A Refined Generalization Analysis for Supervised Contrastive Representation Learning

Nong Minh Hieu, Antoine Ledent

School of Computing and Information Systems - Singapore Management University

Contrastive Representation Learning (CRL) Overview

- Given $f : \mathcal{X} \to \mathcal{R}$, a representation function.

Contrastive Representation Learning (CRL) Overview

- Given $f : \mathcal{X} \rightarrow \mathcal{R}$, a representation function.
- $x, x^+ \sim \mathcal{D}$ and $x_i^- \sim \mathcal{D}^-, \forall i \in [k]$. The goal is to pull $f(x), f(x^+)$ closer together while pushing away anchor-negative pairs $f(x), f(x_i^-)$.

Contrastive Representation Learning (CRL) Overview

- Given $f : \mathcal{X} \to \mathcal{R}$, a representation function.
- $x, x^+ \sim \mathcal{D}$ and $x_i^- \sim \mathcal{D}^-, \forall i \in [k]$. The goal is to pull $f(x), f(x^+)$ closer together while pushing away anchor-negative pairs $f(x), f(x_i^-)$.
- Enforce through contrastive loss:

$$\phi\left(\{\mathbf{v}_i\}_{i=1}^k\right) = \log\left(1 + \sum_{i=1}^{k} e^{\mathbf{v}_i}\right), \quad (1)$$

where $\mathbf{v}_i = f(x)^\top \left[f(x_i^-) - f(x^+)\right]$.

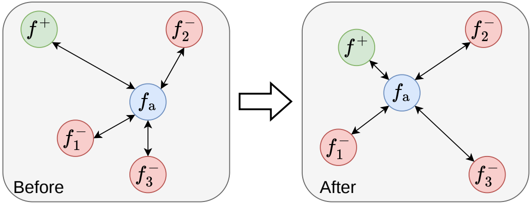


Figure: Visual illustration of N-pair contrastive loss [soh16].

Contrastive Representation Learning (CRL) Overview

- Given $f : \mathcal{X} \to \mathcal{R}$, a representation function.
- $x, x^+ \sim \mathcal{D}$ and $x_i^- \sim \mathcal{D}^-, \forall i \in [k]$. The goal is to pull $f(x), f(x^+)$ closer together while pushing away anchor-negative pairs $f(x), f(x_i^-)$.
- Enforce through contrastive loss:

$$\phi\left(\{\mathbf{v}_i\}_{i=1}^k\right) = \log\left(1 + \sum_{i=1}^k e^{\mathbf{v}_i}\right), \quad (1)$$

where $\mathbf{v}_i = f(x)^\top \left[f(x_i^-) - f(x^+)\right]$.

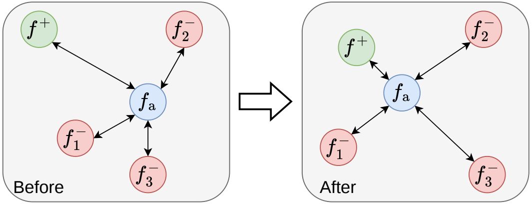
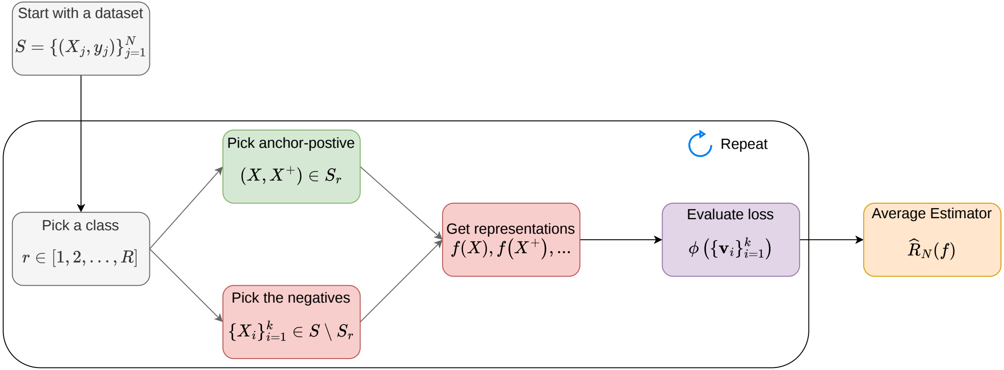


Figure: Visual illustration of N-pair contrastive loss [soh16].

Question: How well do representation models generalize?

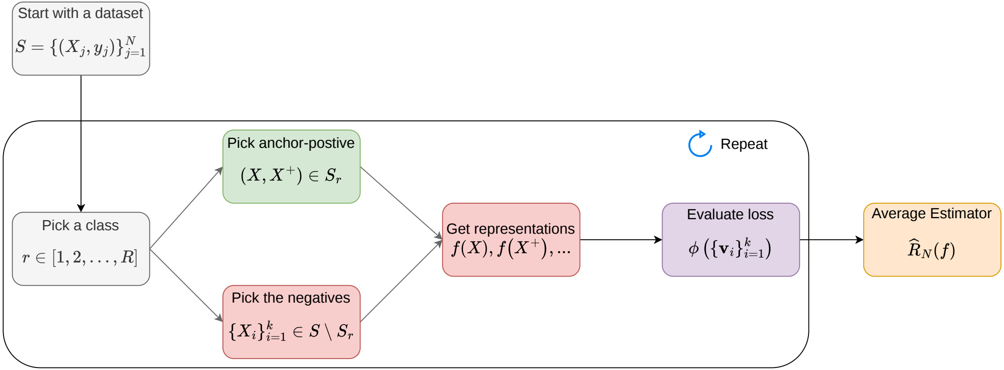
Tuples Sampling for Contrastive Learning

- In real-life *supervised* [kho20, hie25] multi-class contrastive learning (where **labels are fully supplied along with data points**):



Tuples Sampling for Contrastive Learning

- In real-life *supervised* [kho20, hie25] multi-class contrastive learning (where **labels are fully supplied along with data points**):



- **Question**: If we train a representation f on an estimator $\widehat{R}_N(f)$ **constructed on a labeled dataset**, can we provide a generalization certificate?

Problem Statement

Set-up: Given $S = \{(X_j, y_j)\}_{j=1}^{N} \sim \bar{\mathcal{D}}^{\otimes N}$ where $\bar{\mathcal{D}} := \sum_{r=1}^{R} \rho_r \mathcal{D}_r$, our tasks consist of **(i)** constructing an estimator $\widehat{R}_N(f)$ for the following population contrastive risk

$$\mathrm{L}_\phi(f) := \mathop{\mathbb{E}}_{r\sim\rho}[\mathrm{L}_\phi^r(f)] = \sum_{r=1}^{R} \rho_r \mathrm{L}_\phi^r(f), \tag{2}$$

$$\mathrm{L}_\phi^r(f) := \mathop{\mathbb{E}}_{\substack{X, X^+ \sim \mathcal{D}_r^{\otimes 2} \\ \{X_i^-\}_{i=1}^k \sim \bar{\mathcal{D}}_r^{\otimes k}}} \left[\phi\left(\left\{f(X)^\top \left[f(X^+) - f(X_i^-)\right]\right\}_{i=1}^{k}\right)\right], \tag{3}$$

Problem Statement

Set-up: Given $S = \{(X_j, y_j)\}_{j=1}^N \sim \bar{\mathcal{D}}^{\otimes N}$ where $\bar{\mathcal{D}} := \sum_{r=1}^R \rho_r \mathcal{D}_r$, our tasks consist of **(i)** constructing an estimator $\widehat{R}_N(f)$ for the following population contrastive risk

$$\mathrm{L}_\phi(f) := \mathop{\mathbb{E}}_{r\sim\rho}[\mathrm{L}_\phi^r(f)] = \sum_{r=1}^R \rho_r \mathrm{L}_\phi^r(f), \tag{2}$$

$$\mathrm{L}_\phi^r(f) := \mathop{\mathbb{E}}_{\substack{X, X^+ \sim \mathcal{D}_r^{\otimes 2} \\ \{X_i^-\}_{i=1}^k \sim \bar{\mathcal{D}}_r^{\otimes k}}} \left[\phi\left(\left\{f(X)^\top \left[f(X^+) - f(X_i^-)\right]\right\}_{i=1}^k\right)\right], \tag{3}$$

where $\bar{\mathcal{D}}_r := \frac{1}{1-\rho_r}\sum_{q\neq r} \rho_q \mathcal{D}_q$ and **(ii)** establish excess risk guarantee over a fixed hypothesis class $\mathcal{F}$, i.e., obtain a result of the form:

$$\mathbb{P}_{S\sim\bar{\mathcal{D}}^{\otimes N}}\left(\forall f \in \mathcal{F} : \mathrm{L}_\phi(f) \leqslant \widehat{R}_N(f) + r_{N,\delta}\right) \geqslant 1 - \delta, \tag{4}$$

Problem Statement

Set-up: Given $S = \{(X_j, y_j)\}_{j=1}^N \sim \bar{\mathcal{D}}^{\otimes N}$ where $\bar{\mathcal{D}} := \sum_{r=1}^R \rho_r \mathcal{D}_r$, our tasks consist of **(i)** constructing an estimator $\widehat{R}_N(f)$ for the following population contrastive risk

$$\mathrm{L}_\phi(f) := \mathop{\mathbb{E}}_{r\sim\rho}[\mathrm{L}_\phi^r(f)] = \sum_{r=1}^R \rho_r \mathrm{L}_\phi^r(f), \tag{2}$$

$$\mathrm{L}_\phi^r(f) := \mathop{\mathbb{E}}_{\substack{X,X^+\sim\mathcal{D}_r^{\otimes 2} \\ \{X_i^-\}_{i=1}^k\sim\bar{\mathcal{D}}_r^{\otimes k}}} \left[\phi\left(\left\{f(X)^\top \left[f(X^+) - f(X_i^-)\right]\right\}_{i=1}^k\right)\right], \tag{3}$$

where $\bar{\mathcal{D}}_r := \frac{1}{1-\rho_r}\sum_{q\neq r}\rho_q\mathcal{D}_q$ and **(ii)** establish excess risk guarantee over a fixed hypothesis class $\mathcal{F}$, i.e., obtain a result of the form:

$$\mathbb{P}_{S\sim\bar{\mathcal{D}}^{\otimes N}}\left(\forall f\in\mathcal{F} : \mathrm{L}_\phi(f) \leqslant \widehat{R}_N(f) + r_{N,\delta}\right) \geqslant 1-\delta, \tag{4}$$

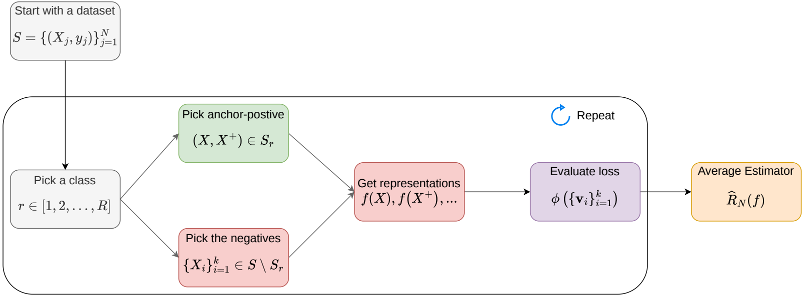
for all $\delta\in(0,1)$ and $r_{N,\delta}\to 0$ as $N\to\infty$.

Problem with RC in Supervised CRL

- A technically convenient choice is to choose $\widehat{R}_N(f)$ as average over a set of i.i.d. tuples as conducted in prior works.

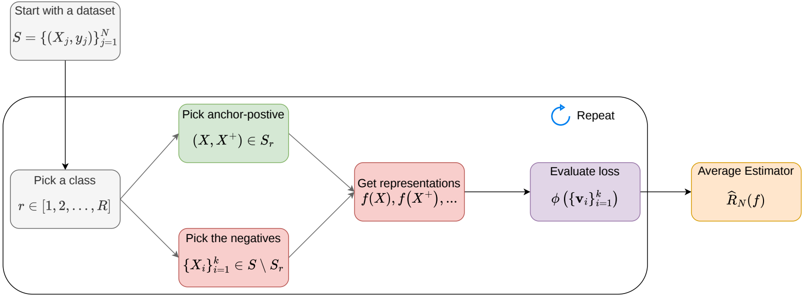
Problem with RC in Supervised CRL

- A technically convenient choice is to choose $\widehat{R}_N(f)$ as average over a set of i.i.d. tuples as conducted in prior works.
- However, recall from the data-collection procedure:



Problem with RC in Supervised CRL

- A technically convenient choice is to choose $\widehat{R}_N(f)$ as average over a set of i.i.d. tuples as conducted in prior works.
- However, recall from the data-collection procedure:



Issue: The above tuples sampling process breaks any hope for independence, i.e., we cannot straightforwardly apply Rademacher bound.

Comparison with Prior Work

	Prior Work [aro19, lei23]	This Work
Sampling	$T = \left\{(X_j, X_j^+, X_{j1:k}^-)\right\}_{j=1}^n$ drawn directly from a product distribution	$S = \{X_j\}_{j=1}^N \sim \bar{\mathcal{D}}^{\otimes N}$, i.e., tuples are constructed from S
Difficulties	No difficulty from tuples construction	Tuples construction leads to a series of bias handling issues
Techniques	Direct application of RC	U-Statistics decoupling

Class-wise U-Statistics

- The risk $\mathrm{L}_\phi(f)$ can be decomposed class-wise as $\sum_{r=1}^{R} \rho_r \mathrm{L}_\phi^r(f)$.

Class-wise U-Statistics

- The risk $\mathrm{L}_\phi(f)$ can be decomposed class-wise as $\sum_{r=1}^R \rho_r \mathrm{L}_\phi^r(f)$.
- Our first attempt - estimate risk for each class: For $r \in [R]$, let $S_r = \{Z_j\}_{j=1}^{N_r}$ and $\bar{S}_r = \{Z_j^*\}_{j=1}^{N-N_r}$, we formulate the class-wise estimator as

$$U_N(f) := \sum_{r=1}^{R} \widehat{\rho}_r U_N^r(f), \tag{5}$$

$$U_N^r(f) := \frac{1}{\binom{N_r}{2}\binom{N-N_r}{k}} \sum_{\substack{p,q \in \binom{[N_r]}{2} \\ \ell_1,\ldots,\ell_k \in \binom{[N-N_r]}{k}}} \phi\left(\left\{ f(Z_p)^\top \left[f(Z_q) - f(Z_{\ell_i}^*) \right] \right\}_{i=1}^{k}\right), \tag{6}$$

- where for each $r \in [R]$, N_r denotes the size of S_r and $\widehat{\rho}_r := \frac{N_r}{N}$, i.e., the empirical estimate of true class-r probability ρ_r.

Class-wise U-Statistics

Suppose that $\mathcal{F}$ is a class of representation functions $f : \mathcal{X} \to \mathcal{R}$ and:

$$\mathcal{F}_\phi := \left\{ (\mathbf{x}, \mathbf{x}^+, \{\mathbf{x}_i^-\}_{i=1}^k) \mapsto \phi\Big(\Big\{ f(\mathbf{x})^\top \left[f(\mathbf{x}^+) - f(\mathbf{x}_i^-) \right] \Big\}_{i=1}^k \Big) \right\}. \tag{7}$$

Theorem 1 (Hieu&Ledent'2026): Suppose that $\rho_r \leqslant \frac{1}{2}$ for all $r \in [R]$. Then, as long as $N \geqslant \widetilde{\mathcal{O}}(k^2)$, for all $f \in \mathcal{F}$ and any $\delta \in (0,1)$, with probability of at least $1 - \delta$:

$$\mathrm{L}_\phi(f) \leqslant U_N(f) + \mathcal{O}\left(\mathfrak{C}_N(\mathcal{F}_\phi) \sqrt{\frac{\max(k, R)}{N}} + \mathcal{B}\sqrt{\frac{R \ln R/\delta}{N}} \right). \tag{8}$$

$\rightarrow$ This is an improvement over [hie25] where the bound depends on $\rho_{\min}^{-\frac{1}{2}}$.

Class-wise U-Statistics

Suppose that $\mathcal{F}$ is a class of representation functions $f : \mathcal{X} \to \mathcal{R}$ and:

$$\mathcal{F}_\phi := \left\{ (\mathbf{x}, \mathbf{x}^+, \{\mathbf{x}_i^-\}_{i=1}^k) \mapsto \phi\Big(\Big\{ f(\mathbf{x})^\top \left[f(\mathbf{x}^+) - f(\mathbf{x}_i^-) \right] \Big\}_{i=1}^k \Big) \right\}. \tag{7}$$

Theorem 1 (Hieu&Ledent'2026): Suppose that $\rho_r \leqslant \frac{1}{2}$ for all $r \in [R]$. Then, as long as $N \geqslant \widetilde{\mathcal{O}}(k^2)$, for all $f \in \mathcal{F}$ and any $\delta \in (0,1)$, with probability of at least $1 - \delta$:

$$\mathrm{L}_\phi(f) \leqslant U_N(f) + \mathcal{O}\left(\mathfrak{C}_N(\mathcal{F}_\phi) \sqrt{\frac{\max(k, R)}{N}} + \mathcal{B}\sqrt{\frac{R \ln R/\delta}{N}} \right). \tag{8}$$

$\rightarrow$ This is an improvement over [hie25] where the bound depends on $\rho_{\min}^{-\frac{1}{2}}$.
$\rightarrow$ An $\mathcal{O}(R)$ scale in sample complexity is unnatural for a k-tuple learning problem.

Class-wise U-Statistics

Suppose that $\mathcal{F}$ is a class of representation functions $f : \mathcal{X} \to \mathcal{R}$ and:

$$\mathcal{F}_\phi := \left\{ (\mathbf{x}, \mathbf{x}^+, \{\mathbf{x}_i^-\}_{i=1}^k) \mapsto \phi\Big(\Big\{ f(\mathbf{x})^\top \left[f(\mathbf{x}^+) - f(\mathbf{x}_i^-) \right] \Big\}_{i=1}^k \Big) \right\}. \tag{7}$$

Theorem 1 (Hieu&Ledent'2026): Suppose that $\rho_r \leqslant \frac{1}{2}$ for all $r \in [R]$. Then, as long as $N \geqslant \widetilde{\mathcal{O}}(k^2)$, for all $f \in \mathcal{F}$ and any $\delta \in (0,1)$, with probability of at least $1 - \delta$:

$$\mathrm{L}_\phi(f) \leqslant U_N(f) + \mathcal{O}\left(\mathfrak{C}_N(\mathcal{F}_\phi) \sqrt{\frac{\max(k, R)}{N}} + \mathcal{B}\sqrt{\frac{R \ln R/\delta}{N}} \right). \tag{8}$$

→ This is an improvement over [hie25] where the bound depends on $\rho_{\min}^{-\frac{1}{2}}$.
→ An $\mathcal{O}(R)$ scale in sample complexity is unnatural for a k-tuple learning problem.
→ The $\mathcal{O}(R)$ is a fundamental gap in the design choice of U-Statistics and sub-optimal for extreme multi-class scenarios.

Debiased U-Statistics

- The population risk $\mathrm{L}_\phi(f)$ can also be decomposed as:

$$\mathrm{L}_\phi(f) = \frac{1}{1-\tau}\mathrm{L}_\phi^\omega(f) - \frac{\tau}{1-\tau}\mathrm{L}_\phi^\lambda(f), \tag{9}$$

where $\tau := 1 - \sum_{r=1}^{R} \rho_r(1-\rho_r)^k$ is called the class-collision probability and:

Debiased U-Statistics

- The population risk $\mathrm{L}_\phi(f)$ can also be decomposed as:

$$\mathrm{L}_\phi(f) = \frac{1}{1-\tau}\mathrm{L}_\phi^\omega(f) - \frac{\tau}{1-\tau}\mathrm{L}_\phi^\lambda(f), \tag{9}$$

where $\tau := 1 - \sum_{r=1}^R \rho_r(1-\rho_r)^k$ is called the class-collision probability and:

$$\mathrm{L}_\phi^\omega(f) := \mathop{\mathbb{E}}_{r\sim\rho} \mathop{\mathbb{E}}_{\substack{X,X^+\sim\mathcal{D}_r^{\otimes 2} \\ \{X_i^-\}_{i=1}^k\sim\bar{\mathcal{D}}^{\otimes k}}} \left[\phi\left(\left\{f(X)^\top\left[f(X^+)-f(X_i^-)\right]\right\}_{i=1}^k\right)\right], \tag{10}$$

$$\mathrm{L}_\phi^\lambda(f) := \mathop{\mathbb{E}}_{r\sim\rho} \mathop{\mathbb{E}}_{\substack{X,X^+,X_1^-\sim\mathcal{D}_r^{\otimes 3} \\ \{X_i^-\}_{i=2}^k\sim\bar{\mathcal{D}}^{\otimes k-1}}} \left[\phi\left(\left\{f(X)^\top\left[f(X^+)-f(X_i^-)\right]\right\}_{i=1}^k\right)\right]. \tag{11}$$

Debiased U-Statistics

- The population risk $\mathrm{L}_\phi(f)$ can also be decomposed as:

$$\mathrm{L}_\phi(f) = \frac{1}{1-\tau}\mathrm{L}_\phi^\omega(f) - \frac{\tau}{1-\tau}\mathrm{L}_\phi^\lambda(f), \tag{9}$$

where $\tau := 1 - \sum_{r=1}^R \rho_r(1-\rho_r)^k$ is called the class-collision probability and:

$$\mathrm{L}_\phi^\omega(f) := \mathop{\mathbb{E}}_{r\sim\rho} \mathop{\mathbb{E}}_{\substack{X,X^+\sim\mathcal{D}_r^{\otimes 2} \\ \{X_i^-\}_{i=1}^k\sim\bar{\mathcal{D}}^{\otimes k}}} \left[\phi\left(\left\{f(X)^\top\left[f(X^+)-f(X_i^-)\right]\right\}_{i=1}^k\right)\right], \tag{10}$$

$$\mathrm{L}_\phi^\lambda(f) := \mathop{\mathbb{E}}_{r\sim\rho} \mathop{\mathbb{E}}_{\substack{X,X^+,X_1^-\sim\mathcal{D}_r^{\otimes 3} \\ \{X_i^-\}_{i=2}^k\sim\bar{\mathcal{D}}^{\otimes k-1}}} \left[\phi\left(\left\{f(X)^\top\left[f(X^+)-f(X_i^-)\right]\right\}_{i=1}^k\right)\right]. \tag{11}$$

- Here L_ϕ^ω defines the risk where negative samples are **arbitrarily allowed to collide** with the anchor-positive pair.

Debiased U-Statistics

- The population risk $\mathrm{L}_\phi(f)$ can also be decomposed as:

$$\mathrm{L}_\phi(f) = \frac{1}{1-\tau}\mathrm{L}^\omega_\phi(f) - \frac{\tau}{1-\tau}\mathrm{L}^\lambda_\phi(f), \tag{9}$$

where $\tau := 1 - \sum_{r=1}^R \rho_r(1-\rho_r)^k$ is called the class-collision probability and:

$$\mathrm{L}^\omega_\phi(f) := \mathop{\mathbb{E}}_{r\sim\rho} \mathop{\mathbb{E}}_{\substack{X,X^+\sim\mathcal{D}_r^{\otimes 2} \\ \{X_i^-\}_{i=1}^k\sim\bar{\mathcal{D}}^{\otimes k}}} \left[\phi\left(\left\{f(X)^\top\left[f(X^+)-f(X_i^-)\right]\right\}_{i=1}^k\right)\right], \tag{10}$$

$$\mathrm{L}^\lambda_\phi(f) := \mathop{\mathbb{E}}_{r\sim\rho} \mathop{\mathbb{E}}_{\substack{X,X^+,X_1^-\sim\mathcal{D}_r^{\otimes 3} \\ \{X_i^-\}_{i=2}^k\sim\bar{\mathcal{D}}^{\otimes k-1}}} \left[\phi\left(\left\{f(X)^\top\left[f(X^+)-f(X_i^-)\right]\right\}_{i=1}^k\right)\right]. \tag{11}$$

- Here L^ω_ϕ defines the risk where negative samples are **arbitrarily allowed to collide** with the anchor-positive pair.
- L^λ_ϕ defines the risk where negative samples **must collide at least once** with the anchor-positive pair.

Debiased U-Statistics

Our second attempt - estimate the risk jointly with debiasing:

$$\widetilde{U}_N(f) := \frac{1}{1-\widehat{\tau}} U_N^{\omega}(f) - \frac{\widehat{\tau}}{1-\widehat{\tau}} U_N^{\lambda}(f), \tag{12}$$

where we have:

- U_N^{ω}, U_N^{λ} are estimators for $\mathrm{L}_{\phi}^{\omega}$ and $\mathrm{L}_{\phi}^{\lambda}$, respectively.
- $\widehat{\tau} := 1 - \sum_{r=1}^{R} \widehat{\rho}_r (1-\widehat{\rho}_r)^k$, i.e., the plug-in estimator of τ.

Debiased U-Statistics

Theorem 2 (Hieu&Ledent'2026): Suppose that $R \geqslant k$ for all $r \in [R]$. Then, as long as $N \geqslant \widetilde{\mathcal{O}}\left(\max\left\{Rk, k^2 + \frac{1}{1-\tau}\right\}\right)$, for all $f \in \mathcal{F}$ and any $\delta \in (0,1)$, with probability of at least $1-\delta$:

$$\mathrm{L}_\phi(f) \leqslant \widetilde{U}_N(f) + \mathcal{O}\left(\frac{\mathfrak{C}_N(\mathcal{F}_\phi)}{1-\tau}\sqrt{\frac{k}{N}} + \frac{\mathcal{B}}{(1-\tau)^2}\sqrt{\frac{R \ln R/\delta}{N}}\right). \tag{13}$$

Debiased U-Statistics

Theorem 2 (Hieu&Ledent'2026): Suppose that $R \geqslant k$ for all $r \in [R]$. Then, as long as $N \geqslant \widetilde{\mathcal{O}}\left(\max\left\{Rk, k^2 + \frac{1}{1-\tau}\right\}\right)$, for all $f \in \mathcal{F}$ and any $\delta \in (0,1)$, with probability of at least $1-\delta$:

$$\mathrm{L}_\phi(f) \leqslant \widetilde{U}_N(f) + \mathcal{O}\left(\frac{\mathfrak{C}_N(\mathcal{F}_\phi)}{1-\tau}\sqrt{\frac{k}{N}} + \frac{\mathcal{B}}{(1-\tau)^2}\sqrt{\frac{R\ln R/\delta}{N}}\right). \tag{13}$$

Then, as long as $1-\tau \in \Omega(1)$, the above bound reduces to the following:

$$\mathrm{L}_\phi(f) \leqslant \widetilde{U}_N(f) + \mathcal{O}\left(\mathfrak{C}_N(\mathcal{F}_\phi)\sqrt{\frac{k}{N}} + \mathcal{B}\sqrt{\frac{R\ln R/\delta}{N}}\right). \tag{14}$$

References

[aro19] *A Theoretical Analysis of Contrastive Unsupervised Representation Learning*, 2019.

[hie25] *Generalization Analysis for Supervised Contrastive Representation Learning under Non-IID Settings*, 2025.

[kho20] *Supervised Contrastive Learning*, 2020.

[lei23] *Generalization Analysis for Contrastive Representation Learning*, 2023.

[soh16] *Improved Deep Metric Learning with Multi-class N-pair Loss Objective*, 2016.