Adaptive Robust Learning using Latent Bernoulli Variables

Aleksandr Karakulev[1] Dave Zachariah[1] Prashant Singh[1,2]

[1]Uppsala University, Sweden

[2]Science for Life Laboratory, Sweden

UPPSALA UNIVERSITET

SciLifeLab

Motivation

Standard setting: given i.i.d. observations $z_1, z_2, \ldots, z_n$, find the true distribution $p(z \,|\, \theta_\mathbf{o})$ in the family $p(z \,|\, \theta)$, $\theta \in \Theta$

$$\theta_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p(z_i \,|\, \theta) = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \underbrace{-\ln p(z_i \,|\, \theta)}_{\ell_\theta(z_i)}$$

Real world: observations are corrupted (Huber's model)

$$z \sim (1 - \varepsilon) p(z) + \varepsilon\, q(z)$$

- errors in data collection,
- oversights in labeling,
- inaccurate measurements,
- malicious attacks,
- etc.

Motivation

Standard setting: given i.i.d. observations $z_1, z_2, \ldots, z_n$, find the true distribution $p(z \,|\, \theta_\mathbf{o})$ in the family $p(z \,|\, \theta)$, $\theta \in \Theta$

$$\theta_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p(z_i \,|\, \theta) = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \underbrace{-\ln p(z_i \,|\, \theta)}_{\ell_\theta(z_i)}$$

Real world: observations are corrupted (Huber's model)

$$z \sim (1 - \varepsilon) p(z) + \varepsilon q(z)$$

- errors in data collection,
- oversights in labeling,
- inaccurate measurements,
- malicious attacks,
- etc.

RLVI: Robust Learning via Variational Inference

Latent variables: for each z_i define

$$t_i = \begin{cases} 1 & \text{if } z_i \sim p(z) \\ 0 & \text{if } z_i \sim q(z) \end{cases} \quad \Longrightarrow \quad \max_{\substack{\theta \in \Theta \\ t_i \in \{0,1\}}} \prod_{i=1}^{n} p(z_i \,|\, \theta)^{t_i}$$

Marginalize likelihood using $p(t_i \,|\, 1{-}\varepsilon) = (1-\varepsilon)^{t_i}\, \varepsilon^{1-t_i}$:

$$\max_{\substack{\theta \in \Theta \\ \varepsilon \in (0,1)}} \underbrace{\sum_{t} \prod_{i=1}^{n} p(z_i \,|\, \theta)^{t_i} p(t_i \,|\, 1{-}\varepsilon)}_{p(z_1,\ldots,z_n \,|\, \theta, \varepsilon)}$$

Variational bound using $r(t_i \,|\, \pi_i) = \pi_i^{t_i}\, (1-\pi_i)^{1-t_i}$:

$$\ln p(z_1, \ldots, z_n \,|\, \theta, \varepsilon) \geq \mathrm{ELBO}(\theta, \pi, \varepsilon)$$

$$= \boxed{-\sum_{i=1}^{n} \pi_i \ell_\theta(z_i) - \mathrm{KL}\left[r(t \,|\, \pi) \,||\, p(t \,|\, 1{-}\varepsilon)\right]}$$

RLVI: Robust Learning via Variational Inference

Latent variables: for each z_i define

$$t_i = \begin{cases} 1 & \text{if } z_i \sim p(z) \\ 0 & \text{if } z_i \sim q(z) \end{cases} \quad \Longrightarrow \quad \max_{\substack{\theta \in \Theta \\ t_i \in \{0,1\}}} \prod_{i=1}^{n} p(z_i \,|\, \theta)^{t_i}$$

Marginalize likelihood using $p(t_i \,|\, 1{-}\varepsilon) = (1-\varepsilon)^{t_i}\, \varepsilon^{1-t_i}$:

$$\max_{\substack{\theta \in \Theta \\ \varepsilon \in (0,1)}} \underbrace{\sum_{t} \prod_{i=1}^{n} p(z_i \,|\, \theta)^{t_i} p(t_i \,|\, 1{-}\varepsilon)}_{p(z_1,\ldots,z_n \,|\, \theta, \varepsilon)}$$

Variational bound using $r(t_i \,|\, \pi_i) = \pi_i^{t_i}\,(1-\pi_i)^{1-t_i}$:

$$\ln p(z_1, \ldots, z_n \,|\, \theta, \varepsilon) \geq \mathrm{ELBO}(\theta, \pi, \varepsilon)$$

$$= \boxed{-\sum_{i=1}^{n} \pi_i \ell_\theta(z_i) - \mathrm{KL}\left[r(t \,|\, \pi) \,||\, p(t \,|\, 1{-}\varepsilon)\right]}$$

RLVI: Robust Learning via Variational Inference

Latent variables: for each z_i define

$$t_i = \begin{cases} 1 & \text{if } z_i \sim p(z) \\ 0 & \text{if } z_i \sim q(z) \end{cases} \quad \Longrightarrow \quad \max_{\substack{\theta \in \Theta \\ t_i \in \{0,1\}}} \prod_{i=1}^{n} p(z_i \,|\, \theta)^{t_i}$$

Marginalize likelihood using $p(t_i \,|\, 1{-}\varepsilon) = (1-\varepsilon)^{t_i}\, \varepsilon^{1-t_i}$:

$$\max_{\substack{\theta \in \Theta \\ \varepsilon \in (0,1)}} \underbrace{\sum_{t} \prod_{i=1}^{n} p(z_i \,|\, \theta)^{t_i} p(t_i \,|\, 1{-}\varepsilon)}_{p(z_1,\ldots,z_n \,|\, \theta, \varepsilon)}$$

Variational bound using $r(t_i \,|\, \pi_i) = \pi_i^{t_i}\, (1-\pi_i)^{1-t_i}$:

$$\ln p(z_1, \ldots, z_n \,|\, \theta, \varepsilon) \geq \mathrm{ELBO}(\theta, \pi, \varepsilon)$$

$$= \boxed{-\sum_{i=1}^{n} \pi_i \ell_\theta(z_i) - \mathrm{KL}\left[r(t \,|\, \pi) \,||\, p(t \,|\, 1{-}\varepsilon)\right]}$$

Numerical Optimization

Thus we obtain the objective:

$$\mathcal{L}(\theta, \pi) := \sum_{i=1}^{n} \left(\pi_i \ell_\theta(z_i) + \pi_i \ln \frac{\pi_i}{\pi_{\text{avg}}} + (1 - \pi_i) \ln \frac{1 - \pi_i}{1 - \pi_{\text{avg}}} \right)$$

$$\implies \theta_{\text{RLVI}} = \arg \min_{\theta \in \Theta} \min_{\pi \in (0,1)^n} \mathcal{L}(\theta, \pi)$$

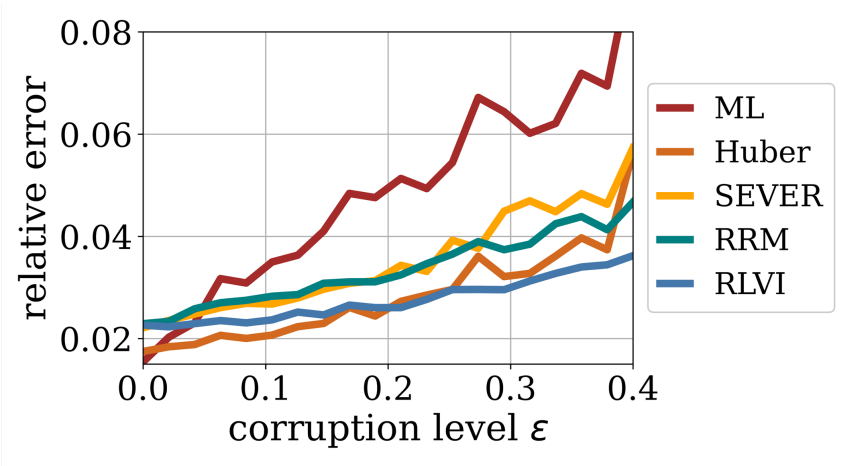
EM algorithm

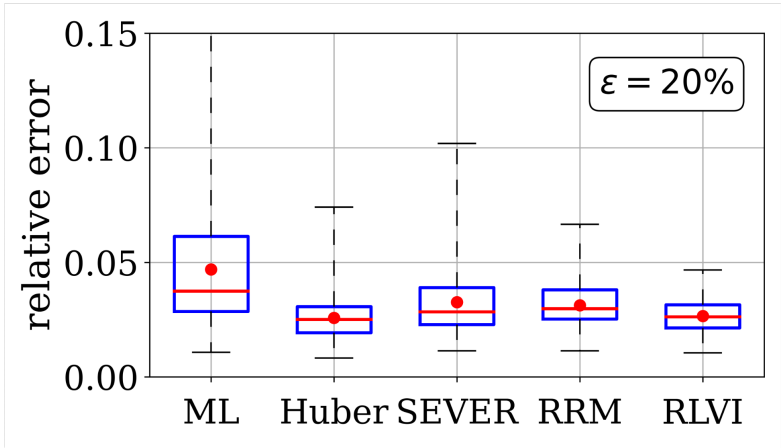
input: data $z_1, \ldots, z_n$
repeat
  update π by convex optimization // E-step
  update θ by $\min_{\theta \in \Theta} \sum_i \pi_i \ell_\theta(z_i)$ // M-step
until convergence
output: θ

Standard Statistical Learning

Linear regression from corrupted data

$$z \sim (1-\varepsilon)p(z\,|\,\theta_{\mathrm{o}}) + \varepsilon q(z)$$

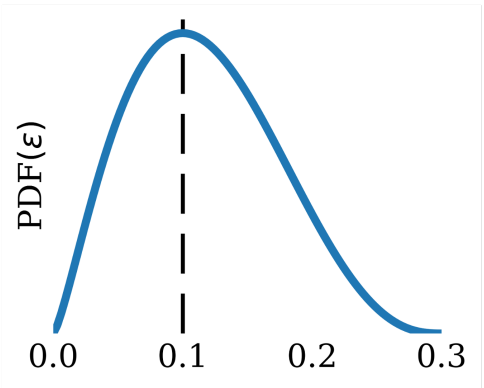


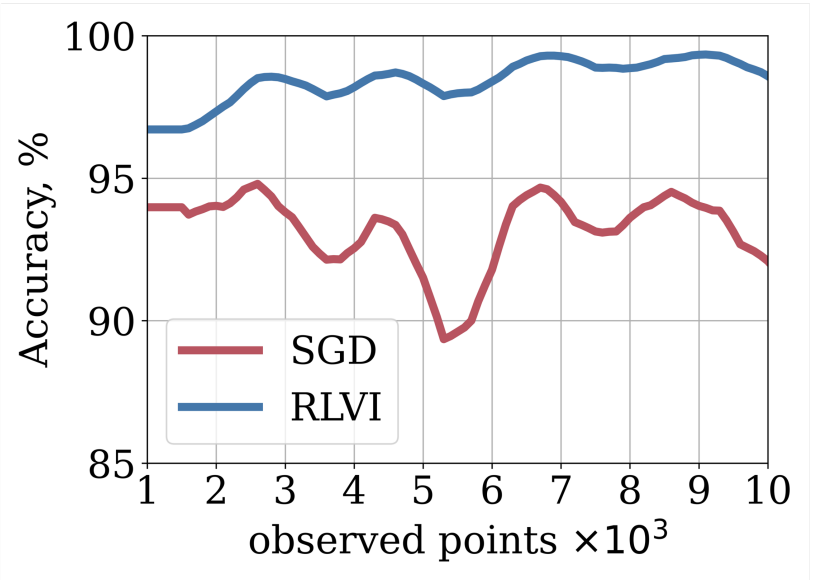


- RLVI does not depend on the estimate $\tilde{\varepsilon} \geq \varepsilon$
- Can apply to other problems (e.g., classification, PCA)

Online Learning

Learning classifier from data that arrives continually:





- Different number of labels are corrupted in each batch
- RLVI infers ε and learns the model **adaptively**

Deep Learning: Overfitting

Overparameterized models overfit – use hard truncation:

$$\pi_i < \tau \implies \pi_i \leftarrow 0 \quad \text{(by the type II error criterion)}$$

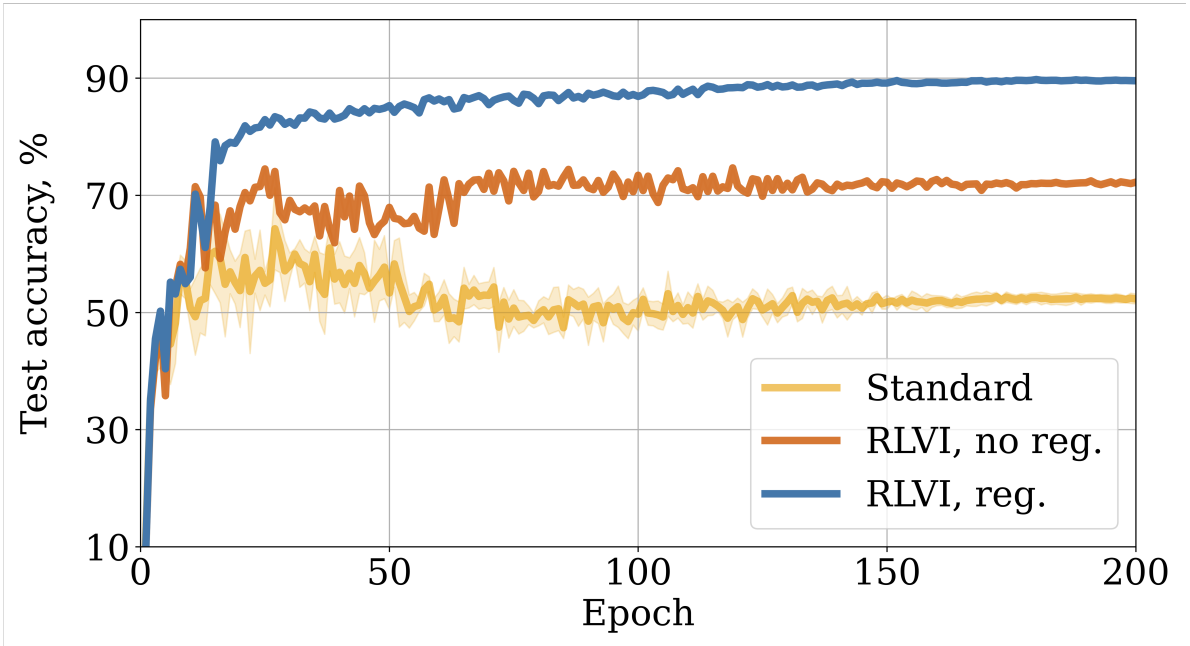


Figure: CIFAR-10 with 45% of training labels corrupted (ResNet-18)

- Truncation improves generalization
- See benchmarks in the paper: MNIST, CIFAR-10, CIFAR-100
- Real noise example: Food-101

Conclusion

Our method (RLVI) is

- widely applicable given any likelihood function
- robust to a wide class of corruption sources
- adaptive and parameter tuning-free
- scalable for large data sets and deep learning models

Paper

Code

Contact: aleksandr.karakulev@it.uu.se