Optimizing Black-box Metrics with Adaptive Surrogates

Qijia Jiang[1], Olaoluwa (Oliver) Adigun[2],
Harikrishna Narasimhan[3], Mahdi M. Fard[3], Maya Gupta[3]

[1]Stanford, [2]USC, [3]Google Research

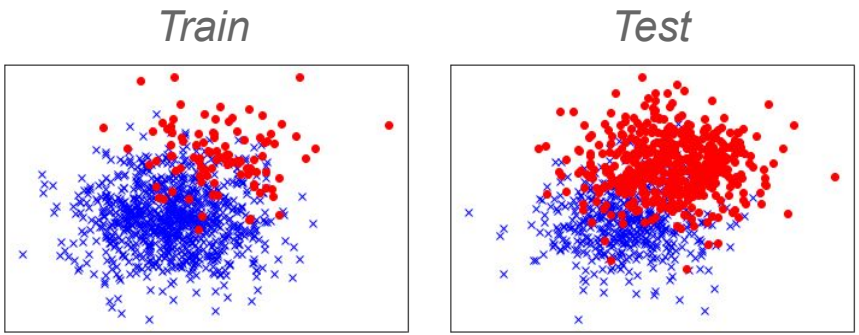
Misaligned Train-Test Metrics

Training objective often mis-aligned with the test evaluation metric

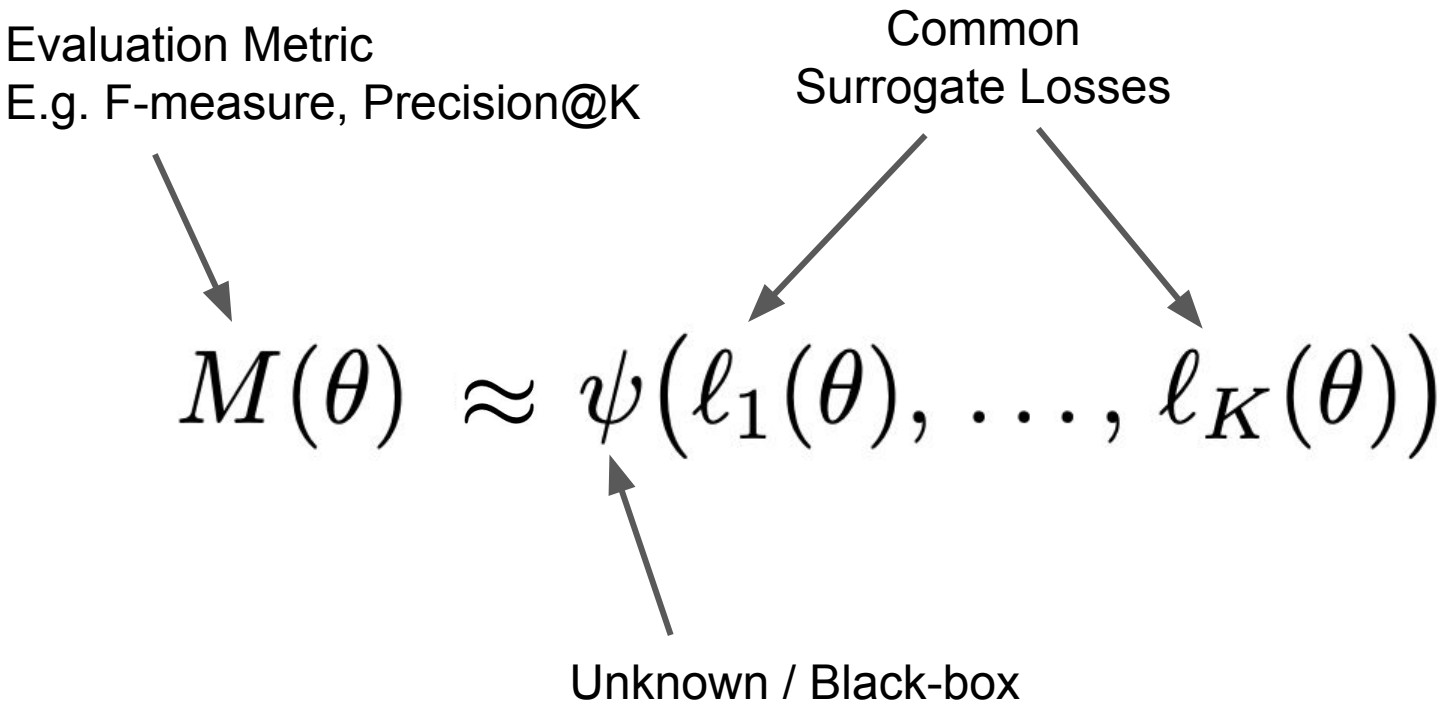
Evaluation metric is complex and is difficult to approximate with a smooth loss

F-measure	Prec@*k*
AUC-PR	Recall@*k*
G-mean	NDCG
H-mean	MAP
PRBEP	MRR

Training data drawn from a different distribution than the test data



Blackbox Metric w/ Compositional Structure



Classification with Noisy Labels

Evaluation metric on true labels (e.g. ratings)
(Small validation data)

Losses on cheap noisy labels (e.g. clicks)
(Training data)

$$M(\theta) \approx \psi\big(\ell_1(\theta), \ldots, \ell_K(\theta)\big)$$

Unknown / Black-box

Complex Ranking Metrics

Precision@10

Different smooth surrogates for the metric

$$M(\theta) \approx \psi\big(\ell_1(\theta), \ldots, \ell_K(\theta)\big)$$

Unknown / Black-box

Main Contributions

- Equivalent optimization problem in lower-dimensional space:

$$\min_{\theta \in \mathbb{R}^d} M(\theta) \longrightarrow$$ Optimization over K-dim surrogate space

- Solve reformulated problem using projected gradient descent with zeroth-order gradient estimates
- We show convergence to a stationary point of *M*
- Experiments on classification and ranking problems

Related Work

- **Optimizing closed-form metrics**
  - e.g. Joachims (2005), Kar et al. (2014), Narasimhan et al. (2015), Yan et al. (2018)
- **Optimizing black-box metrics**
  - Example-weighting (Zhou et al., 2019), Reinforcement learning (Huang et al., 2019), Teacher model (Wu et al., 2018)
  - Limited theoretical guarantees

Related Work

- **Optimizing closed-form metrics**
  - e.g. Joachims (2005), Kar et al. (2014), Narasimhan et al. (2015), Yan et al. (2018)
- **Optimizing black-box metrics**
  - Example-weighting (Zhou et al., 2019), Reinforcement learning (Huang et al., 2019), Teacher model (Wu et al., 2018)
  - Limited theoretical guarantees
- **This Paper**
  - Simple approach to combine a small set of useful surrogates to optimize a metric
  - Directly estimates only the local gradients needed for gradient descent training
  - Rigorous theoretical guarantees

$$M(\theta) \approx \psi(\boldsymbol{\ell}(\theta))$$

Reformulate as Optimization over Surrogate Space

- Space of achievable surrogate profiles:

$$\mathcal{L} := \{(\ell_1(\theta), \ldots, \ell_K(\theta)) \mid \theta \in \mathbb{R}^d\}$$

$$M(\theta) \approx \psi(\boldsymbol{\ell}(\theta))$$

Reformulate as Optimization over Surrogate Space

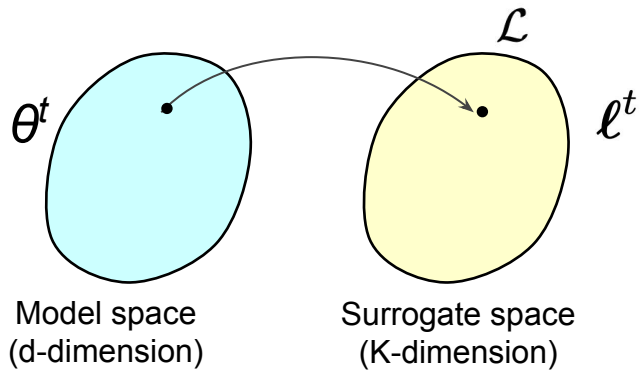
- Space of achievable surrogate profiles:

$$\mathcal{L} := \{(\ell_1(\theta), \ldots, \ell_K(\theta)) \,|\, \theta \in \mathbb{R}^d\}$$

- Reformulate as a constrained optimization over K-dim surrogate space:

$$\min_{\theta \in \mathbb{R}^d} M(\theta) \simeq \min_{\ell \in \mathcal{L}} \psi(\ell)$$

- Lower dim problem as usually $K \ll d$



$$M(\theta) \approx \psi(\boldsymbol{\ell}(\theta))$$

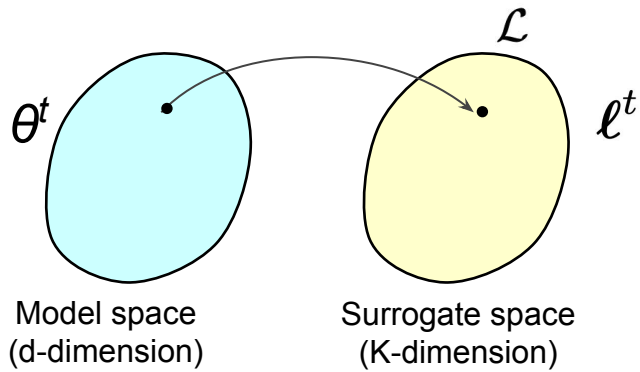
Projected Gradient Descent over Surrogate Space

- Apply projected gradient descent to solve reformulated problem

$$\min_{\ell \in \mathcal{L}} \psi(\ell)$$

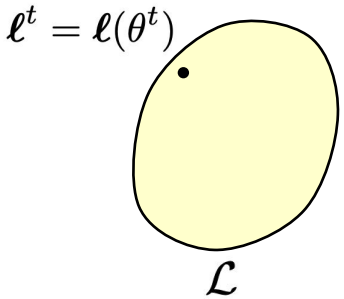
- Challenges:
  - ψ is not known
  - $\mathcal{L}$ is not explicitly available

How do you estimate gradients for ψ?
How do you project onto $\mathcal{L}$?



$$M(\theta) \approx \psi(\boldsymbol{\ell}(\theta))$$

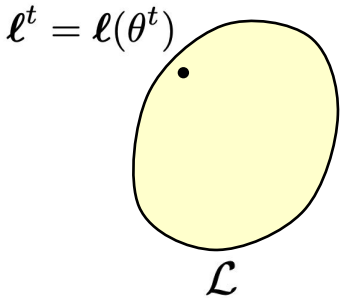
Simplified PGD Algorithm



$$M(\theta) \approx \psi(\boldsymbol{\ell}(\theta))$$

Simplified PGD Algorithm

- Estimate local gradient $\widehat{\boldsymbol{g}} \in \mathbb{R}^K$ for ψ at $\boldsymbol{\ell}(\theta^t)$

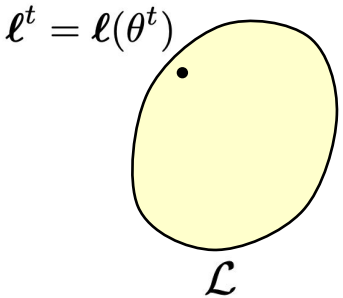


$$M(\theta) \approx \psi(\boldsymbol{\ell}(\theta))$$

Simplified PGD Algorithm

- Estimate local gradient $\widehat{\boldsymbol{g}} \in \mathbb{R}^K$ for ψ at $\boldsymbol{\ell}(\theta^t)$
  - Perturb model θ^t and compute linear fit from losses to metric

$$\begin{bmatrix} \boldsymbol{\ell}(\theta^t + \epsilon_1) \\ \boldsymbol{\ell}(\theta^t + \epsilon_2) \\ \vdots \end{bmatrix} \widehat{\boldsymbol{g}} \approx \begin{bmatrix} M(\theta^t + \epsilon_1) \\ M(\theta^t + \epsilon_2) \\ \vdots \end{bmatrix}$$



$$M(\theta) \approx \psi(\boldsymbol{\ell}(\theta))$$

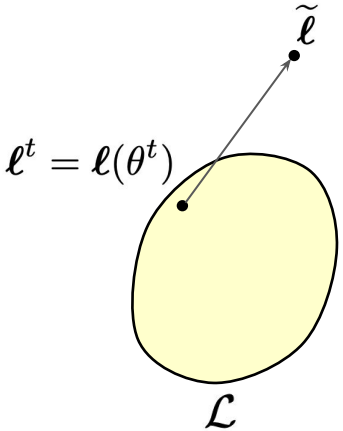
Simplified PGD Algorithm

- Estimate local gradient $\widehat{\boldsymbol{g}} \in \mathbb{R}^K$ for ψ at $\boldsymbol{\ell}(\theta^t)$
  - Perturb model θ^t and compute linear fit from losses to metric

$$\begin{bmatrix} \boldsymbol{\ell}(\theta^t + \epsilon_1) \\ \boldsymbol{\ell}(\theta^t + \epsilon_2) \\ \vdots \end{bmatrix} \widehat{\boldsymbol{g}} \approx \begin{bmatrix} M(\theta^t + \epsilon_1) \\ M(\theta^t + \epsilon_2) \\ \vdots \end{bmatrix}$$

- Gradient update on surrogate profile:

$$\widetilde{\boldsymbol{\ell}} = \boldsymbol{\ell}^t - \eta \widehat{\boldsymbol{g}}$$



$$M(\theta) \approx \psi(\boldsymbol{\ell}(\theta))$$

Simplified PGD Algorithm

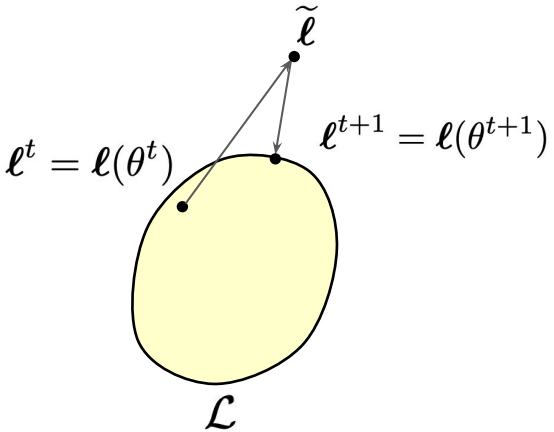
- Estimate local gradient $\widehat{\boldsymbol{g}} \in \mathbb{R}^K$ for ψ at $\boldsymbol{\ell}(\theta^t)$
  - Perturb model θ^t and compute linear fit from losses to metric

$$\begin{bmatrix} \boldsymbol{\ell}(\theta^t + \epsilon_1) \\ \boldsymbol{\ell}(\theta^t + \epsilon_2) \\ \vdots \end{bmatrix} \widehat{\boldsymbol{g}} \approx \begin{bmatrix} M(\theta^t + \epsilon_1) \\ M(\theta^t + \epsilon_2) \\ \vdots \end{bmatrix}$$

- Gradient update on surrogate profile:

$$\widetilde{\boldsymbol{\ell}} = \boldsymbol{\ell}^t - \eta \widehat{\boldsymbol{g}}$$

- Project $\widetilde{\boldsymbol{\ell}}$ to set of achievable surrogate profiles $\mathcal{L}$



$$M(\theta) \approx \psi(\boldsymbol{\ell}(\theta))$$

Simplified PGD Algorithm

- Estimate local gradient $\widehat{\boldsymbol{g}} \in \mathbb{R}^K$ for ψ at $\boldsymbol{\ell}(\theta^t)$
  - Perturb model θ^t and compute linear fit from losses to metric

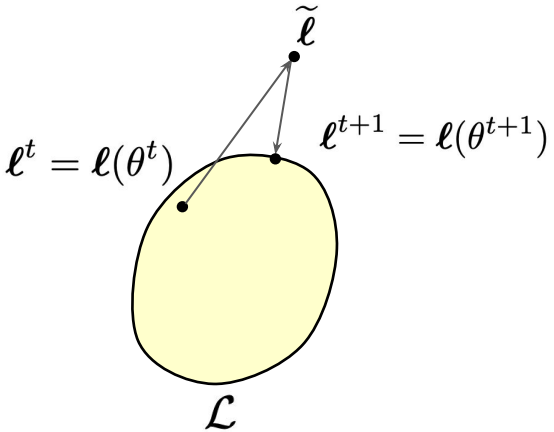
$$\begin{bmatrix} \boldsymbol{\ell}(\theta^t + \epsilon_1) \\ \boldsymbol{\ell}(\theta^t + \epsilon_2) \\ \vdots \end{bmatrix} \widehat{\boldsymbol{g}} \approx \begin{bmatrix} M(\theta^t + \epsilon_1) \\ M(\theta^t + \epsilon_2) \\ \vdots \end{bmatrix}$$

- Gradient update on surrogate profile:

$$\widetilde{\boldsymbol{\ell}} = \boldsymbol{\ell}^t - \eta \widehat{\boldsymbol{g}}$$

- Project $\widetilde{\boldsymbol{\ell}}$ to set of achievable surrogate profiles $\mathcal{L}$: solve a regression problem in θ to match target profile

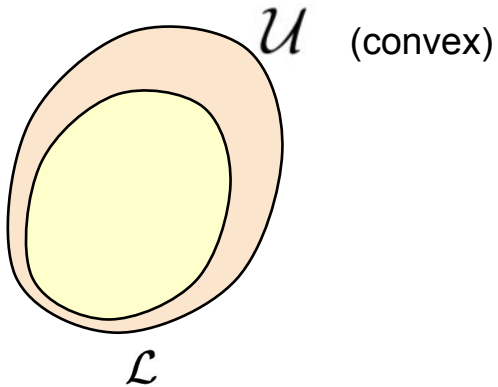
$$\theta^{t+1} = \underset{\theta \in \Theta}{\operatorname{argmin}} \|\boldsymbol{\ell}(\theta) - \tilde{\boldsymbol{\ell}}\|^2$$



$$M(\theta) \approx \psi(\boldsymbol{\ell}(\theta))$$

Convex Projection and Convergence

- Our actual algorithm works with surrogates $\ell_k(\theta)$ that are convex
- Even with convex surrogates, $\mathcal{L}$ is not necessarily a convex set
- So we optimize over a convex superset of the surrogate space $\mathcal{L}$
- We show that the projection onto this set can performed *inexactly* as a convex regression problem in θ

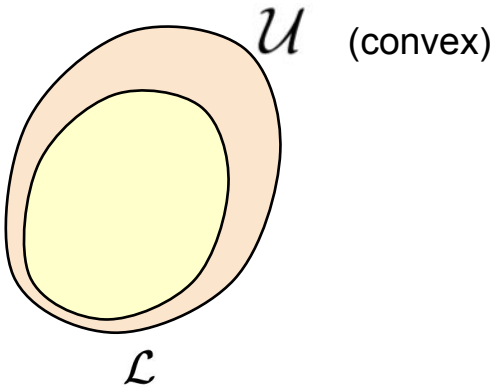


$$M(\theta) \approx \psi(\boldsymbol{\ell}(\theta))$$

Convex Projection and Convergence

- Our actual algorithm works with surrogates $\ell_k(\theta)$ that are convex
- Even with convex surrogates, $\mathcal{L}$ is not necessarily a convex set
- So we optimize over a convex superset of the surrogate space $\mathcal{L}$
- We show that the projection onto this set can performed *inexactly* as a convex regression problem in θ
- **Guarantee:** Converges to a near stationary point of the metric under smoothness/monotonicity assumptions, i.e.,

$$\min_{1 \le t \le T} \mathbb{E}[\|\nabla \psi(\ell(\theta^t))\|^2] \le \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \textit{constant}$$



Classification with Proxy Labels

- Minimize classification error with proxy labels, small validation set with true labels
- Sigmoid losses on the positive and negative examples used as surrogates

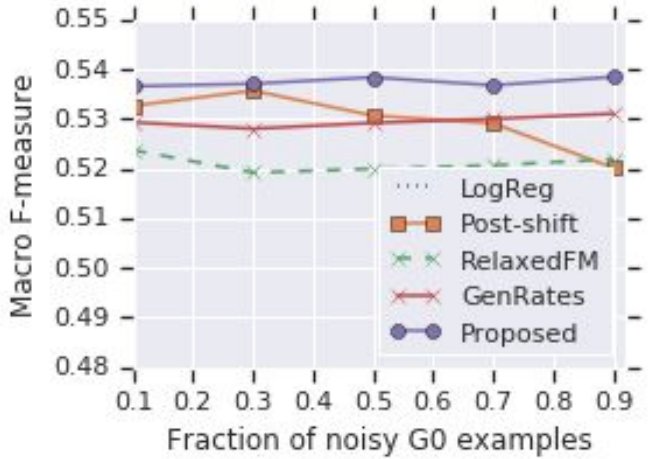
Dataset	Label	Proxy	LogReg	PostShift	Proposed
Adult	Gender	Marital Status Wife	0.333	0.322	**0.314**
Business	Same Business	Same Phone No	0.340	0.251	**0.236**

(lower values are better)

F-measure with Noisy Features

- Maximize F-measure with features from one group of examples being noisy, small validation sample with clean features
- Surrogates: **hinge loss** averaged over either the positive or negative examples, calculated separately for each of the two groups

Credit Default dataset
Predict if a customer would default
Noisy features for male customers



(higher values are better)

Ranking with PRBEP

- Maximize Precision-Recall Break-Even Point:
  - Precision at the threshold where precision and recall are equal
- Surrogates: **Precision at Recalls 0.25, 0.5, 0.75**

KDD Cup 2008 Dataset

	Kar et al. (2015)	Proposed
Train	0.473	**0.546**
Test	0.441	**0.480**

(higher values are better)

Conclusions

- Optimize a black-box metric by adaptively combining a small set of useful surrogates.
- Proposed method applies projected gradient descent over a surrogate space, and enjoys convergence guarantees.
- Experiments on classification tasks with noisy labels and features, and ranking tasks with complex metrics.