Linear and nonlinear gradient-based dimension reduction

RT-UQ PhD day 2025, Grenoble

Romain Verdière

PhD supervisors: Clémentine Prieur, Olivier Zahm

AIRSEA Team
INRIA Grenoble - LJK

22/04/2025

Model approximation

Notations

Let u be a computational model defined on an open set $\mathcal{X} \subseteq \mathbb{R}^d$:

$$u: \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & u(x) \end{array}$$

where $d \gg 1$.

In many real case scenarios (scientific and engineering problems) u is:

- computationally expensive and slow to evaluate,
- ∇u can be evaluated for the same computational cost as an evaluation of u using the adjoint method (Plessix, 2006).

Model approximation

Goal

Given a set $\mathcal{S} = \{(x_k, u(x_k), \nabla u(x_k))\}_{1 \leq k \leq n_{\mathrm{train}}}$ and a tolerance ϵ, build an accurate and fast to evaluate approximation $\tilde{u}$ such that:

$$\mathbb{E}[(u(\boldsymbol{X}) - \tilde{u}(\boldsymbol{X}))^2] \leq \epsilon$$

with $\boldsymbol{X}$ a random vector.

Due to the **curse of dimensionality**, classical approximation methods requires n_{train} to grow **exponentially** with d.

Need for dimension reduction

Work around: Exploit low dimensional structures, if exists.

Problem formulation

Find a **feature map** $g : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}^m$ and a **profile function** $f : \mathbb{R}^m \to \mathbb{R}$ with $m \ll d$ such that

$$\mathbb{E}[(u(\boldsymbol{X}) - f \circ g(\boldsymbol{X}))^2] \leq \epsilon,$$

for some prescribed tolerance $\epsilon > 0$.

- Approximation class for f and g ?
- How to use ∇u to build g ?

How to use ∇u to learn g ?

$$u = f \circ g$$
$$\Downarrow$$
$$\nabla u(x) = \nabla g(x)^\top \nabla f(g(x))$$
$$\Downarrow$$
$$\nabla u(x) \in \mathrm{range}(\nabla g(x)^\top)$$
$$\Downarrow$$
$$\mathcal{J}_m(g) := \mathbb{E}[\|\nabla u(x) - \Pi_{\mathrm{range}(\nabla g(x)^\top)} \nabla u(x)\|^2] = 0$$

How to use ∇u to learn g ?

$$u = f \circ g$$
$$\Downarrow$$
$$\nabla u(x) = \nabla g(x)^\top \nabla f(g(x))$$
$$\Downarrow$$
$$\nabla u(x) \in \mathrm{range}(\nabla g(x)^\top)$$
$$\Downarrow$$
$$\mathcal{J}_m(g) := \mathbb{E}[\|\nabla u(x) - \Pi_{\mathrm{range}(\nabla g(x)^\top)} \nabla u(x)\|^2] = 0$$

- Is the reciprocal $\Uparrow$ true ?

How to use ∇u to learn g ?

$$u = f \circ g$$
$$\Downarrow$$
$$\nabla u(x) = \nabla g(x)^\top \nabla f(g(x))$$
$$\Downarrow$$
$$\nabla u(x) \in \mathrm{range}(\nabla g(x)^\top)$$
$$\Downarrow$$
$$\mathcal{J}_m(g) := \mathbb{E}[\|\nabla u(x) - \Pi_{\mathrm{range}(\nabla g(x)^\top)} \nabla u(x)\|^2] = 0$$

- Is the reciprocal $\Uparrow$ true ?
- Does $\mathcal{J}_m(g) \approx 0 \implies u \approx f \circ g$?

Poincaré inequality

- Is the reciprocal $\Uparrow$ true ? **Yes, if**:
  - either $g(x) = U_m^\top x$ with $U_m \in \mathbb{R}^{d\times m}$ a matrix with orthogonal columns (Zahm et al., 2019; Bigoni et al., 2022),
  - or, more generally, $g(x) = (\varphi_1(x), \dots, \varphi_m(x))$ where $\varphi(x) := (\varphi_1(x), \dots, \varphi_d(x))$ is a C^1 diffeomorphism in $\mathbb{R}^d$. (Verdière, Prieur, and Zahm, 2023)

Poincaré inequality

- Is the reciprocal $\Uparrow$ true ? **Yes, if**:
  - either $g(x) = U_m^\top x$ with $U_m \in \mathbb{R}^{d \times m}$ a matrix with orthogonal columns (Zahm et al., 2019; Bigoni et al., 2022),
  - or, more generally, $g(x) = (\varphi_1(x), \dots, \varphi_m(x))$ where $\varphi(x) := (\varphi_1(x), \dots, \varphi_d(x))$ is a C^1 diffeomorphism in $\mathbb{R}^d$. (Verdière, Prieur, and Zahm, 2023)
- Does $\mathcal{J}_m(g) \approx 0 \implies u \approx f \circ g$?

Poincaré inequality

- Is the reciprocal $\Uparrow$ true ? **Yes, if**:
  - either $g(x) = U_m^\top x$ with $U_m \in \mathbb{R}^{d\times m}$ a matrix with orthogonal columns (Zahm et al., 2019; Bigoni et al., 2022),
  - or, more generally, $g(x) = (\varphi_1(x), \ldots, \varphi_m(x))$ where $\varphi(x) := (\varphi_1(x), \ldots, \varphi_d(x))$ is a C^1 diffeomorphism in $\mathbb{R}^d$. (Verdière, Prieur, and Zahm, 2023)
- Does $\mathcal{J}_m(g) \approx 0 \implies u \approx f \circ g$?

Poincaré inequality (Bakry et al., 2008)

For $\boldsymbol{X}$ a continuous random variable in $\mathbb{R}^d$, the Poincaré constant $\mathbb{C}(\boldsymbol{X})$ is defined as the smallest constant such that:

$$\mathbb{E}[(h(\boldsymbol{X}) - \mathbb{E}(h(\boldsymbol{X})))^2] \leq \mathbb{C}(\boldsymbol{X})\mathbb{E}[\|\nabla h(\boldsymbol{X})\|_2^2],$$

holds for any continuously differentiable function $h : supp(\boldsymbol{X}) \to \mathbb{R}$. We say that $\boldsymbol{X}$ satisfies Poincaré inequality (6) if $\mathbb{C}(\boldsymbol{X}) < +\infty$.

In particular, if $\boldsymbol{X} \sim \mathcal{N}(0, I_d)$ then $\mathbb{C}(\boldsymbol{X}) = 1$.

Gradient based dimension reduction

For $\mathcal{G}_m$ a given class of functions for g we have:

Proposition

If $\mathbb{C}(\boldsymbol{X}|\mathcal{G}_m) := \sup_{g\in\mathcal{G}_m} \sup_{z_m\in\mathbb{R}^m} \mathbb{C}(\boldsymbol{X}|g(\boldsymbol{X}) = z_m) < +\infty$, the reconstruction error satisfies

$$\min_{f:\mathbb{R}^m\to\mathbb{R}} \mathbb{E}[(u(\boldsymbol{X}) - f\circ g(\boldsymbol{X}))^2] \leq \mathbb{C}(\boldsymbol{X}|\mathcal{G}_m)\underbrace{\mathbb{E}[\|\nabla u(x) - \Pi_{\mathrm{range}(\nabla g(x)^\top)}\nabla u(x)\|^2]}_{:=\mathcal{J}_m(g)}.$$

Linear case: Active subspace

- Linear case (Active Subspace) (Constantine, Dow, and Wang, 2014; Zahm et al., 2019)

$$\mathcal{G}_m = \left\{ g(x) = U_m^\top x \,\middle|\, U_m \in \mathbb{R}^{d\times m} \text{ with orthogonal columns} \right\},$$
$$\mathcal{J}_m(g) = \mathbb{E}[\|(I_d - U_m U_m^\top)\nabla u(\boldsymbol{X})\|_2^2],$$

moreover $\boldsymbol{X} \sim \mathcal{N}(0, I_d) \Rightarrow \mathbb{C}(\boldsymbol{X}|\mathcal{G}_m) = 1$.

Linear case (Active Subspace) (Constantine, Dow, and Wang, 2014; Zahm et al., 2019)

For $\boldsymbol{X} \sim \mathcal{N}(0, I_d)$ and for $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$ the eigenvalues of the **active subspace matrix**:

$$\boxed{H(u) := \mathbb{E}[\nabla u(\boldsymbol{X}) \nabla u(\boldsymbol{X})^\top] \in \mathbb{R}^{d\times d}},$$

we have:

$$\min_{g\in\mathcal{G}_m} \mathcal{J}_m(g) = \min_{U_m} \mathbb{E}[\|(I_d - U_m U_m^\top)\nabla u(\boldsymbol{X})\|_2^2] = \sum_{i=m}^{d} \lambda_i.$$

Dimension reduction strategy: For a given tolerance $\epsilon > 0$, choose m such that $\sum_{i=m}^{d} \lambda_i \leq \epsilon$. The feature map U_m is given by the m first eigenvectors of $H(u)$.

Active Subspace limitations

When does Active Subspace fails?

Active Subspace limitations

When does Active Subspace fails?

- Models with nonlinear low dimensional structures, for instance:

$$u(x) = \sin(\|x\|^2).$$

Isotropic function $\Rightarrow$ flat $H(u)$ spectrum.

Active Subspace limitations

When does Active Subspace fails?

- Models with nonlinear low dimensional structures, for instance:

$$u(x) = \sin(\|x\|^2).$$

Isotropic function $\Rightarrow$ flat $H(u)$ spectrum.

- Models with masked linear low dimensional structures due to high-frequency, low-amplitude components, for instance:

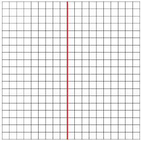
$$u(x_1, x_2) = \sin(x_1) + \frac{1}{6}\sin(10x_2).$$

$\Rightarrow$ Selection errors.

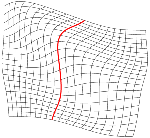
Nonlinear dimension reduction

Non linear dimension reduction

Linear dimension reduction



Non linear dimension reduction



In (Verdière, Prieur, and Zahm, 2023), we propose to build g as the solution to:

$$\min_{g \in \mathcal{G}_m(\mathbb{R}^d)} \mathcal{J}_m(g) := \mathbb{E}[\|\nabla u(x) - \Pi_{\mathrm{range}(\nabla g(x)^\top)} \nabla u(x)\|^2]$$

where

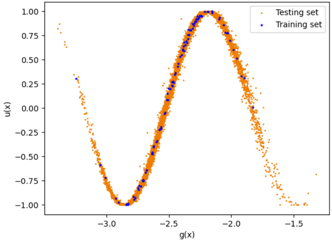
$$\mathcal{G}_m(\mathbb{R}^d) = \left\{ \begin{array}{cccc} g: & \mathbb{R}^d & \to & \mathbb{R}^m \\ & x & \mapsto & (\varphi_1(x), \dots, \varphi_m(x)) \end{array} \middle| \varphi \in \mathcal{D} \right\},$$

for $\mathcal{D}$ a set of $\mathcal{C}^1$-diffeomorphisms parametrized with an **invertible neural network**.

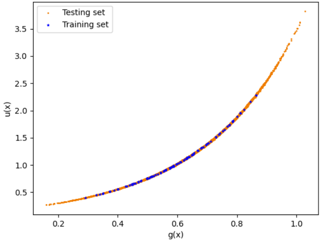
- $\mathcal{J}_m(g)$ is minimized with a gradient descent type algorithm (ADAM),
- $\nabla g(x)$ is computed with automatic differentiation (Griewank et al., 1989; Baydin et al., 2018).

Non linear dimension reduction examples

$$u_1(x) = \sin(\|\mathbf{x}\|^2), \qquad u_2(x) = \exp\left(\frac{1}{d}\sum_{i=1}^{d} \sin(x_i) e^{\cos(x_i)}\right)$$



(a) u_1 on $\Omega = [0, 1]^{20}$ with $n_{\mathtt{train}} = 100$



(b) u_2 on $\Omega = [-1, 1]^8$ with $n_{\mathtt{train}} = 100$

Figure: Scatter plot $\{(g(x^i), u(x^i))\}_{i \geq 1}$ for a random testing set of 10000 points and for $m = 1$

Non linear dimension reduction summary

- Advantages:
  - A nonlinear extension to Active Subspace
  - Outperforms Active Subspace
  - Allows to discover non linear low dimensional **active manifolds**
- Drawbacks:
  - No control on $\mathbb{C}(\boldsymbol{X}|\mathcal{G}_m)$ yet
  - No closed-form minimizer of $\mathcal{J}_m(g)$
  - More complex to use (need to train an invertible neural network)

More details in the following papers:

Romain Verdière, Clémentine Prieur, and Olivier Zahm (Dec. 2023). "Diffeomorphism-based feature learning using Poincaré inequalities on augmented input space". working paper or preprint. URL: https://hal.science/hal-04364208

Daniele Bigoni et al. (2022). "Nonlinear dimension reduction for surrogate modeling using gradient information". In: **Information and Inference: A Journal of the IMA** 11.4, pp. 1597–1639

Mollified Active Subspace

Linear dimension reduction: selection error on an example

Consider the **high frequency, low amplitude** component model:

$$u: \begin{array}{ccc} \mathbb{R}^2 & \to & \mathbb{R} \\ (x_1, x_2) & \mapsto & \sin(x_1) + \frac{1}{6}\sin(10x_2). \end{array}$$

where $\boldsymbol{X} \sim \mathcal{N}(0, I_2)$.

If we only consider coordinate selection:

Coordinate selected	Approximation error	AS upper-bound
$g(x) = x_1$	$\approx$ **0.14**	≈ 1.39
$g(x) = x_2$	≈ 0.43	$\approx$ **0.57**

$\Longrightarrow$ **Selection error**.

Mollification to avoid selection error

Idea : Replace u with $P_{t,M}(u)$

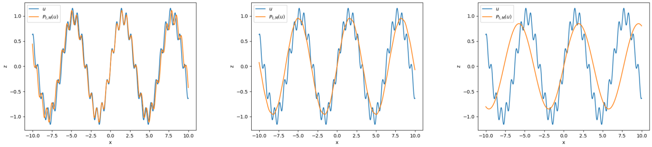
Definition (Mollifying operator)

For $t \in \mathbb{R}^{+*}$ and $M \in \mathbb{R}^{d\times d}$ a positive semi-definite matrix we define:

$$P_{t,M}(u)(x) = \mathbb{E}[u(e^{-tM}x + \sqrt{I_d - e^{-2tM}}\boldsymbol{Z})]$$

where $\boldsymbol{Z} \sim \mu$ is independent of $\boldsymbol{X}$. $P_{t,M}$ is a generalization of the semigroup associated with the Ornstein–Uhlenbeck process.

$P_{t,M}(u)$ kills high-frequency, low-amplitude components of the model.



(a) $t = 0.02$, $M = I_2$ (b) $t = 0.05$, $M = I_2$ (c) $t = 0.2$, $M = I_2$

Figure: Plots of u and $P_{t,M}(u)$ for $x_1 = x_2$ and for different values of t.

Mollifying operator property

- $M = I_d$: isotropic mollification
- M semi-definite positive: anisotropic mollification. Each direction of the eigenvector basis of M is mollified according t times the corresponding eigenvalue.

How to use $P_{t,M}(u)$ instead of u to perform linear dimension reduction ?

Deriving the upper bound

Proposition

For $U_m \in \mathbb{R}^{d\times m}$ a matrix with orthogonal columns and $M \in \mathbb{R}^{d\times d}$ a semi definite positive matrix such that $(U_m U_m^\top)M = M(U_m U_m^\top)$, we have for all $t > 0$:

$$\min_{f\colon \text{measurable}} \mathbb{E}[(u(\boldsymbol{X}) - f(U_m^\top \boldsymbol{X}))^2] \leq \min_{f\colon \text{measurable}} \mathbb{E}[(P_{t,M}(u)(\boldsymbol{X}) - f(U_m^\top \boldsymbol{X}))^2] + \frac{1 - e^{-2\lambda_{\min} t}}{\lambda_{\min}} \mathbb{E}[\|\nabla u(\boldsymbol{X})\|_M^2]$$

where $\boldsymbol{X} \sim \mathcal{N}(0, I_d)$ and where $\|\nabla u(\boldsymbol{X})\|_M^2 = \nabla u(\boldsymbol{X})^\top M \nabla u(\boldsymbol{X})$. Here $\lambda_{\min}$ is **the smallest non zero eigenvalue** of M.

Mollified Active Subspace

- Choice of M:
  - $M = I_d$: isotropic mollification.
  - $M = H(u)$: anisotropic mollification, each direction of the AS basis is mollified according to the corresponding eigenvalue.
  - $M = U_{m_0} U_{m_0}^\top$ where $U_{m_0} \in \mathbb{R}^{d \times m_0}$ contains the m_0 first eigenvectors of $H(u)$. Truncated isotropic mollification: mollify the m_0 first directions of the AS basis.
- Choice of t:

  For a given tolerance $\epsilon > 0$ set the residual error to $\frac{\epsilon}{2}$, i.e **set t such that** :
  $\frac{1-e^{-2\lambda_{\min} t}}{\lambda_{\min}} \mathbb{E}[\|\nabla u(\boldsymbol{X})\|_M^2] = \frac{\epsilon}{2}$
- Mollified Active Subspace (MAS) algorithm:
  1. Compute the AS matrix $H(u)$
  2. Set parameter M and t and compute the MAS matrix $H(P_{t,M}(u))$
  3. Minimize the AS bound and the MAS bound over the feature map $g(x) = U_m^\top x$ and choose the one that gives the **lowest bound**.

How to estimate $H(P_{t,M}(u))$?

For $(x_1, \dots, x_{n_{\text{train}}})$ samples of $\boldsymbol{X} \sim \mathcal{N}(0, I_d)$. We estimate for $H(u)$ with:

$$\widehat{H}(u) = \frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} \nabla u(x_i) \nabla u(x_i)^\top$$

Proposition

For $u \in \mathbb{L}^2(\mathbb{R}^d)$, $t > 0$ and $M \in \mathbb{R}^{d \times d}$ a semi definite positive matrix we have :

$$H(P_{t,M}(u)) = \mathbb{E}[\nabla P_{t,M}(u)(\boldsymbol{X}) \nabla P_{t,M}(u)(\boldsymbol{X})^\top] = e^{-tM} \mathbb{E}_{\boldsymbol{Y},\boldsymbol{Y}'} \left[\nabla u(\boldsymbol{Y}) \nabla u(\boldsymbol{Y}')^\top \right] e^{-tM}$$

for $(\boldsymbol{Y}, \boldsymbol{Y}')^\top \sim \mathcal{N}(0, \Gamma)$, where $\Gamma = \begin{pmatrix} I_d & e^{-2tM} \\ e^{-2tM} & I_d \end{pmatrix}$.

For $(y_1, \dots, y_{n_{\text{add}}})$ samples of $\boldsymbol{Y} \sim \mathcal{N}(0, I_d)$ independent of $\boldsymbol{X}$. We have:

$$\widehat{H}(P_{t,M}(u)) = \frac{e^{-tM}}{n_{\text{train}} n_{\text{add}}} \left(\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \nabla u(x_i) \nabla u(e^{-2tM} x_i + \sqrt{1 - e^{-4tM}} y_j)^\top \right)_{\text{Sym}} e^{-tM}.$$

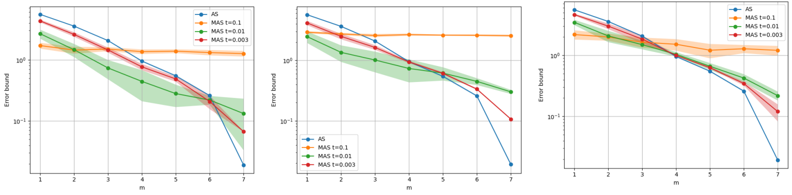
$\Longrightarrow$ Requires n_{add} times more samples than AS.

Numerical example

Given vectors $a = (a_1, \ldots, a_d) \in \mathbb{R}^d$ and $\omega = (\omega_1, \ldots, \omega_d) \in \mathbb{R}^d$, we define:

$$u : (x_1, \ldots, x_d) \longmapsto \sum_{i=1}^{d} a_i \sin(\omega_i x_i).$$

Here we set $d = 8$, $a = \left(2, 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}\right)$ and $\omega = (1, 1, 4, 7, 9, 2, 7, 9)$.

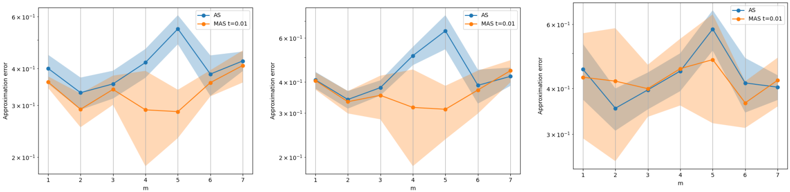


(a) $M = I_2$ (b) $M = H(u)$ (c) $M = U_4 U_4^\top$

Figure: Error bound according to m for different choices of M and t. Here $n_x = 500$ and $n_y = 10$. The solid line represents the mean value over 10 runs, while the shaded area denotes the region of mean $\pm$ standard deviation. Here $U_4 U_4^\top$ is the projection onto the 4th leading eigenvectors of $H(u)$.

Numerical example - Approximation error

f is a FCNN with 3 hidden layers of 20 neurons each and a ReLU activation function. f is trained by minimizing $\frac{1}{n_x}\sum_{i=1}^{n_x}(u(x_i) - f \circ g(x_i))^2$ using ADAM optimizer.



(a) $M = I_2$ (b) $M = H(u)$ (c) $M = U_4 U_4^\top$

Figure: Approximation error according to m for $t = 0.01$ and different choices of M. Here $n_x = 500$ and $n_y = 10$. The approximation error is estimated on a testing set of 10000 samples. The solid line represents the mean value over 10 runs, while the shaded area denotes the region of mean $\pm$ standard deviation.

Conclusion

- Active subspace:
  - To apply first, works well for many models
  - Limitations for functions with nonlinear low dimension structure and for functions with high frequency low amplitude components
- Diffeomorphism feature learning:
  - A non linear extension to Active Subspace: allows to tackle a wider range of models
  - No closed-form minimizer of the bound, requires stochastic optimization
- Mollified active subspace:
  - Allows to correct selection errors for models with oscillatory behaviors
  - Closed-form minimizer of the bound
  - Requires more samples compared to Active Subspace
- **Perspectives**:
  - Importance sampling to explore different values of t
  - Randomized linear algebra to tackle very high dimension ($d \geq 1000$) in the nonlinear setting
  - Application to neural networks compression
  - Application to a complex biogeochemical model

Conclusion

Thank you for your attention !

Romain Verdière, Clémentine Prieur, and Olivier Zahm (Dec. 2023). "Diffeomorphism-based feature learning using Poincaré inequalities on augmented input space". working paper or preprint. URL: https://hal.science/hal-04364208

Romain Verdière, Clémentine Prieur, and Olivier Zahm (May 2025). "Mollifiers to enhance gradient based dimension reduction". working paper or preprint (coming soon)

Bakry, Dominique et al. (2008). "A simple proof of the Poincaré inequality for a large class of probability measures". In: **Electronic Communications in Probability** 13, pp. 60 –66.

Baydin, Atilim Gunes et al. (2018). "Automatic differentiation in machine learning: a survey". In: **Journal of Marchine Learning Research** 18, pp. 1–43.

Bigoni, Daniele et al. (2022). "Nonlinear dimension reduction for surrogate modeling using gradient information". In: **Information and Inference: A Journal of the IMA** 11.4, pp. 1597–1639.

Constantine, Paul G, Eric Dow, and Qiqi Wang (2014). "Active subspace methods in theory and practice: applications to kriging surfaces". In: **SIAM Journal on Scientific Computing** 36.4, A1500–A1524.

Griewank, Andreas et al. (1989). "On automatic differentiation". In: **Mathematical Programming: recent developments and applications** 6.6, pp. 83–107.

Plessix, R-E (2006). "A review of the adjoint-state method for computing the gradient of a functional with geophysical applications". In: **Geophysical Journal International** 167.2, pp. 495–503.

Verdière, Romain, Clémentine Prieur, and Olivier Zahm (Dec. 2023). "Diffeomorphism-based feature learning using Poincaré inequalities on augmented input space". working paper or preprint. URL: https://hal.science/hal-04364208.

— (May 2025). "Mollifiers to enhance gradient based dimension reduction". working paper or preprint.

Zahm, Olivier et al. (2019). "Gradient-based dimension reduction of multivariate vector-valued functions". In: arXiv: 1801.07922 [math.AP]. URL: https://arxiv.org/abs/1801.07922.