
Appendix: Multi-Class Learning: From Theory to Algorithm

1 Appendix A : Proof of Theorem 1

Theorem 1. *With probability at least $1 - \delta$, we have*

$$\mathcal{R}(\mathcal{L}^r) \leq \frac{c_{d,\vartheta} \xi(K) \sqrt{\zeta r} \log^{\frac{3}{2}}(n)}{\sqrt{n}} + \frac{4 \log(1/\delta)}{n},$$

where

$$\xi(K) = \begin{cases} \sqrt{e}(4 \log K)^{1 + \frac{1}{2 \log K}}, & \text{if } q \geq 2 \log K, \\ (2q)^{1 + \frac{1}{q}} K^{\frac{1}{q}}, & \text{otherwise,} \end{cases}$$

$c_{d,\vartheta}$ is a constant depends on d and ϑ .

To prove theorem 1, we first introduce the following two lemmas.

Lemma 1. [5] *Suppose that \mathcal{L} is defined in equation (2) in the paper, then we have*

$$\hat{\mathcal{R}}(\mathcal{L}) \leq \frac{\sqrt{2\pi}}{n} \mathbb{E}_{\mathbf{g}} \sup_{h=(h_1, \dots, h_K) \in \mathcal{H}_{p,\kappa}} \sum_{i=1}^n \sum_{j=1}^K g_{(j-1)n+i} h_j(\mathbf{x}_i),$$

where g_1, \dots, g_{nK} are independent $N(0, 1)$ random variables.

Proof. The empirical Gaussian complexities of $\mathcal{H}_{p,\kappa}$ is denote as

$$\hat{\mathcal{G}}(\mathcal{H}_{p,\kappa}) = \mathbb{E}_{\mathbf{g}} \left[\sup_{h \in \mathcal{H}_{p,\kappa}} \frac{1}{n} \sum_{i=1}^n g_i h(\mathbf{x}_i) \right],$$

where g_1, \dots, g_n are independent $N(0, 1)$ random variables. For any $\gamma > 0$, let $\rho_{\gamma,h}(\mathbf{x}, y)$ be

$$\begin{aligned} \rho_{\gamma,h}(\mathbf{x}, y) &= h(\mathbf{x}, y) - \max_{y' \in \mathcal{Y}} [h(\mathbf{x}, y') - \gamma 1_{y'=y}] \\ &= \min_{y' \in \mathcal{Y}} [h(\mathbf{x}, y) - h(\mathbf{x}, y') + \gamma 1_{y'=y}]. \end{aligned}$$

It is easy to checked that $\rho_{\gamma,h}(\mathbf{x}, y) = \min(\rho_h(\mathbf{x}, y), \gamma)$. For the fixed parameter $\gamma = c_\ell$, we observe that $\rho_{\gamma,h}(\mathbf{x}, y) = \min(\rho_h(\mathbf{x}, y), c_\ell)$. If $\rho_h(\mathbf{x}, y) > c_\ell$, we get

$$\ell(\rho_{\gamma,h}(\mathbf{x}, y)) = \ell(c_\ell) = 0 = \ell(\rho_h(\mathbf{x}, y)).$$

Thus, we have $\rho_{\gamma,h}(\mathbf{x}, y) = \rho_h(\mathbf{x}, y)$. Therefore, for any $z = (\mathbf{x}, y) \in \mathcal{Z}$ we have $\ell(\rho_{\gamma,h}(\mathbf{x}, y)) = \ell(\rho_h(\mathbf{x}, y))$, and

$$\mathcal{L}_\gamma := \{\rho_{\gamma,h}(\mathbf{x}, y) : h \in \mathcal{H}_{p,\kappa}\} = \mathcal{L}.$$

Thus, \mathcal{L}_γ satisfies the following inequality:

$$\begin{aligned}
\hat{\mathcal{R}}(\mathcal{L}) &= \hat{\mathcal{R}}(\mathcal{L}_\gamma) \\
&= \frac{1}{n} \mathbb{E}_\sigma \sum_{i=1}^n \sigma_i \left(h(\mathbf{x}_i, y_i) - \max_{y \in \mathcal{Y}} (h(\mathbf{x}_i, y) - \gamma 1_{y=y_i}) \right) \\
&\leq \frac{1}{n} \mathbb{E}_\sigma \sup_{h \in \mathcal{H}_{p,\kappa}} \left[\sum_{i=1}^n \sigma_i h(\mathbf{x}_i, y_i) \right] + \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}_{p,\kappa}} \sum_{i=1}^n \sigma_i \max_{y \in Y} (h(\mathbf{x}_i, y) - \gamma 1_{y=y_i}) \right] \\
&= \hat{\mathcal{R}}(\mathcal{H}_{p,\kappa}) + \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}_{p,\kappa}} \frac{1}{n} \sum_{i=1}^n \sigma_i \max_{y \in Y} (h(\mathbf{x}_i, y) - \gamma 1_{y=y_i}) \right]
\end{aligned}$$

From [4], we know that

$$\hat{\mathcal{R}}(\mathcal{H}_{p,\kappa}) \leq \sqrt{\frac{\pi}{2}} \hat{\mathcal{G}}(\mathcal{H}_{p,\kappa}).$$

According to the Lemma 4 of [5], we have

$$\hat{\mathcal{G}}(\mathcal{H}_{p,\kappa}) \leq \frac{1}{n} \mathbb{E}_g \sup_{h=(h_1, \dots, h_K) \in H} \sum_{i=1}^n \sum_{j=1}^K g_{(j-1)n+i} h_j(\mathbf{x}_i)$$

Thus, we have

$$\begin{aligned}
\hat{\mathcal{R}}(\mathcal{L}) &\leq \frac{1}{n} \sqrt{\frac{\pi}{2}} \mathbb{E}_g \sup_{h=(h_1, \dots, h_K) \in H} \sum_{i=1}^n \sum_{j=1}^K g_{(j-1)n+i} h_j(\mathbf{x}_i) + \\
&\quad \frac{1}{n} \sqrt{\frac{\pi}{2}} \underbrace{\mathbb{E}_g [g_i \max(h_1(\mathbf{x}_i) - \gamma 1_{y_i=1}, \dots, h_c(\mathbf{x}_i) - \gamma 1_{y_i=c})]}_{:=A_3}.
\end{aligned}$$

In the next, we bound A_3 :

$$\begin{aligned}
A_3 &\leq \mathbb{E}_g \sup_{h=(h_1, \dots, h_K) \in \mathcal{H}_{p,\kappa}} \sum_{i=1}^n \sum_{j=1}^K g_{(j-1)n+i} (h_j(\mathbf{x}_i) - \gamma 1_{y_i=c}) \\
&= \mathbb{E}_g \sup_{h=(h_1, \dots, h_K) \in \mathcal{H}_{p,\kappa}} \sum_{i=1}^n \sum_{j=1}^K g_{(j-1)n+i} (h_j(\mathbf{x}_i)) - \underbrace{\mathbb{E}_g \sum_{i=1}^n \sum_{j=1}^K g_{(j-1)n+i} \gamma 1_{y_i=j}}_{=0} \\
&= \mathbb{E}_g \sup_{h=(h_1, \dots, h_K) \in \mathcal{H}_{p,\kappa}} \sum_{i=1}^n \sum_{j=1}^K g_{(j-1)n+i} (h_j(\mathbf{x}_i)).
\end{aligned}$$

With this inequality, we immediately derive the following bound on $\hat{\mathcal{R}}(\mathcal{L})$:

$$\hat{\mathcal{R}}(\mathcal{L}) \leq \frac{\sqrt{2\pi}}{n} \mathbb{E}_g \sup_{h=(h_1, \dots, h_K) \in \mathcal{H}} \sum_{i=1}^n \sum_{j=1}^K g_{(j-1)n+i} h_j(\mathbf{x}_i).$$

□

Lemma 2. [5] Let g be $N(0, 1)$ distributed. For any $p > 0$, the p -th moment of g can be bounded by

$$[\mathbb{E}|g|^p]^{\frac{1}{p}} \leq (2p)^{\frac{1}{2} + \frac{1}{p}}. \quad (1)$$

Proof. Let $\Gamma(n) = (n-1)!$ be the Gamma function. The p -th moment of a $N(0, 1)$ distributed random variable can be exactly expressed via Gamma function [12]:

$$[\mathbb{E}_g |g|^p]^{\frac{1}{p}} = \frac{2^{\frac{p}{2}}}{\sqrt{\pi}} \Gamma\left(\frac{p+1}{2}\right) \leq \frac{2^{\frac{p}{2}}}{\sqrt{\pi}} \left\lceil \frac{p-1}{2} \right\rceil!.$$

According to the Stirling's approximation in [8], we can obtain that

$$\begin{aligned} [\mathbb{E}_g |g|^p]^{\frac{1}{p}} &\leq \frac{2^{\frac{p}{2}}}{\sqrt{\pi}} \left\lceil \frac{p-1}{2} \right\rceil! \\ &\stackrel{\text{Stirling's approximation}}{\leq} \frac{2^{\frac{p}{2}}}{\sqrt{\pi}} \sqrt{2\pi} \left\lceil \frac{p-1}{2} \right\rceil^{\left\lceil \frac{p-1}{2} \right\rceil! + \frac{1}{2}} \\ &\leq (2p)^{\frac{p}{2}+1}. \end{aligned}$$

□

Proposition 1. [5] Suppose that \mathcal{L} is defined in equation (2) in the paper, then we have

$$\hat{\mathcal{R}}(\mathcal{L}) \leq \frac{\sqrt{\vartheta}}{\sqrt{n}} \times \begin{cases} \sqrt{e}(4 \log K)^{1+\frac{1}{2 \log K}}, & \text{if } p^* \geq 2 \log K, \\ (2p^*)^{1+\frac{1}{p^*}} K^{\frac{1}{p^*}}, & \text{otherwise.} \end{cases}$$

Proof. From Lemma 1, we know that

$$\hat{\mathcal{R}}(\mathcal{L}) \leq \frac{\sqrt{2\pi}}{n} \mathbb{E}_{\mathbf{g}} \underbrace{\sup_{h=(h_1, \dots, h_K) \in \mathcal{H}_{p, \kappa}} \sum_{i=1}^n \sum_{j=1}^K g_{(j-1)n+i} h_j(\mathbf{x}_i)}_{:= A_1}, \quad (2)$$

where g_1, \dots, g_{nK} are independent $N(0, 1)$ random variables. In the next, we will bound A_1 . To this end, we denote by $\|\cdot\|_*$ dual norm of $\|\cdot\|$, i.e.,

$$\|w\|_* := \sup_{\|w'\| \leq 1} \langle w, w' \rangle.$$

For a convex function f , we denote by f^* its Fenchel conjugate, i.e.,

$$f^*(\nu) := \sup_w [\langle w, \nu \rangle - f(w)].$$

From Corollary 4 in [2], we know that, if f is η -strongly convex w.r.t $\|\cdot\|$ and $f^*(\mathbf{0}) = 0$, then for any sequence ν_1, \dots, ν_n and for any μ we have

$$\sum_{i=1}^n \langle \nu_i, \mu \rangle - f(\mu) \leq \sum_{i=1}^n \langle \nabla f^*(\nu_{1:i-1}), \nu \rangle + \frac{1}{2\eta} \sum_{i=1}^n \|\nu_i\|_*^2,$$

where $\nu_{1:i}$ denotes the sum $\sum_{j=1}^i \nu_j$. Let q be any number satisfying $p \leq q \leq 2$. Introduce the function $f_q(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|_{2,q}^2$. It is easy to verify that

$$f_q(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|_{2,q}^2 \leq \frac{1}{2} \|\mathbf{w}\|_{2,p}^2 \leq \frac{1}{2}.$$

Since $f_q(\mathbf{w})$ is $\frac{1}{q^*}$ -strongly convex w.r.t $\|\cdot\|_{2,q^*}$. Let

$$\begin{aligned} \nu_i &= (g_i \phi(\mathbf{x}_i), g_{n+i} \phi(\mathbf{x}_i), \dots, g_{(K-1)n+i} \phi(\mathbf{x}_i)), \\ \mu &= (\mathbf{w}_1, \dots, \mathbf{w}_K), \end{aligned}$$

we can obtain that

$$\begin{aligned} &\lambda \sup_{h \in \mathcal{H}_{q, \kappa}} \sum_{i=1}^n \sum_{j=1}^K g_{(j-1)n+i} \langle \mathbf{w}_j, \phi(\mathbf{x}_i) \rangle \\ &= \sup_{h \in \mathcal{H}_{q, \kappa}} \sum_{i=1}^n \langle (\mathbf{w}_1, \dots, \mathbf{w}_K), \lambda \nu_i \rangle \\ &\leq \sup_{h \in \mathcal{H}_{q, \kappa}} f_q(\mathbf{w}) + \sum_{i=1}^n \langle f^*(\nu_{1:i-1}), \lambda \nu_i \rangle + \sum_{i=1}^n \frac{q^* \lambda^2}{2} \|\nu_i\|_{2,q^*}^2. \end{aligned}$$

Taking expectation on both sides w.r.t. g_1, \dots, g_{nK} , we can obtain that

$$\begin{aligned}
& \mathbb{E}_{\mathbf{g}} \sup_{h \in \mathcal{H}_{q,\kappa}} \sum_{i=1}^n \sum_{j=1}^K g_{(j-1)n+i} \langle \mathbf{w}_j, \phi(\mathbf{x}_i) \rangle \\
& \leq \frac{1}{2\lambda} + \underbrace{\mathbb{E}_{\mathbf{g}} \sum_{i=1}^n \langle f^*(\nu_{1:i-1}), \nu_i \rangle}_{=0} + \sum_{i=1}^n \frac{q^* \lambda}{2} \|\nu_i\|_{2,q^*}^2 \\
& = \frac{1}{2\lambda} + \frac{q^* \lambda}{2} \sum_{i=1}^n \|\nu_i\|_{2,q^*}^2.
\end{aligned}$$

Choosing $\lambda = \frac{1}{\sqrt{q^* \sum_{i=1}^n \|\nu_i\|_{2,q^*}^2}}$, the above inequality translates to

$$\begin{aligned}
& \mathbb{E}_{\mathbf{g}} \sup_{h \in \mathcal{H}_{q,\kappa}} \sum_{i=1}^n \sum_{j=1}^K g_{(j-1)n+i} \langle \mathbf{w}_j, \phi(\mathbf{x}_i) \rangle \\
& \leq \sqrt{q^* \sum_{i=1}^n \|\nu_i\|_{2,q^*}^2} \\
& \leq \sqrt{q^* \sum_{i=1}^n \underbrace{\mathbb{E}_{\mathbf{g}} \|(g_i \phi(\mathbf{x}_i), \dots, g_{(K-1)n+i} \phi(\mathbf{x}_i))\|_{2,q^*}^2}_{:=A_2}}. \tag{3}
\end{aligned}$$

In the next, we bound A_2 :

$$\begin{aligned}
A_2 &= \sum_{i=1}^n \mathbb{E}_{\mathbf{g}} \left[\sum_{j=1}^K \|g_{(j-1)n+i} \phi(\mathbf{x}_i)\|_2^{q^*} \right]^{\frac{2}{q^*}} \\
&= \sum_{i=1}^n \mathbb{E}_{\mathbf{g}} \left[\sum_{j=1}^K |g_{(j-1)n+i}|^{q^*} \right]^{\frac{2}{q^*}} k(\mathbf{x}_i, \mathbf{x}_i) \\
&\stackrel{\text{symmetry}}{=} \mathbb{E}_{\mathbf{g}} \left[\sum_{j=1}^K |g_j|^{q^*} \right]^{\frac{2}{q^*}} \sum_{i=1}^n k(\mathbf{x}_i, \mathbf{x}_i) \\
&\stackrel{\text{Jensen's Inequality}}{\leq} \left[\mathbb{E}_{\mathbf{g}} \sum_{j=1}^K |g_j|^{q^*} \right]^{\frac{2}{q^*}} \sum_{i=1}^n k(\mathbf{x}_i, \mathbf{x}_i) \\
&= \left[K \mathbb{E}_{g_1} |g_1|^{q^*} \right]^{\frac{2}{q^*}} \sum_{i=1}^n k(\mathbf{x}_i, \mathbf{x}_i) \\
&\leq K^{\frac{2}{q^*}} \left[\mathbb{E}_{g_1} |g_1|^{q^*} \right]^{\frac{2}{q^*}} n\vartheta.
\end{aligned}$$

Substituting the above inequality in (10) in the paper, we have

$$\mathbb{E}_{\mathbf{g}} \sup_{h \in \mathcal{H}_{q,\kappa}} \sum_{i=1}^n \sum_{j=1}^K g_{(j-1)n+i} \langle \mathbf{w}_j, \lambda \phi(\mathbf{x}_i) \rangle \leq K^{\frac{1}{q^*}} \sqrt{q^*} \left[\mathbb{E}_{g_1} |g_1|^{q^*} \right]^{\frac{1}{q^*}} \sqrt{n\vartheta}.$$

From the trivial inequality $\|\mathbf{w}\|_{2,p} \geq \|\mathbf{w}\|_{2,q}$, we know that $\mathcal{H}_{p,\kappa} \subset \mathcal{H}_{q,\kappa}$. Combining the above equation and (9) in the paper, we have

$$\hat{\mathcal{R}}(\mathcal{L}) \leq \inf_{p \leq q \leq 2} \frac{\sqrt{2\vartheta\pi}}{\sqrt{n}} K^{\frac{1}{q^*}} \sqrt{q^*} \left[\mathbb{E}_{g_1} |g_1|^{q^*} \right]^{\frac{1}{q^*}}.$$

It can be directly checked that the function $t \rightarrow \sqrt{t}c^{1/t}$ is decreasing along the interval $(0, 2 \log K)$ and increasing along the interval $(2 \log K, \infty)$. Therefore, if $p^* \geq 2 \log K$,

$$\begin{aligned}\hat{\mathcal{R}}(\mathcal{L}) &\leq \frac{\sqrt{2\vartheta\pi}}{\sqrt{n}} K^{\frac{1}{2\log K}} \sqrt{2 \log K} [\mathbb{E}_{g_1} |g_1|^{2 \log K}]^{\frac{1}{2\log K}} \\ &\stackrel{\text{Lemma 2}}{\leq} \frac{\sqrt{2\vartheta\pi}}{\sqrt{n}} K^{\frac{1}{2\log K}} \sqrt{2 \log K} (4 \log K)^{\frac{1}{2} + \frac{1}{\log K}} \\ &\leq \frac{\sqrt{e\vartheta\pi}}{\sqrt{n}} (\log 4K)^{1 + \frac{1}{2\log K}}.\end{aligned}\tag{4}$$

If $0 < p^* \leq 2 \log K$, we have

$$\begin{aligned}\hat{\mathcal{R}}(\mathcal{L}) &\leq \frac{\sqrt{2\vartheta\pi}}{\sqrt{n}} K^{\frac{1}{p^*}} \sqrt{p^*} [\mathbb{E}_{g_1} |g_1|^{p^*}]^{\frac{1}{p^*}} \\ &\stackrel{\text{Lemma 2}}{\leq} \frac{\sqrt{\vartheta\pi}}{\sqrt{n}} K^{\frac{1}{p^*}} (2p^*)^{1 + \frac{1}{p^*}}.\end{aligned}\tag{5}$$

Combining (4) and (5) completes the proof. \square

Proof of Theorem 1. According to the Lemma 3.6 of [6], with probability $1 - \delta$, we have

$$\mathcal{R}(\mathcal{L}^r) \leq \hat{\mathcal{R}}(\mathcal{L}^r) + \sqrt{\frac{2 \log(1/\delta) \mathcal{R}(\mathcal{L}^r)}{n}}.$$

Note that $\forall a, b \geq 0$, $\sqrt{ab} \leq \frac{a}{2} + \frac{b}{2}$. Thus, we have

$$\mathcal{R}(\mathcal{L}^r) \leq \hat{\mathcal{R}}(\mathcal{L}^r) + \frac{\mathcal{R}(\mathcal{L}^r)}{2} + \frac{\log(1/\delta)}{n}.$$

So, we can obtain that

$$\mathcal{R}(\mathcal{L}^r) \leq 2\hat{\mathcal{R}}(\mathcal{L}^r) + \frac{2 \log(1/\delta)}{n}\tag{6}$$

From the Lemma 2.2 of [11], we know that if ℓ is a ζ -smooth loss function,

$$\hat{\mathcal{R}}(\mathcal{L}^r) \leq c_d \sqrt{\zeta r} \log^{\frac{3}{2}}(n) \hat{\mathcal{R}}(\mathcal{L}),\tag{7}$$

where c_d is a constant depends on d . Substituting (7) into (6), we have

$$\hat{\mathcal{R}}(\mathcal{L}^r) \leq 2c_d \sqrt{\zeta r} \log^{\frac{3}{2}}(n) \hat{\mathcal{R}}(\mathcal{L}) + \frac{2 \log(1/\delta)}{n}.\tag{8}$$

From Proposition 1, we have

$$\hat{\mathcal{R}}(\mathcal{L}) \leq \frac{1}{\sqrt{n}} \times \begin{cases} \sqrt{e\vartheta}(4 \log K)^{1 + \frac{1}{2\log K}}, & \text{if } q \geq 2 \log K, \\ \sqrt{\vartheta}(2q)^{1 + \frac{1}{q}} K^{\frac{1}{q}}, & \text{otherwise.} \end{cases}$$

Substituting the above result into (8), the proof is over. \square

2 Appendix B: Proof of Theorem 2

Theorem 2. If ℓ is a ζ -smooth loss. Then, $\forall h \in \mathcal{H}_{p,\kappa}$ and $\forall k > \max(1, \frac{\sqrt{2}}{2d})$, with probability $1 - \delta$, we have

$$L(h) \leq \max \left\{ \frac{k}{k-1} \hat{L}(\ell_h), \hat{L}(\ell_h) + \frac{c_{d,\vartheta,\zeta,k} \xi^2(K) \log^3 n}{n} + \frac{c_\delta}{n} \right\},$$

where

$$\xi(K) = \begin{cases} \sqrt{e}(4 \log K)^{1 + \frac{1}{2\log K}}, & \text{if } q \geq 2 \log K, \\ (2q)^{1 + \frac{1}{q}} K^{\frac{1}{q}}, & \text{otherwise,} \end{cases}$$

$c_{d,\vartheta}$ is a constant depending on d, ϑ, ζ, k , and c_δ is a constant depending on δ .

To prove Theorem 2, we first introduce the following two lemmas.

Lemma 3. *Let $\bar{\mathcal{L}}$ be the normalized loss space*

$$\bar{\mathcal{L}} = \left\{ \frac{r}{\max(L(\ell_h^2), r)} \ell_h \mid \ell_h \in \mathcal{L} \right\}. \quad (9)$$

Suppose that $\forall k > 1$,

$$\hat{U}_n(\bar{\mathcal{L}}) := \sup_{\bar{\ell}_h \in \bar{\mathcal{L}}} \left\{ L(\bar{\ell}_h) - \hat{L}(\bar{\ell}_h) \right\} \leq \frac{r}{Mk}.$$

Then, $\forall h \in \mathcal{H}_{p,\kappa}$, we have

$$L(\ell_h) \leq \max \left\{ \left(\frac{k}{k-1} \hat{L}(\ell_h) \right), \left(\hat{L}(\ell_h) + \frac{r}{Mk} \right) \right\}.$$

Proof. Note that, $\forall \bar{\ell}_h \in \bar{\mathcal{L}}$:

$$L(\bar{\ell}_h) \leq \hat{L}_n(\bar{\ell}_h) + \hat{U}_n(\bar{\mathcal{L}}) \leq \hat{L}_n(\bar{\ell}_h) + \frac{r}{Mk}. \quad (10)$$

Let us consider the two cases:

- 1) $L(\ell_h^2) \leq r, \ell_h \in \mathcal{L}$.
- 2) $L(\ell_h^2) > r, \ell_h \in \mathcal{L}$.

In the first case $\bar{\ell}_h = \ell_h$, by (10), we have

$$L(\ell_h) = L(\bar{\ell}_h) \leq \hat{L}_n(\bar{\ell}_h) + \frac{r}{Mk} = \hat{L}(\ell_h) + \frac{r}{Mk}. \quad (11)$$

In the second case, $\bar{\ell}_h = \frac{r}{L(\ell_h^2)} \ell_h$, then

$$L(\ell_h) - \hat{L}(\ell_h) \leq \hat{U}_n(\mathcal{L}) = \frac{L(\ell_h^2)}{r} \hat{U}_n(\bar{\mathcal{L}}) \leq \frac{M \cdot L(\ell_h)}{r} \frac{r}{Mk} = \frac{L(\ell_h)}{k}. \quad (12)$$

By combining the results of Eqs (11) and (12), the proof is over. \square

Lemma 4. *Suppose that $\bar{\mathcal{L}}$ is defined in Equation (9) in the paper, then*

$$\bar{\mathcal{L}} \subseteq \mathcal{L}^r.$$

Proof. Let us consider \mathcal{L}^r in the two cases:

- 1) $L(\ell_h^2) \leq r, \ell_h \in \mathcal{L}$.
- 2) $L(\ell_h^2) > r, \ell_h \in \mathcal{L}$.

In the first case, $\bar{\ell}_h = \ell_h$ and then:

$$L(\ell_h^2) = L(\bar{\ell}_h^2) \leq r.$$

In the second case, $L(\ell_h^2) > r$, so we have that

$$\bar{\ell}_h = \left\lfloor \frac{r}{L(\ell_h^2)} \right\rfloor \ell_h, \frac{r}{L(\ell_h^2)} \leq 1,$$

and the following bound holds:

$$L(\bar{\ell}_h^2) = \left\lfloor \frac{r}{L(\ell_h^2)} \right\rfloor^2 L(\ell_h^2) \leq \left\lfloor \frac{r}{L(\ell_h^2)} \right\rfloor L(\ell_h^2) = r.$$

Thus, the lemma is proved. \square

Lemma 5. $\psi(r) = \mathcal{R}(\mathcal{L}^r)$ is a sub-root function.

Proof. In order to prove this lemma, we can show the following: 1) $\psi_n(r)$ is positive; 2) $\psi_n(r)$ is non-decreasing; 3) $\psi_n(r)/\sqrt{r}$ is non-increasing.

By the definition of $\mathcal{R}(\mathcal{L}^r)$, it is easy to verify that $\mathcal{R}(\mathcal{L}^r)$ is positive.

Concerning the second property, we have that, for $0 \leq r_1 \leq r_2$: $\mathcal{L}^{r_1} \subseteq \mathcal{L}^{r_2}$, therefore

$$\begin{aligned}\psi(r_1) &= \mathbb{E}_{\mathcal{S}, \sigma} \left[\sup_{\ell_h \in \mathcal{L}^{r_1}} \frac{2}{n} \sum_{i=1}^n \sigma_i \ell_h(z_i) \right] \\ &\leq \mathbb{E}_{\mathcal{S}, \sigma} \left[\sup_{\ell_h \in \mathcal{L}^{r_2}} \frac{2}{n} \sum_{i=1}^n \sigma_i \ell_h(z_i) \right] = \psi(r_2).\end{aligned}$$

Finally, concerning the third property, for $0 \leq r_1 \leq r_2$, let

$$\ell_{hr_2} = \arg \sup_{\ell_h \in \mathcal{L}^{r_2}} \mathbb{E}_{\mathcal{S}, \sigma} \left[\sup_{\ell_h \in \mathcal{L}^{r_2}} \frac{2}{n} \sum_{i=1}^n \sigma_i \ell_h(z_i) \right].$$

Note that, since $\frac{r_1}{r_2} \leq 1$, we have that $\sqrt{\frac{r_1}{r_2}} \ell_{hr_2} \in \mathcal{L}^{r_1}$. Consequently: $L \left[\left(\sqrt{\frac{r_1}{r_2}} \ell_{hr_2} \right) \right]^2 = \frac{r_1}{r_2} L \left[(\ell_{hr_2})^2 \right] \leq r_1$. Thus, we have that:

$$\begin{aligned}\psi(r_1) &= \mathbb{E}_{\mathcal{S}, \sigma} \left[\sup_{\ell_h \in \mathcal{L}^{r_1}} \frac{2}{n} \sum_{i=1}^n \sigma_i \ell_h(z_i) \right] \\ &\geq \mathbb{E}_{\mathcal{S}, \sigma} \left[\frac{2}{n} \sum_{i=1}^n \sigma_i \sqrt{\frac{r_1}{r_2}} \ell_{hr_2}(z_i) \right] \\ &= \sqrt{\frac{r_1}{r_2}} \mathbb{E}_{\mathcal{S}, \sigma} \left[\sup_{\ell_h \in \mathcal{L}^{r_2}} \frac{2}{n} \sum_{i=1}^n \sigma_i \ell_h(z_i) \right] = \sqrt{\frac{r_1}{r_2}} \psi(r_2),\end{aligned}$$

which allows proving the claim since $\frac{\psi(r_2)}{\sqrt{r_2}} \leq \frac{\psi(r_1)}{\sqrt{r_1}}$. \square

Proposition 2. Let us consider a sub-root function $\psi(r)$, with fixed point r^* , and suppose that $\forall r > r^*$,

$$\mathcal{R}(\mathcal{L}^r) \leq \psi(r). \quad (13)$$

Then, $\forall h \in \mathcal{H}_{p, \kappa}$ and $\forall k > \max(1, \frac{\sqrt{2}}{2M})$, with probability $1 - \delta$, we have,

$$L(\ell_h) \leq \max \left\{ \frac{k}{k-1} \hat{L}(\ell_h), \hat{L}(\ell_h) + c_M r^* + \frac{c_\delta}{n} \right\},$$

where $c_M = 18Mk$, $c_\delta = \frac{(12k+14) \log(1/\delta)}{3}$

Proof. From the Theorem 2.1 of [1], we have

$$\begin{aligned}\hat{U}(\bar{\mathcal{L}}) &= \sup_{\bar{\ell}_h \in \bar{\mathcal{L}}} \left\{ L(\bar{\ell}_h) - \hat{L}(\bar{\ell}_h) \right\} \\ &\leq \inf_{\alpha > 0} \left(2(1 + \alpha) \mathcal{R}(\bar{\mathcal{L}}) + \sqrt{\frac{2r \log(1/\delta)}{n}} \right. \\ &\quad \left. + M \left(\frac{1}{3} + \frac{1}{\alpha} \right) \frac{\log(1/\delta)}{n} \right) \\ &\stackrel{\text{Lemma 4}}{\leq} \inf_{\alpha > 0} \left(2(1 + \alpha) \mathcal{R}(\mathcal{L}^r) + \sqrt{\frac{2r \log(1/\delta)}{n}} \right. \\ &\quad \left. + M \left(\frac{1}{3} + \frac{1}{\alpha} \right) \frac{\log(1/\delta)}{n} \right) \\ &\stackrel{(13), \text{ set } \alpha = \frac{1}{2}}{\leq} 3\psi(r) + \sqrt{\frac{2r \log(1/\delta)}{n}} + \frac{7M \log(1/\delta)}{3n} \\ &\stackrel{\text{sub-root}}{\leq} 3\sqrt{rr^*} + \sqrt{\frac{2r \log(1/\delta)}{n}} + \frac{7M \log(1/\delta)}{3n}.\end{aligned}$$

The next step of the proof consists in showing that r can be chosen, such that $\hat{U}(\bar{\mathcal{L}}) \leq \frac{r}{Mk}$ and $r \geq r^*$, so that we can exploit Lemma 3 and conclude the proof. For this purpose, we set

$$A = 3\sqrt{r^*} + \sqrt{\frac{2\log(1/\delta)}{n}}, B = \frac{7M\log(1/\delta)}{3n}.$$

Thus, we have to find the solution of

$$A\sqrt{r} + B = \frac{r}{Mk},$$

which is

$$r = \frac{\left[\left(\frac{2B}{kM} + A^2 \right) + \sqrt{\left(\frac{2B}{kM} + A^2 \right)^2 - \frac{4B^2}{M^2k^2}} \right]}{\frac{2}{M^2k^2}}. \quad (14)$$

Since $k \geq \max(1, \frac{\sqrt{2}}{2M})$, $k^2M^2 \geq \frac{1}{2}$. Therefore, from (14), we have

$$r \geq A^2M^2k^2 \geq \frac{A^2}{2} = r^*, r \leq A^2M^2k^2 + 2BMk.$$

Thus, we have

$$\begin{aligned} \frac{r}{Mk} &\leq A^2Mk + 2B \\ &= \left(3\sqrt{r^*} + \sqrt{2\log(1/\delta)/n} \right)^2 Mk + \frac{14M\log(1/\delta)}{3n}. \end{aligned}$$

Note that, $\forall a, b > 0$, $(a+b)^2 \leq 2a^2 + 2b^2$, so we have that

$$\frac{r}{Mk} \leq 18Mkr^* + \frac{(12k+14)\log(1/\delta)}{3n}$$

By substituting the above inequality into Lemma 3, the proof is over. \square

Proof of Theorem 2. From Proposition 2, with probability $1 - \delta$, we have

$$L(\ell_h) \leq \max \left\{ \frac{k}{k-1} \hat{L}(\ell_h), \hat{L}(\ell_h) + c_d r^* + \frac{c_\delta}{n} \right\}, \quad (15)$$

where r^* is a fixed point of $\mathcal{R}(\mathcal{L}^r)$. From Lemma 2, we know that the $\mathcal{R}(\mathcal{L}^r)$ is a sub-root function, so the fixed point r^* of $\mathcal{R}(\mathcal{L}^r)$ is uniquely exists.

According to Theorem 1, we know that

$$\mathcal{R}(\mathcal{L}^r) \leq \frac{c_{d,\vartheta}\xi(K)\sqrt{\zeta r}\log^{\frac{3}{2}}(n)}{\sqrt{n}} + \frac{4\log(1/\delta)}{n}.$$

Thus, if we set $A = \frac{c_{d,\vartheta}\xi(K)\sqrt{\zeta}\log^{\frac{3}{2}}(n)}{\sqrt{n}}$, $B = \frac{4\log(1/\delta)}{n}$, the fixed point r^* is smaller than the solution of $A\sqrt{r} + B = r$, which is

$$\begin{aligned} r^s &= \frac{2B + A^2 + \sqrt{(2B + A^2)^2 - 4B^2}}{2} \\ &\leq 2B + A^2 = \frac{c_{d,\vartheta}^2\xi^2(K)\zeta\log^3(n)}{n} + \frac{4\log(1/\delta)}{n}. \end{aligned}$$

Substituting the above inequality into (15) finishes the proof. \square

3 Appendix C: Proof of Theorem 3

Theorem 3. Let $\boldsymbol{\nu} = [\|\boldsymbol{\theta}_1\| - \beta r_1, \dots, \|\boldsymbol{\theta}_M\| - \beta r_M]$, then the component m -th of $\nabla \Omega^*(\boldsymbol{\theta})$ is

$$\frac{\text{sgn}(\nu_m) \boldsymbol{\theta}_m |\nu_m|^{q-1}}{\alpha \|\boldsymbol{\theta}_m\| \|\boldsymbol{\nu}\|_q^{q-2}},$$

where $\text{sgn}(x)$ is defined as $\text{sgn}(x) = 1$, if $x > 0$, $\text{sgn}(x) = -1$, if $x < 0$, $\text{sgn}(x) \in [-1, +1]$, if $x = 0$.

Lemma 6. Let $p \in (1, 2]$ and $q = p/(p-1)$, and then the norms $\|\mathbf{c}\|_p$ and $\|\mathbf{c}^*\|_q$ are dual to each other. Define the mapping $f : \mathcal{M} \rightarrow \mathcal{M}$ with

$$c_i^* = f_i(\mathbf{c}) = \nabla_i \left(\frac{1}{2} \|\mathbf{c}\|_p^2 \right) = \frac{\text{sgn}(c_i) |c_i|^{p-1}}{\|\mathbf{c}\|_p^{p-2}}, i = 1, \dots, n,$$

and the inverse mapping f^{-1} with

$$c_i = f_i^{-1}(\mathbf{c}^*) = \nabla_i \left(\frac{1}{2} \|\mathbf{c}^*\|_q^2 \right) = \frac{\text{sgn}(c_i^*) |c_i^*|^{q-1}}{\|\mathbf{c}^*\|_q^{q-2}}, i = 1, \dots, n,$$

These mapping are often called *link functions* in machine learning (e.g., [3]).

Lemma 7. For ℓ_1 -regularization, there is a scalar minimization problem

$$\min_{\mathbf{w} \in \mathbf{R}} \eta_t \mathbf{w} + \lambda_t |\mathbf{w}| + \frac{\gamma_t}{2} \mathbf{w}^2,$$

And an optimal solution \mathbf{w}^* can be summarized as

$$\mathbf{w}^* = \begin{cases} 0 & \text{if } |\eta_t| \leq \lambda_t, \\ -\frac{1}{\gamma_t} (\eta_t - \lambda_t \text{sgn}(\eta_t)) & \text{otherwise.} \end{cases}$$

Proof. The minimization problem is an unconstrained nonsmooth optimization problem. Its optimality condition [9] states that \mathbf{w}^* is an optimal solution if and only if there exists $\xi \in \partial |\mathbf{w}^*|$ such that

$$\eta_t + \lambda_t \xi + \gamma_t \mathbf{w}^* = 0.$$

There are more discussions in Appendix A of [13]. Finally, we can get the closed-form for ℓ_1 -regularization. \square

Proof of Theorem 3. According to standard Legendre-Fenchel duality, we can get

$$\begin{aligned} \nabla \Omega^*(\boldsymbol{\theta}) &= \arg \max_{\mathbf{w}} \mathbf{w} \cdot \boldsymbol{\theta} - \Omega(\mathbf{w}) \\ &= \arg \max_{\mathbf{w}} \mathbf{w} \cdot \boldsymbol{\theta} - \frac{\alpha}{2} \|\mathbf{w}\|_{2,p}^2 - \beta \boldsymbol{\mu} \cdot \mathbf{r}. \end{aligned}$$

To reach the above argmax, the derivative of argmax should be zero, so \mathbf{w} must be proportional to $\boldsymbol{\theta}$. As in UFO-MKL [7], we explicitly give a tricky link function $\mathbf{w}_m = \mu_m \boldsymbol{\theta}_m$ for different kernels. By this explicit link function, the algorithm can update both \mathbf{w} and $\boldsymbol{\mu}$ by $\nabla \Omega^*(\boldsymbol{\theta})$. Then, for the convenience of computation, we focus on $c_m = \mu_m \|\boldsymbol{\theta}_m\|$, rewriting the argmax:

$$\arg \min_{\mathbf{c}} (\beta \mathbf{r} - \mathbf{a}) \cdot \mathbf{c} + \frac{\alpha}{2} \|\mathbf{c}\|_p^2 \quad (16)$$

where $\mathbf{a} = [\|\boldsymbol{\theta}_1\|, \dots, \|\boldsymbol{\theta}_M\|]$.

The optimality condition of the above minimization problem [9] states that \mathbf{c}^* is an optimal solution of (16). And we set the derivative of above argmin being zero

$$\beta \mathbf{r} - \mathbf{a} + \alpha \mathbf{c}^* = 0 \quad (17)$$

And then we can get $\mathbf{c}^* = \frac{1}{\alpha}(\mathbf{a} - \beta \mathbf{r})$. Following similar arguments of Lemma 6 and Lemma 7, we find that it has a closed-form solution

$$c_m = f^{-1}(c_m^*) = \nabla_m \left(\frac{1}{2} \|c_m^*\|_q^2 \right) = \frac{\text{sgn}(c_m^*) |c_m^*|^{q-1}}{\alpha \|\mathbf{c}^*\|_q^{q-2}}.$$

Let $\boldsymbol{\nu} = [\|\boldsymbol{\theta}_1\| - \beta \mathbf{r}_1, \dots, \|\boldsymbol{\theta}_M\| - \beta \mathbf{r}_M]$, and use $\mu_m = c_m / \|\boldsymbol{\theta}_m\|$ and $\mathbf{w}_m = \mu_m \boldsymbol{\theta}_m$, We can get

$$\mu_m = \frac{\text{sgn}(\nu_m) |\nu_m|^{q-1}}{\alpha \|\boldsymbol{\theta}_m\| \|\boldsymbol{\nu}\|_q^{q-2}}, \quad \mathbf{w}_m = \frac{\text{sgn}(\nu_m) \boldsymbol{\theta}_m |\nu_m|^{q-1}}{\alpha \|\boldsymbol{\theta}_m\| \|\boldsymbol{\nu}\|_q^{q-2}}.$$

Similar argmax has been analysis in Section 7.2 of [13]. \square

4 Appendix D: Convergence Analysis

4.1 Conv-MKL

Convergence rate of the proposed Conv-MKL is decided by which ℓ_p MC-MKL algorithm it uses. In experiments, we following implement Conv-MKL based on UFO-MKL [7]. Thus, convergence rate of Conv-MKL is same with UFO-MKL in Section 4.1 of [7].

4.2 SMSD-MKL

Denote by $z^t = \partial \ell(\mathbf{w}, \phi_{\boldsymbol{\mu}}(\mathbf{x}^t), y^t)$, we now state the convergence theorem for any loss function that satisfies the following hypothesis

$$\|z_m\| \leq L \|\phi_m(\cdot)\|_2, \forall t = 1, \dots, M. \quad (18)$$

Theorem 4. Denote by

$$f(\mathbf{w}) = \Omega(\mathbf{w}) + \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{w}, \phi_{\boldsymbol{\mu}}(\mathbf{x}_i), y_i)$$

and by \mathbf{w}^* the solution. Suppose that $\|\phi_m(\cdot)\|_2 \leq 1$, and the loss function ℓ satisfies (18). Let $\delta \in (0, 1)$, then with probability at least $1 - \delta$ over the choices of the random samples we have that after T iterations of the SMSD-MKL algorithm

$$f(\mathbf{w}^{T+1}) - f(\mathbf{w}^*) \leq \frac{eL^2(1 + \log T) \log M}{\alpha \delta T},$$

where e is the Euler's number.

Proof. Using (18) and $\|\phi_m(\cdot)\|_2 \leq 1$, we have

$$\begin{aligned} & \|\partial(\mathbf{w}^t, \phi_{\boldsymbol{\mu}}(\mathbf{x}^t), y^t)\|_{2,q} \\ & \leq LM^{1/q} \max_{j=1, \dots, M} \|\phi_m(\mathbf{x}^t)\|_2 \leq LF^{1/q} \end{aligned}$$

The function $\Omega(\mathbf{w}) = \frac{\alpha}{2} \|\mathbf{w}\|_{2,p}^2 + \beta \boldsymbol{\mu} \cdot \mathbf{r}$ is α -strongly convex w.r.t. the norm $\|\cdot\|_{2,q}$. Hence, using according to Theorem 1 in [10], using $\eta = 1$ and $g = \Omega$, and using Markov inequality as in [10] we prove the stated result. \square

References

- [1] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.
- [2] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and Support Vector Machines. *Advances in Computational Mathematics*, 13:1–50, 2000.
- [3] C. Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003.

- [4] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.
- [5] Y. Lei, U. D. A. Binder, and M. Kloft. Multi-class SVMs: From tighter data-dependent generalization bounds to novel algorithms. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 2035–2043, 2015.
- [6] L. Oneto, A. Ghio, D. Anguita, and S. Ridella. An improved analysis of the rademacher data-dependent bound using its self bounding property. *Neural Networks*, 44:107–111, 2013.
- [7] F. Orabona and J. Luo. Ultra-fast optimization algorithm for sparse multi kernel learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 249–256, 2011.
- [8] H. Robbins. A remark on stirling’s formula. *The American Mathematical Monthly*, 62(1):26–29, 1955.
- [9] R. T. Rockafellar. *Convex analysis*. Princeton university press, 1970.
- [10] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 807–814, 2007.
- [11] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems 22 (NIPS)*, pages 2199–2207, 2010.
- [12] A. Winkelbauer. Moments and absolute moments of the normal distribution. *arXiv preprint arXiv:1209.4340*, 2012.
- [13] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.