A Auxiliary Lemmas

In this section, we introduce auxiliary lemmas used in our analysis. The first one is Hoeffding's inequality.

Lemma A (Hoeffding's inequality)**.** *Let $Z_1, \ldots, Z_s$ be i.i.d. random variables to $[-a, a]$ for $a > 0$. Denote by A_s the sample average $\sum_{i=1}^{s} Z_i / s$. Then, for any $\epsilon > 0$, we get*

$$\mathbb{P}[A_s + \epsilon \leq \mathbb{E}[A_s]] \leq \exp\left(-\frac{\epsilon^2 s}{2a^2}\right).$$

Note that this statement can be reinterpreted as follows: it follows that for $\delta \in (0, 1)$ with probability at least $1 - \delta$

$$A_s + a\sqrt{\frac{2}{s}\log\frac{1}{\delta}} \geq \mathbb{E}[A_s].$$

We next introduce the uniform bound by Rademacher complexity. For a set $\mathcal{G}$ of functions from $\mathcal{Z}$ to $[-a, a]$ and a dataset $S = \{z_i\}_{i=1}^{s} \subset \mathcal{Z}$, we denote empirical Rademacher complexity by $\hat{\Re}_S(\mathcal{G})$ and denote Rademacher complexity by $\Re_s(\mathcal{G})$; let $\sigma = (\sigma_i)_{i=1}^{s}$ be i.i.d random variables taking -1 or 1 with equal probability and let S be distributed according to a distribution μ^s,

$$\hat{\Re}_S(\mathcal{G}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{G}} \frac{1}{s}\sum_{i=1}^{s}\sigma_i f(x_i)\right], \ \ \Re_s(\mathcal{G}) = \mathbb{E}_{\mu^s}[\hat{\Re}_S(\mathcal{G})].$$

Lemma B. *Let $Z_1, \ldots, Z_s$ be i.i.d random variables to $\mathcal{Z}$. Denote by $A_s(f)$ the sample average $\sum_{i=1}^{s} f(Z_i)/s$. Then, for any $\delta \in (0, 1)$, we get with probability at least $1 - \delta$ over the choice of S,*

$$\sup_{f\in\mathcal{G}} |A_s(f) - \mathbb{E}[A_s(f)]| \leq 2\Re_s(\mathcal{G}) + a\sqrt{\frac{2}{s}\log\frac{2}{\delta}}.$$

When a function class is VC-class (for the definite see [vdVW96]), its Rademacher complexity is uniformly bounded as in the following lemma which can be easily shown by Dudley's integral bound [Dud99] and the bound on the covering number by VC-dimension (pseudo-dimension) [vdVW96].

Lemma C. *Let $\mathcal{G}$ be VC-class. Then, there exists positive value M depending on $\mathcal{G}$ such that $\Re_s(\mathcal{G}) \leq M/\sqrt{m}$.*

The following lemma is useful in estimating Rademacher complexity.

Lemma D. *(i) Let $h_i : \mathbb{R} \to \mathbb{R}$ $(i \in \{1, \ldots, s\})$ be L-Lipschitz functions. Then it follows that*

$$\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^{s}\sigma_i h_i \circ f(x_i)\right] \leq L\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{G}}\sum_{i=1}^{s}\sigma_i \circ f(x_i)\right].$$

(ii) We denote by $\mathrm{conv}(\mathcal{G})$ the convex hull of $\mathcal{G}$. Then, we have $\hat{\Re}_S(\mathrm{conv}(\mathcal{G})) = \hat{\Re}_S(\mathcal{G})$.

The following lemma gives the generalization bound by the margin distribution, which is originally derived by [KP02]. Let $\mathcal{G}$ be the set of predictors; $\mathcal{G} \subset \{f : \mathcal{X} \to \mathbb{R}^c\}$ and denote $\Pi\mathcal{G} = \{f_y(\cdot) : \mathcal{X} \to | \ f \in \mathcal{G}, y \in \mathcal{Y}\}$, then the following holds.

Lemma E. *Fix $\delta > 0$. Then, for $\forall\rho > 0$, with probability at least $1 - \rho$ over the random choice of S from ν^n, we have $\forall f \in \mathcal{G}$,*

$$\mathbb{P}_\nu[m_f(X, Y) \leq 0] \leq \mathbb{P}_{\nu_n}[m_f(X, Y) \leq \delta] + \frac{2c^2}{\delta}\Re_n(\Pi\mathcal{G}) + \sqrt{\frac{1}{2n}\log\frac{1}{\rho}}.$$

B Proofs

In this section, we provide missing proofs in the paper.

B. 1 Proofs of Section 3 and 4

We first prove Proposition 1 that states Lipschitz smoothness of the risk function.

Proof of Proposition 1 . Because $l(z,y,w)$ is $\mathcal{C}^2$-function with respect to z, w, there exist semi-positive definite matrices $A_{x,y}^{\phi,\psi}, B_{x,y}^{\phi,\psi}$ such that

$$\begin{aligned} l(\psi(x),y,w_\phi) &= l(\phi(x),y,w_\phi) + \partial_z l(\phi(x),y,w_\phi)^\top (\psi(x)-\phi(x)) \\ &\quad + \frac{1}{2}(\psi(x)-\phi(x))^\top A_{x,y}^{\phi,\psi}(\psi(x)-\phi(x)), \qquad (1) \\ l(\psi(x),y,w_\phi) + \frac{\lambda}{2}\|w_\phi\|_2^2 &= l(\psi(x),y,w_\psi) + \frac{\lambda}{2}\|w_\psi\|_2^2 \\ &\quad + (\partial_w l(\psi(x),y,w_\psi) + \lambda w_\psi)^\top (w_\phi - w_\psi) \\ &\quad + \frac{1}{2}(w_\phi - w_\psi)^\top B_{x,y}^{\phi,\psi}(w_\phi - w_\psi). \qquad (2) \end{aligned}$$

Note that we regard w_ϕ and w_ψ are flattened into column vectors if necessary. By Assumption 1, we find spectral norms of $A_{x,y}^{\phi,\psi}$ is uniformly bounded with respect to x, y, ϕ, ψ, hence eigen-values are also uniformly bounded. In particular, since $\frac{\lambda}{2}\|w_\phi\|_2^2 \le \mathcal{R}(\phi, w_\phi) \le \mathcal{R}(\phi, 0) \le l_0$, we see $-A_{c_\lambda} I \preceq A_{x,y}^{\phi,\psi} \preceq A_{c_\lambda} I$.

By taking the expectation $\mathbb{E}_\nu$ of the equality (1), we get

$$\mathcal{R}(\psi, w_\phi) = \mathcal{R}(\phi, w_\phi) + \langle \nabla_\phi \mathcal{R}(\phi), \psi - \phi \rangle_{L_2^d(\nu_X)} + \frac{1}{2}\mathbb{E}_\nu[(\psi(x)-\phi(x))^\top A_{x,y}^{\phi,\psi}(\psi(x)-\phi(x))] \qquad (3)$$

and by taking the expectation $\mathbb{E}_\nu$ of the equality (2), we get

$$\mathcal{R}(\psi, w_\phi) = \mathcal{R}(\psi, w_\psi) + \frac{1}{2}(w_\phi - w_\psi)^\top \mathbb{E}_\nu[B_{x,y}^{\phi,\psi}](w_\phi - w_\psi), \qquad (4)$$

where we used $\partial_w \mathcal{R}(\psi, w_\psi) = 0$. By combining equalities (3) and (4), we have

$$\mathcal{R}(\psi) = \mathcal{R}(\phi) + \langle \nabla_\phi \mathcal{R}(\phi), \psi - \phi \rangle_{L_2^d(\nu_X)} + H_\phi(\psi),$$

where

$$H_\phi(\psi) = \frac{1}{2}\mathbb{E}_\nu[(\psi(x)-\phi(x))^\top A_{x,y}^{\phi,\psi}(\psi(x)-\phi(x))] - \frac{1}{2}(w_\phi - w_\psi)^\top \mathbb{E}_\nu[B_{x,y}^{\phi,\psi}](w_\phi - w_\psi).$$

By the uniformly boundedness of $A_{x,y}^{\phi,\psi}$ and the semi-positivity of $B_{x,y}^{\phi,\psi}$, we find $H_\phi(\psi) \le \frac{A_{c_\lambda}}{2}\|\phi - \psi\|_{L_2^d(\nu_X)}^2$.

The other cases can be shown in the same manner, thus, we finish the proof. $\square$

We next show the consistency of functional gradient norms.

Proof of Proposition 2. We now prove the first inequality. Note that the integrand of y'-th element of $\nabla_f \mathcal{L}(f)(x)$ for multiclass logistic loss can be written as

$$\partial_{\zeta_{y'}} l(f(x), y) = -\mathbf{1}[y = y'] + \frac{\exp(f_{y'}(x))}{\sum_{\overline{y}\in\mathcal{Y}} \exp(f_{\overline{y}}(x))}.$$

Therefore, we get

$$\begin{aligned} \|\nabla_f \mathcal{L}(f)\|_{L_1^c(\nu_X)} &= \mathbb{E}_{\nu_X}\|\nabla_f \mathcal{L}(f)(X)\|_2 \\ &= \mathbb{E}_{\nu_X}\|\mathbb{E}_{\nu(Y|X)}[\partial_\zeta(f(X),Y)]\|_2 \\ &= \mathbb{E}_{\nu_X}\left[\sqrt{\sum_{y'\in\mathcal{Y}} (\mathbb{E}_{\nu(Y|X)}[\partial_{\zeta_{y'}}(f(X),Y)])^2}\right] \\ &\ge \frac{1}{\sqrt{c}} \sum_{y'\in\mathcal{Y}} \mathbb{E}_{\nu_X}\left[\left|\mathbb{E}_{\nu(Y|X)}[\partial_{\zeta_{y'}}(f(X),Y)]\right|\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{c}} \sum_{y' \in \mathcal{Y}} \mathbb{E}_{\nu_X} \left[\left| \nu(y'|X) \left(-1 + \frac{\exp(f_{y'}(X))}{\sum_{\overline{y} \in \mathcal{Y}} \exp(f_{\overline{y}}(X))} \right) + \sum_{y \neq y'} \nu(y|X) \frac{\exp(f_{y'}(X))}{\sum_{\overline{y} \in \mathcal{Y}} \exp(f_{\overline{y}}(X))} \right| \right] \\
&= \frac{1}{\sqrt{c}} \sum_{y' \in \mathcal{Y}} \mathbb{E}_{\nu_X} \left[\left| \nu(y'|X) \left(-1 + \frac{\exp(f_{y'}(X))}{\sum_{\overline{y} \in \mathcal{Y}} \exp(f_{\overline{y}}(X))} \right) + (1 - \nu(y'|X)) \frac{\exp(f_{y'}(X))}{\sum_{\overline{y} \in \mathcal{Y}} \exp(f_{\overline{y}}(X))} \right| \right] \\
&= \frac{1}{\sqrt{c}} \sum_{y' \in \mathcal{Y}} \mathbb{E}_{\nu_X} \left[\left| -\nu(y'|X) + \frac{\exp(f_{y'}(X))}{\sum_{\overline{y} \in \mathcal{Y}} \exp(f_{\overline{y}}(X))} \right| \right] \\
&= \frac{1}{\sqrt{c}} \sum_{y' \in \mathcal{Y}} \| - \nu(y'|\cdot) + p_f(y'|\cdot) \|_{L_1(\nu_X)},
\end{aligned}
$$

where for the first inequality we used $(\sum_{i=1}^{c} a_i)^2 \leq c \sum_{i=1}^{c} a_i^2$. Noting that the second inequality in Proposition 2 can be shown in the same way by replacing ν by ν_n, we finish the proof. $\square$

We here give the proof of the following inequality concerning choice of embedding introduced in section 4.

$$
\| T_{k_t,n} \partial_\phi \mathcal{R}_n(\phi_t, w_{t+1}) \|_{k_t}^2 \geq \frac{1}{d} \| \partial_\phi \mathcal{R}_n(\phi_t, w_{t+1}) \|_{L_1^d(\nu_{n,X})}^2 \tag{5}
$$

Proof of (5) . For notational simplicity, we denote by $G_t = \partial_\phi \mathcal{R}_n(\phi_t, w_{t+1})(\cdot)$ and by G_t^i the i-the element of G_t. Then, we get

$$
\begin{aligned}
\| T_{k_t,n} G_t \|_{k_t}^2 &= \langle G_t, T_{k_t,n} G_t \rangle_{L_2^d(\nu_{n,X})} \\
&= \mathbb{E}_{(X,X') \sim \nu_{n,X}^2} [G_t(X)^\top G_t(X') G_t(X')^\top G_t(X) / (\| G_t(X) \|_2 \| G_t(X') \|_2)] \\
&= \sum_{i,j=1}^{d} (\mathbb{E}_{\nu_{n,X}} [G_t^i(X) G_t^j(X) / \| G_t(X) \|_2])^2 \\
&\geq \sum_{i=1}^{d} (\mathbb{E}_{\nu_{n,X}} [G_t^i(X)^2 / \| G_t(X) \|_2])^2 \\
&\geq \frac{1}{d} \mathbb{E}_{\nu_{n,X}} [\| G_t(X)) \|_2]^2 = \frac{1}{d} \| G_t \|_{L_1^d(\nu_{n,X})}^2,
\end{aligned}
$$

where we used $(\sum_{i=1}^{c} a_i)^2 \leq c \sum_{i=1}^{c} a_i^2$. $\square$

B. 2 Empirical risk minimization and generalization bound

In this section, we give the proof of convergence of Algorithm 1 for the empirical risk minimization. We here briefly introduce the kernel function that provides useful bound in our analysis. A kernel function k is a symmetric function $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for arbitrary $s \in \mathbb{N}$ and points $\forall (x_i)_{i=1}^s$, a matrix $(k(x_i, x_j))_{i,j=1}^s$ is positive semi-definite. This kernel defines a reproducing kernel Hilbert space $\mathcal{H}_k$ of functions on $\mathcal{X}$, which has two characteristic properties: (i) for $\forall x \in \mathcal{X}$, a function $k(x, \cdot) : \mathcal{X} \to \mathbb{R}$ is an element of $\mathcal{H}_k$, (ii) for $\forall f \in \mathcal{H}_k$ and $\forall x \in \mathcal{X}$, $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}_k}$, where $\langle , \rangle_{\mathcal{H}_k}$ is the inner-product in $\mathcal{H}_k$. These properties are very important and the latter one is called reproducing property. We extend the inner-product into the product space $\mathcal{H}_k^d$ in a straightforward way, i.e., $\langle f, g \rangle_{\mathcal{H}_k^d} = \sum_{i=1}^{d} \langle f^i, g^i \rangle_{\mathcal{H}_k}$.

The following proposition is useful in our analysis. The first property mean that the notation $\| T_{k_t,n} \nabla \mathcal{R}_n(\phi_t) \|_{k_t}$ provided in the paper is nothing but the norm of $T_{k_t,n} \nabla \mathcal{R}_n(\phi_t)$ by the inner-product $\langle , \rangle_{\mathcal{H}_{k_t}^d}$.

Proposition A. *For a kernel function k, the following hold.*

- $\langle f, g \rangle_{L_2(\nu_X)} = \langle T_k f, g \rangle_{\mathcal{H}_k^d}$ *for* $f \in L_2^d(\nu_X),\ g \in \mathcal{H}_k^d$ *where* $T_k f = \mathbb{E}_{\nu_X}[f(X) k(X, \cdot)]$,
  $\langle f, g \rangle_{L_2(\nu_{n,X})} = \langle T_{k,n} f, g \rangle_{\mathcal{H}_k^d}$ *for* $f \in L_2^d(\nu_{n,X}),\ g \in \mathcal{H}_k^d$ *where* $T_{k,n} f = \mathbb{E}_{\nu_{n,X}}[f(X) k(X, \cdot)]$,
- $\| f \|_{L_2(\nu_X)}^2 \leq \mathbb{E}_{\nu_X}[k(X, X)] \| f \|_{\mathcal{H}_k^d}^2$ *for* $f \in \mathcal{H}_k^d$,
  $\| f \|_{L_2(\nu_{n,X})}^2 \leq \mathbb{E}_{\nu_{n,X}}[k(X, X)] \| f \|_{\mathcal{H}_k^d}^2$ *for* $f \in \mathcal{H}_k^d$.

Proof. We show only the case of ν_X because we can prove the other case in the same manner. For $f \in L_2(\nu_X), g \in \mathcal{H}_k^d$, we get the first property by using reproducing property,

$$\langle f, g \rangle_{L_2(\nu_X)} = \mathbb{E}_{\nu_X}[f(X)\top \langle g, k(X,\cdot)\rangle_{\mathcal{H}_k^d}] = \langle g, T_k f\rangle_{\mathcal{H}_k^d}.$$

We next show the second property as follows. For $\forall f \in \mathcal{H}_k^d$, we get

$$\begin{aligned}
\|f\|_{L_2(\nu_X)}^2 &= \mathbb{E}_{\nu_X}\|f(X)\|_2^2 \\
&= \mathbb{E}_{\nu_X}\| \langle f(\cdot), k(X,\cdot)\rangle_{\mathcal{H}_k^d} \|_2^2 \\
&\leq \mathbb{E}_{\nu_X}\|k(X,\cdot)\|_{\mathcal{H}_k}^2 \|f\|_{\mathcal{H}_k^d}^2 \\
&= \mathbb{E}_{\nu_X}[k(X,X)]\|f\|_{\mathcal{H}_k^d}^2.
\end{aligned}$$

□

We give the proof of Theorem 1 concerning the convergence of functional gradient norms.

Proof of Theorem 1 . When $\eta \leq \frac{1}{A_{c_\lambda} K}$, we have from Proposition 1 and Proposition A,

$$\mathcal{R}_n(\phi_{t+1}, w_{t+2}) \leq \mathcal{R}_n(\phi_t, w_{t+1}) - \frac{\eta}{2}\|T_{k_t,n}\partial_\phi \mathcal{R}_n(\phi_t, w_{t+1})\|_{k_t}^2.$$

By Summing this inequality over $t \in \{0,\ldots,T-1\}$ and dividing by T, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\|T_{k_t,n}\partial_\phi \mathcal{R}_n(\phi_t, w_{t+1})\|_{k_t}^2 \leq \frac{2}{\eta T}\mathcal{R}_n(\phi_0, w_1), \tag{6}$$

where we used $\mathcal{R}_n \geq 0$.

On the other hand, since $\partial_z l(z,y,w) = \partial_z l(w^\top z, y) = w\partial_\zeta l(w^\top z, y)$, it follows that

$$\begin{aligned}
\partial_\phi \mathcal{R}_n(\phi, w)(x) &= \mathbb{E}_{\nu_n(Y|x)}[\partial_z l(\phi(x), y, w)] \\
&= \mathbb{E}_{\nu_n(Y|x)}[w\partial_\zeta l(w^\top \phi(x), y)] \\
&= w\nabla_f \mathcal{L}_n(w^\top \phi)(x).
\end{aligned}$$

Thus, by the assumption on $({w_t}^\top w_t)_{t=0}^{T_0}$, we get for $t \in \{0,\ldots,T-1\}$

$$\begin{aligned}
\|\partial_\phi \mathcal{R}_n(\phi_t, w_{t+1})\|_{L_p^d(\nu_{n,X})} &= \mathbb{E}_{\nu_{n,X}}[\|w_{t+1}\nabla_f \mathcal{L}_n(w_{t+1}^\top \phi_t)(X)\|_2^p]^{1/p} \\
&\geq \sigma \mathbb{E}_{\nu_{n,X}}[\|\nabla_f \mathcal{L}_n(w_{t+1}^\top \phi_t)(X)\|_2^p]^{1/p} \\
&= \sigma\|\nabla_f \mathcal{L}_n(w_{t+1}^\top \phi_t)\|_{L_p^c(\nu_{n,X})}.
\end{aligned} \tag{7}$$

Combining inequalities (6) (7) and Assumption 2, we get

$$\min_{t\in[T]}\|\nabla_f \mathcal{L}_n(w_{t+1}^\top \phi_t)\|_{L_p^c(\nu_X)}^q \leq \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_f \mathcal{L}_n(w_{t+1}^\top \phi_t)\|_{L_p^c(\nu_X)}^q \leq \frac{2}{\eta\gamma\sigma^q T}\mathcal{R}_n(\phi_0, w_1) + \frac{\epsilon}{\sigma^q}.$$

Since $p \geq 1$, we observe $\|\nabla_f \mathcal{L}_n(w_{t+1}^\top \phi_t)\|_{L_1^c(\nu_{n,X})} \leq \|\nabla_f \mathcal{L}_n(w_{t+1}^\top \phi_t)\|_{L_p^c(\nu_{n,X})}$ and we finish the proof. □

We next show Theorem 2 that gives the generalization bound by the margin distribution. To do that, we give an upper-bound on the margin distribution by the functional gradient norm.

Proposition B. *For $\forall \delta > 0$, the following bound holds.*

$$\mathbb{P}_{\nu_n}[m_f(X,Y) \leq \delta] \leq \left(1 + \frac{1}{\exp(-\delta)}\right)\sqrt{c}\|\nabla_f \mathcal{L}_n(f)\|_{L_1^c(\nu_{n,X})}.$$

Proof. If $m_f(x,y) \leq \delta$, then, we see

$$\sum_{y' \neq y} \exp(f_{y'}(x) - f_y(x)) \geq \exp\left(\max_{y' \neq y} f_{y'}(x) - f_y(x)\right) = \exp(-m_f(x,y)) \geq \exp(-\delta).$$

This implies,

$$p_f(y|x) = \frac{1}{1 + \sum_{y' \neq y} \exp(f_{y'}(x) - f_y(x))} \leq \frac{1}{1 + \exp(-\delta)}.$$

Thus, we get by Markov inequality and Proposition 2,

$$\begin{aligned}
\mathbb{P}_{\nu_n}[m_f(X,Y) \leq \delta] &\leq \mathbb{P}_{\nu_n}\left[p_f(Y|X) \leq \frac{1}{1+\exp(-\delta)}\right] \\
&= \mathbb{P}_{\nu_n}\left[1 - p_f(Y|X) \geq \frac{\exp(-\delta)}{1+\exp(-\delta)}\right] \\
&\leq \left(1 + \frac{1}{\exp(-\delta)}\right) \mathbb{E}_{\nu_n}[1 - p_f(Y|X)] \\
&= \left(1 + \frac{1}{\exp(-\delta)}\right) \mathbb{E}_{\nu_n}[\nu_n(Y|X) - p_f(Y|X)] \\
&\leq \left(1 + \frac{1}{\exp(-\delta)}\right) \sum_{y \in \mathcal{Y}} \|\nu_n(y|\cdot) - p_f(y|\cdot)]\|_{L_1(\nu_{n,X})} \\
&\leq \left(1 + \frac{1}{\exp(-\delta)}\right) \sqrt{c} \|\nabla_f \mathcal{L}_n(f)\|_{L_1^c(\nu_{n,X})}.
\end{aligned}$$

□

We prove here Theorem 2.

Proof of Theorem 2 . To proof the theorem, we give the network structure. Note that the connection at the t-th layer is as follows.

$$\phi_{t+1}(x) = \phi_t(x) - \eta D_t \sigma(C_t \phi_t(x)).$$

We define recursively the family of functions $\mathcal{H}_t$ and $\hat{\mathcal{H}}_t$ where each neuron belong: We denote by $P_j \in \mathbb{R}^d$ the projection vector to j-th coordinate.

$$\begin{aligned}
\mathcal{H}_0 &\overset{def}{=} \{P_j : \mathcal{X} \to \mathbb{R} \mid j \in \{1, \ldots, d\}\}, \\
\hat{\mathcal{H}}_t &\overset{def}{=} \{\sigma(c_t^\top \phi_t) : \mathcal{X} \to \mathbb{R} \mid \phi_t \in \mathcal{H}_t^d, c_{t-1} \in \mathbb{R}^d, \|c_{t-1}\|_1 \leq \Lambda\}, \\
\mathcal{H}_{t+1} &\overset{def}{=} \{\phi_t^j - \eta d_t^\top \psi_t : \mathcal{X} \to \mathbb{R} \mid \phi_t^j \in \mathcal{H}_t, \psi_t \in \hat{\mathcal{H}}_t^d, d_t \in \mathbb{R}^d, \|d_t\|_1 \leq \Lambda'\}.
\end{aligned}$$

Then, the family of predictors of $y \in \mathcal{Y}$ can be written as

$$\mathcal{G}_{T-1,y} \overset{def}{=} \{w_y^\top \phi_{T-1} : \mathcal{X} \to \mathbb{R} \mid \phi \in \mathcal{H}_{T-1}^d, w_y \in \mathbb{R}^d, \|w_y\|_1 \leq \Lambda_w\}.$$

Note that $\mathcal{G}_{T-1} = \{(f_y)_{y \in \mathcal{Y}} \mid f_y \in \mathcal{G}_{T-1,y}, y \in \mathcal{Y}\}$.

From these relationships and Lemma D, we get

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{H}_t) &\leq \hat{\mathfrak{R}}_S(\mathcal{H}_{t-1}) + \eta \Lambda' \hat{\mathfrak{R}}_S(\hat{\mathcal{H}}_{t-1}) \\
&\leq (1 + \eta \Lambda' \Lambda L_\sigma) \hat{\mathfrak{R}}_S(\mathcal{H}_{t-1}), \\
\hat{\mathfrak{R}}_S(\mathcal{G}_{T-1,y}) &\leq \Lambda_w \hat{\mathfrak{R}}_S(\mathcal{H}_{T-1}).
\end{aligned}$$

The Rademacher complexity of $\mathcal{H}_0$ is obtained as follows. Since $\|P_j\|_2 = 1$, we have

$$\hat{\mathfrak{R}}_S(\mathcal{H}_0) = \frac{1}{n} \mathbb{E}_{(\sigma_i)_{i=1}^n}\left[\sup_{j \in \{1, \ldots, d\}} \sum_{i=1}^n \sigma_i P_j x_i\right]$$

$$
\begin{aligned}
&\leq \frac{1}{n}\mathbb{E}_{(\sigma_i)_{i=1}^n}\left[\sup_{j\in\{1,\ldots,d\}}\|P_j\|_2\left\|\sum_{i=1}^n \sigma_i x_i\right\|_2\right] \\
&= \frac{1}{n}\mathbb{E}_{(\sigma_i)_{i=1}^n}\left[\left\|\sum_{i=1}^n \sigma_i x_i\right\|_2\right] \\
&\leq \frac{1}{n}\left(\mathbb{E}_{(\sigma_i)_{i=1}^n}\left[\left\|\sum_{i=1}^n \sigma_i x_i\right\|_2^2\right]\right)^{\frac{1}{2}} \\
&= \frac{1}{n}\left(\sum_{i=1}^n \|x_i\|_2^2\right)^{\frac{1}{2}} \leq \frac{\Lambda_\infty}{\sqrt{n}},
\end{aligned}
$$

where we used the independence of σ_i when taking the expectation.

We set $\Pi\mathcal{G}_{T-1} = \{f_y(\cdot) : \mathcal{X} \to | \ f \in \mathcal{G}_{T-1}, y \in \mathcal{Y}\}$. Noting that $\hat{\Re}_S(\Pi\mathcal{G}_{T-1}) \leq \sum_{y\in\mathcal{Y}} \hat{\Re}_S(\mathcal{G}_{T-1,y})$, we get

$$
\hat{\Re}_S(\Pi\mathcal{G}_{T-1}) \leq c\Lambda_w\Lambda_\infty(1+\eta\Lambda\Lambda' L_\sigma)^{T-1}/\sqrt{n}.
$$

Thus, we can finish the proof by applying Proposition B and Lemma E. □

B. 3 Sample-splitting technique

In this subsection, we provide proofs for the convergence analysis of the sample-splitting variant of the method for the expected risk minimization. We first give the statistical error bound on the gap between the empirical and expected functional gradients.

Proof of Proposition 3 . For the probability measure ν, we denote by $\phi_\sharp\nu$ the push-forward measure $(\phi, id)_\sharp\nu$, namely, $(\phi, id)_\sharp\nu$ is the measure that the random variable $(\phi(X), Y)$ follows. We also define $\phi_\sharp\nu_m$ in the same manner. Then, we get

$$
\begin{aligned}
&\|T_k\partial_\phi\mathcal{R}(\phi, w_0) - T_{k,m}\partial_\phi\mathcal{R}_m(\phi, w_0)\|_{L_2^d(\mu)} \\
&\quad= \sqrt{\mathbb{E}_{X'\sim\mu}\|\mathbb{E}_\nu[\partial_z l(\phi(X), Y, w_0)k(X, X')] - \mathbb{E}_{\nu_m}[\partial_z l(\phi(X), Y, w_0)k(X, X')]\|_2^2}. \\
&\quad= \sqrt{\sum_{j=1}^d \mathbb{E}_{X'\sim\mu}|(\mathbb{E}_\nu[\partial_{z_j} l(\phi(X), Y, w_0)\iota(\phi(X)))] - \mathbb{E}_{\nu_m}[\partial_{z_j} l(\phi(X), Y, w_0)\iota(\phi(X))])^\top\iota(\phi(X'))|^2}. \\
&\quad\leq \sqrt{K\sum_{j=1}^d \|\mathbb{E}_\nu[\partial_{z_j} l(\phi(X), Y, w_0)\iota(\phi(X)))] - \mathbb{E}_{\nu_m}[\partial_{z_j} l(\phi(X), Y, w_0)\iota(\phi(X))]\|_2^2} \\
&\quad\leq \sqrt{K\sum_{j=1}^d\sum_{i=1}^D \left|\mathbb{E}_{\phi_\sharp\nu}[\partial_{z_j} l(X, Y, w_0)\iota^i(X))] - \mathbb{E}_{\phi_\sharp\nu_m}[\partial_{z_j} l(X, Y, w_0)\iota^i(X)]\right|^2}. \qquad (8)
\end{aligned}
$$

To derive an uniform bound on (8), we estimate Rademacher complexity of

$$
\mathcal{G}_{ij} \stackrel{def}{=} \{\partial_{z_j} l(x, y, w_0)\iota^i(x) : \mathcal{X}\times\mathcal{Y} \to \mathbb{R} \mid \iota^i \in \mathcal{F}^i\}.
$$

For $(x_l, y_l)_{l=1}^m \subset \mathcal{X}\times\mathcal{Y}$, we set $h_l(r) = r\partial_{z_j} l(x_l, y_l, w_0)$. Since, $|\partial_{z_j} l(x_l, y_l, w_0)| \leq \beta_{\|w_0\|_2}$ by Assumption 3, h_l is $\beta_{\|w_0\|_2}$-Lipschitz continuous. Thus, from Lemma C and Lemma D, there exists M such that for all $i \in \{1, \ldots, D\},\ j \in \{1, \ldots, d\}$,

$$
\hat{\Re}_m(\mathcal{G}_{ij}) = \mathbb{E}_\sigma\left[\sup_{\iota^i\in\mathcal{F}^i}\sum_{l=1}^m \sigma_l h_l(\iota^i(x_l))\right]
$$

$$\leq \beta_{\|w_0\|_2} \mathbb{E}_\sigma \left[\sup_{\iota^i \in \mathcal{F}^i} \sum_{l=1}^m \sigma_l \iota^i(x_l) \right]$$
$$\leq \beta_{\|w_0\|_2} \frac{M}{\sqrt{m}}.$$

Therefore, by applying Lemma B with $\delta = \frac{\rho}{dD}$ for $\forall i, j$ simultaneously, it follows that with probability at least $1-\rho$ for $\forall i, j$

$$\sup_{\iota^i \in \mathcal{F}^i} \left| \mathbb{E}_{\phi_\sharp \nu}[\partial_{z_j} l(X, Y, w_0)\iota^i(X))] - \mathbb{E}_{\phi_\sharp \nu_m}[\partial_{z_j} l(X, Y, w_0)\iota^i(X)] \right| \leq \frac{\beta_{\|w_0\|_2}}{\sqrt{m}} \left(2M + \sqrt{2K \log \frac{2dD}{\rho}} \right). \quad (9)$$

Putting (9) int (8), we get with probability at least $1-\rho$

$$\sup_{\iota \in \mathcal{F}} \|T_k \partial_\phi \mathcal{R}(\phi, w_0) - T_{k,m} \partial_\phi \mathcal{R}_m(\phi, w_0)\|_{L_2^d(\mu)} \leq \beta_{\|w_0\|_2} \sqrt{\frac{KdD}{m}} \left(2M + \sqrt{2K \log \frac{2dD}{\rho}} \right).$$

□

We here prove Theorem 3 by using statistical guarantees of empirical functional gradients.

Proof of Theorem 3 . For notational simplicity, we set $m \leftarrow \lfloor n/T \rfloor$ and $\delta \leftarrow \rho/T$. We first note that

$$\begin{aligned}
\langle \partial_\phi \mathcal{R}(\phi_t, w_0), T_{k_t,m} \partial_\phi \mathcal{R}_m(\phi_t, w_0) \rangle_{L_2^d(\nu_X)} &= \frac{1}{m} \sum_{j=1}^m \mathbb{E}_{\nu_X} [\partial_\phi \mathcal{R}(\phi_t, w_0)(X)^\top \partial_\phi \mathcal{R}_m(\phi_t, w_0)(x_j) k_t(X, x_j)] \\
&= \frac{1}{m} \sum_{j=1}^m T_{k_t} \partial_\phi \mathcal{R}(\phi_t, w_0)(x_j)^\top \partial_\phi \mathcal{R}_m(\phi_t, w_0)(x_j) \\
&= \langle T_{k_t} \partial_\phi \mathcal{R}(\phi_t, w_0), \partial_\phi \mathcal{R}_m(\phi_t, w_0) \rangle_{L_2^d(\nu_{m,X})}.
\end{aligned}$$

Noting that $\|\partial_z l(\phi_t(x_j), y_j, w_0)\|_2 \leq \beta_{\|w_0\|_2}$ by Assumption 1, and applying Proposition 3 for all $t \in \{0, \ldots, T-1\}$ independently, it follows that with probability at least $1 - T\delta$ (i.e., $1-\rho$) for $\forall t \in \{0, \ldots, T-1\}$

$$\begin{aligned}
&\left| \langle \partial_\phi \mathcal{R}(\phi_t, w_0), T_{k_t,m} \partial_\phi \mathcal{R}_m(\phi_t, w_0) \rangle_{L_2^d(\nu_X)} - \langle T_{k_t,m} \partial_\phi \mathcal{R}_m(\phi_t, w_0), \partial_\phi \mathcal{R}_m(\phi_t, w_0) \rangle_{L_2^d(\nu_{m,X})} \right| \\
&\quad \leq \|T_{k_t} \partial_\phi \mathcal{R}(\phi_t, w_0) - T_{k_t,m} \partial_\phi \mathcal{R}_m(\phi_t, w_0)\|_{L_2^d(\nu_{m,X})} \|\partial_\phi \mathcal{R}_m(\phi_t, w_0)\|_{L_2^d(\nu_{m,X})} \\
&\quad \leq \beta_{\|w_0\|_2} \epsilon(m, \delta). \qquad (10)
\end{aligned}$$

We next give the following bound.

$$\|T_{k_t} \partial_\phi \mathcal{R}_m(\phi_t, w_0)\|_{L_2^d(\nu_X)}^2 = \mathbb{E}_{\nu_X} \left\| \frac{1}{m} \sum_{j=1}^m \partial_z l(\phi_t(x_i), y_i, w_0) k_t(x_i, X) \right\|_2^2 \leq \beta_{\|w_0\|_2}^2 K^2. \quad (11)$$

On the other hand, we get by Proposition 1

$$\mathcal{R}(\phi_{t+1}, w_0) \leq \mathcal{R}(\phi_{t+1}, w_0) - \eta \langle \partial_\phi \mathcal{R}(\phi_t, w_0), T_{k_t,m} \partial_\phi \mathcal{R}_m(\phi_t, w_0) \rangle_{L_2^d(\nu_X)} + \frac{\eta^2 A_{\|w_0\|_2}}{2} \|T_{k_t} \partial_\phi \mathcal{R}_m(\phi_t, w_0)\|_{L_2^d(\nu_X)}^2. \quad (12)$$

Combining inequalities (10), (11), and (12), we have with probability at least $1 - T\delta$ for $t \in \{0, \ldots, T-1\}$,

$$\mathcal{R}(\phi_{t+1}, w_0) \leq \mathcal{R}(\phi_{t+1}, w_0) - \eta \|T_{k_t,m} \partial_\phi \mathcal{R}_m(\phi_t, w_0)\|_{k_t}^2 + \eta \beta_{\|w_0\|_2} \epsilon(m, \delta) + \frac{\eta^2 \beta_{\|w_0\|_2}^2 K^2 A_{\|w_0\|_2}}{2}.$$

By Summing this inequality over $t \in \{0, \ldots, T-1\}$ and dividing by T, we get with probability $1 - T\delta$

$$\frac{1}{T}\sum_{t=0}^{T-1} \|T_{k_t,m}\partial_\phi \mathcal{R}_m(\phi_t, w_0)\|_{k_t}^2 \le \frac{\mathcal{R}(\phi_0, w_0)}{\eta T} + \beta_{\|w_0\|_2}\epsilon(m,\delta) + \frac{\eta \beta_{\|w_0\|_2}^2 K^2 A_{\|w_0\|_2}}{2}.$$

Thus by Assumption 2 and the assumption on $w_0^\top w_0$, we get

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla_f \mathcal{L}_m(w_0^\top \phi_t)\|_{L_p^d(\nu_{m,X})}^q \le \frac{1}{\gamma\sigma^q}\left\{\frac{\mathcal{R}(\phi_0, w_0)}{\eta T} + \beta_{\|w_0\|_2}\epsilon(m,\delta) + \frac{\eta \beta_{\|w_0\|_2}^2 K^2 A_{\|w_0\|_2}}{2} + \gamma\epsilon\right\}. \tag{13}$$

To clarify the relationship between $\|\nabla_f \mathcal{L}_m(f)\|_{L_1^c(\nu_{m,X})}$ and $\|\nabla_f \mathcal{L}(f)\|_{L_1^c(\nu_X)}$, we take an expectation of the former term with respect to samples $(X_j, Y_j)_{j=1}^m \sim \nu^m$. Since $\|\partial_\zeta l(\zeta, y)\|_2 \le B$, we obtain

$$\begin{aligned}
\mathbb{E}_{(X_j,Y_j)_{j=1}^m \sim \nu^m} \|\nabla_f \mathcal{L}_m(f)\|_{L_1^c(\nu_{m,X})} &= \mathbb{E}_{(X,Y)\sim\nu_m} \|\partial_\zeta l(f(X), Y)\|_2 \\
&\ge \frac{1}{B}\mathbb{E}_{(X,Y)\sim\nu_m} \|\partial_\zeta l(f(X), Y)\|_2^2 \\
&\ge \frac{1}{B}\mathbb{E}_{\nu_{m,X}} \|\mathbb{E}_{\nu(Y|X)}[\partial_\zeta l(f(X), Y)]\|_2^2 \\
&= \frac{1}{B}\mathbb{E}_{\nu_{m,X}} \|\nabla_f \mathcal{L}(f)(X)\|_2^2 \\
&= \frac{1}{B}\|\nabla_f \mathcal{L}(f)\|_{L_2^c(\nu_X)}^2.
\end{aligned}$$

Hence, applying Hoeffding's inequality with $\delta \leftarrow \rho/T$ to $\mathbb{E}_{(X_j,Y_j)_{j=1}^m\sim\nu^m}\|\nabla_f \mathcal{L}_m(w_0^\top \phi_t)\|_{L_1^c(\nu_{m,X})}$ for all $t \in \{0, \ldots, T-1\}$ independently, we find that with probability $1 - T\delta$ for $\forall t \in \{0, \ldots, T-1\}$,

$$\|\nabla_f \mathcal{L}_m(w_0^\top \phi_t)\|_{L_1^c(\nu_{m,X})} + B\sqrt{\frac{2}{m}\log\frac{1}{\delta}} \ge \mathbb{E}_{\sim\nu^m}\|\nabla_f \mathcal{L}_m(w_0^\top \phi_t)\|_{L_1^c(\nu_{m,X})} \ge \frac{1}{B}\|\nabla_f \mathcal{L}(w_0^\top \phi_t)\|_{L_1^c(\nu_X)}^2, \tag{14}$$

where we used for the last inequality $\|\cdot\|_{L_2^c(\nu_X)}^2 \ge \|\cdot\|_{L_1^c(\nu_X)}^2$.

We set $t_* = \arg\min_{t\in\{0,\ldots,T-1\}} \|\nabla_f \mathcal{L}_m(w_0^\top \phi_t)\|_{L_p^d(\nu_{m,X})}$. Combining inequalities (13) and (14) and noting $p \ge 1$, we get with probability at least $1 - 2T\delta$,

$$\frac{1}{B}\|\nabla_f \mathcal{L}(w_0^\top \phi_{t_*})\|_{L_1^c(\nu_X)}^2 \le B\sqrt{\frac{2}{m}\log\frac{1}{\delta}} + \frac{1}{\gamma^{1/q}\sigma}\left\{\frac{\mathcal{R}(\phi_0, w_0)}{\eta T} + \beta_{\|w_0\|_2}\epsilon(m,\delta) + \frac{\eta \beta_{\|w_0\|_2}^2 K^2 A_{\|w_0\|_2}}{2} + \gamma\epsilon\right\}^{\frac{1}{q}}.$$

Noting that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b > 0$, we finally obtain

$$\|\nabla_f \mathcal{L}(w_0^\top \phi_{t_*})\|_{L_1^c(\nu_X)} \le B\left(\frac{2}{m}\log\frac{1}{\delta}\right)^{\frac{1}{4}} + \sqrt{\frac{B}{\gamma^{1/q}\sigma}}\left\{\frac{\mathcal{R}(\phi_0, w_0)}{\eta T} + \beta_{\|w_0\|_2}\epsilon(m,\delta) + \frac{\eta \beta_{\|w_0\|_2}^2 K^2 A_{\|w_0\|_2}}{2} + \gamma\epsilon\right\}^{\frac{1}{2q}}.$$

Recalling that $m \leftarrow \lfloor n/T \rfloor$ and $\delta \leftarrow \rho/T$, the proof is finished. □

References

[Dud99] Richard M Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

[KP02] Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

[vdVW96] AW van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.