

Meta-Learning as Prediction Map Approximation

Wessel Bruinsma

University of Cambridge and Invenia Labs

Research Talk at Sheffield Machine Learning Group, 24 Feb 2022

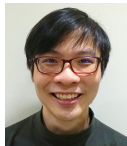
Collaborators



Wessel
Bruinsma



Jonathan
Gordon



Andrew
Foong



James
Requeima



Stratis
Markou



Anna
Vaughan



Yann
Dubois

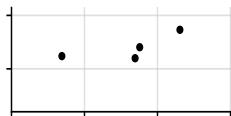


Rich
Turner

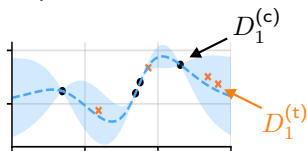
π : data sets \mathcal{D}

\rightarrow

predictions \mathcal{P}



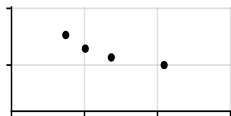
$\xrightarrow{\pi}$



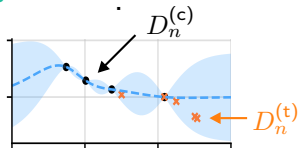
\vdots

neural process

\vdots



$\xrightarrow{\pi}$

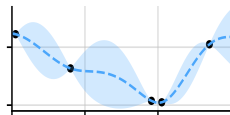


training

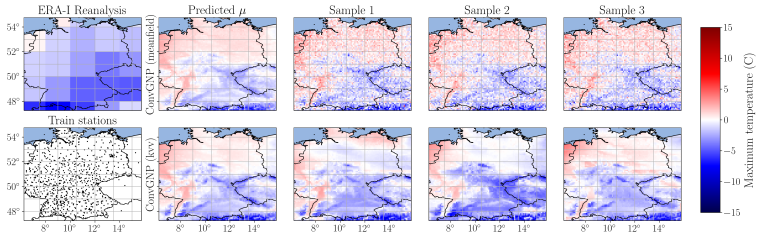
test



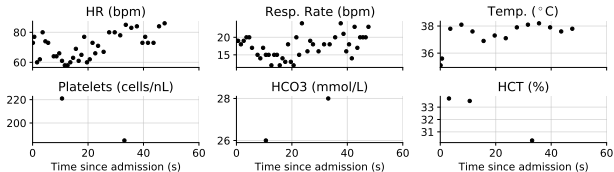
$\xrightarrow{\pi}$



- Climate model downscaling (Markou et al., 2022):

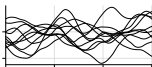
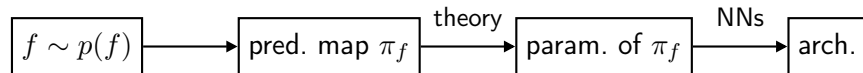
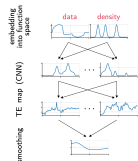
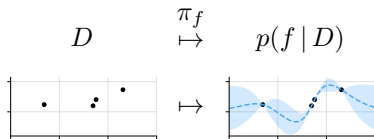


- ICU monitoring (Silva et al., 2012; Shysheya, 2020):



Today: Prediction Map Approximation

3/18



$$m(D) = \rho\left(\sum_{(x,y) \in D} \phi(x,y)\right)$$

- ✓ Theoretical framework
- ✓ Architectures with universal approximation properties
- ✓ Properties of $f \Rightarrow$ symmetries of $\pi_f \Rightarrow$ param. efficient archs!

Prediction Map Approximation

↙ e.g., a sawtooth wave

- Let f be some ground-truth stochastic process.
- **Posterior prediction map:** $\pi_f: \mathcal{D} \rightarrow \mathcal{P}$, $\pi_f(D) = p(f | D)$.
- **Goal:** find **Gaussian** approximation $\tilde{\pi}: \mathcal{D} \rightarrow \mathcal{P}_G$.
- **Approach:**

$$\tilde{\pi}(D) \in \arg \min_{\mu \in \mathcal{P}_G} \text{KL}(\pi_f(D), \mu).$$

- ✗ Approximate f and perform inference in approximation.
- ✓ Directly approximate posteriors of f .
- $\text{KL}(\mathcal{GP}(0, 1 \cdot e^{-|\cdot|}), \mathcal{GP}(0, \sigma^2 e^{-|\cdot|})) = \infty$ unless $\sigma^2 = 1$!
- If $\text{KL}(\pi_f(D), \mu_0) < \infty$ for some $\mu_0 \in \mathcal{P}_G$, then

$$\tilde{\pi}(D) = \pi_{\text{MM}}(D) := \mathcal{GP}(m_{f|D}, k_{f|D}).$$

- Practical objective:

if $f \sim \pi(D)$, then $(f(x_1), \dots, f(x_n)) \sim P_{\mathbf{x}}\pi(D)$

à la variational family $\longrightarrow \tilde{\pi} \in \arg \min_{\pi \in \mathcal{Q}} \mathcal{L}(\pi)$

$$\mathcal{L}(\pi) = \mathbb{E}_{p(D)p(\mathbf{x})} \text{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi(D))$$

$$\approx -\frac{1}{N} \sum_{n=1}^N \log q(D_n^{(t)} | D_n^{(c)}) := \mathcal{L}_n(\pi)$$

↑ density of $\pi(D_n^{(c)})$

- Call π **continuous** if $D_i \rightarrow D$ implies $\pi(D_i) \rightarrow \pi(D)$.
- Setting \mathcal{Q} to

$$\mathcal{M}_G = \{\pi : \mathcal{D} \rightarrow \mathcal{P}_G : \pi \text{ continuous}\},$$

minimiser exists, is unique, and coincides with original problem!

GPs without correlations,
 \Downarrow i.e. $k(x, x') = 0$ if $x \neq x'$

- For now, consider $\mathcal{Q}_{G, MF} = \{\pi: \mathcal{D} \rightarrow \mathcal{P}_{G, MF}\}$.
- Separately parametrise **mean map** and **variance map**:

$$m: \mathcal{D} \rightarrow C(\mathbb{R}, \mathbb{R}), \quad \sigma^2: \mathcal{D} \rightarrow C(\mathbb{R}, (0, \infty)).$$

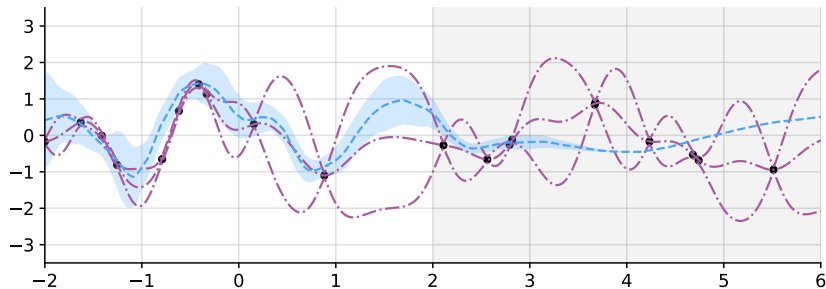
Thm (Zaheer et al., 2017; Wagstaff et al., 2019). A continuous function $f: \mathcal{D}_{\leq M} \rightarrow Z$ has the form of a **deep set**:

$$f(D) = \rho\left(\sum_{(x,y) \in D} \phi(x, y)\right)$$

where $\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^M$ and $\rho: \mathbb{R}^M \rightarrow Z$ are continuous.

- **Conditional neural process** (Garnelo et al., 2018):

$$\mathcal{L} + \mathcal{Q}_{G, MF} + \text{deep sets for } \pi = \text{CNP}$$

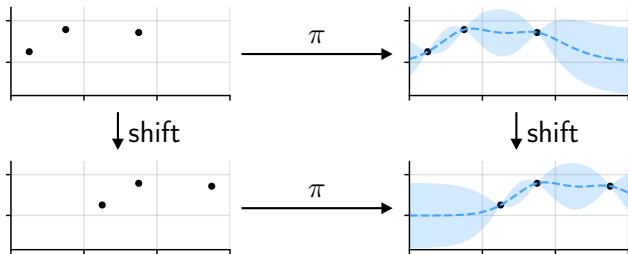


- ✗ Learns very slowly
- ✗ Underfits
- ✗ Generalises poorly

Exploiting Stationarity

- Let T_τ represent a translation by τ .
- A prediction map $\pi: \mathcal{D} \rightarrow \mathcal{P}$ is **translation equivariant (TE)** if

$$\pi(T_\tau D) = T_\tau \pi(D).$$



Prop (Foong et al., 2020). f is stationary $\iff \pi_f$ is TE.

Deep Set

(Zaheer et al., 2017)

$f: \mathcal{D}_{\leq M} \rightarrow Z$ is cont.

\iff

$\downarrow T_\tau?!$

encoder

$$E: \mathcal{D}_{\leq M} \rightarrow \mathbb{R}^M,$$

$$E(D) = \sum_{(x,y) \in D} \phi(x,y)$$

decoder

$$\rho: \mathbb{R}^M \rightarrow Z,$$

$$f(D) = \rho(E(D))$$

Convolutional Deep Set

(Gordon et al., 2020)

$f: \mathcal{D}_{\leq M} \rightarrow Z$ is cont. and TE

\iff

\downarrow functional embedding (RKHS)

$$E: \mathcal{D}_{\leq M} \rightarrow \mathbb{H},$$

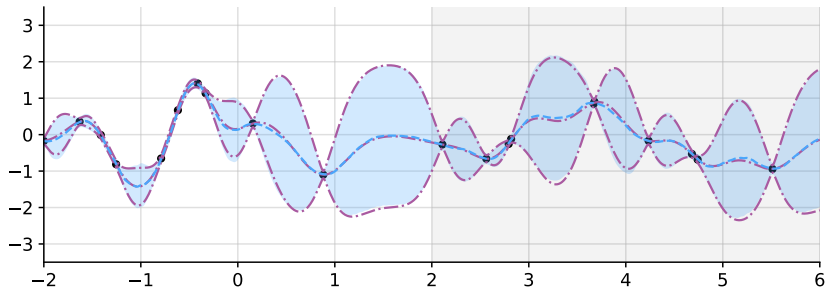
$$E(D) = \sum_{(x,y) \in D} k(\cdot - x)\phi(y)$$

TE map between function spaces $\rightarrow \rho: \mathbb{H} \rightarrow Z,$
 \approx CNN

$$f(D) = \rho(E(D))$$

- Gives convolutional CNP (Gordon et al., 2020):

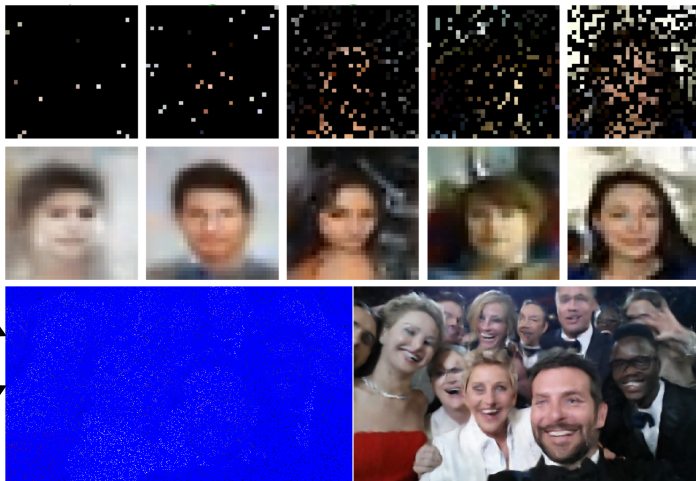
$$\mathcal{L} + \mathcal{Q}_{G, MF} + \text{conv. deep sets for } \pi = \text{ConvCNP}$$



- ✓ Learns pretty quickly
- ✓ Recovers target (diagonalised ground-truth GP)
- ✓ Generalises well

The Convolutional CNP (2)

11/18



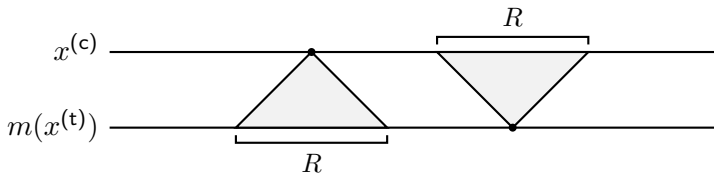
Gordon et al. (2020)

Why Does TE Help Generalise?

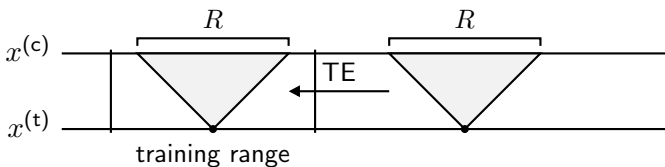
12/18

→ like $k(\tau) = 0$ for $|\tau| \geq \frac{1}{2}R$

- CNNs have **receptive field** $R > 0$:

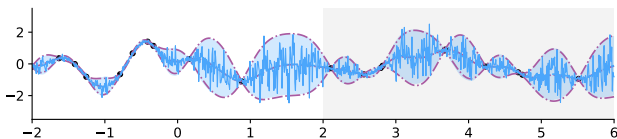


- In combination with TE, helps ConvCNP to generalise:



Thm. Suppose that $\mathcal{L}(\pi) \leq \varepsilon$ for data sampled from $[0, R + \ell]$. Then $\mathcal{L}(\pi) \leq \lceil M/\ell \rceil \varepsilon$ for data from any interval of width M .

Further Improvements



✗ (Conv)CNP fails to model correlations.

- $\mathcal{Q}_G = \{\pi: \mathcal{D} \rightarrow \mathcal{P}_G\}$ instead of $\mathcal{Q}_{G, MF} = \{\pi: \mathcal{D} \rightarrow \mathcal{P}_{G, MF}\}$?
- Bruinsma et al. (2021) establishes repr. thm for **kernel map**:

$$k: \mathcal{D} \rightarrow C^{\text{p.s.d.}}(\mathbb{R} \times \mathbb{R}, \mathbb{R})$$

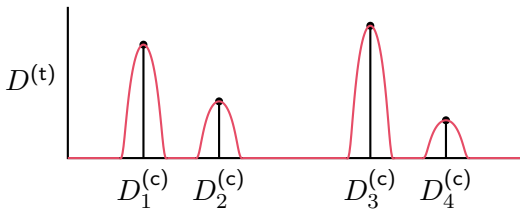
- ✓ Exploits TE using CNNs, learns quickly, and generalises well
- ✗ d -dimensional inputs require $2d$ -dimensional convolutions

- Markou et al. (2022) provide practical params for $d > 1$:

$$k(x, x', D) = \langle \mathbf{r}(x, D), \mathbf{r}(x', D) \rangle.$$

\uparrow TE

Consistency of Prediction Map Approximation



$$\mathcal{L}_n(\pi) = -\frac{1}{N} \sum_{n=1}^N \log q(D_n^{(t)} | D_n^{(c)}) \approx \frac{1}{N} \sum_{n=1}^N (D_n^{(t)} - f(D_n^{(c)}))^2$$

⇒ Cannot optimise $\mathcal{L}_n(\pi)$ over $\pi \in \mathcal{M}_G$: **overfitting!**

- **Practice:** tune NN capacity using black magic.
- Will show that we can reasonably restrict to **compact** $\mathcal{Q} \subset \mathcal{M}_G$.

Let $\mathcal{D} \subseteq \bigcup_{n=0}^{\infty} (\mathcal{X} \times \mathbb{R})^n$ be a collection of data sets of interest.

Assumptions:

- \mathcal{X} is compact.
- There exist $p \geq 2$, $q > 1$, $c > 0$, and $r > 0$ such that

$$\mathbb{E}[|f(x) - f(y)|^p] \leq c|x - y|^q \quad \text{whenever} \quad |x - y| < r.$$

- \mathcal{D} is bounded: $\|\mathcal{D}\| := \sup \{|\mathbf{x}| \vee \|\mathbf{y}\|_{\infty} : (\mathbf{x}, \mathbf{y}) \in \mathcal{D}\} < \infty$.
- $M := \sup_{x \in \mathcal{X}} [|f(x)|^{2+\gamma}] < \infty$ for some $\gamma > 0$.
- Observations under Gaussian noise with $\sigma^2 \in [\underline{\sigma}^2, \bar{\sigma}^2]$.

- Identify every $\pi \in \mathcal{M}_G$ with

$$m: \mathcal{X} \times \mathcal{D} \rightarrow \mathbb{R}, \quad k: \mathcal{X} \times \mathcal{X} \times \mathcal{D} \rightarrow \mathbb{R}, \quad \sigma^2 \in [\underline{\sigma}^2, \bar{\sigma}^2].$$

- Then exist $L^*: [0, \infty)^2 \rightarrow [0, \infty)$ and $M^* > 0$ such that

$$\pi_{\text{MM}} \in \left\{ \pi \in \mathcal{M}_G \left| \begin{array}{l} |m(x_1, D_1) - m(x_2, D_2)| \leq L^*(|x_1 - x_2|, \|D_1 - D_2\|) \\ |k(x_1, D_1) - k(x_2, D_2)| \leq L^*(|x_1 - x_2|, \|D_1 - D_2\|) \\ \|m\|_\infty, \|k\|_\infty \leq M^* \end{array} \right. \right\}.$$

- Call this collection \mathcal{Q}^* . Define a metric on \mathcal{Q}^* :

$$d(\pi_1, \pi_2) = \|m_1 - m_2\|_\infty + \|k_1 - k_2\|_\infty + |\sigma_1^2 - \sigma_2^2|.$$

- Arzelà–Ascoli theorem: (\mathcal{Q}^*, d) is compact.

Thm. Let

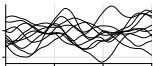
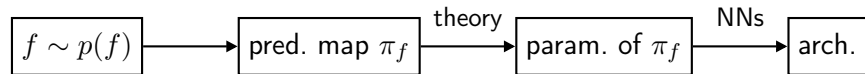
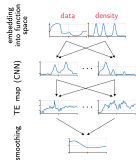
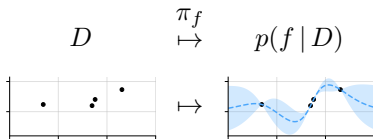
$$\pi_n \in \arg \min_{\pi \in \mathcal{Q}^*} \mathcal{L}_n(\pi), \quad \mathcal{L}_n(\pi) = -\frac{1}{N} \sum_{n=1}^N \log q(D_n^{(t)} | D_n^{(c)}).$$

Then, almost surely, $\pi_n(D) \rightarrow \pi_{\text{MM}}(D)$ for all $D \in \mathcal{D}$.

Pending questions:

- $\mathcal{Q}_{\text{NN}} = \{(m_\theta, k_\theta, \sigma^2) : \theta \in \mathbb{R}^P\}$?
- **How much data:** finite-sample bounds / rates of convergence?

Wrapping Up



$$m(D) = \rho\left(\sum_{(x,y) \in D} \phi(x,y)\right)$$

- ✓ Theoretical framework
- ✓ Architectures with universal approximation properties
- ✓ Properties of $f \Rightarrow$ symmetries of $\pi_f \Rightarrow$ param. efficient archs!

These slides: <https://wesselb.github.io/pdf/predmap>.

Appendix

References

- Bruinsma, Wessel P., James Requeima, Andrew Y. K. Foong, Jonathan Gordon, and Richard E. Turner (2021). “The Gaussian Neural Process”. In: *Proceedings of the 3rd Symposium on Advances in Approximate Bayesian Inference*. eprint: <https://arxiv.org/abs/2101.03606>.
- Foong, Andrew Y. K., Wessel P. Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard E. Turner (2020). “Meta-Learning Stationary Stochastic Process Prediction With Convolutional Neural Processes”. In: *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc. eprint: <https://arxiv.org/abs/2007.01332>.

References (2)

- Garnelo, M., D. Rosenbaum, C. J. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. Whye Teh, D. J. Rezende, and S. M. A. Eslami (2018). “Conditional Neural Processes”. In: *Proceedings of 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR. eprint: <https://arxiv.org/abs/1807.01613>.
- Gordon, Jonathan, Wessel P. Bruinsma, Andrew Y. K. Foong, James Requeima, Yann Dubois, and Richard E. Turner (2020). “Convolutional Conditional Neural Processes”. In: *Proceedings of the 8th International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=Skey4eBYP5>.
- Markou, Stratis, James Requeima, Wessel P. Bruinsma, and Richard E. Turner (2022). “Practical Conditional Neural Processes for Tractable Dependent Predictions”. In: *Proceedings of the 10th International Conference on Learning Representations*.

References (3)

- Shysheya, Aliaksandra (2020). “Neural Models for Non-Uniformly Sampled Data”. MA thesis. Department of Engineering, University of Cambridge.
- Silva, Ikaro, George Moody, Daniel J. Scott, Leo A. Celi, and Roger G. Mark (2012). “Predicting In-Hospital Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012”. In: *Computing in Cardiology* 39, pp. 245–248.
- Wagstaff, E., F. B. Fuchs, M. Engelcke, I. Posner, and M. Osborne (2019). “On the Limitations of Representing Functions on Sets”. In: *Proceedings of 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR. eprint: <https://arxiv.org/abs/1901.09006>.

References (4)

Zaheer, M., S. Kottur, S. Ravanbakhsh, B. Póczos, R. Salakhutdinov, and A. Smola (2017). “Deep Sets”. In: *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. eprint: <https://arxiv.org/abs/1703.06114>.