A Theoretical Framework for Self-Play Theorem Proving Algorithms

Thomas Chen, Zhiyuan Li

Abstract

Self-play, a type of training algorithm that enables a model to self-improve, has recently shown promising empirical results in the context of formal theorem proving using Large Language Models (LLMs). [DM25] instantiate self-play with two cooperating agents: a prover, which proves theorems, and a conjecturer, which generates new theorems as a curriculum to the prover. In this paper, we provide a theoretical framework for understanding the self-improvement capabilities of self-play algorithms for theorem proving. First, we formalize the set of theorems as a graph, with nodes as theorems and edges between pairs of theorems with similar semantics. We introduce a set of primitive assumptions that characterize the guarantees of a trained prover and how a conjecturer can access the structure of the graph. Second, we show that if the underlying graph of theorems is well-connected, then a prover-conjecturer system, where the conjecturing algorithm is based on a reversible random walk, is sufficient to grow the set of proved theorems exponentially. Third, motivated by an issue encountered empirically by self-play algorithms, where the conjecturer tends to generate artificially complex and non-fundamental theorems, we propose a diversity measure for a training distribution of theorems generated by a conjecturer and an improved conjecturing algorithm that locally maximizes this diversity measure, by computing the diffusion similarity between neighboring theorems in the theorem graph. Finally, we describe a method to compute the diffusion similarity by using contrastive learning to embed nodes into Euclidean space and then computing the inner-product between embeddings.

1 Introduction

Formal theorem proving and mathematical reasoning with Large Language Models (LLMs) have made rapid progress, including systems that reached IMO gold-medal-level performance and systems reported to assist with open mathematical research problems [LL25, BCE+25]. Self-play algorithms are a promising new method to train LLM-based theorem provers, involving two cooperating models: a prover model and a conjecturer model. Self-play can be viewed as an automated synthetic data generation mechanism, where a system of two models generate and solve gradually-more-difficult instances of a task, improving both models with no manually annotated data required. Self-play has been used to train models to play Go [SHS+17]. In the context of theorem proving, self-play algorithms have been studied empirically. STP [DM25] trains two models: a prover, which maps theorem statements to their proofs, and a conjecturer, which maps theorem statements to slightly more difficult theorem statements adaptively based on the prover. The prover and conjecturer models cooperate to gradually enable the prover to prove harder theorems.

To our knowledge, this is the first theoretical analysis of self-play training algorithms for formal theorem proving. In this work, we introduce a theoretical framework, based on a graph of theorems, where the primary objective is to design a training algorithm that yields a prover with a large knowledge set—the set of theorems it can prove (Section 4). The framework introduces a set of primitive assumptions, characterizing how a prover can generalize in-domain when trained on a distribution of theorems and how a conjecturer can access the structure of the theorem graph. The

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

space of training algorithms includes prover-conjecturer systems analogous to that in STP, where the conjecturer generates a sequence of distributions over theorems, which the prover is trained on to gradually increase its knowledge. Under a connectivity condition on the underlying theorem graph, we show that the prover-conjecturer system in Algorithm 1, using a simple baseline conjecturing algorithm based on a reversible random walk, guarantees that the size of the prover's knowledge set at iteration $t \in \mathbb{N}$ grows exponentially in t (Theorem 4.2).

However, the size of the prover's knowledge set alone is insufficient as an objective since the knowledge set can contain many related, non-fundamental theorems (e.g. increasingly complex arithmetic problems). Empirically, STP [DM25] use heuristic elegancy filters to filter out inelegant, artificially-complicated generated-conjectures. Minimo [PBHG24] encounters an issue where the conjecturer generates longer and more complicated conjectures rather than deeper conjectures. We focus on how to encourage the conjecturer to generate theorems whose fundamental proof ideas are more *diverse*, which we believe is related to the issues encountered empirically.

Towards quantifying *diversity* of a prover's knowledge set and proposing algorithmic improvements to the conjecturer, the secondary objective of a training algorithm within our framework is: given a hyperparameter $\tau \in \mathbb{N}$, maximize the size of the τ-step expansion of the prover's knowledge set in the theorem graph, a quantity we identify with the diversity of the prover's knowledge. This prover-centric diversity is of practical use itself because it is heuristically the number of theorems with a τ-step reduction to a theorem the prover knows how to prove, which the prover can prove at inference time given τ steps of "reasoning." This prover-centric diversity can also be thought of as quantifying the size of the boundary of the prover's knowledge set. Towards maximizing this prover-centric diversity, we propose a diversity measure for the training distribution of theorems generated by a conjecturer (which the prover is trained on) and show it lower bounds the prover-centric diversity (Lemma 5.3). We propose an improved conjecturing algorithm (Algorithm 3) that generates training distributions that locally maximize this diversity measure, which requires computing a quantity called the diffusion similarity. Finally, we propose a new way to compute the diffusion similarity via contrastive learning (Theorem 5.5).

1.1 Technical Overview

Theorem Graph. Let $V \subset \{0,1\}^*$ be the set of all theorems, encoded as strings according to a formal language like Lean [dMU21]. Let $\mathcal{F}_{\text{prover}}$ be the hypothesis class of prover models (e.g. the class of pre-trained LLMs), where every prover $\mathcal{F}_{\text{prover}} \ni f : V \to \Delta(\{0,1\}^*)$ maps a theorem to a distribution over proofs. A verifier $\text{Ver} : V \times \{0,1\}^* \to \{0,1\}$ verifies if a proof of a theorem is correct. Using Ver, for any prover $f \in \mathcal{F}_{\text{prover}}$ and theorem $x \in V$ define the success rate $\text{succ}(x,f) \in [0,1]$ as the probability that f on input x generates a correct proof of x. For any $\varepsilon \in (0,1)$, define $E(\varepsilon, \mathcal{F}_{\text{prover}}) \subset V \times V$ as the set of pairs of theorems that are semantically similar, with respect to $\mathcal{F}_{\text{prover}}$:

$$E(\varepsilon, \mathcal{F}_{\text{prover}}) := \{(x,y) \in V \times V : \forall f \in \mathcal{F}_{\text{prover}}, |\text{succ}(x,f) - \text{succ}(y,f)| \leq \varepsilon\} \tag{1.1}$$

Assumption 3 and 4 of our framework require that the edge set $E \subset E(\varepsilon, \mathcal{F}_{\text{prover}})$ and for any theorem $x \in V$, the set $N(x) := \{y \in V : (x,y) \in E\}$ is efficiently computable. The neighbor oracle $N : V \to 2^V$, mapping $x \to N(x)$ is a key algorithmic primitive for the conjecturing algorithms we propose later. The theorem graph is the undirected graph $G = (V, E)$ with nodes V and edges E.

Further Assumptions. We define a conjecturer as a mapping from a distribution $D \in \Delta(V)$ of theorems (represented as a vector in $[0,1]^{|V|}$) to a new distribution $D' \in \Delta(V)$ of theorems. One type of conjecturer is a single step of a random walk with transition matrix $P \in [0,1]^{|V| \times |V|}$, where $(D')^\top = D^\top P \in [0,1]^{1 \times |V|}$. Our first result assumes that $|V| = \infty$ and that the random-walk-based conjecturer P satisfies a connectivity condition, Assumption 5. When $P = P_{\text{s.r.w.}}$ is a simple random walk, where a node transitions uniformly to its neighbors, then Assumption 5 amounts to a property of G: that there exists $\kappa > 0$ where for any finite subset of nodes $A \subset V$ then $\kappa \cdot |\partial A| \geq \deg(A)$ where $|\partial A|$ are the number of edges with one endpoint in A and the other in $V - A$, and $\deg(A)$ is the sum of the degrees of nodes in A. Finally, our framework includes Assumption 1 and Assumption 2, which are statistical-learning-like assumptions which describe how a prover trained on a finite sample of i.i.d. (theorem, proof) pairs from some training distribution D can generalize in-domain to D.

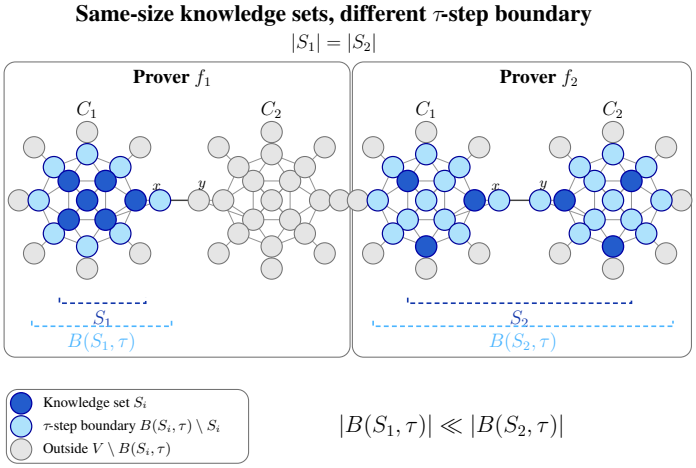

Figure 1: Let $\tau = 1$. (Left) Prover f_1's τ-expanded knowledge set is only supported on the left cluster. (Right) Prover f_2's τ-expanded knowledge set is well-supported on both clusters. Their knowledge sets S_1, S_2 are equal in size (6), but S_2 is more diverse since its τ-step expansion is larger.

A Baseline Conjecturing Algorithm. First we show if the theorem graph G satisfies a connectivity assumption, Assumption 5, then a conjecturer, based on a nearest-neighbor-type reversible random walk, enables the prover to prove an exponential in t number of theorems at every iteration $t \in \mathbb{N}$.

Theorem 1.1. *(Theorem 4.2, Informal) Let $G = (V, E)$ be the theorem graph, and (V, P) a nearest-neighbor-type random walk, constituting a conjecturing algorithm, that satisfies the connectivity condition of Assumption 5 with Isoperimetric constant $\kappa > 0$. Suppose the framework Assumptions 1-4 hold, with universal constant $p \in (0, 1)$ defined in Assumption 2. For any training-error threshold $\epsilon < \frac{1}{2\kappa^2}$, the prover-conjecturer system (Algorithm 1) satisfies that for all $t \in \mathbb{N}$, the number of theorems that the prover can prove with success rate at least p at iteration t is exponential in t.*

An Improved Conjecturing Algorithm for the Diversity Issue. Suppose $f \in \mathcal{F}_{\text{prover}}$ is the final prover trained by a self-play algorithm like Algorithm 1. Whereas Theorem 4.2, under Assumption 5, provides guarantees of $|K^p(f)| := |\{x \in V : \text{succ}(x, f) \geq p\}|$, the size of the prover's knowledge, Theorem 4.2 does not guarantee the diversity of the prover's knowledge. For hyperparameter $\tau \in \mathbb{N}$, we study how to optimize a proxy for diversity of the prover's knowledge, $|K^p_\tau(f)| := |\{x' \in V : \exists x \in V : \text{succ}(x, f) \geq p \text{ and } \text{dist}(x, x') \leq \tau\}|$. Figure 1 depicts two provers, f_1 and f_2, that both can prove the same number of theorems ($|K^p(f_1)| = |K^p(f_2)|$), but where f_1's knowledge set is much less diverse, in the sense that $|K^p_\tau(f_1)| \ll |K^p_\tau(f_2)|$.

Suppose prover f is trained on theorem distribution $D \in \Delta(V)$ in the sense that $\mathbb{P}_{x \sim D}[\text{succ}(x, f) \geq p] \geq 1 - \epsilon$ (as in Assumption 2). Lemma 5.3 shows that in order to maximize $|K^p_\tau(f)|$, it is reasonable to maximize the following property of f's training distribution: $\text{Div}^P_\tau(D) := \frac{1}{||D^\top P^\tau||^2_{\ell^2(1/\pi)}} = \frac{1}{\sum_{x \in V} (D^\top P^\tau e_x)^2 \frac{1}{\pi(x)}}$, where P is the transition matrix of a reversible, nearest-neighbor Markov chain. With this observation, the goal is to design a conjecturing algorithm that generates training distributions $\{D_t\}_{t \in \mathbb{N}}$ where $\text{Div}^P_\tau(D_t)$ is large for each $t \in \mathbb{N}$. We propose Algorithm 3, an improved conjecturer that defines a push-forward map $C : V \to \Delta(V)$ where $\forall x \in V, C(x) \in \Delta(N(x))$. The weighting of $C(x)$ is determined by maximizing $\text{Div}^P_\tau(C(x))$.

$$\forall x \in V, C(x) \leftarrow \arg \max_{\mu \in \Delta(N(x))} \text{Div}^P_\tau(\mu) = \arg \min_{\mu \in \Delta(N(x))} \mu^\top P^\tau \text{diag}(1/\pi)(P^\tau)^\top \mu \tag{1.2}$$

Algorithm 3 needs to compute the diffusion similarity $\phi^{\text{diff}}_\tau(y, y') := e_y^\top P^\tau \text{diag}(1/\pi)(P^\tau)^\top e_{y'}$ for every pair of neighbors $y, y' \in N(x)$ of x. We show the diffusion similarity can be computed by embedding each node in G into Euclidean space via an embedding model learned with contrastive learning, and then computing the inner product between embeddings. This can be used as a subroutine in a conjecturing algorithm like Algorithm 3.

Theorem 1.2. *(Theorem 5.5, Informal) Let $G = (V, E)$ be a finite undirected, connected theorem graph. Let $\mathcal{F}_V^{(|V|)}$ be the class of all mappings from V to $\mathbb{R}^{|V|}$, and $\mathcal{F}_{\text{emb}} \subset \mathcal{F}_V^{(|V|)}$ a class of embedding models. For any $\tau \in \mathbb{N}$, $\mathcal{L}_{\text{cont}}^{\tau} : \mathcal{F}_V^{(|V|)} \to \mathbb{R}$ is a contrastive loss with time-scale parameter τ (Equation (5.3)). Assume $\mathcal{F}_{\text{emb}}$ contains a global minimizer of $\mathcal{L}_{\text{cont}}^{\tau}(\cdot)$. For any such global minimizer $h_{\theta_*} \in \mathcal{F}_{\text{emb}}$, then $\forall u, v \in V$, $\phi_{\tau}^{\text{diff}}(u, v) = h_{\theta_*}(u)^{\top} h_{\theta_*}(v)$.*

Under additional assumptions, we prove Theorem C.17, which is an extension to the result of Theorem 5.5 that applies for classes of embedding models $\mathcal{F}_{\text{emb}}$ where models $\mathcal{F}_{\text{emb}} \ni h_\theta : V \to \mathbb{R}^k$ have small embedding dimension $k \ll |V|$. Theorem C.17 provides approximation-guarantees for an embedding model that is learned via empirical risk minimization (ERM).

2 Related Work

Formal theorem proving with LLMs in Lean [dMU21] is divided into two paradigms: step-level interaction and whole-proof generation. Step-level methods interact with Lean incrementally, proposing one tactic at a time [A+25, H+25, WHZ+24, X+25b, XZN+25]. Whole-proof methods generate a complete Lean proof in a single interaction [C+25, JLW+25, LTL+25, L+25, X+25a, R+25, SWP+25, W+25, Z+25]. Regarding self-play theorem proving systems, STP instantiates a prover-conjecturer self-play loop from pretrained LLMs, where the conjecturer is trained to generate increasingly-difficult synthetic theorems [DM25]. MINIMO instantiates a prover-conjecturer self-play loop to learn a prover from scratch, where the conjecturer generates conjectures via constrained decoding [PBHG24]. Further related works are in Section A.

3 Preliminaries

Preliminary Notation. $0 \in \mathbb{N}$. $[n] := \{1, 2, \ldots, n\}$. For a set V, 2^V is its power set. For $v \in \mathbb{R}^n$, $\text{diag}(v) \in \mathbb{R}^{n \times n}$ is the diagonal matrix with v on its diagonal. For $M \in \mathbb{R}^{n \times n}$, $i, j \in [n]$, denote $M[i, j]$ the (i, j)th entry, $M[i]$ the ith row, and $M[:, j]$ the jth column of M. $O^{k \times k}$ is the set of $(k \times k)$ orthogonal matrices. $\mathbf{1}$ is the all-1's vector. $\forall x \in V$, $e_x \in \mathbb{R}^{|V|}$ is the unit vector along dimension indexed by $x \in V$. For an undirected, unweighted graph $G = (V, E)$, for any $x, y \in V$, $\text{dist}(x, y)$ is the minimum number of edges of any path connecting x and y in G. $\deg(x)$ denotes the degree of node $x \in V$. For $x \in V, r \in \mathbb{N}$, denote $B(x, r) := \{x' \in V : \text{dist}(x, x') \leq r\}$ and $\partial B(x, r) := B(x, r) - B(x, r-1)$, $B(x, -1) := \emptyset$. For set $S \subset V$, let $B(S, r) := \bigcup_{v \in S} B(v, r)$. For $\pi : V \to (0, \infty)$ and $f, f' : V \to \mathbb{R}$, let $\langle f, f' \rangle_{\ell^2(1/\pi)} := \sum_{x \in V} f(x) f'(x) \frac{1}{\pi(x)}$ and $||f||^2_{\ell^2(1/\pi)} := \langle f, f \rangle_{\ell^2(1/\pi)}$. Let $\Delta(V)$ denote the set of probability distributions whose support is contained in the set V. For $D \in \Delta(V)$, $\text{supp}(D)$ is its support. For $g : V \to \Delta(V)$, $g \# D \in \Delta(V)$ is the distribution of y, drawn according to this process: $x \sim D, y \sim g(x)$.

Preliminary: Markov Chains on Graphs and Isoperimetric Inequality. Section B.1 has more-detailed preliminaries. Let $G = (V, E)$ be a connected graph with a countable set of vertices. Let (V, P) denote an irreducible, reversible Markov chain with transition matrix $P \in [0, 1]^{|V| \times |V|}$ and invariant measure $\pi : V \to (0, \infty)$ where $\forall x, y \in V, \pi(x) P[x, y] = \pi(y) P[y, x]$. Let $E(P)$ be the set of (x, y) where $P[x, y] > 0$. If $E(P) \subset E$, then (V, P) is nearest-neighbor-type random walk. The simple random walk $P_{\text{s.r.w.}}$ is a Markov chain where every node x transitions uniformly to its neighbors $N(x)$. For any $A \subset V$, let ∂A be the set of all edges in $E(P)$ with one endpoint in A and the other in $V - A$. Let $\pi(A) := \sum_{x \in A} \pi(x)$ and $a(\partial A) := \sum_{(e^-, e^+) \in \partial A} \pi(e^-) P[e^-, e^+]$.

Definition 3.1 (Isoperimetric Inequality)**.** *A Markov Chain (V, P) satisfies Strong Isoperimetric Inequality,* IS*, if there exists a constant $\kappa > 0$ such that for every finite $A \subset V$, $\pi(A) \leq \kappa \cdot a(\partial A)$.*

4 Framework and Random-Exploration Baseline

4.1 Primitive Assumptions

With $t \in \mathbb{N}$ denoting the iteration index, the primary goal of the training algorithm is to produce a prover f (or a sequence of provers $\{f_t\}_{t \in \mathbb{N}}$) which can prove as many theorems as possible. The following assumptions define a set of core primitives for designing such a training algorithm.

Assumption 1 assumes i.i.d. sample access to an initial training distribution D_0 of theorems where for every theorem x in the support of D_0, the proof of x is known. For a proof to be known, we mean that we have access to a prover f where $\mathrm{succ}(x, f) \geq \Omega(1)$ and can efficiently sample the proof in $O(1)$ expected calls to f. An example of D_0 would be a distribution of elementary-school-level math problems, where a weak model can solve all the problems.

Assumption 1. *(Cold-Start Dataset) We are given access to an oracle that provides i.i.d. samples from $D_0 \in \Delta(V)$ where the proof of every theorem $x \in \mathrm{supp}(D_0)$ is known.*

Assumption 2 describes the ability of a model, trained on a distribution D of theorems via i.i.d. samples, to generalize in-domain to D. We justify it with statistical learning theory in Section E.

Assumption 2. *(Training) Let $p \in (0,1), p = \Theta(1)$ be a universal constant. Let $D \in \Delta(V)$ be a distribution of theorems where $\forall x \in \mathrm{supp}(D)$, the proof of x is known. Then, $\forall\, \epsilon, \delta \in (0,1)$, with $n = \Theta(\frac{\mathcal{C}(\mathcal{F}_{\text{prover}}) \log \frac{1}{\delta}}{\epsilon^2 (1-p)^2})$ where $\mathcal{C}(\mathcal{F}_{\text{prover}})$ is an $\mathcal{F}_{\text{prover}}$-dependent constant, we can train a prover f using n i.i.d. samples from D such that with probability at least $1-\delta$, we have that f satisfies: $\mathbb{P}_{x \sim D}[\mathrm{succ}(x, f) \geq p] \geq 1 - \epsilon$.*

The requirement of Assumption 2 that the proof of every theorem in $\mathrm{supp}(D)$ is known can also be thought of as requiring a training distribution of (theorem, proof) pairs, where every theorem in the support of the training distribution must have a correct proof associated with it. This requirement is necessary, as otherwise it would be possible to train a model to have high success rate on arbitrarily hard theorems for which we do not know the proof.

Let $G = (V, E)$ be the theorem graph. Assumption 3 says that the edge set E is where the success rate of any prover in $\mathcal{F}_{\text{prover}}$ is not too different between two theorems connected by an edge.

Assumption 3. *(Continuity of Success Rate) With $p \in (0,1)$ defined as in Assumption 2, assume for all $f \in \mathcal{F}_{\text{prover}}$, for all $(x, y) \in E$, that $|\mathrm{succ}(x, f) - \mathrm{succ}(y, f)| \leq \frac{p}{2}$*

In Assumption 4, we assume access to a neighbor oracle N as a primitive. We discuss realistic instantiations of this assumption after defining a conjecturer within our framework (Definition 4.1).

Assumption 4. *(Neighbor Access) For any $x \in V$, let $N(x) := \{y \in V : (x, y) \in E\}$. We assume oracle access to a neighbor oracle $N : V \to 2^V$, which on input of any $x \in V$, returns $N(x)$.*

Assumptions 2 – 4 enable a training algorithm to run the following at each iteration. First, expand the current set of known theorems by one step via the neighbor oracle N (Assumption 4). Second, Assumption 3 ensures that unseen theorems have success rate at least $p/2$ under the current prover, enabling the proofs of the unseen theorems to be efficiently sampled. In general, note that if we have access to a prover f whose success rate on theorem x is reasonably high ($\geq p/2$, where $p \geq \Omega(1)$ by Assumption 2), then the proof of x can be efficiently sampled and become known. Finally, a new prover is trained on the expanded set of theorems via Assumption 2. Assumption 3 is necessary to our analysis because standard statistical learning theory arguments rely on exchangeability between training and test distributions and therefore do not by themselves justify this out-of-distribution step.

One example instantiation of the theorem graph is where V is the set of discrete logarithm instances in a fixed cyclic group. Fix a cyclic group $\mathbb{G}$ of prime order q with generator $g \in \mathbb{G}$. For each $a \in \mathbb{Z}_q$, let $h_a = g^a$, and let $x_a \in V$ denote the theorem:

"Given cyclic group $\mathbb{G}$ of prime order q, and $g, g^{-1}, h_a \in \mathbb{G},$ there exists $z \in \mathbb{Z}_q$ such that $g^z = h_a$."

The proof of x_a is the unique exponent $f_*(x_a) = a$. Define the edge set E by $(x_a, x_b) \in E \iff b - a \in \{-1, +1\} \pmod q$. This edge set satisfies Assumption 4, since given the theorem $x_a = (\mathbb{G}, g, g^{-1}, h_a)$, the neighbor oracle can compute $N(x_a) = \{x_{a+1}, x_{a-1}\}$, where x_{a+1} has target $h_a g = g^{a+1}$, and x_{a-1} has target $h_a g^{-1} = g^{a-1}$. Assumption 3 amounts to the property that the prover class exhibits similar success rates for adjacent discrete-log instances: $(x_a, x_{a+1}) \in E \implies \forall f \in \mathcal{F}_{\text{prover}}, |\mathrm{succ}(x_a, f) - \mathrm{succ}(x_{a+1}, f)| \leq \frac{p}{2}$. Given g and its inverse g^{-1}, an instance with target h_{a+1} can be reduced to the adjacent instance with target h_a by multiplying by g^{-1}, i.e. $h_{a+1} g^{-1} = g^{a+1} g^{-1} = g^a = h_a$. Thus, for prover classes that can exploit such reductions, it is conceivable that neighboring instances x_a and $x_{a \pm 1}$ have similar success rates.

Algorithm 1 Prover-Conjecturer System

Require: Prover class $\mathcal{F}_{\text{prover}}$, reversible conjecturer $C_w : V \to \Delta(V)$ as in Definition 4.1, parameters $\epsilon, \delta \in (0,1)$, threshold $p > 0$

1: Acquire theorem distribution $D_0 \in \Delta(V)$ using Assumption 1
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: Given i.i.d. sample access to D_t, train prover $f_t \in \mathcal{F}_{\text{prover}}$ on D_t to accuracy ϵ and failure probability δ using Assumption 2
4: Define i.i.d. sample access to D_{t+1} by applying Algorithm 2 with input $(D_t, f_t, C_w, p/2)$

4.2 Random-Walk Conjecturing Algorithm

Conjecturer. In STP [DM25], the conjecturer is a mapping that takes as input a theorem x, its proof $f_*(x)$, a lemma l used in $f_*(x)$, and a system prompt; it returns a theorem c related to x. We study a simplification where the conjecturer maps a theorem to a distribution of theorems.

Definition 4.1. *(Conjecturer) For graph $G = (V, E)$, a conjecturer is a mapping $C : V \to \Delta(V)$, mapping a theorem $x \in V$ to a distribution over its neighbors in G, $N(x)$. Given i.i.d. sample access to input distribution $D \in \Delta(V)$, the conjecturer returns i.i.d. sample access to $D' = C\#D \in \Delta(V)$. A reversible conjecturer is specified by a set of symmetric, non-negative edge weights $w : E \to \mathbb{R}_+$, where $\forall x, y \in V, w(x,y) = w(y,x) \geq 0$ and $\sum_z w(x,z) \in (0, \infty)$. Denote $P_w \in [0,1]^{|V| \times |V|}$ as the matrix representation of w, defined as: $\forall x, y \in V, P_w[x,y] = \frac{w(x,y)}{\sum_{z \in N(x)} w(x,z)}$. Let $C_w : V \to \Delta(V)$ be an alternate representation: $\forall x \in V, C_w(x) = P_w[x]$.*[1]

For instance, when $\forall x \in V, z \in N(x), w(x,z) = 1$ and for $z \notin N(x), w(x,z) = 0$, then $C_w(x) = \text{Unif}(N(x))$, so that (C_w, P_w, w) defines a simple random walk over G, with $\pi_w(x) \propto \deg(x)$.

Returning to the discussion of the access model of the theorem graph (Assumption 4), in this framework we view a conjecturer as defining a reversible, nearest-neighbor-type random walk on a theorem graph. Querying the conjecturer offers some limited access to the underlying theorem graph. Since C_w maps a theorem x to a distribution over its neighbors $N(x)$, one could approximate $N(x)$ by approximating $\text{supp}(C_w(x))$. Thus, a potential graph-access assumption weaker than Assumption 4 would be access to some initial conjecturer C_w, a realistic object. The goal of our framework becomes how to improve an initial conjecturer (however it may have been trained), provided it satisfies our reversibility conditions in Definition 4.1. The important conceptual contributions of our framework are the same whether we use this weaker assumption or Assumption 4.

Prover-Conjecturer System. This section provides sufficient conditions for a prover-conjecturer system (Algorithm 1) to learn to prove more theorems across iterations. Theorem 4.2 assumes the conjecturer (C_w, P_w, w), defining Markov chain (V, P_w), satisfies a connectivity assumption.

Assumption 5. *Let $G = (V, E)$ be a connected graph where the number of vertices is countably infinite ($|V| = \infty$) and each vertex has finite degree. Assume that the Markov chain (V, P) for G satisfies the strong isoperimetric inequality,* IS*, defined in Definition 3.1.*

At iteration $t \in \mathbb{N}$, the prover-conjecturer system (Algorithm 1), initialized with conjecturer (C_w, P_w, w), has i.i.d. sample access to $D_t \in \Delta(V)$, a training distribution over theorems whose proofs are known. It trains prover $f_t \in \mathcal{F}_{\text{prover}}$ on D_t via Assumption 2. Finally, it constructs i.i.d. sample access to a new distribution over theorems, $D_{t+1} \in \Delta(V)$ using Algorithm 2. Given D_t, Algorithm 2 applies the conjecturer C_w to D_t, yielding $C_w\#D_t$. Theorems with low pass rate under f_t are filtered out, yielding D_{t+1}.

Due to the push-forward construction of D_{t+1} from D_t in Algorithm 1 for every $t \in \mathbb{N}$, sampling from D_t requires first sampling $x_0 \sim D_0$, then applying t push-forward transformations to yield $x_j \sim D_j$, for all $j \in [t]$. In particular, first draw a large number N_0 of i.i.d. samples $x_0^{(i)} \sim D_0$, $i \in [N_0]$. At every iteration $j \in \{0\} \cup [t]$, suppose inductively that there are N_j remaining unused i.i.d. samples

[1] The transition matrix P_w defines a nearest-neighbor-type reversible random walk over state-set V, with stationary measure $\pi_w : V \to (0, \infty)$, where $\forall x \in V, \pi_w(x) \propto \sum_{z \in N(x)} w(x,z) =: w(x)$. These three representations C_w, P_w, w are equivalent and uniquely determined given any of the others up to a positive common factor under irreducibility. We write (C_w, P_w, w) (or just C_w) to denote a conjecturer.

Algorithm 2 Pass-Rate-Oracle-Based Sampler

Require: i.i.d. sample access to $D \in \Delta(V)$, prover $f \in \mathcal{F}_{\text{prover}}$, conjecturer $C_w : V \to \Delta(V)$, threshold $q > 0$, verifier $\text{Ver} : V \times \{0,1\}^* \to \{0,1\}$, and oracle access to $\text{succ}(\cdot, f)$

Ensure: i.i.d. sample access to $D' \in \Delta(V)$

i.i.d. sampler for D':
  repeat
    Draw theorem $x \sim D$
    Sample $x' \sim C_w(x)$
  until $\text{succ}(x', f) \geq q$
  Sample and record a correct proof of x' from f, checking correctness using Ver.
  return x'

from D_j. Use $\Theta(\frac{\mathcal{C}(\mathcal{F}_{\text{prover}}) \log \frac{1}{\delta}}{\epsilon^2(1-p)^2})$ of the N_j samples to train prover f_j via Assumption 2. Of the unused samples, apply the conjecturer to map $x_j^{(i)}$ to $y_j^{(i)} \sim C_w(x_j^{(i)})$. Discard $y_j^{(i)}$ if $\text{succ}(y_j^{(i)}, f_j) < p/2$, else let $x_{j+1}^{(i)} := y_j^{(i)}$. Let N_{j+1} denote the number of remaining samples. Note that if prover f_j is successfully trained according to Assumption 2 to where $\mathbb{P}_{x \sim D_j}[\text{succ}(x, f_j) \geq p] \geq 1 - \epsilon$, then by Assumption 3, $\mathbb{P}_{x \sim D_j, x' \sim C_w(x)}[\text{succ}(x', f_j) \geq p/2] \geq 1 - \epsilon$, so that by standard concentration inequalities, $N_{j+1} \gtrapprox (1-\epsilon)N_j$. In the worst case, we have that $\forall j \in [t], N_j \approx N_0(1-\epsilon)^j$, so that the number of oracle calls to D_0 required to sample from D_j grows as $(\frac{1}{1-\epsilon})^j$.

4.3 Expansion Guarantee for Algorithm 1.

Theorem 4.2 shows that on an infinite graph, with a reversible conjecturer (Definition 4.1) that induces a random walk that satisfies the Strong Isoperimetric Inequality (Assumption 5), and for training error threshold $\epsilon < \frac{1}{2\kappa^2}$, Algorithm 1 generates a sequence of provers $\{f_t\}_{t \in \mathbb{N}}$ such that $\forall t \in \mathbb{N}$, the size of the t-th prover's knowledge set $|\{x \in V : \text{succ}(x, f_t) \geq p\}|$ is exponential in t.

Theorem 4.2 (Analysis of Algorithm 1)**.** *Let $G = (V, E)$ be a theorem graph with $|V| = \infty$ and (C_w, P_w, w) be a reversible conjecturer as in Definition 4.1. Suppose (V, P_w) satisfies Assumption 5 with Isoperimetric constant $\kappa > 0$. Let $\pi : V \to (0, \infty)$ be an invariant measure of (V, P_w). Suppose Assumptions 1-4 hold, with universal constant $p \in (0,1)$ as in Assumption 2, and that π is chosen so that $\sup_{z \in V} \pi(z) < \infty$ and $0 < ||D_0||^2_{\ell^2(1/\pi)} < \infty$. For any $\delta, \epsilon \in (0,1)$, Algorithm 1 produces a sequence of provers $\{f_t\}_{t \in \mathbb{N}}$ where $\forall t \in \mathbb{N}$, with probability at least $1 - (t+1)\,\delta$,*

$$|\{x \in V : \text{succ}(x, f_t) \geq p\}| \geq \frac{(1-\epsilon)^2}{(\sup_{z \in V} \pi(z)) \cdot ||D_0||^2_{\ell^2(1/\pi)}} \cdot (\frac{1-\epsilon}{1 - \frac{1}{2\kappa^2}})^{2t} \tag{4.1}$$

To prove Theorem 4.2, we show that for every $t \in \mathbb{N}$, $||D_{t+1}||^2_{\ell^2(1/\pi)} \leq (\frac{||P_w^\top||_{\ell^2(1/\pi) \to \ell^2(1/\pi)}}{1-\epsilon})^2 \cdot ||D_t||^2_{\ell^2(1/\pi)}$, where $||P_w^\top||_{\ell^2(1/\pi) \to \ell^2(1/\pi)}$ is the operator norm of $P_w^\top$ in the Hilbert space $\ell^2(1/\pi)$. By Corollary B.15, since (V, P_w) satisfies Assumption 5, then $||P_w^\top||_{\ell^2(1/\pi) \to \ell^2(1/\pi)} \leq 1 - \frac{1}{2\kappa^2}$. Finally, Theorem 4.2 follows from Lemma F.1, which lower bounds $|\{x \in V : \text{succ}(x, f_t) \geq p\}|$ using $\frac{1}{||D_t||^2_{\ell^2(1/\pi)}}$. The full proof is in Section B.

Thus, if the simple random walk $P_{\text{s.r.w.}}$ over the theorem graph satisfies IS (i.e. for any finite $A \subset V$, $\frac{|\{(e^-, e^+) \in E : e^- \in A,\, e^+ \notin A\}|}{\sum_{x \in A} \deg(x)} \geq \frac{1}{\kappa}$), then Algorithm 1 with a conjecturer, whose transition matrix is $P_{\text{s.r.w.}}$, generates a sequence of provers $\{f_t\}_{t \in \mathbb{N}}$ whose knowledge set grows exponentially.

5 Diversity Metric and an Improved Conjecturing Algorithm

5.1 Diversity Measure

Unlike Section 4, suppose for simplicity the theorem graph has a finite but large number of nodes and $||\pi||_1 = 1$. The following ideas can be instantiated with any $\tau \in \mathbb{N}$, but practically τ should

be a small non-zero natural number. The main object of interest in this section is the τ-expanded knowledge set of a prover f.

Definition 5.1. *With universal $p \in (0,1)$ in Assumption 2, $\forall \tau \in \mathbb{N}$, $f \in \mathcal{F}_{\text{prover}}$, let $K^p_\tau(f) := B(\{x \in V : \text{succ}(x,f) \geq p\}, \tau) = \{x' \in V : \exists x \in V : \text{succ}(x,f) \geq p \text{ and } \text{dist}(x,x') \leq \tau\}$. We call $K^p(f) := K^p_0(f)$ the knowledge set of f and $K^p_\tau(f)$ the τ-expanded knowledge set.*

We define an operational notion of diversity of f's knowledge as the size of $K^p_\tau(f)$. $K^p_\tau(f)$ can be interpreted as the set of theorems that f can prove by performing τ steps of "reasoning" at inference time to reduce the input theorem into one it knows how to prove. Figure 1 depicts two provers f_1, f_2 whose knowledge sets are the same size $|K^p(f_1)| = |K^p(f_2)|$, but where their τ-step expansions are of different sizes: $|K^p_\tau(f_1)| \ll |K^p_\tau(f_2)|$. Thus, f_2's knowledge is more diverse than f_1's. Identifying diversity with $|K^p_\tau(f)|$ reflects the intuition that knowing a diverse set of proof ideas is useful to a prover f as it increases the set of theorems that f can prove by combining known proof ideas.

Let (C_w, P_w, w) be a baseline conjecturer satisfying Definition 4.1 and for convenience denote $P := P_w$. Our goal is to design an improved conjecturer that generates a sequence of training distributions $\{D_t\}_{t \in \mathbb{N}}$ where at each $t \in \mathbb{N}$ prover f_t is trained on D_t, and where $|K^p_\tau(f_t)|$ increases quickly with t. In fact, Lemma 5.3 (proved in Section F) shows that for each $t \in \mathbb{N}$ it is reasonable to maximize $\text{Div}^P_\tau(D_t)$ (Definition 5.2), a diversity measure of the training distribution D_t.

Definition 5.2 (Diversity Measure)**.** *Let $G = (V, E)$ be a theorem graph with P being the transition matrix of any reversible, nearest-neighbor-type random walk with stationary measure π. For any parameter $\tau \in \mathbb{N}$, and any $\mu \in \Delta(V)$, define* $\text{Div}^P_\tau(\mu) := \frac{1}{||\mu^\top P^\tau||^2_{\ell^2(1/\pi)}} = \frac{1}{\sum_{x \in V} (\mu^\top P^\tau e_x)^2 \frac{1}{\pi(x)}}$.

Lemma 5.3 (Large Div^P_τ implies large K^p_τ)**.** *Let $p \in [0,1]$ be a universal constant in Assumption 2. Instantiate $\text{Div}^P_\tau : \Delta(V) \to \mathbb{R}$ with the transition matrix P of a reversible, nearest-neighbor-type random walk. For any $\tau \in \mathbb{N}$, prover $f \in \mathcal{F}_{\text{prover}}$, and theorem distribution $D \in \Delta(V)$, the τ-expanded knowledge set of f satisfies $|K^p_\tau(f)| \geq \frac{(P_{x \sim D}[\text{succ}(x,f) \geq p])^2}{\sup_{x \in V} \pi(x)} \text{Div}^P_\tau(D)$.*

With prover f trained on theorem distribution D via Assumption 2, then $P_{x \sim D}[\text{succ}(x,f) \geq p] \geq 1 - \epsilon$. By Lemma 5.3, $|K^p_\tau(f)| \geq \frac{(1-\epsilon)^2}{\sup_{x \in V} \pi(x)} \text{Div}^P_\tau(D)$, so it is reasonable to maximize $\text{Div}^P_\tau(D)$.

We note that in the setting of Section 4, where Assumption 5 holds for random walk (V, P), then maximizing $|B(K^p(f), \tau)|$ reduces to maximizing $|K^p(f)|$. When $P = P_{\text{s.r.w.}}$, then Assumption 5 implies for bounded-degree graphs that there is a reasonable constant c where for any finite $A \subset V, |B(A, \tau)| \geq \exp(c\tau)|A|$. That is, any finite subset of theorems is diverse. More generally, if Assumption 5 holds for P (not necessarily $P_{\text{s.r.w.}}$), then for $\{D_t\}_{t \in \mathbb{N}}$ generated as $D^\top_{t+1} = D^\top_t P$, we have $\text{Div}^P_\tau(D_t) = \frac{1}{||D_{t+\tau}||^2_{\ell^2(1/\pi)}}$, where $\pi^\top P = \pi^\top$. Because the proof of Theorem 4.2 implies fast exponential decay of $||D_t||^2_{\ell^2(1/\pi)}$ as a function of $t \in \mathbb{N}$ for every t, it follows that $\text{Div}^P_\tau(D_t)$ grows exponentially for each t. In short, when Assumption 5 holds, then the two objectives $K^p(f)$ and $K^p_\tau(f)$ are aligned. As such, the more interesting setting for studying diversity is when Assumption 5 does not hold for the baseline random walk P. When $P = P_{\text{s.r.w.}}$, some intuitive graphs where Assumption 5 does not hold are graphs consisting of multiple poorly-interconnected clusters, such as in Figure 1 or Figure 2. For a graph like in Figure 1, the choice of which nodes to include in $K^p(f)$ will greatly impact $K^p_\tau(f)$, making the problem of maximizing both $K^p(f)$ and $K^p_\tau(f)$ non-trivial.

5.2 Improved Conjecturing Algorithm

Algorithm 3 is a conjecturer that generates a sequence of training distributions $\{D_t\}_{t \in \mathbb{N}}$, with the aim to maximize $\text{Div}^P_\tau(D_t)$ at each $t \in \mathbb{N}$, for a chosen hyper-parameter $\tau \in \mathbb{N}$. This maximization cannot easily be done in closed form, as the conjecturer is restricted to generate D_{t+1} from D_t by one step of a nearest-neighbor-type random walk. Instead, fixing $\tau \in \mathbb{N}$, Algorithm 3 defines a push-forward map $C : V \to \Delta(V)$ where for any $x \in V$, $C(x) \in \Delta(N(x))$ is supported on neighbors $N(x)$ of x. The weighting of $C(x)$ is determined by maximizing $\text{Div}^P_\tau(C(x))$.

$$\forall x \in V, C(x) \leftarrow \arg \max_{\mu \in \Delta(N(x))} \text{Div}^P_\tau(\mu) = \arg \min_{\mu \in \Delta(N(x))} \mu^\top P^\tau \text{diag}(1/\pi)(P^\tau)^\top \mu \tag{5.1}$$

Algorithm 3 Improved Conjecturer

Require: $\tau \in \mathbb{N}$, $\eta \in [0, 0.5]$, reversible conjecturer (C_w, P_w, w) satisfying Definition 4.1, and i.i.d. sample access to $D \in \Delta(V)$

Ensure: i.i.d. sample access to $D' = C\#D \in \Delta(V)$

1: **procedure** $C(x)$
2: Compute $p^\tau(u)$ for all $u \in N(x)$ using Algorithm 4 with (C_w, P_w, w)
3: Let $M'(x,\tau) \in \mathbb{R}^{|N(x)|\times|N(x)|}$, with (y,y')-entry $\sum_z p^\tau(y,z)p^\tau(y',z)\frac{1}{w(z)}$, $\forall y, y' \in N(x)$
4: $\mu_x \leftarrow \arg \min_{\substack{\mu\in\Delta(N(x)) \\ \mu(y)\geq\eta/\deg(x)\ \forall y\in N(x)}} \mu^\top M'(x,\tau)\mu$
5: Sample $y' \sim \mu_x$ and **return** y'
6: **i.i.d. sampler for** $D' = C\#D$**:**
7: Draw $x \sim D$, $x' \sim C(x)$, and **return** x'

Where $\forall z \in V, \text{diag}(1/\pi)[z,z] = \frac{1}{\pi(z)}$. This optimization problem can be solved given the matrix, $M(x,\tau) \in \mathbb{R}^{|N(x)|\times|N(x)|}$ where for any $y, y' \in N(x)$, the (y,y')-th entry is: $M(x,\tau)[y,y'] := e_y^\top P^\tau \text{diag}(1/\pi)(P^\tau)^\top e_{y'}$. We can characterize $M(x,\tau)$ in terms of diffusion embeddings of the neighbors of x, where a diffusion embedding embeds any node $z \in V$ to a vector $p^\tau(z)$ representing the probability distribution of a τ-step random walk started at z according to transition matrix P.

Definition 5.4. *Let $G = (V,E)$ be a connected, undirected graph, with P being the transition matrix of a reversible random walk over V, with stationary measure π. $\forall \tau \in \mathbb{N}, \forall x \in V$, define $p^\tau(x)$ as the scale-τ diffusion embedding of x where: $p^\tau(x) := (P^\tau)^\top e_x \in [0,1]^{|V|}$ and $\forall z \in V$, $p^\tau(x,z) := e_z^\top (P^\tau)^\top e_x \in [0,1]$. $\forall x, y \in V$, define the scale-τ diffusion similarity as $\phi_\tau^{\text{diff}}(x,y) := \langle p^\tau(x), p^\tau(y)\rangle_{\ell^2(1/\pi)} := \sum_{z\in V} p^\tau(x,z)\, p^\tau(y,z)\, \frac{1}{\pi(z)}$ and diffusion distance as $\rho_\tau^{\text{diff}}(x,y) := ||p^\tau(x) - p^\tau(y)||_{\ell^2(1/\pi)} := \sqrt{\sum_{z\in V}(p^\tau(x,z) - p^\tau(y,z))^2 \frac{1}{\pi(z)}}$ [CL06, Yao11].*

For $y, y' \in N(x)$, the (y,y')-th entry of $M(x,\tau)$ is $\phi_\tau^{\text{diff}}(y,y')$. Computing the diffusion similarity requires $\frac{1}{\pi(z)}$ for relevant $z \in V$, but $\pi(z)$ need only be computed up to a normalizing constant for the optimization problem in Equation (5.1). For a conjecturer (C_w, P_w, w) satisfying Definition 4.1, with $P := P_w$, $\forall z \in V, \pi(z) \propto w(z)$. Let $M'(x,\tau)$ be the principal submatrix of $P^\tau \text{diag}(1/w)(P^\tau)^\top$ indexed by $N(x)$ and $\forall z \in V$, $\text{diag}(1/w)[z,z] = \frac{1}{w(z)}$. With regularization strength $\eta \in [0, 0.5]$, $\forall x \in V$, Algorithm 3 computes:

$$C(x) \leftarrow \arg \max_{\substack{\mu\in\Delta(N(x)) \\ \mu(y)\geq\eta/\deg(x),\ \forall y\in N(x)}} \text{Div}_\tau^P(\mu) = \arg \min_{\substack{\mu\in\Delta(N(x)) \\ \mu(y)\geq\eta/\deg(x),\ \forall y\in N(x)}} \mu^\top M'(x,\tau)\mu \quad (5.2)$$

This mapping $C : V \to \Delta(V)$ defined in Algorithm 3 can be substituted in for C_w in line 4 of Algorithm 2, called in line 4 of the prover-conjecturer system (Algorithm 1). If G is finite, connected, undirected, and non-bipartite, the regularization with $\eta > 0$ ensures (via Lemma F.2) the random walk induced by C converges to a stationary measure, eventually exploring the entire graph. Algorithm 3 uses Algorithm 4 to compute diffusion embeddings.

As a concrete guarantee, Proposition D.2 shows that Algorithm 3 provably explores the theorem graph G faster than a simple random walk when G is in a special class of clustered graphs.

In general, τ should not be set too small or too large. If $\tau = 0$, the diffusion embedding is not informative since every node $x \in V$ is embedded as $e_x \in \mathbb{R}^{|V|}$. If τ is larger than the mixing time of P, then $\forall x \in V, (P^\top)^\tau e_x \approx \pi$, so the diffusion embedding is also not informative.

5.3 Computing Diffusion Similarity via Contrastive Learning

Theorem 5.5 shows the diffusion similarity (used in step 4 of Algorithm 3) can be computed via learned contrastive embeddings. A self-contained summary of contrastive learning is in Section C.

Algorithm 4 Dynamic Program for $p^\tau(x)$

Require: $x \in V$, $\tau \in \mathbb{N}$, neighbor oracle $N : V \to 2^V$, and reversible conjecturer (C_w, P_w, w)
Ensure: $p^\tau(x)$
1: $B_{-1} \leftarrow \emptyset$, $B_0 \leftarrow \{x\}$
2: **for** $s = 1, \ldots, \tau$ **do**
3: $\quad \partial B_{s-1} \leftarrow B_{s-1} \setminus B_{s-2}$
4: $\quad B_s \leftarrow B_{s-1} \cup \bigcup_{v \in \partial B_{s-1}} N(v)$
5: $S \leftarrow B_\tau$; initialize $p^0(x, y) \leftarrow \mathbf{1}\{y = x\}$ for $y \in S$
6: **for** $s = 1, \ldots, \tau$ **do**
7: $\quad p^s(x, y) \leftarrow \sum_{z \in S} p^{s-1}(x, z) P_w[z, y]$ for all $y \in S$
8: **return** $p^\tau(x)$

Theorem 5.5. *Let $G = (V, E)$ be a finite, undirected, connected theorem graph. Let conjecturer (C, P, w) satisfy Definition 4.1, and $\pi^\top = \pi^\top P$. Let $\mathcal{F}_{\text{emb}}$ be a class of embedding models $\mathcal{F}_{\text{emb}} \ni h_\theta : V \to \mathbb{R}^{|V|}$. For any $\tau \in \mathbb{N}$, define the scale-τ contrastive loss as:*

$$\mathcal{L}^\tau_{\text{cont}}(h_\theta) := \mathbb{E}_{(x,x^-)\sim D^\tau_-} (h_\theta(x)^\top h_\theta(x^-))^2 - 2 \cdot \mathbb{E}_{(x,x^+)\sim D^\tau_+} h_\theta(x)^\top h_\theta(x^+) \tag{5.3}$$

Where $D^\tau_+, D^\tau_- \in \Delta(V \times V)$ are distributions of positive and negative pairs: $(a, b) \sim D^\tau_+$ where $z \sim \pi, a \sim C^\tau(z), b \sim C^\tau(z)$ (C^τ denotes applying C τ times) and $(a, b) \sim D^\tau_-$ where $a \sim \pi, b \sim \pi$. *Fix any $\tau \in \mathbb{N}$. Assume $\mathcal{F}_{\text{emb}}$ contains a global minimizer of $\mathcal{L}^\tau_{\text{cont}}$ over functions $V \to \mathbb{R}^{|V|}$. For any global minimizer $\mathcal{F}_{\text{emb}} \ni h_{\theta_*} : V \to \mathbb{R}^{|V|}$, then $\forall u, v \in V, \langle p^\tau(u), p^\tau(v) \rangle_{\ell^2(1/\pi)} = h_{\theta_*}(u)^\top h_{\theta_*}(v)$.*

The proof of Theorem 5.5 is in Section C.2. Theorem 5.5 shows that the diffusion similarity of any two nodes on the graph (a quantity computed in step 3 and used in step 4 of Algorithm 3) can instead be computed from the Euclidean inner product of their learned embeddings. Given n samples $\{x_i\}_{i \in [n]} \subset V$ drawn i.i.d. from the stationary measure $\pi \in \Delta(V)$ of P, an embedding model $\mathcal{F}_{\text{emb}} \ni h_{\hat{\theta}} : V \to \mathbb{R}^{|V|}$ can be learned by computing the empirical risk minimizer (ERM, Definition C.1) of the empirical contrastive loss $\hat{\mathcal{L}}_{\text{cont},n}$ in Equation (C.15). Under additional assumptions, we show in Theorem C.17 an end-to-end result where ERM, in a finite-sample regime with small embedding dimension, yields embeddings that can be used to approximate the diffusion similarity. Also, in Section C.1.5, we discuss how to instantiate Theorem 5.5 and Theorem C.17 for a *local* subgraph of G around the theorems of interest, rather than needing to take G to be the entire theorem graph, which is impractical.

6 Conclusion and Limitations

We formalized the space of theorems as a graph and a conjecturer as a nearest-neighbor-type Markov chain over the theorem graph. Theorem 4.2 shows that under Assumption 5, the knowledge of the prover trained via Algorithm 1 grows exponentially with the number of iterations. For the case where Assumption 5 does not hold, we propose a diversity measure (Definition 5.2) for a training distribution of theorems generated by a conjecturer and an improved conjecturing algorithm (Algorithm 3) that locally maximizes this diversity measure, by computing the diffusion similarity (Definition 5.4) between neighboring theorems in the theorem graph. Finally, we describe a method to compute the diffusion similarity using contrastive learning (Theorem 5.5 and Theorem C.17). One limitation of the current framework is that the conjecturer is not adaptive toward the particular prover f_t at iteration t. Another intuition that is not captured by the current framework is that of theorem compositionality. One can imagine a conjecturer that has knowledge of two theorems $X \to Y$ and $Y \to Z$ may be able to conjecture $X \to Z$, but this is not easily compatible with the view that a conjecturer maps a single theorem to a distribution of theorems. Finally, we note that this theoretical framework could potentially serve as a framework for understanding automated synthetic data generation for other verifiable domains besides formal theorem proving.

Acknowledgements

TC acknowledges funding from an NSF Graduate Research Fellowship. This work was supported in part by an OpenAI Superalignment Fast Grant and by DARPA AIQ under Agreement No. HR00112520023.

We are deeply grateful to Tengyu Ma for many insightful discussions. We also thank Kaiyue Wen for discussions.

References

[A+25] Tudor Achim et al. Aristotle: IMO-level automated theorem proving. *arXiv preprint arXiv:2510.01346*, 2025.

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[BBLS26] Maria-Florina Balcan, Avrim Blum, Zhiyuan Li, and Dravyansh Sharma. On learning verifiers and implications to chain-of-thought reasoning, 2026.

[BCE+25] Sébastien Bubeck, Christian Coester, Ronen Eldan, Timothy Gowers, Yin Tat Lee, Alexandru Lupsasca, Mehtaab Sawhney, Robert Scherrer, Mark Sellke, Brian K. Spears, Derya Unutmaz, Kevin Weil, Steven Yin, and Nikita Zhivotovskiy. Early science acceleration experiments with gpt-5, 2025.

[BDF26] Maissam Barkeshli, Michael R. Douglas, and Michael H. Freedman. Artificial intelligence and the structure of mathematics, 2026.

[BEGK02] Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability and low lying spectra in reversible markov chains. *Communications in Mathematical Physics*, 2002.

[BJY20] Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play, 2020.

[C+25] Luoxin Chen et al. Seed-Prover: Deep and broad reasoning for automated theorem proving. *arXiv preprint arXiv:2507.23726*, 2025.

[CL06] Ronald R. Coifman and Stéphane Lafon. Diffusion maps, 2006.

[DM25] Kefan Dong and Tengyu Ma. Stp: Self-play llm theorem provers with iterative conjecturing and proving, 2025.

[dMU21] Leonardo de Moura and Sebastian Ullrich. The lean 4 theorem prover and programming language. *Automated Deduction–CADE 28: 28th International Conference on Automated Deduction*, 2021.

[EY36] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank, 1936.

[Fan50] Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations, 1950.

[GL06] Herman Geuvers and Iris Loeb. Natural deduction via graphs: Formal definition and computation rules, 2006.

[H+25] Thomas Hubert et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 2025. Published online Nov. 12, 2025.

[HWGM22] Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss, 2022.

[JLW+25] Xingguang Ji, Yahui Liu, Qi Wang, Jingyuan Zhang, Yang Yue, Rui Shi, Chenxi Sun, Fuzheng Zhang, Guorui Zhou, and Kun Gai. Leanabell-Prover-V2: Verifier-integrated reasoning for formal theorem proving via reinforcement learning. *arXiv preprint arXiv:2507.08649*, 2025.

[Kre78] Erwin Kreyszig. Introductory functional analysis with applications, 1978.

[KWLS25] Juno Kim, Denny Wu, Jason Lee, and Taiji Suzuki. Metastable dynamics of chain-of-thought reasoning: Provable benefits of search, rl and distillation, 2025.

[L+25] Yong Lin et al. Goedel-Prover-V2: Scaling formal theorem proving with scaffolded data synthesis and self-correction. *arXiv preprint arXiv:2508.03613*, 2025.

[LL25] Thang Luong and Edward Lockhart. Advanced version of gemini with deep think officially achieves gold-medal standard at the international mathematical olympiad, 2025.

[LPW17] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. Markov chains and mixing times, 2017.

[LTL+25] Yong Lin, Shange Tang, Bohan Lyu, Jiayun Wu, Hongzhou Lin, Kaiyu Yang, Jia Li, Mengzhou Xia, Danqi Chen, Sanjeev Arora, and Chi Jin. Goedel-Prover: A frontier model for open-source automated theorem proving. *arXiv preprint arXiv:2502.07640*, 2025.

[Ma22] Tengyu Ma. Lecture notes for machine learning theory, 2022.

[OW91] Michael L. Overton and Robert S. Womersley. On the sum of the largest eigenvalues of a symmetric matrix, 1991.

[PBHG24] Gabriel Poesia, David Broman, Nick Haber, and Noah D. Goodman. Learning formal mathematics from intrinsic motivation, 2024.

[R+25] Z. Z. Ren et al. DeepSeek-Prover-V2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2504.21801*, 2025.

[SHS+17] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017.

[SWP+25] Shijie Shang, Ruosi Wan, Yue Peng, Yutong Wu, Xiong hui Chen, Jie Yan, and Xiangyu Zhang. StepFun-Prover Preview: Let's think and verify step by step. *arXiv preprint arXiv:2507.20199*, 2025.

[TBS+16] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow, 2016.

[VCLR13] Vincent Q. Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: Near-optimal convex relaxation of sparse pca, 2013.

[W+25] Haiming Wang et al. Kimina-Prover Preview: Towards large formal reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.11354*, 2025.

[WHZ+24] Zijian Wu, Suozhi Huang, Zhejian Zhou, Huaiyuan Ying, Jiayu Wang, Dahua Lin, and Kai Chen. InternLM2.5-stepprover: Advancing automated theorem proving via expert iteration on large-scale LEAN problems. *arXiv preprint arXiv:2410.15700*, 2024.

[Woe04] Wolfgang Woess. Random walks on infinite graphs and groups, 2004.

[X+25a] Huajian Xin et al. DeepSeek-Prover-V1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. In *International Conference on Learning Representations*, 2025.

[X+25b] Ran Xin et al. BFS-prover: Scalable best-first tree search for LLM-based automatic theorem proving. *arXiv preprint arXiv:2502.03438*, 2025.

[XZN+25] Ran Xin, Zeyu Zheng, Yanchen Nie, Kun Yuan, and Xia Xiao. Scaling up multi-turn off-policy RL and multi-agent tree search for LLM step-provers. *arXiv preprint arXiv:2509.06493*, 2025.

[Yao11] Yuan Yao. Lecture 6. diffusion distance, 2011.

[Z+25] Yichi Zhou et al. Solving formal math problems by decomposition and iterative reflection. *arXiv preprint arXiv:2507.15225*, 2025.

Contents

1 Introduction **1**

1.1 Technical Overview 2

2 Related Work **4**

3 Preliminaries **4**

4 Framework and Random-Exploration Baseline **4**

4.1 Primitive Assumptions 4

4.2 Random-Walk Conjecturing Algorithm 6

4.3 Expansion Guarantee for Algorithm 1. 7

5 Diversity Metric and an Improved Conjecturing Algorithm **7**

5.1 Diversity Measure 7

5.2 Improved Conjecturing Algorithm 8

5.3 Computing Diffusion Similarity via Contrastive Learning 9

6 Conclusion and Limitations **10**

A Further Related Works **15**

B Proof of Theorem 4.2 **15**

B.1 Preliminaries. 15

B.2 Proof of Theorem 4.2 20

C Theorems about Diffusion Similarity and Contrastive Learning **22**

C.1 Preliminary Lemmas 22

C.1.1 Setup 22

C.1.2 Lemmas on Contrastive Learning 25

C.1.3 The Connection Between Diffusion Embeddings and Contrastive Learning 26

C.1.4 Matrix Factorization Lemmas 31

C.1.5 Learning Contrastive Embeddings on Local Subgraphs 35

C.2 Main Theorems 35

D Instantiating Algorithm 3 on Clustered Graphs **39**

D.1 End-to-End Result of Algorithm 3 on Clique Graphs 39

E Justification of Assumption 2 **42**

E.1 Statistical Learning Theory Inspired Assumptions 43

E.2 Proof of Lemma E.1 44

F Auxiliary Lemmas **45**

G Figures 47

A Further Related Works

Theoretical Analyses of Self-play Algorithms. Existing theoretical analyses of self-play largely focus on adversarial settings, such as zero-sum Markov games. The theorem-proving setting we study is different: the prover and conjecturer have asymmetric but cooperative roles. [BJY20] provide a theoretical analysis of self-play algorithms in zero-sum Markov Games, showing convergence to a Nash Equilibrium [BJY20]. The Wasserstein GAN training algorithm [ACB17] is a theoretically-grounded asymmetric self-play algorithm, but still where the agents are adversarial to each other. Our work also differs from works that study how to learn a model to verify theorems in a purely statistical setting [BBLS26]. Our setting is non-statistical since the goal of the training algorithm is to allow the final model to generalize beyond its starting distribution, not just to generalize well on its starting distribution given i.i.d. samples. However, we will assume some statistical-learning primitives (Assumption 2) in our framework.

Theorem Graphs. Prior works formalize the graph or hyper-graph structure of mathematical theorems [GL06, BDF26]. In [GL06], each node represents a statement; the statement corresponding to a node can be deduced in one step via a deduction rule from the statements corresponding to its descendants. [BDF26] formalizes nodes as statements, and each (p, q) hyper-edge corresponds to applying a deduction rule on p input statements to yield q output conclusions. Axioms of the logical system are root nodes. In relation to these works, our notion of a theorem graph still represents mathematical statements as nodes. However, the definition of an edge is not defined in terms of deduction rules in a particular logical system. Rather, edges are determined by the class of prover models $\mathcal{F}_{\text{prover}}$, where an edge between two theorems (x, y) indicates that for every prover $f \in \mathcal{F}_{\text{prover}}$, the success rate of f on theorem x is not too different from the success rate of f on theorem y.

Graph Exploration Algorithms, Graph Diffusion, Contrastive Learning. Given a graph, the question of how to design an algorithm to explore the graph efficiently has been studied in [KWLS25]. They formalize chain-of-thought reasoning as a random walk over a graph of ideas, where the goal of the reasoner is to hit a particular node of the graph, starting at a designated start node. The underlying graphs they study are metastable graphs [BEGK02]. The question we study in Section 5 is conceptually related to [KWLS25], in that we want to find a graph algorithm that can efficiently explore a graph with poor connectivity, such as metastable graphs. However, our proposed graph-exploration algorithms are different from [KWLS25]. In particular, our algorithm uses a similarity measure between nodes in the graph, called the diffusion similarity, that is defined in terms of a random walk on the graph. The diffusion similarity measure is closely related to the diffusion distance, a graph metric which is useful for characterizing the clustering-structure of a graph [CL06]. Finally, we propose a new way to compute the diffusion similarity, using a version of contrastive learning formalized by [HWGM22], who provide a theoretical characterization of the global minimizers of a particular contrastive loss in terms of the minimizers of a related matrix factorization objective.

B Proof of Theorem 4.2

The purpose of this section is to eventually prove Theorem 4.2 in Section B.2.

B.1 Preliminaries.

Isoperimetric Inequality. Let $G = (V, E)$ be an undirected graph with a countable set of vertices, V, and edges E. G is connected, and is locally finite (every vertex has finite degree). Let P be a Markov chain transition matrix over V (We will refer to the pair (V, P) as a Markov chain). The simple random walk is a Markov chain where every node x transitions uniformly to one of its neighbors in the graph, $N(x) := \{x' \in V : (x, x') \in E\}$. In the following, we define some basic properties of Markov chains over graphs.

For $n \in \mathbb{N}$, let $p^{(n)}(x, y)$ denote the probability the Markov chain (V, P) initialized at x is at y at the n-th step.

Definition B.1. *(Irreducibility) Markov chain (V, P) is irreducible if $\forall x, y \in V$, there is some $n \in \mathbb{N}$ such that $p^{(n)}(x, y) > 0$.*

For the remaining definitions, we will assume the Markov chain is irreducible. We define aperiodicity as follows.

Definition B.2. *(Aperiodicity) A Markov chain (V, P) is aperiodic if for any $x \in V$, $\gcd\{n \in \mathbb{N} : p^{(n)}(x, x) > 0\} = 1$.*

Note that if a Markov chain is irreducible and for some $x \in V$, $\gcd\{n \in \mathbb{N} : p^{(n)}(x, x) > 0\} = 1$, then for any $x' \in V$, $\gcd\{n \in \mathbb{N} : p^{(n)}(x', x') > 0\} = 1$.

A stationary measure $\pi : V \to (0, \infty)$ where $\pi^\top = \pi^\top P$. If $||\pi||_1 = 1$, then it is a stationary distribution. The following Theorem shows that irreducibility and aperiodicity imply mixing for finite Markov chains.

Lemma B.3. *(Theorem 4.9 of [LPW17]) Suppose that Markov chain (V, P) is finite ($|V| < \infty$), irreducible and aperiodic with stationary distribution π. Then there exist constants $C > 0$ and $\alpha \in (0, 1)$ where*

$$\max_{x \in V} ||p^{(n)}(x, \cdot) - \pi||_{\text{TV}} \leq C\alpha^n \tag{B.1}$$

The following definitions introduce the notion of reversibility and conductance.

Definition B.4. *(Reversibility) Let V be an infinite, countable set. Markov chain (V, P) is reversible if there exists a (stationary) measure $\pi : V \to (0, \infty)$ such that*

$$\pi(x)P[x, y] = \pi(y)P[y, x] \quad \forall x, y \in V \tag{B.2}$$

Definition B.5. *(Conductance) Let $a(x, y) = \pi(x)P[x, y] = a(y, x)$ be the conductance between x and y. $\pi(x)$ is the total conductance at x. For a Markov chain transition matrix P, denote its edge set $E(P)$ where:*

$$(x, y) \in E(P) \iff a(x, y) > 0 \tag{B.3}$$

Note, a nearest-neighbor-type random walk is a Markov chain (V, P) where $E(P) = E$ (the edge-set of the graph G).

Now, we define the Isoperimetric Inequality for a Markov chain over a graph, which will generalize Definition 3.1. Let (V, P) be a reversible, irreducible Markov chain with total conductance $\pi(\cdot)$, edge conductance $a(\cdot, \cdot)$. Define the following quantities.

$$\pi(A) := \sum_{x \in A} \pi(x), \quad \forall A \subset V \tag{B.4}$$

$$a(D) := \sum_{e \in D} a(e^-, e^+), \quad \forall D \subset E(P) \tag{B.5}$$

$$\partial A := \text{set of all edges in } E(P) \text{ with} \tag{B.6}$$

$$\text{one endpoint in } A \text{ and the other in } V - A \tag{B.7}$$

For the definition of ∂A, note that the graphs we study in this work are all undirected. We now define the Isoperimetric Inequality.

Definition B.6 (Isoperimetric Inequality)**.** *Let $\mathscr{F} : \mathbb{R}_+ \to \mathbb{R}_+$. A Markov chain (V, P) satisfies the $\mathscr{F}$-Isoperimetric Inequality, if there exists a constant $\kappa > 0$ such that for every finite $A \subset V$, $\mathscr{F}(\pi(A)) \leq \kappa\, a(\partial A)$.*

The particular function $\mathscr{F} : \mathbb{R}_+ \to \mathbb{R}_+$ we care about in this work is $\mathscr{F}(t) = t$, corresponding to the Strong Isoperimetric Inequality defined in Definition 3.1. However, for the proofs of some of the following Lemmas, it is useful to define, for $1 \leq d \leq \infty$, the d-dimensional Isoperimetric Inequality IS_d as Definition B.6 with $\mathscr{F}(t) = t^{1-1/d}$. The strong Isoperimetric Inequality IS can be thought of as IS_d with $d = \infty$.

Hilbert Spaces and Norms. With (V, P) being an irreducible, reversible Markov chain with stationary measure $\pi : V \to (0, \infty)$, let $\ell_0(V)$ denote the linear space of finitely supported functions on V (i.e. functions that are non-zero only on a finite number of elements of V). For an (undirected) edge $e = (e^-, e^+) \in E$, define the resistance as $r(e) := \frac{1}{a(e^-,e^+)}$. The following defines Hilbert spaces of functions, relevant to Theorem 4.2.

Definition B.7. *Define the Hilbert space $\ell^2(V, \pi)$ with inner product*

$$\langle f, g\rangle_{\ell^2(V,\pi)} = \sum_{v\in V} f(v)g(v)\pi(v) \tag{B.8}$$

$\ell^2(V, \pi)$ is the set of functions $f : V \to \mathbb{R}$ with $\sum_{v\in V} |f(v)|^2\pi(v) < \infty$.

Define the Hilbert space $\ell^2(V, 1/\pi)$ with inner product

$$\langle f, g\rangle_{\ell^2(V,1/\pi)} = \sum_{v\in V} f(v)g(v)\frac{1}{\pi(v)} \tag{B.9}$$

$\ell^2(V, 1/\pi)$ is the set of functions $f : V \to \mathbb{R}$ with $\sum_{v\in V} |f(v)|^2 \frac{1}{\pi(v)} < \infty$. When clear from context, we will write $\ell^2(\pi)$ and $\ell^2(1/\pi)$ for $\ell^2(V, \pi)$ and $\ell^2(V, 1/\pi)$, respectively.

Analogously, the Hilbert space $\ell^2(E, r)$ has inner product

$$\langle u, v\rangle_{\ell^2(E,r)} = \sum_{e\in E} u(e)v(e)r(e) \tag{B.10}$$

$\ell^2(E, r)$ is the set of functions $f : E \to \mathbb{R}$ with $\sum_{e\in E} |f(e)|^2 r(e) < \infty$.

We introduce the following difference operator

$$\nabla : \ell^2(V, \pi) \to \ell^2(E, r), \quad \nabla f(e) = \frac{f(e^+) - f(e^-)}{r(e)} \tag{B.11}$$

Define the following norms:

Definition B.8. *(Norms) For a function $f : V \to \mathbb{R}$, define its Sobolev norm*

$$S_P(f) := \frac{1}{2} \sum_{x,y\in V} |f(x) - f(y)|a(x, y) \tag{B.12}$$

and its norm in $\ell^p(V, \pi)$

$$||f||_p = \Big(\sum_{x\in V} |f(x)|^p \pi(x) \Big)^{1/p} \tag{B.13}$$

whenever these sums converge. Let $\mathcal{D}(\mathcal{P})$ be the space of all functions f on V such that $\nabla f \in \ell^2(E, r)$. If f is such a function, define its Dirichlet norm

$$D_P(f) = \langle \nabla f, \nabla f\rangle_{\ell^2(E,r)} \tag{B.14}$$

$$= \sum_{e\in E} \frac{(f(e^+) - f(e^-))^2}{r(e)} \tag{B.15}$$

$$= \frac{1}{2} \sum_{x,y\in X} (f(x) - f(y))^2 a(x, y) \tag{B.16}$$

Preliminary Lemmas. The next few Lemmas will be used to prove Theorem B.14, which is key to proving the final Theorem 4.2.

Lemma B.9. *(Proposition 4.3 of [Woe04]) Markov chain (V, P) satisfies IS_d $(1 \le d \le \infty)$ if and only if*

$$||f||_{\frac{d}{d-1}} \le \kappa\, S_P(f) \quad \text{for every } f \in \ell_0(V) \tag{B.17}$$

Proof. See Proposition 4.3 of [Woe04] on page 40. □

Lemma B.10. *(Lemma 2.4 of [Woe04]) If $f \in \mathcal{D}(P)$, then $\nabla^*(\nabla f) = -(P - I)f$*

Proof. See Lemma 2.4 of [Woe04] on page 16. □

Note that $\ell_0(V) \subset \mathcal{D}(P)$.

Now, we recall some facts of linear operators. Let $Q = (q(i,j))_{i,j\in I}$ be a non-negative real matrix over a countable set. Q acts on functions (vectors) $I \to \mathbb{R}$ in the usual way by matrix multiplication on the left. Define the operator norm:

Definition B.11. *(Definition 2.7-1 of [Kre78]) Let X, Y be normed spaces and $Q : D(Q) \to Y$ a linear operator, where $D(Q) \subset X$. Q is bounded if there is a real number c s.t.*

$$\forall x \in D(Q), ||Qx|| \leq c||x|| \tag{B.18}$$

Moreover, define $||Q|| = \sup_{x\in D(Q), x\neq 0} \frac{||Qx||}{||x||}$

Define the adjoint of a linear operator:

Lemma B.12. *(Theorem 3.9-2 of [Kre78]) The Hilbert-adjoint operator Q^* of a bounded linear operator $Q : H_1 \to H_2$ such that*

$$Q^* : H_2 \to H_1 \tag{B.19}$$

$$\forall x \in H_1, y \in H_2, \langle Qx, y\rangle = \langle x, Q^* y\rangle \tag{B.20}$$

is unique and is a bounded linear operator with norm:

$$||Q^*|| = ||Q|| \tag{B.21}$$

If $Q = Q^*$, it is self-adjoint, and its operator norm has the following Rayleigh characterization.

Lemma B.13. *(Theorem 9.2-2 of [Kre78]) For any bounded, self-adjoint Q on a complex Hilbert space H, we have*

$$||Q|| = \sup_{||x||=1} |\langle Qx, x\rangle| \tag{B.22}$$

Proof. See Theorem 9.2-2 of [Kre78]. □

Key Theorem. Theorem B.14 is the key theorem which enables us to prove Theorem 4.2.

Theorem B.14. *(IS implies Spectral Expansion)*

Consider an irreducible, reversible Markov chain (V, P) that satisfies the strong Isoperimetric inequality with Isoperimetric constant κ:

$$\text{For every finite } A \subset V, \quad \pi(A) \leq \kappa \cdot a(\partial A) \tag{B.23}$$

Then, the operator norm of P in $\ell^2(V, \pi)$ is at most $1 - \frac{1}{2\kappa^2}$. That is, $||P||_{\ell^2(V,\pi)\to\ell^2(V,\pi)} \leq 1 - \frac{1}{2\kappa^2}$.

Proof. This follows from the proof of Theorem 10.3 of [Woe04], but we write it out here for completeness.

By Lemma B.9, IS implies the Sobolev inequality $||f||_1 \leq \kappa\, S_P(f)$ for all $f \in \ell_0(V)$. First, we show that this implies the Dirichlet inequality, $||f||_2^2 \leq 2\kappa^2 D_P(f)$, because:

$$
\begin{aligned}
||f||_2^4 &= ||f^2||_1^2 && \text{(B.24)}\\
&\leq \kappa^2 S_P(f^2)^2 && \text{(B.25)}\\
&\leq \kappa^2 \big(\frac{1}{2}\sum_{x,y\in V} a(x,y)|f(x)-f(y)|(|f(x)|+|f(y)|)\big)^2 && \text{(B.26)}\\
&\leq \kappa^2 D_P(f) \sum_{x,y\in V} a(x,y)(|f(x)|+|f(y)|)^2/2 && \text{(B.27)}\\
&\leq \kappa^2 D_P(f) \sum_{x,y\in V} a(x,y)(f(x)^2+f(y)^2) && \text{(B.28)}\\
&\leq 2\kappa^2 D_P(f)||f||_2^2 && \text{(B.29)}\\
\implies ||f||_2^2 &\leq 2\kappa^2 D_P(f) && \text{(B.30)}
\end{aligned}
$$

By Lemma B.10,

$$
\begin{aligned}
D_P(f) &= \langle f, (I-P)f\rangle_{\ell^2(\pi)} && \text{(B.31)}\\
&= ||f||_2^2 - \langle f, Pf\rangle_{\ell^2(\pi)} \quad \text{for } f\in\ell_0(V) && \text{(B.32)}\\
\implies \langle f, Pf\rangle_{\ell^2(\pi)} &= ||f||_2^2 - D_P(f) && \text{(B.33)}\\
&\leq (1-\frac{1}{2\kappa^2})||f||_2^2 && \text{(B.34)}
\end{aligned}
$$

Where the $||\cdot||_2^2 = ||\cdot||_{\ell^2(\pi)}^2$ is the weighted 2-norm defined in Definition B.8. By Lemma B.13,

$$
\begin{aligned}
||P||_{\ell^2(\pi)\to\ell^2(\pi)} &= \sup_{f\in\ell^2(\pi), f\neq 0} \frac{|\langle f, Pf\rangle_{\ell^2(\pi)}|}{||f||_2^2} && \text{(B.35)}\\
&\leq \sup_{f\in\ell^2(\pi), f\neq 0} \frac{|\sum_{x,y\in V} f(x)f(y)\pi(x)P[x,y]|}{||f||_2^2} && \text{(B.36)}\\
&\leq \sup_{f\in\ell^2(\pi), f\neq 0} \frac{\sum_{x,y\in V} \pi(x)P[x,y]\,|f(x)||f(y)|}{||f||_2^2} && \text{(B.37)}\\
&= \sup_{f\in\ell^2(\pi), f\neq 0} \frac{\langle |f|, P|f|\rangle_{\ell^2(\pi)}}{||f||_2^2} && \text{(B.38)}\\
&= \sup_{f\in\ell^2(\pi), f\neq 0} \frac{\langle f, Pf\rangle_{\ell^2(\pi)}}{||f||_2^2} && \text{(B.39)}\\
&= \sup_{f\in\ell_0(V), f\neq 0} \frac{\langle f, Pf\rangle_{\ell^2(\pi)}}{||f||_2^2} && \text{(Since } \ell_0(V) \text{ is dense in } \ell^2(V,\pi))\\
&\leq 1-\frac{1}{2\kappa^2} && \text{(By Equation (B.34))}
\end{aligned}
$$

We conclude that $||P||_{\ell^2(\pi)\to\ell^2(\pi)} \leq 1-\frac{1}{2\kappa^2}$, where $||P||_{\ell^2(\pi)\to\ell^2(\pi)}$ denotes the operator norm of P over $\ell^2(\pi)$.

□

The following Corollary establishes a one-step contraction due to the random walk according to P on a graph satisfying IS.

Corollary B.15. *Let $G=(V,E)$ and let (V,P) be a reversible random walk on G, with stationary measure π, satisfying Assumption 5 with Isoperimetric constant $\kappa>0$. Then, for any $D\in\ell^2(1/\pi)$, we have*

$$||P^\top D||_{\ell^2(1/\pi)} \leq (1 - \frac{1}{2\kappa^2})||D||_{\ell^2(1/\pi)} \tag{B.40}$$

Proof. The core claim is that $||P^\top||_{\ell^2(1/\pi)\to\ell^2(1/\pi)} = ||P||_{\ell^2(\pi)\to\ell^2(\pi)}$. To see this, because P is reversible with stationary measure π, we have

$$\text{diag}(\pi)P = P^\top \text{diag}(\pi) \tag{B.41}$$
$$P^\top = \text{diag}(\pi)P\text{diag}(\pi)^{-1} \tag{B.42}$$

Thus,

$$||P^\top||_{\ell^2(1/\pi)\to\ell^2(1/\pi)} = \sup_{x\,:\,||x||_{\ell^2(1/\pi)}=1} ||P^\top x||_{\ell^2(1/\pi)} \tag{B.43}$$
$$= \sup_{x\,:\,||x||_{\ell^2(1/\pi)}=1} ||\text{diag}(\pi)P\text{diag}(\pi)^{-1}x||_{\ell^2(1/\pi)} \tag{B.44}$$
$$= \sup_{x\,:\,||x||_{\ell^2(1/\pi)}=1} ||P\text{diag}(\pi)^{-1}x||_{\ell^2(\pi)} \tag{B.45}$$
$$= \sup_{x'\,:\,||x'||_{\ell^2(\pi)}=1} ||Px'||_{\ell^2(\pi)} \tag{B.46}$$
$$= ||P||_{\ell^2(\pi)\to\ell^2(\pi)} \tag{B.47}$$

Equation (B.46) follows from the fact that for any $x \in \ell^2(1/\pi)$, $x' = \text{diag}(\pi)^{-1}x \in \ell^2(\pi)$. Conversely, for any $x' \in \ell^2(\pi)$, $x = \text{diag}(\pi)x' \in \ell^2(1/\pi)$.

Finally, by Theorem B.14, $||P||_{\ell^2(\pi)\to\ell^2(\pi)} \leq 1 - \frac{1}{2\kappa^2}$. Since $||P||_{\ell^2(\pi)\to\ell^2(\pi)} = ||P^\top||_{\ell^2(1/\pi)\to\ell^2(1/\pi)}$, then for any $D \in \ell^2(1/\pi)$, we have

$$||P^\top D||_{\ell^2(1/\pi)} \leq ||P^\top||_{\ell^2(1/\pi)\to\ell^2(1/\pi)}\,||D||_{\ell^2(1/\pi)} \leq (1 - \frac{1}{2\kappa^2})||D||_{\ell^2(1/\pi)} \tag{B.48}$$

□

B.2 Proof of Theorem 4.2

Theorem 4.2 (Analysis of Algorithm 1)**.** *Let $G = (V, E)$ be a theorem graph with $|V| = \infty$ and (C_w, P_w, w) be a reversible conjecturer as in Definition 4.1. Suppose (V, P_w) satisfies Assumption 5 with Isoperimetric constant $\kappa > 0$. Let $\pi : V \to (0, \infty)$ be an invariant measure of (V, P_w). Suppose Assumptions 1-4 hold, with universal constant $p \in (0, 1)$ as in Assumption 2, and that π is chosen so that $\sup_{z\in V} \pi(z) < \infty$ and $0 < ||D_0||^2_{\ell^2(1/\pi)} < \infty$. For any $\delta, \epsilon \in (0, 1)$, Algorithm 1 produces a sequence of provers $\{f_t\}_{t\in\mathbb{N}}$ where $\forall t \in \mathbb{N}$, with probability at least $1 - (t + 1)\,\delta$,*

$$|\{x \in V : \text{succ}(x, f_t) \geq p\}| \geq \frac{(1-\epsilon)^2}{(\sup_{z\in V} \pi(z)) \cdot ||D_0||^2_{\ell^2(1/\pi)}} \cdot (\frac{1-\epsilon}{1-\frac{1}{2\kappa^2}})^{2t} \tag{4.1}$$

Proof. We reason inductively, by analyzing the behavior of Algorithm 1 each iteration.

First, suppose at iteration $t \in \mathbb{N}$, Algorithm 1 has i.i.d. sample access to $D_t \in \Delta(V)$, and trains prover $f_t \in \mathcal{F}_{\text{prover}}$ to where $\mathbb{P}_{x\sim D_t}[\text{succ}(x, f_t) \geq p] \geq 1 - \epsilon$ (an event which occurs with probability at least $1 - \delta$ according to Assumption 2). Algorithm 1 (using the subroutine, Algorithm 2) generates $C_w \# D_t$ from D_t and then generates D_{t+1} from $C_w \# D_t$ by filtering out theorems whose pass rate is below $p/2$. By Corollary B.15,

$$||C_w \# D_t||^2_{\ell^2(1/\pi)} := ||(P^\top) D_t||^2_{\ell^2(1/\pi)} \leq (1 - \frac{1}{2\kappa^2})^2 ||D_t||^2_{\ell^2(1/\pi)} \tag{B.49}$$

Under the event that $\mathbb{P}_{x \sim D_t}[\text{succ}(x, f_t) \geq p] \geq 1 - \epsilon$, let

$$\begin{aligned} \alpha_t &:= \mathbb{P}_{x \sim C_w \# D_t}[\text{succ}(x, f_t) \geq p/2] && \text{(B.50)} \\ &\geq \mathbb{P}_{x \sim D_t}[\text{succ}(x, f_t) \geq p] && \text{(B.51)} \\ &\geq 1 - \epsilon && \text{(B.52)} \end{aligned}$$

Let D_{t+1} be a new distribution whose sampling process involves sampling $x \sim C_w \# D_t$ until $\text{succ}(x, f_t) \geq p/2$ and then returning such an x (i.e. the rejection sampling step in Algorithm 2). This process provides i.i.d. sample access to D_{t+1} supported on the α_t-fraction of theorems in $C_w \# D_t$ where $\text{succ}(x, f_t) \geq p/2$. The probability of any theorem under D_{t+1} is equal to its respective probability under $C_w \# D_t$ multiplied by $\frac{1}{\alpha_t} \leq \frac{1}{1-\epsilon}$. Since for $f \in \ell^2(1/\pi)$, $||f||^2_{\ell^2(1/\pi)} = \sum_{v \in V} |f(v)|^2 \frac{1}{\pi(v)}$, then

$$||D_{t+1}||^2_{\ell^2(1/\pi)} \leq (\frac{1}{1-\epsilon})^2 ||C_w \# D_t||^2_{\ell^2(1/\pi)} \tag{B.53}$$

Combining with Equation (B.49), we conclude

$$||D_{t+1}||^2_{\ell^2(1/\pi)} \leq (\frac{1 - \frac{1}{2\kappa^2}}{1-\epsilon})^2 ||D_t||^2_{\ell^2(1/\pi)} \tag{B.54}$$

We now describe how these per-iteration guarantees can be put together inductively. We claim the following iteration invariant: for every iteration $t \in \mathbb{N}$, conditional on the event that all previous iterations did not fail, then Algorithm 1 has access to a theorem distribution D_t satisfying the premise of Assumption 2, that $\forall x \in \text{supp}(D_t)$, the proof of x is known. Moreover, the tth iteration fails with probability at most δ.

As the base case, Assumption 1 ensures Algorithm 1 has i.i.d. sample access to cold-start theorem distribution D_0, and that all proofs of theorems in $\text{supp}(D_0)$ are known. The iteration invariant holds.

Regarding the inductive step, suppose at iteration t, the algorithm has i.i.d. sample access to D_t, where for all $x \in \text{supp}(D_t)$, the proof of x is known. Suppose also that all previous iterations $\{0\} \cup [t-1]$ did not fail. By Assumption 2, Algorithm 1 can train a new prover f_t using i.i.d. sample access to D_t, which guarantees that $\mathbb{P}_{x \sim D_t}[\text{succ}(x, f_t) \geq p] \geq 1 - \epsilon$ with failure probability at most δ. By Assumption 3, $\mathbb{P}_{x \sim D_t}[\text{succ}(x, f_t) \geq p] \geq 1 - \epsilon$ implies that $\mathbb{P}_{x \sim C_w \# D_t}[\text{succ}(x, f_t) \geq p/2] \geq 1 - \epsilon$. The rejection sampling procedure in Algorithm 2, which generates D_{t+1} from $C_w \# D_t$, ensures that $\forall x \in \text{supp}(D_{t+1})$, $\text{succ}(x, f_t) \geq p/2$, so that it is efficient to sample the proof of any theorem in $\text{supp}(D_{t+1})$ from f_t. Thus, the proof of every theorem in $\text{supp}(D_{t+1})$ is known, and D_{t+1} satisfies the premise of Assumption 2. This iteration fails with probability at most δ, which occurs if step that trains f_t from D_t fails.

Thus, by induction over the iteration number $t \in \mathbb{N}$, we conclude from Equation (B.54) that for every $t \in \mathbb{N}$, conditional on the event that every iteration $i \in \{0\} \cup [t-1]$ does not fail (which occurs with probability at least $1 - t\,\delta$ since each iteration fails with probability at most δ), then

$$||D_t||^2_{\ell^2(1/\pi)} \leq (\frac{1 - \frac{1}{2\kappa^2}}{1-\epsilon})^{2t} ||D_0||^2_{\ell^2(1/\pi)} \tag{B.55}$$

Finally, Algorithm 1 trains the final prover $f_t \in \mathcal{F}_{\text{prover}}$ on D_t at iteration t. With failure probability at most δ, we have $D_t(\{x \in V : \text{succ}(x, f_t) \geq p\}) \geq 1 - \epsilon$.

At iteration $t \in \mathbb{N}$, suppose none of the training guarantees of the $t+1$ provers $\{f_0, f_1, \ldots, f_t\}$ fail, which occurs with probability at least $1 - (t+1)\,\delta$. Applying Lemma F.1 with $D' = D_t$,

$A' = \{x \in V : \text{succ}(x, f_t) \geq p\}$ and the guarantee that $D_t(\{x \in V : \text{succ}(x, f_t) \geq p\}) \geq 1 - \epsilon$, we have

$$|\{x \in V : \text{succ}(x, f_t) \geq p\}| \geq \frac{(1-\epsilon)^2}{\sup_{z \in V} \pi(z)} \cdot \frac{1}{||D_t||^2_{\ell^2(1/\pi)}} \tag{B.56}$$

By Equation (B.55), the conclusion follows.

$$|\{x \in V : \text{succ}(x, f_t) \geq p\}| \geq \frac{(1-\epsilon)^2}{(\sup_{z \in V} \pi(z)) \cdot ||D_0||^2_{\ell^2(1/\pi)}} \cdot (\frac{1-\epsilon}{1 - \frac{1}{2\kappa^2}})^{2t}. \tag{B.57}$$

Note also that since $\mathbb{P}_{x \sim C_w \# D_t}[\text{succ}(x, f_t) \geq p/2] \geq 1 - \epsilon$ at iteration t, the probability of rejection each iteration is at most ϵ, so that the rejection sampling procedure per iteration is fairly efficient. However, the probability of rejection compounds over iterations, so that the expected number of samples from D_0 needed to sample from D_t is proportional to $(\frac{1}{1-\epsilon})^t$.

□

C Theorems about Diffusion Similarity and Contrastive Learning

The purpose of this section is to show the connection between contrastive learning and diffusion similarity, and show how an embedding model learned by minimizing a contrastive loss can be used to compute the diffusion similarity (and diffusion distance) between points on a graph. Our main results, in Section C.2, include Theorem 5.5, an infinite-sample and large embedding dimension learning result, and Theorem C.17, a finite-sample and small embedding dimension learning result.

C.1 Preliminary Lemmas

C.1.1 Setup

Contrastive learning defines an algorithm for learning an embedding model that maps elements from a domain V (e.g. the set of real-world images) into Euclidean space $\mathbb{R}^k$ for some $k \in \mathbb{N}$. It is most notably used for training neural networks as embedding models, and we let $\mathcal{F}_{\text{emb}}$ denote the hypothesis class of models $h_\theta : V \to \mathbb{R}^k$ which we assume contains the ground-truth embedding model that we must learn. The core idea of contrastive learning is to define two distributions of pairs of elements in V: a positive pair distribution $D_+ \in \Delta(V \times V)$, and a negative pair distribution $D_- \in \Delta(V \times V)$. Given i.i.d. samples from both distributions, the trained model is encouraged to embed two elements sampled from a positive pair $(x, y) \sim D_+$ close together in Euclidean space, while embedding two elements sampled from a negative pair $(x, x') \sim D_-$ far apart in Euclidean space. These two distributions of pairs of elements are used to define a contrastive loss, and training the model amounts to minimizing an empirical estimate of the contrastive loss from finite samples.

In fact, contrastive learning can be described mathematically through the lens of matrix factorization, as articulated in [HWGM22]. The benefit of this viewpoint of contrastive learning is that it enables us to provably characterize the learned embeddings and the sample complexity of learning them. In the following, we will describe and use the mathematical setup for contrastive learning of [HWGM22].

Let the domain V consist of $N \in \mathbb{N}$ elements, where N is large but finite. We will identify V with $[N]$ here for simplicity. Define $\mu \in \Delta([N])$ as the *base distribution* and $Q \in \mathbb{R}^{N \times N}$ as an augmentation scheme. Alternatively, overloading notation, we can view $Q : [N] \to \Delta(N)$ as a mapping from an element x to a distribution over augmentations of x, $Q(x)$. For every $x \in [N]$, this distribution of augmentations, $Q(x)$, equals the x-th row of the matrix-representation of Q.

We will first introduce the matrix factorization viewpoint of contrastive learning, and we will from this how to define the positive and negative pair distributions $D_+, D_- \in \Delta([N] \times [N])$ from μ and Q and the final learning algorithm. Let $A \in \mathbb{R}^{N \times N}$ and its normalization $\overline{A}$ be positive semi-definite (p.s.d.) matrices of the form,

$$A = Q^\top \operatorname{diag}(\mu)\, Q \tag{C.1}$$
$$D = \operatorname{diag}(A\mathbf{1}) \tag{C.2}$$
$$\overline{A} = D^{-1/2} A D^{-1/2} \tag{C.3}$$

For rank-k factor $F_\theta \in \mathbb{R}^{N\times k}$, let $\mathcal{L}_{\text{mf}}(F_\theta) := ||\overline{A} - F_\theta F_\theta^\top||_F^2$ be the matrix factorization loss for matrix $\overline{A} \in \mathbb{R}^{N\times N}$. Let $\mathcal{F}_{\text{emb}}$ be the class of embedding models. We will parameterize F_θ by an embedding model $h_\theta : [N] \to \mathbb{R}^k$. Towards this, if we let σ_{xy} denote the (x, y)-th entry of A,

$$A = [\sigma_{xy}]_{x,y\in[N]} \tag{C.4}$$
$$\text{With } \sigma_x := \sum_{y\in[N]} \sigma_{xy} \tag{C.5}$$

and we parameterize $F_\theta \in \mathbb{R}^{N\times k}$ where $\forall x \in [N], F_\theta[x] = \sqrt{\sigma_x} h_\theta(x)^\top$ for $h_\theta : [N] \to \mathbb{R}^k$, then $\mathcal{L}_{\text{mf}}(F_\theta) = \mathcal{L}_{\text{mf}}(h_\theta)$ (overloading notation) is expressible in the following form.

$$\mathcal{L}_{\text{mf}}(h_\theta) := ||\overline{A} - F_\theta F_\theta^\top||_F^2 \tag{C.6}$$
$$= \sum_{x,y\in[N]} \left(\frac{\sigma_{xy}}{\sqrt{\sigma_x}\sqrt{\sigma_y}} - \sqrt{\sigma_x}\sqrt{\sigma_y} h_\theta(x)^\top h_\theta(y)\right)^2 \tag{C.7}$$
$$= \text{const} + \sum_{x,y\in[N]} \sigma_x\sigma_y (h_\theta(x)^\top h_\theta(y))^2 - 2\sum_{x,y\in[N]} \sigma_{xy} h_\theta(x)^\top h_\theta(y) \tag{C.8}$$
$$= \text{const} + \mathbb{E}_{x,x^-}(h_\theta(x)^\top h_\theta(x^-))^2 - 2\cdot \mathbb{E}_{x,x^+} h_\theta(x)^\top h_\theta(x^+) \tag{C.9}$$
$$= \text{const} + \mathcal{L}_{\text{cont}}(h_\theta) \tag{C.10}$$
$$\text{Where } \mathcal{L}_{\text{cont}} : \mathcal{F}_V^{(k)} \to \mathbb{R} \tag{C.11}$$
$$\mathcal{L}_{\text{cont}}(h_\theta) := \mathbb{E}_{x,x^-}(h_\theta(x)^\top h_\theta(x^-))^2 - 2\cdot \mathbb{E}_{x,x^+} h_\theta(x)^\top h_\theta(x^+) \tag{C.12}$$

Where $\mathcal{F}_V^{(k)}$ denotes the class of all functions mapping V to $\mathbb{R}^k$. Amazingly, $\mathcal{L}_{\text{cont}}(h_\theta) := \mathbb{E}_{x,x^-}(h_\theta(x)^\top h_\theta(x^-))^2 - 2\cdot \mathbb{E}_{x,x^+} h_\theta(x)^\top h_\theta(x^+)$ is one version of the (population) contrastive loss function, similar to what is used empirically.

Equation (C.8) in the derivation above naturally defines the positive pair distribution as where $(x, y) \in [N] \times [N]$ has marginal probability of σ_{xy} of being sampled. This corresponds to the sampling process of first sampling an element from the base distribution first, $z \sim \mu$, and then sampling the pair elements as independent augmentations $x \sim Q(z)$, $y \sim Q(z)$, returning (x, y). Similarly, since a negative pair $(x, y) \in [N]\times[N]$ has marginal probability of $\sigma_x\sigma_y$ in Equation (C.8), the natural negative pair distribution corresponds to first sampling $z_x, z_y \sim \mu$ independently, and then sampling $x \sim Q(z_x), y \sim Q(z_y)$, returning (x, y).

Crucially, since $\mathcal{L}_{\text{cont}}(h_\theta)$ and $\mathcal{L}_{\text{mf}}(h_\theta)$ differ by only an additive constant independent of h_θ, we can understand the global minimizers of $\mathcal{L}_{\text{cont}}$ via the global minimizers of $\mathcal{L}_{\text{mf}}$. In particular, the Eckart-Young-Mirsky Theorem characterizes the global minimizers of $\mathcal{L}_{\text{mf}}$ via the spectrum of $\overline{A}$ [EY36]. In the case $\overline{A} = U\Sigma U^\top$ is p.s.d., then the global minimizer $F_k \in \mathbb{R}^{N\times k}$ is of the form

$$F_k = U_k \Sigma_k^{1/2} R \tag{C.13}$$

Where $R \in O^{k\times k}$ is any $(k \times k)$ orthogonal matrix.

In addition to having an analytic characterization of the global minima, one other useful property of $\mathcal{L}_{\text{cont}}$ is that it is written as an expectation, and so can be approximated by an empirical risk. For a dataset $\{\overline{x}_i\}_{i\in[n]} \subset V$, define the empirical contrastive loss as:

Algorithm 5 Learning Embeddings via ERM

Require: i.i.d. sample access to base distribution $D \in \Delta(V)$, augmentation function $Q : V \to \Delta(V)$, parameters $n, k \in \mathbb{N}$, and embedding model class $\mathcal{F}_{\text{emb}}$ of models that embed into $\mathbb{R}^k$

Ensure: An embedding model $h_\theta \in \mathcal{F}_{\text{emb}}$

1: Draw i.i.d. samples $\bar{x}_1, \ldots, \bar{x}_n \sim D$

2: Define the empirical contrastive loss

$$\hat{\mathcal{L}}_{\text{cont},n}(h_{\theta'}) := -\frac{2}{n} \sum_{i \in [n]} \mathbb{E}_{\substack{x \sim Q(\cdot|\bar{x}_i) \\ x^+ \sim Q(\cdot|\bar{x}_i)}} \left[h_{\theta'}(x)^\top h_{\theta'}(x^+) \right] + \frac{1}{n(n-1)} \sum_{\substack{i,j \in [n] \\ i \neq j}} \mathbb{E}_{\substack{x \sim Q(\cdot|\bar{x}_i) \\ x^- \sim Q(\cdot|\bar{x}_j)}} \left[\left(h_{\theta'}(x)^\top h_{\theta'}(x^-) \right)^2 \right].$$

3: Compute $h_\theta \leftarrow \arg\min_{h_{\theta'} \in \mathcal{F}_{\text{emb}}} \hat{\mathcal{L}}_{\text{cont},n}(h_{\theta'})$

4: **return** h_θ

$$\hat{\mathcal{L}}_{\text{cont},n}(h_{\theta'}) := \tag{C.14}$$

$$-\frac{2}{n} \sum_{i \in [n]} \mathbb{E}_{\substack{x \sim Q(\cdot|\bar{x}_i), \\ x^+ \sim Q(\cdot|\bar{x}_i)}} (h_{\theta'}(x)^\top h_{\theta'}(x^+)) + \frac{1}{n(n-1)} \sum_{i,j \in [n], i \neq j} \mathbb{E}_{\substack{x \sim Q(\cdot|\bar{x}_i), \\ x^- \sim Q(\cdot|\bar{x}_j)}} (h_{\theta'}(x)^\top h_{\theta'}(x^-))^2 \tag{C.15}$$

Claim D.2 of [HWGM22] shows that $\mathbb{E}_{\{\bar{x}_i\}_{i \in [n]} \sim \mu} \hat{\mathcal{L}}_{\text{cont},n}(h_\theta) = \mathcal{L}_{\text{cont}}(h_\theta)$, i.e. that $\hat{\mathcal{L}}_{\text{cont},n}$ is unbiased. Define the empirical risk minimizer (ERM) as follows.

Definition C.1 (ERM, ϵ-minimizer)**.** *The empirical risk minimizer (ERM) is any embedding function of $h_\theta \in \mathcal{F}_{\text{emb}}$ that minimizes $\hat{\mathcal{L}}_{\text{cont},n}$ over $\mathcal{F}_{\text{emb}}$.*

$$h_\theta \leftarrow \arg \min_{h_{\theta'} \in \mathcal{F}_{\text{emb}}} \hat{\mathcal{L}}_{\text{cont},n}(h_{\theta'}) \tag{C.16}$$

For $\epsilon > 0$, an ϵ-minimizer of $\hat{\mathcal{L}}_{\text{cont},n}$ is any embedding function of $h_\theta \in \mathcal{F}_{\text{emb}}$ where:

$$\hat{\mathcal{L}}_{\text{cont},n}(h_\theta) \leq \min_{h_{\theta'} \in \mathcal{F}_{\text{emb}}} \hat{\mathcal{L}}_{\text{cont},n}(h_{\theta'}) + \epsilon \tag{C.17}$$

We will also call an ϵ-minimizer of $\hat{\mathcal{L}}_{\text{cont},n}$ an ϵ-ERM.

Algorithm 5 describes a procedure to use empirical risk minimization to learn an embedding model from finite samples from the base distribution.

As we will later see in Lemma C.3 and Lemma C.4, under some assumptions, the empirical risk minimizer (ERM) will be also approximately minimize the population contrastive loss. The assumptions necessary for that result are expressed in terms of Rademacher complexity, a core concept of Statistical Learning Theory. For $k \in \mathbb{N}$, we use the definition of empirical Rademacher complexity of a function class $\mathcal{F}$ of models $\mathcal{F} \ni h : V \to \mathbb{R}^k$ used in [HWGM22].

Definition C.2 (Rademacher Complexity)**.** *The Rademacher complexity of function class $\mathcal{F}$ of models $\mathcal{F} \ni h : V \to \mathbb{R}^k$ on n datapoints is:*

$$\hat{\mathcal{R}}_n(\mathcal{F}) := \max_{x_1, \ldots x_n \in V} \mathbb{E}_{\sigma \in \{\pm 1\}^n} \sup_{h \in \mathcal{F}, i \in [k]} \frac{1}{n} \left(\sum_{j \in [n]} \sigma_j h_i(x_j) \right) \tag{C.18}$$

Note this definition takes a $\max$ over the n datapoints $x_1, \ldots, x_n$ rather than an expectation, as is usual [Ma22]. Lemma C.3 analyzes the generalization error of the ERM, using Statistical Learning Theory.

C.1.2 Lemmas on Contrastive Learning

Lemma C.3. *(Theorem 4.1 of [HWGM22]) Fix a base distribution $\mu \in \Delta(V)$, augmentation function $Q : V \to \Delta(V)$, and parameters $k \in \mathbb{N}$, $k \leq |V|$ and $p_{\text{erm fails}} \in (0,1)$. Let $\mathcal{F}_{\text{emb}} \subset \mathcal{F}_V^{(k)}$ be a class of embedding models h_θ that map a domain V into dimension k Euclidean space: $\mathcal{F}_{\text{emb}} \ni h_\theta : V \to \mathbb{R}^k$. Let $\mathcal{L}_{\text{cont}} : \mathcal{F}_V^{(k)} \to \mathbb{R}$ be the population contrastive loss, instantiated with μ and Q, as in Equation (C.12), and $\hat{\mathcal{L}}_{\text{cont},n} : \mathcal{F}_V^{(k)} \to \mathbb{R}$ be the empirical contrastive loss over an i.i.d. dataset of size n drawn from μ, as in Equation (C.15). Assume:*

1. *For some $\chi > 0$ that $\forall h_\theta \in \mathcal{F}_{\text{emb}}, \forall x \in V$, $||h_\theta(x)||_\infty \leq \chi$.*
2. *$\mathcal{F}_{\text{emb}}$ contains at least one global minimizer of $\mathcal{L}_{\text{cont}}$ over functions $h : V \to \mathbb{R}^k$*

Then, with $h_{\theta_*} \in \mathcal{F}_{\text{emb}}$ any global minimizer of $\mathcal{L}_{\text{cont}}$ and $h_{\hat{\theta}} \in \mathcal{F}_{\text{emb}}$ any global minimizer of $\hat{\mathcal{L}}_{\text{cont},n}$, for which a training dataset of size n was drawn i.i.d. from μ, then with probability at least $1 - p_{\text{erm fails}}$ over the randomness of the training dataset,

$$\mathcal{L}_{\text{cont}}(h_{\hat{\theta}}) \leq \mathcal{L}_{\text{cont}}(h_{\theta_*}) + \text{poly}(k,\chi)\,\hat{\mathcal{R}}_{n/2}(\mathcal{F}_{\text{emb}}) + \text{poly}(k,\chi)\,(\sqrt{\frac{\log\frac{2}{p_{\text{erm fails}}}}{n}} + p_{\text{erm fails}}) \quad \text{(C.19)}$$

Proof. Same as Theorem 4.1 of [HWGM22]. □

[HWGM22] say, as a comment immediately after their Theorem 4.1, that $\hat{\mathcal{R}}_n(\mathcal{F})$ typically looks like $\sqrt{\frac{\mathcal{C}(\mathcal{F})}{n}}$ where $\mathcal{C}(\mathcal{F})$ is an $\mathcal{F}$-dependent constant. Assuming this is the case, then as a consequence of Lemma C.3, the minimizer of the empirical contrastive loss, with $n = \Theta(\frac{\mathcal{C}(\mathcal{F}_{\text{emb}})\,\text{poly}(k,\chi)}{\epsilon^2}\log\frac{1}{p_{\text{erm fails}}})$ i.i.d. samples, is ϵ-minimizer embedding model with probability at least $1 - p_{\text{erm fails}}$. We write this explicitly in the following Lemma.

Lemma C.4 (Generalization of ERM)**.** *Fix a base distribution $\mu \in \Delta(V)$, augmentation function $Q : V \to \Delta(V)$, and parameters $k \in \mathbb{N}$, $k \leq |V|$, and $\epsilon \in (0,1)$. Let $\mathcal{F}_{\text{emb}} \subset \mathcal{F}_V^{(k)}$ be a class of embedding models h_θ that map a domain V into dimension k Euclidean space: $\mathcal{F}_{\text{emb}} \ni h_\theta : V \to \mathbb{R}^k$. Let $\mathcal{L}_{\text{cont}} : \mathcal{F}_V^{(k)} \to \mathbb{R}$ be the population contrastive loss, instantiated with μ and Q, as in Equation (C.12), and $\hat{\mathcal{L}}_{\text{cont},n} : \mathcal{F}_V^{(k)} \to \mathbb{R}$ be the empirical contrastive loss over an i.i.d. dataset of size n drawn from μ, as in Equation (C.15). Assume:*

1. *For some $\chi > 0$ that $\forall h_\theta \in \mathcal{F}_{\text{emb}}, \forall x \in V$, $||h_\theta(x)||_\infty \leq \chi$.*
2. *$\mathcal{F}_{\text{emb}}$ contains at least one global minimizer of $\mathcal{L}_{\text{cont}}$ over functions $h : V \to \mathbb{R}^k$*
3. *$\hat{\mathcal{R}}_n(\mathcal{F}_{\text{emb}}) \leq \sqrt{\frac{\mathcal{C}(\mathcal{F}_{\text{emb}})}{n}}$, where $\mathcal{C}(\mathcal{F}_{\text{emb}})$ is an $\mathcal{F}_{\text{emb}}$-dependent constant*

Set $p_{\text{erm fails}} = \frac{\epsilon}{\text{poly}(k,\chi)}$. Then, with $h_{\theta_*} \in \mathcal{F}_{\text{emb}}$ any global minimizer of $\mathcal{L}_{\text{cont}}$ and $h_{\hat{\theta}} \in \mathcal{F}_{\text{emb}}$ any global minimizer of $\hat{\mathcal{L}}_{\text{cont},n}$, for which a training dataset of size $n = \Theta(\frac{\mathcal{C}(\mathcal{F}_{\text{emb}})\,\text{poly}(k,\chi)}{\epsilon^2}\log\frac{1}{p_{\text{erm fails}}})$ was drawn i.i.d. from μ, then with probability at least $1 - p_{\text{erm fails}}$ over the randomness of the training dataset,

$$\mathcal{L}_{\text{cont}}(h_{\hat{\theta}}) \leq \mathcal{L}_{\text{cont}}(h_{\theta_*}) + \epsilon \quad \text{(C.20)}$$

Equivalently, with $F_{\hat{\theta}} \in \mathbb{R}^{|V| \times k}$ where $\forall x \in [N], F_{\hat{\theta}}[x] = \sqrt{\sigma_x} h_{\hat{\theta}}(x)^\top$ (with σ_x defined in Equation (C.5)), then

$$\mathcal{L}_{\text{mf}}(F_{\hat{\theta}}) := ||\overline{A} - F_{\hat{\theta}} F_{\hat{\theta}}^\top||_F^2 \leq \min_{F:\ \text{rank}(F)=k} ||\overline{A} - FF^\top||_F^2 + \epsilon \quad \text{(C.21)}$$

Proof. Apply Lemma C.3, plugging in $\hat{\mathcal{R}}_{n/2} \leq \sqrt{\frac{\mathcal{C}(\mathcal{F}_{\text{emb}})}{n/2}}$, $p_{\text{erm fails}} = \frac{\epsilon}{\text{poly}(k,\chi)}$, and $n = \Theta(\frac{\mathcal{C}(\mathcal{F}_{\text{emb}})\,\text{poly}(k,\chi)}{\epsilon^2}\log\frac{1}{p_{\text{erm fails}}})$, with the constant scaled so that the bound from Lemma C.3 becomes

$$\mathcal{L}_{\text{cont}}(h_{\hat{\theta}}) \leq \mathcal{L}_{\text{cont}}(h_{\theta_*}) + \text{poly}(k,\chi)\,\hat{\mathcal{R}}_{n/2}(\mathcal{F}_{\text{emb}}) + \text{poly}(k,\chi)\,(\sqrt{\frac{\log\frac{2}{p_{\text{erm fails}}}}{n}} + p_{\text{erm fails}}) \tag{C.22}$$

$$\leq \mathcal{L}_{\text{cont}}(h_{\theta_*}) + \epsilon \tag{C.23}$$

This establishes Equation (C.20). Equation (C.21) follows from Equation (C.20) and Equation (C.10), which we showed above:

$$\forall h_\theta \in \mathcal{F}_{\text{emb}}, \mathcal{L}_{\text{mf}}(h_\theta) = c + \mathcal{L}_{\text{cont}}(h_\theta) \tag{C.24}$$

Where constant c is independent of h_θ. □

Note that as the proof of Lemma C.3 bounds the excess test loss of the ERM with uniform convergence, this result can actually be strengthened: any $\frac{\epsilon}{2}$-minimizer of the empirical contrastive loss, with $n = \Theta(\frac{\mathcal{C}(\mathcal{F}_{\text{emb}})\,\text{poly}(k,\chi)}{\epsilon^2}\log\frac{1}{p_{\text{erm fails}}})$ i.i.d. samples, is ϵ-minimizer embedding model with probability at least $1 - p_{\text{erm fails}}$.

C.1.3 The Connection Between Diffusion Embeddings and Contrastive Learning

In the following, assume the graph $G = (V, E)$ is finite ($|V| < \infty$), undirected, and connected.

Definition C.5. *(Spectral Diffusion Embedding) Let $G = (V, E)$ be a connected, undirected graph. Let (V, P) be a reversible Markov chain with transition matrix $P \in [0,1]^{|V|\times|V|}$ and stationary measure π. Define $\{\phi_i\}_{i\in[|V|]}$ as the right eigenvectors of P such that $P\phi_i = \lambda_i\phi_i$, with $1 = \lambda_1 \geq |\lambda_2| \geq \ldots \geq |\lambda_{|V|}| \geq 0$. The normalization of the eigenvectors is where $\forall i \in [|V|]$, $\text{diag}(\pi)^{1/2}\phi_i$ is a unit vector.*[2] *Denote $\phi_i(x)$ as the x-th entry of $\phi_i \in \mathbb{R}^{|V|}$. For any $t \in \mathbb{N}$, define the scale-t spectral diffusion embedding as*

$$\forall x \in V,\ \Phi_t(x) = (\lambda_i^t\phi_i(x))_{i\in[|V|]} \in \mathbb{R}^{|V|} \tag{C.25}$$

Moreover, for any $k \in [|V|]$, define the k-truncated scale-t spectral diffusion embedding as

$$\forall x \in V,\ \Phi_t^{\leq k}(x) = (\lambda_i^t\phi_i(x))_{i\in[k]} \in \mathbb{R}^k \tag{C.26}$$

Lemma C.6 establishes a connection between diffusion similarity and Euclidean inner product of the spectral diffusion embeddings.

Lemma C.6 (Diffusion Embeddings)**.** *Let $G = (V, E)$ be a connected, undirected graph. Let (V, P) be a reversible Markov chain with transition matrix $P \in [0,1]^{|V|\times|V|}$ and stationary measure π. For any $t \in \mathbb{N}$ and $x \in V$, with $\Phi_t(x) = (\lambda_i^t\phi_i(x))_{i\in[|V|]} \in \mathbb{R}^{|V|}$ denoting the spectral diffusion embedding of x as in Definition C.5, we have*

$$\forall x, y \in V, \langle\Phi_t(x), \Phi_t(y)\rangle = \langle p^t(x), p^t(y)\rangle_{\ell^2(1/\pi)} \tag{C.27}$$

where $\langle\cdot,\cdot\rangle$ is the usual Euclidean inner product and $\langle\cdot,\cdot\rangle_{\ell^2(1/\pi)}$ is the inner product where $\forall i \in [|V|]$, the ith term is weighted by $\frac{1}{\pi(i)}$.

[2] Alternatively, note that since P is reversible, $\text{diag}(\pi)P$ is symmetric. Let symmetric matrix $\overline{A} = \text{diag}(\pi)^{1/2}P\text{diag}(\pi)^{-1/2} = \sum_i \lambda_i v_i v_i^\top$ with $\{v_i\}_{i\in[|V|]}$ being orthonormal. Then $\forall i \in [|V|]$, $\phi_i = \text{diag}(\pi)^{-1/2}v_i$.

Proof. Because (V, P) is reversible, let $P = D^{-1}W \in \mathbb{R}^{|V|\times|V|}$ where $D = \text{diag}(W\mathbf{1}) = \text{diag}(\pi)$ and $W = \text{diag}(\pi)P \in \mathbb{R}^{|V|\times|V|}$ is symmetric. The symmetric matrix $\overline{A} = D^{-1/2}WD^{-1/2}$ is the normalized adjacency matrix, with eigendecomposition $\overline{A} = \sum_{i\in[|V|]} \lambda_i u_i u_i^\top$. Since $P = D^{-1/2}\overline{A}D^{1/2}$, then $P = \sum_{i\in[|V|]} \lambda_i (D^{-1/2}u_i)(D^{1/2}u_i)^\top$, so that $\forall i \in [|V|], \phi_i = D^{-1/2}u_i$. Now, for any $t \in \mathbb{N}$ and $x \in V$,

$$
\begin{aligned}
p^t(x) &:= (P^\top)^t e_x && \text{(C.28)}\\
&= D^{1/2}\overline{A}^t D^{-1/2} e_x && \text{(C.29)}\\
&= D^{1/2} \sum_{i\in[|V|]} \lambda_i^t u_i \phi_i(x) && \text{(C.30)}\\
&= D^{1/2} U \Phi_t(x) && \text{(C.31)}
\end{aligned}
$$

Where U is an orthogonal transformation $U = [u_1, \ldots, u_{|V|}] \in \mathbb{R}^{|V|\times|V|}$. It follows that for any $x, y \in V$,

$$
\begin{aligned}
\langle p^t(x), p^t(y)\rangle_{\ell^2(1/\pi)} &:= (p^t(x))^\top D^{-1} p^t(y) && \text{(C.32)}\\
&= \Phi_t(x)^\top U^\top U \Phi_t(y) && \text{(C.33)}\\
&= \Phi_t(x)^\top \Phi_t(y) && \text{(C.34)}
\end{aligned}
$$

□

As a corollary, for any $u, v \in V$, the diffusion distance $||p^t(u) - p^t(v)||_{\ell^2(1/\pi)} = ||\Phi_t(u) - \Phi_t(v)||_2$. For more about diffusion distance, see [CL06] and [Yao11].

The following Lemma describes the approximation properties of a diffusion embedding with small embedding dimension $k \in [|V| - 1]$ (rather than the full $|V|$-dimensional embedding), when λ_{k+1}^{2t} is small.

Lemma C.7. *(Small Embedding Dimension Approximation; Lemma 2 of [Yao11]) Let $G = (V, E)$ be a connected, undirected graph. Let (V, P) be a reversible Markov chain with transition matrix $P \in [0, 1]^{|V|\times|V|}$ and stationary measure π. Let $\pi_{\min} := \min_{v\in V} \pi(v)$ and $1 = \lambda_1 \geq |\lambda_2| \geq \ldots, \geq |\lambda_{|V|}| \geq 0$ be arranged in decreasing magnitude. Given $t \in \mathbb{N}$ and $k \in [|V| - 1]$, for any $u, v \in V$,*

$$
\begin{aligned}
||\Phi_t(u) - \Phi_t(v)||_2^2 - \frac{2\lambda_{k+1}^{2t}}{\pi_{\min}} \cdot 1[u \neq v] \leq ||\Phi_t^{\leq k}(u) - \Phi_t^{\leq k}(v)||_2^2 \leq ||\Phi_t(u) - \Phi_t(v)||_2^2 && \text{(C.35)}\\
||\Phi_t(v)||_2^2 - \frac{\lambda_{k+1}^{2t}}{\pi_{\min}} \leq ||\Phi_t^{\leq k}(v)||_2^2 \leq ||\Phi_t(v)||_2^2 && \text{(C.36)}\\
\langle \Phi_t(u), \Phi_t(v)\rangle - \frac{\lambda_{k+1}^{2t}}{\pi_{\min}} \leq \langle \Phi_t^{\leq k}(u), \Phi_t^{\leq k}(v)\rangle \leq \langle \Phi_t(u), \Phi_t(v)\rangle + \frac{\lambda_{k+1}^{2t}}{\pi_{\min}} && \text{(C.37)}
\end{aligned}
$$

Proof. As in the proof of Lemma C.8, with $D = \text{diag}(\pi)$ and matrix $D^{1/2}PD^{-1/2}$ having eigendecomposition $U\Lambda U^\top$, then $P = (D^{-1/2}U)\Lambda(D^{1/2}U)^\top$.

The matrix of right eigenvectors $\Phi := [\phi_1, \ldots, \phi_n] = D^{-\frac{1}{2}}U$. Thus,

$$
\begin{aligned}
\Phi\Phi^\top &= D^{-\frac{1}{2}} U U^\top D^{-\frac{1}{2}} && \text{(C.38)}\\
&= D^{-1} && \text{(C.39)}
\end{aligned}
$$

Denoting $n := |V|$, for any $u, v \in V$,

$$\sum_{i=1}^{n} \phi_i(u)\phi_i(v) = \frac{1[u=v]}{\pi(v)} \tag{C.40}$$

$$\sum_{i=1}^{n} \big(\phi_i(u) - \phi_i(v)\big)^2 = \begin{cases} 0, & u = v, \\ \frac{1}{\pi(u)} + \frac{1}{\pi(v)}, & u \neq v, \end{cases} \leq \frac{2}{\pi_{\min}} \mathbf{1}_{\{u \neq v\}}. \tag{C.41}$$

Now

$$||\Phi_t^{\leq k}(u) - \Phi_t^{\leq k}(v)||_2^2 = ||\Phi_t(u) - \Phi_t(v)||_2^2 - \sum_{i \geq k+1} \lambda_i^{2t} \big(\phi_i(u) - \phi_i(v)\big)^2 \tag{C.42}$$

$$\geq ||\Phi_t(u) - \Phi_t(v)||_2^2 - \lambda_{k+1}^{2t} \sum_{i=1}^{n} \big(\phi_i(u) - \phi_i(v)\big)^2 \tag{C.43}$$

$$\geq ||\Phi_t(u) - \Phi_t(v)||_2^2 - \frac{2\lambda_{k+1}^{2t}}{\pi_{\min}} \mathbf{1}_{\{u \neq v\}}. \tag{C.44}$$

The upper bound $||\Phi_t^{\leq k}(u) - \Phi_t^{\leq k}(v)||_2^2 \leq ||\Phi_t(u) - \Phi_t(v)||_2^2$ is immediate, since truncation only removes nonnegative terms. This proves Equation (C.35).

Similarly,

$$||\Phi_t^{\leq k}(v)||_2^2 = ||\Phi_t(v)||_2^2 - \sum_{i \geq k+1} \lambda_i^{2t} \phi_i(v)^2 \tag{C.45}$$

$$\geq ||\Phi_t(v)||_2^2 - \lambda_{k+1}^{2t} \sum_{i=1}^{n} \phi_i(v)^2 \tag{C.46}$$

$$\geq ||\Phi_t(v)||_2^2 - \frac{\lambda_{k+1}^{2t}}{\pi_{\min}}. \tag{C.47}$$

Again, since truncation only removes nonnegative terms, $||\Phi_t^{\leq k}(v)||_2^2 \leq ||\Phi_t(v)||_2^2$. This proves Equation (C.36).

Finally,

$$\langle \Phi_t^{\leq k}(u), \Phi_t^{\leq k}(v) \rangle = \langle \Phi_t(u), \Phi_t(v) \rangle - \sum_{i \geq k+1} \lambda_i^{2t} \phi_i(u)\phi_i(v) \tag{C.48}$$

$$\geq \langle \Phi_t(u), \Phi_t(v) \rangle - \sqrt{\Big(\sum_{i \geq k+1} \lambda_i^{2t} \phi_i(u)^2 \Big) \Big(\sum_{i \geq k+1} \lambda_i^{2t} \phi_i(v)^2 \Big)}$$

(Cauchy Schwartz)

$$\geq \langle \Phi_t(u), \Phi_t(v) \rangle - \lambda_{k+1}^{2t} \sqrt{\Big(\sum_{i=1}^{n} \phi_i(u)^2 \Big) \Big(\sum_{i=1}^{n} \phi_i(v)^2 \Big)} \tag{C.49}$$

$$\geq \langle \Phi_t(u), \Phi_t(v) \rangle - \frac{\lambda_{k+1}^{2t}}{\pi_{\min}} \tag{C.50}$$

$$\text{Similarly, } \langle \Phi_t^{\leq k}(u), \Phi_t^{\leq k}(v) \rangle \leq \langle \Phi_t(u), \Phi_t(v) \rangle + \frac{\lambda_{k+1}^{2t}}{\pi_{\min}} \tag{C.51}$$

□

We will now describe how to learn the embeddings $\{\Phi_t^{\leq k}(u)\}_{u \in V}$ for general $k \in [|V|]$ and $t \in \mathbb{N}$. Lemma C.8 shows the global minimizers of the contrastive loss recover $\{\Phi_t^{\leq k}(u)\}_{u \in V}$ up to an orthogonal transformation.

Lemma C.8 (Global Minimizers for Contrastive Loss)**.** *Let $G = (V, E)$ be any connected, undirected graph. Let (V, P) be a reversible Markov chain with transition matrix $P \in [0,1]^{|V|\times|V|}$ and stationary measure $\pi \in \Delta(V)$. Let $\{\lambda_i(P)\}_{i\in[|V|]}$ be the eigenvalues of P, ordered by decreasing absolute value. Fix any $t \in \mathbb{N}$ and $k \in [|V|]$. Suppose $|\lambda_k(P)| - |\lambda_{k+1}(P)| > 0$ if $k < |V|$ and suppose the population contrastive loss (Equation (C.12)) with base distribution π, augmentation function P^t, and embedding dimension k has a global minimizer in $\mathcal{F}_{\text{emb}}$. Then, for any such global minimizer $\mathcal{F}_{\text{emb}} \ni h_{\theta_*} : V \to \mathbb{R}^k$, there exists an orthogonal matrix $R_* \in O^{k\times k}$ where*

$$\forall u \in V, h_{\theta_*}(u) = R_* \, \Phi_t^{\le k}(u) \in \mathbb{R}^k \tag{C.52}$$

With mapping $C : V \to \Delta(V)$ defined as the sampler-representation of P, where $\forall x \in V, C(x) = P[x]$, then the positive and negative pair distribution corresponding to base distribution π and augmentation function P^t are where positive pairs (x, y) are generated by sampling $z \sim \pi$, then $x \sim C^t(z), y \sim C^t(z)$ independently. Negative pairs (x, y) are generated via $x, y \sim \pi$ independently. Note that σ_x (defined in Equation (C.5)) equals $\pi(x)$ for every $x \in V$.

Proof. Fix $t \in \mathbb{N}$ and $k \in [|V|]$. The plan is to follow the derivation process of the contrastive loss $\mathcal{L}_{\text{cont}}$, but with $\mu = \pi$ and $Q = P^t$. First, we have

$$\begin{aligned}
A &= Q^\top \text{diag}(\mu) Q && \text{(C.53)}\\
&= (P^t)^\top \text{diag}(\pi) P^t && \text{(C.54)}\\
&= \text{diag}(\pi) P^{2t} && \text{(By reversibility of } P\text{)}\\
D &:= \text{diag}(A\mathbf{1}) && \text{(C.55)}\\
&= \text{diag}(\pi) && \text{(Row sums of } P^{2t} \text{ are 1)}\\
\overline{A} &:= D^{-1/2} A D^{-1/2} && \text{(C.56)}\\
&= \text{diag}(\pi)^{1/2} P^{2t} \text{diag}(\pi)^{-1/2} && \text{(C.57)}
\end{aligned}$$

Let the eigendecomposition of $\overline{A}$ be $\overline{A} = \sum_{i\in[|V|]} \lambda_i(\overline{A})\, u_i u_i^\top = U\Lambda U^\top$. By the Eckart-Young-Mirsky Theorem [EY36], since $\overline{A}$ is p.s.d. and $\lambda_k(\overline{A}) - \lambda_{k+1}(\overline{A}) = \lambda_k(P)^{2t} - \lambda_{k+1}(P)^{2t} > 0$ if $k < |V|$, the rank-k minimizers of the matrix factorization loss are:

$$\{U\Lambda^{1/2} R : R \in O^{k\times k}\} = \arg\min_{F\in\mathbb{R}^{|V|\times k}} ||\overline{A} - FF^\top||_F^2 \tag{C.58}$$

Because for any $h_\theta \in \mathcal{F}_{\text{emb}}$, $\mathcal{L}_{\text{mf}}(h_\theta) = \mathcal{L}_{\text{cont}}(h_\theta) + c$ for some constant c independent of h_θ (Equation (C.10)), then if $\mathcal{F}_{\text{emb}}$ contains some $h_{\theta_*} : V \to \mathbb{R}^k$ which is a global minimizer of $\mathcal{L}_{\text{cont}}$, then the induced matrix $F_{\theta_*} \in \mathbb{R}^{|V|\times k}$, where the x-th row is:

$$\forall x \in V, F_{\theta_*}[x] = \sqrt{\pi(x)} h_{\theta_*}(x)^\top \tag{C.59}$$

Must be in the set $\arg\min_{F\in\mathbb{R}^{|V|\times k}} ||\overline{A} - FF^\top||_F^2 = \{U\Lambda^{1/2}R : R \in O^{k\times k}\}$. In particular, there exists $R \in O^{k\times k}$ where

$$F_{\theta_*} = U\Lambda^{1/2} R \tag{C.60}$$

The x-th row of $U\Lambda^{1/2}R$ is $(\lambda_i(\overline{A})^{1/2} u_i(x))_{i\in[k]}^\top R \in \mathbb{R}^{1\times k}$, where $u_i(x)$ is the x-th entry of u_i. Thus,

$$\begin{aligned}
\forall x \in V,\ h_{\theta_*}(x) &= (\pi(x))^{-1/2} R \, (\lambda_i(\overline{A})^{1/2} u_i(x))_{i\in[k]} && \text{(C.61)}\\
&= R \left(\lambda_i(\overline{A})^{1/2} (\pi(x))^{-1/2} u_i(x)\right)_{i\in[k]} && \text{(C.62)}
\end{aligned}$$

Finally, we note that since $\{(\lambda_i(\overline{A}), u_i)\}_{i\in[k]}$ are the top k eigenvalue and eigenvectors of $\overline{A} = \text{diag}(\pi)^{1/2}P^{2t}\text{diag}(\pi)^{-1/2}$, then the eigendecomposition of P^{2t} is:

$$P^{2t} = \sum_{i\in[|V|]} \lambda_i(\overline{A})\big(\text{diag}(\pi)^{-1/2}u_i\big)\big(\text{diag}(\pi)^{1/2}u_i\big)^\top \tag{C.63}$$

$$= \sum_{i\in[|V|]} (\lambda_i(P))^{2t}\big(\text{diag}(\pi)^{-1/2}u_i\big)\big(\text{diag}(\pi)^{1/2}u_i\big)^\top \tag{C.64}$$

Where $\{\lambda_i(P)\}_{i\in[|V|]}$ are the eigenvalues of P arranged in decreasing order of absolute value, so that $\lambda_i(\overline{A})^{1/2} = |\lambda_i(P)|^t$. Finally, we see that the right eigenvectors of P^{2t} are exactly $\{\text{diag}(\pi)^{-1/2}u_i\}_{i\in[|V|]}$. Thus, Equation (C.62) can be written in terms of the k truncated spectral diffusion embedding in Definition C.5.

$$\forall x \in V,\ h_{\theta_*}(x) = R\left(\lambda_i(\overline{A})^{1/2}(\pi(x))^{-1/2}u_i(x)\right)_{i\in[k]} \tag{C.65}$$

$$= R\,\text{diag}(\text{sign}(\{\lambda_i(P)^t\}_{i\in[k]}))\,\Phi_t^{\le k}(x) \tag{C.66}$$

$$= R_*\,\Phi_t^{\le k}(x) \tag{C.67}$$

Where $\text{diag}(\text{sign}(\{\lambda_i(P)^t\}_{i\in[k]})) \in O^{k\times k}$ is a diagonal matrix where the (i,i)-th entry is 1 if $\lambda_i(P)^t$ is non-negative, and -1 if $\lambda_i(P)^t$ is negative, for $i \in [k]$. Since R, $\text{diag}(\text{sign}(\{\lambda_i(P)^t\}_{i\in[k]})) \in O^{k\times k}$, then $R_* := R\,\text{diag}(\text{sign}(\{\lambda_i(P)^t\}_{i\in[k]})) \in O^{k\times k}$.

□

Corollary C.9. *Let $G = (V, E)$ be any connected, undirected graph. Let (V, P) be a reversible Markov chain with transition matrix $P \in [0,1]^{|V|\times|V|}$ and stationary measure π. Fix any $t \in \mathbb{N}$ and $k \in [|V|]$. Let $\{\lambda_i(P)\}_{i\in[|V|]}$ be the eigenvalues of P, ordered by decreasing absolute value. Suppose $|\lambda_k(P)| - |\lambda_{k+1}(P)| > 0$ if $k < |V|$. Denote $\mathcal{F}_{\text{emb}}^{(k)} \subset \mathcal{F}_V^{(k)}$ as a class of embedding models from V to $\mathbb{R}^k$ and $\mathcal{F}_{\text{emb}}^{(|V|)} \subset \mathcal{F}_V^{(|V|)}$ a class of embedding models from V to $\mathbb{R}^{|V|}$. Suppose the population contrastive loss (Equation (C.12)) with base distribution π, augmentation function P^t, and embedding dimension k (dimension $|V|$) has a global minimizer in $\mathcal{F}_{\text{emb}}^{(k)}$ ($\mathcal{F}_{\text{emb}}^{(|V|)}$).*

Let θ_* be any global minimizer of the population contrastive loss with embedding dimension $|V|$, and $\theta_*^{(k)}$ be any global minimizer of the population contrastive loss with embedding dimension with embedding dimension k. There exist orthogonal matrices $R \in O^{|V|\times|V|}$ and $R^{(k)} \in O^{k\times k}$ such that $\forall u \in V$,

$$h_{\theta_*^{(k)}}(u) = R^{(k)}\,\Phi_t^{\le k}(u) \in \mathbb{R}^k \tag{C.68}$$

$$h_{\theta_*}(u) = R\,\Phi_t(u) \in \mathbb{R}^{|V|} \tag{C.69}$$

In particular, $\forall u, v \in V$, with $\pi_{\min} := \min_{v\in V}\pi(v)$,

$$||h_{\theta_*}(u) - h_{\theta_*}(v)||^2 - \frac{2(\lambda_{k+1}(P))^{2t}}{\pi_{\min}} = ||p^t(u) - p^t(v)||^2_{\ell^2(1/\pi)} - \frac{2(\lambda_{k+1}(P))^{2t}}{\pi_{\min}} \tag{C.70}$$

$$\le ||h_{\theta_*^{(k)}}(u) - h_{\theta_*^{(k)}}(v)||_2^2 = ||\Phi_t^{\le k}(u) - \Phi_t^{\le k}(v)||_2^2 \tag{C.71}$$

$$\le ||p^t(u) - p^t(v)||^2_{\ell^2(1/\pi)} = ||h_{\theta_*}(u) - h_{\theta_*}(v)||^2 \tag{C.72}$$

$$||h_{\theta_*}(v)||_2^2 - \frac{(\lambda_{k+1}(P))^{2t}}{\pi_{\min}} \le ||h_{\theta_*^{k_*}}(v)||_2^2 \le ||h_{\theta_*}(v)||_2^2 \tag{C.73}$$

Proof. We apply Lemma C.8 with graph G and with embedding dimension k (resp. embedding dimension $|V|$) to get Equation (C.68) (resp. Equation (C.69)). Equation (C.72) is obtained by combining Equation (C.68), Equation (C.69), Lemma C.6, and Lemma C.7. Equation (C.73) is obtained by combining Lemma C.7 with Equation (C.68) and Equation (C.69). □

C.1.4 Matrix Factorization Lemmas

Lemma C.6 and Lemma C.8 motivate the idea of using embeddings learned via contrastive learning to compute the diffusion similarity and distance. Lemma C.4 shows that finding the ERM (or $\frac{\epsilon}{2}$-ERM) suffices, with high probability, to find an ϵ-minimizer of the population contrastive loss $\mathcal{L}_{\text{cont}}$. Lemma C.13, a Lemma about matrix-factorization, is useful in quantifying how close an embedding corresponding to such an ϵ-minimizer is to a ground-truth embedding, up to a rotation.

To start, Lemma C.12, Lemma C.11, and Lemma C.10 are introduced here to prove Lemma C.13. Note that for two matrices $A, B \in \mathbb{R}^{n\times n}$, $\langle A, B\rangle := \text{tr}(A^\top B)$ denotes the Hilbert-Schmidt inner product of the two matrices, with $||A||_F^2 = \langle A, A\rangle$.

Lemma C.10. *(Lemma 5.4 of [TBS+16]) For any natural numbers $n \geq k$ and matrices $U, X \in \mathbb{R}^{n\times k}$,*

$$\text{dist}(U, X)^2 := \min_{R\in O^{k\times k}} ||U - XR||_F^2 \leq \frac{1}{2(\sqrt{2}-1)\sigma_k(X)^2}||UU^\top - XX^\top||_F^2. \tag{C.74}$$

Lemma C.11. *(Ky Fan Maximum Principle [Fan50]; Theorem 1 of [OW91]) Let A be a real, symmetric $n \times n$ matrix, with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$. For any $k \in [n]$,*

$$\max_{X^\top X = I_k} \text{tr}(X^\top AX) = \sum_{i=1}^{k} \lambda_i \tag{C.75}$$

Lemma C.12. *(Lemma 3.1 of [VCLR13]) Let $A \in R^{n\times n}$ be a symmetric matrix, with eigenvalues ordered as $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$. For $k \in [n-1]$, let P_k be the projector onto the subspace spanned by the top-k eigenvectors of A, denoted $\lambda_1, \ldots, \lambda_k$. Assume that $\lambda_k^2 - \lambda_{k+1}^2 > 0$. For any symmetric rank-k projector $P \in \mathbb{R}^{n\times n}$, $P^\top = P$, $P^2 = P$, $\text{tr}(P) = k$, we have:*

$$\langle A^2, P_k - P\rangle \geq \frac{\lambda_k^2 - \lambda_{k+1}^2}{2}||P_k - P||_F^2. \tag{C.76}$$

Proof. Diagonalize $A^2 = U\text{diag}(\{\lambda_i^2\}_{i\in[n]})U^\top$ and write $\widetilde{P} := U^\top PU$. Partition $\widetilde{P}$ as

$$\widetilde{P} = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} \quad \text{with } P_{11} \in \mathbb{R}^{k\times k}. \tag{C.77}$$

Then $\widetilde{P}_k = U^\top P_k U = \text{diag}(I_k, 0)$ in the eigenbasis, and

$$\begin{aligned} \langle A^2, P_k - P\rangle &= \text{tr}\big(\text{diag}(\{\lambda_i^2\}_{i\in[n]})(\widetilde{P}_k - \widetilde{P})\big) && \text{(C.78)} \\ &= \text{tr}\big(\Lambda_k^2(I - P_{11})\big) - \text{tr}\big(\Lambda_\perp^2 P_{22}\big) && \text{(C.79)} \end{aligned}$$

where $\Lambda_\perp := \text{diag}(\{\lambda_{k+1}, \ldots, \lambda_n\})$. Using $\Lambda_k^2 \succeq \lambda_k^2 I_k$ and $\Lambda_\perp^2 \preceq \lambda_{k+1}^2 I$ gives

$$\langle A^2, P_k - P\rangle \geq \lambda_k^2 \text{tr}(I - P_{11}) - \lambda_{k+1}^2 \text{tr}(P_{22}). \tag{C.80}$$

Since $\widetilde{P}$ is a rank-k projector, $\text{tr}(\widetilde{P}) = k$, so $\text{tr}(P_{11}) + \text{tr}(P_{22}) = k$ and thus $\text{tr}(I - P_{11}) = \text{tr}(P_{22})$. Also, for rank-$k$ projectors,

$$\begin{aligned} ||P_k - P||_F^2 &= \text{tr}\big((P_k - P)^2\big) && \text{(C.81)} \\ &= \text{tr}(P_k) + \text{tr}(P) - 2\text{tr}(P_k P) && \text{(C.82)} \\ &= 2k - 2\text{tr}(P_{11}) && \text{(C.83)} \\ &= 2\text{tr}(I - P_{11}). && \text{(C.84)} \end{aligned}$$

Therefore $\text{tr}(I - P_{11}) = \frac{1}{2}||P_k - P||_F^2$, and we obtain Equation (C.76). □

Lemma C.13 (Approximate minimizers are close to a global minimizer)**.** *Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive semidefinite with eigendecomposition $A = U\text{diag}(\{\lambda_i\}_{i\in[n]})U^\top$ where $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$. Fix $1 \leq k < n$ and assume $\lambda_k > 0$ and the squared eigengap $\lambda_k^2 - \lambda_{k+1}^2 > 0$. Define, for matrix $F \in \mathbb{R}^{n\times k}$ of rank k,*

$$\mathcal{L}(F) := ||A - FF^\top||_F^2 \tag{C.85}$$

Suppose $F \in \mathbb{R}^{n\times k}$ satisfies $\mathcal{L}(F) \leq \min_{\text{rank}(Z)=k} \mathcal{L}(Z) + \epsilon$ for some $0 < \epsilon \leq 1$. Define the constant $\mathcal{C}(A,k) := \frac{\left(1+4||A||_2\sqrt{\frac{2}{\lambda_k^2-\lambda_{k+1}^2}}\right)^2}{2(\sqrt{2}-1)\lambda_k}$. Then, there exists a global minimizer $F^* \in \arg\min_{\text{rank}(Z)=k} \mathcal{L}(Z)$ such that:

$$||F - F^*||_F^2 \leq \mathcal{C}(A,k)\,\epsilon \tag{C.86}$$

Proof. First, we recall the Eckart-Young-Mirsky Theorem [EY36], which characterizes the global minima of the loss function $\mathcal{L}(F)$. Let $U_k \in \mathbb{R}^{n\times k}$ be the first k eigenvectors and let $\Lambda_k := \text{diag}(\lambda_1, \ldots, \lambda_k)$. Since $A \succeq 0$ and $\lambda_k^2 - \lambda_{k+1}^2 > 0$, the best rank-$k$ approximation in Frobenius norm is $A_k := U_k\Lambda_k U_k^\top$, which is unique.

$$\min_{\text{rank}(Z)=k} \mathcal{L}(Z) = ||A - A_k||_F^2 = \sum_{i>k} \lambda_i^2 = ||A||_F^2 - ||\Lambda_k||_F^2. \tag{C.87}$$

Because of this, the set of global minimizers of $\mathcal{L}(\cdot)$ over rank-k, $(n \times k)$ matrices are $S := \{U_k\Lambda_k^{1/2}R : R \in O^{k\times k}\}$. With this, the proof plan is to upper bound $||FF^\top - A_k||_F$, and then apply Lemma C.10 to get the final result. To bound, $||FF^\top - A_k||_F$, define P as the rank-k projector onto $\text{col}(F)$.

$$P := F(F^\top F)^{-1}F^\top \in \mathbb{R}^{n\times n} \tag{C.88}$$

$$Q := F(F^\top F)^{-1/2} \in \mathbb{R}^{n\times k} \tag{C.89}$$

Where $P = QQ^\top$. We will use the following decomposition, as a result of the triangle inequality.

$$||FF^\top - A_k||_F \leq ||FF^\top - PAP||_F + ||PAP - A_k||_F \tag{C.90}$$

The remaining proof will be dedicated to bounding the terms $||FF^\top - PAP||_F$ and $||PAP - A_k||_F$ individually.

Bounding $||FF^\top - PAP||_F$.

First, we will show that $||FF^\top - PAP||_F \leq \sqrt{\epsilon}$. Towards this, we derive a useful identity for $\mathcal{L}(F)$, culminating in Equation (C.96). A direct expansion of $\mathcal{L}(F)$ gives the identity

$$\mathcal{L}(F) = ||(FF^\top - PAP) + (PAP - A)||_F^2 \tag{C.91}$$

$$= ||FF^\top - PAP||_F^2 + ||PAP - A||_F^2 + 2\langle FF^\top - PAP, PAP - A\rangle \tag{C.92}$$

$$= ||FF^\top - PAP||_F^2 + ||PAP - A||_F^2 \tag{C.93}$$

Where the last equality is because $\langle FF^\top - PAP, PAP - A\rangle = \text{tr}(FF^\top PAP - FF^\top A) + \text{tr}(-(PAP)^2 + PAPA) = 0$, as $\text{tr}(FF^\top PAP) = \text{tr}(FF^\top A)$ and $\text{tr}((PAP)^2) = \text{tr}(PAPA)$. Subtracting $\min_{\text{rank}(Z)=k} \mathcal{L}(Z) = ||A||_F^2 - ||\Lambda_k||_F^2$ yields

$$\mathcal{L}(F) - \min_{\text{rank}(Z)=k} \mathcal{L}(Z) = ||FF^\top - PAP||_F^2 + ||PAP - A||_F^2 - (||A||_F^2 - ||\Lambda_k||_F^2) \tag{C.94}$$

$$= ||FF^\top - PAP||_F^2 + (||\Lambda_k||_F^2 + ||PAP||_F^2 - 2\text{tr}(PAPA)) \tag{C.95}$$

$$= ||FF^\top - PAP||_F^2 + (||\Lambda_k||_F^2 - ||Q^\top AQ||_F^2) \tag{C.96}$$

Where in Equation (C.96), we used that $||Q^\top AQ||_F^2 = ||PAP||_F^2 = \text{tr}(PAPA)$.

Because $\mathcal{L}(F) - \min_{\text{rank}(Z)=k} \mathcal{L}(Z) \leq \epsilon$ by assumption, then by Equation (C.96), to show that $||FF^\top - PAP||_F^2 \leq \epsilon$, it suffices to show that $||\Lambda_k||_F^2 - ||Q^\top AQ||_F^2 \geq 0$. Towards this, let $Q_\perp$ complete Q to an orthonormal basis, so that $I = QQ^\top + Q_\perp Q_\perp^\top$. Then

$$Q^\top A^2 Q = Q^\top A(I)AQ \tag{C.97}$$

$$= Q^\top AQQ^\top AQ + Q^\top AQ_\perp Q_\perp^\top AQ \tag{C.98}$$

$$\text{tr}(Q^\top A^2 Q) = ||Q^\top AQ||_F^2 + ||Q_\perp^\top AQ||_F^2 \geq ||Q^\top AQ||_F^2 \tag{C.99}$$

On the other hand, since $P = QQ^\top$ has rank k with $Q^\top Q = I_k$, by Lemma C.11,

$$\text{tr}(A^2 P) = \text{tr}(Q^\top A^2 Q) \leq \sum_{i=1}^k \lambda_i(A^2) = \sum_{i=1}^k \lambda_i^2 = ||\Lambda_k||_F^2 \tag{C.100}$$

Combining Equation (C.100) with Equation (C.99) gives $||Q^\top AQ||_F^2 \leq ||\Lambda_k||_F^2$. By Equation (C.96), and the fact that $\mathcal{L}(F) - \min_{\text{rank}(Z)=k} \mathcal{L}(Z) \leq \epsilon$, we conclude that

$$||FF^\top - PAP||_F^2 \leq \epsilon \tag{C.101}$$

which upper bounds the first term in our decomposition, Equation (C.90).

Bounding $||PAP - A_k||_F$.

With $P_k := U_k U_k^\top$ denoting the projector onto the top-k eigenspace of A, we will first argue that projectors P and P_k are close in Frobenius norm. We have

$$\text{tr}(A^2 P_k) - \text{tr}(A^2 P) = ||\Lambda_k||_F^2 - \text{tr}(A^2 P) \tag{C.102}$$

$$\leq ||\Lambda_k||_F^2 - ||Q^\top AQ||_F^2 \qquad \text{(By Equation (C.99) and } P = QQ^\top\text{)}$$

$$\leq \epsilon \tag{C.103}$$

Where the last inequality is by Equation (C.96), the non-negativity of $||FF^\top - PAP||_F^2$, and the fact that $\mathcal{L}(F) - \min_{\text{rank}(Z)=k} \mathcal{L}(Z) \leq \epsilon$.

The squared eigengap $\lambda_k^2 - \lambda_{k+1}^2 > 0$ is positive by assumption. Since $\langle A^2, P_k - P\rangle = \text{tr}(A^2 P_k) - \text{tr}(A^2 P) \leq \epsilon$, Lemma C.12 (Equation (C.76)) implies that

$$||P_k - P||_F^2 \leq \frac{2\epsilon}{\lambda_k^2 - \lambda_{k+1}^2} \tag{C.104}$$

Write $\Delta := P - P_k$. Note that $A_k = P_k A P_k$. We have

$$PAP - P_k A P_k = P_k A\Delta + \Delta A P_k + \Delta A \Delta \tag{C.105}$$

$$||PAP - P_k A P_k||_F \leq 2||A||_2 \, ||P_k||_2 \, ||\Delta||_F + ||A||_2 \, ||\Delta||_2 \, ||\Delta||_F \tag{C.106}$$

$$\leq 4||A||_2 ||\Delta||_F \qquad \text{(Since } ||P_k||_2 = 1 \text{ and } ||\Delta||_2 \leq ||P||_2 + ||P_k||_2 = 2\text{)}$$

With $||\Delta||_F = ||P - P_k||_F \leq \sqrt{\frac{2\epsilon}{\lambda_k^2 - \lambda_{k+1}^2}}$ from (C.104), we have

$$||PAP - A_k||_F = ||PAP - P_k A P_k||_F \leq 4||A||_2 \sqrt{\frac{2\epsilon}{\lambda_k^2 - \lambda_{k+1}^2}} \tag{C.107}$$

Combining the Bounds.

Plugging in Equation (C.101) and Equation (C.107) into Equation (C.90),

$$||FF^\top - A_k||_F \leq ||FF^\top - PAP||_F + ||PAP - A_k||_F \tag{C.108}$$

$$\leq \sqrt{\epsilon} + 4||A||_2\sqrt{\frac{2\epsilon}{\lambda_k^2 - \lambda_{k+1}^2}} \tag{C.109}$$

$$= \left(1 + 4||A||_2\sqrt{\frac{2}{\lambda_k^2 - \lambda_{k+1}^2}}\right)\sqrt{\epsilon} \tag{C.110}$$

Finally, noting that any global minimizer $F_0 \in \arg\min_{\text{rank}(Z)=k} \mathcal{L}(Z)$ satisfies that $F_0F_0^\top = A_k$, we can view $||FF^\top - A_k||_F$ as the difference between the gram matrices of F and a global minimizer F_0. Lemma C.10 enables us to convert an upper bound on the Gram error $||FF^\top - A_k||_F$ to factor error, between F and the closest global minimizer to it in Frobenius norm (specified by a $(k \times k)$ orthogonal matrix). Let $F_0 := U_k\Lambda_k^{1/2} \in S$ so that $F_0F_0^\top = A_k$ and $(\sigma_k(F_0))^2 = \lambda_k$. Apply Lemma C.10 with $U = F$ and $X = F_0$, and then combine the result with $(\sigma_k(F_0))^2 = \lambda_k$ and Equation (C.110) to obtain:

$$\text{dist}(F, F_0)^2 := \min_{R \in O^{k\times k}} ||F - F_0R||_F^2 \tag{C.111}$$

$$\leq \frac{1}{2(\sqrt{2}-1)\lambda_k}||FF^\top - A_k||_F^2 \tag{C.112}$$

$$\leq \frac{\left(1 + 4||A||_2\sqrt{\frac{2}{\lambda_k^2 - \lambda_{k+1}^2}}\right)^2}{2(\sqrt{2}-1)\lambda_k}\epsilon. \tag{C.113}$$

Finally, choose $R^* \in O^{k\times k}$ that attains the minimum in $\text{dist}(F, F_0)$ and set $F^* := F_0R^* \in S$. Then $||F - F^*||_F^2 = \text{dist}(F, F_0)^2 \leq \frac{\left(1+4||A||_2\sqrt{\frac{2}{\lambda_k^2 - \lambda_{k+1}^2}}\right)^2}{2(\sqrt{2}-1)\lambda_k}\epsilon$, completing the proof. □

Lemma C.14 is a simple Lemma that converts the guarantee from Lemma C.13 into one about the embedding functions parameterizing matrices F_θ and F_*.

Lemma C.14. *Let $n \geq k$ be natural numbers, and matrices $F_\theta, F_* \in \mathbb{R}^{n\times k}, R \in O^{k\times k}$ to where $\text{dist}(F_\theta, F_*) = ||F_\theta - F_*R||_F$. Suppose $\pi \in \Delta([n])$, $h_\theta : [n] \to \mathbb{R}^k, h_{\theta_*} : [n] \to \mathbb{R}^k$ parameterize F_θ, F_* as follows. For all $x \in [n]$,*

$$F_\theta[x] = \sqrt{\pi(x)}h_\theta(x) \tag{C.114}$$

$$F_*[x] = \sqrt{\pi(x)}h_{\theta_*}(x) \tag{C.115}$$

Then, $\text{dist}(F_\theta, F_*)^2 = \mathbb{E}_{x\sim\pi}||h_\theta - R\,h_{\theta_*}||_2^2$.

Proof.

$$\text{dist}(F_\theta, F_*)^2 = ||F_\theta - F_*R||_F^2 \tag{C.116}$$

$$= \sum_{x\in[n]} \pi(x)||h_\theta - R\,h_{\theta_*}||_2^2 \tag{C.117}$$

$$= \mathbb{E}_{x\sim\pi}||h_\theta - R\,h_{\theta_*}||_2^2 \tag{C.118}$$

□

C.1.5 Learning Contrastive Embeddings on Local Subgraphs

In addition, we would like the embeddings to be learned locally, not for the entire graph simultaneously. For a theorem graph $G = (V, E)$, define the following notion of a subgraph of G supported on the T-neighborhood of a node $x \in V$. In the following, we will denote $(C, P, w) := (C_w, P_w, w)$ as a reversible conjecturer satisfying Definition 4.1, to simplify notation.

Definition C.15 (Local Subgraph)**.** *Let $G = (V, E)$ be a connected, undirected graph, with reversible, nearest-neighbor-type Markov chain (V, P). For any $x \in V$, $T \in \mathbb{N}$, let $G_{x,T} = (V_{x,T}, E_{x,T})$ be a subgraph of G supported on $B(x, T)$, with edges (a, b) from $B(x, T) \ni a$ to $V - B(x, T) \ni b$ replaced by a self-loop, (a, a) in $G_{x,T}$. Let Markov chain $(V_{x,T}, P_{x,T})$, be the restriction of (V, P) to $G_{x,T}$, where any transition probability of (V, P) along edges from $B(x, T)$ to $V - B(x, T)$ are redirected back to their source via the aforementioned self-loops. Let $\pi_{x,T}$ be its stationary measure of $(V_{x,T}, P_{x,T})$. For $T \in N$, $\tau \in N$, $y \in B(x, T)$, denote $p^{\tau}_{G_{x,T}}(y)$ as the distribution of a τ step random walk on $G_{x,T}$ according to Markov chain $(V_{x,T}, P_{x,T})$.*

On a $T = \tau + 1$ local subgraph centered at $x \in V$, the τ-step random walk distribution of x and any neighbor $y \in N(x)$ is the same as that on the entire graph.

Lemma C.16. *Let $G = (V, E)$ be a connected, undirected graph with reversible, nearest-neighbor-type Markov chain (V, P). Let $T = \tau + 1$. Let $P_{x,T}$ be the restriction of P to $G_{x,T}$. For any $x \in V$ and any $y \in N(x)$, we have $p^{\tau}_{G_{x,T}}(x) = p^{\tau}(x)$ and $p^{\tau}_{G_{x,T}}(y) = p^{\tau}(y)$, where $p^{\tau}(x), p^{\tau}(y)$ is the usual scale-τ diffusion embeddings defined in Definition 5.4.*

Proof. Since $G_{x,\tau+1}$ is restricted to $B(x, \tau + 1)$, then $G_{x,\tau+1}$ and G look identical to a nearest-neighbor-type random walk starting at x or any $y \in N(x)$ of length at most τ. □

Lemma C.16 shows the scale-τ diffusion embedding can be computed locally, in a $(\tau + 1)$ neighborhood around the queried node. In addition, for any reversible Markov chain (V, P) on the entire graph G with stationary measure π, for any $x \in V$ and $T \in \mathbb{N}$, the stationary measure $\pi_{x,T}$ is equal to π restricted to $B(x, T)$, up to a multiplicative constant, by the self-loop construction of $G_{x,T}$. Thus, the diffusion similarity and distance computed locally differs from that computed globally by a multiplicative constant, which means the locally-computed diffusion similarity is sufficient for solving the optimization problem in Equation (5.1).

To learn the spectral diffusion embedding for reversible Markov chain (V, P), of conjecturer (C, P, w) in a local T-neighborhood around a node $x \in V$ via Algorithm 5, one will require access to the following oracles, defined in Definition C.15.

1. The push-forward map $C_{x,T} : B(x, T) \to \Delta(B(x, T))$ of the restricted Markov chain $(V_{x,T}, P_{x,T})$, where $\forall y \in B(x, T), C_{x,T}(y) = P_{x,T}[y] \in \Delta(B(x, T))$,
2. i.i.d. sample access to $\pi_{x,T} \in \Delta(B(x, T))$.

For $(V_{x,T}, P_{x,T})$, it suffices to have access to the original conjecturer $(C, P, w) := (C_w, P_w, w)$ on G and the neighbor oracle $N_{G_{x,T}} : V \to 2^V$ of the local graph $G_{x,T}$ supported on $B(x, T)$. This is because for any $y \in B(x, T)$, to sample $z \sim C_{x,T}(y)$, one can first sample $z' \sim C(y)$. Then, if $z' \in N_{G_{x,T}}(y)$, set $z = z'$. Else, set $z = y$. z will be a sample from $C_{x,T}(y)$ with the correct distribution.

Algorithm 6 describes how to compute $N_{G_{x,T}} : V \to 2^V$ and $\pi_{x,T}$. It involves first performing a T-step BFS starting from node x. Once the set of all theorems in the T-neighborhood of x is collected, it is straightforward to create an oracle for $N_{G_{x,T}}$. Regarding the sampler for $\pi_{x,T}$, because $\forall y \in B(x, T), \pi_{x,T}(y) \propto \pi(y) \propto w(y)$, then $\pi_{x,T}(y) = \frac{w(y)}{\sum_{z \in B(x,T)} w(z)}$. It suffices to compute the vector $\pi_{x,T}$ for all $y \in B(x, T)$.

C.2 Main Theorems

Theorem 5.5. *Let $G = (V, E)$ be a finite, undirected, connected theorem graph. Let conjecturer (C, P, w) satisfy Definition 4.1, and $\pi^{\top} = \pi^{\top} P$. Let $\mathcal{F}_{\text{emb}}$ be a class of embedding models $\mathcal{F}_{\text{emb}} \ni h_\theta : V \to \mathbb{R}^{|V|}$. For any $\tau \in \mathbb{N}$, define the scale-τ contrastive loss as:*

Algorithm 6 Compute Local Neighbor Oracle and Stationary Measure

Require: $T \in \mathbb{N}$, source theorem $x \in V$, neighbor oracle $N : V \to 2^V$, and conjecturer $(C, P, w) := (C_w, P_w, w)$

Ensure: Local neighbor oracle $N_{G_{x,T}} : B(x,T) \to 2^{B(x,T)}$ and local stationary measure $\pi_{x,T} \in \Delta(B(x,T))$

Set $B_{-1} \leftarrow \emptyset$ and $B_0 \leftarrow \{x\}$
for $s = 1, \ldots, T$ **do**
    Set $\partial B_{s-1} \leftarrow B_{s-1} \setminus B_{s-2}$
    Set $B_s \leftarrow B_{s-1} \cup \bigcup_{v \in \partial B_{s-1}} N(v)$.
Set $S \leftarrow B_T$

Define $N_{G_{x,T}}$ by the following subroutine:
    Local Neighbor Oracle $N_{G_{x,T}}$**:**
        Input: $y \in S$
        Query $N(y)$
        Set $m(y) \leftarrow |N(y) \cap S|$
        return $N(y) \cap S$ together with $\deg(y) - m(y)$ copies of y
Set

$$\pi_{x,T}(y) \leftarrow \frac{w(y)}{\sum_{z \in S} w(z)} \qquad \text{for all } y \in S.$$

return $N_{G_{x,T}}$ and $\pi_{x,T}$

$$\mathcal{L}^{\tau}_{\text{cont}}(h_\theta) := \mathbb{E}_{(x,x^-) \sim D^{\tau}_{-}} (h_\theta(x)^\top h_\theta(x^-))^2 - 2 \cdot \mathbb{E}_{(x,x^+) \sim D^{\tau}_{+}} h_\theta(x)^\top h_\theta(x^+) \tag{5.3}$$

Where $D^{\tau}_{+}, D^{\tau}_{-} \in \Delta(V \times V)$ are distributions of positive and negative pairs: $(a,b) \sim D^{\tau}_{+}$ where $z \sim \pi, a \sim C^\tau(z), b \sim C^\tau(z)$ (C^τ denotes applying C τ times) and $(a,b) \sim D^{\tau}_{-}$ where $a \sim \pi, b \sim \pi$. *Fix any $\tau \in \mathbb{N}$. Assume $\mathcal{F}_{\text{emb}}$ contains a global minimizer of $\mathcal{L}^{\tau}_{\text{cont}}$ over functions $V \to \mathbb{R}^{|V|}$. For any global minimizer $\mathcal{F}_{\text{emb}} \ni h_{\theta_*} : V \to \mathbb{R}^{|V|}$, then $\forall u, v \in V, \langle p^\tau(u), p^\tau(v) \rangle_{\ell^2(1/\pi)} = h_{\theta_*}(u)^\top h_{\theta_*}(v)$.*

Proof. By Lemma C.8, there exists an orthogonal matrix $R \in O^{|V| \times |V|}$ where

$$\forall u \in V, h_{\theta_*}(u) = R\,\Phi_\tau(u) \in \mathbb{R}^{|V|} \tag{C.119}$$

Thus,

$$\begin{aligned} \forall u, v \in V, \langle h_{\theta_*}(u), h_{\theta_*}(v) \rangle &= \langle R\,\Phi_\tau(u), R\,\Phi_\tau(v) \rangle && \text{(C.120)} \\ &= \langle \Phi_\tau(u), \Phi_\tau(v) \rangle && \text{(C.121)} \\ &= \langle p^\tau(u), p^\tau(v) \rangle_{\ell^2(1/\pi)} && \text{(By Lemma C.6)} \end{aligned}$$

□

Below is our end-to-end learning result, which can be viewed as a finite-sample, small-embedding dimension version of Theorem 5.5.

Theorem C.17 (End-to-End Contrastive Learning Result)**.** *Let $G = (V, E)$ be any connected, undirected graph. Let (V, P) be an irreducible, reversible Markov chain with transition matrix $P \in [0,1]^{|V| \times |V|}$ and stationary distribution $\pi \in \Delta(V)$.*

For any $\tau \in \mathbb{N}$, instantiate the contrastive loss $\mathcal{L}_{\text{cont}}$ with base distribution π, augmentation function $P^\tau : V \to \Delta(V)$, which augments $x \in V$ into a distribution of nodes given by $e_x^\top P^\tau$.

For any $k \in [|V| - 1]$, let $\mathcal{F}_{\text{emb}} \subset \mathcal{F}_V^{(k)}$ be a class of embedding models h_θ that map a domain V into dimension k Euclidean space: $\mathcal{F}_{\text{emb}} \ni h_\theta : V \to \mathbb{R}^k$. Let $\mathcal{L}_{\text{cont}} : \mathcal{F}_V^{(k)} \to \mathbb{R}$ be the

population contrastive loss, instantiated with base distribution π and augmentation function P^τ, as in Equation (C.12), and $\hat{\mathcal{L}}_{\text{cont},n} : \mathcal{F}_V^{(k)} \to \mathbb{R}$ be the empirical contrastive loss over an i.i.d. dataset of size n drawn from π, as in Equation (C.15). Assume:

1. *For some $\chi > 0$, we have that $\forall h_\theta \in \mathcal{F}_{\text{emb}}, \forall x \in V$, $||h_\theta(x)||_\infty \leq \chi$.*
2. *$\mathcal{F}_{\text{emb}}$ contains at least one global minimizer of $\mathcal{L}_{\text{cont}}$ over $\mathcal{F}_V^{(k)}$ (i.e. over functions $h : V \to \mathbb{R}^k$)*
3. *$\hat{\mathcal{R}}_n(\mathcal{F}_{\text{emb}}) \leq \sqrt{\frac{\mathcal{C}(\mathcal{F}_{\text{emb}})}{n}}$, where $\mathcal{C}(\mathcal{F}_{\text{emb}})$ is an $\mathcal{F}_{\text{emb}}$-dependent constant*
4. *With $\{\lambda_i\}_{i\in[|V|]} := \{\lambda_i(P^{2\tau})\}_{i\in[|V|]}$ being the eigenvalues of $P^{2\tau}$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{|V|} \geq 0$, assume $\lambda_k^2 - \lambda_{k+1}^2 > 0$.*
5. *There exists $\beta > 0$ such that $\forall y \in V$, $\sum_{x\in N(y)} \pi(x) \leq \beta\, \pi(y)$. (E.g. if P is the simple random walk, then $\beta = d$ the maximum degree of any node in G).*

Set any $\epsilon, \Delta \in (0,1)$ and set $p_{\text{erm fails}} = \frac{\epsilon}{\text{poly}(k,\chi)}$. Then, with probability at least $1 - p_{\text{erm fails}}$ over the randomness of a training dataset of size $n = \Theta(\frac{\mathcal{C}(\mathcal{F}_{\text{emb}})\,\text{poly}(k,\chi)}{\epsilon^2} \log \frac{1}{p_{\text{erm fails}}})$ drawn i.i.d. from π, the following guarantee holds for any ERM $h_{\hat{\theta}} \in \mathcal{F}_{\text{emb}}$ of $\hat{\mathcal{L}}_{\text{cont},n}$, where $\text{rank}(F_{\hat{\theta}}) = \text{rank}([\sqrt{\pi(x)} h_{\hat{\theta}}(x)^\top]_{x\in V}) = k$ (where $F_{\hat{\theta}}$ is a $|V| \times k$ matrix whose x-th row is $\sqrt{\pi(x)} h_{\hat{\theta}}(x)^\top$).

$$\mathbb{P}_{x\sim\pi}\left[\exists y, z \in N(x) \cup \{x\} \,:\, |\phi_\tau^{\text{diff}}(y,z) - \langle h_{\hat{\theta}}(y), h_{\hat{\theta}}(z)\rangle| > \frac{\lambda_{k+1}}{\pi_{\min}} + 2\chi\sqrt{k}\cdot\sqrt{\Delta} + \Delta\right] \tag{C.122}$$

$$\leq \frac{\epsilon}{\Delta} \cdot \frac{(\beta+1)\cdot\left(1 + 4\sqrt{\frac{2}{\lambda_k^2 - \lambda_{k+1}^2}}\right)^2}{2(\sqrt{2}-1)\lambda_k} \tag{C.123}$$

Proof. By Lemma C.4, with probability at least $1 - p_{\text{erm fails}}$,

$$\mathcal{L}_{\text{cont}}(h_{\hat{\theta}}) \leq \mathcal{L}_{\text{cont}}(h_{\theta_*}) + \epsilon \tag{C.124}$$

$$||\overline{A} - F_{\hat{\theta}} F_{\hat{\theta}}^\top||_F^2 \leq ||\overline{A} - F_{\theta_*} F_{\theta_*}^\top||_F^2 + \epsilon \tag{C.125}$$

Under that event, since $\text{rank}(F_{\hat{\theta}}) = k$ by assumption, by Lemma C.13 and Lemma C.14,

$$\mathbb{E}_{x\sim\pi}||h_{\hat{\theta}}(x) - R\, h_{\theta_*}(x)||_2^2 \leq \epsilon \cdot \frac{\left(1 + 4\sqrt{\frac{2}{\lambda_k^2 - \lambda_{k+1}^2}}\right)^2}{2(\sqrt{2}-1)\lambda_k} \tag{C.126}$$

Where we've used that $||\overline{A}||_2 = \lambda_1(\overline{A}) = \lambda_1(P^{2\tau}) = 1$. By Markov's inequality,

$$\mathbb{P}_{x\sim\pi}\left[||h_{\hat{\theta}}(x) - R\, h_{\theta_*}(x)||_2^2 \geq \Delta\right] \leq \frac{\epsilon}{\Delta} \cdot \frac{\left(1 + 4\sqrt{\frac{2}{\lambda_k^2 - \lambda_{k+1}^2}}\right)^2}{2(\sqrt{2}-1)\lambda_k} \tag{C.127}$$

In fact, using the premise that $\exists \beta > 0 \,:\, \forall y \in V, \sum_{x\in N(y)} \pi(x) \leq \beta\, \pi(y)$, we can extend this closeness guarantee to a neighborhood around $x \sim \pi$.

$$\mathbb{P}_{x\sim\pi}\Big[\exists y \in N(x) : ||h_{\hat{\theta}}(y) - R\, h_{\theta_*}(y)||_2^2 \geq \Delta\Big] \tag{C.128}$$

$$\leq \mathbb{E}_{x\sim\pi} \sum_{y\in N(x)} 1\Big[||h_{\hat{\theta}}(y) - R\, h_{\theta_*}(y)||_2^2 \geq \Delta\Big] \tag{C.129}$$

$$= \sum_{x\in V} \pi(x) \sum_{y\in N(x)} 1\Big[||h_{\hat{\theta}}(y) - R\, h_{\theta_*}(y)||_2^2 \geq \Delta\Big] \tag{C.130}$$

$$\leq \sum_{y\in V} \beta\, \pi(y)\, 1\Big[||h_{\hat{\theta}}(y) - R\, h_{\theta_*}(y)||_2^2 \geq \Delta\Big] \tag{C.131}$$

$$= \beta \cdot \mathbb{P}_{y\sim\pi}\Big[||h_{\hat{\theta}}(y) - R\, h_{\theta_*}(y)||_2^2 \geq \Delta\Big] \tag{C.132}$$

$$\leq \beta \cdot \frac{\epsilon}{\Delta} \cdot \frac{\left(1 + 4\sqrt{\frac{2}{\lambda_k^2 - \lambda_{k+1}^2}}\right)^2}{2(\sqrt{2}-1)\lambda_k} \tag{C.133}$$

Thus,

$$\mathbb{P}_{x\sim\pi}\Big[\exists y \in N(x) : ||h_{\hat{\theta}}(y) - R\, h_{\theta_*}(y)||_2^2 \geq \Delta \text{ or } ||h_{\hat{\theta}}(x) - R\, h_{\theta_*}(x)||_2^2 \geq \Delta\Big] \tag{C.134}$$

$$\leq \frac{\epsilon}{\Delta} \cdot \frac{(\beta+1) \cdot \left(1 + 4\sqrt{\frac{2}{\lambda_k^2 - \lambda_{k+1}^2}}\right)^2}{2(\sqrt{2}-1)\lambda_k} \tag{C.135}$$

Under the event that $x \sim \pi$ satisfies

$$\forall y \in N(x) : ||h_{\hat{\theta}}(y) - R\, h_{\theta_*}(y)||_2^2 \leq \Delta \text{ and } ||h_{\hat{\theta}}(x) - R\, h_{\theta_*}(x)||_2^2 \leq \Delta \tag{C.136}$$

Then, $\forall y, z \in N(x) \cup \{x\}$,

$$\langle h_{\hat{\theta}}(y), h_{\hat{\theta}}(z)\rangle \leq \langle R\, h_{\theta_*}(y), R\, h_{\theta_*}(z)\rangle + 2 \max_{h_\theta \in \mathcal{F}_{\text{emb}}} \max_{\hat{y}\in V} ||h_\theta(\hat{y})||_2 \cdot \sqrt{\Delta} + \Delta \tag{C.137}$$

$$\leq \langle h_{\theta_*}(y), h_{\theta_*}(z)\rangle + 2\chi\sqrt{k} \cdot \sqrt{\Delta} + \Delta \tag{C.138}$$

$$\text{Similarly, } \langle h_{\hat{\theta}}(y), h_{\hat{\theta}}(z)\rangle \geq \langle h_{\theta_*}(y), h_{\theta_*}(z)\rangle - 2\chi\sqrt{k} \cdot \sqrt{\Delta} - \Delta \tag{C.139}$$

By Lemma C.8 (which says $\langle h_{\theta_*}(x), h_{\theta_*}(y)\rangle = \langle \Phi_t^{\leq k}(x), \Phi_t^{\leq k}(y)\rangle$) and Lemma C.7, then $\forall x, y \in V$,

$$\langle \Phi_t(x), \Phi_t(y)\rangle - \frac{\lambda_{k+1}}{\pi_{\min}} \leq \langle h_{\theta_*}(x), h_{\theta_*}(y)\rangle \leq \langle \Phi_t(x), \Phi_t(y)\rangle + \frac{\lambda_{k+1}}{\pi_{\min}} \tag{C.140}$$

It follows that when $x \sim \pi$ satisfies Equation (C.136), then with $\xi := \frac{\lambda_{k+1}}{\pi_{\min}} + 2\chi\sqrt{k} \cdot \sqrt{\Delta} + \Delta$, we have

$$\langle \Phi_t(y), \Phi_t(z)\rangle - \xi \leq \langle h_{\hat{\theta}}(y), h_{\hat{\theta}}(z)\rangle \leq \langle \Phi_t(y), \Phi_t(z)\rangle + \xi \tag{C.141}$$

Since by Lemma C.6 we have $\phi_\tau^{\text{diff}}(y, z) = \langle \Phi_t(y), \Phi_t(z)\rangle$, then with $\xi := \frac{\lambda_{k+1}}{\pi_{\min}} + 2\chi\sqrt{k} \cdot \sqrt{\Delta} + \Delta$, we have

$$\mathbb{P}_{x\sim\pi}\Big[\exists y,z\in N(x)\cup\{x\}\ :\ |\phi_\tau^{\text{diff}}(y,z)-\langle h_{\hat\theta}(y),h_{\hat\theta}(z)\rangle|>\xi\Big] \tag{C.142}$$

$$\le \frac{\epsilon}{\Delta}\cdot\frac{(\beta+1)\cdot\left(1+4\sqrt{\frac{2}{\lambda_k^2-\lambda_{k+1}^2}}\right)^2}{2(\sqrt{2}-1)\lambda_k} \tag{C.143}$$

In short, for sufficiently small ϵ, the learned embedding model $h_{\hat\theta}: V\to\mathbb{R}^k$ can be used to approximate $\phi_\tau^{\text{diff}}(y,z)$ of neighboring points of *most* $x\in V$ by the inner product, $\langle h_{\hat\theta}(y),h_{\hat\theta}(z)\rangle$.

□

D Instantiating Algorithm 3 on Clustered Graphs

In this section, we first describe a concrete end-to-end result for how Algorithm 3 improves over a simple random walk, in terms of diversity (Definition 5.2), on a special class of clustered graphs.

D.1 End-to-End Result of Algorithm 3 on Clique Graphs

Proposition D.2 is an end-to-end result for Algorithm 3, showing that on a special class of theorem graphs (Definition D.1), that Algorithm 3 explores the graph faster than a simple random walk. The final guarantee is in terms of the spectral gaps of the random walks, and the diversity of the distributions they generate, according to Definition 5.2.

Definition D.1. *(Clique-Graph) Let parameters M be the clique size, r be the meta-graph degree, and K be the number of cliques. Let $Q=([K],E(Q))$ be a finite connected, undirected, r-regular graph. For $i\in[K]$, denote $N_Q(i)\subset[K]$ as the neighbors of i in Q. Suppose there exists $c\in[r]$ where for every $i\in[K]$ and every two distinct neighbors $j,k\in N_Q(i)$, $|N_Q(j)\cap N_Q(k)|=c$. For clique size $M\ge 2$, define the graph $G=G(Q,M)$ by $V=[K]\times[M]$, where each $C_i:=\{i\}\times[M]$ is a clique of size M. For every meta-edge $ij\in E(Q)$ and $a\in[M]$, add the inter-clique edge $(i,a)\sim(j,a)$.*

Proposition D.2. *(End-to-End Guarantee of Algorithm 3) Let $Q=([K],E(Q))$ be a finite connected, undirected, r-regular graph. Let $G=G(Q,M)$ be a Clique-Graph satisfying Definition D.1 with meta-graph Q. Suppose $M\ge r+3$. Let P_Q denote the simple random walk over Q. Algorithm 3, with parameters $\tau=1,\eta=0$, defines a random walk over the meta-graph Q with transition matrix $P^{(Q)}_{\text{improved}}=(\frac{1}{r+1}+O(\frac{1}{M}))I+(\frac{r}{r+1}+O(\frac{1}{M}))P_Q$. The simple random walk over G defines a random walk over the meta-graph Q with transition matrix $P^{(Q)}_{\text{baseline}}=\frac{M-1}{M-1+r}I+\frac{r}{M-1+r}P_Q$. Define the spectral gap of transition matrix P as $\gamma(P):=1-\max(\lambda_2(P),|\lambda_K(P)|)$. Then, there exists a universal constant $c>0$ such that*

$$\gamma(P^{(Q)}_{\text{improved}})\ge c\,\frac{M}{r^2}\,\gamma(P^{(Q)}_{\text{baseline}}) \tag{D.1}$$

In particular, for $M\gg\max(\frac{r^2}{c},r+3)$, we have $\gamma(P^{(Q)}_{\text{improved}})\gg\gamma(P^{(Q)}_{\text{baseline}})$.

As a consequence, suppose $D_0\in\Delta([K]\times[M])$ is uniform on a starting clique $\{i\}\times[M]$. For all $t\in\mathbb{N}$, let D_t^{improved} be the result of applying the random walk defined in Algorithm 3 for t steps on D_0, and D_t^{baseline} the result of applying a simple random walk over G for t steps on D_0. Then, for any $\epsilon\in(0,1)$, for $t_{\text{baseline}}:=\frac{1}{2\gamma(P^{(Q)}_{\text{baseline}})}\log\frac{K}{\epsilon}$, $\text{Div}_1^{P_{\text{s.r.w.}}}(D^{\text{baseline}}_{t_{\text{baseline}}})\ge 1-\epsilon$, with maximum value 1. Whereas, for $t_{\text{improved}}:=\frac{1}{2\gamma(P^{(Q)}_{\text{improved}})}\log\frac{K}{\epsilon}\le O(\frac{r^2}{M})\,t_{\text{baseline}}$, $\text{Div}_1^{P_{\text{s.r.w.}}}(D^{\text{improved}}_{t_{\text{improved}}})\ge 1-\epsilon$, with maximum value 1.

Proof. Fix a node $x=(i,a)\in C_i$. Its neighbors split into two classes:

$$I_x:=\{(i,b):b\in[M],\ b\ne a\} \tag{D.2}$$

$$E_x:=\{(j,a):j\in N_Q(i)\}, \tag{D.3}$$

with $|I_x| = M-1$, $|E_x| = r$, and total degree $d = M-1+r$. For $\tau = 1$, since G is d-regular, $\forall x, p^1(x) = \frac{1}{d}\mathbf{1}_{N(x)}$. The matrix M'_x in Algorithm 3 equals, up to a constant multiplicative factor of d^{-3}, $M'_x[y,z] \propto |N(y) \cap N(z)|, y, z \in N(x)$. Therefore Algorithm 3 minimizes

$$F_x(\mu) := \sum_{y,z \in N(x)} \mu(y)\mu(z)\,|N(y) \cap N(z)| \qquad \text{over } \mu \in \Delta(N(x)). \tag{D.4}$$

Write $u_b := (i,b) \in I_x$ and $v_j := (j,a) \in E_x$. We compute the relevant neighborhood intersections.

(i) Internal/internal. For any $b \in [M], b \neq a$, $|N(u_b) \cap N(u_b)| = d$. For any distinct $b, b' \neq a$, $|N(u_b) \cap N(u_{b'})| = M-2$. Indeed, u_b and $u_{b'}$ share exactly the $M-2$ vertices in C_i other than u_b and $u_{b'}$.

(ii) Internal/external. For any $b \neq a$ and any $j \in N_Q(i)$, $|N(u_b) \cap N(v_j)| = 2$. The two common neighbors are $x = (i,a)$ and (j,b).

(iii) External/external. For any $j \in N_Q(i)$, $|N(v_j) \cap N(v_j)| = d$. For any distinct $j, k \in N_Q(i)$, $|N(v_j) \cap N(v_k)| = c$. Indeed, the common neighbors are exactly the vertices $\{(\ell, a) : \ell \in N_Q(j) \cap N_Q(k)\}$ whose number is c by assumption.

Now let $\mu_b := \mu(u_b),\ \nu_j := \mu(v_j),\ s := \sum_{b \neq a} \mu_b$. Then $\sum_{j \in N_Q(i)} \nu_j = 1 - s$, and the objective becomes

$$F_x(\mu) = d\sum_{b \neq a} \mu_b^2 + (M-2) \sum_{\substack{b,b' \neq a \\ b \neq b'}} \mu_b \mu_{b'} + d \sum_{j \in N_Q(i)} \nu_j^2 + c \sum_{\substack{j,k \in N_Q(i) \\ j \neq k}} \nu_j \nu_k + 2 \cdot \sum_{j \in N_Q(i),\, b \neq a} \nu_j \mu_b \tag{D.5}$$

$$= \big(d - (M-2)\big) \sum_{b \neq a} \mu_b^2 + (M-2)s^2 + (d-c) \sum_{j \in N_Q(i)} \nu_j^2 + c(1-s)^2 + 2 \cdot 2s(1-s) \tag{D.6}$$

Since $d - (M-2) = r + 1 > 0$, $d - c > 0$, for fixed s the objective is minimized by taking μ_b uniform on I_x and ν_j uniform on E_x, i.e.

$$\mu_b = \frac{s}{M-1} \tag{D.7}$$

$$\nu_j = \frac{1-s}{r} \tag{D.8}$$

Thus it suffices to minimize the one-variable quadratic

$$f_M(s) = A_M s^2 + 4s(1-s) + B_M(1-s)^2 \tag{D.9}$$

where $A_M = M - 2 + \frac{r+1}{M-1}$ and $B_M = c + \frac{d-c}{r}$. Expanding,

$$f_M(s) = (A_M + B_M - 4)s^2 + (4 - 2B_M)s + B_M. \tag{D.10}$$

Because $M \geq 3$, we have $A_M + B_M - 4 > 0$, so f_M has the unique minimizer

$$s = \eta_M := \frac{B_M - 2}{A_M + B_M - 4}. \tag{D.11}$$

Hence the unique optimizer of Algorithm 3 at $x = (i,a)$ is

$$\mu_x^* = \frac{\eta_M}{M-1} \sum_{b \neq a} e_{(i,b)} + \frac{1-\eta_M}{r} \sum_{j \in N_Q(i)} e_{(j,a)} \tag{D.12}$$

In particular, Algorithm 3 sends total mass η_M to the current cluster C_i, and total mass $(1-\eta_M)/r$ to each neighboring cluster C_j, $j \in N_Q(i)$. These transition dynamics depend only on the cluster index i, not on the label a. Thus, we can describe the dynamics of the Markov chain in terms of cluster-level transition probabilities, i.e. as a random walk over the meta-graph Q, where each cluster $\{C_i\}_{i=1}^K$

corresponds to one state. As a random walk over the meta-graph Q, Algorithm 3 is described by a transition matrix $P^{(Q)}_{\text{improved}} \in [0,1]^{K\times K}$.

$$P^{(Q)}_{\text{improved}} = \eta_M I_K + (1-\eta_M)P_Q \in [0,1]^{K\times K} \tag{D.13}$$

Similarly, the simple random walk on G can also be described as a random walk over the meta-graph Q, whose transition matrix we denote $P^{(Q)}_{\text{baseline}} \in [0,1]^{K\times K}$. From any $x = (i,a) \in C_i$ there are $M-1$ neighbors inside C_i and exactly one neighbor in C_j for each $j \in N_Q(i)$. Therefore,

$$P^{(Q)}_{\text{baseline}} = \frac{M-1}{M-1+r} I_K + \frac{r}{M-1+r} P_Q \in [0,1]^{K\times K} \tag{D.14}$$

Regarding notation, define $P_{\text{baseline}} \in [0,1]^{KM\times KM}$ and $P_{\text{improved}} \in [0,1]^{KM\times KM}$ as the simple random walk over G and the random walk induced by Algorithm 3 over G (not over the meta-graph Q), respectively. $P^{(Q)}_{\text{baseline}}$ and $P^{(Q)}_{\text{improved}}$ are the meta-graph analogs of P_{baseline} and P_{improved}.

Additionally, for transition matrix P of a Markov chain, define the spectral gap as $\gamma(P) := 1 - \max(\lambda_2(P), |\lambda_K(P)|)$. Then,

$$\begin{aligned}
\gamma(P^{(Q)}_{\text{baseline}}) &= 1 - \lambda_2(P^{(Q)}_{\text{baseline}}) && \text{(Since when } M-1 \geq r\text{, then } P^{(Q)}_{\text{baseline}} \succeq 0) \\
&= 1 - [\frac{M-1}{M-1+r} + \frac{r}{M-1+r}\lambda_2(P_Q)] && \text{(D.15)} \\
&= \frac{r}{M-1+r}(1-\lambda_2(P_Q)) && \text{(D.16)} \\
\gamma(P^{(Q)}_{\text{improved}}) &= \min\left(1-\lambda_2(\eta_M I + (1-\eta_M)P_Q), 1 - |\lambda_K(\eta_M I + (1-\eta_M)P_Q)|\right) && \text{(D.17)} \\
&= \min\left((1-\eta_M)(1-\lambda_2(P_Q)), 1 - |\lambda_K(\eta_M I + (1-\eta_M)P_Q)|\right) && \text{(D.18)} \\
&\geq \min\left((1-\eta_M)(1-\lambda_2(P_Q)), 1 - |2\eta_M - 1|\right) && \text{(Since } \lambda_K(P_Q) \geq -1\text{)} \\
\implies \frac{\gamma(P^{(Q)}_{\text{improved}})}{\gamma(P^{(Q)}_{\text{baseline}})} &\geq \frac{M-1+r}{r}\min(1-\eta_M, \frac{1-|2\eta_M-1|}{1-\lambda_2(P_Q)}) && \text{(D.19)} \\
&\geq \frac{M-1+r}{r}\min(1-\eta_M, \frac{1-|2\eta_M-1|}{2}) && (1-\lambda_2(P_Q) \leq 2) \\
&\geq \frac{M-1+r}{r}\min(1-\eta_M, \eta_M) && \text{(D.20)} \\
&\geq \frac{M-1+r}{r}\cdot\frac{c+\frac{d-c}{r}-2}{M-2+\frac{r+1}{M-1}+c+\frac{d-c}{r}-4} && \text{(D.21)} \\
&\geq \Omega(\frac{M}{r^2}) && \text{(D.22)}
\end{aligned}$$

This proves the first claim.

Towards the second claim, by assumption, D_0 is uniform over its starting cluster:

$$\forall a, a' \in [M], \forall i \in [K], D_0((i,a)) = D_0((i,a')) \tag{D.23}$$

By an inductive argument we have that for all $t \in \mathbb{N}$, D^{improved}_t and D^{baseline}_t are uniform over clusters:

$$\begin{aligned}
\forall a, a' \in [M], \forall i \in [K], D^{\text{improved}}_t((i,a)) = D^{\text{improved}}_t((i,a')) \text{ and} && \text{(D.24)} \\
D^{\text{baseline}}_t((i,a)) = D^{\text{baseline}}_t((i,a')) && \text{(D.25)}
\end{aligned}$$

Thus, for all $t \in \mathbb{N}$, D_t^{improved} (resp. D_t^{baseline}) is characterized by a probability distribution $q_t^{\text{improved}} \in \Delta([K])$ (resp. q_t^{baseline}) over clusters, where each node in cluster i has probability $\frac{q_t^{\text{improved}}(i)}{M}$ (resp. $\frac{q_t^{\text{baseline}}(i)}{M}$). The stationary measure π is uniform over the KM nodes of G since G is regular. Thus,

$$\forall t \in \mathbb{N}, ||D_t^{\text{improved}}||^2_{\ell^2(1/\pi)} = KM \cdot \frac{1}{M} \cdot ||q_t^{\text{improved}}||_2^2 \tag{D.26}$$

$$\forall t \in \mathbb{N}, ||D_t^{\text{baseline}}||^2_{\ell^2(1/\pi)} = KM \cdot \frac{1}{M} \cdot ||q_t^{\text{baseline}}||_2^2 \tag{D.27}$$

Let $u = \frac{1}{K}\mathbf{1}$ be the stationary measure of both $P^{(Q)}_{\text{baseline}}$ and $P^{(Q)}_{\text{improved}}$, random walks over the meta-graph Q. With $\tau = 1$ and setting $P := P_{\text{s.r.w.}} = P_{\text{baseline}}$, for any $t \in \mathbb{N}$ we have:

$$\forall t \in \mathbb{N}, ||D_t^{\text{improved}}||^2_{\ell^2(1/\pi)} = 1 + K \cdot ||q_t^{\text{improved}} - u||_2^2 \tag{D.28}$$

$$\implies \forall t \in \mathbb{N}, \text{Div}_\tau^P(D_t^{\text{improved}}) = \frac{1}{||(D_t^{\text{improved}})^\top P_{\text{baseline}}||^2_{\ell^2(1/\pi)}} \tag{D.29}$$

$$= \frac{1}{1 + K \cdot ||(q_t^{\text{improved}} - u)^\top P^{(Q)}_{\text{baseline}}||_2^2} \tag{D.30}$$

$$= \frac{1}{1 + K \cdot ||(q_0 - u)^\top (P^{(Q)}_{\text{improved}})^t P^{(Q)}_{\text{baseline}}||_2^2} \tag{D.31}$$

For $t = \frac{1}{2\gamma(P^{(Q)}_{\text{improved}})} \log \frac{K}{\epsilon}$, then

$$K\, ||(q_0 - u)^\top (P^{(Q)}_{\text{improved}})^t P^{(Q)}_{\text{baseline}}||_2^2 \tag{D.32}$$

$$\leq K\, [(1 - \gamma(P^{(Q)}_{\text{improved}}))^t (1 - \gamma(P^{(Q)}_{\text{baseline}}))]^2 ||q_0 - u||_2^2 \tag{D.33}$$

$$\leq K\, (1 - \gamma(P^{(Q)}_{\text{improved}}))^{2t} \tag{D.34}$$

$$\leq \epsilon \tag{D.35}$$

Since $\epsilon < 1$, when $t = \frac{1}{2\gamma(P^{(Q)}_{\text{improved}})} \log \frac{K}{\epsilon}$, we have that $\text{Div}_\tau^P(D_t^{\text{improved}}) \geq \frac{1}{1+\epsilon} \geq 1 - \epsilon$.

By an analogous argument for D_t^{baseline}, for any $t \in \mathbb{N}$ we have

$$\forall t \in \mathbb{N}, ||D_t^{\text{baseline}}||^2_{\ell^2(1/\pi)} = 1 + K \cdot ||q_t^{\text{baseline}} - u||_2^2 \tag{D.36}$$

$$\implies \forall t \in \mathbb{N}, \text{Div}_\tau^P(D_t^{\text{baseline}}) = \frac{1}{||(D_t^{\text{baseline}})^\top P_{\text{baseline}}||^2_{\ell^2(1/\pi)}} \tag{D.37}$$

$$= \frac{1}{1 + K \cdot ||(q_t^{\text{baseline}} - u)^\top P^{(Q)}_{\text{baseline}}||_2^2} \tag{D.38}$$

$$= \frac{1}{1 + K \cdot ||(q_0 - u)^\top (P^{(Q)}_{\text{baseline}})^t P^{(Q)}_{\text{baseline}}||_2^2} \tag{D.39}$$

For $t = \frac{1}{2\gamma(P^{(Q)}_{\text{baseline}})} \log \frac{K}{\epsilon}$, then $\text{Div}_\tau^P(D_t^{\text{baseline}}) \geq \frac{1}{1+\epsilon} \geq 1 - \epsilon$.

□

E Justification of Assumption 2

The core result of this section is Lemma E.1, which justifies Assumption 2 by showing that finer-grained assumptions inspired from statistical learning theory implies Assumption 2.

Lemma E.1 (Finer-Grained Assumptions)**.** *Assumptions 6, 7, 8 imply Assumption 2.*

E.1 Statistical Learning Theory Inspired Assumptions

First, we introduce Assumptions 6 - 8. Let Ver : $V \times \{0,1\}^* \to \{0,1\}$ be a formal verifier where for theorem $x \in V$, then $\text{Ver}(x, y) = 1$ iff $y \in \{0,1\}^*$ is a correct proof of x. The first assumption assumes the function class $\mathcal{F}_{\text{prover}}$ of provers $\mathcal{F}_{\text{prover}} \ni f : V \to \Delta(\{0,1\}^*)$ can express the ground-truth function.

Assumption 6. *(Expressivity) There exists $f^* \in \mathcal{F}_{\text{prover}}$ which is a ground-truth prover, so that for every theorem $x \in V$, $f^*(x) \in \Delta(\{0,1\}^*)$ is distribution over correct proofs of x.*

Each $f \in \mathcal{F}_{\text{prover}}$ maps a theorem to a distribution of proofs ($f : V \to \Delta(\{0,1\}^*)$, where V is the set of theorems). Thus we can define a notion of the success rate of a prover on a theorem as the probability that the prover outputs a correct proof. Formally, let $f(x, r)$ be the output of prover f on theorem x and random seed r, where $r \sim D_r$. Define the following loss function over $\mathcal{F}_{\text{prover}}$ (which abstracts away the randomness $r \sim D_r$ into a success rate):

Definition E.2. *(Loss Functions) Define the following loss functions on provers in $\mathcal{F}_{\text{prover}}$.*

$$\ell(f, x) := \mathbb{E}_{r \sim D_r}[1 - \text{Ver}(x, f(x, r))] \tag{E.1}$$

$$= 1 - \text{succ}(x, f) \tag{E.2}$$

$$\mathcal{L}(f, D) := \mathbb{E}_{x \sim D}\, \ell(f, x) \tag{E.3}$$

$$\mathcal{L}_n(f, \{x_i\}_{i \in [n]}) := \frac{1}{n} \sum_{i \in [n]} \ell(f, x_i) \tag{E.4}$$

With the understanding that $S_n = \{x_i\}_{i \in [n]}$ are drawn i.i.d. from D. Let $\ell(\mathcal{F}_{\text{prover}}) := \{x \to \ell(f, x) : f \in \mathcal{F}_{\text{prover}}\}$ be the family of loss functions induced by $\mathcal{F}_{\text{prover}}$ and ℓ.

We define the empirical risk minimizer (ERM) as follows.

Definition E.3. *(Minimizers of Population and Empirical Risk) Let $D \in \Delta(V)$ be any distribution over theorems. For any $n \in \mathbb{N}_+$, suppose $\{x_i\}_{i \in [n]}$ are drawn i.i.d. from D. First, note that f^* is the population minimizer,*

$$\mathcal{L}(f^*, D) = 0 \tag{E.5}$$

Now, define the ERM $\hat{f}^{(n)}$.

$$\hat{f}^{(n)} \leftarrow \arg \min_{f \in \mathcal{F}_{\text{prover}}} \mathcal{L}_n(f, \{x_i\}_{i \in [n]}) \tag{E.6}$$

The Rademacher complexity[3] is defined as follows.

Definition E.4. *For a function class $\mathcal{F}$, and data distribution D, the average Rademacher complexity is defined as*

$$\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_{\{x_i\}_{i \in [n]} \overset{\text{iid}}{\sim} D} \mathbb{E}_{\{\sigma_i\}_{i \in [n]} \overset{\text{iid}}{\sim} \{\pm 1\}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right] \tag{E.7}$$

Upper bounding the Rademacher complexity of $\ell(\mathcal{F}_{\text{prover}})$ will upper bound the generalization error of the ERM. In particular,

Lemma E.5. *(Corollary 4.19 of [Ma22]) For any $\delta > 0, n \in \mathbb{N}$, let $S_n := \{x_i\}_{i \in [n]}$ be a training set of size n drawn i.i.d. from D, and let $\hat{f}^{(n)}$ minimize the empirical loss over S_n (via Definition E.3). Then, with probability at least $1 - \delta$ over the draw of S_n,*

$$\mathcal{L}(\hat{f}^{(n)}, D) - \mathcal{L}(f^*, D) \leq 2\mathcal{R}_n(\ell(\mathcal{F}_{\text{prover}})) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \tag{E.8}$$

[3]This definition is different than Definition C.2, previously defined and used in the context of [HWGM22] in Section C

Proof.

$$\mathcal{L}(\hat{f}^{(n)}, D) - \mathcal{L}(f^*, D) \leq \mathcal{L}(\hat{f}^{(n)}, D) - \mathcal{L}_n(\hat{f}^{(n)}, S_n) + \mathcal{L}_n(f^*, S_n) - \mathcal{L}(f^*, D) \tag{E.9}$$

$$\leq 2 \cdot \sup_{f \in \mathcal{F}_{\text{prover}}} |\mathcal{L}(f, D) - \mathcal{L}_n(f, \{x_i\})| \tag{E.10}$$

Then, apply Corollary 4.19 of [Ma22]. □

In this context, we assume an upper bound on the Rademacher Complexity of the function class $\ell(\mathcal{F}_{\text{prover}})$.

Assumption 7. *(Generalization) Given prover class $\mathcal{F}_{\text{prover}}$ assume there exists a $\mathcal{F}_{\text{prover}}$-dependent quantity $\mathcal{C}(\mathcal{F}_{\text{prover}})$ where for any theorem distribution $D \in \Delta(V)$, with function class $\ell(\mathcal{F}_{\text{prover}})$ defined as in Definition E.2, we have:*

$$\mathcal{R}_n(\ell(\mathcal{F}_{\text{prover}})) \leq \sqrt{\frac{\mathcal{C}(\mathcal{F}_{\text{prover}})}{n}} \tag{E.11}$$

The next assumption describes under what conditions it is efficient to find the ERM from the model class $\mathcal{F}_{\text{prover}}$, on a size n dataset of i.i.d. theorems from a distribution of theorems. In particular, a necessary condition for such a procedure to find the ERM is that each theorem in the training distribution has a proof that can be efficiently found.

Assumption 8. *(Optimization) Suppose we have i.i.d. sample access to theorem distribution $D \in \Delta(V)$, and for any $x \in \text{supp}(D)$, the proof of x is known. Then, for any n, it is computationally tractable to find the ERM $\hat{f}^{(n)}$ (Equation (E.6)) of a size n dataset drawn i.i.d. from D.*

Note that in practice, we may only be able to efficiently find an ϵ_{opt}-ERM, for any reasonable $\epsilon_{\text{opt}} > 0$, where an ϵ_{opt}-ERM is a prover $\tilde{f}^{(n)}$ where

$$\mathcal{L}_n(\tilde{f}^{(n)}, \{x_i\}_{i\in[n]}) \leq \min_{f \in \mathcal{F}_{\text{prover}}} \mathcal{L}_n(f, \{x_i\}_{i\in[n]}) + \epsilon_{\text{opt}} \tag{E.12}$$

However, the remaining analysis only requires $\mathcal{L}(\tilde{f}^{(n)}, D) - \mathcal{L}(f^*, D) \leq \epsilon_{\text{opt}} + 2 \cdot \sup_{f \in \mathcal{F}_{\text{prover}}} |\mathcal{L}(f, D) - \mathcal{L}_n(f, \{x_i\})|$. With $n = \Theta(\frac{\mathcal{C}(\mathcal{F}_{\text{prover}})}{\epsilon_{\text{gen}}^2})$ samples, we get $L(\tilde{f}^{(n)}, D) - L(f^*, D) \leq \epsilon_{\text{opt}} + \epsilon_{\text{gen}}$ with high probability. Thus, for any reasonable desired accuracy level ϵ, we can set $\epsilon_{\text{opt}}, \epsilon_{\text{gen}} = \epsilon/2$, finding the ϵ_{opt}-ERM with generalization gap of ϵ_{gen}. The necessary sample complexity would only increase by a factor of 4, a constant factor which we neglect. With this understanding, we keep the current phrasing of Assumption 8 of assuming the ERM is tractable.

E.2 Proof of Lemma E.1

We now prove the main result of this section.

Lemma E.1 (Finer-Grained Assumptions)**.** *Assumptions 6, 7, 8 imply Assumption 2.*

Proof. Since $\forall x \in \text{supp}(D)$, the proof of x is known, then by Assumption 8, given $\delta > 0$ and $n \in \mathbb{N}_+$ i.i.d. samples from D, we can compute $\hat{f}^{(n)}$ via Equation (E.6). By Lemma E.5, Assumption 6, and Assumption 7, we have that with probability at least $1 - \delta$,

$$\mathbb{E}_{x\sim D}[1 - \text{succ}(x, \hat{f}^{(n)})] = L(\hat{f}^{(n)}, D) \leq \frac{4 \cdot \sqrt{\mathcal{C}(\mathcal{F}_{\text{prover}})}}{\sqrt{n}} + 2\sqrt{\frac{\log \frac{2}{\delta}}{2n}} \tag{E.13}$$

By Markov's inequality, for any $p \in (0, 1)$,

$$\mathbb{P}_{x\sim D}[\text{succ}(x, \hat{f}^{(n)}) \geq p] \geq 1 - \frac{\frac{4\sqrt{\mathcal{C}(\mathcal{F}_{\text{prover}})} + 2\sqrt{\frac{1}{2}\log\frac{2}{\delta}}}{\sqrt{n}}}{1 - p} \tag{E.14}$$

Thus, it suffices to set n to be $(\frac{4\sqrt{\mathcal{C}(\mathcal{F}_{\text{prover}})}+2\sqrt{\frac{1}{2}\log\frac{2}{\delta}}}{\epsilon(1-p)})^2 \leq \Theta(\frac{\mathcal{C}(\mathcal{F}_{\text{prover}})\log\frac{1}{\delta}}{\epsilon^2(1-p)^2})$ in order for $\mathbb{P}_{x\sim D}[\text{succ}(x,\hat{f}^{(n)}) \geq p] \geq 1-\epsilon$. □

F Auxiliary Lemmas

Lemma 5.3 (Large Div_τ^P implies large K_τ^p)**.** *Let $p \in [0,1]$ be a universal constant in Assumption 2. Instantiate $\text{Div}_\tau^P : \Delta(V) \to \mathbb{R}$ with the transition matrix P of a reversible, nearest-neighbor-type random walk. For any $\tau \in \mathbb{N}$, prover $f \in \mathcal{F}_{\text{prover}}$, and theorem distribution $D \in \Delta(V)$, the τ-expanded knowledge set of f satisfies $|K_\tau^p(f)| \geq \frac{(P_{x\sim D}[\text{succ}(x,f)\geq p])^2}{\sup_{x\in V}\pi(x)}\text{Div}_\tau^P(D)$.*

Proof. For a distribution $D \in \Delta(V)$ and subset $A \subset V$, denote $D(A)$ as the measure of A under D. Since P is the transition matrix of a nearest-neighbor-type random walk, then for any set $A \subset V$, and any distribution $D \in \Delta(V)$, with $(P^\top)^\tau D$ representing the distribution of a τ-step random walk according to P initialized with D, we have

$$((P^\top)^\tau D)\,(B(A,\tau)) \geq D(A) \tag{F.1}$$

In particular, for the set $A = \{x \in V : \text{succ}(x,f) \geq p\}$, we have

$$((P^\top)^\tau D)\,(B(\{x \in V : \text{succ}(x,f) \geq p\},\tau)) \geq \mathbb{P}_{x\sim D}[\text{succ}(x,f) \geq p] \tag{F.2}$$

Next, for any set $A' \subset V$ and any distribution $D' \in \Delta(V)$, the cardinality of A' is lower bounded in terms of its measure under D' and the (weighted) two-norm of D'. In particular, we apply Lemma F.1 with $A' = B(\{x \in V : \text{succ}(x,f) \geq p\},\tau)$ and $D' = (P^\top)^\tau D$.

$$\begin{aligned}
|K_\tau^p(f)| &:= |B(\{x \in V : \text{succ}(x,f) \geq p\},\tau)| && \text{(F.3)}\\
&\geq \frac{((P^\top)^\tau D)\,(B(\{x \in V : \text{succ}(x,f) \geq p\},\tau))^2}{\sup_{x\in V}\pi(x)}\frac{1}{||(P^\top)^\tau D||^2_{\ell^2(1/\pi)}} && \text{(F.4)}\\
&\geq \frac{(\mathbb{P}_{x\sim D}[\text{succ}(x,f) \geq p])^2}{\sup_{x\in V}\pi(x)}\text{Div}_\tau^P(D) && \text{(By Equation (F.2))}
\end{aligned}$$

□

Lemma F.1. *For graph $G = (V,E)$, let $\pi : V \to (0,\infty)$. For any distribution $D' \in \Delta(V)$ and any $A' \subset V$, we have*

$$|A'| \geq \frac{(D'(A'))^2}{\sup_{x\in V}\pi(x)}\frac{1}{||D'||^2_{\ell^2(1/\pi))}} \tag{F.5}$$

Proof. For any distribution $D' \in \Delta(V)$, any $\alpha \in [0,1]$, and any $A' \subset V$, we have:

$$\begin{aligned}
(D'(A'))^2 &\leq (\sum_{x\in V}\mathbf{1}_{A'}(x)D'(x))^2 && \text{(F.6)}\\
&\leq [\sum_{x\in V}(\mathbf{1}_{A'}(x)\sqrt{\pi(x)})^2]\cdot[\sum_{x\in V}(D'(x)\frac{1}{\sqrt{\pi(x)}})^2] && \text{(F.7)}
\end{aligned}$$

Thus,

$$|A'| = \sum_{x \in V} \mathbf{1}_{A'}(x) \tag{F.8}$$

$$\geq \frac{\sum_{x \in V} (\mathbf{1}_{A'}(x) \sqrt{\pi(x)})^2}{\sup_{x \in V} \pi(x)} \tag{F.9}$$

$$\geq \frac{(D'(A'))^2}{\sup_{x \in V} \pi(x)} \frac{1}{||D'||^2_{\ell^2(1/\pi))}} \tag{F.10}$$

□

Lemma F.2. *Let $G = (V, E)$ be a finite, undirected, connected, and non-bipartite graph. Let (V, P) be a nearest-neighbor-type Markov chain on G. For every $x \in V$, let $N(x)$ denote its neighbors in G. If $\forall x \in V, \forall y \in N(x), P[x, y] > 0$, then P is irreducible and aperiodic, and there exist constants $C > 0$ and $\alpha \in (0, 1)$ where*

$$\max_{x \in V} ||p^{(n)}(x, \cdot) - \pi||_{\text{TV}} \leq C\alpha^n \tag{F.11}$$

Proof. Because G is connected, then $\forall x \in V, \forall y \in N(x), P[x, y] > 0$, so P is irreducible.

Because G is undirected, $\gcd\{n \in \mathbb{N} : p^{(n)}(x, x) > 0\} \leq 2$. Because G is non-bipartite, $\gcd\{n \in \mathbb{N} : p^{(n)}(x, x) > 0\} \neq 2$. Thus, P is aperiodic.

Because V is finite and the rows of $P \in [0, 1]^{|V| \times |V|}$ sum to 1, then $P\mathbf{1} = \mathbf{1}$, so that P has an eigenvalue of 1. The left eigenvector corresponding to eigenvalue 1 is the stationary distribution π. Since P is irreducible, π is unique.

By Lemma B.3, there exist constants $C > 0$ and $\alpha \in (0, 1)$ where

$$\max_{x \in V} ||p^{(n)}(x, \cdot) - \pi||_{\text{TV}} \leq C\alpha^n \tag{F.12}$$

□

G Figures

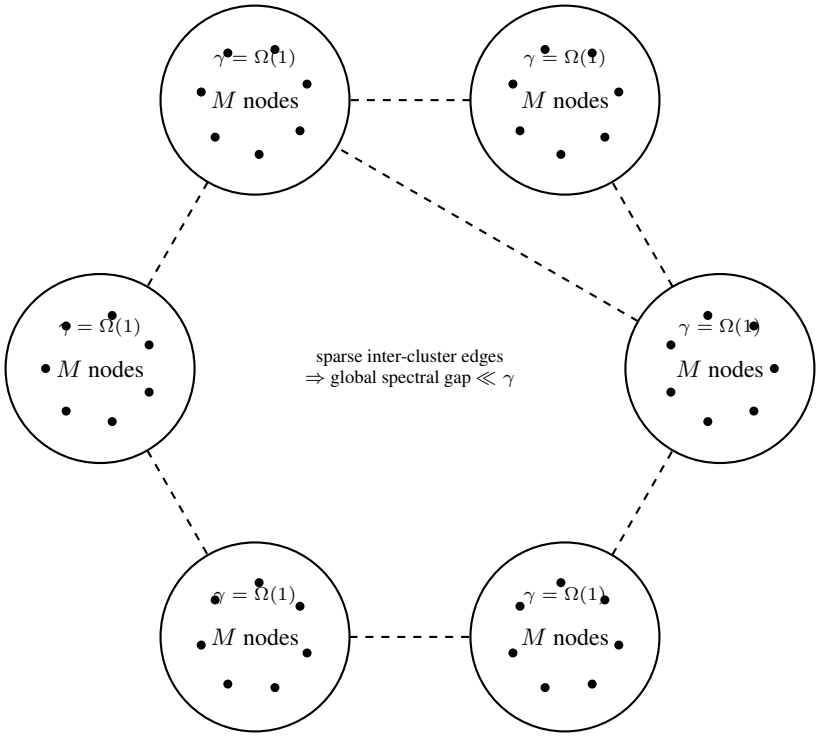

Figure 2: Schematic of a K-cluster graph: each cluster is internally well-connected (spectral gap γ), but only a sparse set of inter-cluster edges connects clusters, yielding a much smaller global spectral gap.

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 describe a theoretical framework for prover-conjecturer self-play algorithms for theorem proving, a baseline conjecturing algorithm and prover-conjecturer system, a definition of a diversity measure, an improved conjecturing algorithm and method to compute diffusion similarity via contrastive learning, which is what the paper delivers.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: The limitations are discussed in Section 6, Conclusion and Limitations. Also, in the main content of the paper itself, the motivation, practicality, and limitations of the assumptions are discussed.

Guidelines:

- The answer [N/A] 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: We carefully lay out our assumptions. Each theorem lists the assumptions it requires. The proofs of each theorem are comprehensive and in the appendix.

   Guidelines:

   - The answer [N/A] 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: [N/A]

Justification: This paper has no experiments

Guidelines:

- The answer [N/A] 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: [N/A]

   Justification: This paper does not have any data or code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [N/A]

   Justification: This paper does not have any experiments.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Justification: This paper does not have any experiments.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Justification: This paper does not have any experiments.

   Guidelines:

   - The answer [N/A] 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: The paper has no human subjects, data, or code. There are no Safety, Security, Discrimination, Surveillance, Deception, Environment, Human Rights, and Bias and Fairness consequences of this research work, as this research does not build a particular technology but rather aims to understand an existing technology: LLMs which prove mathematical theorems and the way in which they are trained.

   Guidelines:

   - The answer [N/A] 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: [N/A]

    Justification: There is no immediate societal impact of the work performed. It is a theoretical paper that aims to understand a scientific technique for training theorem-proving LLMs.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: There are no data or models applicable.

Guidelines:

- The answer [N/A] 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: [N/A]

Justification: The paper does not use existing assets.

Guidelines:

- The answer [N/A] 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: [N/A]

Justification: The paper does not release new assets.

Guidelines:

- The answer [N/A] 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: [N/A]

Justification: We did no crowdsourcing experiments in this work.

Guidelines:

- The answer [N/A] 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: [N/A]

Justification: We performed no studies with human participants.

Guidelines:

- The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.