Bayes Consistency vs. $\mathcal{H}$-Consistency:

The Interplay between Surrogate Loss Functions and the Scoring Function Class

Appendix

A  Proof of Lemma[1]

Proof. This essentially follows from the definition of $\mathcal{F}_{spwlin}$. In particular, we have:
\[
 f_y(x) \geq 0 \iff \min_{y' \neq y} \{(w_{y} - w_{y'})^\top x + (b_{y} - b_{y'})\} \geq 0 \\
 \iff \min_{y' \neq y} \{(w_{y}^\top x + b_{y}) - (w_{y'}^\top x + b_{y'})\} \geq 0 \\
 \iff (w_{y}^\top x + b_{y}) \geq (w_{y'}^\top x + b_{y'}) \quad \forall y', y' \neq y \\
 \iff y \in \text{argmax}_{y' \in [n]} w_{y'}^\top x + b_{y'}.
\]

\[\square\]

B  Proof of Theorem[2]

Proof. Let $D$ be a $\mathcal{H}_{lin}$-realizable distribution. Then $\exists h^* \in \mathcal{H}_{lin}$ such that $P_{(X,Y) \sim D}(Y = h^*(X)) = 1$, and therefore $\epsilon_D^{\mathcal{H}_{lin}}[\mathcal{H}_{lin}] = 0$. Thus our goal is to show that $\exists$ a strictly increasing function $g : \mathbb{R}_+ \to \mathbb{R}_+$ that is continuous at 0 with $g(0) = 0$ such that for all $f \in \mathcal{F}_{spwlin}$,
\[
 \epsilon_D^{D}[\text{argmax} \circ f] \leq g\left(\epsilon_D^{\text{OvA,log}}[f] - \epsilon_D^{\text{OvA,log}}[\mathcal{F}_{spwlin}]\right).
\]

We will do this in two parts:

(1) We will show that $\epsilon_D^{\text{OvA,log}}[\mathcal{F}_{spwlin}] = 0$.

(2) We will show that for all $f \in \mathcal{F}_{spwlin}$, $\epsilon_D^{D}[\text{argmax} \circ f] \leq \frac{1}{\ln(2)} \epsilon_D^{\text{OvA,log}}[f]$.

Putting these together will then give that for all $f \in \mathcal{F}_{spwlin}$,
\[
 \epsilon_D^{D}[\text{argmax} \circ f] \leq \frac{1}{\ln(2)} \left(\epsilon_D^{\text{OvA,log}}[f] - \epsilon_D^{\text{OvA,log}}[\mathcal{F}_{spwlin}]\right).
\]

Part 1. We will show that for any sufficiently small $\epsilon > 0$, $\exists f^* \in \mathcal{F}_{spwlin}$ such that $\epsilon_D^{\text{OvA,log}}[f^*] < \epsilon$; this will establish that $\epsilon_D^{D}[\mathcal{F}_{spwlin}] = 0$.

Let $0 < \epsilon < 2n \ln(2)$. Since $h^* \in \mathcal{H}_{lin}$, we have $\exists \{w_{y^*}, b_{y^*}\}_{y=1}^n$ such that
\[
 h^*(x) = \text{argmax}_{y \in [n]} (w_{y^*})^\top x + b_{y^*} \quad \forall x.
\]

Define $f^* \in \mathcal{F}_{spwlin}$ as
\[
 f_y^*(x) = \min_{y' \neq y} \{(w_{y^*} - w_{y^*})^\top x + (b_{y^*} - b_{y^*})\} \\
 = \min_{y' \neq y} \{(w_{y^*}^\top x + b_{y^*}) - ((w_{y^*})^\top x + b_{y^*})\}.
\]

Then we have
\[
 P_{(X,Y) \sim D}(f_y^*(X) > 0) = 1.
\]

Therefore $\exists \kappa > 0$ such that
\[
 P_{(X,Y) \sim D}(f_y^*(X) < \kappa) \leq \frac{\epsilon}{2n \ln(2)}.
\]
Define $f^e \in \mathcal{F}_{\text{spwlin}}$ as
\[
f^e_y(x) = \frac{f^e_y(x)}{\kappa} \ln \left( \frac{1}{e^\kappa/2n - 1} \right).
\]
Then it can be verified that
\[
f^e_y(x) > 0 \implies f^e_y(x) > 0 \implies \psi_{\text{OvA,log}}(y, f^e(x)) \leq n \ln(2),
\]
and moreover,
\[
f^e_y(x) \geq \kappa \implies f^e_y(x) \geq \ln \left( \frac{1}{e^\kappa/2n - 1} \right) \implies \psi_{\text{OvA,log}}(y, f^e(x)) \leq \frac{\epsilon}{2}.
\]
This gives
\[
er^\text{OvA,log}_{\mathcal{D}}[f^e] = E_{(X,Y) \sim \mathcal{D}} \left[ \psi_{\text{OvA,log}}(Y, f^e(X)) \right]
\leq P_{(X,Y) \sim \mathcal{D}} \left( 0 < f^e_Y(X) < \kappa \right) \cdot E \left[ \psi_{\text{OvA,log}}(Y, f^e(X)) \mid 0 < f^e_Y(X) < \kappa \right]
+ P_{(X,Y) \sim \mathcal{D}} \left( f^e_Y(X) \geq \kappa \right) \cdot E \left[ \psi_{\text{OvA,log}}(Y, f^e(X)) \mid f^e_Y(X) \geq \kappa \right]
\leq \frac{\epsilon}{2n \ln(2)} \cdot n \ln(2) + 1 \cdot \frac{\epsilon}{2}
= \epsilon.
\]

**Part 2.** Let $f \in \mathcal{F}_{\text{spwlin}}$, and let $\{w_y, b_y\}_{y=1}^n$ be such that
\[
f_y(x) = \min_{y' \neq y} \left\{ (w_y - w_{y'})^\top x + (b_y - b_{y'}) \right\} \quad \forall x.
\]
Define $h : \mathcal{X} \to \mathcal{Y}$ such that
\[
h(x) \in \arg\max_{y \in [n]} f_y(x) \quad \forall x.
\]
Then we have
\[
er_D^{\text{OvA,log}}[h] = E_{(X,Y) \sim \mathcal{D}} \left[ \ell_{0:1}(Y, h(X)) \right]
= E_{(X,Y) \sim \mathcal{D}} \left[ 1(h(X) \neq Y) \right]
\leq E_{(X,Y) \sim \mathcal{D}} \left[ \sum_{y \neq Y} 1(h(X) = y) \right]
\leq \frac{1}{\ln(2)} E_{(X,Y) \sim \mathcal{D}} \left[ \sum_{y \neq Y} \ln \left( 1 + e^{f_y(X)} \right) \right]
\leq \frac{1}{\ln(2)} E_{(X,Y) \sim \mathcal{D}} \left[ \ln \left( 1 + e^{-f_y(X)} \right) + \sum_{y \neq Y} \ln \left( 1 + e^{f_y(X)} \right) \right]
\quad \text{(since $\ln(1 + e^{-f_y(x)}) \geq 0 \quad \forall (x, y)$)}
= \frac{1}{\ln(2)} E_{(X,Y) \sim \mathcal{D}} \left[ \ell_{\text{OvA,log}}(Y, f(X)) \right]
= \frac{1}{\ln(2)} er^\text{OvA,log}_{\mathcal{D}}[f].
\]

**C Proof of Theorem**

**Proof.** Let $w_1, \ldots, w_n \in \mathbb{R}^d, b_1, \ldots, b_n \in \mathbb{R}$, and let $f \in \mathcal{F}_{\text{spwlin}}$ be parametrized by $\{w_y, b_y\}_{y=1}^n$, so that
\[
f_y(x) = \min_{y' \neq y} \left\{ (w_y - w_{y'})^\top x + (b_y - b_{y'}) \right\} \quad \forall x.
\]
We will show that
\[
\arg\max_{y \in [n]} f_y(x) = \arg\max_{y \in [n]} w_y^\top x + b_y ;
\]
this will establish the result.

To see that the above claim is true, notice that we can write
\[
f_y(x) = (w_y^\top x + b_y) - \max_{y' \neq y} \{w_{y'}^\top x + b_{y'}\}.
\]
In other words, \(f_y(x)\) is the difference between \((w_y^\top x + b_y)\) and the largest value of \((w_{y'}^\top x + b_{y'})\) among \(y' \neq y\). Clearly, this difference is largest when \((w_y^\top x + b_y) \geq (w_{y'}^\top x + b_{y'}) \forall y' \neq y\) (in particular, in this case the difference is non-negative; in all other cases, the difference is negative, and therefore smaller). Thus
\[
f_y(x) \geq f_{y'}(x) \forall y' \neq y \iff (w_y^\top x + b_y) \geq (w_{y'}^\top x + b_{y'}) \forall y' \neq y .
\]
This proves the claim. \(\square\)

D Proof of Corollary 4

This follows directly from the proof of Theorem 5.

E Details of Real Data Sets Used in Experiments in Section 5.2

<table>
<thead>
<tr>
<th>Data set</th>
<th># train</th>
<th># validation</th>
<th># test</th>
<th># classes</th>
<th># features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covertype (50K)</td>
<td>30000</td>
<td>10000</td>
<td>10000</td>
<td>7</td>
<td>54</td>
</tr>
<tr>
<td>Digits</td>
<td>5620</td>
<td>1874</td>
<td>3498</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>USPS</td>
<td>5468</td>
<td>1823</td>
<td>2007</td>
<td>10</td>
<td>256</td>
</tr>
<tr>
<td>MNIST (70K)</td>
<td>45000</td>
<td>15000</td>
<td>10000</td>
<td>10</td>
<td>780</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>37500</td>
<td>12500</td>
<td>10000</td>
<td>10</td>
<td>3072</td>
</tr>
<tr>
<td>Sensorless</td>
<td>35105</td>
<td>11702</td>
<td>11702</td>
<td>11</td>
<td>48</td>
</tr>
<tr>
<td>Letter</td>
<td>10500</td>
<td>4500</td>
<td>5000</td>
<td>26</td>
<td>16</td>
</tr>
</tbody>
</table>

Notes:

Subsampling: For Covertype, we used a random subsample of the original data set containing 50,000 examples (the original data set has 581,012 examples).

Image data sets with pixel features: The versions of the USPS and MNIST datasets that we used came with features scaled to the ranges \([-1, 1]\] and \([0, 1]\), respectively. For CIFAR10, we similarly scaled the features to the range \([0, 1]\) by dividing all features by 255.