Optimizing Generalized Rate Metrics with Three Players

Harikrishna Narasimhan, Andrew Cotter, Maya Gupta
Constrained Learning Problems

\[
\min_{\theta} \mathcal{P}(\theta)
\]

s.t. \[ \mathcal{P}^j(\theta) \leq \epsilon, \quad j = 1, 2, \ldots \]

General evaluation metric

Complex policy / fairness goal
Constrained Learning Problems

\[
\min_\theta \mathcal{P}(\theta) \\
\text{s.t. } \mathcal{P}^j(\theta) \leq \epsilon, \quad j = 1, 2, \ldots
\]

Example: Fair Hiring
Constrained Learning Problems

\[
\min_\theta \mathcal{P}(\theta)
\]

s.t. \( \mathcal{P}^j(\theta) \leq \epsilon, \quad j = 1, 2, ... \)

F-measure
\[
\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

Example: Fair Hiring

G1

G2
Constrained Learning Problems

\[
\min_{\theta} \mathcal{P}(\theta)
\]

\[
\text{s.t. } \mathcal{P}^j(\theta) \leq \epsilon, \quad j = 1, 2, \ldots
\]

**Example: Fair Hiring**

- **G1**: [Images of people]
- **G2**: [Images of people]

**F-measure**

\[
\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

**Equal opportunity**

\[
|\text{Recall}(G_1) - \text{Recall}(G_2)| \leq \epsilon
\]
Constrained Learning Problems

\[
\min_{\theta} \mathcal{P}(\theta)
\]

s.t. \( \mathcal{P}^j(\theta) \leq \epsilon, \quad j = 1, 2, \ldots \)

F-measure

\[
\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

Example: Fair Hiring

Equal precision

\[
|\text{Precision}(G_1) - \text{Precision}(G_2)| \leq \epsilon
\]
Constrained Learning Problems

\[
\min_{\theta} \mathcal{P}(\theta)
\]

s.t. \[ \mathcal{P}^j(\theta) \leq \epsilon, \quad j = 1, 2, \ldots \]

Example: Fair Hiring

G1

G2

F-measure

\[
\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

G-mean

\[
\sqrt{\text{TPR} \times \text{TNR}}
\]

Equal precision

\[
|\text{Precision}(G_1) - \text{Precision}(G_2)| \leq \epsilon
\]
Constrained Learning Problems

\[
\min_{\theta} \mathcal{P}(\theta)
\]
\[
\text{s.t.} \quad \mathcal{P}^j(\theta) \leq \epsilon, \quad j = 1, 2, \ldots
\]

F-measure

\[
2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

G-mean

\[
\sqrt{\text{TPR} \times \text{TNR}}
\]

Example: Fair Hiring

Equal precision

\[
|\text{Precision}(G_1) - \text{Precision}(G_2)| \leq \epsilon
\]

Match distribution

\[
\text{KL} (p_{\text{target}} \parallel \hat{p}_{\text{pred}}) \leq \epsilon
\]
Constrained Learning Problems

\[ \min_\theta \mathcal{P}(\theta) \]

F-measure
\[
\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

How does one design learning algorithms to handle general performance metrics and constraints?

\[ |\text{Precision}(G_1) - \text{Precision}(G_2)| \leq \epsilon \]

Match Distribution
\[ \text{KL} \left( p_{\text{target}} \parallel \hat{p}_{\text{pred}} \right) \leq \epsilon \]
Problem Setup

$$\min_{\theta} \psi(R_1(\theta), \ldots, R_K(\theta))$$

s.t.  $$\psi^j(R_1(\theta), \ldots, R_K(\theta)) \leq 0, \quad j = 1, 2, \ldots$$

where there are $K$ prediction rates
Problem Setup

\[
\min_{\theta} \psi(R_1(\theta), \ldots, R_K(\theta))
\]

s.t. \( \psi^j(R_1(\theta), \ldots, R_K(\theta)) \leq 0, \quad j = 1, 2, \ldots \)

where there are \( K \) prediction rates:

\[
R_k(\theta) = \mathbb{E}_{(X,Y) \sim D_k} \left[ \mathbb{I}(Y \neq \text{sign}(f_\theta(X))) \right]
\]

for some \( f_\theta : \mathcal{X} \rightarrow \mathbb{R} \)

Expectations of counts
E.g. false positive rate, coverage
Problem Setup

\[
\min_{\theta} \psi(R_1(\theta), \ldots, R_K(\theta))
\]

s.t. \( \psi^j(R_1(\theta), \ldots, R_K(\theta)) \leq 0, \quad j = 1, 2, \ldots \)

where there are \( K \) prediction rates:

\[
R_k(\theta) = \mathbb{E}_{(X,Y) \sim D_k} \left[ \mathbb{I}(Y \neq \text{sign}(f_\theta(X))) \right]
\]

for some \( f_\theta : \mathcal{X} \to \mathbb{R} \)

Non-continuous in \( \theta \)

Non-decomposable: not simple averages of pointwise errors

Expectations of counts
E.g. false positive rate, coverage
Prior Methods

- **Solution 1: Use Surrogates†**
  - Relax indicators with continuous surrogates

†Joachims, 15; Kar et al. 14; 16; N et al. 15; Goh et al. 16; Zafar et al. 17
Prior Methods

● Solution 1: Use Surrogates‡
  ○ Relax indicators with continuous surrogates
  ○ Relaxing constraints with surrogates may make the problem infeasible
  ○ Surrogates may output values outside the range of $\psi$

\[
KL (p \parallel \hat{p})
\]

Defined for $[0, 1]$

‡Joachims, 15; Kar et al. 14; 16; N et al. 15; Goh et al. 16; Zafar et al. 17
Prior Methods

● **Solution 1: Use Surrogates**
  ○ Relax indicators with continuous surrogates
  ○ Relaxing constraints with surrogates may make the problem infeasible
  ○ Surrogates may output values outside the range of \( \psi \)

\[
\text{KL}(p \parallel \hat{p})
\]

- Defined for \([0, 1]\)

● **Solution 2: Cost-weighted Minimization Oracle**
  ○ Sequence of weighted objectives \( \sum_k \alpha_k R_k(\theta) \) and use an oracle to solve sub-problem. **Strong oracle assumption!**

*Parambath et al. 14; Koyejo et al. 14; N et al. 15; Yan et al. 18; N 18; Alabi et al. 18*
Prior Methods

- Solution 1: Use Surrogates
  - Relax indicators with continuous surrogates

This Paper

- **General framework** that recovers prior methods as special cases
- Practical algorithms with **minimal use of surrogates** and **tighter handling** of constraints

- Sequence of weighted objectives $\sum_k \alpha_k R_k(\theta)$ and use an oracle to solve sub-problem

*Parambath et al. 14; Koyejo et al. 14; N et al. 15; Yan et al. 18; N 18; Alabi et al. 18*
Min-max Game Formulation

\[
\min_{\theta} \psi(R_1(\theta), \ldots, R_K(\theta))
\]

Convex, monotonic \(\psi\)

Unconstrained problem; same ideas apply with constraints
Min-max Game Formulation

\[
\min_{\theta} \psi (R_1(\theta), \ldots, R_K(\theta))
\]

\[\equiv\]

Slack variables to decouple rates

\[
\min_{\theta, \xi} \psi (\xi_1, \ldots, \xi_K) \quad \text{s.t.} \quad \xi_k \geq R_k(\theta), \forall k
\]
Min-max Game Formulation

\[
\min_{\theta} \psi (R_1(\theta), \ldots, R_K(\theta))
\]

\[\equiv\]

\[
\min_{\theta, \xi} \psi (\xi_1, \ldots, \xi_K) \quad \text{s.t.} \quad \xi_k \geq R_k(\theta), \forall k
\]

\[\equiv\]

Lagrangian min-max problem

\[
\min_{\theta, \xi} \max_{\lambda \geq 0} \psi (\xi_1, \ldots, \xi_K) - \sum_k \lambda_k \xi_k + \sum_k \lambda_k R_k(\theta)
\]
Three-player Viewpoint

\[
\min_{\theta, \xi} \max_{\lambda \geq 0} \psi(\xi_1, \ldots, \xi_K) - \sum_k \lambda_k \xi_k + \sum_k \lambda_k R_k(\theta)
\]

- \(\lambda\)-player: Linear
Three-player Viewpoint

\[
\min_{\theta, \xi} \ \max_{\lambda \geq 0} \ \psi (\xi_1, \ldots, \xi_K) - \sum_{k} \lambda_k \xi_k + \sum_{k} \lambda_k R_k(\theta)
\]

- \(\lambda\)-player: Linear
- \(\xi\)-player: Convex
Three-player Viewpoint

\[
\min_{\theta, \xi} \max_{\lambda \geq 0} \psi (\xi_1, \ldots, \xi_K) - \sum_k \lambda_k \xi_k + \sum_k \lambda_k R_k(\theta)
\]

- \(\lambda\)-player: Linear
- \(\xi\)-player: Convex
- \(\theta\)-player: Non-continuous due to indicators
Choosing Player Strategies

<table>
<thead>
<tr>
<th>Oracle-based Alg.</th>
<th>( \xi )-player</th>
<th>( \lambda )-player</th>
<th>( \theta )-player</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>(analytical solution)</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\min_{\theta, \xi} \max_{\lambda \geq 0} \psi (\xi_1, \ldots, \xi_K) - \sum_k \lambda_k \xi_k + \sum_k \lambda_k R_k(\theta)
\end{align*}
\]
Choosing Player Strategies

<table>
<thead>
<tr>
<th>ξ-player</th>
<th>θ-player</th>
<th>θ-player</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>Best Response with Indicators</td>
</tr>
</tbody>
</table>

Oracle-based Alg. (analytical solution)

Cost-weighted minimization oracle

\[
\min_{\theta} \sum_{k} \lambda_{k}^{t} R_{k}(\theta)
\]

\[
\min_{\theta, \xi} \max_{\lambda \geq 0} \psi(\xi_1, \ldots, \xi_K) - \sum_{k} \lambda_{k} \xi_{k} + \sum_{k} \lambda_{k} R_{k}(\theta)
\]
Choosing Player Strategies

<table>
<thead>
<tr>
<th></th>
<th>(\xi)-player</th>
<th>(\lambda)-player</th>
<th>(\theta)-player</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle-based Alg.</td>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>Best Response with Indicators</td>
</tr>
<tr>
<td>Surrogate-based Alg.</td>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>SGD with Surrogates</td>
</tr>
</tbody>
</table>

**Idea:** Use original objective for \(\lambda\) and surrogates for \(\theta\)

\[
\min_{\theta, \xi} \max_{\lambda \geq 0} \psi(\xi_1, \ldots, \xi_K) - \sum_k \lambda_k \xi_k + \sum_k \lambda_k \tilde{R}_k(\theta)
\]

Extends work of Cotter et al.‘19
Players Find Equilibrium

<table>
<thead>
<tr>
<th></th>
<th>ξ-player</th>
<th>λ-player</th>
<th>θ-player</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle-based Alg.</td>
<td>Best Response</td>
<td>SGD \ with Indicators</td>
<td>Best Response \ with Indicators</td>
</tr>
<tr>
<td>Surrogate-based Alg.</td>
<td>Best Response</td>
<td>SGD \ with Indicators</td>
<td>SGD \ with Surrogates</td>
</tr>
</tbody>
</table>

- Iterative approach returns a **stochastic classifier**
Players Find Equilibrium

<table>
<thead>
<tr>
<th></th>
<th>$\xi$-player</th>
<th>$\lambda$-player</th>
<th>$\theta$-player</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Oracle-based Alg.</strong></td>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>Best Response with Indicators</td>
</tr>
<tr>
<td><strong>Surrogate-based Alg.</strong></td>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>SGD with Surrogates</td>
</tr>
</tbody>
</table>

- Iterative approach returns a stochastic classifier
  - Near-optimality & near-feasibility guarantees (convex $\psi, \psi^j$)
Players Find Equilibrium

<table>
<thead>
<tr>
<th></th>
<th>$\xi$-player</th>
<th>$\lambda$-player</th>
<th>$\theta$-player</th>
<th>Equilibrium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle-based Alg.</td>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>Best Response with Indicators</td>
<td>Mixed Nash</td>
</tr>
<tr>
<td>Surrogate-based Alg.</td>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>SGD with Surrogates</td>
<td>Mixed Coarse Correlated</td>
</tr>
</tbody>
</table>

- Iterative approach returns a **stochastic classifier**
  - Near-optimality & near-feasibility guarantees (convex $\psi$, $\psi^j$)
Players Find Equilibrium

<table>
<thead>
<tr>
<th></th>
<th>ξ-player</th>
<th>λ-player</th>
<th>θ-player</th>
<th>Equilibrium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle-based Alg.</td>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>Best Response with Indicators</td>
<td>Mixed Nash</td>
</tr>
<tr>
<td>Surrogate-based Alg.</td>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>SGD with Surrogates</td>
<td>Mixed Coarse Correlated</td>
</tr>
</tbody>
</table>

- Iterative approach returns a stochastic classifier
  - Near-optimality & near-feasibility guarantees (convex $\psi, \psi^j$)
- Surrogates: Weaker optimality guarantee with a smaller comparator class
Framework Generalizes Prior Methods

<table>
<thead>
<tr>
<th></th>
<th>(\xi)-player</th>
<th>(\lambda)-player</th>
<th>(\theta)-player</th>
<th>Equilibrium</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Oracle-based Alg.</strong></td>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>Best Response with Indicators</td>
<td>Mixed Nash</td>
</tr>
<tr>
<td><strong>Surrogate-based Alg.</strong></td>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>SGD with Surrogates</td>
<td>Mixed Coarse Correlated</td>
</tr>
<tr>
<td>SPADE [N et al. 15]</td>
<td>Best Response</td>
<td>SGD with Surrogates</td>
<td>SGD with Surrogates</td>
<td>Pure Nash</td>
</tr>
<tr>
<td>NEMSIS [Kar et al. 16]</td>
<td>-</td>
<td>FTL with Surrogates</td>
<td>SGD with Surrogates</td>
<td>Pure Nash</td>
</tr>
<tr>
<td>Frank-Wolfe [N et al. 15]</td>
<td>-</td>
<td>FTL* with Indicators</td>
<td>Best Response with Rates</td>
<td>Mixed Nash</td>
</tr>
</tbody>
</table>

*Uses result of Abernethy & Wang ‘19

*Do not handle constraints
Heuristic Algorithm for Non-convex $\psi$

<table>
<thead>
<tr>
<th></th>
<th>$\xi$-player</th>
<th>$\lambda$-player</th>
<th>$\theta$-player</th>
<th>Equilibrium</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Oracle-based Alg.</strong></td>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>Best Response with Indicators</td>
<td>Mixed Nash</td>
</tr>
<tr>
<td><strong>Surrogate-based Alg.</strong></td>
<td>Best Response</td>
<td>SGD with Indicators</td>
<td>SGD with Surrogates</td>
<td>Mixed Coarse Correlated</td>
</tr>
<tr>
<td><strong>Heuristic Alg.</strong></td>
<td>SGD</td>
<td>SGD with Indicators</td>
<td>SGD with Surrogates</td>
<td>-</td>
</tr>
</tbody>
</table>

Non-convex $\psi, \psi^j$: e.g. sums or differences of ratios
Experiment

\[
\max \quad \text{F-measure} \\
\text{s.t. } \text{F-measure}_{\text{minority}} \geq \text{F-measure}_{\text{other}} - \epsilon
\]
Experiment

\[ \max \ F\text{-measure} \]

\[ \text{s.t. } F\text{-measure}_{\text{minority}} \geq F\text{-measure}_{\text{other}} - \epsilon \]

Law School (black / others)

Unconstrained Baselines

Trades-off objective for constraints
Optimizing Generalized Rate Metrics with Three Players

Harikrishna Narasimhan, Andrew Cotter, Maya Gupta

Poster: East Exhibition Hall B + C #51

Open-source Library: TensorFlow Constrained Optimization (TFCO)
https://github.com/google-research/tensorflow_constrained_optimization