

Informed Meta-Learning Learning the inductive bias from human expert knowledge

Katarzyna Kobalczyk Mihaela van der Schaar

Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, United Kingdom

Cons: Handcrafted biases require significant engineering efforts and need to be customized for every new learning problem. A.1. Architectural and training determining details for $\frac{1}{2}$

Informed machine-learning: from knowledge to inductive biases. Machine learning practitioners map expert knowledge to prior probabilities over functions by manual model design. Regions of the function space with a non-zero prior probability define the hypothesis space and the relative probabilities of prior solutions dictate the inductive biases of the model. Knowledge about different learning tasks is associated with distinct inductive biases. **Pros:** Models with well-designed task-specific biases require less training data and are more robust.

Pros: Meta-learning enables flexible learning of during met complex inductive biases otherwise difficult to design by hand.

Cons: Success of meta-learners depends on the *a*(*t*) *r* similarity of tasks at meta-train and test time. Heterogenous task distributions fail to provide strong inductive biases. inductive biases.

 h_1

Informed Neural Processes. We instantiate an informed meta-learner by extending the family of Neural Processes—probabilistic, fully amortized meta-learners. tions given observed data, and in the informed meta-learning **scenario informed ineural Processes.** We instantiate an informed meta-tearner by external processes. As distinctive full association density decays are also have a section of the executive decays are also have a section o -propapilistic, rutty arriortized ineta-teamers. Informed Neural Processes, We instantiate an informed me

i 2 *C i* 2 *T* $\vert k \vert$

tasks. Moreover, the fact that NPs model a distribution over

functions, instead of returning a single, MLE prediction en-

ables us to measure the reduction in uncertainty about solu-

 \mathcal{K} h_2 **Figure 2.** *Figure 2. Figure 2. Figure 2. Comparison of Temarchical Comparison of Temarism 2. The Second 2* knowledge to inductive biases, $\mathcal{K} \to p(f|\mathcal{K})$, is **is one ta-learned by training over multiple learning** tasks and their representations of knowledge. The stask-relevant regions. learners, forming the foundation $\mathbf f$ this view, the meta-knowledge, \mathcal{C} , can be represented with \mathcal{C} , can be represented with \mathcal{C}

 \cdots and the expertison on the expert \cdots $\$ concentrating the meta-distribution around the task-relevant regions. $\sqrt{2\pi}$ $\sqrt{4\pi\epsilon_1\epsilon_2\epsilon_3\epsilon_4}$

Meta-learning as inductive bias learning. Instead of manual inductive bias specification, the paradigm meta-learned by training over a distribution of *ⁱ h*1(*xi, yi*). related tasks. *r* = P

p(*y | x, z*)*q*(*z | r^C*)*dz.*

*ⁱ*2*^C ^h*(*xi, yi*). The variational distribu-

Method with a single forward pass through networks \mathbf{h} and \mathbf{h} parameters of these two networks are estimated by extensive are estimated by episodic structure are estimated b 2. concat & MLP: *a*(*r, k*) = MLP([*r||k*]), 3. MLP & FiLM: *a*(*r, k*) = FiLM(*k*) [MLP(*r*)]. We use the idea of modulation parameters introduced by (Perez et al.,

Training. INPs are trained in an episodic fashion over a distribution of learning to and their associated knowledge representations ${\mathcal K}_j$. Denoting by r_C and r_T the context and target data representations and by k the knowledge embedding vector of a single task, parameters of the model are learned by maximising the expectation of ELBO over all training tasks, **Training.** INPs are trained in an episodic fashion over a distribution of learning task ${\cal T}_j = \{{\cal D}_C, {\cal D}_T\}$ = *p*(*y^T | x^T , z*) *q*(*z | r^T , k*) σ *a f* learning took τ (0 0) *q*(*z | r^T , k*)

 $\int_0^{\infty} \frac{\eta(\eta)}{\eta(\eta)} d\eta$ from a variational distribution *q*. That is, each sample *z* ⇠ *q* $t \rightarrow \pi$ **f** $(1 \quad \langle 1 \quad 1 \quad 1 \quad D \quad (\langle 1 \quad 1 \rangle) \rangle$ $\log p(y_T | x_T, r_C, k) \geq \mathbb{E}_{q(z|r_T, k)} [\log p(y_T | x_T, z)] - D_{\text{KL}}(q(z | r_T, k) || q(z | r_C, k))$

The night will start off cold with temperatures falling to -8.9°C The night will start off cold with temperatures falling to -8.9°C $b = 0$ by late morning, and then gradually rise to a high of 1.6°C in the late afternoon. Temperatures will start to drop again in the
evening reaching -3 1°C by midnight evening, reaching -3.1°C by midnight. T_{he} night will stay of cold with tamperatures folling to $0.09C$ when $\frac{1}{\sqrt{K}}$ is one of $\frac{1}{\sqrt{K}}$ in $\frac{1}{\sqrt{K}}$ is one of $\frac{1}{\sqrt{K}}$ in $\frac{1}{\sqrt{K}}$ in $\frac{1}{\sqrt{K}}$ in $\frac{1}{\sqrt{K}}$ is $\frac{1}{\sqrt{K}}$ in $\frac{1}{\sqrt{K}}$ in $\frac{1}{\sqrt{K}}$ is $\frac{1}{\sqrt{K}}$ in $\frac{1}{\sqrt{K}}$ in $\frac{1}{\sqrt{K}}$

yⁱ z yⁱ

xi

The night will start off chilly with temperatures around 0.5°C, The night will start off chilly with temperatures around 0.5°C, The night will start by chilly with temperatures around 0.5 °C,
but it will drop to -1.7 °C by early morning. The day will $\frac{d}{dx}$ gradually warm up, reaching a high of 5.1°C in the afternoon before cooling off to 1.0° C by midnight. when it is one of the *start off* shills with town our two special 0.500

> *The night will start off cold with temperatures falling to -16.8°C The night will start off cold with temperatures falling to -16.8°C* The night will start by cold with temperatures juiting to -10.8 C
by dawn, and the day will continue to get colder, reaching a by dawn, and the day will continue to get colder, reacting a
chilly -23.0°C by midnight. Afternoon temperatures will hover α around -18.5°C, so bundle up if you're heading out. *K* = printet will at out of explorate to prove continue fulling to 16,000 start off cold with temperatures falling to -16.8° C

 \Box The night will be bitterly cold with temperatures around -18 degrees, gradually increasing to -14 degrees by late afternoon. *The temperature will slightly drop again to -15 degrees in the* The temperature with stightly arop again to -13 degrees in the evening, warming up a bit to -14 degrees at midnight. Rueden et al., 2023a), where "-likelihood" is defined as: $\frac{1}{20}$ 4.1.1. DATA EFFICIENCY AND TASK DISTRIBUTION ϵ 20 ϵ even de bitterly cold with temperatures around -18
ually inexessing to *14 degrees* by late afternoon *ning up a bit to -14 degrees at midnight.*

respect to the union to the union terms in the union terms in the union of the union \mathcal{A}

respect to the union to the union $\mathcal{L}_{\mathcal{A}}$

to computing the posterior *p*(*f | D^C*), which is obtained

Model. (I)NPs model the distribution over functions through a fixed dimensional latent variable z sampled from a variational distribution q . INPs extend NPs by additionally conditioning q on the z sampled nom a vanational distribution *q*. hvrs extend ivrs by additionally c
information contained in expert knowledge. We model the predictive posterior distr predictive distribution and the distribution of the distribution of the distribution of the distribution of the
The distribution of the distri information contained in expert knowledge. We model the predictive posterior distribution as:

training over a distribution of tasks.

k = *h*2(*K*) and *r*⁰

^C = *a*(*r^C , k*), INPs model (4) as:

= *p*(*y | x, r*⁰

^C) = ^Z

p(*y | x, z*)*q*(*z | r*⁰

^C)*dz.* (6)

$$
p(y \mid x, \mathcal{D}_C, \mathcal{K}) := \int p(y \mid x, z) q(z \mid \mathcal{D}_C, \mathcal{K}) dz.
$$

2018). Here *a* is an MLP whose parameters are modulated with a modulated with the outputs of *h*2.

10

 θ

Introduction Results **Data efficiency, robustness and uncertainty.** *n* $n = 0$? *{a} {b} {a, b} n* $n = 1$ $f(x) = ax + sin(bx) + c$ Number of context points representations *K^j* . To train and evaluate an INP model we sample training, validation and testing collections of tasks. Each task, (omitting the dependence on *j* for clarity), consists of a labeled context dataset *D^C* . and the target dataset **Drawing the target dataset are the target of the target dataset are the target of the target dataset are the street of the goal of the target of the goal of the** \mathbb{H} \bullet ⁰ 0 $\frac{1}{200}$ \mathcal{I} $\mathbb{H}^ \mathbb{L}$ 0 $\frac{1}{00}$ \overline{I} representations *K^j* . To train and evaluate an INP model we sample training, validation and testing collections of tasks. Each task, (omitting the dependence on *j* for clarity), consists of a labeled context dataset *D^C* . and the target dataset **D** $\frac{3}{5}$. The set of the target dataset are the goal of the -100 -50 θ Log likelihood -400 -200 θ $\overline{}$ representations *K^j* . To train and evaluate an INP model we \mathbb{Z}^2 sample training, validation and the tasks. $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ $s_{\frac{1}{2}} = s_{\frac{1}{2}} + s_{\frac{1}{2}}$ \sum_{α} $\frac{1}{\sqrt{2}}$ 50 Log likelihood 400 200 $\overline{0}$ Log likelihood

Beyond 1D regression. We extend INPs to the task of few-shot classification with expert knowledge in the form class-level features or descriptions in natural language. Successful alignment of knowledge and data rer \inf ansser states, zero shot statesmeater, and improved for shot present 5 randomly sampled in a randomly sampled in a set of one captions per tangent GPT-4 to generate short descriptions of \mathcal{L}_1 present 5 randomly sampled in a randomly sampled in a series of the class and prompt GPT-4 to generate short descriptions of the class and prompt GPT-4 to generate short descriptions of the class and prompt GPT-4 to genera Informed meta-learning data representations facilitates robust generalization to new, previously when and improved lew-shot classes, zero-shot classification and improved lew-shot and analyzed the parameters i
The plassification accuracy sampled according to: **a** *U*[1*,* 1], *b u*₁, *u*₁, ✏*ⁱ* ⇠ *N* (0*,* 0*.*2). The parameters *a, b, c* are randomly sampled according to: *a* ⇠ *U*[1*,* 1], *b* ⇠ *U*[0*,* 6], or none (*K* =?) of the parameters *a*, *b*, or *c*. The *f*(*x*) = *aximum* $\frac{1}{2}$ = *aximum* sample $\frac{1}{2}$ = *b*_c, *c*, *t* for some *f* should propled values of the parameters *a, b, c*. We introduce a Classification with expert knowledge in the form class-tevel reatures of and *descriptions in petural lenguage*. Successful elignment of knowledge and abscriptions in natural language. Successiul alignment or knowledge and
data_representations_facilitates_robust_ceneralization_to_new_previously sampled according to: **a**nd the according to: **a**
Ince in the according to: *Dace of the class* unseen classes, zero-shot classification and improved few-shot
classification.accuracy **Beyond 1D regression.** We extend INPs to the task of few-shot data representations facilitates robust generalization to new, previously classification with expert knowledge in the form class-level features or descriptions in natural language. Successful alignment of knowledge and classification accuracy.

 24^{00} 06^{00} 12^{00} 18^{00} 24^{00}

ing white eyes and a dull coat of gray.

by its black body, webbed feet, a bright orange bill with an inverted feather curl at the base, piercing white eyes with a distinctive stripe, and a dull grey

On a series of controlled synthetic experiments, we:

- on a conce or controlled synthetic exportments, we.
• Demonstrate how successful integration of oracle knowledge about the data generating process improves data efficiency and mitigate the **adverse effects of train-test distribution shift.** Rugian Of plant NPs and Input \mathbf{b} between meta-training tasks. Knowledge integration tasks. Knowledge integration \mathbf{b} en a series of controlled symmetric experiments, we.
• Demonstrate how successful integration of oracle knowledge about the achitogration of placed knowledge about the hances data efficiency. And integate the between meta-training and testing tasks. Knowledge integration of the second tasks. Knowledge integration of the second tasks. And the second tasks of the second tasks of the second tasks. And the second tasks of the secon data generating process improves data efficiency and mitigate the OF OF actual NPS and INPS. The plant of plant integration of plants in the set of plants of plants of plants in $\mathsf n$ shift. Knowledge integration tasks. Knowledge integration tasks. Knowledge integration
- **•** We derive methods to quantitatively measure the impact of expert knowledge on the reduction in epistemic uncertainty. ratively measure the **impact of expert** At the Mederive methods to andiducty measure the **impact of expert** At the number of context and the number of context and target and ta knowledge representations by setting *k* = 0. This allows reasure the **impact of expert** the straining and testing tasks.

hefit of meta-learned prior is sed
Exam patural language to functional priors la prestigal seeperios **The experimental section is dividend into two parts.** Figures of into two parts on into two parts of into two parts **phartial, included** sucharios, predictive functions are difficult to modet with closed-form mathematical learn how to put a strong prior on the function's slope, alli to map accompted This setup simulates a scenario, in which *K* contains **partial, included** in production bounding, learn to put to the month prior on the function. level of oscillations and bias. experiments of the prediction of the set of
Experiments of the set **flexibility**. We show how INPs learn to map descriptions in natural From natural language to functional priors. In practical scenarios, predictive functions are difficult to model with closed-form mathematical

 \exists^0

 θ

 $\frac{-20}{2}$

 -20

BARRY -

SHIPS OF THE SHIP

 $\Rightarrow 0$

 \longrightarrow

 12^{00}

points are sampled according to the following pro-

0 and 10; the number of targets is set to *m* = 100.

points are sampled according to the following pro-

 $200 \t 06^{00} \t 12^{00} \t 18^{00} \t 24^{00} \t 24^{00} \t 06^{00} \t 12^{00} \t 18^{00} \t 24^{00}$

 24° 06 $^{\circ}$ 12 $^{\circ}$ 18 $^{\circ}$ 24 $^{\circ}$ 24 $^{\circ}$ 06 $^{\circ}$ 12 $^{\circ}$ 18 $^{\circ}$ 24 $^{\circ}$

0 and 10; the number of targets is set to *m* = 100.

cess. A function *f* is sampled from the family of

 $\equiv 0$

Temperature [10]

0 and 10; the number of targets is set to *m* = 100.