Bayesian Context Aggregation for Neural Processes

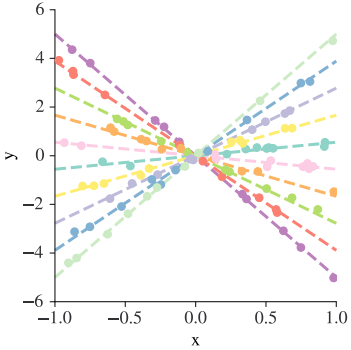
ICLR 2021

Michael Volpp[1,2], Fabian Flürenbrock[1], Lukas Grossberger[1], Christian Daniel[1], Gerhard Neumann[2]

[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]Karlsruhe Institute of Technology, Karlsruhe, Germany

Probabilistic Regression as a Multi-task Learning Problem

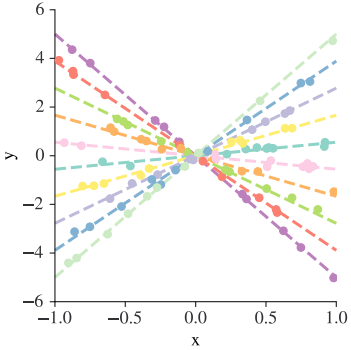
- Family $\mathbf{F}$ of functions $f \sim \mathbb{P}(\mathbf{R}^{d_x} \to \mathbf{R}^{d_y})$ ("tasks") with some form of shared structure



ICLR 2021 *j* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***

Probabilistic Regression as a Multi-task Learning Problem

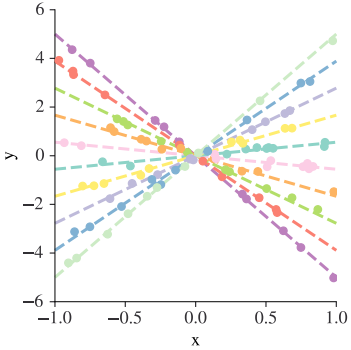
- Family $\mathbf{F}$ of functions $f_i \sim \mathbf{F}: \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ ("tasks") with some form of shared structure
- Noisy evaluations $D_i = \{(x_{i,j}, y_{i,j})\}_j$ with $y_{i,j} = f_i(x_{i,j}) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma_n^2)$



ICLR 2021 *j* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***

Probabilistic Regression as a Multi-task Learning Problem

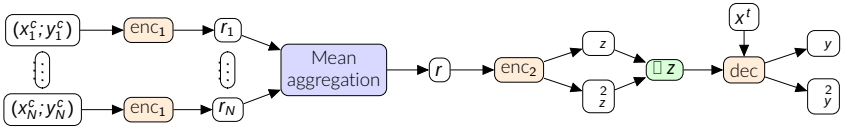
- Family $\mathbf{F}$ of functions $f_i : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ ("tasks") with some form of shared structure
- Noisy evaluations $D_i = \{(x_{i,j}, y_{i,j})\}_j$ with $y_{i,j} = f_i(x_{i,j}) + \varepsilon, \ \varepsilon \sim N(0, \sigma_n^2)$
- Learn the predictive distribution $p(y^t | x^t, D^c)$ over target outputs, conditioned on a context set $D^c \subseteq D_i$



ICLR 2021 *j* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***

The Neural Process[1]

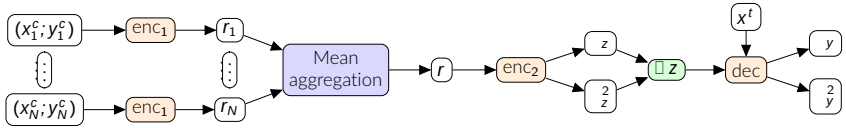
Neural network (NN)-based multi-task learning architecture



[1] Garnelo et al., ***Conditional Neural Processes***, ICML 2018, Garnelo et al. ***Neural Processes***. arxiv/1807.01622 2018

The Neural Process[1]

Neural network (NN)-based multi-task learning architecture



1. Infer latent representation $p(z|D^c)$ of the target function from $D^c = \{(x_n^c; y_n^c)\}_{n=1}^N$

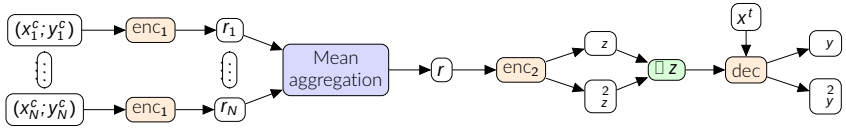
[1] Garnelo et al., ***Conditional Neural Processes***, ICML 2018, Garnelo et al. ***Neural Processes***. arxiv/1807.01622 2018

ICLR 2021 *j* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***

KIT BOSCH

The Neural Process[1]

Neural network (NN)-based multi-task learning architecture

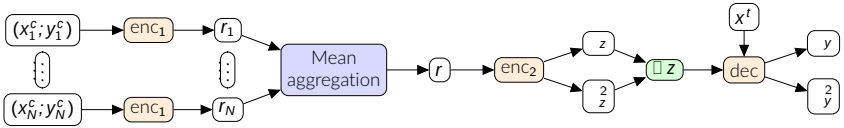


1. Infer latent representation $p(z|D^c)$ of the target function from $D^c = \{(x_n^c; y_n^c)\}_{n=1}^N$
   a. Map each context tuple $(x_n^c; y_n^c)$ onto a latent observation $r_n = \text{enc}_1(x_n^c; y_n^c)$

[1] Garnelo et al., ***Conditional Neural Processes***, ICML 2018, Garnelo et al. ***Neural Processes***. arxiv/1807.01622 2018

The Neural Process[1]

Neural network (NN)-based multi-task learning architecture

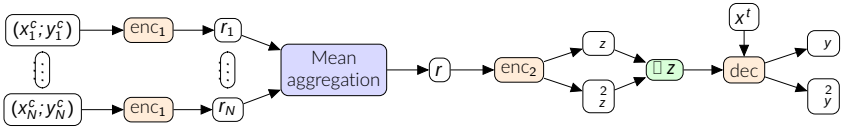


1. Infer latent representation $p(z|D^c)$ of the target function from $D^c = \{(x_n^c; y_n^c)\}_{n=1}^N$
   a. Map each context tuple $(x_n^c; y_n^c)$ onto a latent observation $r_n = \text{enc}_1(x_n^c; y_n^c)$
   b. Form an aggregated latent observation r using ***mean aggregation***: $r = \frac{1}{N} \sum_{n=1}^{N} r_n$

[1] Garnelo et al., ***Conditional Neural Processes***, ICML 2018, Garnelo et al. ***Neural Processes***. arxiv/1807.01622 2018

The Neural Process[1]

Neural network (NN)-based multi-task learning architecture



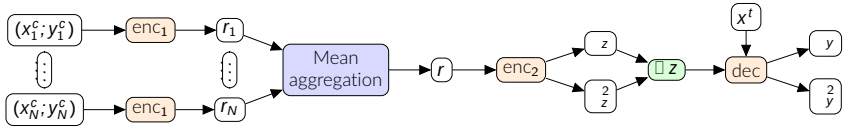
1. Infer latent representation $p(z|D^c)$ of the target function from $D^c = \{(x_n^c; y_n^c)\}_{n=1}^N$
   a. Map each context tuple $(x_n^c; y_n^c)$ onto a latent observation $r_n = \text{enc}_1(x_n^c; y_n^c)$
   b. Form an aggregated latent observation r using ***mean aggregation***: $r = \frac{1}{N} \sum_{n=1}^{N} r_n$
   c. Map r onto the parameters of the latent distribution: $(\mu_z; \sigma_z^2) = \text{enc}_2(r)$

[1] Garnelo et al., ***Conditional Neural Processes***, ICML 2018, Garnelo et al. ***Neural Processes***. arxiv/1807.01622 2018

ICLR 2021 *j* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***

The Neural Process[1]

Neural network (NN)-based multi-task learning architecture



1. Infer latent representation $p(z|\mathbf{D}^c)$ of the target function from $\mathbf{D}^c = \{(x_n^c; y_n^c)\}_{n=1}^N$
   a. Map each context tuple $(x_n^c; y_n^c)$ onto a latent observation $r_n = \text{enc}_1(x_n^c; y_n^c)$
   b. Form an aggregated latent observation r using ***mean aggregation***: $r = \frac{1}{N} \sum_{n=1}^{N} r_n$
   c. Map r onto the parameters of the latent distribution: $(\mu_z; \sigma_z^2) = \text{enc}_2(r)$
2. Map samples $z \sim p(z|\mathbf{D}^c)$ onto a Gaussian output distribution: $(\mu_y; \sigma_y^2) = \text{dec}(z; x^t)$

[1] Garnelo et al., ***Conditional Neural Processes***, ICML 2018, Garnelo et al. ***Neural Processes***. arxiv/1807.01622 2018

Context Aggregation and Task Ambiguity

- Different areas in the $(x; y)$-space can have different ***task ambiguity*** (TA)

Context Aggregation and Task Ambiguity

- Different areas in the $(x; y)$-space can have different ***task ambiguity*** (TA)
- Context tuples $(x_n^c; y_n^c)$ with
  - high TA should have little influence on $p(z \mid \mathbf{D}^c)$
  - low TA should have large influence on $p(z \mid \mathbf{D}^c)$

Context Aggregation and Task Ambiguity

- Different areas in the $(x; y)$-space can have different ***task ambiguity*** (TA)
- Context tuples $(x_n^c; y_n^c)$ with
  - high TA should have little influence on $p(z|\mathbf{D}^c)$
  - low TA should have large influence on $p(z|\mathbf{D}^c)$
- Mean aggregation assigns the same weight to all context tuples: $r = \frac{1}{N} \sum_{n=1}^{N} r_n$

ICLR 2021 *j* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***

Context Aggregation and Task Ambiguity

- Different areas in the $(x; y)$-space can have different ***task ambiguity*** (TA)
- Context tuples $(x_n^c; y_n^c)$ with
  - high TA should have little influence on $p(z|\mathbf{D}^c)$
  - low TA should have large influence on $p(z|\mathbf{D}^c)$
- Mean aggregation assigns the same weight to all context tuples: $r = \frac{1}{N} \sum_{n=1}^{N} r_n$

How to efficiently incorporate task ambiguity into NP parameter inference?

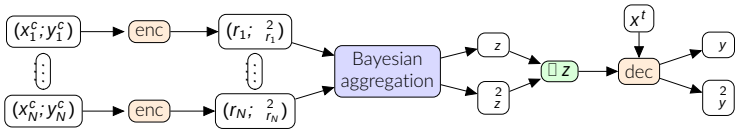
Bayesian Context Aggregation

Context aggregation and parameter inference should be treated as one holistic mechanism!
Directly aggregate the context data into the statistical description of z!

Bayesian Context Aggregation

Context aggregation and parameter inference should be treated as one holistic mechanism!
Directly aggregate the context data into the statistical description of z!

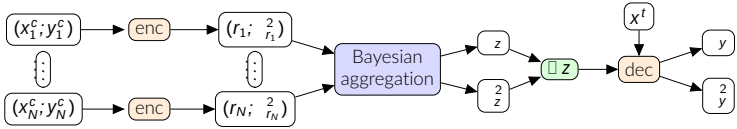
- NP with our ***Bayesian aggregation (BA)***:



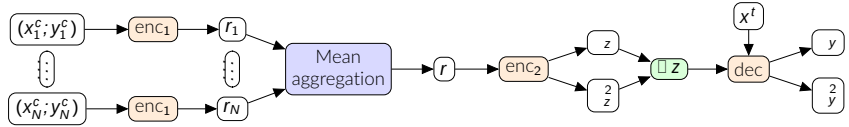
Bayesian Context Aggregation

Context aggregation and parameter inference should be treated as one holistic mechanism!
Directly aggregate the context data into the statistical description of z!

- NP with our ***Bayesian aggregation (BA)***:



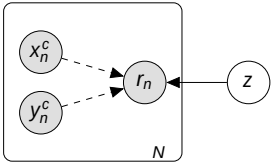
- Compare: NP with traditional mean aggregation (MA):



ICLR 2021 *Workshop* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***

Bayesian Context Aggregation

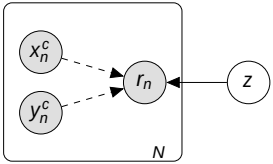
- Context aggregation as Bayesian inference



ICLR 2021 *j* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***

Bayesian Context Aggregation

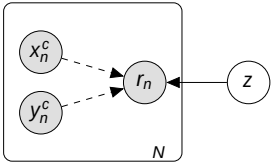
- Context aggregation as Bayesian inference
- Observation model: $p(r_n|z) = \mathcal{N}(r_n|z; \mathrm{diag}(\sigma_{r_n}^2))$



ICLR 2021 *j* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***

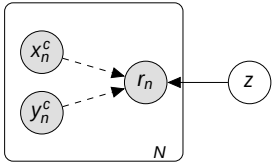
Bayesian Context Aggregation

- Context aggregation as Bayesian inference
- Observation model: $p(r_n \mid z) = \mathcal{N}(r_n \mid z; \mathrm{diag}(\sigma_{r_n}^2))$
- Encoder learns: $(r_n; \sigma_{r_n}^2) = \mathrm{enc}(x_n^c; y_n^c)$



Bayesian Context Aggregation

- Context aggregation as Bayesian inference
- Observation model: $p(r_n|z) = \mathcal{N}(r_n|z; \mathrm{diag}(\sigma^2_{r_n}))$
- Encoder learns: $(r_n; \sigma^2_{r_n}) = \mathrm{enc}(x^c_n; y^c_n)$
- Latent posterior: $p(z|\{r_n\}) = \mathcal{N}(z|\mu_z; \mathrm{diag}(\sigma^2_z))$

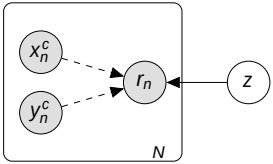


$$\sigma^2_z = \left[\sum_{n=1}^{N} \left(\sigma^2_{r_n}\right)^{-1}\right]^{-1} \quad ; \quad \mu_z = \sum_{n=1}^{N} \frac{\sigma^2_z}{\sigma^2_{r_n}} r_n$$

(for each latent dim.)

Bayesian Context Aggregation

- Context aggregation as Bayesian inference
- Observation model: $p(r_n|z) = \mathcal{N}(r_n|z; \mathrm{diag}(\sigma^2_{r_n}))$
- Encoder learns: $(r_n; \sigma^2_{r_n}) = \mathrm{enc}(x^c_n; y^c_n)$
- Latent posterior: $p(z|\{r_n\}) = \mathcal{N}(z|\mu_z; \mathrm{diag}(\sigma^2_z))$



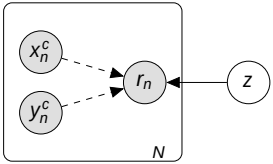
$$\sigma^2_z = \left[\sum_{n=1}^{N} \left(\sigma^2_{r_n}\right)^{-1}\right]^{-1}; \qquad \mu_z = \sum_{n=1}^{N} \frac{\sigma^2_z}{\sigma^2_{r_n}} r_n \qquad \text{(for each latent dim.)}$$

- r_n enters μ_z with learned weight $\sigma^2_z / \sigma^2_{r_n}$

ICLR 2021 *j* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***

Bayesian Context Aggregation

- Context aggregation as Bayesian inference
- Observation model: $p(r_n|z) = \mathcal{N}(r_n|z; \mathrm{diag}(\sigma^2_{r_n}))$
- Encoder learns: $(r_n; \sigma^2_{r_n}) = \mathrm{enc}(x^c_n; y^c_n)$
- Latent posterior: $p(z|\{r_n\}) = \mathcal{N}(z|\mu_z; \mathrm{diag}(\sigma^2_z))$



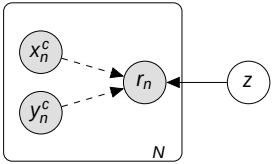
$$\sigma^2_z = \left[\sum_{n=1}^{N} \left(\sigma^2_{r_n}\right)^{-1}\right]^{-1}; \qquad \mu_z = \sum_{n=1}^{N} \frac{\sigma^2_z}{\sigma^2_{r_n}} r_n \qquad \text{(for each latent dim.)}$$

- r_n enters μ_z with learned weight $\sigma^2_z / \sigma^2_{r_n}$
- Principled quantification of task ambiguity

ICLR 2021 *j* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***

Bayesian Context Aggregation

- Context aggregation as Bayesian inference
- Observation model: $p(r_n \mid z) = \mathcal{N}(r_n \mid z; \mathrm{diag}(\sigma^2_{r_n}))$
- Encoder learns: $(r_n; \sigma^2_{r_n}) = \mathrm{enc}(x^c_n; y^c_n)$
- Latent posterior: $p(z \mid \{r_n\}) = \mathcal{N}(z \mid \mu_z; \mathrm{diag}(\sigma^2_z))$



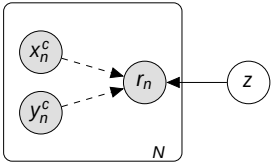
$$\sigma^2_z = \left[\sum_{n=1}^{N} \left(\sigma^2_{r_n}\right)^{-1}\right]^{-1}; \qquad \mu_z = \sum_{n=1}^{N} \frac{\sigma^2_z}{\sigma^2_{r_n}} r_n \qquad \text{(for each latent dim.)}$$

- r_n enters μ_z with learned weight $\sigma^2_z / \sigma^2_{r_n}$
- Principled quantification of task ambiguity
- Only marginal computational overhead

ICLR 2021 *j* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***

Bayesian Context Aggregation

- Context aggregation as Bayesian inference
- Observation model: $p(r_n|z) = \mathcal{N}(r_n|z; \mathrm{diag}(\sigma^2_{r_n}))$
- Encoder learns: $(r_n; \sigma^2_{r_n}) = \mathrm{enc}(x^c_n; y^c_n)$
- Latent posterior: $p(z|\{r_n\}) = \mathcal{N}(z|\mu_z; \mathrm{diag}(\sigma^2_z))$



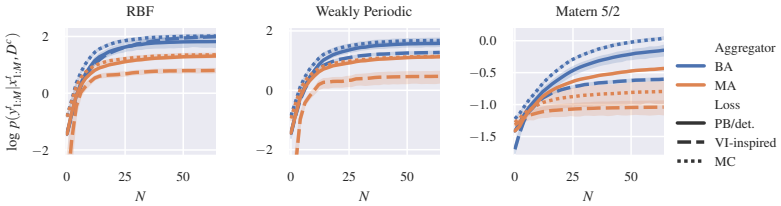
$$\sigma^2_z = \left[\sum_{n=1}^{N} \left(\sigma^2_{r_n}\right)^{-1}\right]^{-1} \quad ; \quad \mu_z = \sum_{n=1}^{N} \frac{\sigma^2_z}{\sigma^2_{r_n}} r_n \qquad \text{(for each latent dim.)}$$

- r_n enters μ_z with learned weight $\sigma^2_z / \sigma^2_{r_n}$
- Principled quantification of task ambiguity
- Only marginal computational overhead
- Compatible with existing NP architectures

ICLR 2021 *j* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***

Experiments

	PB/det.		VI		MC	
	BA	MA (CNP)	BA	MA (LP-NP)	BA	MA
RBF GP	**1:37 ± 0:15**	0:94 ± 0:04	**1:40 ± 0:04**	0:45 ± 0:12	**1:62 ± 0:05**	1:07 ± 0:05
Weakly Periodic GP	**1:13 ± 0:08**	0:76 ± 0:02	**0:89 ± 0:03**	0:07 ± 0:14	**1:30 ± 0:06**	0:85 ± 0:04
Matern-5/2 GP	**−0:50 ± 0:07**	−0:68 ± 0:01	**−0:79 ± 0:01**	−1:09 ± 0:10	**−0:33 ± 0:01**	−0:90 ± 0:15
Furuta Dynamics	**7:50 ± 0:27**	7:06 ± 0:12	**7:32 ± 0:18**	5:57 ± 0:21	**8:25 ± 0:33**	7:55 ± 0:24



ICLR 2021 *j* M. Volpp, F. Flürenbrock, L. Grossberger, C. Daniel, G. Neumann, ***Bayesian Context Aggregation for Neural Processes***