Sharp Calibrated Gaussian Processes

Alexandre Capone[1], Sandra Hirche[1], Geoff Pleiss[2]

1- Technical University of Munich

2- University of British Columbia/Vector Institute

Contact: alexandre.capone@tum.de

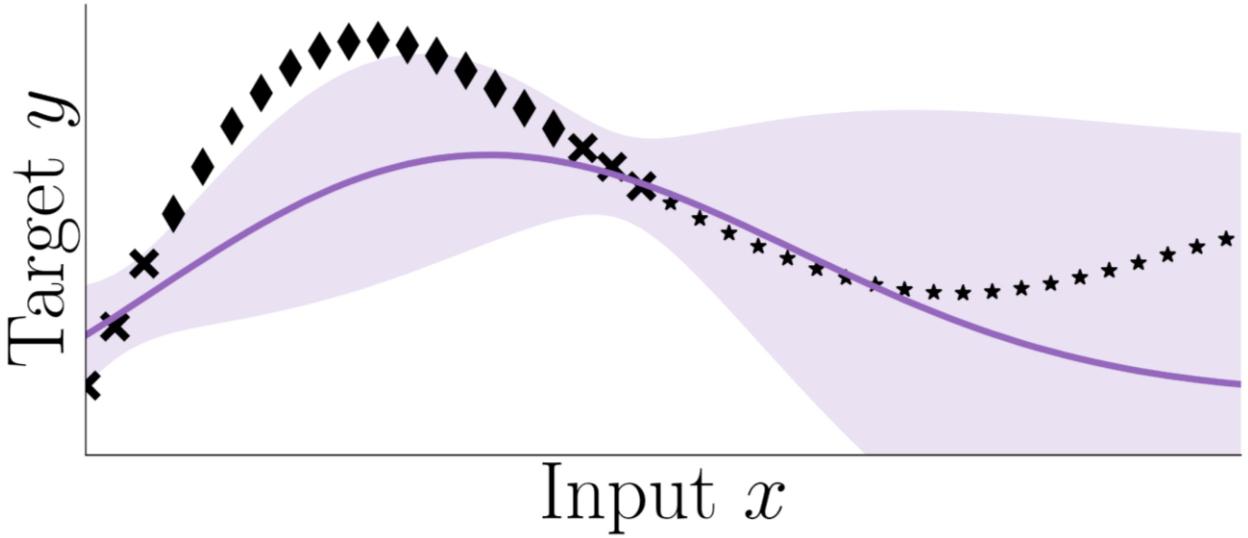
Neurips, 14 December 2021

Gaussian Processes are Restrictive – Even with Good Kernel

Gaussian process distribution is rigid

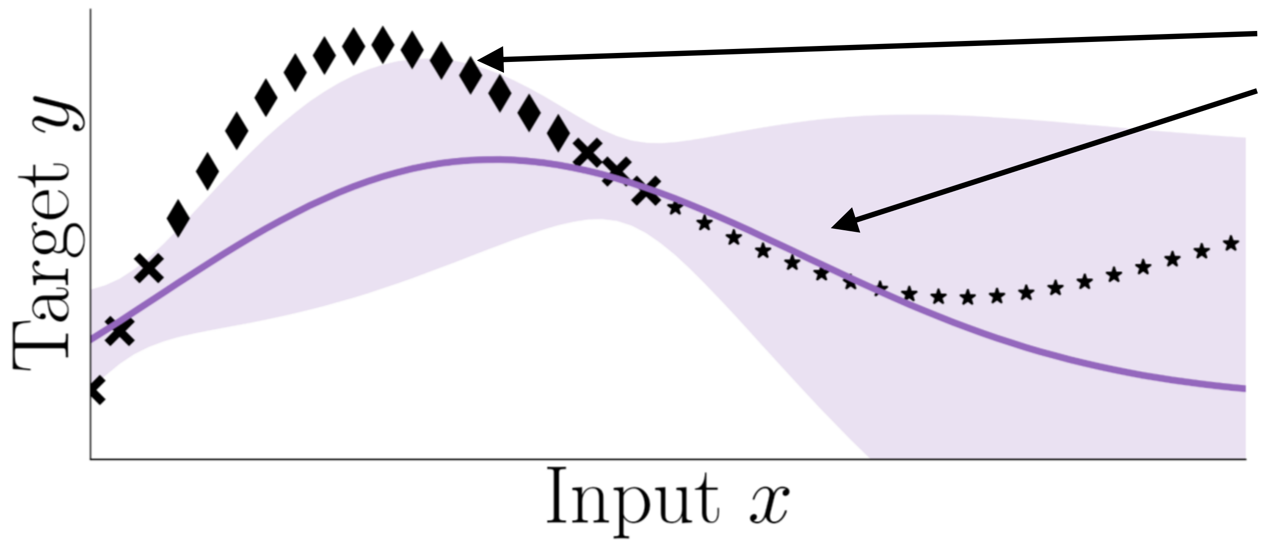
Gaussian Processes are Restrictive – Even with Good Kernel

Gaussian process distribution is rigid



Gaussian Processes are Restrictive – Even with Good Kernel

Gaussian process distribution is rigid

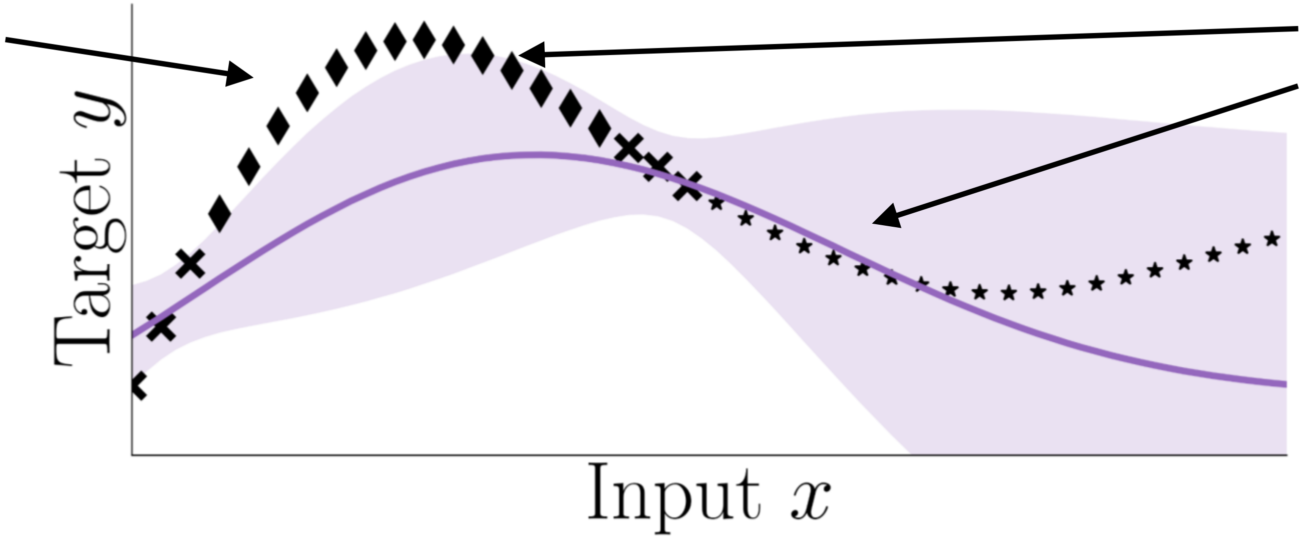


Data mostly above mean – asymmetric!

Gaussian Processes are Restrictive – Even with Good Kernel

Gaussian process distribution is rigid

Misclassified data

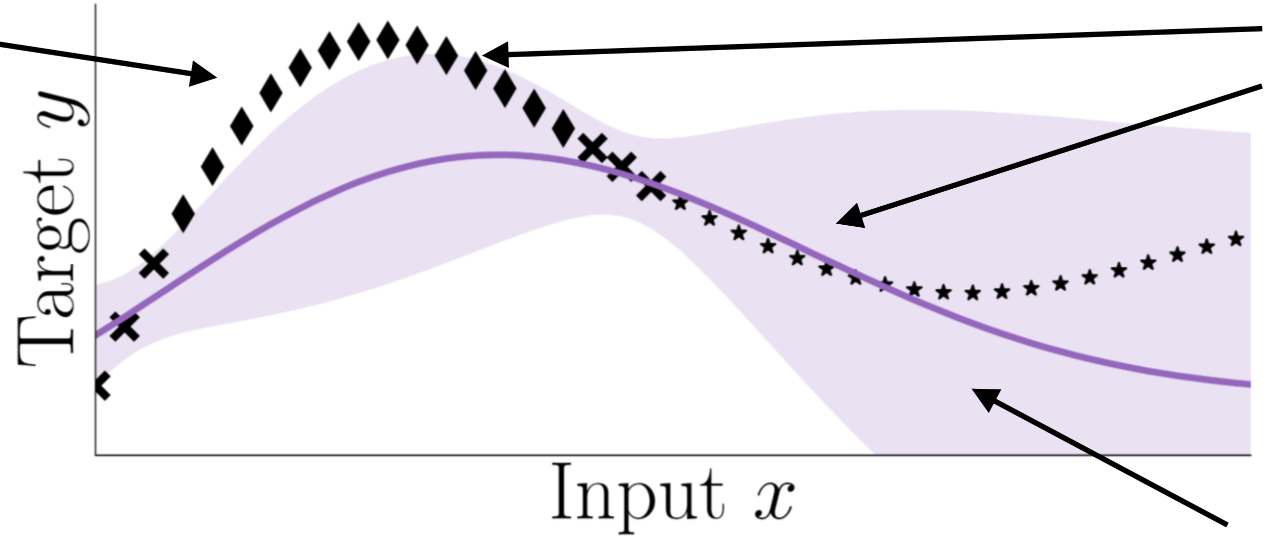


Data mostly above mean – asymmetric!

Gaussian Processes are Restrictive – Even with Good Kernel

Gaussian process distribution is rigid

Misclassified data



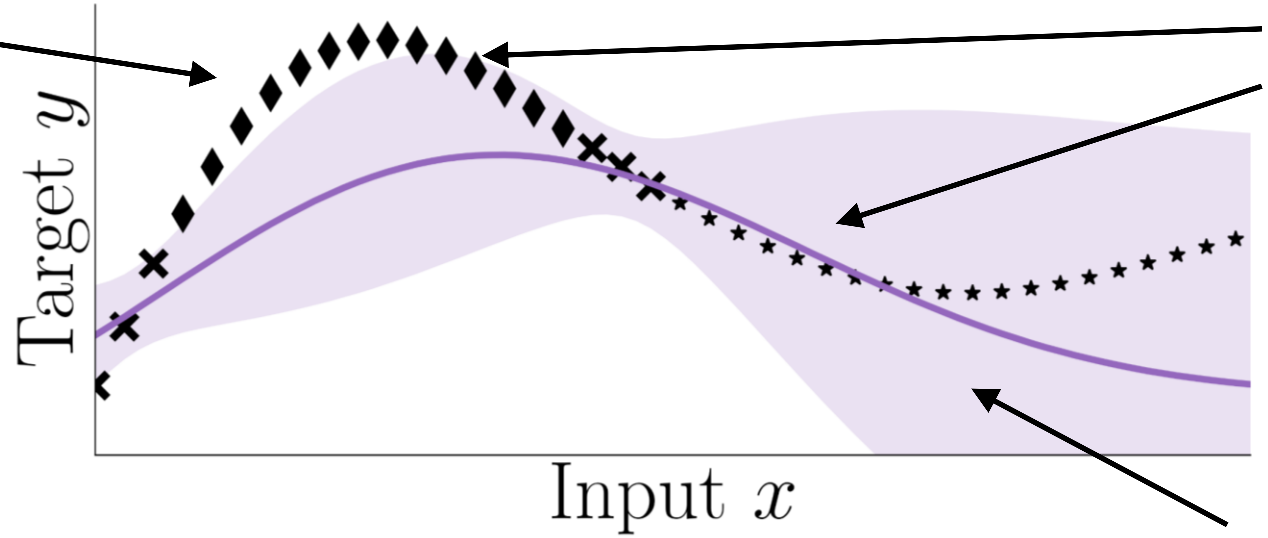
Data mostly above mean – asymmetric!

Bound is not tight

Gaussian Processes are Restrictive – Even with Good Kernel

Gaussian process distribution is rigid

Misclassified data

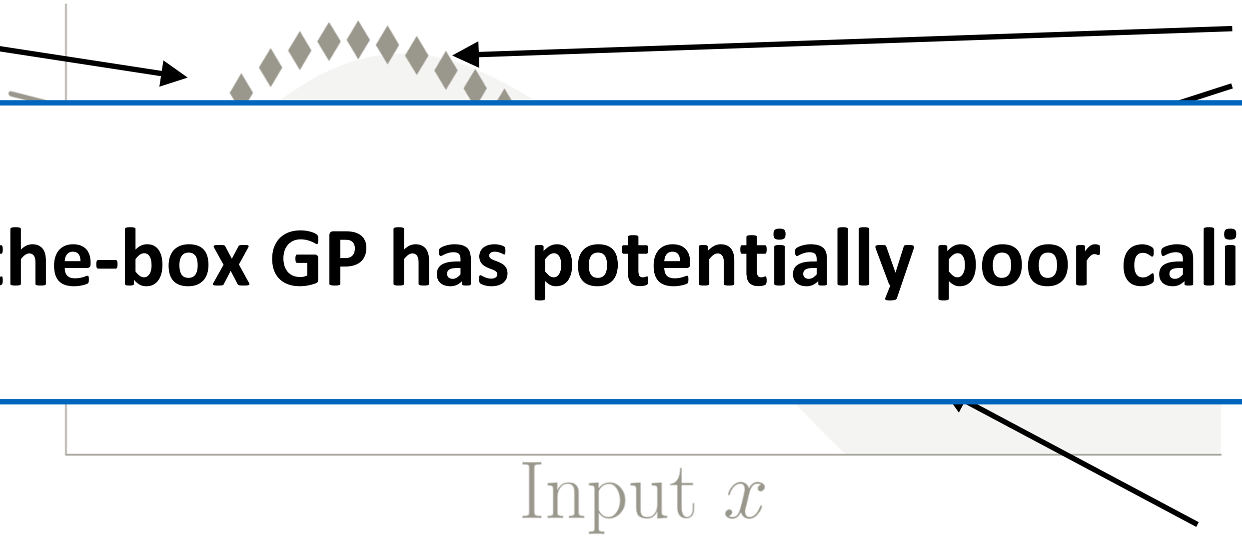


Data mostly above mean – asymmetric!

Bound is not tight

Gaussian Processes are Restrictive – Even with Good Kernel

Gaussian process distribution is rigid



Out-of-the-box GP has potentially poor calibration

Calibrated Gaussian Processes – Related Work

- **Recalibration approaches:** Kuleshov et al. (2018), Vovk et al. (2020), Marx et al. (2022)
  ➡ **Theoretical guarantees**, but **confidence intervals too coarse**

- **Check-score-based:** Song et al. (2019), Kuleshov and Deshpande (2022)

Calibrated Gaussian Processes – Related Work

- **Recalibration approaches:** Kuleshov et al. (2018), Vovk et al. (2020), Marx et al. (2022)
  ➡ **Theoretical guarantees**, but **confidence intervals too coarse**

- **Check-score-based:** Song et al. (2019), Kuleshov and Deshpande (2022)
  ➡ **Tight intervals**, but **less accurate** and **weaker guarantees**

Calibrated Gaussian Processes – Related Work

- **Recalibration approaches:** Kuleshov et al. (2018), Vovk et al. (2020), Marx et al. (2022)
  ➡ **Theoretical guarantees**, but **confidence intervals too coarse**

- **Check-score-based:** Song et al. (2019), Kuleshov and Deshpande (2022)
  ➡ **Tight intervals**, but **less accurate** and **weaker guarantees**

Open Problem: Accurate models with **tight intervals** + **strong guarantees**

Calibrated Gaussian Processes – Our Approach

- **Step 1:** Leverage **flexibility** of separate hyperparameters to compute tight confidence intervals

Calibrated Gaussian Processes – Our Approach

- **Step 1:** Leverage **flexibility** of separate hyperparameters to compute tight confidence intervals

- **Step 2:** Leverage **monotonicity** properties of posterior variance to compute model for **any confidence level**

Calibrated Gaussian Processes – Our Approach

- **Step 1:** Leverage **flexibility** of separate hyperparameters to compute tight confidence intervals

Calibrated Gaussian Processes – Our Approach

- **Step 1:** Leverage **flexibility** of separate hyperparameters to compute tight confidence intervals

$$\min_{\substack{\beta_\delta \in \mathbb{R} \\ \boldsymbol{\theta}_\delta \in \boldsymbol{\Theta}}} \sum_{i=1}^{N_{\text{cal}}} \beta_\delta^2 \sigma_{\mathcal{D}_{\text{tr}}}^2 \left(\boldsymbol{\theta}_\delta, \boldsymbol{x}_{\text{cal}}^i\right)$$

$$\text{s.t.} \sum_{i=1}^{N_{\text{cal}}} \frac{\mathbb{I}_{\geq 0}\left(\Delta y_{\text{cal}}^i - \beta_\delta \sigma_{\mathcal{D}_{\text{tr}}}\left(\boldsymbol{\theta}_\delta, \boldsymbol{x}_{\text{cal}}^i\right)\right)}{N_{\text{cal}} + 1} = \delta$$

Calibrated Gaussian Processes – Our Approach

- **Step 1:** Leverage **flexibility** of separate hyperparameters to compute tight confidence intervals

$$\min_{\substack{\beta_\delta \in \mathbb{R} \\ \boldsymbol{\theta}_\delta \in \boldsymbol{\Theta}}} \sum_{i=1}^{N_{\mathrm{cal}}} \beta_\delta^2 \sigma_{\mathcal{D}_{\mathrm{tr}}}^2 \left(\boldsymbol{\theta}_\delta, \boldsymbol{x}_{\mathrm{cal}}^i\right) \quad \longleftarrow \quad \text{Tightness}$$

$$\text{s.t.} \sum_{i=1}^{N_{\mathrm{cal}}} \frac{\mathbb{I}_{\geq 0}\left(\Delta y_{\mathrm{cal}}^i - \beta_\delta \sigma_{\mathcal{D}_{\mathrm{tr}}}\left(\boldsymbol{\theta}_\delta, \boldsymbol{x}_{\mathrm{cal}}^i\right)\right)}{N_{\mathrm{cal}} + 1} = \delta$$

Calibrated Gaussian Processes – Our Approach

- **Step 1:** Leverage **flexibility** of separate hyperparameters to compute tight confidence intervals

$$\min_{\substack{\beta_\delta \in \mathbb{R} \\ \boldsymbol{\theta}_\delta \in \boldsymbol{\Theta}}} \sum_{i=1}^{N_{\text{cal}}} \beta_\delta^2 \sigma_{\mathcal{D}_{\text{tr}}}^2 \left(\boldsymbol{\theta}_\delta, \boldsymbol{x}_{\text{cal}}^i\right)$$

← Tightness

Guarantees δ of the data is contained →

$$\text{s.t.} \sum_{i=1}^{N_{\text{cal}}} \frac{\mathbb{I}_{\geq 0}\left(\Delta y_{\text{cal}}^i - \beta_\delta \sigma_{\mathcal{D}_{\text{tr}}}\left(\boldsymbol{\theta}_\delta, \boldsymbol{x}_{\text{cal}}^i\right)\right)}{N_{\text{cal}} + 1} = \delta$$

Calibrated Gaussian Processes – Our Approach

- **Step 1:** Leverage **flexibility** of separate hyperparameters to compute tight confidence intervals

$$\min_{\substack{\beta_\delta \in \mathbb{R} \\ \boldsymbol{\theta}_\delta \in \boldsymbol{\Theta}}} \sum_{i=1}^{N_{\text{cal}}} \beta_\delta^2 \sigma_{\mathcal{D}_{\text{tr}}}^2 \left(\boldsymbol{\theta}_\delta, \boldsymbol{x}_{\text{cal}}^i\right)$$ ← Tightness

Guarantees δ of the data is contained →

$$\text{s.t.} \sum_{i=1}^{N_{\text{cal}}} \frac{\mathbb{I}_{\geq 0}\left(\Delta y_{\text{cal}}^i - \beta_\delta \sigma_{\mathcal{D}_{\text{tr}}}\left(\boldsymbol{\theta}_\delta, \boldsymbol{x}_{\text{cal}}^i\right)\right)}{N_{\text{cal}} + 1} = \delta$$

- Unconstrained reformulation:

Calibrated Gaussian Processes – Our Approach

- **Step 1:** Leverage **flexibility** of separate hyperparameters to compute tight confidence intervals

$$\min_{\substack{\beta_\delta \in \mathbb{R} \\ \boldsymbol{\theta}_\delta \in \boldsymbol{\Theta}}} \sum_{i=1}^{N_{\text{cal}}} \beta_\delta^2 \sigma_{\mathcal{D}_{\text{tr}}}^2 \left(\boldsymbol{\theta}_\delta, \boldsymbol{x}_{\text{cal}}^i\right)$$

← Tightness

Guarantees δ of the data is contained →

$$\text{s.t.} \sum_{i=1}^{N_{\text{cal}}} \frac{\mathbb{I}_{\geq 0}\left(\Delta y_{\text{cal}}^i - \beta_\delta \sigma_{\mathcal{D}_{\text{tr}}}\left(\boldsymbol{\theta}_\delta, \boldsymbol{x}_{\text{cal}}^i\right)\right)}{N_{\text{cal}} + 1} = \delta$$

- Unconstrained reformulation:

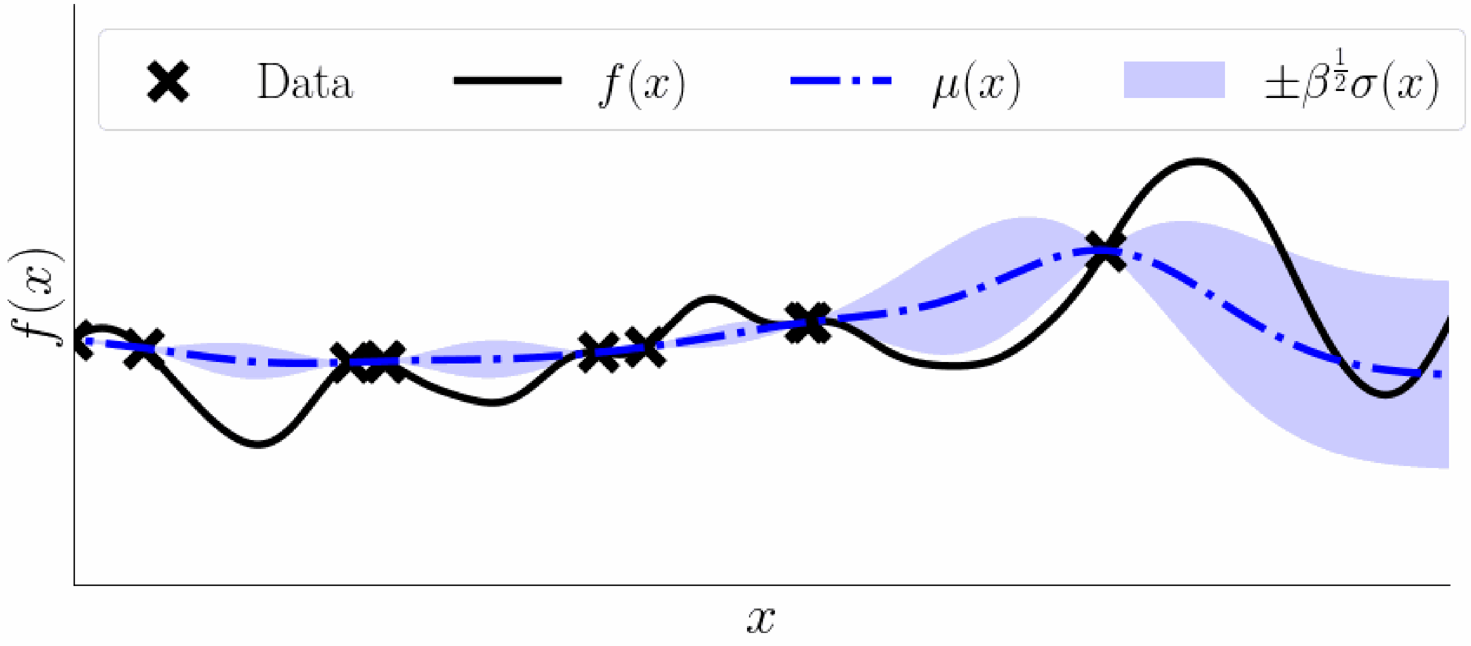
$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N_{\text{cal}}} \left[q_{\text{lin}}(\delta, \boldsymbol{\Sigma}_{\mathcal{D}_{\text{tr}}}^{-1} \boldsymbol{\Delta y}_{\text{cal}}) \sigma_{\mathcal{D}_{\text{tr}}} \left(\boldsymbol{\theta}_\delta, \boldsymbol{x}_{\text{cal}}^i\right)\right]^2$$

Calibrated Gaussian Processes – Our Approach

- **Step 2:** Leverage **monotonicity** properties of posterior variance to compute model for **any confidence level**

Calibrated Gaussian Processes – Our Approach

- **Step 2:** Leverage **monotonicity** properties of posterior variance to compute model for **any confidence level**



Calibrated Gaussian Processes – Our Approach

- **Step 2:** Leverage **monotonicity** properties of posterior variance to compute model for **any confidence level**

for $i = 1$ **to** M **do**

Compute $\beta_{\delta_1}, \boldsymbol{\theta}_{\delta_1}, ..., \beta_{\delta_{N_{\text{cal}}}}, \boldsymbol{\theta}_{\delta_{N_{\text{cal}}}}$ by solving (6) and (7) subject to

$$\boldsymbol{\theta}_{\delta_i} \leq \boldsymbol{\theta}_{\delta_j}, \quad \text{if } \delta_i < \delta_j \text{ and } \beta_{\delta_i} \geq 0,$$
$$\boldsymbol{\theta}_{\delta_i} \geq \boldsymbol{\theta}_{\delta_j}, \quad \text{if } \delta_i < \delta_j \text{ and } \beta_{\delta_j} \leq 0.$$

end for

Calibrated Gaussian Processes – Our Approach

- **Step 2:** Leverage **monotonicity** properties of posterior variance to compute model for **any confidence level**

for $i = 1$ **to** M **do**

Compute $\beta_{\delta_1}, \boldsymbol{\theta}_{\delta_1}, \ldots, \beta_{\delta_{N_{\mathrm{cal}}}}, \boldsymbol{\theta}_{\delta_{N_{\mathrm{cal}}}}$ by solving (6) and (7) subject to

$$\boldsymbol{\theta}_{\delta_i} \leq \boldsymbol{\theta}_{\delta_j}, \quad \text{if } \delta_i < \delta_j \text{ and } \beta_{\delta_i} \geq 0,$$
$$\boldsymbol{\theta}_{\delta_i} \geq \boldsymbol{\theta}_{\delta_j}, \quad \text{if } \delta_i < \delta_j \text{ and } \beta_{\delta_j} \leq 0.$$

end for

Calibrated Gaussian Processes – Our Approach

- **Step 2:** Leverage **monotonicity** properties of posterior variance to compute model for **any confidence level**

for $i = 1$ **to** M **do**
  Compute $\beta_{\delta_1}, \boldsymbol{\theta}_{\delta_1}, ..., \beta_{\delta_{N_{\text{cal}}}}, \boldsymbol{\theta}_{\delta_{N_{\text{cal}}}}$ by solving (6) and (7) subject to

$$\boldsymbol{\theta}_{\delta_i} \leq \boldsymbol{\theta}_{\delta_j}, \quad \text{if } \delta_i < \delta_j \text{ and } \beta_{\delta_i} \geq 0,$$

$$\boldsymbol{\theta}_{\delta_i} \geq \boldsymbol{\theta}_{\delta_j}, \quad \text{if } \delta_i < \delta_j \text{ and } \beta_{\delta_j} \leq 0.$$

end for

- **Enforce monotonicity** in hyperparameters **with increasing confidence** level δ

Calibrated Gaussian Processes – Our Approach

- **Step 2:** Leverage **monotonicity** properties of posterior variance to compute model for **any confidence level**

for $i = 1$ **to** M **do**
Compute $\beta_{\delta_1}, \boldsymbol{\theta}_{\delta_1}, ..., \beta_{\delta_{N_{\mathrm{cal}}}}, \boldsymbol{\theta}_{\delta_{N_{\mathrm{cal}}}}$ by solving (6) and (7) subject to

$$\boldsymbol{\theta}_{\delta_i} \leq \boldsymbol{\theta}_{\delta_j}, \quad \text{if } \delta_i < \delta_j \text{ and } \beta_{\delta_i} \geq 0,$$
$$\boldsymbol{\theta}_{\delta_i} \geq \boldsymbol{\theta}_{\delta_j}, \quad \text{if } \delta_i < \delta_j \text{ and } \beta_{\delta_j} \leq 0.$$

end for

- **Enforce monotonicity** in hyperparameters **with increasing confidence** level δ

  ➡ **Monotonicity in confidence regions**

Calibrated Gaussian Processes – Our Approach

Theorem

If data is iid, then, for all confidence levels : δ

$$\mathbb{P}_{\boldsymbol{x},y\sim\Pi}\left(y-\mu_{\mathcal{D}_{tr}}(\hat{\boldsymbol{\theta}}(\delta),\boldsymbol{x})\leq\hat{\beta}(\delta)\sigma_{\mathcal{D}_{tr}}(\hat{\boldsymbol{\theta}}(\delta),\boldsymbol{x})\right)\in\left[\delta-\frac{1}{N_{cal}+1},\delta+\frac{1}{N_{cal}+1}\right].$$

Number of calibration data

Calibrated Gaussian Processes – Our Approach

Theorem

If data is iid, then, for all confidence levels : δ

$$\mathbb{P}_{\boldsymbol{x},y\sim\Pi}\left(y-\mu_{\mathcal{D}_{tr}}(\hat{\boldsymbol{\theta}}(\delta),\boldsymbol{x})\leq\hat{\beta}(\delta)\sigma_{\mathcal{D}_{tr}}(\hat{\boldsymbol{\theta}}(\delta),\boldsymbol{x})\right)\in\left[\delta-\frac{1}{N_{cal}+1},\delta+\frac{1}{N_{cal}+1}\right].$$

Number of
calibration data

Calibrated Gaussian Processes – Our Approach

Theorem

If data is iid, then, for all confidence levels : δ

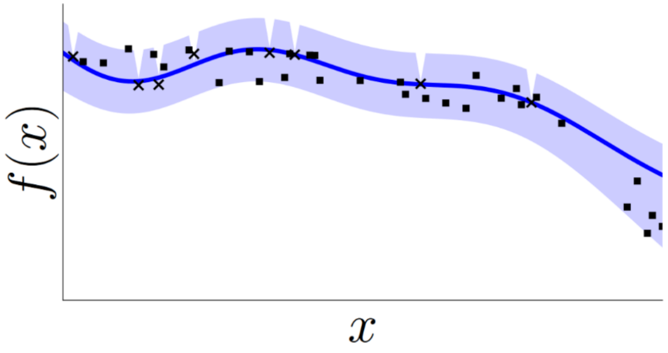
$$\mathbb{P}_{\boldsymbol{x},y\sim\Pi}\left(y - \mu_{\mathcal{D}_{tr}}(\hat{\boldsymbol{\theta}}(\delta), \boldsymbol{x}) \leq \hat{\beta}(\delta)\sigma_{\mathcal{D}_{tr}}(\hat{\boldsymbol{\theta}}(\delta), \boldsymbol{x})\right) \in \left[\delta - \frac{1}{N_{cal}+1}, \delta + \frac{1}{N_{cal}+1}\right].$$

Number of calibration data

Enforces calibration AND optimizes sharpness

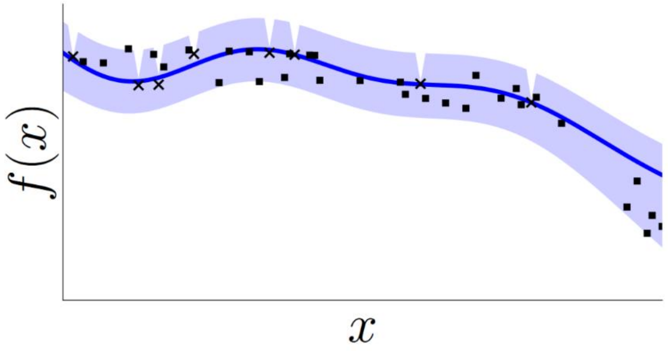
Calibrated Gaussian Processes – Results Toy Problem

Calibrated Gaussian Processes – Results Toy Problem

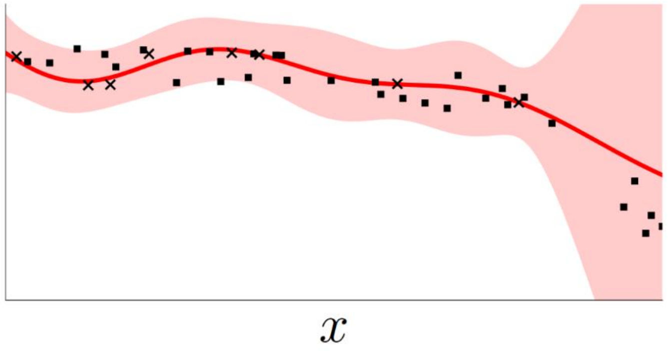


(a) Our approach.

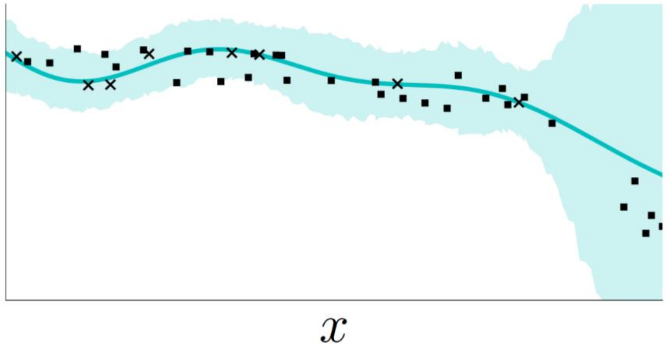
Calibrated Gaussian Processes – Results Toy Problem



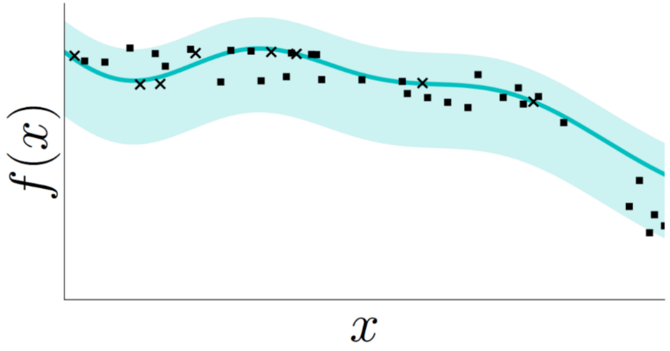
(a) Our approach.



(b) Kuleshov et al. (2018).

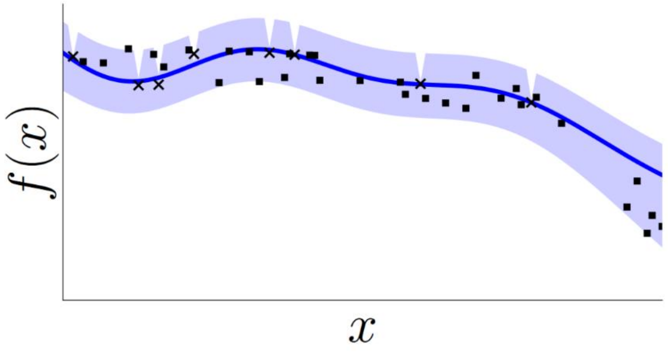


(c) Vovk et al. (2020).

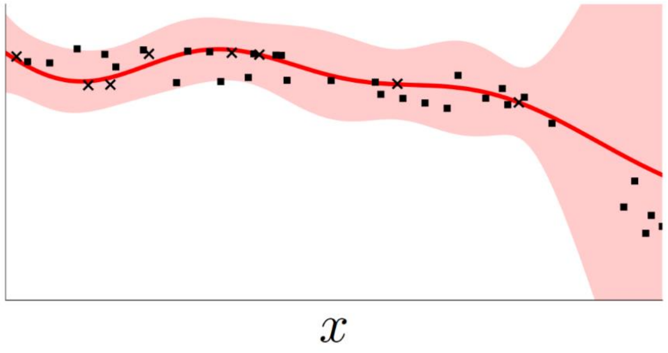


(d) Mean ± constant (Marx et al., 2022).

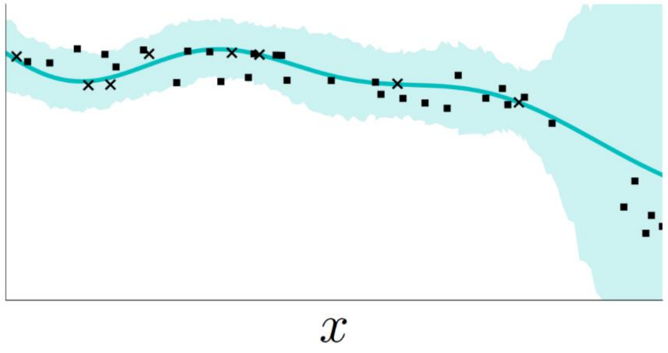
Calibrated Gaussian Processes – Results Toy Problem



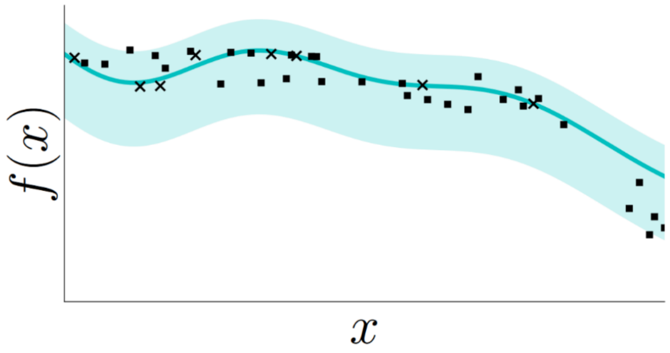
(a) Our approach.



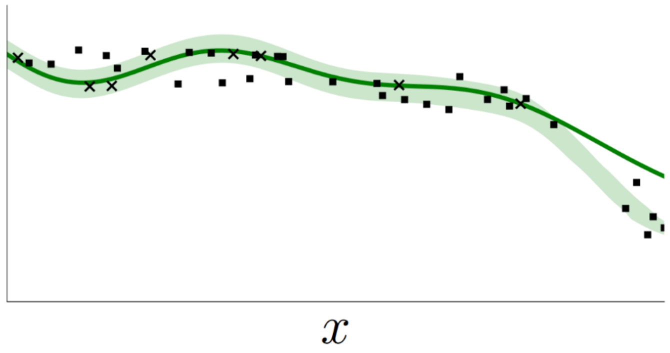
(b) Kuleshov et al. (2018).



(c) Vovk et al. (2020).

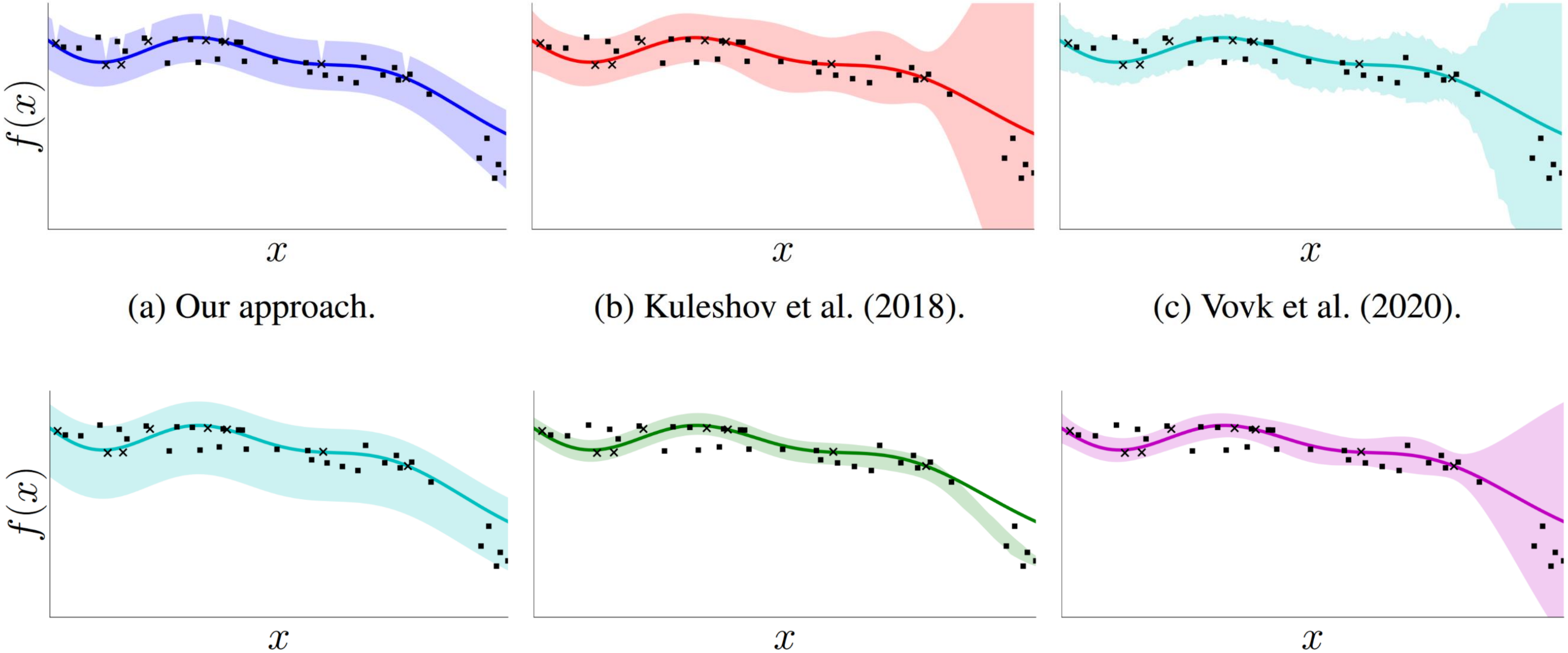


(d) Mean $\pm$ constant (Marx et al., 2022).



(e) Neural network-based recalibrator (Kuleshov & Deshpande, 2022).

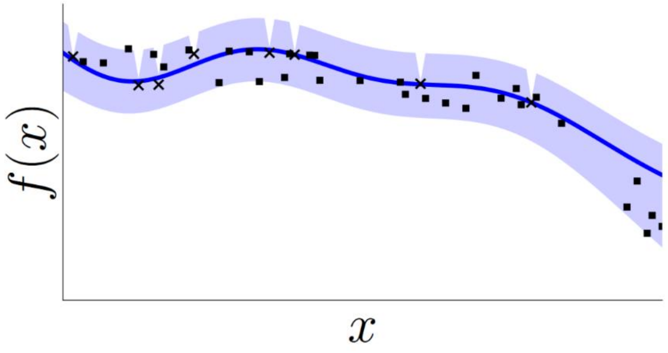
Calibrated Gaussian Processes – Results Toy Problem



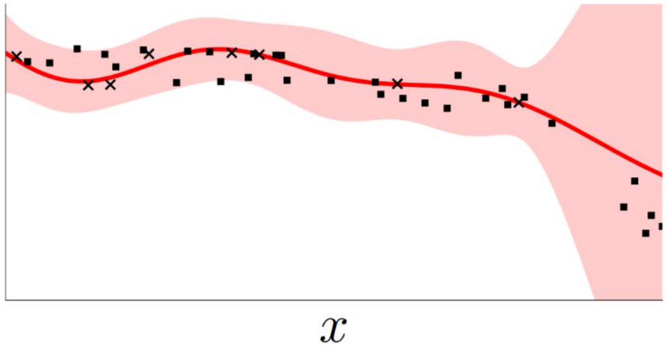
(a) Our approach. (b) Kuleshov et al. (2018). (c) Vovk et al. (2020).

(d) Mean ± constant (Marx et al., 2022). (e) Neural network-based recalibrator (Kuleshov & Deshpande, 2022). (f) Base model (vanilla GP).

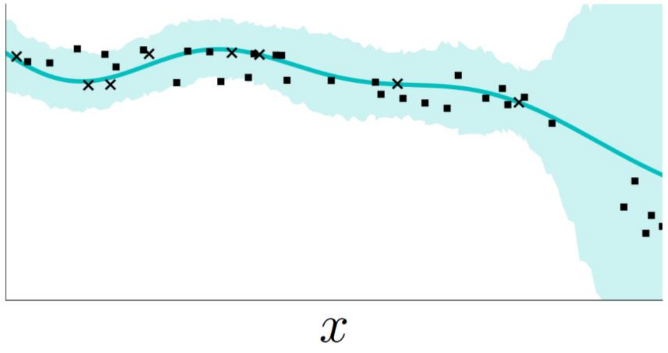
Calibrated Gaussian Processes – Results Toy Problem



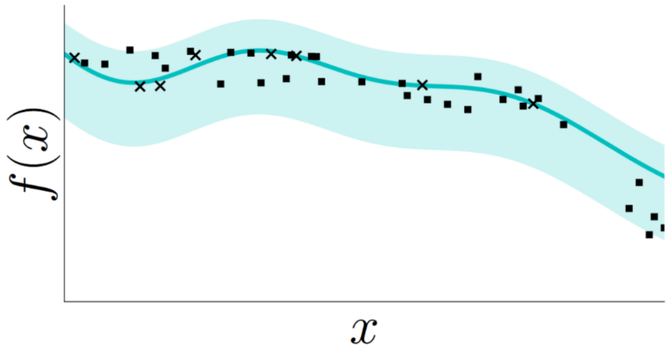
(a) Our approach.



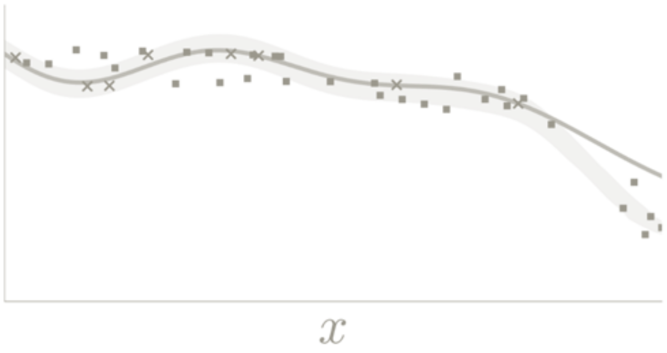
(b) Kuleshov et al. (2018).



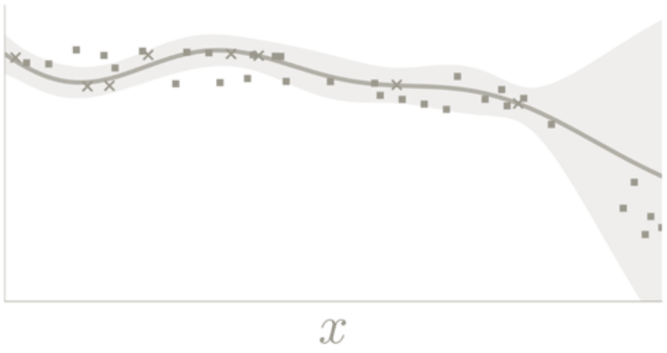
(c) Vovk et al. (2020).



(d) Mean $\pm$ constant (Marx et al., 2022).

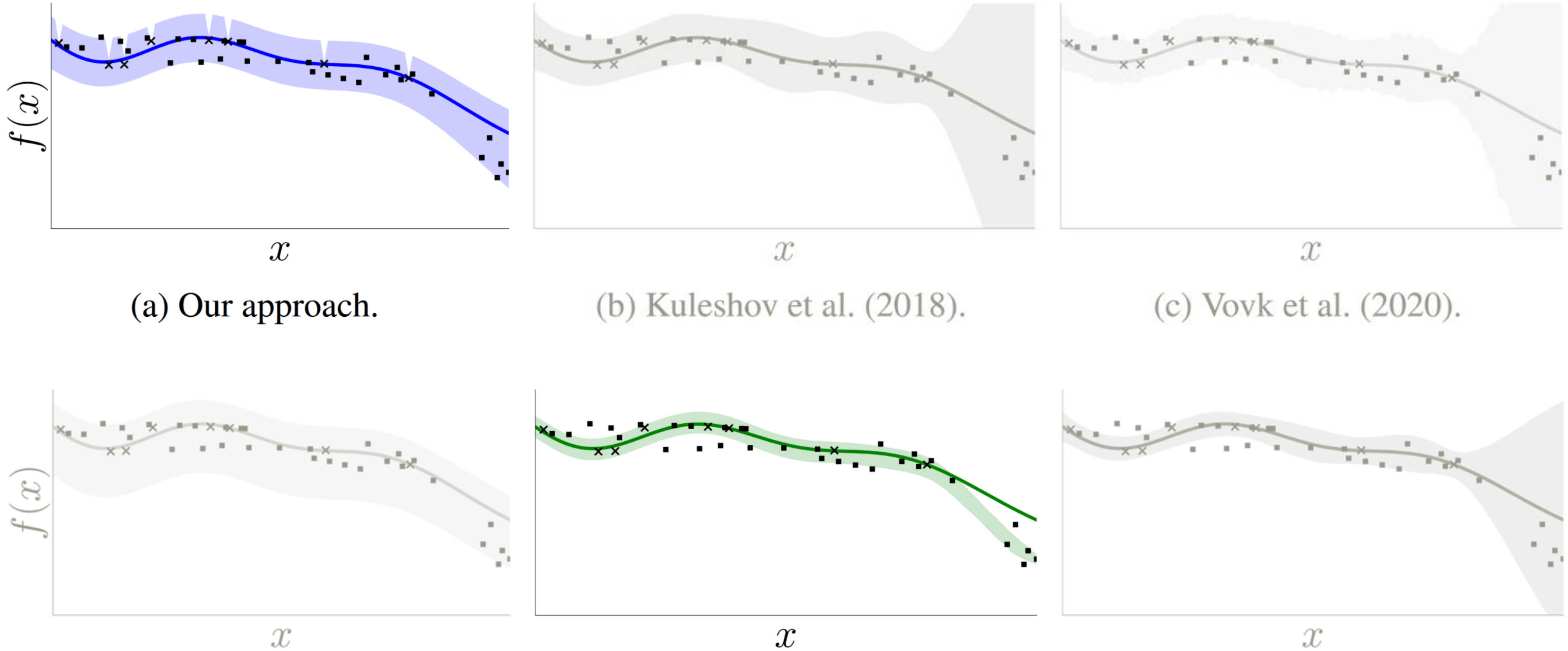


(e) Neural network-based recalibrator (Kuleshov & Deshpande, 2022).



(f) Base model (vanilla GP).

Calibrated Gaussian Processes – Results Toy Problem



(a) Our approach.

(b) Kuleshov et al. (2018).

(c) Vovk et al. (2020).

(d) Mean ± constant (Marx et al., 2022).

(e) Neural network-based recalibrator (Kuleshov & Deshpande, 2022).

(f) Base model (vanilla GP).

Calibrated Gaussian Processes – Results Benchmark Datasets

Data set	Metric	Ours	RK	RV	RM	NN	B
Boston	ECE	0.003	**0.0029**	**0.0029**	**0.0029**	0.0056	0.041
	STD	**0.16**	0.31	0.3	0.33	0.22	1.9
	NLL	0.21	0.39	0.4	0.42	**-0.24**	1.6
	95% CI	0.76	1.4	1.4	1.4	**0.73**	7.4
yacht	ECE	0.0044	**0.0043**	0.0044	**0.0043**	0.0081	0.039
	STD	0.16	0.5	0.47	0.5	**0.14**	2.8
	NLL	0.26	0.68	0.69	0.68	**-2**	2
	95% CI	0.76	2.3	2.3	2.3	**0.3**	11
mpg	ECE	0.0036	**0.0035**	**0.0035**	**0.0035**	0.0053	0.044
	STD	**0.13**	0.38	0.38	0.38	0.29	2.8
	NLL	0.032	0.63	0.64	0.63	**0.02**	2
	95% CI	**0.6**	1.7	1.8	1.7	0.96	11

Data set	Metric	Ours	RK	RV	RM	NN	B
wine	ECE	**0.00047**	**0.00047**	**0.00047**	**0.00047**	0.0067	0.0058
	STD	**0.54**	1	1	0.88	0.72	1.4
	NLL	1.2	1.3	1.3	1.3	**-0.36**	1.4
	95% CI	**2.1**	3.8	3.8	3.9	2.8	5.4
concrete	ECE	0.00071	**0.00064**	**0.00064**	**0.00064**	0.00076	0.032
	STD	**0.25**	0.64	0.64	0.64	0.57	2.8
	NLL	**0.72**	1.1	1.1	1.1	0.85	2
	95% CI	**0.93**	2.5	2.5	2.5	2.1	11
kin8nm	ECE	**0.00016**	**0.00016**	**0.00016**	**0.00016**	0.00053	0.028
	STD	**0.074**	0.12	0.12	0.12	0.098	0.4
	NLL	-0.54	-0.65	-0.65	-0.63	**-0.76**	0.1
	95% CI	**0.26**	0.47	0.47	0.48	0.44	1.6
Facebook2	ECE	0.00044	**0.00043**	**0.00043**	0.00045	0.0089	0.044
	STD	**0.068**	0.18	0.18	0.18	0.18	1.2
	NLL	3.6	-1.3	-1.3	-1.2	**-2.3**	1.2
	95% CI	**0.6**	1.7	1.7	1.7	3.4	4.6

Calibrated Gaussian Processes – Results Benchmark Datasets

Data set	Metric	Ours	RK	RV	RM	NN	B
Boston	ECE	0.003	**0.0029**	**0.0029**	**0.0029**	0.0056	0.041
	STD	**0.16**	0.31	0.3	0.33	0.22	1.9
	NLL	0.21	0.39	0.4	0.42	**-0.24**	1.6
	95% CI	0.76	1.4	1.4	1.4	**0.73**	7.4
yacht	ECE	0.0044	**0.0043**	0.0044	**0.0043**	0.0081	0.039
	STD	0.16	0.5	0.47	0.5	**0.14**	2.8
	NLL	0.26	0.68	0.69	0.68	**-2**	2
	95% CI	0.76	2.3	2.3	2.3	**0.3**	11
mpg	ECE	0.0036	**0.0035**	**0.0035**	**0.0035**	0.0053	0.044
	STD	**0.13**	0.38	0.38	0.38	0.29	2.8
	NLL	0.032	0.63	0.64	0.63	**0.02**	2
	95% CI	**0.6**	1.7	1.8	1.7	0.96	11

Data set	Metric	Ours	RK	RV	RM	NN	B
wine	ECE	**0.00047**	**0.00047**	**0.00047**	**0.00047**	0.0067	0.0058
	STD	**0.54**	1	1	0.88	0.72	1.4
	NLL	1.2	1.3	1.3	1.3	**-0.36**	1.4
	95% CI	**2.1**	3.8	3.8	3.9	2.8	5.4
concrete	ECE	0.00071	**0.00064**	**0.00064**	**0.00064**	0.00076	0.032
	STD	**0.25**	0.64	0.64	0.64	0.57	2.8
	NLL	**0.72**	1.1	1.1	1.1	0.85	2
	95% CI	**0.93**	2.5	2.5	2.5	2.1	11
kin8nm	ECE	**0.00016**	**0.00016**	**0.00016**	**0.00016**	0.00053	0.028
	STD	**0.074**	0.12	0.12	0.12	0.098	0.4
	NLL	-0.54	-0.65	-0.65	-0.63	**-0.76**	0.1
	95% CI	**0.26**	0.47	0.47	0.48	0.44	1.6
Facebook2	ECE	0.00044	**0.00043**	**0.00043**	0.00045	0.0089	0.044
	STD	**0.068**	0.18	0.18	0.18	0.18	1.2
	NLL	3.6	-1.3	-1.3	-1.2	**-2.3**	1.2
	95% CI	**0.6**	1.7	1.7	1.7	3.4	4.6

Calibrated Gaussian Processes – Results Benchmark Datasets

	OURS	RK	RV	RM	NN	B
ECE	0,00071	0,00064	0,00064	0,00064	0,00560	0,03900
STD	0,16	0,38	0,38	0,38	0,22	1,90
NLL	0,26	0,63	0,64	0,63	-0,36	1,60
95% CI	0,76	1,70	1,80	1,70	0,96	7,40

Calibrated Gaussian Processes – Results Benchmark Datasets

	OURS	RK	RV	RM	NN	B
ECE	0,00071	0,00064	0,00064	0,00064	0,00560	0,03900
STD	0,16	0,38	0,38	0,38	0,22	1,90
NLL	0,26	0,63	0,64	0,63	-0,36	1,60
95% CI	0,76	1,70	1,80	1,70	0,96	7,40

DATA SET
BOSTON
YACHT
MPG

95% CI 0.6 1.7 1.7 1.7 3.4 4.6

- **Best calibration** or **marginally worse**

Calibrated Gaussian Processes – Results Benchmark Datasets

	OURS	RK	RV	RM	NN	B
ECE	0,00071	0,00064	0,00064	0,00064	0,00560	0,03900
STD	0,16	0,38	0,38	0,38	0,22	1,90
NLL	0,26	0,63	0,64	0,63	-0,36	1,60
95% CI	0,76	1,70	1,80	1,70	0,96	7,40

- **Best calibration** or **marginally worse**
- **Best sharpness** of those that **enforce calibration**

Thank you!