Learning Neural Optimal Interpolation Models and Solvers

Maxime Beauchamp[1 [0000−0003−0605−8935]], Quentin Febvre[1], Joseph Thompson[1], Hugo Georgenthum[1], and Ronan Fablet[1]

IMT Atlantique Bretagne Pays de la Loire maxime.beauchamp@imt-atlantique.fr

Abstract. The reconstruction of gap-free signals from observation data is a critical challenge for numerous application domains, such as geoscience and space-based earth observation, when the available sensors or the data collection processes lead to irregularly-sampled and noisy observations. Optimal interpolation (OI), also referred to as kriging, provides a theoretical framework to solve interpolation problems for Gaussian processes (GP). The associated computational complexity being rapidly intractable for n-dimensional tensors and increasing numbers of observations, a rich literature has emerged to address this issue using ensemble methods, sparse schemes or iterative approaches. Here, we introduce a neural OI scheme. It exploits a variational formulation with convolutional auto-encoders and a trainable iterative gradient-based solver. Theoretically equivalent to the OI formulation, the trainable solver asymptotically converges to the OI solution when dealing with both stationary and non-stationary linear spatio-temporal GPs. Through a bi-level optimization formulation, we relate the learning step and the selection of the training loss to the theoretical properties of the OI, which is an unbiased estimator with minimal error variance. Numerical experiments for 2D+t synthetic GP datasets demonstrate the relevance of the proposed scheme to learn computationally-efficient and scalable OI models and solvers from data. As illustrated for a real-world interpolation problems for satellite-derived geophysical dynamics, the proposed framework also extends to non-linear and multimodal interpolation problems and significantly outperforms state-of-the-art interpolation methods, when dealing with very high missing data rates.

Keywords: optimal interpolation, differentiable framework, variational model, optimizer learning

1 Introduction

Interpolation problems are critical challenges when dealing with irregularly-sampled observations. Among others, Space earth observation, geoscience, ecology, fisheries generally monitor a process of interest through partial observations due to the characteristics of the sensors and/or the data collection process. As illustrated in Fig.3 for satellite-based earth observation, missing

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

data rates may be greater than 90%, which makes the interpolation problem highly challenging.

Optimal Interpolation (OI) also referred to as kriging [6], provides a theoretical framework to address such interpolation problems for Gaussian processes. Given the covariance structure of the process of interest along with the covariance of the observation noise, one can derive the analytical OI solution. For high-dimensional states, such as space-time processes, the computation of this analytical solution rapidly becomes intractable as it involves the inversion of a $N \times N$ matrix with N the number of observation points. When dealing with space-time processes, OI also relates to data assimilation [2]. In this context, Kalman methods, including ensemble-based extensions, exploit the sequential nature of the problem to solve the OI problem allowing for dynamical flow propagation of the uncertainties.

Data-driven and learning-based approaches have also received a growing interest to address interpolation problems [3, 15], while Image and video inpainting are popular interpolation problems in computer vision [22]. As they typically relate to object removal applications or restoration problems, they usually involve much lower missing data rates than the ones to be delt with in natural images, which are likely not representative of space-time dynamics addressed in geoscience, meteorology, ecology... A recent literature has also emerged to exploit deep learning methods to solve inverse problems classically stated as the minimization of a variational cost. This includes neural architectures based on the unrolling of minimization algorithms [17, 21].

Here, we introduce a neural OI framework. Inspired by the neural method introduced in [8] for data assimilation, we develop a variational formulation based on convolutional auto-encoders and introduce an associated trainable iterative gradient-based solver. Our key contributions are four-fold:

- We show that our variational formulation is equivalent to OI when dealing with Gaussian processes driven by linear dynamics. Under these assumptions, our trainable iterative gradient-based solver converges asymptotically towards the OI solution;
- Regarding the definition of the training losses, we relate the learning step of the proposed neural architecture to the properties of the OI solution as an unbiased estimator with minimal error variance;
- Our framework extends to learning optimal interpolation models and solvers for non-linear/non-Gaussian processes and multimodal observation data;
- Numerical experiments for a 2D+t Gaussian process support the theoretical equivalence between OI and our neural scheme for linear Gaussian case-studies. They also illustrate the targeted scalable acceleration of the interpolation.
- We report a real-world application to the interpolation of sea surface dynamics from satellite-derived observations. Our neural OI scheme significantly outperforms the state-of-the-art methods and can benefit from multimodal observation data to further improve the reconstruction performance.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

To make easier the reproduction of our results, an open-source version of our code is available [1].

This paper is organized as follows. Section 2 formally introduces optimal interpolation and related work. We present the proposed neural OI framework in Section 3. Section 4 reports numerical experiments for both synthetic GP datasets and real-world altimetric sea surface observations. We discuss further our main contributions in Section 5.

2 Problem statement and Related work

For a n-dimensional Gaussian process $\mathbf{x}$ with mean μ and covariance $\mathbf{P}$, the optimal interpolation states the reconstruction of state $\mathbf{x}$ from noisy and partial observations $\mathbf{y}$ as the minimization of a variational cost:

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - H_\Omega \cdot \mathbf{x}\|^2_{\mathbf{R}} + \|\mathbf{x} - \mu\|^2_{\mathbf{P}} \tag{1}$$

with H_Ω denotes the observation matrix to map state $\mathbf{x}$ over domain $\mathcal{D}$ to the observed domain Ω. $\|\cdot\|^2_{\mathbf{R}}$ is the Mahanalobis norm w.r.t the covariance of the observation noise $\mathbf{R}$ and $\|\cdot\|^2_{\mathbf{P}}$ the Mahanalobis distance with *a priori* covariance $\mathbf{P}$. The latter decomposes as a 2-by-2 block matrix $[\mathbf{P}_{\Omega,\Omega}\mathbf{P}_{\Omega,\overline{\Omega}}; \mathbf{P}_{\overline{\Omega},\overline{\Omega}}\mathbf{P}^T_{\Omega,\overline{\Omega}}]$ with $\mathbf{P}_{\mathcal{A},\mathcal{A}'}$ the covariance between subdomains $\mathcal{A}$ and $\mathcal{A}$ of domain $\mathcal{D}$.

The OI variational cost (1) being linear quadratic, the solution of the optimal interpolation problem is given by:

$$\widehat{\mathbf{x}} = \mu + \mathbf{K} \cdot \mathbf{y} \tag{2}$$

with $\mathbf{K}$ referred to as the Kalman gain $\mathbf{P}H^\intercal_\Omega(H_\Omega\mathbf{P}H^\intercal_\Omega + \mathbf{R})^{-1}$, where $\mathbf{P}H^\intercal_\Omega$ is the (grid,obs) prior covariance matrix, $H_\Omega\mathbf{P}H^\intercal_\Omega$ is the (obs,obs) prior covariance matrix. For high-dimensional states, such as nD and nD+t states, and large observation domains, the computation of the Kalman gain becomes rapidly intractable due to the inversion of a $|\Omega| \times |\Omega|$ covariance matrix. This has led to a rich literature to solve minimization (1) without requiring the above-mentioned $|\Omega| \times |\Omega|$ matrix inversion, among others gradient-based solvers using matrix-vector multiplication (MVMs) reformulation [18], methods based on sparse matrix decomposition with tapering [19] or precision-based matrix parameterizations [16].

Variational formulations have also been widely explored to solve inverse problems. Similarly to (1), the general formulation involves the sum of a data fidelity term and of a prior term [2]. In a model-driven approach, the latter derives from the governing equations of the considered processes. For instance, data assimilation in geoscience generally exploits PDE-based terms to state

[1] To be made available in a final version

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

the prior on some hidden dynamics from observations. In signal processing and computational imaging, similar formulations cover a wide range of inverse problems, including inpainting issues [4]:

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \mathcal{J}_{\Phi}(\mathbf{x}, \mathbf{y}, \mathbf{\Omega}) = \arg\min_{\mathbf{x}} \mathcal{J}^{o}(\mathbf{x}, \mathbf{y}) + \lambda \left\| \mathbf{x} - \Phi(\mathbf{x}) \right\|^2$$

$\mathcal{J}^{o}(\mathbf{x}, \mathbf{y})$ is the data fidelity term which is problem-dependent. The prior regularization term $\|\mathbf{x} - \Phi(\mathbf{x})\|^2$ can be regarded as a projection operator. This parameterization of the prior comprises both gradient-based priors using finite-difference approximations, proximal operators as well as plug-and-play priors [17]. As mentioned above, these formualtions have also gained interest in the deep learning literature for the definition of deep learning schemes based on the unrolling of minimization algorithms [1] for (**??**). Here, we further explore the latter category of approaches to solve optimal problems stated as (1), including when covariance $\mathbf{P}$ is not known a priori.

3 Neural OI framework

This Section presents the proposed trainable OI framework. We first introduce the proposed neural OI solver (Section 3.1) and the associated learning setting (Section 3.2). We then describe extensions to non-linear and multimodal interpolation problems.

3.1 Neural OI model and solver

Let us introduce the following variational formulation to reconstruct state $\mathbf{x}$ from partial observations $\mathbf{y}$:

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \mathcal{J}_{\Phi}(\mathbf{x}, \mathbf{y}, \Omega) = \arg\min_{\mathbf{x}} \left\| \mathbf{y} - H_{\Omega} \cdot \mathbf{x} \right\|^2 + \lambda \left\| \mathbf{x} - \Phi(\mathbf{x}) \right\|^2 \qquad (3)$$

where λ is a positive scalar to balance the data fidelity term and the prior regularization. $\Phi(\cdot)$ is a linear neural auto-encoder which states the prior onto the solution.

Variational formulation (3) is equivalent to optimal interpolation problem (1) when considering a matrix parameterization of the prior $\Phi(x) = (\mathbf{I} - \mathbf{L})\mathbf{x}$ with $\mathbf{L}$ the square-root (as a Cholesky decomposition) of $\mathbf{P}$ and a spherical observation covariance, *i.e.* $\mathbf{R} = \sigma^2 \mathbb{1}$. The proof comes immediately when noting that the regularization term of the variational cost also writes: $\mathbf{x}^{\mathsf{T}}\mathbf{P}^{-1}\mathbf{x} = \mathbf{x}^{\mathsf{T}}\mathbf{L}^{\mathsf{T}}\mathbf{L}\mathbf{x} = ||\mathbf{L}\mathbf{x}||^2 = ||\mathbf{x} - \Phi\mathbf{x}||^2$.

Lemma 1. *For a stationary Gaussian process and a Gaussian observation noise with* $\mathbf{R} = \sigma^2\mathbf{I}$*, we can restate the associated optimal interpolation problem (1) as minimization problem (3) with neural operator* $\Phi(\cdot)$ *being a linear convolutional network.*

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

The proof results from the translation-invariant property of the covariance of stationary Gaussian processes. Computationally, we can derive Φ as the inverse Fourier transform of the square-root of the Fourier transform of covariance $\mathbf{P}$ in (1) as exploited in Gaussian texture synthesis [10].

Lemma 1 provides the basis to learn a solver of variational formulation (3) and address optimal interpolation problem (1). We benefit from automatic differentiation tools associated with neural operators to investigate iterative gradient-based solvers as introduced in meta-learning [12], see Algorithm 1. The latter relies on an iterative gradient-based update where neural operator $\mathcal{G}$ combines an LSTM cell [20] and a linear layer to map the hidden state of the LSTM cell to the space spanned by state $\mathbf{x}$. Through the LSTM may capture long-term dependencies, operator $\mathcal{G}$ defines a momentum-based gradient descent. Overall, this class of generic learning-based methods was explored and referred to as 4DVarNet schemes in [8] for data assimilation problems. Here, as stated in Lemma 2, we parameterize weighting factors $a(\cdot)$ and $\omega(\cdot)$ such that the LSTM-based contribution dominates for the first iterations while for a greater number of iterations the iterative update reduces to a simple gradient descent. Hereafter, we refer to the proposed neural OI framework as 4DVarNet-OI.

Algorithm 1 Iterative gradient-based solver for (3) given initial condition $\mathbf{x}^{(0)}$, observation $\mathbf{y}$ and sampled domain Ω. Let $a(\cdot)$ and $\omega(\cdot)$ be positive scalar functions and $\mathcal{G}$ a LSTM-based neural operator.

$\mathbf{x} \leftarrow \mathbf{x}^{(0)}$
$i \leftarrow 0$
while $i \leq K$ **do**
  $i \leftarrow i + 1$
  $\mathbf{g} \leftarrow \nabla_{\mathbf{x}} \mathcal{J}_{\Phi}\left(\mathbf{x}^{(i)}, \mathbf{y}, \Omega\right)$
  $\mathbf{x} \leftarrow \mathbf{x} - a(i) \cdot [\omega(i) \cdot \mathbf{g} + (1 - \omega(i)) \cdot \mathcal{G}(\mathbf{g})]$
end while

Lemma 2. *Let us consider the following parameterizations for functions $a(\cdot)$ and $\omega(\cdot)$*

$$a(i) = \frac{\nu \cdot K_0}{K_0 + i} \; ; \omega(i) = \tanh\left(\alpha \cdot (i - K_1)\right) \tag{4}$$

where ν and α are positive scalars, and $K_{0,1}$ positive integers. If $\mathcal{G}(\cdot)$ is a bounded operator and $\Phi(\cdot)$ is a linear operator given by $\mathbf{I} - \mathbf{P}^{1/2}$, then Algorithm 1 converges towards the solution (2) of the minimization of optimal interpolation cost (3).

The proof of this lemma derives as follows. As $\mathcal{G}(\cdot)$ is bounded, the considered parameterization of the gradient step in Algorithm 1 is asymptotically equivalent to a simple gradient descent with a decreasing step size. Therefore, it satisfies the convergence conditions towards the global minimum for a linear-quadratic variational cost [5]. We may highlight that the same applies with a stochastic version of Algorithm 1 and a convex variational cost [5].

The boundedness of operator $\mathcal{G}$ derives from that of the LSTM cell. Therefore, Lemma 2 guarantees that Algorithm 1 with a LSTM-based parameterization

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

for operator $\mathcal{G}$ converges to the minimum of optimal interpolation cost (3) whatever the parameters of $\mathcal{G}$.

In this setting, operator $\mathcal{G}$ aims at accelerating the convergence rate of the gradient descent towards analytical solution (2). Overall, we define $\Theta = \{\Phi, \mathcal{G}\}$ and $\Psi_{\Theta}^{K}(\mathbf{x}^{(0)}, \mathbf{y}, \Omega)$ the interpolated state resulting from the application of Algorithm 1 with K iterations from initial condition $\mathbf{x}^{(0)}$ given observation data $\{\mathbf{y}, \Omega\}$.

3.2 Learning setting

Formally, we state the training of the considered neural OI scheme (3) according to a bi-level optimization problem

$$\widehat{\Theta} = \arg\min_{\Theta} \mathcal{L}\left(\mathbf{x}_k, \mathbf{y}_k, \Omega_k, \widehat{\mathbf{x}}_k\right) \quad \text{s.t. } \widehat{\mathbf{x}}_k = \arg\min_{\mathbf{x}_k} \mathcal{J}_{\Phi}\left(\mathbf{x}_k, \mathbf{y}_k, \Omega_k\right) \tag{5}$$

where $\mathcal{L}(\{\mathbf{x}_k, \mathbf{y}_k, \widehat{\mathbf{x}}_k\})$ defines a training loss and k denotes the time index along the data assimilation window (DAW) $[t - k\Delta t; t + k\Delta t]$.

Let us consider Optimal Interpolation problem (1) with a spherical observation covariance $\mathbf{R} = \sigma^2 \cdot \mathbf{I}$ and prior covariance $\mathbf{P}$. Let us parameterize trainable operator Φ in (3) as a linear convolution operator. Optimal interpolation (2) is then solution of a bi-level optimization problem (5) with $\Phi = \mathbf{I} - \mathbf{P}^{1/2}$ for each of the following training losses:

$$\begin{aligned} \mathcal{L}_1\left(\mathbf{x}_k, , \widehat{\mathbf{x}}_k\right) &= \sum_{k=1}^{N} \left\| \mathbf{x}_k - \widehat{\mathbf{x}}_k \right\|^2 \\ \mathcal{L}_2\left(\mathbf{x}_k, \mathbf{y}_k, \widehat{\mathbf{x}}_k\right) &= \sum_{k=1}^{N} \left\| \mathbf{y}_k - H_{\Omega} \cdot \mathbf{x}_k \right\|_{\mathbf{R}}^2 + \left\| \mathbf{x}_k \right\|_{\mathbf{P}}^2 \end{aligned} \tag{6}$$

where $\mathcal{L}_1$ denotes the mean squared error (MSE) w.r.t true states and $\mathcal{L}_2$ stands for the OI variational cost. This results from the equivalence between variational formulations (1) and (3) under parameterization $\Phi = \mathbf{I} - \mathbf{P}^{1/2}$ and the property that OI solution (2) is a minimum-variance unbiased estimator.
Such a formulation motivates the following training setting for the proposed scheme:

$$\widehat{\Theta} = \arg\min_{\Theta} \mathcal{L}\left(\mathbf{x}_k, \mathbf{y}_k, \widehat{\mathbf{x}}_k\right) \quad \text{s.t. } \widehat{\mathbf{x}}_k = \Psi_{\Theta}^{K}\left(\mathbf{x}_k^{(0)}, \mathbf{y}_k, \Omega_k\right) \tag{7}$$

where $\mathcal{L}$ is either $\mathcal{L}_1$ or $\mathcal{L}_2$ and $\mathbf{x}_k^{(0)}$ is an initial condition. Let stress that $\mathcal{L}_2$ relates to unsupervised learning but requires the explicit definition and parameterization of prior covariance $\mathbf{P}$. In such situations, the proposed training framework aims at delivering a fast and scalable computation of (2). If using training loss $\mathcal{L}_1$, it only relies on the true states with no additional hypothesis on the underlying covariance, which makes it more appealing for most supervised

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

learning-based applications, see the experimental conclusions of Section 4. To train jointly the solver component $\mathcal{G}$ and operator Φ, we vary initial conditions between some initialization $\mathbf{x}_k^{(0)}$ of the state and detached outputs of Algorithm 1 for a predefined number of total iteration steps. This strategy also provides a practical solution to the memory requirement, which rapidly increases with the number of iterations during the training phase due to the resulting depth of the computational graph. In all the reported experiments, we use Adam optimizer over 200 epochs.

3.3 Extension to non-linear and multimodal optimal interpolation

While the analytical derivation of solution (2) requires to consider a linear-quadratic formulation in both (1) and (3), Algorithm 1 applies to any differentiable parameterization of operator Φ. This provides the basis to investigate optimal interpolation models and solvers for non-linear and/or non-Gaussian processes through a non-linear parameterization for operator Φ. Here, we benefit from the variety of neural auto-encoder architectures introduced in the deep learning literature, such as simple convolutional auto-encoders, U-Nets [7], ResNets [11]... For such parameterization, the existence of a unique global minimum for minimization (3) may not be guaranteed and Algorithm 1 will converge to a local minimum depending on the considered initial condition.

Multimodal interpolation represents another appealing extension of the proposed framework. Let us assume that some additional gap-free observation data $\mathbf{z}$ is available such that $\mathbf{z}$ is expected to partially inform state $\mathbf{x}$. We then introduce the following multimodal variational cost:

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \lambda_1 \left\| \mathbf{y} - H_\Omega \cdot \mathbf{x} \right\|^2 + \lambda_2 \left\| g(\mathbf{z}) - h(\mathbf{x}) \right\|^2 + \lambda_3 \left\| \mathbf{x} - \Phi(\mathbf{x}) \right\|^2 \quad (8)$$

where $g(\cdot)$ and $h(\cdot)$ are trainable neural operators which respectively extract features from state $\mathbf{x}$ and observation $\mathbf{z}$. In this multimodal setting, Θ in (7) comprises the trainable parameters of operators Φ, $\mathcal{G}$, g and h. Given this reparameterization of the variational cost, we can exploit the exact same architecture for the neural solver defined by Algorithm 1 and the same learning setting.

4 Experiments

We report numerical experiments for the interpolation of 2D+t Gaussian process (GP) for which we can compute the analytical OI solution (2), as well as a real-world case-study for the reconstruction of sea surface dynamics from irregularly-sampled satellite-derived observations.

4.1 2D+t GP case-study

We use as synthetic dataset the stochastic partial differential equation (SPDE) approach introduced by [16] to generate a spatio-temporal Gaussian Process

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

(GP). Let $\mathbf{x}$ denote the SPDE solution, we draw from the classic isotropic formulation to introduce some diffusion in the general fractional operator:

$$\left\{\frac{\partial}{\partial t} + \left\{\kappa^2(\mathbf{s},t) - \nabla \cdot \mathbf{H}(\mathbf{s},t)\nabla\right\}^{\alpha/2}\right\}\mathbf{x}(\mathbf{s},t) = \tau \mathbf{z}(\mathbf{s},t) \tag{9}$$

with parameters $\kappa = 0.33$ and regularization variance $\tau = 1$. To ensure the GP to be smooth enough, we use a value of $\alpha = 4$. Such a formulation enables to generate GPs driven by local anisotropies in space leading to non stationary spatio-temporal fields with eddy patterns. Let denote this experiment GP-DIFF2 where $\mathbf{H}$ is a 2-dimensional diffusion tensor. We introduce a generic decomposition of $\mathbf{H}(\mathbf{s},t)$, see e.g. [9], through the equation $\mathbf{H} = \gamma \mathbf{I}_2 + \beta \mathbf{v}(\mathbf{s})^\mathsf{T}\mathbf{v}(\mathbf{s})$ with $\gamma = 1$, $\beta = 25$ and $\mathbf{v}(\mathbf{s}) = (v_1(\mathbf{s}), v_2(\mathbf{s}))^\mathsf{T}$ using a periodic formulation of its two vector fields components: it decomposes the diffusion tensor as the sum of an isotropic and anisotropic effects, the latter being described by its amplitude and magnitude. This is a valid decomposition for any symmetric positive-definite 2×2 matrix.
We use the Finite Difference Method (FDM) in space coupled with an Implicit Euler scheme (IES) in time to solve for the equation. Let $\mathcal{D} = [0, 100] \times [0, 100]$ be the square spatial domain of simulation and $\mathcal{T} = [0, 500]$ the temporal domain. Both spatial and temporal domains are discretized so that the simulation is made on a uniform Cartesian grid consisting of points (x_i, y_j, t_k) where x_i=$i\Delta x$, y_j=$j\Delta j$, t_k=$k\Delta t$ with Δx, Δy and Δt all set to one.
To be consistent with the second dataset produced in Section 4.2, we sample pseudo-observations similar to along-track patterns produced by satellite datasets, with a periodic sampling leading to spatial observational rate similar to the along-track case-study. Observational noise is negligible and taken as $\mathbf{R} = \sigma^2\mathbf{I}$ with $\sigma^2 = 1\mathrm{E}-3$ to compute the observational term of the variational cost.

For the dataset GP-DIFF2, we involve spatio-temporal sequences of length 5 as data assimilation window (DAW) to apply our framework and benchmark the following methods: analytical OI, as a solution of the linear system, the gradient descent OI solution, a direct CNN/UNet interpolation using a zero-filling initialization and different flavors of 4DVarnet using either a UNet trainable prior or a known precision-based prior coupled with LSTM-based solvers. As already stated in Section 3.2, we use two training losses: the mean squared error (MSE) w.r.t to the groundtruth and the OI variational cost of Eq. 1. Regarding the performance metrics, we assess the quality of a model based on both OI cost value for the known SPDE precision matrix and the MSE score w.r.t to the groundtruth. We also provide the computational GPU time of all the benchmarked models on the test period. For training-free models (analytical and gradient-based OI), there is no training time. The training period goes from timestep 100 to 400 and the optimization is made on 20 epochs with Adam optimizer, with no significant improvements if trained longer. During the training procedure, we select the best model according to metrics computed

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

over the validation period from timestep 30 to 80. Overall, the set of metrics is computed on a test period going from timestep 450 to 470. Let note that by construction, the analytical OI solution is optimal regarding the OI variational cost.

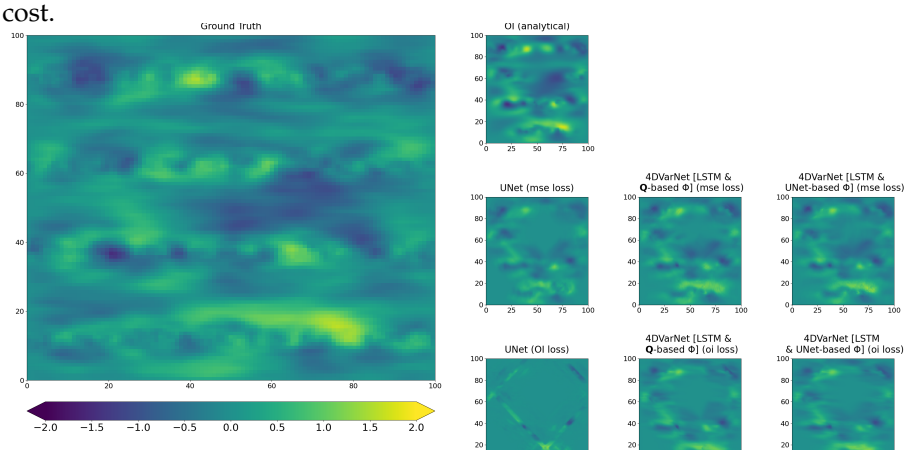


Fig. 1: SPDE-based GP spatio-temporal field (Ground Truth) and its reconstructions based on a 5 time lag assimilation window

Table 1 displays the performance evaluation for the experiment GP-DIFF2. Figure 1 shows the interpolation obtained at the middle of the test period for all the benchmarked models. Because we use spatio-temporal assimilation windows of length 5, we only display the reconstructions at the center of the DAW. We also provide in Figure 2a) the scatterplot of the global MSE w.r.t OI variational cost throughout the iteration process, and in Figure 2b) the OI variational cost vs the number of iterations of the algorithm. Both figures represent how the methods behaves once the training phase is finished: for learning-based gradient-descent approaches, their corresponding line plot then illustrates how the trained recurrent schemes is efficient to mimic and speed up the traditional gradient descent. The mapping clearly indicates that direct inversion by CNN schemes is not efficient for this reconstruction task. On the opposite, LSTM-based iterative solvers are all consistent with the optimal solution, with potential variations that can be explained by the training loss used. While using MSE w.r.t true states quickly converges in a very few number of iterations, 20 typically, involving the OI variational cost as a training loss implies to increase the number of gradient steps, up to about a hundred, to reach satisfactory performance. This makes sense because using global MSE relates to supervised learning while the variational cost-based training is not.

In addition, when a similar LSTM-based solver is used, the trainable prior considerably speeds up the convergence towards the optimal solution compared to the known precision matrix-based prior. This leads to three key conclusions:

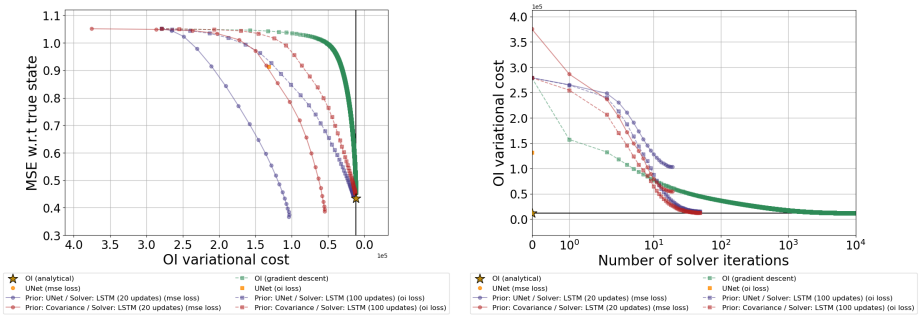
- using MSE as training loss with trainable neural priors is enough for a reconstruction task and can even speed up the iterative convergence compared to the known statistical prior parametrization;

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

- the GP-based experiments demonstrates the relevance of an LSTM-based solver to speed up and accurate iterative solutions of minimization problems;
- looking for an optimal solution within the bi-level neural optimization of prior and solver may lead to deviate from the original variational cost to minimize.

Table 1: **Interpolation performance for the synthetic 2D+T GP case-study:** For each benchmarked model, we report the considered performance metrics for the three training loss strategies (MSE w.r.t true states and OI variational cost)

GP	Approach	Prior	Training loss	MSE_x	OI-score	Comp. time (mins)
GP-DIFF2	OI	Covariance		**2.72**	**9.8E+03**	2.41
	UNet	N/A	MSE loss	3.50	6.10E+05	0.08
			OI loss	5.99	1.08E+05	0.49
	4DVarNet-LSTM	Covariance	MSE loss	2.84	4.52E+04	2.4
			OI loss	3.26	1.06E+04	2.68
		UNet	MSE loss	2.74	1.04E+05	0.25
			OI loss	3.17	1.27E+04	0.48



(a) OI cost vs MSE (b) OI cost vs number of iterations

Fig. 2: Optimal Interpolation derived variational cost vs Mean Squared Error (MSE) loss (a) and OI variational cost vs number of iterations (b) for all the benchmarked methods at timestep 16 of the test period throughout their iterations process. For the analytical Optimal Interpolation and direct UNet neural formulations, there is no iteration, so a single point is displayed. Direct UNet trained with MSE loss does not appear in the Figures because it does not scale properly compared to the other methods

4.2 Satellite altimetry dataset

We also apply our neural OI scheme to a real-world dataset, namely the interpolaton of sea surface height (SSH) fields from irregularly-sampled satellite altimetry observations. The SSH relates to sea surface dynamics and satellite altimetry data which are characterized by an average missing data rate

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

above 90%. We exploit the experimental setting defined in [14] [2]. It relies on a groundtruthed dataset given by the simulation of realistic satellite altimetry observations from numerical ocean simulations. Overall, this dataset refers to 2D+t states for a $10^\circ \times 10^\circ$ domain with $1/20^\circ$ resolution corresponding to a small area in the Western part of the Gulf Stream. Regarding the evaluation framework, we refer the reader to SSH mapping data challenge above mentioned for a detailed presentation of the datasets and evaluation metrics. The latter comprises the average RMSE-scores μ(RMSE) (the higher the better), the minimal spatial scales resolved λ_x (degree) (the lower the better) and the minimal temporal scales resolved λ_t (days) (the lower the better). We also look for the relative gains τ_{SSH} (%) and $\tau_{\nabla SSH}$ (%) w.r.t DUACS OI for SSH and its gradient. For learning-based approaches, the training dataset spans from mid-February 2013 to October 2013, while the validation period refers to January 2013. All methods are tested on the test period from October 22, 2012 to December 2, 2012.

For benchmarking purposes, we consider the operational baseline (DUACS) based on an optimal interpolation, multi-scale OI scheme MIOST, model-driven interpolation schemes BFN and DYMOST. We also include a state-of-the-art UNet architecture to train a direct inversion scheme. For all neural schemes, we consider 29-day space-time sequences to account for time scales considered in state-of-the-art OI schemes. Regarding the parameterization of our framework, we consider a bilinear residual architecture for prior Φ, a classic UNet flavor as well as a simple linear convolutional prior. Similarly to the GP case-study, we use a 2D convolutional LSTM cell with 150-dimensional hidden states. Besides the interpolation scheme using only altimetry data, we also implement a multimodal version of our interpolation framework. It uses sea surface temperature (SST) field as complementary gap-free observations. SST fields are widely acknowledge to convey information on sea surface dynamics though the derivation of an explicit relationship between SSH and SST fields remain a challenge, except for specific dynamical regimes [13]. Our multimodal extension exploits simple ConvnNets for the parameterization of operators $g(\cdot)$ and $h(\cdot)$ in Eq.8.

Figure 3 displays the reconstructions of the SSH field and the corresponding gradients on 2012-11-05 for all the benchmarked models. It clearly stresses how our scheme improves the reconstruction when considering a non-linear prior. Especially, we greatly sharpen the gradient along the main meander of the Gulf Stream compared with other interpolation schemes. Oceanic eddies are also better retrieved. Table 2 further highlights the performance gain of the proposed scheme. The relative gain is greater than 50% compared to the operational satellite altimetry processing. We outperform by more than 20% in terms of relative gain to the baseline MIOST and UNet schemes, which are the second best interpolation schemes. Interestingly, our scheme is the only one to retrieve time scales below 10 days when considering only altimetry data.

[2] SSH Mapping Data Challenge 2020a: `https://github.com/ocean-data-challenges/2020a_SSH_mapping_NATL60`

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

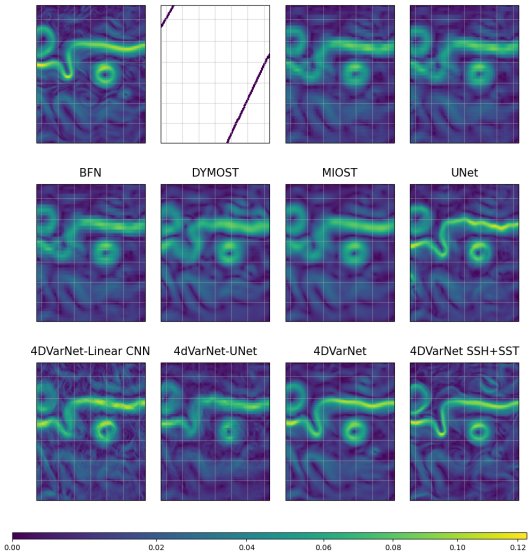


Fig. 3: Gradient SSH reconstructions (2012-11-05) for all benchmarked models based on 4 along-track nadirs pseudo-observations in the Gulf Stream domain. Trom top left to bottom right: Ground Truth, Observations, naive OI, DUACS, BFN, DYMOST, MIOST, UNet, 4DVarNet with a linear CNN-based prior, 4DVarNet with UNet prior, 4DVarNet with BiLin-Res prior and multimodal 4DVarNet with BiLin-Res prior embedding SSH-SST synergies in the variational cost

Table 2: **Interpolation performance for the satellite altimetry case-study:** For each benchmarked models, we report the considered performance metrics averaged on the test period when learning-based methods are trained on the MSE loss (true states and its gradient). Metrics obtained from SOTA DA methods (top lines in the Table) can be found in the BOOST-SWOT 2020a SSH Mapping Data Challenge: `https://github.com/ocean-data-challenges/2020a_SSH_mapping_NATL60`

Approach	Prior	**MSE**	λ_x (degree)	λ_t (days)	τ_{SSH} (%)	$\tau_{\nabla SSH}$ (%)
DUACS	-	0.92	1.42	12.13	-	-
BFN	-	0.92	1.23	10.82	7.93	23.69
DYMOST	-	0.91	1.36	11.91	-10.88	0.38
MIOST	-	0.93	1.35	10.41	25.63	11.16
UNet	-	0.924	1.25	11.33	20.13	26.16
4DVarNet-LSTM	Linear CNN	0.89	1.46	12.63	-84.14	-10.24
	UNet	0.89	1.4	12.45	0.24	0.01
	BiLin-Res	**0.94**	**1.17**	**6.86**	**54.79**	**55.14**
Multimodal interpolation models (SSH+SST)						
UNet	-	0.55	2.36	35.72	-2741.29	-355.24
4DVarNet-LSTM	BiLin-Res	**0.96**	**0.66**	**2.97**	**79.80**	**75.71**

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

As stressed by last line and map of Table 2 and Figure 3, the multimodal version of the proposed interpolation scheme further improves the interpolation performance. Our trainable OI solver learns how to extract fine-scale features from SST fields to best reconstruct the fine-scale structure of SSH fields and brings a significant improvement on all performance metrics.

5 Conclusion

This paper addresses the end-to-end learning of neural schemes for optimal interpolation. We extend the neural scheme introduced in [8] for data assimilation to optimal interpolation with theoretical guarantees so that the considered trainable solvers asymptotically converge towards the analytical OI solution. The computation of the analytical OI solution is challenging when dealing with high-dimensional states. Whereas classic gradient-based iterative methods may suffer from a relatively low convergence rate, our experiments support the relevance of the proposed trainable solvers to speed up the convergence and reach good interpolation performance with only 10 to 100 gradient steps. Importantly, the convolutional architecture of the trainable solver also guarantees their scalability and a linear complexity with respect to the size of the spatial domain as well as the number of observations. Our GP experiment highlight the relevance of the bi-level formulation of the OI problem.We greatly speed up the interpolation time, when considering a UNet-based parameterization of the inner cost and the interpolation error as the outer performance metrics. The latter strategy greatly simplifies the application of the proposed framework to real datasets, where the underlying covariance model is not known and/or a Gaussian process approximation does not apply. As illustrated for our application to ocean remote sensing data, the proposed framework greatly outperforms all SOTA techniques, especially when benefitting from additional multimodal observations. Whereas in the GP case, we know the variational OI cost to be the optimal variational formulation to solve the interpolation, no such theoretical result exists in most non-Gaussian/non-linear cases. The proposed end-to-end learning framework provides new means to explore the reduction of estimation biases in Bayesian setting. Especially, our experiments on ocean remote sensing data suggest that the prior term in the inner variational formulation shall be adapted to the observation configuration rather than considering generic plug-and-play priors. This works also supports new avenues thanks to the connection made between neural Optimal Interpolation and trainable solvers. Indeed, while the GP experiment used in the paper is entirely controlled, in the sense that the parameters of the stochastic PDE driving the GP are known, future works may consider to also train the SPDE parameters so that the prior operator Φ would be linear, though stochastic. It would open the gate to uncertainty quantification and fast huge ensemble-based formulations. In addition, it would pave the way to a full stochastic neural formulation of the framework, when making the explicit link between diffusion-based generative models and SPDE that are in fact linear diffusion models.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

Bibliography

[1] Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M.W., Pfau, D., Schaul, T., Shillingford, B., De Freitas, N.: Learning to learn by gradient descent by gradient descent. In: Advances in neural information processing systems. pp. 3981–3989 (2016)

[2] Asch, M., Bocquet, M., Nodet, M.: Data Assimilation. Fundamentals of Algorithms, Society for Industrial and Applied Mathematics (Dec 2016). https://doi.org/10.1137/1.9781611974546, https://doi.org/10.1137/1.9781611974546

[3] Barth, A., Alvera-Azcárate, A., Troupin, C., Beckers, J.M.: Dincae 2.0: multivariate convolutional neural network with error estimates to reconstruct sea surface temperature satellite and altimetry observations. Geoscientific Model Development **15**(5), 2183–2196 (2022). https://doi.org/10.5194/gmd-15-2183-2022, https://gmd.copernicus.org/articles/15/2183/2022/

[4] Bertalmio, M., Bertozzi, A.L., Sapiro, G.: Navier-Stokes, fluid dynamics, and image and video inpainting. In: IEEE CVPR. pp. 355–362 (2001)

[5] Borkar, V.: Stochastic Approximation, Texts and Readings in Mathematics, vol. 48 (2008), http://link.springer.com/10.1007/978-93-86279-38-5

[6] Chilès, J., Delfiner, P.: Geostatistics : modeling spatial uncertainty. Wiley, New-York, second edn. (2012)

[7] Cicek, O., Abdulkadir, A., Lienkamp, S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Proc. MICCAI. pp. 424–432 (2016)

[8] Fablet, R., Beauchamp, M., Drumetz, L., Rousseau, F.: Joint interpolation and representation learning for irregularly sampled satellite-derived geophysical fields. Frontiers in Applied Mathematics and Statistics **7**, 25 (2021). https://doi.org/10.3389/fams.2021.655224, https://www.frontiersin.org/article/10.3389/fams.2021.655224

[9] Fuglstad, G.A., Lindgren, F., Simpson, D., Rue, H.: Exploring a new class of non-stationary spatial gaussian random fields with varying local anisotropy. Statistica Sinica **25**(1), 115–133 (2015). https://doi.org/10.5705/ss.2013.106w

[10] Galerne, B., Gousseau, Y., Morel, J.: Random Phase Textures: Theory and Synthesis. IEEE Transactions on Image Processing **20**(1), 257–267 (Jan 2011). https://doi.org/10.1109/TIP.2010.2052822, conference Name: IEEE Transactions on Image Processing

[11] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (Jun 2016). https://doi.org/10.1109/CVPR.2016.90

[12] Hospedales, T., Antoniou, A., Micaelli, P., Storkey, A.: Meta-learning in neural networks: A survey. arXiv:2004.05439 (2020)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28

[13] Isern-Fontanet, J., Chapron, B., Lapeyre, G., Klein, P.: Potential use of microwave sea surface temperatures for the estimation of ocean currents. GEOPHYSICAL RESEARCH LETTERS **33** (2006), doi:10.1029/2006GL027801, l24608

[14] Le Guillou, F., Metref, S., Cosme, E., Ubelmann, C., Ballarotta, M., Verron, J., Le Sommer, J.: Mapping altimetry in the forthcoming swot era by back-and-forth nudging a one-layer quasi-geostrophic model. Earth and Space Science Open Archive p. 15 (2020). https://doi.org/10.1002/essoar.10504575.1, https://doi.org/10.1002/essoar.10504575.1

[15] Lguensat, R., Huynh Viet, P., Sun, M., Chen, G., Fenglin, T., Chapron, B., Fablet, R.: Data-driven Interpolation of Sea Level Anomalies using Analog Data Assimilation (Oct 2017), https://hal.archives-ouvertes.fr/hal-01609851

[16] Lindgren, F., Rue, H., Lindström, J.: An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**(4), 423–498 (2011). https://doi.org/10.1111/j.1467-9868.2011.00777.x, https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2011.00777.x

[17] McCann, M., Jin, K., Unser, M.: Convolutional Neural Networks for Inverse Problems in Imaging: A Review. IEEE SPM **34**(6), 85–95 (2017). https://doi.org/10.1109/MSP.2017.2739299

[18] Pleiss, G., Jankowiak, M., Eriksson, D., Damle, A., Gardner, J.R.: Fast matrix square roots with applications to gaussian processes and bayesian optimization (2020). https://doi.org/10.48550/ARXIV.2006.11267, https://arxiv.org/abs/2006.11267

[19] Romary, T., Desassis, N.: Combining covariance tapering and lasso driven low rank decomposition for the kriging of large spatial datasets (2018). https://doi.org/10.48550/ARXIV.1806.01558, https://arxiv.org/abs/1806.01558

[20] Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015), https://proceedings.neurips.cc/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf

[21] Wei, K., Aviles-Rivero, A., Liang, J., Fu, Y., Schönlieb, C.R., Huang, H.: Tuning-free Plug-and-Play Proximal Algorithm for Inverse Imaging Problems. In: ICML. pp. 10158–10169 (2020), https://proceedings.mlr.press/v119/wei20b.html, iSSN: 2640-3498

[22] Xu, R., Li, X., Zhou, B., Loy, C.C.: Deep flow-guided video inpainting. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_28