Appendix

The proofs and analysis of the main theorems

Proof of Theorem 1.

Consider a ground truth $\boldsymbol{y}$, a target operator h, an optimal MLP operator f^*_{mlp}, and an VQC-MLPNet operator $f^*_{\boldsymbol{\theta}}$. The approximation error ϵ_{app} can be upper bounded using the triangle inequality as:

$$\epsilon_{\text{app}} = \mathcal{R}(f_{\boldsymbol{\theta}^*}) - \mathcal{R}(h^*) \leq |\mathcal{R}(f_{\boldsymbol{\theta}^*}) - \mathcal{R}(f^*_{\text{mlp}})| + |\mathcal{R}(f^*_{\text{mlp}}) - \mathcal{R}(h^*)|. \tag{1}$$

Moreover, the expected risk $\mathcal{R}$ is taken as the cross-entropy loss function such that:

$$|\mathcal{R}(f_{\boldsymbol{\theta}^*}) - \mathcal{R}(f^*_{\text{mlp}})| + |\mathcal{R}(f^*_{\text{mlp}}) - \mathcal{R}(h^*)| \leq L_{\text{ce}}\|f_{\boldsymbol{\theta}^*} - f^*_{\text{mlp}}\|_2 + L_{\text{ce}}\|f^*_{\text{mlp}} - h^*\|_2, \tag{2}$$

where L_{ce} is a Lipschitz constant induced by the gradient of the cross-entropy loss function with a softmax activation.

Leveraging Cybenco's universal approximation theory, the term we further derive:

$$\|f_{\boldsymbol{\theta}^*} - f^*_{\text{mlp}}\|_2 \leq \frac{C_1}{\sqrt{M}}. \tag{3}$$

Next, we define f^*_{mlp} and $f^*_{\boldsymbol{\theta}}$ operators explicitly:

$$f^*_{\text{mlp}}(\mathbf{x}) = \mathbf{W}^{(2)\top}\sigma(\mathbf{W}^{(1)\top}_*\mathbf{x}), \quad f_{\boldsymbol{\theta}^*}(\mathbf{x}) = \mathbf{W}^{(2)\top}\sigma(\hat{\mathbf{W}}^{(1)\top}\mathbf{x}), \tag{4}$$

and assuming a Lipschitz-continuous activation function $\sigma(\cdot)$, we apply the Cauchy-Swartz inequality, resulting in:

$$\begin{aligned}
\|f_{\boldsymbol{\theta}^*} - f^*_{\text{mlp}}\|_2 &= \left\|\mathbf{W}^{(2)\top}\sigma(\hat{\mathbf{W}}^{(1)\top}\mathbf{x}) - \mathbf{W}^{(2)\top}\sigma(\mathbf{W}^{(1)\top}_*\mathbf{x})\right\|_2 \\
&\leq \max_{1\leq j\leq J}\left(\sum_{m=1}^{M}\left|\mathbf{w}^{(2)}_m(j)\right|\right)\left\|\sigma(\hat{\mathbf{W}}^{(1)}\mathbf{x}) - \sigma(\hat{\mathbf{W}}^{(1)}_*\mathbf{x})\right\|_2 \\
&\leq \max_{1\leq j\leq J}\left(\sum_{m=1}^{M}\left|\mathbf{w}^{(2)}_m(j)\right|\right)\|\mathbf{x}\|_2\left\|\hat{\mathbf{W}}^{(1)} - \mathbf{W}^{(1)}_*\right\|_2,
\end{aligned} \tag{5}$$

where $\left\|\hat{\mathbf{W}}^{(1)} - \mathbf{W}^{(1)}_*\right\|_2$ is an induced norm of matrix as:

$$\left\|\hat{\mathbf{W}}^{(1)} - \mathbf{W}^{(1)}_*\right\|_2 = \sup_{\|\mathbf{x}\|_2\leq 1}\left\|(\hat{\mathbf{W}}^{(1)} - \mathbf{W}^{(1)}_*)\mathbf{x}\right\|_2. \tag{6}$$

Next, we consider $\hat{\mathbf{W}}^{(1)}_Q$ as:

$$\hat{\mathbf{W}}^{(1)}_Q = \left[|\hat{\mathbf{w}}^{(1)}_1\rangle, |\hat{\mathbf{w}}^{(1)}_2\rangle, ..., |\hat{\mathbf{w}}^{(1)}_D\rangle\right] \tag{7}$$

Then, using the triangle inequality, we have:

$$\begin{aligned}\left\|\hat{\mathbf{W}}^{(1)}-\mathbf{W}_{*}^{(1)}\right\|_2 &= \left\|\hat{\mathbf{W}}^{(1)}-\hat{\mathbf{W}}_{Q}^{(1)}+\hat{\mathbf{W}}_{Q}^{(1)}-\mathbf{W}_{*}^{(1)}\right\|_2 \\ &\leq \underbrace{\left\|\hat{\mathbf{W}}^{(1)}-\hat{\mathbf{W}}_{Q}^{(1)}\right\|_2}_{r_{\text{enc}}}+\underbrace{\left\|\hat{\mathbf{W}}_{Q}^{(1)}-\mathbf{W}_{*}^{(1)}\right\|_2}_{r_{\text{exp}}}.\end{aligned} \tag{8}$$

On the one hand, the term $\left\|\hat{\mathbf{W}}^{(1)}-\hat{\mathbf{W}}_{Q}^{(1)}\right\|_2$ corresponds to the error associated with amplitude encoding. The $L_{2,2}$ group norm of $\left\|\hat{\mathbf{W}}^{(1)}-\hat{\mathbf{W}}_{Q}^{(1)}\right\|_2$ is:

$$\left\|\hat{\mathbf{W}}^{(1)}-\hat{\mathbf{W}}_{Q}^{(1)}\right\|_2=\left[\sum_{d=1}^{D}\left\|\hat{\mathbf{w}}_{d}^{(1)}-|\hat{\mathbf{w}}_{d}^{(1)}\rangle\right\|_2^2\right]^{\frac{1}{2}}. \tag{9}$$

By the definition of amplitude encoding, it requires normalized input vectors that cannot preserve all the information from the classical data, leading to the encoding error of each $\left\|\hat{\mathbf{w}}_{d}^{(1)}-|\hat{\mathbf{w}}_{d}^{(1)}\rangle\right\|_2^2$ that scales as:

$$\left\|\hat{\mathbf{w}}_{d}^{(1)}-|\hat{\mathbf{w}}_{d}^{(1)}\rangle\right\|_2^2\leq\mathcal{O}\left(\frac{1}{2^{\beta_1 U}}\right), \tag{10}$$

where the constant $\beta_1\in(0,1]$ is a scaling factor. The above term means that for a scaling factor $\beta\in(0,\frac{1}{2}]$, we have:

$$r_{\text{enc}}=\left[\sum_{d=1}^{D}\left\|\hat{\mathbf{w}}_{d}^{(1)}-|\hat{\mathbf{w}}_{d}^{(1)}\rangle\right\|_2^2\right]^{\frac{1}{2}}\leq\mathcal{O}\left(\frac{1}{2^{\beta U}}\right). \tag{11}$$

On the other hand, we consider the expressive error r_{exp} as the term of $L_{2,2}$ group norm:

$$r_{\text{exp}}=\left\|\hat{\mathbf{W}}_{Q}^{(1)}-\mathbf{W}_{*}^{(1)}\right\|_2=\left[\sum_{d=1}^{D}\left\||\hat{\mathbf{w}}_{d}^{(1)}\rangle-\mathbf{w}_{*}^{(1)}\right\|_2^2\right]^{\frac{1}{2}}. \tag{12}$$

For each $\left\||\hat{\mathbf{w}}_{d}^{(1)}\rangle-\mathbf{w}_{*}^{(1)}\right\|_2^2$, given a decaying factor $\alpha_1>0$, since the target matrix W^* is smooth, we leverage the quantum Fourier-like expansion technique and exponential decay of coefficients, we have:

$$\left\||\hat{\mathbf{w}}_{d}^{(1)}\rangle-\mathbf{w}_{*}^{(1)}\right\|_2^2\leq\sum_{l>L}|c_l|\leq\mathcal{O}\left(e^{-\alpha_1 L}\right). \tag{13}$$

Then, for a constant $C_2>0$ and $\alpha=2\alpha_1>0$, we have:

$$r_{\text{exp}}=\left[\sum_{d=1}^{D}\left\||\hat{\mathbf{w}}_{d}^{(1)}\rangle-\mathbf{w}_{*}^{(1)}\right\|_2^2\right]^{\frac{1}{2}}\leq\mathcal{O}\left(e^{-\alpha L}\right). \tag{14}$$

To sum up, for two constants $C_2 > 0$ and $C_3 > 0$, we have:

$$\begin{aligned} \|f_{\boldsymbol{\theta}^*} - f^*_{\text{mlp}}\|_2 &\leq \max_{1\leq j\leq J}\left(\sum_{m=1}^{M}\left|\mathbf{w}_m^{(2)}(j)\right|\right)\|\mathbf{x}\|_2\left\|\hat{\mathbf{W}}^{(1)} - \mathbf{W}_*^{(1)}\right\|_2 \\ &\leq C_2 e^{-\alpha L} + \frac{C_3}{2^{\beta U}}. \end{aligned} \tag{15}$$

Finally, we obtain the upper bound as:

$$\epsilon_{\text{app}} \leq \|f_{\boldsymbol{\theta}^*} - f^*_{\text{mlp}}\|_2 + \|f^*_{\text{mlp}} - h^*\|_2 \leq \frac{C_1}{\sqrt{M}} + C_2 e^{-\alpha L} + \frac{C_3}{2^{\beta U}}. \tag{16}$$

The NTK Technique for VQC-MLPNet's Trainability. To analyze the trainability of the MLP-VQC model, we consider the operator f_{vm} with a quantum-enhanced weight $\hat{\mathbf{W}}^{(1)} = f_{\text{lin}} \circ f_{\text{vqc}} \circ f_{\text{ae}}(\mathbf{W}^{(1)})$, where f_{vqc} is parameterized by $\boldsymbol{\theta}_{\text{vqc}} = \{\alpha_{1:U}, \beta_{1:U}, \gamma_{1:U}\}$, and the final layer uses classical weights $\boldsymbol{\theta}_{W^{(2)}}$.

The MLP-VQC operator is given by:

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{\sqrt{M}}\sum_{m=1}^{M}\mathbf{w}_m^{(2)}\sigma\left(\langle\hat{\mathbf{w}}_m^{(1)}, \mathbf{x}\rangle\right). \tag{17}$$

As for the gradient descent of f_{vm} w.r.t. α_u, β_u, γ_u and $\mathbf{w}_m^{(2)}$, we have:

$$\frac{\partial f_{\boldsymbol{\theta}}}{\partial \alpha_u} = \frac{1}{\sqrt{M}}\sum_{m=1}^{M}\mathbf{w}_m^{(2)}\sigma'\left(\langle\hat{\mathbf{w}}_m^{(1)}, \mathbf{x}\rangle\right)\left(\left\langle\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\alpha_u}, \mathbf{x}\right\rangle\right) \tag{18}$$

$$\frac{\partial f_{\boldsymbol{\theta}}}{\partial \beta_u} = \frac{1}{\sqrt{M}}\sum_{m=1}^{M}\mathbf{w}_m^{(2)}\sigma'\left(\langle\hat{\mathbf{w}}_m^{(1)}, \mathbf{x}\rangle\right)\left(\left\langle\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\beta_u}, \mathbf{x}\right\rangle\right) \tag{19}$$

$$\frac{\partial f_{\boldsymbol{\theta}}}{\partial \gamma_u} = \frac{1}{\sqrt{M}}\sum_{m=1}^{M}\mathbf{w}_m^{(2)}\sigma'\left(\langle\hat{\mathbf{w}}_m^{(1)}, \mathbf{x}\rangle\right)\left(\left\langle\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\gamma_u}, \mathbf{x}\right\rangle\right) \tag{20}$$

$$\frac{\partial f_{\text{vm}}}{\partial \boldsymbol{\theta}_{W^{(2)}}} = \frac{1}{\sqrt{M}}\sigma\left(\langle\hat{\mathbf{w}}_m^{(1)}, \mathbf{x}\rangle\right) \tag{21}$$

In particular, $\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\alpha_u} = f_{\text{lin}} \circ \frac{\partial f_{\text{vqc}}}{\partial\alpha_u} \circ f_{\text{ae}}(\mathbf{w}_m^{(1)})$ relies on the VQC's architecture and is not related to the weight parameters $\mathbf{w}_m^{(1)}$. Accordingly, we define a constant $C_{\boldsymbol{\alpha}}$ and obtain:

$$\sum_{u=1}^{U}\left(\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\alpha_u}\right)^{\top}\left(\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\alpha_u}\right) = C_{\boldsymbol{\alpha}}. \tag{22}$$

Similarly, given the constants $C_{\boldsymbol{\beta}}$ and $C_{\boldsymbol{\gamma}}$, we obtain the following terms:

$$\sum_{u=1}^{U}\left(\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\beta_u}\right)^{\top}\left(\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\beta_u}\right) = C_{\boldsymbol{\beta}}, \tag{23}$$

$$\sum_{u=1}^{U}\left(\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\gamma_u}\right)^{\top}\left(\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\gamma_u}\right)=C_{\boldsymbol{\gamma}}. \tag{24}$$

Next, we define the NTK $\mathcal{K}_{\text{vm}}$ as:

$$\mathcal{K}_{\text{vm}}=\mathcal{K}_{\text{vqc}}+\mathcal{K}_{W^{(2)}}. \tag{25}$$

Since the VQC component of the NTK can be decomposed into contributions from the parameterized quantum gates, for any two data vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, the related NTK $\mathcal{K}_{\text{vqc}}(\mathbf{x}_1,\mathbf{x}_2)$ can be further decomposed as:

$$\mathcal{K}_{\text{vqc}}(\mathbf{x}_1,\mathbf{x}_2)=\mathcal{K}_{\text{vqc}}^{(\boldsymbol{\alpha})}(\mathbf{x}_1,\mathbf{x}_2)+\mathcal{K}_{\text{vqc}}^{(\boldsymbol{\beta})}(\mathbf{x}_1,\mathbf{x}_2)+\mathcal{K}_{\text{vqc}}^{(\boldsymbol{\gamma})}(\mathbf{x}_1,\mathbf{x}_2). \tag{26}$$

Each component is computed over the quantum-enhanced features. For instance, given a constant $C_{\boldsymbol{\alpha}}$, the $\boldsymbol{\alpha}$-parameter NTK contribution is:

$$\begin{aligned}\mathcal{K}_{\text{vqc}}^{(\boldsymbol{\alpha})}(\mathbf{x}_1,\mathbf{x}_2)&=\sum_{u=1}^{U}\left\langle\frac{\partial f_{\text{vm}}}{\partial\alpha_u},\frac{\partial f_{\text{vm}}}{\partial\alpha_u}\right\rangle\\&=\frac{1}{M}\sum_{m=1}^{M}(\mathbf{w}_m^{(2)})^2\sigma'(\langle\hat{\mathbf{w}}_m^{(1)},\mathbf{x}_1\rangle)\sigma'(\langle\hat{\mathbf{w}}_m^{(1)},\mathbf{x}_2\rangle)\sum_{u=1}^{U}\left\langle\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\alpha_u}\mathbf{x}_1,\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\alpha_u}\mathbf{x}_2\right\rangle\\&=\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{w}_k^{(2)}\right)^2\sigma'\left(\left\langle\hat{\mathbf{w}}_m^{(1)},\mathbf{x}_1\right\rangle\right)\sigma'\left(\left\langle\hat{\mathbf{w}}_m^{(1)},\mathbf{x}_2\right\rangle\right)\langle\mathbf{x}_1,C_{\boldsymbol{\alpha}}\mathbf{x}_2\rangle,\end{aligned} \tag{27}$$

With similar expressions for $\mathcal{K}_{\text{vqc}}^{(\boldsymbol{\beta})}$ and $\mathcal{K}_{\text{vqc}}^{(\boldsymbol{\gamma})}$ using constants $C_{\boldsymbol{\beta}}$ and $C_{\boldsymbol{\gamma}}$, we obtain

$$\begin{aligned}\mathcal{K}_{\text{vqc}}^{(\boldsymbol{\beta})}(\mathbf{x}_1,\mathbf{x}_2)&=\sum_{u=1}^{U}\left\langle\frac{\partial f_{\text{vm}}}{\partial\beta_u},\frac{\partial f_{\text{vm}}}{\partial\beta_u}\right\rangle\\&\frac{1}{M}\sum_{m=1}^{M}(\mathbf{w}_m^{(2)})^2\sigma'(\langle\hat{\mathbf{w}}_m^{(1)},\mathbf{x}_1\rangle)\sigma'(\langle\hat{\mathbf{w}}_m^{(1)},\mathbf{x}_2\rangle)\sum_{u=1}^{U}\left\langle\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\beta_u}\mathbf{x}_1,\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\beta_u}\mathbf{x}_2\right\rangle\\&=\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{w}_m^{(2)}\right)^2\sigma'\left(\left\langle\hat{\mathbf{w}}_m^{(1)},\mathbf{x}_1\right\rangle\right)\sigma'\left(\left\langle\hat{\mathbf{w}}_m^{(1)},\mathbf{x}_2\right\rangle\right)\langle\mathbf{x}_1,C_{\boldsymbol{\beta}}\mathbf{x}_2\rangle,\end{aligned} \tag{28}$$

$$\begin{aligned}\mathcal{K}_{\text{vqc}}^{(\boldsymbol{\gamma})}(\mathbf{x}_1,\mathbf{x}_2)&=\sum_{u=1}^{U}\left\langle\frac{\partial f_{\text{vm}}}{\partial\gamma_u},\frac{\partial f_{\text{vm}}}{\partial\gamma_u}\right\rangle\\&=\frac{1}{M}\sum_{m=1}^{M}(\mathbf{w}_m^{(2)})^2\sigma'(\langle\hat{\mathbf{w}}_m^{(1)},\mathbf{x}_1\rangle)\sigma'(\langle\hat{\mathbf{w}}_m^{(1)},\mathbf{x}_2\rangle)\sum_{u=1}^{U}\left\langle\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\gamma_u}\mathbf{x}_1,\frac{\partial\hat{\mathbf{w}}_m^{(1)}}{\partial\gamma_u}\mathbf{x}_2\right\rangle\\&=\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{w}_m^{(2)}\right)^2\sigma'\left(\left\langle\hat{\mathbf{w}}_m^{(1)},\mathbf{x}_1\right\rangle\right)\sigma'\left(\left\langle\hat{\mathbf{w}}_m^{(1)},\mathbf{x}_2\right\rangle\right)\langle\mathbf{x}_1,C_{\boldsymbol{\gamma}}\mathbf{x}_2\rangle.\end{aligned} \tag{29}$$

Moreover, the classical linear layer NTK is given by:

$$\mathcal{K}_{W^{(2)}}(\mathbf{x}_1, \mathbf{x}_2) = \left\langle \frac{\partial f_{\text{vm}}}{\partial \boldsymbol{\theta}_{W^{(2)}}}, \frac{\partial f_{\text{vm}}}{\partial \boldsymbol{\theta}_{W^{(2)}}} \right\rangle = \frac{1}{M} \sum_{m=1}^{M} \sigma^{'} \left(\langle \hat{\mathbf{w}}_m^{(1)}, \mathbf{x}_1 \rangle \right) \sigma^{'} \left(\langle \hat{\mathbf{w}}_m^{(1)}, \mathbf{x}_2 \rangle \right). \tag{30}$$

In the infinite-width limit $M \to \infty$, the NTKs converge to their expected values under the initialization distribution. This yields constant kernels for each component (Eqs. 31-32), enabling analytical tractability of convergence behavior.

$$\mathcal{K}_{\text{vqc}}^{(\boldsymbol{\alpha})}(\mathbf{x}_1, \mathbf{x}_2) \xrightarrow{M \to \infty} \mathbb{E}_{(\boldsymbol{\theta}_{\text{vqc}}, \boldsymbol{\theta}_{W^{(2)}})} \left[(\mathbf{w}_m^{(2)})^2 \sigma^{'}(\langle \hat{\mathbf{w}}_m^{(1)}, \mathbf{x}_1 \rangle) \sigma^{'}(\langle \hat{\mathbf{w}}_m^{(1)}, \mathbf{x}_2 \rangle) \langle \mathbf{x}_1, C_{\boldsymbol{\alpha}} \mathbf{x}_2 \rangle \right], \tag{31}$$

$$\mathcal{K}_{\text{vqc}}^{(\boldsymbol{\beta})}(\mathbf{x}_1, \mathbf{x}_2) \xrightarrow{M \to \infty} \mathbb{E}_{(\boldsymbol{\theta}_{\text{vqc}}, \boldsymbol{\theta}_{W^{(2)}})} \left[(\mathbf{w}_m^{(2)})^2 \sigma^{'}(\langle \hat{\mathbf{w}}_m^{(1)}, \mathbf{x}_1 \rangle) \sigma^{'}(\langle \hat{\mathbf{w}}_m^{(1)}, \mathbf{x}_2 \rangle) \langle \mathbf{x}_1, C_{\boldsymbol{\beta}} \mathbf{x}_2 \rangle \right], \tag{32}$$

$$\mathcal{K}_{\text{vqc}}^{(\boldsymbol{\gamma})}(\mathbf{x}_1, \mathbf{x}_2) \xrightarrow{M \to \infty} \mathbb{E}_{(\boldsymbol{\theta}_{\text{vqc}}, \boldsymbol{\theta}_{W^{(2)}})} \left[(\mathbf{w}_m^{(2)})^2 \sigma^{'}(\langle \hat{\mathbf{w}}_m^{(1)}, \mathbf{x}_1 \rangle) \sigma^{'}(\langle \hat{\mathbf{w}}_m^{(1)}, \mathbf{x}_2 \rangle) \langle \mathbf{x}_1, C_{\boldsymbol{\gamma}} \mathbf{x}_2 \rangle \right], \tag{33}$$

$$\mathcal{K}_{W^{(2)}}(\mathbf{x}_1, \mathbf{x}_2) \xrightarrow{M \to \infty} \mathbb{E}_{\boldsymbol{\theta}_{\text{vqc}}} \left[\sigma^{'} \left(\langle \hat{\mathbf{w}}_m^{(1)}, \mathbf{x}_1 \rangle \right) \sigma^{'} \left(\langle \hat{\mathbf{w}}_m^{(1)}, \mathbf{x}_2 \rangle \right) \right]. \tag{34}$$

Thus, the over-parameterized regime (large-width) ensures the NTK $\mathcal{K}_{\text{vm}}$ remains nearly constant throughout training. We then compute the minimum eigenvalue $\lambda_{\min}(\mathcal{K}_{\text{vm}})$. Since classical kernels often have better-conditioned NTKs, introducing the classical component can boost the lowest eigenvalue: $\lambda_{\min}(\mathcal{K}_{\text{vm}}) \gg \lambda_{\min}(\mathcal{K}_{W^{(2)}})$.

Next, we prove the upper bound on the optimization error based on the constant NTK $\mathcal{K}_{\text{vm}}$ and $\lambda_{\min}(\mathcal{K}_{\text{vm}})$. Given the set of training data $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)\}$, we aim to minimize the cross-entropy loss:

$$\hat{\mathcal{R}}(f_{\boldsymbol{\theta}}) = -\frac{1}{N} \sum_{n=1}^{N} y_n \log \left(\sigma_c(f_{\boldsymbol{\theta}}(\mathbf{x}_n)) \right), \tag{35}$$

where $\sigma_c(\cdot)$ denotes the softmax probabilities. Since the parameter update follows gradient flow dynamics:

$$\frac{d\boldsymbol{\theta}}{dt} = -\nabla_{\boldsymbol{\theta}} \hat{\mathcal{R}}(f_{\boldsymbol{\theta}}), \tag{36}$$

we have:

$$\frac{d}{dt} f_{\boldsymbol{\theta}}(\mathbf{X}) = \frac{df_{\boldsymbol{\theta}}}{d\boldsymbol{\theta}} \frac{d\boldsymbol{\theta}}{dt} = -\mathcal{K}_{\text{vm}} \cdot \left(\sigma_c(f_{\boldsymbol{\theta}}(\mathbf{X})) - \mathbf{y} \right), \tag{37}$$

where we define the data matrix $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 ... \mathbf{x}_N]$ and the label vector $\mathbf{y} = [y_1 y_2 ... y_N]^\top$.

Furthermore, due to the nonlinearity of the activation function $\sigma_{\text{c}}(\cdot)$, we further simplify it by expanding the softmax around the initial predictions. Typically, this linearization approximation is:

$$\sigma_c(f_{\boldsymbol{\theta}_t}(\mathbf{X})) \approx \sigma_c(f_{\boldsymbol{\theta}_0}(\mathbf{X})) + \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X})^\top \left(f_{\boldsymbol{\theta}_t}(\mathbf{X}) - f_{\boldsymbol{\theta}_0}(\mathbf{X}) \right), \tag{38}$$

Now we have a linear, time-invariant ordinary differential equation (ODE):

$$\frac{d}{dt}f_{\boldsymbol{\theta}_t}(\mathbf{X}) \approx -\mathcal{K}_{\text{vm}} \cdot \left([\sigma_c(f_{\boldsymbol{\theta}_0}(\mathbf{X})) - \mathbf{y}] + \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X})^\top (f_{\boldsymbol{\theta}_t}(\mathbf{X}) - f_{\boldsymbol{\theta}_0}(\mathbf{X})) \right). \tag{39}$$

Define the perturbation around initialization as $u(t) = f_{\boldsymbol{\theta}_t}(\mathbf{X}) - f_{\boldsymbol{\theta}_0}(\mathbf{X})$. Then, the linearized gradient flow dynamics can be written explicitly:

$$\frac{d}{dt}u(t) = -\mathcal{K}_{\text{vm}} \cdot \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}) u(t) - \mathcal{K}_{\text{vm}} \cdot (\sigma_c(f_{\boldsymbol{\theta}_0}(\mathbf{X})) - \mathbf{y}), \tag{40}$$

which becomes a linear, time-invariant ODE system:

- Initial condition: $u(0) = 0$.
- Define the matrices clearly:

$$A = \mathcal{K}_{\text{vm}} \cdot \nabla_{\boldsymbol{\theta}}(f_{\boldsymbol{\theta}_0}(\mathbf{X})), \qquad b = \mathcal{K}_{\text{vm}} \cdot (\sigma_c(f_{\boldsymbol{\theta}_0}(\mathbf{X})) - \mathbf{y}). \tag{41}$$

Then, the solution for the prediction evolution is:

$$u(t) = -A^{-1}(I - e^{-At})b. \tag{42}$$

Thus, we obtain:

$$f_{\boldsymbol{\theta}_t}(\mathbf{X}) = f_{\boldsymbol{\theta}_0}(\mathbf{X}) - (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}))^{-1} \left(I - e^{-\mathcal{K}_{\text{vm}} \cdot \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}) t} \right) (\sigma_c(f_{\boldsymbol{\theta}_0}(\mathbf{X})) - \mathbf{y}). \tag{43}$$

To minimize $\tilde{f}_{\boldsymbol{\theta}_t}(\mathbf{X})$, we obtain the optimal $\boldsymbol{\theta}^*$ such that

$$f_{\boldsymbol{\theta}^*}(\mathbf{X}) = f_{\boldsymbol{\theta}_\infty}(\mathbf{X}) = f_{\boldsymbol{\theta}_0}(\mathbf{X}) - (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}))^{-1} (\sigma_c(f_{\boldsymbol{\theta}_0}(\mathbf{X})) - \mathbf{y}). \tag{44}$$

As for the optimization error ϵ_{opt}, at epoch t, we set $\boldsymbol{\theta}_t = \hat{\boldsymbol{\theta}}$ and further derive that:

$$\begin{aligned} \sup_{\hat{\boldsymbol{\theta}} \in \Theta} \left(\hat{\mathcal{R}}(f_{\hat{\boldsymbol{\theta}}}) - \hat{\mathcal{R}}(f_{\boldsymbol{\theta}^*}) \right) &\leq L_{\text{ce}} \| f_{\boldsymbol{\theta}_t}(\mathbf{X}) - f_{\boldsymbol{\theta}^*}(\mathbf{X}) \|_2 \\ &= L_{\text{ce}} \left\| (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}))^{-1} e^{-\mathcal{K}_{\text{vm}} \cdot \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}) t} (\sigma_c(f_{\boldsymbol{\theta}_0}(\mathbf{X})) - \mathbf{y}) \right\|_2 \\ &\leq L_{\text{ce}} \left\| (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}))^{-1} \right\|_2 \left\| e^{-\mathcal{K}_{\text{vm}} \cdot \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}) t} \right\|_2 \| \sigma_c(f_{\boldsymbol{\theta}_0}(\mathbf{X})) - \mathbf{y} \|_2 . \end{aligned} \tag{45}$$

Using the technique of eigen-decomposition: $e^{-\mathcal{K}_{\text{vm}} \cdot \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}) t} = U e^{-\Lambda t} U^\top$ and $e^{-\Lambda t} = \text{diag}(e^{-\lambda_1 t}, .., e^{-\lambda_n t})$, our expression becomes simpler:

$$\sup_{\hat{\boldsymbol{\theta}} \in \Theta} \left(\hat{\mathcal{R}}(f_{\hat{\boldsymbol{\theta}}}) - \hat{\mathcal{R}}(f_{\boldsymbol{\theta}^*}) \right) \leq L_{\text{ce}} \left\| (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}))^{-1} \right\|_2 \left\| U e^{-\Lambda t} U^\top \right\|_2 \| \sigma_c(f_{\boldsymbol{\theta}_0}(\mathbf{X})) - \mathbf{y} \|_2 . \tag{46}$$

Since U is orthogonal, it doesn't affect the operator norm ($\|U\|_2 = 1$), and we use the tricks:

$$\left\| (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}))^{-1} \right\|_2 = \frac{1}{\lambda_{\min}(\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}))}, \tag{47}$$

$$\left\|e^{-\mathcal{K}_{\text{vm}}\cdot\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X})t}\right\|_2 = \max_i e^{-\lambda_i\left(\mathcal{K}_{\text{vm}}\cdot\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X})\right)t} = e^{-\lambda_{\min}\left(\mathcal{K}_{\text{vm}}\cdot\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X})\right)t}. \tag{48}$$

Then, we can upper bound the optimization error ϵ_{opt} as:

$$\epsilon_{\text{opt}} = \sup_{\hat{\boldsymbol{\theta}}\in\Theta}\left(\hat{\mathcal{R}}(f_{\hat{\boldsymbol{\theta}}}) - \hat{\mathcal{R}}(f_{\boldsymbol{\theta}^*})\right) \leq L_{\text{ce}}\frac{\|\sigma_c(f_{\boldsymbol{\theta}_0}(\mathbf{X})) - \mathbf{y}\|_2}{\lambda_{\min}(\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}))}e^{-\lambda_{\min}\left(\mathcal{K}_{\text{vm}}\cdot\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X})\right)t}. \tag{49}$$

In practice, since the term $\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{x})$ is typically bounded and stable, the primary determinant of convergence rate is the minimum eigenvalue $\lambda_{\min}(\mathcal{K}_{\text{vm}})$. Thus, we simplify the above term as:

$$\epsilon_{\text{opt}} = \sup_{\hat{\boldsymbol{\theta}}\in\Theta}\left(\hat{\mathcal{R}}(f_{\hat{\boldsymbol{\theta}}}) - \hat{\mathcal{R}}(f_{\boldsymbol{\theta}^*})\right) \leq C_0 e^{-\lambda_{\min}(\mathcal{K}_{\text{vm}})t}, \tag{50}$$

where the constant $C_0 = \frac{L_{\text{ce}}\|\sigma_c(f_{\boldsymbol{\theta}_0}(\mathbf{X}))-\mathbf{y}\|_2}{\lambda_{\min}(\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{X}))}$.