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]

[1]University of Athens [2]Archimedes/Athena RC [3]NCSR "Demokritos" [4]The Cyprus Institute

What do different contrastive losses actually optimize for?

- InfoNCE variants and Kernel Contrastive Losses (KCL) **share the same minimisers** when optimising either their **batch objectives** or their expectations **asymptotically**.
- InfoNCE variants exhibit **unknown non-asymptotic behavior**
- Kernel Contrastive Losses are (i) **non-asymptotically** minimised by perfectly aligned and uniform encoders, and (ii) their expected loss is **independent of the batch size**.

Can we optimise for both alignment and uniformity?

- Our theoretical results suggest that there can be a perfectly aligned encoder that is uniform on the negative samples
- InfoNCE variants demonstrate direct and indirect **coupling between the alignment and uniformity** terms thus hurting optimisation
- We introduce the **Decoupled Hyperspherical Energy Loss (DHEL)** that completly **decouples alignment from uniformity**
- Kernel Contrastive Losses (KCL) also decouple these terms

HELLENIC REPUBLIC
National and Kapodistrian University of Athens

ARCHIMEDES

DEMOKRITOS

THE CYPRUS INSTITUTE
RESEARCH • TECHNOLOGY • INNOVATION

H.F.R.I.
Hellenic Foundation for Research & Innovation

ICML
International Conference On Machine Learning

InfoNCE variants share the same mini-batch minimisers

Corollary from Theorems 4.1 & 5.1: When the number of samples is $1 < M \leq d+1$ the mini-batch CL loss functions $\mathbf{L}_{\mathbf{infoNCE}}$, $\mathbf{L}_{\mathbf{SimCLR}}$, $\mathbf{L}_{\mathbf{DCL}}$ and $\mathbf{L}_{\mathbf{DHEL}}$ are all minimised by a point configuration where (i) the positive samples are perfectly aligned, and (ii) the **negative samples form a simplex ETF** on the unit sphere $\mathbb{S}^{d-1}$.

InfoNCE variants share the same minimisers asymptotically

Proposition: The expectations of all the batch-level $\mathbf{L}_{\mathbf{infoNCE}}$, $\mathbf{L}_{\mathbf{SimCLR}}$, $\mathbf{L}_{\mathbf{DCL}}$ and $\mathbf{L}_{\mathbf{DHEL}}$ have the **same asymptotic behaviour** when subtracting appropriate normalising constants. Therefore, (from Wang & Isola 2020 ICML) they are all asymptotically minimised by a point configuration where (i) the positive samples are perfectly aligned, and (ii) the negative samples are uniformly distributed on the sphere $U(\mathbb{S}^{d-1})$.

Main takeaways

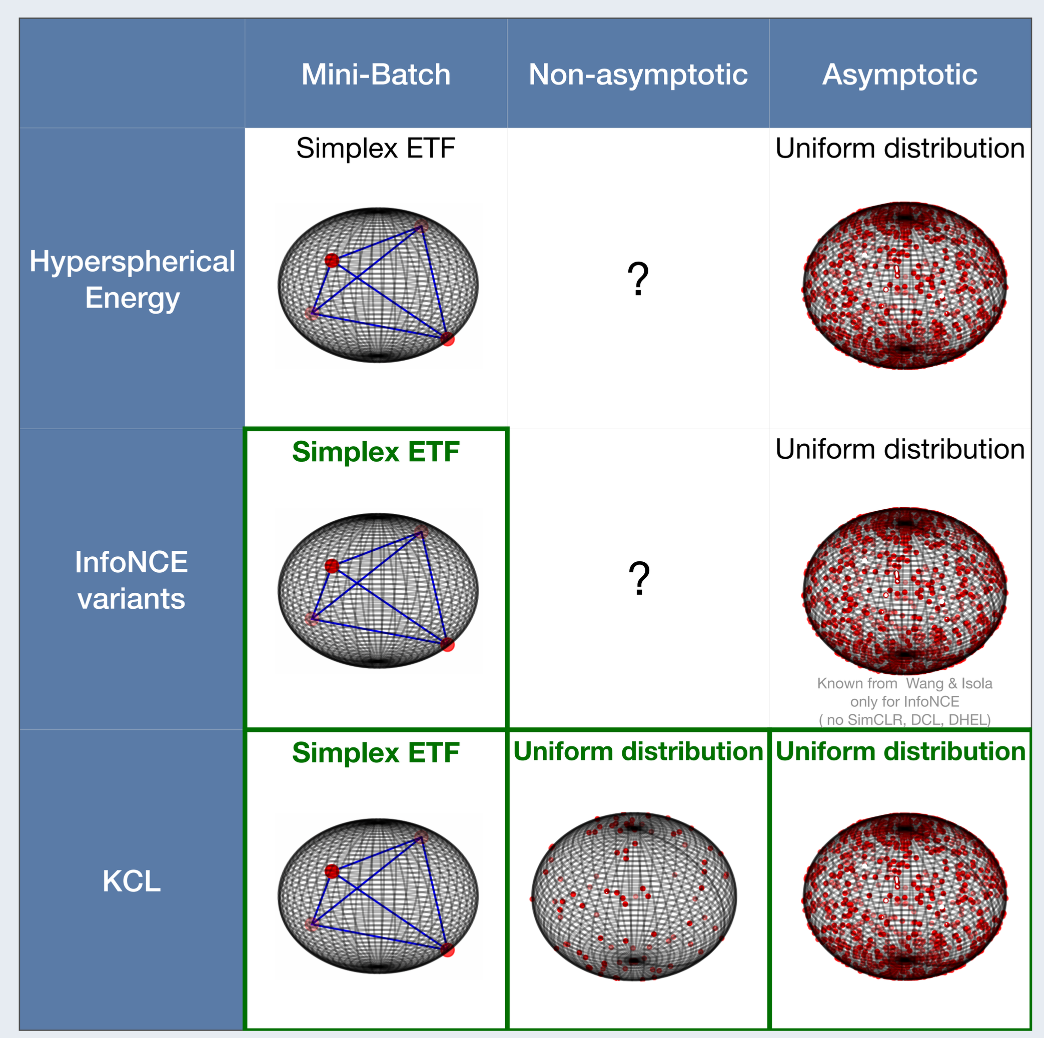

Figure 1:Minimisers of CL ojectives

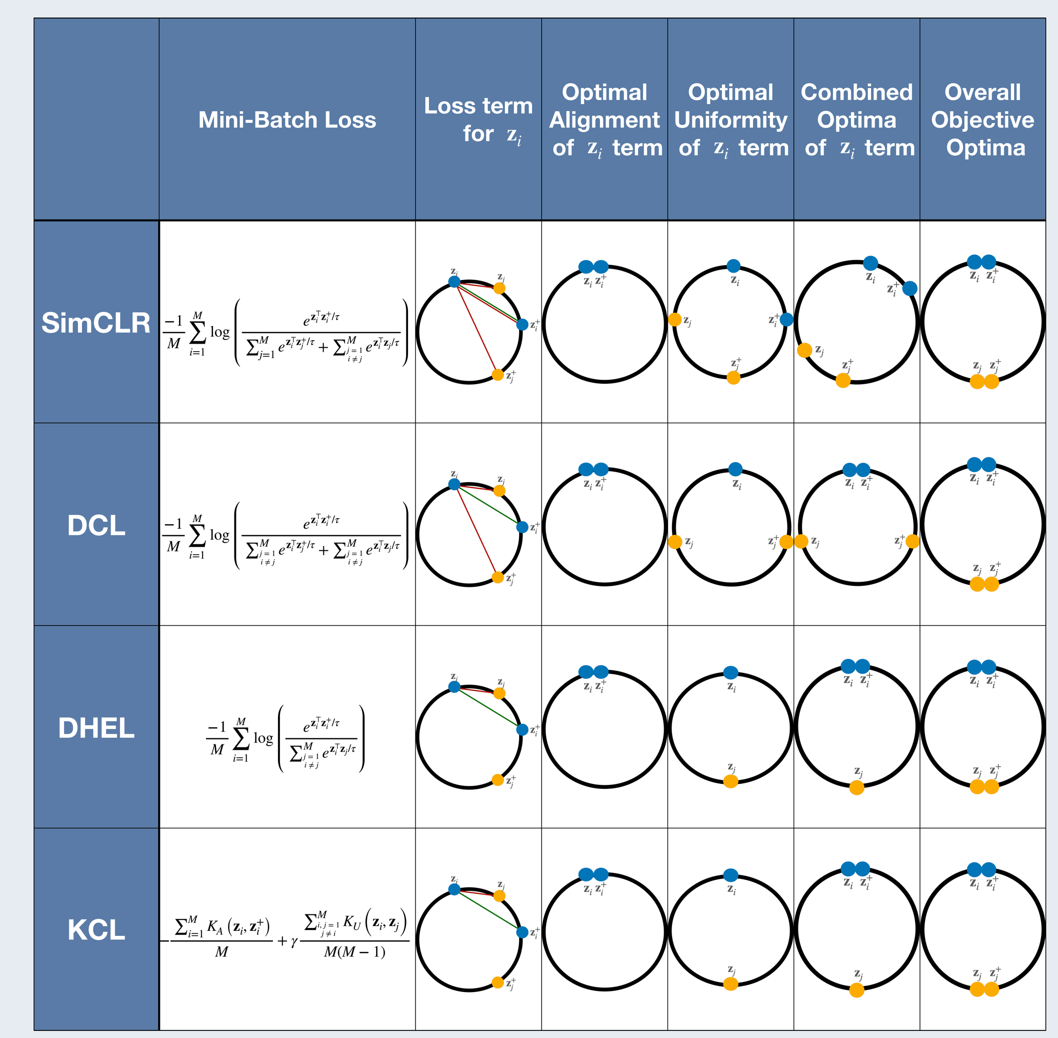

Figure 2:Alignment and uniformity coupling across CL ojectives

Kernel Contrastive Losses share the same minimisers as InfoNCE

Mini-Batch: From **Theorem 6.1** Kernel-based losses are minimised for the same point cofiguration as the infoNCE variants.

Asymptotically: Known result from Hyperspherical Energy Minimisation

KCL are minimised by the uniform distribution non-asymptotically

Proposition: The expectation of the batch-level kernel contrastive loss functions is **independent of the size of the batch**. Therefore, the batch-level loss is an **unbiased estimator** of the (asymptotic) expected loss.

Results

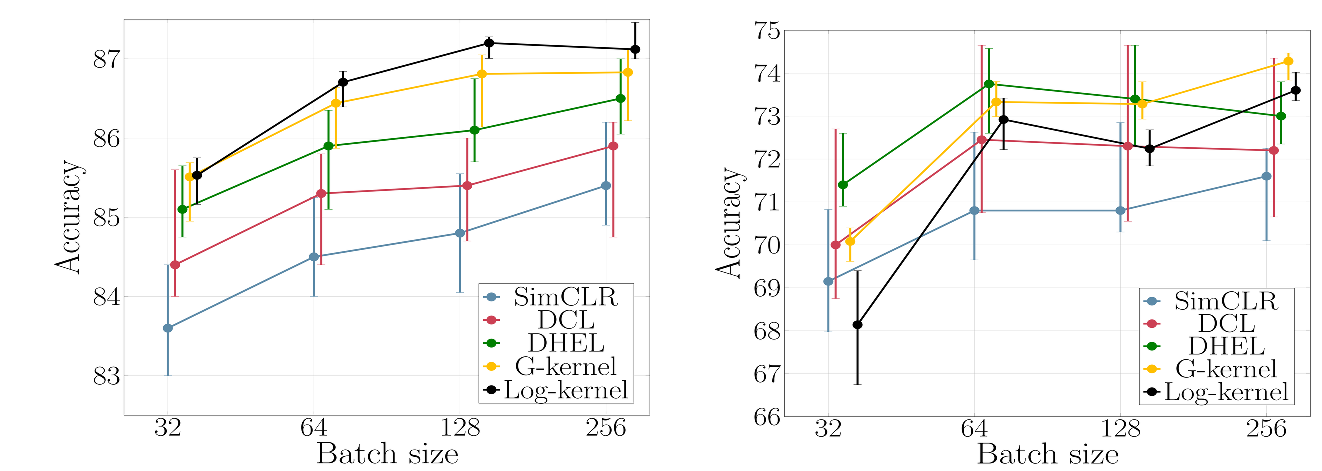

Figure 3:Median performance for different batch sizes on CIFAR10 (left) and ImageNet-100 (right). Errors against each methods hyperparameters are calculated using the 25% and 75% quantiles.

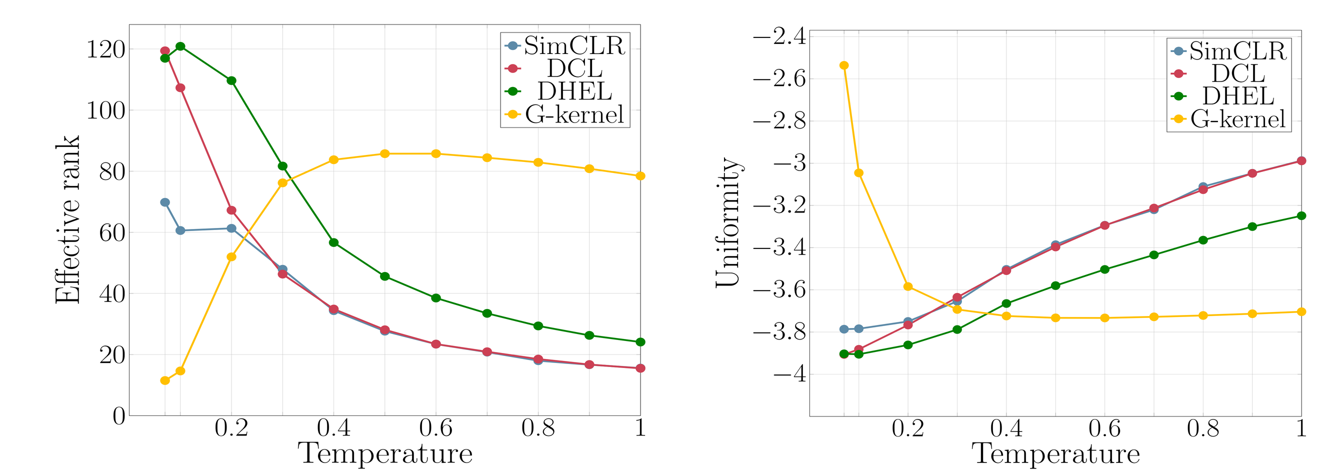

Figure 4:Mean value of effective rank (left) and uniformity (right) vs temperature calculated on CIFAR10

Pros of DHEL and KCL

- **Outperform InfoNCE variants** even with **smaller batch sizes**
- Demonstrate **robustness against hyperparameters**
- Effectively **utilize more dimensions**, mitigating the dimensionality collapse problem
- Learn representations that are consistently **more uniformly distributed** across temperature values
- Achieve an **alignment-uniformity balance** that benefits downstream performance

DHEL vs KCL: DHEL (i) is **consistent** across datasets and (ii) requires **fewer hyperparameters** by naturally balancing alignment and uniformity. KCL is more **robust** in both performance and properties.