Bridging Mini-Batch and Asymptotic Analysis in Contrastive Learning: From InfoNCE to Kernel-Based Losses

Panagiotis Koromilas [*1] **Giorgos Bouritsas** [*1 2] **Theodoros Giannakopoulos** [3] **Mihalis A. Nicolaou** [4]
Yannis Panagakis [1 2]

Abstract

What do different contrastive learning (CL) losses actually optimize for? Although multiple CL methods have demonstrated remarkable representation learning capabilities, the differences in their inner workings remain largely opaque. In this work, we analyse several CL families and prove that, under certain conditions, they admit the same minimisers when optimizing either their batch-level objectives or their expectations asymptotically. In both cases, an intimate connection with the hyperspherical energy minimisation (HEM) problem resurfaces. Drawing inspiration from this, we introduce a novel CL objective, coined Decoupled Hyperspherical Energy Loss (DHEL). DHEL simplifies the problem by decoupling the target hyperspherical energy from the alignment of positive examples while preserving the same theoretical guarantees. Going one step further, we show the same results hold for another relevant CL family, namely kernel contrastive learning (KCL), with the additional advantage of the expected loss being independent of batch size, thus identifying the minimisers in the non-asymptotic regime. Empirical results demonstrate improved downstream performance and robustness across combinations of different batch sizes and hyperparameters and reduced dimensionality collapse, on several computer vision datasets.
Code: github.com/pakoromilas/DHEL-KCL.git

1. Introduction

Contrastive learning has revolutionised self-supervised learning of representations from unlabelled data. Nonetheless, optimising contrastive losses exhibits significant challenges in practice. These include the need for *large batches of negative samples* leading to memory issues (He et al., 2020b; Tian et al., 2020a; Chen et al., 2020), *high sensitivity to the temperature hyperparameter* affecting model performance (Wang & Liu, 2021; Zhang et al., 2021), the propensity for *dimensionality collapse* in learned representations (Hua et al., 2021; Jing et al., 2022), and a reliance on *sophisticated hard-negative sampling strategies* (Robinson et al., 2021).

Although there are approaches to understand and address some of the challenges above in isolation, theoretical analyses of CL often use loss functions and assumptions that diverge from those effective in practice (e.g., SimCLR) or depend on conditions often unrealistic in real-world settings, e.g. infinite batch sizes, conditional independence, or simplified network architectures (Saunshi et al., 2019; Wang & Isola, 2020; Jing et al., 2022; Balestriero & LeCun, 2022; Ji et al., 2023).

This work poses a first step towards bridging the gap between different variants of the classical InfoNCE loss (Oord et al., 2018). That is, we examine their optimal solutions within two regimes: *the finite regime* concerning losses evaluated on a sampled mini-batch, and *the asymptotic regime of their expectation* (i.e. in the limit of infinite batch size). In the finite case, under a batch size condition, we show multiple InfoNCE variants share the same unique optimal solution attained when (i) positive pairs align perfectly and (ii) representations form a regular simplex inscribed in the sphere, with all pairwise distances equal. Additionally, we show that they have the same asymptotic behaviour, and in turn, the same minimisers: those identified in (Wang & Isola, 2020), i.e. (i) perfect alignment and (ii) uniform distribution on the unit sphere. Interestingly, in both cases, outcome (ii) coincides with the notion of *minimal hyperspherical energy*.

However, despite commonalities in optima, many variants exhibit notable performance discrepancies. This suggests

[*] Equal contribution [1] Department of Informatics and Telecommunications, National and Kapodistrian University of Athens [2] Archimedes AI/Athena Research Center [3] NCSR ”Demokritos” [4] The Cyprus Institute. Correspondence to: Panagiotis Koromilas <pakoromilas@di.uoa.gr>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

that optima are difficult to attain in practice. To facilitate optimisation, we introduce a new variant coined as Decoupled Hypershperical Energy Loss that fully decouples the two terms reflecting desired properties to optimise. Specifically, we propose simply replacing the classical InfoNCE denominator—see Table 1, left—with a denominator involving only negative samples, eliminating dependence on positive counterparts per Eq. (5). *The resulting alignment and uniformity terms are independent and can in principle be optimised separately*, contrary to existing variants where it is unclear if possible since the terms are coupled.

Moving one step further in the direction of Hyperspherical Energy Minimisation (HEM) in CL, we examine the optima of another family of CL losses, i.e. Kernel Contrastive Learning (KCL). Kernels first appeared in the CL literature in (Li et al., 2021), where the authors introduce a new CL objective based on kernel dependence maximisation and establish a connection with InfoNCE minimisation. In this work, we investigate a general family similar to the one of (Li et al., 2021), and discover that under certain conditions, *mini-batch KCL loss, as well as its expectation, have the same optima as all the analysed InfoNCE variants*. Importantly, KCL enjoys several interesting properties: (1) the expected loss is *independent of the number of negative samples*, and (2) *we can identify its minima non-asymptotically*.

We conducted empirical tests on DHEL and KCL using different kernel functions meeting necessary conditions. Results show both methods (i) *maintain superior performance across various and small batch sizes*, (ii) *are robust to temperature hyperparameter changes*, and (iii) *utilise more dimensions effectively, addressing th*e dimensionality collapse issue.

Our contributions can be summarized as follows:

- We prove that different general CL loss families share the same unique optimal solution in the single mini-batch regime when the batch size is no larger than the ambient dimension + 1, as well as in the asymptotic expected case.
- We introduce a novel CL loss family that decouples positive from negative samples in the uniformity term, preserves the desired properties and achieves considerable empirical improvements across various metrics.
- We establish a connection between Kernel Contrastive Learning and Hyperspherical Energy Minimisation, highlight its theoretical advantages and empirically validate that KCL can be used in place of InfoNCE variants.

2. Related Work

Contrastive Learning. Contrastive learning was formally introduced by (Chopra et al., 2005) and was later generalised to the (N+1) tuple loss (Sohn, 2016) before the popular InfoNCE loss was introduced in contrastive predictive coding (Oord et al., 2018). InfoNCE combined with a range of engineering tricks (sampled augmentations, large batch sizes, etc) is the workhorse of modern CL methods (Chen et al., 2020; Dwibedi et al., 2021; Yeh et al., 2022).

However, a range of limitations have been identified. Downstream performance is sensitive to the temperature hyperparameter, necessitating extensive tuning (Wang & Liu, 2021; Zhang et al., 2021). Empirical evidence shows that performance improves with an increased number of negative samples, leading to the requirement for large batch sizes and the incorporation of hard-negative sampling (Chen et al., 2020; Tian et al., 2020b; He et al., 2020a; Robinson et al., 2021). Additionally, there is a tendency for learned representations to use only a fraction of dimensions, not fully exploiting the capacity of the representation space (Hua et al., 2021; Jing et al., 2022).

Kernels in CL. Kernels have been used in a CL for different purposes, including incorporating prior knowledge, conditional sampling of positives and analysing the induced representation space kernels when optimising SLL objectives (Dufumier et al., 2023; Kiani et al., 2022; Tsai et al.; Johnson et al., 2023; Waida et al., 2023). Most relevant to our work is the loss of (Li et al., 2021) (a regularised version of the Hilbert-Schmidt Independence Criterion - HSIC (Gretton et al., 2005) - which reducess to a two-term KCL loss under certain conditions) which motivated the theoretical study of the CL generalisation error on downstream tasks through the lens of kernels (Waida et al., 2023).

Optima of CL Objectives. It is well known that the CL objective is asymptotically minimised for encoders that produce perfectly aligned and uniformly distributed representations (Wang & Isola, 2020). This is in line with continuous HEM which is also achieved by the uniform distribution (Liu et al., 2022). Sreenivasan et al. (2023) showed that in the mini-batch regime, the optimal solution of InfoNCE is achieved when positive representations are perfectly aligned and negatives are placed on a regular simplex (equivalent to an equiangular tight frame - ETF (Benedetto & Fickus, 2003)), which connects the solution to discrete HEM and is a special case of our Theorem 4.1. Graf et al. (2021) show that the Supervised CL loss is also minimised when each class embeddings collapse to the vertices of an ETF. Projections, a concept that is included in the contrastive learning pipeline, is also shown to help better minimise the energy (Lin et al., 2020).

Neural Collapse & Hyperspherical Energy Minimisation. Neural Collapse, where intra-class embeddings have zero variability and class means align with classifier weights in a simplex ETF during overtraining, was first identified in (Papyan et al., 2020) and explored under various training conditions (e.g., MSE, Cross Entropy, data imbalance) in (Han et al., 2021; Thrampoulidis et al., 2022; Zhu et al., 2021; Lu & Steinerberger, 2022; Zhou et al., 2022). Liu et al. (2023) generalised the notion of Neural Collapse by showing that the class means converge to the uniform distribution in the asymptotic case, thus further enhancing the connection between optimization of supervised deep learning methods and the energy minimisation problem. Moreover, energy minimisation has been effective in NN regularisation, promoting neuron diversity on a hypersphere to avoid correlated neurons (Liu et al., 2018; 2021; Lin et al., 2020).

The above are strikingly similar to the minima of mini-batch CL objectives with the unifying umbrella being the minimisation of hyperspherical energy, a problem deeply studied (see (Borodachov et al., 2019)). Notably, solutions have been identified for specific scenarios, in the discrete context, such as when the number of points M is less than or equal to the ambient dimension d+1, and for $M = 2d$, as well as continuous (Liu et al., 2022). We will discuss these cases several times throughout the paper.

3. Preliminaries and Notation

Contrastive Learning setup. *Self-supervised contrastive learning (SSCL)* is a paradigm aiming to learn data representations without having access to labels, but based solely on prior knowledge about similarities between inputs, or more strictly speaking, about *downstream task invariances*.

Formally, let $\mathcal{X}$ be a (measurable) space, i.e. the *input space* where our data reside and another (measurable) space $\mathcal{Z}$, the *embedding space*. Let $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Z}$ be an encoder (e.g. a neural network) parametrised by a set of parameters $\boldsymbol{\theta} \in \Theta$ mapping datapoints to representations. In our setup $\mathcal{Z} = \mathbb{S}^{d-1} = \{|\mathbf{u}| \in \mathbb{R}^d \mid \|\mathbf{u}\| = 1\}$ the unit sphere. We will be using the symbols $\mathbf{x}, \mathbf{y}$ for input datapoints and $\mathbf{u}, \mathbf{v}$ for representations.

Additionally, denote the (unknown) underlying data distribution with p (on $\mathcal{X}$). Further, consider a distribution of *positive pairs* with p_+ (on $\mathcal{X} \times \mathcal{X}$ and marginals equal to p), which incorporates all the data symmetries, i.e. its support are all pairs of data that are considered equivalent w.r.t. downstream tasks. We will denote the pushforward measures induced by f with $f_{\#}p$ and (with slight abuse of notation) $f_{\#}p_+$ where f is applied element-wise to $\mathbf{x}$ and $\mathbf{y}$.

Denote with $\mathbf{X} = [\mathbf{x}_1; \dots; \mathbf{x}_M] \in \mathcal{X}^M$ a collection of M input datapoints and with $\mathbf{U} = [\mathbf{u}_1; \dots; \mathbf{u}_M] \in \mathbb{R}^{d \times M}$, a collection of M representations. We will be also using the following shorthand $f_{\boldsymbol{\theta}}(\mathbf{X}) = [f_{\boldsymbol{\theta}}(\mathbf{x}_1); \dots; f_{\boldsymbol{\theta}}(\mathbf{x}_M)] = \mathbf{U}$. Also, when we sample $(\mathbf{X}, \mathbf{Y}) \sim p_+^M$ we will occasionally write $\hat{\mathbf{X}} \sim p_+^M$ instead, with $\hat{\mathbf{x}}_i = \mathbf{x}_i$ iff $i \in \{1, \dots, M\}$ and $\hat{\mathbf{x}}_i = \mathbf{y}_{i-M}$ iff $i \in \{M+1, 2M\}$ (smilarly for $(\mathbf{U}, \mathbf{V}) \sim f_{\#}p_+$ and $\hat{\mathbf{U}}$). In SSCL, the encoder is trained by optimising an objective that encourages the representations of positive pairs to be close in $\mathcal{Z}$ and those of negatives to be further. In practice, this is performed by iteratively obtaining a sample $(\mathbf{X}, \mathbf{Y})$ of M positives from p_+ and computing a *mini-batch loss* denoted with $L_{\text{CL}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))$, the gradients of which are used to update the parameters $\boldsymbol{\theta}$. This common process can be perceived as aiming to optimise an *expected loss* $\underset{(\mathbf{X},\mathbf{Y})\overset{\text{i.i.d}}{\sim} p_{\text{pos}}^M}{\mathbb{E}} [L_{\text{CL}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))]$ with gradient descent by estimating the gradients with Monte Carlo (in this case a single sample is used). For reasons that will become clear later, our interest will revolve around these two viewpoints of the loss, along with a third one, i.e. the *asymptotic expected loss* $\lim_{M\to\infty} \underset{(\mathbf{X},\mathbf{Y})\overset{\text{i.i.d}}{\sim} p_{\text{pos}}^M}{\mathbb{E}} [L_{\text{CL}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))]$.

4. Reconciling Contrastive Loss Variants

Mini-batch optimisation. We will start our investigation of the minima of different CL variants from mini-batch losses, whose gradients are used to update the parameters of the encoder. We initially focus on the following two formulas of single sample (mini-batch) contrastive losses:

$$
\begin{aligned}
L_{\text{a}}(\mathbf{U}, \mathbf{V}; \phi, \psi) &= \frac{1}{M} \sum_{i=1}^{M} \psi \left(\sum_{j=1, j\neq i}^{M} \phi \left((\mathbf{v}_j - \mathbf{v}_i)^\top \mathbf{u}_i \right) \right), \\
L_{\text{b}}(\mathbf{U}, \mathbf{V}; \phi, \psi) &= \frac{1}{M} \sum_{i=1}^{M} \psi \left(\sum_{\substack{j=1, \\ j\neq i, i+M}}^{2M} \phi \left((\hat{\mathbf{u}}_j - \mathbf{v}_j)^\top \mathbf{u}_i \right) \right),
\end{aligned}
\tag{1}
$$

where $\psi, \phi : \mathbb{R} \to \mathbb{R}$. These generalise many practical variants such as the original InfoNCE (Gutmann & Hyvärinen, 2010; Oord et al., 2018; Wu et al., 2018; Sohn, 2016; Chen et al., 2020) (L_{a}), SimCLR (or NT-Xent loss) (Chen et al., 2020) and DCL (Yeh et al., 2022) (L_{b}). The two variants differ in the datapoints that are used to compute the denominator that normalises the similarity between the pair of positive datapoints $(\mathbf{u}_i, \mathbf{v}_i)$; the former considers half of the datapoints in the batch, while the latter all of them, except $\mathbf{u}_i$ itself. Its positive counterpart $\mathbf{v}_i$ may or may not be considered. The exact formulas for each particular method can be found in Table 1 and Appendix B.1, Eq. (10).

Frequently, a symmetric version of the losses in Eq. (1) is used, defined as $L_{\text{CL-sym}}(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \left(L_{\text{CL}}(\mathbf{U}, \mathbf{V}) + L_{\text{CL}}(\mathbf{V}, \mathbf{U}) \right)$, where we omitted ϕ and ψ for brevity. In a very recent work Sreenivasan et al. (2023) studied the optima of InfoNCE for the $M \leq d+1$ case. Here, we generalise their results for all the losses of Eq. (1) - proof in Appendix B.1. Formally:

Theorem 4.1. *Consider the following optimisation problem:*

$$\underset{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M}{\operatorname{argmin}} L_{\text{CL-sym}}(\mathbf{U},\mathbf{V}), \tag{2}$$

*where $\mathbf{U},\mathbf{V}$ are tuples of M vectors on the unit $d-1$-sphere and $L_{\text{CL-sym}}$ is the symmetric version of any of the loss functions $L_{\text{a}}(\cdot,\cdot;\phi,\psi)$, $L_{\text{b}}(\cdot,\cdot;\phi,\psi)$ as defined in Eq. (1). Further, suppose the following conditions: (1) $\phi:\mathbb{R}\to\mathbb{R}$ is **increasing & convex**, (2) $\psi:\mathbb{R}\to\mathbb{R}$ is **increasing &** $\tilde{\psi}(x;\alpha)=\psi(\alpha\phi(x))$ **is convex for** $\alpha>0$ and (3) $1<M\leq d+1$. Then, the problem of Eq. (2) obtains its optimal value when $(\mathbf{U},\mathbf{V})=(\mathbf{U}^*,\mathbf{V}^*)$ with:*

$$\mathbf{U}^*=\mathbf{V}^* \text{ and } \mathbf{U}^*: \text{ regular } M-1 \text{ simplex}. \tag{3}$$

*Additionally, (4) if ψ,ϕ are **strictly increasing** and $\tilde{\psi}$ is **strictly convex** then all the $(\mathbf{U}^*,\mathbf{V}^*)$ that satisfy Eq. (3) are the **unique** optima.*

Corollary 4.2. *The mini-batch CL loss functions L_{InfoNCE}, L_{SimCLR}, L_{DCL} have the **same unique minima** on the unit sphere when $1<M\leq d+1$, i.e. all the optimal solutions of Eq. (2) will satisfy the properties of Eq. (3).*

Despite the multitude of variations of InfoNCE that have been proposed, Theorem 4.1 asserts that having the same optimal solution is conditioned solely on the monotonicity and convexity of the functions ϕ and $\tilde{\psi}$. This result might be counterintuitive given that L_{a} and L_{b} allow for different couplings of the representations $\mathbf{u}_i$ and $\mathbf{v}_i$. Additionally, it provides a first step towards clarifying the CL landscape and gives a general strategy for designing losses without compromising the optimality of the above solutions.

The discovered minima are themselves typically considered desirable in the representation learning literature (Papyan et al., 2020; Kothapalli, 2023; Wang & Isola, 2020) since the L.H.S. of Eq. (3) implies *perfect alignment*, i.e. pairs of equivalent points according to p_+ are mapped to the same representation, and the R.H.S. implies *perfect uniformity*, i.e. maximum spreading of the points in the unit sphere, a property that usually simplifies the downstream function to-be-learned. Finally, it is well known (Borodachov et al., 2019; Liu et al., 2022) that the regular $M-1$ simplex is a (unique minimiser) of the *Hypershperical Energy* for a wide variety of kernels, which illustrates the connection between mini-batch CL, neural collapse and hyperspherical energy minimisation. Note that such a connection has been previously pinpointed by Wang & Isola (2020), but only for the asymptotic behaviour of the expected mini-batch CL loss, as we discuss below.

Asymptotic behaviour of the expectation. Theorem 4.1 gives us a qualitative understanding of the similarities among CL variants and of the direction of the gradient at each training iteration, but it does not reveal the bigger picture, i.e. the objective that we are actually trying to optimise. For example, an obvious limitation of the optimum of Eq. (3) is that we can have at most two perfectly aligned points since adding one extra would compromise uniformity.

To understand the true objective, observe that the gradient of the mini-batch loss is an *unbiased estimate* of the gradient of the *expected loss* $\mathbb{E}\left[L_{\text{CL}}(f_{\boldsymbol{\theta}}(\mathbf{X}),f_{\boldsymbol{\theta}}(\mathbf{Y}))\right]$ (due to linearity of expectation and gradient) using a single sample. In other words, the expected loss is the *true loss* that we are optimising using gradient estimates. It is, therefore, more appropriate to analyse the optima of the latter. However, as we see in Table 1, the expected loss for three common CL losses: *InfoNCE, SimCLR and DCL*, depends on the batch size M even after normalising with an appropriate (missing from the original objective) constant (proof in Lemma B.4 in the Appendix using simple derivations). Thus, we resort to examining the asymptotic behaviour similarly to (Wang & Isola, 2020). Using similar arguments, it is straightforward to see that the asymptotic behaviour of the above variants is the same (proof in Appendix B.2). Formally:

Proposition 4.3. *The expectations of the following batch-level contrastive loss functions: $L_{\text{InfoNCE}}(\cdot,\cdot)$, $L_{\text{SimCLR}}(\cdot,\cdot)$, $L_{\text{DCL}}(\cdot,\cdot)$ have the **same asymptotic behaviour** when subtracting appropriate normalising constants ($\log(M-1)$ for the first and $\log(2M-2)$ for two latter), i.e. when $M\to\infty$ they converge to the asymptotic formula of InfoNCE (Wang & Isola, 2020):*

$$\mathop{\mathbb{E}}_{(\mathbf{u},\mathbf{v})\sim f_{\#}p_+}\left[-\mathbf{v}^\top\mathbf{u}/\tau\right]+\mathop{\mathbb{E}}_{\mathbf{u}\sim f_{\#}p}\left[\log\mathop{\mathbb{E}}_{\mathbf{u}'\sim f_{\#}p}\left[e^{\mathbf{u}^\top\mathbf{u}'/\tau}\right]\right]. \tag{4}$$

Therefore, the conclusions of Theorem 1 in (Wang & Isola, 2020) hold for all three variants; the first term is minimised if there exists f such that all positive pairs are perfectly aligned and the second is minimised if there exists f such that $f_{\#}p$ is the uniform distribution on the sphere $U(\mathbb{S}^{d-1})$.

5. Decoupled Hypershperical Energy Loss

Expected loss: What happens when the batch size is finite? Closely examining the equations in Table 1, we can see that the true objectives are a sum of a common alignment term and a uniformity one that varies. As previously discussed we aim to achieve perfect alignment and perfect uniformity. *The latter does not depend on p_+ (supposing that p_+ is such that perfectly optimising the alignment term does not prohibit the ability to achieve perfect uniformity)*. However, this does not straightforwardly seem to be the case for InfoNCE, SimCLR and DCL, since in all cases we observe a dependence of the uniformity term on p_+ that vanishes only asymptotically. This imposes a coupling between the two terms that can potentially hinder optimisation.

Table 1: Comparison of InfoNCE variants.

Loss name		InfoNCE	SimCLR	DCL
MB		$\frac{1}{M}\sum_{i=1}^{M}\log\left(1+\sum_{\substack{j=1\\j\neq i}}^{M} e^{(\mathbf{v}_j-\mathbf{v}_i)^\top\mathbf{u}_i/\tau}\right)$	$\frac{1}{M}\sum_{i=1}^{M}\log\left(1+\sum_{\substack{j=1\\j\neq i,M+i}}^{2M} e^{(\hat{\mathbf{u}}_j-\mathbf{v}_i)^\top\mathbf{u}_i/\tau}\right)$	$\frac{1}{M}\sum_{i=1}^{M}\log\left(\sum_{\substack{j=1\\j\neq i,M+i}}^{2M} e^{(\hat{\mathbf{u}}_j-\mathbf{v}_i)^\top\mathbf{u}_i/\tau}\right)$
EMB	$\mathbb{E}_{(\mathbf{u},\mathbf{v})\sim f_{\#}p_+}\left[-\mathbf{v}^\top\mathbf{u}\right]+$	$\mathbb{E}_{\substack{(\mathbf{u},\mathbf{v})\sim f_{\#}p_+\\ \mathbf{V}'\overset{\text{i.i.d}}{\sim} f_{\#}p^{M-1}}}\left[\log\left(e^{\mathbf{v}^\top\mathbf{u}}+\sum_{j=1}^{M-1}e^{\mathbf{u}^\top\mathbf{v}'_j}\right)\right]$	$\mathbb{E}_{\substack{(\mathbf{u},\mathbf{v})\sim f_{\#}p_+\\ \hat{\mathbf{U}}\overset{\text{i.i.d}}{\sim} f_{\#}p_+^{M-1}}}\left[\log\left(e^{\mathbf{v}^\top\mathbf{u}}+\sum_{j=1}^{2M-2}e^{\hat{\mathbf{u}}_j^\top\mathbf{u}}\right)\right]$	$\mathbb{E}_{\substack{\mathbf{u}\sim f_{\#}p\\ \hat{\mathbf{U}}\overset{\text{i.i.d}}{\sim} f_{\#}p_+^{M-1}}}\left[\log\left(\sum_{j=1}^{2M-2}e^{\hat{\mathbf{u}}_j^\top\mathbf{u}}\right)\right]$
Asymptotic	Normalising constant	$-\log(M-1)$	$-\log(2M-2)$	$-\log(2M-2)$
	Limit	$\mathbb{E}_{(\mathbf{u},\mathbf{v})\sim f_{\#}p_+}\left[\mathbf{u}^\top\mathbf{v}\right]+\mathbb{E}_{\mathbf{u}\sim f_{\#}p}\left[\log\mathbb{E}_{\mathbf{u}'\sim f_{\#}p}\left[e^{\mathbf{u}^\top\mathbf{u}'}\right]\right]$		
argmin MB ($M\leq d+1$)		$M-1$ regular simplex		
argmin EMB / Asymptotic		Unknown / $U(\mathbb{S}^{d-1})$		

Table 2: Comparison of DHEL and KCL variants.

Loss name		DHEL	KCL
MB		$\frac{1}{M}\sum_{i=1}^{M}\log\left(\sum_{\substack{j=1\\j\neq i}}^{M} e^{(\mathbf{u}_j-\mathbf{v}_i)^\top\mathbf{u}_i}\right)$	$-\frac{1}{M}\sum_{i=1}^{M}K_A\left(\mathbf{u}_i,\mathbf{v}_i\right)+\frac{\gamma}{M(M-1)}\sum_{\substack{i,j=1\\j\neq i}}^{M}K_U\left(\mathbf{u}_i,\mathbf{u}_j\right)$
EMB		$\mathbb{E}_{(\mathbf{u},\mathbf{v})\sim f_{\#}p_+}\left[-\mathbf{v}^\top\mathbf{u}\right]+\mathbb{E}_{\substack{\mathbf{u}\sim f_{\#}p\\ \mathbf{U}'\overset{\text{i.i.d}}{\sim} f_{\#}p^{M-1}}}\left[\log\left(\sum_{j=1}^{M-1}e^{\mathbf{u}^\top\mathbf{u}'_j}\right)\right]$	$\mathbb{E}_{(\mathbf{u},\mathbf{v})\sim f_{\#}p_+}\left[-K_A\left(\mathbf{u},\mathbf{v}\right)\right]+\gamma\,\mathbb{E}_{\substack{\mathbf{u}\sim f_{\#}p\\ \mathbf{u}'\sim f_{\#}p}}\left[K_U(\mathbf{u},\mathbf{u}')\right]$
Asymptotic	Normalising constant	$-\log(M-1)$	0
	Limit	$\mathbb{E}_{(\mathbf{u},\mathbf{v})\sim f_{\#}p_+}\left[-\mathbf{v}^\top\mathbf{u}\right]+\mathbb{E}_{\mathbf{u}\sim f_{\#}p}\left[\log\mathbb{E}_{\mathbf{u}'\sim f_{\#}p}\left[e^{\mathbf{u}^\top\mathbf{u}'}\right]\right]$	$\mathbb{E}_{(\mathbf{u},\mathbf{v})\sim f_{\#}p_+}\left[-K_A\left(\mathbf{u},\mathbf{v}\right)\right]+\gamma\,\mathbb{E}_{\substack{\mathbf{u}\sim f_{\#}p\\ \mathbf{u}'\sim f_{\#}p}}\left[K_U(\mathbf{u},\mathbf{u}')\right]$
argmin MB ($M\leq d+1$)		$M-1$ regular simplex	
argmin EMB / Asymptotic		Unknown / $U(\mathbb{S}^{d-1})$	$U(\mathbb{S}^{d-1})$ / $U(\mathbb{S}^{d-1})$

Decoupling uniformity from alignment Motivated by this observation, we make a simple modification on InfoNCE and propose a new CL objective that allows for an expected uniformity term that is only dependent on p:

$$L_{\text{DHEL}}(\mathbf{U},\mathbf{V})=\frac{1}{M}\sum_{i=1}^{M}-\log\left(\frac{e^{\mathbf{u}_i^\top\mathbf{v}_i/\tau}}{\sum_{\substack{j=1\\i\neq j}}^{M}e^{\mathbf{u}_i^\top\mathbf{u}_j/\tau}}\right), \quad (5)$$

which is a special case of the generalised Decoupled Hyperspherical Energy Loss:

$$L_{\text{c}}(\mathbf{U},\mathbf{V})=\frac{1}{M}\sum_{i=1}^{M}\psi\left(\sum_{\substack{j=1,\\j\neq i}}^{M}\phi\left(\left(\mathbf{u}_j-\mathbf{v}_i\right)^\top\mathbf{u}_i\right)\right) \quad (6)$$

Key advantage of DHEL. The dependence of our uniformity only on p can be also understood intuitively: DHEL is based on the observation that for perfect uniformity, it suffices to contrast a datapoint $\mathbf{x}_i$ against a *single* positive view of a negative $\mathbf{x}_j$. Adding more views, as in (Chen et al., 2020; Yeh et al., 2022), does not only seem unnecessary but might also have undesired repercussions since such a uniformity term would aim to uniformly distribute *all* points on the sphere, ignoring that half of them are positives of the other half.[1] Therefore, even though in theory the minima do not seem to be affected, previous InfoNCE variants have two *competing terms*, an issue that we overcome with DHEL.

Theoretical properties of DHEL. First off, we also analysed DHEL w.r.t. its optima in the mini-batch and the asymptotic expectation case. The following theorem shows that under the same conditions, Theorem 4.1 and Proposition 4.3 continue to hold (proofs in Appendix B.1, B.2):

Theorem 5.1. *Consider the optimisation problem of Eq.* (2) *with* $L_{\text{CL-sym}}(\cdot,\cdot)$ *being the symmetric version of the loss function* L_{c}. *Further, suppose that conditions (1) and (2) of Theorem 4.1 hold, e.g. as for our loss DHEL. Then, when* $1<M\leq d+1$, *the mini-batch CL optimisation of Eq.* (2) *obtains its optimal value* $(\mathbf{U}^*,\mathbf{V}^*)$ *as in Eq.* (3). *Moreover, the expectation of* $L_{\text{DHEL}}(\cdot,\cdot)$ *asymptotically converges to Eq.* (4) *when subtracting a normalising constant equal to* $\log(M-1)$. *Therefore, the asymptotic expectation of DHEL is minimised by any encoder* f *that is perfectly aligned and distributes representations uniformly on the sphere, i.e.* $f_{\#}p=U(\mathbb{S}^{d-1})$, *if such an encoder exists.*

6. Minima of Kernel Contrastive Learning

As discussed, for all CL variants considered, the uniform distribution on the unit sphere is known to be a minimiser of the true loss only asymptotically. Motivated by this, we seek an alternative loss whose expectation will admit the same minimiser in the non-asymptotic regime. To achieve this, we first observe that the logarithm makes the characterisation of

[1] A formal analysis of the uniformity of the symmetric SimCLR loss, using a lower bound obtained with Jensen's inequality, reveals the hyperspherical energy of a linear kernel, which is minimised when *all* points are uniformly distributed when $1<M\leq d+1$.

the optima difficult.[2] Removing the logarithm and dividing by an appropriate normalisation constant indeed provides us with a batch-level loss whose expectation is independent of the batch size:

$$\underset{(\mathbf{u},\mathbf{v})\sim f_{\#}p_+}{\mathbb{E}}\left[-\mathbf{v}^\top\mathbf{u}/\tau\right]+\underset{\substack{\mathbf{u}\sim f_{\#}p\\ \mathbf{u}'\sim f_{\#}p}}{\mathbb{E}}\left[e^{\mathbf{u}^\top\mathbf{u}_j'}/\tau\right]. \tag{7}$$

But are the desired minima preserved by this objective? To answer this question, we observe that the second term is equivalent to minimising the energy potential of the *gaussian kernel*. This is a well-studied problem (Borodachov et al., 2019), discussed also in (Wang & Isola, 2020) and is known that the minimiser is once again the uniform distribution on the sphere. Drawing inspiration from this, we examine a more general case, that of *Kernel Contrastive Learning* (Li et al., 2021; Waida et al., 2023); the mini-batch objective is as follows:

$$L_{\text{KCL}}(\mathbf{U},\mathbf{V})=-\frac{\sum\limits_{i=1}^{M}K_A\left(\mathbf{u}_i,\mathbf{v}_i\right)}{M}+\gamma\frac{\sum\limits_{\substack{i,j=1\\ j\neq i}}^{M}K_U\left(\mathbf{u}_i,\mathbf{u}_j\right)}{M(M-1)}, \tag{8}$$

where both kernels are of the form $K(\mathbf{x},\mathbf{y})=\kappa(\|\mathbf{x}-\mathbf{y}\|^2)$, with $\kappa:(0,4]\to\mathbb{R}$ and the limit $\lim\limits_{x\to 0^+}\kappa(x)$ exists and is bounded, and $\gamma>0$ is a weighting coefficient. Using known results for the hyperspherical energy minimisation problem, in the following theorem we provide the conditions that guarantee the preservation of the already discussed minima:

Theorem 6.1. *Consider the optimisation problem of Eq.* (2) *with* $L_{\text{CL-sym}}(\cdot,\cdot)$ *being the symmetric version of the loss function* L_{KCL}*. Further, suppose the following conditions: (1) the function* k_A *corresponding to kernel* K_A *is* ***decreasing****, (2)* k_U*, the function corresponding to* K_U *is* ***decreasing and convex*** *and (3)* $1<M\leq d+1$*. Then, the problem of Eq.* (2) *obtains its optimal value when* $(\mathbf{U},\mathbf{V})=(\mathbf{U}^*,\mathbf{V}^*)$ *as in Eq.* (3)*. Additionally, (4) if* κ_A *is* ***strictly decreasing*** *and* κ_U *is* ***strictly decreasing and strictly convex*** *then all the* $(\mathbf{U}^*,\mathbf{V}^*)$ *that satisfy Eq.* (3) *are the* ***unique*** *optima.*

In appendix B.3 we extend the above theorem for the case $M=2d$, where using known results from the HEM literature, we show that a minimiser of $L_{\text{KCL-sym}}$ is the *cross-polytope*. Moreover, the following proposition states that the expectation is always independent of the batch size, thus we do not have to resort to asymptotic analyses and provides the necessary conditions for the minimiser to be the uniform distribution on the sphere.

Proposition 6.2. *The expectation of the batch-level kernel contrastive loss functions* $L_{\text{KCL}}(\cdot,\cdot)$ *is* ***independent of the size of the batch****. Therefore, the batch-level loss is an* ***unbiased estimator*** *of the (asymptotic) expected loss:*

$$\underset{(\mathbf{u},\mathbf{v})\sim f_{\#}p_+}{\mathbb{E}}\left[-K_A\left(\mathbf{u},\mathbf{v}\right)\right]+\gamma\underset{\substack{\mathbf{u},\sim f_{\#}p\\ \mathbf{u}'\sim f_{\#}p}}{\mathbb{E}}\left[K_U\left(\mathbf{u},\mathbf{u}'\right)\right]. \tag{9}$$

If (1) κ_A *is (strictly) decreasing and if (2)* $\exists\,\boldsymbol{\theta}^*$ *such that* $\mathbb{P}_{(\mathbf{x},\mathbf{y})\sim p_+}\left[f_{\boldsymbol{\theta}}(\mathbf{x})=f_{\boldsymbol{\theta}}(\mathbf{y})\right]=1$*, then the set of* $\boldsymbol{\theta}^*$ *for which (2) holds are (unique) minimisers of the first term of Eq.* (9)*. Additionally, if (3) -*κ_U' *(first derivative) is* ***strictly completely monotone*** *in* $(0,4]$*, (4) the expectation defined in the l.h.s. of Eq.* (9) *is finite and (5)* $\exists\,\boldsymbol{\theta}^*$ *such that the pushforward measure* $f_{\#}p=U(\mathbb{S}^{d-1})$*, then* $\boldsymbol{\theta}^*$ *is a unique minimiser of the second term of Eq.* (9)*. Finally, if (6)* $\exists\,\boldsymbol{\theta}^*$ *such that conditions (2) and (3) can be satisfied simultaneously, then* $\boldsymbol{\theta}^*$ *is a unique minimiser of Eq.* (9)*.*

Remark. In (Li et al., 2021) it is shown that in certain cases (i.e. for a discrete distibution) a two-term kernel contrastive loss as the one in Eq. (9) arises as proportional to a kernel dependence measure they aim to maximise (HSIC). However, in their paper a different loss is used in practice; first, they use a biased estimator different from Eq. (8) and second they add a regulariser. Additionally, they identify a connection only with they asymptotic version of InfoNCE and they do not study the minima of HSIC as we do in Theorem 6.1 and Proposition 6.2.

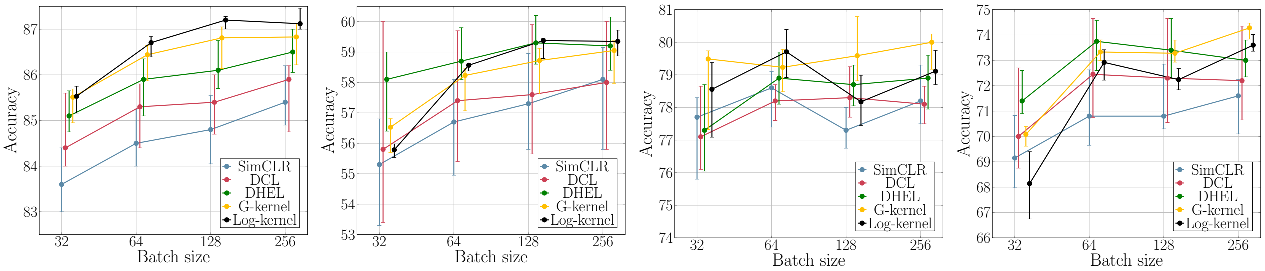
7. Experimental Evaluation

In this section, we empirically verify our theoretical results. Our methods are compared to two popular techniques in the literature: (i) **SimCLR** (Chen et al., 2020) the most used implementation of contrastive pretaining that also demonstrates consistency in terms of performance and (ii) **DCL** (Yeh et al., 2022) the only method in the literature that demonstrates robust performance for various and small batch sizes. We implement (iii) **DHEL** and (iv) **two KCL losses for the Gaussian and Logarithmic kernel** (see Appendix A).

Following common practices (Wang et al., 2021; Yeh et al., 2022; Zhang et al., 2022; Wang & Isola, 2020), we conduct experiments on four popular image classification datasets, namely *CIFAR10, CIFAR100, STL-10, and ImageNet-100*. To illustrate robustness, we validate the performance for a range of each method's hyperparameters and different batch sizes. In addition, to understand the quality of the learned representations, we demonstrate the behaviour of several desired properties.

We choose ResNet50 as the encoder architecture for the ImageNet-100 dataset and ResNet18 for the other datasets. We train our models for 200 epochs on four batch sizes

[2] Using Jensen's inequality we can attempt to minimise a lower bound as in Theorem 4.1, but equality can hold only for DCL and DHEL and only if for all $\mathbf{u}$ and any M negatives $\mathbf{u}_j'$, the inner products $\mathbf{u}^\top\mathbf{u}_j'$ are equal $\forall j$. If the minimiser of the bound is the uniform distribution this can only happen for $d=2$ (Cho, 2009).



(a) CIFAR10

(b) CIFAR100

(c) STL10

(d) ImageNet-100

Figure 1: Median performance for different batch sizes. Errors against each methods hyperparameters are calculated using the 25% and 75% quantiles. DHEL and KCL showcase improved both performance and robust against hyperparameters.

(32, 64, 128, 256) and optimise them using SGD. Downstream performance is measured using the classical linear evaluation benchmarking technique: we train a linear layer on the learned representations for 200 epochs. Following (Wang et al., 2021), 11 temperatures (regarding the methods that use temperature as a hyperparameter) are tested, while for kernel methods, along with their hyperparameter, we additionally run experiments for different weighting coefficients γ. Further details on learning rates, schedulers, augmentations etc. are provided in Appendix C.3.

7.1. Downstream performance and robustness

The error diagram in Figure 1 illustrates both performance and robustness; for each method and batch size, we include the median and 25% (lower error) and 75% (upper error) quantiles calculated on the accuracy for different hyperparameters (the median was preferred to the mean due to the presence of a few outliers across all methods).

Performance. First off, *DHEL significantly outperforms SimCLR across all datasets and batch sizes*, with the upper performance of the latter being smaller than that of the former. Second, *the median of DHEL is always higher than that of DCL*, while their upper performance are comparative (in CIFAR10 and STL10 DHEL's upper performance is always higher than DCL). Additionally, *kernel methods outperform both SimCLR and DCL* competitors, while in several cases, *kernels improve further upon DHEL*. Overall, *both DHEL and KCL methods showcase significant improvements in median performance, with their upper quantiles being, in most cases, comparable or better than DCL.*

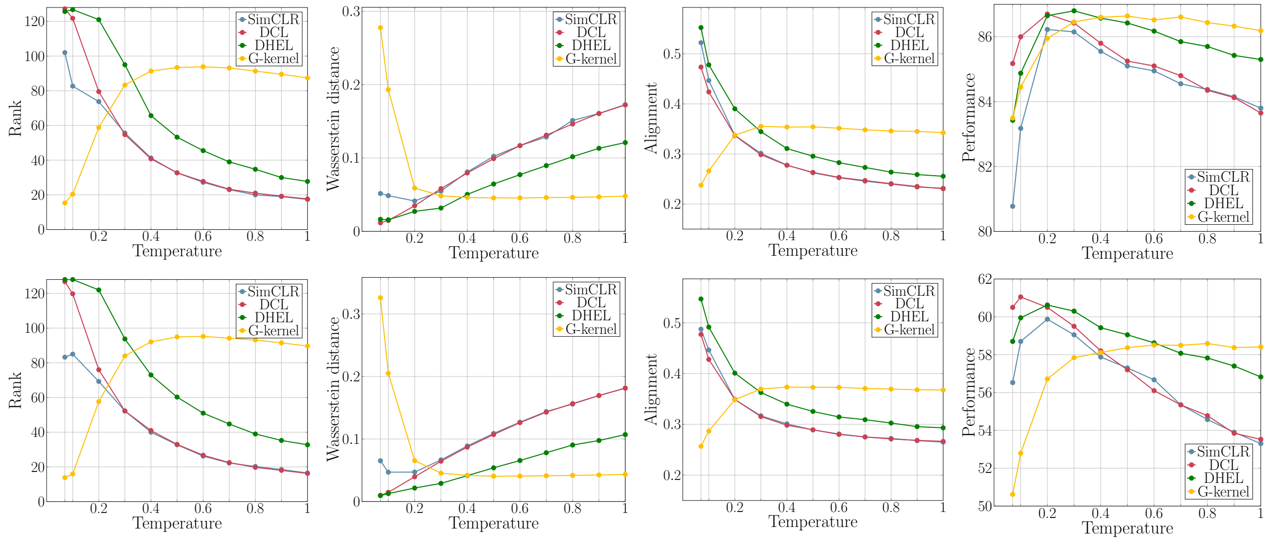
Performance w.r.t. batch size. DHEL and KCL *largely outperform competitors for small batch sizes* in terms of median performance. It is inferred that *our methods enable high downstream performance Contrastive Learning pretraining for a small number of negative samples*. Note that, typically in the literature much larger batch sizes are used, e.g. *SimCLR needs batch sizes greater than 512 (Chen et al., 2020; Yeh et al., 2022)* and He et al. (2020a) use batch sizes as large as 64K.

Robustness w.r.t. hyperparameters. In addition to the fact that median accuracy of both DHEL and kernel methods consistently outperform SimCLR and DCL, *their performance deviates in a much smaller range*, thus empirically proving robustness w.r.t temperature and γ for KCL. Importantly, observe that G-kernel and Log-kernel, in most cases demonstrate small spread around the median, a property that hints that are easier to optimise, in accordance to our result that kernel mini-batch losses are unbiased estimators of an objective minimised by perfect alignment & perfect HE.

7.2. Ablation studies

In the following section, we ablate the aforementioned methods w.r.t. various metrics: **(1) Alignment:** An estimate of the expected distance between the representations of a positive pair. **(2) Uniformity**: The logarithm of an estimate of the expected pairwise Gaussian energy potential of the distribution of the learned representations **(3) Rank**: The rank of a matrix of representations sampled from p; reflects the number of dimensions utilised and thus the ability to linearly separate our data (Cover, 1965; Garrido et al., 2023). **(4) Effective rank**: A smooth rank approximation (Roy & Vetterli, 2007; Garrido et al., 2023), that is less prone to numerical errors; has been found in practice to correlate well with downstream performance. Please refer to Appendix C for more details.

Novel metric: Wasserstein distance between similarity distributions. We introduce **(5)** a novel metric. The motivation is the fact that, although the uniformity metric is minimised when the representations are uniformly distributed on the unit sphere, it relies on a specific kernel (gaussian) and requires selecting a parameter t. Here we propose instead a metric that measures the distance between the ideal inner product (or equivalently L2 distance) distribution and the one that our algorithm yields. In particular, we



(a) Rank

(b) Wasserstein distance

(c) Alignment

(d) Performance

Figure 2: Mean value of properties vs temperature calculated on CIFAR10 (top) & CIFAR100 (bottom) dataset

estimate the *1-Wasserstein distance* $W_1(q_{\text{sim}}, p_{\text{sim}})$ where p_{sim} is the p.d.f of the inner products when $\mathbf{u}, \mathbf{u}' \sim U(\mathbb{S}^{d-1})$ and q_{sim} is the corresponding one when $\mathbf{u}, \mathbf{u}' \sim f_{\#}p$. According to (Cho, 2009) the former has a closed-form expression that we can use to obtain samples and estimate the distance from data. Importantly, in Appendix C, we show that the *uniformity metric underestimates the closeness of q_{sim} to the ideal distribution of similarities*, and therefore our metric paints a more complete picture of the learned distribution of representations. In Figure 2 we demonstrate the mean across batch sizes (and γ for g-kernel) for 3 representative properties (alignment, Wasserstein distance & rank) and performance for different temperatures (including all methods that use this hyperparameter and are comparable) in the CIFAR10 and CIFAR100 datasets. The traditional uniformity metric as well as the effective rank are presented in the Appendix C.4.

Dimensionality collapse. With the maximum number of available dimensions being 128, *DHEL consistently utilises a greater number of dimensions*, e.g.in CIFAR100 uses more than double the dimensions as compared to competitors. Additionally, *gaussian-KCL demonstrates once again its stability*, without compromising the rank, albeit not reaching the highest values of DHEL. **Uniformity.** *DHEL manages to learn representations that are consistently more uniformly distributed across temperature values.* It is also verified that in the low-temperature regime, all methods learn uniformly distributed features, a behaviour that is known as the uniformity-tolerance dilemma (Wang & Liu, 2021). **Alignment.** SimCLR and DCL learn more aligned representations, which along with the aforementioned findings seem to imply that uniformity is preferential for DHEL (see also Section 8). It is still not clear why this happens, but we may speculate that our modification indeed facilitates optimisation of the second term, and therefore a weighting coefficient might alleviate this modest imbalance. *Nevertheless, the current balancing seems to benefit downstream performance more.*

Performance. Downstream performance is decreased with respect to temperature for all methods, except G-kernel, but with *DHEL enjoying a greater range of effective temperatures and a smaller rate of decrease in accuracy*. Once again, observe the *remarkable stability of the G-kernel across different hyperparameter values*.

8. Discussion

Non-asymptotic optima. As demnonstrated in Table 1 and Table 2 all examined loss variants share the same minimisers in both the mini batch and the asymptotic scenarios. However, there's a practical discrepancy: the optimal solutions for the former case are attainable by optimizing each batch separately, while the latter scenario is not feasible due to the finite size of the dataset.

In the only practical scenario, where the non-asymptotic expected loss is optimised, only the optimal solution of the kernel based methods is known. This means that in practice contrastive methods may or may not have the same optima. This, along the difference in the bias of the estimator (Proposition 6.2), may explain the inconsistency in both performance and properties between DHEL and KCL methods. Of course, assuming that the target of Contrastive

Learning is indeed perfect alignment and uniformity, then *the fact that KCL optimises for it in the non-asymptotic regime is favourable*.

Batch size dependence. Proposition 6.2 suggests that the expectation of the mini-batch KCL loss is independent of the batch size. In other words, *different batch sizes yield the same expected loss*, contrary to the InfoNCE methods, where essentially, when one changes the batch size, the loss that is optimised changes as well. Does this mean that KCL should have stable performance across batch sizes? In practice one does actually compute an estimation of the expected loss which is affected by the batch size, while the same holds for the gradient of the expected loss. That is the batch size affects the InfoNCE losses by changing both the actual value of the expected loss and its estimate, while the optimisation of kernel losses is affected only by the second.

Too small batch sizes might lead to suboptimal solutions or slow optimisation, due to high-variance gradients. This holds for all methods, regardless of the loss they are optimising for, which might in part explain why all methods tend to improve when increasing the batch size. However, except for the gradient variance, one should also consider the gradient bias. This is zero for KCL, but not zero for the other methods (Chen et al., 2022).

Overall, when increasing the batch size over a threshold below which the gradients are too noisy, KCL obtains better gradient estimates than its counterparts, due to the zero gradient bias. This probably explains why KCL achieves better performance in absolute numbers. Further increasing the batch size, improves the gradient estimates even more for KCL, but also for the other methods since both their gradient variance and their gradient bias are reduced, which might explain why performance keeps improving across all methods.

Optimisation vs downstream performance. Both InfoNCE and kernel-based losses seek to optimise for uniformity and alignment. Our methods do not in all cases achieve to better optimise for both these desired properties. Instead they achieve a balance that better reflects on downstream performance. They also tend to favour uniformity more (Figure 2) which is probably desired. Recent works (Gupta et al., 2023; Xie et al., 2022) have argued that perfect alignment might not be ideal for downstream performance, since several downstream tasks might not actually be invariant to the augmentations from which we obtain the positive samples. For additional experimental results on this matter, please refer to Figure 3.

Connection to supervised learning. When performing supervised training beyond zero error the class means either form a simplex ETF in the non-asymptotic case or follow a uniform distribution asymptotically, with zero in-class variability (Liu et al., 2023). Our analysis shows that the same results hold for contrastive learning. However, in this case, the results apply to individual data points rather than classes. By considering contrastive learning as instance discrimination (Wu et al., 2018) —where each data point represents a unique class— we can identify connections for both optimisation and the representation spaces learned through self-supervised and supervised methods.

Limitations. Our analysis of InfoNCE loss variants focuses on the mini-batch and asymptotic optima. However, it does not address the non-asymptotic optima, which is the scenario typically encountered in practice. Adding that InfoNCE loss variants are fundamentally different from most Machine Learning objectives, where there is no influence of the batch size on the expected loss, the practical behavior of such loss functions requires further research in order to enhance our understanding of contrastive learning optimisation. In contrast, our work does provide the non-asymptotic optima of kernel methods. While these methods serve as unbiased estimators of their expected loss, examining the variance of these estimators can offer valuable insights that can guide the design of methods that are even more robust across different batch sizes.

The experiments in this work aimed to provide empirical results comparing InfoNCE and kernel-based methods across various properties, including robustness to batch sizes and hyperparameters, rank, uniformity, and alignment. However, a more comprehensive evaluation is necessary to better understand the applicability of these methods. To thoroughly assess the superiority of kernel methods, they need to be tested on large-scale datasets and examined under practical conditions such as large batch sizes, memory banks, and momentum contrast (He et al., 2020b).

9. Conclusion

In this paper, we made a step towards bridging theory and practice in CL by proving InfoNCE variants share fundamental finite sample and asymptotic optimal solutions. To better attain these optima exhibiting alignment and uniformity, we proposed Decoupled Hyperspherical Energy Loss. Furthermore, establishing kernel CL as equivalent to hyperspherical energy minimization provides optimization advantages. Both new methods empirically demonstrate consistent improvements in downstream performance across different hyperparameters and small batch sizes, as well as mitigation of dimensionality collapse.

Impact Statement

This paper advances the understanding of contrastive learning (CL) optimisation goals, aiming not just to boost model performance but to clarify the underpinnings of

CL losses and their relation to hyperspherical energy minimization (HEM). While our focus is on theoretical insights and introducing the novel Decoupled Hyperspherical Energy Loss (DHEL), this work also lays the groundwork for developing state-of-the-art models with improved robustness and reduced dimensionality collapse. We acknowledge the potential dual-use of our findings and advocate for responsible application and the development of safeguards against misuse. To facilitate further research, we make our code plublicly available at https://github.com/pakoromilas/DHEL-KCL.git.

Acknowledgements

Panagiotis Koromilas was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the 4th Call for HFRI PhD Fellowships (Fellowship Number: 10816). Giorgos Bouritsas and Yannis Panagakis were supported by the project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. This research was partially supported by a grant from The Cyprus Institute on Cyclone.

References

Balestriero, R. and LeCun, Y. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *Advances in Neural Information Processing Systems*, 35:26671–26685, 2022.

Benedetto, J. J. and Fickus, M. Finite normalized tight frames. *Advances in Computational Mathematics*, 18: 357–385, 2003.

Borodachov, S. V., Hardin, D. P., and Saff, E. B. *Discrete energy on rectifiable sets*. Springer, 2019.

Chen, C., Zhang, J., Xu, Y., Chen, L., Duan, J., Chen, Y., Tran, S., Zeng, B., and Chilimbi, T. Why do we need large batchsizes in contrastive learning? a gradient-bias perspective. *Advances in Neural Information Processing Systems*, 35:33860–33875, 2022.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on Machine Learning (ICML)*, pp. 1597–1607. PMLR, 2020.

Cho, E. Inner product of random vectors. *International Journal of Pure and Applied Mathematics*, 56(2):217–221, 2009.

Chopra, S., Hadsell, R., and LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pp. 539–546. IEEE, 2005.

Cover, T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

Dufumier, B., Barbano, C. A., Louiset, R., Duchesnay, E., and Gori, P. Integrating prior knowledge in contrastive learning with kernel. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8851–8878. PMLR, 23–29 Jul 2023.

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., and Zisserman, A. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9588–9597, 2021.

Garrido, Q., Balestriero, R., Najman, L., and Lecun, Y. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning*, pp. 10929–10974. PMLR, 2023.

Graf, F., Hofer, C., Niethammer, M., and Kwitt, R. Dissecting supervised contrastive learning. In *International Conference on Machine Learning (ICML)*, pp. 3821–3830. PMLR, 2021.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Gupta, S., Robinson, J., Lim, D., Villar, S., and Jegelka, S. Structuring representation geometry with rotationally equivariant contrastive learning. In *The Twelfth International Conference on Learning Representations*, 2023.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Han, X., Papyan, V., and Donoho, D. L. Neural collapse under mse loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations (ICLR)*, 2021.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020a.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. Computer Vision Foundation / IEEE, 2020b.

Hua, T., Wang, W., Xue, Z., Ren, S., Wang, Y., and Zhao, H. On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9598–9608, 2021.

Ji, W., Deng, Z., Nakada, R., Zou, J., and Zhang, L. The power of contrast for feature learning: A theoretical analysis. *Journal of Machine Learning Research*, 24(330): 1–78, 2023.

Jing, L., Vincent, P., LeCun, Y., and Tian, Y. Understanding dimensional collapse in contrastive self-supervised learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Johnson, D. D., Hanchi, A. E., and Maddison, C. J. Contrastive learning can find an optimal basis for approximately view-invariant functions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Kiani, B. T., Balestriero, R., Chen, Y., Lloyd, S., and LeCun, Y. Joint embedding self-supervised learning in the kernel regime. *arXiv preprint arXiv:2209.14884*, 2022.

Kothapalli, V. Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=QTXocpAP9p.

Li, Y., Pogodin, R., Sutherland, D. J., and Gretton, A. Self-supervised learning with kernel dependence maximization. *Advances in Neural Information Processing Systems*, 34:15543–15556, 2021.

Lin, R., Liu, W., Liu, Z., Feng, C., Yu, Z., Rehg, J. M., Xiong, L., and Song, L. Regularizing neural networks via minimizing hyperspherical energy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6917–6927, 2020.

Liu, W., Lin, R., Liu, Z., Liu, L., Yu, Z., Dai, B., and Song, L. Learning towards minimum hyperspherical energy. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Liu, W., Lin, R., Liu, Z., Xiong, L., Schölkopf, B., and Weller, A. Learning with hyperspherical uniformity. In *International Conference On Artificial Intelligence and Statistics (AISTATS)*, pp. 1180–1188. PMLR, 2021.

Liu, W., Yu, L., Weller, A., and Schölkopf, B. Generalizing and decoupling neural collapse via hyperspherical uniformity gap. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2022.

Liu, W., Yu, L., Weller, A., and Schölkopf, B. Generalizing and decoupling neural collapse via hyperspherical uniformity gap. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Lu, J. and Steinerberger, S. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022. Special Issue on Harmonic Analysis and Machine Learning.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Robinson, J. D., Chuang, C., Sra, S., and Jegelka, S. Contrastive learning with hard negative samples. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Roy, O. and Vetterli, M. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H. A theoretical analysis of contrastive unsupervised representation learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5628–5637. PMLR, 2019.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Sreenivasan, K., Lee, K., Lee, J.-G., Lee, A., Cho, J., yong Sohn, J., Papailiopoulos, D., and Lee, K. Mini-batch optimization of contrastive loss. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Thrampoulidis, C., Kini, G. R., Vakilian, V., and Behnia, T. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27225–27238, 2022.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J. (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pp. 776–794. Springer, 2020a.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020b.

Tsai, Y. H., Li, T., Ma, M. Q., Zhao, H., Zhang, K., Morency, L., and Salakhutdinov, R. Conditional contrastive learning with kernel. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Waida, H., Wada, Y., Andéol, L., Nakagawa, T., Zhang, Y., and Kanamori, T. Towards understanding the mechanism of contrastive learning via similarity structure: A theoretical analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 709–727. Springer, 2023.

Wang, F. and Liu, H. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2495–2504, 2021.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, pp. 9929–9939. PMLR, 2020.

Wang, X., Liu, Z., and Yu, S. X. Unsupervised feature learning by cross-level instance-group discrimination. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12586–12595, 2021.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Xie, Y., Wen, J., Lau, K. W., Rehman, Y. A. U., and Shen, J. What should be equivariant in self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4111–4120, 2022.

Yeh, C.-H., Hong, C.-Y., Hsu, Y.-C., Liu, T.-L., Chen, Y., and LeCun, Y. Decoupled contrastive learning. In *European Conference on Computer Vision*, pp. 668–684. Springer, 2022.

Zhang, C., Zhang, K., Zhang, C., Pham, T. X., Yoo, C. D., and Kweon, I. S. How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. In *International Conference on Learning Representations*, 2021.

Zhang, C., Zhang, K., Pham, T. X., Niu, A., Qiao, Z., Yoo, C. D., and Kweon, I. S. Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14441–14450, 2022.

Zhou, J., Li, X., Ding, T., You, C., Qu, Q., and Zhu, Z. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In *International Conference on Machine Learning*, pp. 27179–27202. PMLR, 2022.

Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., and Qu, Q. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:29820–29834, 2021.

A. Additional Preliminaries and Notations

Detailed Formulas for InfoNCE Variants. It is not hard to see that all the losses considered in the main text are special cases of the above two losses:

$$\begin{aligned}
L_{\text{InfoNCE}}(\mathbf{U}, \mathbf{V}) &= \frac{1}{M}\sum_{i=1}^{M} -\log\left(\frac{e^{\mathbf{u}_i^\top \mathbf{v}_i/\tau}}{\sum_{j=1}^{M} e^{\mathbf{u}_i^\top \mathbf{v}_j/\tau}}\right) = \frac{1}{M}\sum_{i=1}^{M}\log\left(1 + \sum_{j=1, j\neq i}^{M} e^{(\mathbf{v}_j - \mathbf{v}_i)^\top \mathbf{u}_i/\tau}\right) \\
L_{\text{DHEL}}(\mathbf{U}, \mathbf{V}) &= \frac{1}{M}\sum_{i=1}^{M} -\log\left(\frac{e^{\mathbf{u}_i^\top \mathbf{v}_i/\tau}}{\sum_{\substack{j=1\\ i\neq j}}^{M} e^{\mathbf{u}_i^\top \mathbf{u}_j/\tau}}\right) = \frac{1}{M}\sum_{i=1}^{M}\log\left(\sum_{j=1, j\neq i}^{M} e^{(\mathbf{u}_j - \mathbf{v}_i)^\top \mathbf{u}_i/\tau}\right) \\
L_{\text{SimCLR}}(\mathbf{U}, \mathbf{V}) &= \frac{-1}{M}\sum_{i=1}^{M}\log\left(\frac{e^{\mathbf{u}_i^\top \mathbf{v}_i/\tau}}{\sum_{j=1}^{M} e^{\mathbf{u}_i^\top \mathbf{v}_j/\tau} + \sum_{\substack{j=1\\ i\neq j}}^{M} e^{\mathbf{u}_i^\top \mathbf{u}_j/\tau}}\right) \\
&= \frac{1}{M}\sum_{i=1}^{M}\log\left(1 + \sum_{j=1, j\neq i}^{M}\left(e^{(\mathbf{v}_j - \mathbf{v}_i)^\top \mathbf{u}_i/\tau} + e^{(\mathbf{u}_j - \mathbf{v}_i)^\top \mathbf{u}_i/\tau}\right)\right) \\
L_{\text{DCL}}(\mathbf{U}, \mathbf{V}) &= \frac{-1}{M}\sum_{i=1}^{M}\log\left(\frac{e^{\mathbf{u}_i^\top \mathbf{v}_i/\tau}}{\sum_{\substack{j=1\\ j\neq i}}^{M} e^{\mathbf{u}_i^\top \mathbf{v}_j/\tau} + \sum_{\substack{j=1\\ i\neq j}}^{M} e^{\mathbf{u}_i^\top \mathbf{u}_j/\tau}}\right) \\
&= \frac{1}{M}\sum_{i=1}^{M}\log\left(\sum_{j=1, j\neq i}^{M}\left(e^{(\mathbf{v}_j - \mathbf{v}_i)^\top \mathbf{u}_i/\tau} + e^{(\mathbf{u}_j - \mathbf{v}_i)^\top \mathbf{u}_i/\tau}\right)\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
L_{\text{InfoNCE}}(\mathbf{U}, \mathbf{V}) &= L_{\text{a}}\left(\mathbf{U}, \mathbf{V}; \exp(x/\tau); \log(1+x)\right) \\
L_{\text{DHEL}}(\mathbf{U}, \mathbf{V}) &= L_{\text{c}}\left(\mathbf{U}, \mathbf{V}; \exp(x/\tau); \log(x)\right) \\
L_{\text{SimCLR}}(\mathbf{U}, \mathbf{V}) &= L_{\text{b}}\left(\mathbf{U}, \mathbf{V}; \exp(x/\tau); \log(1+x)\right) \\
L_{\text{DCL}}(\mathbf{U}, \mathbf{V}) &= L_{\text{b}}\left(\mathbf{U}, \mathbf{V}; \exp(x/\tau); \log(x)\right)
\end{aligned} \tag{10}$$

Kernels. Notable examples of kernels that obey the conditions that we encounter in this paper are the following:

- *Linear*: $K_t^{\text{lin}}(\mathbf{x}, \mathbf{y}) = -t\|\mathbf{x} - \mathbf{y}\|^2 = \kappa_t^{\text{lin}}(\|\mathbf{x} - \mathbf{y}\|^2)$, where $\kappa_t^{\text{lin}}(x) = -tx$.
- *Gaussian*: $K_t^{\text{gauss}}(\mathbf{x}, \mathbf{y}) = e^{-t\|\mathbf{x}-\mathbf{y}\|^2} = \kappa_t^{\text{gauss}}(\|\mathbf{x} - \mathbf{y}\|^2)$, where $\kappa_t^{\text{gauss}}(x; t) = e^{-tx}$.
- *Riesz*: $K_s^{\text{riesz}}(\mathbf{x}, \mathbf{y}) = \text{sign}(s)\|\mathbf{x} - \mathbf{y}\|^{-s} = \kappa_s^{\text{riesz}}(\|\mathbf{x} - \mathbf{y}\|^2)$, where $\kappa_s^{\text{riesz}}(x) = \text{sign}(s)x^{-s/2}$.
- *Inverse Multiquadric (IMQ)*: $K_c^{imq}(\mathbf{x}, \mathbf{y}) = -\frac{1}{2}\log\left(s\|\mathbf{x} - \mathbf{y}\|^2 + \beta\right) = \kappa_{s,\beta}^{\log}(\|\mathbf{x} - \mathbf{y}\|^2)$, where $\kappa_{s,\beta}^{\log}(x) = -\frac{1}{2}\log\left(sx + \beta\right)$.
- *Logarithmic*: $K_{s,\beta}^{\log}(\mathbf{x}, \mathbf{y}) = -\frac{1}{2}\log\left(s\|\mathbf{x} - \mathbf{y}\|^2 + \beta\right) = \kappa_{s,\beta}^{\log}(\|\mathbf{x} - \mathbf{y}\|^2)$, where $\kappa_{s,\beta}^{\log}(x) = -\frac{1}{2}\log\left(sx + \beta\right)$.

The properties of kernel functions that arise in the theoretical results are the below: (1) (strict) *monotonicity*, (2) (strict) *convexity*, (3) (strict) *absolute monotonicity*, i.e. derivatives of all orders $f^{(n)}$ exist and are non-negative (positive in the strict case) everywhere)and (4) *complete monotonicity*, i.e. derivatives of all orders exist and $(-1)^n f^{(n)} \geq 0$ everywhere (> 0 for the strict case).

With elementary derivations, it is easy to see the following for $t > 0$, κ_t^{lin} and κ_t^{gauss} are strictly decreasing and convex, while the latter is also strictly convex. For $s > -2$, κ_s^{riesz} is strictly decreasing and strictly convex, while the same holds for $\kappa_{s,\beta}^{\log}$ when $s, \beta > 0$. Additionally, for $t > 0$, κ_t^{lin} is completely monotone, κ_t^{gauss} is strictly completely monotone, while the same holds for their negative first derivatives $-(\kappa_t^{\text{lin}})^{(1)}$, $-(\kappa_t^{\text{gauss}})^{(1)}$. κ_s^{riesz} is strictly completely monotone for $s > 0$, while its negative first derivative is strictly completely monotone for $s > -2$. $\kappa_{s,\beta}^{\log}$ is strictly completely monotone for $s, \beta > 0$, while the same holds for its negative first derivative.

B. Deferred Proofs

B.1. Minima of Mini-Batch Contrastive Losses

Now we can proceed in proving our first theorem which encapsulates Theorems 4.1 and 5.1 of the main paper.

Theorem B.1. *Consider the following optimisation problem:*

$$\underset{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M}{\operatorname{argmin}} L_{\text{CL-sym}}(\mathbf{U},\mathbf{V}), \tag{11}$$

with $L_{\text{CL-sym}}(\mathbf{U},\mathbf{V}) = \frac{1}{2}\left(L_{\text{CL}}(\mathbf{U},\mathbf{V}) + L_{\text{CL}}(\mathbf{V},\mathbf{U})\right)$*, where* $\mathbb{S}^{d-1} = \{\mathbf{u} \in \mathbb{R}^d \mid \|\mathbf{u}\|_2 = 1\}$ *a unit sphere of* d *dimensions,* $\mathbf{U},\mathbf{V}$ *are tuples of* M *vectors on the unit sphere and* $L_{\text{CL}}(\cdot,\cdot)$ *is any of the loss functions* $\{L_{\text{a}}(\cdot,\cdot;\phi,\psi), L_{\text{b}}(\cdot,\cdot;\phi,\psi), L_{\text{c}}(\cdot,\cdot;\phi,\psi),\}$ *as defined in Eq.* (1) *and* (6)*. Further, suppose the following conditions: (1)* $\phi:\mathbb{R}\to\mathbb{R}$ *is* ***increasing & convex****, (2)* $\psi:\mathbb{R}\to\mathbb{R}$ *is* ***increasing &*** $\tilde{\psi}(x;\alpha) = \psi\left(\alpha\phi\left(x\right)\right)$ ***is convex for*** $\alpha > 0$ *and (3)* $1 < M \leq d+1$*. Then, the optimisation problem of Eq.* (2) *obtains its optimal value* $(\mathbf{U}^*,\mathbf{V}^*)$ *when:*

$$\mathbf{U}^* = \mathbf{V}^* \quad \textit{and} \quad \mathbf{U}^* = [\mathbf{u}_1^*,\ldots,\mathbf{u}_M^*] \textit{ form a regular } M-1 \textit{ simplex centered at the origin.} \tag{12}$$

Additionally, (4) if ψ,ϕ *are* ***strictly increasing*** *and* $\tilde{\psi}$ *is* ***strictly convex*** *then all the* $(\mathbf{U}^*,\mathbf{V}^*)$ *that satisfy Eq.* (12) *are the* ***unique*** *optima.*

Proof. Let us start from L_{a}. Our proof will follow similar steps as in (Sreenivasan et al., 2023) but in a more general fashion.

Part I: L_{a}**.**

First, we will use the convexity of ϕ to lower bound the inner sum.

$$\begin{aligned}\sum_{j=1,j\neq i}^{M}\phi\big((\mathbf{v}_j-\mathbf{v}_i)^\top\mathbf{u}_i\big) &\overset{(a)}{\geq} (M-1)\phi\left(\frac{1}{M-1}\sum_{j=1,j\neq i}^{M}\left(\mathbf{v}_j^\top\mathbf{u}_i-\mathbf{v}_i^\top\mathbf{u}_i\right)\right)\\ &= (M-1)\phi\left(\frac{\mathbf{v}^\top\mathbf{u}_i-\mathbf{v}_i^\top\mathbf{u}_i-(M-1)(\mathbf{v}_i^\top\mathbf{u}_i)}{M-1}\right) = (M-1)\phi\left(\frac{\mathbf{v}^\top\mathbf{u}_i-M\mathbf{v}_i^\top\mathbf{u}_i}{M-1}\right),\end{aligned}$$

where (a) follows from Jensen's inequality (Condition (1)) and $\mathbf{v} = \sum_{j=1}^{M}\mathbf{v}_j$. Then, we will use the convexity of $\tilde{\psi}$ (Condition (2)) to bound the outer sum.

$$\begin{aligned}L_{\text{a}}(\mathbf{U},\mathbf{V};\phi,\psi) &\overset{(b)}{\geq} \frac{1}{M}\sum_{i=1}^{M}\psi\left((M-1)\phi\left(\frac{\mathbf{v}^\top\mathbf{u}_i-M\mathbf{v}_i^\top\mathbf{u}_i}{M-1}\right)\right)\\ &= \frac{1}{M}\sum_{i=1}^{M}\tilde{\psi}\left(\frac{\mathbf{v}^\top\mathbf{u}_i-M\mathbf{v}_i^\top\mathbf{u}_i}{M-1};M-1\right)\\ &\overset{(c)}{\geq} \tilde{\psi}\left(\frac{1}{M}\sum_{i=1}^{M}\left(\frac{\mathbf{v}^\top\mathbf{u}_i-M\mathbf{v}_i^\top\mathbf{u}_i}{M-1}\right);M-1\right)\\ &= \psi\left((M-1)\phi\left(\frac{1}{M}\sum_{i=1}^{M}\left(\frac{\mathbf{v}^\top\mathbf{u}_i-M\mathbf{v}_i^\top\mathbf{u}_i}{M-1}\right)\right)\right)\\ &= \psi\left((M-1)\phi\left(\frac{1}{M(M-1)}\left(\mathbf{v}^\top\mathbf{u}-M\sum_{i=1}^{M}\mathbf{v}_i^\top\mathbf{u}_i\right)\right)\right)\\ &= \psi\left((M-1)\phi\left(\frac{1}{M}\sum_{i=1}^{M}\left(\frac{\mathbf{v}^\top\mathbf{u}_i-M\mathbf{v}_i^\top\mathbf{u}_i}{M-1}\right)\right)\right)\\ &\overset{(d)}{\geq} \psi\left((M-1)\phi\left(\frac{1}{M(M-1)}\left(\mathbf{v}^{*\top}\mathbf{u}^*-M\sum_{i=1}^{M}\mathbf{v}_i^{*\top}\mathbf{u}_i^*\right)\right)\right),\end{aligned}$$

where $\mathbf{u} = \sum_{j=1}^{M} \mathbf{u}_j$ and $\mathbf{u}_i^*, \mathbf{v}_i^*$ are the minima of the inner argument, (b) follows from the fact that ψ is increasing (Condition (2)), (c) follows from the fact that $M-1>0$ and Jensen's inequality (Condition (2)) and (d) from the fact that ψ and ϕ are increasing (Conditions (1) and (2)) and $M-1>0$. Thus, it suffices to solve the following optimisation problem:

$$\underset{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M}{\operatorname{argmin}} \frac{1}{M(M-1)}\left(\mathbf{v}^\top\mathbf{u} - M\sum_{i=1}^{M}\mathbf{v}_i^\top\mathbf{u}_i\right) \Leftrightarrow \underset{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M}{\operatorname{argmax}} M\sum_{i=1}^{M}\mathbf{v}_i^\top\mathbf{u}_i - \left(\sum_{i=1}^{M}\mathbf{v}_i\right)^\top\left(\sum_{i=1}^{M}\mathbf{u}_i\right)$$

or equivalently, as shown in Eq. (12) in the Appendix of (Sreenivasan et al., 2023),

$$\underset{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M}{\operatorname{argmax}} \quad \mathbf{v}_{\text{stack}}^\top \left(\left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)\otimes\mathbb{I}_d\right)\mathbf{u}_{\text{stack}}\,, \tag{13}$$

where $\mathbf{v}_{\text{stack}} = \left[\mathbf{v}_1^\top,\ldots,\mathbf{v}_M^\top\right]^\top \in \mathbb{R}^{Md\times 1}$, $\mathbf{u}_{\text{stack}} = \left[\mathbf{u}_1^\top,\ldots,\mathbf{u}_M^\top\right]^\top \in \mathbb{R}^{Md\times 1}$, $\mathbb{I}_M\in\mathbb{R}^{M\times M}$ the identity matrix, $\mathbf{1}_M\in\mathbb{R}^{M\times 1}$ a vector whose elements are all equal to 1 and $\otimes$ denotes the Kronecker product. The optimisation problem of Eq. (13) has been studied in (Lu & Steinerberger, 2022) and (Sreenivasan et al., 2023). We repeat the result here for completeness.

It is shown that the eigenvalues of $\left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)$ are equal to M with multiplicity $M-1$ and 0 with multiplicity 1, with corresponding eigenvectors $\mathbf{p}\in\mathbb{R}^M$ such that $\mathbf{p}^\top\mathbf{1}_M = 0$ and $\mathbf{p} = k\mathbf{1}_M$ respectively. Therefore, since the set of eigenvalues of the Kronecker product of two matrices contains all the possible pair-wise products of the eigenvalues of the individual matrices, the eigenvalues of $\left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)\otimes\mathbb{I}_d$ are M with multiplicity $M(M-1)$ and 0 with multiplicity M. Since the matrix of interest is symmetric, its singular values coincide with the absolute of its eigenvalues, and therefore $\|\left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)\otimes\mathbb{I}_d\|_2 = M$. Concluding:

$$\begin{aligned}
\mathbf{v}_{\text{stack}}^\top\left(\left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)\otimes\mathbb{I}_d\right)\mathbf{u}_{\text{stack}} &\overset{(e)}{\leq} \|\mathbf{v}_{\text{stack}}\|_2\|\left(\left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)\otimes\mathbb{I}_d\right)\mathbf{u}_{\text{stack}}\|_2 \\
&\overset{(f)}{\leq} \|\mathbf{v}_{\text{stack}}\|_2\|\left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)\otimes\mathbb{I}_d\|_2\|\mathbf{u}_{\text{stack}}\|_2 \\
&= (\sqrt{M})M\sqrt{M} = M^2,
\end{aligned} \tag{14}$$

where (e) follows from the Cauchy–Schwarz inequality, (f) from the definition of the spectral norm and the last equality from the fact that $\|\mathbf{u}_i\| = \|\mathbf{v}_i\| = 1, \forall i\in[M]$. Now, moving backwards for every inequality (a)-(e) we have that:

- (f) holds with equality **iff** $\mathbf{u}_{\text{stack}}$ is an eigenvector corresponding to the maximum eigenvalue, i.e. $\left(\left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)\otimes\mathbb{I}_d\right)\mathbf{u}_{\text{stack}} = M\mathbf{u}_{\text{stack}}$. But using a common property of the Kronecker product and the fact that $\text{vec}(\mathbf{U}) = \mathbf{u}_{\text{stack}}$, it follows that $\left(\left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)\otimes\mathbb{I}_d\right)\mathbf{u}_{\text{stack}} = \text{vec}\left(\mathbb{I}_d\mathbf{U}\left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)^\top\right) = \text{vec}\left([\mathbf{u}_1\ldots\mathbf{u}_M]\left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)^\top\right)$ and thus:

$$\begin{aligned}
\text{vec}\left([\mathbf{u}_1\ldots\mathbf{u}_M]\left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)^\top\right) = M\mathbf{u}_{\text{stack}} &\Leftrightarrow [\mathbf{u}_1^{(\ell)}\ldots\mathbf{u}_M^{(\ell)}]\left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)^\top = M[\mathbf{u}_1^{(\ell)}\ldots\mathbf{u}_M^{(\ell)}] \\
&\Leftrightarrow \left(M\mathbb{I}_M - \mathbf{1}_M\mathbf{1}_M^\top\right)[\mathbf{u}_1^{(\ell)},\ldots.\mathbf{u}_M^{\ell}]^\top = M[\mathbf{u}_1^{(\ell)},\ldots,\mathbf{u}_M^{(\ell)}]^\top \\
&\Leftrightarrow [\mathbf{u}_1^{(\ell)},\ldots,\mathbf{u}_M^{(\ell)}]\mathbf{1}_M = 0, \forall\ell\in[d] \\
&\Leftrightarrow \sum_{i=1}^{M}\mathbf{u}_i^* = \mathbf{u}^* = \mathbf{0}.
\end{aligned} \tag{15}$$

- (e) holds with equality **iff** $\mathbf{v}_{\text{stack}} = k\mathbf{u}_{\text{stack}}$ for any $k>0$. But since $\|\mathbf{v}_{\text{stack}}\| = \|\mathbf{u}_{\text{stack}}\| = \sqrt{M}$, then it must be that $k=1$ and therefore:

$$\mathbf{v}_i^* = \mathbf{u}_i^*,\ \forall i\in[M]. \tag{16}$$

- (d) holds with equality only if $\mathbf{u}_i = \mathbf{u}_i^*$ and $\mathbf{v}_i = \mathbf{v}_i^*$ when ϕ,ψ **are strictly increasing** (Condition (4)). If the latter doesn't hold, then we might obtain the optimum for other input values besides the ones that were found in the previous two conditions.

- (c) holds with equality **iff** $\frac{\mathbf{v}^\top \mathbf{u}_i - M\mathbf{v}_i^\top \mathbf{u}_i}{M-1} = c_1$ constant $\forall i \in [M]$ or if $\tilde{\psi}$ is linear. The former is already satisfied by the conditions (15), (16) and $c_1 = -\frac{M}{M-1}$. Thus, no extra condition arises here.

- (b) holds with equality only if $\mathbf{u}_i, \mathbf{v}_i$ minimise the arguments of ψ (which happens when the conditions of (a) are satisfied) when ψ **is strictly increasing** (Condition (4)). Again, if the latter doesn't hold, we might obtain the optimum for other input values besides those found by the rest of the conditions.

- (a) holds with equality if $(\mathbf{v}_j - \mathbf{v}_i)^\top \mathbf{u}_i = \tilde{c}_2$ constant $\forall i, j \in [M]$. If ϕ is **strictly convex** (Condition (4)), then this will be the only condition allowing equality to hold. Given condition (16), the former implies that $\mathbf{u}_j^\top \mathbf{u}_i = 1 + \tilde{c}_2 = c_2$ and since from condition (15) we have that $\mathbf{u} = 0$, then:

$$
\begin{aligned}
0 = \mathbf{u}^\top \mathbf{u} = \left(\sum_{i=1}^M \mathbf{u}_i^\top\right)\left(\sum_{i=1}^M \mathbf{u}_i\right) = \sum_{i=1}^M \mathbf{u}_i^\top \mathbf{u}_i + \sum_{\substack{i,j=1\\ j\neq i}}^M \mathbf{u}_i^\top \mathbf{u}_j \Leftrightarrow 0 = M + M(M-1)c_2 \\
\Leftrightarrow \mathbf{u}_i \mathbf{u}_j = -\frac{1}{M-1} \forall i \neq j \in [M]. \qquad (17)
\end{aligned}
$$

Conditions (15), (16), (17) prove that any point configuration that satisfies Eq. (3) is an optimum of Eq. (2) for the L_a loss and with the additional conditions of strict monotonicity of ϕ, ψ and strict convexity of ϕ we also obtain that these point configurations are the unique optima.

It remains to show that these optima can be indeed attained. In particular, Eq. (15) and Eq. (16) are easy to attain for any twin (identical) point configurations that are centred at the origin. Eq. (17), or more precisely the fact that all angles are equal, is exactly the definition of a *regular $M-1$-dimensional simplex inscribed in the sphere*. However, for the regular $M-1$-dimensional simplex to exist the ambient dimension must be at least as large as $M-1$, i.e.$d \geq M-1$, which justifies Condition (3).

Part II: L_b. Using a similar rationale for L_b we obtain:

$$
\sum_{j=1, j\neq i}^M \left(\phi\left((\mathbf{v}_j - \mathbf{v}_i)^\top \mathbf{u}_i\right) + \phi\left((\mathbf{u}_j - \mathbf{v}_i)^\top \mathbf{u}_i\right)\right) \overset{(a')}{\geq} 2(M-1)\phi\left(\frac{\mathbf{v}^\top \mathbf{u}_i - (2M-1)\mathbf{v}_i^\top \mathbf{u}_i + \mathbf{u}^\top \mathbf{u}_i - 1}{2(M-1)}\right),
$$

and subsequently:

$$
\begin{aligned}
L_\text{b}(\mathbf{U}, \mathbf{V}) &\overset{(b')}{\geq} \frac{1}{M}\sum_{i=1}^M \psi\left(2(M-1)\phi\left(\frac{\mathbf{v}^\top \mathbf{u}_i - (2M-1)\mathbf{v}_i^\top \mathbf{u}_i + \mathbf{u}^\top \mathbf{u}_i - 1}{2(M-1)}\right)\right) \\
&= \frac{1}{M}\sum_{i=1}^M \tilde{\psi}\left(\frac{\mathbf{v}^\top \mathbf{u}_i - (2M-1)\mathbf{v}_i^\top \mathbf{u}_i + \mathbf{u}^\top \mathbf{u}_i - 1}{2(M-1)}; 2(M-1)\right) \\
&\overset{(c')}{\geq} \psi\left(2(M-1)\phi\left(\frac{1}{M}\sum_{i=1}^M\left(\frac{\mathbf{v}^\top \mathbf{u}_i - (2M-1)\mathbf{v}_i^\top \mathbf{u}_i + \mathbf{u}^\top \mathbf{u}_i - 1}{2(M-1)}\right)\right)\right) \\
&= \psi\left(2(M-1)\phi\left(\frac{1}{2M(M-1)}\left(\mathbf{v}^\top \mathbf{u} - (2M-1)\sum_{i=1}^M\left(\mathbf{v}_i^\top \mathbf{u}_i\right) + \mathbf{u}^\top \mathbf{u} - M\right)\right)\right) \\
&\overset{(d')}{\geq} \psi\left(2(M-1)\phi\left(\frac{1}{2M(M-1)}\left(\mathbf{v}^{*\top} \mathbf{u}^* - (2M-1)\sum_{i=1}^M\left(\mathbf{v}_i^{*\top} \mathbf{u}_i^*\right) + \mathbf{u}^{*\top} \mathbf{u}^* - M\right)\right)\right),
\end{aligned}
$$

where once again (a') follows from Jensen's inequality, (b') follows from the fact that ψ is increasing, (c') follows from Jensen's inequality and the fact that $M-1 > 0, 2(M-1) > 0$ and (d') from the fact that ψ and ϕ are increasing and

$M-1>0, 2M(M-1)>0$. Again it suffices to solve the following optimisation problem.

$$
\begin{aligned}
&\underset{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M}{\operatorname{argmin}} \frac{1}{2M(M-1)}\left(\mathbf{v}^\top\mathbf{u}-(2M-1)\sum_{i=1}^M \mathbf{v}_i^\top\mathbf{u}_i+\mathbf{u}^\top\mathbf{u}-M\right)\\
&= \underset{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M}{\operatorname{argmax}} (2M-1)\sum_{i=1}^M \mathbf{v}_i^\top\mathbf{u}_i-(\mathbf{v}^\top+\mathbf{u}^\top)\mathbf{u}\\
&= \underset{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M}{\operatorname{argmax}} (2M-1)\sum_{i=1}^M \mathbf{v}_i^\top\mathbf{u}_i-(\mathbf{v}^\top+\mathbf{u}^\top)\mathbf{u}+(2M-1)\sum_{i=1}^M \mathbf{u}_i^\top\mathbf{u}_i-M(2M-1)\\
&= \underset{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M}{\operatorname{argmax}} (2M-1)\sum_{i=1}^M (\mathbf{v}_i^\top+\mathbf{u}_i^\top)\mathbf{u}_i-(\mathbf{v}^\top+\mathbf{u}^\top)\mathbf{u}\\
&= \underset{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M}{\operatorname{argmax}} (\mathbf{v}_{\text{stack}}+\mathbf{u}_{\text{stack}})^\top\left(\left((2M-1)\mathbb{I}_M-\mathbf{1}_M\mathbf{1}_M^\top\right)\otimes\mathbb{I}_d\right)\mathbf{u}_{\text{stack}}\,.
\end{aligned}
\tag{18}
$$

Similarly with (Lu & Steinerberger, 2022), we will find the eigendecomposition of $(2M-1)\mathbb{I}_M-\mathbf{1}_M\mathbf{1}_M^\top$. Select $\mathbf{p}\in\mathbb{R}^M$ with $\mathbf{p}^\top\mathbf{1}_M=0$. Then, $\left((2M-1)\mathbb{I}_M-\mathbf{1}_M\mathbf{1}_M^\top\right)\mathbf{p}=(2M-1)\mathbf{p}-\mathbf{1}_M\mathbf{1}_M^\top\mathbf{p}=(2M-1)\mathbf{p}$ and so we proved that $(2M-1)$ is an eigenvalue for all the vectors with $\mathbf{p}^\top\mathbf{1}_M=0$, i.e. with multiplicity $M-1$. Additionally, select $\mathbf{p}\in\mathbb{R}^M$ with $\mathbf{p}=k\mathbf{1}_M$. Then, $\left((2M-1)\mathbb{I}_M-\mathbf{1}_M\mathbf{1}_M^\top\right)\mathbf{p}=(2M-1)\mathbf{p}-\mathbf{1}_M\mathbf{1}_M^\top\mathbf{p}=k(2M-1)\mathbf{1}_M-kM\mathbf{1}_M=k(M-1)\mathbf{1}_M=(M-1)\mathbf{p}$ and so we proved that $(M-1)$ is an eigenvalue for all the vectors with $\mathbf{p}=k\mathbf{1}_M$, i.e. with multiplicity 1. Thus, as in Part I, $\|\left((2M-1)\mathbb{I}_M-\mathbf{1}_M\mathbf{1}_M^\top\right)\otimes\mathbb{I}_d\|_2=2M-1$. Concluding:

$$
\begin{aligned}
(\mathbf{u}_{\text{stack}}^\top+\mathbf{v}_{\text{stack}}^\top)\left(\left((2M-1)\mathbb{I}_M-\mathbf{1}_M\mathbf{1}_M^\top\right)\otimes\mathbb{I}_d\right)\mathbf{u}_{\text{stack}} &\overset{(e')}{\leq} \|\mathbf{u}_{\text{stack}}+\mathbf{v}_{\text{stack}}\|_2\|\left(\left((2M-1)\mathbb{I}_M-\mathbf{1}_M\mathbf{1}_M^\top\right)\otimes\mathbb{I}_d\right)\mathbf{u}_{\text{stack}}\|_2\\
&\overset{(f')}{\leq} \|\mathbf{u}_{\text{stack}}+\mathbf{v}_{\text{stack}}\|_2\|\left((2M-1)\mathbb{I}_M-\mathbf{1}_M\mathbf{1}_M^\top\right)\otimes\mathbb{I}_d\|_2\|\mathbf{u}_{\text{stack}}\|_2\\
&\overset{(g')}{\leq} (\|\mathbf{u}_{\text{stack}}\|_2+\|\mathbf{v}_{\text{stack}}\|_2)\|\left((2M-1)\mathbb{I}_M-\mathbf{1}_M\mathbf{1}_M^\top\right)\\
&\qquad\qquad\otimes\mathbb{I}_d\|_2\|\mathbf{u}_{\text{stack}}\|_2\\
&= (2\sqrt{M})(2M-1)\sqrt{M}=2(2M-1)M^2,
\end{aligned}
\tag{19}
$$

where again (e') follows from the Cauchy–Schwarz inequality, (f') from the definition of the spectral norm and (g') from the triangle inequality. Now, moving backwards for every inequality (a')-(g') as in Part I[3] we will obtain the same conditions as in Part I which will prove the desideratum for L_{b}.

Part III: L_{c}**.** Following the exact same steps as before we obtain: First for the inner summands,

$$
\begin{aligned}
&\sum_{j=1,j\neq i}^M \phi\left((\mathbf{u}_j-\mathbf{v}_i)^\top\mathbf{u}_i\right)\overset{(a'')}{\geq}(M-1)\phi\left(\frac{\sum_{j=1,j\neq i}^M\mathbf{u}_j^\top\mathbf{u}_i-(M-1)\mathbf{v}_i^\top\mathbf{u}_i}{M-1}\right)\\
&=(M-1)\phi\left(\frac{\mathbf{u}^\top\mathbf{u}_i-(M-1)\mathbf{v}_i^\top\mathbf{u}_i-1}{M-1}\right),
\end{aligned}
$$

[3] (g') holds with equality for the same conditions as with (f')

then for the total loss:

$$
\begin{aligned}
L_{\mathrm{c}}(\mathbf{U},\mathbf{V}) &\overset{(b'')}{\geq} \frac{1}{M}\sum_{i=1}^{M}\psi\left((M-1)\phi\left(\frac{\mathbf{u}^\top\mathbf{u}_i-(M-1)\mathbf{v}_i^\top\mathbf{u}_i-1}{M-1}\right)\right) \\
&= \frac{1}{M}\sum_{i=1}^{M}\tilde{\psi}\left(\frac{\mathbf{u}^\top\mathbf{u}_i-(M-1)\mathbf{v}_i^\top\mathbf{u}_i-1}{M-1};M-1\right) \\
&\overset{(c'')}{\geq} \psi\left((M-1)\phi\left(\frac{1}{M}\sum_{i=1}^{M}\left(\frac{\mathbf{u}^\top\mathbf{u}_i-(M-1)\mathbf{v}_i^\top\mathbf{u}_i-1}{M-1}\right)\right)\right) \\
&= \psi\left((M-1)\phi\left(\frac{1}{M(M-1)}\left(\mathbf{u}^\top\mathbf{u}-(M-1)\sum_{i=1}^{M}\left(\mathbf{v}_i^\top\mathbf{u}_i\right)-M\right)\right)\right) \\
&\overset{(d'')}{\geq} \psi\left((M-1)\phi\left(\frac{1}{M(M-1)}\left(\mathbf{u}^{*\top}\mathbf{u}^*-(M-1)\sum_{i=1}^{M}\left(\mathbf{v}_i^{*\top}\mathbf{u}_i^*\right)-M\right)\right)\right),
\end{aligned}
$$

and finally for the inner argument optimisation problem:

$$
\underset{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M}{\operatorname{argmin}}\frac{1}{M(M-1)}\left(\mathbf{u}^\top\mathbf{u}-(M-1)\sum_{i=1}^{M}\mathbf{v}_i^\top\mathbf{u}_i-M\right)=\underset{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M}{\operatorname{argmax}}(M-1)\sum_{i=1}^{M}\mathbf{v}_i^\top\mathbf{u}_i-\|\mathbf{u}\|^2. \tag{20}
$$

The last part of the proof here is slightly different. In particular, the two terms in the above equation can be maximised independently. For the first term, we have that each summand in the sum can be optimised independently and that:

$$
\mathbf{v}_i^\top\mathbf{u}_i \overset{(e'')}{\leq} \|\mathbf{v}_i\|\|\mathbf{u}_i\|=1, \tag{21}
$$

where again we used Cauchy–Schwarz. Now (e”) holds with equality, as before, iff $\mathbf{v}_i=\mathbf{u}_i$. For the second term, we need to minimise $\|\mathbf{u}\|^2$, which evidently happens iff $\mathbf{u}=\mathbf{0}$. Therefore, we arrived at the same conditions as in Part I and Part II, while the rest of the equalities in (a”)-(d”) while be satisfied as before. This concludes the proof.

□

The below corollary follows directly from Theorem B.1.

Corollary B.2. *Consider the following optimisation problem:*

$$
\underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} L_{\text{CL-sym}}\left(f_{\boldsymbol{\theta}}(\mathbf{X}),f_{\boldsymbol{\theta}}(\mathbf{Y})\right), \tag{22}
$$

*where $f_{\boldsymbol{\theta}}:\mathcal{X}\to\mathbb{S}^{d-1}$ is an encoder function parametrised by a tuple of parameters $\boldsymbol{\theta}$ and $L_{\text{CL-sym}}$ is a contrastive loss function defined as in Theorem B.1. Suppose that the conditions (1)-(3) set in Theorem B.1 hold. Further, suppose that (5) $\exists\,\boldsymbol{\theta}^*\in\Theta$ such that $f_{\boldsymbol{\theta}}$ achieves **simultaneously perfect alignment and perfect uniformity**, i.e. that:*

$$
f_{\boldsymbol{\theta}^*}(\mathbf{X})=f_{\boldsymbol{\theta}^*}(\mathbf{Y}) \quad \textit{and} \quad f_{\boldsymbol{\theta}^*}(\mathbf{X})\,\textit{form a regular } M-1 \textit{ simplex.} \tag{23}
$$

*Then, the optimisation problem of Eq. (22) obtains its optimal value for all $\boldsymbol{\theta}^*$ that satisfy Eq. (23). Additionally, if the condition (4) set in Theorem B.6 holds, then all the $\boldsymbol{\theta}^*$ that satisfy Eq. (23) are the **unique** optima.*

Corollary B.3. *The following mini-batch CL loss functions: $L_{\text{InfoNCE}}(\cdot,\cdot)$, $L_{\text{DHEL}}(\cdot,\cdot)$, $L_{\text{SimCLR}}(\cdot,\cdot)$, $L_{\text{DCL}}(\cdot,\cdot)$ have the **same unique minima** on the unit sphere when $1<M\leq d+1$, i.e. all the optimal solutions of Eq. (11) will satisfy the properties of Eq. (12).*

Proof. Recall from Eq. (10) that the above losses are special cases of $L_{\text{a}},L_{\text{b}},L_{\text{c}}$ with $\phi(x)=\exp(x/\tau)$, and $\psi(x)=\log(x)$ or $\psi(x)=\log(1+x)$. We know that for $\tau>0$, $\exp(x/\tau)$ is strictly increasing and strictly convex, while $\log(1+\alpha\exp(x/\tau))$ and $\log(\alpha\exp(x/\tau))=\log\alpha+x/\tau$ are strictly increasing and convex for $\alpha,\tau>0$. Therefore, all Conditions (1)-(4) of Theorem 4.1 are satisfied. □

B.2. Expected (True) Contrastive Losses and Asymptotic Behaviour

Lemma B.4. *The expectations of the following mini-batch contrastive loss functions: $L_{\text{InfoNCE}}, L_{\text{SimCLR}}, L_{\text{DCL}}, L_{\text{DHEL}}$ are the ones given in Tables 1 and 2.*

Proof. It is straightforward to see that the expectations of the mini-batch losses are as follows:

$$\begin{aligned}
\underset{(\mathbf{X},\mathbf{Y})\overset{\text{i.i.d}}{\sim}p_+^M}{\mathbb{E}}\left[L_{\text{a}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))\right] &= \underset{(\mathbf{X},\mathbf{Y})\overset{\text{i.i.d}}{\sim}p_+^M}{\mathbb{E}}\left[\frac{1}{M}\sum_{i=1}^{M}\psi\left(\sum_{j=1,j\neq i}^{M}\phi\left(\left(f_{\boldsymbol{\theta}}\left(\mathbf{y}_j\right)-f_{\boldsymbol{\theta}}\left(\mathbf{y}_i\right)\right)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}_i\right)\right)\right)\right] \\
&= \frac{1}{M}\sum_{i=1}^{M}\underset{(\mathbf{X},\mathbf{Y})\overset{\text{i.i.d}}{\sim}p_+^M}{\mathbb{E}}\left[\psi\left(\sum_{j=1,j\neq i}^{M}\phi\left(\left(f_{\boldsymbol{\theta}}\left(\mathbf{y}_j\right)-f_{\boldsymbol{\theta}}\left(\mathbf{y}_i\right)\right)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}_i\right)\right)\right)\right] \\
&= \frac{1}{M}\sum_{i=1}^{M}\underset{\substack{(\mathbf{x}_i,\mathbf{y}_i)\sim p_+ \\ \{\mathbf{y}_j\}_{j=1}^{M-1}\overset{\text{i.i.d}}{\sim}p}}{\mathbb{E}}\left[\psi\left(\sum_{j=1,j\neq i}^{M}\phi\left(\left(f_{\boldsymbol{\theta}}\left(\mathbf{y}_j\right)-f_{\boldsymbol{\theta}}\left(\mathbf{y}_i\right)\right)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}_i\right)\right)\right)\right] \\
&= \underset{\substack{(\mathbf{x},\mathbf{y})\sim p_+ \\ \{\mathbf{y}_j\}_{j=1}^{M-1}\overset{\text{i.i.d}}{\sim}p}}{\mathbb{E}}\left[\psi\left(\sum_{j=1}^{M-1}\phi\left(\left(f_{\boldsymbol{\theta}}\left(\mathbf{y}_j\right)-f_{\boldsymbol{\theta}}\left(\mathbf{y}\right)\right)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}\right)\right)\right)\right] \qquad (24)
\end{aligned}$$

$$\begin{aligned}
\underset{(\mathbf{X},\mathbf{Y})\overset{\text{i.i.d}}{\sim}p_+^M}{\mathbb{E}}\left[L_{\text{b}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))\right] = \underset{\substack{(\mathbf{x},\mathbf{y})\sim p_+ \\ \{(\mathbf{x}_j,\mathbf{y}_j)\}_{j=1}^{M-1}\overset{\text{i.i.d}}{\sim}p_+}}{\mathbb{E}}\Bigg[\psi\Bigg(\sum_{j=1}^{M-1}\phi\left(\left(f_{\boldsymbol{\theta}}\left(\mathbf{y}_j\right)-f_{\boldsymbol{\theta}}\left(\mathbf{y}\right)\right)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}\right)\right) \\
+\phi\left(\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}_j\right)-f_{\boldsymbol{\theta}}\left(\mathbf{y}\right)\right)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}\right)\right)\Bigg)\Bigg] \qquad (25)
\end{aligned}$$

$$\underset{(\mathbf{X},\mathbf{Y})\overset{\text{i.i.d}}{\sim}p_+^M}{\mathbb{E}}\left[L_{\text{c}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))\right] = \underset{\substack{(\mathbf{x},\mathbf{y})\sim p_+ \\ \{\mathbf{x}_j\}_{j=1}^{M-1}\overset{\text{i.i.d}}{\sim}p}}{\mathbb{E}}\left[\psi\left(\sum_{j=1}^{M-1}\phi\left(\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}_j\right)-f_{\boldsymbol{\theta}}\left(\mathbf{y}\right)\right)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}\right)\right)\right)\right] \qquad (26)$$

Expanding Eq. (24), (25) (twice) and (26) for $L_{\text{InfoNCE}}, L_{\text{SimCLR}}, L_{\text{DCL}}, L_{\text{DHEL}}$ respectively, we obtain:

$$\begin{aligned}
\underset{(\mathbf{X},\mathbf{Y})\overset{\text{i.i.d}}{\sim}p_+^M}{\mathbb{E}}\left[L_{\text{InfoNCE}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))\right] &= \underset{(\mathbf{u},\mathbf{v})\sim f_{\boldsymbol{\theta}\#}p_+}{\mathbb{E}}\left[-\mathbf{v}^\top\mathbf{u}/\tau\right] + \underset{\substack{(\mathbf{u},\mathbf{v})\sim f_{\boldsymbol{\theta}\#}p_+ \\ \mathbf{V}'\overset{\text{i.i.d}}{\sim}f_{\boldsymbol{\theta}\#}p^{M-1}}}{\mathbb{E}}\left[\log\left(e^{\mathbf{v}^\top\mathbf{u}/\tau}+\sum_{j=1}^{M-1}e^{\mathbf{u}^\top\mathbf{v}'_j/\tau}\right)\right] \\
\underset{(\mathbf{X},\mathbf{Y})\overset{\text{i.i.d}}{\sim}p_+^M}{\mathbb{E}}\left[L_{\text{SimCLR}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))\right] &= \underset{(\mathbf{u},\mathbf{v})\sim f_{\boldsymbol{\theta}\#}p_+}{\mathbb{E}}\left[-\mathbf{v}^\top\mathbf{u}/\tau\right] + \underset{\substack{(\mathbf{u},\mathbf{v})\sim f_{\boldsymbol{\theta}\#}p_+ \\ \hat{\mathbf{U}}\overset{\text{i.i.d}}{\sim}f_{\boldsymbol{\theta}\#}p_+^{M-1}}}{\mathbb{E}}\left[\log\left(e^{\mathbf{v}^\top\mathbf{u}/\tau}+\sum_{j=1}^{2M-2}e^{\hat{\mathbf{u}}_j^\top\mathbf{u}}\right)\right] \\
\underset{(\mathbf{X},\mathbf{Y})\overset{\text{i.i.d}}{\sim}p_+^M}{\mathbb{E}}\left[L_{\text{DCL}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))\right] &= \underset{(\mathbf{u},\mathbf{v})\sim f_{\boldsymbol{\theta}\#}p_+}{\mathbb{E}}\left[-\mathbf{v}^\top\mathbf{u}/\tau\right] + \underset{\substack{\mathbf{u}\sim f_{\boldsymbol{\theta}\#}p \\ \hat{\mathbf{U}}\overset{\text{i.i.d}}{\sim}f_{\boldsymbol{\theta}\#}p_+^{M-1}}}{\mathbb{E}}\left[\log\left(\sum_{j=1}^{2M-2}e^{\hat{\mathbf{u}}_j^\top\mathbf{u}/\tau}\right)\right] \\
\underset{(\mathbf{X},\mathbf{Y})\overset{\text{i.i.d}}{\sim}p_+^M}{\mathbb{E}}\left[L_{\text{DHEL}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))\right] &= \underset{(\mathbf{u},\mathbf{v})\sim f_{\boldsymbol{\theta}\#}p_+}{\mathbb{E}}\left[-\mathbf{v}^\top\mathbf{u}/\tau\right] + \underset{\substack{\mathbf{u}\sim f_{\boldsymbol{\theta}\#}p \\ \mathbf{U}'\overset{\text{i.i.d}}{\sim}f_{\boldsymbol{\theta}\#}p^{M-1}}}{\mathbb{E}}\left[\log\left(\sum_{j=1}^{M-1}e^{\mathbf{u}^\top\mathbf{u}'_j/\tau}\right)\right]
\end{aligned} \qquad (27)$$

□

Proposition B.5. *The expectations of the following batch-level contrastive loss functions: $L_{\text{InfoNCE}}(\cdot,\cdot)$, $L_{\text{DHEL}}(\cdot,\cdot)$, $L_{\text{SimCLR}}(\cdot,\cdot)$, $L_{\text{DCL}}(\cdot,\cdot)$ have the **same asymptotic behaviour** when normalised by appropriate normalising constants, i.e.*

when $M \to \infty$ we have that:

$$\begin{aligned}
\lim_{M\to\infty} \mathop{\mathbb{E}}_{(\mathbf{X},\mathbf{Y})\overset{i.i.d}{\sim} p_+^M} \left[L_{\text{InfoNCE}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))\right] - \log(M-1) &= \lim_{M\to\infty} \mathop{\mathbb{E}}_{(\mathbf{X},\mathbf{Y})\overset{i.i.d}{\sim} p_+^M} \left[L_{\text{SimCLR}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))\right] - \log(2M-2) \\
&= \lim_{M\to\infty} \mathop{\mathbb{E}}_{(\mathbf{X},\mathbf{Y})\overset{i.i.d}{\sim} p_+^M} \left[L_{\text{DCL}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))\right] - \log(2M-2) \\
&= \mathop{\mathbb{E}}_{(\mathbf{X},\mathbf{Y})\overset{i.i.d}{\sim} p_+^M} \left[L_{\text{DHEL}}(f_{\boldsymbol{\theta}}(\mathbf{X}), f_{\boldsymbol{\theta}}(\mathbf{Y}))\right] - \log(M-1) \\
&= \mathop{\mathbb{E}}_{(\mathbf{x}',\mathbf{y}')\sim p_+} \left[-f_{\boldsymbol{\theta}}\left(\mathbf{y}'\right)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)\right] \\
&\qquad + \mathop{\mathbb{E}}_{\mathbf{x}'\sim p} \left[\log \mathop{\mathbb{E}}_{\mathbf{x}\sim p} \left(e^{f_{\boldsymbol{\theta}}(\mathbf{x})^\top f_{\boldsymbol{\theta}}(\mathbf{x}')}\right)\right]
\end{aligned} \tag{28}$$

Proof. For a fixed $\mathbf{x}'$, since $\lim_{M\to\infty} \frac{1}{M-1} e^{f_{\boldsymbol{\theta}}\left(\mathbf{y}'\right)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)} \lim_{M\to\infty} \frac{1}{2M-2} e^{f_{\boldsymbol{\theta}}\left(\mathbf{y}'\right)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)} = 0$, because $e^{f_{\boldsymbol{\theta}}\left(\mathbf{y}'\right)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)}$ is bounded and since

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{X}\sim p^{M-1}} \left[\frac{1}{M-1} \sum_{j=1}^{M-1} e^{f_{\boldsymbol{\theta}}(\mathbf{y}_j)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)}\right] &= \mathop{\mathbb{E}}_{\mathbf{X}\sim p^{M-1}} \left[\frac{1}{M-1} \sum_{j=1}^{M-1} e^{f_{\boldsymbol{\theta}}(\mathbf{x}_j)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)}\right] \\
&= \mathop{\mathbb{E}}_{\hat{\mathbf{X}}\sim p_+^{M-1}} \left[\frac{1}{2M-2} \sum_{j=1}^{2M-2} e^{f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_j)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)}\right] = \mathop{\mathbb{E}}_{\mathbf{x}\sim p} \left[e^{f_{\boldsymbol{\theta}}(\mathbf{x})^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)}\right],
\end{aligned}$$

then with probability 1 over the respective sample spaces, due to the strong law of large numbers, it holds that:

$$\begin{aligned}
\lim_{M\to\infty} \frac{1}{M} e^{f_{\boldsymbol{\theta}}\left(\mathbf{y}'\right)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)} + \frac{1}{M} \sum_{j=1}^{M-1} e^{f_{\boldsymbol{\theta}}(\mathbf{y}_j)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)} &= \lim_{M\to\infty} \frac{1}{2M-2} e^{f_{\boldsymbol{\theta}}\left(\mathbf{y}'\right)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)} + \frac{1}{2M-2} \sum_{j=1}^{2M-2} e^{f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_j)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)} \\
&= \lim_{M\to\infty} \frac{1}{2M-2} \sum_{j=1}^{2M-2} e^{f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_j)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)} \\
&= \lim_{M\to\infty} \frac{1}{M-1} \sum_{j=1}^{M-1} e^{f_{\boldsymbol{\theta}}(\mathbf{x}_j)^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)} = \mathop{\mathbb{E}}_{\mathbf{x}\sim p} \left[e^{f_{\boldsymbol{\theta}}(\mathbf{x})^\top f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)}\right].
\end{aligned}$$

Now the desideratum follows directly using the same steps as in the proof of Theorem 1 in (Wang & Isola, 2020). Briefly, the same limit holds for the $\log$ (continuous function) of the above quantities due to the Continuous Mapping Theorem, and therefore when taking the limit of each loss variant (after first subtracting the right normalisation constant $M-1$ or $2M-2$), since the quantities inside the expectation are bounded, we can invoke the Dominated Convergence Theorem and switch the limit with the expectation, thus arriving at the desideratum. □

B.3. Minima of Mini-Batch Kernel Contrastive Losses

Theorem B.6. *Consider the following optimisation problem:*

$$\operatorname*{argmin}_{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M} L_{\text{KCL-sym}}(\mathbf{U},\mathbf{V}), \tag{29}$$

with $L_{\text{KCL-sym}}(\mathbf{U},\mathbf{V}) = \frac{1}{2}\left(L_{\text{KCL}}(\mathbf{U},\mathbf{V}) + L_{\text{KCL}}(\mathbf{V},\mathbf{U})\right)$*, where* $\mathbb{S}^{d-1} = \{\mathbf{u} \in \mathbb{R}^d \mid \|\mathbf{u}\|_2 = 1\}$ *a unit sphere of* d *dimensions,* $\mathbf{U},\mathbf{V}$ *are tuples of* M *vectors on the unit sphere and* $L_{\text{KCL}}(\cdot,\cdot)$ *is a kernel loss function of the form:*

$$L_{\text{KCL}}(\mathbf{U},\mathbf{V}) = -\frac{1}{M} \sum_{i=1}^{M} K_A(\mathbf{u}_i,\mathbf{v}_i) + \gamma \frac{1}{M(M-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{M} K_U(\mathbf{u}_i,\mathbf{u}_j), \tag{30}$$

where $K_A(\mathbf{x},\mathbf{y}) = \kappa_A(\|\mathbf{x}-\mathbf{y}\|^2)$ *and* $K_U(\mathbf{x},\mathbf{y}) = \kappa_U(\|\mathbf{x}-\mathbf{y}\|^2)$ *with* $\kappa_A, \kappa_U : (0,4] \to \mathbb{R}$*, the limits* $\lim_{x\to 0^+}\kappa_A(x)$*,* $\lim_{x\to 0^+}\kappa_U(x)$ *exist and are bounded in both cases and* $\gamma > 0$*. Further, for case (a)* $1 < M \leq d+1$*, suppose the following conditions: (1)* k_A *is* ***decreasing*** *and (2)* k_U *is* ***decreasing and convex****. Then, the optimisation problem of Eq. (2) obtains its optimal value* $(\mathbf{U}^*, \mathbf{V}^*)$ *when:*

$$\mathbf{U}^* = \mathbf{V}^* \quad \textit{and} \quad \mathbf{U}^* = [\mathbf{u}_1^*, \ldots, \mathbf{u}_M^*] \ \textit{form a regular } M-1 \textit{ simplex centered at the origin.} \tag{31}$$

Additionally, (3) if κ_A *is* ***strictly decreasing*** *and* κ_U *is* ***strictly decreasing and strictly convex*** *then all the* $(\mathbf{U}^*, \mathbf{V}^*)$ *that satisfy Eq. (3) are the* ***unique*** *optima. For case (b)* $M = 2d$*, suppose again that* k_A *is* ***decreasing*** *and that (4)* k_U *is* ***completely monotone****. Then Eq. (2) obtains its optimal value when:*

$$\mathbf{U}^* = \mathbf{V}^* \quad \textit{and} \quad \mathbf{U}^* = [\mathbf{u}_1^*, \ldots, \mathbf{u}_M^*] \ \textit{form a cross-polytope.} \tag{32}$$

Proof. Our strategy in this proof will be to analyse the two terms independently and show that it is possible to simultaneously attain their minima, i.e. for the same input arguments. Let us start with the first term.

Part I: Alignment term $-\frac{1}{M}\sum_{i=1}^{M} K_A(\mathbf{u}_i, \mathbf{v}_i)$.

Observe that in this term, the summands are independent of each other, so we can optimise each of them independently. Let $(\mathbf{u}_i^*, \mathbf{v}_i^*)$ be a minimiser for $-K_A(\mathbf{u}_i, \mathbf{v}_i) = -\kappa_A(\|\mathbf{u}_i - \mathbf{v}_i\|^2)$. Therefore,

$$-\kappa_A(\|\mathbf{u}_i - \mathbf{v}_i\|^2) \geq -\kappa_A(\|\mathbf{u}_i^* - \mathbf{v}_i^*\|^2) \Leftrightarrow \kappa_A(\|\mathbf{u}_i - \mathbf{v}_i\|^2) \leq \kappa_A(\|\mathbf{u}_i^* - \mathbf{v}_i^*\|^2) \overset{(a)}{\Leftrightarrow} \|\mathbf{u}_i - \mathbf{v}_i\|^2 \geq \|\mathbf{u}_i^* - \mathbf{v}_i^*\|^2,$$

where *(a) follows from the fact that* κ_A *is decreasing*, and thus:

$$\operatorname*{argmin}_{\mathbf{u}_i,\mathbf{v}_i\in\mathbb{S}^{d-1}} - K_A(\mathbf{u}_i, \mathbf{v}_i) \supseteq \operatorname*{argmin}_{\mathbf{u}_i,\mathbf{v}_i\in\mathbb{S}^{d-1}} \|\mathbf{u}_i - \mathbf{v}_i\|^2 = \{(\mathbf{u}_i^*, \mathbf{v}_i^*) \mid \mathbf{u}_i^* = \mathbf{v}_i^*\}.$$

So, since the above holds $\forall i \in \{1,\ldots,M\}$ we obtain $-K_A(\mathbf{u}_i, \mathbf{v}_i) \geq -K_A(\mathbf{u}_i^*, \mathbf{v}_i^*) \Leftrightarrow -\frac{1}{M}\sum_{i=1}^{M} K_A(\mathbf{u}_i, \mathbf{v}_i) \geq -\frac{1}{M}\sum_{i=1}^{M} K_A(\mathbf{u}_i^*, \mathbf{v}_i^*), \forall\, \mathbf{U}, \mathbf{V} \in (\mathbb{S}^{d-1})^M$ and thus:

$$\operatorname*{argmin}_{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M} -\frac{1}{M}\sum_{i=1}^{M} K_A(\mathbf{u}_i, \mathbf{v}_i) \supseteq \{(\mathbf{U}^*, \mathbf{V}^*) \mid \mathbf{u}_i^* = \mathbf{v}_i^*, \forall i \in \{1,\ldots,M\}\}.$$

Part II: Uniformity term $\frac{\gamma}{M(M-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{M} K_U(\mathbf{u}_i, \mathbf{u}_j)$.

First, observe that optimising the second term depends only on $\mathbf{U}$. Further, note that finding its minimiser is a classical hyperspherical energy minimisation problem, which is straightforward when κ_U is convex and decreasing and $1 < M \leq d+1$. We can invoke Theorem 1 from (Liu et al., 2023), which asserts that if the aforementioned conditions hold, then:

$$\operatorname*{argmin}_{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M} \frac{\gamma}{M(M-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{M} K_U(\mathbf{u}_i, \mathbf{u}_j) \supseteq \{(\mathbf{U}^*, \mathbf{V}^*) \mid \mathbf{U}^*\text{: regular } M-1 \text{ simplex on } \mathbb{S}^{d-1} \text{ centered at the origin}\}.$$

Similarly, for the case $M = 2d$, we can invoke Theorem 5.7.2 (Borodachov et al., 2019) or Theorem 2 (Liu et al., 2023), which imply that:

$$\operatorname*{argmin}_{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M} \frac{\gamma}{M(M-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{M} K_U(\mathbf{u}_i, \mathbf{u}_j) \supseteq \{(\mathbf{U}^*, \mathbf{V}^*) \mid \mathbf{U}^*\text{: cross-polytope}\},$$

when the function $\tilde{k}_U(x) = k_U(2-2x)$, i.e. the corresponding function expressing the kernel w.r.t. the inner product, is *absolutely monotone* in $[-1,1)$. But (Borodachov et al., 2019) show that this equivalent to k_U being completely monotone in $(0,4]$ (Condition (4)).

In the intersection of the two sets of minimisers, both terms will be minimised, i.e. $-\frac{1}{M}\sum_{i=1}^{M} K_A(\mathbf{u}_i,\mathbf{v}_i) \geq -\frac{1}{M}\sum_{i=1}^{M} K_A(\mathbf{u}_i^*,\mathbf{v}_i^*)$ and $\frac{\gamma}{M(M-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{M} K_U(\mathbf{u}_i,\mathbf{u}_j) \geq \frac{\gamma}{M(M-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{M} K_U(\mathbf{u}_i^*,\mathbf{u}_j^*)$, $\forall\, \mathbf{U},\mathbf{V} \in (\mathbb{S}^{d-1})^M$. Therefore, the intersection is a minimiser of the total objective:

$$\{(\mathbf{U}^*,\mathbf{V}^*) \mid \mathbf{U}^* = \mathbf{V}^*\text{: regular } M-1 \text{ simplex on } \mathbb{S}^{d-1} \text{ centered at the origin}\} \subseteq \underset{\mathbf{U},\mathbf{V}\in(\mathbb{S}^{d-1})^M}{\operatorname{argmin}} L_{\text{KCL}}(\mathbf{U},\mathbf{V}),$$

and similarly for the cross-polytope. It is easy to see that the same set will be a minimiser of $L_{\text{KCL}}(\mathbf{V},\mathbf{U})$ and thus it is also a minimiser of $L_{\text{KCL-sym}}(\mathbf{V},\mathbf{U})$.

Finally, if κ_A is strictly decreasing, then (a) holds with equality only when $\mathbf{u}_i = \mathbf{v}_i$, while if κ_U is strictly convex and strictly decreasing then, again by Theorem 1 in (Liu et al., 2022), we know that the regular $M-1$ simplex is the only minimiser of the second term, and thus Eq. (3) is the unique minimiser of the kernel contrastive loss for $1 < M \leq d+1$.

□

B.4. Expected (True) Kernel Contrastive Losses

Proposition B.7. *The expectation of the mini-batch kernel contrastive loss functions $L_{\text{KCL}}(\cdot,\cdot)$ is **independent of the size of the batch** and therefore equal to the asymptotic expected loss. In other words, mini-batch kernel loss is an **unbiased estimator** of the asymptotic expected loss and in particular, we have that:*

$$\underset{(\mathbf{X},\mathbf{Y})\sim p_+^M}{\mathbb{E}}\left[L_{\text{KCL-sym}}\left(f_{\boldsymbol{\theta}}\left(\mathbf{X}\right), f_{\boldsymbol{\theta}}\left(\mathbf{Y}\right)\right)\right] = -\underset{(\mathbf{x},\mathbf{y})\sim p_+}{\mathbb{E}}\left[K_A\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}\right), f_{\boldsymbol{\theta}}\left(\mathbf{y}\right)\right)\right] + \gamma \underset{\substack{\mathbf{x}\sim p\\ \mathbf{x}'\sim p}}{\mathbb{E}}\left[K_U\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}\right), f_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)\right)\right]. \tag{33}$$

*If (1) κ_A is (strictly) decreasing and if (2) $\exists\,\boldsymbol{\theta}^*$ such that $\mathbb{P}_{(\mathbf{x},\mathbf{y})\sim p_+}\left[f_{\boldsymbol{\theta}}(\mathbf{x}) = f_{\boldsymbol{\theta}}(\mathbf{y})\right] = 1$, then the set of $\boldsymbol{\theta}^*$ for which (2) holds are (unique) minimisers of the first term of Eq. (9). Additionally, if (3) -$\kappa_U^{(1)}$ (first derivative) is **strictly completely monotone** in $(0,4]$, (4) the expectation defined in the l.h.s. of Eq. (9) is finite and (5) $\exists\,\boldsymbol{\theta}^*$ such that the pushforward measure $f_{\boldsymbol{\theta}\#}p = U(\mathbb{S}^{d-1})$, then $\boldsymbol{\theta}^*$ is a unique minimiser of the second term of Eq. (9). Finally, if (6) $\exists\,\boldsymbol{\theta}^*$ such that conditions (2) and (3) can be satisfied simultaneously, then $\boldsymbol{\theta}^*$ is a unique minimiser of Eq. (9).*

Proof. The first part of the proposition is obvious since:

$$\begin{aligned}
\underset{(\mathbf{X},\mathbf{Y})\sim p_+^M}{\mathbb{E}}\left[L_{\text{KCL}}\left(f_{\boldsymbol{\theta}}\left(\mathbf{X}\right), f_{\boldsymbol{\theta}}\left(\mathbf{Y}\right)\right)\right] &= \underset{(\mathbf{X},\mathbf{Y})\sim p_+^M}{\mathbb{E}}\Bigg[-\frac{1}{M}\sum_{i=1}^{M} K_A(f_{\boldsymbol{\theta}}(\mathbf{x}_i), f_{\boldsymbol{\theta}}(\mathbf{y}_i)) \\
&\qquad\qquad + \frac{\gamma}{M(M-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{M} K_U(f_{\boldsymbol{\theta}}(\mathbf{x}_i), f_{\boldsymbol{\theta}}(\mathbf{y}_j))\Bigg] \\
&= \underset{(\mathbf{X},\mathbf{Y})\sim p_+^M}{\mathbb{E}}\left[-\frac{1}{M}\sum_{i=1}^{M} K_A(f_{\boldsymbol{\theta}}(\mathbf{x}_i), f_{\boldsymbol{\theta}}(\mathbf{y}_i))\right] \\
&\qquad\qquad + \underset{(\mathbf{X},\mathbf{Y})\sim p_+^M}{\mathbb{E}}\left[\frac{\gamma}{M(M-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{M} K_U(f_{\boldsymbol{\theta}}(\mathbf{x}_i), f_{\boldsymbol{\theta}}(\mathbf{y}_j))\right] \\
&= -\frac{1}{M}\sum_{i=1}^{M} \underset{(\mathbf{x}_i,\mathbf{y}_i)\sim p_+}{\mathbb{E}}\left[K_A(f_{\boldsymbol{\theta}}(\mathbf{x}), f_{\boldsymbol{\theta}}(\mathbf{y}))\right] \\
&\qquad\qquad + \frac{\gamma}{M(M-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{M} \underset{\substack{\mathbf{x}_i\sim p\\ \mathbf{x}_j\sim p}}{\mathbb{E}}\left[K_U(f_{\boldsymbol{\theta}}(\mathbf{x}_i), f_{\boldsymbol{\theta}}(\mathbf{x}_j))\right] \\
&= -\underset{(\mathbf{x},\mathbf{y})\sim p_+}{\mathbb{E}}\left[K_A(f_{\boldsymbol{\theta}}(\mathbf{x}), f_{\boldsymbol{\theta}}(\mathbf{y}))\right] + \gamma \underset{\substack{\mathbf{x}\sim p\\ \mathbf{x}'\sim p}}{\mathbb{E}}\left[K_U(f_{\boldsymbol{\theta}}(\mathbf{x}), f_{\boldsymbol{\theta}}(\mathbf{x}'))\right]
\end{aligned}$$

Regarding the minimiser of the alignment term, we have that for every $(\mathbf{x}, \mathbf{y})$:

$$\|f_{\boldsymbol{\theta}}(\mathbf{x}) - f_{\boldsymbol{\theta}}(\mathbf{y})\|^2 \geq 0 \Longleftrightarrow -\kappa_A\left(\|f_{\boldsymbol{\theta}}(\mathbf{x}) - f_{\boldsymbol{\theta}}(\mathbf{y})\|^2\right) \geq -\kappa_A(0),$$

since κ_A is decreasing. It follows that $-\kappa_A(0) \leq \underset{(\mathbf{x},\mathbf{y})\sim p_+}{\mathbb{E}}\left[K_A\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}\right), f_{\boldsymbol{\theta}}\left(\mathbf{y}\right)\right)\right]$. But, we know that $\|f_{\boldsymbol{\theta}^*}(\mathbf{x}) - f_{\boldsymbol{\theta}^*}(\mathbf{y})\|^2 = 0$ almost surely and thus also $K_A\left(f_{\boldsymbol{\theta}^*}\left(\mathbf{x}\right), f_{\boldsymbol{\theta}^*}\left(\mathbf{y}\right)\right) = \kappa_A\left(\|f_{\boldsymbol{\theta}^*}(\mathbf{x}) - f_{\boldsymbol{\theta}^*}(\mathbf{y})\|^2\right) = \kappa_A(0)$ almost surely. Therefore, $-\underset{(\mathbf{x},\mathbf{y})\sim p_+}{\mathbb{E}}\left[K_A\left(f_{\boldsymbol{\theta}^*}\left(\mathbf{x}\right), f_{\boldsymbol{\theta}^*}\left(\mathbf{y}\right)\right)\right] \leq \underset{(\mathbf{x},\mathbf{y})\sim p_+}{\mathbb{E}}\left[K_A\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}\right), f_{\boldsymbol{\theta}}\left(\mathbf{y}\right)\right)\right]$, which proves that $\boldsymbol{\theta}^*$ is a minimiser. If κ_A is strictly decreasing (Condition (1)), then equality holds only if the set $(\mathbf{x}, \mathbf{y})$ for which $\|f_{\boldsymbol{\theta}}(\mathbf{x}) - f_{\boldsymbol{\theta}}(\mathbf{y})\|^2 > 0$ has measure zero under p_+, i.e. for $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ (unique minimiser).

Regarding the minimiser of the uniformity term, we can invoke Theorem 6.2.1 from (Borodachov et al., 2019) (restated as Lemma 2 in (Wang & Isola, 2020)), that under conditions (3) and (4) asserts that the unique measure $\mu \in \mathcal{M}(\mathbb{S}^{d-1})$ minimising $\underset{\substack{\mathbf{u}\sim\mu \\ \mathbf{v}\sim\mu}}{\mathbb{E}}\left[K_U\left(\mathbf{u}, \mathbf{v}\right)\right]$ is the uniform measure on the sphere $U(\mathbb{S}^{d-1})$. Then, condition (5) simply guarantees the existence of a parameter $\boldsymbol{\theta}$ so that this measure is attainable. □

C. Experimental Details

In the following section, we provide a detailed description of the experimental setup.

C.1. Detailed Sampling Process.

The distributions of interest p, p_+ throughout all the experiments are formed from the following two-step sampling process. First, we sample a datapoint $\mathbf{x}_{\text{init}} \in \mathcal{X}$ from an (unknown) *initial distribution* p_{init} on $\mathcal{X}$ (i.e. the one from which we sample the datapoints in our dataset) and subsequently we independently sample a *transformation operator* $T : \mathcal{X} \to \mathcal{X}$ from a (usually known) distribution p_T on a space of available transformations $\mathcal{T}$. Then, p is the distribution of the datapoint $T(x_{\text{init}})$ and the p.d.f. is given by $p(\mathbf{x}) = \int_{T\in\mathcal{T}} p_{\text{init}}(x)p_T(T)dT$.[4] Additionally, we sample positive pairs by first sampling a datapoint $\mathbf{x}_{\text{init}}$ and then transforming it by two independently sampled operators T_1, T_2. We define the distribution of positive pairs as the distribution of the tuples $\left(T_1\left(\mathbf{x}_{\text{init}}\right), T_2\left(\mathbf{x}_{\text{init}}\right)\right)$ and the p.d.f. is given by $p_+(\mathbf{x}, \mathbf{y}) = \int_{\substack{T_1, T_2\in\mathcal{T}, x_{\text{init}}\in\mathcal{X} \\ y=T_2(x_{\text{init}}), x=T_1(x_{\text{init}})}} p_{\text{init}}(x_{\text{init}})p_T(T_1)p_T(T_2)dx_{\text{init}}dT_1dT_2$. The transformation operators encode the symmetries of the data, i.e. it is expected that the downstream tasks will be invariant to them.

In practice, batches of data are sampled from a fixed finite dataset of $N > M$ samples $\mathcal{D} \sim p_{\text{init}}^N$ as follows: M samples are obtained by sampling uniformly at random from the dataset, i.e. $(\mathbf{x}_{\text{init},i})_{i=1}^M \sim U(\mathcal{D}) = \tilde{p}_{\text{init}}$ and $2M$ transformations $T_{1,i}, T_{2,i}$ are independently sampled from p_T, resulting in a batch of M positive pairs $\left((\mathbf{x}_i, \mathbf{y}_i)\right)_{i=1}^M \sim p_+^M$, where $\mathbf{x}_i = T_{1,i}(\mathbf{x}_{\text{init},i})$, $\mathbf{y}_i = T_{2,i}(\mathbf{x}_{\text{init},i})$.

C.2. Performance Metrics.

Below we provide more details on the metrics used in Figure 2, Section 7.2 of the main paper.

- **Alignment**. It estimates the expected L2 distance between a pair of positive samples:

$$L_{\text{alignment}}(f_{\#}p_+) = \underset{(\mathbf{u},\mathbf{v})\sim f_{\#}p_+}{\mathbb{E}}\left[\|\mathbf{u} - \mathbf{v}\|_2^2\right] \approx \frac{1}{M}\sum_{i=1}^{M}\|f(\mathbf{x}_i) - f(\mathbf{y}_i)\|_2^2 = \hat{L}_{\text{alignment}}(f, \mathbf{X}, \mathbf{Y}) \tag{34}$$

- **Uniformity**. The logarithm of an estimation of the expected pairwise Gaussian potential as in (Wang & Isola, 2020):

$$\begin{aligned} E_{\text{uniformity}}(f_{\#}p; t) &= \underset{\substack{\mathbf{u}\sim f_{\#}p \\ \mathbf{u}'\sim f_{\#}p}}{\mathbb{E}}\left[e^{-t\|\mathbf{u}-\mathbf{u}'\|_2^2}\right] \approx \frac{1}{M(M-1)}\sum_{\substack{i,j=1 \\ j\neq i}}^{M} e^{-t\|f(\mathbf{x}_i)-f(\mathbf{x}_j)\|_2^2} = \hat{E}_{\text{uniformity}}(f, \mathbf{X}; t) \\ L_{\text{uniformity}}(f_{\#}p; t) &= \log E_{\text{uniformity}}(f_{\#}p; t) \approx \hat{L}_{\text{uniformity}}(f, \mathbf{X}; t) = \log \hat{E}_{\text{uniformity}}(f, \mathbf{X}; t), \end{aligned} \tag{35}$$

[4] The transformations are usually parameterised by parameters residing in a measurable space. Here we slightly abuse notation and the integration over T implies an integration over the transformation parameters.

where the last approximation holds for large M (strong law of large numbers and continuous mapping theorem) and $t = 2$ as in (Wang & Isola, 2020).

- **Wasserstein distance between similarity distributions**. Our novel metric estimates the *1-Wasserstein distance* $W_1(q_{\text{sim}}, p_{\text{sim}})$, where p_{sim} is the p.d.f of the inner products when $\mathbf{u}, \mathbf{u}' \sim U(\mathbb{S}^{d-1})$ and q_{sim} is the corresponding one when $\mathbf{u}, \mathbf{u}' \sim f_{\#}p$. According to Cho (2009), p_{sim} is equal to $\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-1}{2})\sqrt{\pi}} \left(\sqrt{1-s^2}\right)^{d-3}$ for $s \in (-1, 1)$ and 0 elsewhere. We chose 1-Wasserstein distance because it can be easily calculated by the following formula

  $$W_1(q_{\text{sim}}, p_{\text{sim}}) = \int_{-\infty}^{+\infty} |F_{q_{\text{sim}}}(s) - F_{p_{\text{sim}}}(s)| ds, \tag{36}$$

  where $F_{p_{\text{sim}}}, F_{q_{\text{sim}}}$ are the c.d.fs corresponding to $p_{\text{sim}}, q_{\text{sim}}$ respectively. Therefore, one can estimate the latter from samples and approximate the integral numerically. In our implementation, we use the method `scipy.stats.wasserstein_distance` from the SciPy library (Virtanen et al., 2020), which implements precisely the aforementioned process.

  To understand the connection between this metric and the uniformity metric used in (Wang & Isola, 2020), we will use an equivalent definition of 1-Wasserstein, using the Kantorovich-Rubinstein dual (see (Peyré et al., 2019) for more details): $W_1(q_{\text{sim}}, p_{\text{sim}}) = \frac{1}{K} \sup_{\|g\|_{\text{Lip}} \leq K} \mathop{\mathbb{E}}_{s \sim q_{\text{sim}}} [g(s)] - \mathop{\mathbb{E}}_{s \sim p_{\text{sim}}} [g(s)]$, where g continuous, $g : \mathcal{S} \to \mathbb{R}$ and $\|g\|_{\text{Lip}} \leq K$ means that the Lipschitz constant of g is at most K. The domain $\mathcal{S}$ in our case is the interval $[-1, 1]$ Now revisiting the definition of the uniformity metric (ignoring the logarithm) we get:

  $$E_{\text{uniformity}}(f_{\#}p; t) = \mathop{\mathbb{E}}_{\substack{\mathbf{u} \sim f_{\#}p \\ \mathbf{u}' \sim f_{\#}p}} \left[e^{-t(2 - 2\mathbf{u}^\top \mathbf{u}')} \right] = \mathop{\mathbb{E}}_{\substack{\mathbf{u} \sim f_{\#}p \\ \mathbf{u}' \sim f_{\#}p}} \left[e^{-2t + 2t\mathbf{u}^\top \mathbf{u}'} \right] = \mathop{\mathbb{E}}_{s \sim q_{\text{sim}}} \left[e^{-2t + 2ts} \right] = \mathop{\mathbb{E}}_{s \sim q_{\text{sim}}} \left[g_1(s; t) \right],$$

  where $g_1(s; t) = e^{-2t+2ts}$ and its Lipschitz constant in $[-1, 1]$ is at most equal to the maximum value of its derivative, i.e. $\text{Lip}(g_1(\cdot; t)) = 2te^{-2t+2t} = 2t$. Therefore:

  $$\begin{aligned} E_{\text{uniformity}}(f_{\#}p; t) - E_{\text{uniformity}}(U(\mathbb{S}^{d-1}); t) &= \mathop{\mathbb{E}}_{s \sim q_{\text{sim}}} [g_1(s; t)] - \mathop{\mathbb{E}}_{s \sim p_{\text{sim}}} [g_1(s; t)] \\ &\leq \sup_{\|g\|_{\text{L}} \leq 2t} \mathop{\mathbb{E}}_{s \sim q_{\text{sim}}} [g(s)] - \mathop{\mathbb{E}}_{s \sim q_{\text{sim}}} [g(s)] = 2tW_1(q_{\text{sim}}, p_{\text{sim}}) \Leftrightarrow \\ L_{\text{uniformity}}(f_{\#}p; t) &\leq \log \left(2tW_1(q_{\text{sim}}, p_{\text{sim}}) + E_{\text{uniformity}} \left(U \left(\mathbb{S}^{d-1} \right); t \right) \right). \end{aligned} \tag{37}$$

  Given that $E_{\text{uniformity}} \left(U \left(\mathbb{S}^{d-1} \right); t \right)$ is fixed for a given t, the above implies that $L_{\text{uniformity}}$ underestimates the closeness of $f_{\#}p$ to a uniform distribution.

- **Rank**. The rank of a given matrix of representations $\text{rank}(\mathbf{U}) \leq \min(M, d)$, where $\mathbf{U} \in \mathbb{R}^{M \times d}$. This gives a measurement of the dimensions that are utilised. To account for numerical errors it is computed as follows:

  $$\widehat{\text{rank}}(\mathbf{U}) = |\{\sigma_i(\mathbf{U}) > \epsilon \mid \sigma_i(\mathbf{U}) : i-\text{th singular value of } \mathbf{U}\}|, \tag{38}$$

  where ϵ was chosen to $1e-5$.

- **Effective rank**. A smooth approximation of the rank (Roy & Vetterli, 2007), that is less prone to numerical errors and has been found in practice to correlate well with downstream performance (Garrido et al., 2023). It is equal to the entropy of the normalised singular values:

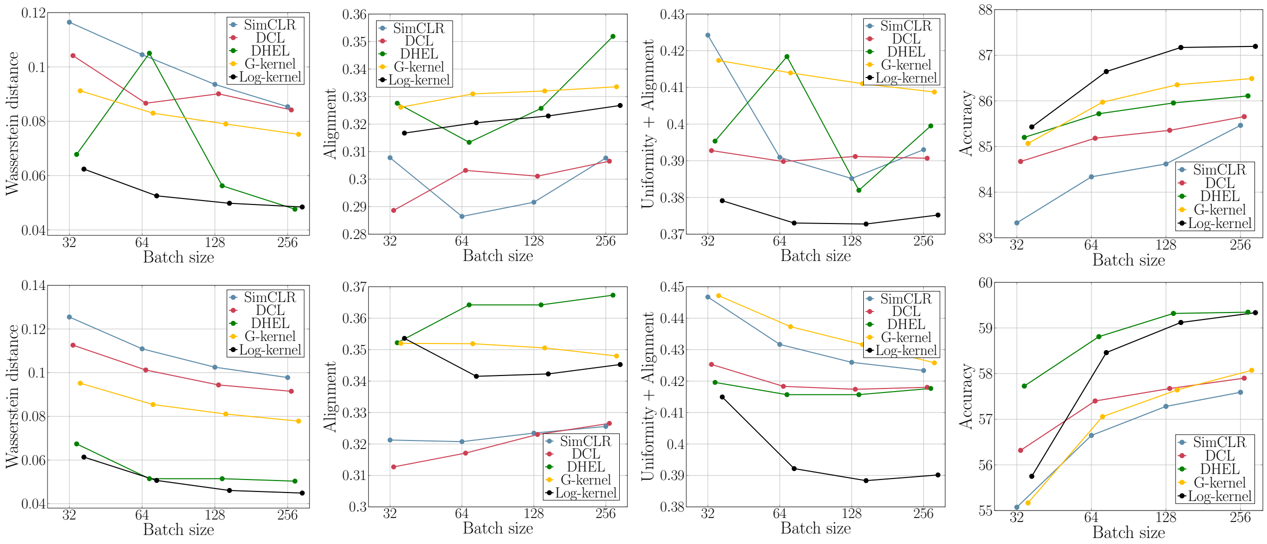
  $$\begin{aligned} \hat{\sigma}_i(\mathbf{U}) &= \frac{\sigma_i(\mathbf{U})}{\sum_{i=1}^{\min(M,d)} |\sigma_i(\mathbf{U})|} + \epsilon \\ \text{eff-rank}(\mathbf{U}) &= -\sum_{i=1}^{\min(M,d)} \hat{\sigma}_i(\mathbf{U}) \log \hat{\sigma}_i(\mathbf{U}), \end{aligned} \tag{39}$$

  where we chose $\epsilon = 1e-7$ as in (Garrido et al., 2023).

C.3. Implementation Details

Code The implementation of the experimental pipeline (networks, augmentations, training, evaluation functions etc) were based on `https://github.com/AndrewAtanov/simclr-pytorch.git`, while our implementation of the proposed loss functions and metrics can be found at `https://github.com/pakoromilas/DHEL-KCL.git`

CIFAR10, CIFAR100 and STL10 ResNet-18 is employed as the encoder architecture for CIFAR10, CIFAR100, and STL10 datasets. Training spans 200 epochs with the SGD optimizer and the cosine annealing learning rate schedule, using a base learning rate of (batch size) / 256. It's worth mentioning that STL10 includes both the train and unlabeled sets for pre-training the model. Augmentations include resizing, cropping, horizontal flipping, color jittering, and random grayscale conversion. Linear evaluation is conducted by training a single linear layer on the learned embeddings, with an additional



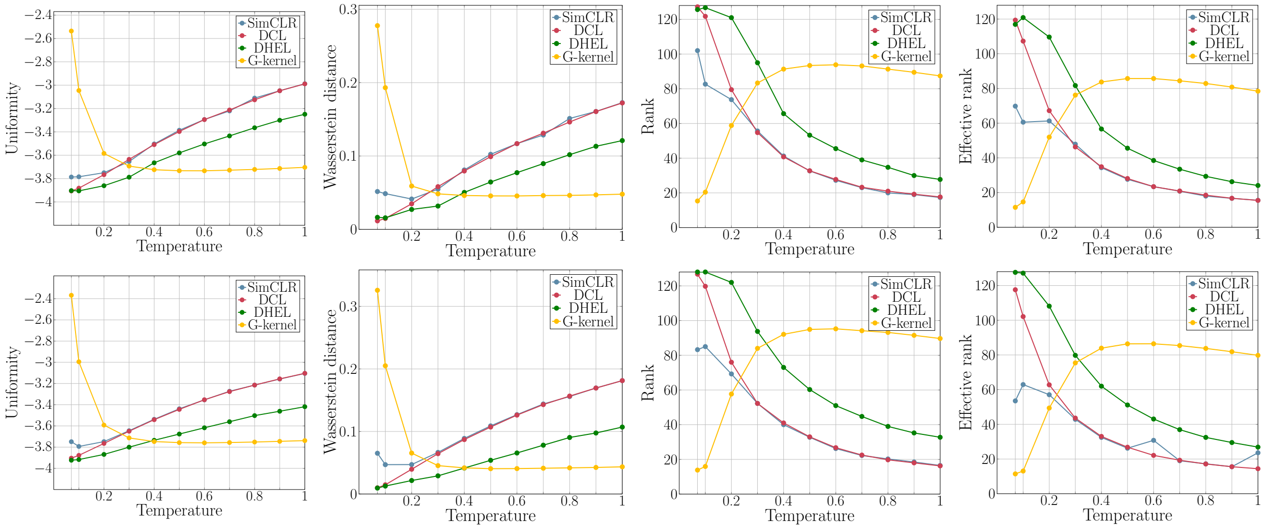
(a) Alignment

(b) Wasserstein distance

(c) Alignment + Wasserstein

(d) Performance

Figure 3: Mean value of properties vs batch size calculated on CIFAR10 (top) & CIFAR100 (bottom) datasets.



(a) Uniformity

(b) Wasserstein distance

(c) Rank

(d) Effective rank

Figure 4: Comparison of two uniformity and rank metrics calculated on CIFAR-10 (top) & CIFAR-100 (bottom) dataset

200 epochs using SGD and a learning rate of 0.1.

ImageNet-100 ResNet-50 is employed as the encoder architecture for ImageNet-100. Training spans 200 epochs with the SGD optimizer and the cosine annealing learning rate schedule, using a base learning rate of 1.4 * (batch size) / 256. We use the same augmentations as in the above datasets and extend them to include gaussian blur. Linear evaluation is conducted by training a single linear layer on the learned embeddings, with an additional 200 epochs using SGD and a learning rate of 0.5.

Hyperparameters For the InfoNCE methods we run experiments for temperatures $[0.07, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$. For the G-kernel we use the same temperature parameters along $w = [8, 16, 32]$. For the Log-kernel we use $s = [1, 1.5, 2, 3]$ and $w = [16, 32, 64]$.

C.4. Further Results

In this section we present further experimental results on (i) the optimisation capabilities of the examined loss functions, as well as (ii) the comparison between different measures of uniformity and rank.

Optimisation capabilities As demonstrated in Appendix B.2, CL objectives are optimised for both perfect alignment and uniformity. In Figure 3 we compare the mean value of such properties for all the examined methods across different batch sizes for the CIFAR10 and CIFAR100 datasets. The methods that decouple uniformity from alignment (DHEL and KCL) achieve superior optimization of uniformity, although they are inferior to baseline methods in alignment optimization. This shortfall may not be problematic, as recent studies (Gupta et al., 2023; Xie et al., 2022) suggest that perfect alignment might not be ideal for downstream performance, since many downstream tasks may not be invariant to the augmentations used to generate positive samples, implying that perfect alignment could be less critical in these scenarios.

Despite alignment and Wasserstein distance observed values being on different scales, both metrics have the same range [0, 1] and monotonicity.Therefore, we examine the balance of the overall metric (alignment + uniformity) to gain insights into how these properties interact. Methods that optimize this combination tend to perform better in downstream tasks. For example, the Log-kernel method performs well for both CIFAR10 and CIFAR100, and the DHEL method excels in CIFAR100. However, the G-kernel achieves the second-best performance in CIFAR10 despite having the poorest optimization. This discrepancy can be attributed to the possibly undesired optimal alignment (Gupta et al., 2023; Xie et al., 2022), and the tolerance-uniformity dilemma (Wang & Liu, 2021). The pretraining stage can be benefited by designing a proper weighting function that correlates these two properties to downstream task performance.

Uniformity metric vs Wasserstein distance In Figure 4, we elucidate the difference between the conventional uniformity metric and the Wasserstein distance introduced in our study. Although both metrics are generally consistent, discernible differences emerge at temperatures 0.07, 0.1, and 0.2. The Wasserstein distance more effectively highlights the superior uniformity of DCL and the compromised uniformity of SimCLR. This distinction is also observed in downstream task performance. Despite DCL and SimCLR having the same alignment and uniformity for all temperatures, DCL achieves superior performance at these specific temperatures ([0.07, 0.1, 0.2]) and matches SimCLR's performance at other temperatures, as shown in Figure 2 . Our metric captures the actual uniformity difference, which reflects to different downstream task performance. Notably, this behavior is observed at optimal uniformity levels, underscoring the discerning power of the introduced metric.

Rank vs Effective rank In Figure 4 we can see that both the rank and the effective rank metrics demonstrate the same trend. In most cases, the rank does correlate with downstream performance as (Garrido et al., 2023) mention, but note that for small temperatures, this correlation seems to die out, which probably comes from the fact that alignment is quite poor, since the latter is not captured by either rank metrics.