Optimizing Black-box Metrics with Iterative Example Weighting

Gaurush Hiranandani (UIUC)
Jatin Mathur (UIUC)
Harikrishna Narasimhan (Google Research)
Mahdi Milani Fard (Google Research)
Oluwasanmi Koyejo (Google Research & UIUC)

MOTIVATION

- **Distribution Shift**
  - (Large) training data and (small) test data are from different time frames (Domain Shift)
  - Proxy label in training data (clicks vs relevance)
    - Instance and Label-dependent noise (test metric transforms to an instance-weighted metric for training)
- **Fairness:** Metric and the protected group attribute can only be used for evaluating fairness and can not be used for training
- **General Metrics:**
  - Revenue, Purchase, etc.
  - Other online metrics

PROBLEM STATEMENT

We aim to optimize an evaluation measure (metric) of the form

$$\mathcal{E}(h) = \psi\big(C(\mathrm{h})\big),$$

where ψ is unknown (black-box), and

$C(\mathrm{h})$ is the confusion matrix of a classifier h,

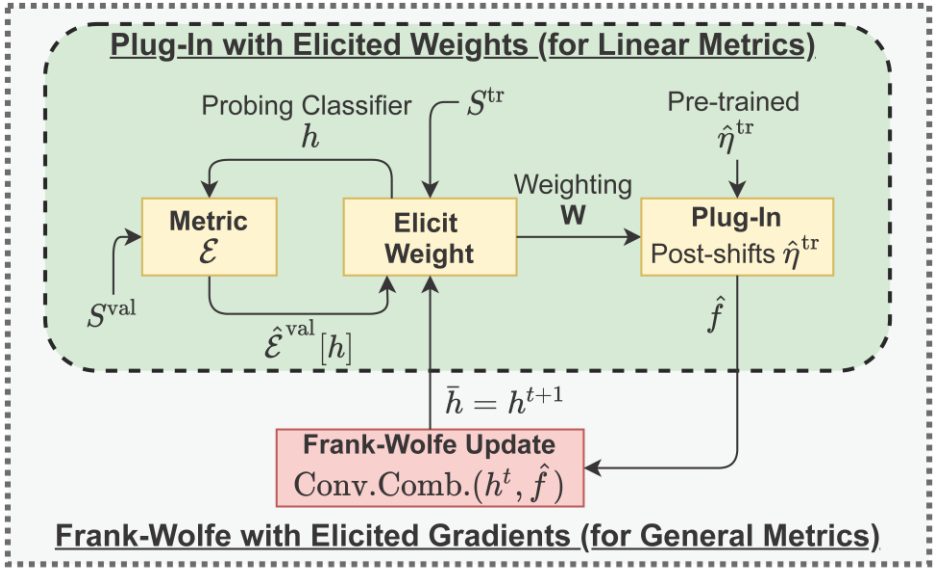
in the presence of an oracle that when queried for a classifier "h" responds with $\mathcal{E}(h)$.

e.g., under distribution shift, training set is sampled from a different distribution than the validation set, which has clean labels/features

APPROACH: FRANK WOLFE WITH ELICITED GRADIENTS

Iterative approach :

- Optimize a local-linear objective at each iteration
  - Obtain $w(x, y) \approx \nabla\psi(C(h))$ **Weight/Gradient Elicitation**
  - Once w is elicited, appropriately threshold conditional class estimator $\eta_i^{tr}(x) \approx \mathbb{P}(y = i \mid x)$ **Plug-in (Post-shift)**
- Combine previous iterates of classifiers
- Repeat until convergence



KEY STEP: WEIGHT/GRADIENT ELICITATION

We just perturb predictions using the trivial classifiers!

-- can be based on the modeling of the weighing function.

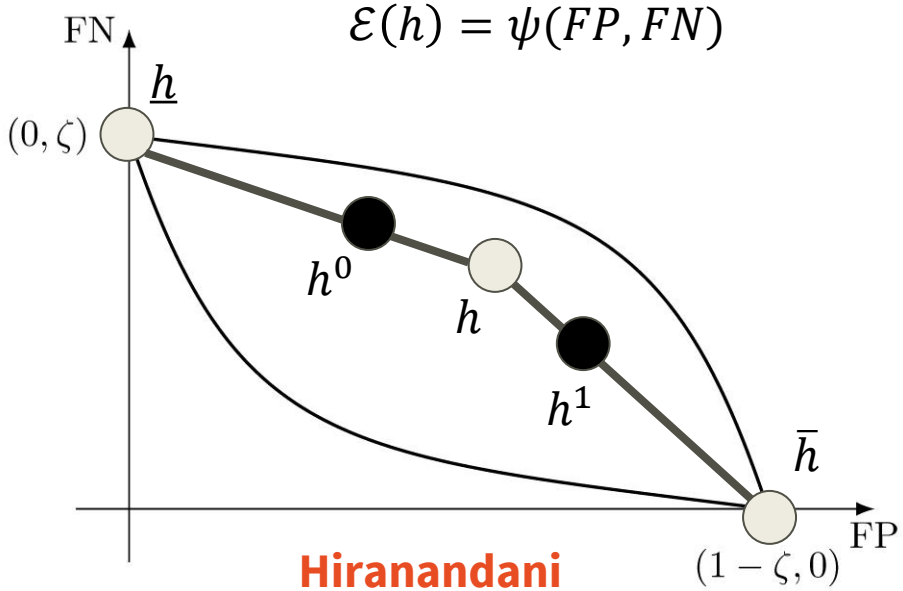
$$h^1(x) = (1 - \epsilon(x))h(x) + \epsilon(x)\,\bar{h}(x)$$

$$h^0(x) = (1 - \epsilon(x))h(x) + \epsilon(x)\,\underline{h}(x)$$

- Query $\mathcal{E}(h^1), \mathcal{E}(h^0)$ on the validation side
- Solve the system of equations to elicit weights

Suppose

$$\mathcal{E}(h) = \psi(FP, FN)$$



Hiranandani et al. (2019)

BENEFITS OVER PRIOR WORK

	Our Approach	Prior work such as Jiang et al. 2020
No Retraining of Model (Time Efficient)	✓	✗
Elegant Gradient Elicitation (Amenable to Deep Networks)	✓	✗
Constant Number of Perturbations (Query Efficient)	✓	✗
Theoretical Guarantees (Statistical Consistency)	✓	✗

EXPERIMENTS: OPTIMIZING BLACK-BOX METRICS

- Cifar10 dataset: Optimize accuracy under asymmetric label flip setting (Patrini et al. 2017)
  - $P(Y = k|X) = \eta(x) \rightarrow$ Resnet
  - $\approx 2.4\%$ improvement
- Adience Face-Image dataset: Optimize F-measure under domain shift based on age
  - $P(Y = k|X) = \eta(x) \rightarrow$ Resnet
  - $\approx 3.4\%$ improvement (statistically significant)
- Adult dataset: Optimize G-mean under proxy label setting (Jiang et al. 2020)
  - $P(Y = 1|X) = \eta(x) \rightarrow$ linear logistic regression
  - $\approx 0.3\%$ improvement (statistically significant) and also 5x faster
- Adult dataset: Optimize G-mean of group-wise confusion entries:

$$\mathcal{E}(h) = \left(TP_{male}TN_{male}TP_{female}TN_{female}\right)^{\frac{1}{4}}$$

  - Only validation data has sensitive group info that can only be used to evaluate classifier
  - $\approx 1\%$ improvement (statistically significant)

CONTRIBUTIONS

- Proposed the FW-EG method for optimizing black-box metrics of the confusion matrix
  - Framework includes common distribution shift settings as special cases
  - Able to handle general non-linear metrics
- Elegant estimation of the example weights, use them to iteratively post-shift a class probability estimator
- Agnostic to the choice of base model because does not require re-training -- thus amenable to deep networks
- Analyze the procedure's statistical properties
- Exhaustive experiments on various label noise, domain shift, and fair classification setups

THANK YOU!

Source Code: https://github.com/koyejolab/fweg