\mu_t) = \int x^T \nabla f(y) \nabla f(y)^T x \, d\mu_t(x) d\mu_t(y) = \int x^T \lozenge_t x \, d\mu_t = \mathbb{E}_{\mu_t \otimes \mu_t}[\partial_X f(Y)^2]$, is the integral of the EGOP form $x^T \lozenge_t x$. One can show that

$$\partial_t W(\mu_t) \approx -2W(\mu_t) \Longrightarrow \partial_t \log W(\mu_t) \approx -2,$$

or more directly, we can simply observe

$$W(\mu_t) = \exp(-2t) \int \nabla f^T \Sigma_0 \nabla f \, d\mu_t \propto \exp(-2t) \Longrightarrow \partial_t \log W(\mu_t) \approx -2,$$

assuming our initial covariance is non-degenerate. In the following, we have the "variance" term of the bias-variance trade-off.

Proposition 2 (Isotropic Variance)**.**

$$\mathbb{E}\left[\int (\mathcal{P}_{\Sigma_t} f - \hat{\mathcal{P}}_{\Sigma_t} f)^2 d\mu_t\right] = O\left(1/[n \det(\Sigma_t^{1/2})]\right) = O\left(1/[n \exp(-Dt)]\right).$$

Hence, setting $h := W(\mu_t)$ results in the usual curse of dimensionality.

Proposition 3 (Isotropic Bias-Variance Trade-off)**.** $\mathbb{E}[\int (f - \hat{\mathcal{P}}_{\Sigma_t} f)^2 d\mu_t] \leq O\left(h + h^{-D/2}/n\right).$

Drawing our attention back to the EGOP functional, we see that, as a Dirichlet form, it expresses the smoothness of f in the domain μ_t under the metric Σ_t^{-1}. This leads to our approach, where we localize by following the gradient of $W(\mu)$, which we estimate approximately via an appropriate rescaling of the velocity field $\dot{x}_t = -\lozenge_t x_t$. In this sense, we are shifting our distribution to optimize the smoothness, or rather, minimize the variation of, f.

We note that the isotropic flow is not entirely arbitrary, as it is the gradient flow with respect to the functional $\int \|x - x^*\|^2 d\mu = W_2^2(\mu, \delta_{x^*})$. In this sense, it is exactly the localization procedure best suited for the function $f = \|x\|^2$.

3 Learning a Localized Kernel

In this section, we identify an appropriate velocity field for data adaptive local linear regression. In particular, we seek a flow that rapidly decreases $x^T \lozenge_t x = \mathbb{E}_{\mu_t}[(x^T \nabla f)^2]$, and thus the directions of maximal local variation relative to our target function. We desire μ_t to approximately correspond to sub-level sets of the form $\{x : x^T \lozenge_t x \leq C \exp(-t)\}$. As indicated in Section 2, in the continuum this corresponds to a flow in the direction of $-\lozenge_t x$, which we relate to a practical iterative scheme for real data.

3.1 The EGOP Flow

We define the EGOP flow to be the continuous time flow with velocity field

$$v_t(x_t) = -\frac{{x_t}^T \lozenge_t x_t}{x_t^T \lozenge_t^2 x_t} \lozenge_t x_t.$$

To motivate this choice, from Section 2, we see that it is desirable to minimize the EGOP functional $W(\mu_t) = \int x_t^T \lozenge_t x_t \, d\mu_t$. This functional is given by point-wise integration of the EGOP form $F_t(x) = x^T \lozenge_t x$. Hence, to most sharply minimize this function, we seek to follow the velocity field induced by its gradient, $\nabla F_t(x) \propto \lozenge_t x$. The velocity field $v_t(x_t) = -\frac{x_t^T \lozenge_t x_t}{x_t^T \lozenge_t^2 x_t} \lozenge_t x_t = -\frac{\nabla F_t(x_t)}{\|\nabla F_t(x_t)\|^2} F_t(x_t)$ is precisely the rescaling that allows for a proportionate rate of decay $\partial_s F_t(x_s)|_{s=t} = -2F_t(x_t)$.

3.2 The AGOP Descent Algorithm

The typical approach to discretizing flows in optimization algorithms such as gradient descent involves updating the point considered at a given step by moving it incrementally along the prescribed velocity field. Mathematically, this is described by defining $x_{t+1} = x_t - \alpha_t v_t(x_t)$. In our setting of empirically observed data, however, we cannot actually shift the data points. Instead, by adjusting the region of localization, we mimic this process, contracting the domain to match the image of the previous gradient steps.

Our discretized algorithm is structured similarly to Radhakrishnan et al. [2022], as we iteratively estimate the function of interest f and the corresponding AGOP matrix with a Mahananoblis-metrized kernel regressor. Our key innovation is to exclude, at each iteration, data points for which $x^T \hat{\lozenge}_i x$ is particularly large, where $\hat{\lozenge}_i$ is the estimated AGOP at the ith iterate. As a notable difference from this previous work, we use the inverse covariance matrix rather than the AGOP matrix as a local metric, and in Appendix A we compare their asymptotic behavior. The full algorithm is below.

3.3 Setting the tuning parameters Samp, initial neighborhood size, and α

Initial neighborhood size Theoretically, the algorithm starts with the entire sample, but in practice it should start from a spherical neighborhood large enough for estimating the local covariance matrix, i.e. the initial M. Practically, this can be done using regular kernel regression for f, and choosing the kernel width h by Cross-Validation (CV); then, the neighborhood radius should be $\approx 3h^{CV}$.

Algorithm 1 AGOP Descent

Initialize data (X, Y), $n \leftarrow |X|$, select a basis such that $\Sigma_0 = I$, k kernel function
Set $x_0 \leftarrow x^*$
Set target sample size Samp
Fix $\alpha \in (0, 1)$ Removal proportion
MISE$\leftarrow 0$
while $n >$ Samp **do**
  $n \leftarrow |X|$
  $M \leftarrow \frac{1}{n} \sum_{x_i \in X} (x_i - x^*)(x_i - x^*)^T$
  for i, j in $1 : n \times 1 : n$ **do**
    $W_{ij} \leftarrow k(M^{-1/2}(x_i - x_j))$
  end for
  $W_{ii} \leftarrow 0$ for $i = 1 : n$
  $W \leftarrow$ RowNormalize(W) % $\sum_j W[i, j] = 1$, for $i = 1 : n$
  prediction $\leftarrow$ zeros(n)
  gradients $\leftarrow$ zeros(n, D)
  for i in $1 : n$ **do**
    $L_i, c_i \leftarrow$ LocalLinearRegression$(Y, X - X[i, :], W[i, :])$ % Linear fit and intercept
    prediction$[i] \leftarrow c_i$
    gradients$[i, :] \leftarrow L_i$
  end for
  $\lozenge \leftarrow$ gradients^T gradients $/n$
  $m \leftarrow (1 - \alpha)$-quantile$([x_i - x^*]^T \lozenge [x_i - x^*])$
  Remove x_i from X if $[x_i - x^*]^T \lozenge [x_i - x^*] > m$
end while
$W_{0j} \leftarrow k(M^{-1/2}(x_0 - x_j))$ for $j = 1 : n$
$W[0, :] \leftarrow$ RowNormalize$(W[0, :])$
$L_0, c_0 \leftarrow$ LocalLinearRegression$(Y, X, W[0, :])$
return c_0

Samp This is the stopping parameter of the `while` loop in Algorithm 1. In our implementation, we choose it by CV.

α The parameter α controls the time discretization of the EGOP flow. We note that Samp $\approx n(1 - \alpha)^{\#iterations}$; hence, we can choose a #iterations sufficiently large, then set $\alpha = 1 - \exp\left(\frac{\ln n/\text{Samp}}{\#\text{iterations}}\right)$. In our experiments, we found that the algorithm results are not sensitive to α and set $\alpha = 0.2$ unless otherwise mentioned.

4 Convergence rate analysis under EGOP kernel regression

In this section, we verify fundamental results for localizations generated by flows. All proofs are in Appendix B.

Key to our analysis will be the following generic result on localizations induced by flows.

Lemma 1. *Let μ_t be a flow satisfying $\partial_t \mu_t = -\nabla(v_t \mu_t)$, and $\mathcal{P}_{\Sigma_t}$ the normalized k-convolutional* *operator on μ_t in Mahalanobis-metric Σ_t^{-1}, $\hat{\mathcal{P}}_{\Sigma_t}$ its empirical version and*

$$c_v := \lim_{t\to\infty} \log\det(\Sigma_t)/\log W(\mu_t).$$

Then, for t_n such that $W(t_n) = O(n^{-1/(1+c_v/2)})$,

$$\mathbb{E}\left[\int (f - \hat{\mathcal{P}}_{\Sigma_{t_n}} f)^2 \, d\mu_t\right] = O(n^{-1/(1+c_v/2)})$$

By L'Hôpital's rule,

$$\lim_{t\to\infty} \log\det(\Sigma_t)/\log W(\mu_t) = \lim_{t\to\infty} \partial_t \log\det(\Sigma_t)/\partial_t \log W(\mu_t),$$

hence it suffices to consider these equations at first order.

Lemma 2. *Let μ_t be a flow satisfying $\partial_t \mu_t = -\nabla(v_t \mu_t)$. Then,*

$$\begin{aligned} \partial_t \log\det(\Sigma_t) &= -2\int \langle \Sigma_t^{-1} x, v_t(x)\rangle \, d\mu_t \\ \partial_t W(\mu_t) &= -2\int \langle \lozenge_t x, v_t(x)\rangle \, d\mu_t - 2\int \langle \nabla^2 f(x)\Sigma_t \nabla f(x), v_t(x)\rangle \, d\mu_t \end{aligned}$$

We decompose $\partial_t W(\mu_t)$ into the *contraction* and *twist* components

$$\mathcal{C}_t := \int \langle \lozenge_t x, v_t(x)\rangle \, d\mu_t, \quad \mathcal{T}_t := \int \langle \nabla^2 f(x)\Sigma_t \nabla f(x), v_t(x)\rangle \, d\mu_t,$$

with the contractive component corresponding to the decrease in the AGOP form $F_t(x_s)$ with t fixed and s varying, and the twist being induced by the shift of the measure μ_t that perturbs the matrix $\lozenge_t$.

Lemma 3. *Letting μ_t denote the EGOP flow,*

$$C_t = W(\mu_t), \quad \lim_{t\to\infty} \mathcal{T}_t > 0.$$

In our setting of interest, we can additionally bound the log-volume.

Lemma 4. *Let (X, Y) satisfy the noisy manifold assumption in (d, D), and μ_t denote the EGOP flow. Then,*

$$\lim_{t\to\infty} \partial_t \log\det \Sigma_t \geq -2d.$$

We use these results to verify intrinsic dimensional learning for arbitrary high-dimensional noise in the noisy manifold setting.

Theorem 1. *Let (X, Y) satisfy the noisy manifold assumption in (d, D), and μ_t denote the EGOP flow. Then $c_v \leq d$, in particular*

$$\mathbb{E}\left[\int (f - \hat{\mathcal{P}}_{\Sigma_{t_n}} f)^2 \, d\mu_t\right] = O(n^{-1/(1+d/2)}).$$

In other words, the kernel regressor $\hat{\mathcal{P}}_{\Sigma_{t_n}} f$ converges to the target f at a rate that depends only on the intrinsic dimension d and not on the noise or ambient dimension.

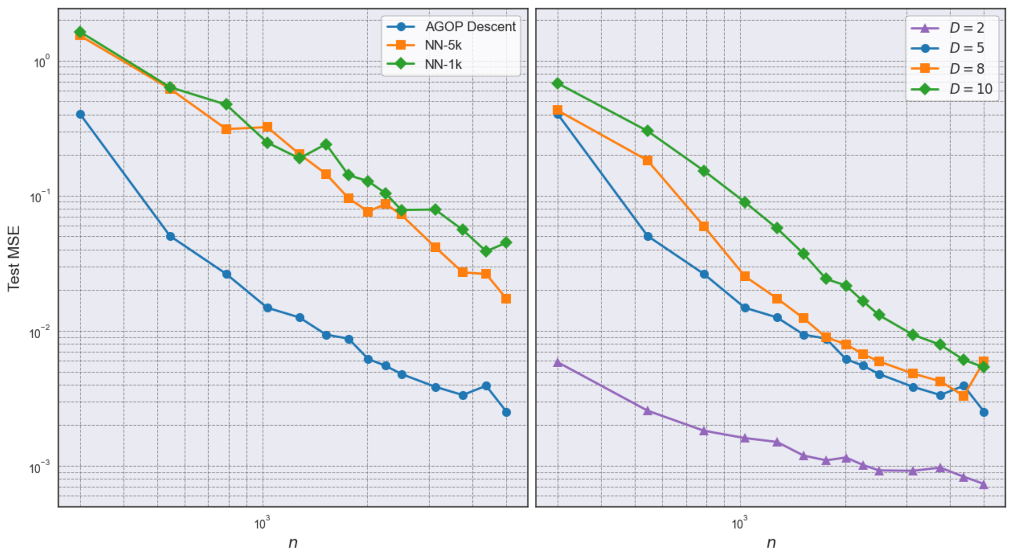

Figure 4: (Left) Comparison of AGOP Descent to performance of two-layer neural network architectures trained on helical data in ambient dimension $D = 5$. (Right) AGOP Descent trained on helical data in various ambient dimensions D.

5 Simulations

In our simulations, we consider helical data, parameterized by a curve $\theta(t) = (\sin(t + w_1), \cos(t + w_1), \sin(t+w_2), \cos(t+w_2), \ldots, g(t))$, where $g(t) = t$ is a linear term included if D is odd dimensional, and the w_i are constant offsets taken as a mesh from 0 to 2π. We rescale this data by a constant τ, then contaminate it with uniform, orthogonal noise of radius r. In our simulations, we set $\tau = 0.8$, $r = 0.5$, and sample t from 0 to 2π. See Figure 7 for a visualization. A benefit of this choice of curve is that the tangent directions are diverse and the curvature is stable. For the outcomes y, we generate a 3rd degree polynomial with coefficients uniformly sampled from $(-3, 3)$, then evaluate it at the projection point onto $\theta(t)$. See Appendix D for the precise implementation of AGOP Descent used in these examples.

5.1 Learning Rate

We generate helical data in a variety of dimensions, testing the AGOP Descent algorithm. As seen in Figure 4, the learning rate is constant, although it is affected by dimensional constants. The estimates for the MSE were computed over 100 withheld test samples, repeated 1000 times for each training data size n.

5.2 Feature Learning

In this section we compare the local feature learning capabilities of a deep, transformer based neural network [Gorishniy et al., 2023] and AGOP descent. We consider data generated from a 1-sphere under the supervised noisy manifold hypothesis. Motivated by the recent work [Anonymous, 2025] where it was shown that low-dimensional spectral embeddings can imprecisely recover intrinsic structure in the noisy manifold setting, we demonstrate that AGOP Descent allows for this gap to be bridged. In particular, on this simple dataset, the features are nearly completely denoised, achieving a localization of comparable quality to the transformer embedding, as shown in Figures 2 and 3. In Appendix D, we provide additional unsupervised embeddings for comparison.

5.3 Two-layer Neural Network

We assess the performance of two-layer neural networks in the continuous index setting, see Appendix D for architecture details. That low intrinsic dimensionality of datasets can accelerate learning has been frequently observed in the machine learning literature (Kiani et al. [2024], Liu et al. [2021], etc.). We show that these guarantees are diminished for learning f with low local intrinsic dimension (continuous single-index), even when the features follow an approximate manifold structure. This is demonstrated in Figure 4. Further, the far lower test MSE achieved by AGOP Descent illustrates the significant suboptomality of these algorithms.

5.4 Predicting the backbone angles in Molecular Dynamics (MD) data

This example comes from the analysis of molecular geometries. Raw data consist of X, Y, Z coordinates for each of the N_a atoms of a molecule, which, due to interatomic interactions, lie near a low-dimensional manifold [Das et al., 2006]. While the governing equations of the simulated dynamics are unknown, for small organic molecules, it has been observed that certain backbone angles [Das et al., 2006] vary along the aforementioned low-dimensional manifold. Specifically, for the malonaldehyde molecule, the two backbone angles denoted $\tau_{1,2}$ are shown in Figure 8. We used a subsample of molecular configurations of size $n = 10^4$ from the MD simulation data of Chmiela et al. [2017] as input data.[2] The configuration data, pre-processed as in Koelle et al. [2022], consists of $D = 50$ dimensional vectors and lies near a 2-dimensional surface with a torus topology (see Figure 8). On a hold-out set of 500 test points, AGOP Descent yields an MSE of 0.0011, compared to 0.012 for Gaussian kernel Nadaraya-Watson with cross-validated bandwidth selection.

6 Discussion

Our work presents a localization scheme motivated by Radhakrishnan et al. [2022]. It is important to note that some of our results are *in population*, as we consider an estimator resulting from the continuous EGOP flow. In particular, we do not consider the rate at which one can learn the EGOP matrix itself given only empirical data, an essential next step to verify efficient estimation. In Appendix E, we go into further detail on this problem, and we give both theoretical and numerical evidence indicating that this can be overcome in a future analysis.

It is also of key interest to further develop the connection between our discretized flow and the EGOP velocity field. While these two procedures are comparable in their level-set descents, there are key differences in the resulting distributions of datapoints. In particular, as discussed in Appendix A, these procedures result in different densities on the prescribed regions, and thus the AGOPs will have different values. A simple solution is to impose an inverse propensity weighting on the discretized data to better match that of its continuous counterpart. However, it is unclear whether

[2] Made available at `montlake.github.io` along with the backbone angles $\tau_{1,2}$ for each sample.

this is computationally justified, as we see tremendous results for the uncorrected implementation presented in this work. In future research, we hope to further explore this disparity, either adjusting the EGOP flow to capture more information regarding this discretization, or to refine our estimation procedure for enhanced theoretical guarantees.

Though not explored in this work, of additional interest is an adaptation to more carefully extend the localization procedure induced by AGOP descent. In particular, we prove only adaptive learning for flat level sets of the function f, which can be achieved via ellipsoidal localizations. In Appendix D, we enhance our method with a diffusion maps-style affinity matrix [Belkin and Niyogi, 2003, Coifman and Lafon, 2006], and show that this allows for curvature in the localization.

References

Eddie Aamari and Clément Levrard. Non-asymptotic rates for manifold, tangent space, and curvature estimation, 2018. URL `https://arxiv.org/abs/1705.00989`.

Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks, 2024. URL `https://arxiv.org/abs/2202.08658`.

L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2006. ISBN 9783764373092. URL `https://books.google.com/books?id=Hk_wNp0sc4gC`.

Anonymous. Xxx. In M. Fazel, D. Hsu, S. Lacoste-Julien, and V. Smith, editors, *Proceedings of the International Conference on Machine Learning (ICML)*, page (to appear). PMLR, 2025.

Annalisa Barla, Francesca Odone, and Alessandro Verri. Hausdorff kernel for 3d object acquisition and detection. In *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part IV 7*, pages 20–33. Springer, 2002.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2018.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in neural information processing systems*, 35:9768–9783, 2022.

Enric Boix-Adsera, Etai Littwin, Emmanuel Abbe, Samy Bengio, and Joshua Susskind. Transformers learn through gradual rank increase. *Advances in Neural Information Processing Systems*, 36: 24519–24551, 2023.

Leo Breiman and William S Meisel. General estimates of the intrinsic variability of data in nonlinear regression models. *Journal of the American Statistical Association*, 71(354):301–307, 1976.

Jeffrey Bush. C4l image dataset, 2021. URL `https://dx.doi.org/10.21227/bc9m-f507`.

Clément Chadebec and Stéphanie Allassonnière. A geometric perspective on variational autoencoders, 2022. URL `https://arxiv.org/abs/2209.07370`.

Olivier Chapelle, Patrick Haffner, and Vladimir N Vapnik. Support vector machines for histogram-based image classification. *IEEE transactions on Neural Networks*, 10(5):1055–1064, 1999.

Stefan Chmiela, Alexandre Tkatchenko, Huziel Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, March 2017.

Alexander Cloninger and Timo Klock. A deep network construction that adapts to intrinsic dimensionality beyond the domain, 2021. URL `https://arxiv.org/abs/2008.02545`.

Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

Tiangang Cui, Xin Tong, and Olivier Zahm. Optimal riemannian metric for poincar {\'e} inequalities and how to ideally precondition langevin dymanics. *arXiv preprint arXiv:2404.02554*, 2024.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems*, 36:752–784, 2023.

P. Das, M. Moll, H. Stamati, L.E. Kavraki, and C. Clementi. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proceedings of the National Academy of Sciences*, 103(26):9885–9890, 2006.

Michael Elad. On the origin of the bilateral filter and ways to improve it. *IEEE Transactions on image processing*, 11(10):1141–1151, 2002.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959.

Jerome H Friedman. A tree-structured approach to nonparametric multiple regression. In *Smoothing Techniques for Curve Estimation: Proceedings of a Workshop held in Heidelberg, April 2–4, 1979*, pages 5–22. Springer, 1979.

Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry A. Wasserman. Minimax manifold estimation. *Journal of Machine Learning Research*, 13:1263–1291, 2012. URL `http://dl.acm.org/citation.cfm?id=2343687`.

Marc G Genton. Classes of kernels for machine learning: a statistics perspective. *Journal of machine learning research*, 2(Dec):299–312, 2001.

Albert Gong, Kyuseong Choi, and Raaz Dwivedi. Supervised kernel thinning. *arXiv preprint arXiv:2410.13749*, 2024.

Yury Gorishniy, Ivan Rubachev, and Artem Babenko. On embeddings for numerical features in tabular deep learning, 2023. URL `https://arxiv.org/abs/2203.05556`.

David R Heise. Multivariate model building: The validation of a search strategy., 1971.

Marian Hristache, Anatoli Juditsky, Jörg Polzehl, and Vladimir Spokoiny. Structure adaptive approach for dimension reduction. *Annals of Statistics*, pages 1537–1566, 2001.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

Bobak Kiani, Jason Wang, and Melanie Weber. Hardness of learning neural networks under the manifold hypothesis. *Advances in Neural Information Processing Systems*, 37:5661–5696, 2024.

Samson Koelle, Hanyu Zhang, Marina Meilă, and Yu-Chia Chen. Manifold coordinates with physical meaning. *Journal of Machine Learning Research*, 23, 2022.

Alex Kokot and Alex Luedtke. Coreset selection for the sinkhorn divergence and generic smooth divergences. *arXiv preprint arXiv:2504.20194*, 2025.

Risi Kondor and Tony Jebara. A kernel between sets of vectors. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 361–368, 2003.

Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *Advances in Neural Information Processing Systems*, 37:58716–58756, 2024.

Hao Liu, Minshuo Chen, Tuo Zhao, and Wenjing Liao. Besov function approximation and binary classification on low-dimensional manifolds using convolutional residual networks. In *International Conference on Machine Learning*, pages 6770–6780. PMLR, 2021.

David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with sgd. *arXiv preprint arXiv:2209.14863*, 2022.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

Francesca Odone, Annalisa Barla, and Alessandro Verri. Building kernels from binary strings for image matching. *IEEE Transactions on Image Processing*, 14(2):169–180, 2005.

Michael Osborne. *Bayesian Gaussian processes for sequential prediction, optimisation and quadrature.* PhD thesis, Oxford University, UK, 2010.

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism of feature learning in deep fully connected networks and kernel machines that recursively learn features. *arXiv preprint arXiv:2212.13881*, 2022.

Adityanarayanan Radhakrishnan, Mikhail Belkin, and Dmitriy Drusvyatskiy. Linear recursive feature machines provably recover low-rank matrices. *Proceedings of the National Academy of Sciences*, 122(13):e2411325122, 2025.

BLS Prakasa Rao. *Nonparametric functional estimation*. Academic press, 2014.

Alexander M Samarov. Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association*, 88(423):836–847, 1993.

Cordelia Schmid and Roger Mohr. Local grayvalue invariants for image retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 19(5):530–535, 1997.

Bernhard Schölkopf, Patrice Simard, Alex Smola, and Vladimir Vapnik. Prior knowledge in support vector kernels. *Advances in neural information processing systems*, 10, 1997.

Erwan Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500, 2016.

Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. Kernel regression for image processing and reconstruction. *IEEE Transactions on image processing*, 16(2):349–366, 2007.

Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. Deblurring using regularized locally adaptive kernel regression. *IEEE transactions on image processing*, 17(4):550–563, 2008.

Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 839–846. IEEE, 1998.

Shubhendu Trivedi, Jialei Wang, Samory Kpotufe, and Gregory Shakhnarovich. A consistent estimator of the expected gradient outerproduct. In *UAI*, pages 819–828, 2014.

Alexandre B. Tsybakov. *Nonparametric estimators*, pages 1–76. Springer New York, New York, NY, 2009. ISBN 978-0-387-79052-7. doi: 10.1007/978-0-387-79052-7_1. URL https://doi.org/10.1007/978-0-387-79052-7_1.

S Vichy N Vishwanathan, Nicol N Schraudolph, Risi Kondor, and Karsten M Borgwardt. Graph kernels. *The Journal of Machine Learning Research*, 11:1201–1242, 2010.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Wallraven, Caputo, and Graf. Recognition with local features: the kernel recipe. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 257–264. IEEE, 2003.

Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

Yingcun Xia, Howell Tong, Wai Keung Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3): 363–410, 2002.

Gan Yuan, Mingyue Xu, Samory Kpotufe, and Daniel Hsu. Efficient estimation of the central mean subspace via smoothed gradient outer products. *arXiv preprint arXiv:2312.15469*, 2023.