Meta-Learning as Prediction Map Approximation

Wessel Bruinsma

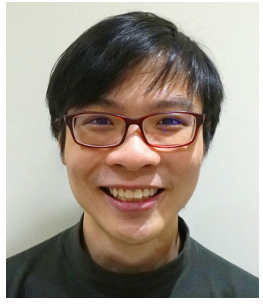
University of Cambridge and Invenia Labs

Research Talk at Sheffield Machine Learning Group, 24 Feb 2022

Collaborators

Wessel Bruinsma

Jonathan Gordon



Andrew Foong

James Requeima

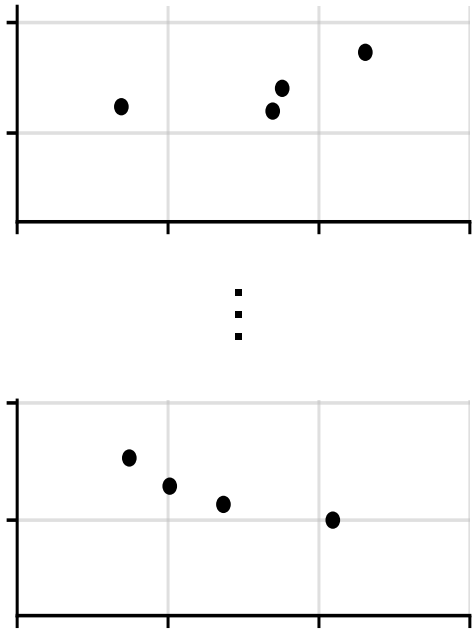
Stratis Markou

Anna Vaughan

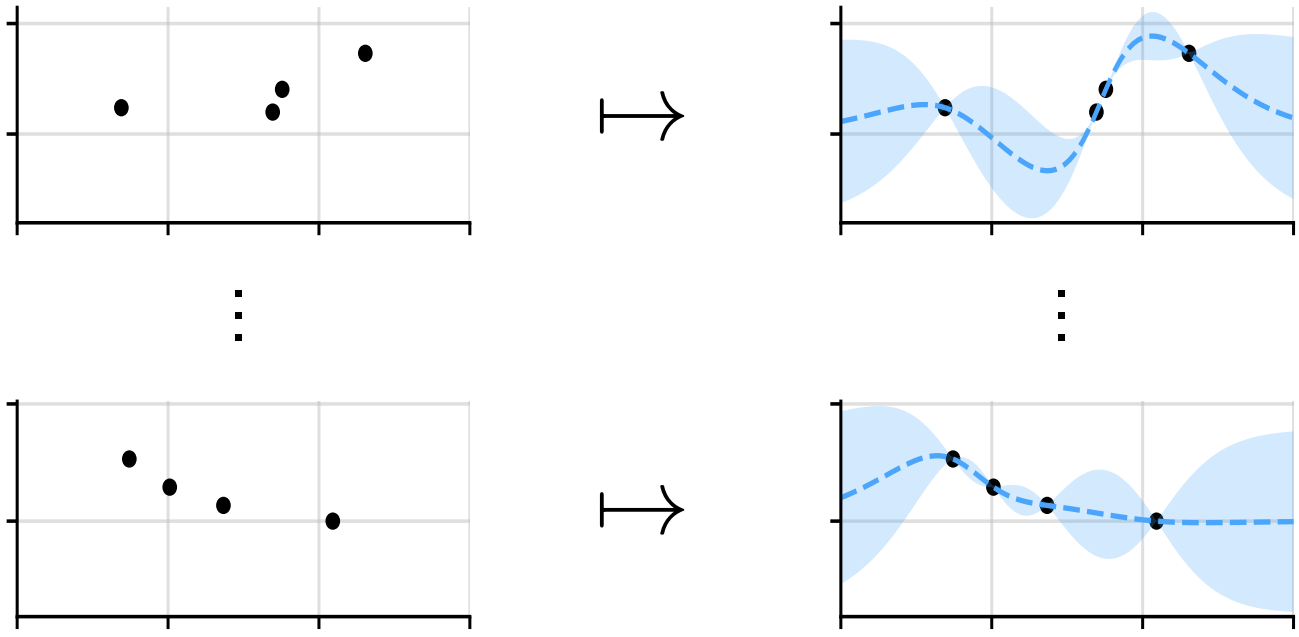
Yann Dubois

Rich Turner

Meta-Learning and Neural Processes

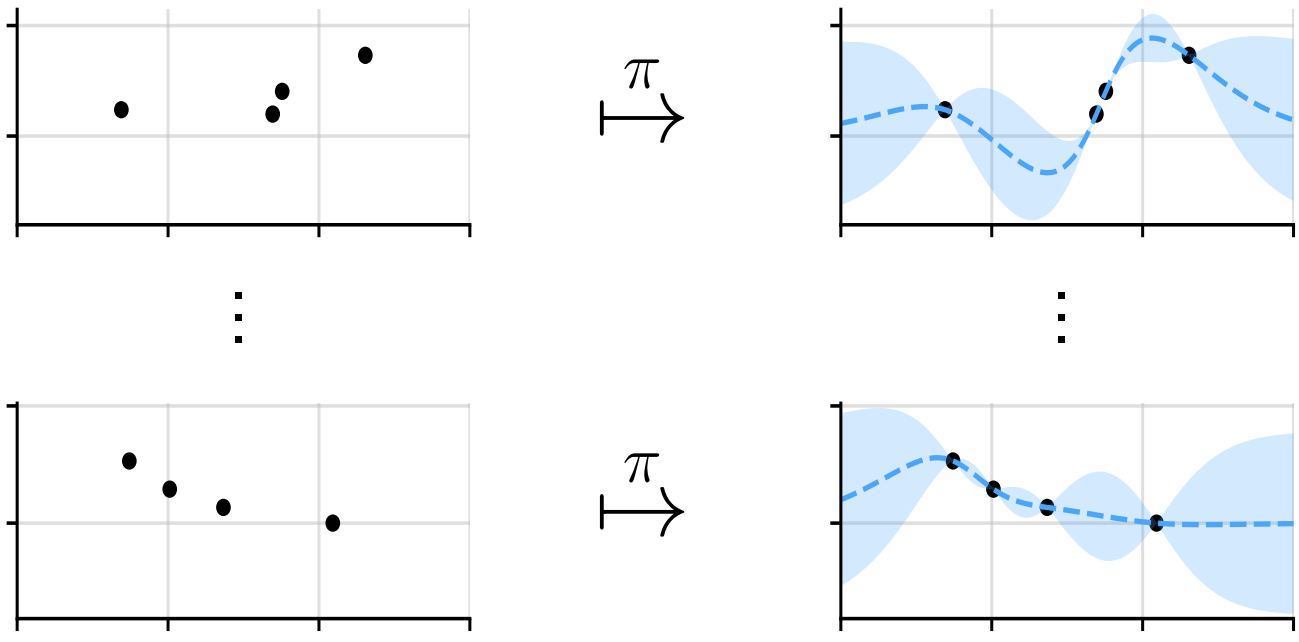


Meta-Learning and Neural Processes



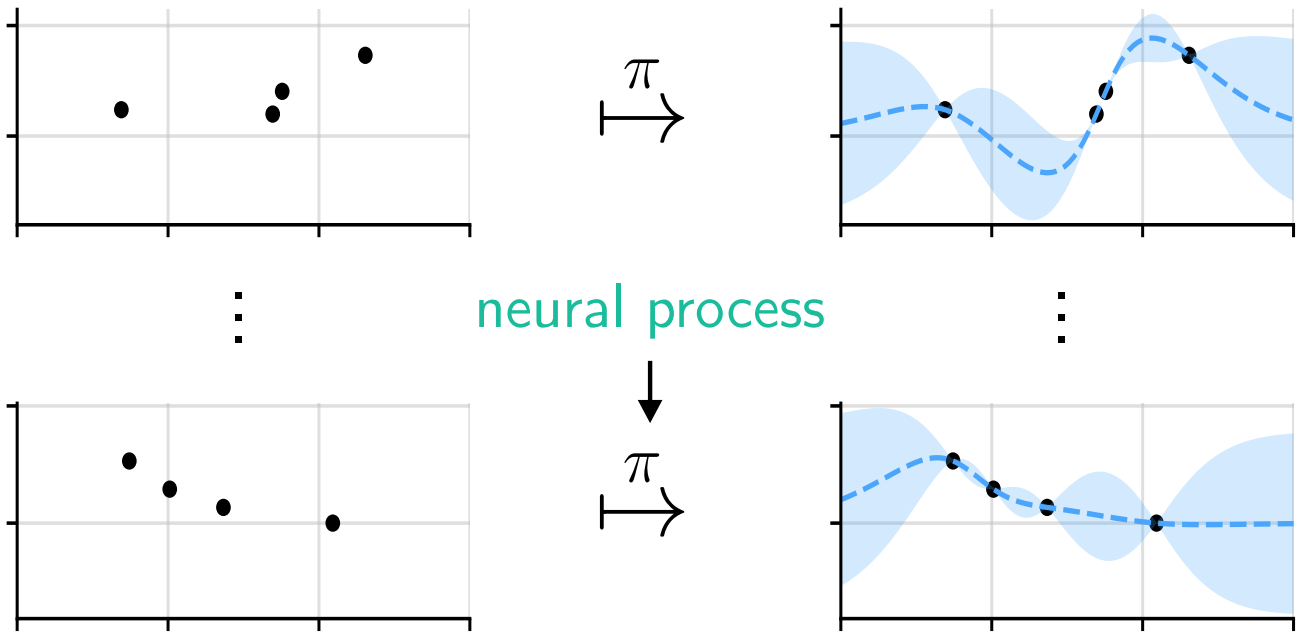
Meta-Learning and Neural Processes

π : data sets $\mathcal{D}$ $\longrightarrow$ predictions $\mathcal{P}$

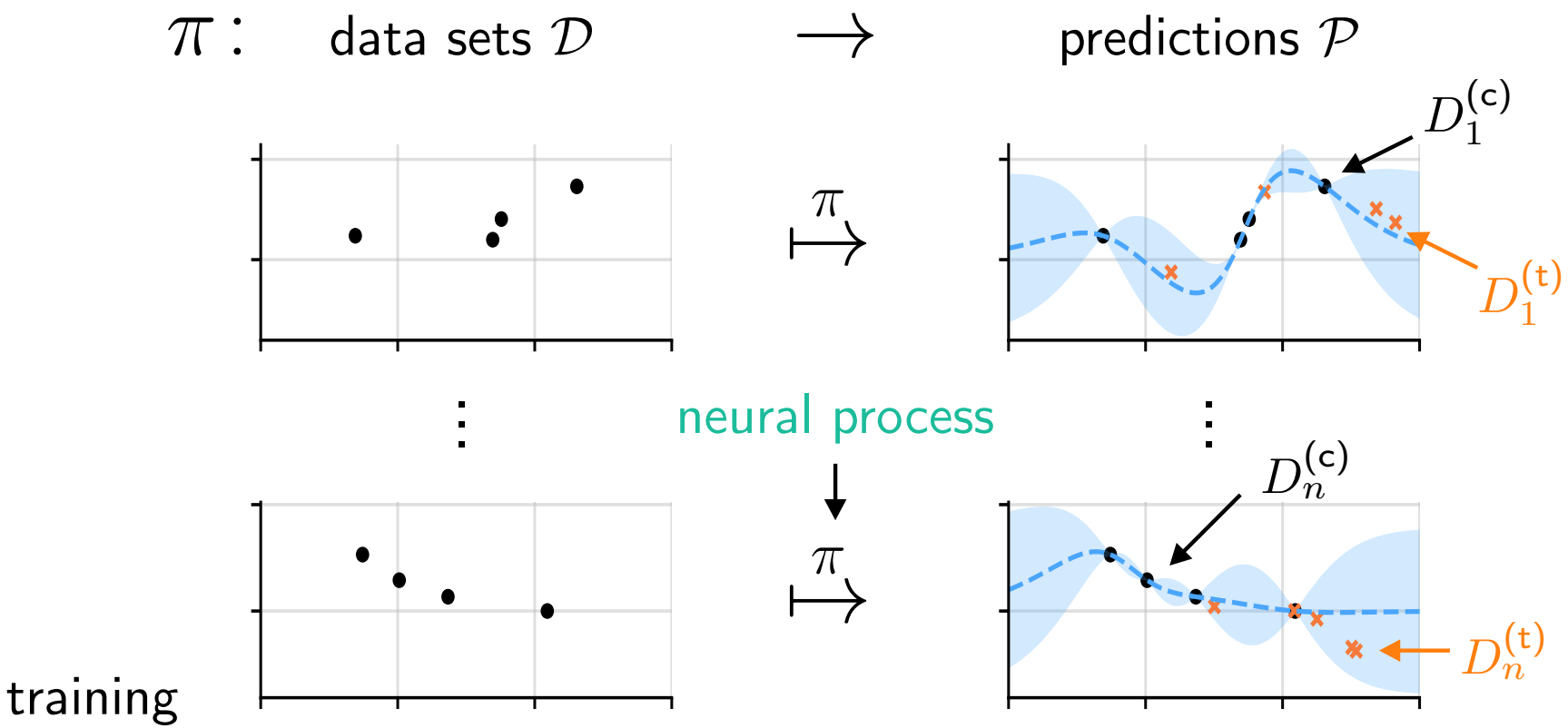


Meta-Learning and Neural Processes

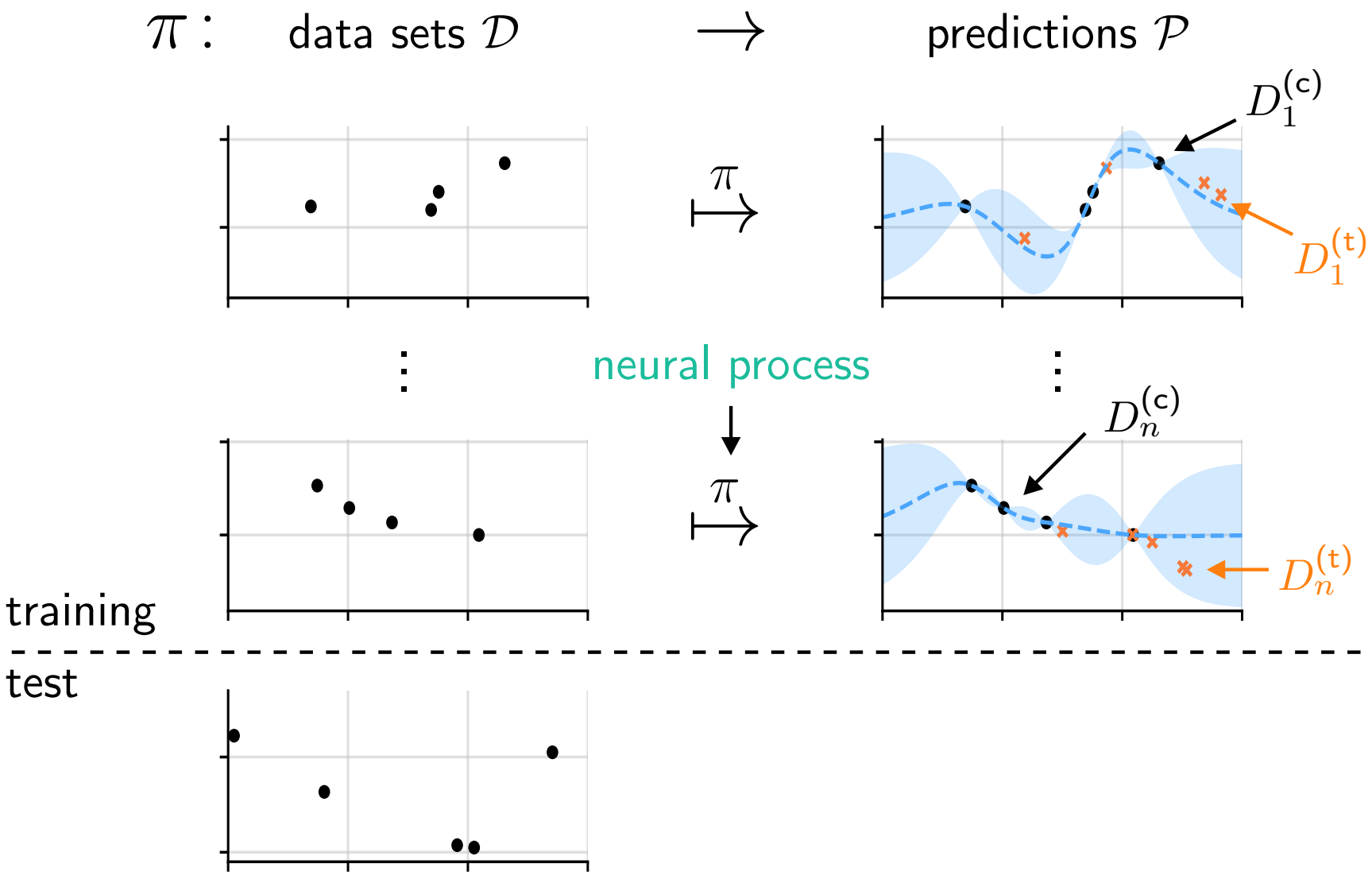
π : data sets $\mathcal{D}$ $\longrightarrow$ predictions $\mathcal{P}$



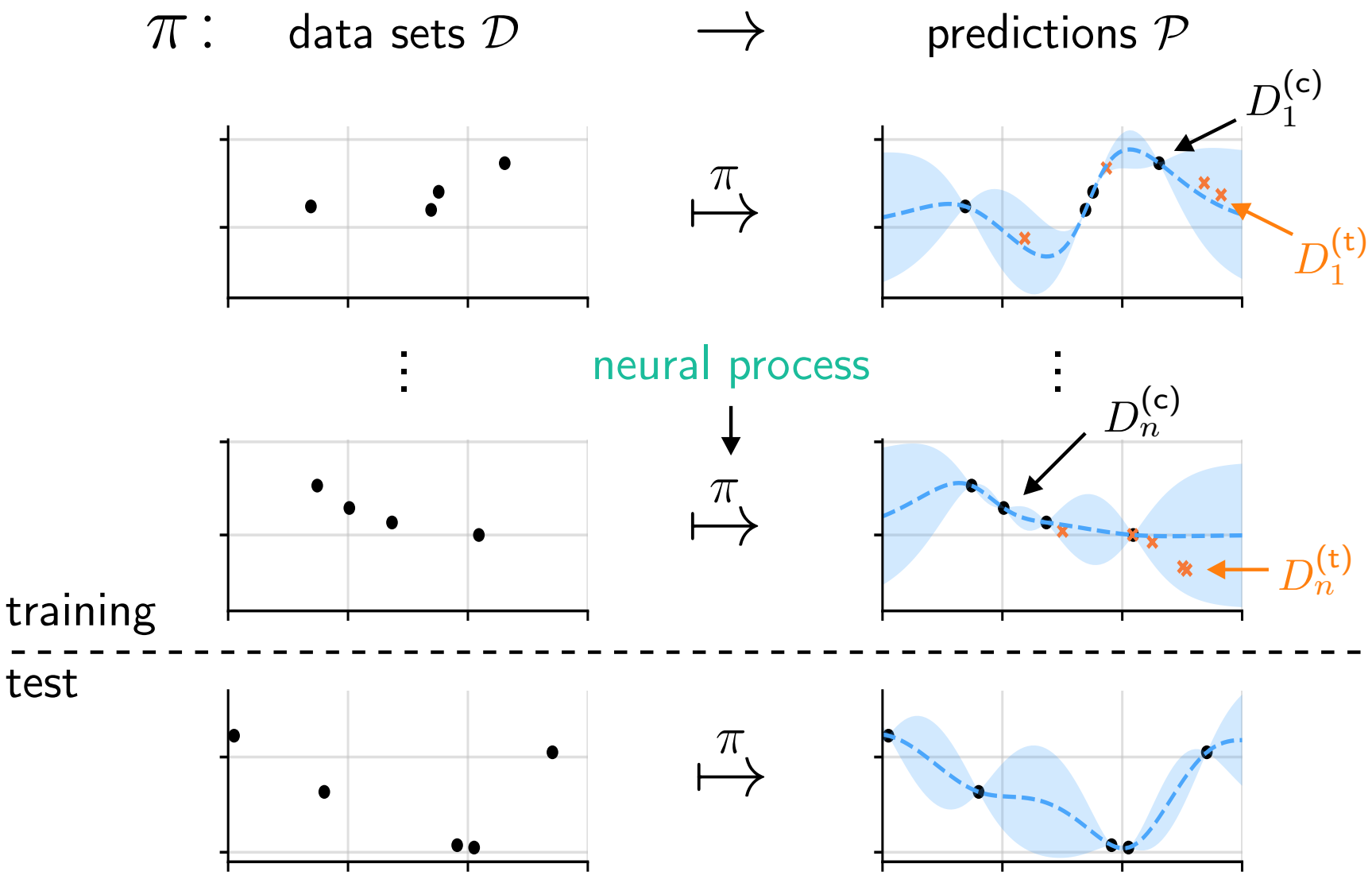
Meta-Learning and Neural Processes



Meta-Learning and Neural Processes

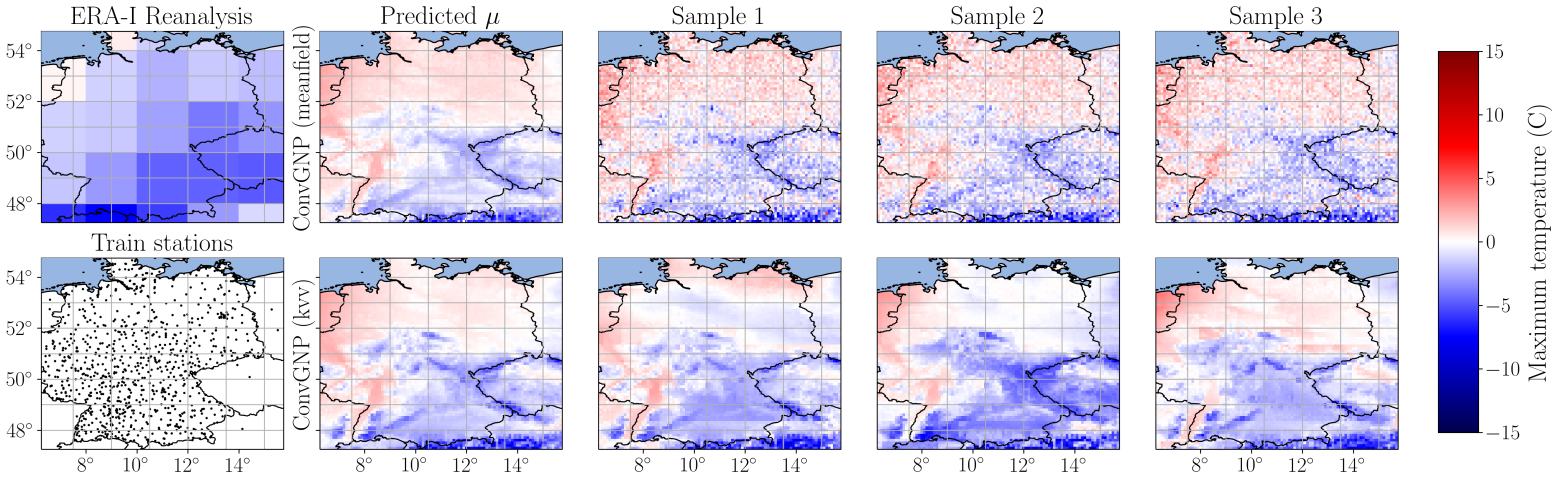


Meta-Learning and Neural Processes



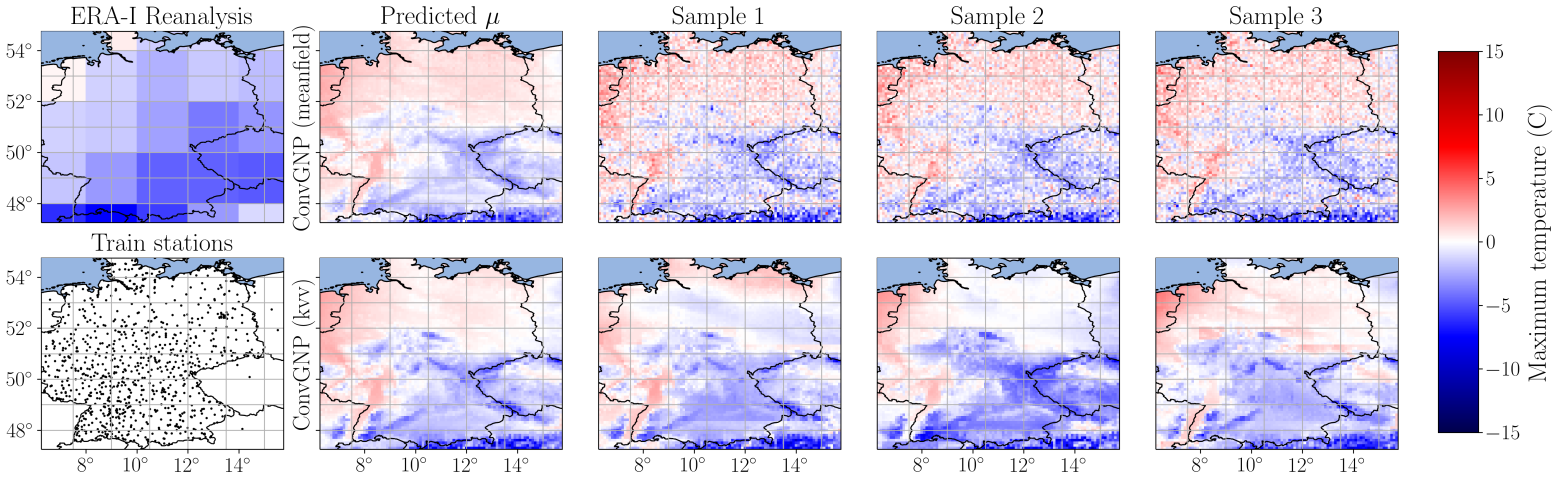
Applications of Neural Processes

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

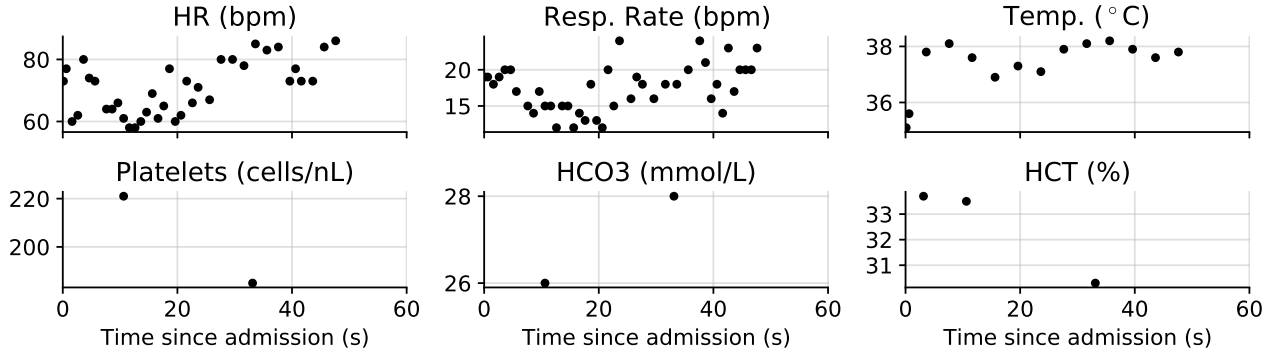


Applications of Neural Processes

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



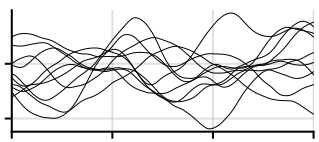
- ICU monitoring (Silva et al., 2012; Shysheya, 2020):



Today: Prediction Map Approximation

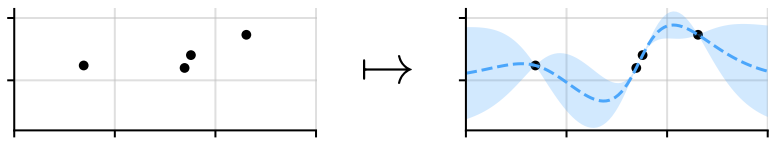
Today: Prediction Map Approximation

$$f \sim p(f)$$

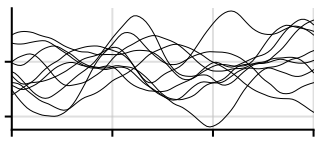


Today: Prediction Map Approximation

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

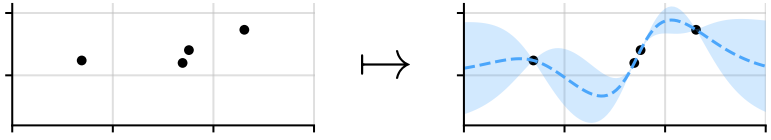


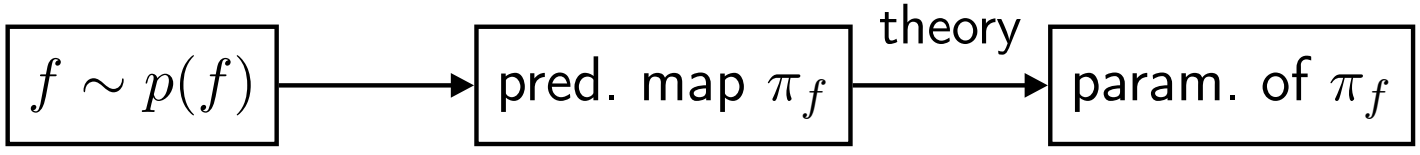
$f \sim p(f)$ → pred. map π_f

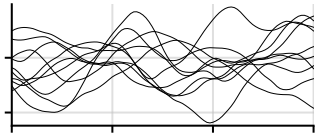


Today: Prediction Map Approximation

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



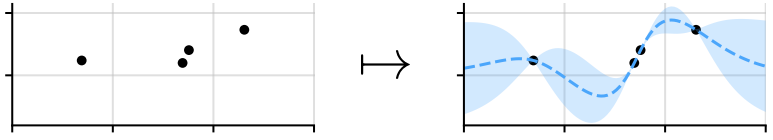


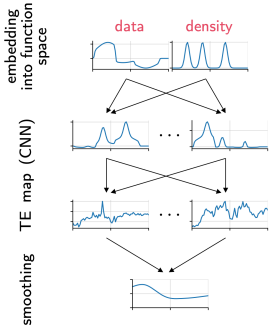


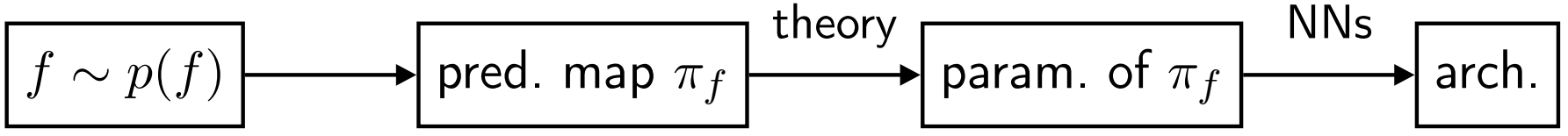
$$m(D) = \rho\big(\sum_{(x,y)\in D} \phi(x,y)\big)$$

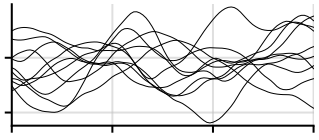
Today: Prediction Map Approximation

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



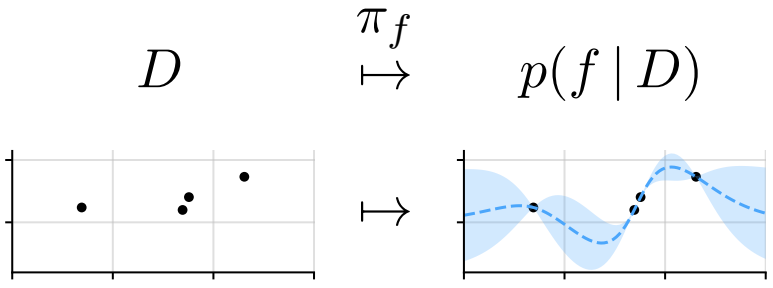


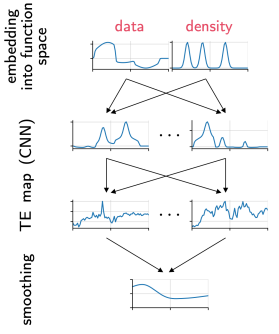


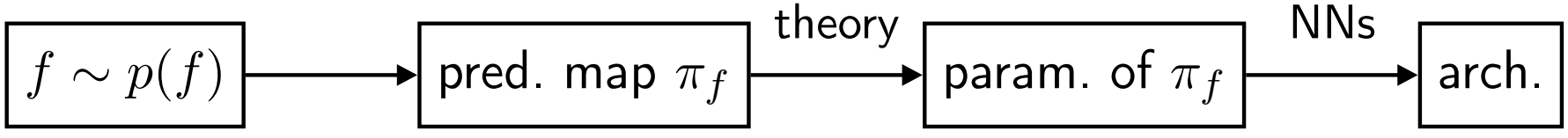


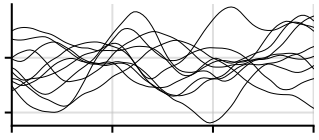
$$m(D) = \rho\big(\sum_{(x,y)\in D} \phi(x,y)\big)$$

Today: Prediction Map Approximation





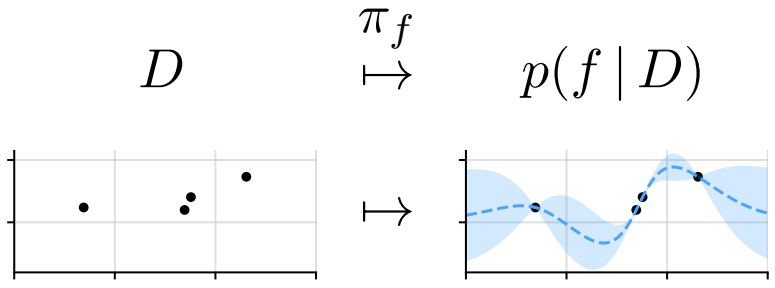


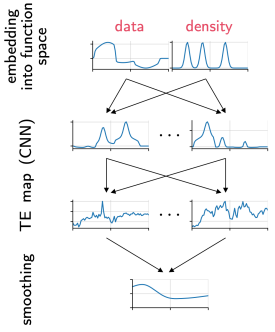


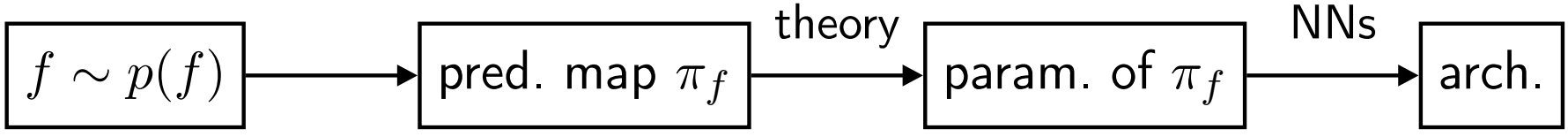
$$m(D) = \rho\Big(\sum_{(x,y)\in D} \phi(x,y)\Big)$$

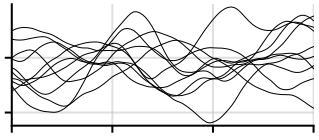
- ✓ Theoretical framework

Today: Prediction Map Approximation





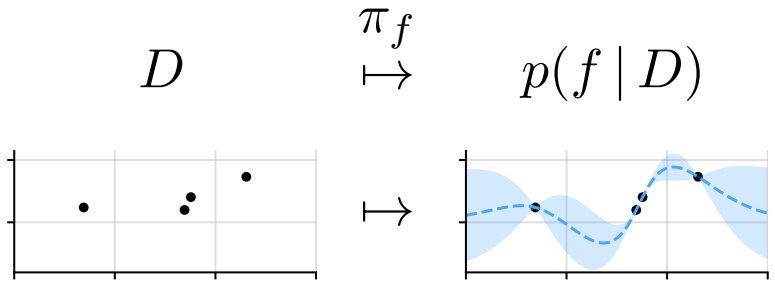


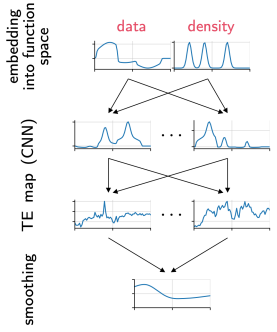


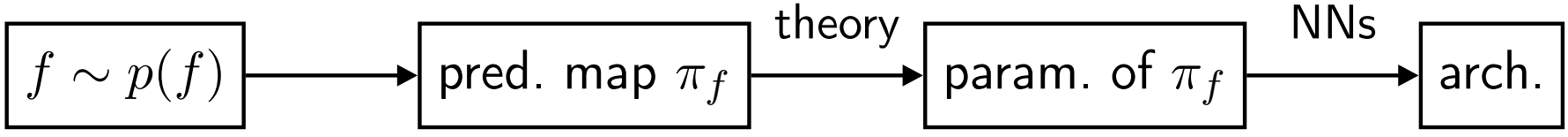
$$m(D) = \rho\big(\textstyle\sum_{(x,y)\in D} \phi(x,y)\big)$$

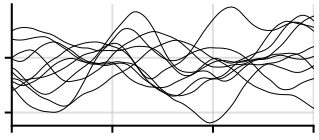
- ✓ Theoretical framework
- ✓ Architectures with universal approximation properties

Today: Prediction Map Approximation









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

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

Prediction Map Approximation

Setup

- Let f be some ground-truth stochastic process.

Setup

e.g., a sawtooth wave

- Let f be some ground-truth stochastic process.

Setup

e.g., a sawtooth wave

- Let f be some ground-truth stochastic process.
- Posterior prediction map: $\pi_f : \mathcal{D} \to \mathcal{P}$, $\pi_f(D) = p(f \mid D)$.

Setup

e.g., a sawtooth wave

- Let f be some ground-truth stochastic process.
- Posterior prediction map: $\pi_f : \mathcal{D} \to \mathcal{P}$, $\pi_f(D) = p(f \mid D)$.
- Goal: find Gaussian approximation $\tilde{\pi} : \mathcal{D} \to \mathcal{P}_{\mathsf{G}}$.

Setup

e.g., a sawtooth wave

- Let f be some ground-truth stochastic process.
- Posterior prediction map: $\pi_f : \mathcal{D} \to \mathcal{P}$, $\pi_f(D) = p(f \mid D)$.
- Goal: find Gaussian approximation $\tilde{\pi} : \mathcal{D} \to \mathcal{P}_{\mathsf{G}}$.
- Approach:

$$\tilde{\pi}(D) \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_{\mathsf{G}}} \mathrm{KL}(\pi_f(D), \mu).$$

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: find Gaussian approximation $\tilde{\pi} \colon \mathcal{D} \to \mathcal{P}_{\mathsf{G}}$.
- Approach:

$$\tilde{\pi}(D) \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_{\mathsf{G}}} \mathrm{KL}(\pi_f(D), \mu).$$

✗ Approximate f and perform inference in approximation.

Setup

e.g., a sawtooth wave

- Let f be some ground-truth stochastic process.
- Posterior prediction map: $\pi_f : \mathcal{D} \to \mathcal{P}$, $\pi_f(D) = p(f \mid D)$.
- Goal: find Gaussian approximation $\tilde{\pi} : \mathcal{D} \to \mathcal{P}_{\mathsf{G}}$.
- Approach:

$$\tilde{\pi}(D) \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_{\mathsf{G}}} \mathrm{KL}(\pi_f(D), \mu).$$

✗ Approximate f and perform inference in approximation.

✓ Directly approximate posteriors of f.

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: find Gaussian approximation $\tilde{\pi} \colon \mathcal{D} \to \mathcal{P}_{\mathsf{G}}$.
- Approach:

$$\tilde{\pi}(D) \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_{\mathsf{G}}} \mathrm{KL}(\pi_f(D), \mu).$$

✗ Approximate f and perform inference in approximation.

✓ Directly approximate posteriors of f.

- $\mathrm{KL}(\mathcal{GP}(0, 1 \cdot e^{-|\cdot|}),$ $)$

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: find Gaussian approximation $\tilde{\pi} \colon \mathcal{D} \to \mathcal{P}_{\mathsf{G}}$.
- Approach:

$$\tilde{\pi}(D) \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_{\mathsf{G}}} \mathrm{KL}(\pi_f(D), \mu).$$

✗ Approximate f and perform inference in approximation.

✓ Directly approximate posteriors of f.

- $\mathrm{KL}(\mathcal{GP}(0, 1 \cdot e^{-|\cdot|}), \mathcal{GP}(0, \sigma^2 e^{-|\cdot|}))$

Setup

e.g., a sawtooth wave

- Let f be some ground-truth stochastic process.
- Posterior prediction map: $\pi_f : \mathcal{D} \to \mathcal{P}$, $\pi_f(D) = p(f \mid D)$.
- Goal: find Gaussian approximation $\tilde{\pi} : \mathcal{D} \to \mathcal{P}_{\mathsf{G}}$.
- Approach:

$$\tilde{\pi}(D) \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_{\mathsf{G}}} \mathrm{KL}(\pi_f(D), \mu).$$

✗ Approximate f and perform inference in approximation.

✓ Directly approximate posteriors of f.

- $\mathrm{KL}(\mathcal{GP}(0, 1 \cdot e^{-|\cdot|}), \mathcal{GP}(0, \sigma^2 e^{-|\cdot|})) = \infty$ unless $\sigma^2 = 1$!

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: find Gaussian approximation $\tilde{\pi} \colon \mathcal{D} \to \mathcal{P}_{\mathsf{G}}$.
- Approach:

$$\tilde{\pi}(D) \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_{\mathsf{G}}} \mathrm{KL}(\pi_f(D), \mu).$$

✗ Approximate f and perform inference in approximation.

✓ Directly approximate posteriors of f.

- $\mathrm{KL}(\mathcal{GP}(0, 1 \cdot e^{-|\cdot|}), \mathcal{GP}(0, \sigma^2 e^{-|\cdot|})) = \infty$ unless $\sigma^2 = 1$!
- If $\mathrm{KL}(\pi_f(D), \mu_0) < \infty$ for some $\mu_0 \in \mathcal{P}_{\mathsf{G}}$, then

$$\tilde{\pi}(D) = \pi_{\mathsf{MM}}(D)$$

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: find Gaussian approximation $\tilde{\pi} \colon \mathcal{D} \to \mathcal{P}_{\mathsf{G}}$.
- Approach:

$$\tilde{\pi}(D) \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_{\mathsf{G}}} \mathrm{KL}(\pi_f(D), \mu).$$

✗ Approximate f and perform inference in approximation.

✓ Directly approximate posteriors of f.

- $\mathrm{KL}(\mathcal{GP}(0, 1 \cdot e^{-|\cdot|}), \mathcal{GP}(0, \sigma^2 e^{-|\cdot|})) = \infty$ unless $\sigma^2 = 1$!
- If $\mathrm{KL}(\pi_f(D), \mu_0) < \infty$ for some $\mu_0 \in \mathcal{P}_{\mathsf{G}}$, then

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

- Practical objective:

$$\tilde{\pi} \in \arg\min_{\pi \in \mathcal{Q}} \mathcal{L}(\pi),$$
$$\mathcal{L}(\pi) = \mathbb{E}_{p(D)p(\mathbf{x})} \mathrm{KL}(P_{\mathbf{x}}\pi_f(D), P_{\mathbf{x}}\pi(D))$$

Neural Processes

- Practical objective:

$$\tilde{\pi} \in \arg\min_{\pi \in \mathcal{Q}} \mathcal{L}(\pi),$$

à la variational family $\longrightarrow$

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

Neural Processes

- Practical objective:

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

$$\tilde{\pi} \in \arg\min_{\pi \in \mathcal{Q}} \mathcal{L}(\pi),$$

à la variational family

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

Neural Processes

- Practical objective:

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

$$\tilde{\pi} \in \arg\min_{\pi \in \mathcal{Q}} \mathcal{L}(\pi),$$

à la variational family

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

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

Neural Processes

- Practical objective:

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

à la variational family $\longrightarrow$

$$\tilde{\pi} \in \arg\min_{\pi \in \mathcal{Q}} \mathcal{L}(\pi),$$

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

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

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

Neural Processes

- Practical objective:

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

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

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

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

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

- Call π continuous if $D_i \to D$ implies $\pi(D_i) \rightharpoonup \pi(D)$.

Neural Processes

- Practical objective:

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

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

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

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

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

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

$$\mathcal{M}_{\mathsf{G}} = \{\pi \colon \mathcal{D} \to \mathcal{P}_{\mathsf{G}} : \pi \text{ continuous}\},$$

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

Parametrising Prediction Maps

- For now, consider $\mathcal{Q}_{\mathsf{G,MF}} = \{\pi : \mathcal{D} \to \mathcal{P}_{\mathsf{G,MF}}\}$.

Parametrising Prediction Maps

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

- For now, consider $\mathcal{Q}_{\mathsf{G,MF}} = \{\pi : \mathcal{D} \to \mathcal{P}_{\mathsf{G,MF}}\}$.

Parametrising Prediction Maps

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

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

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

Parametrising Prediction Maps

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

- For now, consider $\mathcal{Q}_{\text{G,MF}} = \{\pi \colon \mathcal{D} \to \mathcal{P}_{\text{G,MF}}\}$.
- 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.

Parametrising Prediction Maps

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

- For now, consider $\mathcal{Q}_{\text{G,MF}} = \{\pi \colon \mathcal{D} \to \mathcal{P}_{\text{G,MF}}\}$.
- 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:

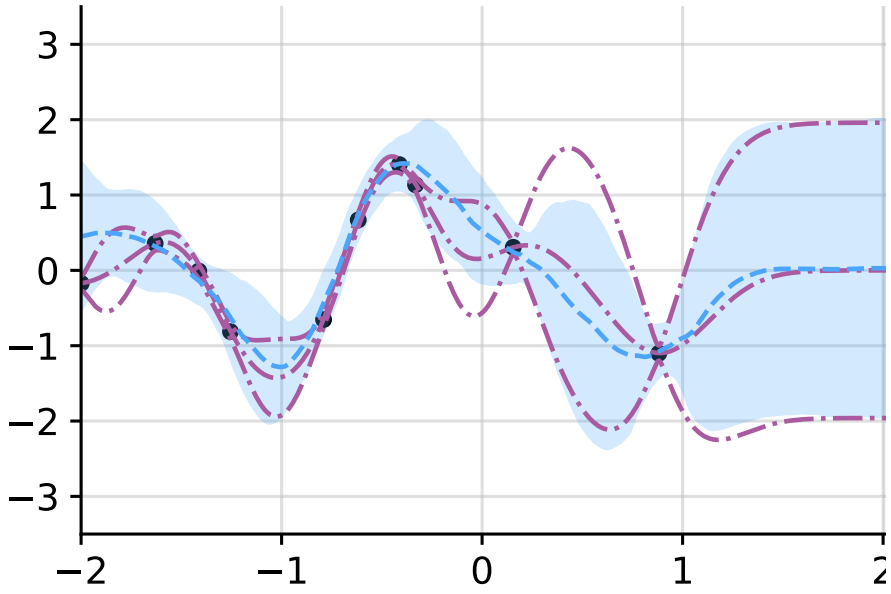
$$f(D) = \rho\big(\textstyle\sum_{(x,y)\in D} \phi(x, y)\big)$$

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

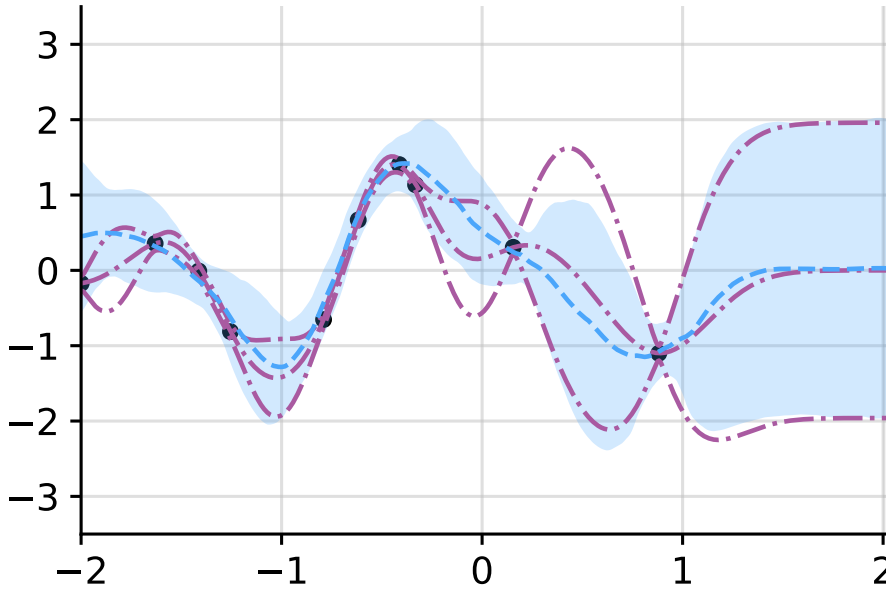
- Conditional neural process (Garnelo et al., 2018):

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

The Conditional Neural Process

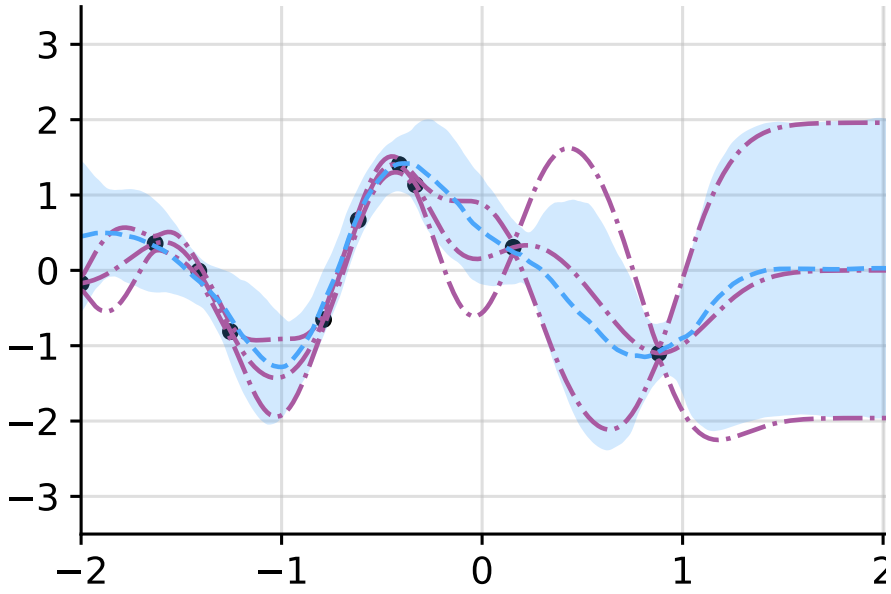


The Conditional Neural Process



✗ Learns very slowly

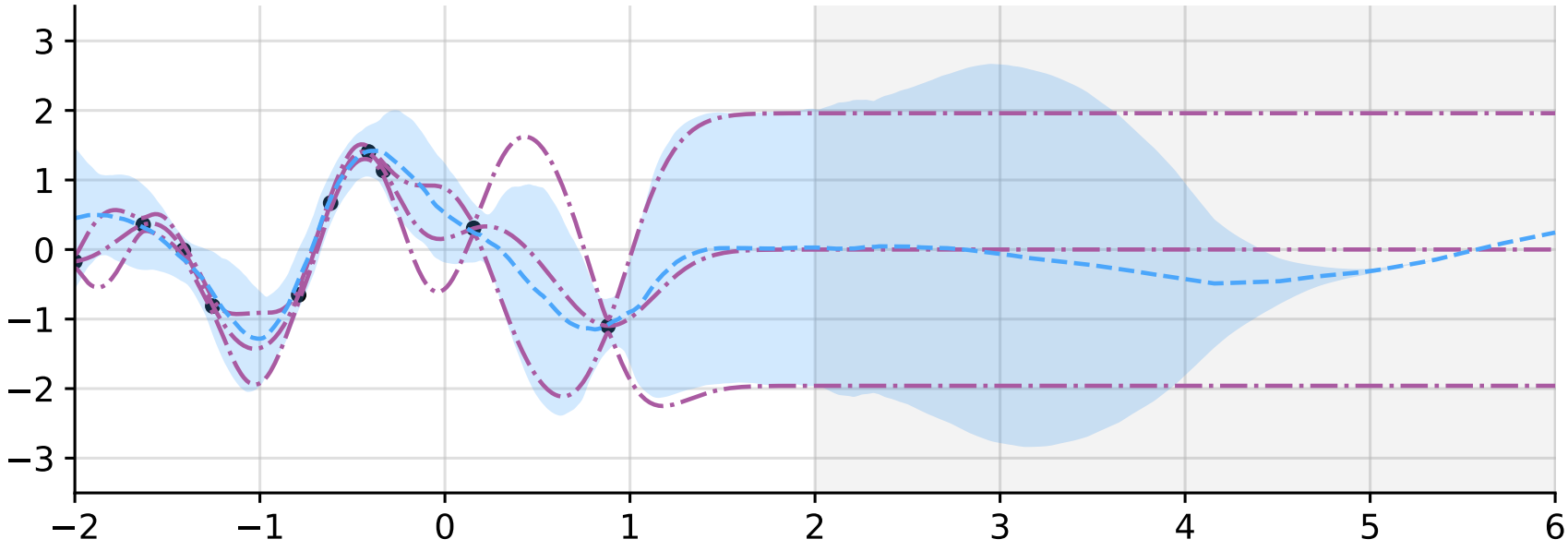
The Conditional Neural Process



✗ Learns very slowly

✗ Underfits

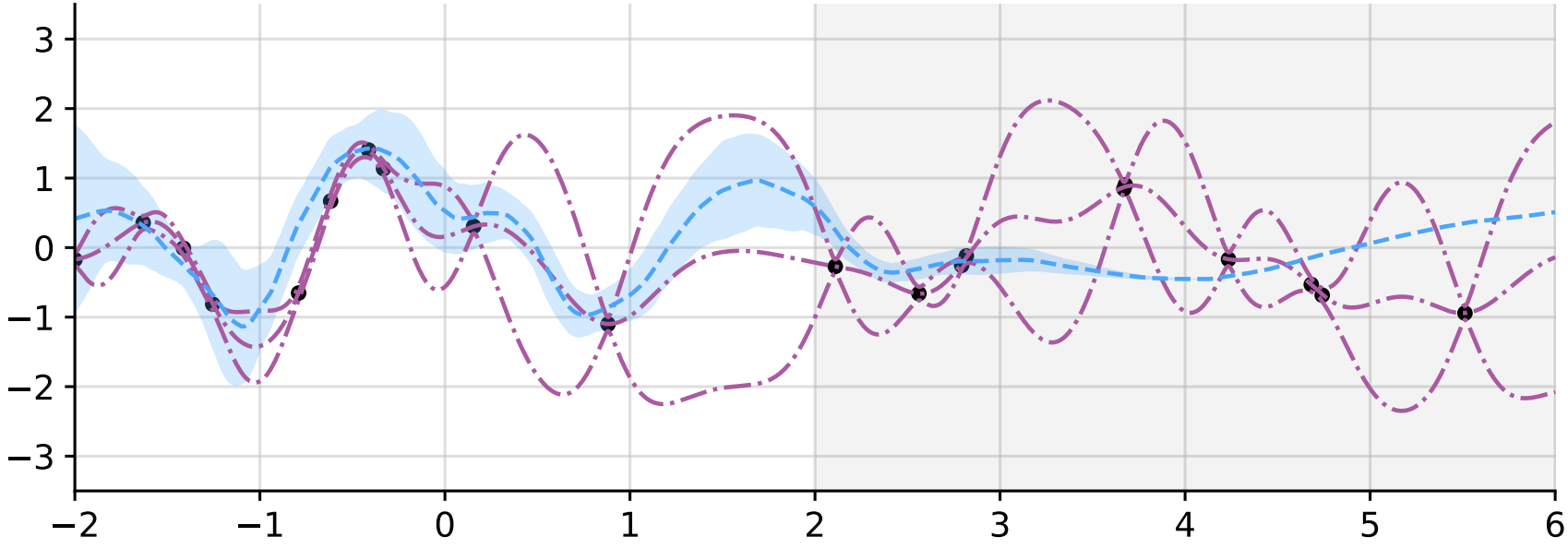
The Conditional Neural Process



✗ Learns very slowly

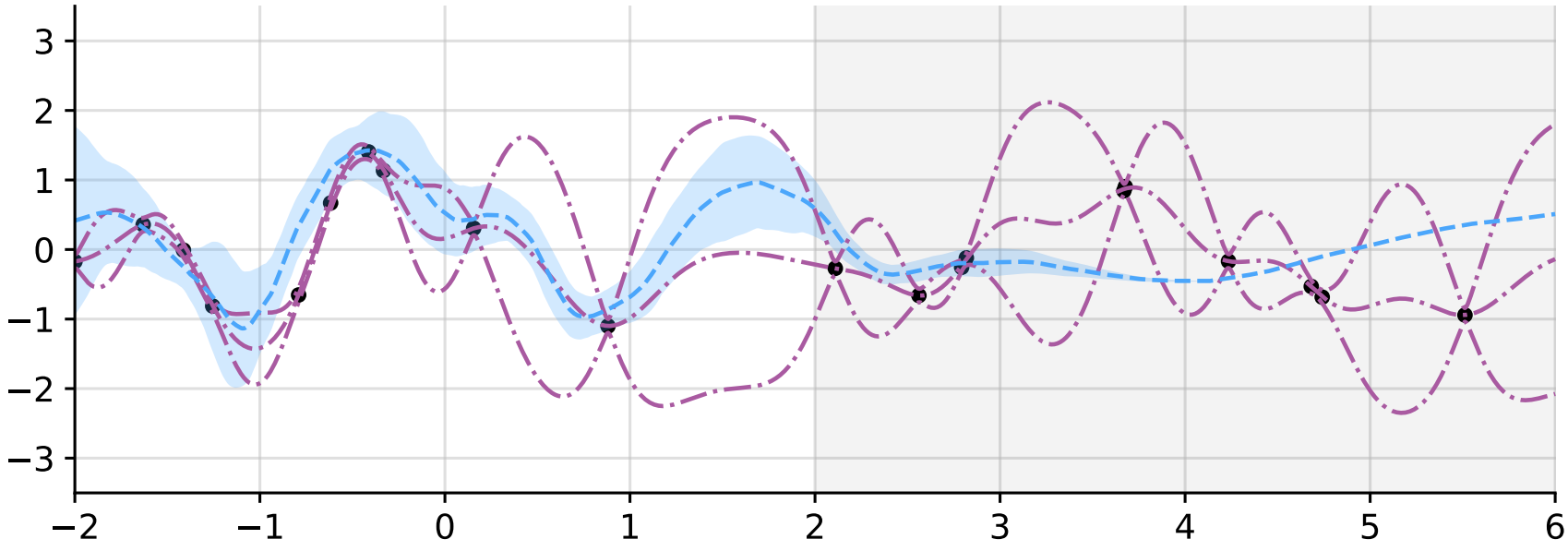
✗ Underfits

The Conditional Neural Process



- ✗ Learns very slowly
- ✗ Underfits

The Conditional Neural Process



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

Exploiting Stationarity

Translation-Equivariant Prediction Maps

- Let T_τ represent a translation by τ.

Translation-Equivariant Prediction Maps

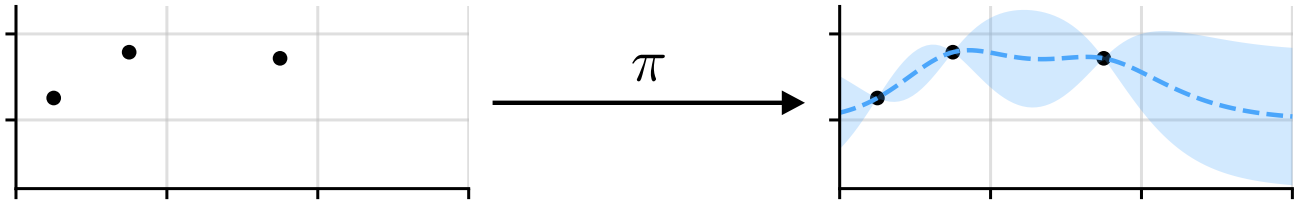
- Let T_τ represent a translation by τ.
- A prediction map $\pi\colon \mathcal{D} \to \mathcal{P}$ is translation equivariant (TE) if

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

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

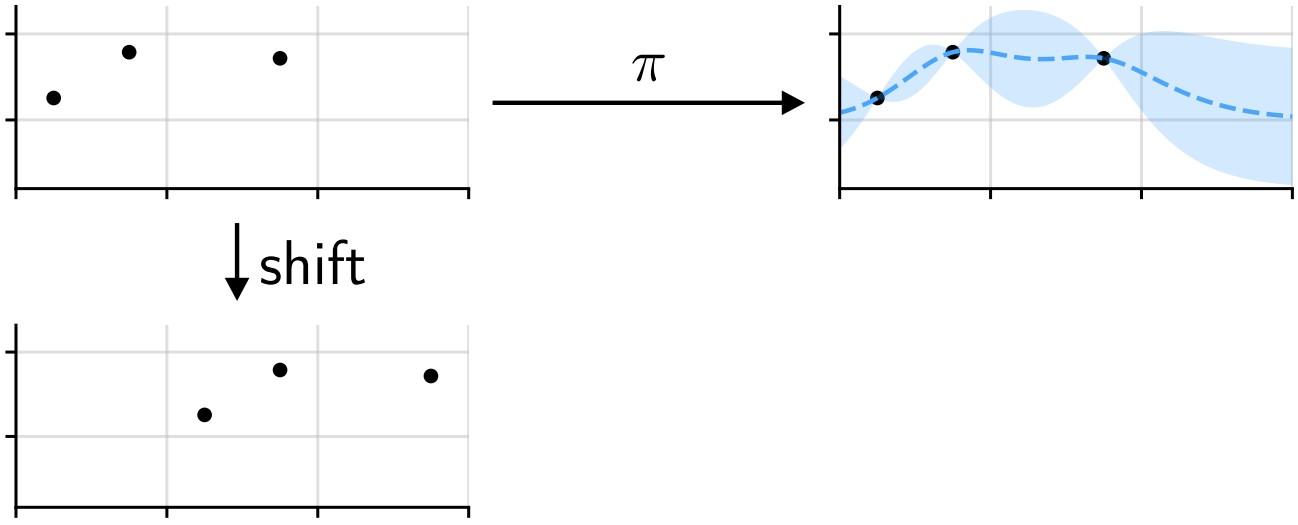
$$\pi(\mathsf{T}_\tau D) = \mathsf{T}_\tau \pi(D).$$



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

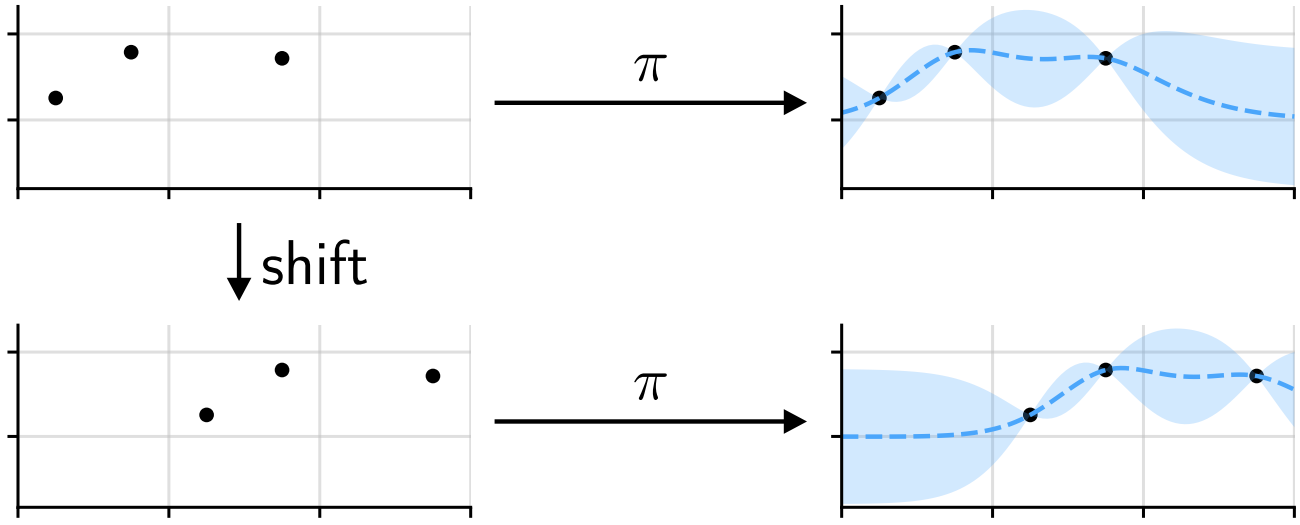
$$\pi(\mathsf{T}_\tau D) = \mathsf{T}_\tau \pi(D).$$



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

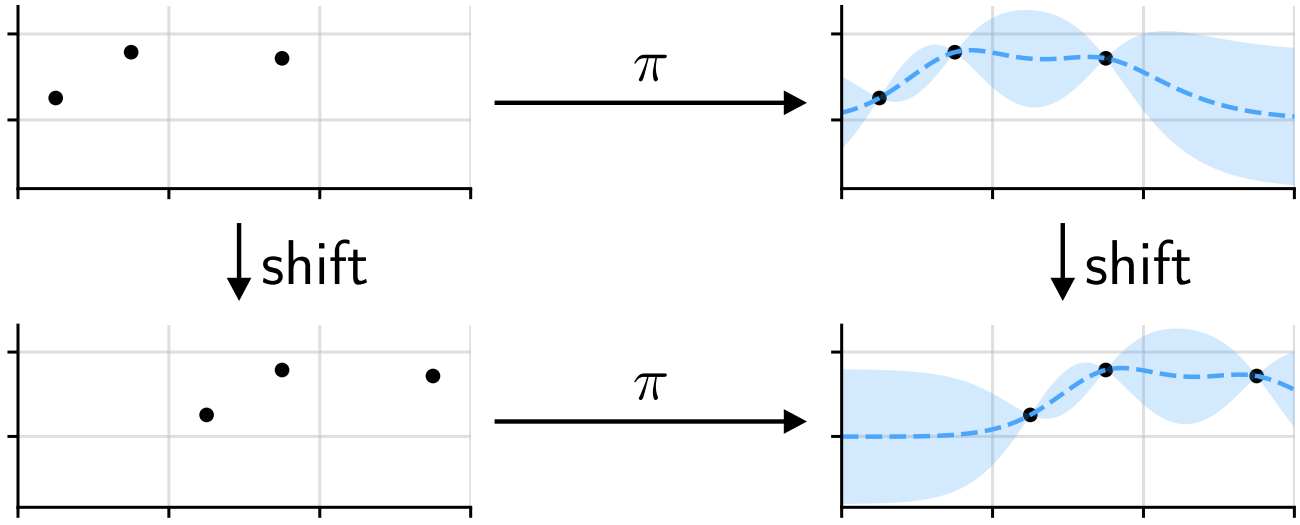
$$\pi(\mathsf{T}_\tau D) = \mathsf{T}_\tau \pi(D).$$



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

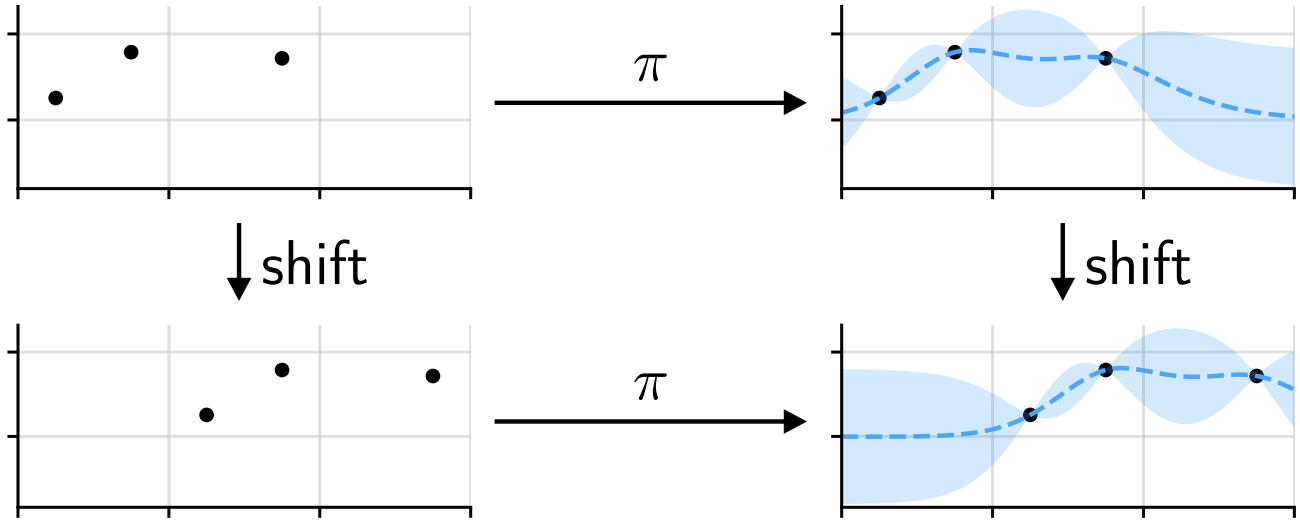
$$\pi(\mathsf{T}_\tau D) = \mathsf{T}_\tau \pi(D).$$



Translation-Equivariant Prediction Maps

- Let T_τ represent a translation by τ.
- A prediction map $\pi : \mathcal{D} \to \mathcal{P}$ 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)

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

Parametrising TE Prediction Maps

Deep Set

(Zaheer et al., 2017)

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

$\iff$

$$E : \mathcal{D}_{\leq M} \to \mathbb{R}^M,$$
$$E(D) = \sum_{(x,y)\in D} \phi(x, y)$$

$$\rho : \mathbb{R}^M \to Z,$$
$$f(D) = \rho(E(D))$$

Parametrising TE Prediction Maps

Deep Set

(Zaheer et al., 2017)

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

$\iff$

encoder

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

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

decoder

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

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

Parametrising TE Prediction Maps

Deep Set

(Zaheer et al., 2017)

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

$\Longleftrightarrow$

encoder

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

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

decoder

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

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

Parametrising TE Prediction Maps

Deep Set

(Zaheer et al., 2017)

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

$\iff$ T_τ?!

encoder

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

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

decoder

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

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

Parametrising TE Prediction Maps

Deep Set (Zaheer et al., 2017)	Convolutional Deep Set (Gordon et al., 2020)
$f\colon \mathcal{D}_{\leq M} \to Z$ is cont.	$f\colon \mathcal{D}_{\leq M} \to Z$ is cont. and TE
$\iff$	T_τ?!
encoder $E\colon \mathcal{D}_{\leq M} \to \mathbb{R}^M$, $E(D) = \sum_{(x,y)\in D} \phi(x,y)$	
decoder $\rho\colon \mathbb{R}^M \to Z$, $f(D) = \rho(E(D))$	

Parametrising TE Prediction Maps

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

Parametrising TE Prediction Maps

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

Parametrising TE Prediction Maps

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

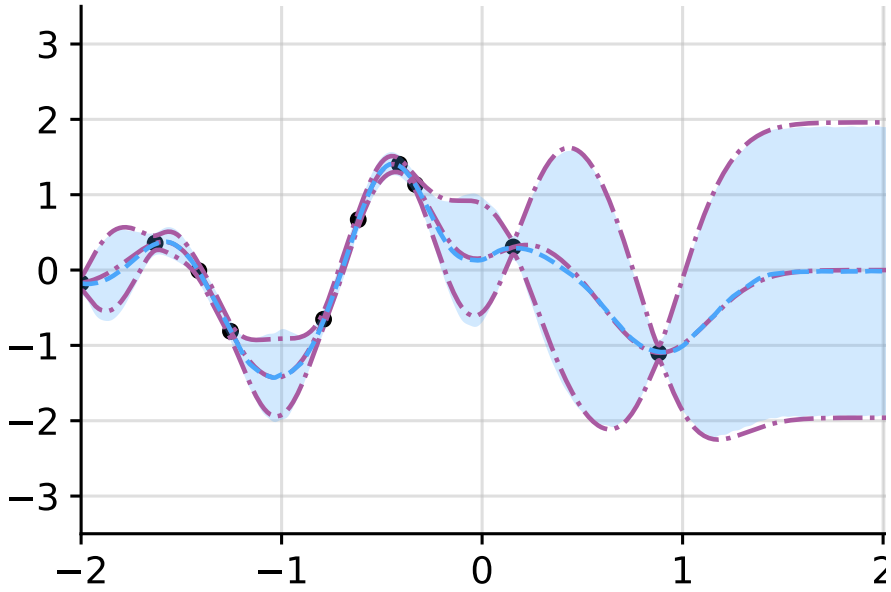
Parametrising TE Prediction Maps

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

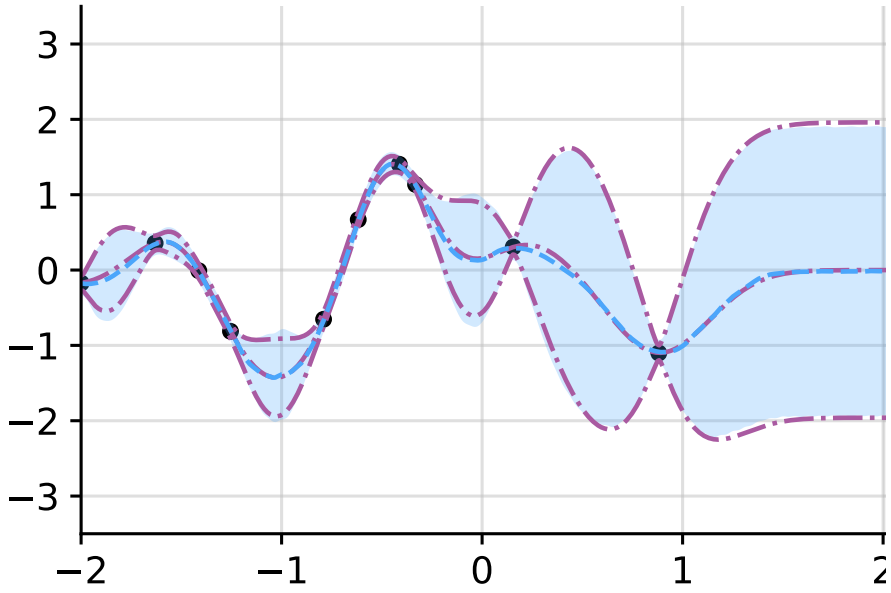
- Gives convolutional CNP (Gordon et al., 2020):

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

The Convolutional CNP

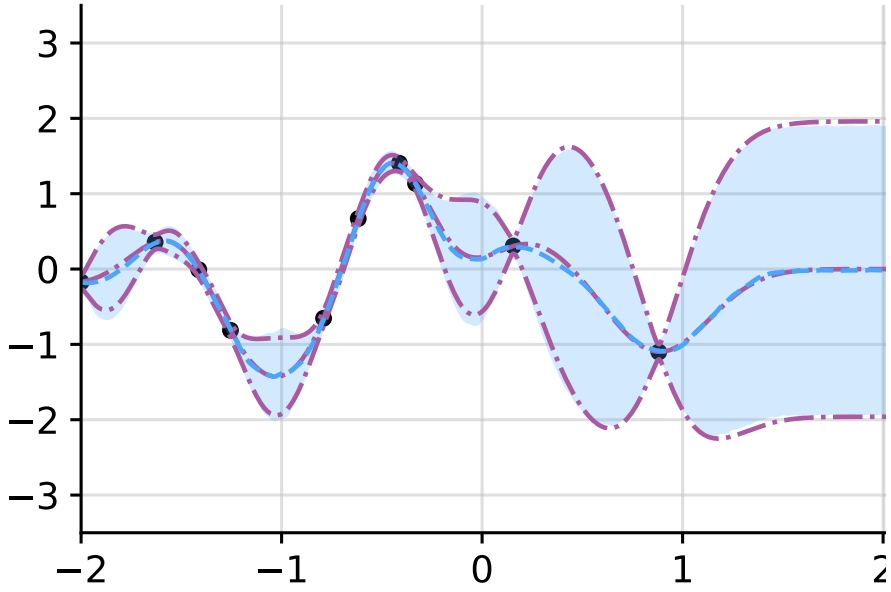


The Convolutional CNP



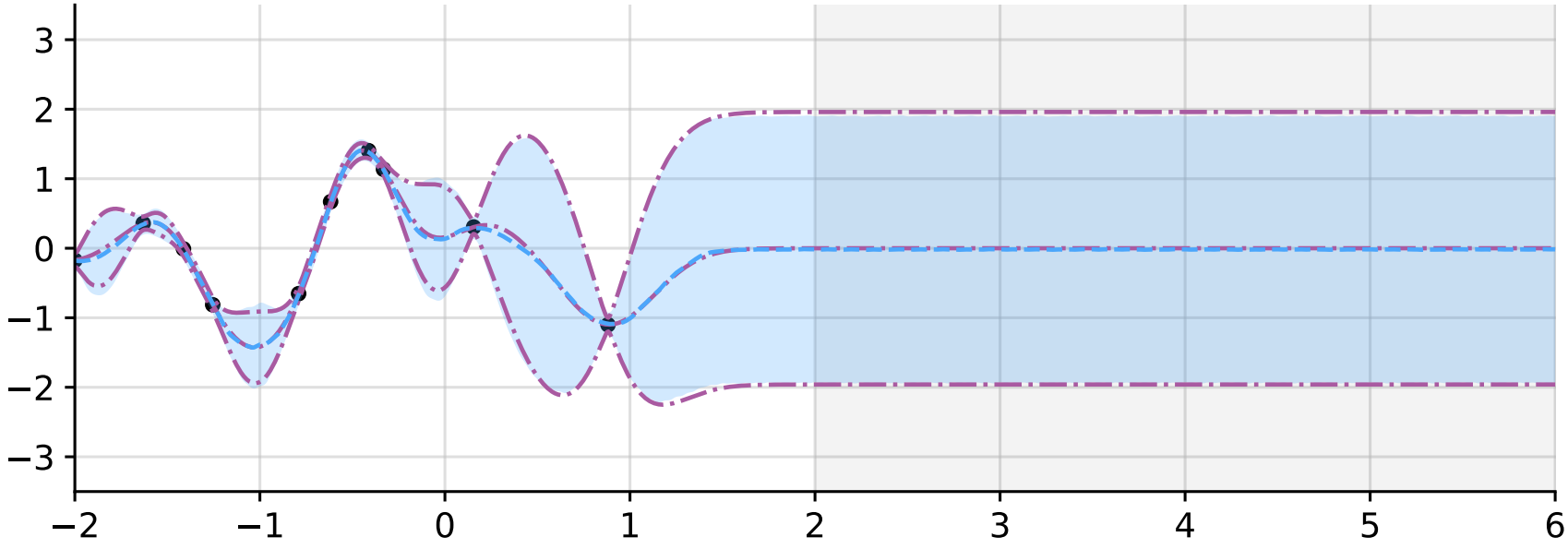
✓ Learns pretty quickly

The Convolutional CNP



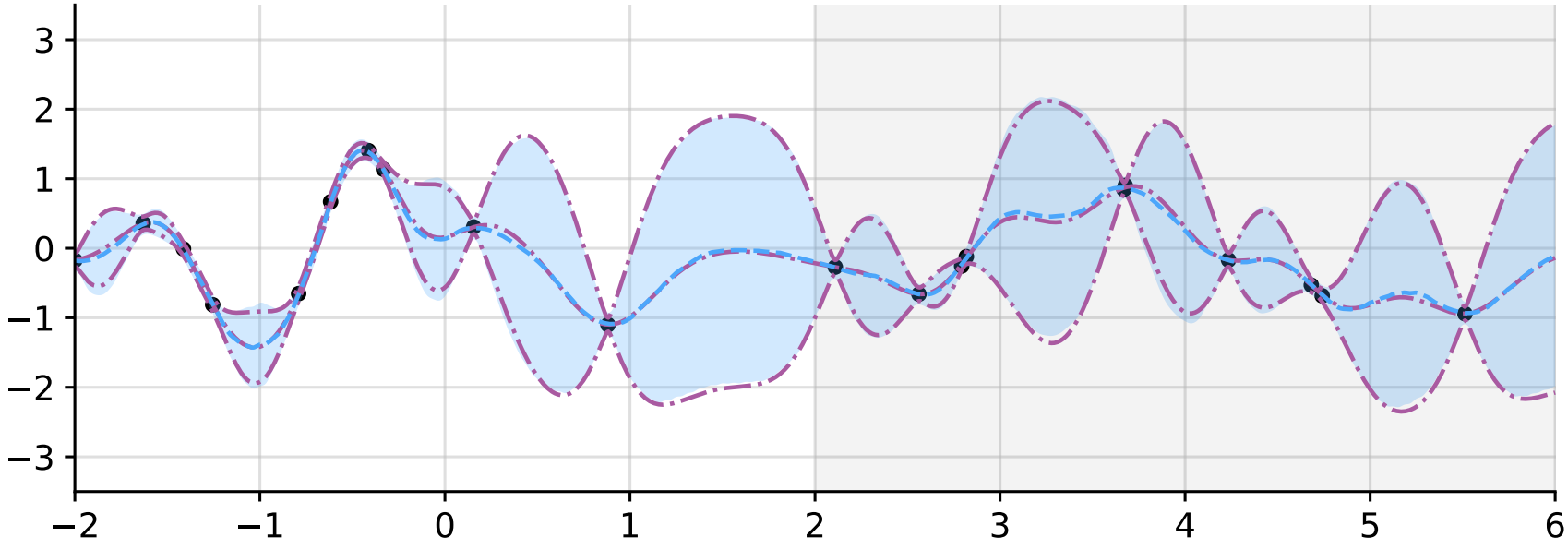
- ✓ Learns pretty quickly
- ✓ Recovers target (diagonalised ground-truth GP)

The Convolutional CNP



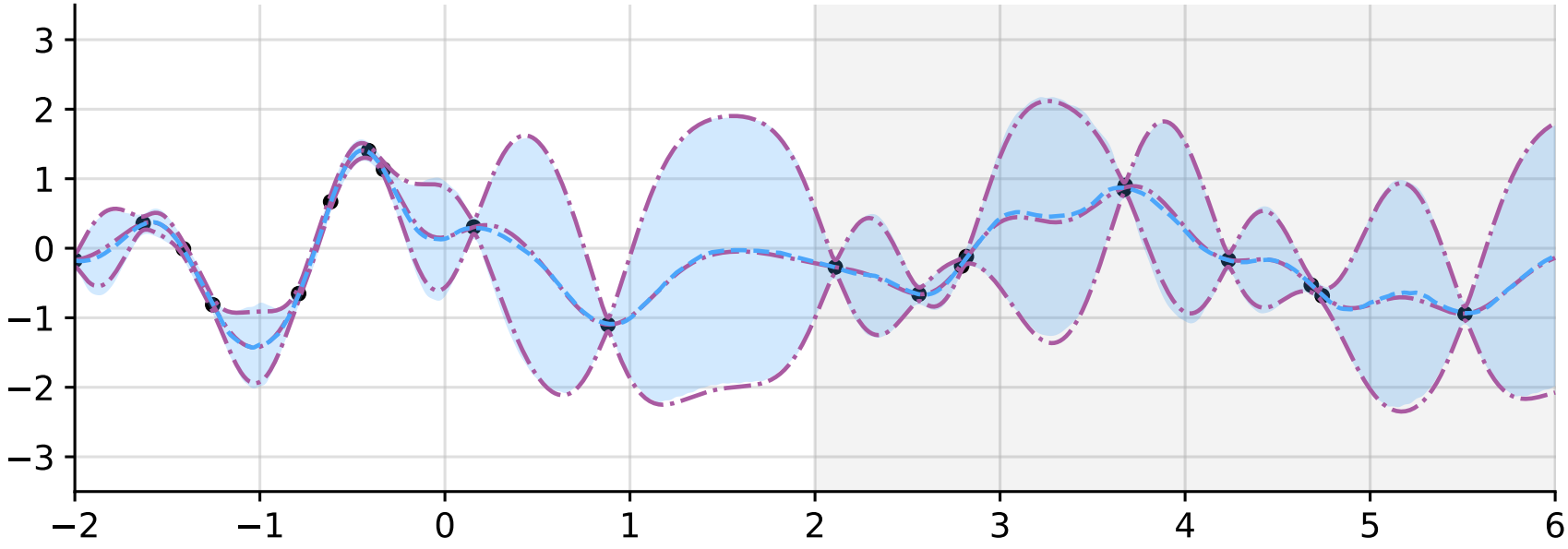
- ✓ Learns pretty quickly
- ✓ Recovers target (diagonalised ground-truth GP)

The Convolutional CNP



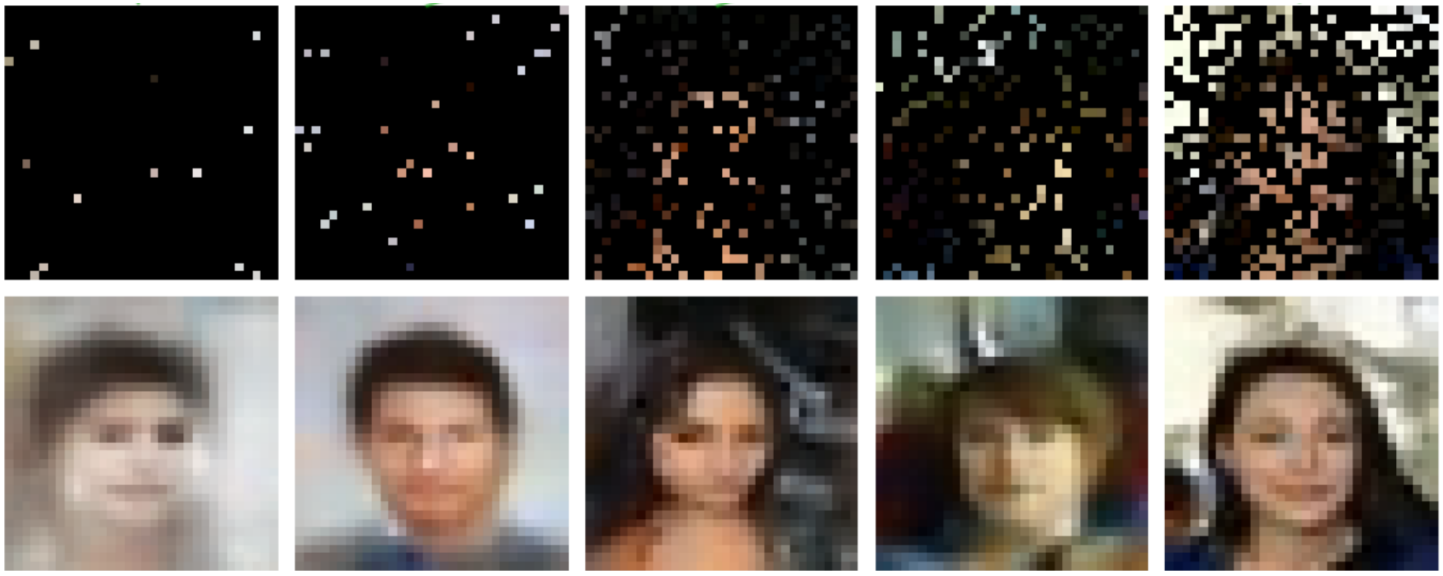
- ✓ Learns pretty quickly
- ✓ Recovers target (diagonalised ground-truth GP)

The Convolutional CNP



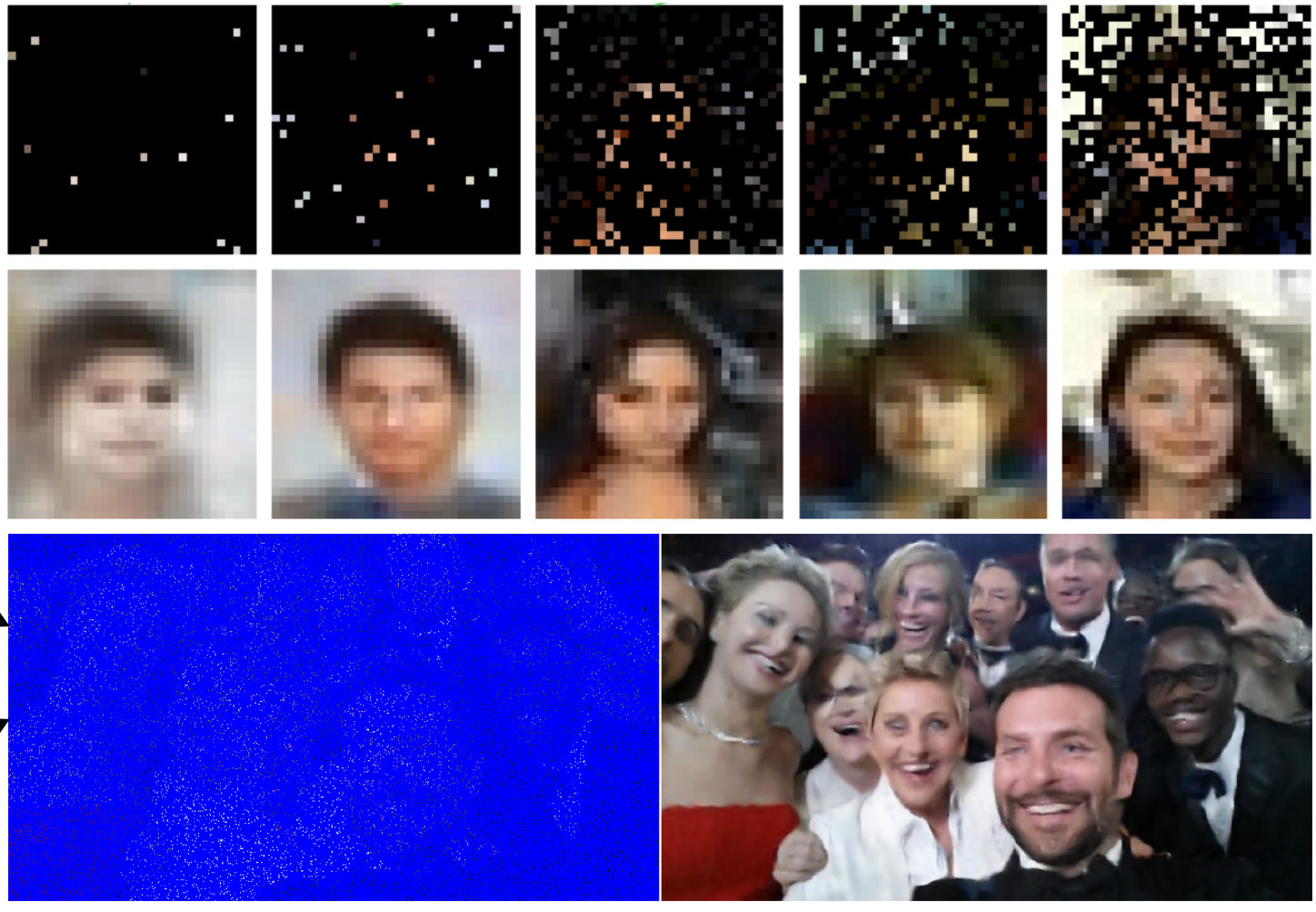
- ✓ Learns pretty quickly
- ✓ Recovers target (diagonalised ground-truth GP)
- ✓ Generalises well

The Convolutional CNP (2)



Gordon et al. (2020)

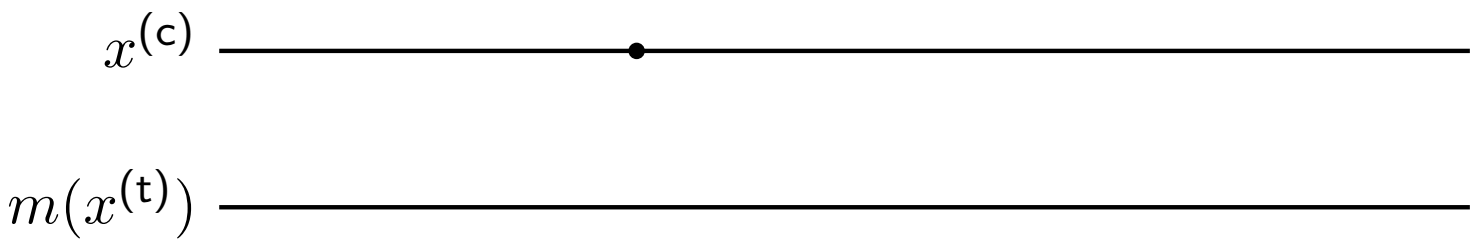
The Convolutional CNP (2)



Gordon et al. (2020)

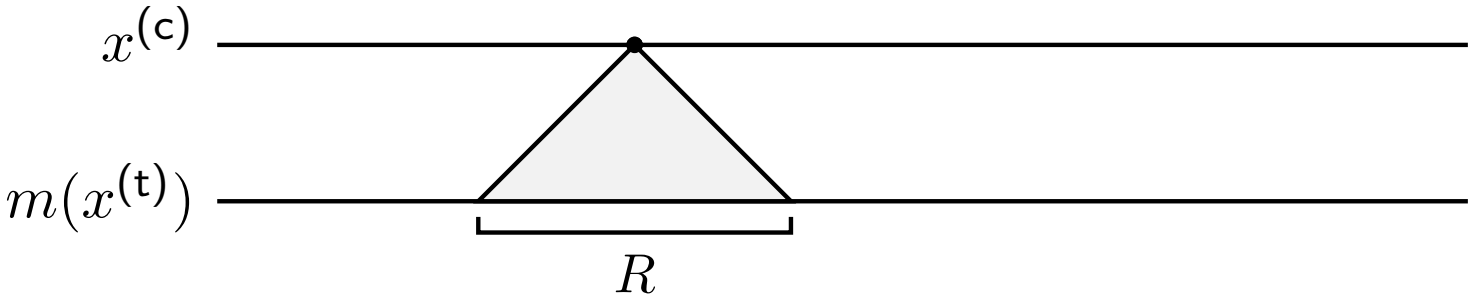
Why Does TE Help Generalise?

- CNNs have receptive field $R > 0$:



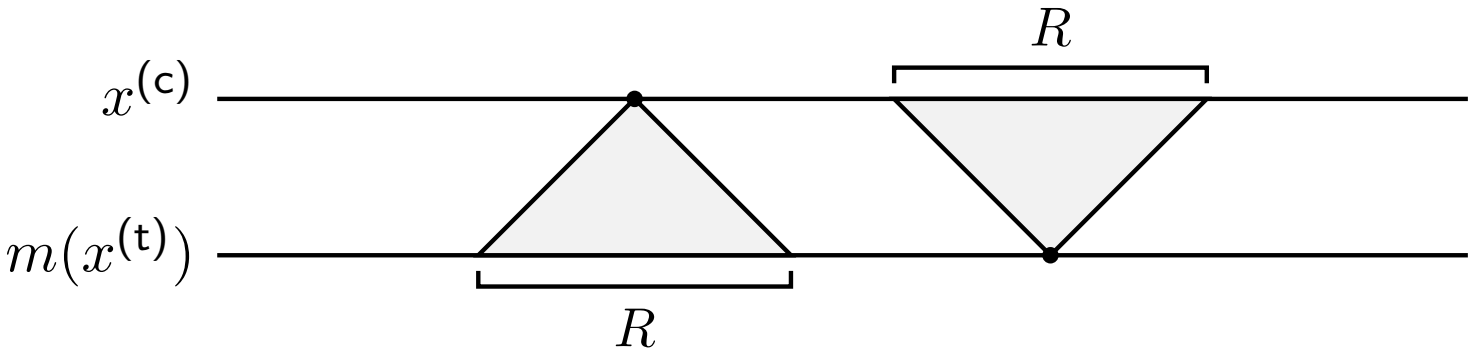
Why Does TE Help Generalise?

- CNNs have receptive field $R > 0$:



Why Does TE Help Generalise?

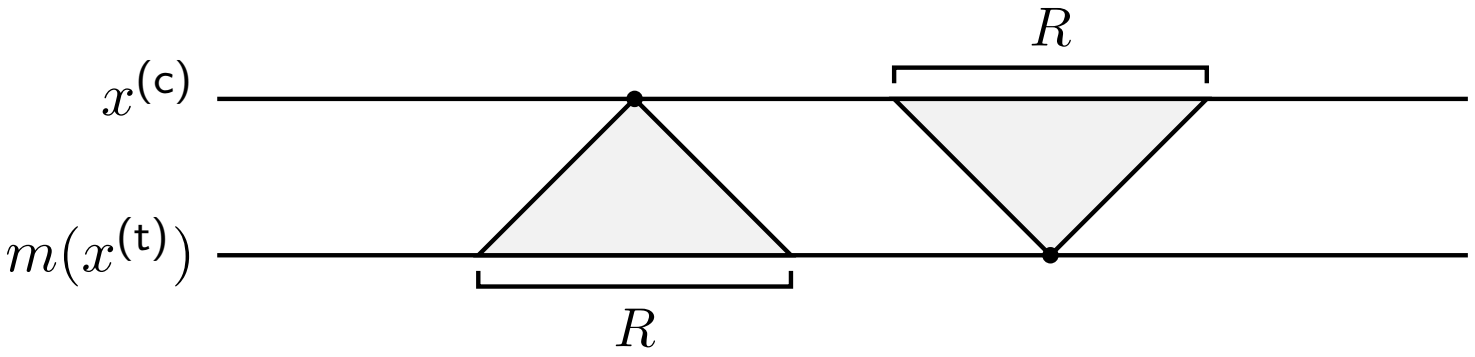
- CNNs have receptive field $R > 0$:



Why Does TE Help Generalise?

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

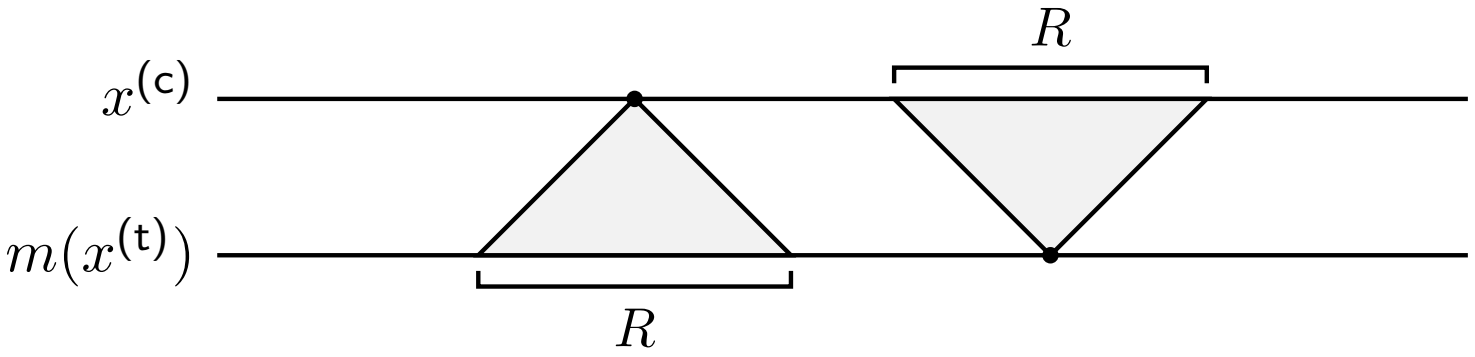
- CNNs have receptive field $R > 0$:



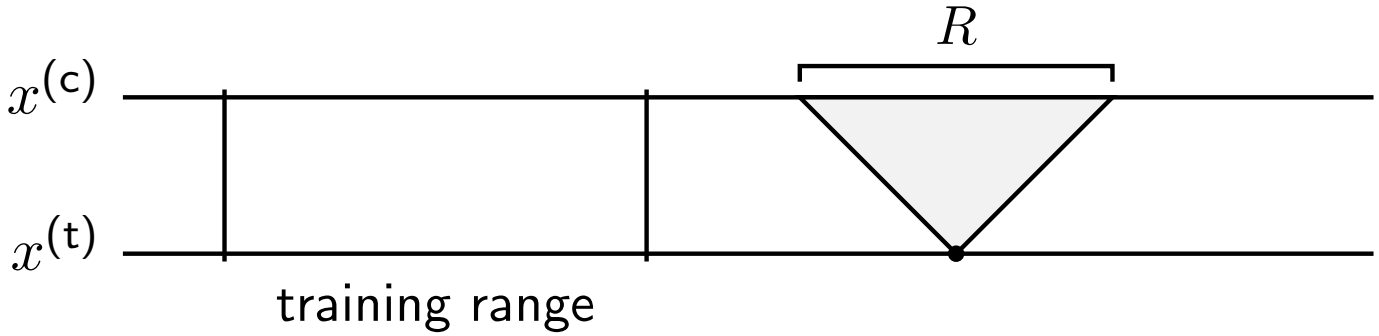
Why Does TE Help Generalise?

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

- CNNs have receptive field $R > 0$:



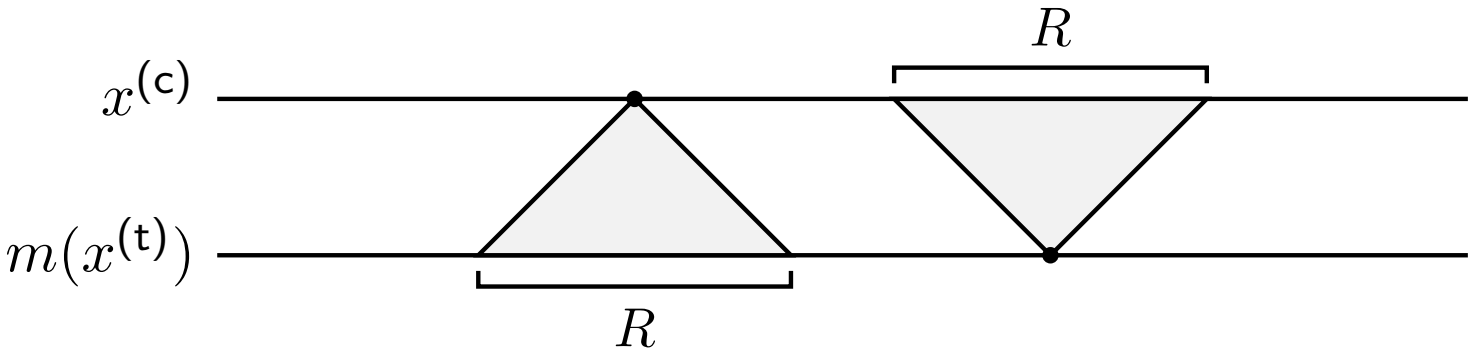
- In combination with TE, helps ConvCNP to generalise:



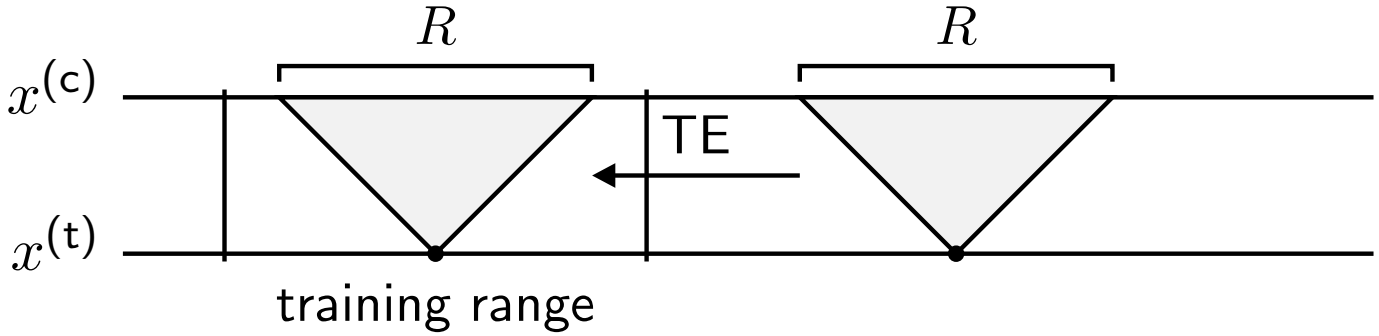
Why Does TE Help Generalise?

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

- CNNs have receptive field $R > 0$:



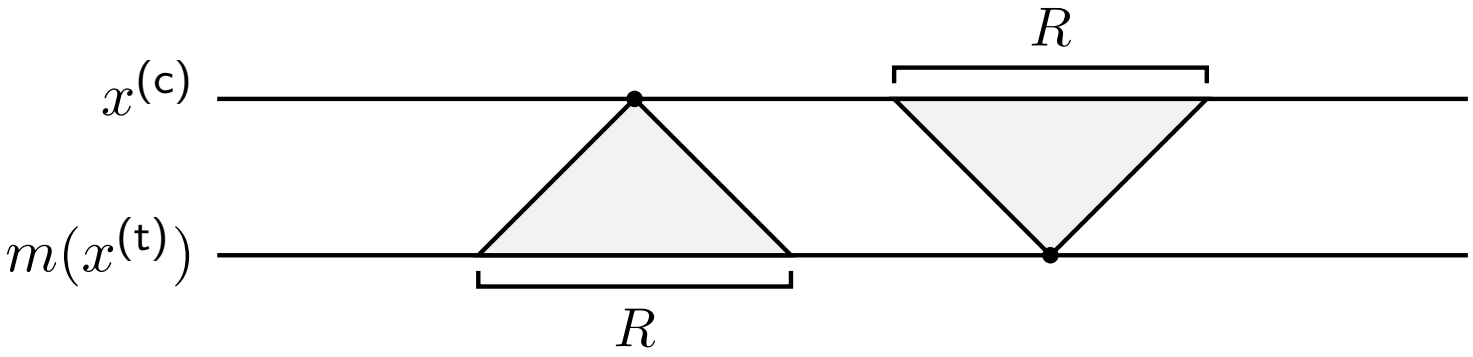
- In combination with TE, helps ConvCNP to generalise:



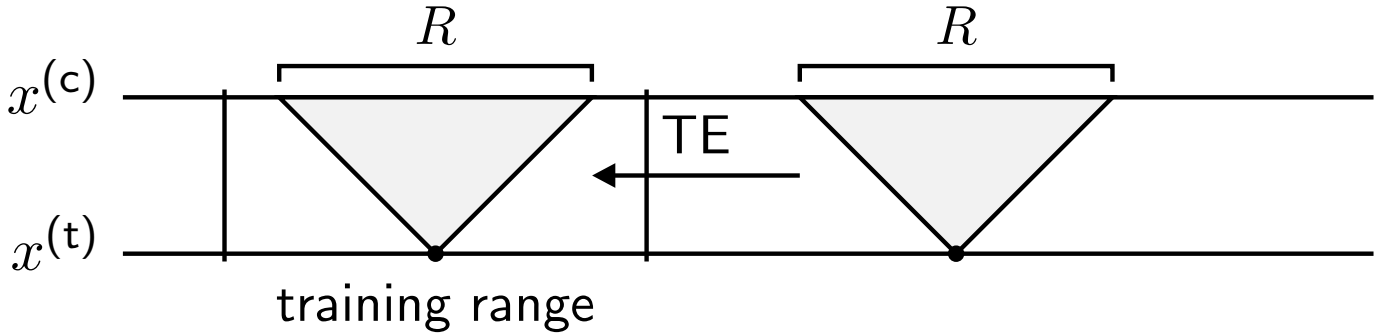
Why Does TE Help Generalise?

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

- CNNs have receptive field $R > 0$:



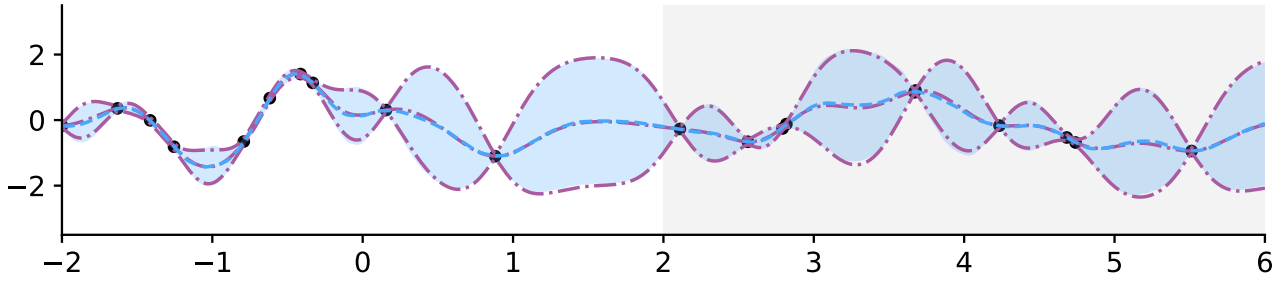
- In combination with TE, helps ConvCNP to generalise:



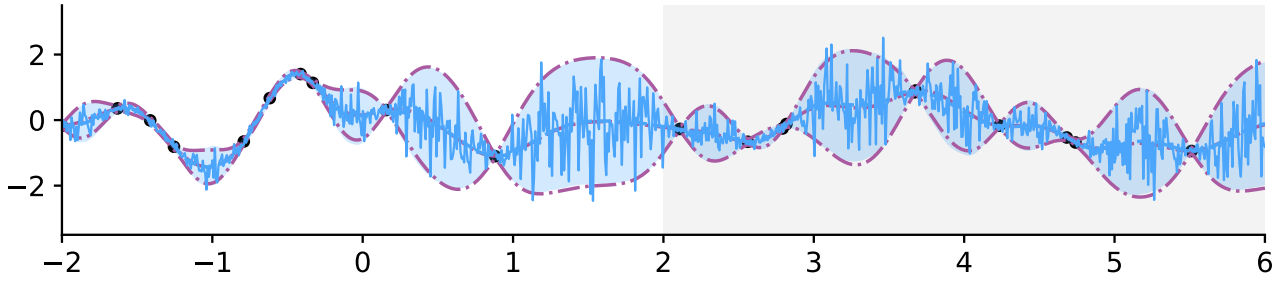
Thm. Suppose that $\mathcal{L}(\pi) \leq \varepsilon$ for data sampled from $[0, R + \ell]$. Then $\mathcal{L}(\pi) \leq \lceil M/\ell \rceil \varepsilon$ for data from any interval of width M.

Further Improvements

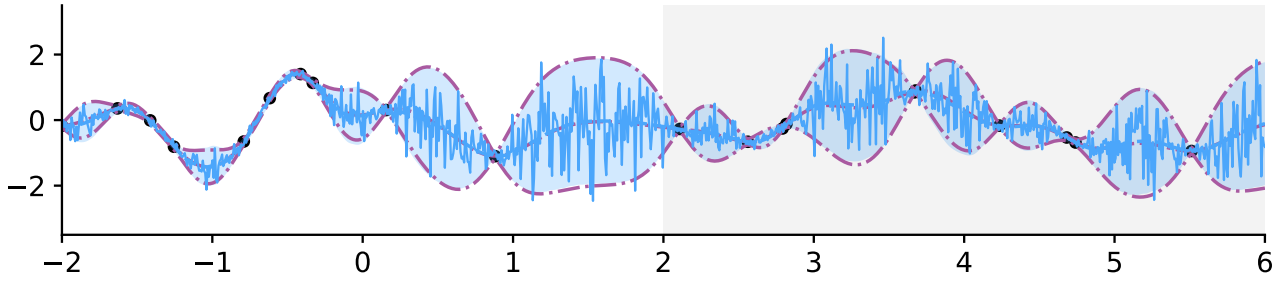
Gaussian TE Prediction Maps



Gaussian TE Prediction Maps

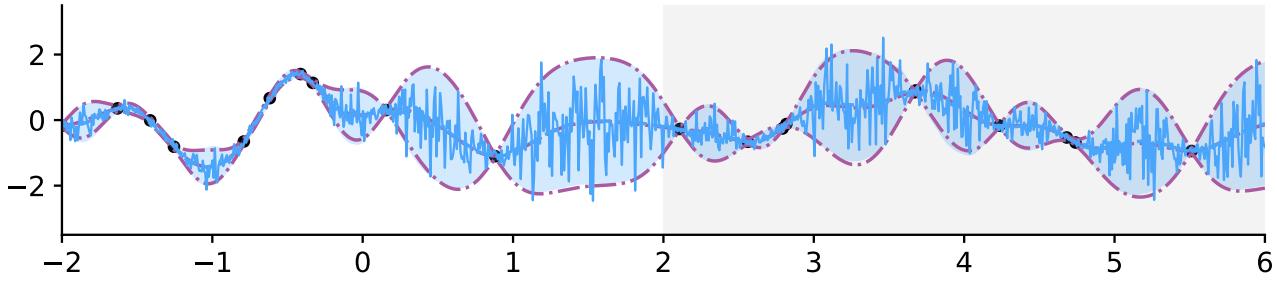


Gaussian TE Prediction Maps



✗ (Conv)CNP fails to model correlations.

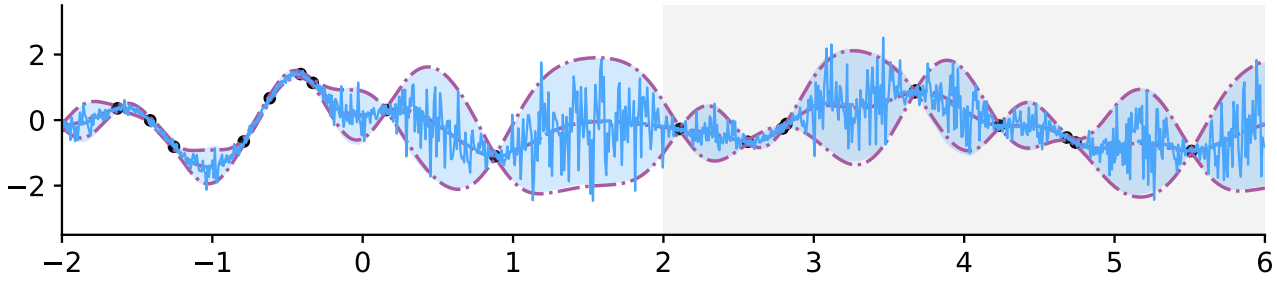
Gaussian TE Prediction Maps



✗ (Conv)CNP fails to model correlations.

- $\mathcal{Q}_{\mathsf{G}} = \{\pi : \mathcal{D} \to \mathcal{P}_{\mathsf{G}}\}$ instead of $\mathcal{Q}_{\mathsf{G,MF}} = \{\pi : \mathcal{D} \to \mathcal{P}_{\mathsf{G,MF}}\}$?

Gaussian TE Prediction Maps

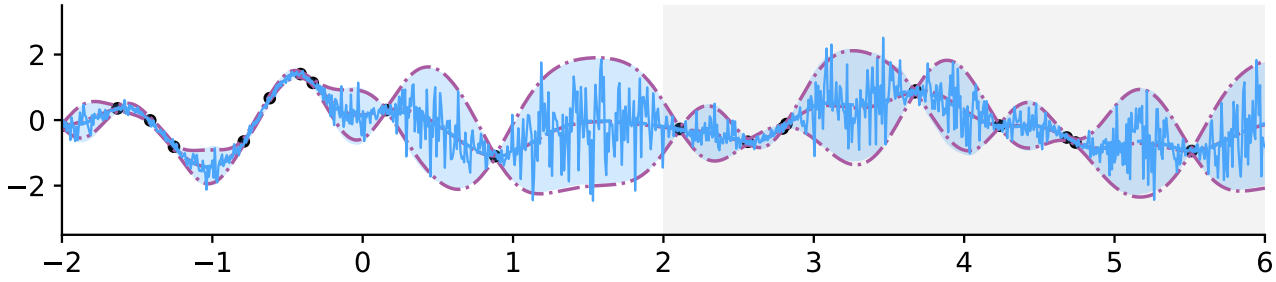


✗ (Conv)CNP fails to model correlations.

- $\mathcal{Q}_{\mathsf{G}} = \{\pi \colon \mathcal{D} \to \mathcal{P}_{\mathsf{G}}\}$ 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:

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

Gaussian TE Prediction Maps



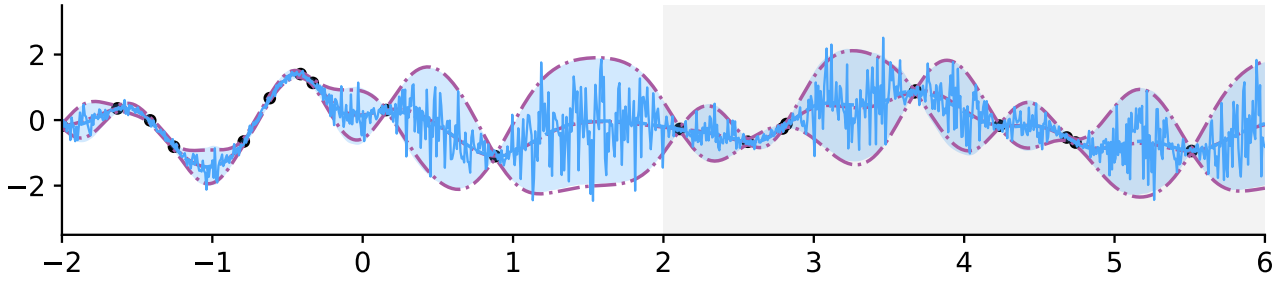
✗ (Conv)CNP fails to model correlations.

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

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

✓ Exploits TE using CNNs, learns quickly, and generalises well

Gaussian TE Prediction Maps



✗ (Conv)CNP fails to model correlations.

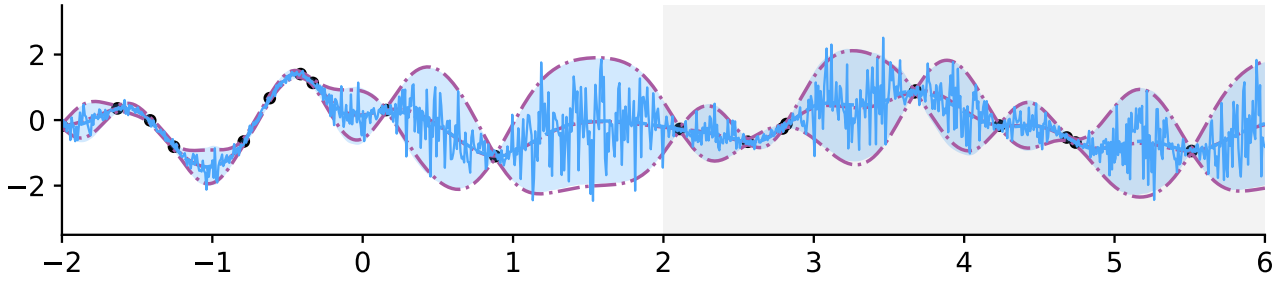
- $\mathcal{Q}_{\mathsf{G}} = \{\pi \colon \mathcal{D} \to \mathcal{P}_{\mathsf{G}}\}$ instead of $\mathcal{Q}_{\mathsf{G,MF}} = \{\pi \colon \mathcal{D} \to \mathcal{P}_{\mathsf{G,MF}}\}$?
- Bruinsma et al. (2021) establishes repr. thm for kernel map:

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

✓ Exploits TE using CNNs, learns quickly, and generalises well

✗ d-dimensional inputs require $2d$-dimensional convolutions

Gaussian TE Prediction Maps



✗ (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. (2022) provide practical params for $d > 1$:

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

TE

Consistency of Prediction Map Approximation

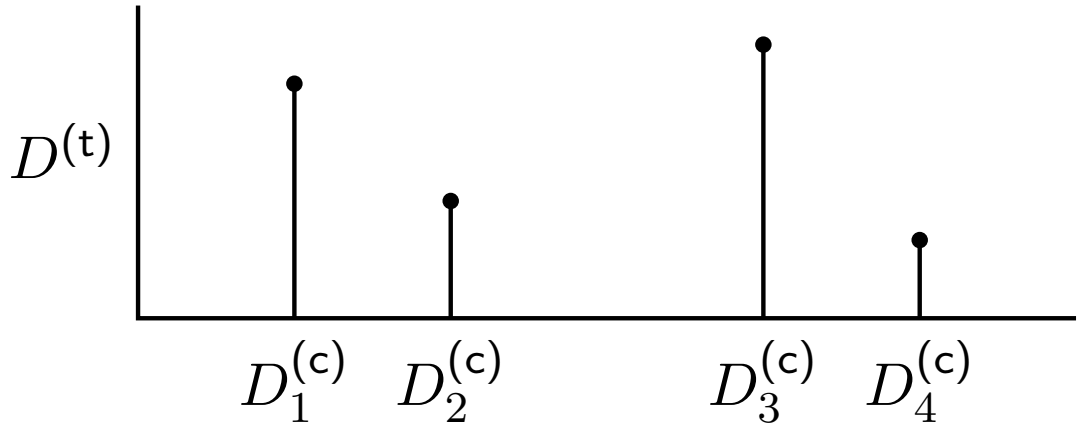
The Problem

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

The Problem

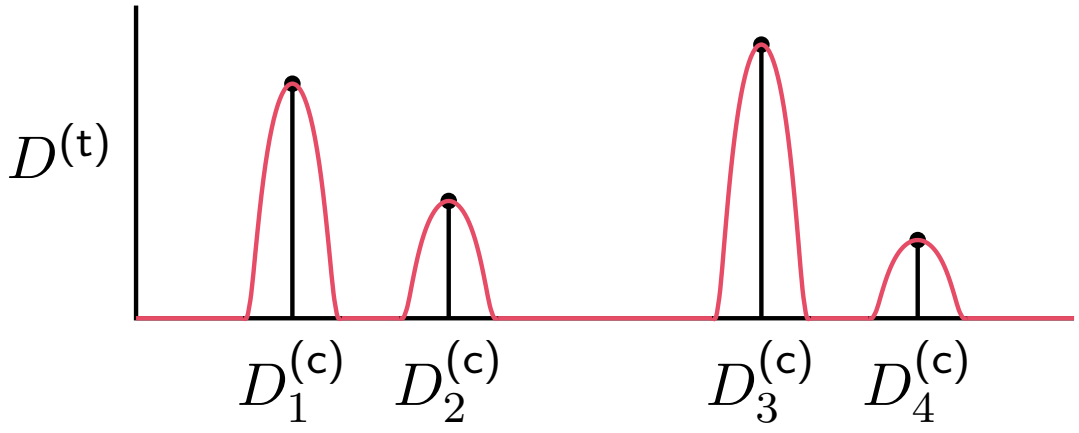
$$\mathcal{L}_n(\pi) = -\frac{1}{N}\sum_{n=1}^{N} \log q(D_n^{(\mathsf{t})} \mid D_n^{(\mathsf{c})}) \text{ “} \approx \frac{1}{N}\sum_{n=1}^{N} (D_n^{(\mathsf{t})} - f(D_n^{(\mathsf{c})}))^2 \text{ ”}$$

The Problem



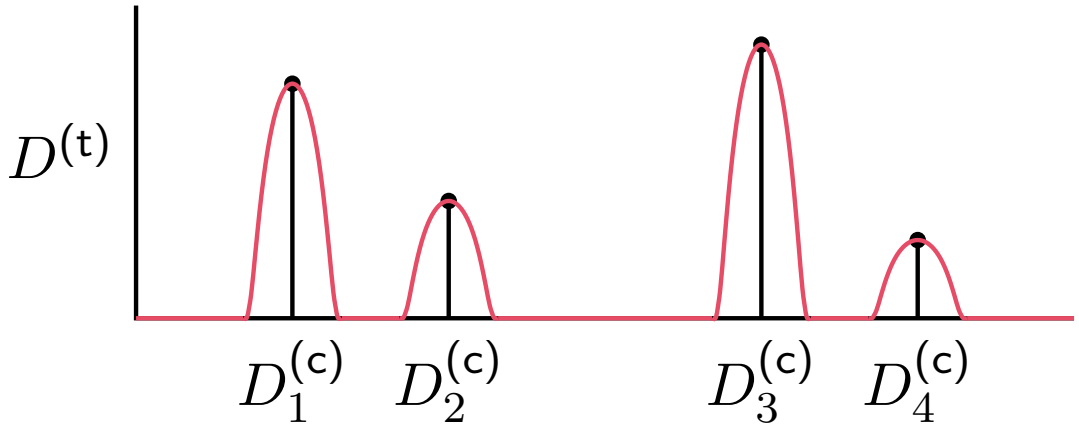
$$\mathcal{L}_n(\pi) = -\frac{1}{N}\sum_{n=1}^{N} \log q(D_n^{(\mathsf{t})} \mid D_n^{(\mathsf{c})}) \text{ “}\approx \frac{1}{N}\sum_{n=1}^{N} (D_n^{(\mathsf{t})} - f(D_n^{(\mathsf{c})}))^2 \text{ ”}$$

The Problem



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

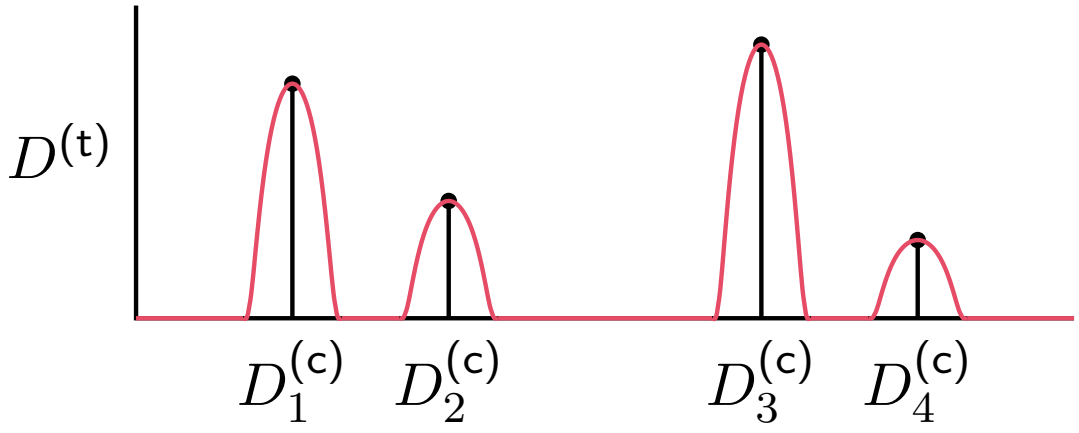
The Problem



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

$\Rightarrow$ Cannot optimise $\mathcal{L}_n(\pi)$ over $\pi \in \mathcal{M}_{\mathsf{G}}$: overfitting!

The Problem

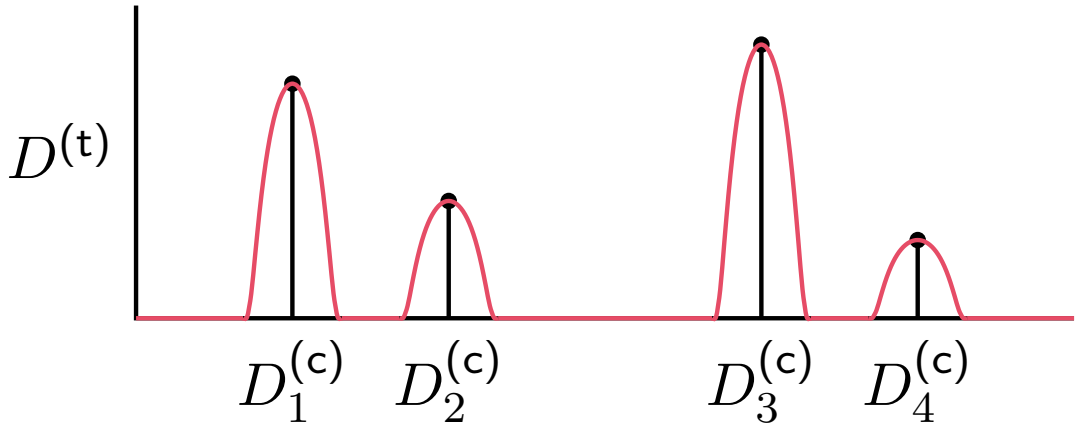


$$\mathcal{L}_n(\pi) = -\frac{1}{N}\sum_{n=1}^{N} \log q(D_n^{(t)} \mid D_n^{(c)}) \text{ “}\approx \frac{1}{N}\sum_{n=1}^{N}(D_n^{(t)} - f(D_n^{(c)}))^2\text{ ”}$$

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

- Practice: tune NN capacity using black magic.

The Problem



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

$\Rightarrow$ Cannot optimise $\mathcal{L}_n(\pi)$ over $\pi \in \mathcal{M}_{\mathsf{G}}$: overfitting!

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

Assumptions on Ground-Truth Process

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

Assumptions:

Assumptions on Ground-Truth Process

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.

Assumptions on Ground-Truth Process

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

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

Assumptions on Ground-Truth Process

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: $|||\mathcal{D}||| := \sup\{|\mathbf{x}| \vee \|\mathbf{y}\|_\infty : (\mathbf{x}, \mathbf{y}) \in \mathcal{D}\} < \infty$.

Assumptions on Ground-Truth Process

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 $\gamma > 0$.

Assumptions on Ground-Truth Process

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 $\sigma^2 \in [\underline{\sigma}^2, \overline{\sigma}^2]$.

A Compact Variational Family

- Identify every $\pi \in \mathcal{M}_{\mathsf{G}}$ with

$$m\colon \mathcal{X} \times \mathcal{D} \to \mathbb{R}, \quad k\colon \mathcal{X} \times \mathcal{X} \times \mathcal{D} \to \mathbb{R}, \quad \sigma^2 \in [\underline{\sigma}^2, \overline{\sigma}^2].$$

A Compact Variational Family

- Identify every $\pi \in \mathcal{M}_{\mathsf{G}}$ with

$$m\colon \mathcal{X} \times \mathcal{D} \to \mathbb{R}, \quad k\colon \mathcal{X} \times \mathcal{X} \times \mathcal{D} \to \mathbb{R}, \quad \sigma^2 \in [\underline{\sigma}^2, \overline{\sigma}^2].$$

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

$$\pi_{\mathsf{MM}} \in \left\{ \pi \in \mathcal{M}_{\mathsf{G}} \;\middle|\; \begin{array}{c} |m(x_1,D_1) - m(x_2,D_2)| \le L^*(|x_1 - x_2|, \|D_1 - D_2\|) \\ |k(x_1,D_1) - k(x_2,D_2)| \le L^*(|x_1 - x_2|, \|D_1 - D_2\|) \\ \|m\|_\infty, \|k\|_\infty \le M^* \end{array} \right\}.$$

A Compact Variational Family

- Identify every $\pi \in \mathcal{M}_{\mathsf{G}}$ with

$$m\colon \mathcal{X} \times \mathcal{D} \to \mathbb{R}, \quad k\colon \mathcal{X} \times \mathcal{X} \times \mathcal{D} \to \mathbb{R}, \quad \sigma^2 \in [\underline{\sigma}^2, \overline{\sigma}^2].$$

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

$$\pi_{\mathsf{MM}} \in \left\{ \pi \in \mathcal{M}_{\mathsf{G}} \;\middle|\; \begin{array}{c} |m(x_1,D_1) - m(x_2,D_2)| \le L^*(|x_1 - x_2|, \|D_1 - D_2\|) \\ |k(x_1,D_1) - k(x_2,D_2)| \le L^*(|x_1 - x_2|, \|D_1 - D_2\|) \\ \|m\|_\infty, \|k\|_\infty \le M^* \end{array} \right\}.$$

- Call this collection $\mathcal{Q}^*$.

A Compact Variational Family

- Identify every $\pi \in \mathcal{M}_{\mathsf{G}}$ with

$$m\colon \mathcal{X} \times \mathcal{D} \to \mathbb{R}, \quad k\colon \mathcal{X} \times \mathcal{X} \times \mathcal{D} \to \mathbb{R}, \quad \sigma^2 \in [\underline{\sigma}^2, \overline{\sigma}^2].$$

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

$$\pi_{\mathsf{MM}} \in \left\{ \pi \in \mathcal{M}_{\mathsf{G}} \;\middle|\; \begin{array}{c} |m(x_1,D_1) - m(x_2,D_2)| \le L^*(|x_1 - x_2|, \|D_1 - D_2\|) \\ |k(x_1,D_1) - k(x_2,D_2)| \le L^*(|x_1 - x_2|, \|D_1 - D_2\|) \\ \|m\|_\infty, \|k\|_\infty \le M^* \end{array} \right\}.$$

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

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

A Compact Variational Family

- Identify every $\pi \in \mathcal{M}_{\mathsf{G}}$ with

$$m\colon \mathcal{X} \times \mathcal{D} \to \mathbb{R}, \quad k\colon \mathcal{X} \times \mathcal{X} \times \mathcal{D} \to \mathbb{R}, \quad \sigma^2 \in [\underline{\sigma}^2, \overline{\sigma}^2].$$

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

$$\pi_{\mathsf{MM}} \in \left\{ \pi \in \mathcal{M}_{\mathsf{G}} \;\middle|\; \begin{array}{c} |m(x_1,D_1)-m(x_2,D_2)| \le L^*(|x_1-x_2|, \|D_1-D_2\|) \\ |k(x_1,D_1)-k(x_2,D_2)| \le L^*(|x_1-x_2|, \|D_1-D_2\|) \\ \|m\|_\infty, \|k\|_\infty \le M^* \end{array} \right\}.$$

- 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.

Consistency of Prediction Map Approximation

Thm. Let

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

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

Consistency of Prediction Map Approximation

Thm. Let

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

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

Pending questions:

Consistency of Prediction Map Approximation

Thm. Let

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

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

Pending questions:

- $\mathcal{Q}_{\mathsf{NN}} = \{(m_\theta, k_\theta, \sigma^2) : \theta \in \mathbb{R}^P\}$?

Consistency of Prediction Map Approximation

Thm. Let

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

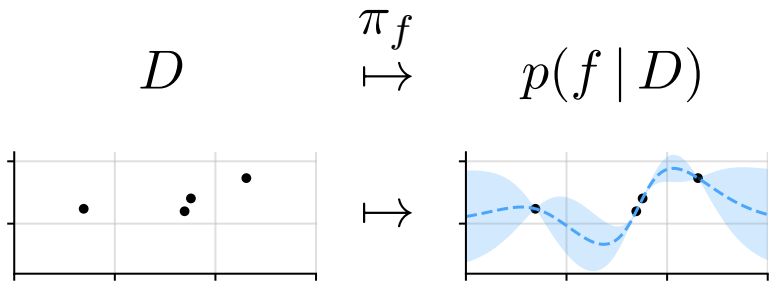
Then, almost surely, $\pi_n(D) \rightharpoonup \pi_{\mathsf{MM}}(D)$ for all $D \in \mathcal{D}$.

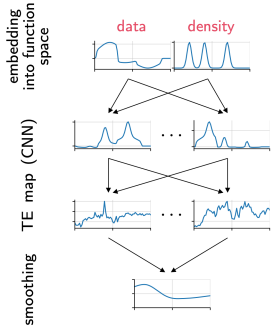
Pending questions:

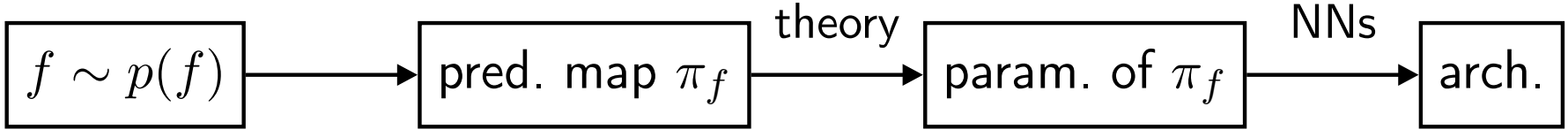
- $\mathcal{Q}_{\mathsf{NN}} = \{(m_\theta, k_\theta, \sigma^2) : \theta \in \mathbb{R}^P\}$?
- How much data: finite-sample bounds / rates of convergence?

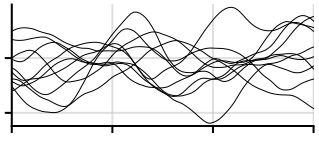
Wrapping Up

Prediction Map Approximation









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

- ✓ 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=Skey4eBYPS.

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. 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.