Supplementary Material of Towards Understanding the Mechanism of Contrastive Learning via Similarity Structure: A Theoretical Analysis

Hiroki Waida[1(✉)], Yuichiro Wada[2,3], Léo Andéol[4,5,6,7], Takumi Nakagawa[1,3], Yuhui Zhang[1], and Takafumi Kanamori[1,3]

[1] Tokyo Institute of Technology, Tokyo, Japan
(✉) waida.h.aa@m.titech.ac.jp
[2] Fujitsu, Kanagawa, Japan
[3] RIKEN AIP, Tokyo, Japan
[4] Institut de Mathématiques de Toulouse, Toulouse, France
[5] SNCF, Saint-Denis, France
[6] Université de Toulouse, Toulouse, France
[7] CNRS, Toulouse, France

A Proof in Section 2.2

First we prove Proposition 1.

Proof (Proof of Proposition 1). The proof of the claim closely follows the proof of Lemma 3.1 in Muandet et al. [30] (see also Smola et al. [37]), which shows that if $\mathbb{E}_P[\sqrt{k(x,x)}] < +\infty$ where $x \sim P$, then $\mathbb{E}_P[k(\cdot,x)] \in \mathcal{H}_k$. For the sake of completeness, we provide the proof of Proposition 1 by modifying the proof of Muandet et al. [30] slightly.

Let $\mathbb{M}$ be a measurable set in $\mathbb{X}$. Define $\mu_{\mathbb{M}}(f) := \mathbb{E}[h(f(x))|\mathbb{M}]$. Our goal is to show that $\mu_{\mathbb{M}}(f) \in \mathcal{H}_k$ holds. To this end, for $\phi \in \mathcal{H}_k$, we compute

$$\begin{aligned}
|\mathbb{E}\left[\phi(f(x))|\mathbb{M}\right]| &= |\mathbb{E}\left[\langle \phi, k(\cdot, f(x))\rangle_{\mathcal{H}_k}|\mathbb{M}\right]| \\
&\leq \mathbb{E}\left[|\langle \phi, k(\cdot, f(x))\rangle_{\mathcal{H}_k}|\,|\mathbb{M}\right] \\
&\leq \mathbb{E}\left[\|\phi\|_{\mathcal{H}_k}\|k(\cdot, f(x))\|_{\mathcal{H}_k}|\mathbb{M}\right] \qquad \text{(Cauchy-Schwarz ineq.)} \\
&= \|\phi\|_{\mathcal{H}_k}\mathbb{E}\left[\sqrt{k(f(x),f(x))}|\mathbb{M}\right].
\end{aligned}$$

Since $\sup_{z,z' \in \mathbb{S}^{d-1}} k(z,z') < \infty$ holds, we have $\mathbb{E}[\sqrt{k(f(x),f(x))}|\mathbb{M}] < +\infty$. Hence, the map $\phi \mapsto \mathbb{E}[\phi(f(x))|\mathbb{M}]$ is a bounded linear functional on $\mathcal{H}_k$, and thus from Riesz's representation theorem, there exists some $\xi \in \mathcal{H}_k$ such that $\mathbb{E}[\phi(f(x))|\mathbb{M}] = \langle \xi, \phi\rangle_{\mathcal{H}_k}$. However, let $\phi = k(\cdot, z)$, then $\xi(z) = \langle \xi, k(\cdot,z)\rangle_{\mathcal{H}_k} = \mathbb{E}[k(f(x),z)|\mathbb{M}]$. This implies $\xi = \mathbb{E}[k(f(x),\cdot)|\mathbb{M}] \in \mathcal{H}_k$. Since k is symmetric, we have $\mu_{\mathbb{M}}(f) = \mathbb{E}[k(\cdot, f(x))|\mathbb{M}] \in \mathcal{H}_k$. □

B Proofs in Section 5.1

In this section, we prove Theorem 1 and Corollary 1.

B.1 Useful Lemmas for the Proof of Theorem 1

Before showing Theorem 1, we give several basic and useful lemmas that are used in the proof of the theorem. Since the definition of $\mu_{\mathbb{M}}(f)$, where $\mathbb{M}$ is a measurable subset of $\mathbb{X}$ and $f \in \mathcal{F}$, is slightly different from the kernel mean embedding of the usual form [5,30] due to the existence of the encoder function f, we provide the proof for each lemma for the sake of completeness.

Lemma 1. *Let $\{e_j\}$ be an orthonormal basis of $\mathcal{H}_k$, and let $\mathbb{M}$ be a measurable set. Let $f \in \mathcal{F}$. Then, the following identity holds for each j:*

$$\int_{\mathbb{M}} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) = \langle \mu_{\mathbb{M}}(f), e_j \rangle_{\mathcal{H}_k}.$$

Proof. We calculate,

$$\begin{aligned}
\langle \mu_{\mathbb{M}}(f), e_j \rangle_{\mathcal{H}_k} &= \left\langle \int_{\mathbb{M}} h(f(x)) P_{\mathbb{X}}(dx|\mathbb{M}), e_j \right\rangle_{\mathcal{H}_k} \\
&= \left\langle \int_{\mathbb{M}} \sum_{j'} \langle h(f(x)), e_{j'} \rangle_{\mathcal{H}_k} e_{j'} P_{\mathbb{X}}(dx|\mathbb{M}), e_j \right\rangle_{\mathcal{H}_k} \\
&= \left\langle \sum_{j'} e_{j'} \int_{\mathbb{M}} \langle h(f(x)), e_{j'} \rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}), e_j \right\rangle_{\mathcal{H}_k} \\
&= \int_{\mathbb{M}} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}),
\end{aligned}$$

where in the third line, we use the Dominated Convergence Theorem for the Bochner integral (e.g., see Theorem 1.1.8 in Arendt et al. [1]). Hence, we obtain the claim. □

Lemma 2. *Let $\mathbb{M}, \mathbb{M}'$ be measurable subsets of $\mathbb{X}$. Let $f \in \mathcal{F}$. Then, we have*

$$\int_{\mathbb{M}} \int_{\mathbb{M}'} \langle h(f(x)), h(f(x')) \rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) P_{\mathbb{X}}(dx'|\mathbb{M}') = \langle \mu_{\mathbb{M}}(f), \mu_{\mathbb{M}'}(f) \rangle_{\mathcal{H}_k}.$$

Proof. Let $\{e_j\}$ be an orthonormal basis of $\mathcal{H}_k$. Then we have,

$$
\begin{aligned}
&\int_{\mathbb{M}}\int_{\mathbb{M}'} \langle h(f(x)), h(f(x'))\rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) P_{\mathbb{X}}(dx'|\mathbb{M}') \\
&= \int_{\mathbb{M}}\int_{\mathbb{M}'} \left\langle \sum_j \langle h(f(x)), e_j\rangle_{\mathcal{H}_k} e_j, \sum_j \langle h(f(x')), e_j\rangle_{\mathcal{H}_k} e_j \right\rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) P_{\mathbb{X}}(dx'|\mathbb{M}') \\
&= \int_{\mathbb{M}}\int_{\mathbb{M}'} \sum_j \langle h(f(x)), e_j\rangle_{\mathcal{H}_k} \langle h(f(x')), e_j\rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) P_{\mathbb{X}}(dx'|\mathbb{M}') \\
&= \sum_j \int_{\mathbb{M}}\int_{\mathbb{M}'} \langle h(f(x)), e_j\rangle_{\mathcal{H}_k} \langle h(f(x')), e_j\rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) P_{\mathbb{X}}(dx'|\mathbb{M}') \qquad (1) \\
&= \sum_j \left(\int_{\mathbb{M}} \langle h(f(x)), e_j\rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) \right) \left(\int_{\mathbb{M}'} \langle h(f(x')), e_j\rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx'|\mathbb{M}') \right) \\
&= \sum_j \langle \mu_{\mathbb{M}}(f), e_j\rangle_{\mathcal{H}_k} \langle \mu_{\mathbb{M}'}(f), e_j\rangle_{\mathcal{H}_k} \qquad \text{(Lemma 1)} \\
&= \left\langle \sum_j \langle \mu_{\mathbb{M}}(f), e_j\rangle_{\mathcal{H}_k} e_j, \sum_j \langle \mu_{\mathbb{M}'}(f), e_j\rangle_{\mathcal{H}_k} e_j \right\rangle_{\mathcal{H}_k} \\
&= \langle \mu_{\mathbb{M}}(f), \mu_{\mathbb{M}'}(f)\rangle_{\mathcal{H}_k},
\end{aligned}
$$

where in (1) we use the Dominated Convergence Theorem. Hence we obtain the claim. □

B.2 Proof of Theorem 1

The following notation is used in the proof of Theorem 1.

Definition 3. *Denote* $M_k = \sup_{z,z'\in\mathbb{S}^{d-1}} \|k(\cdot,z) - k(\cdot,z')\|^2_{\mathcal{H}_k}$. *We define,*

$$
\begin{aligned}
R(\lambda) := \frac{M_k}{2} \sum_{i\neq j} P_+((\mathbb{M}_i \cap \mathbb{M}_j) \times (\mathbb{M}_i \cap \mathbb{M}_j)) \\
+ \lambda\psi(1) \sum_{i=1}^{K} P_{\mathbb{X}}(\mathbb{M}_i)\,(1 - P_{\mathbb{X}}(\mathbb{M}_i)) + (1-\lambda)\psi(1),
\end{aligned}
$$

where $dP_+(x,x') = w(x,x') d\nu_{\mathbb{X}}^{\otimes 2}(x,x')$.

Note that under Assumption 1, $k(z) := k(z,z) = \psi(z^\top z) = \psi(1)$ is a constant function on $\mathbb{S}^{d-1}$. We are now ready to present the proof of Theorem 1.

Proof (Proof of Theorem 1). It is convenient to analyze the following form instead of the kernel contrastive loss:

$$\widetilde{L}_{\mathrm{KCL}}(f;\lambda) := \underbrace{\mathbb{E}_{x,x^+}\left[\|h(f(x))-h(f(x^+))\|_{\mathcal{H}_k}^2\right]}_{\text{the positive term}} - \lambda \underbrace{\mathbb{E}_{x,x^-}\left[\|h(f(x))-h(f(x^-))\|_{\mathcal{H}_k}^2\right]}_{\text{the negative term}}. \quad (2)$$

Note that, $\widetilde{L}_{\mathrm{KCL}}(f;\lambda) = 2(1-\lambda)\psi(1) + 2L_{\mathrm{KCL}}(f;\lambda)$ holds since $f(x) \in \mathbb{S}^{d-1}$ for all $x \in \mathbb{X}$. For the positive term of $\widetilde{L}_{\mathrm{KCL}}(f)$, we can evaluate that,

$$\begin{aligned}
&\mathbb{E}_{x,x^+}\left[\|h(f(x))-h(f(x^+))\|_{\mathcal{H}_k}^2\right] \\
&\geq \int_{\bigcup_{i=1}^K \mathbb{M}_i\times\mathbb{M}_i} \|h(f(x))-h(f(x'))\|_{\mathcal{H}_k}^2 w(x,x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x') \quad (3)\\
&\geq \sum_{i=1}^K \int_{\mathbb{M}_i\times\mathbb{M}_i} \|h(f(x))-h(f(x'))\|_{\mathcal{H}_k}^2 w(x,x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x') \\
&\quad - \sum_{j\neq i} \int_{(\mathbb{M}_i\cap\mathbb{M}_j)\times(\mathbb{M}_i\cap\mathbb{M}_j)} \|h(f(x))-h(f(x'))\|_{\mathcal{H}_k}^2 w(x,x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x') \quad (4)\\
&\geq \sum_{i=1}^K \left(\int_{\mathbb{M}_i\times\mathbb{M}_i} \|h(f(x))-h(f(x'))\|_{\mathcal{H}_k}^2 w(x,x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x') \right. \\
&\quad \left. - M_k \sum_{j\neq i} P_+((\mathbb{M}_i\cap\mathbb{M}_j)\times(\mathbb{M}_i\cap\mathbb{M}_j)) \right) \quad (5)
\end{aligned}$$

where in the second inequality we use the fact that

$$Q(\bigcup_{i=1}^K \mathbb{M}_i\times\mathbb{M}_i) \geq \sum_{i=1}^K Q(\mathbb{M}_i\times\mathbb{M}_i) - \sum_{i\neq j} Q((\mathbb{M}_i\times\mathbb{M}_i)\cap(\mathbb{M}_j\times\mathbb{M}_j)),$$

for any probability measure Q in $\mathbb{X}\times\mathbb{X}$, and in the last inequality we use the definition $M_k = \sup_{z,z'\in\mathbb{S}^{d-1}} \|k(\cdot,z)-k(\cdot,z')\|_{\mathcal{H}_k}^2$. The first term of the above lower bound can be bounded as

$$\begin{aligned}
&\sum_{i=1}^K \int_{\mathbb{M}_i\times\mathbb{M}_i} \|h(f(x))-h(f(x'))\|_{\mathcal{H}_k}^2 w(x,x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x') \\
&\geq \sum_{i=1}^K \int_{\mathbb{M}_i\times\mathbb{M}_i} \|h(f(x))-h(f(x'))\|_{\mathcal{H}_k}^2 \cdot (\lambda+\delta)w(x)w(x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x'), \quad (6)
\end{aligned}$$

where we utilize the definition of $\mathbb{M}_i$ for each $i \in \{1,\cdots,K\}$; recall that due to the condition **(B)** in Assumption 2, for every $x, x' \in \mathbb{M}_i$ we have $\mathrm{sim}(x,x';\lambda) \geq \delta$.

On the other hand, for the negative term we can compute as follows:

$$
\begin{aligned}
&- \mathbb{E}_{x,x^-}\left[\|h(f(x)) - h(f(x^-))\|_{\mathcal{H}_k}^2\right] \\
&= -\int_{\bigcup_{i=1}^K \mathbb{M}_i \times \mathbb{M}_i} \|h(f(x)) - h(f(x'))\|_{\mathcal{H}_k}^2 w(x)w(x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x') \\
&\quad - \int_{\mathbb{X}\times\mathbb{X}\setminus\left(\bigcup_{i=1}^K \mathbb{M}_i \times \mathbb{M}_i\right)} \|h(f(x)) - h(f(x'))\|_{\mathcal{H}_k}^2 w(x)w(x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x') \\
&\geq -\sum_{i=1}^K \int_{\mathbb{M}_i \times \mathbb{M}_i} \|h(f(x)) - h(f(x'))\|_{\mathcal{H}_k}^2 w(x)w(x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x') \\
&\quad - \int_{\mathbb{X}\times\mathbb{X}\setminus\left(\bigcup_{i=1}^K \mathbb{M}_i \times \mathbb{M}_i\right)} \|h(f(x)) - h(f(x'))\|_{\mathcal{H}_k}^2 w(x)w(x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x'),
\end{aligned} \tag{7}
$$

where the last inequality is due to the union bound. For the second term in the right hand side of the inequality above, we have

$$
\begin{aligned}
&- \int_{\mathbb{X}\times\mathbb{X}\setminus\left(\bigcup_{i=1}^K \mathbb{M}_i \times \mathbb{M}_i\right)} \|h(f(x)) - h(f(x'))\|_{\mathcal{H}_k}^2 w(x)w(x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x') \\
&\geq -\sum_{i\neq j} \int_{\mathbb{M}_i \times \mathbb{M}_j} \|h(f(x)) - h(f(x'))\|_{\mathcal{H}_k}^2 w(x)w(x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x') \\
&\quad - \int_{\mathbb{X}\times\mathbb{X}\setminus\left(\bigcup_{i,j=1}^K \mathbb{M}_i \times \mathbb{M}_j\right)} \|h(f(x)) - h(f(x'))\|_{\mathcal{H}_k}^2 w(x)w(x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x')
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
&\geq -\sum_{i\neq j} \int_{\mathbb{M}_i \times \mathbb{M}_j} \|h(f(x)) - h(f(x'))\|_{\mathcal{H}_k}^2 w(x)w(x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x') \\
&\quad - M_k P_{\mathbb{X}}^{\otimes 2}\left(\mathbb{X}\times\mathbb{X}\setminus\left(\bigcup_{i,j=1}^K \mathbb{M}_i \times \mathbb{M}_j\right)\right).
\end{aligned} \tag{9}
$$

Here, the second term of (9) vanishes since Assumption 2 implies $\mathbb{X}\times\mathbb{X} = \bigcup_{i,j=1}^K \mathbb{M}_i \times \mathbb{M}_j$. The first term of (9) is further lower bounded as,

$$
\begin{aligned}
&- \sum_{i \neq j} \int_{\mathbb{M}_i \times \mathbb{M}_j} \|h(f(x)) - h(f(x'))\|_{\mathcal{H}_k}^2 w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
&\geq 2 \sum_{i \neq j} \left\{ - \sup_{z \in \mathbb{S}^{d-1}} k(z,z) P_{\mathbb{X}}(\mathbb{M}_i) P_{\mathbb{X}}(\mathbb{M}_j) \right. \\
&\qquad \left. + \int_{\mathbb{M}_i \times \mathbb{M}_j} \langle h(f(x)), h(f(x')) \rangle_{\mathcal{H}_k} w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \right\} \qquad (10) \\
&= 2 \sum_{i \neq j} \left\{ - \sup_{z \in \mathbb{S}^{d-1}} k(z,z) P_{\mathbb{X}}(\mathbb{M}_i) P_{\mathbb{X}}(\mathbb{M}_j) + P_{\mathbb{X}}(\mathbb{M}_i) P_{\mathbb{X}}(\mathbb{M}_j) \langle \mu_i(f), \mu_j(f) \rangle_{\mathcal{H}_k} \right\} \\
&\qquad \text{(Lemma 2)} \\
&= -2 \sup_{z \in \mathbb{S}^{d-1}} k(z,z) \sum_{i=1}^{K} P_{\mathbb{X}}(\mathbb{M}_i) \left(1 - P_{\mathbb{X}}(\mathbb{M}_i)\right) + 2\mathfrak{c}(f).
\end{aligned}
$$

Thus for the negative term we obtain the inequality,

$$
\begin{aligned}
&- \mathbb{E}_{x,x^-} \left[\|h(f(x)) - h(f(x^-))\|_{\mathcal{H}_k}^2 \right] \\
&\geq - \sum_{i=1}^{K} \int_{\mathbb{M}_i \times \mathbb{M}_i} \|h(f(x)) - h(f(x'))\|_{\mathcal{H}_k}^2 w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
&\quad + 2\mathfrak{c}(f) - 2 \sup_{z \in \mathbb{S}^{d-1}} k(z,z) \sum_{i=1}^{K} P_{\mathbb{X}}(\mathbb{M}_i) \left(1 - P_{\mathbb{X}}(\mathbb{M}_i)\right). \qquad (11)
\end{aligned}
$$

Combining (2),(5),(6), and (11), we have

$$
\begin{aligned}
&L_{\mathrm{KCL}}(f; \lambda) + (1 - \lambda)\psi(1) \\
&\geq \frac{\delta}{2} \sum_{i=1}^{K} \mathbb{E}_{x,x^-} \left[\|h(f(x)) - h(f(x^-))\|_{\mathcal{H}_k}^2 ; \mathbb{M}_i \times \mathbb{M}_i \right] + \lambda \mathfrak{c}(f) \\
&\quad - \frac{M_k}{2} \sum_{i \neq j} P_+((\mathbb{M}_i \cap \mathbb{M}_j) \times (\mathbb{M}_i \cap \mathbb{M}_j)) - \lambda \psi(1) \sum_{i=1}^{K} P_{\mathbb{X}}(\mathbb{M}_i) \left(1 - P_{\mathbb{X}}(\mathbb{M}_i)\right). \qquad (12)
\end{aligned}
$$

Therefore, we complete the proof. □

B.3 Proof of Corollary 1

Proof (Proof of Corollary 1). The proof of Corollary 1 is completed by checking whether the equality holds in each inequality that appears in the proof of Theorem 1. We list the detail of the checks below:

(3): Since $w(x,x')=0$ for any $(x,x')\in\mathbb{M}_i\times\mathbb{M}_j$ $(i\neq j)$, we have $\int_{\mathbb{M}_i\times\mathbb{M}_j}\|h(f(x))-h(f(x'))\|^2_{\mathcal{H}_k}w(x,x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x')=0$. Here, we have the decomposition $\mathbb{X}\times\mathbb{X}=(\bigcup_{i=1}^K\mathbb{M}_i)\times(\bigcup_{j=1}^K\mathbb{M}_j)=\bigcup_{i,j=1}^K\mathbb{M}_i\times\mathbb{M}_j$, where $(\mathbb{M}_i\times\mathbb{M}_j)\cap(\mathbb{M}_{i'}\times\mathbb{M}_{j'})=\emptyset$ for any (i,j,i',j') such that $i\neq i'$ or $j\neq j'$, from the assumption that $\mathbb{M}_1,\cdots,\mathbb{M}_K$ are disjoint. Hence, using the additivity of a probability measure yields the equality.

(4): Since $\mathbb{M}_i\cap\mathbb{M}_j=\emptyset$ for $i,j\in[K]$ such that $i\neq j$, the first term in the right-hand-side of (4) is equal to the first term in the left-hand-side of (4). On the other hand, the second term in the right-hand-side of (4) is equal to 0. Hence, the equality holds.

(5): Since the second term of the right-hand-side of (5) is 0 under the assumption that $\mathbb{M}_i\cap\mathbb{M}_j=\emptyset$ for i,j $(i\neq j)$, the equality holds.

(6): The equality holds from the assumption that for any $x,x'\in\mathbb{M}_i$ $(i\in[K])$, $\mathrm{sim}(x,x';\lambda)=\delta$ holds. Indeed, this assumption implies that $w(x,x')=(\lambda+\delta)w(x)w(x')$ for any $x,x'\in\mathbb{M}_i$.

(7): Since $\mathbb{M}_1\times\mathbb{M}_1,\cdots,\mathbb{M}_K\times\mathbb{M}_K$ are disjoint, the equality holds.

(8): Since $\bigcup_{i,j=1}^K\mathbb{M}_i\times\mathbb{M}_j=\mathbb{X}\times\mathbb{X}$, the second term of the right-hand-side of (8) is equal to 0. Thus, the equality holds.

(9): The equality holds due to the same reason as (8) above.

(10): Since $\|h(f(x))\|^2_{\mathcal{H}_k}=k(f(x),f(x))=\psi(f(x)^\top f(x))=\psi(1)$ for any $x\in\mathbb{X}$ and $f\in\mathcal{F}$, the equality holds.

(11): Since (11) is the combination of (7), (8), (9), and (10), the equality in (11) holds in this case.

(12): Since (12) is obtained by combining (5), (6), and (11), the equality holds.

Therefore, we obtain the result. □

C Proof in Section 5.2

We present the proof of a generalized version of Theorem 2. The generalized theorem is presented below.

Theorem 5 (The generalization of Theorem 2). *Suppose that Assumption 1 and 2 hold. Take $K\in\mathbb{N}$ and $\mathbb{M}_1,\cdots,\mathbb{M}_K$ such that the condition* **(A)** *in Assumption 2 is satisfied. Let $\widetilde{\mathbb{M}}_1,\cdots\widetilde{\mathbb{M}}_K$ be a disjoint partition of $\mathbb{X}$ satisfying $\widetilde{\mathbb{M}}_i\subset\mathbb{M}_i$ for each $i\in[K]$. Define $\widetilde{y}:\mathbb{X}\to[K]$ as $\widetilde{y}(x)=i$ for every $x\in\widetilde{\mathbb{M}}_i$. Then, for each meaningful encoder $f\in\mathcal{F}$, we have*

$$L_{\mathrm{Err}}(f,W_\mu,\beta_\mu;\widetilde{y})\leq\frac{8(K-1)}{\Delta_{\min}(f)\cdot\min_{i\in[K]}P_{\mathbb{X}}(\mathbb{M}_i)}\mathfrak{a}(f)$$

where $\Delta_{\min}(f)=\min_{i\neq j}\|\mu_i(f)-\mu_j(f)\|^2_{\mathcal{H}_k}$.

Proof (Proof of Theorem 5). From the definition, we have $\widetilde{\mathbb{M}}_i\subset\mathbb{M}_i$ and $\widetilde{\mathbb{M}}_i\cap\widetilde{\mathbb{M}}_j=\emptyset$ for all the pairs of distinct indices $i,j\in[K]$. Let us recall the definition of $L_{\mathrm{Err}}(f,W_\mu,\beta_\mu;\widetilde{y})$:

$$L_{\mathrm{err}}(f,W_\mu,\beta_\mu;\widetilde{y})=P_{\mathbb{X}}\left(g_{f,W_\mu,\beta_\mu}(x)\neq\widetilde{y}(x)\right).$$

Here recall that we let arg max, arg min also breaks tie arbitrary. For instance, if there are distinct integers $i_1, \cdots, i_j \in [K]$ such that $g_{f,W_\mu,\beta_\mu}(x) = \{i_1, \cdots, i_j\}$, then we define $g_{f,W_\mu,\beta_\mu}(x) = \widetilde{y}(x)$ if $\widetilde{y}(x) \in \{i_1, \cdots, i_j\}$, and $g_{f,W_\mu,\beta_\mu}(x) = i_1$ if $\widetilde{y}(x) \notin \{i_1, \cdots, i_j\}$. The event $\mathbb{A} := \{x \mid g_{f,W_\mu,\beta_\mu}(x) \neq \widetilde{y}(x)\} = \{x \mid g_{f,W_\mu,\beta_\mu}(x) \neq \widetilde{y}(x)\} \cap \bigcup_{i=1}^K \widetilde{\mathbb{M}}_i \subset \mathbb{X}$ is a subset of the event $\mathbb{D} := \bigcup_{i=1}^K \bigcup_{j\neq i} \{x \mid \|h(f(x)) - \mu_i(f)\|_{\mathcal{H}_k} \geq \|h(f(x)) - \mu_j(f)\|_{\mathcal{H}_k}\} \cap \widetilde{\mathbb{M}}_i$, since

$$
\begin{aligned}
& x \in \mathbb{A} \\
\iff & \arg\min_{i\in[K]} \|h(f(x)) - \mu_i(f)\|_{\mathcal{H}_k} \neq \widetilde{y}(x) \quad \text{and} \quad x \in \bigcup_{i=1}^K \widetilde{\mathbb{M}}_i \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(def. of } g_{f,W_\mu,\beta_\mu} \text{ and } g_{1\text{-NN}}) \\
\iff & x \in \bigcup_{j\neq\widetilde{y}(x)} \{x' \mid \|h(f(x')) - \mu_{\widetilde{y}(x)}(h_f)\|_{\mathcal{H}_k} \geq \|h(f(x')) - \mu_j(f)\|_{\mathcal{H}_k}\} \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{and} \quad x \in \bigcup_{i=1}^K \widetilde{\mathbb{M}}_i \\
\iff & x \in \bigcup_{j\neq\widetilde{y}(x)} \{x' \mid \|h(f(x')) - \mu_{\widetilde{y}(x)}(h_f)\|_{\mathcal{H}_k} \geq \|h(f(x')) - \mu_j(f)\|_{\mathcal{H}_k}\} \cap \widetilde{\mathbb{M}}_{\widetilde{y}(x)} \\
\implies & x \in \bigcup_{i=1}^K \bigcup_{j\neq i} \{x' \mid \|h(f(x')) - \mu_i(f)\|_{\mathcal{H}_k} \geq \|h(f(x')) - \mu_j(f)\|_{\mathcal{H}_k}\} \cap \widetilde{\mathbb{M}}_i = \mathbb{D}.
\end{aligned}
$$

Define $\mathbb{L}_{ij} := \{c(\mu_j(f) - \mu_i(f)) \mid c \in \mathbb{R}\} \subset \mathcal{H}_k$ for every $i, j \in [K], i \neq j$. Since each $\mathbb{L}_{ij}$ is a closed subspace of $\mathcal{H}_k$, for every $z \in \mathcal{H}_k$ there exists some $z_1 \in \mathbb{L}_{ij}$ and $z_2 \in \mathbb{L}_{ij}^\perp$ (where $\mathbb{L}_{ij}^\perp$ is the orthogonal complement space of $\mathbb{L}_{ij}$) such that z admits the unique decomposition $z = z_1 + z_2$. Here define the projection $\widetilde{\pi}_{ij} : \mathcal{H}_k \to \mathbb{L}_{ij}$ as $\widetilde{\pi}_{ij}(z) = z_1$, and define the shifted projection π_{ij} as $\pi_{ij} : \mathcal{H}_k \to \mathcal{H}_k, \pi_{ij}(z) := \widetilde{\pi}_{ij}(z - \mu_i(f)) + \mu_i(f)$. From the definition, we have that $\|\pi_{ij}(z) - \mu_i(f)\|_{\mathcal{H}_k} \leq \|z - \mu_i(f)\|_{\mathcal{H}_k}$ and $\|\pi_{ij}(z) - \mu_j(f)\|_{\mathcal{H}_k} \leq \|z - \mu_j(f)\|_{\mathcal{H}_k}$.

Hereafter, we use the abbreviation $\Delta_{ij} := \Delta_{ij}(f) = \|\mu_i(f) - \mu_j(f)\|_{\mathcal{H}_k}^2$ for the sake of convenience. Using π_{ij}, $i, j \in [K], i \neq j$, the event $\mathbb{D}$ can be decomposed into,

$$
\begin{aligned}
\mathbb{D} = & \underbrace{\left(\mathbb{D} \cap \left(\bigcup_{i=1}^K \bigcup_{j\neq i} \left\{x \mid \|\pi_{ij}(h(f(x))) - \mu_j(f)\|_{\mathcal{H}_k} \leq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cap \widetilde{\mathbb{M}}_i\right)\right)}_{=\mathbb{D}_1} \\
& \cup \underbrace{\left(\mathbb{D} \cap \left(\bigcup_{i=1}^K \bigcup_{j\neq i} \left\{x \mid \|\pi_{ij}(h(f(x))) - \mu_j(f)\|_{\mathcal{H}_k} \leq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cap \widetilde{\mathbb{M}}_i\right)^c\right)}_{=\mathbb{D}_2}.
\end{aligned}
$$

For $\mathbb{D}_1$, we have

$$
\begin{aligned}
& P_{\mathbb{X}}(\mathbb{D}_1) \\
& \leq P_{\mathbb{X}}\left(\bigcup_{i=1}^{K}\bigcup_{j\neq i}\left\{x \mid \|\pi_{ij}(h(f(x))) - \mu_j(f)\|_{\mathcal{H}_k} \leq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cap \widetilde{\mathbb{M}}_i\right) \\
& \leq \sum_{i=1}^{K}\sum_{j\neq i} P_{\mathbb{X}}\left(\left\{x \mid \|\pi_{ij}(h(f(x))) - \mu_j(f)\|_{\mathcal{H}_k} \leq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cap \widetilde{\mathbb{M}}_i\right) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(the union bound)} \\
& \leq \sum_{i=1}^{K}\sum_{j\neq i} P_{\mathbb{X}}\Big(\Big\{x \mid -\|\pi_{ij}(h(f(x))) - \mu_i(f)\|_{\mathcal{H}_k} \\
& \qquad\qquad\qquad\qquad +\|\mu_i(f) - \mu_j(f)\|_{\mathcal{H}_k} \leq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\Big\} \cap \widetilde{\mathbb{M}}_i\Big) \quad \text{(triangle ineq.)} \\
& = \sum_{i=1}^{K}\sum_{j\neq i} P_{\mathbb{X}}\left(\left\{x \mid \|\pi_{ij}(h(f(x))) - \mu_i(f)\|_{\mathcal{H}_k} \geq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cap \widetilde{\mathbb{M}}_i\right) \\
& \leq \sum_{i=1}^{K}\sum_{j\neq i} \frac{4}{\Delta_{ij}}\mathbb{E}\left[\|\pi_{ij}(h(f(x))) - \mu_i(f)\|_{\mathcal{H}_k}^2; \widetilde{\mathbb{M}}_i\right] \qquad \text{(Markov's ineq.)} \\
& \leq \sum_{i=1}^{K}\sum_{j\neq i} \frac{4}{\Delta_{ij}}\mathbb{E}\left[\|h(f(x)) - \mu_i(f)\|_{\mathcal{H}_k}^2; \widetilde{\mathbb{M}}_i\right] \qquad \text{(def. of } \pi_{ij}) \\
& \leq \sum_{i=1}^{K}\sum_{j\neq i} \frac{4}{\Delta_{ij}}\mathbb{E}\left[\|h(f(x)) - \mu_i(f)\|_{\mathcal{H}_k}^2; \mathbb{M}_i\right]. \qquad \text{(def. of } \widetilde{\mathbb{M}}_i)
\end{aligned}
$$

For $\mathbb{D}_2$, we note that we can rewrite as,

$$
\begin{aligned}
&P_{\mathbb{X}}(\mathbb{D}_2)\\
&= P_{\mathbb{X}}\left(\mathbb{D}\cap\left(\bigcup_{i=1}^{K}\bigcup_{j\neq i}\left\{x \mid \|\pi_{ij}(h(f(x)))-\mu_j(f)\|_{\mathcal{H}_k}\leq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\}\cap\widetilde{\mathbb{M}}_i\right)^c\right)\\
&= P_{\mathbb{X}}\left(\left(\bigcup_{i=1}^{K}\bigcup_{j\neq i}\{x \mid \|h(f(x))-\mu_i(f)\|_2\geq\|h(f(x))-\mu_j(f)\|_2\}\cap\widetilde{\mathbb{M}}_i\right)\cap\right.\\
&\qquad\qquad\left.\left(\bigcap_{i=1}^{K}\bigcap_{j\neq i}\left\{x \mid \|\pi_{ij}(h(f(x)))-\mu_j(f)\|_{\mathcal{H}_k}> \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\}\cup\widetilde{\mathbb{M}}_i^c\right)\right)\\
&= P_{\mathbb{X}}\left(\bigcup_{i=1}^{K}\bigcup_{j\neq i}\bigcap_{i'=1}^{K}\bigcap_{j'\neq i'}\left(\{x \mid \|h(f(x))-\mu_i(f)\|_2\geq\|h(f(x))-\mu_j(f)\|_2\}\cap\widetilde{\mathbb{M}}_i\cap\right.\right.\\
&\qquad\qquad\left.\left.\left(\left\{x \mid \|\pi_{i'j'}(h(f(x)))-\mu_{j'}(h_f)\|_{\mathcal{H}_k}> \frac{1}{2}\Delta_{i'j'}^{\frac{1}{2}}\right\}\cup\widetilde{\mathbb{M}}_{i'}^c\right)\right)\right)
\end{aligned}
$$

By using above, we have

$$
\begin{aligned}
&P_{\mathbb{X}}(\mathbb{D}_2)\\
&\leq P_{\mathbb{X}}\left(\bigcup_{i=1}^{K}\bigcup_{j\neq i}(\{x \mid \|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k}\geq\|h(f(x))-\mu_j(f)\|_{\mathcal{H}_k}\right.\\
&\qquad\qquad\left.\text{and } \|\pi_{ij}(h(f(x)))-\mu_j(f)\|_{\mathcal{H}_k}> \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\}\cap\widetilde{\mathbb{M}}_i)\right) \qquad (13)\\
&\leq P_{\mathbb{X}}\left(\bigcup_{i=1}^{K}\bigcup_{j\neq i}\left(\{x \mid \|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k}\geq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\}\cap\widetilde{\mathbb{M}}_i\right)\right) \qquad \text{(def. of } \pi_{ij})\\
&\leq \sum_{i=1}^{K}\sum_{j\neq i}P_{\mathbb{X}}\left(\{x \mid \|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k}\geq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\}\cap\widetilde{\mathbb{M}}_i\right) \quad \text{(the union bound)}\\
&\leq \sum_{i=1}^{K}\sum_{j\neq i}\frac{4}{\Delta_{ij}}\mathbb{E}\left[\|\pi_{ij}(h(f(x)))-\mu_i(f)\|_{\mathcal{H}_k}^2;\widetilde{\mathbb{M}}_i\right] \qquad \text{(Markov's ineq.)}\\
&\leq \sum_{i=1}^{K}\sum_{j\neq i}\frac{4}{\Delta_{ij}}\mathbb{E}\left[\|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k}^2;\widetilde{\mathbb{M}}_i\right] \qquad \text{(def. of } \pi_{ij})\\
&\leq \sum_{i=1}^{K}\sum_{j\neq i}\frac{4}{\Delta_{ij}}\mathbb{E}\left[\|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k}^2;\mathbb{M}_i\right]. \qquad \text{(def. of } \widetilde{\mathbb{M}_i})
\end{aligned}
$$

Here, let us show (13). First let us fix $i,j\in[K]$, where $i\neq j$. For $i',j'\in[K]$ satisfying $i'\neq j'$, we consider the following two cases.

– If $i' = i$ and $j' = j$, then $\widetilde{\mathbb{M}}_i \cap \widetilde{\mathbb{M}}_i^c = \emptyset$, which implies

$$\begin{aligned}
&\{x \mid \|h(f(x)) - \mu_i(f)\|_2 \geq \|h(f(x)) - \mu_j(f)\|_2\} \\
&\quad \cap \widetilde{\mathbb{M}}_i \cap \left(\left\{ x \mid \|\pi_{i'j'}(h(f(x))) - \mu_{j'}(h_f)\|_{\mathcal{H}_k} > \frac{1}{2} \Delta_{i'j'}^{\frac{1}{2}} \right\} \cup \widetilde{\mathbb{M}}_{i'}^c \right) \\
&= \{x \mid \|h(f(x)) - \mu_i(f)\|_{\mathcal{H}_k} \geq \|h(f(x)) - \mu_j(f)\|_{\mathcal{H}_k} \\
&\quad \text{and } \|\pi_{ij}(h(f(x))) - \mu_j(f)\|_{\mathcal{H}_k} > \frac{1}{2} \Delta_{ij}^{\frac{1}{2}} \right\} \cap \widetilde{\mathbb{M}}_i.
\end{aligned}$$

– if $i' \neq i$ or $j' \neq j$, then

$$\begin{aligned}
&\{x \mid \|h(f(x)) - \mu_i(f)\|_2 \geq \|h(f(x)) - \mu_j(f)\|_2\} \cap \widetilde{\mathbb{M}}_i \cap \\
&\qquad \left(\left\{ x \mid \|\pi_{i'j'}(h(f(x))) - \mu_{j'}(h_f)\|_{\mathcal{H}_k} > \frac{1}{2} \Delta_{i'j'}^{\frac{1}{2}} \right\} \cup \widetilde{\mathbb{M}}_{i'}^c \right) \\
&\subset \widetilde{\mathbb{M}}_i.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\bigcap_{i'=1}^{K} \bigcap_{j' \neq i'} \left(\{x \mid \|h(f(x)) - \mu_i(f)\|_2 \geq \|h(f(x)) - \mu_j(f)\|_2\} \right. \\
&\qquad \left. \cap \widetilde{\mathbb{M}}_i \cap \left(\left\{ x \mid \|\pi_{i'j'}(h(f(x))) - \mu_{j'}(h_f)\|_{\mathcal{H}_k} > \frac{1}{2} \Delta_{i'j'}^{\frac{1}{2}} \right\} \cup \widetilde{\mathbb{M}}_{i'}^c \right) \right) \\
&\subset \{x \mid \|h(f(x)) - \mu_i(f)\|_{\mathcal{H}_k} \geq \|h(f(x)) - \mu_j(f)\|_{\mathcal{H}_k} \\
&\qquad \text{and } \|\pi_{ij}(h(f(x))) - \mu_j(f)\|_{\mathcal{H}_k} > \frac{1}{2} \Delta_{ij}^{\frac{1}{2}} \} \cap \widetilde{\mathbb{M}}_i.
\end{aligned}$$

By combining all the results, we obtain

$$\begin{aligned}
P_{\mathbb{X}}(\mathbb{A}) &\leq P_{\mathbb{X}}(\mathbb{D}) \\
&\leq P_{\mathbb{X}}(\mathbb{D}_1) + P_{\mathbb{X}}(\mathbb{D}_2) \\
&\leq \sum_{i=1}^{K} \sum_{j \neq i} \frac{8}{\Delta_{ij}} \mathbb{E}\left[\|h(f(x)) - \mu_i(f)\|_{\mathcal{H}_k}^2; \mathbb{M}_i \right] \\
&\leq \frac{8(K-1)}{\Delta_{\min}(f)} \sum_{i=1}^{K} \mathbb{E}[\|h(f(x)) - \mu_i(f)\|_{\mathcal{H}_k}^2; \mathbb{M}_i] \\
&\leq \frac{8(K-1)}{\Delta_{\min}(f)} \sum_{i=1}^{K} \frac{1}{P_{\mathbb{X}}(\mathbb{M}_i)} \mathbb{E}_{x,x^-}[\|h(f(x)) - h(f(x^-))\|_{\mathcal{H}_k}^2; \mathbb{M}_i \times \mathbb{M}_i] \\
&\qquad \text{(Jensen's inequality)} \\
&\leq \frac{8(K-1)}{\Delta_{\min}(f) \cdot \min_{i \in [K]} P_{\mathbb{X}}(\mathbb{M}_i)} \mathfrak{a}(f),
\end{aligned}$$

and we complete the proof. □

Proof (Proof of Theorem 2). From the definition of y, it is guaranteed that the sets $\{x \in \mathbb{X} \mid y(x) = i\}$ for $i = 1, \cdots, K$ are disjoint and satisfy the relation $\{x \in \mathbb{X} \mid y(x) = i\} \subseteq \mathbb{M}_i$ for every $i \in [K]$. Thus, Theorem 5 can apply to this case, and we obtain the result. □

D Proofs in Section 5.3

In this section, we show Theorem 3 and present additional technical contents.

D.1 Proof of Theorem 3

First, we prove Theorem 3. Before that, we present the following theorem, which is a part of the proof of Theorem 3.

Theorem 6. *Let $(X_1, X_1'), \cdots, (X_n, X_n')$ be random variables introduced in Section 5.3. Suppose that Assumption 1 holds, and suppose that n is even. Then, with probability at least $1 - \varepsilon$, the following inequality holds:*

$$\sup_{f \in \mathcal{F}} \left(-\frac{1}{n(n-1)} \sum_{i \neq j} k(f(X_i), f(X_j')) + \mathbb{E}_{X, X^-} \left[k(f(X), f(X^-)) \right] \right) \leq 2\rho \mathfrak{R}^-_{n/2}(\mathcal{Q}; s^*) + \sqrt{\frac{10b^2 \log(1/\varepsilon)}{n}},$$

where we define $\mathfrak{R}^-_{n/2}(\mathcal{Q}; s^*)$ with the symmetric group S_n of degree n:

$$\mathfrak{R}^-_{n/2}(\mathcal{Q}; s^*) := \max_{s \in S_n} \mathbb{E}_{\substack{X, X' \\ \sigma_{1:(n/2)}}} \left[\sup_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^{n/2} \sigma_i f(X_{s(2i-1)})^\top f(X'_{s(2i)}) \right].$$

We remark that in Theorem 6, we need to deal with more delicate technical matters compared to the typical generalization error bounds (e.g., Theorem 3.3 of Mohri et al. [29]), since in our setup $X_1, X_1', \cdots, X_n, X_n'$ are not necessarily independent to each other. We give the proof of Theorem 6 in Appendix D.4.

Now, we can show Theorem 3.

Proof (Proof of Theorem 3).

First observe that,

$$
\begin{aligned}
&\sup_{f\in\mathcal{F}}\left(-\widehat{L}_{\mathrm{KCL}}(f;\lambda)+L_{\mathrm{KCL}}(f;\lambda)\right)\\
&=\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}k(f(X_i),f(X_i^{'}))-\frac{\lambda}{n(n-1)}\sum_{i\neq j}k(f(X_i),f(X_j'))\right.\\
&\qquad\qquad\left.-\mathbb{E}_{X,X^+}\left[k(f(X),f(X^+))\right]+\lambda\mathbb{E}_{X,X^-}\left[k(f(X),f(X^-))\right]\right)\\
&\leq\underbrace{\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}k(f(X_i),f(X_i^{'}))-\mathbb{E}_{X,X^+}\left[k(f(X),f(X^+))\right]\right)}_{\text{(i)}}\\
&\qquad+\lambda\underbrace{\sup_{f\in\mathcal{F}}\left(-\frac{1}{n(n-1)}\sum_{i\neq j}k(f(X_i),f(X_j'))+\mathbb{E}_{X,X^-}\left[k(f(X),f(X^-))\right]\right)}_{\text{(ii)}}.
\end{aligned}
$$

Let us define the function space $\mathcal{K}:=\{k(f(\cdot),f(\cdot)):\mathbb{X}\times\mathbb{X}\to\mathbb{R}\mid f\in\mathcal{F}\}$. Then $\mathcal{K}$ is uniformly bounded with constant $b=\sup_{z,z\in\mathbb{S}^{d-1}}|k(z,z')|$. Here we note that $b<+\infty$ holds since k is continuous and $\mathbb{S}^{d-1}$ is compact; see Section 2.1. From the ULLNs (Theorem 3.3 in Mohri et al. [29]), with probability at least $1-\varepsilon/2$, we have

$$
\text{(i)}\leq 2\mathfrak{R}_n^+(\mathcal{K})+\sqrt{\frac{2b^2\log(2/\varepsilon)}{n}}.
$$

Since k is represented by $k(x,x')=\psi(x^\top x')$ for some ρ-Lipshitz function ψ from Assumption 1, by applying Talagrand's lemma (Lemma 26.9 in Shalev-Shwartz and Ben-David [35]) we have $\mathfrak{R}_n^+(\mathcal{K})\leq\rho\mathfrak{R}_n^+(\mathcal{Q})$. Hence, with probability at least $1-\varepsilon/2$, we have

$$
\text{(i)}\leq 2\rho\mathfrak{R}_n^+(\mathcal{Q})+\sqrt{\frac{2b^2\log(2/\varepsilon)}{n}}.
$$

For (ii), from Theorem 6, with probability at least $1-\varepsilon/2$ we have

$$
\text{(ii)}\leq 2\rho\mathfrak{R}_{n/2}^-(\mathcal{Q};s^*)+\sqrt{\frac{10b^2\log(2/\varepsilon)}{n}}.
$$

Therefore, with probability at least $1-\varepsilon$ we have,

$$
\begin{aligned}
&\sup_{f\in\mathcal{F}}\left(-\widehat{L}_{\mathrm{KCL}}(f;\lambda)+L_{\mathrm{KCL}}(f;\lambda)\right)\\
&\leq 2\rho\mathfrak{R}_n(\mathcal{Q})+\sqrt{\frac{2b^2\log(2/\varepsilon)}{n}}+\lambda\sqrt{\frac{10b^2\log(2/\varepsilon)}{n}},
\end{aligned}
\tag{14}
$$

where $\mathfrak{R}_n(\mathcal{Q}) := \mathfrak{R}_n^+(\mathcal{Q}) + \lambda \mathfrak{R}_{n/2}^-(\mathcal{Q}; s^*)$.

Note that in the same way as the proof of the above probability bound, we have the following inequality: with probability at least $1 - \varepsilon$,

$$\begin{aligned} &\sup_{f \in \mathcal{F}} \left(\widehat{L}_{\mathrm{KCL}}(f; \lambda) - L_{\mathrm{KCL}}(f; \lambda) \right) \\ &\leq 2\rho \mathfrak{R}_n(\mathcal{Q}) + \sqrt{\frac{2b^2 \log(2/\varepsilon)}{n}} + \lambda \sqrt{\frac{10 b^2 \log(2/\varepsilon)}{n}}. \end{aligned} \tag{15}$$

Hence, let $\widehat{f}$ be the minimizer of $\widehat{L}_{\mathrm{KCL}}(f; \lambda)$, then from (14) and (15), with probability at least $1 - 2\varepsilon$ we have

$$L_{\mathrm{KCL}}(\widehat{f}; \lambda) \leq L_{\mathrm{KCL}}(f; \lambda) + 4\rho \mathfrak{R}_n(\mathcal{Q}) + 2\sqrt{\frac{2b^2 \log(2/\varepsilon)}{n}} + 2\lambda \sqrt{\frac{10 b^2 \log(2/\varepsilon)}{n}},$$

where we note that $\widehat{L}_{\mathrm{KCL}}(\widehat{f}; \lambda) \leq \widehat{L}_{\mathrm{KCL}}(f; \lambda)$ from the definition of $\widehat{f}$. Therefore, we complete the proof. □

D.2 An Upper Bound of the Rademacher Complexity

In this section for the sake of simplicity, we consider the case in which for every $f \in \mathcal{F}$, there exists the unique function $f_0 \in \mathcal{F}_0$ such that $f(x) = f_0(x)/\|f_0(x)\|_2$ for every $x \in \mathbb{X}$. First let us recall the definition of a sub-Gaussian process:

Definition 4 (Quoted from Definition 5.16 in Wainwright [47]). *A collection of zero-mean random variables $\{X_\theta, \theta \in \mathbb{T}\}$ is a sub-Gaussian process with respect to a metric ρ_X on $\mathbb{T}$ if*

$$\mathbb{E}\left[e^{\lambda(X_\theta - X_{\widetilde{\theta}})} \right] \leq e^{\frac{\lambda^2 \rho_X^2(\theta, \widetilde{\theta})}{2}} \quad \text{for all } \theta, \widetilde{\theta} \in \mathbb{T}, \text{ and } \lambda \in \mathbb{R}.$$

Here, recall the following quantity: $\mathfrak{R}_n(\mathcal{Q}) = \mathfrak{R}_n^+(\mathcal{Q}) + \lambda \mathfrak{R}_{n/2}^-(\mathcal{Q}; s^*)$. We next upper bound the Rademacher complexity via the chaining technique (Theorem 5.22 in Wainwright [47]).

Proposition 2. *Suppose n is even. For $\mathfrak{R}_n(\mathcal{Q})$, we have the upper bound,*

$$\mathfrak{R}_n(\mathcal{Q}) \leq \frac{64(1 + \sqrt{2}\lambda)}{\mathfrak{m}(\mathcal{F}_0)\sqrt{n}} \int_0^{Cd} \sqrt{\log \mathfrak{C}(u; \mathcal{F}_0, \|\cdot\|_\infty)} du,$$

where $\|f_0\|_\infty := \sup_{x \in \mathbb{X}} \|f_0(x)\|_2$ for $f_0 \in \mathcal{F}_0$, $\mathfrak{m}(\mathcal{F}_0)$ is defined in Section 2.1, C is a constant independent of d, n, λ, and $\mathfrak{C}(u; \mathcal{F}_0, \|\cdot\|_\infty)$ is the u-covering number of $(\mathcal{F}_0, \|\cdot\|_\infty)$ (for the definition of covering number, see e.g., Definition 5.1 in Wainwright [47]).

Proof (Proof of Proposition 2). In this proof, we follow the proof idea of Tu et al. [45] (see Lemma 5 in Tu et al. [45]). Since our setup is different from Tu et al. [45], we need to modify the proof and add several new techniques. Define

$$Z_{f_0} := \frac{\mathfrak{m}(\mathcal{F}_0)}{2\sqrt{n}} \sum_{i=1}^{n} \sigma_i q(f_0(X_i), f_0(X_i')),$$

where $\sigma_1, \cdots, \sigma_n$ are Rademacher random variables that are independent to each other and to each (X_i, X_i'), $i \in [n]$, $f_0 \in \mathcal{F}_0$, $(X_1, X_1), \cdots, (X_n, X_n')$ are the random vectors defined in Section 5.3, and

$$q(z, z') = \frac{z^\top z'}{\|z\|_2 \cdot \|z'\|_2} \quad z, z' \in \mathbb{S}^{d-1}.$$

Also, let us recall the assumption for $\mathcal{F}_0$ introduced in Section 2.1: for every $f \in \mathcal{F}$, there exists the unique function $f_0 \in \mathcal{F}_0$ such that $f(x) = f_0(x)/\|f_0(x)\|_2$ for every $x \in \mathbb{X}$. We show that $\{Z_{f_0}\}_{f_0 \in \mathcal{F}_0}$ is a sub-Gaussian process as follows: note that, for every $f_{1,0}, f_{2,0} \in \mathcal{F}_0$,

$$\begin{aligned}
&\frac{\mathfrak{m}(\mathcal{F}_0)}{2\sqrt{n}} \left|\sigma_i(q(f_{1,0}(X_i), f_{1,0}(X_i')) - q(f_{2,0}(X_i), f_{2,0}(X_i')))\right| \\
&\leq \frac{\mathfrak{m}(\mathcal{F}_0)}{2\sqrt{n}} \left|f_1(X_i)^\top f_1(X_i') - f_2(X_i)^\top f_2(X_i')\right| && \text{(def. of } \sigma_i\text{)} \\
&\leq \frac{\mathfrak{m}(\mathcal{F}_0)}{2\sqrt{n}} \left(\left|f_1(X_i)^\top f_1(X_i') - f_1(X_i)^\top f_2(X_i')\right|\right. \\
&\qquad \left. + \left|f_1(X_i)^\top f_2(X_i') - f_2(X_i)^\top f_2(X_i')\right|\right) && \text{(triangle ineq.)} \\
&\leq \frac{\mathfrak{m}(\mathcal{F}_0)}{2\sqrt{n}} \left(\|f_1(X_i)\|_2 \|f_1(X_i') - f_2(X_i')\|_2 + \|f_1(X_i) - f_2(X_i)\|_2 \|f_2(X_i')\|_2\right) \\
&\qquad && \text{(Cauchy-Schwarz ineq.)} \\
&\leq \frac{\mathfrak{m}(\mathcal{F}_0)}{2\sqrt{n}} \left(\|f_1(X_i') - f_2(X_i')\|_2 + \|f_1(X_i) - f_2(X_i)\|_2\right) && \text{(def. of } f_1, f_2\text{)} \\
&\leq \frac{\mathfrak{m}(\mathcal{F}_0)}{\sqrt{n}} \sup_{x \in \mathbb{X}} \|f_1(x) - f_2(x)\|_2 \\
&= \frac{\mathfrak{m}(\mathcal{F}_0)}{\sqrt{n}} \sup_{x \in \mathbb{X}} \left\| \frac{f_{1,0}(x)}{\|f_{1,0}(x)\|_2} - \frac{f_{2,0}(x)}{\|f_{2,0}(x)\|_2} \right\|_2 && \text{(the uniqueness of } f_{1,0}, f_{2,0}\text{)} \\
&\leq \frac{1}{\sqrt{n}} \|f_{1,0} - f_{2,0}\|_\infty. && \text{(def. of } \mathfrak{m}(\mathcal{F}_0)\text{)}
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\mathbb{E}_{X_{1:n}, X'_{1:n}, \sigma_{1:n}} \left[\exp\left(t(Z_{f_{1,0}} - Z_{f_{2,0}})\right)\right] &\leq \exp\left(\frac{t^2}{2n} \|f_{1,0} - f_{2,0}\|_\infty^2\right)^n \\
&= \exp\left(\frac{t^2}{2} \|f_{1,0} - f_{2,0}\|_\infty^2\right).
\end{aligned}$$

This indicates that $\{Z_{f_0}\}_{f\in\mathcal{F}_0}$ is a sub-Gaussian process with the norm $\|\cdot\|_\infty$. Here note that $\sup_{f_{1,0},f_{2,0}\in\mathcal{F}_0}\|f_{1,0}-f_{2,0}\|_\infty \leq C\sqrt{d}$ for some constant $C\in\mathbb{R}$ that is independent of d, since $\mathcal{F}_0$ is uniformly bounded. By using the chaining theorem (Theorem 5.22 in Wainwright [47]), we have

$$\mathfrak{R}_n^+(\mathcal{Q}) \leq \frac{64}{\mathfrak{m}(\mathcal{F}_0)\sqrt{n}}\int_0^{C\sqrt{d}}\sqrt{\log \mathfrak{C}(u;\mathcal{F}_0,\|\cdot\|_\infty)}du.$$

For $\mathfrak{R}_{n/2}^-(\mathcal{Q})$, in a similar way we obtain,

$$\mathfrak{R}_{n/2}^-(\mathcal{Q}) \leq \frac{64\sqrt{2}}{\mathfrak{m}(\mathcal{F}_0)\sqrt{n}}\int_0^{C\sqrt{d}}\sqrt{\log \mathfrak{C}(u;\mathcal{F}_0,\|\cdot\|_\infty)}du.$$

Thus, we have

$$\mathfrak{R}_n(\mathcal{Q}) \leq \frac{64(1+\sqrt{2}\lambda)}{\mathfrak{m}(\mathcal{F}_0)\sqrt{n}}\int_0^{C\sqrt{d}}\sqrt{\log \mathfrak{C}(u;\mathcal{F}_0,\|\cdot\|_\infty)}du,$$

and complete the proof. □

The integral in the above upper bound is often called Dudley entropy integral [47]. Proposition 2 makes it easier to derive a generalization bound via chaining, since it is enough to evaluate the Dudley entropy integral for the function space $\mathcal{F}_0$ instead of the space of critic functions $\mathcal{Q}$.

Here, denote by $\mathfrak{D}(\mathcal{F}_0,\|\cdot\|_\infty)$, the Dudley entropy integral w.r.t. $(\mathcal{F}_0,\|\cdot\|_\infty)$, i.e.,

$$\mathfrak{D}(\mathcal{F}_0,\|\cdot\|_\infty) = \int_0^{C\sqrt{d}}\sqrt{\log \mathfrak{C}(u;\mathcal{F}_0,\|\cdot\|_\infty)}du.$$

It is shown by Tu et al. [45] that if $\mathcal{F}_0$ is a function space of feedforward (deep) neural networks, where each neural networks have weight matrices whose norms are bounded by some universal constant, and Lipschitz activation functions that vanish at the origin, then $\mathfrak{D}(\mathcal{F}_0,\|\cdot\|_\infty) < +\infty$ holds. Based on this fact, we introduce:

Assumption 3. The Dudley entropy integral $\mathfrak{D}(\mathcal{F}_0,\|\cdot\|_\infty)$ is finite, and $\mathfrak{R}_n(\mathcal{Q}) \leq O((1+\lambda)/\sqrt{n})$ holds.

Consequently, we obtain the generalization error bound.

Corollary 2. *Suppose that Assumption 1, 3 hold, and n is even. Then, with probability at least $1-\varepsilon$ where $\varepsilon>0$, we have*

$$L_{\mathrm{KCL}}(f;\lambda) \leq \widehat{L}_{\mathrm{KCL}}(f;\lambda) + O\left(\frac{(1+\lambda)\left(1+\sqrt{\log(2/\varepsilon)}\right)}{\sqrt{n}}\right).$$

Proof. Due to Theorem 3 and Assumption 3. □

D.3 Useful Results on McDiarmid's Inequality for Dependent Random Variables

Before showing Theorem 6, we need to prepare several definitions and an existing result. The following three definitions are quoted from Zhang et al. [53].

Definition 5 (Dependency Graph, quoted from Definition 3.1 in Zhang et al. [53]). *An undirected graph G is called a dependency graph of a random vector $\mathbf{X} = (X_1, \cdots, X_n)$ if*

1. *$V(G) = [n]$*
2. *if $I, J \subset [n]$ are non-adjacent in G, then $\{X_i\}_{i\in I}$ and $\{X_j\}_{j\in J}$ are independent.*

Definition 6 (Forest Approximation, quoted from Definition 3.4 in Zhang et al. [53]). *Given a graph G, a forest F, and a mapping $\phi : V(G) \to V(F)$, if $\phi(u) = \phi(v)$ or $\langle \phi(u), \phi(v)\rangle \in E(F)$ for any $\langle u, v\rangle \in E(G)$, we say that (ϕ, F) is a forest approximation of G. Let $\Phi(G)$ denote the set of forest approximations of G.*

Definition 7 (Forest Complexity, quoted from Definition 3.5 in Zhang et al. [53]). *Given a graph G and any forest approximation $(\phi, F) \in \Phi(G)$ with F consisting of trees $\{T_i\}_{i\in[k]}$, let*

$$\lambda_{(\phi,F)} = \sum_{\langle u,v\rangle\in E(F)} \left(|\phi^{-1}(u)| + |\phi^{-1}(v)|\right)^2 + \sum_{i=1}^{k} \min_{u\in V(T_i)} |\phi^{-1}(u)|^2.$$

We call

$$\Lambda(G) = \min_{(\phi,F)\in\Phi(G)} \lambda_{(\phi,F)}$$

the forest complexity of the graph G.

Zhang et al. [53] have shown the following result, which is an extension of McDiarmid's inequality [28] for dependent random variables.

Theorem 7 (Quoted from Theorem 3.6 in Zhang et al. [53]). *Suppose that $f : \boldsymbol{\Omega} \to \mathbb{R}$ is a $\mathbf{c}$-Lipschitz function and G is a dependency graph of a random vector $\mathbf{X}$ that takes values in $\boldsymbol{\Omega}$. For any $t > 0$, the following inequality holds:*

$$\Pr(f(\mathbf{X}) - \mathbf{E}[f(\mathbf{X})] \geq t) \leq \exp\left(-\frac{2t^2}{\Lambda(G)\|\mathbf{c}\|_\infty^2}\right).$$

Note that, in the above theorem $f : \boldsymbol{\Omega} \to \mathbb{R}$ is said to be $\mathbf{c}$-Lipschitz if $|f(\mathbf{x}) - f(\mathbf{x}')| \leq \sum_{i=1}^{p} \mathbf{c}_i \mathbb{1}_{\{\mathbf{x}_i \neq \mathbf{x}_i'\}}$ for every $\mathbf{x}, \mathbf{x}' \in \boldsymbol{\Omega}$, where $\boldsymbol{\Omega} \subset \mathbb{R}^p$ for some $p \in \mathbb{N}$.

D.4 Proof of Theorem 6

We show Theorem 6 by utilizing the contents in Appendix D.3. Recall the definition of the random variables introduced in Section 5.3: $(X_1, X_1'), \cdots, (X_n, X_n')$ are pairs of random variables sampled independently according to the joint probability distribution with density $w(x, x')$, where X_i and X_j' are independent for each pair of distinct indices $i, j \in [K]$. From the definition, the following claim holds.

Lemma 3. *Let G_n be a dependency graph that is defined with a random vector $(X_1, X_1', \cdots, X_n, X_n')$, where the edges in G_n are defined as follows: for any $i, j \in [n]$, X_i and X_j are not connected, and X_i and X_j' are connected by an edge if and only if $i = j$. Then, we have $\Lambda(G_n) \leq 5n$.*

Proof. Let $\phi : G_n \to G_n$ be the identity map. From the definition, G_n can be decomposed into trees $\{T_i\}_{i\in[n]}$ where $V(T_i) = \{X_i, X_i'\}$ for each $i \in [n]$. Let F be the forest consisting of the trees $\{T_i\}_{i\in[n]}$. Then, we have $\lambda_{(\phi,F)} = 5n$, which implies $\Lambda(G_n) \leq \lambda_{(\phi,F)} \leq 5n$. □

Proof (Proof of Theorem 6). The goal of this proof is to upper bound the following quantity with high probability:

$$\sup_{f\in\mathcal{F}} \left(-\frac{1}{n(n-1)} \sum_{i\neq j} k(f(X_i), f(X_j')) + \mathbb{E}_{X,X^-} \left[k(f(X_i), f(X^-)) \right] \right).$$

However, as explained before, the standard argument (see e.g., Theorem 3.3 in Mohri et al. [29]) cannot apply to this case since $k(f(X_i), f(X_j'))$, $i, j \in [n], i \neq j$ are not necessarily independent to each other from our problem setup. We instead utilize the McDiarmid's inequality for dependent random variables, which is shown by Zhang et al. [53], to avoid this problem. Our proof below is mainly based on Theorem 3.3 in Mohri et al. [29], but it includes some modification due to the application of the results by Zhang et al. [53]. Let $\widetilde{X}_1, \widetilde{X}_1', \cdots, \widetilde{X}_n, \widetilde{X}_n'$ be i.i.d. random variables to the original random variables $X_1, X_1'. \cdots, X_n, X_n'$. Define the measurable function $F(f) := F(f)(x_1, x_1', \cdots, x_n, x_n')$ on $\mathbb{X}^{2n}$ as

$$F(f) := \frac{1}{n(n-1)} \sum_{i\neq j} k(f(x_i), f(x_j')) - \mathbb{E}_{X,X^-} \left[k(f(X), f(X^-)) \right].$$

For simplicity, denote

$$F(f)_{x_\ell} := \frac{1}{n(n-1)} \left(\sum_{j:j\neq\ell} k(f(\widetilde{x}_\ell), f(x_j')) + \sum_{\substack{i,j:i\neq j \\ i\neq\ell}} k(f(x_i), f(x_j')) \right) - \mathbb{E}_{X,X^-} \left[k(f(X), f(X^-)) \right].$$

In a similar way, we also use the notation $F(f)_{x'_\ell}$. Let $j \in [n]$. Then, for every $f \in \mathcal{F}$, we have

$$\begin{aligned} F(f) - \sup_{f\in\mathcal{F}} F(f)_{X_j} &\le F(f) - F(f)_{X_j} \\ &\le \left|F(f) - F(f)_{X_j}\right| \\ &\le \left|\frac{1}{n(n-1)} \sum_{i\in[n], i\neq j} \Big(k(f(X_j), f(X'_i)) - k(f(\widetilde{X}_j), f(X'_i))\Big)\right| \\ &\le \frac{1}{n(n-1)} \cdot 2(n-1)b = \frac{2b}{n}, \end{aligned}$$

where $b := \sup_{z,z'\in\mathbb{S}^{d-1}} |k(z,z')|$. Hence, $\sup_{f\in\mathcal{F}} F(f) - \sup_{f\in\mathcal{F}} F(f)_{X_j} \le \frac{2b}{n}$. By applying the same argument several times, $\sup_{f\in\mathcal{F}} F(f)$ satisfies the assumption of Theorem 7. Therefore, from Theorem 7 (i.e., Theorem 3.6 in Zhang et al. [53]) and Lemma 3, with probability at least $1-\varepsilon$ we have

$$\sup_{f\in\mathcal{F}} F(f) \le \mathbb{E}\left[\sup_{f\in\mathcal{F}} F(f)\right] + \sqrt{\frac{10b^2 \log(1/\varepsilon)}{n}}. \tag{16}$$

Let $\sigma_{1:n/2} := (\sigma_1, \cdots, \sigma_{n/2})$ be a random vector that consists of a Rademacher random variable (i.e., a random variable taking ± 1 with probability $1/2$ each) for each entry, and let $\widetilde{X}_{1:n}, \widetilde{X}'_{1:n}$ be i.i.d. copies of the random vectors $X_{1:n}, X'_{1:n}$,

respectively. Denote $m = n/2 \in \mathbb{N}$. Then,

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}} F(f)\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n(n-1)}\sum_{i\neq j} k(f(X_i), f(X'_j)) - \mathbb{E}_{X,X^-}\left[k(f(X), f(X^-))\right]\right)\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n(n-1)}\sum_{i\neq j} k(f(X_i), f(X'_j)) - \mathbb{E}_{\widetilde{X}_{1:n},\widetilde{X}'_{1:n}}\left[\frac{1}{n(n-1)}\sum_{i\neq j} k(f(\widetilde{X}_i), f(\widetilde{X}'_j))\right]\right)\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n!m}\sum_{s\in S_n}\left(\sum_{i=1}^{m} k(f(X_{s(2i-1)}), f(X'_{s(2i)})) - \mathbb{E}_{\widetilde{X}'_{1:n},\widetilde{X}'_{1:n}}\left[\sum_{i=1}^{m} k(f(\widetilde{X}_{s(2i-1)}), f(\widetilde{X}'_{s(2i)}))\right]\right)\right)\right] \qquad (17) \\
&\leq \frac{1}{n!}\sum_{s\in S_n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m} k(f(X_{s(2i-1)}), f(X'_{s(2i)})) - \mathbb{E}_{\widetilde{X}_{1:n},\widetilde{X}'_{1:n}}\left[\frac{1}{m}\sum_{i=1}^{m} k(f(\widetilde{X}_{s(2i-1)}), f(\widetilde{X}'_{s(2i)}))\right]\right)\right] \\
&\leq \frac{1}{n!}\sum_{s\in S_n}\mathbb{E}_{\substack{X_{1:n},X'_{1:n}\\ \widetilde{X}_{1:n},\widetilde{X}'_{1:n}}}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\left(k(f(X_{s(2i-1)}), f(X'_{s(2i)})) - k(f(\widetilde{X}_{s(2i-1)}), f(\widetilde{X}'_{s(2i)}))\right)\right)\right] \\
&= \frac{1}{n!}\sum_{s\in S_n}\mathbb{E}_{\substack{X_{1:n},X'_{1:n}\\ \widetilde{X}_{1:n},\widetilde{X}'_{1:n}\\ \sigma_{1:m}}}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(k(f(X_{s(2i-1)}), f(X'_{s(2i)})) - k(f(\widetilde{X}_{s(2i-1)}), f(\widetilde{X}'_{s(2i)}))\right)\right)\right] \qquad (18) \\
&\leq \frac{2}{n!}\sum_{s\in S_n}\mathbb{E}_{\substack{X_{1:n},X'_{1:n}\\ \sigma_{1:m}}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i k(f(X_{s(2i-1)}), f(X'_{s(2i)}))\right] \\
&\leq \frac{2\rho}{n!}\sum_{s\in S_n}\mathbb{E}_{\substack{X_{1:n},X'_{1:n}\\ \sigma_{1:m}}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(X_{s(2i-1)})^\top f(X'_{s(2i)})\right] \qquad (19) \\
&\leq 2\rho\max_{s\in S_n}\mathbb{E}_{\substack{X_{1:n},X'_{1:n}\\ \sigma_{1:m}}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(X_{s(2i-1)})^\top f(X'_{s(2i)})\right] \\
&= 2\rho\mathfrak{R}^-_m(\mathcal{Q}; s^*),
\end{aligned}
$$

where in (17) we define S_n as the symmetric group of degree n (see Remark 3 for the relation to the *average of "sums-of-i.i.d." blocks* technique for U-statistics which is explained in Clémençon et al. [10]). Besides in (18), for every $s \in S_n$ the random vectors $(X_{s(2i-1)}, X_{s(2i)})$, $(\widetilde{X}_{s(2i-1)}, \widetilde{X}_{s(2i)})$ for $i = 1, \cdots, m$ are independent and identically distributed, which implies that the standard symmetrization argument (Theorem 4.10 of Wainwright [47]) is applicable. Finally, in (19), under Assumption 1, we apply Talagrand's lemma (Lemma 26.9 in Shalev-Shwartz

and Ben-David [35]). Therefore, we obtain with probability at least $1-\varepsilon$,

$$\sup_{f\in\mathcal{F}} F(f) \leq 2\rho\mathfrak{R}^{-}_{n/2}(\mathcal{Q};s^*) + \sqrt{\frac{10b^2\log(1/\varepsilon)}{n}}.$$

Thus, we obtain the claim.

Remark 3. In (17) of the proof of Theorem 6, we use the identity,

$$\frac{1}{n(n-1)}\sum_{i\neq j} k(f(X_i),f(X'_j)) = \frac{1}{n!m}\sum_{s\in S_n}\sum_{i=1}^{m} k(f(X_{s(2i-1)},f(X'_{s(2i)})).$$

We notice that the above identity is closely related to the *average of "sum-of-i.i.d." blocks* technique explained in Appendix A of Clémençon et al. [10]. As well as the technique presented in Clémençon et al. [10], in (17) of our paper we also decompose the sum $\sum_{i\neq j} k(f(X_i),f(X'_j))$ into the sums of the i.i.d. random variables. However, we remark that the definition of the sum $\sum_{i\neq j} k(f(X_i),f(X'_j))$ is different from that presented in Clémençon et al. [10]: indeed, in our case, the random variables $f(X_1),f(X'_1),\cdots,f(X_n),f(X'_n)$ are not necessarily independent of each other. To address this problem, we decompose our sum in (17) as follows: for $2n$ random variables $X_1,X'_1,\cdots,X_n,X'_n$, we create the tuples $(X_{s(1)},X'_{s(2)},\cdots,X_{s(n-1)},X'_{s(n)})$ where $s\in S_n$, then sum up all the components $\{\sum_{i=1}^{n} k(f(X_{s(2i-1)}),f(X'_{s(2i)}))\}_{s\in S_n}$.

E Proof in Section 5.4

Proof (Proof of Theorem 4). First applying Theorem 2 to the empirical loss minimizer $\widehat{f}$, we have

$$L_{\mathrm{Err}}(\widehat{f},W_\mu,\beta_\mu;y) \leq \frac{8(K-1)}{\Delta_{\min}(\widehat{f})\cdot\min_{i\in[K]}P_{\mathbb{X}}(\mathbb{M}_i)}\mathfrak{a}(\widehat{f}). \tag{20}$$

Using Theorem 1, we have the inequality,

$$\mathfrak{a}(\widehat{f}) \leq L_{\mathrm{KCL}}(\widehat{f};\lambda) + (1-\frac{\delta}{2})\mathfrak{a}(\widehat{f}) - \lambda\mathfrak{c}(\widehat{f}) + R(\lambda). \tag{21}$$

Combining (20) and (21), we obtain

$$\begin{aligned} &L_{\mathrm{Err}}(\widehat{f},W_\mu,\beta_\mu;y)\\ &\leq \frac{8(K-1)}{\Delta_{\min}(\widehat{f})\cdot\min_{i\in[K]}P_{\mathbb{X}}(\mathbb{M}_i)}\left(L_{\mathrm{KCL}}(\widehat{f};\lambda) + (1-\frac{\delta}{2})\mathfrak{a}(\widehat{f}) - \lambda\mathfrak{c}(\widehat{f}) + R(\lambda)\right). \end{aligned} \tag{22}$$

Here, using the standard technique for upper bounding the optimal classification loss or error [3,4], the classification error $L_{\mathrm{Err}}(\widehat{f},W_\mu,\beta_\mu;y)$ is lower bounded as

$$L_{\mathrm{Err}}(\widehat{f},W^*\beta^*;y) = \inf_{W,\beta} L_{\mathrm{Err}}(\widehat{f},W,\beta;y) \leq L_{\mathrm{Err}}(\widehat{f},W_\mu,\beta_\mu;y). \tag{23}$$

From (22) and (23),

$$L_{\mathrm{Err}}(\widehat{f}, W^*, \beta^*; y) \leq \frac{8(K-1)}{\Delta_{\min}(\widehat{f}) \cdot \min_{i\in[K]} P_{\mathbb{X}}(\mathbb{M}_i)} \left(L_{\mathrm{KCL}}(\widehat{f}; \lambda) + (1 - \frac{\delta}{2})\mathfrak{a}(\widehat{f}) - \lambda \mathfrak{c}(\widehat{f}) + R(\lambda) \right). \tag{24}$$

Applying Theorem 3 to (24), we obtain: with probability at least $1 - 2\varepsilon$,

$$L_{\mathrm{Err}}(\widehat{f}, W_\mu, \beta_\mu; y) \lesssim L_{\mathrm{KCL}}(f; \lambda) + (1 - \frac{\delta}{2})\mathfrak{a}(\widehat{f}) - \lambda \mathfrak{c}(\widehat{f}) + R(\lambda) + 2\mathrm{Gen}(n, \lambda, \varepsilon),$$

where $\lesssim$ omits the coefficient $\frac{8(K-1)}{\Delta_{\min}(\widehat{f}) \cdot \min_{i\in[K]} P_{\mathbb{X}}(\mathbb{M}_i)}$. Therefore, we obtain the result. □

F Additional Information, Results, and Discussion

This section includes additional information, results, and discussion that are not presented in the main body.

F.1 Examples Satisfying Assumption 2

Proofs in Example 1 We show the several claims that appear in Example 1 as a proposition.

Proposition 3. *Let $r > 0$, $K \in \mathbb{N}$, and $v_1, \cdots, v_K \in \mathbb{R}^p$. For each $i \in [K]$, let $\mathbb{B}_i \subset \mathbb{R}^p$ be the open ball of radius r centered at a point v_i. Suppose $\mathbb{B}_1, \cdots, \mathbb{B}_K$ are disjoint to each other. Define $\overline{\mathbb{X}} = \bigcup_{i=1}^K \mathbb{B}_i$, $\mathbb{X} = \overline{\mathbb{X}}$, and the conditional probability $a(x|\overline{x}) = \mathrm{vol}(\mathbb{B}_1)^{-1} \sum_{i=1}^K \mathbb{1}_{\mathbb{B}_i \times \mathbb{B}_i}(x, \overline{x})$, where $\mathrm{vol}(\mathbb{B}_1)$ be the volume of $\mathbb{B}_i$ in $\mathbb{R}^p$. Let $p_{\overline{\mathbb{X}}}(\overline{x}) := (K\mathrm{vol}(\mathbb{B}_1))^{-1}$ be a probability density function of $P_{\overline{\mathbb{X}}}$. Define $y : \mathbb{X} \to [K]$ as $y(x) = i$ if $x \in \mathbb{B}_i$. Then, we have the following properties:*

1. *$w(x) > 0$ for every $x \in \mathbb{X}$.*
2. *$\mathrm{sim}(x, x'; \lambda) = K\mathbb{1}_{\bigcup_{i\in[K]} \mathbb{B}_i \times \mathbb{B}_i}(x, x') - \lambda$ for every $x, x' \in \mathbb{X}$.*
3. *Let $\delta \in (-\lambda, K - \lambda]$. Then, δ, K, $\mathbb{B}_1, \cdots, \mathbb{B}_K$, and y satisfy Assumption 2.*

Proof. We first show the claim 1. From the definition of $w(x)$, for every $x \in \mathbb{B}_1$ we have

$$w(x) = \int_{\overline{\mathbb{X}}} a(x|\overline{x}) p_{\overline{\mathbb{X}}}(\overline{x}) d\overline{x} = \int_{\mathbb{B}_1} \frac{1}{K\left(\mathrm{vol}(\mathbb{B}_1)\right)^2} d\overline{x} = \frac{1}{K\mathrm{vol}(\mathbb{B}_1)}.$$

Similarly, for each $i \in [K]$ we obtain $w(x) = (K\mathrm{vol}(\mathbb{B}_1))^{-1}$ for every $x \in \mathbb{B}_i$. Since $\mathbb{X} = \overline{\mathbb{X}} = \bigcup_{i=1}^K \mathbb{B}_i$, we have that $w(x) > 0$ for every $x \in \mathbb{X}$.

Next, let us show the claim 2. From the claim 1, the function $\mathrm{sim}(x, x'; \lambda)$ is well-defined. To compute $\mathrm{sim}(x, x'; \lambda)$, we need to know the function $w(x, x')$. The computation of $w(x, x')$ is done as follows:

$$\begin{aligned} w(x, x') &= \int_{\overline{\mathbb{X}}} a(x|\overline{x}) a(x'|\overline{x}) p_{\overline{\mathbb{X}}}(\overline{x}) d\overline{x} \\ &= \begin{cases} \int_{\mathbb{B}_i} \frac{1}{K(\mathrm{vol}(\mathbb{B}_1))^3} d\overline{x} & \text{if } x, x' \in \mathbb{B}_i \text{ for some } i \in [K] \\ 0 & \text{if } x \in \mathbb{B}_i \text{ and } x' \in \mathbb{B}_j \text{ for some } i \neq j \end{cases} \\ &= \begin{cases} \frac{1}{K(\mathrm{vol}(\mathbb{B}_1))^2} & \text{if } x, x' \in \mathbb{B}_i \text{ for some } i \in [K] \\ 0 & \text{if } x \in \mathbb{B}_i \text{ and } x' \in \mathbb{B}_j \text{ for some } i \neq j. \end{cases} \end{aligned}$$

Hence, it is obvious that the claim 2 holds.

Finally, let us prove the claim 3. However, from the claim 2 we see that $\mathrm{sim}(x, x'; \lambda) \geq \delta$ if and only if $x, x' \in \mathbb{B}_i$ for some $i \in [K]$. Furthermore, y is well-defined and the set $\{x \in \mathbb{X} \mid y(x) = i\} = \mathbb{B}_i$ is measurable for every $i \in [K]$. Thus, the claim 3 is also true, and we end the proof. □

An Example When Clusters Overlap Here, we also deal with an example where the clusters in $\mathbb{X}$ have some overlap. In the following proposition, for the sake of simplicity, we consider the case that there are two clusters in $\mathbb{X}$.

Proposition 4. *Let $r > 0$, and $v_1, v_2 \in \mathbb{R}^p$. For each $i \in \{1, 2\}$, let $\mathbb{B}(v_i; r) \subset \mathbb{R}^p$ be the open ball of radius r centered at point v_i. Suppose that $\|v_1 - v_2\|_2 = 3r$. Define $\overline{\mathbb{X}} = \mathbb{B}(v_1; r) \cup \mathbb{B}(v_2; r)$, $\mathbb{X} = \mathbb{B}(v_1; 2r) \cup \mathbb{B}(v_2; 2r)$, and $a(x|\overline{x}) = \mathrm{vol}(\mathbb{B}(v_1; 2r))^{-1} \sum_{i=1}^{2} \mathbb{1}_{\mathbb{B}(v_i;2r) \times \mathbb{B}(v_i;r)}(x, \overline{x})$. Let $p_{\overline{\mathbb{X}}}(\overline{x}) := (2 \cdot \mathrm{vol}(\mathbb{B}(v_1; r)))^{-1}$ be a probability density function of $P_{\overline{\mathbb{X}}}$. Define $y : \mathbb{X} \to \{1, 2\}$ as $y(x) = 1$ if $x \in \mathbb{B}(v_1; 2r)$ and $y(x) = 2$ if $x \in \mathbb{B}(v_2; 2r) \setminus \mathbb{B}(v_1; 2r)$. Then, we have the following results:*

1. *$w(x) > 0$ for every $x \in \mathbb{X}$.*
2. *$\mathrm{sim}(x, x'; \lambda) = 2 - \lambda$ if $x, x' \in \mathbb{B}(v_1; 2r) \setminus \mathbb{B}(v_2; 2r)$ or $x, x' \in \mathbb{B}(v_2; 2r) \setminus \mathbb{B}(v_1; 2r)$, $\mathrm{sim}(x, x'; \lambda) = -\lambda$ if $(x, x') \in (\mathbb{B}(v_1; 2r) \setminus \mathbb{B}(v_2; 2r)) \times (\mathbb{B}(v_2; 2r) \setminus \mathbb{B}(v_1; 2r))$ or $(x, x') \in (\mathbb{B}(v_2; 2r) \setminus \mathbb{B}(v_1; 2r)) \times (\mathbb{B}(v_1; 2r) \setminus \mathbb{B}(v_2; 2r))$, and $\mathrm{sim}(x, x'; \lambda) = 1 - \lambda$ otherwise.*
3. *Let $\delta \in (-\lambda, 1 - \lambda]$. Then, δ, K, $\mathbb{B}_1, \cdots, \mathbb{B}_K$, and y satisfy Assumption 2.*

Proof. Let $\overline{\mathbb{X}}_1$ (resp. $\overline{\mathbb{X}}_2$) denote $\mathbb{B}(v_1; r)$, (resp. $\mathbb{B}(v_2; r)$). Then,

$$\begin{aligned} w(x) &= \mathbb{E}\left[a(x|\overline{x})\right] \\ &= \int_{\overline{\mathbb{X}}} a(x|\overline{x}) p_{\overline{\mathbb{X}}}(\overline{x}) d\overline{x} \\ &= \int_{\overline{\mathbb{X}}_1} a(x|\overline{x}) p_{\overline{\mathbb{X}}}(\overline{x}) d\overline{x} + \int_{\overline{\mathbb{X}}_2} a(x|\overline{x}) p_{\overline{\mathbb{X}}}(\overline{x}) d\overline{x} \\ &= p_{\overline{\mathbb{X}}}(\overline{x}) \left\{ \int_{\overline{\mathbb{X}}_1} a(x|\overline{x}) d\overline{x} + \int_{\overline{\mathbb{X}}_2} a(x|\overline{x}) d\overline{x} \right\} \qquad (p_{\overline{\mathbb{X}}} \text{ is a constant function}) \end{aligned}$$

Here, we consider Case 1 and Case 2. Firstly, Case 1 is when either $x \in \mathbb{B}(v_1; 2r) \setminus \mathbb{B}(v_2; 2r)$ or $x \in \mathbb{B}(v_2; 2r) \setminus \mathbb{B}(v_1; 2r)$ holds. Since in this case, it is sufficient to prove for the case that $x \in \mathbb{B}(v_1; 2r) \setminus \mathbb{B}(v_2; 2r)$ holds, we may assume this condition. Then, $\int_{\overline{\mathbb{X}}_1} a(x|\overline{x})d\overline{x} = \mathrm{vol}(\mathbb{B}(v_1; r))\mathrm{vol}(\mathbb{B}(v_1; 2r))^{-1}$ and $\int_{\overline{\mathbb{X}}_2} a(x|\overline{x})d\overline{x} = 0$. Thus, $w(x) = 1/(2\mathrm{vol}(\mathbb{B}(v_1; 2r)))$. Secondly, Case 2 is when $x \in \mathbb{B}(v_1; 2r) \cap \mathbb{B}(v_2; 2r)$. Then, $\int_{\overline{\mathbb{X}}_1} a(x|\overline{x})d\overline{x} = \int_{\overline{\mathbb{X}}_2} a(x|\overline{x})d\overline{x} = \mathrm{vol}(\mathbb{B}(v_1; r))\mathrm{vol}(\mathbb{B}(v_1; 2r))^{-1}$. Thus, $w(x) = 1/\mathrm{vol}(\mathbb{B}(v_1; 2r))$. Since $r > 0$, it implies $\mathrm{vol}(\mathbb{B}(v_1; 2r)) > 0$. Thus $w(x) > 0$ for both cases.

Next, we compute

$$\begin{aligned} w(x, x') &= \mathbb{E}_{\overline{x}}\left[a(x|\overline{x})a(x'|\overline{x})\right] \\ &= \int_{\overline{\mathbb{X}}} a(x|\overline{x})a(x'|\overline{x})p_{\overline{\mathbb{X}}}(\overline{x})d\overline{x} \\ &= \int_{\overline{\mathbb{X}}_1} a(x|\overline{x})a(x'|\overline{x})p_{\overline{\mathbb{X}}}(\overline{x})d\overline{x} \ + \int_{\overline{\mathbb{X}}_2} a(x|\overline{x})a(x'|\overline{x})p_{\overline{\mathbb{X}}}(\overline{x})d\overline{x} \\ &= p_{\overline{\mathbb{X}}}(\overline{x}) \left\{ \int_{\overline{\mathbb{X}}_1} a(x|\overline{x})a(x'|\overline{x})d\overline{x} + \int_{\overline{\mathbb{X}}_2} a(x|\overline{x})a(x'|\overline{x})d\overline{x} \right\}. \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (p_{\overline{\mathbb{X}}} \text{ is a constant function}) \end{aligned}$$

Here, we consider Case A, Case B, Case C, and Case D. Firstly Case A is that both x and x' belong to $\mathbb{B}(v_1; 2r) \setminus \mathbb{B}(v_2; 2r)$ (note that the computation for the case that both x and x' belong to $\mathbb{B}(v_2; 2r) \setminus \mathbb{B}(v_1; 2r)$ is the same). Then, $\int_{\overline{\mathbb{X}}_1} a(x|\overline{x})a(x'|\overline{x})d\overline{x} = \mathrm{vol}(\mathbb{B}(v_1; r))/\mathrm{vol}(\mathbb{B}(v_1; 2r))^2$ and $\int_{\overline{\mathbb{X}}_2} a(x|\overline{x})a(x'|\overline{x})d\overline{x} = 0$. Hence, $w(x, x') = \{2(\mathrm{vol}(\mathbb{B}(v_1; 2r)))^2\}^{-1}$. Here recall that $w(x) = w(x') = 1/(2\mathrm{vol}(\mathbb{B}(v_1; 2r)))$, then we have $\mathrm{sim}(x, x'; \lambda) = 2 - \lambda$. Secondly Case B is that $x \in \mathbb{B}(v_1; 2r) \setminus \mathbb{B}(v_2; 2r)$ and $x' \in \mathbb{B}(v_2; 2r) \setminus \mathbb{B}(v_1; 2r)$ (the calculation for the case that $x \in \mathbb{B}(v_2; 2r) \setminus \mathbb{B}(v_1; 2r)$ and $x' \in \mathbb{B}(v_1; 2r) \setminus \mathbb{B}(v_2; 2r)$ is the same). Then, $\int_{\overline{\mathbb{X}}_1} a(x|\overline{x})a(x'|\overline{x})d\overline{x} = \int_{\overline{\mathbb{X}}_2} a(x|\overline{x})a(x'|\overline{x})d\overline{x} = 0$. Therefore, $\mathrm{sim}(x, x'; \lambda) = 0 - \lambda = -\lambda$. Thirdly Case C is that both x and x' belong to $\mathbb{B}(v_1; 2r) \cap \mathbb{B}(v_2; 2r)$. Then, $\int_{\overline{\mathbb{X}}_1} a(x|\overline{x})a(x'|\overline{x})d\overline{x} = \int_{\overline{\mathbb{X}}_2} a(x|\overline{x})a(x'|\overline{x})d\overline{x} = \mathrm{vol}(\mathbb{B}(v_1; r))/\mathrm{vol}(\mathbb{B}(v_1; 2r))^2$. Since $w(x) = w(x') = 1/\mathrm{vol}(\mathbb{B}(v_1; 2r))$, $\mathrm{sim}(x, x'; \lambda) = 1 - \lambda$. Finally in Case D, consider the complementary of the union of the other cases. From the setting, we may assume that x belongs to $\mathbb{B}(v_1; 2r) \cap \mathbb{B}(v_2; 2r)$ and x' to $\mathbb{B}(v_1; 2r) \setminus \mathbb{B}(v_2; 2r)$. Then, $\int_{\overline{\mathbb{X}}_1} a(x|\overline{x})a(x'|\overline{x})d\overline{x} = \mathrm{vol}(\mathbb{B}(v_1; r))/\mathrm{vol}(\mathbb{B}(v_1; 2r))^2$ and $\int_{\overline{\mathbb{X}}_2} a(x|\overline{x})a(x'|\overline{x})d\overline{x} = 0$. Since $w(x) = 1/\mathrm{vol}(\mathbb{B}(v_1; 2r))$ and $w(x') = 1/(2\mathrm{vol}(\mathbb{B}(v_1; 2r)))$, we have $\mathrm{sim}(x, x'; \lambda) = 1 - \lambda$. As a result,

$$\mathrm{sim}(x, x'; \lambda) = \begin{cases} 2 - \lambda, \text{ if Case A holds,} \\ -\lambda, \ \text{ if Case B holds,} \\ 1 - \lambda, \text{ if Case C holds,} \\ 1 - \lambda, \text{ if Case D holds.} \end{cases}$$

Finally, take $\delta \in (-\lambda, 1 - \lambda]$. Then, from the computation for $\mathrm{sim}(x, x'; \lambda)$ above, the conditions in Assumption 2 are satisfied. □

F.2 SSL-HSIC Revisit

Li et al. [26] propose the framework termed SSL-HSIC, which is defined using the notion Hilbert-Schmidt Independence Criterion (HSIC, see e.g. [15,37]). They show that under some conditions, for a random variable Z (resp. Y) that represents the feature vector (resp. the label), one obtains

$$\mathrm{HSIC}(Z,Y) = c\left(\mathbb{E}_{x,x^+}\left[k(f(x),f(x^+))\right] - \mathbb{E}_{x,x^-}\left[k(f(x),f(x^-))\right]\right),$$

where $c > 0$. Li et al. [26] define the loss of SSL-HSIC as,

$$L_{\text{SSL-HSIC}}(f;\kappa) = -\mathrm{HSIC}(Z,Y) + \kappa\sqrt{\mathrm{HSIC}(Z,Z)},$$

where $\kappa \in \mathbb{R}$.

In the case that $\kappa > 0$, we have

$$L_{\mathrm{KCL}}(f;1) \lesssim L_{\text{SSL-HSIC}}(f;\kappa).$$

F.3 Supplementary Information of Section 3

Relations to Variants of InfoNCE We first define variants of InfoNCE [6,32]:

- Decoupled InfoNCE loss, which is a variant of the decoupled NT-Xent loss of Chen et al. [7]:

$$\begin{aligned}&\widetilde{L}_{\mathrm{NCE}}(f;\tau,\lambda)\\ &= -\mathbb{E}_{x,x^+}\left[\frac{f(x)^\top f(x^+)}{\tau}\right] + \lambda\mathbb{E}_{\substack{x,x^+\\ \{x_i^-\}}}\left[\log\left(e^{\frac{f(x)^\top f(x^+)}{\tau}} + \sum_{i=1}^{M} e^{\frac{f(x)^\top f(x_i^-)}{\tau}}\right)\right].\end{aligned}$$

- Asymptotic of contrastive loss [48] decoupled by following the way of Chen et al. [7]:

$$\widetilde{L}_{\infty\text{-NCE}}(f;\tau,\lambda) = -\mathbb{E}_{x,x^+}\left[\frac{f(x)^\top f(x^+)}{\tau}\right] + \lambda\mathbb{E}_{x}\left[\log\mathbb{E}_{x'}\left[e^{\frac{f(x)^\top f(x')}{\tau}}\right]\right],$$

- InfoNCE loss as a variant of decoupled contrastive learning loss [51]:

$$\widetilde{L}_{\mathrm{NCE}}(f;\tau,1) = -\mathbb{E}_{x,x^+}\left[\frac{f(x)^\top f(x^+)}{\tau}\right] + \mathbb{E}_{x,\{x_i^-\}}\left[\log\left(\sum_{i=1}^{M} e^{\frac{f(x)^\top f(x_i^-)}{\tau}}\right)\right].$$

- InfoNCE loss as a variant of decoupled contrastive learning loss with additional weight parameter, following Chen et al. [7]:

$$\widetilde{L}_{\mathrm{NCE}}(f;\tau,\lambda) = -\mathbb{E}_{x,x^+}\left[\frac{f(x)^\top f(x^+)}{\tau}\right] + \lambda\mathbb{E}_{x,\{x_i^-\}}\left[\log\left(\sum_{i=1}^{M} e^{\frac{f(x)^\top f(x_i^-)}{\tau}}\right)\right].$$

Note that $L_{\mathrm{NCE}}(f;\tau)$ and $L_{\infty\text{-NCE}}(f;\tau)$ in Section 2 coincide with $\widetilde{L}_{\mathrm{NCE}}(f;\tau,1)$ and $\widetilde{L}_{\infty\text{-NCE}}(f;\tau,1)$ in this subsection, respectively. We show the following facts:

Proposition 5. *The following relations hold:*

$$\tau^{-1}L_{\mathrm{LinKCL}}(f;\lambda) \leq \widetilde{L}_{\mathrm{NCE}}(f;\tau,\lambda) + \lambda \log M^{-1}, \tag{25}$$

$$\tau^{-1}L_{\mathrm{LinKCL}}(f;\lambda) \leq \widetilde{L}_{\infty\text{-NCE}}(f;\tau,\lambda), \tag{26}$$

$$\tau^{-1}L_{\mathrm{LinKCL}}(f;\lambda) \leq \widetilde{L}_{\mathrm{NCE}}(f;\tau,\lambda) + \lambda \log M^{-1}. \tag{27}$$

Proof. From the definition of $L_{\mathrm{NCE}}(f;\tau,\lambda)$, we have

$$\begin{aligned}
&\widetilde{L}_{\mathrm{NCE}}(f;\tau,\lambda) + \lambda \log \frac{1}{M} \\
&= -\tau^{-1}\mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right] \\
&\qquad + \lambda \mathbb{E}_{x,x^+,\{x_i^-\}}\left[\log\left(\frac{1}{M}e^{f(x)^\top f(x^+)/\tau} + \frac{1}{M}\sum_{i=1}^{M} e^{f(x)^\top f(x_i^-)/\tau}\right)\right] \\
&\geq -\tau^{-1}\mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right] + \lambda \mathbb{E}_{x,\{x_i^-\}}\left[\log\left(\frac{1}{M}\sum_{i=1}^{M} e^{f(x)^\top f(x_i^-)/\tau}\right)\right] \\
&\geq -\tau^{-1}\mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right] + \tau^{-1}\lambda \mathbb{E}_{x,\{x_i^-\}}\left[\frac{1}{M}\sum_{i=1}^{M} f(x)^\top f(x_i^-)\right] \\
&= \tau^{-1}L_{\mathrm{LinKCL}}(f;\lambda),
\end{aligned}$$

where in the first inequality we use the fact that $M^{-1}e^{f(x)^\top f(x^+)/\tau} \geq 0$ for any $x, x^+ \in \mathbb{X}$, and in the second inequality we use Jensen's inequality. Note that when $\lambda = 1$, we obtain (1).

The proofs of (27) are almost the same as the proof of (25). The equation (26) is obtained by applying Jensen's inequality. □

Relations to SCL Let us define the quadratic kernel contrastive loss as:

$$L_{\mathrm{QKCL}}(f;\lambda) = -\mathbb{E}_{x,x^+}\left[\left(f(x)^\top f(x^+)\right)^2\right] + \lambda \mathbb{E}_{x,x^-}\left[\left(f(x)^\top f(x^-)\right)^2\right].$$

The spectral contrastive loss $L_{\mathrm{SCL}}(f)$ [18] is defined as,

$$L_{\mathrm{SCL}}(f) = -2\mathbb{E}_{x,x^+}[f(x)^\top f(x^+)] + \mathbb{E}_{x,x^-}[(f(x)^\top f(x^-))^2]. \tag{28}$$

The following proposition is an elementary result.

Proposition 6. *We have,*

$$L_{\mathrm{QKCL}}(f;2^{-1}) \leq \frac{1}{2}L_{\mathrm{SCL}}(f) + \frac{1}{4}.$$

Proof. Since $t^2 + 1/4 \geq t$ for every $t \in \mathbb{R}$, we obtain the claim. □

F.4 Comparison of Assumption 2 of our work to Assumption 3 in HaoChen and Ma [17]

Let $\mathbb{M}$ be a measurable subset of $\mathbb{X}$, and let $g : \mathbb{X} \to \mathbb{R}$ be a function. HaoChen and Ma [17] introduce the following notion that quantifies the inner-connectivity of clusters (see (4) in HaoChen and Ma [17]):

$$Q_{\mathbb{M}}(g) := \frac{\mathbb{E}_{x,x^+}[(g(x) - g(x^+))^2 | \mathbb{M} \times \mathbb{M}]}{\mathbb{E}_{x,x^-}[(g(x) - g(x^-))^2 | \mathbb{M} \times \mathbb{M}]}.$$

Here, the expectations above are defined as

$$\mathbb{E}_{x,x^+}[(g(x) - g(x^+))^2 | \mathbb{M} \times \mathbb{M}] = \int_{\mathbb{X}\times\mathbb{X}} (g(x) - g(x'))^2 P_+(dx, dx' | \mathbb{M} \times \mathbb{M}),$$
$$\mathbb{E}_{x,x^-}[(g(x) - g(x^-))^2 | \mathbb{M} \times \mathbb{M}] = \int_{\mathbb{X}} \int_{\mathbb{X}} (g(x) - g(x'))^2 P_{\mathbb{X}}(dx | \mathbb{M}) P_{\mathbb{X}}(dx' | \mathbb{M}),$$

where we use the notation $dP_+ = w(x, x')d\nu_{\mathbb{X}}^{\otimes 2}$. We focus on the following notion, where HaoChen and Ma [17] denote their subsets by $\{S_1, \cdots, S_m\}$.

Assumption 4 ($\mathcal{F}$-implementable inner-cluster connection larger than β, quoted from Assumption 3 in HaoChen and Ma [17]). For any function $f \in \mathcal{F}$ and any linear head $w \in \mathbb{R}^k$, let function $g(x) = w^\top f(x)$. For any $i \in [m]$ we have that:

$$Q_{S_i}(g) \geq \beta.$$

In summary, the relation between Assumption 3 of HaoChen and Ma [17] and Assumption 2 of our work is given below:

Proposition 7. *Suppose that Assumption 2 holds. Take $\delta \in \mathbb{R}$, $K \in \mathbb{N}$, and $\mathbb{M}_1, \cdots, \mathbb{M}_K$ such that the conditions* **(A)** *and* **(B)** *are satisfied. Suppose also that: 1) there exists some $c > 0$ such that for every $i \in [K]$, $c \cdot P_+(\mathbb{M}_i \times \mathbb{M}_i) \leq P_{\mathbb{X}}(\mathbb{M}_i)^2$ holds; 2) $\delta + \lambda \geq 0$ holds. Then, the function class $\widetilde{\mathcal{F}}$ including all the maps from $\mathbb{X}$ to $\mathbb{S}^{d-1}$ satisfies Assumption 3 in HaoChen and Ma [17].*

Proof. Take an arbitrary $f \in \widetilde{\mathcal{F}}$ and $w \in \mathbb{R}^d$. For each $i \in [K]$, for any $x, x' \in \mathbb{M}_i$ we have that $w(x, x') \geq (\delta + \lambda) w(x) w(x')$. Since $\delta + \lambda \geq 0$,

$$\int_{\mathbb{M}_i \times \mathbb{M}_i} (g(x) - g(x'))^2 w(x, x') \nu_{\mathbb{X}}^{\otimes 2}(dx, dx')$$
$$\geq (\delta + \lambda) \int_{\mathbb{M}_i \times \mathbb{M}_i} (g(x) - g(x'))^2 w(x) w(x') \nu_{\mathbb{X}}^{\otimes 2}(dx, dx').$$

Here, using $c \cdot P_+(\mathbb{M}_i \times \mathbb{M}_i) \leq P_{\mathbb{X}}(\mathbb{M}_i)^2$, we have

$$\frac{1}{P_+(\mathbb{M}_i \times \mathbb{M}_i)} \int_{\mathbb{M}_i \times \mathbb{M}_i} (g(x) - g(x'))^2 w(x, x') \nu_{\mathbb{X}}^{\otimes 2}(dx, dx')$$
$$\geq c(\delta + \lambda) \frac{1}{P_{\mathbb{X}}(\mathbb{M}_i)^2} \int_{\mathbb{M}_i \times \mathbb{M}_i} (g(x) - g(x'))^2 w(x) w(x') \nu_{\mathbb{X}}^{\otimes 2}(dx, dx').$$

The above inequality means that,

$$\int_{\mathbb{X}\times\mathbb{X}} (g(x)-g(x'))^2 P_+(dx,dx'|\mathbb{M}_i\times\mathbb{M}_i)$$
$$\geq c(\delta+\lambda)\int_{\mathbb{X}}\int_{\mathbb{X}} (g(x)-g(x'))^2 P_{\mathbb{X}}(dx|\mathbb{M}_i)P_{\mathbb{X}}(dx'|\mathbb{M}_i).$$

Thus, we obtain $Q_{\mathbb{M}_i}(g)\geq c(\delta+\lambda)$, where $g(x)=w^\top f(x)$. □

The above proposition indicates that the inner-connectivity $Q_{\mathbb{M}_i}(g)$ for any $i\in[K]$ and $g(x)=w^\top f(x)$ with $f\in\widetilde{\mathcal{F}}$ is lower bounded by $c(\delta+\lambda)$ under the assumptions. Therefore, Assumption 2 of our work is a sufficient condition of Assumption 3 in HaoChen and Ma [17] if $c\cdot P_+(\mathbb{M}_i\times\mathbb{M}_i)\leq P_{\mathbb{X}}(\mathbb{M}_i)^2$ (for every $i\in[K]$) and $\delta+\lambda\geq 0$ hold.

Remark 4. We can construct a positive value c in the above statement explicitly. In this remark, we show a simple way to do so. Let X,Y be random variables on a probability space (Ω,P_Ω) with the joint probability distribution P_+ and the marginal distribution $P_{\mathbb{X}}$. Denote $p_1=P_\Omega(X\in\mathbb{M}_i)$ and $p=P_\Omega(X\in\mathbb{M}_i,Y\in\mathbb{M}_i)$. Here, let V be the covariance matrix of the random variables $\mathbb{1}_{\{X\in\mathbb{M}_i\}}$ and $\mathbb{1}_{\{Y\in\mathbb{M}_i\}}$. The positive semi-definiteness of V implies the inequality $(p_1-p_1^2)^2-(p-p_1^2)^2\geq 0$. This inequality is valid when $2p_1^2-p_1\leq p\leq p_1$. Combining this fact with the property $p\geq 0$, we obtain

$$\max\{2p_1^2-p_1,0\}\leq p\leq p_1,$$

which implies,

$$P_{\mathbb{X}}(\mathbb{M}_i)\cdot\max\{2P_{\mathbb{X}}(\mathbb{M}_i)-1,0\}\cdot P_+(\mathbb{M}_i\times\mathbb{M}_i)\leq P_{\mathbb{X}}(\mathbb{M}_i)^2.$$

Thus, if $P_{\mathbb{X}}(\mathbb{M}_i)>1/2$ holds for every $i\in[K]$, then we can take

$$c=\min_{i\in[K]}\{P_{\mathbb{X}}(\mathbb{M}_i)(2P_{\mathbb{X}}(\mathbb{M}_i)-1)\}.$$

F.5 More Discussion about Generalization Bounds in Section 5.3

In this section, we discuss the differences between our generalization error bound and the results presented by other works on contrastive learning. We summarize the differences below.

- Arora et al. [3]; Ash et al. [4]; Lei et al. [25]; Zou and Liu [55] consider the case that for a pair (x,x^+), M-tuple samples $(x_1^-,\cdots,x_M^-)$ independent from other random variables are available. Thus, our problem setup is different from them. Especially, in our analysis, it is also necessary to tackle the cases in which X_i,X_i', $i\in[n]$, are not necessarily independent and the standard techniques (e.g., see Mohri et al. [29]) cannot be applied. We instead utilized the results of McDiarmid's inequality for dependent random variables shown by Zhang et al. [53].

- The empirical loss considered in Zhang et al. [52] is defined in a different way from our empirical kernel contrastive loss. Also, our proof technique is different from Zhang et al. [52].
- HaoChen et al. [18]; Nozawa et al. [31] consider the case in which the augmented samples are not necessarily independent. Nozawa et al. [31] utilize the theory on PAC-Bayes bounds [16], and HaoChen et al. [18] provide the high probability bound. Our analysis is different from Nozawa et al. [31] since our analysis is based on several concentration inequalities. HaoChen et al. [18] consider the empirical spectral contrastive loss that is defined by raw samples and expressed in the expectation w.r.t. the augmented samples that are drawn according to the conditional distribution given the raw samples (see Section 4.1 in HaoChen et al. [18]). On the other hand, we derive a generalization error bound for the empirical kernel contrastive loss defined by only augmented samples.
- Wang et al. [49] establish the generalization error bound for the spectral contrastive loss [18], where their analysis improves the convergence rate of HaoChen et al. [18]. In their analysis, they decompose the second term of the spectral contrastive loss in a different way from us (for the detail of their decomposition, see the proof of Proposition D.1 of Wang et al. [49]). Also, they utilize the concentration inequality shown by Clémençon et al. [10] (see equation (51) in Wang et al. [49]), while we use the results proved by Zhang et al. [53]. Thus, the techniques we use in the proof of Theorem 6 are different from those of Wang et al. [49].

F.6 Detailed Comparison to Robinson et al. [34]

Robinson et al. [34] tackle the hard negative sampling problem in contrastive learning from both the theoretical and empirical perspectives. They also establish generalization bounds for their hard negative objectives by introducing the 1-NN classifier (for their definition of the 1-NN classifier, see the statement of Theorem 5 in Robinson et al. [34]). We give a detailed comparison between our results and the theoretical analysis by Robinson et al. [34]. The main differences are listed below:

- The problem setup of the theoretical results by Robinson et al. [34] is based on that of Arora et al. [3], i.e., they rely on the conditional independence assumption. On the other hand, we does not rely on it, where we utilize the similarity function $\mathrm{sim}(\cdot, \cdot; \lambda)$ instead.
- In the proof of Theorem 5 of Robinson et al. [34] (see also Theorem 8 and the proof of their work), the supervised loss is upper-bounded by the term $\mathbb{E}_c \mathbb{E}_{x, x^+ \sim_{iid} p(\cdot|c)} \|f(x) - f(x^+)\|^2$. In summary, the differences between Theorem 5 of Robinson et al. [34] and Theorem 2 of our work are: (i) In the numerator of the upper bound in the proof of Theorem 8 of Robinson et al. [34], the term mentioned above appears. On the other hand, in Theorem 2 of our work, the quantity $\mathfrak{a}(f)$ appears. (ii) Our upper bound includes the quantity $\Delta_{\min}(f)$. (iii) We also note that the proof techniques used in Theorem 2 in our work are different from Robinson et al. [34].

– Note that the label employed in the analysis by Robinson et al. [34] is a random variable, while our analysis employ the deterministic labeling function.

F.7 Detailed Comparison to Huang et al. [21]; Zhao et al. [54]

Huang et al. [21] present the generalization bounds that utilizes the 1-NN classifier (for the definition of the 1-NN classifier introduced in Huang et al. [21], see Section 2 in their paper). Besides, Zhao et al. [54] extend the results of Huang et al. [21]. Thus, it is worth discussing the differences between the results by Huang et al. [21]; Zhao et al. [54] and our Theorem 2. We summarize the differences below:

– Huang et al. [21] show that if the centers of clusters in the feature space are sufficiently apart from each other (note that they call it *divergence*), then their supervised error function is upper bounded by the *alignment* term up to several constants and parameters. Hence, their results do not show that the *divergence* relates directly to the supervised error, i.e., the *divergence* term does not appear in their upper bounds of the supervised error. On the other hand, we show that the quantities related to the *divergence* in the RKHS can also contribute to upper-bounding the supervised error (see Theorem 2).
– In Huang et al. [21]; Zhao et al. [54], it is little investigated to what range of encoder models their results can apply. On the other hand, our Theorem 2 requires only the meaningfulness (Definition 2) of encoders belonging to $\mathcal{F}$. Especially, suppose k is the linear kernel, then Theorem 2 in our study refines Theorem 1 of Huang et al. [21] in this sense.
– Huang et al. [21]; Zhao et al. [54] utilize the notion termed (σ, δ)-*augmentation*, while our analysis utilizes Assumption 2 based on the definition of the similarity function $\mathrm{sim}(\cdot, \cdot; \lambda)$.
– Huang et al. [21]; Zhao et al. [54] often use the assumption that the encoder f is a Lipschitz function. Meanwhile, our main result does not require that f should be a Lipschitz function.
– Zhao et al. [54] consider the squared loss for the downstream classification task (see Theorem 3.2 in their paper). On the other hand, we consider the classification error.

G Connections between KCL and Normalized Cut

In this section, we present supplementary information of Section 4.1. Throughout this section, we assume that $\inf_{x \in \mathbb{X}} w(x) > 0$ and $\sup_{\overline{x} \in \overline{\mathbb{X}}} \sup_{x \in \mathbb{X}} a(x|\overline{x}) < \infty$ hold. Note that the assumption $\sup_{\overline{x} \in \overline{\mathbb{X}}} \sup_{x \in \mathbb{X}} a(x|\overline{x}) < \infty$ implies $\sup_{x \in \mathbb{X}} w(x) < \infty$.

G.1 The Problem Setup of Normalized Cut

In this section, we first explain the population-level normalized cut problem based on Shi and Malik [36]; von Luxburg [46]; Terada and Yamamoto [40]. Suppose that there are total K clusters in $\mathbb{X}$. Following Terada and Yamamoto [40],

the optimization problem of the population-level normalized cut is given as:

$$\min_{\mathbb{V}_1,\cdots,\mathbb{V}_K} \sum_{i=1}^{K} \frac{W(\mathbb{V}_i, \mathbb{V}_i^c)}{\mathrm{vol}(\mathbb{V}_i)} \tag{29}$$

where the minimum in the above problem is taken over all the possible combinations of K disjoint non-empty measurable subsets $\mathbb{V}_1, \cdots, \mathbb{V}_K$ satisfying $\bigcup_{i=1}^{K} \mathbb{V}_i = \mathbb{X}$, and W and $\mathrm{vol}(\cdot)$ are defined as,

$$W(\mathbb{V}_i, \mathbb{V}_i^c) = \int_{(x,x') \in \mathbb{V}_i \times \mathbb{V}_i^c} \mathrm{sim}(x, x'; \lambda) w(x) w(x') d\nu_{\mathbb{X}}^{\otimes 2}(x, x'), \tag{30}$$

$$\mathrm{vol}(\mathbb{V}_i) = \int_{\mathbb{V}_i} w(x) d\nu_{\mathbb{X}}(x), \tag{31}$$

where $\nu_{\mathbb{X}}^{\otimes 2} := \nu_{\mathbb{X}} \otimes \nu_{\mathbb{X}}$ is the product measure. Here also note that $\mathrm{vol}(\cdot)$ is the volume of a set $\mathbb{V}_i$. Terada and Yamamoto [40] consider the case that a reproducing kernel is used as similarity measurement: see Theorem 7 in Terada and Yamamoto [40]. Note that some existing work deals with the measurable partition problems such as the ratio cut and Cheeger cut [44].

Denote by $L^2(\mathbb{X}, P_{\mathbb{X}})$, the Hilbert space over the field $\mathbb{R}$ consisting of real-valued and squared-integrable function defined on $\mathbb{X}$ for $P_{\mathbb{X}}$-a.e., with its inner product $\langle f, g \rangle_{L^2(\mathbb{X}, P_{\mathbb{X}})} = \int f(x) g(x) dP_{\mathbb{X}}(x)$. Let $U : \mathbb{R}^K \to L^2(\mathbb{X}, P_{\mathbb{X}})$ be a linear operator defined as,

$$(Uz)(\cdot) = \sum_{i=1}^{K} \frac{\mathbb{1}_{\mathbb{V}_i}(\cdot)}{\sqrt{\mathrm{vol}(\mathbb{V}_i)}} z_i, \quad z = (z_1, \cdots, z_K)^\top, \tag{32}$$

where $\mathbb{1}_{\mathbb{V}_i}(x) = 1$ if $x \in \mathbb{V}_i$ and 0 if $x \notin \mathbb{V}_i$, and $z_i := \langle z, e_i \rangle_{\mathbb{R}^K}$ for each $i \in [K]$ with an orthonormal basis $\{e_i\}_{i=1}^K$ of $\mathbb{R}^K$. Note that under the setting that $|\mathbb{X}| < \infty$, the linear operator U is equal to $(\mathbb{1}_{\mathbb{V}_j}(x_i)/\sqrt{\mathrm{vol}(\mathbb{V}_j)})_{ij}$. Therefore, the definition of U matches that of the classical theory of normalized cut [36]. Moreover, every augmented data $x \in \mathbb{X}$ belongs to one of the subsets $\mathbb{V}_1, \cdots, \mathbb{V}_K$. Since $\mathbb{V}_1, \cdots, \mathbb{V}_k$ are disjoint, linear operator U is bounded, and the adjoint operator $U^\dagger$ exists uniquely. Here, von Luxburg [46] explain that the objective function of the normalized cut problem can be rewritten as a combinatorial optimization problem. Applying the arguments presented by von Luxburg [46] to our setup, we have

$$\sum_{i=1}^{K} \frac{W(\mathbb{V}_i, \mathbb{V}_i^c)}{\mathrm{vol}(\mathbb{V}_i)} = -\mathrm{Tr}(U^\dagger A U) + (1 - \lambda) K, \tag{33}$$

where $A : L^2(\mathbb{X}, P_{\mathbb{X}}) \to L^2(\mathbb{X}, P_{\mathbb{X}})$ is a Hilbert-Schmidt integral operator defined as

$$A\psi(\cdot) = \int \mathrm{sim}(\cdot, x; \lambda) \psi(x) w(x) d\nu_{\mathbb{X}}(x) \quad \psi \in L^2(\mathbb{X}, P_{\mathbb{X}}), \tag{34}$$

and $-\mathrm{Tr}(U^\dagger AU) = -\sum_{i=1}^K \langle U^\dagger AUe_i, e_i\rangle_{\mathbb{R}^K}$. The proof of (33) closely follows that of von Luxburg [46]; in Appendix G.3, we present the proof of an extended version. Here the following proposition shows the well-definedeness of A.

Proposition 8. *Suppose the setting described in Section 2.1 holds,* $\inf_{x\in\mathbb{X}} w(x) > 0$, *and* $\sup_{\overline{x}\in\overline{\mathbb{X}}} \sup_{x\in\mathbb{X}} a(x|\overline{x}) < \infty$ *holds. Then, the integral operator A is well-defined.*

Proof. We can evaluate

$$\begin{aligned}
&\left|\int \mathrm{sim}^2(x,x;\lambda)w(x)d\nu_{\mathbb{X}}(x)\right| \\
&= \left|\int \left(\frac{w(x,x)}{w(x)w(x)} - \lambda\right)^2 w(x)d\nu_{\mathbb{X}}(x)\right| \\
&\leq \int \left(\left(\frac{w(x,x)}{w(x)w(x)}\right)^2 + 2\lambda\frac{w(x,x)}{w(x)w(x)} + \lambda^2\right) w(x)d\nu_{\mathbb{X}}(x) \\
&< +\infty,
\end{aligned}$$

where we use the assumptions that $\inf_{x\in\mathbb{X}} w(x) > 0$, $\sup_{\overline{x}\in\overline{\mathbb{X}}} \sup_{x\in\mathbb{X}} a(x|\overline{x}) < \infty$. □

Suppose the dimension of the RKHS $\mathcal{H}_k$ associated with the kernel function k is greater than or equal to K. In the following section, it is convenient to redefine (29) as,

$$\min_{\mathbb{V}_1,\cdots,\mathbb{V}_K} \sum_{i=1}^{\infty} \frac{W(\mathbb{V}_i, \mathbb{V}_i^c)}{\mathrm{vol}(\mathbb{V}_i)},$$

where we define $\mathbb{V}_j = \emptyset$ for every $j > K$. Also, let us redefine (32) as the linear operator $U : \mathcal{H}_k \to L^2(\mathbb{X}, P_{\mathbb{X}})$,

$$(U\psi)(\cdot) = \sum_{i=1}^{\infty} \frac{\mathbb{1}_{\mathbb{V}_i}(\cdot)}{\sqrt{\mathrm{vol}(\mathbb{V}_i)}} \langle \psi, e_i\rangle_{\mathcal{H}_k},$$

where $\{e_j\}_{j=1}^{\infty}$ is an orthonormal basis of $\mathcal{H}_k$ (if $\mathcal{H}_k$ is finite dimensional, then we understand that $\{e_j\}_{j=1}^{\infty}$ consists of finitely many non-zero elements), and we define $\mathbb{1}_{\mathbb{V}_i}(\cdot)/\sqrt{\mathrm{vol}(\mathbb{V}_i)} = 0$ for every $i > K$ as notations. Then, we have the following identity that is analogous of (33):

$$\sum_{i=1}^{\infty} \frac{W(\mathbb{V}_i, \mathbb{V}_i^c)}{\mathrm{vol}(\mathbb{V}_i)} = -\mathrm{Tr}(U^\dagger AU) + (1-\lambda)K. \tag{35}$$

For the sake of completeness, we provide the proof of the identity (35) in Appendix G.3.

G.2 Connecting KCL and Normalized Cut via RKHS

Let $k : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ be a continuous, symmetric, and positive-definite kernel function whose RKHS $\mathcal{H}_k$ is K-dimensional Hilbert space (K is either finite or ∞); For the theory of reproducing kernels, see e.g., [2,5,38]. Many kernel functions satisfy these conditions, e.g. the Gaussian kernel, the polynomial kernel, and the linear kernel. Since $\mathbb{S}^{d-1}$ is separable, the RKHS $\mathcal{H}_k$ has an orthonormal basis that is at most countable (e.g., see Berlinet and Thomas-Agnan [5]). Let $\{e_j\}_{j=1}^{\infty}$ be a countable orthonormal basis of $\mathcal{H}_k$, where $\{e_j\}_{j=1}^{\infty}$ includes only finitely many non-zero elements if $\mathcal{H}_k$ is finite-dimensional. Note that our construction is valid regardless of the choice of $\{e_j\}_{j=1}^{\infty}$. Recall the problem setup presented in Section 2.1. Then, the linear operator $H : \mathcal{H}_k \to L^2(\mathbb{X}, P_{\mathbb{X}})$ is defined as

$$(H\varphi)(\cdot) = \langle h(f(\cdot)), \varphi \rangle_{\mathcal{H}_k}, \tag{36}$$

for $\varphi \in \mathcal{H}_k$. Let $\|\cdot\|_{\mathcal{H}_k}$ be the norm of the RKHS $\mathcal{H}_k$. Then the following holds for the linear operator H defined in (36):

Proposition 9. *The linear operator H is well-defined, i.e., $H\varphi \in L^2(\mathbb{X}, P_{\mathbb{X}})$ for every $\varphi \in \mathcal{H}_k$. Also, H is continuous, i.e., for a sequence φ_j converging strongly to φ in $\mathcal{H}_k$, we have that $H\varphi_j$ is convergent to $H\varphi$.*

Proof. For any $\varphi \in \mathcal{H}_k$, we have

$$\begin{aligned}\int |(H\varphi)(x)|^2 w(x) d\nu_{\mathbb{X}}(dx) &= \int |\langle h(f(x)), \varphi \rangle_{\mathcal{H}_k}|^2 w(x) d\nu_{\mathbb{X}}(x) \\ &\leq \int \|h(f(x))\|_{\mathcal{H}_k}^2 \|\varphi\|_{\mathcal{H}_k}^2 w(x) d\nu_{\mathbb{X}}(x) \\ &\leq \|\varphi\|_{\mathcal{H}_k}^2 \mathbb{E}_x \left[k(f(x), f(x))\right] < +\infty.\end{aligned}$$

Here in the second inequality we use the Cauchy-Schwarz inequality, and in the last equality we use the fact that $\mathbb{S}^{d-1}$ is compact and k is continuous. Furthermore, if $\varphi_j \to \varphi$ in the sense of strongly convergence in $\mathcal{H}_k$, then we have

$$\begin{aligned}&\|\langle h(f(\cdot)), \varphi_j \rangle_{\mathcal{H}_k} - \langle h(f(\cdot)), \varphi \rangle_{\mathcal{H}_k}\|_{L^2(\mathbb{X}, P_{\mathbb{X}})}^2 \\ &= \int |\langle h(f(x)), \varphi_j - \varphi \rangle_{\mathcal{H}_k}|^2 w(x) d\nu_{\mathbb{X}}(x) \\ &\leq \|\varphi_j - \varphi\|_{\mathcal{H}_k}^2 \mathbb{E}_x \left[k(f(x), f(x))\right] \\ &\longrightarrow 0 \quad (j \to \infty).\end{aligned}$$

Thus $H\varphi_j$ converges to $H\varphi$, and we end the proof. □

Proposition 9 implies that H is bounded. Therefore, the adjoint operator $H^\dagger : L^2(\mathbb{X}, P_{\mathbb{X}}) \to \mathcal{H}_k$ exists uniquely.

Now let us recall the definition of the similarity function $\mathrm{sim}(\cdot, \cdot; \lambda)$ with the fixed λ in (2), and we consider to relax the combinatorial problem (29) using the linear operator H defined in (36) as follows: we replace the linear operator U in (33) with H, which results in the objective function $-\mathrm{Tr}(H^\dagger A H)$. Then, the following proposition holds.

Proposition 10. *We have*

$$-\mathrm{Tr}(H^\dagger AH) = -\mathbb{E}_{x,x^+}\left[k(f(x), f(x^+))\right] + \lambda \mathbb{E}_{x,x^-}\left[k(f(x), f(x^-))\right].$$

Proof. From the definition of $\mathrm{sim}(x, x'; \lambda)$,

$$\begin{aligned}(A\psi)(x) &= \int \mathrm{sim}(x, x'; \lambda)\psi(x')w(x')d\nu_{\mathbb{X}}(x') \\ &= \int \left(\frac{w(x, x')}{w(x)w(x')} - \lambda\right)\psi(x')w(x')d\nu_{\mathbb{X}}(x') \\ &= \underbrace{\int \frac{w(x, x')}{w(x)w(x')}\psi(x')w(x')d\nu_{\mathbb{X}}(x')}_{:=(A_{\mathrm{pos}}\psi)(x)} - \lambda \underbrace{\int \psi(x')w(x')d\nu_{\mathbb{X}}(x')}_{:=(A_{\mathrm{neg}}\psi)(x)}.\end{aligned}$$

Firstly, let us proof the identity

$$\mathrm{Tr}(H^\dagger A_{\mathrm{pos}} H) = \mathbb{E}_{x,x^+}\left[k(f(x), f(x^+))\right].$$

The proof is described as follows: From the definition of H,

$$(He_i)(x) = \langle h(f(x)), e_i\rangle_{\mathcal{H}_k} \quad x \in \mathbb{X}.$$

Then we have

$$\begin{aligned}(A_{\mathrm{pos}}He_i)(x) &= \int \frac{w(x, x')}{w(x)w(x')}w(x')\langle h(f(x')), e_i\rangle_{\mathcal{H}_k}d\nu_{\mathbb{X}}(x') \\ &= \int \frac{w(x, x')}{w(x)}\langle h(f(x')), e_i\rangle_{\mathcal{H}_k}d\nu_{\mathbb{X}}(x').\end{aligned}$$

Here, the adjoint operator $H^\dagger$ satisfies the following identity; For $\psi \in L^2(\mathbb{X}, P_{\mathbb{X}})$,

$$\langle He_i, \psi\rangle_{L^2(\mathbb{X},P_{\mathbb{X}})} = \langle e_i, H^\dagger\psi\rangle_{\mathcal{H}_k}.$$

Utilizing this relation yields the following representation:

$$\begin{aligned}&H^\dagger A_{\mathrm{pos}}He_j \\ &= \sum_{i=1}^{\infty}\langle H^\dagger A_{\mathrm{pos}}He_j, e_i\rangle_{\mathcal{H}_k}e_i \\ &= \sum_{i=1}^{\infty}\langle A_{\mathrm{pos}}He_j, He_i\rangle_{L^2(\mathbb{X},P_{\mathbb{X}})}e_i \\ &= \sum_{i=1}^{\infty}\left(\int w(x)\langle h(f(x)), e_i\rangle_{\mathcal{H}_k}\int \frac{w(x, x')}{w(x)}\langle h(f(x')), e_j\rangle_{\mathcal{H}_k}d\nu_{\mathbb{X}}(x')d\nu_{\mathbb{X}}(x)\right)e_i \\ &= \sum_{i=1}^{\infty}\left(\int\int w(x, x')\langle h(f(x)), e_i\rangle_{\mathcal{H}_k}\langle h(f(x')), e_j\rangle_{\mathcal{H}_k}d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x')\right)e_i.\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Tr}(H^\dagger A_{\mathrm{pos}} H) &= \sum_{j=1}^{\infty} \langle H^\dagger A_{\mathrm{pos}} H e_j, e_j \rangle_{\mathcal{H}_k} \\
&= \sum_{j=1}^{\infty} \int\int w(x,x') \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
&= \int\int w(x,x') \sum_{j=1}^{\infty} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
&= \int\int w(x,x') \langle h(f(x)), h(f(x')) \rangle_{\mathcal{H}_k} d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
&= \mathbb{E}_{x,x^+} \left[k(f(x), f(x^+)) \right].
\end{aligned}$$

Note that the third equality above is due to the Dominated Convergence Theorem. Indeed, the sum $\sum_{j=1}^{n} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k}$ converges pointwisely to $\langle h(f(x)), h(f(x)) \rangle_{\mathcal{H}_k}$ on $\mathbb{X} \times \mathbb{X}$, and

$$\begin{aligned}
&\left| \sum_{j=1}^{n} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} \right| \\
&\leq \sum_{j=1}^{n} \left| \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} \right| \\
&\leq \left(\sum_{j=1}^{n} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k}^2 \right)^{1/2} \left(\sum_{j=1}^{n} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k}^2 \right)^{1/2} \\
&\leq \left(\sum_{j=1}^{\infty} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k}^2 \right)^{1/2} \left(\sum_{j=1}^{\infty} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k}^2 \right)^{1/2} \\
&= \| h(f(x)) \|_{\mathcal{H}_k} \| h(f(x')) \|_{\mathcal{H}_k} \\
&\leq \sup_{x \in \mathbb{X}} k(f(x), f(x)) < +\infty.
\end{aligned}$$

On the other hand, it is obvious that,

$$\begin{aligned}
&\langle H^\dagger A_{\mathrm{neg}} H e_j, e_j \rangle_{\mathcal{H}_k} \\
&= \langle A_{\mathrm{neg}} H e_j, H e_j \rangle_{L^2(\mathbb{X}, P_{\mathbb{X}})} \\
&= \int w(x) \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} \int w(x') \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} d\nu_{\mathbb{X}}(x') d\nu_{\mathbb{X}}(x) \\
&= \int\int \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x').
\end{aligned}$$

Hence we obtain,

$$\begin{aligned}\mathrm{Tr}(H^\dagger A_{\mathrm{neg}} H) &= \sum_{j=1}^{\infty} \langle H^\dagger A_{\mathrm{neg}} H e_j, e_j \rangle_{\mathcal{H}_k} \\ &= \int\int \langle h(f(x)), h(f(x')) \rangle_{\mathcal{H}_k} w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\ &= \mathbb{E}_{x,x^-} \left[k(f(x), f(x^-)) \right].\end{aligned}$$

Hence, we obtain the desired results and end the proof. □

Comparison with Related Work from the Graph Cut Viewpoint HaoChen et al. [18] has already investigated links between the population-level spectral clustering and contrastive learning. However, our integral kernel (2) introduced in Section 4.1 is slightly different from that of HaoChen et al. [18], since 1) we divide $w(x, x')$ by $w(x)w(x')$ in the first term rather than by $\sqrt{w(x)w(x')}$ (see Appendix F in HaoChen et al. [18]), 2) we also incorporate the hyperparemeter λ.

Note that Tian [42] introduces a unified framework termed α-CL, which connects various contrastive losses from the coordinate-wise optimization perspective. In Tian [42], the contrastive covariance plays a central role in the theoretical analysis. On the other hand, we use the similarity function defined in Section 4, and thus the approach of our analysis is different from the contrastive covariance of Tian [42].

G.3 Proof of (35)

Proof (Proof of (35)). For the proof of (35), we closely follow the approaches presented in Section 5 of von Luxburg [46]. Since we consider the population-level normalized cut, we present the proof of (35) for the sake of completeness.

Let us define the identity operator $D : L^2(\mathbb{X}, P_{\mathbb{X}}) \to L^2(\mathbb{X}, P_{\mathbb{X}})$ as $D\psi = \psi$ for $\psi \in L^2(\mathbb{X}, P_{\mathbb{X}})$. From the definitions of D and U, we have

$$DUe_i = \begin{cases} \frac{\mathbb{1}_{\mathbb{V}_i}(\cdot)}{\sqrt{\mathrm{vol}(\mathbb{V}_i)}} & (i \leq K), \\ 0 & (i > K). \end{cases}$$

Hence, we have the following for $i \leq K$:

$$\langle U^\dagger DUe_i, e_i \rangle_{\mathcal{H}_k} = \int \frac{w(x)\mathbb{1}_{\mathbb{V}_i}(x)^2}{\mathrm{vol}(\mathbb{V}_i)} d\nu_{\mathbb{X}}(x) = 1.$$

On the other hand, for $i \leq K$ we have

$$\begin{aligned}&\langle U^\dagger AUe_i, e_i \rangle_{\mathcal{H}_k} \\ &= \int\int (w(x, x') - \lambda w(x)w(x')) \frac{\mathbb{1}_{\mathbb{V}_i}(x)}{\sqrt{\mathrm{vol}(\mathbb{V}_i)}} \frac{\mathbb{1}_{\mathbb{V}_i}(x')}{\sqrt{\mathrm{vol}(\mathbb{V}_i)}} d\nu_{\mathbb{X}}(x') d\nu_{\mathbb{X}}(x)\end{aligned}$$

Therefore, we obtain the following:

$$
\begin{aligned}
&\mathrm{Tr}(U^\dagger(D-A)U)\\
&=\sum_{i=1}^{\infty}\langle U^\dagger(D-A)Ue_i,e_i\rangle_{\mathcal{H}_k}\\
&=\sum_{i=1}^{K}\langle U^\dagger(D-A)Ue_i,e_i\rangle_{\mathcal{H}_k}\\
&=\lambda K\\
&\quad+\frac{1}{2}\sum_{i=1}^{K}\int\int(w(x,x')-\lambda w(x)w(x'))\left(\frac{\mathbb{1}_{\mathbb{V}_i}(x)}{\sqrt{\mathrm{vol}(\mathbb{V}_i)}}-\frac{\mathbb{1}_{\mathbb{V}_i}(x')}{\sqrt{\mathrm{vol}(\mathbb{V}_i)}}\right)^2 d\nu_{\mathbb{X}}^{\otimes 2}(x,x')\\
&=\lambda K+\sum_{i=1}^{K}\int_{x\in\mathbb{V}_i}\int_{x'\in\mathbb{V}_i^c}\frac{(w(x,x')-\lambda w(x)w(x'))}{\mathrm{vol}(\mathbb{V}_i)}d\nu_{\mathbb{X}}(x')d\nu_{\mathbb{X}}(x)\\
&=\lambda K+\sum_{i=1}^{K}\int_{x\in\mathbb{V}_i}\int_{x'\in\mathbb{V}_i^c}\frac{\mathrm{sim}(x,x';\lambda)}{\mathrm{vol}(\mathbb{V}_i)}w(x)w(x')d\nu_{\mathbb{X}}(x')d\nu_{\mathbb{X}}(x)\\
&=\lambda K+\sum_{i=1}^{K}\frac{W(\mathbb{V}_i,\mathbb{V}_i^c)}{\mathrm{vol}(\mathbb{V}_i)}.
\end{aligned}
$$

Hence we end the proof. □

H Experiments

In this section, we present the experimental setup and the results.

H.1 Experimental setup

We provide the setting of the experiments presented in this paper. The code used in our experiments is based on the official implementation of SimSiam[3] and written with PyTorch [33]. We follow the experimental setting of Chen and He [9]. For the sake of completeness, we provide the detail of the setup used in our experiments. During the stage of pretraining, we construct a trainable encoder model as follows: following Chen and He [9], we use a backbone architecture whose parameters are initialized, followed by the MLP that consists of linear layers, batch normalization [23], and the ReLU activation function. Note that this type of MLP is called *projection head* [6]. The output of a trainable encoder model is normalized using the Euclidean norm as several works do [6, 13]. On the other hand, during the stage of linear evaluation [6, 9], the additional MLP is removed from a trained encoder model, and then a linear classification head is added to the encoder model. The parameters of the trained encoder model are

[3] https://github.com/facebookresearch/simsiam (Last accessed: March 25, 2023)

frozen in this stage, and only the linear head is trained. In all of the experiments reported in this paper, we use ResNet-18 [20] as the backbone architecture. We use the 2-layer multi-layer perceptron for the projection head, where the first linear layer is bias-free, and the last linear layer has the bias term.

For the pretraining, we use the same data augmentation techniques as Chen et al. [8]; Chen and He [9], where two augmented images are sampled from each image as well as these works. Note that following Chen and He [9], for the CIFAR-10 experiments, we exclude the Gaussian blur augmentation. For the linear evaluation, we also follow the data augmentation techniques of Chen and He [9]. Note that in both the stage of pretraining and linear evaluation, we set drop_last to True in the training data loader.

For optimization during both the stage of pretraining and linear evaluation, following Chen and He [9], we use the SGD optimizer. Inspired by HaoChen et al. [18], we use the cosine-decay learning rate scheduler [27] with warmup [14]. Note that we use the cosine-decay learning rate scheduler in both the pretraining and linear evaluation and apply warmup in the pretraining. Following the implementation of the learning rate scheduler of the official implementation of SCL[4], we also define our learning rate scheduler by the number of iterations.

Configurations In all of the experiments reported in this paper, we use the following configurations:

Pretraining For the learning rate, we set the initial learning rate to 0.0005, the base learning rate to 0.05, and the warmup epochs to 10. Following Chen and He [9], we use the linear scaling [14] for the learning rate. For the setting of the SGD optimizer, we also follow the setting of Chen and He [9] used for their CIFAR-10 experiments (see Appendix D in their paper): the momentum is set 0.9, and the weight decay is set 0.0005. For the output dimension of encoders, we set 512.

Linear Evaluation We follow the configurations of Chen and He [9] for linear evaluation: for the SGD optimizer, the momentum is 0.9, the weight decay is 0, the batch size is fixed to 256, and the learning rate is 30.0, where the linear scaling [14] is applied to the learning rate. Note that we train the linear head for 100 epochs.

Kernel Functions In the experiments, we use the following kernel functions:

Gaussian Kernel. The Gaussian kernel k_{Gauss} is defined as,

$$k_{\mathrm{Gauss}}(z, z') = \exp\left(-\frac{\|z - z'\|_2^2}{\sigma^2}\right),$$

where $\sigma^2 > 0$ is the bandwidth parameter.

[4] https://github.com/jhaochenz/spectral_contrastive_learning/blob/ee431bdba9bb62ad00a7e55792213ee37712784c/optimizers/lr_scheduler.py (Last accessed: March 25, 2023)

Quadratic Kernel. The Quadratic kernel k_{Quad} is defined as,

$$k_{\mathrm{Quad}}(z, z') = \left(z^\top z'\right)^2.$$

Loss Functions For the implementation of the kernel contrastive loss, we implement the empirical kernel contrastive loss (4). Note that in our experiments, the KCL frameworks with the Gaussian kernel and quadratic kernel are called Gaussian KCL (GKCL), and Quadratic KCL (QKCL), respectively.

For comparison, we also perform several reproducing experiments for SimCLR [6] and SCL [18]. For the implementation of the objective function of SimCLR, we use lightly.loss.NTXentLoss[5] of Lightly [39]. For the implementation of the spectral contrastive loss of SCL, we adapt the implementation of the official SCL code[6].

Datasets In the experiments, we use the following datasets: CIFAR-10 [24], STL-10 [11], and ImageNet-100 [41]. Note that ImageNet-100 is a subset of the ImageNet-1K dataset [12], where the ImageNet-100 dataset contains images categorized in 100 classes [41]. When extracting images from the original the ImageNet-1K dataset to create the ImageNet-100 dataset, we select the 100 classes used in Tian et al. [41]. We also remark that for the experiments with the STL-10 dataset, we use the mixed dataset that consists of the unlabeled images and the labeled training images for pretraining, the labeled training images for the training of the linear head in the stage of linear evaluation, and the labeled test images for computing the accuracy in linear evaluation. Throughout the experiments, we use the following image size for each dataset: 32×32 pixels for CIFAR-10, 96×96 pixels for STL-10, and 224×224 pixels for ImageNet-100, where the image sizes of CIFAR-10 and STL-10 are the same as the sizes of the original images, respectively, and the image sizes for ImageNet-100 are inspired by those for the ImageNet-1K dataset used in Chen et al. [6]; Chen and He [9].

Detail of Architectures for the CIFAR-10 Experiments In the experiments with the CIFAR-10 dataset, following the settings of He et al. [20]; Chen and He [9]; HaoChen et al. [18], we modify the original ResNet-18 [20] as follows: in the implementation code of ResNet[7] of torchvision [43], we replace the first convolution layer with that whose kernel size is 3, stride is 1, and padding is 1, and the maxpool layer with that whose kernel size is 1 and stride is 1.

[5] Note that in our experiments, we use Lightly 1.2.25.

[6] https://github.com/jhaochenz/spectral_contrastive_learning/blob/ee431bdba9bb62ad00a7e55792213ee37712784c/models/spectral.py (Last accessed: March 25, 2023)

[7] https://github.com/pytorch/vision/blob/eac3dc7bab436725b0ba65e556d3a6ffd43c24e1/torchvision/models/resnet.py (Last accessed: March 26, 2023)

Table 1. Top-1 and Top-5 accuracy (%) in linear evaluation for each method. For the CIFAR-10 and STL-10 experiments, we perform three trials of "pretraining+linear evaluation," and the results indicate the mean±standard deviation. For the ImageNet-100 experiments, we perform one trial of "pretraining+linear evaluation." All the results reported below are obtained as follows: we trained the linear heads for 100 epochs and evaluated the final classification accuracy of the models with the corresponding validation or test dataset. The word "repro." is the abbreviation for "reproducing," meaning that we performed several reproducing experiments to compare to the performance of KCL.

	CIFAR-10		STL-10		ImageNet-100	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
SimCLR (repro.)	90.09±0.05	99.70±0.01	87.23±0.35	99.56±0.04	77.26	94.06
SCL (repro.)	91.53±0.10	99.71±0.05	86.68±0.12	99.49±0.07	75.22	93.36
GKCL	90.87±0.08	99.63±0.00	86.69±0.09	99.38±0.01	76.40	93.20
QKCL	90.62±0.10	99.59±0.05	87.07±0.20	99.37±0.03	77.12	93.96

Supplementary Information of the Implementation We use the following packages for the experiments: PyTorch [33], torchvision [43], NumPy [19], Lightly [39], Matplotlib [22], and seaborn [50].

H.2 Results of Linear Evaluation

We perform pretraining and linear evaluation with the CIFAR-10, STL-10, and ImageNet-100 datasets. In the stage of pretraining, we train the encoder models for 800 epochs. For the experiments with the CIFAR-10 and STL-10 datasets, the batch sizes are set 256. Besides, for the experiments with the ImageNet-100 dataset, we set 512 for the batch sizes. We select the following hyperparameters of the KCL frameworks for all the experiments reported in this subsection: $\sigma^2 = 1$ and $\lambda = 8$ for GKCL, and $\lambda = 4$ for QKCL. For SCL, inspired by HaoChen et al. [18], we select 3 for the radius parameter. For SimCLR, inspired by Chen et al. [6], we select 0.1 for the temperature parameter. These hyperparameters are also used for all the experiments in this subsection.

The results are shown in Table 1. In Table 1, the experiments with the CIFAR-10 dataset are performed using one Quadro P6000 GPU. Besides, the experiments with STL-10 and ImageNet-100 are performed using one Tesla V100S GPU.

H.3 More Experiments

Ablation Study on the Weight Parameters and Batch Sizes We investigate how the selection of λ and batch sizes affect the quality of learned representations. In the experiments, we use the CIFAR-10 dataset. We select

$\{1, 2, 4, 8, 16, 32\}$ for λ, and $\{64, 128, 256, 512, 1024\}$ for the batch sizes. In each run, we pretrain an encoder model for 200 epochs. We use both GKCL and QKCL and evaluate those results.

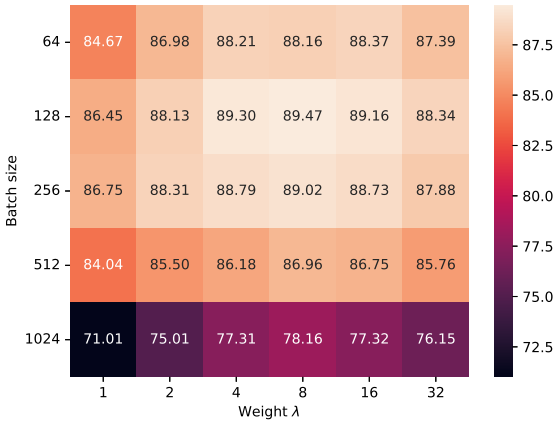

Fig. 2. The results of ablation study for GKCL. The number in each cell indicates the Top-1 accuracy (%).

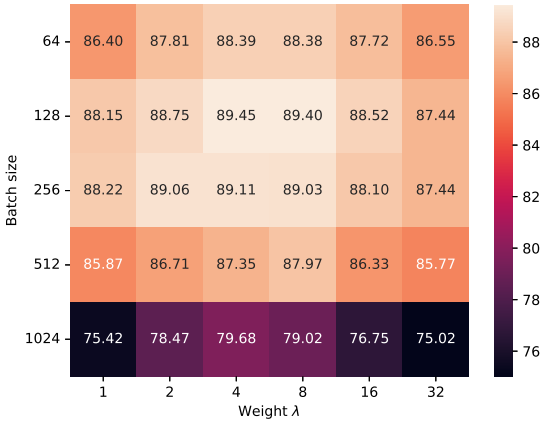

Fig. 3. The results of ablation study for QKCL. The number in each cell indicates the Top-1 accuracy (%).

The Top-1 accuracy computed at the end of linear evaluation for GKCL and QKCL are shown in Fig. 2 and 3, respectively. Note that the experiments reported in Fig. 2 and 3 are performed by using one Tesla V100S GPU. The results of the experiments indicate that 1) the selection of the value for λ affects the quality of the representations learned by KCL, 2) the small batch sizes (e.g., 128 and 256) are more efficient when pretraining encoders, while the large batch sizes (e.g., 1024) degrade the performance. Note that Chen et al. [7] showed similar findings to the first point for the generalized NT-Xent loss. Besides, Chen and He [9] point out the efficiency of SimSiam with small batch sizes.

How Does λ Influence the Geometry of Representations Learned? From Theorem 1, minimization of the kernel contrastive loss makes $\lambda \cdot \mathfrak{c}(f)$ smaller, which can imply that the means of the clusters tend to be scattered as λ increases. Motivated by this result, in this subsubsection, we simulate how the means of the feature vectors derived from augmented data whose corresponding raw data belong to the same class distribute. In the experiments, we use the STL-10 dataset. In the stage of unsupervised pretraining, we use the combination of the unlabeled images and the labeled training images in the STL-10 dataset. We use GKCL for the pretraining. The weights used in the experiments are $\{1, 2, 4, 8, 16, 32, 64\}$. We pretrain the encoder model for 400 epochs in each run. We set the batch sizes to 256. After the stage of pretraining, we compute the mean for each class and calculate the cosine similarities between those means. Since the clusters $\mathbb{M}_1, \cdots, \mathbb{M}_K$ are hard to obtain for the STL-10 dataset, we instead use the labels included in the labeled training images of the STL-10 dataset to compute the mean over the feature vectors of augmented data transformed from raw data in each class. Note that we draw an augmented image from each raw image when computing the means. The experiments in this subsubsection are performed by using one Tesla V100S GPU.

The results are summarized in Fig. 4 as a box plot. The results indicate that the variation becomes smaller as λ increases. Thus, larger λ makes the means scattered more in this experimental setting. Note that this result may imply that the clusters $\mathbb{M}_1, \cdots, \mathbb{M}_K$ and the subsets defined with labels have some relation. We leave the investigation of this question as future work.

H.4 Too Large λ Degrades the Performance in Downstream Classification Tasks

In this subsection, we report the results of the experiments with different values for λ. In the experiments, we use $\{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$ for the weight λ. For the contrastive learning framework, we use GKCL. We use two datasets, CIFAR-10 and STL-10, and pretrain the encoder during 400 epochs in each run. In the stage of pretraining, we set 128 for the batch size. Each experiment reported in this subsection is performed using one Tesla V100S GPU.

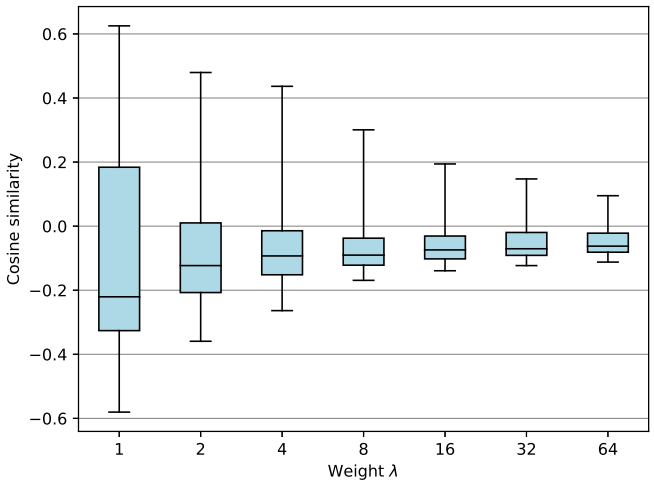

Fig. 4. The box plot of the cosine similarities between the means of the different classes for each encoder model pretrained with different λ. Note that the horizontal lines in each bar represent, in the order from bottom to top, the minimum value, the first quartile, the median, the third quartile, and the maximum value, respectively.

Table 2. Top-1 accuracy (%) in the results of linear evaluation, where the encoder is pretrained with different λ for each run.

	Top-1 Accuracy	
λ	CIFAR-10	STL-10
1	89.06	83.20
2	90.29	84.54
4	90.77	85.36
8	90.66	85.33
16	90.52	84.19
32	89.78	83.39
64	88.85	81.50
128	86.93	79.71
256	85.24	77.89
512	82.43	76.26

The results on the Top-1 accuracy at the end of the linear evaluation are presented in Table 2. The results indicate that too large λ, such as 512, degrades the performance in the downstream task.

References

1. Arendt, W., Batty, C.J., Hieber, M., Neubrander, F.: Vector-valued Laplace transforms and Cauchy problems. Birkhäuser Basel (2011). https://doi.org/10.1007/978-3-0348-0087-7 2
2. Aronszajn, N.: Theory of reproducing kernels. Transactions of the American Mathematical Society **68**(3), 337–404 (1950). https://doi.org/10.1090/s0002-9947-1950-0051437-7 33
3. Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., Saunshi, N.: A theoretical analysis of contrastive unsupervised representation learning. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 5628–5637. PMLR (2019) 21, 28, 29
4. Ash, J., Goel, S., Krishnamurthy, A., Misra, D.: Investigating the role of negatives in contrastive representation learning. In: Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 151, pp. 7187–7209. PMLR (2022) 21, 28
5. Berlinet, A., Thomas-Agnan, C.: Reproducing kernel Hilbert spaces in probability and statistics. Springer Science & Business Media (2004). https://doi.org/10.1007/978-1-4419-9096-9 2, 33
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (2020) 25, 37, 39, 40
7. Chen, T., Luo, C., Li, L.: Intriguing properties of contrastive losses. In: Advances in Neural Information Processing Systems. vol. 34, pp. 11834–11845. Curran Associates, Inc. (2021) 25, 42
8. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297v1 (2020) 38
9. Chen, X., He, K.: Exploring simple siamese representation learning. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15745–15753 (2021). https://doi.org/10.1109/CVPR46437.2021.01549 37, 38, 39, 42
10. Clémençon, S., Lugosi, G., Vayatis, N.: Ranking and Empirical Minimization of U-statistics. The Annals of Statistics **36**(2), 844 – 874 (2008). https://doi.org/10.1214/009052607000000910 20, 21, 29
11. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 15, pp. 215–223. PMLR (2011) 39
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. Ieee (2009). https://doi.org/10.1109/CVPR.2009.5206848 39
13. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9568–9577 (2021). https://doi.org/10.1109/ICCV48922.2021.00945 37
14. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677v2 (2017) 38

15. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. In: Algorithmic Learning Theory. pp. 63–77. Springer Berlin Heidelberg (2005). https://doi.org/10.1007/11564089_7 25
16. Guedj, B.: A primer on pac-bayesian learning. arXiv preprint arXiv:1901.05353v3 (2019) 29
17. HaoChen, J.Z., Ma, T.: A theoretical study of inductive biases in contrastive learning. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=AuEgNlEAmed 27, 28
18. HaoChen, J.Z., Wei, C., Gaidon, A., Ma, T.: Provable guarantees for self-supervised deep learning with spectral contrastive loss. In: Advances in Neural Information Processing Systems. vol. 34, pp. 5000–5011. Curran Associates, Inc. (2021) 26, 29, 36, 38, 39, 40
19. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**(7825), 357–362 (Sep 2020). https://doi.org/10.1038/s41586-020-2649-2 40
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90 38, 39
21. Huang, W., Yi, M., Zhao, X., Jiang, Z.: Towards the generalization of contrastive self-supervised learning. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=XDJwuEYHhme 30
22. Hunter, J.D.: Matplotlib: A 2d graphics environment. Computing in Science & Engineering **9**(3), 90–95 (2007). https://doi.org/10.1109/MCSE.2007.55 40
23. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 448–456. PMLR (2015) 37
24. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep. (2009) 39
25. Lei, Y., Yang, T., Ying, Y., Zhou, D.X.: Generalization analysis for contrastive representation learning. arXiv preprint arXiv:2302.12383v2 (2023) 28
26. Li, Y., Pogodin, R., Sutherland, D.J., Gretton, A.: Self-supervised learning with kernel dependence maximization. In: Advances in Neural Information Processing Systems. vol. 34, pp. 15543–15556. Curran Associates, Inc. (2021) 25
27. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=Skq89Scxx 38
28. McDiarmid, C.: On the method of bounded differences, p. 148–188. London Mathematical Society Lecture Note Series, Cambridge University Press (1989). https://doi.org/10.1017/CBO9781107359949.008 17
29. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of machine learning. MIT press (2018) 12, 13, 18, 28
30. Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B.: Kernel mean embedding of distributions: A review and beyond. Foundations and Trends® in Machine Learning **10**(1-2), 1–141 (2017) 1, 2
31. Nozawa, K., Germain, P., Guedj, B.: Pac-bayesian contrastive unsupervised representation learning. In: Proceedings of the 36th Conference on Uncertainty in

Artificial Intelligence (UAI). Proceedings of Machine Learning Research, vol. 124, pp. 21–30. PMLR (2020) 29
32. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748v2 (2018) 25
33. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019) 37, 40
34. Robinson, J.D., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=CR1XOQ0UTh- 29, 30
35. Shalev-Shwartz, S., Ben-David, S.: Understanding machine learning: From theory to algorithms. Cambridge University Press (2014). https://doi.org/10.1017/CBO9781107298019 13, 21
36. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(8), 888–905 (2000). https://doi.org/10.1109/34.868688 30, 31
37. Smola, A., Gretton, A., Song, L., Schölkopf, B.: A hilbert space embedding for distributions. In: Algorithmic Learning Theory. pp. 13–31. Springer Berlin Heidelberg (2007). https://doi.org/10.1007/978-3-540-75225-7_5 1, 25
38. Steinwart, I., Christmann, A.: Support vector machines. Springer Science & Business Media (2008). https://doi.org/10.1007/978-0-387-77242-4 33
39. Susmelj, I., Helle, M., Wirth, P., Prescott, J., Ebner et al., M.: Lightly (2020), https://github.com/lightly-ai/lightly 39, 40
40. Terada, Y., Yamamoto, M.: Kernel normalized cut: a theoretical revisit. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6206–6214. PMLR (2019) 30, 31
41. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Computer Vision – ECCV 2020. pp. 776–794. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-58621-8_45 39
42. Tian, Y.: Understanding deep contrastive learning via coordinate-wise optimization. In: Advances in Neural Information Processing Systems. vol. 35, pp. 19511–19522. Curran Associates, Inc. (2022) 36
43. TorchVision maintainers and contributors: Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision (2016) 39, 40
44. Trillos, N.G., Slepčev, D., von Brecht, J., Laurent, T., Bresson, X.: Consistency of cheeger and ratio graph cuts. Journal of Machine Learning Research **17**(181), 1–46 (2016), http://jmlr.org/papers/v17/14-490.html 31
45. Tu, Z., Zhang, J., Tao, D.: Theoretical analysis of adversarial learning: A minimax approach. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019) 15, 16
46. Von Luxburg, U.: A tutorial on spectral clustering. Statistics and computing **17**(4), 395–416 (2007). https://doi.org/10.1007/s11222-007-9033-z 30, 31, 32, 36
47. Wainwright, M.J.: High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press (2019). https://doi.org/10.1017/9781108627771 14, 16, 20
48. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: Proceedings of the 37th In-

ternational Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 9929–9939. PMLR (2020) 25
49. Wang, Z., Luo, Y., Li, Y., Zhu, J., Schölkopf, B.: Spectral representation learning for conditional moment models. arXiv preprint arXiv:2210.16525v2 (2022) 29
50. Waskom, M.L.: seaborn: statistical data visualization. Journal of Open Source Software **6**(60), 3021 (2021). https://doi.org/10.21105/joss.03021 40
51. Yeh, C.H., Hong, C.Y., Hsu, Y.C., Liu, T.L., Chen, Y., LeCun, Y.: Decoupled contrastive learning. In: Computer Vision – ECCV 2022. pp. 668–684. Springer Nature Switzerland (2022). https://doi.org/10.1007/978-3-031-19809-0_38 25
52. Zhang, G., Lu, Y., Sun, S., Guo, H., Yu, Y.: f-mutual information contrastive learning (2022), https://openreview.net/forum?id=3kTt_W1_tgw 29
53. Zhang, R.R., Liu, X., Wang, Y., Wang, L.: Mcdiarmid-type inequalities for graph-dependent variables and stability bounds. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019) 17, 18, 19, 28, 29
54. Zhao, X., Du, T., Wang, Y., Yao, J., Huang, W.: ArCL: Enhancing contrastive learning with augmentation-robust representations. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=n0Pb9T5kmb 30
55. Zou, X., Liu, W.: Generalization bounds for adversarial contrastive learning. Journal of Machine Learning Research **24**(114), 1–54 (2023) 28