The model parameters $\theta$ for multiclass zero-one adversarial classification are equivalently obtained from empirical risk minimization under the adversarial zero-one loss function: 

$$AL^{\theta}(x, y) = \frac{1}{|S|} \sum_{i \in S} \|y_i(x, y) - f_i(x)\|_1 - 1,$$

where $S$ is any non-empty member of the powerset of classes $\{1, 2, \ldots, |Y|\}$.

Plots of $AL^{\theta}$ in binary and 3-class classification:

- $AL^{\theta}$ is the maximum value over $2^{|Y|} - 1$ linear hyperplanes.
- Binary classification: similar with hinge loss, but with two hinges at -1 and 1 (as shown in figure on the right).
- Three class classification: the loss has seven facets.
- Comparison with WW and CS surrogates (as shown below).

**Figure:** Loss function contour plots over the space of potential differences for the prediction task with three classes when $y = 1$ under $AL^{\theta}$, WW, and CS.

### Results and Experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Linear Kernel</th>
<th>Gaussian Kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>iris</td>
<td>96.9 (3.1)</td>
<td>96.4 (2.4)</td>
</tr>
<tr>
<td>glass</td>
<td>62.5 (6.0)</td>
<td>62.3 (3.9)</td>
</tr>
<tr>
<td>redwine</td>
<td>58.8 (5.9)</td>
<td>58.6 (3.7)</td>
</tr>
<tr>
<td>vehicle</td>
<td>78.8 (2.5)</td>
<td>78.4 (2.3)</td>
</tr>
<tr>
<td>segment</td>
<td>94.9 (5.7)</td>
<td>94.7 (5.5)</td>
</tr>
<tr>
<td>sat</td>
<td>90.7 (6.0)</td>
<td>90.4 (6.0)</td>
</tr>
<tr>
<td>optdigits</td>
<td>96.6 (6.0)</td>
<td>96.5 (6.0)</td>
</tr>
<tr>
<td>pagestacks</td>
<td>96.5 (5.5)</td>
<td>96.4 (5.5)</td>
</tr>
<tr>
<td>libras</td>
<td>90.1 (6.0)</td>
<td>90.0 (6.0)</td>
</tr>
<tr>
<td>vertbal</td>
<td>95.5 (2.9)</td>
<td>95.4 (2.9)</td>
</tr>
<tr>
<td>avg</td>
<td>81.59</td>
<td>81.32</td>
</tr>
</tbody>
</table>

**Optimization**

- **Primal Optimization using Stochastic Sub-gradient Descent**
  - The sub-gradient of $AL^{\theta}$ includes the mean of feature differences.

- **Dual Optimization using Quadratic Programming (QP)**
  - Constrained OP of $AL^{\theta}$ plus L2 regularization.
  - Minimizes $\frac{1}{|S|} \sum_{i \in S} \sum_{j \in S} \alpha_i \alpha_j L_{ij}$.

### Acknowledgments

This research was supported as part of the Future of Life Institute (futureoflife.org) FLI-RFP-PAI program, grant#2016-156780 and by NSF grant RI-1536379.

### References

- Crammer and Singer (2002).
- Tewari and Bartlett (2004).
- Doğan et al. (2016).

### Related Work

- **Multiclass Support Vector Machine**
  - Three main formulations:
    1. WW by Weston and Watkins (1999): $\text{max} \sum_{i=1}^{|Y|} \sum_{j \in S} \alpha_i \hat{y}_i(x, y) - \sum_{i \in S} \alpha_i y_i(x, y)$.
    2. CS by Crammer and Singer (2002): $\text{max} \sum_{i=1}^{|Y|} \sum_{j \in S} \alpha_i \hat{y}_i(x, y) - \sum_{i \in S} \alpha_i y_i(x, y)$.
    3. LLW by Lee, Lin, and Wahba (2004): $\text{max} \sum_{i=1}^{|Y|} \sum_{j \in S} \alpha_i \hat{y}_i(x, y) - \sum_{i \in S} \alpha_i y_i(x, y)$.

- **Adversarial Prediction Games**
  - Two player zero-sum games:
    1. Adversary player - controls conditional label distribution $P(x)$.
    2. Estimator player - controls $P(y|x)$ and seeks to minimize expected loss.

- **Fisher Consistency**
  - Minimizing a Fisher consistent loss will yield the Bayes optimal decision boundary given the true distribution, $P(x, y)$.
  - Multiclass: it requires $\arg \max f_i(x) \subseteq \arg \max f_i(x)$.
  - $P(x) = \sum_{y \in S} P(y|x) = 1$.

- **Universal Consistency**
  - $AL^{\theta}$ is a Lipschitz loss with constant 1, optimizing it with universal kernel and reasonably small regularization on any distribution yields Bayes optimal classifier. (Steinwart et al. 2008)