Meta-Learning as Prediction Map Approximation

Wessel Bruinsma

University of Cambridge and Invenia Labs

Job Research Talk at Microsoft Research, 6 Jan 2022

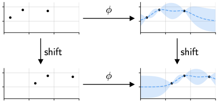
Hello! 👋

- PhD student at Cambridge MLG, supervised by Rich Turner.
- Researcher at Invenia Labs.

- MPhil in MLMI (Cam).
- BSc in EE (Delft).

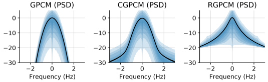
Neural processes:

- equivariance,
- correlated predictions



Gaussian processes:

- multiple outputs,
- nonpar. kernel priors



Theory:

- MFVI BNNs $\Rightarrow$ prior,
- PAC Bayes

$$
\begin{aligned}
&\|\mathbb{E}_Q[\mathbf{f}(\mathbf{x})] - \mathbb{E}_P[\mathbf{f}(\mathbf{x})]\|_2 \\
&\quad \le c_1 c_2^{L-1} \frac{1 + \frac{1}{\sqrt{D_{\mathrm{i}}}}\|\mathbf{x}\|_2}{\sqrt{M}} \mathrm{KL}(Q,P)(\mathrm{KL}(Q,P)^{\frac{L-1}{2}} \vee 1), \\
&\|\mathbb{E}_Q[\mathbf{f}^2(\mathbf{x})] - \mathbb{E}_P[\mathbf{f}^2(\mathbf{x})]\|_\infty \\
&\quad \le c_3 c_4^{L-1} \frac{1 + \frac{1}{D_{\mathrm{i}}}\|\mathbf{x}\|_2^2}{\sqrt{M}} \mathrm{KL}(Q,P)(\mathrm{KL}(Q,P)^{L-1} \vee 1)
\end{aligned}
$$

Hello! 👋 (2)

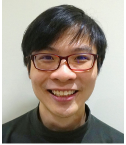
- Stheno (`github.com/wesselb/stheno`)
- Plum (`github.com/wesselb/plum`)
- Matrix (`github.com/wesselb/matrix`)

```
>>> from matrix import Diagonal, LowRank, Woodbury, Kronecker, LowerTriangular, ...

>>> d = Diagonal(np.array([1, 2, 3]))

>>> B.inv(d + 1)
<Woodbury matrix: batch=(), shape=(3, 3), dtype=float64
 diag=<diagonal matrix: batch=(), shape=(3, 3), dtype=float64
       diag=[1.    0.5   0.333]>
 lr=<low-rank matrix: batch=(), shape=(3, 3), dtype=float64, rank=1
                 ...
                 [0.333]]>>>

>>> B.inv(B.inv(d + 1)) - 1
<diagonal matrix: batch=(), shape=(3, 3), dtype=float64
 diag=[1. 2. 3.]>
```

Collaborators


Wessel Bruinsma


Jonathan Gordon


Andrew Foong


James Requeima

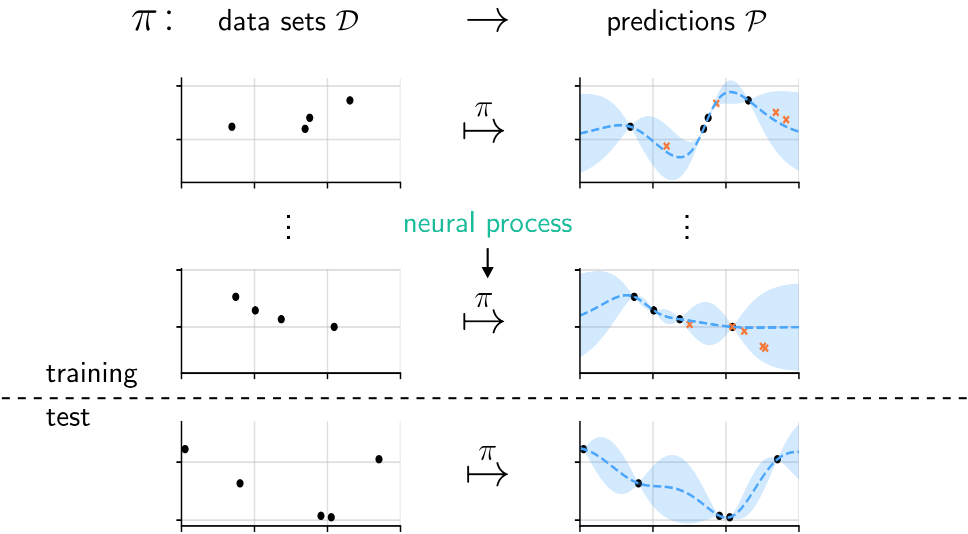

Stratis Markou


Yann Dubois


Anna Vaughan

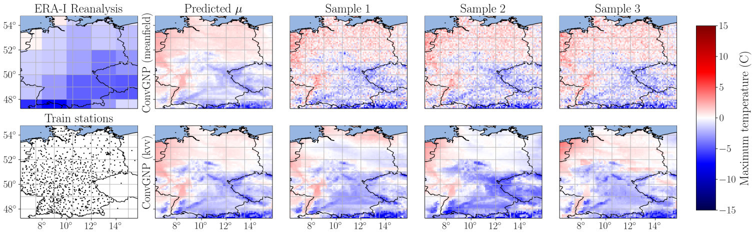

Rich Turner

Meta-Learning and Neural Processes

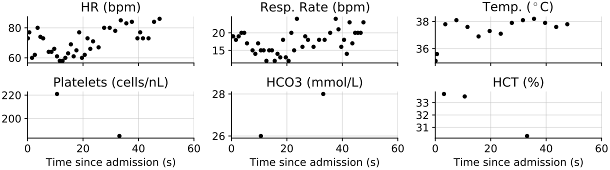


Applications of Neural Processes

- Climate model downscaling (Markou et al., 2021):



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



Applications of Neural Processes (2)

- Potential energy surface prediction:

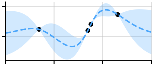
```
C    0.0072 -0.5687  0.0
C   -1.2854  0.2499  0.0
O    1.1304  0.3147  0.0
H    0.0392 -1.1972  0.89
H    0.0392 -1.1972 -0.89
H   -1.3175  0.8784  0.89
H   -1.3175  0.8784 -0.89
H   -2.1422 -0.4239  0.0
H    1.9857 -0.1365  0.0
```

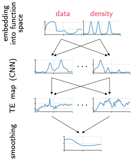
$$\mapsto \quad U_{\mathsf{BO}}(\mathbf{r}_1, \ldots, \mathbf{r}_{n_{\mathsf{atoms}}}) \text{ and/or } (\mathbf{f}_1, \ldots, \mathbf{f}_{n_{\mathsf{atoms}}})$$

1. Variable number and permutation invariance of atoms
2. Forces are conversative
3. Rototranslation invariance (energy) or equivariance (forces)

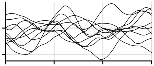
Today: Prediction Map Approximation

$$D \quad \overset{\pi_f}{\mapsto} \quad p(f \mid D)$$

 $\mapsto$ 



$$\boxed{f \sim p(f)} \longrightarrow \boxed{\text{pred. map } \pi_f} \xrightarrow{\text{theory}} \boxed{\text{param. of } \pi_f} \xrightarrow{\text{NNs}} \boxed{\text{arch.}}$$



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

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

Prediction Map Approximation

Setup

e.g., a sawtooth wave

- Let f be some ground-truth stochastic process.
- Posterior prediction map: $\pi_f\colon \mathcal{D} \to \mathcal{P}$, $\pi_f(D) = p(f \mid D)$.
- Goal: approximate π_f with $\tilde{\pi} \in \mathcal{Q}$ (à la variational family).
- Approach: minimise loss function:

$$\tilde{\pi} \in \operatorname*{arg\,min}_{\pi \in \mathcal{Q}} \mathcal{L}_{\mathrm{Div}}(\pi), \quad \mathcal{L}_{\mathrm{Div}}(\pi) = \sup_{D \in \mathcal{D}} \mathrm{Div}(\pi_f(D), \pi(D)).$$

some divergence; *e.g.*, $\mathrm{Div} = \mathrm{KL}$

- Regularity of π_f? Existence of $\tilde{\pi}$?
  - $\mathrm{KL}(\mathcal{GP}(0, 1 \cdot e^{-|\cdot|}), \mathcal{GP}(0, \sigma^2 e^{-|\cdot|})) = \infty$ unless $\sigma^2 = 1$!

✗ Approximate f and perform inference in approximation.

✓ Directly approximate posteriors of f.

Parametrising Prediction Maps

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

- For now, consider $\mathcal{Q}_{\mathsf{G,MF}} = \{\pi \colon \mathcal{D} \to \mathcal{P}_{\mathsf{G,MF}}\}$.
- How do we parametrise a $\pi \colon \mathcal{D} \to \mathcal{P}_{\mathsf{G,MF}}$?

$\Rightarrow$ Separately parametrise mean map and variance map:

$$m \colon \mathcal{D} \to C(\mathbb{R}, \mathbb{R}), \qquad \sigma^2 \colon \mathcal{D} \to C(\mathbb{R}, (0, \infty)).$$

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

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

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

A Practical Objective

- Appealing objective, but theoretical issues:

$$\mathcal{L}_{\mathsf{KL}}(\pi) = \sup_{D \in \mathcal{D}} \mathrm{KL}(\pi_f(D), \pi(D)).$$

- Practical objective (Garnelo, Rosenbaum, et al., 2018; Gordon, Bronskill, et al., 2019):

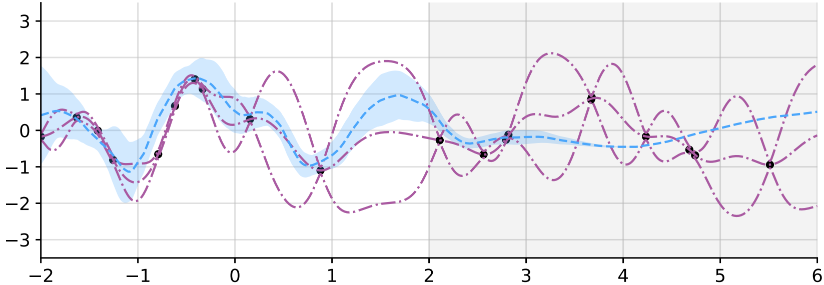
density of $\pi(D^{(\mathsf{c})})$

$$\begin{aligned}\mathcal{L}_{\mathsf{ML}}(\pi) &= \mathbb{E}_{p(D)}[\log q(D^{(\mathsf{t})} \mid D^{(\mathsf{c})})] \\ &\approx \tfrac{1}{M} \textstyle\sum_{m=1}^{M} \log q(D_m^{(\mathsf{t})} \mid D_m^{(\mathsf{c})}).\end{aligned}$$

- Under conditions, minimisers coincide (Bruinsma et al., 2021)!
- Conditional neural process (Garnelo, Rosenbaum, et al., 2018):

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

The Conditional Neural Process



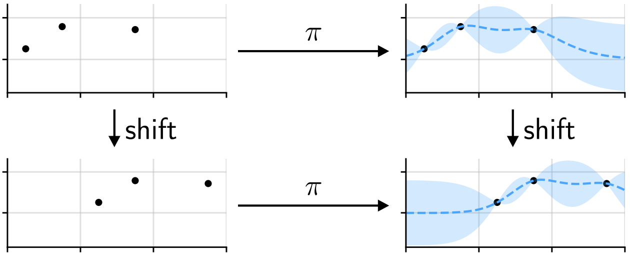
- ✗ Learns very slowly
- ✗ Underfits
- ✗ Generalises poorly

Exploiting Stationarity

Translation-Equivariant Prediction Maps

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

$$\pi(\mathsf{T}_\tau D) = \mathsf{T}_\tau \pi(D).$$



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

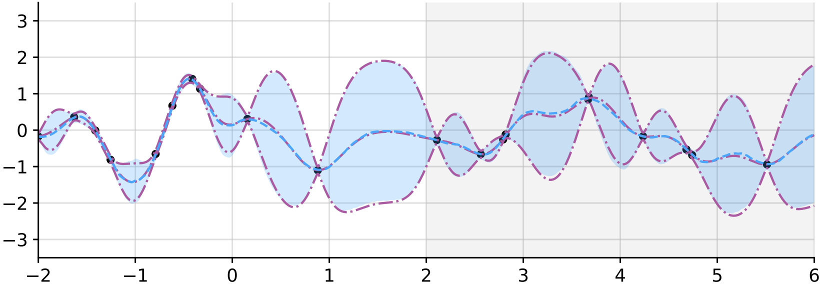
Parametrising TE Prediction Maps

	Deep Set (Zaheer et al., 2017)		Convolutional Deep Set (Gordon, Bruinsma, et al., 2020)
	$f\colon \mathcal{D}_{\le M} \to Z$ is cont.		$f\colon \mathcal{D}_{\le M} \to Z$ is cont. and TE
	$\iff$		$\iff$
	T_τ?!		functional embedding (RKHS)
encoder	$E\colon \mathcal{D}_{\le M} \to \mathbb{R}^M$, $E(D) = \sum_{(x,y)\in D} \phi(x, y)$		$E\colon \mathcal{D}_{\le M} \to \mathbb{H}$, $E(D) = \sum_{(x,y)\in D} k(\bullet - x)\phi(y)$
decoder	$\rho\colon \mathbb{R}^M \to Z$, $f(D) = \rho(E(D))$	TE map between function spaces $\approx$ CNN	$\rho\colon \mathbb{H} \to Z$, $f(D) = \rho(E(D))$

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

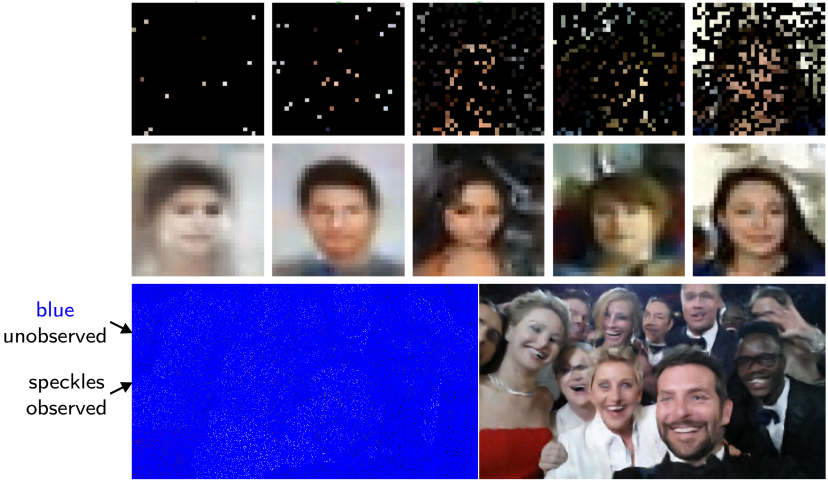
$$\boxed{\mathcal{Q}_{\text{G,MF}} + \text{conv. deep sets for } \pi + \mathcal{L}_{\text{ML}} = \text{ConvCNP}}$$

The Convolutional CNP



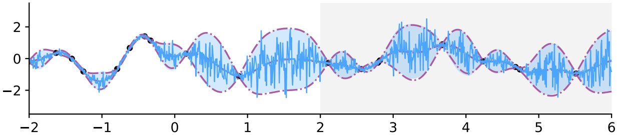
- ✓ Learns pretty quickly
- ✓ Recovers target (diagonalised ground-truth GP)
- ✓ Generalises well

The Convolutional CNP (2)



Gordon, Bruinsma, et al. (2020)

Further Improvements



- ✗ (Conv)CNP fails to model correlations.
- $\mathcal{Q}_{\mathsf{G}} = \{\pi \colon \mathcal{D} \to \mathcal{P}_{\mathsf{G}}\}$ instead of $\mathcal{Q}_{\mathsf{G,MF}} = \{\pi \colon \mathcal{D} \to \mathcal{P}_{\mathsf{G,MF}}\}$?
- Bruinsma et al. (2021) establishes repr. thm for kernel map:

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

  - ✓ Exploits TE using CNNs, learns quickly, and generalises well
  - ✗ D-dimensional inputs require $2D$-dimensional convolutions
- Markou et al. (2021) provide practical params for $D > 1$.
  - ✓ Approximation with TE basis functions works well

Future Directions (For This Project)

non-Gaussian $\mathcal{Q}$

- Combination of latent variable with correlated prediction maps
- Generalisation to other symmetry groups
  - Kawano et al. (2021) and Holderrieth et al. (2021) extend ConvCNP construction to more general symmetries
- Approximate symmetries:

Conjecture. Under conditions, a cont. function $f\colon \mathcal{D}_{\leq M} \to Z$ is approximately an augmented G-conv. deep set:

$$f(D) \approx \rho\big(\textstyle\sum_{(x,y)\in D} k(\,\bullet\,, x)\phi(y), h_1, \ldots, h_Q\big)$$

where

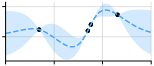
- $\phi\colon \mathbb{R} \to \mathbb{R}^{K+1}$ is continuous,
- $k\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a G-invariant kernel,
- $h_1, \ldots, h_Q \in \mathbb{H}$ are additional channels, and
- $\rho\colon \mathbb{H} \to Z$ is continuous and G-equivariant.

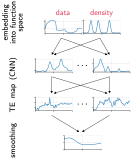
Wrapping Up

Prediction Map Approximation

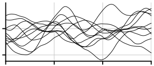
$$D \quad \overset{\pi_f}{\mapsto} \quad p(f \mid D)$$

$\mapsto$





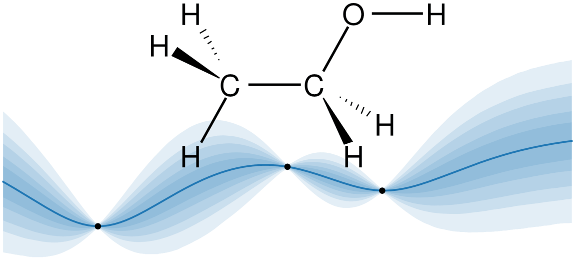
$f \sim p(f)$ → pred. map π_f $\xrightarrow{\text{theory}}$ param. of π_f $\xrightarrow{\text{NNs}}$ arch.



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

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

Wesel at Microsoft Research



- Theoretical background and experience in ML
- Practical experience in software engineering (Invenia Labs, OSS)
- Very keen to apply ML to scientific problems!

These slides: `https://wesselb.github.io/pdf/msr`.

Appendix

Setup: Example

collection of all GPs

- Consider $\mathcal{Q} = \{\pi\colon \mathcal{D} \to \mathcal{P}_{\mathsf{G}}\}$, $\mathrm{D} = \mathrm{KL}$, and f a GP mixture:

$$f \mid z \sim \mathcal{GP}(0, k_z), \quad z \sim \mathrm{Cat}(\pi).$$

✗ Prior approx. then inference: condition $\sum_i \pi_i f_i$ on D.

✓ Posterior approximation: $\sum_i p(z = i \mid D) f_{i,D}$.

retain inference of which component!

Non-Gaussian TE Prediction Maps

- Neural process (Garnelo, Schwarz, et al., 2018) uses latent variable $\mathbf{z}$ to model correlations: $\mathcal{Q}$ consists of maps π such that

$$f \sim \pi(D) \quad \iff \quad \mathbf{z} \sim q(\mathbf{z} \,|\, D), \quad f \,|\, \mathbf{z} \sim p(f \,|\, \mathbf{z}).$$

  ✓ Non-Gaussian and models correlations
  ✗ $\mathcal{L}_{\mathsf{ML}}$ now intractable
  ✗ Two ways to approximate $f \,|\, D^{(\mathsf{c})}, D^{(\mathsf{t})}$:

$$q(\mathbf{z} \,|\, D^{(\mathsf{c})}, D^{(\mathsf{t})}) \quad \text{and} \quad \tfrac{1}{Z} q(\mathbf{z} \,|\, D^{(\mathsf{c})}) p(D^{(\mathsf{t})} \,|\, \mathbf{z})$$

- NP objective tackles both simultaneously (Foong et al., 2020):

$$\mathcal{L}_{\mathsf{NP}}(\pi) = \underbrace{\mathcal{L}_{\mathsf{ML}}(\pi)}_{\approx \log p(\mathbf{y})} \underbrace{- \mathbb{E}_{p(D)}[\mathrm{KL}(q(\mathbf{z} \,|\, D^{(\mathsf{c})}, D^{(\mathsf{t})}), \tfrac{1}{Z} q(\mathbf{z} \,|\, D^{(\mathsf{c})}) p(D^{(\mathsf{t})} \,|\, \mathbf{z}))]}_{\approx \mathrm{KL}(q(\mathbf{z}), p(\mathbf{z} \,|\, \mathbf{y})),\ \text{encourages Bayes' consistency}}$$

Non-Gaussian TE Prediction Maps (2)

- Latent variable + ConvCNP = ConvNP (Foong et al., 2020)

✗ All constructions of NPs use mean-field approximation!

- Promising future direction:

$\boxed{\text{Combine latent variables with Gaussian TE prediction maps.}}$

  ✓ Improved uncertainty estimates
  ✓ Enables approximation of models like Bernoulli random fields

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`.

Garnelo, M., J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, S. M. A. Eslami, and Y. Whye Teh (2018). "Neural Processes". In: *Proceedings of 35th International Conference on Machine Learning. Theoretical Foundations and Applications of Deep Generative Models Workshop*. eprint: `https://arxiv.org/abs/1807.01622`.

Gordon, Jonathan, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E. Turner (2019). "Meta-Learning Probabilistic Inference for Prediction". In: *Proceedings of the 7th International Conference on Learning Representations*. eprint: `https://arxiv.org/abs/1805.09921`.

References (3)

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=Skey4eBYPS`.

Holderrieth, Peter, Michael Hutchinson, and Yee Whye Teh (2021). "Equivariant Learning of Stochastic Fields: Gaussian Processes And Steerable Conditional Neural Processes". In: *Proceedings of 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR. eprint: `https://arxiv.org/abs/2011.12916`.

Kawano, Makoto, Wataru Kumagai, Akiyoshi Sannai, Yusuke Iwasawa, and Yutaka Matsuo (2021). "Group Equivariant Conditional Neural Processes". In: *Proceedings of the 9th International Conference on Learning Representations*. URL: `https://openreview.net/forum?id=e8W-hsu_q5`.

References (4)

Markou, Stratis, James Requeima, Wessel P. Bruinsma, and Richard E. Turner (2021). *Practical Conditional Neural Processes for Tractable Dependent Predictions*. Under review for ICLR 2022.

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 (5)

Zaheer, M., S. Kottur, S. Ravanbakhsh, B. Poczos, 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`.