# Addressing Label Shift in Distributed Learning via Entropy Regularization


UNIVERSITY OF OSLO  integreat

Zhiyuan Wu[1*], Changkyu Choi[3*], Xiangcheng Cao[4], Volkan Cevher[4], Ali Ramezani-Kebrya[1,2]

[1]University of Oslo  [2]Integreat, Norwegian Centre for Knowledge-driven Machine Learning  [3]UiT The Arctic University of Norway  [4]LIONS, EPFL

VISUAL INTELLIGENCE  lions@epfl  THE ARCTIC UNIVERSITY OF NORWAY · UiT


## Motivation

### Label Shift

In real-world applications, the label distribution $p(\boldsymbol{y})$ may change from training and test, while the conditional distribution $p(\boldsymbol{x}|\boldsymbol{y})$ remains unchanged.

### Challenge in Distributed Learning

In multi-node setups, each node may face both intra-node and inter-node label shifts. Standard ERM is insufficient to address this distributional heterogeneity.

### Goal

- Estimate local label shift ratios from few unlabeled test samples, with statistical guarantees.
- Train a global model robust to label shift, in a privacy-preserving distributed framework.

### Key Idea

- **VRLS:** Entropy-regularized predictor and density ratio estimation.
- **IW-ERM:** Weighted ERM using local ratio estimates to optimize true test risk.

## Problem Setup under Single-Node Case

### Importance Ratio under Label Shift

Given training data $(\boldsymbol{x}_i, \boldsymbol{y}_i) \sim p^{\mathrm{tr}}$ and test data $(\boldsymbol{x}_i, \boldsymbol{y}_i) \sim p^{\mathrm{te}}$, define the importance ratio as:

$$\boldsymbol{r}(\boldsymbol{y}) = \frac{p^{\mathrm{te}}(\boldsymbol{y})}{p^{\mathrm{tr}}(\boldsymbol{y})}.$$

### Entropy Regularization

Train $f_\theta(\boldsymbol{x})$ to approximate $p^{\mathrm{tr}}(\boldsymbol{y}|\boldsymbol{x})$ using cross-entropy loss with entropy-based regularization:

$$\Omega(f_\theta) = \sum_{c=1}^{m} \mathrm{softmax}[f_{\boldsymbol{\theta}}(\boldsymbol{x})]_c \log\left(\mathrm{softmax}[f_{\boldsymbol{\theta}}(\boldsymbol{x})]_c\right).$$

### Why Entropy?

Helps mitigate overconfidence, improving calibration for ratio estimation.

## VRLS Algorithm

**Input:** Training set $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n_{\mathrm{tr}}}$ and unlabeled test set $\{\boldsymbol{x}_j\}_{j=1}^{n_{\mathrm{te}}}$.

**Train predictor:**

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \left[\frac{1}{n_{\mathrm{tr}}}\sum_{i=1}^{n_{\mathrm{tr}}} \ell_{\mathrm{CE}}\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \boldsymbol{y}_i\right) + \zeta\,\Omega(f_{\boldsymbol{\theta}})\right]$$

**Estimate importance ratio:**

$$\hat{\boldsymbol{r}} = \arg\max_{\boldsymbol{r}\in\mathbb{R}_+^m} \frac{1}{n_{\mathrm{te}}}\sum_{j=1}^{n_{\mathrm{te}}} \log\left[f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}_j)^\top \boldsymbol{r}\right]$$

**Apply $\hat{\boldsymbol{r}}$:** Distributed training with importance weighting.

## Multi-Node Extension

**Setup:** For a system of $K$ nodes, each node $k$ has training distribution $p_k^{\mathrm{tr}}$ and test distribution $p_k^{\mathrm{te}}$. A global model $h_{\boldsymbol{w}}: \mathcal{X} \to \mathcal{Y}$ is trained across all nodes. Under label shift:

$$r_k(\boldsymbol{y}) = \frac{\sum_{j=1}^{K} p_j^{\mathrm{te}}(\boldsymbol{y})}{p_k^{\mathrm{tr}}(\boldsymbol{y})}.$$

**Local Risk:** Each node $k$ aims to minimize its true expected test risk:

$$R_k(h_w) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p_k^{\mathrm{te}}}\left[\ell(h_w(\boldsymbol{x}), \boldsymbol{y})\right].$$

## IW-ERM with VRLS in Distributed Learning

### Phase 1: Local Ratio Estimation (VRLS)

- Each node $k \in [K]$ estimates its local label shift ratio $\hat{\boldsymbol{r}}_k$ by performing the following steps:
  - ▶ Train a local predictor $\hat{f}_{k,\boldsymbol{\theta}}$ with entropy regularization.
  - ▶ Estimate the importance ratio $\hat{\boldsymbol{r}}_k$ by optimizing the VRLS objective on local test data.

### Phase 2: Ratio Aggregation

- Each node transmits its local label shift estimate $\hat{\boldsymbol{r}}_k$ to a central aggregator. The final estimate is:

$$\hat{\boldsymbol{r}}_k(\boldsymbol{y}) = \frac{1}{\hat{p}_k^{\mathrm{tr}}(\boldsymbol{y})}\sum_{j=1}^{K}\hat{p}_j^{\mathrm{te}}(\boldsymbol{y}).$$

### Phase 3: Global IW-ERM Training

- Train global model $h_{\boldsymbol{w}}$ with reweighted loss that incorporates the estimated label shift ratios:

$$\min_{\boldsymbol{w}} \sum_{k=1}^{K}\frac{1}{n_k^{\mathrm{tr}}}\sum_{i=1}^{n_k^{\mathrm{tr}}} \hat{\boldsymbol{r}}_k(\boldsymbol{y}_{k,i})\,\ell(h_{\boldsymbol{w}}(\boldsymbol{x}_{k,i}), \boldsymbol{y}_{k,i}).$$

**Output:** A global model $h_{\boldsymbol{w}}$ adapted to heterogeneous label shifts across distributed nodes.

## Theoretical Highlights

### Ratio Estimation:

Under mild assumptions, the VRLS estimator $\hat{\boldsymbol{r}}$ converges to the true optimum $\boldsymbol{r}_f^\star$ with high probability:

$$\left\|\hat{\boldsymbol{r}} - \boldsymbol{r}_f^\star\right\| \le O\left(\frac{1}{\sqrt{n^{\mathrm{te}}}}\right) + \mathrm{calib.\ error}(\boldsymbol{\theta}, \boldsymbol{\theta}^*),$$

due to the MLE structure and bounded softmax outputs.

### Distributed Convergence:

Let $h(T)$ denote the baseline convergence rate (e.g., $O(\frac{1}{\sqrt{T}})$ for SGD). IW-ERM with VRLS satisfies:

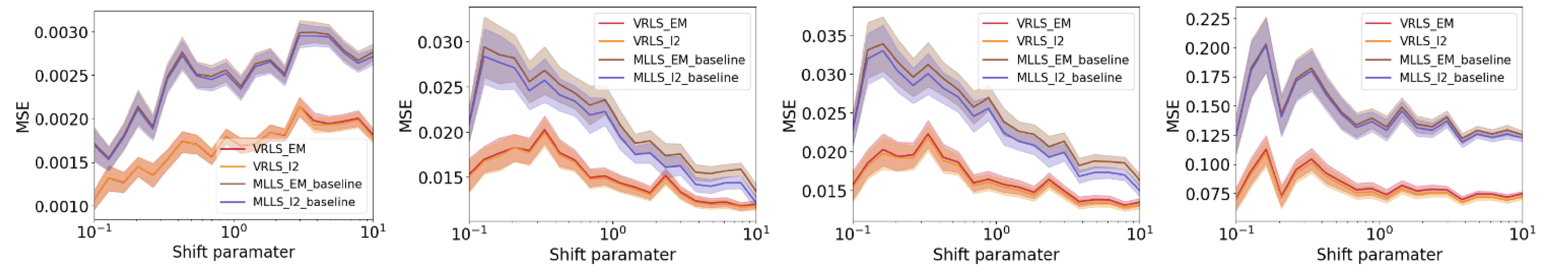
$$\ell(h_{\boldsymbol{w}}) - \ell^* \le O(R_{\max}) \cdot h(T).$$

Communication overhead is minimal: each node transmits only a one-shot $\hat{\boldsymbol{r}}_k$, preserving privacy and incurring no extra per-iteration cost.

## Experiments and Results

### Single-Node Ratio Estimation:

- Datasets: MNIST and CIFAR-10, with synthetic label shift severity $\alpha$.
- VRLS achieves lower MSE compared to EM-based baseline.




MNIST  CIFAR-10  CIFAR-10, Relaxed  CIFAR-10, Relax-m

### Multi-Node Setup:

- Up to 200 nodes, each with local distributions $p_k^{\mathrm{tr}}$ and $p_k^{\mathrm{te}}$.
- Our IW-ERM with VRLS outperforms FedAvg, FedProx, etc.
- Each node transmits its local $\hat{\boldsymbol{r}}_k$ once to the server; no raw data is shared.

Accuracy across 100 Nodes on Fashion-MNIST:

| Method | Accuracy |
|---|---|
| IW-ERM | **0.7520 ± 0.0209** |
| IW-ERM (small) | 0.7376 ± 0.0099 |
| FedAvg | 0.5472 ± 0.0297 |
| FedBN | 0.5359 ± 0.0306 |
| FedProx | 0.5606 ± 0.0070 |
| SCAFFOLD | 0.5774 ± 0.0036 |
| Upper Bound | 0.8273 ± 0.0041 |

Accuracy across varying node numbers on CIFAR-10:

| Nodes | IW-ERM | FedAvg | FedBN |
|---|---|---|---|
| 100 | **0.5354** | 0.3915 | 0.1537 |
| 200 | **0.6216** | 0.5942 | 0.1753 |

## Future Directions

Beyond $p(\boldsymbol{x}|\boldsymbol{y})$ invariance: handling mixed or relaxed shift conditions.

Improving ratio estimation: tighter generalization bounds and faster statistical convergence.

Scaling VRLS to large-scale and privacy-preserving distributed systems.

## Acknowledgments


Experiments were done on UiO Sigma2 resources. This work was supported by the Research Council of Norway through Integreat, Norwegian Centre for Knowledge-driven Machine Learning (332645) and Visual Intelligence, Norwegian Centre for Research-based Innovation (309439). The work of Volkan Cevher was supported by the Hasler Foundation Program: Hasler Responsible AI (21043), the Army Research Office (W911NF-24-1-0048), and the Swiss National Science Foundation (200021-205011).