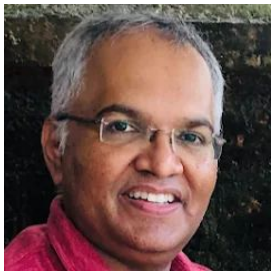
# Learning Model Uncertainty as Variance Minimizing Instance Weights


Nishant Jain

Karthikeyan Shanmugham



Pradeep Shenoy

Google Research

# Introduction

- In sensitive use-cases, models are required to be highly accurate on examples they are confident about .
- This leads to a requirement of also modelling the uncertainty in the model predictions.
- Problems like selective classification, neural network calibration, label uncertainty, etc. involve modelling this uncertainty estimate.

# Formulating a Selective Classification Problem

- Data Distribution : $P(X,Y) \quad X \in \mathcal{X} \quad Y \in \mathcal{Y}$
- Prediction Model $f : \mathcal{X} \mapsto \mathcal{Y}$
- Risk associated : $\mathbb{E}_{P(X,Y)}[\ell(f(x), y)]$
- Selection function $g$ : $\mathcal{X} \mapsto \{0, 1\}$
- Selective classifier is a pair ($f$,$g$):

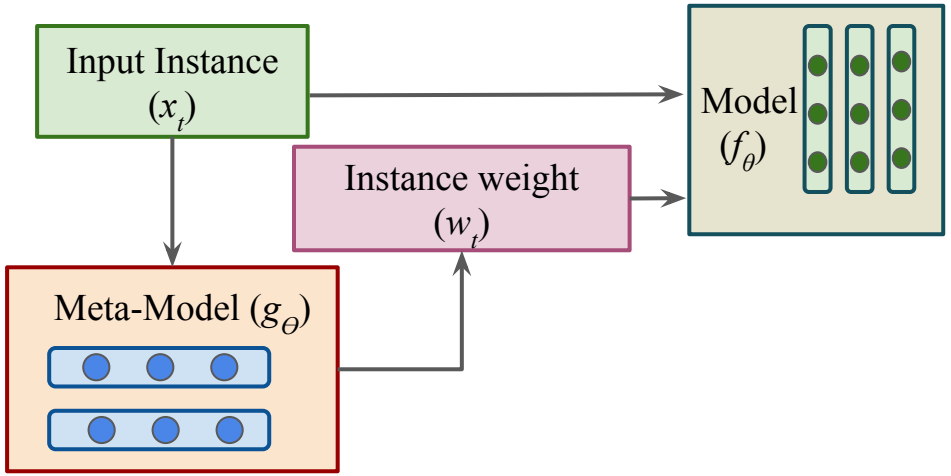
$$(f, g)(x) = \begin{cases} f(x) & \text{if } g(x)=1; \\ \text{can't predict} & \text{if } g(x)=0; \end{cases}$$

- Coverage : $\mathbb{E}_{P(X,Y)}$ [$g(x)$]

# **Re**-weighting for dropout **Va**riance **R**eduction

# Instance Conditional Weights in ReVaR

- Train Set : $\{x_t, y_t\}$
- Special Validation Set: $(X^s, Y^s)$
- Classifier : $f_\theta$
- Uncertainty-Scorer Network : $g_\Theta$




Objective:

$$\theta^* = \arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} g_\Theta(x_i) \cdot l(y_i, f_\theta(x_i)) \quad s.t.\ \Theta^* = \arg\min_\Theta \mathcal{L}_{meta}(X^s, Y^s, \theta^*)$$

# Variance Minimization as Meta Regularization

- We propose the following objective for our U-Scorer to capture uncertainty:

$$\mathcal{L}_{meta} = \mathcal{L}_c(X_s, Y_s) + \mathcal{L}_{eps}(\theta, X_s) = \sum_{j=1}^{M} l(y_j^s, f_\theta(x_j^s)) + l_{eps}(\theta, x_j^s)$$

$$l_{eps}(\theta, x) \approx \frac{1}{K}\left(\sum_{k=1}^{K}(f_{\mathcal{D}_k \odot \theta}(x) - E[f_{\mathcal{D}_k \odot \theta}(x)])^2\right)$$

$K$ : total forward passes per example  $M$ : size of validation set

$D_k$ : dropout mask sampled in $k^{\text{th}}$ pass

- High U-Score → High Uncertainty.

# Analysis on Various Uncertainty Sources

- **Three kinds of uncertainty:**
  a) Samples that are atypical with respect to train but typical with respect to validation.
  b) Samples where label noise is present.
  c) Samples where uncertainty in the label is due to some unobserved latent features that affect the label.

  a, c : Epistemic Uncertainty.
  b : Aleatoric Uncertainty.

# Generative Model for the synthetic data

$$Y = W_{\text{data}}^T X + (\mathcal{N}(0,1) \cdot [c + G^T X])$$

$$X \in \mathbb{R}^{72\times 1} \quad X = [X_c X_e], X_c \in \mathbb{R}^{48\times 1}, X_e \in \mathbb{R}^{24\times 1}$$

$$X^{\text{train}} \sim \mathcal{N}(\mu, \Sigma) \quad X^{\text{val}} \sim \mathcal{N}(\mu', \Sigma)$$

$$\mu' = \mu + s\mathcal{N}(\mu_s, \Sigma_s) \quad \text{s: scalar}$$

$$W_{\text{data}}^T = [W_c^T \; W_e^T] \quad W_c \in \mathbb{R}^{48\times 1}, \; W_e \in \mathbb{R}^{24\times 1}$$

# Baselines

- **MWN**: loss input to the meta-network and meta objective and train objective are just MLE.
- **IBR**: Our method except for the meta-regularizer involving variance minimization.

# Various Scenarios

**Scenario 1 - Sample Dependent Label Noise and No Shift:**. $c = 0, s = 0, G \neq 0.$

Label noise scales as $|G^T X|^2$

**Scenario 2 - Sample Dependent Label noise and Covariate Shift:** $c = 0, G \neq 0, s \neq 0.$

**Scenario 3 - Hardness due to missing relevant features:** We set $c = 1, G = 0, s = 0$

only $X_c$ is available to the learner in both train and validation

| S | Target | MWN | IBR | Ours |
|---|---|---|---|---|
| 1 | $\frac{\lambda_1}{\lvert G^T X\rvert^2}$ | 0.77 | 0.78 | 0.84 |
| 2 | $\frac{\lambda_1}{\lvert G^T X\rvert^2} + \lambda_2 \cdot h$ | 0.58 | 0.62 | 0.80 |
| 3 | $\frac{\lambda_1}{W_e^T \sum(X_e \mid X_c) W_e}$ | 0.46 | 0.52 | 0.81 |
| 4 | $\frac{\lambda_1}{W_e^T \sum(X_e \mid X_c) W_e} + \lambda_2 \cdot h$ | 0.51 | 0.57 | 0.82 |
| 5 | $\lambda_1 \cdot \mathcal{U}(0, 1)$ | 0.44 | 0.58 | 0.84 |

$h = (x - \mu)^2$ $R^2$ metric. $\lambda_1, \lambda_2$ are fitting coefficients.

# Various Scenarios

**Scenario 4 - Dropping Features and covariate shift in validation set:**. We set $c = 1, G = 0, s > 0$ only $X_c$ is available to the learner.

**Scenario 5 - Spurious Feature Shift:**. $c = 1, G = 0, s > 0$. Further $W_e = 0$. learner sees $X$ for both test and validation.

**Val Set:** $\mathcal{N}(\mu', \sum')$ such that the distribution of $X_c$ remains same and the distribution of $X_e$ changes.

| S | Target | MWN | IBR | Ours |
|---|---|---|---|---|
| 1 | $\frac{\lambda_1}{\lvert G^T X \rvert^2}$ | 0.77 | 0.78 | 0.84 |
| 2 | $\frac{\lambda_1}{\lvert G^T X \rvert^2} + \lambda_2 \cdot h$ | 0.58 | 0.62 | 0.80 |
| 3 | $\frac{\lambda_1}{W_e^T \sum(X_e \lvert X_c) W_e}$ | 0.46 | 0.52 | 0.81 |
| 4 | $\frac{\lambda_1}{W_e^T \sum(X_e \lvert X_c) W_e} + \lambda_2 \cdot h$ | 0.51 | 0.57 | 0.82 |
| 5 | $\lambda_1 \cdot \mathcal{U}(0, 1)$ | 0.44 | 0.58 | 0.84 |

# Real-World Scenarios

# Results

- **Selective Classification:** Area under accuracy rejection curve (AUARC)

| | Selective Classification Baselines | | | | | New Baselines | | ReVaR |
|---|---|---|---|---|---|---|---|---|
| | SR | MCD | DG | SN | SAT | VR | MBR | Ours |
| DR(In-Dist.) | 92.87 ± 0.1 | 93.44 ± 0.0 | 93.07 ± 0.1 | 93.13 ± 0.1 | 93.56 ± 0.1 | 92.55 ± 0.1 | 92.95 ± 0.2 | **94.12 ± 0.1** |
| DR(OOD) | 87.67 ± 0.1 | 88.27 ± 0.1 | 88.07 ± 0.2 | 88.56 ± 0.1 | 88.97 ± 0.2 | 87.91 ± 0.1 | 88.06 ± 0.3 | **89.94 ± 0.1** |
| CIFAR-100 | 92.30 ± 0.1 | 92.71 ± 0.1 | 92.22 ± 0.2 | 82.10 ± 0.1 | 92.80 ± 0.3 | 92.17 ± 0.1 | 92.50 ± 0.1 | **93.20 ± 0.1** |
| ImageNet-100 | 93.10 ± 0.0 | 94.20 ± 0.0 | 93.50 ± 0.1 | 93.60 ± 0.1 | 94.12 ± 0.2 | 93.25 ± 0.1 | 93.88 ± 0.2 | **94.95 ± 0.1** |
| ImageNet-1K | 86.20 ± 0.1 | 87.30 ± 0.0 | 86.90 ± 0.2 | 86.80 ± 0.1 | 87.10 ± 0.3 | 86.95 ± 0.1 | 86.35 ± 0.1 | **88.20 ± 0.2** |

- **Model Calibration:** Expected Calibration Error (ECE)

| | Calibration Baselines | | | | | New Baselines | | ReVaR |
|---|---|---|---|---|---|---|---|---|
| | CE | MMCE | Brier | FLSD-53 | AdaFocal | VR | MBR | Ours |
| DR(In-Dist.) | 7.7 ± 0.1 | 6.7 ± 0.0 | 5.8 ± 0.1 | 5.0 ± 0.1 | **3.6 ± 0.1** | 7.4 ± 0.1 | 7.1 ± 0.1 | 3.8 ± 0.1 |
| DR(OOD) | 9.1 ± 0.1 | 7.9 ± 0.1 | 6.8 ± 0.1 | 6.1 ± 0.1 | **5.9 ± 0.2** | 8.6 ± 0.1 | 8.4 ± 0.3 | 6.4 ± 0.1 |
| CIFAR-100 | 16.6 ± 0.1 | 15.3 ± 0.1 | 6.9 ± 0.1 | 5.9 ± 0.1 | **2.3 ± 0.1** | 9.1 ± 0.1 | 10.7 ± 0.1 | 3.1 ± 0.1 |
| ImageNet-100 | 9.6 ± 0.0 | 9.1 ± 0.0 | 6.7 ± 0.1 | 5.8 ± 0.1 | **2.7 ± 0.2** | 8.2 ± 0.1 | 7.9 ± 0.1 | **2.7 ± 0.1** |
| ImageNet-1K | 3.0 ± 0.1 | 9.0 ± 0.0 | 3.4 ± 0.1 | 16.1 ± 0.1 | **2.1 ± 0.1** | 3.5 ± 0.1 | 3.2 ± 0.1 | 2.6 ± 0.1 |

# Results

- **Input Dependent Label Noise:**

Accuracy

| | MCD | MWN | L2R | FSR | Ours |
|---|---|---|---|---|---|
| Inst.CIFAR-100 | 61.12 | 65.89 | 67.12 | 70.21 | **71.87** |
| Clothing1M | 68.78 | 73.56 | 72.97 | 73.86 | **73.97** |

KL-Divergence (uncertain labels)

| | Plex | Plex+ours |
|---|---|---|
| IN-100H | 0.75 | **0.71** |
| CF-100H | 0.49 | **0.47** |

- **Shifted Test, Validation Sets:** AUARC

| ImageNet-A | | | ImageNet-C | | | ImageNet-R | | |
|---|---|---|---|---|---|---|---|---|
| Ours | MCD | SAT | Ours | MCD | SAT | Ours | MCD | SAT |
| **9.98** | 8.44 | 8.91 | **65.9** | 63.7 | 64.2 | **68.8** | 66.8 | 67.1 |

| Data | MCD | SAT | Revar | Revar-PV |
|---|---|---|---|---|
| Camelyon | 74.99 | 75.16 | 76.32 | 78.12 |
| iWildCam | 76.07 | 76.17 | 77.98 | 79.86 |

Thank You!