Generalization Analysis for Multi-Dimensional Classification

1. Tight $\widetilde{O}(\sqrt{\frac{q}{n}})$ Bounds for Multi-Dimensional Classification with ℓ_2 Lipschitz Loss

With the relevant definitions in the paper, we develop the following novel tight vector-contraction inequality for ℓ_2 Lipschitz loss:

Lemma 1.1. *Let $\mathcal{F}$ be the class of the multi-dimensional classification defined by (1) and (2). Let Assumptions 3.1 and 3.2 hold. Given a dataset D of size n. Then, we have*

$$\hat{\Re}_D(\mathcal{L}) \leq \frac{576B\sqrt{q}}{\sqrt{n}} + 2q\mu 48^2\sqrt{k}\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))\left(1+\left(1+\log_2(8en^2q^2k^2)\log\sqrt{nq}\right)\log^{\frac{1}{2}}(nq)\log\frac{M\sqrt{n}}{\mu B}\right),$$

where $\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))$ is the worst-case Rademacher complexity of the inner decomposition function class.

Proof. Here we no longer use Sudakov's minoration and the relationship between Rademacher and Gaussian complexity to prove the $\widetilde{O}(\sqrt{q}/\sqrt{n})$ bound for Multi-Dimensional Classification with ℓ_2 Lipschitz Loss, because they prevent us from obtaining $\widetilde{O}(\sqrt{q}/\sqrt{n})$ bounds with no dependency on k. The main reason is that we find that a factor of $\sqrt{k}$ in the radius of the empirical ℓ_2 cover of the inner decomposition function class cannot be eliminated by Sudakov's minoration. We can eliminate the $\sqrt{k}$ factor and improve the dependency on k to be independent by the following lemma and the inner decomposition function class.

Lemma 1.2 (Theorem 12.8 in (Anthony & Bartlett, 2009))**.** *For any function in $\mathcal{F}$ takes values in $[-B, B]$ and any S with sample of size n, $\epsilon > 0$, $n \geq d$,*

$$\log \mathcal{N}_\infty(\epsilon, \mathcal{F}, S) \leq 1 + \mathrm{fat}_{\epsilon/4}(\mathcal{F})\log_2\frac{4eBn}{d\epsilon}\log\frac{4nB^2}{\epsilon^2}.$$

The first part of the proof of Lemma 1.1 is similar to steps 1-3 of the proof of Lemma 4.1 in the main paper, because we find that the square-root dependency on q mainly comes from the $\sqrt{q}$ factor in the radius of the empirical ℓ_2 cover of the outer decomposition function class $\mathcal{R}(\mathcal{F})$, which is inevitable for ℓ_2 Lipschitz loss. However, the dependency on k can be further improved, and we prove it in detail as follows:

For the dataset $D = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ with n i.i.d. examples:

$$\begin{aligned}
&\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{q}\sum_{j=1}^{q}\left(r_j(\boldsymbol{f}(\boldsymbol{x}_i)) - r_j(\boldsymbol{f}'(\boldsymbol{x}_i))\right)^2} \\
=&\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{q}\sum_{j=1}^{q}\left(f_j(\boldsymbol{x}_i) - f'_j(\boldsymbol{x}_i)\right)^2} \\
\leq&\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{q}\sum_{j=1}^{q}\max_{i\in[n],j\in[q]}\left(f_j(\boldsymbol{x}_i) - f'_j(\boldsymbol{x}_i)\right)^2} \\
\leq&\max_{i\in[n],j\in[q]}|f_j(\boldsymbol{x}_i) - f'_j(\boldsymbol{x}_i)| \\
=&\max_i\max_j|g(\boldsymbol{h}^j(\boldsymbol{x}_i)) - g(\boldsymbol{h}^{j'}(\boldsymbol{x}_i))| \\
\leq&\max_i\max_j\|\boldsymbol{h}^j(\boldsymbol{x}_i) - \boldsymbol{h}^{j'}(\boldsymbol{x}_i)\|_\infty
\end{aligned}$$

$$
\begin{aligned}
&= \max_i \max_j \max_s \left| h_s^j(\boldsymbol{x}_i) - h_s^{j'}(\boldsymbol{x}_i) \right| \\
&= \max_i \max_j \max_s |t_s(\boldsymbol{h}^j(\boldsymbol{x}_i)) - t_s(\boldsymbol{h}^{j'}(\boldsymbol{x}_i))|. \quad \text{(The definition of the inner decomposition function class } \mathcal{T}(\mathcal{H}^j))
\end{aligned}
$$

Then, according to the definition of the empirical covering numbers, we have that an empirical ℓ_∞ norm cover of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$ at radius ϵ is also an empirical ℓ_2 norm cover of the outer decomposition function class $\mathcal{R}(\mathcal{F})$ at radius ϵ, and we can conclude that:

$$
\mathcal{N}_2(\epsilon, \mathcal{R}(\mathcal{F}), [q] \times D) \le \mathcal{N}_\infty\left(\epsilon, \mathcal{T}(\mathcal{H}^j), [k] \times [q] \times D\right). \tag{1}
$$

According to the above Lemma 1.2, we have

$$
\begin{aligned}
&\log \mathcal{N}_\infty\left(\epsilon, \mathcal{T}(\mathcal{H}^j), [k] \times [q] \times D\right) \\
\le& 1 + \mathrm{fat}_{\epsilon/4}(\mathcal{T}(\mathcal{H}^j)) \log_2^2 \frac{4eB^2nqk}{\epsilon^2} \\
\le& 1 + \frac{64nqk\widetilde{\mathfrak{R}}_{nqk}^2(\mathcal{T}(\mathcal{H}^j))}{\epsilon^2} \log_2^2 \frac{4eB^2nqk}{\epsilon^2}. \quad \text{(Use inequality } \mathrm{fat}_\epsilon(\mathcal{T}(\mathcal{H}^j)) \le \frac{4nqk\widetilde{\mathfrak{R}}_{nqk}^2(\mathcal{T}(\mathcal{H}^j))}{\epsilon^2})
\end{aligned}
$$

Then, according to Dudley's entropy integral inequality and combined with inequality (8) in the appendix of the paper, we have

$$
\begin{aligned}
&\hat{\mathfrak{R}}_D(\mathcal{L}) \\
\le& \inf_{\alpha>0}\left(4\alpha + 48q\mu\hat{\mathfrak{R}}_{[q]\times D}(\mathcal{R}(\mathcal{F}))\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right) \\
\le& \inf_{\alpha>0}\left(4\alpha + 48q\mu\inf_{\beta>0}\left(4\beta + \frac{12}{\sqrt{nq}}\int_\beta^B \sqrt{\log\mathcal{N}_2(\epsilon, \mathcal{R}(\mathcal{F}), [q]\times D)}d\epsilon\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right) \\
\le& \inf_{\alpha>0}\left(4\alpha + 48q\mu\inf_{\beta>0}\left(4\beta + \frac{12}{\sqrt{nq}}\int_\beta^B \sqrt{\log\mathcal{N}_\infty(\epsilon, \mathcal{T}(\mathcal{H}^j), [k]\times[q]\times D)}d\epsilon\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right) \\
\le& \inf_{\alpha>0}\left(4\alpha + 48q\mu\inf_{\beta>0}\left(4\beta + \frac{12}{\sqrt{nq}}\int_\beta^B \sqrt{1 + \frac{64nqk\widetilde{\mathfrak{R}}_{nqk}^2(\mathcal{T}(\mathcal{H}^j))}{\epsilon^2}\log_2^2\frac{4eB^2nqk}{\widetilde{\mathfrak{R}}_{nqk}^2(\mathcal{T}(\mathcal{H}^j))}}d\epsilon\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right) \\
\le& \inf_{\alpha>0}\left(4\alpha + 48q\mu\inf_{\beta>0}\left(4\beta + \frac{12}{\sqrt{nq}}\int_\beta^B \sqrt{1 + \frac{64nqk\widetilde{\mathfrak{R}}_{nqk}^2(\mathcal{T}(\mathcal{H}^j))}{\epsilon^2}\log_2^2(8en^2q^2k^2)}d\epsilon\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right) \\
\le& \inf_{\alpha>0}\left(4\alpha + 48q\mu\inf_{\beta>0}\left(4\beta + \frac{12B}{\sqrt{nq}} + 96\sqrt{k}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))\log_2(8en^2q^2k^2)\int_\beta^B \epsilon^{-1}d\epsilon\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right) \\
\le& \inf_{\alpha>0}\left(4\alpha + \frac{576B\sqrt{q}}{\sqrt{n}} + 48q\mu\inf_{\beta>0}\left(4\beta + 96\sqrt{k}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))\log_2(8en^2q^2k^2)\log\frac{B}{\beta}\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right) \\
\le& \inf_{\alpha>0}\left(4\alpha + \frac{576B\sqrt{q}}{\sqrt{n}} + 2q\mu 48^2\sqrt{k}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))(1 + \log_2(8en^2q^2k^2)\log\frac{B}{24\sqrt{k}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))})\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right) \\
&\text{(Choose } \beta = 24\sqrt{k}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))) \\
\le& \frac{576B\sqrt{q}}{\sqrt{n}} + \inf_{\alpha>0}\left(4\alpha + 2q\mu 48^2\sqrt{k}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))\left(1 + \log_2(8en^2q^2k^2)\log\sqrt{nq}\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right) \\
&\text{(Use } \widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j)) \ge \frac{B}{\sqrt{2nqk}})
\end{aligned}
$$

$$\leq \frac{576B\sqrt{q}}{\sqrt{n}} + 2q\mu 48^2\sqrt{k}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j)) \left(1 + \left(1 + \log_2(8en^2q^2k^2)\log\sqrt{nq}\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{24\cdot 48\mu q\sqrt{k}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))}\right)$$

(Choose $\alpha = 24\cdot 48\mu q\sqrt{k}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j)))$

$$\leq \frac{576B\sqrt{q}}{\sqrt{n}} + 2q\mu 48^2\sqrt{k}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j)) \left(1 + \left(1 + \log_2(8en^2q^2k^2)\log\sqrt{nq}\right)\log^{\frac{1}{2}}(nq)\log\frac{M\sqrt{n}}{\mu B}\right).$$

(Use $\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j)) \geq \frac{B}{\sqrt{2nqk}}$)

□

With the vector-contraction inequality in Lemma 1.1, we can derive the following tight bound for ℓ_2 Lipschitz loss:

Theorem 1.3. *Let $\mathcal{F}$ be the class of the multi-dimensional classification defined by (1) and (2). Let Assumptions 3.1 and 3.2 hold. Given a dataset D of size n. Then, for any $0 < \delta < 1$, with probability at least $1-\delta$, the following holds for any $\boldsymbol{f} \in \mathcal{F}$:*

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{2\cdot 24^2 B\sqrt{q}}{\sqrt{n}} + \frac{96^2\mu\sqrt{q}}{\sqrt{n}}\left(1 + \left(1 + \log_2(8en^2q^2k^2)\log\sqrt{nq}\right)\log^{\frac{1}{2}}(nq)\log\frac{M\sqrt{n}}{\mu B}\right).$$

Proof Sketch. We first upper bound the worst-case Rademacher complexity $\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))$ of the inner decomposition function class as $\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j)) \leq B/\sqrt{nqk}$, and then combined with Lemma 1.1, the desired bound can be derived. □

The bound in Theorem 1.3 is tighter than the improved $\widetilde{O}(\sqrt{\frac{qk}{n}})$ bound in Theorem 4.3 of the main paper with a faster convergence rate $\widetilde{O}(\sqrt{\frac{q}{n}})$. This confirms our statement in the rebuttal that "our discovery of the essence of a factor on k in the radius of the empirical cover of the decomposition class allows us to improve the dependency on k to be independent rather than just a factor of $\sqrt{k}$ even for ℓ_2 Lipschitz loss".

References

Anthony, M. and Bartlett, P. L. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.