# Global Convergence and Geometry of Contrastive Learning through Temperature Annealing

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Contrastive learning with the InfoNCE loss critically depends on the temperature parameter, yet its principled scheduling remains poorly understood, with many analyses leaving it fixed. We present a theory for asymptotic global convergence in InfoNCE with temperature annealing by casting each local InfoNCE term as a Gibbs free-energy on a compact Reimaniann manifold with embeddings modeled under Langevin dynamics. Under mild smoothness and energy-barrier assumptions, and a strong structural condition on the similarity gaps and energy barriers, we prove that key results from classical simulated annealing hold. In particular, under sufficiently slow logarithmic inverse temperature schedules, embeddings converge in probability to the global minimizers of the limiting contrastive potential; conversely, inverse temperature schedules that grow asymptotically faster than a critical rate risk being trapped in suboptimal minima. A further geometric analysis discusses how manifold compactness, thermally driven exploration, and the Hessian sharpening jointly enable escape from local basins. Small-scale empirical evaluation on CIFAR-10 with ResNet-18 verifies that slow annealing schedules can avoid the pitfalls of fixed temperatures. Our results offer a principled foundation for designing and tuning temperature schedules in modern contrastive representation learning.

## 1 Introduction

Contrastive learning has emerged as a crucial paradigm in representation learning. A common formulation in this domain is the Information Noise-Contrastive Estimation (InfoNCE) loss, which encourages an anchor embedding $z$ to be close to its positive embedding $z^+$ while repelling negative embeddings $z^-$ [1, 2]. Central to these formulations is the temperature parameter $\tau$ (or, equivalently, the inverse temperature $\beta = 1/\tau$), which modulates the sharpness of the similarity distribution.

For a given anchor $z_i$, its positive $z_j$, and negatives $\{z_k : k \neq i\}$, the InfoNCE loss is defined as

$$\ell_{i,j} = -\log \frac{\exp\left(\beta \operatorname{sim}\left(z_i, z_j\right)\right)}{\sum_{k \neq i} \exp\left(\beta \operatorname{sim}\left(z_i, z_k\right)\right)} \tag{1}$$

where $\operatorname{sim}(\cdot, \cdot)$ is a similarity function (such as cosine similarity). The overall contrastive loss is then defined as an average over a set $\mathcal{P}$ of positive pairs:

$$\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \ell_{i,j}. \tag{2}$$

A key observation is that as $\tau \to 0$ (i.e., $\beta \to \infty$), the loss increasingly penalizes mismatches and thus enforces stronger separation. However, if the temperature $\tau$ is dropped too rapidly, the optimizer

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

may “freeze” in a suboptimal local minimum. This tradeoff is reminiscent of simulated annealing, where a carefully controlled temperature schedule is essential for both exploration in the early stages and eventual convergence to a global minimum. Despite the empirical use of fixed or heuristically decayed temperatures in many contrastive learning frameworks [3, 4] and some initial studies on adaptive schedules [5, 6], a theoretical understanding of how and why tuning temperature guarantees convergence to high-quality representations has been lacking.

In this work, we address this gap by recasting InfoNCE into a local Gibbs-like form to apply arguments analogous to simulated annealing. We model the evolution of embeddings via a stochastic differential equation (SDE) with an inverse temperature schedule $\beta(t)$. We prove that if $\beta(t)$ increases sufficiently slowly (logarithmically, $\beta(t) = c\ln(t+K)$ with $c \le c^*$), the dynamics converge to a global optimum; conversely, if $\beta(t)$ grows too quickly ($c > c^*$), convergence can fail. Here, $c^* = 1/\Delta E_{\max}$ is the critical schedule constant derived from the maximum energy barrier required to escape local minima [7]. A strong assumption on the similarity gap structure $\Delta s_{\min}$ relative to the maximum energy barrier $\Delta E_{\max}$ is required by our current proof technique. We discuss this condition and its potential relaxation in the proof. We then relate this to a geometric interpretation involving Riemannian structures, clarifying how curvature evolves under annealing. Finally, we perform small-scale empirical validation on CIFAR-10 with finite-time schedules for illustration.

**Contributions.** Our main contributions are:

- We formulate InfoNCE annealing as an SDE on a compact Reimannian manifold and connect it to simulated annealing theory.
- We prove that under a logarithmic schedule $\beta(t) = c\ln(t+K)$ with $c \le c^*$ (where $c^*$ is the critical schedule constant) and a specific landscape condition ($\Delta s_{\min} > \Delta E_{\max}$), the SDE dynamics converge in probability to the global minimizers of the limiting potential (Theorem 3.1).
- We provide a matching non-convergence result (Proposition 3.1), showing th at if $\beta(t)$ grows too quickly (i.e., with $c > c^*$), the dynamics can get stuck in sub-optimal basins.
- We provide a geometric analysis on the hypersphere product manifold $\mathcal{M} = (\mathbb{S}^{d-1})^N$, including deriving the InfoNCE Hessian (Appendix B), showing at least linear $O(\beta)$ sharpening of the landscape away from optima, which aids convergence.
- We empirically check our theoretical insights on CIFAR-10 with a ResNet-18 backbone, illustrating that the shape of the annealing schedule impacts convergence speed of the loss function.

### 1.1 Related Work

Contrastive learning has been extensively explored both empirically and theoretically in recent years. Built on the Noise-Contrastive Estimation principle [2], early instance-discrimination approaches [8] paved the way for contemporary methods such as SimCLR [3], MoCo [9], and BYOL [10]. These frameworks typically use a fixed temperature parameter, but recent works have begun to investigate adaptive schedules or even eliminate it. For instance, Kukleva et al. [5] propose temperature schedules specifically tailored to long-tail data distributions, showing improved performance on underrepresented classes. Qiu et al. [6] develop an automatic temperature individualization strategy, arguing that different semantic classes require different scales of separation. Kim and Kim [11] go a step further by proposing a temperature-free loss, eliminating the need for explicit hyperparameter tuning altogether.

On the theoretical side, prior works have analyzed the properties and generalization aspects of contrastive objectives under fixed temperature assumptions [12]. In contrast, our work studies the role of a time-varying temperature schedule, drawing inspiration from classical simulated annealing results [13, 7]. This connection is further underpinned by stochastic approximation theory [14] and perspectives that view stochastic gradient descent (SGD) as approximate Bayesian inference [15] or, more broadly, a form of Langevin dynamics.

Our analysis is also informed by recent advances in understanding the geometry of parameter spaces in deep learning. In particular, the notion that the Hessian of a loss function induces a natural Riemannian metric, central to natural gradient methods [16], and the related literature on Wasserstein gradient flows [17, 18] provide valuable insights into the dynamics of contrastive learning. Additionally, recent work by Wang et al. [19] explores how augmentations interact with the loss landscape in contrastive training and the importance of carefully controlling temperature as a parameter of sharpness.

In summary, while prior literature has touched upon aspects of contrastive learning, temperature tuning, and the geometry of deep networks, our work integrates these themes to answer the question: "when and why does temperature scheduling yield global (or near-global) optima in contrastive learning?". By modeling the training dynamics via a time-varying inverse temperature schedule and SDEs, we connect classical annealing theory with the modern practice of contrastive learning to provide asymptotic global convergence guarantees. Moreover, we reveal the underlying geometric structure governing the evolution of embeddings under temperature change, complementing ongoing efforts to bridge the gap between theory and practice [20].

# 2 Problem Setup

In this section, we formalize the problem and establish the notation that will be used throughout the paper. The goal is to motivate a reformulation of the InfoNCE loss as a local Gibbs distribution, which naturally leads to an annealing approach.

## 2.1 Data Generating Process and Latent Space

To motivate our manifold-based approach, one can conceptualize an underlying latent variable model. Let $Z \subset \mathbb{R}^d$ be the latent space. In many practical scenarios, it is convenient to assume that the latent space is a $(d-1)$-dimensional hypersphere, i.e., $Z = \mathbb{S}^{d-1}$, as embeddings can easily be L2-normalized to provide compactness. Let $X \subset \mathbb{R}^D$ denote the observation space, and assume there exists a generative mapping $g : Z \to X$ from the latent space to the observation space that allows us to use representations to encode data points. Furthermore, assume that $g$ is invertible (or approximately invertible) to ensure that latent representations can be recovered.

Further assume that the ground-truth latent $z \in Z$ follows a distribution $p(z)$ (often uniform on $\mathbb{S}^{d-1}$). Positive pairs in contrastive learning are generated by applying data augmentations. In computer vision, these could be cropping, zooming, or rotating, for example. We model the effect of augmentations as perturbations in the latent space. Specifically, we assume that positive samples $z^+$ are drawn from a conditional distribution

$$p(z^+|z) \propto \exp\left(-(z^+ - z)^\top \Lambda (z^+ - z)\right),$$

where $\Lambda$ is a diagonal matrix capturing different concentration parameters for each latent dimension. This formulation conveys that perturbations are close together on the manifold, which is precisely the intuition behind the InfoNCE loss function. The use of $\Lambda$ allows for anisotropy: some latent dimensions may be perturbed more strongly than others.

## 2.2 Contrastive Loss as a Local Gibbs Free Energy

**Local Gibbs perspective.** Define an "energy" for each anchor-candidate pair $(i, k)$ where $k \neq i$:

$$E_i(k) \;=\; -\operatorname{sim}(z_i, z_k).$$

The local partition function for anchor $i$ over its candidates is $Z_i(\beta) = \sum_{k \neq i} \exp(-\beta\, E_i(k))$, yielding a probability distribution over candidates:

$$p_i(k \mid \beta) = \frac{\exp(-\beta\, E_i(k))}{Z_i(\beta)} = \frac{\exp(\beta \operatorname{sim}(z_i, z_k))}{\sum_{l \neq i} \exp(\beta \operatorname{sim}(z_i, z_l))}.$$

This is precisely a Gibbs-Boltzmann distribution. The InfoNCE loss term relates directly to the probability of selecting the positive sample $j$, as $\exp[-\ell_{i,j}] = p_i(j \mid \beta)$. As $\beta \to \infty$, this distribution concentrates sharply on the candidate(s) $k^*$ maximizing similarity (minimizing energy) with $z_i$.

**Free-energy form.** The loss term can be expressed in a form analogous to the Helmholtz free energy ($F = E - TS$). Specifically, the loss scaled by temperature ($T = 1/\beta$) reveals this structure:

$$\frac{\ell_{i,j}}{\beta} \;=\; \underbrace{-\operatorname{sim}(z_i, z_j)}_{\text{Energy } E_i(j)} + \underbrace{\frac{1}{\beta} \log Z_i(\beta)}_{\substack{\text{Entropic/Log-Partition} \\ \approx -\text{Temperature}\times\text{Entropy}}} .$$

The first term minimizes energy by pulling the positive $z_j$ closer, while the second term penalizes configurations where $z_i$ is highly similar to many negatives.

**Connection to Annealing.** The overall contrastive objective $\mathcal{L}(Z, \beta) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \ell_{i,j}$ is a sum of local free energies. Note, the dynamics of training are driven by the gradient of $\mathcal{L}$, not the scaled version $\mathcal{L}/\beta$. This local-free-energy view suggests training as a time-varying Gibbs sampler: growing $\beta = \beta(t)$ while injecting Langevin-like noise via small Gaussian perturbations to the gradient yields precisely the SDE of Section 3, allowing us to import classical simulated annealing convergence theory.

### 2.3 Modeling Assumptions

For our theoretical analysis, we impose several key assumptions regarding the similarity function, the latent space structure, and the modeled training dynamics. The precise mathematical formulations required for the convergence proofs are detailed in Appendix A.1 (Assumptions A1-A7). Here we provide an overview:

1. **Bounded Similarity:** We assume $\mathrm{sim}(z, z')$ is bounded. For instance, if we use cosine similarity, we have $\mathrm{sim}(z, z') \in [-1, 1]$. This boundedness guarantees that the exponential terms in the InfoNCE loss, $\exp\big(\beta\,\mathrm{sim}(z, z')\big)$, remain finite for all $\beta$, thereby ensuring that each local Gibbs distribution is well-defined.
2. **Embedding Manifold:** We assume the full state of $N$ embeddings, $Z = (z_1, \ldots, z_N)$, lies on a general compact, connected Riemannian manifold $\mathcal{M}$. This compactness is crucial for guaranteeing that energy barriers are finite. For our specific geometric analysis in Section 4 and to connect with common practice of $\ell_2$-normalization, we will consider the important special case where this manifold is the product of unit hyperspheres, $\mathcal{M} = (\mathbb{S}^{d-1})^N$ to explicitly discuss the geometry.
3. **Smoothness:** We further assume $\mathrm{sim}(\cdot, \cdot)$ is sufficiently smooth (specifically $C^2$, see Assumption A3) and is itself Lipschitz continuous. In other words, small changes in the embeddings lead to bounded changes in the similarity and its gradients. This property is crucial for applying gradient-based convergence analyses and controlling the log-sum-exp partition terms.
4. **Anisotropy in the Positive Conditional:** As discussed, we model augmentations via an anisotropic conditional distribution $p(z^+|z)$, parameterized by a diagonal matrix $\Lambda$ with strictly positive entries. This ensures each latent dimension can be perturbed differently, aligning with the diversity of real-world augmentations (e.g. some dimensions may be more "sensitive" to transformations than others).
5. **Langevin Dynamics:** Training dynamics are modeled using Langevin dynamics to approximate the discrete-time process of training with mini-batch SGD or its variants. This perspective is a standard and powerful technique in theoretical machine learning. However, this modeling choice inherently abstracts away certain details of practical optimization, such as the precise covariance structure of mini-batch gradient noise (often anisotropic and data-dependent) compared to the assumed Brownian motion, as well as the specific update rules and momentum/adaptive aspects of optimizers like Adam [21].
6. **Structural Landscape Condition:** Our main convergence proof requires a technical condition on the structure of the loss landscape, relating the geometry of similarity scores to the energy barriers of the system. We formalize this via a minimum similarity gap, denoted $\Delta s_{\min}$, which assumes a minimum separation between the most similar negative example and the next-most similar one (see Assumption A7 for the precise definition).

These assumptions collectively enable us to recast the InfoNCE loss in a local Gibbs framework and apply classical simulated annealing arguments for global convergence. In practice, many are often met or closely approximated via normalization techniques (for bounding norms), careful choice of $\mathrm{sim}(\cdot, \cdot)$, and standard data-augmentation pipelines.

## 3 Temperature Annealing and Convergence

Section 2 established that the InfoNCE objective $\mathcal{L}(Z, \beta)$ can be viewed as a sum of local free energies, where the inverse temperature $\beta$ controls the sharpness of the underlying local Gibbs distributions. This perspective naturally motivates using techniques from simulated annealing, where $\beta$ is increased over time, to guide the system towards a global minimum asymptotically. Our main goal in this section is to formalize this connection and state the resulting convergence guarantees. We model the embedding evolution via continuous-time stochastic gradient flow and then present the key theorems regarding convergence under specific annealing schedules.

### 3.1 Modeling the Stochastic Gradient Flow

While practical training uses discrete updates, it is standard and conceptually simpler for theoretical analysis to consider the continuous-time limit. We model the evolution of the full embedding state $Z_t \in \mathcal{M}$ using the overdamped Langevin diffusion on the manifold $\mathcal{M}$ (Assumption A2), driven by the InfoNCE loss potential $\mathcal{L}(Z, \beta(t))$ and subject to thermal noise scaled by the time-varying temperature $\beta(t) = 1/\tau(t)$:

$$\mathrm{d}Z_t \;=\; -\mathrm{grad}\, \mathcal{L}(Z_t, \beta(t))\, \mathrm{d}t \;+\; \sqrt{2/\beta(t)}\, \mathrm{d}\mathbf{W}_t^{\mathcal{M}}. \tag{3}$$

Here, $\mathrm{grad}$ is the Riemannian gradient on $\mathcal{M}$ and $\mathbf{W}_t^{\mathcal{M}}$ is standard Brownian motion on $\mathcal{M}$ (Assumption A1). This SDE models the exploration-exploitation dynamics inherent in annealing.

**Equilibrium at Fixed Temperature.** Before considering time-varying $\beta(t)$, we recall the equilibrium behavior for a fixed inverse temperature $\beta > 0$. As formally stated and proven in Appendix A.4, under Assumptions A2 and A3, the SDE (3) with constant $\beta$ admits a unique stationary Gibbs-Boltzmann distribution:

$$\pi_\beta(\mathrm{d}Z) \;=\; \frac{1}{\mathcal{Z}_\beta} \exp\!\big[-\beta\, \mathcal{L}(Z, \beta)\big]\, d\mu(Z). \tag{4}$$

As $\beta \to \infty$, this distribution concentrates its mass on the global minimizers of the potential $\mathcal{L}(Z, \beta)$ (or its limiting form $U_0(Z)$), as formally characterized in Appendix A.5.

**Time-Varying Annealing Schedule.** Simulated annealing leverages this concentration property by slowly increasing $\beta(t)$. If $\beta(t)$ increases slowly enough, the system can escape local minima while the temperature is high ($\beta$ is small) and then "freeze" into the global minimum as the temperature drops ($\beta \to \infty$). The critical question is how slowly $\beta(t)$ must increase.

### 3.2 Convergence Guarantees for Annealing Schedules

Classical simulated annealing theory provides precise conditions on the schedule $\beta(t)$ for guaranteed convergence. Adapting these results to our specific time-inhomogeneous SDE (3) yields the following key results.

**Theorem 3.1** (Global Convergence for Logarithmic Annealing)**.** *Under Assumptions A1 through A7 (detailed in Appendix A.1), let $Z_t$ be the solution to the SDE* (3)*. If the landscape satisfies the structural condition $\Delta s_{\min} > \Delta E_{\max}$ (where $\Delta E_{\max} = 1/c^*$ is from Assumption A5), and the schedule (Assumption A6) uses a coefficient c chosen such that $1/\Delta s_{\min} < c \leq c^*$, then $Z_t$ converges in probability to the set $U^*$ of global minimizers of $U_0(Z)$: for any $\epsilon > 0$,*

$$\lim_{t \to \infty} \mathbb{P}\big(Z_t \in \mathcal{N}(U^*, \epsilon)\big) = 1.$$

Here, $\mathcal{N}(U^*, \epsilon) = \{Z \in \mathcal{M} \mid \inf_{Y \in U^*} d(Z, Y) < \epsilon\}$ denotes an $\epsilon$-neighborhood of the set $U^*$ under the Riemannian metric $d(\cdot, \cdot)$ of the manifold $\mathcal{M}$. This theorem provides the central guarantee: a sufficiently slow logarithmic inverse temperature schedule ensures the SDE dynamics find the optimal configuration corresponding to perfect contrastive separation. Conversely, annealing inverse temperature too quickly violates the conditions needed to guarantee escape from all local minima.

**Proposition 3.1** (Non-Convergence for Rapid Annealing)**.** *Let Assumptions A1 through A5 hold. If the logarithmic annealing schedule $\beta(t)$ grows too quickly, specifically* $\liminf_{t \to \infty} \frac{\beta(t)}{\ln t} = c' > c^*$, *then there exists a set of initial conditions with positive measure from which the process $Z_t$ defined by the SDE* (3) *converges to a suboptimal local minimum basin of $U_0(Z)$ with positive probability. That is for any sufficiently small $\epsilon > 0$:*

$$\limsup_{t\to\infty} \mathbb{P}\big(Z_t \notin \mathcal{N}(U^*, \epsilon)\big) > 0.$$

This result highlights the precarious choice of annealing rate; faster schedules risk premature convergence to suboptimal representations. The proofs in Appendix A detail how these results adapt classical annealing arguments [22, 7, 23] to handle the specific time-varying potential $\mathcal{L}(Z, \beta(t))$.

### 3.3 Connection to Discrete Optimization

While our main theoretical results concern the continuous-time SDE (3), practical training uses discrete updates (e.g., SGD/Adam). A corresponding annealed SGD update can be formulated as:

$$Z_{k+1} = \Pi_{\mathcal{M}}\Big[Z_k - \eta_k\, \widehat{\nabla}\mathcal{L}(Z_k, \beta_k) + \sqrt{2\,\eta_k/\beta_k}\,\xi_k\Big],$$

where $\eta_k$ is the learning rate, $\beta_k = \beta(t_k)$ the discrete annealing schedule, and $\xi_k \sim \mathcal{N}(0, I)$. Classical stochastic approximation theory (e.g., [24, 25, 26]) provides conditions under which such discrete recursions track their limiting SDEs. These typically involve standard learning rate decay (e.g., $\sum_k \eta_k = \infty, \sum_k \eta_k^2 < \infty$), appropriate scaling of the injected noise (e.g., such that $\eta_k\beta_k \to 0$), and regularity of the gradient noise. Although a full adaptation for our specific manifold InfoNCE setting is beyond this paper's scope (see Section 6.1), these established results offer strong theoretical guidance for designing effective discrete annealing schedules that approximate the SDE dynamics analyzed herein.

## 4 Geometric Structure of Contrastive Embeddings

Having established the connection to simulated annealing and the convergence guarantees under the SDE model in Section 3, we now delve deeper into the geometric aspects of the process. Specifically, we examine how constraining embeddings to the hypersphere $\mathbb{S}^{d-1}$ influences the dynamics and how the geometry interacts with the annealing schedule $\beta(t)$. This geometric perspective provides further intuition for why annealing helps escape local minima.

### 4.1 Spherical Geometry and Embedding Constraints

Having established our convergence results on a general compact Riemannian manifold, we now specialize our analysis to provide concrete geometric intuition. As outlined in Section 2.3, we focus on the important case where embeddings $z_i$ are constrained to the unit hypersphere, $\|z_i\| = 1$, so $z_i \in \mathbb{S}^{d-1}$. The full state space is thus the product manifold $\mathcal{M} = (\mathbb{S}^{d-1})^N$.

**Manifold Properties.** Equipped with the product Riemannian metric (derived from the standard round metric on each $\mathbb{S}^{d-1}$ factor), $\mathcal{M}$ possesses several crucial properties relevant to the dynamics and analysis. As a product of compact spaces, $\mathcal{M}$ is compact. This guarantees that the loss function $\mathcal{L}(Z, \beta)$ and its limiting form $U_0(Z)$ attain their minima, that energy barriers between basins are finite, and that the SDE dynamics do not diverge to infinity. Furthermore, each $\mathbb{S}^{d-1}$ factor has positive sectional curvature (+1). While the product manifold structure is more complex, this underlying curvature influences geodesic paths and the behavior of gradient flows. Finally, $\mathcal{M}$ is geodesically complete, ensuring that gradient flows (the noiseless dynamics) are well-defined for all time. These geometric properties underpin the applicability of standard results for diffusions on compact manifolds used in our proofs.

### 4.2 InfoNCE Dynamics as a Riemannian System

We now consider the InfoNCE objective $\mathcal{L}(Z, \beta)$ as a potential function defined on the Riemannian manifold $\mathcal{M}$.

**Riemannian Gradient and SDE.** The driving force for the dynamics must respect the manifold constraint, meaning the drift vector must lie in the tangent space $T_Z\mathcal{M}$ at each point $Z \in \mathcal{M}$. This is achieved using the Riemannian gradient, $\mathrm{grad}\,\mathcal{L}$, which is the orthogonal projection of the standard Euclidean gradient $\nabla\mathcal{L}$ onto the tangent space:

$$\mathrm{grad}_{z_i}\,\mathcal{L} \;=\; \Pi_{T_{z_i}\mathbb{S}^{d-1}}\big[\nabla_{z_i}\mathcal{L}(Z,\beta)\big] \quad \text{where} \quad \Pi_{T_{z_i}\mathbb{S}^{d-1}}(u) := u - \langle u, z_i\rangle z_i.$$

The SDE governing the annealing process (Assumption A1, Eq. (3)) uses this Riemannian gradient:

$$\mathrm{d}Z_t \;=\; -\mathrm{grad}\,\mathcal{L}(Z_t,\beta(t))\,\mathrm{d}t \;+\; \sqrt{2/\beta(t)}\,\mathrm{d}\mathbf{W}_t^{\mathcal{M}}.$$

This ensures that the trajectory $Z_t$ remains on the manifold $\mathcal{M}$ throughout the annealing process.

**Landscape Curvature and Sharpening.** While $\mathcal{L}(Z,\beta)$ is generally non-convex, its local curvature plays a role in annealing. As shown by the Hessian calculation (Appendix B, Eq. (9)), the curvature around local minima increases as $\beta$ grows. Specifically, away from the global optimum, the dominant term scales Hessian eigenvalues linearly with $\beta$, causing minima basins to become "sharper". This sharpening effect aids the annealing process: as $\beta(t)$ increases (temperature drops), the system is more strongly drawn towards minima, and once it finds the global minimum (facilitated by noise at higher temperatures), the increasing sharpness helps to "lock" it in place, making escape increasingly unlikely.

### 4.3 Geometric Interpretation of Annealing Escape

Theorem 3.1 guarantees convergence under slow cooling. From a geometric viewpoint on $\mathcal{M}$, this convergence arises from the interplay between the manifold structure, the noise, and the time-varying potential:

- **Finite Geodesic Barriers:** The compactness of $\mathcal{M}$ ensures finite energy barriers (thus finite $\Delta E_{\max}$) between local minima of the limiting potential $U_0(Z)$.
- **Noise-Driven Exploration:** At high temperatures (low $\beta(t)$), the diffusion term $\sqrt{2/\beta(t)}\,\mathrm{d}\mathbf{W}_t^{\mathcal{M}}$ is large, allowing the process $Z_t$ to explore widely across the manifold $\mathcal{M}$ and traverse these finite barriers, even if they correspond to "long" paths along the curved sphere surface.
- **Slow Cooling Enables Escape:** The logarithmic schedule (Assumption A6) ensures the noise term diminishes slowly enough ($\tau(t) \sim 1/\ln t$) relative to the landscape's energy barriers, providing time for the diffusion to find paths connecting different basins and escape suboptimal minima before the noise becomes too weak (as formalized by Proposition 3.1).
- **Landscape Sharpening:** As discussed, the increasing curvature for large $\beta(t)$ helps the system settle definitively into the global minimum once found.

Therefore, the combination of manifold compactness, sufficient noise maintained by slow cooling, and landscape sharpening ensures that suboptimal minima can be escaped and the global optimum attained almost surely.

### 4.4 Comparison to Fixed Temperature Geometric Analysis

Prior works analyzing contrastive learning geometrically often assume a fixed temperature $\tau$ (e.g., [27]). While gradient flow on $\mathbb{S}^{d-1}$ can still be considered, the fixed temperature creates a potential issue: if $\tau$ is too small (high $\beta$), the system might freeze in the first minimum it finds (poor exploration); if $\tau$ is too large (low $\beta$), the landscape might be too flat, leading to poor separation between positives and negatives even at equilibrium. The annealing approach, by effectively scanning through temperatures via increasing $\beta(t)$, dynamically adjusts the exploration-exploitation balance and avoids these fixed-$\tau$ pitfalls, robustly navigating the landscape towards the global optimum.

## 5 Empirical Validation

To validate our theory that an improperly chosen fixed temperature can drive the optimizer into suboptimal frozen minima, we conduct small-scale empirical validation comparing temperature annealing approaches against standard baselines on the CIFAR-10 dataset [28]. Our goal is to assess the finite-time performance of different schedule shapes within a practical training setup.

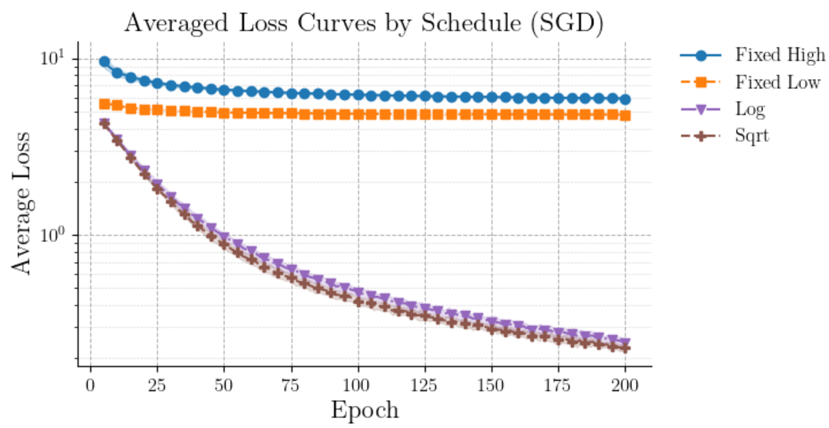


Figure 1: Average InfoNCE loss per epoch during pre-training on CIFAR-10. Curves show mean over 3 seeds.

### 5.1 Experimental Setup

We perform contrastive pre-training on CIFAR-10 using a ResNet-18 backbone [4] with a 2-layer MLP projection head ($d = 128$, L2-normalized), employing SimCLR-style augmentations. We compare four inverse temperature schedules over $T = 200$ epochs: fixed baselines `fixed_low` ($\beta = 1.0$) and `fixed_high` ($\beta = 1000000.0$), and annealing schedules `log` and `sqrt` interpolating between $\beta_{\text{low}} = 1.0$ and $\beta_{\text{high}} = 1000000.0$. For fair comparison over this finite horizon, bounded schedules allow assessment of the shape's impact within a practical range. To best mirror the theory, we use the SGD optimizer (lr $3 \times 10^{-4}$) with gradient clipping (norm 1.0) and batch size 128. Representation quality is measured by linear probe accuracy (logistic regression) on frozen backbone features. Results are averaged over seeds. Full experimental details, including schedule definitions, are provided in Appendix C.

The results comparing the final linear probe accuracy across different temperature schedules are presented in Table 1. The evolution of the average contrastive loss during pre-training is shown in Figure 1.

Table 1: CIFAR-10 Linear Probe Accuracy (%) after contrastive pretraining over 200 epochs (3 seeds, mean, min, max). Schedules anneal between $\beta_{\text{low}} = 1.0$ and $\beta_{\text{high}} = 1000000.0$.

| Schedule | Mean Acc (%) | Min Acc (%) | Max Acc (%) |
|---|---|---|---|
| `Fixed High` | 39.49 | 38.68 | 40.16 |
| `Fixed Low` | 37.02 | 36.46 | 37.27 |
| `Log` | 46.83 | 45.95 | 47.38 |
| `Sqrt` | 46.71 | 45.82 | 47.52 |

### 5.2 Discussion and Practical Guidelines

The results demonstrate two distinct failure modes for fixed-temperature schedules, as predicted. The `fixed_low` schedule plateaus early due to insufficient landscape sharpening leading to poor separation in the embedding space. The `fixed_high` schedule "freezes" almost immediately in a poor local minimum, leading to high final loss and poor accuracy.

In contrast, both the `log` and `sqrt` annealing schedules navigate the landscape better. By starting at a high temperature (low $\beta$) to enable exploration and gradually cooling (increasing $\beta$), they avoid the pitfalls of the fixed schedules and converge to a significantly better solution, achieving a 7-point accuracy improvement. This provides strong empirical evidence for the practical necessity of the annealing principle for non-adaptive optimizers, directly validating the core message of our theoretical analysis. While the `log` and `sqrt` shapes perform comparably in this finite setting, they can be clearly superior to the fixed-temperature baselines in even smaller dataset like CIFAR-10; larger datasets will only have more complex loss landscapes.

However, annealing schedules offer potential benefits beyond marginal accuracy gains. They provide a principled way to balance exploration and exploitation, avoiding potential optimization issues or representation degradation sometimes associated with overly aggressive fixed high temperatures [3, 27] where setting $\tau$ too small causes all features to map to the same point, i.e., representation collapse. The logarithmic schedule, in particular, is theoretically grounded and demonstrated competitive performance in the finite setting. However, in practice, literature has noted that in most applications, the logarithmic schedule is too slow and that a square root schedule empirically outperforms log-cooling: it drops temperature fast enough to converge in reasonable time yet remains slow enough to escape poor local minima [29, 30] and may serve as an effective practical alternative to fixed high inverse temperatures.

Finally, note that every schedule has a hyperparameter in terms of the $c$ factor. Preliminary exploration of schedule sensitivity to annealing rate is presented in Appendix D. A further discussion of the interaction between momentum-based optimizers as annealing can be found in Appendix E.

# 6 Conclusion

Our main theoretical contribution is the proof (Theorem 3.1) that a logarithmic annealing schedule, $\beta(t) = c\ln(t+K)$ where $c$ does not exceed the critical schedule constant $c^*$ of the limiting loss landscape, guarantees convergence in probability of the SDE dynamics to the globally optimal representations $U^*$. This optimal state corresponds to configurations with maximal anchor-positive similarity. Furthermore, our geometric analysis clarified the mechanisms underlying convergence on the compact manifold $\mathcal{M}$. We discussed how finite geodesic energy barriers, exploration driven by thermal noise (scaled by $1/\beta(t)$), the landscape sharpening captured by the Hessian analysis ($O(\beta)$ scaling, Appendix B), and the slow decay of noise under the logarithmic schedule collectively enable the system to escape local minima and reach the global optimum.

Our empirical validation on CIFAR-10, though limited in scale, supported the theory by demonstrating the failure of fixed low temperatures and the robustness achieved by schedules that anneal towards a sufficiently high $\beta$.

## 6.1 Limitations and Future Research Directions

The work rests on assumptions such as exact Langevin dynamics (Assumption A1), hyperspherical embeddings (Assumption A2), and a uniform similarity gap (Assumption A7) that only approximately hold in practice. Furthermore, the structural condition $\Delta s_{\min} > \Delta E_{\max}$ required by our current proof technique is notably very strong. While we have outlined potential relaxations for this worst-case requirement in Appendix A.2, future work might also consider an average-case analysis, aiming for convergence to high-quality representations with high probability, which could still offer significant empirical benefits.

Moreover, our convergence results are inherently asymptotic, and finite-time behavior (e.g., rates of convergence as a function of dimension, dataset size, or the barrier constant $c^*$) remains largely unexplored. Empirically, our validation on CIFAR-10 with ResNet-18 and the SGD optimizer serves as an initial proof of concept; scaling to larger datasets and architectures, potentially with unbounded annealing schedules, would be necessary to fully assess the practical benefits in diverse settings.

Looking ahead, several research directions are promising. Bridging the gap between continuous-time theory and discrete-time training by deriving comparable finite-time convergence results for practical optimizers like SGD or Adam, when coupled with appropriate step-size and annealing schedules, would enhance the direct applicability of these findings. Developing methods for estimating or bounding the critical schedule constant $c^*$ could inform more principled schedule design. Additionally, adaptive schemes that adjust the temperature $\tau$ in response to estimated curvature might offer faster practical convergence. Finally, extending this framework beyond the spherical manifold to other embedding geometries, such as hyperbolic spaces or Stiefel manifolds, or to alternative contrastive loss formulations, promises to disentangle the properties of InfoNCE from those of the manifold.

## References


[1] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL `https://arxiv.org/abs/1807.03748`.

[2] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304. PMLR, May 2010. URL `https://proceedings.mlr.press/v9/gutmann10a.html`.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[5] Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=ejHUr4nfHhD`.

[6] Zi-Hao Qiu, Quanqi Hu, Zhuoning Yuan, Denny Zhou, Lijun Zhang, and Tianbao Yang. Not all semantics are created equal: Contrastive self-supervised learning with automatic temperature individualization. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[7] Bruce Hajek. Cooling schedules for optimal annealing. *Mathematics of Operations Research*, 13(2):311–329, 1988. URL `http://www.jstor.org/stable/3689827`.

[8] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems, Volume 1*, pages 766–774, 2014.

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[11] Bum Jun Kim and Sang Woo Kim. Temperature-free loss function for contrastive learning. *arXiv preprint arXiv:2501.17683*, 2025. URL `https://arxiv.org/abs/2501.17683`.

[12] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. PMLR, Jun 2019. URL `https://proceedings.mlr.press/v97/saunshi19a.html`.

[13] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6): 721–741, 1984.

[14] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586. URL `https://doi.org/10.1214/aoms/1177729586`.

[15] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

[16] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998. URL `https://doi.org/10.1162/089976698300017746`.

[17] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Springer, 2008.

[18] Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1–2):101–174, 2001. URL `https://doi.org/10.1081/PDE-100002243`.

[19] Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=ECvgmYVyeUz`.

[20] Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S. Zimmermann, and Wieland Brendel. Infonce: Identifying the gap between theory and practice. *arXiv preprint arXiv:2407.00143*, 2024. URL `https://arxiv.org/abs/2407.00143`.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. URL `https://arxiv.org/abs/1412.6980`.

[22] Stuart Geman and Chii-Ruey Hwang. Diffusions for global optimization. *SIAM Journal on Control and Optimization*, 24(5):1031–1043, 1986.

[23] Basil Gidas. Nonstationary markov chains and convergence of the annealing algorithm. *Journal of Statistical Physics*, 39(1–2):73–131, 1985.

[24] Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 of *Applications of Mathematics*. Springer-Verlag, 2nd edition, 2003.

[25] Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[26] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*, volume 22. Springer-Verlag, 1990.

[27] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[28] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

[29] Emile Aarts and Jan Korst. *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*. John Wiley & Sons, 1989.

[30] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983. URL `https://www.science.org/doi/abs/10.1126/science.220.4598.671`.

[31] Elton P. Hsu. *Stochastic Analysis on Manifolds*, volume 38 of *Graduate Studies in Mathematics*. American Mathematical Society, 2002.

[32] Richard Holley and Daniel W Stroock. Simulated annealing via Sobolev inequalities. *Communications in Mathematical Physics*, 115(4):553–569, 1988.

[33] Olivier Catoni. Sharp large deviations estimates for simulated annealing algorithms. *Annales de l'I.H.P. Probabilités et Statistiques*, 27(3):291–383, 1991.

[34] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, 1984. ISBN 978-1-4684-0176-9. doi: 10.1007/978-1-4684-0176-9.

[35] Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*. Texts in Applied Mathematics. Springer, 2014. ISBN 978-1-4939-1322-0. doi: 10.1007/978-1-4939-1323-7.

[36] Shigeo Kusuoka and Daniel W. Stroock. Precise asymptotics of certain wiener functionals. *Journal of Functional Analysis*, 99(1):1–74, 1991. doi: 10.1016/0022-1236(91)90051-6. URL `https://www.sciencedirect.com/science/article/pii/0022123691900516`.

[37] Ziyin Liu, Ekdeep Singh Lubana, Masahito Ueda, and Hidenori Tanaka. What shapes the loss landscape of self-supervised learning? In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=3zSn48RUO8M`.

# A Proofs of Theoretical Results

## A.1 Assumptions for Convergence Theorems

**Assumption A1** (Langevin Dynamics Model)**.** *The evolution of the embedding vector $Z_t \in \mathcal{M}$ is modeled by the overdamped Langevin diffusion process on the manifold $\mathcal{M}$, governed by the SDE:*

$$\mathrm{d}Z_t \;=\; -\mathrm{grad}\,\mathcal{L}(Z_t, \beta(t))\,\mathrm{d}t \;+\; \sqrt{2/\beta(t)}\,\mathrm{d}\mathbf{W}_t^{\mathcal{M}}, \tag{5}$$

*where* grad *is the Riemannian gradient on $\mathcal{M}$, $\mathcal{L}(Z,\beta)$ is the InfoNCE loss, $\beta(t)$ is the time-varying inverse temperature, and $\mathbf{W}_t^{\mathcal{M}}$ is standard Brownian motion on $\mathcal{M}$.*

**Assumption A2** (Manifold)**.** *The embeddings $Z = (z_1, \ldots, z_N)$ are constrained to a compact, connected, Riemannian manifold without boundary.*

**Assumption A3** (Smoothness & Boundedness)**.** *The similarity function $\mathrm{sim}(z, z')$ is $C^2$-smooth with respect to its arguments $z, z' \in \mathbb{S}^{d-1}$, and is bounded, i.e., $|\mathrm{sim}(z, z')| \le S_{\max} < \infty$ for all $z, z'$. This ensures the InfoNCE loss $\mathcal{L}(Z,\beta)$ is $C^2$-smooth on $\mathcal{M}$ for finite $\beta$.*

**Assumption A4** (Limiting Potential & Minima)**.** *The limiting potential function*

$$U_0(Z) = \lim_{\beta\to\infty} \mathcal{L}(Z,\beta) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left[-\mathrm{sim}(z_i, z_j) + \max_{k\neq i} \mathrm{sim}(z_i, z_k)\right]$$

*exists, is $C^1$-smooth on $\mathcal{M}$, and possesses a non-empty set $U^* \subset \mathcal{M}$ of global minimizers. Assume $U_0$ has a finite number of critical points on $\mathcal{M}$.*

**Assumption A5** (Energy Barriers)**.** *Let $c^*$ be the critical schedule constant determined by the energy barriers of the potential $U_0(Z)$ on the manifold $\mathcal{M}$. Specifically, $c^* = 1/\Delta E_{\max}$ where $\Delta E_{\max}$ represents the maximum required "escape cost" for the diffusion process to reach $U^*$ from any starting point (related to Hajek's constants or Freidlin-Wentzell theory in terms of the maximum energy to escape any local basins). Assume $0 < \Delta E_{\max} < \inf$ so that $0 < c^* < \infty$. (This is guaranteed by Assumptions A2 and A4).*

**Assumption A6** (Annealing Schedule)**.** *The inverse temperature $\beta(t)$ is $C^1$, non-decreasing for $t \ge t_0$, satisfies $\beta(t) \to \infty$ as $t \to \infty$, and follows a logarithmic-type schedule with $\beta(t) = c\ln(t+K)$ for some $t_0, K \ge 0$ and a constant $0 < c \le c^*$.*

**Assumption A7** (Similarity Gaps - Technical)**.** *For any $z_i \in \mathbb{S}^{d-1}$, assume there exists a minimum similarity gap $\Delta s_{\min} > 0$ such that if $s_{ik^*} = \max_{k\neq i} s_{ik}$, then $s_{ik^*} - s_{ik} \ge \Delta s_{\min}$ for all $k \neq i, k^*$. (This technical assumption simplifies the analysis of gradient convergence rate; see Remark after Proof of Theorem 3.1).*

## A.2 Proof of Theorem 3.1 (Global Convergence)

**Theorem (Global Convergence).** *Under Assumptions A1 through A7, let $Z_t$ be the solution to the SDE* (3). *If the landscape satisfies the structural condition $\Delta s_{\min} > \Delta E_{\max}$ (where $\Delta E_{\max} = 1/c^*$ is from Assumption A5), and the schedule (Assumption A6) uses a coefficient $c$ chosen such that $1/\Delta s_{\min} < c \leq c^*$, then $Z_t$ converges in probability to the set $U^*$ of global minimizers of $U_0(Z)$, i.e., for any $\epsilon > 0$,*

$$\lim_{t\to\infty} \mathbb{P}\big(Z_t \in \mathcal{N}(U^*, \epsilon)\big) = 1.$$

*Proof.* The proof adapts classical results for simulated annealing via diffusion processes on compact manifolds [22, 31], specifically relating the required annealing rate to energy barriers [7]. The key challenge is the time-dependence of the potential $\mathcal{L}(Z, \beta(t))$, making the underlying Markov process time-inhomogeneous. Our strategy is to show that this time-dependence vanishes quickly enough for standard asymptotic SA theory (applied to the limiting potential $U_0(Z)$) to hold.

**Decomposition of Drift and Limiting Potential:** The SDE (5) describes dynamics under the drift $-\operatorname{grad} \mathcal{L}(Z_t, \beta(t))$. Let $U_0(Z)$ be the limiting potential defined in Assumption A4. We can write the drift as $-\operatorname{grad} U_0(Z_t) - \delta\mathbf{F}(Z_t, t)$, where $\delta\mathbf{F}(Z, t) = \operatorname{grad} \mathcal{L}(Z, \beta(t)) - \operatorname{grad} U_0(Z)$. We aim to show that the process asymptotically behaves like annealing under the fixed potential $U_0(Z)$.

**Convergence of the Gradient Perturbation $\delta\mathbf{F}$:** We analyze the convergence of $\operatorname{grad} \mathcal{L}(Z, \beta)$ to $\operatorname{grad} U_0(Z)$ as $\beta \to \infty$. The difference arises from the expectation term $\mathbb{E}_{k\sim p_i(k|\beta)}[\nabla_{z_i} s_{ik}]$ in the gradient of $\mathcal{L}$ compared to the term $\nabla_{z_i}(\max_{k\neq i} s_{ik}) = \nabla_{z_i} s_{ik^*(i)}$ in the gradient of $U_0$ (using Assumption A7 for uniqueness of $k^*(i)$ and differentiability). The probability $p_{ik}(\beta) = \exp(\beta s_{ik})/(\sum_l \exp(\beta s_{il}))$ concentrates exponentially fast on the maximizer $k^*(i)$ as $\beta \to \infty$: $|p_{ik}(\beta) - \delta_{k,k^*(i)}| \leq N e^{-\beta(s_{ik^*} - s_{ik})}$, where $N$ is the number of negatives. Using Assumption A7 ($s_{ik^*} - s_{ik} \geq \Delta s_{\min}$ for $k \neq k^*$), we have $|p_{ik}(\beta) - \delta_{k,k^*(i)}| \leq N e^{-\beta \Delta s_{\min}}$. Let $G_{\max} = \sup_{Z\in\mathcal{M},i,k} \|\nabla_{z_i} s_{ik}\| < \infty$ (which exists by Assumption A3 on the compact manifold A2). Then, for some constant $C_1$:

$$\big\|\mathbb{E}_{k\sim p_i(k|\beta)}[\nabla_{z_i} s_{ik}] - \nabla_{z_i} s_{ik^*(i)}\big\| \leq (N-1)(N e^{-\beta\Delta s_{\min}}) G_{\max} = C_1 e^{-\beta \Delta s_{\min}}.$$

The norm of the overall gradient perturbation is bounded (for some constant $C_2$):

$$\|\delta\mathbf{F}(Z, \beta)\| = \|\operatorname{grad} \mathcal{L}(Z, \beta) - \operatorname{grad} U_0(Z)\| \leq C_2 \beta e^{-\beta \Delta s_{\min}}.$$

(The extra factor of $\beta$ comes from the definition $\nabla \ell_i = \beta(\mathbb{E}[\nabla s_{ik}] - \nabla s_{ij})$). Substituting the annealing schedule $\beta(t) = c \ln(t + K)$ from Assumption A6 (where $0 < c \leq c^*$):

$$\|\delta\mathbf{F}(Z_t, t)\| \leq C_2 \beta(t) e^{-\beta(t)\Delta s_{\min}} \leq C_3 \ln(t+K)(t+K)^{-c\Delta s_{\min}},$$

for some constant $C_3$. Since $c > 0$ and $\Delta s_{\min} > 0$ (from Assumptions A6 and A7 respectively), this upper bound decays faster than any inverse polynomial $t^{-p}$ (for $p < c\Delta s_{\min}$). The term $(t+K)^{-c\Delta s_{\min}}$ ensures that $\|\delta\mathbf{F}(Z_t, t)\|$ decays to 0 as $t \to \infty$, with the overall rate being super-polynomial if $c\Delta s_{\min}$ is sufficently large relative to the $\ln(t + K)$ factor. For the integrability condition (6) discussed next, the requirement is that $c\Delta s_{\min} > 1$.

**Application of Time-Inhomogeneous Simulated Annealing Results:** For the dynamics of SDE (5) to be governed by the limiting potential $U_0(Z)$, convergence results in time-inhomogeneous simulated annealing (e.g., [23, 24]) typically require that the perturbation term $\delta\mathbf{F}(Z, t)$ diminishes sufficiently rapidly. A common strong sufficient condition is the integrability of its supremum norm:

$$\int_{t_0}^{\infty} \sup_Z \|\delta\mathbf{F}(Z, t)\| dt < \infty. \tag{6}$$

From the decay bound $\|\delta\mathbf{F}(Z_t, t)\| \leq C_3 \ln(t+K)(t+K)^{-c\Delta s_{\min}}$, this integral converges if and only if $c\Delta s_{\min} > 1$, or equivalently, $c > 1/\Delta s_{\min}$. Let this be Condition (I).

Independently, classical simulated annealing theory [7] for a fixed potential $U_0(Z)$ establishes that for an inverse temperature schedule $\beta(t) = c \ln(t + K)$, convergence to the global minimizers of $U_0(Z)$ occurs if and only if $c \leq 1/\Delta E_{\max}$, where $\Delta E_{\max}$ is the maximum energy barrier of $U_0(Z)$. From Assumption A5, our critical schedule constant $c^*$ is defined as $c^* = 1/\Delta E_{\max}$. Thus, the condition for convergence of the underlying SA process on $U_0(Z)$ is $c \leq c^*$. This is Condition (II), and it is satisfied by Assumption A6.

**Conclusion:** For the SDE (5) to converge to the minimizers of $U_0(Z)$, our proof strategy requires both the perturbation $\delta\mathbf{F}$ to be negligible in the limit (Condition I: $c > 1/\Delta s_{\min}$) and the annealing schedule for the limiting system $U_0(Z)$ to be sufficiently slow (Condition II: $c \le c^*$, which is $c \le 1/\Delta E_{\max}$). Therefore, under the stated assumptions, if a constant $c$ exists such that

$$\frac{1}{\Delta s_{\min}} < c \le c^* \quad \left(\text{i.e., } \frac{1}{\Delta s_{\min}} < c \le \frac{1}{\Delta E_{\max}}\right),$$

the SDE dynamics converge in probability to $U^*$. The existence of such a $c$ requires $1/\Delta s_{\min} < c^*$, which, by substituting $c^* = 1/\Delta E_{\max}$ (from Assumption A5), implies $\Delta s_{\min} > \Delta E_{\max}$. This constitutes a strong technical condition relating the landscape's similarity gaps to its energy barriers, necessary for this specific proof pathway. □

**Remark 1** (On the Structural Condition $\Delta s_{\min} > \Delta E_{\max}$)**.** *The proof of Theorem 3.1, as presented, relies on satisfying both $c > 1/\Delta s_{\min}$ (Condition I, for integrability of $\sup_Z \|\delta\mathbf{F}(Z,t)\|$) and $c \le c^*(= 1/\Delta E_{\max})$ (Condition II, for classical SA convergence on $U_0(Z)$). The theorem statement explicitly includes the necessary prerequisite $\Delta s_{\min} > \Delta E_{\max}$ for such a value of c to exist. This condition relates the minimum similarity gap (Assumption A7) to the maximum energy barrier of $U_0(Z)$ (related to Assumption A5).*

*This is a strong technical assumption on the landscape, arising because our proof uses the common, strong requirement of absolute integrability for the norm of the perturbation gradient $\delta\mathbf{F}$ (cf. [23, 24]). It is plausible that this constraint could be relaxed. More advanced results in time-inhomogeneous simulated annealing might establish convergence under weaker conditions on the decay of $\|\delta\mathbf{F}(Z,t)\|$. For instance, some frameworks only require the perturbation to vanish uniformly, i.e., $\lim_{t\to\infty}\sup_Z \|\delta\mathbf{F}(Z,t)\| = 0$ (which holds in our case if $c\Delta s_{\min} > 0$, given Assumptions A6 and A7), or satisfy other relative decay rates, when the primary SA condition ($c \le c^*$) is met (see e.g., general principles in Holley & Stroock [32], detailed analyses in Kushner & Yin [24], or Catoni [33] for trapping bounds under uniform perturbation decay). A detailed adaptation of such theorems is a promising direction for future work to potentially remove or weaken the $\Delta s_{\min} > \Delta E_{\max}$ requirement.*

**Remark 2** (Relaxing A7)**.** *Assumption A7 provides a simple way to establish the superpolynomial decay of the gradient perturbation $\delta\mathbf{F}$. This assumption might be relaxed. Even without a uniform gap, under the smoothness condition A3, the convergence $\lim_{\beta\to\infty} \operatorname{grad}\mathcal{L}(Z,\beta) = \operatorname{grad} U_0(Z)$ still holds pointwise. Arguments based on uniform convergence or dominated convergence might establish that $\int_{t_0}^{\infty} \sup_Z \|\delta\mathbf{F}(Z,t)\| dt < \infty$ still holds, ensuring the validity of step 3 without requiring Assumption A7. However, Assumption A7 yields a more direct rate calculation.*

### A.3 Proof of Proposition 3.1 (Non-Convergence for Rapid Annealing)

**Proposition** (**Non-Convergence for Rapid Annealing**)**.** *Let Assumptions A1 through A5 hold. If the logarithmic annealing schedule $\beta(t)$ grows too quickly, specifically $\liminf_{t\to\infty} \frac{\beta(t)}{\ln t} = c' > c^*$, then there exists a set of initial conditions with positive measure from which the process $Z_t$ defined by the SDE* (3) *converges to a suboptimal local minimum basin of $U_0(Z)$ with positive probability. That is for any sufficiently small $\epsilon > 0$:*

$$\limsup_{t\to\infty} \mathbb{P}\big(Z_t \notin \mathcal{N}(U^*, \epsilon)\big) > 0.$$

*Proof.* The proof adapts classical arguments from simulated annealing theory, particularly the necessity of a sufficiently slow cooling rate for convergence [7], to the current diffusion setting. We show that if the inverse temperature schedule $\beta(t)$ increases asymptotically faster than the critical rate (i.e., temperature cools too quickly), the integral representing an upper bound on the expected number of escapes over certain energy barriers converges. This implies trapping with positive probability via a Borel-Cantelli argument [33].

**Critical Annealing Rate:** Classical simulated annealing theory [7] establishes that for a schedule $\beta(t) \propto c\ln(t)$ for large $t$, convergence to the global minimizers $U^*$ of $U_0(Z)$ is guaranteed only if $c \le 1/\Delta E_{\max}$, where $\Delta E_{\max}$ is the maximum energy barrier of $U_0(Z)$. With $c^* = 1/\Delta E_{\max}$ (from Assumption A5), this means convergence requires the asymptotic rate coefficient of $\beta(t)/\ln t$ to be no more than $c^*$. The condition assumed in this proposition, $\liminf_{t\to\infty} \frac{\beta(t)}{\ln t} = c' > c^*$, directly violates this necessary condition for guaranteed convergence, indicating that the system cools too rapidly.

**Escape Rates and Energy Barriers:** Let $U_{local}$ be a suboptimal local minimum of $U_0(Z)$. Let $\Delta E_{trap}$ be the height of an energy barrier that must be overcome to escape the basin of $U_{local}$. We will specifically consider $\Delta E_{trap} = \Delta E_{\max}$, the largest such barrier relevant for reaching $U^*$ from any non-global state. The instantaneous escape rate from this basin across such a barrier scales as $k(t) \propto \exp(-\beta(t)\Delta E_{trap})$ according to Arrhenius-type laws derived from large deviation theory [34].

**Trapping under Rapid Annealing (Fast Cooling):** The proposition assumes $\liminf_{t\to\infty} \frac{\beta(t)}{\ln t} = c' > c^*$. By definition of $c^* = 1/\Delta E_{\max}$ (Assumption A5), this means $c' > 1/\Delta E_{\max}$. Since $c' > 1/\Delta E_{\max}$, we can choose a small $\epsilon_0 > 0$ such that the rate $c_{eff} := c' - \epsilon_0$ satisfies $c_{eff} > 1/\Delta E_{\max}$. By the definition of $\liminf$, there exists a time $t_0$ such that for all $t \geq t_0$:

$$\frac{\beta(t)}{\ln t} \geq c_{eff}.$$

Consider the barrier $\Delta E = \Delta E_{\max}$. Then, for $t \geq t_0$:

$$\beta(t)\Delta E_{\max} \geq c_{eff}\Delta E_{\max} \ln t.$$

Let $p_0 = c_{eff}\Delta E_{\max}$. Since $c_{eff} > 1/\Delta E_{\max}$ and $\Delta E_{\max} > 0$, we have $p_0 > 1$. The integral of an upper bound on the escape rate factor is then:

$$\int_{t_0}^{\infty} K_0 \exp(-\beta(t)\Delta E_{\max})dt \leq K_0 \int_{t_0}^{\infty} \exp(-c_{eff}\Delta E_{\max} \ln t)dt = K_0 \int_{t_0}^{\infty} t^{-p_0} dt,$$

where $K_0$ is a constant related to the attempt frequency. Since $p_0 > 1$, this integral converges.

**Trapping via Borel-Cantelli:** The convergence of the integral $\int_{t_0}^{\infty} K_0 \exp(-\beta(t)\Delta E_{\max})dt$ provides an upper bound on the expected number of escapes. To apply Borel-Cantelli, let time $t \geq t_0$ be partitioned by a sequence $t_k \to \infty$ (e.g., $t_k = k$). Let $E_k$ be the event that an escape over the barrier $\Delta E_{\max}$ occurs within the time interval $[t_k, t_{k+1})$. The probability of such an escape can be bounded by integrating the instantaneous rate:

$$\mathbb{P}[E_k] \leq \int_{t_k}^{t_{k+1}} K_0 \exp(-\beta(s)\Delta E_{\max})ds \leq K_0 \int_{t_k}^{t_{k+1}} s^{-p_0} ds.$$

Since $p_0 > 1$, the sum $\sum_k \int_{t_k}^{t_{k+1}} s^{-p_0} ds = \int_{t_0}^{\infty} s^{-p_0} ds < \infty$. Thus, $\sum_k \mathbb{P}[E_k] < \infty$. By the Borel-Cantelli lemma, this implies that $\mathbb{P}(E_k \text{ i.o.}) = 0$, meaning that almost surely, only a finite number of escapes over the barrier $\Delta E_{\max}$ (or any barrier $\Delta E_{trap}$ for which $c'\Delta E_{trap} > 1$) will occur. Therefore, if the process $Z_t$ is in a basin requiring an escape over such a barrier after the (random) time of the last likely escape, it will remain trapped with positive probability.

**Effect of Time-Varying Potential:** The preceding argument for trapping is based on the limiting potential $U_0(Z)$. For this to hold for the SDE (5) driven by $\mathcal{L}(Z, \beta(t))$, the perturbation term $\delta\mathbf{F}(Z,t) = \operatorname{grad}\mathcal{L}(Z,\beta(t)) - \operatorname{grad} U_0(Z)$ must not prevent trapping. As shown in the proof of Theorem 3.1 (Section A.2), $\|\delta\mathbf{F}(Z_t,t)\| \leq C_3 \ln(t+K)(t+K)^{-c'\Delta s_{\min}}$ uniformly in $Z$. For classical trapping arguments to apply robustly, results for time-inhomogeneous diffusions (e.g., Gidas, 1985 [23]; Kushner & Yin, 2003 [24]) require the influence of $\delta\mathbf{F}$ to be subdominant. A strong condition ensuring this is the integrability of its norm, $\int_{t_0}^{\infty} \sup_Z \|\delta\mathbf{F}(Z,t)\|dt < \infty$, which holds if $c'\Delta s_{\min} > 1$. If this condition is met, the system's dominant behavior is governed by annealing on $U_0(Z)$ with the rapidly cooling schedule $\beta(t) \approx c' \ln t$.

**Conclusion:** If the annealing schedule coefficient $c'$ in $\beta(t) \approx c' \ln t$ exceeds the critical schedule constant $c^*$ (i.e., $\beta(t)$ grows too quickly, $c' > 1/\Delta E_{\max}$), then the integrated escape probability over critical energy barriers (such as $\Delta E_{\max}$) converges. Provided the perturbation $\delta\mathbf{F}$ due to the time-varying potential is sufficiently well-behaved (e.g., its norm is integrable, which occurs if $c'\Delta s_{\min} > 1$), a Borel-Cantelli argument implies that trapping in suboptimal local minima of $U_0(Z)$ occurs with positive probability. Thus, $Z_t$ may fail to converge in probability to the global minimum set $U^*$. □

**Remark 3** (Super-logarithmic schedules and Rapid Annealing)**.** *Proposition 3.1 shows that if the coefficient $c'$ in $\beta(t) \approx c' \ln t$ satisfies $c' > c^*$, the process may become trapped. This highlights that an inverse temperature schedule growing asymptotically faster than the critical logarithmic rate leads to issues. Indeed, schedules where $\beta(t)$ grows even faster than logarithmically (e.g., polynomially like $\beta(t) \propto t^a$ for $a > 0$, or linearly $\beta(t) \propto t$) would also violate the convergence condition $c \leq c^*$ established for logarithmic schedules in Theorem 3.1. Such aggressive increases in $\beta(t)$ (representing very rapid cooling) quench thermal noise too quickly, preventing the necessary exploration to escape local minima. Our convergence guarantees in Theorem 3.1 are specific to the logarithmic schedule form in Assumption A6 where $\beta(t)$ grows sufficiently slowly.*

### A.4 Proof of Proposition 3.2 (Stationary Distribution at Fixed Temperature)

**Proposition** (**Stationary Distribution at Fixed Temperature**)**.** *For any fixed $\beta > 0$, let Assumptions A2 and A3 hold. Consider the time-homogeneous SDE corresponding to Eq.* (5) *with fixed $\beta$:*

$$\mathrm{d}Z_t = -\operatorname{grad} \mathcal{L}(Z, \beta)\, \mathrm{d}t + \sqrt{2/\beta}\, \mathrm{d}\mathbf{W}_t^{\mathcal{M}}. \tag{7}$$

*This process admits a unique stationary probability distribution $\pi_\beta(dZ)$ on $\mathcal{M}$, given by the Gibbs-Boltzmann distribution:*

$$\pi_\beta(dZ) = \frac{1}{\mathcal{Z}_\beta} \exp(-\beta \mathcal{L}(Z, \beta)) d\mu(Z),$$

*where $d\mu$ is the Riemannian volume measure on $\mathcal{M}$ and $\mathcal{Z}_\beta = \int_{\mathcal{M}} \exp(-\beta \mathcal{L}(Z, \beta)) d\mu(Z)$ is the normalization constant (partition function).*

*Proof.* The proof relies on standard results concerning the existence, uniqueness, and form of stationary distributions for non-degenerate diffusion processes on compact Riemannian manifolds [31, 35, 36].

**Properties of the SDE and Manifold:** The state space $\mathcal{M} = (\mathbb{S}^{d-1})^N$ is a compact, connected, smooth Riemannian manifold without boundary (Assumption A2). The drift term $b(Z) = -\operatorname{grad} \mathcal{L}(Z, \beta)$ is smooth ($C^1$) due to Assumption A3. The diffusion tensor associated with $\sqrt{2/\beta}\, \mathrm{d}\mathbf{W}_t^{\mathcal{M}}$ is $D = (1/\beta) g^{-1}$, where $g^{-1}$ is the inverse metric tensor. Since $\beta > 0$ and the metric is positive definite, the diffusion is uniformly elliptic (non-degenerate).

**Existence and Uniqueness of Stationary Distribution:** Uniformly elliptic diffusion processes with smooth drift on compact, connected manifolds are known to be strong Feller and topologically irreducible [31]. By standard ergodic theory for Markov processes, a strong Feller, topologically irreducible diffusion on a compact manifold admits a unique invariant probability measure (stationary distribution) $\pi_\beta$. The process is ergodic with respect to $\pi_\beta$.

**Identification of the Stationary Distribution:** The SDE (7) is a form of Langevin dynamics on the manifold $\mathcal{M}$ with potential energy function $V(Z) = \mathcal{L}(Z, \beta)$ and constant temperature $T = 1/\beta$. It is a well-established result that the unique stationary distribution for such dynamics, satisfying detailed balance, is the Gibbs-Boltzmann distribution [36, 35]:

$$\pi_\beta(dZ) \propto \exp\left(-\frac{V(Z)}{T}\right) d\mu(Z) = \exp\left(-\frac{\mathcal{L}(Z, \beta)}{1/\beta}\right) d\mu(Z) = \exp(-\beta \mathcal{L}(Z, \beta)) d\mu(Z).$$

The normalization constant $\mathcal{Z}_\beta = \int_{\mathcal{M}} \exp(-\beta \mathcal{L}(Z, \beta)) d\mu(Z)$ ensures $\int_{\mathcal{M}} \pi_\beta(dZ) = 1$. Finiteness of $\mathcal{Z}_\beta$ is guaranteed because $\mathcal{L}(Z, \beta)$ is continuous (by Assumption A3) on the compact manifold $\mathcal{M}$ (Assumption A2), hence bounded, making its exponential bounded and integrable over the finite volume of $\mathcal{M}$. □

### A.5 Proof of Proposition 3.3 (Characterization of Global Minima)

**Proposition** (**Characterization of Global Minima**)**.** *Let $S_{\max} = \sup_{z, z' \in \mathbb{S}^{d-1}} \operatorname{sim}(z, z')$. Assume this supremum is attainable (which holds if* $\operatorname{sim}$ *is continuous and* $\mathbb{S}^{d-1}$ *is compact). Under Assumption A4, a configuration $Z^* \in \mathcal{M}$ belongs to the set $U^*$ of global minimizers of the limiting potential $U_0(Z)$ if and only if for every positive pair $(i, j) \in \mathcal{P}$, the maximum possible similarity is achieved, i.e.,* $\operatorname{sim}(z_i^*, z_j^*) = S_{\max}$.

*Proof.* Recall the definition of the limiting potential from Assumption A4:

$$U_0(Z) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \underbrace{\left[-\mathrm{sim}(z_i, z_j) + \max_{k\neq i} \mathrm{sim}(z_i, z_k)\right]}_{T_{ij}(Z)}.$$

We seek configurations $Z^*$ that minimize $U_0(Z)$. Since $U_0$ is an average, minimizing $U_0$ is equivalent to minimizing each term $T_{ij}(Z)$ simultaneously for all $(i, j) \in \mathcal{P}$, if possible.

Consider a single term $T_{ij}(Z)$. Let $S_{ij} = \mathrm{sim}(z_i, z_j)$ and $S_{ik} = \mathrm{sim}(z_i, z_k)$. We know $S_{ik} \leq S_{\max}$ for all $i, k$ by definition of $S_{\max}$. The term $\max_{k\neq i} S_{ik}$ considers all similarities involving anchor $z_i$ except potentially $S_{ii}$ (if $i$ could be a negative for itself, which is usually excluded). Importantly, the positive sample $z_j$ is included among the candidates $k \neq i$. Therefore, $\max_{k\neq i} S_{ik} \geq S_{ij}$. This implies $T_{ij}(Z) = -S_{ij} + \max_{k\neq i} S_{ik} \geq -S_{ij} + S_{ij} = 0$. So, each term $T_{ij}(Z)$ is non-negative.

The minimum possible value for $T_{ij}(Z)$ is 0. This minimum is achieved if and only if $-S_{ij} + \max_{k\neq i} S_{ik} = 0$, which requires $\max_{k\neq i} S_{ik} = S_{ij}$. Since we also know $\max_{k\neq i} S_{ik} \leq S_{\max}$, achieving the minimum value of 0 requires $S_{ij} = S_{\max}$. If $S_{ij} = S_{\max}$, then automatically $\max_{k\neq i} S_{ik}$ must also be equal to $S_{\max}$ (as it's bounded by $S_{\max}$ but must be $\geq S_{ij}$). Thus, $T_{ij}(Z)$ achieves its minimum value of 0 if and only if $\mathrm{sim}(z_i, z_j) = S_{\max}$.

The overall potential $U_0(Z)$ is minimized when all terms $T_{ij}(Z)$ are simultaneously minimized, i.e., when $T_{ij}(Z^*) = 0$ for all $(i, j) \in \mathcal{P}$. This occurs if and only if $\mathrm{sim}(z_i^*, z_j^*) = S_{\max}$ for all $(i, j) \in \mathcal{P}$. The minimum value of $U_0(Z)$ is therefore 0. □

# B Hessian of the InfoNCE Loss

We derive the Hessian matrix of the InfoNCE loss for a single anchor embedding $z_i$ with respect to that anchor. This is similar to Ziyin et al. [37] but here the temperature parameter is left explicit. Let $z_j$ be the positive sample and $\{z_k\}$ be the set of all samples available to anchor $z_i$ (including $z_j$). Let $s_{ik} = \mathrm{sim}(z_i, z_k)$ denote the similarity function, and $\beta = 1/\tau$ be the inverse temperature.

The InfoNCE loss for anchor $z_i$ is given by:

$$\begin{aligned} l_i(z_i) &= -\log \frac{\exp(\beta s_{ij})}{\sum_k \exp(\beta s_{ik})} \\ &= \log\left(\sum_k \exp(\beta s_{ik})\right) - \beta s_{ij} \end{aligned}$$

Let $Z_i = \sum_k \exp(\beta s_{ik})$ be the partition function and $p_{ik} = \frac{\exp(\beta s_{ik})}{Z_i}$ be the softmax probability distribution over samples $k$ induced by anchor $i$. The loss can be written as $l_i = \log Z_i - \beta s_{ij}$.

Let $\nabla = \nabla_{z_i}$ denote the gradient operator with respect to $z_i$. The gradient of the similarity is $\nabla s_{ik} = \frac{\partial \mathrm{sim}(z_i, z_k)}{\partial z_i}$.

The gradient of the loss is:

$$\begin{aligned} \nabla l_i &= \nabla(\log Z_i) - \beta \nabla s_{ij} \\ &= \frac{1}{Z_i} \nabla Z_i - \beta \nabla s_{ij} \\ &= \frac{1}{Z_i} \sum_k \exp(\beta s_{ik}) \beta \nabla s_{ik} - \beta \nabla s_{ij} \\ &= \beta \sum_k p_{ik} \nabla s_{ik} - \beta \nabla s_{ij} \\ &= \beta(\mu_i - \nabla s_{ij}) \end{aligned}$$

where $\mu_i = \sum_k p_{ik} \nabla s_{ik} = \mathbb{E}_{k\sim p_i}[\nabla s_{ik}]$ is the expected similarity gradient under the distribution $p_i$.

The Hessian matrix $\mathbf{H} = \nabla(\nabla l_i)^T$ is obtained by differentiating the gradient:

$$\mathbf{H} = \nabla[\beta(\mu_i^T - (\nabla s_{ij})^T)] = \beta[\nabla\mu_i^T - \nabla(\nabla s_{ij})^T] \tag{8}$$

Let $\mathbf{H}_{ik} = \nabla(\nabla s_{ik})^T$ be the Hessian of the similarity function $s_{ik}$ with respect to $z_i$. The second term is simply $-\beta\mathbf{H}_{ij}$. For the first term, we use the product rule and the gradient of the softmax probabilities $\nabla p_{ik} = \beta p_{ik}(\nabla s_{ik} - \mu_i)$:

$$\begin{aligned}
\nabla\mu_i^T &= \nabla\left(\sum_k p_{ik}(\nabla s_{ik})^T\right) \\
&= \sum_k[(\nabla p_{ik})(\nabla s_{ik})^T + p_{ik}\nabla(\nabla s_{ik})^T] \\
&= \sum_k[\beta p_{ik}(\nabla s_{ik} - \mu_i)(\nabla s_{ik})^T + p_{ik}\mathbf{H}_{ik}] \\
&= \beta\sum_k p_{ik}(\nabla s_{ik})(\nabla s_{ik})^T - \beta\mu_i\sum_k p_{ik}(\nabla s_{ik})^T + \sum_k p_{ik}\mathbf{H}_{ik} \\
&= \beta\left(\sum_k p_{ik}(\nabla s_{ik})(\nabla s_{ik})^T - \mu_i\mu_i^T\right) + \sum_k p_{ik}\mathbf{H}_{ik} \\
&= \beta\,\mathrm{Cov}_{k\sim p_i}[\nabla s_{ik}] + \mathbb{E}_{k\sim p_i}[\mathbf{H}_{ik}]
\end{aligned}$$

where $\mathrm{Cov}_{k\sim p_i}[\nabla s_{ik}]$ is the covariance matrix of the similarity gradients under $p_i$, and $\mathbb{E}_{k\sim p_i}[\mathbf{H}_{ik}]$ is the expected Hessian of the similarity function.

Substituting back into the expression for $\mathbf{H}$:

$$\begin{aligned}
\mathbf{H} &= \beta[(\beta\,\mathrm{Cov}_{k\sim p_i}[\nabla s_{ik}] + \mathbb{E}_{k\sim p_i}[\mathbf{H}_{ik}]) - \mathbf{H}_{ij}] \\
&= \beta^2\,\mathrm{Cov}_{k\sim p_i}[\nabla s_{ik}] + \beta(\mathbb{E}_{k\sim p_i}[\mathbf{H}_{ik}] - \mathbf{H}_{ij})
\end{aligned} \tag{9}$$

Equation (9) shows that the Hessian of the InfoNCE loss consists of two terms. The first term involves the covariance of the similarity gradients and scales quadratically with $\beta$. The second term involves the expected Hessian of the similarity function (minus the Hessian for the positive pair) and scales linearly with $\beta$.

Analyzing the asymptotic behavior as $\beta \to \infty$ requires considering the limiting behavior of the distribution $p_i$. As $\beta$ increases, $p_{ik}$ concentrates its mass on the sample(s) $k^*$ maximizing the similarity $s_{ik}$. Consequently, the covariance term $\mathrm{Cov}_{k\sim p_i}[\nabla s_{ik}]$ vanishes because the expectation is taken over a distribution collapsing to one (or a few) points. Simultaneously, the expected Hessian $\mathbb{E}_{k\sim p_i}[\mathbf{H}_{ik}]$ converges to $\mathbf{H}_{ik^*}$.

Therefore, the asymptotic scaling of the Hessian depends on whether the positive sample $j$ is the most similar sample $k^*$:

- If $k^* \neq j$ (i.e., a negative sample is most similar to the anchor $z_i$, indicating a suboptimal configuration), the second term dominates:

  $$\mathbf{H} \approx \beta(\mathbf{H}_{ik^*} - \mathbf{H}_{ij}) \quad \text{as } \beta \to \infty \quad (\text{if } k^* \neq j)$$

  In this regime, the Hessian norm scales linearly, $||\mathbf{H}|| \sim O(\beta)$. This implies that away from the optimum, the local minima sharpen linearly with $\beta$.

- If $k^* = j$ (i.e., the positive sample is the most similar, corresponding to configurations near an optimum where the gradient $\nabla l_i \approx 0$), the second term vanishes. The scaling is then determined by the $\beta^2\mathrm{Cov}[\cdot]$ term. While the covariance vanishes, a more detailed analysis of the rate at which it vanishes relative to $\beta^2$ would be needed to determine the precise scaling at the optimum. However, the dominant scaling away from the optimum is linear.

This linear sharpening ($O(\beta)$) of the loss landscape curvature as temperature decreases contributes to the convergence behavior observed during annealing, complementing the theoretical escape guarantees provided by the slow decay of noise.

# C Experimental Setup Details

This appendix provides supplementary details for the empirical validation experiments presented in Section 5, conducted on the CIFAR-10 dataset. The code used to run these experiments as well as our generated results is available on our GitHub repository. [1]

## C.1 Dataset and Augmentations

We use the standard CIFAR-10 dataset [28], which consists of 50,000 training images and 10,000 test images across 10 classes, each of size 32x32 pixels.

**Contrastive Pre-training Augmentations:** Following a SimCLR-style [3] approach adapted for CIFAR-10 resolution, we generate two distinct augmented views $(v_1, v_2)$ from each input image $x$ during pre-training using the following sequence of transformations from 'torchvision.transforms':

```
T.Compose([
    T.RandomCrop(32, padding=4),
    T.RandomHorizontalFlip(p=0.5),
    T.RandomApply([
        T.ColorJitter(brightness=0.8, contrast=0.8,
                      saturation=0.8, hue=0.2)
    ], p=0.8),
    T.RandomGrayscale(p=0.2),
    T.ToTensor(),
    T.Normalize(mean=[0.4914, 0.4822, 0.4465],
                std=[0.2023, 0.1994, 0.2010])
])
```

The pair $(v_1, v_2)$ constitutes a positive example for the InfoNCE loss.

**Linear Probe Transforms:** For evaluating the learned representations via linear probing, both the training set (used to train the linear classifier) and the test set (used for final accuracy evaluation) are processed using only basic normalization:

```
T.Compose([
    T.ToTensor(),
    T.Normalize(mean=[0.4914, 0.4822, 0.4465],
                std=[0.2023, 0.1994, 0.2010])
])
```

## C.2 Model Architecture

The model used for contrastive pre-training comprises a ResNet backbone and an MLP projection head:

- **Backbone:** A standard ResNet-18 architecture [4], implemented via `torchvision.models.resnet18`, initialized with random weights (`weights=None`). The final fully connected classification layer (`fc`) is replaced by an `nn.Identity()` layer. The output feature dimension from the backbone is 512.
- **Projection Head:** A 2-layer MLP maps the 512-dimensional backbone features to the final 128-dimensional embedding space. It consists of a linear layer (512 $\to$ 512), followed by a ReLU activation, and a final linear layer (512 $\to$ 128).
- **Output Normalization:** The 128-dimensional output vector from the projection head is L2-normalized to lie on the unit hypersphere $\mathbb{S}^{127}$ before being used in the contrastive loss calculation.

[1] `https://anonymous.4open.science/r/contrastive-learning-temperature-schedules-E4D4`

### C.3 Training Hyperparameters

Contrastive pre-training was performed using the following settings:

- **Optimizer:** SGD.
- **Learning Rate:** $3 \times 10^{-4}$.
- **Weight Decay:** $1 \times 10^{-6}$.
- **Batch Size:** $B = 128$.
- **Epochs:** $T = 200$.
- **Gradient Clipping:** Gradients were clipped to have a maximum L2 norm of 1.0 before the optimizer step using `torch.nn.utils.clip_grad_norm_`.
- **Loss Function:** InfoNCE loss between paired views (Eq. 1, calculated using the stable implementation in the supplementary code).
- **Random Seeds:** Experiments were run with 3 different random seeds (3333, 3334, 3335). Results in the main paper report the mean and standard deviation across these seeds.

### C.4 Temperature Annealing Schedules

In our empirical validation (Section 5), we compare several fixed and annealing schedules for the inverse temperature $\beta(t)$, where $t \in \{0, 1, \ldots, T-1\}$ is the training epoch index and $T$ is the total number of epochs.

All annealing schedules are designed to interpolate between a starting inverse temperature $\beta_{\text{low}}$ and a target final inverse temperature $\beta_{\text{high}}$. This approach, while deviating from purely asymptotic schedules, allows for a controlled comparison of different schedule shapes over a finite training horizon commonly used in practice, ensuring numerical stability and a common target sharpness level. We use $\beta_{\text{low}} = 1.0$ and $\beta_{\text{high}} = 1000000.0$.

For the annealing schedules ($\texttt{log}, \texttt{linear}, \texttt{sqrt}$), we introduce a common scaling hyperparameter $c_{\text{factor}}$ (referred to as 'c_factor' in the code configuration) which multiplies the *change* from $\beta_{\text{low}}$. That is, if $\beta_{\text{base}}(t)$ is the base interpolated value for a schedule (progressing from $\beta_{\text{low}}$ to $\beta_{\text{high}}$), the actual beta used is $\beta(t) = \text{clip}(\beta_{\text{low}} + (\beta_{\text{base}}(t) - \beta_{\text{low}}) \cdot c_{\text{factor}}, \beta_{\text{low}}, \beta_{\text{high}})$. Unless otherwise specified (e.g., in sensitivity analysis in Appendix D), we use $c_{\text{factor}} = 0.01$.

The specific schedules tested are:

- `fixed_low`**:** Constant inverse temperature.
$$\beta(t) = \beta_{\text{low}} = 1.0$$
- `fixed_high`**:** Constant inverse temperature.
$$\beta(t) = \beta_{\text{high}} = 10000000.0$$
- `log`**:** Logarithmic increase based on the theoretical schedule, scaled to reach $\beta_{\text{high}}$ at epoch $T-1$.
$$c = \frac{\beta_{\text{high}} - \beta_{\text{low}}}{\log(T+1)} \quad (\text{for } T > 0)$$
$$\beta_{\text{base}}(t) = \beta_{\text{low}} + c \cdot \log(t+2)$$
$$\beta(t) = \text{clip}(\beta_{\text{low}} + (\beta_{\text{base}}(t) - \beta_{\text{low}}) \cdot c_{\text{factor}}, \quad \beta_{\text{low}}, \quad \beta_{\text{high}})$$
- `linear`**:** Linear increase in inverse temperature $\beta$.
$$\text{progress} = (t+1)/T$$
$$\beta_{\text{base}}(t) = \beta_{\text{low}} + (\beta_{\text{high}} - \beta_{\text{low}}) \cdot \text{progress}$$
$$\beta(t) = \text{clip}(\beta_{\text{low}} + (\beta_{\text{base}}(t) - \beta_{\text{low}}) \cdot c_{\text{factor}}, \quad \beta_{\text{low}}, \quad \beta_{\text{high}})$$
- `sqrt`**:** Increase in $\beta$ proportional to the square root of time progress.
$$\text{progress} = \sqrt{t+1}/\sqrt{T}$$
$$\beta_{\text{base}}(t) = \beta_{\text{low}} + (\beta_{\text{high}} - \beta_{\text{low}}) \cdot \text{progress}$$
$$\beta(t) = \text{clip}(\beta_{\text{low}} + (\beta_{\text{base}}(t) - \beta_{\text{low}}) \cdot c_{\text{factor}}, \quad \beta_{\text{low}}, \quad \beta_{\text{high}})$$

The use of bounded schedules allows a controlled comparison focused on the impact of the annealing shape over a finite, practical training for small-scale validation, rather than aiming to demonstrate asymptotic convergence which would require unbounded schedules.

### C.5 Linear Probe Evaluation

After pre-training for 200 epochs, the ResNet-18 backbone weights are frozen. Features (512-dimensional) are extracted for all images in the CIFAR-10 training and test sets using the frozen backbone. A linear classifier is trained on the extracted training features and corresponding labels. We use `sklearn.linear_model.LogisticRegression` with its default parameters ('solver='liblinear'', 'C=1.0', 'max_iter=1000'). The reported linear probe accuracy is the classification accuracy achieved by this trained classifier on the extracted test set features.

### C.6 Software and Hardware

Experiments were implemented using Python 3.11 and PyTorch 2.6.0 with CUDA 12.4. Training was performed on L4 GPUs accessed via Google Colaboratory with an average training time of $40$ seconds per epoch, respectively. Exact figures are reported in the generated data.

## D Preliminary Experiments on the Sensitivity to Annealing Rate

The theoretical annealing schedules ensuring convergence, such as $\beta(t) = c\ln(t+K)$ (Theorem 3.1), depend on two key parameters: an offset $K$ and a rate constant $c > 0$. The choice of the offset $K$ (e.g., $K = 2$ in $\ln(t + 2)$ for 0-indexed time $t$) primarily ensures a positive starting $\beta$ and has negligible impact on the asymptotic convergence properties.

However, the choice of the annealing rate constant $c$ is more critical theoretically and practically. Classical simulated annealing theory relates the minimum required value of $c$ to the maximum energy barrier ($\Delta E_{\max}$) that must be overcome to escape local minima [7]. Specifically, convergence is guaranteed if $c \leq 1/\Delta E_{\max}$. In the context of complex, high-dimensional loss landscapes encountered in representation learning, estimating these energy barriers is generally intractable, making the optimal theoretical choice of $c$ unknown *a priori*.

Nevertheless, we can empirically investigate the sensitivity of finite-time performance to the relative rate of annealing. We reuse the common scaling hyperparameter $c_{\text{factor}}$ which scales the progression from $\beta_{\text{low}} = 1.0$ towards $\beta_{\text{high}} = 100.0$ for the bounded `log` and `linear` schedules. We ran preliminary experiments on CIFAR-10 for a shorter duration of 100 epochs using a single seed (seed 1000) for different values of $c_{\text{factor}} \in \{0.5, 1.0, 2.0, 4.0\}$ and the Adam optimizer.

Table 2 shows the final linear probe accuracy, and Figure 2 shows the corresponding loss curves.

Table 2: CIFAR-10 probe accuracy (%) vs. scaling constant $c$ for `log` and `linear` after contrastive pretraining over 100 epochs.

| Schedule / $c$ | 0.5 | 1.0 | 2.0 | 4.0 |
|---|---|---|---|---|
| `Log` | 49.46 | 49.18 | 49.58 | 50.09 |
| `Linear` | 47.12 | 47.96 | 48.94 | 48.96 |

For both schedule shapes, using a small scaling factor ($c_{\text{factor}} = 0.5$), which corresponds to undercooling (reaching a final $\beta$ lower than $\beta_{\text{high}}$), generally yields slower loss reduction and lower accuracy compared to $c_{\text{factor}} = 1.0$. This aligns with the main finding that reaching a sufficiently high $\beta$ is important. Increasing the rate ($c_{\text{factor}} = 2.0, 4.0$) causes the schedules to reach $\beta_{\text{high}}$ much earlier. For the `log` schedule, faster rates ($c = 2.0, 4.0$) resulted in slightly higher accuracy in this run, while for `linear`, performance seemed to plateau or slightly decrease at the fastest rate ($c = 4.0$).

Overall, these results indicate relative robustness to the specific annealing rate within a reasonable range ($c_{\text{factor}} \in [1.0, 4.0]$) for these bounded schedules over 100 epochs, although extreme undercooling ($c = 0.5$) is detrimental. A default value like $c_{\text{factor}} = 1.0$ provides a sensible middle ground, ensuring the target $\beta_{\text{high}}$ is reached while allowing sufficient time for exploration. While excessively high rates (very large $c_{\text{factor}}$) could potentially lead to issues analogous to a fixed high $\beta$ start (e.g.,

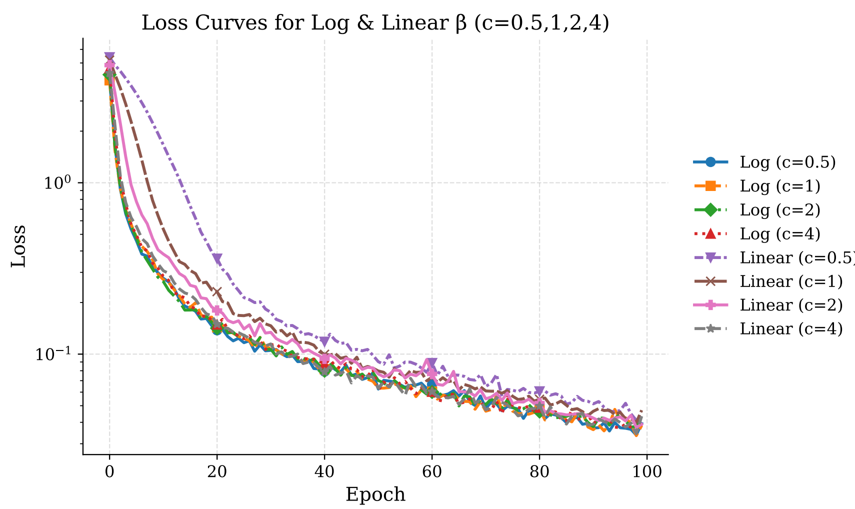


Figure 2: InfoNCE loss per epoch during pre-training on CIFAR-10 (single seed, 100 epochs) for `log` and `linear` schedules with varying rate scaling factors $c_{\text{factor}}$.

getting trapped or representation collapse), our tests did not show catastrophic failure for $c_{\text{factor}}$ up to 4.0.

Ultimately, the precise optimal schedule rate, $c^*$, is tied to the height of the unknown energy barriers of the specific loss landscape. Further investigation with unbounded schedules over much longer training horizons would be needed to fully probe the asymptotic behavior and confirm the robustness conjecture for different schedule shapes and rates. This connection between landscape geometry (barrier heights) and optimal annealing rates remains an important direction for future research. Indeed, it could potentially inform adaptive temperature schemes by estimating this height retroactively.

# E Interaction Between Momentum-Based Optimizers and Annealing

Momentum-based optimizers like Adam implicitly have mechanisms to traverse the loss landscape better, helping them escape from local minima. However, our main theoretical analysis and empirical validation focus on the overdamped Langevin SDE and its discrete-time analogue, vanilla SGD. To provide insight into the significant role of the optimizer, we present a direct comparison of performance on the `fixed_low` temperature schedule.

The setup was identical to the SGD experiments for the `fixed_low` schedule ($\beta = 1.0$) as described in Section 5 and Appendix C, with the sole exception being the use of the Adam optimizer (learning rate $3 \times 10^{-4}$, $\beta_1 = 0.9, \beta_2 = 0.999$, weight decay $10^{-6}$). The results are averaged over a different set of 3 random seeds (42, 42, 44).

## E.1 Results and Discussion.

The performance difference between the optimizers under the same challenging `fixed_low` schedule is substantial, as shown in Table 3.

Table 3: Comparison of final linear probe accuracy (%) on the `fixed_low` ($\beta = 1.0$) schedule using SGD versus Adam. Results are averaged over 3 seeds.

| Optimizer | Mean Acc (%) | Min Acc (%) | Max Acc (%) |
|---|---|---|---|
| SGD | 37.02 | 36.46 | 37.27 |
| Adam | **43.83** | 43.40 | 44.12 |

Even without any temperature annealing, simply switching from SGD to Adam yields a significant $\approx$6.8-point accuracy improvement. This illustrates that momentum and adaptive learning rates provide their own powerful mechanisms for navigating and escaping the local minima of the InfoNCE landscape, especially in a low-$\beta$ regime where the landscape is less sharp.

This finding helps contextualize our main theoretical results. While our work proves that annealing is a necessary principle for guaranteed global convergence in the foundational SGD-like setting, it also highlights that the choice of optimizer is a critical factor. A full theoretical treatment of annealing for momentum-based optimizers, likely requiring an analysis of underdamped Langevin SDEs, is a challenging but important direction for future research. Such an analysis could clarify whether momentum allows for faster annealing schedules or alters the fundamental escape dynamics.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions: the theoretical connection between InfoNCE annealing and simulated annealing via SDEs, the proof of global convergence for logarithmic schedules, the geometric analysis on the hypersphere, and the empirical validation on CIFAR-10.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Theoretical assumptions are outlined in the problem setup and formally in the appendix. The conclusion discusses both theoretical and practical limitations, including the SDE approximation of discrete optimizers, the asymptotic nature of convergence guarantees, the limited empirical scope (CIFAR-10/ResNet-18), and the assumption of exact hypersphere embeddings. The strong strutrucal assumptions is made clear.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Key assumptions are outlined in the problem setup and formally in the appendix. Theoretical results are clearly stated and numbered. Intution is provided in the main text and dealt with formally in the appendix. References to classical simulated annealing literature are provided.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper details the dataset (CIFAR-10), model architecture (ResNet-18 backbone, MLP projection head, L2 norm), optimizer (SGD) and its hyperparameters (LR, betas, epsilon, weight decay), batch size, epochs, gradient clipping strategy, and random seeds used. An appendix on experimental setup is provided the exact data augmentation pipeline and precise mathematical definitions for all annealing schedules tested. This level of detail should allow others to reproduce the experiments. The code is also available on GitHub with a link the repository in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and data provided via a GitHub. The appendix on experimental setup provides detailed experimental setup instructions which should be sufficient to run the provided code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The empirical validation section provides an overview, and the appendix on experimental setup provides comprehensive details on the dataset, data augmentations, model architecture, optimizer settings (SGD with specified LR), batch size, number of epochs, gradient clipping, loss function implementation details (implied by reference to code), and random seeds. Linear probe evaluation details are also given.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 reports mean, minimum, and maximum accuracy over 3 random seeds for the main CIFAR-10 experiments, explicitly stating the source of variability (random seeds). Figure 1 plots mean loss curves over these seeds. While standard error/deviation bars are not used, the min/max values provide information about the run-to-run variation observed across the specified seeds.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix on experimental setup details the software stack (Python, PyTorch, CUDA versions) and the hardware used (NVIDIA T4/L4 GPUs via Google Colaboratory). It also provides approximate training times per epoch for these GPUs, giving context for the required computational cost. Training times per epoch are also recorded in the output. A

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We do not believe we deviate from the Code of Ethics. The work is theoretical and involves standard datasets/practices without foreseeable ethical concerns regarding bias, fairness, or misuse beyond those general to representation learning.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This paper presents foundational theoretical research on optimization techniques for contrastive learning. It does not propose a specific application or model for deployment. As per NeurIPS guidelines, discussion of broader impacts is not required for purely foundational work without a direct path to societal application or potential misuse beyond the general capabilities enabled by improving representation learning algorithms.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This work does not introduce or release high-risk models or datasets. The experiments use the standard CIFAR-10 dataset, and the models trained are for validating theoretical concepts on a small scale, posing no foreseeable misuse risk requiring specific safeguards beyond standard academic code release.

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The only existing asset used is the CIFAR-10 dataset. The paper by Krizhevsky and Hinton (2009) is cited, as asked on the University of Toronto website. No formal license is published on the official site.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The primary new asset is the implementation code and the model performance metrics, which are provided via GitHub. Documentation in form of a README will accompany the public release detailing setup, usage, and how to retrieve the data used to produce the figures and tables in the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as a significant component.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.