GMLS-Nets: Scientific Machine Learning Methods for Unstructured Data

Nathaniel Trask[1,+], **Ravi G. Patel**[1], **Ben J. Gross**[2], **Paul J. Atzberger**[2,†]

[1] Sandia National Laboratories
Center for Computing Research
[+]natrask@sandia.gov

[2] University of California Santa Barbara
[†]atzberg@gmail.com
http://atzberger.org/

Abstract

Data fields sampled on irregularly spaced points arise in many applications in the sciences and engineering. For regular grids, Convolutional Neural Networks (CNNs) have been successfully used to gain benefits from weight sharing and translational invariance. We generalize CNNs by introducing methods for data on unstructured point clouds based on Generalized Moving Least Squares (GMLS). GMLS is a non-parametric technique for estimating linear bounded functionals from scattered data, and has recently emerged as an effective technique for solving partial differential equations. By parameterizing the GMLS estimator, we obtain learning methods for linear and non-linear operators with unstructured stencils. In GMLS-Nets the necessary calculations are local, readily parallelizable, and the estimator is supported by a rigorous approximation theory. We show how the framework may be used for unstructured physical data sets to perform functional regression to identify associated differential operators, develop predictive dynamical models, and obtain feature extractors to predict quantities of interest. The results show the promise of these architectures as foundations for data-driven model development in scientific machine learning applications.

GMLS-Nets: Neural Networks for Scattered Data Sets

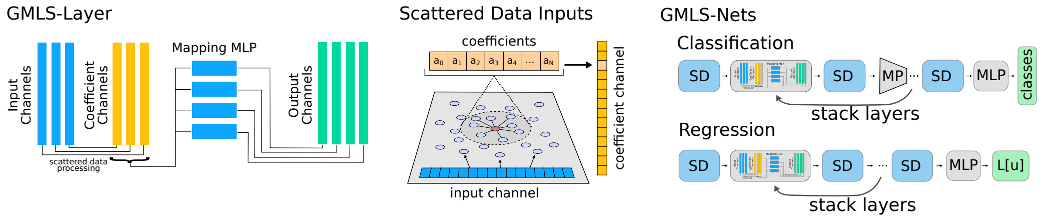

Figure 1: GMLS-Nets. Scattered data inputs are processed by learnable operators $\tau[u]$ parameterized via GMLS estimators. At each location, point data is encoded as coefficient vectors $\mathbf{a}$ by equation 2 input to learnable mapping $q_{\tilde{\mathbf{x}},\xi}(\mathbf{a}(u))$ of equation 3. GMLS-Layers can be stacked to obtain deeper architectures for classification and regression tasks *(inset, SD: scattered data, MP: max-pool, MLP: multi-layer perceptron)*.

Implementations in TensorFlow and PyTorch are available at https://github.com/rgp62/gmls-nets and https://github.com/atzberg/gmls-nets.

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

Many scientific and engineering applications require processing data sets sampled on irregularly spaced points. Such settings include e.g. sensors associated with data sites evolving under unknown or partially known dynamics, or scientific simulation data over unstructured meshes. Recently, there has been an increasing interest in scientific machine learning (SciML) [2] targeting data-driven techniques for the sciences. Here, data is often scarce or highly constrained, suggesting the most successful SciML strategies will leverage prior knowledge to enhance information gains [1, 2]. This could include physical properties and invariances, such as transformation symmetries, conservation structure, or mathematical knowledge such as solution regularity [1, 3, 6, 11]. We introduce methods here based on Generalized Moving Least Squares (GMLS). Several works seek similar extensions of CNNs to unstructured data [4, 5, 13, 14, 19]; a notable feature of the current approach is that it results in efficient, local, highly parallelizable problems.

GMLS is a non-parametric functional regression technique to construct approximations of linear, bounded functionals. On a Banach space $\mathbb{V}$ with dual space $\mathbb{V}^*$, we estimate target functional $\tau_{\tilde{\mathbf{x}}}[u] \in \mathbb{V}^*$ acting on $u = u(\mathbf{x}) \in \mathbb{V}$, where $\mathbf{x}, \tilde{\mathbf{x}}$ denote locations in compact domain $\Omega \subset \mathbb{R}^d$. We assume u is characterized by an unstructured collection of sampling functionals, $\Lambda(u) := \{\lambda_j(u)\}_{j=1}^N \subset \mathbb{V}^*$. We construct the estimate by considering $\mathbb{P} \subset \mathbb{V}$ and seek an element $p^* \in \mathbb{P}$ which provides an optimal reconstruction of the samples in weighted-ℓ_2 sense

$$p^* = \underset{p \in \mathbb{P}}{\text{argmin}} \sum_{j=1}^{N} \left(\lambda_j(u) - \lambda_j(p)\right)^2 \omega(\lambda_j, \tau_{\tilde{\mathbf{x}}}). \tag{1}$$

Here $\omega(\lambda_j, \tau_{\tilde{\mathbf{x}}})$ is a kernel function weighting the spatial correlation between the target functional and sampling set. If one associates locations $\mathbb{X}_h := \{\mathbf{x}_j\}_{j=1}^N \subset \Omega$ with $\Lambda(u)$, one may select a radial kernel $\omega = W_\epsilon(||\mathbf{x}_j - \tilde{\mathbf{x}}||_2)$ with compact support $r < \epsilon$. Using basis $\mathbb{P} = \text{span}\{\phi_1, ..., \phi_{\dim(\mathbb{P})}\}$ and denoting $\Phi(x) = \{\phi_i(x)\}_{i=1,...,dim(\mathbb{P})}$, the optimal reconstruction of u over $\mathbb{P}$ and subsequent GMLS estimate of $\tau_{\tilde{\mathbf{x}}}$ are given by

$$p^* = \Phi(x)^\intercal \mathbf{a}(u), \qquad \tau_{\tilde{\mathbf{x}}}^h[u] = \tau_{\tilde{\mathbf{x}}}(\Phi)^\intercal \mathbf{a}(u). \tag{2}$$

The GMLS estimator thus parameterizes the dual space $\mathbb{V}^*$ by applying the operator to the optimal reconstruction using encoding vector $\boldsymbol{a}$. To construct a framework appropriate for non-linear operators, we consider the more general form

$$\tau_{\tilde{\mathbf{x}}}^h[u] = q_{\tilde{\mathbf{x}},\xi}(\mathbf{a}(u)), \tag{3}$$

where $q_{\tilde{\mathbf{x}},\xi}$ now is a family of possibly nonlinear mappings parameterized by ξ. For simplicity in this work, we specialize by taking: Λ as point evaluations on $\mathbb{X}_h$; $\mathbb{P}$ as $\pi_m(\mathbb{R}^d)$, the space of m^{th}-order polynomials; $W_\epsilon(r) = (1 - r/\epsilon)_+^{\bar{p}}$ for $\bar{p} \in \mathbb{N}$. We emphasize that the framework may be applied more generally using choices of Λ, $\mathbb{P}$, or ω tailored to a given application or physics. For theoretical underpinnings, structure preservation, PDE solvers on manifolds, and other recent applications, we refer readers to [7, 9, 15, 16, 18].

To build SciML architectures, we construct GMLS-Layers whereby GMLS provides an encoder of data over $\mathbb{P}$, expressed via the coefficient vector $\mathbf{a}(u)$, which is then passed into the parameterization Eq. 3. The hyperparameters ξ may be estimated via gradient descent, and we thus obtain an architecture similar to convNets benefiting from weight-sharing that is stackable with appropriate meshfree generalizations of pooling operators acting over $\epsilon-$ball neighborhoods of data (Fig. 1). We explore two possible parameterizations in this work, a linear mapping $q_{\tilde{\mathbf{x}},\xi}(\mathbf{a}(u)) = \xi^\intercal \mathbf{a}(u)$ for $\xi \in \mathbb{R}^{dim(\mathbb{P})}$, and a nonlinear mapping by multilayer perceptrons $q_{\tilde{\mathbf{x}},\xi}(\mathbf{a}(u)) = \mathcal{MLP}_\xi(\mathbf{a}(u))$ where ξ correspond to weights and biases of a dense network with ReLU activation functions.

[1] Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC.,a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

* Work supported by DOE Grant ASCR PhILMS DE-SC0019246.

Data-driven Modeling of Physical Systems

Many scientific data sets are generated by processes for which there are expected governing laws expressible in terms of ordinary or partial differential equations. GMLS-Nets provide natural features to regress such operators from observed state trajectories or responses to fluctuations [17]. We consider learning the following finite difference (FDM) and finite volume (FVM) updates of the dynamics $\frac{\partial u}{\partial t} = \mathcal{L}[u(t,x)]$, where $\mathcal{L}[u]$ can be a linear or non-linear operator and $u^n = u(t^n)$, $t^n = n\Delta t$ are snapshots of the system state at discrete times.

$$\frac{u^{n+1} - u^n}{\Delta t} = \mathcal{L}_{FDM}[\{u^k\}_{k\in\mathcal{K}}; \xi], \qquad \frac{u_i^{n+1} - u_i^n}{\Delta t} = \frac{1}{\mu(c_i)} \sum_{f\in F_i} \int \mathcal{L}_{FVM}[u^{n+1}; \xi] \cdot d\mathbf{A}. \quad (4)$$

Here F_i are cell boundary faces, and $\mu(c_i)$ the volume of cell c_i. The latter is appropriate for extracting models which discretely preserve conservation properties.

The learning capabilities of GMLS-Nets to regress differential operators are shown in Fig. 2, and details may be found in [17]. In these examples, weight-sharing is exploited so that a FDM/FVM scheme may be extracted from either a single snapshot of the analytic solution (advection-diffusion) or from a short burst of several molecular dynamics timesteps (Brownian motion). In both cases, the extracted model provides access to larger timescales without compromising accuracy.

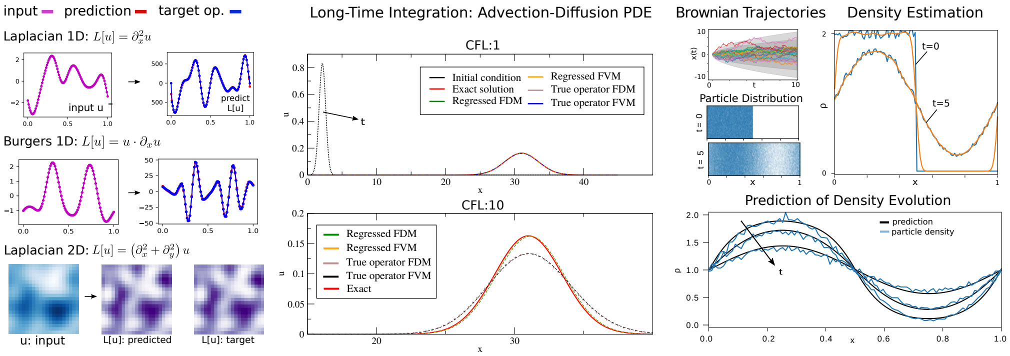

Figure 2: *Left* Regression of Differential Operators. GMLS-Nets can accurately reproduce the action of linear and non-linear operators; 1D/2D Laplacians and Burger's equation are shown. *Middle:* Advection-diffusion solution when $\Delta t = \Delta t_{CFL}$, with $\Delta t_{CFL} \sim 1$ and $\Delta t_{CFL} \gg 1$. An implicit integrator causes FDM/FVM discretizations of true operator $\mathcal{L}$ to be overly dissipative for large timesteps, while regressed long-time FVM operator matches the phase and magnitude of analytic solution almost exactly. *Right:* GMLS-Nets can be trained with molecular-level data to infer continuum dynamical models. Data are simulations of Brownian motion with periodic boundary conditions with the GMLS-Net trained using FVM estimator of equation 4. Predictive continuum model for the density evolution is obtained with good long-term agreement.

Feature Extraction in Fluid Mechanics for Characterization and Prediction

GMLS-Net on unstructured fluid simulation data.

We consider now a SciML application concerning the learning of engineering quantities of interest from unstructured simulation data for the canonical fluid mechanics problem of flow past a cylinder of radius a. To construct a training set, we use a finite volume (FV) code [8] to generate steady state solutions of the Reynolds-Averaged Navier Stokes (RANS) equations with $k - \epsilon$ turbulence model [12], applying velocity inflow conditions ranging over input velocity $U_\infty \in [0.1, 20]$ and viscosity $\nu \in [10^{-2}, 10^8]$, and calculating the drag force F_d experienced by the cylinder. We then extract the velocity field U from cell centers of the FV mesh. In this manner, we obtain an unstructured data set with features U scattered in space, and force characterized by drag coefficient C_d. We do not provide either U_∞ or ν as features. We construct 400 simulations in this manner, use 80% for training and the remainder as a test set (Fig. 3), obtaining a root mean square accuracy of $> 98.5\%$. Note that

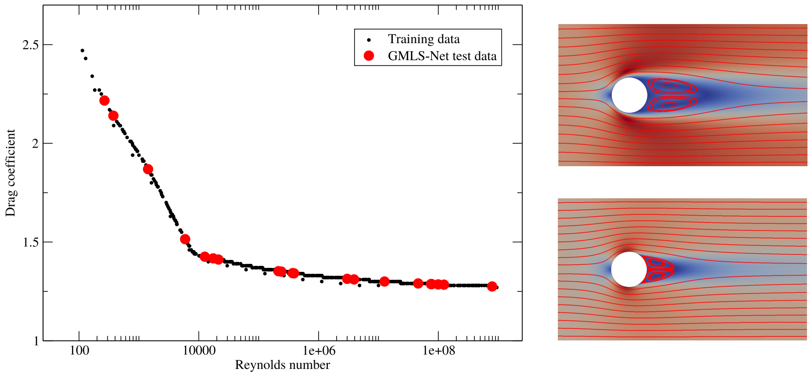

Figure 3: GMLS-Nets are trained on a CFD data set of flow velocity fields while hiding viscosity, inflow condition, and pressure. *Left:* Predicted drag coefficient plotted as a function of Reynolds number for CFD predicted training set (small black dots) and GMLS-Net predicted test set (large red dots). Note that Reynolds number is provided to show data collapse and not used as the training feature. *Right:* Flow velocity fields corresponding to the smallest *(top)* and largest *(bottom)* Reynolds numbers in test set.

while C_d would be post-processed from U, p, and ν in a CFD calculation, we seek here to recover it from only U. This partial characterization is representative of e.g. particle image velocimetry (PIV) calculations.

We highlight several remarkable features of this result. First, drag requires knowledge of both velocity and pressure to calculate the hydrodynamic stresses acting on the cylinder. As we only train on the velocity field, the network is learning to instead characterize drag from flow features. It is well-known that drag may be characterized by the Reynolds number $Re = U_\infty a/\nu$, and that Re correlates with flow features such as the length of recirculation zone [10]. Traditional engineering analysis applies nondimensional analysis to extract a functional relationship of the form

$$\frac{2F_d}{\rho U_\infty^2 A} = C_d\left(\frac{Ua}{\nu}\right), \tag{5}$$

and the drag coefficient $C_d : \mathbb{R} \to \mathbb{R}$ provides collapse of experimental data onto a single curve. This process requires engineering judgement to identify relevant non-dimensional groups, and is typically applicable only to simplified flows. In contrast, this data-driven approach is able to extract an accurate drag prediction without knowledge of ν - suggesting it is able to exploit structure with only partial knowledge of physics. Note that no surrogate velocity field U or p are extracted or approximated; GMLS-Net is able to regress drag directly from the velocity field.

While included here as a simple illustration of GMLS-Nets' ability to naturally handle simulation data relevant to SciML problems, this result translates easily to important engineering problems: e.g. in PIV one may only extract velocity data while pressure data is intractable. This result suggests that important engineering quantities of interest may still be characterized in this setting, despite the lack of complete measurements. While the considered RANS model is a simplistic model with known issues for flows involving recirculation, it provides a simple demonstration of the potential for GMLS-Nets to learn features from more sophisticated computational and experimental flow characterizations.

In conclusion, we demonstrate that GMLS-Nets is capable of obtaining dynamical models for long-time integration beyond the limits of traditional CFL conditions, for making predictions of density evolution of molecular systems, and for predicting directly from flow data quantities of interest in engineering applications. These initial results indicate some promising capabilities of GMLS-Nets for use in data-driven modeling in scientific machine learning applications.

References

[1] P. J. Atzberger. “Importance of the Mathematical Foundations of Machine Learning Methods for Scientific and Engineering Applications”. In: *SciML2018 Workshop, position paper, https://arxiv.org/abs/1808.02213* (2018).

[2] Nathan Baker, Frank Alexander, Timo Bremer, Aric Hagberg, Yannis Kevrekidis, Habib Najm, Manish Parashar, Abani Patra, James Sethian, Stefan Wild, and Karen Willcox. “Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence”. In: (2018).

[3] Yohai Bar-Sinai, Stephan Hoyer, Jason Hickey, and Michael P. Brenner. “Learning data-driven discretizations for partial differential equations”. In: *Proceedings of the National Academy of Sciences* 116.31 (2019), pp. 15344–15349. ISSN: 0027-8424. DOI: 10.1073/pnas.1814058116.

[4] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. “Geometric Deep Learning: Going beyond Euclidean data”. In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 18–42. ISSN: 1053-5888. DOI: 10.1109/MSP.2017.2693418.

[5] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann Lecun. “Spectral networks and locally connected networks on graphs”. English (US). In: *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*. 2014.

[6] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. “Discovering governing equations from data by sparse identification of nonlinear dynamical systems”. In: 113.15 (2016), pp. 3932–3937.

[7] B. J. Gross, N. Trask, P. Kuberry, and P. J. Atzberger. “Meshfree Methods on Manifolds for Hydrodynamic Flows on Curved Surfaces: A Generalized Moving Least-Squares (GMLS) Approach”. In: *arXiv:1905.10469* (2019). URL: https://arxiv.org/abs/1905.10469.

[8] Hrvoje Jasak, Aleksandar Jemcov, Zeljko, and Tukovic Tukovic. *OpenFOAM: A C++ Library for Complex Physics Simulations*. Tech. rep.

[9] Davoud Mirzaei, Robert Schaback, and Mehdi Dehghan. “On generalized moving least squares and diffuse derivatives”. In: *IMA Journal of Numerical Analysis* 32.3 (2012), pp. 983–1000.

[10] John W. Mitchell, Philip J. Pritchard, Alan T. McDonald, Robert W. Fox, and John W. Mitchell. *Fox and McDonald’s introduction to fluid mechanics*. 2014. ISBN: 9781118912652.

[11] Ravi G. Patel and Olivier Desjardins. “Nonlinear integro-differential operator regression with neural networks”. In: *ArXiv* abs/1810.08552 (2018).

[12] Stephen B. Pope. *Turbulent Flows*. Cambridge: Cambridge University Press, 2000. ISBN: 9780511840531.

[13] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 5099–5108.

[14] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. “The Graph Neural Network Model”. In: *Trans. Neur. Netw.* 20.1 (Jan. 2009), pp. 61–80. ISSN: 1045-9227.

[15] Nathaniel Trask, Pavel Bochev, and Mauro Perego. “A conservative, consistent, and scalable meshfree mimetic method”. In: *arXiv preprint arXiv:1903.04621* (2019).

[16] Nathaniel Trask, Mauro Perego, and Pavel Bochev. “A high-order staggered meshless method for elliptic problems”. In: *SIAM Journal on Scientific Computing* 39.2 (2017), A479–A502.

[17] Nathaniel Trask, Ravi G. Patel, Ben J. Gross, and Paul J. Atzberger. “GMLS-Nets: A framework for learning from unstructured data”. In: *arXiv:1909.05371* (2019). URL: https://arxiv.org/abs/1909.05371.

[18] Holger Wendland. *Scattered data approximation*. Vol. 17. Cambridge university press, 2004.

[19] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. “Deep Sets”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 3391–3401.