AIMS Mathematics, 11(8): 26454–26469.
DOI:10.3934/math.20261061
Received: 25 March 2026
Revised: 03 August 2026
Accepted: 12 August 2026
Published: 25 August 2026

https://www.aimspress.com/journal/Math

Research article

Gradient perturbation and decoupled probabilistic alignment in micro-batch contrastive learning: a theoretical analysis

Yuxiao Zhao[1,2], Runtao Duan[2], Xiaomin Ying[2,*] and Guohua Dong[2,*]

[1] School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China

[2] Center for Computational Biology, Beijing Institute of Basic Medical Sciences, Beijing 100850, China

* **Correspondence:** Email: yingxmbio@foxmail.com, dongguohua15@alumni.nudt.edu.cn.

Abstract: Contrastive learning relies heavily on large-scale negative sampling, posing severe optimization challenges when processing high-dimensional sparse features under forced extreme micro-batch conditions. This paper systematically investigates a potential statistical mechanism underlying training instability and numerical collapse under such constraints. Through the multivariate Delta method and asymptotic analysis, we demonstrate that the finite-sample estimation of the empirical partition function introduces structural gradient perturbations that may become more pronounced as the number of sampled negatives decreases. To overcome this limitation, we formalize a decoupled probabilistic alignment theoretical framework governed by a "global intrinsic semantic threshold". By reformulating the alignment task as additive pairwise Bernoulli modeling, the method removes the dependence on a shared partition function, yielding an unbiased finite-batch estimator of the gradient of its explicitly defined population objective. Comprehensive validations across controlled high-dimensional Monte Carlo simulations and a real-world fMRI visual decoding task substantiate our theoretical claims. Under extreme micro-batch limits, where standard contrastive loss triggers numerical divergence, the decoupled objective suppresses gradient variance by nearly 130-fold and maintains stable convergence. This theoretical framework alleviates the structural dependence of high-performance representation learning on massive batches, providing a robust mathematical foundation for resource-constrained scientific computing.

Keywords: contrastive learning; micro-batch; Bernoulli modeling; gradient perturbation; Delta method

Mathematics Subject Classification: 62F12, 68T07, 90C15

1. Introduction

Over the past few years, contrastive learning has achieved breakthrough progress in the fields of unsupervised and multimodal representation learning, giving rise to milestone models with remarkable success in both academia and industry, such as SimCLR [1], MoCo [2], and CLIP [3]. The core optimization driving force of such methods typically relies on a "relative competition" mechanism within the feature space [4], namely, maximizing the similarity of positive sample pairs while minimizing that of negative sample pairs [5] via contrastive loss functions such as InfoNCE. From a mathematical perspective, the effectiveness of this "negative repulsion" mechanism heavily depends on whether the empirical partition function can serve as an accurate benchmark for the "global repulsive force." Under ideal conditions with sufficient computational resources and a large number of sampled negative targets, the normalized empirical partition estimator can more closely approximate its population counterpart, thereby generally facilitating a smoother optimization and convergence process.

However, in many cutting-edge scientific computing and high-dimensional sparse feature learning scenarios [6], such as high-resolution 3D medical volume segmentation [7, 8], functional magnetic resonance imaging-based semantic brain decoding [9, 10], and the alignment of multi-step generative diffusion priors [11], this ideal optimization paradigm relying on massive batch sizes is fundamentally unattainable. First, due to prohibitive physical acquisition costs and subject privacy constraints, high-dimensional biological and medical signals [12] suffer from severe inherent sample scarcity, which directly precludes large-scale sampling from the data source. Second, to break the computational monopoly of a few tech giants and democratize large-scale multimodal models for the broader research community, it is imperative that optimization algorithms [13] can reliably execute massive non-linear computational graphs under the strict memory constraints of consumer-grade hardware. Consequently, the optimization process in these practical domains is frequently forced into the extreme regime of micro-batches (e.g., $B \leq 5$). When only a few negative targets are available, the normalized empirical partition estimator may be poorly representative of its population counterpart, making the parameter updates more sensitive to the particular targets included in the current batch and potentially increasing the risk of optimization instability [14, 15]. While recent outstanding empirical works like DCL [14] and SigLIP [15, 16] have demonstrated the engineering scalability of decoupled contrastive learning in computer vision, they primarily focus on architectural design and large-scale benchmarking. To the best of our knowledge, the academic community still lacks a parameter-gradient-level analytical framework characterizing how finite-negative sampling perturbs InfoNCE optimization and clarifying which random-ratio mechanism is removed by a decoupled pairwise objective.

To fill this theoretical void, starting from the underlying logic of mathematical optimization [17], this paper deeply investigates the statistical perturbation problem of the empirical partition function in micro-batch contrastive learning. Through the multivariate Delta method and asymptotic analysis, we characterize the finite-negative bias and conditional negative-sampling variability introduced into the InfoNCE parameter gradient [18]. These effects may become more pronounced as the number of sampled negatives decreases, making the parameter updates more sensitive to the particular targets included in the current batch. Based on this rigorous mathematical diagnosis, rather than merely proposing another loss variant, we formalize an analytical framework based on a global intrinsic threshold, termed decoupled probabilistic alignment. Inspired by recent empirical advances in

partition-free contrastive learning, we reformulate feature alignment as additive pairwise Bernoulli modeling without shared softmax normalization and analyze the relationship between its finite-batch gradient and its explicitly defined population objective.

The main contributions of this paper are summarized as follows:

- Based on the multivariate Delta method, we characterize the finite-negative bias and conditional negative-sampling covariance of the InfoNCE parameter gradient. Under the stated assumptions, these effects may become more pronounced as the number of sampled negatives decreases.
- By introducing a global semantic threshold, we reformulate feature alignment as additive pairwise Bernoulli modeling. This removes the shared softmax random-ratio structure and yields an unbiased finite-batch gradient estimator for its explicitly defined population objective.
- We conduct empirical validations on both high-dimensional synthetic data and a real-world fMRI visual decoding task. The results indicate that under extreme micro-batch settings (e.g., $B \leq 5$), the decoupled formulation exhibits more stable gradient behavior and convergence trends during optimization compared to standard InfoNCE.

2. Preliminaries

In this section, we formalize the mathematical framework of representation learning under the contrastive paradigm. Let $\mathcal{X} \subset \mathbb{R}^D$ be a high-dimensional sparse data manifold governed by a distribution p_{data}. In a micro-batch optimization step, let $\mathcal{B} = \{(x_1, t_1), \ldots, (x_B, t_B)\}$ be a set of B independent paired samples. A feature extractor (e.g., a deep neural network) maps these samples into a predicted embedding space and a target embedding space, denoted by $\hat{z}_i = \hat{z}_\theta(x_i) \in \mathbb{R}^D$ and $t_k \in \mathbb{R}^D$ respectively, for $i, k \in \{1, \ldots, B\}$.

Both embedding vectors are L_2-normalized such that they reside on a unit hypersphere, that is, $\|\hat{z}_i\|_2 = \|t_k\|_2 = 1$, which implies $\hat{z}_i, t_k \in \mathbb{S}^{D-1}$. The cosine similarity between a predicted embedding and a target embedding is defined as the inner product

$$s_{ik} = \langle \hat{z}_i, t_k \rangle = \hat{z}_i \cdot t_k. \tag{2.1}$$

Conventional contrastive learning paradigms typically achieve feature alignment by minimizing the InfoNCE loss for each anchor i:

$$\mathcal{L}^{(i)}_{InfoNCE} = -\log \frac{\exp(s_{ii}/\tau)}{\sum_{k=1}^{B} \exp(s_{ik}/\tau)} = -\log \frac{\exp(s_{ii}/\tau)}{\hat{Z}_i}, \tag{2.2}$$

where $\tau > 0$ is a temperature hyperparameter that scales the similarity distribution. The denominator, defined as

$$\hat{Z}_i = \sum_{k=1}^{B} \exp(s_{ik}/\tau), \tag{2.3}$$

is termed the **empirical partition function**, which serves as a discrete normalization factor over the batch. Under the ideal asymptotic assumption of an infinite batch size ($B \to \infty$), this empirical factor, when scaled by $1/B$, approximates the **true partition function** defined by the expectation over the target marginal distribution:

$$Z^*_i = \mathbb{E}_{T^- \sim p_T} \left[\exp\left(\langle \hat{z}_i, T^- \rangle / \tau \right) \right]. \tag{2.4}$$

where p_T denotes the marginal distribution of target embeddings and T^- is sampled independently of the fixed anchor. The paired positive target t_i is treated separately in the subsequent analysis.

In subsequent sections, we will demonstrate that, when B is small, finite-negative estimation of the partition function and its parameter derivative can introduce finite-negative bias and conditional negative-sampling variability into the parameter-gradient dynamics [19].

3. Theoretical diagnosis: stochastic gradient perturbation analysis

In memory-constrained or sample-scarce scenarios, the optimization process is frequently forced to operate under extremely small batch sizes B. This section aims to mathematically quantify the structural perturbation inflicted upon the stochastic gradient direction by such extreme constraints via probability theory and asymptotic analysis.

For the following analysis, we condition on a fixed positive pair (x, t^+), assume $B \geq 2$, and let $M = B-1$ denote the number of sampled negative targets. Let $T_1^-, \ldots, T_M^-$ be conditionally independent and identically distributed samples drawn from the target marginal distribution p_T, independently of the fixed anchor x.

Define the positive exponential weight and its corresponding parameter-gradient factor as

$$U_+ = \exp\left(s_\theta(x, t^+)/\tau\right), \qquad h_+ = \frac{1}{\tau}\nabla_\theta s_\theta(x, t^+). \tag{3.1}$$

For each negative target, define

$$U_j = \exp\left(s_\theta(x, T_j^-)/\tau\right), \qquad h_j = \frac{1}{\tau}\nabla_\theta s_\theta(x, T_j^-), \qquad V_j = U_j h_j. \tag{3.2}$$

Here, U_j measures the contribution of the j-th negative target to the partition function, and V_j is its corresponding weighted contribution to the parameter gradient. Unless otherwise specified, $\mathbb{E}_-$, Var_-, and Cov_- denote expectation, variance, and covariance only over the sampled negative targets, conditional on the fixed positive pair and parameter value.

Lemma 3.1 (Stochastic properties of the empirical partition function)**.** *We define the normalized negative partition estimator as the empirical mean of the M negative exponential weights:*

$$\bar{Z}_M^- = \frac{1}{M}\sum_{j=1}^{M} \exp\left(s_\theta(x, T_j^-)/\tau\right) = \frac{1}{M}\sum_{j=1}^{M} U_j. \tag{3.3}$$

Here, the bar denotes empirical averaging, the subscript M denotes the number of sampled negative targets, and the superscript "−" indicates that only negative targets are included. The positive weight U_+ is treated separately. Assume that

$$\mu_U = \mathbb{E}_-[U_j], \qquad \sigma_U^2 = \mathrm{Var}_-(U_j) < \infty. \tag{3.4}$$

Then,

$$\mathbb{E}_-[\bar{Z}_M^-] = \mu_U, \qquad \mathrm{Var}_-(\bar{Z}_M^-) = \frac{\sigma_U^2}{M}. \tag{3.5}$$

Thus, the conditional variance of the negative partition estimator scales as $O(M^{-1})$ and, whenever $\sigma_U^2 > 0$ is fixed, as $\Theta(M^{-1})$.

Proof. By the conditional independence of the sampled negative targets, the variance of the empirical mean is

$$\begin{aligned} \mathrm{Var}_-(\bar{Z}_M^-) &= \mathrm{Var}_- \left(\frac{1}{M} \sum_{j=1}^{M} U_j \right) \\ &= \frac{1}{M^2} \sum_{j=1}^{M} \mathrm{Var}_-(U_j) \\ &= \frac{\sigma_U^2}{M}. \end{aligned} \tag{3.6}$$

This explicitly establishes the inverse scaling law of the estimator variance with respect to the number of sampled negatives. Furthermore, the coefficient

$$\sigma_U^2 = \mathrm{Var}_- \left[\exp\left(s_\theta(x, T^-)/\tau \right) \right] \tag{3.7}$$

depends on the temperature parameter τ. The exponential transformation can therefore make the partition estimator highly sensitive to variations in the sampled similarity scores. □

Theorem 3.2 (Asymptotic expansion of the stochastic gradient under finite-batch sampling)**.** *Assume:*

(i) The pairs $\{(U_j, V_j)\}_{j=1}^M$ *are conditionally independent and identically distributed, with a distribution that does not depend on* M*;*

(ii) The exponential weights and weighted gradient contributions admit finite third absolute moments:

$$\mathbb{E}_- \left[\left(|U_j| + \|V_j\|_2 \right)^3 \right] < \infty; \tag{3.8}$$

(iii) The target marginal distribution p_T *does not depend on* θ*, and differentiation and conditional expectation can be interchanged.*

Define

$$\mu_V = \mathbb{E}_-[V_j], \qquad r = \frac{\mu_V}{\mu_U}. \tag{3.9}$$

The finite-negative InfoNCE parameter gradient is

$$\hat{g}_M = \nabla_\theta \mathcal{L}_{\mathrm{InfoNCE}} = -h_+ + \frac{U_+ h_+ + \sum_{j=1}^M V_j}{U_+ + \sum_{j=1}^M U_j}. \tag{3.10}$$

Define the corresponding infinite-negative population objective, up to an additive constant independent of θ*, as*

$$\mathcal{J}_\infty(\theta \mid x, t^+) = -\frac{s_\theta(x, t^+)}{\tau} + \log \mu_U. \tag{3.11}$$

Since

$$\nabla_\theta \mu_U = \mathbb{E}_- \left[\nabla_\theta U_j \right] = \mathbb{E}_-[V_j] = \mu_V, \tag{3.12}$$

the infinite-negative population gradient is

$$g_\infty = \nabla_\theta \mathcal{J}_\infty = -h_+ + r. \tag{3.13}$$

Then, as $M \to \infty$, the expected finite-negative parameter gradient admits the expansion

$$\begin{aligned}\mathbb{E}_-[\hat{g}_M] - g_\infty = \frac{1}{M}\Bigg[\frac{U_+}{\mu_U}(h_+ - r) - \frac{\mathrm{Cov}_-(V_j, U_j)}{\mu_U^2} \\ + \frac{\mu_V \,\mathrm{Var}_-(U_j)}{\mu_U^3}\Bigg] + R_M,\end{aligned} \tag{3.14}$$

where

$$\|R_M\|_2 = O(M^{-3/2}). \tag{3.15}$$

Moreover, the conditional negative-sampling covariance satisfies

$$\mathrm{Cov}_-(\hat{g}_M) = \frac{1}{M\mu_U^2}\,\mathrm{Cov}_-(V_j - rU_j) + o(M^{-1}). \tag{3.16}$$

Here $\mathrm{Cov}_-(V_j, U_j)$ *is a vector, whereas* $\mathrm{Cov}_-(V_j - rU_j)$ *is a covariance matrix.*

Proof. Define the empirical negative-sample means

$$\bar{U}_M = \frac{1}{M}\sum_{j=1}^{M} U_j, \qquad \bar{V}_M = \frac{1}{M}\sum_{j=1}^{M} V_j. \tag{3.17}$$

The random-ratio component of the finite-negative gradient can be expressed as

$$f_M(\bar{V}_M, \bar{U}_M) = \frac{\bar{V}_M + U_+ h_+/M}{\bar{U}_M + U_+/M}. \tag{3.18}$$

At the population means (μ_V, μ_U),

$$\begin{aligned}f_M(\mu_V, \mu_U) &= \frac{\mu_V + U_+ h_+/M}{\mu_U + U_+/M} \\ &= r + \frac{U_+}{M\mu_U}(h_+ - r) + O(M^{-2}).\end{aligned} \tag{3.19}$$

Applying the second-order multivariate Delta method directly to f_M gives

$$\begin{aligned}\mathbb{E}_-\left[f_M(\bar{V}_M, \bar{U}_M)\right] = r + \frac{U_+}{M\mu_U}(h_+ - r) \\ - \frac{\mathrm{Cov}_-(V_j, U_j)}{M\mu_U^2} + \frac{\mu_V \,\mathrm{Var}_-(U_j)}{M\mu_U^3} + O(M^{-3/2}).\end{aligned} \tag{3.20}$$

Combining this result with the positive alignment term $-h_+$ gives Eq (3.14).

For the covariance, the first-order Delta expansion has the linear random term

$$\frac{1}{\mu_U}\left[(\bar{V}_M - \mu_V) - r(\bar{U}_M - \mu_U)\right]. \tag{3.21}$$

Its conditional covariance is

$$\frac{1}{M\mu_U^2}\,\mathrm{Cov}_-(V_j - rU_j), \tag{3.22}$$

which gives Eq (3.16).

Finally, the L_2-normalization implies $s_\theta \in [-1, 1]$, so the exponential denominator is bounded away from zero. Together with the finite third-moment assumption, this controls the remainder terms. □

Analysis. The expansion in Theorem 3.2 provides a quantitative decomposition of the finite-negative parameter gradient into its infinite-negative population component and finite-negative perturbation terms.

The first perturbation term,

$$\frac{U_+}{M\mu_U}(h_+ - r), \tag{3.23}$$

arises because the positive target remains inside a finite partition function containing only M sampled negatives. The remaining terms arise from the statistical dependence between the random gradient numerator and the random partition denominator. In general, the expectation of a ratio of correlated random quantities is not equal to the ratio of their expectations.

From Lemma 3.1 and Theorem 3.2, both the leading finite-negative bias and the conditional negative-sampling covariance generally scale as $O(M^{-1})$, where $M = B - 1$. Thus, as the number of sampled negatives decreases, the parameter gradient may become increasingly sensitive to the particular negative targets included in the current batch. The leading coefficients are nevertheless distribution-dependent, and the bias may vanish under special conditions.

It is important to note that this is an asymptotic result rather than an exact numerical law for extremely small batch sizes. Moreover, the covariance in Eq (3.16) measures only the variation caused by resampling the negative targets after fixing the positive pair. It does not represent the complete SGD variance across different anchors, positive targets, or training iterations.

The temperature parameter τ affects both the exponential weights and the corresponding gradient factors. It may therefore substantially change the magnitude of the finite-negative perturbations, although this relationship is not universally monotonic.

Overall, finite-negative estimation introduces a random-ratio structure into the InfoNCE parameter gradient. Under extreme micro-batch conditions, the resulting bias and conditional negative-sampling variability may increase the sensitivity of the optimization trajectory and consequently increase the risk of optimization instability or numerical collapse. This result identifies a potential statistical mechanism rather than implying that numerical collapse must occur in every micro-batch configuration.

4. Decoupled probabilistic alignment via intrinsic thresholding

Motivated by the finite-negative random-ratio perturbation characterized in Theorem 3.2, we avoid explicit dependence on the empirical softmax partition function. This section formalizes a decoupled probabilistic modeling mechanism and shows that its finite-batch parameter gradient is unbiased for an explicitly defined pairwise population objective.

To visually elucidate this paradigm shift from "intra-batch competition" to "pairwise alignment", Figure 1 contrasts their computational structures. In the conventional coupled mode, all similarity scores contribute through a shared empirical partition function. In the proposed decoupled mode, the objective is expressed as a sum of pairwise logistic terms without shared softmax normalization, thereby removing the random-ratio structure analyzed in Section 3.

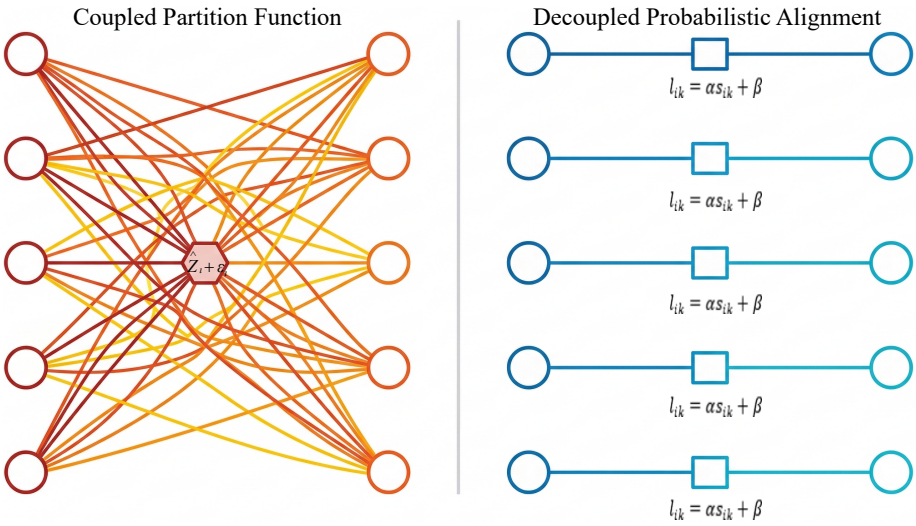

Figure 1. Structural comparison of optimization paradigms under micro-batch constraints. **Left (coupled partition function):** All similarity scores are coupled through a shared empirical softmax partition function. **Right (decoupled probabilistic alignment):** An affine pairwise logit with the induced threshold $s_{\text{th}} = -\beta/\alpha$ reformulates alignment into additive pairwise terms, thereby avoiding the shared softmax random-ratio structure.

Definition 4.1 (Affine pairwise logit and induced semantic threshold)**.** We introduce an affine transformation with a learnable bias β to convert the cosine similarity into a pairwise logit. For each predicted-target pair, the logit l_{ik} is defined as

$$l_{ik} = \alpha s_{ik} + \beta = \alpha\,(\hat{z}_i \cdot t_k) + \beta, \tag{4.1}$$

where the inverse-temperature scale $\alpha > 0$ controls the sharpness of the decision boundary, and the bias term $\beta \in \mathbb{R}$ acts as a global intercept. Since $\alpha > 0$, the probability boundary $\sigma(l_{ik}) = 1/2$ corresponds to the induced similarity threshold

$$s_{\text{th}} = -\frac{\beta}{\alpha}. \tag{4.2}$$

This threshold provides an absolute pairwise activation baseline without requiring comparisons with the other targets in the current batch.

Based on this transformation, we reformulate feature alignment as a collection of additive pairwise Bernoulli losses without shared softmax normalization. For each anchor i, the empirical decoupled alignment objective is defined as

$$\mathcal{L}^{(i)}_{\text{DPA},\lambda} = -\log\sigma(l_{ii}) - \frac{\lambda}{B-1}\sum_{\substack{k=1\\k\neq i}}^{B}\log\left(1-\sigma(l_{ik})\right), \qquad B \geq 2, \tag{4.3}$$

where $\sigma(a) = 1/(1+\exp(-a))$ is the Sigmoid function and $\lambda > 0$ is a fixed negative-to-positive weighting parameter that does not depend on the batch size.

The factor $1/(B-1)$ normalizes the contribution of the negative pairs. Consequently, their total expected weight remains λ when the batch size changes. Without this normalization, increasing

or decreasing B would also change the relative weighting of positive and negative pairs and would therefore correspond to a different population objective.

Definition 4.2 (Population objective of decoupled alignment)**.** Let $\vartheta = (\theta, \alpha, \beta)$ denote all trainable parameters. For a fixed positive pair (x_i, t_i), we define the corresponding conditional population objective as

$$\begin{aligned}\mathcal{J}^{(i)}_{\mathrm{DPA},\lambda}(\vartheta \mid x_i, t_i) = &-\log \sigma\left(l_\vartheta(x_i, t_i)\right) \\ &- \lambda \mathbb{E}_{T^- \sim p_T}\left[\log\left(1 - \sigma\left(l_\vartheta(x_i, T^-)\right)\right)\right],\end{aligned} \tag{4.4}$$

where p_T denotes the marginal distribution of target embeddings and $T^- \sim p_T$ is sampled independently of the fixed anchor x_i.

Theorem 4.3 (Unbiased gradient estimation for the population objective)**.** *Assume that, conditional on the fixed positive pair (x_i, t_i), the negative targets $\{t_k\}_{k\neq i}$ are independently and identically distributed according to p_T and are independent of the fixed anchor x_i. Assume also that λ is fixed independently of B, that p_T is treated as fixed with respect to ϑ during differentiation, and that differentiation and expectation can be interchanged. Then, for every $B \geq 2$,*

$$\mathbb{E}_-\left[\nabla_\vartheta \mathcal{L}^{(i)}_{\mathrm{DPA},\lambda} \mid x_i, t_i\right] = \nabla_\vartheta \mathcal{J}^{(i)}_{\mathrm{DPA},\lambda}(\vartheta \mid x_i, t_i). \tag{4.5}$$

Therefore, the empirical DPA gradient is an unbiased estimator of the gradient of its explicitly defined population objective.

Proof. Define the negative pairwise loss as

$$\ell_-(l) = -\log\left(1 - \sigma(l)\right). \tag{4.6}$$

Since the $B - 1$ negative targets are conditionally identically distributed according to p_T,

$$\begin{aligned}&\mathbb{E}_-\left[\frac{1}{B-1}\sum_{\substack{k=1\\k\neq i}}^{B} \ell_-(l_{ik}) \,\middle|\, x_i, t_i\right] \\ &\qquad = \mathbb{E}_{T^- \sim p_T}\left[\ell_-\left(l_\vartheta(x_i, T^-)\right)\right].\end{aligned} \tag{4.7}$$

Substituting this equality into the empirical objective gives

$$\mathbb{E}_-\left[\mathcal{L}^{(i)}_{\mathrm{DPA},\lambda} \mid x_i, t_i\right] = \mathcal{J}^{(i)}_{\mathrm{DPA},\lambda}(\vartheta \mid x_i, t_i). \tag{4.8}$$

Under the assumed interchange of differentiation and expectation, taking the parameter gradient on both sides yields

$$\mathbb{E}_-\left[\nabla_\vartheta \mathcal{L}^{(i)}_{\mathrm{DPA},\lambda} \mid x_i, t_i\right] = \nabla_\vartheta \mathcal{J}^{(i)}_{\mathrm{DPA},\lambda}(\vartheta \mid x_i, t_i). \tag{4.9}$$

For the positive pair, the score derivative is

$$\frac{\partial \mathcal{L}^{(i)}_{\mathrm{DPA},\lambda}}{\partial s_{ii}} = \alpha\left(\sigma(l_{ii}) - 1\right). \tag{4.10}$$

For each negative pair, it is

$$\frac{\partial \mathcal{L}^{(i)}_{\mathrm{DPA},\lambda}}{\partial s_{ik}} = \frac{\lambda\alpha}{B-1}\sigma(l_{ik}), \qquad k \neq i. \tag{4.11}$$

Thus, each pair contributes additively and no shared softmax partition-function ratio appears in the gradient. □

Compared with InfoNCE, the DPA objective removes the shared softmax random-ratio structure and replaces it with additive pairwise logistic terms. Theorem 4.3 establishes that its finite-batch parameter gradient is unbiased for the explicitly defined pairwise population objective. Pairwise additivity does not imply that all gradient terms are statistically independent, because they still share the anchor representation and model parameters. The result also does not imply zero gradient variance or guarantee stable convergence. Instead, it shows that DPA removes the particular cross-sample normalization mechanism identified in Section 3, which may improve robustness under extreme micro-batch conditions.

In summary, the proposed DPA objective addresses the specific finite-negative batch-coupling mechanism identified in Section 3. Unlike InfoNCE, its parameter gradient contains no shared empirical partition-function ratio. With a fixed negative weight λ and the normalization $1/(B-1)$, the finite-batch gradient is unbiased with respect to the same explicitly defined pairwise population objective for every $B \geq 2$. Consequently, changing the batch size affects the amount of negative-sampling variability but does not change the population gradient targeted in expectation.

5. Numerical validation and empirical analysis

5.1. Simulation setup

During the optimization process of deep models, when training instability or convergence failure occurs, it is often difficult to distinguish whether the root cause originates from engineering implementation issues or structural flaws within the objective function itself. Therefore, this experiment aims to construct a controlled mathematical environment to reduce the influence of engineering factors, focusing on analyzing the impact of micro-batch constraints on the partition function in contrastive learning.

Specifically, we designed a high-dimensional Monte Carlo simulation framework. Considering that real-world neural representations (e.g., fMRI visual responses) typically exhibit high-dimensional sparsity (approximately 1.5×10^4–1.9×10^4 dimensions), we set the semantic target space as a unit hypersphere with a dimensionality of $D = 16384$, and uniformly sampled a batch of target embeddings of size B from it in each optimization step. Furthermore, to simulate the inherently low signal-to-noise ratio (SNR) in biological or sensory measurements, we generated noisy input signals by injecting zero-mean, high-variance isotropic Gaussian noise ($\sigma = 2.5$) into the targets. A linear projection network was employed to map these noisy inputs back to the semantic space, and the model was optimized using stochastic gradient descent (SGD) under varying micro-batch sizes ($B \in \{2, 4, 8\}$). To avoid the structural misleadingness of comparing absolute loss values across different objective functions, we selected three objective physical metrics for continuous tracking: the gradient L_2 norm, the relative normalized loss, and the global positive pair cosine similarity, thereby rigorously evaluating the stability of different optimization objectives on a unified scale.

5.2. Quantitative analysis of optimization stability

Figure 2 illustrates the starkly contrasting mathematical dynamics of the conventional InfoNCE and the formalized decoupled alignment mechanism under the extreme micro-batch setting ($B = 2$) and high-dimensional noise. We conduct a comprehensive diagnosis of this optimization discrepancy from

three independent physical perspectives.

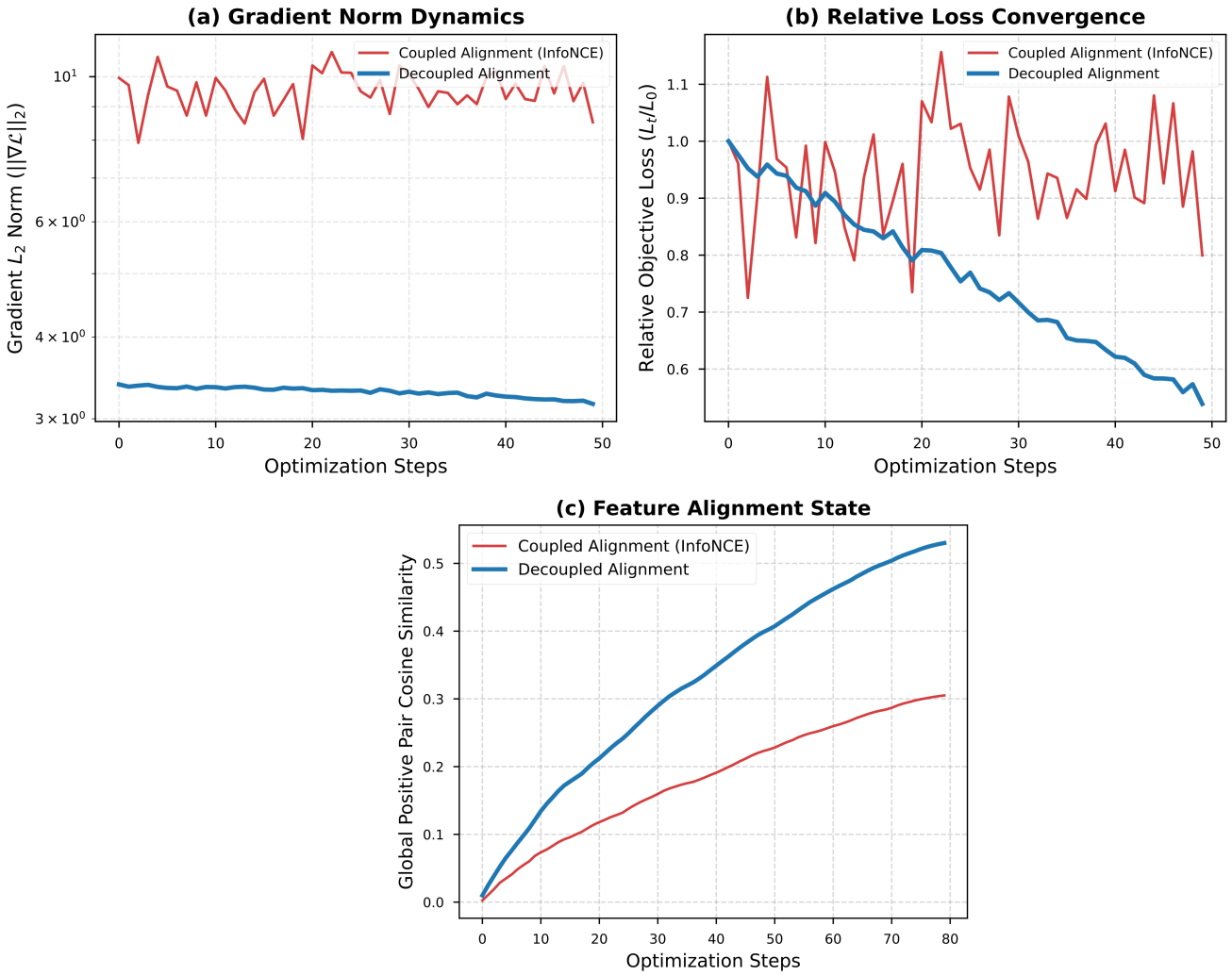

Figure 2. Comprehensive optimization dynamics on a synthetic high-dimensional manifold under extreme micro-batch constraints ($B = 2$). **(a)** Gradient norm dynamics ($\|\nabla\mathcal{L}\|_2$): Standard InfoNCE exhibits severe oscillations and high recorded variability under finite-negative sampling, while the decoupled alignment exhibits a smoother trajectory with lower recorded variability. **(b)** Relative loss dynamics: Tracked using a normalized initial scale (L_t/L_0). InfoNCE displays pronounced oscillations, whereas the decoupled alignment shows a smoother and more stable decreasing trajectory. **(c)** Feature alignment state: Monitored by the mean positive-pair cosine similarity. InfoNCE shows slower alignment progress, whereas the decoupled mechanism increases the positive-pair similarity more rapidly and smoothly.

Regarding gradient field stability (Figure 2a), the L_2 norm of the standard InfoNCE gradient displays violent stochastic oscillations, spanning multiple orders of magnitude on a logarithmic scale. The observed gradient-norm fluctuations provide qualitative empirical support for the finite-negative sensitivity characterized in Theorem 3.2. Conversely, by removing the shared softmax partition function, DPA exhibits a substantially smoother gradient-norm trajectory and greater empirical

resilience to local noise under the same experimental configuration.

In terms of numerical convergence (Figure 2b), the normalized InfoNCE loss exhibits a sawtooth-like unstable trajectory and eventually produces numerical overflow under the evaluated configuration. In contrast, DPA maintains a smoothly decreasing loss trajectory and exhibits stable convergence without producing non-finite values.

Ultimately, regarding the true geometric feature alignment state (Figure 2c), tracked via the pure physical metric of global positive pair cosine similarity, the learning capacity of InfoNCE is severely suppressed. Trapped by the high-variance gradient estimations induced by the extreme micro-batch, its geometric alignment progress is exceedingly sluggish and heavily lags behind. In stark contrast, the decoupled alignment engine, unconstrained by the batch field of view, smoothly and rapidly drives the global feature similarity upward. This geometric visualization demonstrates that the decoupled mechanism maintains superior optimization efficiency under severe mathematical constraints, alleviating the structural bottlenecks of the coupled paradigm.

5.3. Ablation on micro-batch limits

To quantitatively compare the optimization behavior of InfoNCE and DPA across different micro-batch sizes, Table 1 reports the corresponding gradient statistics.

Table 1. Quantitative comparison of gradient dynamics under micro-batch constraints.

Method	Batch size (B)	Peak gradient norm	Gradient variance
Coupled alignment (InfoNCE)	8	6.44×10^{0}	2.91×10^{-3}
Coupled alignment (InfoNCE)	4	8.74×10^{0}	4.00×10^{-2}
Coupled alignment (InfoNCE)	2	1.09×10^{1}	4.10×10^{-1}
DeCoupled alignment	8	$\mathbf{4.21 \times 10^{-1}}$	$\mathbf{1.20 \times 10^{-6}}$
DeCoupled alignment	4	$\mathbf{1.19 \times 10^{0}}$	$\mathbf{4.01 \times 10^{-5}}$
DeCoupled alignment	2	$\mathbf{3.39 \times 10^{0}}$	$\mathbf{3.17 \times 10^{-3}}$

As demonstrated in Table 1, when the batch size B is reduced from 8 to 2, the gradient variance of standard InfoNCE surges by approximately 140 times (from 2.91×10^{-3} to 4.10×10^{-1}). This trend provides qualitative empirical support for the finite-negative sensitivity characterized in Section 3. In contrast, the decoupled alignment mechanism structurally isolates the partition function. Consequently, even under the degraded condition of $B = 2$, it effectively suppresses the gradient variance to a low level of 3.17×10^{-3}. This represents a nearly 130-fold variance reduction compared to InfoNCE, confirming its capability for robust estimation.

5.4. Real-world validation: fMRI-to-image semantic alignment

To validate whether the optimization pathology diagnosed in the synthetic manifold manifests in real-world cutting-edge scientific computing, we conducted an empirical study on a fMRI visual decoding task. The experiment utilizes the natural scenes dataset (NSD) [20], selected for its high-quality fMRI-image pairs and relatively substantial data scale. Crucially, the fMRI voxel features extracted from the visual cortex in the NSD benchmark are also high-dimensional (e.g., 15,724 selected voxel features for Subject 1), while our theoretical simulation uses $D = 16,384$. This dimensional

comparability provides a natural logical connection between the synthetic Monte Carlo analysis and the real-world validation. We adopted the representative MindEye [9, 21] framework as our baseline. Although MindEye has achieved a substantial breakthrough in aligning sparse fMRI signals to CLIP semantic features, the multi-step mapping network it introduces incurs a massive memory footprint. This extensive computational graph forces the practical training batch size to be strictly limited (originally configured as $B = 32$), which typically approaches the physical memory limit of a single A100 GPU. Therefore, a rigorous mathematical investigation into the optimization dynamics of micro-batches under extreme memory constraints is highly necessary for the practical deployment of such frontier applications.

To investigate structural stability under extreme limits, we further compressed the batch size into the micro-batch regime ($B \in \{2, 3, 4, 5, 8, 32\}$) and systematically compared the training dynamics of the standard coupled InfoNCE (MindEye's default configuration) against the decoupled probabilistic alignment (DPA). In all comparative experiments, the models were trained for the default 240 epochs as configured by MindEye. The empirical results are summarized in Table 2.

Table 2. Training stability of fMRI semantic alignment (NSD dataset, 240 epochs) under micro-batch constraints.

Batch size (B)	Baseline (coupled InfoNCE)	Proposed method (decoupled alignment)
32 (original)	Stable convergence (completed)	Stable convergence (completed)
8	Stable convergence (completed)	Stable convergence (completed)
5	**Diverged / NaN (early epochs)**	**Stable convergence (completed)**
4	**Diverged / NaN (early epochs)**	**Stable convergence (completed)**
3	**Diverged / NaN (early epochs)**	**Stable convergence (completed)**
2	**Diverged / NaN (early epochs)**	**Stable convergence (completed)**

The empirical observations further demonstrate the practical relevance of the theoretical analysis. Under identical training conditions, the InfoNCE runs produced numerical overflow during the initial optimization epochs at $B \leq 5$, whereas DPA completed the scheduled 240-epoch training process and exhibited stable convergence across all evaluated micro-batch settings. This contrast provides qualitative empirical support for the finite-negative sensitivity characterized in Section 3 and for the removal of the shared softmax random-ratio structure by DPA. Overall, the results demonstrate the greater numerical robustness of DPA under the evaluated extreme micro-batch configurations.

6. Conclusions

This paper systematically analyzes the optimization instability of contrastive learning under micro-batch conditions from both theoretical and empirical perspectives. Based on the multivariate Delta method and asymptotic analysis, we characterize, under the stated assumptions, the structural perturbations introduced into the InfoNCE parameter gradient by finite-sample estimation of the empirical partition function. This finding provides a rigorous theoretical basis for understanding a potential mechanism underlying small-batch training instability and failure.

To address these issues, we formalize a decoupled probabilistic alignment framework governed by a global semantic threshold. Under the stated assumptions, the method removes the dependence

of the parameter gradient on a shared softmax partition function, and its finite-batch gradient is an unbiased estimator of the gradient of its explicitly defined pairwise population objective. High-dimensional Monte Carlo simulations and a real-world fMRI visual decoding task further evaluate its practical optimization behavior under micro-batch conditions. In the evaluated extreme micro-batch configurations, the InfoNCE implementation produced non-finite values, whereas the decoupled method exhibits lower recorded gradient fluctuations.

Overall, this study characterizes a potential mechanism affecting contrastive-learning optimization under micro-batch conditions and provides a theoretical justification for decoupled approaches. The results have practical value for applications strictly constrained by computational resources or data-acquisition costs and offer theoretical and empirical guidance for constructing more robust cross-modal alignment architectures.

Author contributions

Yuxiao Zhao: Methodology, Writing–review & editing, Validation; Runtao Duan: Conceptualization, Formal analysis; Xiaomin Ying: Writing–review & editing; Guohua Dong: Conceptualization, Methodology, Funding acquisition. All authors have read and approved the final version of the manuscript for publication.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.32300525).

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, In: *International conference on machine learning*, 2020, 1597–1607.
2. K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, In: *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Seattle, WA, USA, 2020, 9726–9735. https://doi.org/10.1109/CVPR42600.2020.00975
3. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., Learning transferable visual models from natural language supervision, In: *Proceedings of the 38th international conference on machine learning*, PMLR, 2021, 8748–8763.

4. Z. Wu, Y. Xiong, S. X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, In: *2018 IEEE/CVF conference on computer vision and pattern recognition*, Salt Lake City, UT, USA, 2018, 3733–3742. https://doi.org/10.1109/CVPR.2018.00393
5. T. Wang, P. Isola, Understanding contrastive representation learning through alignment and uniformity on the hypersphere, In: *Proceedings of the 37th international conference on machine learning*, PMLR, 2020, 9929–9939.
6. W. Yang, L. Pan, J. Wan, Smoothing gradient descent algorithm for the composite sparse optimization, *AIMS Math.*, **9** (2024), 33401–33422. https://doi.org/10.3934/math.20241594
7. A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, D. Xu, Swin UNETR: swin transformers for semantic segmentation of brain tumors in mri images, In: *Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries*, Cham: Springer, 2021, 272–284. https://doi.org/10.1007/978-3-031-08999-2_22
8. J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, *Nat. Commun.*, **15** (2024), 654. https://doi.org/10.1038/s41467-024-44824-z
9. P. Scotti, A. Banerjee, J. Goode, S. Shabalin, A. Nguyen, A. Dempster, et al., Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors, *Advances in Neural Information Processing Systems*, **36** (2023), 24705–24728. https://doi.org/10.52202/075280-1073
10. Y. Zhao, G. Dong, L. Zhu, X. Ying, Memory recall: retrieval-augmented mind reconstruction for brain decoding, *Inform. Fusion*, **123** (2025), 103280. https://doi.org/10.1016/j.inffus.2025.103280
11. A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv:2204.06125.
12. C. Li, D. Zhao, Restoring medical images with combined noise base on the nonlinear inverse scale space method, *Math. Model. Control*, **5** (2025), 216–235. https://doi.org/10.3934/mmc.2025016
13. K. Tansri, P. Chansangiam, Gradient-descent iterative algorithm for solving exact and weighted least-squares solutions of rectangular linear systems, *AIMS Math.*, **8** (2023), 11781–11798. https://doi.org/10.3934/math.2023596
14. C.-H. Yeh, C.-Y. Hong, Y.-C. Hsu, T.-L. Liu, Y. Chen, Y. LeCun, Decoupled contrastive learning, In: *Computer vision — ECCV 2022*, Cham: Springer, 2022, 668–684. https://doi.org/10.1007/978-3-031-19809-0_38
15. X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, In: *2023 IEEE/CVF international conference on computer vision (ICCV)*, Paris, France, 2023, 11941–11952. https://doi.org/10.1109/ICCV51070.2023.01100
16. Y. Zhao, G. Dong, R. Duan, L. Zhu, X. Ying, Brainseek: a neural-driven deep semantic reasoning framework from functional magnetic resonance imaging signals, *Eng. Appl. Artif. Intel.*, **162** (2025), 112740. https://doi.org/10.1016/j.engappai.2025.112740
17. Y. Cheng, Y. Zhang, X. Zha, D. Wang, On stochastic accelerated gradient with non-strongly convexity, *AIMS Math.*, **7** (2022), 1445–1459. https://doi.org/10.3934/math.2022085
18. C. Audet, J. Bigeon, R. Couderc, M. Kokkolaras, Sequential stochastic blackbox optimization with zeroth-order gradient estimators, *AIMS Math.*, **8** (2023) 25922–25956. https://doi.org/10.3934/math.20231321

19. M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, M. Lucic, On mutual information maximization for representation learning, arXiv:1907.13625.

20. E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, et al., A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence, *Nat. Neurosci.*, **25** (2022), 116–126. https://doi.org/10.1038/s41593-021-00962-x

21. F. Ozcelik, R. VanRullen, Natural scene reconstruction from fmri signals using generative latent diffusion, *Sci. Rep.*, **13** (2023), 15666. https://doi.org/10.1038/s41598-023-42891-8

AIMS Press