Performance-Aligned Learning Algorithms with Statistical Guarantees
“New learning algorithms that align with performance/loss metrics and provide the statistical guarantees of Fisher consistency”
Introduction and Motivation
Supervised Learning

Data Distribution $P(x, y)$

Training

$x_1 y_1$

$x_2 y_2$

$\vdots$

$x_n y_n$

Testing

$x_{n+1} \hat{y}_{n+1}$

$x_{n+2} \hat{y}_{n+2}$

$\vdots$

Data

Loss/Performance Metrics: $\text{loss}(\hat{y}, y) / \text{score}(\hat{y}, y)$

Multiclass Classification

- Zero one loss / accuracy metric
- Absolute loss (for ordinal regression)

Multivariate Performance

- F1-score
- Precision@k

Structured Prediction

- Hamming loss (sum of 0-1 loss)
Empirical Risk Minimization (ERM) (Vapnik, 1992)

- Assume a family of parametric hypothesis function \( f \) (e.g. linear discriminator)
- Find the hypothesis \( f^* \) that minimize the empirical risk:

\[
\min_f \frac{1}{n} \sum_{i=1}^{n} \text{loss}(f(x_i), y_i) = \min_f \mathbb{E}_{X,Y \sim p} \left[ \text{loss}(f(X), Y) \right]
\]

Non-convex, non-continuous metrics \( \rightarrow \) Intractable optimization

Convex surrogate loss need to be employed!

A desirable property of convex surrogates:

Fisher Consistency

Under ideal condition: optimize surrogate \( \rightarrow \) minimizes the loss metric
(given the true distribution and fully expressive model)
Two Main Approaches

1. **Probabilistic Approach**
   - Construct prediction probability model
   - Employ the logistic loss surrogate

   **Logistic Regression, Conditional Random Fields (CRF)**

2. **Large-Margin Approach**
   - Maximize the margin that separates correct prediction from the incorrect one
   - Employ the hinge loss surrogate

   **Support Vector Machine (SVM), Structured SVM**

* Pictures are taken from MLPP book (Kevin Murphy)
Multiclass Classification | Logistic Regression vs SVM

1 Multiclass Logistic Regression

- Statistical guarantee of Fisher consistency (minimizes the zero-one loss metric in the limit)
- No dual parameter sparsity

2 Multiclass SVM

- Current multiclass SVM formulations:
  - Lack Fisher consistency property, or
  - Doesn’t perform well in practice
- Computational efficiency (via the kernel trick & dual parameter sparsity)
Structured Prediction | CRF vs Structured SVM

1. Conditional Random Fields (CRF)
   - Statistical guarantee of Fisher consistency
   - No easy mechanism to incorporate customized loss/performance metrics
   - Computation of the normalization term may be intractable

2. Structured SVM
   - No Fisher consistency guarantees
   - Flexibility to incorporate customized loss/performance metrics
   - Relatively more efficient in computation
New Learning Algorithms?

✓ Align better with the loss/performance metric (by incorporating the metric into its learning objective)

✓ Provide Fisher consistency guarantee

✓ Computationally efficient

✓ Perform well in practice

How?

Robust adversarial learning approach

“What predictor best maximizes the performance metric (or minimizes the loss metric) in the worst case given the statistical summaries of the empirical distributions?”
Performance-Aligned Surrogate Losses for General Multiclass Classification

Based on:


Supervised Learning | Multiclass Classification

Data Distribution $P(x, y)$

Training:
- $x_1, y_1$
- $x_2, y_2$
- $\ldots$
- $x_n, y_n$

Testing:
- $x_{n+1}, \hat{y}_{n+1}$
- $x_{n+2}, \hat{y}_{n+2}$
- $\ldots$

Loss/Performance Metrics:
- $\text{loss}(\hat{y}, y) / \text{score}(\hat{y}, y)$

Finite set of possible value of $y$
Multiclass Classification | Zero-One Loss

Example: Digit Recognition

Loss Metric: Zero-One Loss

\[ L = \begin{bmatrix}
0 & 1 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 & 1 \\
1 & 1 & 0 & 1 & 1 \\
1 & 1 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 0
\end{bmatrix} \]

\[ \text{Loss Metric:} \quad \text{loss}(\hat{y}, y) = I(\hat{y} \neq y) \]
Example: Movie Rating Prediction

Predicted vs Actual Label:
Distance → Loss → Loss

Loss Metric: Absolute Loss

Loss Metric:
\[ \text{loss}(\hat{y}, y) = |\hat{y} - y| \]
Multiclass Classification | Classification with Abstention

Predictor can say ‘abstain’

1 2 3

Prediction

Loss Metric: Abstention Loss

\[
\text{loss}(\hat{y}, y) = \begin{cases} 
\alpha & \text{if abstain} \\
I(\hat{y} \neq y) & \text{otherwise}
\end{cases}
\]

\[
L = \begin{bmatrix}
0 & 1 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 & 1 \\
1 & 1 & 0 & 1 & 1 \\
1 & 1 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 0 \\
\alpha & \alpha & \alpha & \alpha & \alpha
\end{bmatrix}
\]
Multiclass Classification | Other Loss Metrics

**Squared loss metric**

$$\text{loss}(\hat{y}, y) = (\hat{y} - y)^2$$

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>4</th>
<th>9</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Taxonomy-based loss metric**

$$\text{loss}(\hat{y}, y) = h - v(\hat{y}, y) + 1$$

![Taxonomy Diagram]

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

**Cost-sensitive loss metric**

$$\text{loss}(\hat{y}, y) = C_{\hat{y}, y}$$

<table>
<thead>
<tr>
<th>0</th>
<th>3</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
Robust Adversarial Learning
Robust Adversarial Learning (Grunwald & Dawid, 2004; Delage & Ye, 2010; Asif et.al, 2015)

Original Loss Metric
Non-convex, non-continuous
\[
\min_f \mathbb{E}_{X,Y \sim \tilde{P}} [\text{loss}(f(X), Y)]
\]

Approximate the loss with convex surrogates
\[
\min_{\hat{P}(\hat{Y}|X)} \mathbb{E}_{X,Y \sim \hat{P},\hat{Y}|X \sim \hat{P}} [\text{loss}(\hat{Y}, Y)]
\]

Probabilistic prediction
\[
\min_{\hat{P}(\hat{Y}|X)} \mathbb{E}_{X,Y \sim \hat{P},\hat{Y}|X \sim \hat{P}} [\text{loss}(\hat{Y}, Y)]
\]

Evaluate against an adversary, instead of using empirical data

Adversary's probabilistic prediction
\[
\min_{\hat{P}(\hat{Y}|X)} \max_{\tilde{P}(\tilde{Y}|X)} \mathbb{E}_{X,Y \sim \tilde{P},\hat{Y}|X \sim \hat{P},\tilde{Y}|X \sim \tilde{P}} [\text{loss}(\hat{Y}, \tilde{Y})]
\]

s.t. \[
\mathbb{E}_{X \sim \hat{P},\tilde{Y}|X \sim \hat{P}}[\phi(X, \tilde{Y})] = \mathbb{E}_{X,Y \sim \tilde{P}}[\phi(X, Y)]
\]

Constraint the statistics of the adversary's distribution to match the empirical statistics
Robust Adversarial Dual Formulation

Primal:
\[
\min_{\hat{P}(Y|X)} \max_{\hat{P}(Y|X)} \mathbb{E}_{X \sim \hat{P}; Y \sim \hat{P}} \left[ \text{loss}(\hat{Y}, \hat{Y}) \right]
\]
subject to:
\[
\mathbb{E}_{X \sim \hat{P}; Y \sim \hat{P}}[\phi(X, \hat{Y})] = \mathbb{E}_{X,Y \sim \hat{P}}[\phi(X, Y)]
\]
Lagrange multiplier, minimax duality

Dual:
\[
\min_{\theta} \mathbb{E}_{X,Y \sim \hat{P}} \max_{\hat{P}(Y|X)} \min_{\hat{P}(Y|X)} \mathbb{E}_{Y \sim \hat{P}; X \sim \hat{P}} \left[ \text{loss}(\hat{Y}, \hat{Y}) + \theta^T (\phi(X, \hat{Y}) - \phi(X, Y)) \right]
\]
ERM with the adversarial surrogate loss (AL):
\[
AL(x, y, \theta) = \max_{\hat{P}(Y|x)} \min_{\hat{P}(Y|x)} \mathbb{E}_{Y \sim \hat{P}; X \sim \hat{P}} \left[ \text{loss}(\hat{Y}, \hat{Y}) + \theta^T (\phi(x, \hat{Y}) - \phi(x, y)) \right]
\]
Convex in \( \theta \)

Simplified notation
\[
AL(f, y) = \max_{q \in \Delta} \min_{p \in \Delta} p^T L q + f^T q - f_y
\]
where:
\[
\begin{align*}
p_i &= \hat{P}(\hat{Y} = i|x) \\
q_i &= \hat{P}(\hat{Y} = i|x) \\
f_i &= \theta^T \phi(x, i)
\end{align*}
\]
Adversarial Surrogate Loss

\[ AL(f, y) = \max_{q \in \Delta} \min_{p \in \Delta} p^T L q + f^T q - f_y \]

Convert to a Linear Program

\[ AL(f, y) = \max_{q, v} v + f^T q - f_y \]

\[ \text{s.t.: } L_{(i,:)} q \geq v \quad \forall i \in [k] \]
\[ q_i \geq 0 \quad \forall i \in [k] \]
\[ q^T 1 = 1 \]

Convex Polytope formed by the constraints

\[ C = \left\{ \begin{bmatrix} q \\ v \end{bmatrix} \right| \begin{bmatrix} q \\ v \end{bmatrix} \geq b, \text{ where } A = \begin{bmatrix} L & -I \\ I & 0 \\ 1^T & 0 \\ -1^T & 0 \end{bmatrix}, \ b = \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix} \right\} \]

Example for a four class classification

\[
\begin{bmatrix}
0 & 1 & 1 & 1 & -1 \\
1 & 0 & 1 & 1 & -1 \\
1 & 1 & 0 & 1 & -1 \\
1 & 1 & 1 & 0 & -1 \\
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
1 & -1 & -1 & -1 & 0
\end{bmatrix}
\]

\[ \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ v \end{bmatrix} \geq \begin{bmatrix} 1 \end{bmatrix} \]

LP Solver \( O(k^{3.5}) \)

Extreme points of the (bounded) polytope

There is always an optimal solution that is an extreme point of the domain.

Computing \( AL = \) finding the best extreme point
Zero-One Loss : $\text{AL}^{0-1}$ | Convex Polytope

Convex Polytope of the $\text{AL}^{0-1}$

$$C = \left\{ \frac{q}{v} \Bigg| A \frac{q}{v} \geq b, \text{ where } A = \begin{bmatrix} L & -1 \\ 1 & 0 \\ 1^T & 0 \\ -1^T & 0 \end{bmatrix}, \ b = \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix} \right\}$$

$$L = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

Extreme points of the polytope

$$D = \left\{ \frac{q}{v} = \frac{1}{|S|} \left[ \sum_{i \in S} e_i \right] \Bigg| \emptyset \neq S \subseteq [k] \right\}$$

$e_i$ is a vector with a single 1 at the $i$-th index, and 0 elsewhere.

$$[k] \triangleq \{1, \ldots, k\} \quad e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The Adversarial Surrogate Loss for Zero-One Loss Metrics ($\text{AL}^{0-1}$)

$$\text{AL}^{0-1}(f, y) = \max_{S \subseteq [k], S \neq \emptyset} \frac{\sum_{i \in S} f_i + |S| - 1}{|S|} - f_y$$

Computation of $\text{AL}^{0-1}$

- Sort $f_i$ in non-increasing order
- Incrementally add potentials to the set $S$, until adding more potential decrease the loss value

$O(k \log k)$, where $k$ is the number of classes
AL$^{0-1}$ | Loss Surface

**Binary Classification**

- Plots over the space of potential differences $\psi_i = f_i - f_y$
- The true label is $y = 1$

**Three Class Classification**
Other Multiclass Loss Metrics

Ordinal Regression with Absolute Loss Metric

Extreme points of the polytope:

\[ D = \left\{ \begin{bmatrix} q_n \\ v \end{bmatrix} = \frac{1}{2} \begin{bmatrix} e_i + e_j \\ j - i \end{bmatrix} \mid i, j \in [k]; i \leq j \right\} \]

\( e_i \) is a vector with a single 1 at the \( i \)-th index, and 0 elsewhere.

Adversarial Surrogate Loss \( AL^{ord} \):

\[
AL^{ord}(f, y) = \max_{i,j \in [k]} \frac{f_i + f_j + j - i}{2} - f_y
\]

Computation cost:
\( O(k) \), where \( k \) is the number of classes
Other Multiclass Loss Metrics

Classification with Abstention \((0 \leq \alpha \leq 0.5)\)

Extreme points of the polytope:

\[
D = \left\{ \begin{bmatrix} q_v \\ v \end{bmatrix} = (1 - \alpha) \begin{bmatrix} e_i \\ 0 \end{bmatrix} + \alpha \begin{bmatrix} e_j \\ 1 \end{bmatrix} \left| i, j \in [k] \right\} \cup \left\{ \begin{bmatrix} q_v \\ v \end{bmatrix} = \begin{bmatrix} e_i \\ 0 \end{bmatrix} \left| i \in [k] \right\} \right. 
\]

\(e_i\) is a vector with a single 1 at the \(i\)-th index, and 0 elsewhere.

Adversarial Surrogate Loss \(AL^{\text{abstain}}\):

\[
AL^{\text{abstain}}(f, y, \alpha) = \max \left\{ \max_{i,j\in[k], i \neq j} \left( \frac{1}{k^2} \right) f_i + \alpha f_j + \alpha, \max_i f_i \right\} - f_y
\]

Computation cost:
\(O(k)\), where \(k\) is the number of classes
Fisher Consistency

Fisher Consistency Requirement in Multiclass Classification

\[ f^* \in \mathcal{F}^* \triangleq \arg\min_f \mathbb{E}_{Y|X \sim P} [AL_f(x, Y)] \]

\[ \Rightarrow \arg\max_y f^*(x, y) \subseteq \mathcal{Y}^\circ \triangleq \arg\min_{y'} \mathbb{E}_{Y|X \sim P} \{\text{loss}(y', Y)\} \]

- \( P(Y|x) \) is the true conditional distribution
- \( f \) is optimized over all measurable functions

Bayes risk minimizer

Minimizer Property

\[ f^* \in \arg\min_{f} \max_{q \in \Delta} \min_{p \in \Delta} \{f^Tq + p^T Lq - d^T f\} = \arg\min_{f} \max_{q \in \Delta} \left\{f^T q + \min_{y} (Lq)_y - d^T f\right\} \]

- \( d \) is the true conditional distribution
- \( y^\circ \) is the Bayes optimal predictor

Under \( f^* \):

\[ f^* + L(y^\circ, :)^T \text{ is a uniform vector} \]

Consistency

\[ f^* + L(y^\circ, :)^T \text{ is a uniform vector} \]

\[ \Rightarrow \arg\max_y f^*(x, y) = \arg\min_y L(y^\circ, y) \]

Fisher consistent
Optimization

Sub-gradient descent

\[ Q^* = \arg\max_{q \in \Delta} \min_{p \in \Delta} \left\{ p^T L q + \theta^T \left[ \sum_j q_j \phi(x, j) - \phi(x, y) \right] \right\} \]

\[ \partial_{\theta} AL(x, y, \theta) = \text{conv} \left\{ \sum_j q_j^* \phi(x, j) - \phi(x, y) \mid q \in Q^* \right\} \]

Example: \( AL^{0-1} \)

\[ \partial_{\theta} AL^{0-1}(x, y, \theta) \geq \frac{1}{|S^*|} \sum_{j \in S^*} \phi(x, j) - \phi(x, y) \]

\( S^* \) is the set that maximize \( AL^{0-1} \)

Incorporate Rich Feature Spaces via the Kernel Trick

input space \( x_i \) ----> rich feature space \( \omega(x_i) \)

Compute the dot products

\[ K(x_i, x_j) = \omega(x_i) \cdot \omega(x_j) \]

1. Dual Optimization (benefit: dual parameter sparsity)
2. Primal Optimization (via PEGASOS (Shalev-Shwartz, 2010))
Experiments:
Example: Multiclass Classification (0-1 loss)
**Multiclass Support Vector Machine (SVM)**

1. **The WW Model** (Weston et al., 2002)
   \[
   \text{loss}_{\text{WW}}(\mathbf{x}_i, y_i) = \sum_{j \neq y_i} [1 - (f_{y_i}(\mathbf{x}_i) - f_j(\mathbf{x}_i))]_+ 
   \]
   Relative Margin Model

2. **The CS Model** (Crammer and Singer, 1999)
   \[
   \text{loss}_{\text{CS}}(\mathbf{x}_i, y_i) = \max_{j \neq y_i} [1 - (f_{y_i}(\mathbf{x}_i) - f_j(\mathbf{x}_i))]_+ 
   \]
   Relative Margin Model

3. **The LLW Model** (Lee et al., 2004)
   \[
   \text{loss}_{\text{LLW}}(\mathbf{x}_i, y_i) = \sum_{j \neq y_i} [1 + f_j(\mathbf{x}_i)]_+ 
   \]
   Absolute Margin Model

**Fisher Consistent?**
- (Tewari and Bartlett, 2007) (Liu, 2007)

**Perform well in low dimensional feature?**
- (Dogan et al., 2016)
## AL$^{0-1}$ | Experiments

### Dataset properties and AL$^{0-1}$ constraints

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Properties</th>
<th>SVM constraints</th>
<th>AL$^{0-1}$ constraints added and active</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Linear kernel</td>
</tr>
<tr>
<td>(1) iris</td>
<td>#class 3</td>
<td>#train 105</td>
<td># test 45</td>
</tr>
<tr>
<td>(2) glass</td>
<td>#class 6</td>
<td>#train 149</td>
<td># test 65</td>
</tr>
<tr>
<td>(3) redwine</td>
<td>#class 10</td>
<td>#train 1119</td>
<td># test 480</td>
</tr>
<tr>
<td>(4) ecoli</td>
<td>#class 8</td>
<td>#train 235</td>
<td># test 101</td>
</tr>
<tr>
<td>(5) vehicle</td>
<td>#class 4</td>
<td>#train 592</td>
<td># test 254</td>
</tr>
<tr>
<td>(6) segment</td>
<td>#class 7</td>
<td>#train 1617</td>
<td># test 693</td>
</tr>
<tr>
<td>(7) sat</td>
<td>#class 7</td>
<td>#train 4435</td>
<td># test 2000</td>
</tr>
<tr>
<td>(8) optdigits</td>
<td>#class 10</td>
<td>#train 3823</td>
<td># test 1797</td>
</tr>
<tr>
<td>(9) pageblocks</td>
<td>#class 5</td>
<td>#train 3831</td>
<td># test 1642</td>
</tr>
<tr>
<td>(10) libras</td>
<td>#class 15</td>
<td>#train 252</td>
<td># test 108</td>
</tr>
<tr>
<td>(11) vertebral</td>
<td>#class 3</td>
<td>#train 217</td>
<td># test 93</td>
</tr>
<tr>
<td>(12) breasttissue</td>
<td>#class 6</td>
<td>#train 74</td>
<td># test 32</td>
</tr>
</tbody>
</table>

12 datasets  dual parameter sparsity
# Results for Linear Kernel and Gaussian Kernel

The mean (standard deviation) of the accuracy. Bold numbers: best or not significantly worse than the best.

## Linear Kernel

<table>
<thead>
<tr>
<th>D</th>
<th>AL⁰⁻¹</th>
<th>Linear Kernel</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AL⁰⁻¹</td>
<td>WW</td>
<td>CS</td>
<td>LLW</td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>96.3 (3.1)</td>
<td>96.0 (2.6)</td>
<td>96.3 (2.4)</td>
<td>79.7 (5.5)</td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td>62.5 (6.0)</td>
<td>62.2 (3.6)</td>
<td>62.5 (3.9)</td>
<td>52.8 (4.6)</td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td>58.8 (2.0)</td>
<td>59.1 (1.9)</td>
<td>56.6 (2.0)</td>
<td>57.7 (1.7)</td>
<td></td>
</tr>
<tr>
<td>(4)</td>
<td>86.2 (2.2)</td>
<td>85.7 (2.5)</td>
<td>85.8 (2.3)</td>
<td>74.1 (3.3)</td>
<td></td>
</tr>
<tr>
<td>(5)</td>
<td>78.8 (2.2)</td>
<td>78.8 (1.7)</td>
<td>78.4 (2.3)</td>
<td>69.8 (3.7)</td>
<td></td>
</tr>
<tr>
<td>(6)</td>
<td>94.9 (0.7)</td>
<td>94.9 (0.8)</td>
<td>95.2 (0.8)</td>
<td>75.8 (1.5)</td>
<td></td>
</tr>
<tr>
<td>(7)</td>
<td>84.9 (0.7)</td>
<td>85.4 (0.7)</td>
<td>84.7 (0.7)</td>
<td>74.9 (0.9)</td>
<td></td>
</tr>
<tr>
<td>(8)</td>
<td>96.6 (0.6)</td>
<td>96.5 (0.7)</td>
<td>96.3 (0.6)</td>
<td>76.2 (2.2)</td>
<td></td>
</tr>
<tr>
<td>(9)</td>
<td>96.0 (0.5)</td>
<td>96.1 (0.5)</td>
<td>96.3 (0.5)</td>
<td>92.5 (0.8)</td>
<td></td>
</tr>
<tr>
<td>(10)</td>
<td>74.1 (3.3)</td>
<td>72.0 (3.8)</td>
<td>71.3 (4.3)</td>
<td>34.0 (6.4)</td>
<td></td>
</tr>
<tr>
<td>(11)</td>
<td>85.5 (2.9)</td>
<td>85.9 (2.7)</td>
<td>85.4 (3.3)</td>
<td>79.8 (5.6)</td>
<td></td>
</tr>
<tr>
<td>(12)</td>
<td>64.4 (7.1)</td>
<td>59.7 (7.8)</td>
<td>66.3 (6.9)</td>
<td>58.3 (8.1)</td>
<td></td>
</tr>
</tbody>
</table>

**Average**

|   | 81.59 | 81.02 | 81.25 | 68.80 |

**#b**

|   | 9 | 6 | 8 | 0 |

**Linear Kernel**

AL⁰¹: slight benefit

LLW: poor perf.

## Gaussian Kernel

<table>
<thead>
<tr>
<th>D</th>
<th>AL⁰⁻¹</th>
<th>Gaussian Kernel</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AL⁰⁻¹</td>
<td>WW</td>
<td>CS</td>
<td>LLW</td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>96.7 (2.4)</td>
<td>96.4 (2.4)</td>
<td>96.2 (2.3)</td>
<td>95.4 (2.1)</td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td>69.5 (4.2)</td>
<td>66.8 (4.3)</td>
<td>69.4 (4.8)</td>
<td>69.2 (4.4)</td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td>63.3 (1.8)</td>
<td>64.2 (2.0)</td>
<td>64.2 (1.9)</td>
<td>64.7 (2.1)</td>
<td></td>
</tr>
<tr>
<td>(4)</td>
<td>86.0 (2.7)</td>
<td>84.9 (2.4)</td>
<td>85.6 (2.4)</td>
<td>86.0 (2.5)</td>
<td></td>
</tr>
<tr>
<td>(5)</td>
<td>84.3 (2.5)</td>
<td>84.4 (2.6)</td>
<td>83.8 (2.3)</td>
<td>84.4 (2.6)</td>
<td></td>
</tr>
<tr>
<td>(6)</td>
<td>96.5 (0.6)</td>
<td>96.6 (0.5)</td>
<td>96.3 (0.6)</td>
<td>96.4 (0.5)</td>
<td></td>
</tr>
<tr>
<td>(7)</td>
<td>91.9 (0.5)</td>
<td>92.0 (0.6)</td>
<td>91.9 (0.5)</td>
<td>91.9 (0.4)</td>
<td></td>
</tr>
<tr>
<td>(8)</td>
<td>98.7 (0.4)</td>
<td>98.8 (0.4)</td>
<td>98.8 (0.3)</td>
<td>98.9 (0.3)</td>
<td></td>
</tr>
<tr>
<td>(9)</td>
<td>96.8 (0.5)</td>
<td>96.6 (0.4)</td>
<td>96.7 (0.4)</td>
<td>96.6 (0.4)</td>
<td></td>
</tr>
<tr>
<td>(10)</td>
<td>83.6 (3.8)</td>
<td>83.8 (3.4)</td>
<td>85.0 (3.9)</td>
<td>83.2 (4.2)</td>
<td></td>
</tr>
<tr>
<td>(11)</td>
<td>86.0 (3.1)</td>
<td>85.3 (2.9)</td>
<td>85.5 (3.3)</td>
<td>84.4 (2.7)</td>
<td></td>
</tr>
<tr>
<td>(12)</td>
<td>68.4 (8.6)</td>
<td>68.1 (6.5)</td>
<td>66.6 (8.9)</td>
<td>68.0 (7.2)</td>
<td></td>
</tr>
</tbody>
</table>

**Average**

|   | 85.14 | 84.82 | 85.00 | 84.93 |

**#b**

|   | 9 | 6 | 6 | 7 |
## Multiclass Zero-One Classification

<table>
<thead>
<tr>
<th>Model</th>
<th>Fisher Consistent?</th>
<th>Perform well in low dimensional feature?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The SVM WW Model (Weston et.al., 2002)</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Relative Margin Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. The SVM CS Model (Crammer and Singer, 1999)</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Relative Margin Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. The SVM LLW Model (Lee et.al., 2004)</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>Absolute Margin Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. The $A_{0-1}$ (Adversarial Surrogate Loss)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Relative Margin Model</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
General Multiclass Classification

1. Zero-One Loss Metric

2. Ordinal Classification with the Absolute Loss Metric

3. Ordinal Classification with the Squared Loss Metric

4. Weighted Multiclass Loss Metrics

5. Classification with Abstention / Reject Option
Performance-Aligned Graphical Models

Based on:

Conditional Graphical Models

Some Popular Graphical Structure in Structured Prediction

Chain Structure

Activity Prediction, Sequence Tagging, NLP tasks: e.g. Named Entity Recognition

Tree Structure

Parse Tree-Based NLP tasks: Semantic Role Labeling and Sentiment Analysis

Lattice Structure

Computer Vision Tasks: e.g. Image Segmentation
Previous Approaches for Conditional Graphical Models

### Conditional Random Fields (CRF)
(Lafferty et. al., 2001)

- **Fisher Consistent**
  Produce Bayes optimal prediction in ideal case.

- **No easy mechanism to incorporate customized loss/performance metrics**
  The algorithm optimized the conditional likelihood. Loss/performance metric-based prediction can be performed after learning process.

### Structured SVM (SSVM)
(Tsochantaridis et. al., 2005)

- **No Fisher consistency guarantee**
  Based on Multiclass SVM-CS. Not consistent for distribution with no majority label.

- **Align with the loss/performance metrics**
  The algorithm accept customized loss/performance metric in its optimization objective.
Adversarial Graphical Models (AGM)

Primal:

\[
\min_{\hat{P}(\hat{Y}|X)} \max_{\tilde{P}(\tilde{Y}|X)} \mathbb{E}_{X \sim \hat{P}; Y \sim \tilde{P} ; X \sim \hat{P}} \left[ \text{loss}(\hat{Y}, \tilde{Y}) \right] \quad \text{s.t.:} \quad \mathbb{E}_{X \sim \hat{P}; Y \sim \tilde{P}} \left[ \Phi(X, \tilde{Y}) \right] = \tilde{\Phi}
\]

- Feature function \( \Phi(X, Y) \) is additively decomposed over cliques, \( \Phi(x, y) = \sum_c \phi(x, y_c) \)
- The loss metric is additively decomposed over each \( y_i \) variables, \( \text{loss}(\hat{y}, \tilde{y}) = \sum_{i=1}^{n} \text{loss}(\hat{y}_i, \tilde{y}_i) \)
- Focus on pairwise graphical models: interactions between label = edges in graphs

Dual:

\[
\min_{\theta_e, \theta_v} \mathbb{E}_{X, Y \sim \hat{P}} \max_{\tilde{P}(\tilde{Y}|X)} \min_{\hat{P}(\hat{Y}|X)} \sum_{\hat{y}, \tilde{y}} \hat{P}(\hat{y}|x) \tilde{P}(\tilde{y}|x) \left[ \sum_{i=1}^{n} \text{loss}(\hat{y}_i, \tilde{y}_i) \right]
\]

\[
+ \theta_e \cdot \sum_{(i,j) \in E} \left[ \phi(x, \hat{y}_i, \hat{y}_j) - \phi(x, y_i, y_j) \right] + \theta_v \cdot \sum_{i=1}^{n} \left[ \phi(x, \hat{y}_i) - \phi(x, y_i) \right]
\]

\( \theta_e \): Lagrange multipliers for constraints with edge features
\( \theta_v \): Lagrange multipliers for constraints with node features

\textbf{size: } \( k^n \times k^n \) 

\textbf{Intractable} 

for modestly-sized \( n \)
AGM | Marginal Formulation

Dual:

\[
\min_{\theta_e, \theta_v} \mathbb{E}_{\mathbf{X}, \mathbf{Y} \sim \hat{P}} \max_{\hat{P}(\hat{\mathbf{y}}|\mathbf{x}) \hat{P}(\hat{\mathbf{y}}|\mathbf{x})} \min_{\hat{\mathbf{y}}, \hat{\mathbf{y}}} \hat{P}(\hat{\mathbf{y}}|\mathbf{x}) \hat{P}(\hat{\mathbf{y}}|\mathbf{x}) \left[ \sum_{i}^{n} \text{loss}(\hat{y}_i, \hat{y}_i) + \theta_e \cdot \sum_{(i, j) \in E} [\phi(\mathbf{x}, \hat{y}_i, \hat{y}_j) - \phi(\mathbf{x}, y_i, y_j)] + \theta_v \cdot \sum_{i}^{n} [\phi(\mathbf{x}, \hat{y}_i) - \phi(\mathbf{x}, y_i)] \right]
\]

Dual | Marginal Formulation:

\[
\min_{\theta_e, \theta_v} \mathbb{E}_{\mathbf{X}, \mathbf{Y} \sim \hat{P}} \max_{\hat{P}(\hat{\mathbf{y}}|\mathbf{x}) \hat{P}(\hat{\mathbf{y}}|\mathbf{x})} \min_{\hat{\mathbf{y}}, \hat{\mathbf{y}}} \hat{P}(\hat{\mathbf{y}}|\mathbf{x}) \hat{P}(\hat{\mathbf{y}}|\mathbf{x}) \left[ \sum_{i}^{n} \sum_{\hat{y}_i} \hat{\mathbf{y}}_i \hat{P}(\hat{\mathbf{y}}_i|\mathbf{x}) \hat{P}(\hat{\mathbf{y}}_i|\mathbf{x}) \text{loss}(\hat{y}_i, \hat{y}_i) + \sum_{(i, j) \in E} \sum_{\hat{y}_i, \hat{y}_j} \hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j \hat{P}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j|\mathbf{x}) [\theta_e \cdot \phi(\mathbf{x}, \hat{y}_i, \hat{y}_j) - \sum_{(i, j) \in E} \theta_e \cdot \phi(\mathbf{x}, y_i, y_j)] + \sum_{i}^{n} \theta_v \cdot \phi(\mathbf{x}, \hat{y}_i)]
\]

The objective depends on \( \hat{P}(\hat{\mathbf{y}}|\mathbf{x}) \) only through its node marginal probability \( \hat{P}(\hat{y}_i|\mathbf{x}) \)

The objective depends on \( \hat{P}(\hat{\mathbf{y}}|\mathbf{x}) \) only through its node and edge marginal probability \( \hat{P}(\hat{y}_i|\mathbf{x}) \) and \( \hat{P}(\hat{y}_i, \hat{y}_j|\mathbf{x}) \)

Similar to CRF and SSVM:
General Graphical Models: Intractable
Focus:
Graphs with low tree-width, e.g.: chain, tree, simple loops.
Tractable optimization
AGM | Optimization

Matrix Notation (Tree Structure AGM):

\[
\min_{\theta_e, \theta_v} \mathbb{E}_{X,Y \sim \widetilde{P}} \max_Q \min_P \sum_i^n \left[ p_i L_i (Q_{pt(i);i}^T 1) + \left\langle Q_{pt(i);i} - Z_{pt(i);i}, \sum_l \theta_e^{(l)} W_{pt(i);i;l} \right\rangle \\
+ (Q_{pt(i);i}^T 1 - z_i^T (\sum_l \theta_v^{(l)} w_{i;l})) \right]
\]

subject to: \(Q_{pt(pt(i);pt(i))}^T 1 = Q_{pt(i);i} 1, \forall i \in \{1, \ldots, n\}\).

Optimization Techniques:

- Stochastic (sub)-gradient descent  
  (outer optimization for \(\theta_e\) and \(\theta_v\))
- Dual decomposition (inner \(Q\) optimization)
- Discrete optimal transport solver (recovering \(Q\))
- Closed-form solution (inner \(p\) optimization)

Runtime (for a single subgradient update):

- Depends on the loss metric used  
- For the additive zero-one loss (Hamming loss)
  \(O(nl k \log k + nk^2)\)  
  \(k\): # classes, \(n\): # nodes,  
  \(l\): # iterations in dual decomposition

CRF \(O(nk^2)\)  
SSVM \(O(nk^2)\)

General graphs low tree-width

\(O(n l w k^{(w+1)} \log k + nk^{2(w+1)})\)  
\(n\): # cliques, \(w\): treewidth of the graph
AGM | Consistency

If the loss function is additive

AGM is consistent
  when $f$ is optimized over all measurable functions on the input space

AGM is also consistent
  when $f$ is optimized over a restricted set of functions:
  all measurable function that are additive over the edge and node potentials.
Facial Emotion Intensity Prediction (Chain Structure, Labels with Ordinal Category)

- Each node: 3 class classification: \textit{neutral} = 1 < \textit{increasing} = 2 < \textit{apex} = 3
- 167 sequences
- Ordinal loss metrics: zero-one loss, absolute loss, and squared loss
- Weighted and unweighted. Weights reflect the focus of prediction (e.g. focus more on latest nodes)

Results: The mean (standard deviation) of the average loss metrics. Bold numbers: best or not significantly worse than the best

<table>
<thead>
<tr>
<th>Loss metrics</th>
<th>AGM</th>
<th>CRF</th>
<th>SSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>zero-one, unweighted</td>
<td>0.34</td>
<td>0.32</td>
<td>0.37</td>
</tr>
<tr>
<td>absolute, unweighted</td>
<td>0.33</td>
<td>0.34</td>
<td>0.40</td>
</tr>
<tr>
<td>quadratic, unweighted</td>
<td>0.38</td>
<td>0.38</td>
<td>0.40</td>
</tr>
<tr>
<td>zero-one, weighted</td>
<td>0.28</td>
<td>0.32</td>
<td>0.29</td>
</tr>
<tr>
<td>absolute, weighted</td>
<td>0.29</td>
<td>0.36</td>
<td>0.29</td>
</tr>
<tr>
<td>quadratic, weighted</td>
<td>0.36</td>
<td>0.40</td>
<td>0.33</td>
</tr>
<tr>
<td>average</td>
<td>0.33</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td># bold</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Semantic Role Labeling (Tree Structure)
- Predict label of each node given known parse tree.
- CoNLL 2005 dataset
- Cost-sensitive loss metric is used reflect the importance of each label

Results:

Table 2: The average loss metrics for the semantic role labeling task.

<table>
<thead>
<tr>
<th>Loss metrics</th>
<th>AGM</th>
<th>CRF</th>
<th>SSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>cost-sensitive loss</td>
<td>0.14</td>
<td>0.19</td>
<td>0.14</td>
</tr>
</tbody>
</table>
# Conditional Graphical Models

<table>
<thead>
<tr>
<th>Conditional Random Field (CRF)</th>
<th>Performance-Aligned?</th>
<th>Consistent?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Lafferty et. al., 2001)</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structured SVM</th>
<th>Performance-Aligned?</th>
<th>Consistent?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Tsochantaridis et. al., 2005)</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Adversarial Graphical Models</th>
<th>Performance-Aligned?</th>
<th>Consistent?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(our approach)</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Bipartite Matching in Graphs

Based on:
Bipartite Matching Task

Maximum weighted bipartite matching:

$$\max_{\pi \in \Pi} \psi(\pi) = \max_{\pi \in \Pi} \sum_i \psi_i(\pi_i)$$

Machine learning task:
Learn the appropriate weights $\psi_i(\cdot)$

Objective:
Minimize a loss metric, e.g., the Hamming loss

$$\text{loss}_{\text{Ham}}(\pi, \pi') = \sum_{i=1}^{n} 1(\pi'_i \neq \pi_i)$$
Learning Bipartite Matching | Applications

1. Word alignment
   (Taskar et. al., 2005; Pado & Lapta, 2006; Mac-Cartney et. al., 2008)

   natürlich ist das haus klein
   
   of course the house is small

2. Correspondence between images
   (Belongie et. al., 2002; Dellaert et. al., 2003)

3. Learning to rank documents
   (Dwork et. al., 2001; Le & Smola, 2007)

A non-bipartite matching task can be converted to a bipartite matching problem
Previous Approaches for Bipartite Matching

1 CRF (Petterson et. al., 2009; Volkovs & Zemel, 2012)

\[ P_\psi(\pi) = \frac{1}{Z_\psi} \exp \left( \sum_{i=1}^{n} \psi_i(\pi_i) \right) \]
\[ Z_\psi = \sum_\pi \prod_{i=1}^{n} \exp(\psi_i(\pi_i)) = \text{perm}(M) \]
where \( M_{i,j} = \exp(\psi_i(j)) \)

- Fisher Consistent
  - Produce Bayes optimal prediction in ideal case
- Computationally intractable
  - Normalization term requires matrix permanent computation (a \#P-hard problem).
  - Approximation is needed for modestly sized problems.

2 Structured SVM (Tsochantaridis et. al., 2005)

\[ \min_{\psi} \mathbb{E}_{\pi \sim \tilde{P}} \left[ \max_{\pi'} \left\{ \text{loss}(\pi, \pi') + \psi(\pi') \right\} - \psi(\pi) \right] \]
\[ \tilde{P} \text{ is the empirical distribution} \]

- Computationally Efficient
  - Hungarian algorithm for computing the maximum violated constraints
- No Fisher consistency guarantee
  - Based on Multiclass SVM-CS
  - Not consistent for distribution with no majority label
Adversarial Bipartite Matching (our approach)

Primal:

$$\min_{\hat{P}} \max_{P(\pi|x)} \mathbb{E}_{x \sim \hat{P}, \pi|x \sim \hat{P}} \left[ \text{loss}(\hat{\pi}, \pi) \right]$$

s.t. $$\mathbb{E}_{x \sim \hat{P}, \pi|x \sim \hat{P}} \left[ \sum_{i=1}^{n} \phi_i(x, \pi_i) \right] = \mathbb{E}_{(x, \pi) \sim \hat{P}} \left[ \sum_{i=1}^{n} \phi_i(x, \pi_i) \right]$$

Dual:

$$\min_{\theta} \mathbb{E}_{x, \pi \sim \hat{P}} \min_{\bar{P}(\pi|x)} \max_{\bar{P}(\pi|x)} \mathbb{E}_{x \sim \bar{P}, \pi|x \sim \bar{P}} \left[ \text{loss}(\hat{\pi}, \pi) + \theta \cdot \sum_{i=1}^{n} \left( \phi_i(x, \pi_i) - \phi_i(x, \pi_i) \right) \right]$$

Augmented Hamming loss matrix for \( n = 3 \) permutations

<table>
<thead>
<tr>
<th>( \hat{\pi} )</th>
<th>( \bar{\pi} = 123 )</th>
<th>( \bar{\pi} = 132 )</th>
<th>( \bar{\pi} = 213 )</th>
<th>( \bar{\pi} = 231 )</th>
<th>( \bar{\pi} = 312 )</th>
<th>( \bar{\pi} = 321 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\pi} = 123 )</td>
<td>0 + ( \delta_{123} )</td>
<td>2 + ( \delta_{132} )</td>
<td>2 + ( \delta_{213} )</td>
<td>3 + ( \delta_{231} )</td>
<td>3 + ( \delta_{312} )</td>
<td>2 + ( \delta_{321} )</td>
</tr>
<tr>
<td>( \hat{\pi} = 132 )</td>
<td>2 + ( \delta_{123} )</td>
<td>0 + ( \delta_{132} )</td>
<td>3 + ( \delta_{213} )</td>
<td>2 + ( \delta_{231} )</td>
<td>2 + ( \delta_{312} )</td>
<td>3 + ( \delta_{321} )</td>
</tr>
<tr>
<td>( \hat{\pi} = 213 )</td>
<td>2 + ( \delta_{123} )</td>
<td>3 + ( \delta_{132} )</td>
<td>0 + ( \delta_{213} )</td>
<td>2 + ( \delta_{231} )</td>
<td>2 + ( \delta_{312} )</td>
<td>3 + ( \delta_{321} )</td>
</tr>
<tr>
<td>( \hat{\pi} = 231 )</td>
<td>3 + ( \delta_{123} )</td>
<td>2 + ( \delta_{132} )</td>
<td>2 + ( \delta_{213} )</td>
<td>0 + ( \delta_{231} )</td>
<td>3 + ( \delta_{312} )</td>
<td>2 + ( \delta_{321} )</td>
</tr>
<tr>
<td>( \hat{\pi} = 312 )</td>
<td>3 + ( \delta_{123} )</td>
<td>2 + ( \delta_{132} )</td>
<td>2 + ( \delta_{213} )</td>
<td>3 + ( \delta_{231} )</td>
<td>0 + ( \delta_{312} )</td>
<td>2 + ( \delta_{321} )</td>
</tr>
<tr>
<td>( \hat{\pi} = 321 )</td>
<td>2 + ( \delta_{123} )</td>
<td>3 + ( \delta_{132} )</td>
<td>3 + ( \delta_{213} )</td>
<td>2 + ( \delta_{231} )</td>
<td>2 + ( \delta_{312} )</td>
<td>0 + ( \delta_{321} )</td>
</tr>
</tbody>
</table>

size: \( n! \times n! \)

Intractable for modestly-sized \( n \)
Polytope of the Permutation Mixtures

Dual:

\[ \min_{\theta} \mathbb{E}_{(x,\pi) \sim \tilde{P}} \min_{\tilde{P}(\pi|x)} \max_{P(\pi|x)} \mathbb{E}_{\tilde{P}|x \sim \tilde{P}} \sum_{i=1}^{n} I(\pi'_i \neq \pi_i) + \theta \cdot \sum_{i=1}^{n} (\phi_i(x, \tilde{\pi}_i) - \phi_i(x, \pi_i)) \]

Marginal Distribution Matrices:

**Predictor**

\[
P = \begin{pmatrix}
\hat{\pi}_1 & p_{1,1} & p_{1,2} & p_{1,3} \\
\hat{\pi}_2 & p_{2,1} & p_{2,2} & p_{2,3} \\
\hat{\pi}_3 & p_{3,1} & p_{3,2} & p_{3,3}
\end{pmatrix}
\]

\[p_{i,j} = \tilde{P}(\hat{\pi}_i = j)\]

**Adversary**

\[
Q = \begin{pmatrix}
\tilde{\pi}_1 & q_{1,1} & q_{1,2} & q_{1,3} \\
\tilde{\pi}_2 & q_{2,1} & q_{2,2} & q_{2,3} \\
\tilde{\pi}_3 & q_{3,1} & q_{3,2} & q_{3,3}
\end{pmatrix}
\]

\[q_{i,j} = \tilde{P}(\tilde{\pi}_i = j)\]

Birkhoff – Von Neumann theorem:

A convex polytope whose points are doubly stochastic matrices

\[P1 = P^T 1 = Q1 = Q^T 1 = 1\]

Reduce the space of optimization: from \(O(n!)\) to \(O(n^2)\)
Marginal Distribution Formulation

Dual:

\[
\min_\theta \mathbb{E}_{(x, \pi) \sim \tilde{P}} \min_{\tilde{P}(\tilde{\pi} | x)} \max_{P(\tilde{\pi} | x)} \mathbb{E}_{\tilde{\pi} | x \sim \tilde{P}, \pi | x \sim \tilde{P}} \left[ \sum_{i=1}^{n} I(\pi_i' \neq \pi_i) + \theta \cdot \sum_{i=1}^{n} (\phi_i(x, \tilde{\pi}_i) - \phi_i(x, \pi_i)) \right]
\]

Marginal Formulation:

Rearrange the optimization order and add regularization and smoothing penalties

\[
\max_{\mathbf{Q} \geq 0} \min_\theta \frac{1}{m} \sum_{i=1}^{m} \min_{\mathbf{P}_i \geq 0} \left[ \langle \mathbf{Q}_i - \mathbf{Y}_i, \sum_k \theta_k \mathbf{X}_{i,k} \rangle - \langle \mathbf{P}_i, \mathbf{Q}_i \rangle + \frac{\mu}{2} \| \mathbf{P}_i \|_{\mathcal{F}}^2 - \frac{\mu}{2} \| \mathbf{Q}_i \|_{\mathcal{F}}^2 \right] + \frac{1}{2} \| \theta \|_2^2
\]

s.t. : \( \mathbf{P}_i 1 = \mathbf{P}_i^T 1 = \mathbf{Q}_i 1 = \mathbf{Q}_i^T 1 = 1, \quad \forall i \)

Optimization Techniques Used:

- Outer (\( Q \)) : projected Quasi-Newton (Schmidt, et.al., 2009)
- Inner (\( \theta \)) : closed-form solution
- Inner (\( P \)) : projection to doubly-stochastic matrix
- Projection to doubly-stochastic matrix : ADMM
Consistency

Empirical Risk Perspective of Adversarial Bipartite Matching

\[
\min_{\theta} \mathbb{E}_{x \sim \tilde{P}} \left[ AL_{f_{\theta}}^{\text{perm}}(x, \pi) \right]
\]

where: \( AL_{f_{\theta}}^{\text{perm}}(x, \pi) \triangleq \min_{\tilde{P}(\tilde{x}|x)} \max_{\tilde{P}(\tilde{\pi}|x)} \mathbb{E}_{\tilde{\pi}|x \sim \tilde{P}} \left[ \text{loss}(\tilde{\pi}, \tilde{\pi}) + f_{\theta}(x, \tilde{\pi}) - f_{\theta}(x, \pi) \right] \)

\( \text{AL}^{\text{perm}} \) is consistent

when \( f \) is optimized over all measurable functions on the input space \((x, \pi)\)

\( \text{AL}^{\text{perm}} \) is also consistent

\( f \) is optimized over a restricted set of functions: \( f(x, \pi) = \sum_i g_i(x, \pi_i) \)

when \( g \) is allowed to be optimized over all measurable functions on the individual input space \((x, \pi_i)\)
Experiments

Application: Video Tracking

Empirical runtime (until convergence)

Table 5. Running time (in seconds) of the model for various number of elements $n$ with fixed number of samples ($m = 50$)

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Elements</th>
<th>Adv Marg.</th>
<th>SSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAMPUS</td>
<td>12</td>
<td>1.0</td>
<td>1.96</td>
</tr>
<tr>
<td>STADTMITTE</td>
<td>16</td>
<td>1.3</td>
<td>2.46</td>
</tr>
<tr>
<td>SUNNYDAY</td>
<td>18</td>
<td>1.5</td>
<td>2.75</td>
</tr>
<tr>
<td>PEDCROSS2</td>
<td>30</td>
<td>2.5</td>
<td>8.18</td>
</tr>
<tr>
<td>BAHNHOF</td>
<td>34</td>
<td>2.8</td>
<td>9.79</td>
</tr>
</tbody>
</table>

relative: 12=1.0  relative: 1.96=1.0

Public Benchmark Datasets

Table 3. Dataset properties

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Elements</th>
<th># Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUD-CAMPUS</td>
<td>12</td>
<td>70</td>
</tr>
<tr>
<td>TUD-STADTMITTE</td>
<td>16</td>
<td>178</td>
</tr>
<tr>
<td>ETH-SUNNYDAY</td>
<td>18</td>
<td>353</td>
</tr>
<tr>
<td>ETH-BAHNHOF</td>
<td>34</td>
<td>999</td>
</tr>
<tr>
<td>ETH-PEDCROSS2</td>
<td>30</td>
<td>836</td>
</tr>
</tbody>
</table>

Adversarial. Marginal Formulation: grows (roughly) quadratically in $n$

CRF: impractical even for $n = 20$

(Petterson et. al., 2009)
Experiment Results

Table 1: The mean and standard deviation (in parenthesis) of the average accuracy (1 - the average Hamming loss) for the adversarial bipartite matching model compared with Structured-SVM.

<table>
<thead>
<tr>
<th>Training/Testing</th>
<th>Adv. Bipartite Matching</th>
<th>Structured SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campus/Stadtmitte</td>
<td>0.662 (0.08)</td>
<td>0.662 (0.08)</td>
</tr>
<tr>
<td>Stadtmitte/Campus</td>
<td>0.667 (0.11)</td>
<td>0.660 (0.12)</td>
</tr>
<tr>
<td>Bahnhof/Sunnyday</td>
<td>0.754 (0.10)</td>
<td>0.729 (0.15)</td>
</tr>
<tr>
<td>Pedcross2/Sunnyday</td>
<td>0.750 (0.10)</td>
<td>0.736 (0.13)</td>
</tr>
<tr>
<td>Sunnyday/Bahnhof</td>
<td>0.751 (0.18)</td>
<td>0.739 (0.20)</td>
</tr>
<tr>
<td>Pedcross2/Bahnhof</td>
<td>0.763 (0.16)</td>
<td>0.731 (0.21)</td>
</tr>
<tr>
<td>Bahnhof/Pedcross2</td>
<td>0.714 (0.16)</td>
<td>0.701 (0.18)</td>
</tr>
<tr>
<td>Sunnyday/Pedcross2</td>
<td>0.712 (0.17)</td>
<td>0.700 (0.18)</td>
</tr>
</tbody>
</table>

6 pairs of dataset significantly outperforms SSVM

2 pairs of dataset competitive with SSVM
<table>
<thead>
<tr>
<th>Model</th>
<th>Efficient?</th>
<th>Consistent?</th>
<th>Perform well?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditional Random Field (CRF)</td>
<td>✗</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>(Petterson et. al., 2009; Volkovs &amp; Zemel, 2012)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structured SVM</td>
<td>✓</td>
<td>✗</td>
<td>—</td>
</tr>
<tr>
<td>(Tsochantaridis et. al., 2005)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adversarial Bipartite Matching</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(our approach)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion
Robust Adversarial Learning Algorithms

- Align better with the loss/performance metric (by incorporating the metric into its learning objective)
- Provide Fisher consistency guarantee
- Computationally efficient
- Perform well in practice
Future Directions
Future Directions (1)

1. Fairness in Machine Learning

*Important* issues in *automated decision* using ML algorithms.

*Requires* the algorithm to *produce fair* prediction.

Our formulation only enforces constraints on the adversary.

$$\min_{\hat{P}(\hat{Y}|X)} \max_{\tilde{P}(\tilde{Y}|X)} \mathbb{E}_{X,Y \sim \hat{P}, \tilde{Y}|X \sim \hat{P}, \tilde{Y}|X \sim \hat{P}} \left[ \text{loss}(\hat{Y}, \tilde{Y}) \right]$$

s.t. $\mathbb{E}_{X \sim \hat{P}, \tilde{Y}|X \sim \hat{P}}[\phi(X, \tilde{Y})] = \mathbb{E}_{X,Y \sim \tilde{P}}[\phi(X, Y)]$

*Add fairness constraints to the predictor?*

2. Statistical Theory of Loss Functions

In *multiclass classification* problem, both $\text{AL}^{0-1}$ and $\text{SVM-LLW}$ are *Fisher consistent*. However, their *performances* are quite different.

Is there any stronger statistical guarantee that can separate the high-performing Fisher consistent algorithm from the low-performing ones?
3. Structured Prediction & Graphical Models

More complex graphical structures are popular in some applications, e.g. computer vision.

Exact learning algorithms for AGM in this case may be intractable.

Can we develop learning algorithms for general graphical models?

What kind of approximation algorithms can be applicable?

4. Deep Learning

Deep learning has been successfully applied to many prediction problems.

Most of deep learning architectures are not designed to optimize customized loss metrics.

How can the robust adversarial learning approach help designing deep learning architectures?
Collaborators

AL
Anqi Liu

KA
Kaiser Asif

MB
Mohammad Bashiri

SB
Sima Behpour

AR
Ashkan Rezaei

WX
Wei Xing

BZ
Prof. Brian Ziebart

XZ
Prof. Xinhua Zhang
Publications

• **Consistent Robust Adversarial Prediction for General Multiclass Classification**  
  **Rizal Fathony**, Kaiser Asif, Anqi Liu, Mohammad Bashiri, Wei Xing, Sima Behpour, Xinhua Zhang, Brian D. Ziebart.  
  Submitted to JMLR.

• **Distributionally Robust Graphical Models**  
  **Rizal Fathony**, Ashkan Rezaei, Mohammad Bashiri, Xinhua Zhang, Brian D. Ziebart.  

• **Efficient and Consistent Adversarial Bipartite Matching**  
  **Rizal Fathony***, Sima Behpour*, Xinhua Zhang, Brian D. Ziebart.  

• **Adversarial Surrogate Losses for Ordinal Regression**  
  **Rizal Fathony**, Mohammad Bashiri, Brian D. Ziebart.  

• **Adversarial Multiclass Classification: A Risk Minimization Perspective**  

• **Kernel Robust Bias-Aware Prediction under Covariate Shift**  
Thank You