Supplementary Materials of "On Reject and Refine Options in Multicategory Classification"

Chong Zhang, Wenbo Wang and Xingye Qiao

January 10, 2017

1 Outlines of the Proofs of Theorems 1, 2 and 3

We first provide brief outlines of the proofs of Theorems 1, 2 and 3.

Theorem 1 and Theorem 2

There are two major steps in the proof of Theorems 1 and 2. The first step is to decompose the excess ℓ-risk into the estimation error and approximation error. Then we show that the probability of the estimation error exceeding $O(skr\log(r^{-1}))$ for the L_1 penalty, or $O(\sqrt{ps}kr\log(r^{-1}))$ for the L_2 penalty, can be written in terms of a concentration inequality indexed by a scaled empirical process. The second step is to obtain a suitable probability upper bound of this concentration inequality. To this end, one can use the chaining technique, which discretizes the functional space of the optimization problem, hence decomposing the corresponding probability into several parts. For each part, the probability can be controlled by established concentration inequalities. See Theorem A.2 in Wang and Shen (2007) for an example.

Therefore, the question boils down to control the complexity of the discretized functional space. A common approach to depict such complexity in the literature is to use the entropy numbers. In the Supplementary Materials, for linear and kernel learning, we introduce Lemmas 2 and 4 respectively, to control the complexity of the corresponding functional spaces for the empirical processes, in terms of their L_2 entropy numbers. In particular, we show that for a small and positive ϵ, the ϵ-entropy numbers for linear and kernel learning are in the order of $O(\epsilon^{-2})$ under mild conditions. Consequently, we can prove the desired concentration inequality.

It should be noted that, although the orders of the entropy numbers for linear and kernel learning are similar, the techniques used are quite different. In particular, in linear learning, we treat the functional space as a convex hull of $2p$ functions, which leads to a bound on the entropy number. For kernel learning, we consider the natural embedding of the kernel function into the regular L_2 functional space consisting of continuous functions on the domain of $\boldsymbol{x}$. Such embedding can be shown to be absolutely 2-summing with 2-summing norm no larger than 1. Hence we can bound the entropy number of this embedding operator (which can be shown to be the same as the entropy number of the original kernel space) by its corresponding approximation numbers, which can be further bounded by Carl's inequality between approximation and entropy numbers.

Theorem 3

Theorem 3 extends the well established results on fast rate of convergence from binary classifiers to multicategory ones. The key to the proof is to find a pseudo-norm that can be used to both upper and lower bound the conditional excess ℓ-risk $g_{\boldsymbol{f}}(\boldsymbol{x}, y) = \sum_{j \neq y} \ell\{\langle \boldsymbol{f}, \mathcal{Y}_j \rangle\} - \sum_{j \neq y} \ell\{\langle \boldsymbol{f}^*, \mathcal{Y}_j \rangle\}$ (up to constants). In Bartlett and Wegkamp (2008), as the modified hinge loss function $\psi(u)$ is piecewise linear, and remains flat for large u, one can use $\rho(f_1, f_2) \propto |f_1 - f_2|$ as the pseudo-norm. However, for more general loss functions, especially differentiable loss functions, an L_1 type pseudo-norm cannot lower bound the conditional excess ℓ-risk. Therefore, we employ the (squared) L_2 type pseudo-norm in this proof. With the low noise assumption, we can show that the class $\{g_{\boldsymbol{f}}(\boldsymbol{x}, y)\}$ is a Bernstein class with the Bernstein exponent $\alpha/(1+\alpha)$. The next step is to apply the symmetrization technique, and show that the estimation error can be (up to a constant) bounded by a tail probability plus a small term that converges to zero at a very fast speed, where the tail probability term is indexed by an empirical process of $\{g_{\boldsymbol{f}}(\boldsymbol{x}, y)\}$. At this stage, we can employ Bernstein's inequality to bound the corresponding tail probability. As $\{g_{\boldsymbol{f}}(\boldsymbol{x}, y)\}$ is a Bernstein class, the variance term in the power of the upper bound in Bernstein's inequality can be bounded by a linear term of $\mathbb{E} g_{\boldsymbol{f}}(\boldsymbol{x}, y)$. Combined with an upper bound on the entropy number for Gaussian kernel space, we can prove the desired result in Theorem 3.

2 Detailed Proofs to Propositions and Theorems

Before the proofs, we first introduce a lemma for simplicity and completeness of further arguments.

Lemma 1 (Zhang and Liu, 2014, Lemma 1)**.** *Suppose we have an arbitrary* $\boldsymbol{f} \in \mathbb{R}^{k-1}$*. For any* $u, v \in \{1, \ldots, k\}$ *such that* $u \neq v$*, define* $\boldsymbol{T}_{u,v} = \mathcal{Y}_u - \mathcal{Y}_v$*. For any scalar* $z \in \mathbb{R}$*,* $\langle (\boldsymbol{f} + z\boldsymbol{T}_{u,v}), \mathcal{Y}_w \rangle = \langle \boldsymbol{f}, \mathcal{Y}_w \rangle$*, where* $w \in \{1, \ldots, k\}$ *and* $w \neq u, v$*. Furthermore, we have that* $\langle (\boldsymbol{f} + z\boldsymbol{T}_{u,v}), \mathcal{Y}_u \rangle - \langle \boldsymbol{f}, \mathcal{Y}_u \rangle = -\langle (\boldsymbol{f} + z\boldsymbol{T}_{v,u}), \mathcal{Y}_v \rangle + \langle \boldsymbol{f}, \mathcal{Y}_v \rangle$*.*

Lemma 1 shows that we can increase $\langle \boldsymbol{f}^*, \mathcal{Y}_i \rangle$ by an arbitrary $\epsilon > 0$, and decrease $\langle \boldsymbol{f}^*, \mathcal{Y}_j \rangle$ by the same ϵ without changing $\langle \boldsymbol{f}^*, \mathcal{Y}_l \rangle$ for $l \notin \{i, j\}$.

Proof of Proposition 1: We aim to find $\boldsymbol{f}^*$ that minimizes the conditional expected loss

$$\sum_{j=1}^{k} P_j \{ \sum_{i \neq y} \ell(\langle \boldsymbol{f}, \mathcal{Y}_i \rangle) \},$$

which is equivalent to find

$$\underset{\boldsymbol{f}}{\operatorname{argmin}} \sum_{j=1}^{k} \ell(\langle \boldsymbol{f}, \mathcal{Y}_j \rangle)(1 - P_j).$$

We assume $P_1 \geq P_2 \geq \cdots \geq P_k$ in this proof for simplicity.

First, we show that $\langle \boldsymbol{f}^*, \mathcal{Y}_1 \rangle \geq \langle \boldsymbol{f}^*, \mathcal{Y}_2 \rangle \geq \cdots \langle \boldsymbol{f}^*, \mathcal{Y}_k \rangle$. We prove this by contradiction. Suppose $\langle \boldsymbol{f}^*, \mathcal{Y}_1 \rangle < \langle \boldsymbol{f}^*, \mathcal{Y}_2 \rangle$. By Lemma 1, we can define $\tilde{\boldsymbol{f}}$ such that $\langle \boldsymbol{f}^*, \mathcal{Y}_1 \rangle = \langle \tilde{\boldsymbol{f}}, \mathcal{Y}_2 \rangle$ and $\langle \boldsymbol{f}^*, \mathcal{Y}_2 \rangle = \langle \tilde{\boldsymbol{f}}, \mathcal{Y}_1 \rangle$. One can verify that $\sum_{j=1}^{k} \ell(\langle \boldsymbol{f}^*, \mathcal{Y}_j \rangle)(1 - P_j) > \sum_{j=1}^{k} \ell(\langle \tilde{\boldsymbol{f}}, \mathcal{Y}_j \rangle)(1 - P_j)$, which is a contradiction to the definition of $\boldsymbol{f}^*$. Notice that this argument holds true for any pairwise comparisons between $\langle \boldsymbol{f}^*, \mathcal{Y}_i \rangle$ and $\langle \boldsymbol{f}^*, \mathcal{Y}_j \rangle$ for $i \neq j$. Therefore, we have $\langle \boldsymbol{f}^*, \mathcal{Y}_1 \rangle \geq \langle \boldsymbol{f}^*, \mathcal{Y}_2 \rangle \geq \cdots \langle \boldsymbol{f}^*, \mathcal{Y}_k \rangle$.

The second step is to start from $\boldsymbol{f} = 0$, and consider the pairwise comparison between $(P_1, \langle \boldsymbol{f}, \mathcal{Y}_1 \rangle)$ and $(P_q, \langle \boldsymbol{f}, \mathcal{Y}_j \rangle)$ for $q > 1$, in order to decrease the conditional expected loss. By Lemma 1 and similar argument as in Section 3.1, one can verify that if $a(1 - P_1) < (1 - P_q)$, we should increase $\langle \boldsymbol{f}, \mathcal{Y}_1 \rangle$ and decrease $\langle \boldsymbol{f}, \mathcal{Y}_q \rangle$ to decrease the conditional expected loss. If $a(1 - P_1) \geq (1 - P_q)$, we should keep $\langle \boldsymbol{f}, \mathcal{Y}_q \rangle$ at 0. After $k - 1$ such comparisons, one can verify that $\boldsymbol{f}$ is such that if the assumption in Proposition 1 holds, then $\langle \boldsymbol{f}, \mathcal{Y}_1 \rangle > 0$, $\langle \boldsymbol{f}, \mathcal{Y}_2 \rangle = \cdots = \langle \boldsymbol{f}, \mathcal{Y}_j \rangle = 0$, and $\langle \boldsymbol{f}, \mathcal{Y}_q \rangle < 0$ for $j + 1 \leq q \leq k$. Note that the fact $\ell'(u)$ is a constant for $u > 0$ is essential for this sequence of pairwise comparisons to hold.

The last step is to check that $\boldsymbol{f}^* = \boldsymbol{f}$, where $\boldsymbol{f}$ is obtained in the second step. To this end, notice that if $\boldsymbol{f}^* \neq \boldsymbol{f}$, we can always perform the pairwise comparison as in the second step to decrease the conditional expected loss. Therefore, Proposition 1 holds. □

Proof of Proposition 2: For any fixed $\boldsymbol{x}$, the conditional expected loss for the reject option is d. To predict class label, clearly $\hat{y} = Y_{(1)}$ is the only admissible decision, whose conditional expected loss is $1 - P_{(1)}$. Therefore, we would predict the label when $1 - P_{(1)} < d$, and we reject when $1 - P_{(1)} \geq d$. □

Proof of Proposition 3: We assume $P_1 \geq P_2 \geq \cdots \geq P_k$ in this proof for simplicity. To prove the lower bound, suppose $a(1-P_1) > (1-P_k)$. Consequently, we have $a(1-P_1) > (1-P_j)$ for $j \geq 2$, which further leads to $a(k-1)(1-P_1) > k-1-(1-P_1)$. With some calculation, this is equivalent to $1 - P_1 > \frac{k-1}{a(k-1)+1}$. Therefore, by letting $\frac{k-1}{a(k-1)+1} = d$, or equivalently, $a = a_1$, we can prove that the lower bound inequality holds.

To prove the upper bound, suppose $1 - P_1 > d$. With $a = a_2 = \frac{(k-1)(1-d)}{d}$, we have $a(1-P_1) > (k-1)(1-d) > (k-1)P_1 > \sum_{j=1}^{k-1} P_j = 1 - P_k$. This proves the upper bound.

To see the tightness of these 2 bounds, one can easily construct numerical counter examples, and we omit the details here. □

Proof of Theorem 1: The key to the proof of this theorem is to bound the tail probability that the deviation of a related empirical process from its expected value exceeds a certain threshold. This consists of two major parts. The first part is to transform the problem into the empirical process, and the second part is to bound the corresponding tail probability.

Recall the definition of $\mathcal{F}(p,k,s)$. Define $t(p,s) = s$ if we use the L_1 penalty, and $t(p,s) = (ps)^{1/2}$ if we use the L_2 penalty. One can verify that for L_1 or L_2 penalized method, and any $j \in \{1,\ldots,k-1\}$, $|\hat{f}_j| = |\hat{\boldsymbol{\beta}}_j^T \boldsymbol{x}| \leq t(p,s)$. Therefore, in future arguments, it suffices to consider $\tilde{\mathcal{F}}(p,k,s) = \mathcal{F}(p,k,s) \cap \{\boldsymbol{f} : \|\boldsymbol{f}\| \leq (k-1)t(p,s)\}$. Furthermore, define $\boldsymbol{f}^{(p,k,s)} = \mathrm{argmin}_{\boldsymbol{f} \in \tilde{\mathcal{F}}(p,k,s)} \mathbb{E}[\sum_{j \neq y} \ell\{\langle \boldsymbol{f}, \mathcal{Y}_j \rangle\}]$,

$$h_{\boldsymbol{f}}(\cdot) = \{2(k-1)\frac{1-d}{d}t(p,s)\}^{-1}\{\sum_{j \neq \cdot} \ell(\langle \boldsymbol{f}, \mathcal{Y}_j \rangle) - \sum_{j \neq \cdot} \ell\big(\langle \boldsymbol{f}^{(p,k,s)}, \mathcal{Y}_j \rangle\big)\},$$

and $\bar{H} = \{h_{\boldsymbol{f}} : \boldsymbol{f} \in \tilde{\mathcal{F}}(p,k,s)\}$. Since ℓ is Lipschitz with Lipschitz constant $\frac{(k-1)(1-d)}{d}$ and $\sum_{j=1}^{k} \langle \boldsymbol{f}, \mathcal{Y}_j \rangle = 0$, $|\sum_{j \neq \cdot} \ell(\langle \boldsymbol{f}, \mathcal{Y}_j \rangle) - \sum_{j \neq \cdot} \ell(\langle \boldsymbol{f}', \mathcal{Y}_j \rangle)| \leq |\frac{(k-1)(1-d)}{d} \langle \boldsymbol{f} - \boldsymbol{f}', \cdot \rangle| \leq 2\frac{(k-1)(1-d)}{d} t(p,s)$. Therefore, we have the $L_2(Q)$ diameter of $\{\sum_{j \neq \cdot} \ell(\langle \boldsymbol{f}, \mathcal{Y}_j \rangle) - \sum_{j \neq \cdot} \ell(\langle \boldsymbol{f}^{(p,k,s)}, \mathcal{Y}_j \rangle)\}$ is bounded by $\{2\frac{(k-1)(1-d)}{d} t(p,s)\}$, and the $L_2(Q)$ diameter of $\bar{H}$ is bounded by 1. Here Q is any arbitrary distribution.

The next lemma bounds the complexity of $\bar{H}$ in terms of its $L_2(Q)$ entropy. For any $\epsilon > 0$, we can define $\mathcal{G}$ to be an ϵ-net of a function class $\mathcal{F}$ if, for any $f \in \mathcal{F}$, there exists $g \in \mathcal{G}$ such that $\|g - f\|_{Q,2} \leq \epsilon$. Let the $L_2(Q)$ covering number $N\{\epsilon, \mathcal{F}, L_2(Q)\}$ be the minimum size of all such possible ϵ-nets, and denote by $H\{\epsilon, \mathcal{F}, L_2(Q)\}$ the logarithm of $N\{\epsilon, \mathcal{F}, L_2(Q)\}$, which is referred to as the $L_2(Q)$ entropy. Define the uniform $L_2(Q)$ covering number, $N(\epsilon, \mathcal{F})$, to be $\sup_Q N\{\epsilon, \mathcal{F}, L_2(Q)\}$, and define the uniform $L_2(Q)$ entropy $H(\epsilon, \mathcal{F})$ in a similar manner. Lemma 2 gives an upper bound on $H(\epsilon, \bar{H})$.

Lemma 2. *For any $\epsilon > 0$, $H(\epsilon, \bar{H}) \leq \frac{2(k-1)}{\epsilon^2} \log(e + 2pe\epsilon^2)$.*

Proof of Lemma 2: To bound the $L_2(Q)$ entropy of $\bar{H}$, we can first bound the $L_2(Q)$ entropy of $\mathcal{G} := \{\sum_{j\neq\cdot} \ell(\langle \boldsymbol{f}, \mathcal{Y}_j\rangle) : \sum_{j=1}^{k-1} \|\boldsymbol{\beta}_j\|_1 \leq t(p,s)\}$, as a $\{2(k-1)\frac{1-d}{d}t(p,s)\epsilon\}$-net on $\mathcal{G}$ naturally introduces an ϵ-net on $\bar{H}$. To this end, we find an ϵ-net on $\mathcal{G}$. Let $g = \sum_{j\neq\cdot} \ell(\langle \boldsymbol{f}, \mathcal{Y}_j\rangle)$, $g' = \sum_{j\neq\cdot} \ell(\langle \boldsymbol{f}', \mathcal{Y}_j\rangle) \in \mathcal{G}$. Notice that

$$\begin{aligned}
\|g - g'\|_{Q,2}^2 &= \mathbb{E}[\sum_{j\neq\cdot} \ell\{\langle \mathcal{Y}_j, \boldsymbol{f}(\boldsymbol{X})\rangle\} - \sum_{j\neq\cdot} \ell\{\langle \mathcal{Y}_j, \boldsymbol{f}'(\boldsymbol{X})\rangle\}]^2 \\
&\leq \mathbb{E}\{\frac{(k-1)(1-d)}{d} \sum_{j\neq\cdot} \langle \mathcal{Y}_j, \boldsymbol{f}(\boldsymbol{X}) - \boldsymbol{f}'(\boldsymbol{X})\rangle\}^2 \\
&\leq \mathbb{E}\{\frac{(k-1)(1-d)}{d} \sum_{j=1}^{k-1} |f_j(\boldsymbol{X}) - f_j'(\boldsymbol{X})|\}^2 \\
&\leq \frac{(k-1)^3(1-d)^2}{d^2} \sum_{j=1}^{k-1} \|f_j - f_j'\|_{Q,2}^2.
\end{aligned}$$

where the last step is from the Cauchy-Schwartz inequality. Next, we define $\vec{\boldsymbol{x}} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_{k-1}^T)^T$ with each $\boldsymbol{x}_j$ a p-dimensional vector. Let $\vec{f}(\vec{\boldsymbol{x}}) = \sum_{j=1}^{k-1} \boldsymbol{\beta}_j^T \boldsymbol{x}_j$. Also let $\vec{Q}$ be the distribution of $\vec{\boldsymbol{X}} = (\delta_1 \boldsymbol{X}_1, \ldots, \delta_{k-1}\boldsymbol{X}_{k-1})$, where $\boldsymbol{X}_j$'s are independent and identically distributed with any arbitrary distribution Q, and $(\delta_1, \ldots, \delta_{k-1})$ has a joint distribution $\mathrm{pr}\{(\delta_1, \ldots, \delta_{k-1})^T = e_j\} = (k-1)^{-1}$. Thus we may conclude that $\sum_{j=1}^{k-1} \|f_j - f_j'\|_{Q,2}^2 = (k-1)\mathbb{E}_{\vec{Q}}(\vec{f} - \vec{f}')^2$, and $\|g - g'\|_{Q,2}^2 \leq \frac{(k-1)^4(1-d)^2}{d^2} \|\vec{f} - \vec{f}'\|_{Q,2}^2$. Consequently, if we can bound $L_2(Q)$ entropy of the function class $\vec{\mathcal{F}} = \{\vec{f} : \vec{f}(\vec{\boldsymbol{x}}) = \sum_{j=1}^{k-1} \sum_{l=1}^{p} \beta_{j,l} x_{j,l};\ \sum_{j=1}^{k-1} \|\boldsymbol{\beta}_j\|_1 \leq t(p,s)\}$, we can bound $\bar{H}$.

To bound the entropy of $\vec{\mathcal{F}}$, we define $w_{j,l}(\vec{\boldsymbol{x}}) = t(p,s)x_{j,l}$. Hence, $\mathcal{J} = \{\pm w_{j,l}\}$ forms a basis for $\vec{\mathcal{F}}$. In other words, each $\vec{f} = \sum_{j=1}^{k-1} \sum_{l=1}^{p} \beta_{j,l} x_{j,l} = \sum_{j=1}^{k-1} \sum_{l=1}^{p} |\beta_{j,l}| \{\mathrm{sign}(\beta_{j,l}) w_{j,l}(\vec{\boldsymbol{x}})\}/t(p,s)$ is a convex combination of $w_{j,l}$. Thus, $\vec{\mathcal{F}}$ is the convex hull of $\mathcal{J}$. By Lemma 2.6.11 in van der Vaart and Wellner (2000), $N\{\epsilon \mathrm{diam}\mathcal{J}, \vec{\mathcal{F}}, L_2(\vec{Q})\} \leq [e + e\{2p(k-1)\}\epsilon^2]^{2/\epsilon^2}$, where $\mathrm{diam}\mathcal{J} = \sup_{J_1, J_2 \in \mathcal{J}} \|J_1 - J_2\|_{\vec{Q},2} \leq 2t(p,s)$. Thus, we conclude that $N\{\epsilon, \bar{H}, L_2(Q)\} =$

$N\{2(k-1)\frac{1-d}{d}t(p,s)\epsilon, \mathcal{G}, L_2(Q)\} \leq N\{2t(p,s)\sqrt{(k-1)^{-1}}\epsilon, \mathcal{J}, L_2(\vec{Q})\} \leq (e+2pe\epsilon^2)^{2(k-1)/\epsilon^2}$. Since the final bound is independent of Q, we have that the bound is uniform for any Q.

■

The next lemma shows that in order to show the result in Theorem 1, we can focus on bounding a tail probability.

Lemma 3. *For given n, p, k, and s, assume that there exists $M > 0$ that satisfies*

$$(\log_2 \frac{16\sqrt{6}\epsilon_0}{M} + 1)^2 \Big\{ \frac{256 \log(e + 2pe\epsilon_0^2)}{n} \Big\} \leq \frac{M^2}{256}, \tag{1}$$

where $\epsilon_0 > 0$ is such that

$$\frac{2(k-1)\log(e + 2pe\epsilon_0^2)}{\epsilon_0^2} = \frac{1}{4} n M^2. \tag{2}$$

Then for $d_{n,p,k} = \inf_{\boldsymbol{f} \in \mathcal{F}(p,k,s)} e_\ell(\boldsymbol{f}, \boldsymbol{f}^{(p,k)})$, we have

$$\mathrm{pr}\{e_\ell\big(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p,k)}\big) \geq 8(k-1)t(p,s)M + d_{n,p,k}\} \leq 6(1 - \frac{1}{16nM^2})^{-1} \exp(-nM^2).$$

Proof of Lemma 3: Define the empirical process $h \to P_n h - Ph$, where $h \in \bar{H}$, $Ph = \int h d\mathbb{P}$ and $P_n h = n^{-1} \sum_{i=1}^n h(y_i)$. We have, by definition of $d_{n,p,k}$,

$$\mathrm{pr}\{e_\ell(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p,k)}) > 8(k-1)t(p,s)M + d_{n,p,k}\} \leq \mathrm{pr}[e_\ell\big(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p,k,s)}\big)\{2(k-1)t(p,s)\}^{-1} > 4M].$$

Since $\hat{\boldsymbol{f}}$ is such that $e_\ell(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p,k,s)})\{2(k-1)t(p,s)\}^{-1} > 4M$, and $\hat{\boldsymbol{f}}$ minimizes the empirical loss $n^{-1}\sum_{i=1}^n \sum_{j \neq y_i} \ell\{\langle \boldsymbol{f}(\boldsymbol{x}_i), \mathcal{Y}_j\rangle\}$, we have $n^{-1}\sum_{i=1}^n\{ \sum_{j\neq y_i} \ell(\langle \boldsymbol{f}^{(p,k,s)}, \mathcal{Y}_j\rangle) - \sum_{j \neq y_i} \ell(\langle \hat{\boldsymbol{f}}, \mathcal{Y}_j\rangle)\} \geq 0$. Hence,

$$\begin{aligned}
&\mathrm{pr}\{e_\ell(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p,k)}) > 8(k-1)t(p,s)M + d_{n,p,k}\} \\
\leq\ & \mathrm{pr}^{outer}\Big[\sup_{\boldsymbol{f} \in \tilde{\mathcal{F}}(p,k,s): e_\ell(\boldsymbol{f}, \boldsymbol{f}^{(p,k,s)})\{2(k-1)t(p,s)\}^{-1} > 4M} \frac{1}{n}\sum_{i=1}^n [\sum_{j \neq y_i} \{\ell(\langle \boldsymbol{f}^{(p,k,s)}, \mathcal{Y}_j\rangle) - \ell(\langle \boldsymbol{f}, \mathcal{Y}_j\rangle)\}] > 0\Big] \\
\leq\ & pr^{outer}\big[\sup_{\boldsymbol{f} \in \tilde{\mathcal{F}}(p,k,s): e_\ell(\boldsymbol{f}, \boldsymbol{f}^{(p,k,s)})\{2(k-1)t(p,s)\}^{-1} > 4M} -\frac{1}{n}\sum_{i=1}^n [h_{\boldsymbol{f}}(y_i) - \mathbb{E}\{h_{\boldsymbol{f}}(Y)\}] \\
&\qquad > \{2(k-1)t(p,s)\}^{-1}\mathbb{E}\{\sum_{j\neq y} \ell(\langle \boldsymbol{f}, \mathcal{Y}_j\rangle) - \sum_{j \neq y} \ell(\langle \boldsymbol{f}^{(p,k,s)}, \mathcal{Y}_j\rangle)\}\big].
\end{aligned}$$

Here pr^{outer} is the outer probability. In the region $\boldsymbol{f} \in \tilde{\mathcal{F}}(p,k,s) : e_\ell(\boldsymbol{f}, \boldsymbol{f}^{(p,k,s)})\{2(k-1)t(p,s)\}^{-1} > 4M$, $\{2(k-1)t(p,s)\}^{-1}\mathbb{E}\{\sum_{j\neq y} \ell(\langle \boldsymbol{f}, \mathcal{Y}_j\rangle) - \sum_{j\neq y} \ell(\boldsymbol{f}^{(p,k,s)}, \mathcal{Y}_j)\}$ is always

larger than $4M$. Hence we have

$$\mathrm{pr}\{e_\ell(\hat{\boldsymbol{f}}, \boldsymbol{f}^{(p,k)}) > 8(k-1)t(p,s)M + d_{n,p,k}\} \leq \mathrm{pr}^{outer}(\sup_{h\in\bar{H}} |P_n h - Ph| > 4M).$$

The rest part of the proof is to bound the tail probability $\sup_{h\in\bar{H}} |P_n h - Ph| > 4M$. Notice that the entropy of $\bar{H}$ is given in Lemma 2, and the entropy is of the order ϵ^{-2}. Thus, by (1), (2), and Theorem A.2 in Wang and Shen (2007), we have $\mathrm{pr}^{outer}(\sup_{h\in\bar{H}} |P_n h - Ph| > 4M) \leq 6\{1-(1/16nM^2)\}^{-1}\exp(-nM^2)$, and this completes the proof. ■

With Lemma 3 proved, we can proceed to prove Theorem 1.

Let $M = 5r\log(r^{-1})$. We need to verify that (1) holds for the choice of M and ϵ_0 in (2). First, note that ϵ_0 goes to 0. Because if ϵ_0 is bounded away from 0 and ∞, the left hand side of (2) is of order $O(\log p)$, and the right hand side of (2) is of order $O\{\log p \log^2(r)^{-1}\}$, which is a contradiction. If $\epsilon_0 \to \infty$, the left hand side of (2) is of order $o(\log p)$, which is still a contradiction. Next, note that (1) is equivalent to

$$(\log_2 \frac{16\sqrt{6}\epsilon_0}{M} + 1)^2 \leq \frac{nM^2}{2^{16}\log(e + 2pe\epsilon_0^2)}. \tag{3}$$

We have $\log_2(16\sqrt{6}\epsilon_0)/M + 1 \propto \log_2(\epsilon_0/M) \preceq \log_2(1/M) \preceq \log(1/r)$, where $\propto$ means "equivalent up to a constant", and $\preceq$ means "less than or equal to up to a constant". As a result, the left hand side of (3) has an order no greater than $O\{\log^2(r^{-1})\}$. For the right hand side of (3), we have $\epsilon_0^{-2} \preceq (nM^2)/\{2^{16}\log(e + 2pe\epsilon_0^2)\}$. The left hand side of (2) has order $O\{\log p \log^2(r)^{-1}\}$. If the order of $1/\epsilon_0$ is less than that of $\log(1/r)$, we have the order of the right hand side of (2) smaller than $O\{\log p \log^2(1/r)\}$, because ϵ_0 goes to 0. Thus, (1) is valid, because the order of left hand side of (3) is less than that of the right hand side.

Finally, note that $nM^2 = 25\log p \log^2(1/r) > 2.5\log p \log(1/r) \geq 2.5\log n > 2\log n$. Hence, we have $\exp(-nM^2) \leq \exp(-2\log n) = n^{-2}$. The desired result in Theorem 1 then follows from the Borel-Cantelli Lemma. □

Proof of Theorem 2: The key to the proof is to show that with any kernel function such that $K(\cdot,\cdot) \leq \infty$, the corresponding entropy number of the function space is approximately in the order ϵ^{-2}.

Let $t(p,s) = s$ for kernel learning. Define $\boldsymbol{f}^{(p,k,s)}$, $h_{\boldsymbol{f}}(\cdot)$, and $\bar{H}$ in a similar manner with respect to the linear learning case in the proof of Theorem 1. Here without loss of generality, assume that the kernel function is upper bounded by 1. Note that the theory

can be naturally generalized to other cases with different upper bounds. Now, with the assumption that the kernel is separable, one can verify that the L_2 diameter of $\bar{H}$ can be bounded by 1. Next, instead of bounding the uniform entropy as in the linear case, we bound the empirical uniform entropy for kernel learning. In particular, let T_X be the empirical measure of a training data set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$, and let the L_2 norm be defined as $\|f\|_{L_2(T_X)} = \left(\frac{1}{m}\sum_{i=1}^m |f(\boldsymbol{x}_i, y_i)|^2\right)^{1/2}$. We can define the $L_2(T_X)$ covering number and entropy number in an obvious manner. In kernel learning, let $H(\epsilon, \bar{H})$ be $\sup_{T_X} H(\epsilon, \bar{H}, L_2(T_X))$, which we call the empirical uniform entropy. Next, we bound $H(\epsilon, \bar{H})$. Notice that C is a constant that may change in different context.

Lemma 4. *For any $\epsilon > 0$, $H(\epsilon, \bar{H}) \leq C\epsilon^{-2}\log(\frac{1}{\epsilon})$.*

Proof of Lemma 4: Let $\mathcal{G} := \{\sum_{j\neq\cdot} \ell(\langle \boldsymbol{f}, \mathcal{Y}_j\rangle) : \sum_{j=1}^{k-1} J(\boldsymbol{f}) \leq s\}$. Let g and g' be defined as in the proof of Lemma 2. One can verify that

$$\begin{aligned}
\|g - g'\|^2_{L_2(T_X)} &= \mathbb{E}[\sum_{j\neq\cdot} \ell\{\langle \mathcal{Y}_j, \boldsymbol{f}(\boldsymbol{X})\rangle\} - \sum_{j\neq\cdot} \ell\{\langle \mathcal{Y}_j, \boldsymbol{f}'(\boldsymbol{X})\rangle\}]^2 \\
&\leq \mathbb{E}\{\frac{(k-1)(1-d)}{d}\sum_{j\neq\cdot}\langle \mathcal{Y}_j, \boldsymbol{f}(\boldsymbol{X}) - \boldsymbol{f}'(\boldsymbol{X})\rangle\}^2 \\
&\leq \frac{(k-1)^2(1-d)^2}{d^2}\mathbb{E}\{\sum_{j=1}^{k-1}|f_j(\boldsymbol{X}) - f'_j(\boldsymbol{X})|\}^2.
\end{aligned}$$

Hence, the $L_2(T_X)$ covering number of $\mathcal{G}$ can be upper bounded through bounding the $L_2(T'_X)$ covering number of $\mathcal{G}'$, which is a set that ranges over all individual classification functions whose norm is upper bounded by s. Here T'_X is the empirical measure of $\{\delta_1\boldsymbol{X}, \delta_2\boldsymbol{X}, \ldots, \delta_{k-1}\boldsymbol{X}\}$, where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, and $(\delta_1, \ldots, \delta_{k-1})$ has a joint distribution $\mathrm{pr}\{(\delta_1, \ldots, \delta_{k-1})^T = e_j\} = (k-1)^{-1}$. Next, by similar arguments as in the proof of Lemma 2 in Zhang et al. (2016) , we have $\sup_{T_X} N\big(\epsilon, \mathcal{G}, L_2(T_X)\big) \leq \frac{5\exp(C\epsilon^{-2})}{\epsilon}$. Therefore, the claim in Lemma 4 holds. ■

The rest of the proof is to notice that the order of the entropy number is $\epsilon^{-2}\log(1/\epsilon)$, which is very close to ϵ^{-2}. Hence, one can verify that (1) and (2) hold in general. By similar arguments as in the proof of linear learning, we can prove Theorem 2. □

Proof of Theorem 3: First, notice that for $j \neq 1$, $(1 - P_{(j)})\ell'(\langle \mathcal{Y}_j, \boldsymbol{f}^*\rangle) = a(1 - P_{(1)})$. Hence, as we assume that the probabilities are bounded away from 0, we can conclude that for a fixed loss function, $\langle \mathcal{Y}_{(j)}, \boldsymbol{f}^*\rangle$ is bounded away from ∞ for all $j \neq 1$. Consequently, we have that $\ell''(\langle \mathcal{Y}_{(j)}, \boldsymbol{f}^*\rangle)$ has a lower bound for any $\langle \mathcal{Y}_j, \boldsymbol{f}^*\rangle < 0$. Denote by ζ this lower bound. Next, define $r(\boldsymbol{f}) = \sum_{j=1}^k (1 - P_j)\{\ell(\langle \boldsymbol{f}, \mathcal{Y}_j\rangle) - \ell(\langle \boldsymbol{f}^*, \mathcal{Y}_j\rangle)\}$ for any $\boldsymbol{f}$. For

brevity in expression, we let $\boldsymbol{f}^* \in \mathcal{F}(p,k)$. Notice that if $\boldsymbol{f}^* \notin \mathcal{F}(p,k)$, the proof becomes slightly more complicated in the approximation error term. Hence we have $r(\hat{\boldsymbol{f}}) \geq 0$, and $\nabla r(\boldsymbol{f}) \mid_{\boldsymbol{f}^*} = \mathbf{0}$.

Without loss of generality, assume that η_0 is small enough such that $a\eta_0(k-1) < 1$. Define $\rho(\boldsymbol{f}, \boldsymbol{f}^*) = \frac{\zeta}{2ak} \max\left(1, \frac{(k-1)\eta_0}{1/a-(k-1)\eta_0}\right) \sum_{j=1}^k (\langle \boldsymbol{f}, \mathcal{Y}_j\rangle - \langle \boldsymbol{f}^*, \mathcal{Y}_j\rangle)^2$. For $a(1-P_{(1)}) > (1-P_{(k)})$ and $\boldsymbol{f}$ close to $\boldsymbol{f}^*$, we have by Taylor's expansion, $r(\boldsymbol{f}) \geq \{(1-P_{(1)}) - \frac{1}{a}(1-P_{(k)})\} \sum_{j=1}^k \frac{\ell''(\langle \boldsymbol{f}^*, \mathcal{Y}_j\rangle)}{2} (\langle \boldsymbol{f}, \mathcal{Y}_j\rangle - \langle \boldsymbol{f}^*, \mathcal{Y}_j\rangle)^2$. Notice that $\sum_{j=1}^k \langle \boldsymbol{f}, \mathcal{Y}_j\rangle = \sum_{j=1}^k \langle \boldsymbol{f}^*, \mathcal{Y}_j\rangle = 0$, and we can conclude that $r(\boldsymbol{f}) \geq \frac{\zeta}{2k}\{(1-P_{(1)}) - \frac{1}{a}(1-P_{(k)})\} \sum_{j=1}^k (\langle \boldsymbol{f}, \mathcal{Y}_j\rangle - \langle \boldsymbol{f}^*, \mathcal{Y}_j\rangle)^2 \geq |a(1-P_{(1)}) - (1-P_{(k)})| \rho(\boldsymbol{f}, \boldsymbol{f}^*)$. On the other hand, if $a(1-P_{(1)}) < (1-P_{(k)})$, one can verify that $\frac{1}{a}(1-P_{(k)}) - (1-P_{(1)}) \leq \frac{1}{a} - (k-1)\eta_0$. Hence, by similar argument as above, we have $r(\boldsymbol{f}) \geq \frac{\zeta}{2k}\{\frac{1}{a}(1-P_{(k)}) - (1-P_{(1)})\} \frac{(k-1)\eta_0}{1/a-(k-1)\eta_0} \geq |a(1-P_{(1)}) - (1-P_{(k)})| \rho(\boldsymbol{f}, \boldsymbol{f}^*)$.

Next, define $g_{\boldsymbol{f}}(\boldsymbol{x}, y) = \sum_{j \neq y} \ell\{\langle \boldsymbol{f}, \mathcal{Y}_j\rangle\} - \sum_{j \neq y} \ell\{\langle \boldsymbol{f}^*, \mathcal{Y}_j\rangle\}$. We prove that $\mathbb{P}g^2 \leq B(\mathbb{P}g)^{\alpha/(1+\alpha)}$ for a constant B that does not depend on n. To this end, notice that for any $\boldsymbol{f}$,

$$
\begin{aligned}
\mathbb{E}\{g_{\boldsymbol{f}}(\boldsymbol{x}, y)\} &= \mathbb{E}\{r(\boldsymbol{f})\} \\
&\geq \mathbb{E}\rho(\boldsymbol{f}, \boldsymbol{f}^*) |a(1-P_{(1)}) - (1-P_{(k)})| \\
&\geq t\mathbb{E}\{\rho(\boldsymbol{f}, \boldsymbol{f}^*)\} I_{|a(1-P_{(1)})-(1-P_{(k)})| \geq t} \\
&= t[\mathbb{E}\{\rho(\boldsymbol{f}, \boldsymbol{f}^*)\} - \mathbb{E}\{\rho(\boldsymbol{f}, \boldsymbol{f}^*) I_{|a(1-P_{(1)})-(1-P_{(k)})| < t}\}] \\
&\geq t[\mathbb{E}\{\rho(\boldsymbol{f}, \boldsymbol{f}^*)\} - C_1(s)t^\alpha],
\end{aligned}
$$

where $C_1(s)$ is a linear function of s, such that $C_1(s) \geq \rho(\boldsymbol{f}, \boldsymbol{f}^*)$ for all $\boldsymbol{f}$. Choose $t = \left[\frac{\mathbb{E}\{\rho(\boldsymbol{f}, \boldsymbol{f}^*)\}}{2C_1(s)}\right]^{1/\alpha}$, and we have $\mathbb{E}\{\rho(\boldsymbol{f}, \boldsymbol{f}^*)\} \leq C_2(s)[\mathbb{E}\{g_{\boldsymbol{f}}(\boldsymbol{x}, y)\}]^{\alpha/(1+\alpha)}$, where $C_2(s)$ is another linear function of s.

On the other hand, notice that

$$
\begin{aligned}
\mathbb{E}\{g_{\boldsymbol{f}}(\boldsymbol{x}, y)\}^2 &= \mathbb{E}[\mathbb{E}\{g_{\boldsymbol{f}}(\boldsymbol{x}, y)\}^2 | \boldsymbol{x}] \\
&\leq C_3 \mathbb{E}\{\rho(\boldsymbol{f}, \boldsymbol{f}^*)\},
\end{aligned}
$$

where C_3 is a universal constant. Hence, combining the above inequalities to obtain that

$$
\mathbb{E}\{g_{\boldsymbol{f}}(\boldsymbol{x}, y)\}^2 \leq C_4(s) \mathbb{E}\{g_{\boldsymbol{f}}(\boldsymbol{x}, y)\}^{\alpha/(1+\alpha)},
$$

where $C_4(s)$ is a linear function of s.

Next, let $\mathbb{P}_n g_{\boldsymbol{f}} = \frac{1}{n}\sum_{i=1}^n g_{\boldsymbol{f}}(\boldsymbol{x}_i, y_i)$, and $\mathbb{P} g_{\boldsymbol{f}} = \mathbb{E}_{\mathbb{P}(\boldsymbol{x},y)} g_{\boldsymbol{f}}$. We have

$$\begin{aligned}
e_\ell(\hat{\boldsymbol{f}}, \boldsymbol{f}^*) &= \mathbb{E}\{2\mathbb{P}_n g_{\hat{\boldsymbol{f}}} + (\mathbb{P} - 2\mathbb{P}_n) g_{\hat{\boldsymbol{f}}}\} \\
&\leq 2\mathbb{E}\{\inf_{\boldsymbol{f}\in\mathcal{F}(p,k,s)} 2\mathbb{P}_n g_{\boldsymbol{f}} + \sup_{\boldsymbol{f}\in\mathcal{F}(p,k,s)} (\mathbb{P} - 2\mathbb{P}_n) g_{\boldsymbol{f}}\} \\
&\leq 2 \inf_{\boldsymbol{f}\in\mathcal{F}(p,k,s)} \mathbb{E}(\mathbb{P}_n g_{\boldsymbol{f}}) + \mathbb{E}\{\sup_{\boldsymbol{f}\in\mathcal{F}(p,k,s)} (\mathbb{P} - 2\mathbb{P}_n) g_{\boldsymbol{f}}\} \\
&\leq 2d_{n,p,k} + 2(k-1)\left[\epsilon_n + \mathrm{pr}\{\sup_{\boldsymbol{f}\in\mathcal{F}_n(p,k,s)} (\mathbb{P} - 2\mathbb{P}_n) g_{\boldsymbol{f}} \geq \epsilon_n\}\right],
\end{aligned}$$

where $\mathcal{F}_n(p,k,s)$ is the space of functions that corresponds to an ϵ_n-net of $\bar{H}$. Furthermore, because the entropy number of $\mathcal{F}_n(p,k,s)$ is the same as that of $\bar{H}$, and is of order $o(\epsilon_n^{-\delta})$ for any $\delta > 0$ (Zhou, 2002), we have, by Bernstein's Inequality,

$$\begin{aligned}
\mathrm{pr}\{\sup_{\boldsymbol{f}\in\mathcal{F}_n(p,k,s)} (\mathbb{P} - 2\mathbb{P}_n) g_{\boldsymbol{f}} \geq \epsilon_n\} &\leq \sum_{\boldsymbol{f}\in\mathcal{F}_n(p,k,s)} \mathrm{pr}\{(\mathbb{P} - \mathbb{P}_n) g_{\boldsymbol{f}} \geq \frac{1}{2}(\mathbb{P} g_{\boldsymbol{f}} + \epsilon_n)\} \\
&\leq |\mathcal{F}_n(p,k,s)| \max_{\boldsymbol{f}\in\mathcal{F}_n(p,k,s)} \exp\left\{-\frac{n}{8}\frac{(\mathbb{P} g_{\boldsymbol{f}} + \epsilon_n)^2}{\mathbb{P} g_{\boldsymbol{f}}^2 + C_1(s)(\mathbb{P} g_{\boldsymbol{f}} + \epsilon_n)/6}\right\} \\
&\leq \exp(C_5 \epsilon_n^{-\delta} - C_6(s) n \epsilon_n^{2-\alpha/(1+\alpha)}),
\end{aligned}$$

where C_5 is a universal constant, and $C_6(s)$ is a linear function of s.

Let $\epsilon_n = M(s) n^{-(1+\alpha)/(2+\alpha)}$, where $M(s)$ is a linear function of s. We choose $M(s)$ such that $C_5 \epsilon_n^{-\delta} = \frac{1}{2} C_6(s) n \epsilon_n^{2-\alpha/(1+\alpha)}$ and $\exp(-n\epsilon_n^{2-\alpha/(1+\alpha)}) = o(\epsilon_n)$. We then have $e_\ell(\hat{\boldsymbol{f}}, \boldsymbol{f}^*) \leq 2d_{n,p,k} + 2C_7(k-1)s n^{-(1+\alpha)/(2+\alpha)}$ for a universal constant C_7. This completes the proof.

For binary loss functions that are flat for large enough u, one can verify that the largest $|\langle \boldsymbol{f}^*, \mathcal{Y}_j\rangle|$ is bounded. In that case, it is possible to obtain similar results by defining $\rho(\boldsymbol{f}, \boldsymbol{f}^*)$ to be $C\sum_{j=1}^k |\langle \boldsymbol{f}, \mathcal{Y}_j\rangle - \langle \boldsymbol{f}^*, \mathcal{Y}_j\rangle|$, where C can be chosen by careful analysis of the loss function. See Bartlett and Wegkamp (2008) for an example in the binary case with SVM as the loss function. □

Proof of Corollary 3: The proof is immediate from Lemma 2 in Wang and Shen (2007) and Theorems 1-3. □

Next, we consider two examples in which the choices of γ_1 and γ_2 differ. In particular, in the first example, the classification problem is nearly separable. In this case, we show that γ_2 can be arbitrarily large. Hence, the convergence rate of the excess risk can be very fast. In the second example, the classes are distributed with weak classification signals. Consequently, we show that $\gamma_1 = 2$ and $\gamma_2 = 1$. Therefore, the convergence rate of the

excess risk is slower than that in the first example.

For brevity, we focus on linear learning, and let ℓ_1 be the reversed DWD loss. We construct two 3-class examples, where the true signal depends only on two predictors $X^{(1)}$ and $X^{(2)}$. The remaining predictors are continuous and uniformly bounded. Assume that the prior proportion of the three classes is the same. Denote by Z_j the line segment between the origin and $\mathcal{Y}_j$; $j = 1, 2, 3$.

Illustrating Example 1: Let the marginal density of $\boldsymbol{X}$ be non-zero only on Z_j's. In particular, for any point $\boldsymbol{x} = (x^{(1)}, x^{(2)})$ in $\mathbb{R}^2$, define $t_j(\boldsymbol{x}) = \langle \mathcal{Y}_j, (x^{(1)}, x^{(2)})^T \rangle$ for $j = 1, 2, 3$. For each observation, suppose that with probability θ, the first two predictor values are $(0, 0)$, and with probability $1 - \theta$, the marginal pdf of $(X^{(1)}, X^{(2)}) \mid Y = j$ be such that $\text{pr}\{t_j(\boldsymbol{x}) \in [a, b]\} \propto \int_a^b t^\beta dt$ for $t_j \in [0, 1]$, where $\beta > 0$. Suppose the cost for rejection d is such that $d < 2/3$. It can be verified that for any β, the best parameters $\{\boldsymbol{\beta}_j,\ j = 1, \ldots, k-1\}$ are uniformly bounded.

Next, we explore the behavior of the linear learning parameters. Define $B = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1})$ to be the p by $k-1$ matrix that contains the $k-1$ parameter vectors. Define $B^{(p,k)}$ in a similar manner as $\boldsymbol{f}^{(p,k)}$. Because p can go to infinity, we study the behavior of B in a neighbourhood of $B^{(p,k)}$ under the uniform metric, $d(B_1, B_2) = \sup_{i,j}\{(B_1)_{ij} - (B_2)_{ij}\}$, where B_{ij} is the (i, j)th element of B. We have the following result for $d(B, B^{(p,k)})$.

Proposition 1. *Let ℓ_1 be the reversed DWD loss. There exists a β and a constant C_1, such that $e_\ell(\boldsymbol{f}, \boldsymbol{f}^{(p,k)}) \geq C_1 d(B, B^{(p,k)})$.*

Proof of Proposition 1: The proof is analogous to that of Theorem 5 in Zhang and Liu (2014). Notice that we can choose β such that the theoretical minimizer $\boldsymbol{f}^*$ does not belong to any $\mathcal{F}(p, k)$. Furthermore, because here we assume that the marginal distribution of the predictors is continuous for $\|\boldsymbol{X}\| > 0$, the assumptions in Theorem 5 of Zhang and Liu (2014) are automatically valid (despite that the loss function we consider in this paper is not differentiable at 0, the corresponding probability is 0 because of the continuous distribution assumption in Illustrating Example 1, hence the proof does not change). ■

As $X^{(1)}$ and $X^{(2)}$ are bounded, one can further verify that $\Delta(\boldsymbol{f}, \boldsymbol{f}^{(p,k)}) \preceq d(B, B^{(p,k)})$. Thus, we have that $\Delta(\boldsymbol{f}, \boldsymbol{f}^{(p,k)}) \preceq e_\ell(\boldsymbol{f}, \boldsymbol{f}^{(p,k)})$ in a small neighbourhood of $\boldsymbol{f}^{(p,k)}$. Consequently, we can choose $\gamma_1 = 1$ in this example.

Next, we calculate γ_2. Without loss of generality, we can restrict our discussion in $\boldsymbol{f} \in \mathcal{F}(2) := \{\boldsymbol{f} : \boldsymbol{f}(\boldsymbol{x}) = (x^{(1)}\beta_1^{(1)} + x^{(2)}\beta_1^{(2)}, x^{(1)}\beta_2^{(1)} + x^{(2)}\beta_2^{(2)})^T\}$. This is because

$X^{(3)}, \ldots, X^{(p_n-1)}$ are irrelevant to the classification problem. The data projected by the classification function vector in $\mathbb{R}^2$ have positive support only on Z_1, Z_2 and Z_3 for $\boldsymbol{f}^{(p,k)}$.

Define $W(\beta_1^{(1)}, \beta_1^{(2)}, \beta_2^{(1)}, \beta_2^{(2)}) = E_{\boldsymbol{f}(x^{(1)}\beta_1^{(1)}+x^{(2)}\beta_1^{(2)}, x^{(1)}\beta_2^{(1)}+x^{(2)}\beta_2^{(2)})^T}\{L(\boldsymbol{f}, Y)\}$, where L is the 0-d-1 loss. For any $(w_1^{(1)}, w_1^{(2)}, w_2^{(1)}, w_2^{(2)}) \in \mathbb{R}^4$, define

$$\Delta W := W(\beta_1^{(1)} + w_1^{(1)}, \beta_1^{(2)} + w_1^{(2)}, \beta_2^{(1)} + w_2^{(1)}, \beta_2^{(2)} + w_2^{(2)}) - W(\beta_1^{(1)}, \beta_1^{(2)}, \beta_2^{(1)}, \beta_2^{(2)}).$$

When the norm of $(w_1^{(1)}, w_1^{(2)}, w_2^{(1)}, w_2^{(2)})$ is small, one can verify that ΔW is upper bounded by, up to a constant, $\{\sup(w_1^{(1)}, w_1^{(2)}, w_2^{(1)}, w_2^{(2)})\}^{(\beta+1)/2}$. Notice that $\sup(|w_1^{(1)}|, |w_1^{(2)}|, |w_2^{(1)}|, |w_2^{(2)}|)$ is a norm defined on $(w_1^{(1)}, w_1^{(2)}, w_2^{(1)}, w_2^{(2)}) \in \mathbb{R}^4$. On the other hand, $\Delta(\boldsymbol{f}, \boldsymbol{f}^*) = \left[E\{f_1(\boldsymbol{x}) - f_1^{(p,k)}(\boldsymbol{x})\}^2 + E\{f_2(\boldsymbol{x}) - f_2^{(p,k)}(\boldsymbol{x})\}^2\right]^{1/2} \succeq (|w_1^{(1)}| + |w_1^{(2)}| + |w_2^{(1)}| + |w_2^{(2)}|)$ as $\boldsymbol{X}$ is bounded. Since $|w_1^{(1)}| + |w_1^{(2)}| + |w_2^{(1)}| + |w_2^{(2)}|$ is also a norm on $(w_1^{(1)}, w_1^{(2)}, w_2^{(1)}, w_2^{(2)}) \in \mathbb{R}^4$, we have that $\sup(|w_1^{(1)}|, |w_1^{(2)}|, |w_2^{(1)}|, |w_2^{(2)}|) \preceq \left(|w_1^{(1)}| + |w_1^{(2)}| + |w_2^{(1)}| + |w_2^{(2)}|\right)$ for all $(w_1^{(1)}, w_1^{(2)}, w_2^{(1)}, w_2^{(2)}) \in \mathbb{R}^4$. Therefore, we may choose $\gamma_2 = (\beta + 1)/2$. Consequently, we obtain $\gamma_2/\gamma_1 = (\beta + 1)/2$ in Corollary 3. When $\beta \to \infty$, the classification signal in this example becomes stronger, and the convergence rate of the excess risk can be arbitrarily fast. Note that a similar result on the excess risk for binary classification was provided in Wang and Shen (2007). □

Illustrating Example 2: This example is constructed similarly as the first example, whereas the difference is on the marginal distribution of $(X^{(1)}, X^{(2)})$. In particular, define $\ell'(u)$ to be the derivative of our loss function in (3) with ℓ_1 the reversed DWD loss for $u \neq 0$, and at $u = 0$, define the derivative to be 1. We still let a proportion θ of the observations remain at the origin, and for the rest $1 - \theta$ observations, let $\text{pr}\{t_j(\boldsymbol{x}) \in [a, b]\} \propto \int_a^b \ell'(|t|)dt$ for $t_j \in [-1/2, 1]$. Assume that when $t_j < 0$, the marginal distribution of $(X^{(1)}, X^{(2)}) \mid Y = j$ is on Z_i, $i \neq j$ and is symmetric with respect to the vector Z_j. In this case, the classification signal is much weaker than the first example. One can verify that $\boldsymbol{f}^* = \boldsymbol{f}^{(p,k)}$. We then have the following result for the relationship between the excess ℓ-risk and $d(B, B^{(p,k)})$.

Proposition 2. *There exists a β and a constant C_2, such that $e_\ell(\boldsymbol{f}, \boldsymbol{f}^{(p,k)}) \geq C_2 d(B, B^{(p,k)})^2$.*

The proof of Proposition 2 is analogous to that of Proposition 1, thus is omitted.

By similar arguments as in the first illustrating example, we can choose $\gamma_1 = 2$ and $\gamma_2 = 1$. Consequently, the convergence rate of the excess risk is much slower than that of the first example. □

Proof of Theorem 4: To prove this theorem, we need to introduce a recent technique in the statistical machine learning literature, namely, the Rademacher complexity (Bartlett and Mendelson, 2002; Koltchinskii and Panchenko, 2002; Shawe-Taylor and Cristianini, 2004; Bartlett et al., 2005; Koltchinskii, 2006; Mohri et al., 2012). Recall the definition of $\mathcal{F}(p,k,s)$ from the main paper. Let $\sigma = \{\sigma_i;\ i = 1,\ldots,n\}$ be independent and identically distributed random variables, that take 1 and -1 with probability $1/2$ each. Denote by S a sample of observations $(\boldsymbol{x}_i, y_i);\ i = 1,\ldots,n$, independent and identically distributed from the underlying distribution $\mathbb{P}(\boldsymbol{X}, Y)$. Consider the continuous indicator function $I_{®,\kappa}\{\boldsymbol{f}(\boldsymbol{x}), y\} = I_{®,\kappa}(\boldsymbol{x})$ with fixed κ. Given S, we define the empirical Rademacher complexity of the function class $\mathcal{F}(p,k,s)$ to be

$$\hat{R}_n\{\mathcal{F}(p,k,s)\} = E_\sigma[\sup_{\boldsymbol{f}\in\mathcal{F}(p,k,s)} \frac{1}{n}\sum_{i=1}^{n} \sigma_i I_{®,\kappa}\{\mathcal{Y}_{y_i}, \boldsymbol{f}(\boldsymbol{x}_i)\}].$$

Here E_σ means taking expectation with respect to the joint distribution of σ. Furthermore, define the Rademacher complexity of $\mathcal{F}(p,k,s)$ to be

$$R_n\{\mathcal{F}(p,k,s)\} = E_{\sigma,S}[\sup_{\boldsymbol{f}\in\mathcal{F}(p,k,s)} \frac{1}{n}\sum_{i=1}^{n} \sigma_i I_{®,\kappa}\{\mathcal{Y}_{y_i}, \boldsymbol{f}(\boldsymbol{x}_i)\}].$$

The proof of Theorem 4 consists of two major steps. In the first step, we show that with probability at least $1-\zeta$ $(0 < \zeta < 1)$, $E\{I_{®,\kappa}(\mathcal{Y}_Y, \hat{\boldsymbol{f}})\}$ is bounded by the summation of its empirical evaluation, the Rademacher complexity of the function class $\mathcal{F}(p,k,s)$, and a penalty term on ζ. In particular, we have the following lemma.

Lemma 5. *Let $R_n\{\mathcal{F}(p,k,s)\}$ and $\hat{R}_n\{\mathcal{F}(p,k,s)\}$ be defined as above. Then with probability at least $1-\zeta$,*

$$E[I_{®,\kappa}\{\mathcal{Y}_Y, \boldsymbol{f}(\boldsymbol{X})\}] \leq \frac{1}{n}\sum_{i=1}^{n} I_{®,\kappa}\{\mathcal{Y}_{y_i}, \boldsymbol{f}(\boldsymbol{x}_i)\} + 2R_n\{\mathcal{F}(p,k,s)\} + T_n(\zeta), \tag{4}$$

where $T_n(\zeta) = \{\log(1/\zeta)/n\}^{1/2}$.

Moreover, with probability at least $1-\zeta$,

$$E[I_{®,\kappa}\{\mathcal{Y}_Y, \boldsymbol{f}(\boldsymbol{X})\}] \leq \frac{1}{n}\sum_{i=1}^{n} I_{®,\kappa}\{\mathcal{Y}_{y_i}, \boldsymbol{f}(\boldsymbol{x}_i)\} + 2\hat{R}_n\{\mathcal{F}(p,k,s)\} + 3T_n(\zeta/2).$$

Proof of Lemma 5: The proof consists of three parts. In the first part, we use the McDiarmid inequality (McDiarmid, 1989) to bound the left hand side of (4) in terms

of its empirical estimation, plus the expectation of their supremum difference, $E(\phi)$, where ϕ is to be defined. In the second part, we show that $E(\phi)$ is bounded by the Rademacher complexity using symmetrization inequalities (van der Vaart and Wellner, 2000). In the third part, we provide a bound on the Rademacher complexity using the empirical Rademacher complexity.

For a given sample S, we define

$$\phi(S) = \sup_{\boldsymbol{f} \in \mathcal{F}(p,k,s)} \left(E[I_{\text{®},\kappa}\{\mathcal{Y}_Y, \boldsymbol{f}(\boldsymbol{X})\}] - \frac{1}{n}\sum_{i=1}^{n} I_{\text{®},\kappa}\{\mathcal{Y}_{y_i}, \boldsymbol{f}(\boldsymbol{x}_i)\}] \right).$$

Let $S^{(i,\boldsymbol{x})} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_i', y_i), \ldots, (\boldsymbol{x}_n, y_n)\}$ be another sample from $\mathbb{P}(\boldsymbol{X}, Y)$, where the difference between S and $S^{(i,\boldsymbol{x})}$ is only on the $\boldsymbol{x}$ value of their ith pair. By definition, we have

$$\begin{aligned} |\phi(S) - \phi(S^{(i,\boldsymbol{x})})| = | &\sup_{\boldsymbol{f} \in \mathcal{F}(p,k,s)} \left(E[I_{\text{®},\kappa}\{\mathcal{Y}_Y, \boldsymbol{f}(\boldsymbol{X})\}] - \frac{1}{n}\sum_{S} I_{\text{®},\kappa}\{\mathcal{Y}_{y_i}, \boldsymbol{f}(\boldsymbol{x}_i)\} \right) \\ &- \sup_{\boldsymbol{f} \in \mathcal{F}(p,k,s)} \left(E[I_{\text{®},\kappa}\{\mathcal{Y}_Y, \boldsymbol{f}(\boldsymbol{X})\}] - \frac{1}{n}\sum_{S^{(i,\boldsymbol{x})}} I_{\text{®},\kappa}\{\mathcal{Y}_{y_i}, \boldsymbol{f}(\boldsymbol{x}_i)\} \right) |. \end{aligned}$$

For simplicity, assume that $\boldsymbol{f}^S$ is the function that achieves the supremum of $\phi(S)$. Note that the case of no function reaching the supremum can be treated analogously, with only some additional discussions on the arbitrarily small difference between $\phi(\boldsymbol{f})$ and its supremum. Thus, we omit the details here. We have that,

$$\begin{aligned} |\phi(S) - \phi(S^{(i,\boldsymbol{x})})| \leq & |E[I_{\text{®},\kappa}\{\mathcal{Y}_Y, \boldsymbol{f}^S(\boldsymbol{X})\}] - \frac{1}{n}\sum_{S} I_{\text{®},\kappa}\{\mathcal{Y}_{y_i}, \boldsymbol{f}^S(\boldsymbol{x}_i)\} \\ & - E[I_{\text{®},\kappa}\{\mathcal{Y}_Y, \boldsymbol{f}^S(\boldsymbol{X})\}] + \frac{1}{n}\sum_{S^{(i,\boldsymbol{x})}} I_{\text{®},\kappa}\{\mathcal{Y}_{y_i}, \boldsymbol{f}^S(\boldsymbol{x}_i)\}|. \\ = & \frac{1}{n}| \sum_{S} I_{\text{®},\kappa}\{\mathcal{Y}_{y_i}, \boldsymbol{f}^S(\boldsymbol{x}_i)\} - \sum_{S^{(i,\boldsymbol{x})}} I_{\text{®},\kappa}\{\mathcal{Y}_{y_i}, \boldsymbol{f}^S(\boldsymbol{x}_i)\}| \\ \leq & \frac{1}{n}. \end{aligned}$$

Next, by the McDiarmid inequality, we have that for any $t > 0$, $\text{pr}[\phi(S) - E\{\phi(S)\} \geq t] \leq \exp[-(2t^2)/\{2n(1/n)^2\}]$, or equivalently, with probability at least $1-\zeta$, $\phi(S) - E\{\phi(S)\} \leq T_n(\zeta)$. Consequently, we have that with probability at least $1-\zeta$, $E\{I_{\text{®},\kappa}(\mathcal{Y}_Y, \hat{\boldsymbol{f}})\} \leq n^{-1}\sum_{i=1}^{n} I_{\text{®},\kappa}\{\mathcal{Y}_{y_i}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i)\} + E\{\phi(S)\} + T_n(\zeta)$. This completes the first part of the proof.

In the second part, we bound $E\{\phi(S)\}$ by the Rademacher complexity. To this end, define $S' = \{(\boldsymbol{x}_i', y_i');\ i = 1, \ldots, n\}$ as an independent, duplicate sample of size n with

identical distribution as S. Denote by E_S the action of taking expectation with respect to the distribution of S, and define $E_{S'}$ analogously. By definition, we have that $E_{S'}\big[n^{-1}\sum_{S'} I_{\circledR,\kappa}\{\mathcal{Y}_{y_i'}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i')\} \mid S\big] = E[I_{\circledR,\kappa}\{\mathcal{Y}_Y, \hat{\boldsymbol{f}}(\boldsymbol{x})\}]$, and $E_{S'}\big[n^{-1}\sum_{S} I_{\circledR,\kappa}\{\mathcal{Y}_{y_i}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i)\} \mid S\big] = n^{-1}\sum_{S} I_{\circledR,\kappa}\{\mathcal{Y}_{y_i}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i)\}$. Then, by Jensen's inequality and the property of σ, we have that

$$
\begin{aligned}
E\{\phi(S)\} &= E_S\big(\sup_{\boldsymbol{f}\in\mathcal{F}(p,k,s)} E_{S'}[\frac{1}{n}\sum_{S'} I_{\circledR,\kappa}\{\mathcal{Y}_{y_i'}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i')\} - \frac{1}{n}\sum_{S} I_{\circledR,\kappa}\{\mathcal{Y}_{y_i}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i)\}] \mid S\big) \\
&\le E_{S,S'}\big[\sup_{\boldsymbol{f}\in\mathcal{F}(p,k,s)} \frac{1}{n}\sum_{S'} I_{\circledR,\kappa}\{\mathcal{Y}_{y_i'}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i')\} - \frac{1}{n}\sum_{S} I_{\circledR,\kappa}\{\mathcal{Y}_{y_i}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i)\}\big] \\
&= E_{S,S',\sigma}\big[\sup_{\boldsymbol{f}\in\mathcal{F}(p,k,s)} \frac{1}{n}\sum_{S'} \sigma_i I_{\circledR,\kappa}\{\mathcal{Y}_{y_i'}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i')\} - \frac{1}{n}\sum_{S} \sigma_i I_{\circledR,\kappa}\{\mathcal{Y}_{y_i}, \hat{\boldsymbol{f}}(\boldsymbol{x}_i)\}\big] \\
&\le 2R_n\{\mathcal{F}(p,k,s)\}.
\end{aligned}
$$

Hence the second part is proved.

In the third step, we need to bound $R_n\{\mathcal{F}(p,k,s)\}$ in terms of $\hat{R}_n\{\mathcal{F}(p,k,s)\}$. This step is analogous to the first part. In particular, one can apply the McDiarmid inequality on $\hat{R}_n\{\mathcal{F}(p,k,s)\}$ and the corresponding expectation $R_n\{\mathcal{F}(p,k,s)\}$. Similar to the first part of this proof, we can show that with probability at least $1-\zeta$, $R_n\{\mathcal{F}(p,k,s)\} \le \hat{R}_n\{\mathcal{F}(p,k,s)\} + 2T_n(\zeta)$. The final results can be obtained by choosing the confidence $1-\zeta/2$ in the first and third steps, and combining the inequalities of the three steps together. ■

The second major step to prove Theorem 4 involves bounding the empirical Rademacher complexities for different learning settings. We have the following lemma.

Lemma 6. *In linear learning, when we use the L_1 penalty, the empirical Rademacher complexity $\hat{R}_n\{\mathcal{F}(p,k,s)\} \le \frac{s}{\kappa}\sqrt{\frac{2\log(2pk-2p)}{n}}$, and when we use the L_2 penalty, $\hat{R}_n\{\mathcal{F}(p,k,s)\} \le \{2(k-1)(ps)^{1/2}\}/(\kappa n^{1/4}) + \{2(ps)^{1/2}\}\Big(\log[e+e\{2p(k-1)\}]/(n^{1/2})\Big)^{1/2}/(\kappa n^{1/4})$. For kernel learning with separable and bounded kernel functions, the empirical Rademacher complexity $\hat{R}_n\{\mathcal{F}(p,k,s)\} \le \frac{s(k-1)}{\kappa\sqrt{n}}$.*

Proof of Lemma 6: For the L_1 penalized learning, note that the Rademacher complexity $\hat{R}_n\{\mathcal{F}(p,k,s)\}$ is upper bounded by the following Rademacher complexity. In particular, by Lemma 4.2 in Mohri et al. (2012), we have that $\hat{R}_n\{\mathcal{F}(p,k,s)\}$ is upper bounded by

$$
\frac{1}{\kappa}\hat{R}_n^*\{\mathcal{F}(p,k,s)\} = \frac{1}{\kappa}E_\sigma\left\{\sup_{\sum_{j=1}^{k-1}\|\boldsymbol{\beta}_j\|_1<s} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\{\sum_{j=1}^{k-1}\boldsymbol{x}_i^T\boldsymbol{\beta}_j\}\right\}, \tag{5}
$$

because the continuous indicator function is Lipschitz with constant $1/\kappa$, and elements in $\mathcal{Y}_j$ are bounded by 1. Without loss of generality, we can rewrite (5) as

$$\frac{1}{\kappa}\hat{R}_n^*\{\mathcal{F}(p,k,s)\} = \frac{1}{\kappa}E_\sigma\left\{\sup_{\|\gamma\|_1<s}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\boldsymbol{\gamma}^T\boldsymbol{x}_i^*\right\},$$

where $\boldsymbol{\gamma}$ can be regarded as a vector that contains all the elements in $\boldsymbol{\beta}_j$ for $j = 1,\ldots,k-1$, and $\boldsymbol{x}_i^*$ is defined accordingly. Next, using Theorem 10.10 in Mohri et al. (2012), we have that $\hat{R}_n^*\{\mathcal{F}(p,k,s)\} \leq s\sqrt{\frac{2\log(2pk-2p)}{n}}$. Thus, $\hat{R}_n\{\mathcal{F}(p,k,s)\} \leq \frac{s}{\kappa}\sqrt{\frac{2\log(2pk-2p)}{n}}$ for L_1 penalized linear learning.

For L_2 penalized learning, the proof is analogous to that of Lemma 8 in Zhang and Liu (2014), and we omit the details here.

For kernel learning, notice that one can include the intercept in the original predictor space (i.e., augment $\boldsymbol{x}$ to include a constant 1 before the other predictors), and define a new kernel function accordingly. This new kernel is also positive definite and separable with bounded kernel function. By Mercer's Theorem, this introduces a new RKHS $\mathcal{H}$. Next, by similar argument as for (5), we have that the original Rademacher complexity is upper bounded by

$$\frac{1}{\kappa}\hat{R}_n^*\{\mathcal{F}(p,k,s)\} = \frac{1}{\kappa}E_\sigma\left\{\sup_{\sum_j\|f_j\|_{\mathcal{H}}^2\leq s}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\{\sum_{j=1}^{k-1}f_j(\boldsymbol{x}_i)\}\right\}, \tag{6}$$

$$\leq \frac{k-1}{\kappa}E_\sigma\left\{\sup_{\|f\|_{\mathcal{H}}^2\leq s}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\boldsymbol{x}_i)\right\}, \tag{7}$$

$$\leq \frac{k-1}{\kappa}\frac{s}{\sqrt{n}}, \tag{8}$$

where the last inequality follows from Theorem 5.5 in Mohri et al. (2012). Hence, we have that for kernel learning, $\hat{R}_n\{\mathcal{F}(p,k,s)\} \leq \frac{s(k-1)}{\kappa\sqrt{n}}$. ■

The proof of Theorem 4 is thus finished by combining Lemmas 5 and 6, and the fact that the continuous indicator function $I_{\circledR,\kappa}$ is an upper bound of the indicator function $I_{\circledR}$ for any κ. □

Proof of Theorem 5: The proof is analogous to that of Theorem 4 and is omitted. □

3 Derivation of Implementations

Derivation of the implementation for linear learning: Introducing Lagrangian variable λ, slack variables ξ_{ij} and η_{ij}, we have that (3) is equivalent to

$$\begin{aligned}
\min_{\boldsymbol{f}\in\mathcal{F}} \quad & \frac{n\lambda}{2}\sum_{q=1}^{k-1}\boldsymbol{\beta}_q^T\boldsymbol{\beta} + \sum_{i=1}^{n}\sum_{j\neq y_i}(\xi_{ij}+\eta_{ij}) \\
subjectto \quad & \xi_{ij} \geq 0, \\
& \eta_{ij} \geq 0, \\
& \xi_{ij} - \langle \boldsymbol{f}(\boldsymbol{x}_i), \mathcal{Y}_j\rangle - 1 \geq 0, \\
& \eta_{ij} - (a-1)\langle \boldsymbol{f}(\boldsymbol{x}_i), \mathcal{Y}_j\rangle \geq 0; \quad i = 1,\ldots,n, \quad j \neq y_i.
\end{aligned}$$

Now define the corresponding Lagrangian function $\mathcal{L}$ as

$$\begin{aligned}
\mathcal{L} = & \frac{n\lambda}{2}\sum_{q=1}^{k-1}\boldsymbol{\beta}_q^T\boldsymbol{\beta} + \sum_{i=1}^{n}\sum_{j\neq y_i}(\xi_{ij}+\eta_{ij}) - \sum_{i=1}^{n}\sum_{j\neq y_i}\tau_{ij}\xi_{ij} - \sum_{i=1}^{n}\sum_{j\neq y_i}\chi_{ij}\eta_{ij} \\
& - \sum_{i=1}^{n}\sum_{j\neq y_i}\alpha_{ij}\{\xi_{ij} - \langle \boldsymbol{f}(\boldsymbol{x}_i), \mathcal{Y}_j\rangle - 1\} - \sum_{i=1}^{n}\sum_{j\neq y_i}\gamma_{ij}\{\eta_{ij} - (a-1)\langle \boldsymbol{f}(\boldsymbol{x}_i), \mathcal{Y}_j\rangle\},
\end{aligned}$$

where α_{ij}, γ_{ij}, τ_{ij}, and χ_{ij}; $i = 1,\ldots,n, j = 1,\ldots,k$ are the Lagrangian multipliers. Define $A_{ij} = I(j \neq y_i)$. Take partial derivative of $\mathcal{L}$ with respect to ξ_{ij}, η_{ij} and $\boldsymbol{\beta}_q$, and we have

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\xi_{ij}} &= A_{ij} - \alpha_{ij} - \tau_{ij} = 0, \\
\frac{\partial\mathcal{L}}{\partial\eta_{ij}} &= A_{ij} - \gamma_{ij} - \chi_{ij} = 0, \\
\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}_q} &= n\lambda\boldsymbol{\beta}_q + \sum_{i=1}^{n}\sum_{j\neq y_i}\alpha_{ij}\mathcal{Y}_{j,q}\boldsymbol{x}_i + \sum_{i=1}^{n}\sum_{j\neq y_i}\gamma_{ij}(a-1)\mathcal{Y}_{j,q}\boldsymbol{x}_i \\
&= n\lambda\boldsymbol{\beta}_q + \sum_{i=1}^{n}\sum_{j\neq y_i}\{\alpha_{ij} + (a-1)\gamma_{ij}\}\mathcal{Y}_{j,q}\boldsymbol{x}_i = \boldsymbol{0},
\end{aligned}$$

where $\mathcal{Y}_{j,q}$ is the q^{th} element of $\mathcal{Y}_j$. Now one can conclude that $0 \leq \alpha_{ij} \leq A_{ij}$, $0 \leq \gamma_{ij} \leq A_{ij}$; $i = 1,\ldots,n$, $j = 1,\ldots,k$, and

$$\boldsymbol{\beta}_q = -\frac{1}{n\lambda}\sum_{i=1}^{n}\sum_{j\neq y_i}\{\alpha_{ij} + (a-1)\gamma_{ij}\}\mathcal{Y}_{j,q}\boldsymbol{x}_i. \tag{9}$$

Plugging $\boldsymbol{\beta}$ in $\mathcal{L}$, one can verify that

$$\mathcal{L} = -\frac{n\lambda}{2}\sum_{q=1}^{k-1}\boldsymbol{\beta}_q^T\boldsymbol{\beta}_q + \sum_{i=1}^{n}\sum_{j\neq y_i}\alpha_{ij}.$$

Derivation of the implementation for kernel learning: Next, we briefly discuss the case of kernel learning. Let the kernel function be $K(\cdot,\cdot)$, and the corresponding gram matrix be $\boldsymbol{K} = \Big(K(\boldsymbol{x}_i, \boldsymbol{x}_{i'})\Big)_{i,i'}$. Without loss of generality, assume that the gram matrix $\boldsymbol{K}$ is invertible. If we penalize the intercepts and choose $J(\boldsymbol{f})$ to be the squared norm of $\boldsymbol{f}$ in the RKHS, the optimization problem (3) can be written as (Kimeldorf and Wahba, 1971).

$$\begin{aligned}
\min_{\boldsymbol{f}\in\mathcal{F}} \quad & \frac{n\lambda}{2}\sum_{q=1}^{k-1}\boldsymbol{\theta}_q^T\boldsymbol{K}\boldsymbol{\theta}_q + \frac{n\lambda}{2}\sum_{q=1}^{k-1}\theta_{q,0}^2 + \sum_{i=1}^{n}\sum_{j\neq y_i}(\xi_{ij}+\eta_{ij}) \\
subjectto \quad & \xi_{ij} \geq 0, \\
& \eta_{ij} \geq 0, \\
& \xi_{ij} - \langle \boldsymbol{f}(\boldsymbol{x}_i), \mathcal{Y}_j\rangle - 1 \geq 0, \\
& \eta_{ij} - (a-1)\langle \boldsymbol{f}(\boldsymbol{x}_i), \mathcal{Y}_j\rangle \geq 0; \quad i = 1,\ldots,n, \quad j \neq y_i,
\end{aligned}$$

where $f_q(\boldsymbol{x}) = \sum_{i=1}^{n} K(\boldsymbol{x}_i, \boldsymbol{x})\theta_{q,i} + \theta_{q,0};\ q = 1,\ldots,k-1$, and $\theta_{q,i}$ is the i^{th} element of $\boldsymbol{\theta}_q$. Now introduce the Lagrangian multipliers α_{ij}, γ_{ij}, τ_{ij}, and χ_{ij} as in the linear case, take the partial derivatives with respect to $\boldsymbol{\theta}_q$, $\theta_{q,0}$, ξ_{ij} and η_{ij} and set to zero, and we have

$$\begin{aligned}
\boldsymbol{\theta}_q &= -\frac{1}{n\lambda}\boldsymbol{K}^{-1}[\sum_{i=1}^{n}\sum_{j\neq y_i}\{\alpha_{ij} + (a-1)\gamma_{ij}\}\mathcal{Y}_{j,q}\boldsymbol{K}_i], \\
\theta_{q,0} &= -\frac{1}{n\lambda}\sum_{i=1}^{n}\sum_{j\neq y_i}\{\alpha_{ij} + (a-1)\gamma_{ij}\}\mathcal{Y}_{j,q},
\end{aligned}$$

where $\boldsymbol{K}_i$ is the i^{th} column of $\boldsymbol{K}$. Therefore, the optimization problem (3) is equivalent to

$$\begin{aligned}
\min_{\alpha_{ij},\gamma_{ij}} \quad & \frac{1}{2n\lambda}\sum_{q=1}^{k-1}[\sum_{i=1}^{n}\sum_{j\neq y_i}\{\alpha_{ij} + (a-1)\gamma_{ij}\}\mathcal{Y}_{j,q}\boldsymbol{K}_i]^T\boldsymbol{K}^{-1}[\sum_{i=1}^{n}\sum_{j\neq y_i}\{\alpha_{ij} + (a-1)\gamma_{ij}\}\mathcal{Y}_{j,q}\boldsymbol{K}_i] \\
& + \frac{1}{2n\lambda}\sum_{q=1}^{k-1}[\sum_{i=1}^{n}\sum_{j\neq y_i}\{\alpha_{ij} + (a-1)\gamma_{ij}\}\mathcal{Y}_{j,q}]^2 - \sum_{i=1}^{n}\sum_{j\neq y_i}\alpha_{ij} \\
subjectto \quad & 0 \leq \alpha_{ij} \leq A_{ij},\ 0 \leq \gamma_{ij} \leq A_{ij};\ i = 1,\ldots,n,\ j = 1,\ldots,k. \qquad (10)
\end{aligned}$$

Because $\boldsymbol{K}_i^T\boldsymbol{K}^{-1}\boldsymbol{K}_j = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, one can verify that (10) can be solved in an analogous

manner as (6) in the main paper.

4 Extended Numerical Results

Example 1, Soft with $a = a_1$			Regular	Reject	R&R	Probability Method	
	Proportion		Error			Proportion	Error
p1	50.91		29.36	28.38	28.38	52.26	30.17
p2	28.35	size 2: 22.64	45.78	44.94	1.673		
		size 3: 5.714					
p3	20.74		69.79	-	-	47.74	-
Overall	100.0		41.92	39.41	28.17	100.0	45.64
Example 1, Soft with $a = a_2$			Regular	Reject	R&R	Probability Method	
	Proportion		Error			Proportion	Error
p1	49.43		28.80	27.58	27.58	52.26	30.17
p2	28.97	size 2: 24.62	45.89	45.35	1.581		
		size 3: 4.349					
p3	21.60		69.61	-	-	47.74	-
Overall	100.0		41.92	39.32	27.47	100.0	45.64

Table 1: Simulation results for Example 1, the Soft loss, with $d = 0.6$

Example 1, DWD with $a = a_1$			Regular	Reject	R&R	Probability Method	
	Proportion		Error			Proportion	Error
p1	52.36		31.19	29.70	29.70	58.71	31.25
p2	27.79	size 2: 21.60	47.07	45.37	2.011		
		size 3: 6.192					
p3	19.85		69.95	-	-	41.29	-
Overall	100.0		42.87	39.96	28.04	100.0	43.93
Example 1, DWD with $a = a_2$			Regular	Reject	R&R	Probability Method	
	Proportion		Error			Proportion	Error
p1	49.45		31.79	31.30	31.30	58.71	31.25
p2	26.62	size 2: 17.91	45.04	46.59	2.980		
		size 3: 8.711					
p3	23.93		66.14	-	-	41.29	-
Overall	100.0		42.87	41.80	30.60	100.0	43.93

Table 2: Simulation results for Example 1, the DWD loss, with $d = 0.6$.

Example 1, Soft with $a = a_1$			Regular	Reject	R&R	Probability Method	
	Proportion		Error			Proportion	Error
p1	47.39		26.56	26.32	26.32	47.92	27.33
p2	26.33	size 2: 22.16	43.92	44.11	1.244		
		size 3: 4.168					
p3	26.28		67.61	-	-	53.08	-
Overall	100.0		41.92	37.22	25.94	100.0	39.64

Example 1, Soft with $a = a_2$			Regular	Reject	R&R	Probability Method	
	Proportion		Error			Proportion	Error
p1	47.01		26.40	26.11	26.11	47.92	27.33
p2	26.72	size 2: 22.90	44.03	44.07	1.201		
		size 3: 3.821					
p3	26.27		67.54	-	-	53.08	-
Overall	100.0		41.92	37.18	25.72	100.0	39.64

Table 3: Simulation results for Example 1, the Soft loss, with $d = 0.5$.

Example 1, DWD with $a = a_1$			Regular	Reject	R&R	Probability Method	
	Proportion		Error			Proportion	Error
p1	50.77		30.02	28.84	28.84	55.13	32.08
p2	25.85	size 2: 20.18	46.16	46.00	1.898		
		size 3: 5.673					
p3	23.38		67.13	-	-	44.87	-
Overall	100.0		42.87	38.22	26.82	100.0	40.12

Example 1, DWD with $a = a_2$			Regular	Reject	R&R	Probability Method	
	Proportion		Error			Proportion	Error
p1	49.78		29.74	28.80	28.80	55.13	32.08
p2	26.14	size 2: 21.56	44.88	45.69	1.731		
		size 3: 4.578					
p3	24.08		67.83	-	-	44.87	-
Overall	100.0		42.87	38.32	26.83	100.0	40.12

Table 4: Simulation results for Example 1, the DWD loss, with $d = 0.5$.

Example 2, Soft with $a = a_1$	Regular	Reject	R&R	Probability Method	
	Proportion	Error			Proportion	Error
p1	57.45	27.10	26.99	26.99	58.61	28.74
p2	size2: 16.81 ⌊ $\{1,2\}$: 72.2%	48.34	48.35	9.141		
p3	25.74	52.72	-	-	41.39	-
Overall	100.0	37.94	37.43	29.92	100.0	38.33
Example 2, Soft with $a = a_2$	Regular	Reject	R&R	Probability Method	
	Proportion	Error			Proportion	Error
p1	55.21	26.58	26.37	26.37	58.61	28.74
p2	size2: 16.11 ⌊ $\{1,2\}$: 72.8%	48.24	48.22	8.979		
p3	28.68	53.01	-	-	41.39	-
Overall	100.0	37.94	37.16	30.32	100.0	38.33

Table 5: Simulation results for Example 2, the Soft loss, with $d = 0.5$.

Example 2, SVM with $a = a_1$	Regular	Reject	R&R	Probability Method	
	Proportion	Error			Proportion	Error
p1	52.53	28.15	27.81	27.81	42.40	25.85
p2	size2: 12.90 ⌊ $\{1,2\}$: 73.1%	48.57	51.00	11.26		
p3	34.57	53.01	-	-	57.60	-
Overall	100.0	39.57	38.91	33.33	100.0	40.66
Example 2, SVM with $a = a_2$	Regular	Reject	R&R	Probability Method	
	Proportion	Error			Proportion	Error
p1	50.40	27.75	28.19	28.19	42.40	25.85
p2	size2: 14.41 ⌊ $\{1,2\}$: 73.3%	47.23	50.61	11.32		
p3	35.19	53.67	-	-	57.60	-
Overall	100.0	39.57	38.86	33.43	100.0	40.66

Table 6: Simulation results for Example 2, the SVM hinge loss, with $d = 0.5$.

Example 2, Soft with $a = a_1$	Regular	Reject	R&R	Probability Method	
	Proportion	Error			Proportion	Error
p1	64.91	29.30	28.36	28.36	63.34	29.12
p2	size2: 19.21 ⌊ $\{1,2\}$: 69.71%	43.00	42.12	9.538		
p3	15.88	67.14	-	-	36.66	-
Overall	100.0	37.94	36.03	29.77	100.0	40.44
Example 2, Soft with $a = a_2$	Regular	Reject	R&R	Probability Method	
	Proportion	Error			Proportion	Error
p1	63.58	29.10	28.01	28.01	63.34	29.12
p2	size2: 18.85 ⌊ $\{1,2\}$: 68.37%	43.65	43.18	10.01		
p3	17.57	63.80	-	-	36.66	-
Overall	100.0	37.94	36.49	30.23	100.0	40.44

Table 7: Simulation results for Example 2, the Soft loss, with $d = 0.6$.

Example 2, SVM with $a = a_1$	Regular	Reject	R&R	Probability Method	
	Proportion	Error		Proportion	Error
p1	66.43	32.15	32.20	32.20	54.68	26.51
p2	size2: 17.25 ⌊ {1, 2}: 69.24%	44.15	42.08	10.70		
p3	16.32	64.92	-	-	45.32	-
Overall	100.0	39.57	38.44	33.02	100.0	41.69
Example 2, SVM with $a = a_2$	Regular	Reject	R&R	Probability Method	
	Proportion	Error		Proportion	Error
p1	65.12	32.06	31.88	31.88	54.68	26.51
p2	size2: 17.57 ⌊ {1, 2}: 67.78%	43.33	43.07	10.16		
p3	17.31	63.99	-	-	45.32	-
Overall	100.0	39.57	38.71	32.93	100.0	41.69

Table 8: Simulation results for Example 2, the SVM hinge loss, with $d = 0.6$.

Example 3, DWD with $a = a_1$		Regular	Reject	R&R	Probability Method	
	Proportion	Error			Proportion	Error
p1	47.08	27.30	27.48	27.48	36.97	24.72
p2	26.25	size 2: 22.42 ⌊ {1, 2}: 42.5% ⌊ {3, 4}: 41.6%	35.29	36.86	3.768	
		size 3: 3.835				
p3	26.67	57.39	-	-	63.03	-
Overall	100.0	36.44	35.76	27.24	100.0	41.78
Example 3, DWD with $a = a_2$		Regular	Reject	R&R	Probability Method	
	Proportion	Error			Proportion	Error
p1	45.58	25.68	26.11	26.11	36.97	24.72
p2	23.26	size 2: 19.71 ⌊ {1, 2}: 40.3% ⌊ {3, 4}: 42.9%	36.45	35.71	1.771	
		size 3: 3.549				
p3	31.16	56.02	-	-	63.03	-
Overall	100.0	36.44	35.97	27.90	100.0	41.78

Table 9: Simulation results for Example 3, the DWD loss, with $d = 0.5$.

Example 3, SVM with $a = a_1$			Regular	Reject	R&R	Probability Method	
	Proportion		Error			Proportion	Error
p1	41.49		27.47	27.29	27.29	33.41	24.72
p2	31.26	size 2: 28.71 ⌊ $\{1,2\}$: 41.1% ⌊ $\{3,4\}$: 41.0%	33.84	33.87	2.494		
		size 3: 2.552					
p3	27.25		57.27	-	-	66.59	-
Overall	100.0		36.69	35.13	25.71	100.0	44.53
Example 3, SVM with $a = a_2$			Regular	Reject	R&R	Probability Method	
	Proportion		Error			Proportion	Error
p1	42.52		26.82	26.67	26.67	33.41	24.72
p2	28.18	size 2: 24.67 ⌊ $\{1,2\}$: 42.2% ⌊ $\{3,4\}$: 41.8%	35.63	36.98	3.619		
		size 3: 3.509					
p3	29.30		56.02	-	-	66.59	-
Overall	100.0		36.69	35.71	26.99	100.0	44.53

Table 10: Simulation results for Example 3, the SVM hinge loss, with $d = 0.5$.

Example 1	$a = a_1 = 1.333$	$a = 1.555$	$a = 1.777$	$a = a_2 = 2$
Soft	39.41	39.39	39.35	39.32
DWD	39.96	40.22	41.77	41.80
Example 2	$a = a_1 = 1.5$	$a = 1.667$	$a = 1.833$	$a = a_2 = 2$
Soft	37.43	37.39	37.26	37.16
SVM	38.91	39.01	38.81	38.86
Example 3	$a = a_1 = 1.667$	$a = 1.889$	$a = 2.111$	$a = a_2 = 2.333$
DWD	35.76	35.77	35.80	35.97
SVM	35.13	35.58	35.55	35.71

Table 11: The average empirical 0-d-1 loss on the test data sets for simulated Examples 1-3 using different loss functions and various a values.

GBM, Soft with $a = a_1$			Regular	Reject	R&R	Probability Method	
	Proportion		Error			Proportion	Error
p1	72.15		13.69	13.69	13.69	72.58	13.78
p2	18.99	size 2: 17.14 ⌊ $\{C, M\}$: 44.3% ⌊ $\{N, P\}$: 33.7%	41.35	39.53	3.724		
		size 3: 1.853					
p3	8.857		43.33	-	-	27.42	-
Overall	100.0		21.85	20.97	14.13	100.0	21.25
GBM, Soft with $a = a_2$			Regular	Reject	R&R	Probability Method	
	Proportion		Error			Proportion	Error
p1	70.15		13.16	12.84	12.84	72.58	13.78
p2	19.85	size 2: 18.42 ⌊ $\{C, M\}$: 44.8% ⌊ $\{N, P\}$: 35.8%	40.09	42.35	4.055		
		size 3: 1.428					
p3	10.00		54.86	-	-	27.42	-
Overall	100.0		21.85	21.00	13.81	100.0	21.25

Table 12: Data analysis results for the GBM data, the Soft loss.

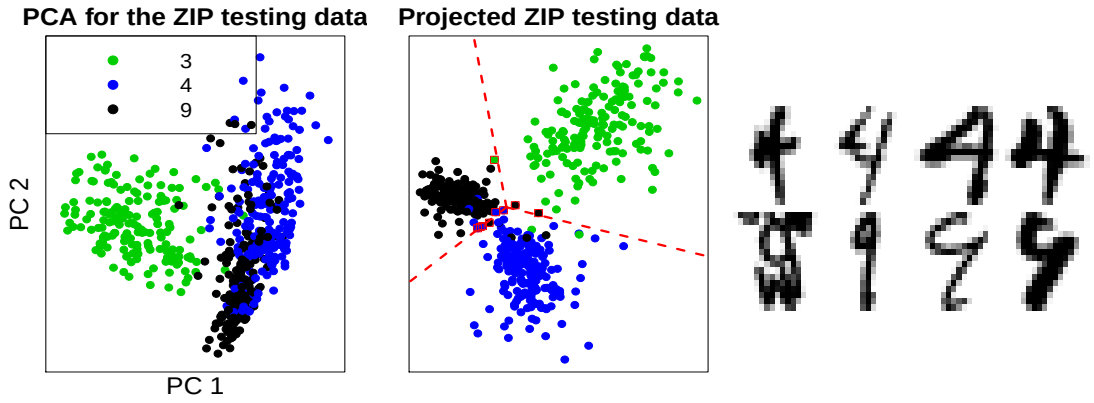

Figure 1: Left: the PCA scatter plot. Middle: the test data mapped to $\mathbb{R}^2$ using $\hat{\boldsymbol{f}}(\boldsymbol{x}) \in \mathbb{R}^2$ in a typical split, where the dashed lines correspond to the classification boundaries, and observations with reject or refine prediction are identified as red squares. Right: some observations that often ($> 80\%$ within the 100 splits) have refined prediction $\{4, 9\}$.

The ZIP data set has been extensively studied by many previous works. We choose categories "3", "4" and "9" to demonstrate the effect of the refine option. For handwritten digits, it is sometimes difficult for machines to classify between "4" and "9", while the difference between "3" and "4" or "3" and "9" is more obvious. For visualization, we draw a PCA plot for the test data on the left panel of Figure 1. In the middle panel, we provide a scatter plot by projecting the sample to the 2D space using $\hat{\boldsymbol{f}}(\boldsymbol{x}) \in \mathbb{R}^2$. In particular, observations with reject or refined set predications are shown in red squares. It can be seen that the observations which are refined are precisely those sitting on 2-way classification boundaries (shown as the dashed red lines), while most of them are between "4" and "9". In the analysis, we use $d = 0.4$, the DWD loss, and the L_2 penalty. We normalize the data set before the analysis. To select the best tuning parameters, we split the training data set into two groups, and use one to train the classifier and the other for tuning. We report the average results of 100 splits.

The results for the ZIP data set are reported in Table 13. For example, for $a = a_1$, note that although there are only a few rejected observations ($< 0.368\%$ on average), their misclassification rate is as high as 95.14%, if not rejected. This stunningly high error rate justifies our reject option. Though there are only 2.578% observations that are refined, the mis-refinement rate is as low as 0.110%, almost always correct. The middle panel of Figure 1 also suggests that the refinement decision is well deserved since the refined data points are in close vicinity to the classification boundaries. Lastly, it can be seen that, for quite a few observations, the classification signal is very vague between "4" and "9", which is consistent with our common sense (see the middle and right panels of Figure 1).

ZIP, DWD with $a = a_1$		Regular	Reject	R&R	Probability Method	
	Proportion	Error			Proportion	Error
p1	97.05	2.087	2.087	2.087	98.90	2.607
p2	size 2: 2.578 ⌊ {4, 9}: 55.9%	35.71	28.57	0.110		
p3	0.368	95.14	-	-	1.104	-
Overall	100.0	3.314	2.909	2.175	100.0	3.020
ZIP, DWD with $a = a_2$		Regular	Reject	R&R	Probability Method	
	Proportion	Error			Proportion	Error
p1	97.47	2.277	2.166	2.166	98.90	2.607
p2	size 2: 2.119 ⌊ {4, 9}: 57.2%	28.57	21.42	0.150		
p3	0.412	97.25	-	-	1.104	-
Overall	100.0	3.314	2.885	2.279	100.0	3.020

Table 13: Data analysis results for the ZIP data, the DWD loss.

References

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005), "Local Rademacher Complexities," *Annals of Statistics*, 33, 1497–1537.

Bartlett, P. L. and Mendelson, S. (2002), "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results," *Journal of Machine Learning Research*, 3, 463–482.

Bartlett, P. L. and Wegkamp, M. H. (2008), "Classification with a Reject Option Using a Hinge Loss," *Journal of Machine Learning Research*, 9, 1823–1840.

Kimeldorf, G. and Wahba, G. (1971), "Some Results on Tchebycheffian Spline Functions," *Journal of Mathematical Analysis and Applications*, 33, 82–95.

Koltchinskii, V. (2006), "Local Rademacher Complexities and Oracle Inequalities in Risk Minimization," *Annals of Statistics*, 34, 2593–2656.

Koltchinskii, V. and Panchenko, D. (2002), "Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifiers," *Annals of Statistics*, 30, 1–50.

McDiarmid, C. (1989), "On the Method of Bounded Differences," in *In Surveys in Combinatorics*, Cambridge University Press, pp. 148–188.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012), *Foundations of Machine Learning*, MIT press.

Shawe-Taylor, J. S. and Cristianini, N. (2004), *Kernel Methods for Pattern Analysis*, Cambridge University Press, 1st ed.

van der Vaart, A. W. and Wellner, J. A. (2000), *Weak Convergence and Empirical Processes with Application to Statistics*, Springer, 1st ed.

Wang, L. and Shen, X. (2007), "On L_1-norm Multi-class Support Vector Machines: Methodology and Theory," *Journal of the American Statistical Association*, 102, 595–602.

Zhang, C. and Liu, Y. (2014), "Multicategory Angle-based Large-margin Classification," *Biometrika*, 101, 625–640.

Zhang, C., Liu, Y., and Wu, Y. (2016), "On Quantile Regression in Reproducing Kernel Hilbert Spaces with Data Sparsity Constraint," *Journal of Machine Learning Research*, 17, 1–45.

Zhou, D.-X. (2002), "The Covering Number in Learning Theory," *Journal of Complexity*, 18, 739–767.