

Meta-Learning as Prediction Map Approximation

Wessel Bruinsma

University of Cambridge and Invenia Labs

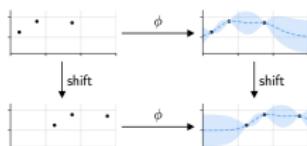
Job Research Talk at Microsoft Research, 6 Jan 2022

- PhD student at Cambridge MLG, supervised by Rich Turner.
- Researcher at Invenia Labs.
- MPhil in MLMI (Cam).
- BSc in EE (Delft).



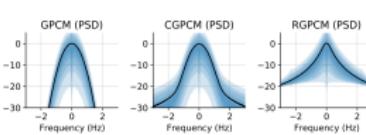
Neural processes:

- equivariance,
- correlated predictions



Gaussian processes:

- multiple outputs,
- nonpar. kernel priors



Theory:

- MFVI BNNs \Rightarrow prior,
- PAC Bayes

$$\begin{aligned} & \|\mathbb{E}_Q[f(\mathbf{x})] - \mathbb{E}_P[f(\mathbf{x})]\|_2 \\ & \leq c_1 c_2^{L-1} \frac{1 + \frac{1}{\sqrt{D_1}} \|\mathbf{x}\|_2}{\sqrt{M}} \text{KL}(Q, P)(\text{KL}(Q, P))^{\frac{L-1}{2}} \vee 1, \\ & \|\mathbb{E}_Q[f^2(\mathbf{x})] - \mathbb{E}_P[f^2(\mathbf{x})]\|_\infty \\ & \leq c_3 c_4^{L-1} \frac{1 + \frac{1}{D_1} \|\mathbf{x}\|_2^2}{\sqrt{M}} \text{KL}(Q, P)(\text{KL}(Q, P)^{L-1} \vee 1) \end{aligned}$$

- 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

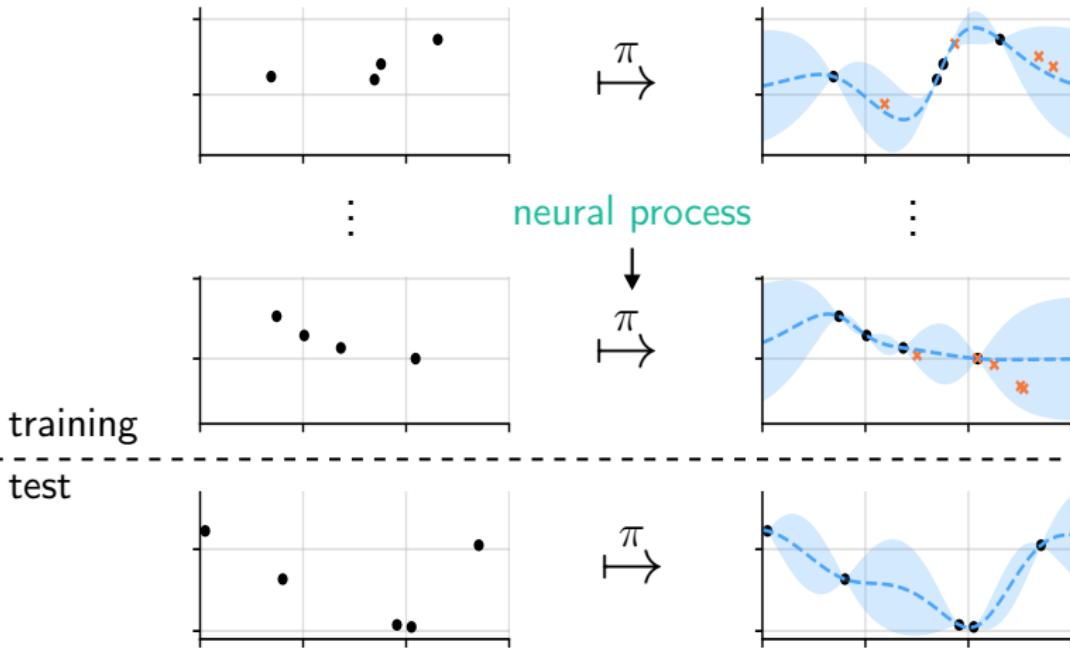


Anna
Vaughan



Rich
Turner

$\pi : \text{data sets } \mathcal{D} \rightarrow \text{predictions } \mathcal{P}$



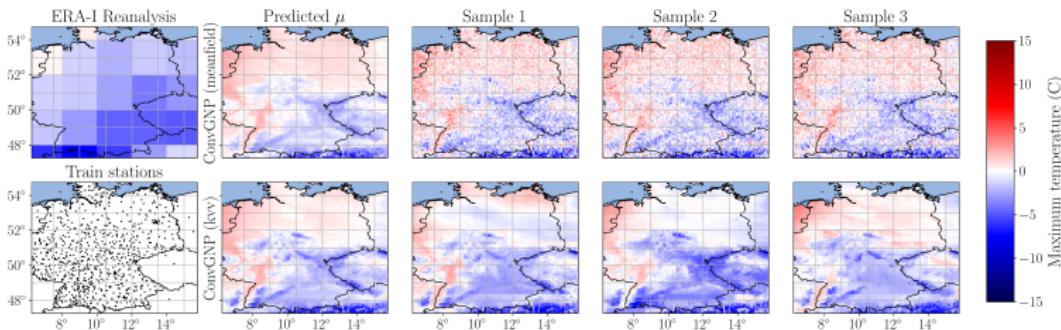
Examples of Neural Processes

4/17

- Image completion (Gordon, Bruinsma, et al., 2020):



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



- Potential energy surface prediction:

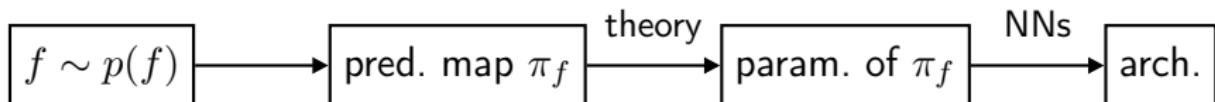
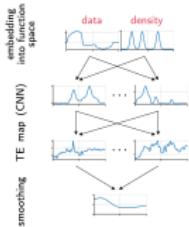
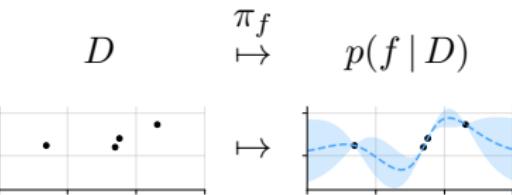
C	0.0072	-0.5687	0.0
C	-1.2854	0.2499	0.0
O	1.1304	0.3147	0.0
H	0.0392	-1.1972	0.89
H	0.0392	-1.1972	-0.89
H	-1.3175	0.8784	0.89
H	-1.3175	0.8784	-0.89
H	-2.1422	-0.4239	0.0
H	1.9857	-0.1365	0.0

$$\begin{array}{c} U_{\text{BO}}(\mathbf{r}_1, \dots, \mathbf{r}_{n_{\text{atoms}}}) \\ \mapsto \quad \text{and/or} \\ (\mathbf{f}_1, \dots, \mathbf{f}_{n_{\text{atoms}}}) \end{array}$$

- ① Variable number and permutation invariance of atoms
- ② Energy conservation of forces
- ③ Rototranslation invariance (energy) or equivariance (forces)

Today: Prediction Map Approximation

6/17



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

- ✓ 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: approximate π_f with $\tilde{\pi} \in \mathcal{Q}$ (à la variational family).
- Approach: Minimise loss function:

$$\tilde{\pi} \in \arg \min_{\pi \in \mathcal{Q}} \mathcal{L}_{\text{Div}}(\pi), \quad \mathcal{L}_{\text{Div}}(\pi) = \sup_{D \in \mathcal{D}} \text{Div}(\pi_f(D), \pi(D)).$$

↑ some divergence;
e.g., $\text{Div} = \text{KL}$

- Regularity of π_f ? Existence of $\tilde{\pi}$?
 - $\text{KL}(\mathcal{GP}(0, 1 \cdot e^{-|\cdot|}), \mathcal{GP}(0, \sigma^2 e^{-|\cdot|})) = \infty$ unless $\sigma^2 = 1$!
- ✗ Approximate f and perform inference in approximation.
- ✓ Directly approximate posteriors of f .

collection of all GPs without correlations,

i.e. $k(x, x') = 0$ if $x \neq x'$



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

- Appealing objective, but theoretical issues:

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

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

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

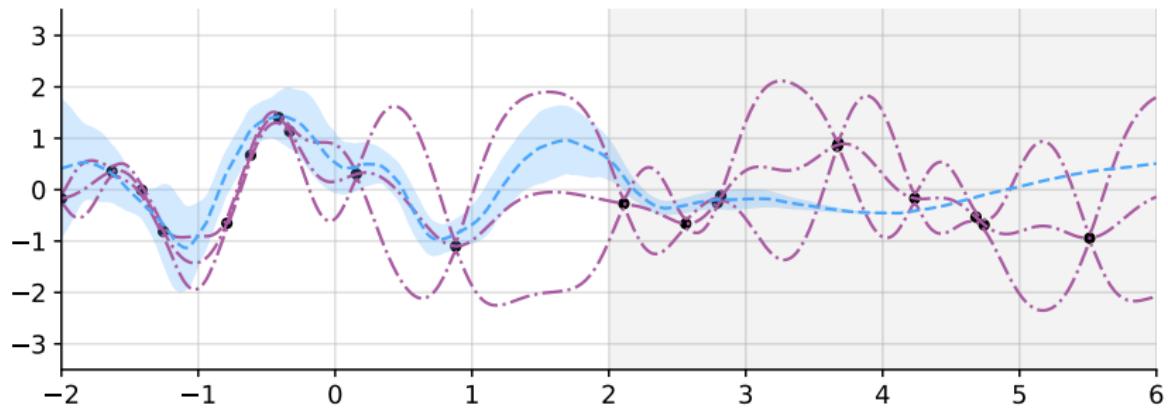
\downarrow density of $\pi(D^{(\text{c})})$

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

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

The Conditional Neural Process

10/17



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

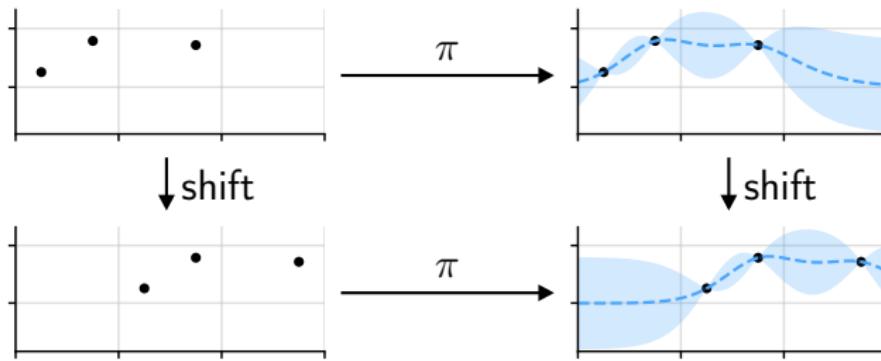
Exploiting Stationarity

Translation-Equivariant Prediction Maps

11/17

- 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, Bruinsma, et al., 2020)

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

\iff

\downarrow functional embedding (RKHS)

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

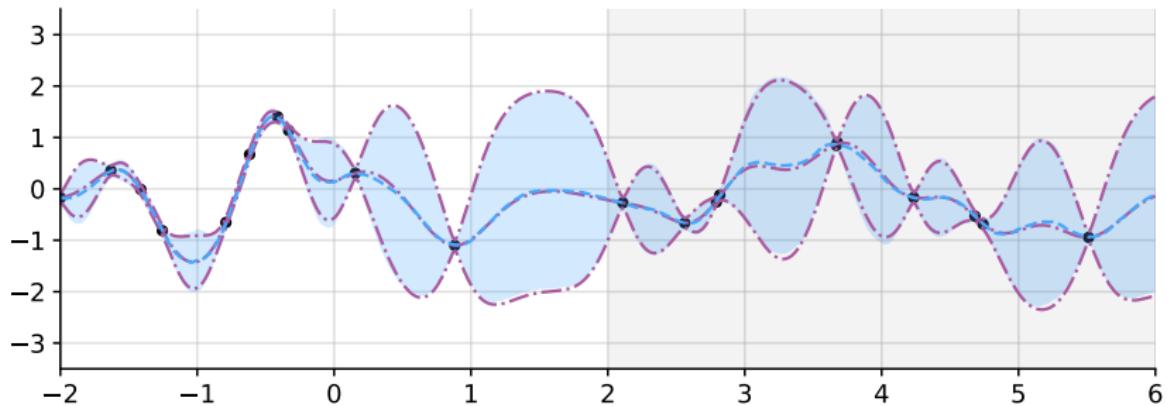
$$E(D) = \sum_{(x,y) \in D} k(\cdot - x)\phi(y)$$

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

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

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

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



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

Further Improvements

✗ (Conv)CNP fails to model correlations.

- $\mathcal{Q} = \{\pi: \mathcal{D} \rightarrow \mathcal{P}_G\}$ instead of $\mathcal{Q} = \{\pi: \mathcal{D} \rightarrow \mathcal{P}_{G, MF}\}$?
- 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. (2021) provide practical params for $D > 1$.
 - ✓ Approximation with TE basis functions works well

→ 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 $f: \mathcal{D}_{\leq M} \rightarrow Z$ is approximately an augmented G -conv. deep set:

$$f(D) \approx \rho \left(\sum_{(x,y) \in D} k(\cdot, x)\phi(y), h_1, \dots, h_Q \right)$$

where

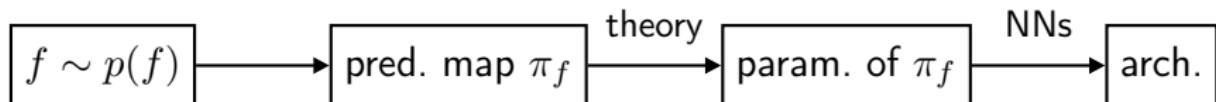
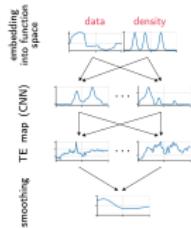
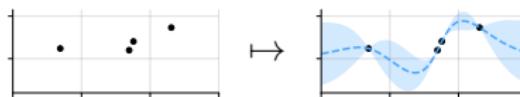
- $\phi: \mathbb{R} \rightarrow \mathbb{R}^{K+1}$ is continuous,
- $k: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is a G -invariant kernel,
- $h_1, \dots, h_Q \in \mathbb{H}$ are additional channels, and
- $\rho: \mathbb{H} \rightarrow Z$ is continuous and G -equivariant.

Wrapping Up

Prediction Map Approximation

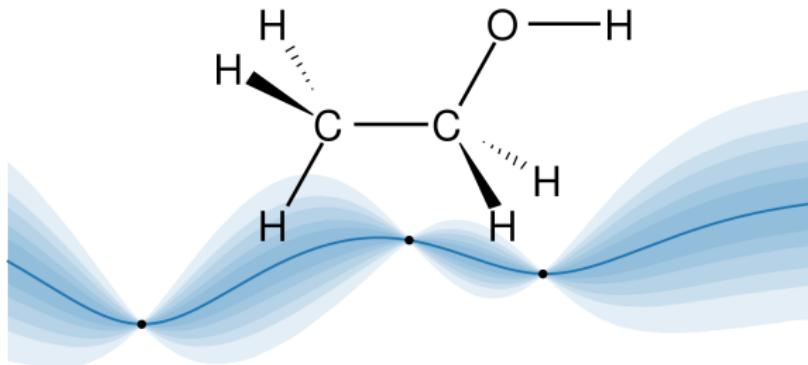
16/17

$$D \xrightarrow{\pi_f} p(f | D)$$



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

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



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

Appendix

Setup: Example

collection of all GPs



- Consider $\mathcal{Q} = \{\pi: \mathcal{D} \rightarrow \mathcal{P}_G\}$, $D = KL$, and f a GP mixture:

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

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

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



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

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

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

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

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

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

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:

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=Skey4eBYPSP>.
- 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.
- 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>.
- 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>.