#527

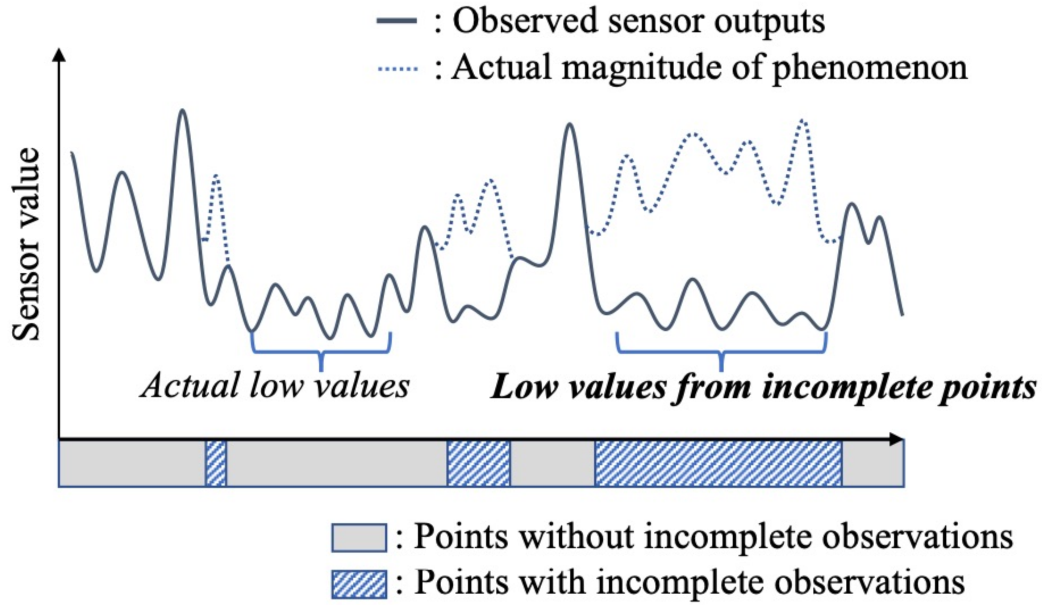
Regression with Sensor Data Containing Incomplete Observations

Takayuki Katsuki, Takayuki Osogami (IBM Research – Tokyo),
kats@jp.ibm.com, osogami@jp.ibm.com

Paper

Learn a regression function to predict the value representing the magnitude of a phenomenon, where the value is observed with a sensor

- Example: Pressure, vibration, and temperature. In healthcare, pulsation, respiration, or body movements.
  - Use case: We can replace intrusive sensors with non-intrusive ones, which will reduce the burden on patients.
- The observed value is not necessarily in agreement with the actual magnitude.
- Especially, ***low values can mean either actual low magnitude or incomplete observation***.
  - This leads to a ***bias toward lower values*** in labels and the resultant learning



Application in healthcare

Explicitly model incomplete observations by asymmetric noise

- Regression from uncorrupted data

$$y = f^*(\boldsymbol{x}) + \epsilon_{\mathrm{s}}$$

Oracle e.g., AWGN

$$\hat{f} \equiv \operatorname*{argmin}_{f \in \mathcal{F}} \mathcal{L}(f), \qquad (2)$$

$$\mathcal{L}(f) \equiv \mathbb{E}_{p(\boldsymbol{x},y)}[L(f(\boldsymbol{x}), y)]$$

Loss function, e.g., squared loss

$\boldsymbol{x} \in \mathbb{R}^D (D \in \mathbb{N})$: Explanatory variable

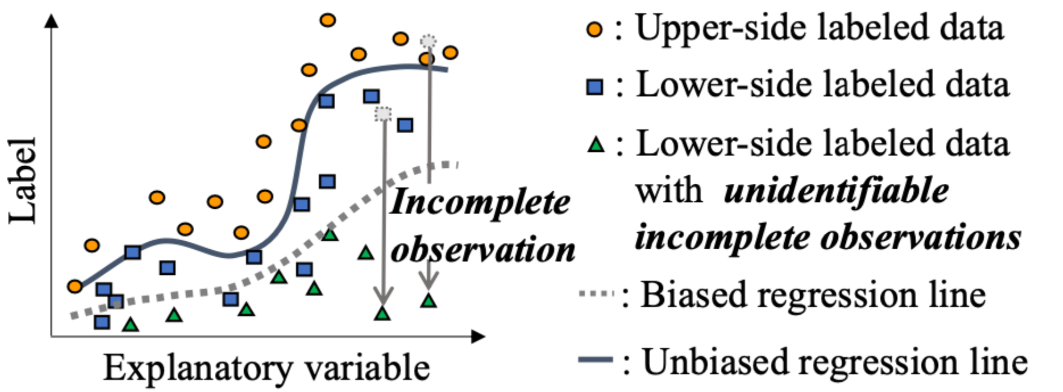
$y \in \mathbb{R}$: Real-valued label

- ***Regression from asymmetrically corrupted data (Our problem)***

$$y' = y + \epsilon_{\mathrm{a}}$$

Asymmetric negative-valued noise

Asymmetrically corrupted data.



Asymmetric noise makes lower-side labeled data particularly unreliable while keeping upper-side labeled data reliable

Assumption 2.1. Assume $\epsilon_s \perp\!\!\!\perp f^*(\boldsymbol{x})$, $\mathbb{E}_{p(\epsilon_s)}[\epsilon_s] = 0$; $\epsilon_a \perp\!\!\!\perp f^*(\boldsymbol{x})$, $\epsilon_a \leq 0$ almost surely (a.s.); $2|\epsilon_s| < |\epsilon_a|$ a.s. when $\epsilon_a < 0$; and $\{(\boldsymbol{x}_n, y_n, y'_n)\}_{n=1}^{N}$ are i.i.d. observations in accordance with Eqs. (1) and (4).

Lemma 2.2. *Let* $\mathcal{F}' \equiv \{f \in \mathcal{F} : |f(\boldsymbol{x}) - f^*(\boldsymbol{x})| \leq |\epsilon_s|$ *a.s.*$\}$. *When* $f \in \mathcal{F}'$, *the following holds under Assumption 2.1:*

$$\mathbb{E}_{p(\boldsymbol{x},y' \mid f(\boldsymbol{x}) \leq y')}[G(\boldsymbol{x}, y')] = \mathbb{E}_{p(\boldsymbol{x},y \mid f(\boldsymbol{x}) \leq y)}[G(\boldsymbol{x}, y)] \quad (5)$$

for any function $G : \mathbb{R}^D \times \mathbb{R} \to \mathbb{R}$ *as long as the expectations exist.*

ϵ_a does not change the expectation for our upper-side labeled data.

Thus, our upper-side labeled data (f (x) ≤ y') is still reliable for regression.

Rewrite the gradient into one that only requires upper-side labeled and unlabeled data in our corrupted data

$\pi_{\mathrm{up}} \equiv p(f_t(\boldsymbol{x}) \leq y)$
$\pi_{\mathrm{lo}} \equiv p(y < f_t(\boldsymbol{x}))$

Treated as hyperparameter

- Gradient for regression from uncorrupted data

$$\hat{f} \equiv \underset{f \in \mathcal{F}}{\operatorname{argmin}} \, \mathcal{L}(f), \tag{2}$$

$$\mathcal{L}(f) \equiv \mathbb{E}_{p(\boldsymbol{x},y)}[L(f(\boldsymbol{x}), y)]$$

$$\nabla \mathcal{L}(f_t) \equiv \mathbb{E}_{p(\boldsymbol{x},y)}[\nabla L(f_t(\boldsymbol{x}), y)], \tag{6}$$

$$\text{where} \quad \nabla L(f_t(\boldsymbol{x}), y) \equiv \left.\frac{\partial L(f(\boldsymbol{x}), y)}{\partial \boldsymbol{\theta}}\right|_{f=f_t}.$$

$$\nabla \mathcal{L}(f_t) = \pi_{\mathrm{up}} \mathbb{E}_{p(\boldsymbol{x},y \mid f_t(\boldsymbol{x}) \leq y)}[\nabla L(f_t(\boldsymbol{x}), y)] + \pi_{\mathrm{lo}} \mathbb{E}_{p(\boldsymbol{x},y \mid y < f_t(\boldsymbol{x}))}[\nabla L(f_t(\boldsymbol{x}), y)]. \tag{7}$$

Unbiased and consistent

- ***Proposed gradient for regression from asymmetrically corrupted data***

Condition 3.1. For $y < f(\boldsymbol{x})$, let $\boldsymbol{g}(f(\boldsymbol{x}))$ be defined as $\nabla L(f(\boldsymbol{x}), y)$. Then $\boldsymbol{g}(f(\boldsymbol{x}))$ depends only on $f(\boldsymbol{x})$ and is conditionally independent of y given $f(\boldsymbol{x})$.

$$\frac{\partial |f(\boldsymbol{x}) - y|}{\partial \boldsymbol{\theta}} = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{\theta}} \qquad \text{when } y < f(\boldsymbol{x}), \tag{8}$$

$$\nabla \tilde{\mathcal{L}}(f_t) \equiv \pi_{\mathrm{up}} \mathbb{E}_{p(\boldsymbol{x},y' \mid f_t(\boldsymbol{x}) \leq y')}\left[\nabla L(f_t(\boldsymbol{x}), y')\right] + \mathbb{E}_{p(\boldsymbol{x})}\left[\boldsymbol{g}(f_t(\boldsymbol{x}))\right] - \pi_{\mathrm{up}} \mathbb{E}_{p(\boldsymbol{x} \mid f_t(\boldsymbol{x}) \leq y')}\left[\boldsymbol{g}(f_t(\boldsymbol{x}))\right]. \tag{9}$$

Upper-side labeled data

Unlabeled data

→ Upper and Unlabeled Regression (***U2 Regression***)

Our algorithm is unbiased as if it were learned from uncorrupted data that does not involve incomplete observations.

Unbiasedness and Consistency of Gradient

Proposition 3.2. *Suppose that Assumption 2.1 holds and the loss function $L(f(\boldsymbol{x}), y)$ satisfies Condition 3.1. Then, the gradient $\nabla\tilde{\mathcal{L}}(f_t)$ in Eq. (9) and its empirical approximation $\nabla\hat{\mathcal{L}}(f_t)$ in Eq. (10) are unbiased and consistent with the gradient $\nabla\mathcal{L}(f_t)$ in Eq. (6) a.s.*

Assumption 3.3. Assume $\epsilon_{\mathrm{a}} \perp\!\!\!\perp \boldsymbol{x}$.

Lemma 3.4. *Let $\nabla\check{\mathcal{L}}(f_t)$ be a variant of the gradient in Eq. (7) replacing $p(\boldsymbol{x}, y)$ with $p(\boldsymbol{x}, y')$, δ be the difference between the expectations of the gradients in the upper side and the lower side $\delta \equiv \left|\mathbb{E}_{p(\boldsymbol{x},y|f(\boldsymbol{x})\leq y)}[\nabla L(f(\boldsymbol{x}), y)] - \mathbb{E}_{p(\boldsymbol{x},y|y<f(\boldsymbol{x}))}[\nabla L(f(\boldsymbol{x}), y)]\right|$, $\eta \in [0,1]$ be the probability of being $0 \leq \epsilon_{\mathrm{s}}$, and $\xi \in [0,1]$ be the probability of $\epsilon_{\mathrm{a}} = 0$. Then, $\nabla\check{\mathcal{L}}(f_t)$ is not consistent with the gradient $\nabla\mathcal{L}(f_t)$ in Eq. (6) a.s., and the difference (bias) between them at step $t+1$ in the gradient descent is*

$$\frac{\eta(1-\eta)(1-\xi)}{1-\eta\xi}\delta \leq |\nabla\check{\mathcal{L}}(f_t) - \nabla\mathcal{L}(f_t)|. \tag{11}$$

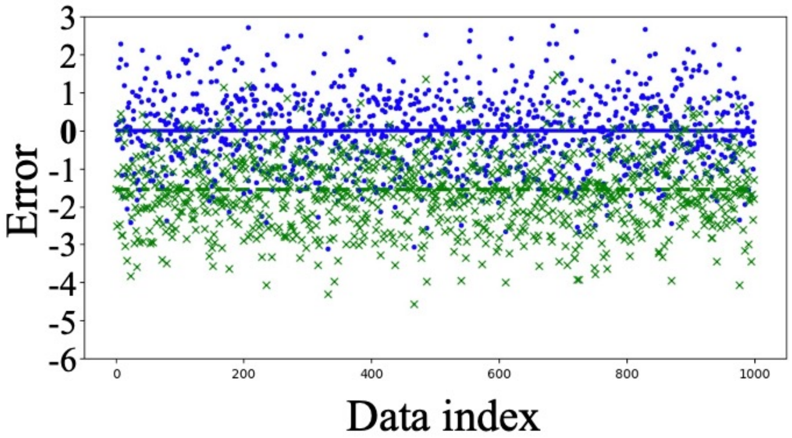
$$\nabla\tilde{\mathcal{L}}(f_t) \equiv \pi_{\mathrm{up}}\mathbb{E}_{p(\boldsymbol{x},y'|f_t(\boldsymbol{x})\leq y')}\left[\nabla L(f_t(\boldsymbol{x}), y')\right] + \mathbb{E}_{p(\boldsymbol{x})}\left[\boldsymbol{g}(f_t(\boldsymbol{x}))\right] - \pi_{\mathrm{up}}\mathbb{E}_{p(\boldsymbol{x}|f_t(\boldsymbol{x})\leq y')}\left[\boldsymbol{g}(f_t(\boldsymbol{x}))\right]. \tag{9}$$

Empirical approximation

$$\nabla\hat{\mathcal{L}}(f_t) = \frac{\pi_{\mathrm{up}}}{n_{\mathrm{up}}}\left[\sum_{(\boldsymbol{x},y')\in\{\boldsymbol{X}_{\mathrm{up}},\boldsymbol{y}'_{\mathrm{up}}\}} \nabla L(f_t(\boldsymbol{x}), y')\right] + \frac{1}{N}\left[\sum_{\boldsymbol{x}\in\boldsymbol{X}_{\mathrm{un}}} \boldsymbol{g}(f_t(\boldsymbol{x}))\right] - \frac{\pi_{\mathrm{up}}}{n_{\mathrm{up}}}\left[\sum_{\boldsymbol{x}\in\boldsymbol{X}_{\mathrm{up}}} \boldsymbol{g}(f_t(\boldsymbol{x}))\right], \tag{10}$$

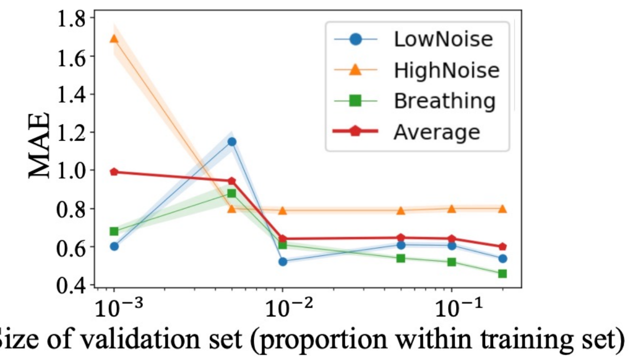
Experiments

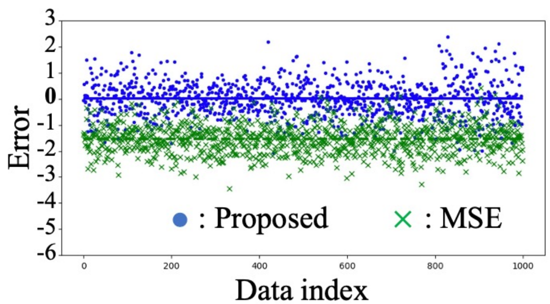
- Demonstration of unbiased learning
  - Errors (predicted value minus true value) by proposed method (blue) and by MSE (green)



Proposed method achieved approximately unbiased learning

- Performance (MAE, lower is better) of proposed method over different sizes of validation set.





Our validation set-based approach for estimating the hyperparameters is robust

Experiments

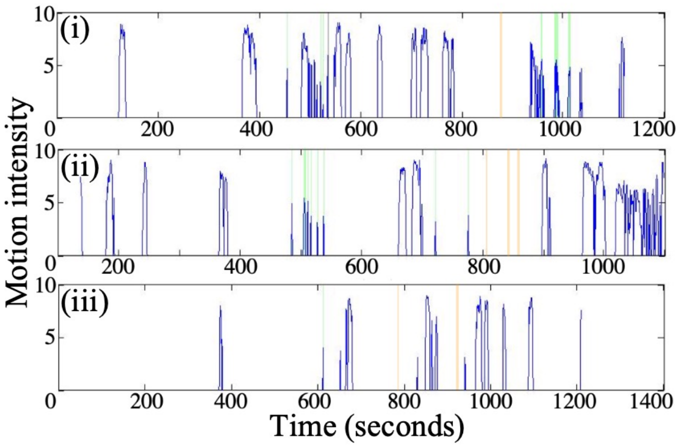
- Comparison between proposed method and baselines in terms of MAE (smaller is better).
  - Best methods are in bold

	Specification	Throwing A	Lifting	Lowering	Throwing B	Avg.
MSE	2.38 ± 0.03	1.54 ± 0.01	1.42 ± 0.01	1.37 ± 0.01	1.21 ± 0.01	1.58
MAE	2.14 ± 0.02	1.46 ± 0.01	1.44 ± 0.01	1.33 ± 0.01	1.31 ± 0.01	1.54
Huber	2.04 ± 0.02	1.66 ± 0.01	1.45 ± 0.01	1.50 ± 0.01	1.32 ± 0.01	1.59
Proposed-1	1.55 ± 0.02	1.18 ± 0.01	1.11 ± 0.01	1.14 ± 0.01	1.03 ± 0.01	1.20
Proposed-2	$\mathbf{1.32 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.94 \pm 0.01}$	$\mathbf{0.86 \pm 0.01}$	$\mathbf{0.97 \pm 0.01}$	**1.02**

Proposed methods significantly outperformed the baselines.

- Real use case for healthcare
  - We estimate intrusive arm sensor output from outputs of non-intrusive bed sensors
  - We use evaluation metrics designed for sleep-wake discrimination

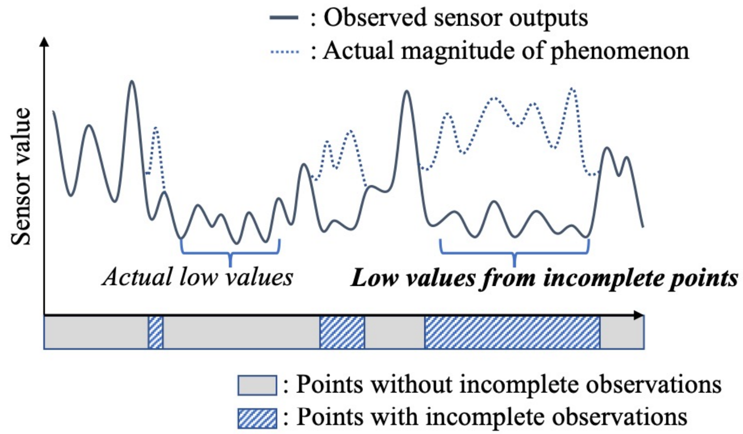
Proportion of correct prediction period	0.89
Rate of false prediction	0.016



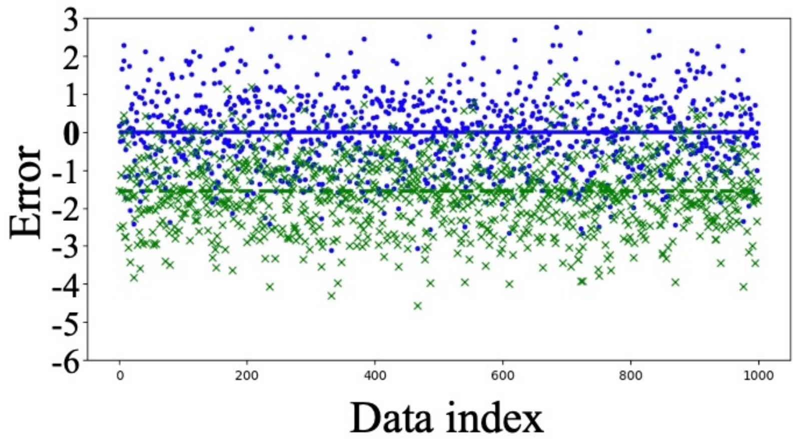
Arm sensor can be replaced with bed sensors

Thank you!

- We formulate a novel problem of learning a regression function for a sensor magnitude with asymmetrically corrupted data. This is vital for applications where the sensor is susceptible to unidentifiable incomplete observations.



- We derive an unbiased and consistent learning algorithm (U2 regression) for this problem with the new class of loss functions.



- Extensive experiments on synthetic and six real-world regression tasks including a real use case for healthcare demonstrate the effectiveness of the proposed method.