DRAW THE GRAPH, DRAW THE NEGATIVES: A LOCAL SAMPLING FRAMEWORK FOR CONTRASTIVE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Noise Contrastive Estimation (NCE) provides a powerful framework for learning unnormalized models by contrasting data with negative samples. While classical NCE draws negatives globally from a reference distribution, Conditional NCE (CNCE) instead generates local negatives around each data point. Existing local constructions, however, are largely restricted to binary discrimination and simple negative-sampling schemes. In this work, we propose a general framework for local negative sampling in NCE that extends the binary view to multi-way classification through graphical negative-sampling schemes. We characterize when the geometry of a generation graph hides the origin of the sampling, in which case the proposal terms cancel and the exact Bayes posterior reduces to an energy-only softmax. Our framework recovers an existing method as a hidden-anchor construction and connects graphical negative sampling to contrastive divergence. Building on this framework, we further introduce a new method, Generative Origin Recognition on cyclic Indices Contrastive Estimation (GORI-CE), which generates negatives through a cyclic Gaussian bridge. Experiments on both synthetic and real data demonstrate the effectiveness of our method and highlight negative-sampling geometry as a useful design principle for energy-based learning.

1 INTRODUCTION

Energy-based models (EBMs) provide a flexible framework for representing complex probability distributions (LeCun et al., 2006; Grathwohl et al., 2020; Ou et al., 2022; Gao et al., 2021; Xie et al., 2016). However, their training is complicated by the typically intractable normalizing constant (Song & Kingma, 2021). Noise Contrastive Estimation (NCE) addresses this difficulty by learning the density ratio between the data distribution and a known noise distribution through classification (Gutmann & Hyvärinen, 2010). NCE and related contrastive objectives have been applied across domains including computer vision, natural language processing, and representation learning (Gao et al., 2020; Mnih & Teh, 2012; van den Oord et al., 2018).

A central difficulty in NCE is the choice of the noise distribution. When the data distribution p_d and the proposal q differ substantially, density-ratio estimation can become inefficient or unreliable. (Rhodes et al., 2020; Chehab et al., 2022; Srivastava et al., 2023; Choi et al., 2022). This motivates NCE methods based on local sampling, such as Conditional Noise Contrastive Estimation (CNCE) (Ceylan & Gutmann, 2018). Given a data sample x_0, CNCE draws a negative $x_1 \sim q(\cdot \mid x_0)$ and classifies which of x_0 and x_1 is the origin of the sampling. However, existing local NCE formulations remain largely pairwise, with each classification problem involving only one conditionally generated negative.

Meanwhile, Energy Discrepancy (ED) (Schröder et al., 2023) provides a score-free and MCMC-free energy objective whose Gaussian perturbation scale interpolates between score matching (Hyvärinen, 2005; Lyu, 2009) and negative log-likelihood. This interpolation addresses the near-sightedness of purely local score-based objectives (Wenliang & Kanagawa, 2020; Zhang et al., 2022; Koehler et al., 2023) while retaining an energy-only training criterion. Interestingly, its practical finite-particle estimator generates multiple local candidates through a shared Gaussian perturbation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Although it can be viewed as a negative-sampling scheme for local NCE, this perspective has not been fully developed in the absence of a general framework.

In this paper, we develop a general local sampling framework for NCE based on multi-way origin classification, by separating a local contrastive task into a sampling kernel, a candidate tuple, and a hypothesis construction. However, multiple negative sampling can itself reveal which candidate is the data origin, making the classification task easier independently of the learned energy. This geometric information appears explicitly in the exact contrastive posterior and we identify sufficient conditions under which it disappears. We use a graphical formulation to design symmetric joint sampling kernels that yield an energy-only NCE objective and derive practical Gaussian samplers from such graphs. Building on this principle, we introduce Generative Origin Recognition on cyclic Indices Contrastive Estimation (GORI-CE), which generates negatives through a cyclic Gaussian bridge. We further generalize the connection between local NCE and contrastive divergence (Hinton, 2002; Olmin et al., 2024; Yair & Michaeli, 2021), and show that the cyclic geometry exposes the classifier to both local and distant negatives within a single contrastive pool. Experiments on density estimation and image generation demonstrate the practical effectiveness of the negative-sampling framework.

2 Preliminaries and background

Let $p_d(x)$ be the unknown data distribution on $\mathbb{R}^d$, and consider the energy-based model

$$p_\theta(x) = \tilde{p}_\theta(x)/Z_\theta,$$

where $\tilde{p}_\theta(x)$ is the unnormalized density model and $Z_\theta = \int \tilde{p}_\theta(x)dx$. The natural objective for estimating θ is the negative log-likelihood (NLL), where for a data point $x_0 \sim p_d$, it is given by $-\log p_\theta(x_0) = -\log \tilde{p}_\theta(x_0) + \log Z_\theta$. However, Z_θ is generally intractable, so directly minimizing NLL is often not computationally feasible.

Noise contrastive estimation (NCE) NCE (Gutmann & Hyvärinen, 2010) avoids evaluating Z_θ through binary classification between the data and noise samples from a proposal distribution $q_n(x)$. However, its performance can depend sensitively on $q_n(x)$, which is often difficult to choose in practice (Chehab et al., 2022). Conditional NCE (Ceylan & Gutmann, 2018) instead draws each negative conditionally on the data point. Given $x_0 \sim p_d$ and a local proposal $q(\cdot \mid x_0)$ on $\mathbb{R}^d$, CNCE draws $x_1 \sim q(\cdot \mid x_0)$ and classifies the sampling origin of the pair (x_0, x_1), i.e., whether $x_0 \to x_1$ or $x_1 \to x_0$. Specifically, let the candidate origin label be $K \in \{0, 1\}$. Then under hypothesis $K = k$, the likelihood is given as

$$\pi_k := p(x_0, x_1 \mid K = k) \propto p_d(x_k)q(x_{1-k} \mid x_k), \quad k \in \{0, 1\}.$$

Assuming a uniform prior over K and fixing the observed label to $K = 0$, the CNCE objective is given as the negative log of the softmax posterior $-\log[\pi_0/(\pi_0 + \pi_1)]$. Furthermore, if we use a symmetric kernel such as a Gaussian kernel, the objective simplifies to $-\log[\tilde{p}_\theta(x_0)/(\tilde{p}_\theta(x_0) + \tilde{p}_\theta(x_1))]$.

Energy Discrepancy Recently, Energy Discrepancy (ED) (Schröder et al., 2023) was introduced to address the nearsightedness of score matching objective, by integrating the Gaussian-perturbed score matching loss over perturbation times up to t. It can be shown that the population ED objective with a Gaussian kernel r_t is given by

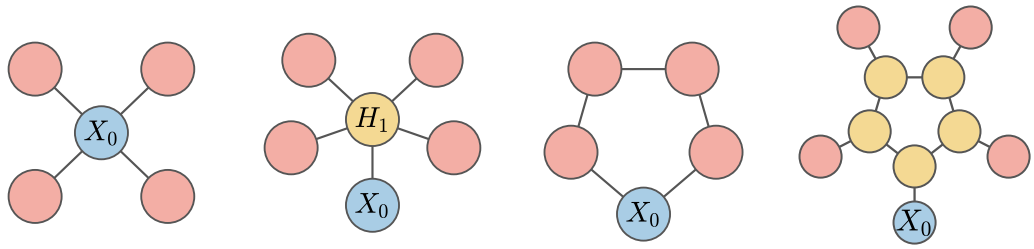
$$ED_{r_t}(\theta) = \mathbb{E}_{p_d(x)}[U_\theta(X)] - \mathbb{E}_{p_d(x)r_t(y|x)}\left[U_\theta^{(t)}(Y)\right],$$

where $U_\theta(x)$ is the energy function and $U_\theta^{(t)}(y) := -\log \int e^{-U_\theta(x)} r_t(y \mid x)dx$ is the energy function after Gaussian perturbation with scale t.

Given the data point $x_0 \sim p_d$, Schröder et al. (2023) approximate the population ED objective with the following finite-particle estimator:

$$\mathcal{L}_{r_t,J}(\theta \mid x_0) := \log\left(\frac{1}{J} + \frac{1}{J}\sum_{j=1}^{J} \exp\Big(U_\theta(x_0) - U_\theta\big(x_0 + \sqrt{t}\,\xi + \sqrt{t}\,\xi_j'\big)\Big)\right), \tag{1}$$

where $\xi, \xi_1', \ldots, \xi_J'$ are drawn independently from $\mathcal{N}(0, I)$. When $J = 1$, the above finite-particle ED objective in Eq. (1) reduces to the CNCE objective with a Gaussian kernel at scale $2t$.



(a) Direct star (b) Anchor (ED) (c) Cycle (GORI) (d) Cyclic anchor

Figure 1: The data point is blue, negatives are red, and auxiliary variables are yellow. Edges specify pairwise Gaussian factors. Exchanging candidate positions preserves the anchor graph; rotating the cycle, or the anchors and their attached candidates together, preserves the cyclic anchor. The direct sampler instead reveals its origin through the sampling geometry.

3 General framework for local sampling in NCE

3.1 Three components of the local NCE framework

In this subsection, we propose a general framework for local sampling in NCE that allows for extending a single negative such as CNCE to multiple negative samples. This framework consists of three components: a sampling kernel, negative samples drawn from the kernel, and a hypothesis construction over the sampling origin.

Sampling kernel. We sample J negatives from a joint sampling density denoted by $Q(X_1, \dots, X_J \mid X_0)$. For simplicity, throughout this paper we restrict Q to distributions proportional to products of pairwise Gaussian kernels r_t, all sharing a common scale t. When $J = 1$, Q reduces to a single symmetric Gaussian kernel $r_t(X_1 \mid X_0)$, recovering CNCE.

Negative samples. Given a data sample X_0, we draw a tuple of J negative samples $(X_1, \dots, X_J)$ jointly from $Q(\cdot \mid X_0)$. However, depending on Q, the relative geometry of the samples may reveal which candidate is the data sample. For instance, if the negative samples are drawn conditionally independently from Gaussian distributions centered at X_0, they tend to surround X_0, making X_0 likely to lie near the center of the candidate pool.

Hypotheses. In binary CNCE, the two hypotheses on the sampling origin—whether the sampling direction is $X_0 \to X_1$ or $X_1 \to X_0$—correspond to the two possible orderings of the pair, (X_0, X_1) and (X_1, X_0). For multiple negative samples, we similarly construct hypotheses by reordering the samples $\mathbf{X} = (X_0, \dots, X_J)$. Specifically, let K be the candidate origin label. Given $\mathbf{X}$, we apply a permutation $\sigma : \{0, \dots, J\} \to \{0, \dots, J\}$ to obtain $(X_{\sigma(0)}, \dots, X_{\sigma(J)})$. This ordering represents the hypothesis $K = \sigma(0)$, i.e., $X_{\sigma(0)}$ is drawn from the data distribution and the remaining samples $(X_{\sigma(1)}, \dots, X_{\sigma(J)})$ are drawn jointly from $Q(\cdot \mid X_{\sigma(0)})$.

Now let $\Sigma' := \{\sigma'_j\}_{j=0}^J$ be the collection of $J+1$ permutations where σ'_0 is the identity and σ'_j is an arbitrary permutation that sends 0 to j for each $j \in \{1, \dots, J\}$. Let K denote the candidate-origin label, assuming a uniform label prior. Then the permutation σ'_k corresponds to hypothesis $K = k$ with the likelihood given by

$$\pi_k(\mathbf{X}) = p_d(X_{\sigma'_k(0)}) Q(X_{\sigma'_k(1)}, \dots, X_{\sigma'_k(J)} \mid X_{\sigma'_k(0)}). \tag{2}$$

Therefore, the induced NCE objective can be written as the negative log-softmax posterior:

$$\mathcal{L}_{NCE}(\theta \mid X_0) = -\log \frac{\pi_0^\theta(\mathbf{X})}{\sum_{\ell=0}^J \pi_\ell^\theta(\mathbf{X})}, \tag{3}$$

where $\pi_k^\theta(\mathbf{X}) = \tilde{p}_\theta(X_k) Q(X_{\sigma'_k(1)}, \dots, X_{\sigma'_k(J)} \mid X_{\sigma'_k(0)})$ is a model likelihood.

3.2 SYMMETRIC SAMPLING KERNEL

Recall that a posterior under the hypothesis $K = 0$ of the binary CNCE is proportional to $p_d(x_0)q(x_1 \mid x_0)$. However, by using a symmetric kernel, the kernel term $q(x_1 \mid x_0)$ cancels in the posterior, leaving an energy-only objective. Similarly, if the joint sampling kernel Q is invariant under some permutations σ_j with $\sigma_j(0) = j$ for each j, i.e., $Q(X_{\sigma_j(1)}, \ldots, X_{\sigma_j(J)} \mid X_{\sigma_j(0)}) = Q(X_1, \ldots, X_J \mid X_0)$ for each j, the kernel terms cancel out, and the induced objective in Eq. (3) remains energy-only. We refer to this property as *symmetry of the joint sampling kernel*.

Definition 1 (Symmetric sampling and natural permutation)**.** Q is symmetric if there exist permutations $\sigma_0 = \mathrm{id}, \sigma_1, \ldots, \sigma_J$, with $\sigma_j(0) = j$, such that

$$Q(X_{\sigma_j(1)}, \ldots, X_{\sigma_j(J)} \mid X_{\sigma_j(0)}) = C(\mathbf{X})$$

for every $j \in \{0, \ldots, J\}$, where $C(\mathbf{X})$ depends only on the sample pool $\mathbf{X} = (X_0, \ldots, X_J)$, not hypothesis index j. For a symmetric Q, a permutation set $\Sigma := \{\sigma_0, \ldots, \sigma_J\}$ satisfying the condition above is called a natural permutation set.

If a joint sampling kernel Q is symmetric, there exists a natural hypothesis set Σ under which the kernel terms are identical across hypotheses, so that the induced NCE objective reduces to an energy-only objective:

$$\mathcal{L}_{NCE}(\theta \mid X_0) = -\log \frac{\tilde{p}_\theta(X_0)}{\sum_{\ell=0}^{J} \tilde{p}_\theta(X_\ell)}. \tag{4}$$

We provide below two concrete examples to illustrate the advantages of symmetric sampling kernels and the geometric cost incurred by non-symmetric kernels.

Anchoring Consider the sampling kernel $Q^{(t)}_{anchor}(\mathbf{X}' \mid X)$: First sample an anchor variable $H_1 \sim r_t(\cdot \mid X)$. Given H_1, draw $X_j \overset{\text{iid}}{\sim} r_t(\cdot \mid H_1)$ for $j = 1, \ldots, J$. Equivalently, $H_1 = X + \sqrt{t}\,\xi$ and $X_j = H_1 + \sqrt{t}\,\xi'_j = X + \sqrt{t}\,\xi + \sqrt{t}\,\xi'_j$. Then the anchoring procedure has a density proportional to

$$Q^{(t)}_{anchor}(X_1, \ldots, X_J \mid X_0) \;\propto\; \int r_t(h \mid X_0) \prod_{\ell=1}^{J} r_t(X_\ell \mid h)\, \mathrm{d}h = \int \prod_{\ell=0}^{J} r_t(X_\ell \mid h)\, \mathrm{d}h, \tag{5}$$

since the Gaussian kernel is symmetric. Therefore the anchoring kernel is permutation invariant and the induced objective is given by Eq. (4). Note that this is equivalent to the finite-particle ED estimator in Eq. (1) up to an additive constant. This shows that the finite-particle ED objective can be viewed as a special case of our local sampling framework, in which the sampling kernel satisfies the condition.

Direct star In contrast, consider the simple sampling kernel shown in Figure 1(a), the direct-star construction, in which all negatives are drawn independently from the observed data point X_0:

$$Q^{(t)}_{dir}(X_1, \ldots, X_J \mid X_0) := \prod_{j=1}^{J} r_t(X_j \mid X_0).$$

For $J \geq 2$, since any permutation exchanging the conditioning variable changes the sampling kernel term, the negative sampling kernel is not symmetric. In particular, the objective from direct star generation is as Eq. (3) with the unnormalized weight π^θ_ℓ given by

$$\pi^\theta_\ell(X) \;=\; \exp\left[-U_\theta(X_\ell) - \frac{J+1}{2t}\|X_\ell - \bar{X}\|^2\right]. \tag{6}$$

The derivation is in Appendix A.1. Observe that the posterior increases as X_ℓ approaches the empirical centroid $\bar{X}$. That is, the sampling geometry leaks information about the origin of the pool: a candidate near the center is more likely to be the data point even before considering its energy. Thus, the model can exploit this geometric shortcut rather than using the energy function to distinguish candidates by their density. Although the direct-star construction reduces to binary CNCE when $J = 1$ due to the symmetry of the Gaussian kernel, for $J > 1$ the geometry of the candidate pool can reveal the identity of the data point.

4 SYMMETRY AS A NEGATIVE-SAMPLING PRINCIPLE

As the direct-star example above illustrates, the geometry of the negative samples may reveal which candidate is the data point, making the contrastive task less informative for density estimation. We therefore seek a generic scheme for constructing symmetric negative-sampling kernels. To this end, we associate a negative-sampling kernel with a graph.

4.1 DESIGNING A SAMPLING KERNEL FROM A GRAPH

Let $G = (V, E)$ be a finite connected undirected graph, and let $S = \{X_0, \dots, X_J\} \subseteq V$ be the set of candidate vertices. In the actual sampling procedure, X_0 is the data point and $X_1, \dots, X_J$ are the negative samples. The remaining vertices $H := V \setminus S = \{H_1, \dots, H_I\}$, where $I = |H|$, are anchor vertices, which are latent variables used to generate the candidates. We call the pair (G, S) a generation graph. In the setting of direct-star, all vertices are candidates, so $H = \varnothing$, while in the anchoring case, the candidates share one latent anchor, so $H = \{H_1\}$.

For each edge $\{i, j\} \in E$, we associate the symmetric Gaussian potential $r_t(V_j \mid V_i) \propto \exp(-\|V_i - V_j\|^2/(2t))$. The product of these edge potentials defines an unnormalized Gaussian potential over all vertices. Let $L_G = D_G - A_G$ be the graph Laplacian, where A_G and D_G are the adjacency and degree matrices of G. Since

$$\sum_{\{i,j\}\in E} \|V_i - V_j\|^2 = \mathbf{V}^\top (L_G \otimes \mathrm{Id}_d)\mathbf{V},$$

where $\mathbf{V} = (V_0^\top, \dots, V_{J+I}^\top)^\top$, the Gaussian potential is characterized by L_G. To obtain a sampling kernel for the candidates, we marginalize the anchor vertices and fix the data vertex $X_0 = x$. Write the graph Laplacian according to $V = S \sqcup H$ as

$$L_G = \begin{pmatrix} L_{SS} & L_{SH} \\ L_{HS} & L_{HH} \end{pmatrix},$$

and define the Schur complement $\widetilde{L}_G := L_{SS} - L_{SH} L_{HH}^{-1} L_{HS}$. If $H = \varnothing$, we simply set $\widetilde{L}_G = L_G$. Let $\widetilde{L}_G^{(0)}$ denote the matrix obtained by removing the row and column corresponding to X_0. Since $\widetilde{L}_G \mathbf{1} = 0$, fixing $X_0 = x$ gives the Gaussian sampling kernel

$$Q_G(x_{1:J} \mid x_0) = \mathcal{N}\Big(x_{1:J}; \mathbf{1}_J \otimes x_0,\, t\big(\widetilde{L}_G^{(0)}\big)^{-1} \otimes \mathrm{Id}_d\Big).$$

Thus, once a generation graph is specified, its Laplacian determines the joint distribution of the negative samples. The derivation is given in Appendices A.2 and B.1.

4.2 GENERATION GRAPH SYMMETRY IMPLIES KERNEL SYMMETRY

We now define symmetry of a generation graph, which provides a sufficient condition for symmetry of the induced kernel. For notational convenience, we index the vertices of G as $V_j := X_j$ for $j = 0, \dots, J$ and $V_{J+i} := H_i$ for $i = 1, \dots, I$. Thus, $S = \{V_0, \dots, V_J\}$ and $H = \{V_{J+1}, \dots, V_{J+I}\}$.

Definition 2 (Symmetric generation graph)**.** With the notation above, a generation graph pair (G, S) is *symmetric* if, for every $j \in \{1, \dots, J\}$, there exists a permutation $\tilde{\sigma}_j : \{0, \dots, J+I\} \to \{0, \dots, J+I\}$ with $\tilde{\sigma}_j(0) = j$ such that

$$\tilde{\sigma}_j(\{0, \dots, J\}) = \{0, \dots, J\}, \qquad \{u, v\} \in E \Leftrightarrow \{\tilde{\sigma}_j(u), \tilde{\sigma}_j(v)\} \in E.$$

That is, for every j, there exists a permutation $\tilde{\sigma}_j$ with $\tilde{\sigma}_j(0) = j$, preserving both the candidate set S and the graph structure. A permutation $\tilde{\sigma}_j$ satisfying the above condition is called *natural* and the collection $\tilde{\Sigma} := \{\tilde{\sigma}_j\}_{j=0}^J$, with $\tilde{\sigma}_0 = id$, is called a *natural permutation set* of (G, S).

The symmetry condition for generation graph is often easy to verify visually. A natural permutation may rotate or rearrange the vertices, but it must leave their adjacency relations unchanged. Thus, if X_0 can be mapped to every candidate X_j in this way while preserving S, then (G, S) satisfies the symmetric-generation condition.

The theorem below further shows that symmetry of the generation graph implies symmetry of the induced sampling kernel, thereby yielding the energy-only NCE objective in Eq. (4).

Theorem 3 (Graph symmetry induces kernel symmetry; informal)**.** *If the generation graph (G, S) is symmetric, then the induced kernel Q_G is symmetric under the corresponding natural permutations. Consequently, the resulting objective reduces to the energy-only form in Eq (4). If $p_d > 0$ almost everywhere and the model is well specified, the resulting NCE task is Fisher-consistent.*

The formal statement and proof are given in Theorem 6. Beyond this general construction, some generation graphs admit simple sampling obtained by composing elementary sampling primitives. For instance, combining cycle and anchoring yields the cyclic anchor graph in Figure 1.

4.3 Connection to Contrastive Divergence

Contrastive Divergence (CD) (Hinton, 2002) approximates the model expectation in the maximum-likelihood gradient using samples from a p_θ-invariant Markov kernel. Olmin et al. (2024) recently showed that RNCE and CNCE can also be interpreted as CD-1 methods: for each method, they construct a p_θ-invariant Markov kernel whose Rao–Blackwellized one-step update matches the gradient of the corresponding NCE objective. We show a similar connection for our symmetric local sampling scheme by constructing, from the jointly generated candidate pool, a p_θ-reversible multi-proposal kernel whose Rao–Blackwellized CD-1 update matches the gradient of our NCE objective.

Theorem 4 (Connection to Contrastive Divergence; informal)**.** *Given a pool $X_{1:J} \sim Q_G(\cdot \mid X_0)$, resampling one candidate according to*

$$w_{\theta,k}(\mathbf{X}) = \frac{p_\theta(X_k)}{\sum_{\ell=0}^{J} p_\theta(X_\ell)}$$

induces a p_θ-reversible Markov kernel $K_{G,\theta}$. Moreover, the gradient of Eq. (4) is the corresponding Rao–Blackwellized CD-1 update. Thus, a symmetric generation graph specifies not only a valid NCE task, but also an associated reversible multiproposal CD kernel.

The formal statement and proof are given in Appendix C. Interestingly, similar ideas have appeared in the sampling literature, where multiproposal MCMC methods generate several candidates jointly and then select one while preserving reversibility (Tjelmeland, 2004; Calderhead, 2014). In our case, this structure arises naturally from the root-hiding symmetry used to obtain an energy-only NCE objective.

5 GORI-CE: Generative Origin Recognition on cyclic Indices Contrastive Estimation

The graph construction allows us to replace the ED anchor while retaining the energy-only loss. We choose a cycle as a simple sparse graph in which all candidates have the same structural role and no auxiliary variables are needed. We call the resulting estimator GORI-CE. Its implementation requires only Gaussian draws, centering, and cumulative sums.

5.1 The cyclic construction with Gaussian bridge

For $J \geq 2$, connect $X_0, X_1, \ldots, X_J$ in a cycle and write $X_{J+1} = X_0$. Using r_t as the edge factor gives

$$Q_{cycle}(X_1, \ldots, X_J \mid X_0) \propto \prod_{r=0}^{J} r_t(X_{r+1} \mid X_r) \propto \exp\left[-\frac{1}{2t}\sum_{r=0}^{J} \|X_{r+1} - X_r\|^2\right].$$

To sample this distribution, we use a Gaussian bridge sampling. More precisely, draw $Z_0, \ldots, Z_J \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$ and center the increments as $\Delta_r = \sqrt{t}\,(Z_r - \bar{Z})$. The resulting increments have the joint Gaussian law conditioned on $\sum_{r=0}^{J} \Delta_r = 0$, so their cumulative sums sample Q_{cycle} exactly(Algorithm 1). By Theorem 3, the resulting negatives can be used directly in Eq (4) and give a consistent objective under its assumptions. The covariance calculation is given in Appendix A.3.

Algorithm 1 uses $O(Jd)$ arithmetic for sampling and $J + 1$ energy evaluations, the same number of energy evaluations as finite-particle ED. Each edge has Gaussian factor variance t, while centering gives each bridge increment marginal covariance $\frac{J}{J+1} t I_d$. Thus, the marginal variance of an individual bridge step is slightly smaller than the edge-factor variance t.

5.2 Comparison at different distances

The anchor and cycle induce different marginal displacement scales from X_0. For the anchor, every negative has the same marginal displacement covariance, $\mathrm{Cov}(X_j - X_0 \mid X_0) = 2tI_d$. For the cycle, it varies with the cyclic index r from X_0 as $\mathrm{Cov}(X_r - X_0 \mid X_0) = tr(J+1-r)/(J+1)I_d$. Thus, at fixed t, adjacent candidates have per-coordinate Root-Mean-Square (RMS) displacement of order $\sqrt{t}$, whereas candidates near the opposite side reach order $\sqrt{Jt}$. These are marginal distance scales; the opposite candidate need not be the farthest in every draw. In contrast, increasing J in the anchor construction only adds negatives at the same marginal displacement scale. The derivation is given in Appendix A.3.

In high dimensions, independent sampling from fixed-scale isotropic Gaussian kernel causes samples to concentrate around a common radius, a phenomenon known as thin-shell concentration (Saremi & Hyvärinen, 2019). As a result, NCE with locally sampled negatives may contrast the data only against negatives lying within a narrow shell region. GORI-CE alleviates this problem by producing samples with varying RMS displacements, whereas CNCE and ED, for which each negative has the same marginal displacement scale from X_0, can suffer from this effect. A detailed analysis is provided in Appendix D.

Algorithm 1: GORI-CE pool and loss for one data point

Input: data point x, negatives J, edge scale t, energy U_θ

$m \leftarrow J + 1$;

Draw $Z_0, \ldots, Z_{m-1} \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$ and set $\bar{Z} \leftarrow \frac{1}{m}\sum_r Z_r$;

$X_0 \leftarrow x$; $\quad \Delta_r \leftarrow \sqrt{t}\,(Z_r - \bar{Z})$;

for $r = 1, \ldots, J$ **do**

  $X_r \leftarrow X_{r-1} + \Delta_{r-1}$;

return $U_\theta(X_0) + \log \sum_{k=0}^{J} \exp[-U_\theta(X_k)]$;

Cyclic sampling with auxiliary variables. The same construction can be applied to auxiliary variables. Attach each candidate X_j to an anchor H_j and connect $H_0, \ldots, H_J$ in a cycle, as in Figure 1 (d). Rotating the anchors and their attached candidates together preserves the graph, so Theorem 3 applies. This extends the ED construction by replacing its single shared anchor with correlated anchors. Appendix B.2 verifies this factorization. The loss is again given by Eq. (4). Auxiliary variables therefore provide another way to change the dependence among negatives while preserving the energy-only classification formula. We call this sampling scheme *cyclic anchoring* and provide the numerical performance of cyclic anchoring in Section 6.

6 Experiments

In this section, we evaluate the effect of negative-sampling geometry in three settings: synthetic density estimation, real-world tabular density estimation, and latent-space image generation. The synthetic experiment showcases the geometric shortcut and multi-scale behavior, while the real-data experiments compare the resulting estimators under matched candidate budgets and RMS displacement.

To compare the two sampling schemes at a common scale, we define v^2 by averaging the squared distances from X_0 per negative and per coordinate:

$$v^2 := \frac{1}{Jd}\sum_{j=1}^{J} \mathbb{E}[\|X_j - X_0\|^2 \,|\, X_0],$$

For the anchor construction, $v^2_{\text{anc}} = 2t$, whereas for the cycle, $v^2_{\text{cyc}} = (J+2)t/6$. Therefore, matching the two schemes at a common RMS displacement v requires $t_{\text{anc}} = v^2/2$ for anchoring and $t_{\text{cyc}} = 6v^2/(J+2)$ for the cycle. The derivation is given in Appendix A.3.

6.1 Density estimation on synthetic datasets

We begin our numerical study with experiments on three synthetic datasets: (a) Uniform 25-Gaussians; (b) Tiered-width 25-Gaussians, in which the component standard deviation decreases with distance from the center; and (c) a 6D uniform hypercube mixture. We compare Direct star,

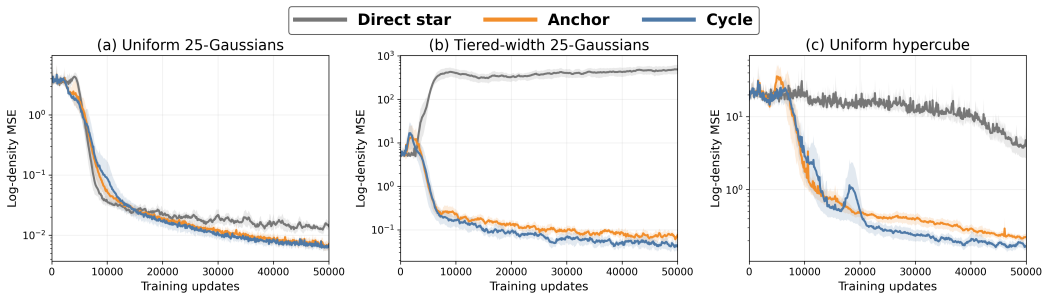


Figure 2: (a) Uniform 25-Gaussian; (b) Tiered width 25-Gaussian; (c) Uniform hypercube. Log-density MSE lines and bands show ten-seed means and 95% nonparametric bootstrap confidence intervals; $J = 16$, $v = 1$.

Table 1: Held-out test log-likelihood for tabular density-estimation benchmarks and BSDS300. Seven-seed mean $\pm$ standard deviation; higher is better.

Method	POWER	GAS	HEPMASS	MINIBOONE	BSDS300
DIRECT STAR	0.17 ± 0.03	3.19 ± 0.28	-34.49 ± 0.56	-74.57 ± 0.15	43.35 ± 0.19
ANCHOR (ED)	-0.01 ± 1.11	4.61 ± 0.34	-19.42 ± 0.04	-41.92 ± 1.02	66.17 ± 1.65
CYCLE (GORI)	$\mathbf{0.49 \pm 0.03}$	$\mathbf{5.76 \pm 0.15}$	$\mathbf{-18.45 \pm 0.04}$	$\mathbf{-33.81 \pm 2.23}$	$\mathbf{86.84 \pm 1.17}$
CA 2×10	0.45 ± 0.03	2.31 ± 6.71	-19.35 ± 0.06	-40.99 ± 0.95	66.60 ± 0.69
CNCE	0.45 ± 0.00	4.65 ± 0.10	-19.43 ± 0.06	-41.04 ± 1.14	65.02 ± 1.20

Anchor (ED), and Cycle (GORI) using $J = 16$ negatives and RMS $v = 1$, with 10 random seeds. Figure 2 first illustrates the geometric cost. The poorer performance of Direct star is consistent with the geometric proposal logit weakening the energy-based learning signal. On the other hand, the result on tiered-width 25-Gaussians highlights robustness to a target distribution with heterogeneous mode widths. In the 6D hypercube experiment, this behavior is consistent with the greater fraction of short-range candidates predicted by Section 5.2.

6.2 Density estimation on real datasets

We further evaluate density estimation on POWER, GAS, HEPMASS, MINIBOONE, and BSDS300, following the data splits and preprocessing of (Papamakarios et al., 2017). For BSDS300, we additionally standardize each coordinate and report log-likelihoods in the original coordinates using the exact change-of-variables correction. We compare Direct star, Anchor (ED), Cycle (GORI), CA 2×10 (cyclic anchoring with two anchors and ten candidates per anchor), and binary CNCE. The candidate pool contains 20 samples. Table 1 shows that GORI-CE achieves the highest mean log-likelihood across all five datasets in this setting. In particular, the larger gains on MINIBOONE and BSDS300 are consistent with the advantage of representing multiple distance scales within a single contrastive pool. Appendix E.2 provides the full experimental setup.

6.3 Image generation

We also assess the performance of our framework on SVHN, CIFAR-10, and CelebA image generation tasks. In this experiment, we adopt the latent-space energy-based prior framework of (Pang et al., 2020), following the architecture and principal training settings of (Schröder et al., 2023). We use $J = 19$ negative samples. The generator and latent EBM are trained jointly, with posterior inference performed using Langevin dynamics. We compare prior-learning objectives using a common candidate budget and

Table 2: FID on SVHN, CIFAR-10, and CelebA. Three-seed mean $\pm$ standard deviation. Lower is better.

Method	SVHN	CIFAR-10	CelebA
DIRECT STAR	34.34 ± 1.45	105.02 ± 20.14	138.52 ± 1.80
ANCHOR (ED)	32.89 ± 2.09	77.63 ± 1.26	47.90 ± 6.28
CYCLE (GORI)	31.69 ± 1.32	$\mathbf{76.07 \pm 0.76}$	52.02 ± 9.52
CA 2×10	$\mathbf{31.12 \pm 1.98}$	76.79 ± 3.50	$\mathbf{46.72 \pm 5.01}$
CNCE	33.83 ± 0.66	88.28 ± 21.86	47.47 ± 4.28

matched root-to-negative RMS displacement. Overall, these results suggest that incorporating structured negative-sampling geometry can improve image generation performance, although the relative advantage between cyclic and anchoring variants depends on the dataset.

7 Related work

Training energy-based models. Training energy-based models is challenging because maximum-likelihood learning requires expectations under an often intractable model distribution (LeCun et al., 2006; Song & Kingma, 2021). A large class of methods therefore relies on short-run MCMC or contrastive divergence (Hinton, 2002; Xie et al., 2016; Nijkamp et al., 2019; Du & Mordatch, 2019), while subsequent work has studied alternative objectives and improvements to contrastive-divergence training (Sutskever & Tieleman, 2010; Yu et al., 2020; Du et al., 2021).

NCE, RNCE, and their extensions. Noise-Contrastive Estimation (NCE) turns the estimation of an unnormalized model into a discrimination problem between data and samples (Gutmann & Hyvärinen, 2010). Subsequent work has studied Bregman-divergence generalizations, asymptotic efficiency, and the choice of the noise distribution (Gutmann & Hirayama, 2011; Uehara et al., 2018; Chehab et al., 2022; Goodfellow, 2014; Xu, 2022). Ranking NCE (RNCE) extends the binary task to multi-way candidate ranking (Jozefowicz et al., 2016; Ma & Collins, 2018), while adaptive schemes such as Flow Contrastive Estimation learn expressive noise distributions jointly with the EBM (Gao et al., 2020). Beyond density estimation, NCE-style classification has also influenced contrastive representation-learning objectives such as InfoNCE (van den Oord et al., 2018; Chen et al., 2020; He et al., 2020). Recently, Olmin et al. (2024) showed that both RNCE and CNCE can be interpreted as instances of contrastive divergence.

NCE with local sampling. Conditional NCE (CNCE) (Ceylan & Gutmann, 2018) avoids a separately specified global noise distribution by drawing a negative from a data-dependent kernel; with local Gaussian perturbations, score matching arises as a limiting case (Ceylan & Gutmann, 2018; Hyvärinen, 2005). Gaussian perturbations also underlie denoising and multi-noise-scale score-matching methods (Vincent, 2011; Song & Ermon, 2019). Recent work extends local sampling through mode jumping and joint spatiotemporal perturbations (Yu et al., 2025; 2026). Energy Discrepancy (ED) independently constructs a local Gaussian candidate pool, and we show that its finite-particle estimator with $w = 1$ admits a multi-way NCE interpretation (Schröder et al., 2023). Relatedly, Yair & Michaeli (2021) interpret contrastive divergence as a classification problem over the forward and time-reversed directions of a Markov chain, while Olmin et al. (2024) show that CNCE can be viewed as a Rao–Blackwellized CD-1 method. Our work extends this kernel-centric view from pairwise transitions to the joint geometry of an entire negative pool.

8 Conclusion

In this work, we introduced a general framework for local NCE that separates a contrastive task into a sampling kernel, a sample tuple, and a hypothesis construction. This framework exposes task-irrelevant geometric information in the contrastive posterior and characterizes the symmetry condition under which it cancels out. We then developed a graph-based principle for constructing such kernels and introduced GORI-CE, which generates contrastive pools through a cyclic Gaussian bridge while retaining an energy-only NCE objective. Our analysis shows that this geometry provides contrast across multiple distance scales. Experiments on synthetic, tabular, and image data demonstrate its practical effectiveness.

Several limitations also suggest directions for future work. First, our current graph construction is restricted to undirected graphs with uniformly weighted Gaussian edge factors, leaving weighted graphs, non-Gaussian kernels, and more general asymmetric constructions unexplored. Second, we focus on finite J and do not characterize the population objective arising in the limit $J \to \infty$. Such a limit may provide a clearer connection between negative-sampling geometry and maximum-likelihood learning. In particular, characterizing this limit may further clarify the relation between multiproposal local contrast, maximum-likelihood learning, and longer-run CD.

AI USE STATEMENT

LLMs were used as an interactive research and writing assistant throughout this work. They were used to brainstorm and refine conceptual formulations, assist with mathematical derivations and proof development, discuss experimental design and the interpretation of experimental diagnostics, assist with code writing and debugging, and revise the manuscript for clarity, structure, and grammar.

We did not use generative Al for translation. The literature review was led by the authors, with generative AI used interactively to help identify additional potentially relevant literature. All cited works were checked by the authors against the original sources. Generative AI was also used for drafting, editing, and polishing portions of the manuscript for clarity and readability.

All AI-assisted mathematical arguments, code, interpretations of experimental results, and manuscript text were reviewed and modified by the authors as necessary. The experiments were executed by the authors, and the authors made the final decisions regarding the methodology, theoretical claims, interpretation of results, and presentation. The authors take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

We provide complete proofs of the theoretical results in the appendices and describe the experimental settings, architectures, preprocessing procedures, and evaluation protocols in Appendix E. The supplementary material includes anonymized source code and scripts for reproducing the synthetic, density-estimation, and image-generation experiments reported in the paper.

REFERENCES

Ben Calderhead. A general construction for parallelizing metropolis- hastings algorithms. *Proceedings of the National Academy of Sciences*, 111(49):17408–17413, 2014.

Ciwan Ceylan and Michael U Gutmann. Conditional noise-contrastive estimation of unnormalised models. In *International Conference on Machine Learning*, pp. 726–734. PMLR, 2018.

Omar Chehab, Alexandre Gramfort, and Aapo Hyvärinen. The optimal noise in noise-contrastive learning is not what you think. In *Uncertainty in Artificial Intelligence*, pp. 307–316. PMLR, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Kristy Choi, Chenlin Meng, Yang Song, and Stefano Ermon. Density ratio estimation via infinitesimal classification. In *International Conference on Artificial Intelligence and Statistics*, pp. 2552–2573. PMLR, 2022.

Florian Dorfler and Francesco Bullo. Kron reduction of graphs with applications to electrical networks. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 60(1):150–163, 2013.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.

Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy-based models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2837–2848. PMLR, 2021.

Ruiqi Gao, Erik Nijkamp, Diederik P. Kingma, Zhen Xu, Andrew M. Dai, and Ying Nian Wu. Flow contrastive estimation of energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7518–7528, 2020.

Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P. Kingma. Learning energy-based models by diffusion recovery likelihood. In *International Conference on Learning Representations*, 2021.

Ian J Goodfellow. On distinguishability criteria for estimating generative models. *arXiv preprint arXiv:1412.6515*, 2014.

Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=Hkxzx0NtDB`.

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

Michael U. Gutmann and Jun-ichiro Hirayama. Bregman divergence as general framework to estimate unnormalized statistical models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 283–290, 2011.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9726–9735. IEEE, 2020.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. In *The Eleventh International Conference on Learning Representations*, 2023.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Siwei Lyu. Interpretation and generalization of score matching. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pp. 359–366, 2009.

Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3698–3707, 2018.

Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. In *International Conference on Machine Learning*, 2012. URL https://api.semanticscholar.org/CorpusID:6633369.

Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. *Advances in Neural Information Processing Systems*, 32, 2019.

Amanda Olmin, Jakob Lindqvist, Lennart Svensson, and Fredrik Lindsten. On the connection between noise-contrastive estimation and contrastive divergence. In *International Conference on Artificial Intelligence and Statistics*, pp. 3016–3024. PMLR, 2024.

Zijing Ou, Tingyang Xu, Qinliang Su, Yingzhen Li, Peilin Zhao, and Yatao Bian. Learning neural set functions under the optimal subset oracle. *Advances in Neural Information Processing Systems*, 35:35021–35034, 2022.

Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. Learning latent space energy-based prior model. *Advances in Neural Information Processing Systems*, 33:21994–22008, 2020.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

Benjamin Rhodes, Kai Xu, and Michael U. Gutmann. Telescoping density-ratio estimation. *Advances in Neural Information Processing Systems*, 33, 2020.

Saeed Saremi and Aapo Hyvärinen. Neural empirical bayes. *Journal of Machine Learning Research*, 20(181):1–23, 2019. URL http://jmlr.org/papers/v20/19-216.html.

Tobias Schröder, Zijing Ou, Jen Lim, Yingzhen Li, Sebastian Vollmer, and Andrew Duncan. Energy discrepancies: a score-independent loss for energy-based models. *Advances in Neural Information Processing Systems*, 36:45300–45338, 2023.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

Akash Srivastava, Seungwook Han, Kai Xu, Benjamin Rhodes, and Michael U. Gutmann. Estimating the density ratio between distributions with high discrepancy using multinomial logistic regression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Ilya Sutskever and Tijmen Tieleman. On the convergence properties of contrastive divergence. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 789–795. JMLR Workshop and Conference Proceedings, 2010.

Hakon Tjelmeland. Using all metropolis–hastings proposals to estimate mean values. *Statistics*, 4, 2004.

Masatoshi Uehara, Takeru Matsuda, and Fumiyasu Komaki. Analysis of noise contrastive estimation from the perspective of asymptotic variance, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Li K Wenliang and Heishiro Kanagawa. Blindness of score-based methods to isolated components and mixing proportions. *arXiv preprint arXiv:2008.10087*, 2020.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *International conference on machine learning*, pp. 2635–2644. PMLR, 2016.

Nathaniel Xu. Self-adapting noise-contrastive estimation for energy-based models. *arXiv preprint arXiv:2211.02650*, 2022.

Omer Yair and Tomer Michaeli. Contrastive divergence learning is a time reversal adversarial game. In *International Conference on Learning Representations*, 2021.

Hanlin Yu, Michael U. Gutmann, Arto Klami, and Omar Chehab. Conditional noise-contrastive estimation of energy-based models by jumping between modes. In *EurIPS 2025 Workshop on Principles of Generative Modeling (PriGM)*, 2025.

Hanlin Yu, RuiKang OuYang, Partha Kaushik, Arto Klami, Michael U. Gutmann, and Omar Chehab. Learning energy-based models from stochastic interpolants using spatiotemporal differences, 2026.

Lantao Yu, Yang Song, Jiaming Song, and Stefano Ermon. Training deep energy-based models with f-divergence minimization. In *International Conference on Machine Learning*, pp. 10957–10967. PMLR, 2020.

Mingtian Zhang, Oscar Key, Peter Hayes, David Barber, Brooks Paige, and Francois-Xavier Briol. Towards healing the blindness of score matching. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022. URL `https://openreview.net/forum?id=Ij8G_k0iuL`.

A Proof and derivation

A.1 Derivation of the geometric term at direct star sampling

The direct star kernel is given as $Q^{(t)}_{direct}(X_1, \cdots, X_J|X_0) = \prod_{\ell\neq 0} r_t(X_\ell \mid X_0)$, so the likelihood is the softmax of $-U_\theta(X_k) + g^{\mathrm{dir}}_k(X)$ with

$$g^{\mathrm{dir}}_k(X) = \log D_k(X) = -\frac{1}{2t}\sum_{\ell\neq k}\|X_\ell - X_k\|^2 - \frac{Jd}{2}\log(2\pi t).$$

Expanding each summand around the centroid and using $\sum_\ell(X_\ell - \bar{X}) = 0$,

$$\sum_{\ell=0}^{J}\|X_\ell - X_k\|^2 = \sum_{\ell=0}^{J}\|X_\ell - \bar{X}\|^2 + m\,\|X_k - \bar{X}\|^2, \tag{7}$$

in which only the second term depends on k; the $\ell = k$ summand is zero and may be added freely. Hence $g^{\mathrm{dir}}_k(X) = -\gamma_0\|X_k - \bar{X}\|^2 + \mathrm{const}(X)$ with $\gamma_0 = m/(2t)$, and the class-independent remainder cancels in the softmax. At $J = 1$ one has $\|X_0 - \bar{X}\| = \|X_1 - \bar{X}\|$, so $g^{\mathrm{dir}}_0 = g^{\mathrm{dir}}_1$ and the correction disappears.

A.2 Formal statement of graph to NCE task theorem

In a generation graph, each edge corresponds to a Gaussian pairwise factor. Multiplying these factors and marginalizing the latent vertices yields the joint sampling factor

$$\Psi_G(\mathbf{X}) := \int_{(\mathbb{R}^d)^H}\prod_{\{a,b\}\in E} r_t(x_b \mid x_a)\,dH.$$

This nonnegative factor assigns a relative likelihood to each candidate tuple. To interpret it as a conditional sampling density, we first show that it is normalizable after pinning any candidate vertex.

Lemma 5 (Conditional normalizability of the sampling factor)**.** *Let the pair (G, S), with $G = (V, E)$, be a generation graph with candidate vertices $S = \{X_0, \cdots, X_J\}$. Then*

$$c_G(X_0) := \int_{(\mathbb{R}^d)^{S\setminus\{X_0\}}}\Psi_G(\mathbf{X})\,d\mathbf{X}_{-0} < \infty.$$

Moreover, $c_G(X_0)$ is invariant under natural permutations of G, so we write it simply as c_G.

Proof. Since the integrand is nonnegative, Tonelli's theorem allows us to combine the marginalization over the latent vertices and the remaining candidate vertices. Hence, fixing $X_0 = x$,

$$\begin{aligned} c_G(x) &= \int_{(\mathbb{R}^d)^{S\setminus\{X_0\}}}\Psi_G(\mathbf{X})\,d\mathbf{X}_{-0} \\ &= \int_{(\mathbb{R}^d)^{V\setminus\{X_0\}}}\prod_{\{a,b\}\in E} r_t(x_b \mid x_a)\,d\mathbf{x}_{V\setminus\{X_0\}}. \end{aligned}$$

Using $r_t(x_b \mid x_a) = (2\pi t)^{-d/2}\exp\Big(-\frac{\|x_a - x_b\|^2}{2t}\Big)$, we obtain

$$c_G(x) = (2\pi t)^{-\frac{d}{2}|E|}\int\exp\left[-\frac{1}{2t}\sum_{\{a,b\}\in E}\|x_a - x_b\|^2\right]d\mathbf{x}_{V\setminus\{X_0\}}.$$

Let L_G be the graph Laplacian and let $L^{(0)}_G$ denote the reduced Laplacian obtained by deleting the row and column corresponding to X_0. Since G is connected, $L^{(0)}_G$ is positive definite. After

translating every free variable by the pinned value x, the quadratic form is independent of x, and the integral becomes a standard Gaussian integral:

$$c_G(x) = (2\pi t)^{-\frac{d}{2}|E|}(2\pi t)^{\frac{d}{2}(|V|-1)} \det(L_G^{(0)})^{-d/2}$$
$$= (2\pi t)^{-\frac{d}{2}(|E|-|V|+1)} \det(L_G^{(0)})^{-d/2}.$$

By the Matrix–Tree theorem, $\det(L_G^{(0)}) = \kappa(G)$, where $\kappa(G)$ is the number of spanning trees of G. Writing $\mu(G) := |E| - |V| + 1$ for the cycle rank, we obtain

$$c_G(x) = (2\pi t)^{-\mu(G)d/2}\kappa(G)^{-d/2} < \infty.$$

In particular, the normalizing constant is independent of the pinned value x.

Moreover, the same argument applies after pinning any vertex X_k: every reduced Laplacian cofactor has determinant $\kappa(G)$. Consequently the pinned normalizer is independent of the candidate chosen as the root. Since both $\mu(G)$ and $\kappa(G)$ are preserved under relabeling of G, the normalizer is also invariant under any natural permutation of the graph. We may therefore denote the common constant simply by c_G. □

Theorem 6 (Graph symmetry induces kernel symmetry)**.** *For a symmetric generation graph (G, S), the sampling kernel Q_G induced by the joint factor Ψ_G, is symmetric. Assume that the data distribution p_d has a positive density and that the model is well specified, i.e., $p_{\theta^\star} = p_d$ for some $\theta^\star$. Then the induced NCE task, equipped with the objective in Eq. (4), is consistent.*

Proof. We will prove Theorem 6 in two parts.

Part 1: Symmetry. By Lemma 5, the sampling factor Ψ_G can be normalized after pinning $X_0 = x$. We therefore define $Q_G(X_1, \cdots, X_J|X_0) := \Psi_G(X_0, X_1, \cdots, X_J)/c_G$. As natural permutation does not change the graph, the factor does not change, thus the joint kernel is symmetric.

Part 2: Consistency. Let $m = J + 1$. By symmetry of Q_G, there exists a common function $C_G(\mathbf{X})$ such that, for every k,

$$Q_G\big(X_{\sigma_k(1)}, \ldots, X_{\sigma_k(J)} \mid X_{\sigma_k(0)}\big) = C_G(\mathbf{X}), \qquad \sigma_k(0) = k.$$

Hence, under hypothesis $K = k$, the joint density of the candidate pool is

$$P_k(\mathbf{X}) = p_d(X_k)C_G(\mathbf{X}).$$

The natural permutation σ_k maps P_0 to P_k, while also mapping the class-0 loss to the class-k loss. Therefore the population objective can equivalently be written as

$$\mathcal{L}(\theta) = \frac{1}{m}\sum_{k=0}^{J} \mathbb{E}_{\mathbf{X}\sim P_k}[-\log \Pi_\theta(k \mid \mathbf{X})].$$

Let

$$M(\mathbf{X}) := \frac{1}{m}\sum_{k=0}^{J} P_k(\mathbf{X}).$$

The Bayes posterior under this mixture is

$$\Pi^\star(k \mid \mathbf{X}) = \frac{p_d(X_k)}{\sum_{\ell=0}^{J} p_d(X_\ell)},$$

whereas the model posterior is $\Pi_\theta(k \mid \mathbf{X}) = \frac{\widetilde{p}_\theta(X_k)}{\sum_{\ell=0}^{J}\widetilde{p}_\theta(X_\ell)}$. Thus,

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^\star) = \mathbb{E}_{\mathbf{X}\sim M}\left[D_{\mathrm{KL}}\big(\Pi^\star(\cdot \mid \mathbf{X}) \,\|\, \Pi_\theta(\cdot \mid \mathbf{X})\big)\right] \geq 0.$$

Equality holds only if $\Pi_\theta(\cdot \mid \mathbf{X}) = \Pi^\star(\cdot \mid \mathbf{X})$ for M-almost every $\mathbf{X}$. Since Q_G is a nondegenerate Gaussian kernel, its density is strictly positive everywhere. Consequently, for almost every pair x, y with $p_d(x)p_d(y) > 0$,

$$\frac{\widetilde{p}_\theta(x)}{\widetilde{p}_\theta(y)} = \frac{p_d(x)}{p_d(y)}.$$

It follows that $\widetilde{p}_\theta(x)/p_d(x)$ is constant almost everywhere, i.e.,

$$\widetilde{p}_\theta = c_\theta p_d.$$

After normalization, $p_\theta = p_d$, proving Fisher consistency. □

A.3 Derivation of RMS for negative samplings

As we defined, the squared RMS displacement of a negative pool is:

$$v^2 := \frac{1}{Jd}\sum_{j=1}^{J}\mathbb{E}\big[\|X_j - X_0\|^2 \mid X_0\big].$$

For the anchor sampling kernel, $X_j - X_0 = \sqrt{t}\,\xi + \sqrt{t}\,\xi'_j$ where $\xi, \xi'_j \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$, and hence $\mathrm{Cov}(X_j - X_0) = 2tI_d$. Therefore,

$$v^2_{\text{anc}} = \frac{1}{Jd}\sum_{j=1}^{J} 2td = 2t.$$

For the cyclic sampling kernel, the Gaussian bridge satisfies

$$\mathrm{Cov}(X_r - X_0) = t\,\frac{r(J+1-r)}{J+1} I_d, \qquad r = 1, \dots, J.$$

Thus

$$\begin{aligned} v^2_{\text{cyc}} &= \frac{t}{J(J+1)}\sum_{r=1}^{J} r(J+1-r) \\ &= \frac{t}{J(J+1)}\,\frac{(J+1)((J+1)^2-1)}{6} = \frac{J+2}{6}t. \end{aligned}$$

For $r = 1$,

$$\mathrm{Var}(X_1 - X_0) = t\frac{m-1}{m} I_d = \Theta(t) I_d,$$

so candidates adjacent to the root lie at RMS scale $\Theta(\sqrt{t})$.

Since $r(m-r)$ is maximized at $r \approx m/2$,

$$\max_r \mathrm{Var}(X_r - X_0) = t\,\frac{\lfloor m^2/4 \rfloor}{m} I_d = \Theta(Jt) I_d.$$

Hence the farthest candidate reaches RMS scale $\Theta(\sqrt{Jt})$.

B Generalized sampling construction algorithm from a generation graph

B.1 Statement and proof

Appendix A.2 directly provides a sampling construction for a symmetric generation graph. Let (G, S) be a symmetric generation graph with candidate vertices $S = \{X_0, \dots, X_J\}$ and latent vertices $H = V \setminus S$. Partition the graph Laplacian according to $V = S \sqcup H$ as:

$$L_G = \begin{pmatrix} L_{SS} & L_{SH} \\ L_{HS} & L_{HH} \end{pmatrix}.$$

To obtain a closed-form representation of the induced procedure, we eliminate the latent vertices using Schur complement, also known as Kron reduction (Dorfler & Bullo, 2013), which we will describe below.

Lemma 7 (Gaussian marginalization using Schur complement)**.** *Let $G = (V, E)$ be a connected graph with $V = S \sqcup H$. Let $\mathbf{X}_S$ and $\mathbf{H}$ denote the stacked variables on S and H, respectively. Then*

$$\begin{aligned} &\int \exp\left[-\frac{1}{2t}\begin{pmatrix}\mathbf{X}_S \\ \mathbf{H}\end{pmatrix}^{\top}(L_G \otimes \mathbf{I}_d)\begin{pmatrix}\mathbf{X}_S \\ \mathbf{H}\end{pmatrix}\right] d\mathbf{H} \\ &= (2\pi t)^{|H|d/2}(\det L_{HH})^{-d/2}\exp\left[-\frac{1}{2t}\mathbf{X}_S^{\top}(\widetilde{L} \otimes \mathbf{I}_d)\mathbf{X}_S\right], \end{aligned}$$

where $\otimes$ is a Kronecker product and

$$\widetilde{L} := L_{SS} - L_{SH}L_{HH}^{-1}L_{HS}.$$

In particular, marginalizing the latent vertices induces the Schur-complement Laplacian $\widetilde{L}$ on the candidate vertices.

Proof. Since G is connected and $S \neq \varnothing$, $L_{HH} \succ 0$. Expanding the quadratic form and completing the square in $\mathbf{H}$ gives

$$\begin{aligned}\mathbf{H}^\top(L_{HH} \otimes \mathbf{I}_d)\mathbf{H} + 2\mathbf{X}_S^\top(L_{SH} \otimes \mathbf{I}_d)\mathbf{H} =& (\mathbf{H}+\mathbf{m})^\top(L_{HH} \otimes \mathbf{I}_d)(\mathbf{H}+\mathbf{m}) \\ &- \mathbf{X}_S^\top(L_{SH}L_{HH}^{-1}L_{HS} \otimes \mathbf{I}_d)\mathbf{X}_S,\end{aligned}$$

where $\mathbf{m} = (L_{HH}^{-1}L_{HS} \otimes \mathbf{I}_d)\mathbf{X}_S$. The Gaussian integral over $\mathbf{H}$ gives $(2\pi t)^{|H|d/2}(\det L_{HH})^{-d/2}$, leaving the Schur complement $L_{SS} - L_{SH}L_{HH}^{-1}L_{HS}$. □

The induced conditional distribution therefore has the following closed form.

Proposition 8. *The joint negative sampling distribution Q_G induced by a generation graph pair (G, S) is given by:*

$$X_{-0} \mid X_0 = x \;\sim\; \mathcal{N}\Big(\mathbf{1}_J \otimes x,\; t\left(\widetilde{L}^{(0)}\right)^{-1} \otimes \mathbf{I}_d\Big). \tag{8}$$

where $\widetilde{L} := L_{SS} - L_{SH}L_{HH}^{-1}L_{HS}$ is the Schur complement of L_G onto the candidate vertices.

Proof. Let $\mathbf{X}_S$ denote the stacked candidate variables, that is, $\mathbf{X}_S := (X_0^\top, \ldots, X_J^\top)^\top \in \mathbb{R}^{(J+1)d}$. Similarly, let $\mathbf{H}$ collect the latent vertices. Since each edge contributes the Gaussian factor r_t,

$$\prod_{\{a,b\}\in E} r_t(x_b \mid x_a) \propto \exp\left[-\frac{1}{2t}\begin{pmatrix}\mathbf{X}_S\\ \mathbf{H}\end{pmatrix}^\top (L_G \otimes \mathbf{I}_d)\begin{pmatrix}\mathbf{X}_S\\ \mathbf{H}\end{pmatrix}\right].$$

By Lemma 7, integrating out $\mathbf{H}$ is a Gaussian marginalization and yields

$$\Psi_G(\mathbf{X}) \propto \exp\left[-\frac{1}{2t}\mathbf{X}_S^\top(\widetilde{L} \otimes \mathbf{I}_d)\mathbf{X}_S\right].$$

Since $\widetilde{L}\mathbf{1} = 0$, the quadratic form is invariant under a common translation of all candidate variables.

$$\begin{aligned}\mathbf{X}_S^\top(\widetilde{L} \otimes \mathbf{I}_d)\mathbf{X}_S &= \begin{pmatrix}0\\ \mathbf{X}_{-0} - \mathbf{1}_J \otimes x\end{pmatrix}^\top (\widetilde{L} \otimes \mathbf{I}_d)\begin{pmatrix}0\\ \mathbf{X}_{-0} - \mathbf{1}_J \otimes x\end{pmatrix} \\ &= (\mathbf{X}_{-0} - \mathbf{1}_J \otimes x)^\top(A \otimes \mathbf{I}_d)(\mathbf{X}_{-0} - \mathbf{1}_J \otimes x),\end{aligned}$$

where $A := \widetilde{L}^{(0)}$ is the principal minor obtained by deleting the row and column of X_0. Therefore, after conditioning on $X_0 = x$, the quadratic form reduces to

$$Q_G(\mathbf{X}_{-0} \mid X_0 = x) \propto \exp\left[-\frac{1}{2t}(\mathbf{X}_{-0} - \mathbf{1}_J \otimes x)^\top(A \otimes \mathbf{I}_d)(\mathbf{X}_{-0} - \mathbf{1}_J \otimes x)\right].$$

Since G is connected, $A \succ 0$. Thus the conditional density is

$$X_{-0} \mid X_0 = x \sim \mathcal{N}\big(\mathbf{1}_J \otimes x,\; tA^{-1} \otimes \mathbf{I}_d\big),$$

which gives Equation 8. □

B.2 Examples of Generalized sampling construction

We next give several symmetric generation graphs and their induced kernels. Let $m = |S| = J + 1$.

Anchor. Let $V = S \cup \{H\}$ with edges $\{H, X_j\}$. With the candidates ordered before H,

$$L_G = \begin{pmatrix} I_m & -\mathbf{1} \\ -\mathbf{1}^\top & m \end{pmatrix},$$

and hence its Schur complement onto the candidate set S is

$$\widetilde{L} = I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^\top = \frac{1}{m}L_{K_m}.$$

Here, we can observe that the anchor sampling is equivalent to the graph sampling from complete graph with edge scale mt.

After pinning $X_0 = x$,

$$\widetilde{L}^{(0)} = I_J - \frac{1}{m}\mathbf{1}\mathbf{1}^\top, \qquad (\widetilde{L}^{(0)})^{-1} = I_J + \mathbf{1}\mathbf{1}^\top.$$

Therefore, by Proposition 8,

$$X_{-0} \mid X_0 = x \sim \mathcal{N}\big(\mathbf{1}_J \otimes x,\, t(I_J + \mathbf{1}\mathbf{1}^\top) \otimes \mathbf{I}_d\big),$$

which is equivalently sampled as

$$X_j = x + \sqrt{t}\,\xi + \sqrt{t}\,\xi'_j, \qquad \xi, \xi'_1, \ldots, \xi'_J \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_d).$$

Cycle. Let $V = S$ with edges $\{X_r, X_{r+1}\}$, $r = 0, \ldots, m-1$, where $X_m := X_0$. Since there are no latent vertices,

$$\widetilde{L} = L_{C_m}.$$

After pinning $X_0 = x$,

$$A := \widetilde{L}^{(0)} = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix},$$

whose inverse satisfies

$$(A^{-1})_{rq} = \frac{\min(r,q)\,[m - \max(r,q)]}{m}.$$

Hence, by Proposition 8,

$$X_{1:J} \mid X_0 = x \sim \mathcal{N}\big(\mathbf{1}_J \otimes x,\, tA^{-1} \otimes \mathbf{I}_d\big),$$

so in particular

$$\mathrm{Var}(X_r - X_0) = t\frac{r(m-r)}{m}\mathbf{I}_d.$$

Equivalently, draw $Z_0, \ldots, Z_{m-1} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$, set

$$\Delta_r = \sqrt{t}\,(Z_r - \bar{Z}), \qquad \bar{Z} = \frac{1}{m}\sum_{a=0}^{m-1} Z_a,$$

and recursively define $X_r = X_{r-1} + \Delta_{r-1}$. Since $\sum_r \Delta_r = 0$, this is exactly the pinned Gaussian bridge on the cycle.

Cyclic anchoring (One candidate per anchor). Let each candidate X_j be attached to a latent anchor H_j, and let $H_0, \ldots, H_{m-1}$ form a cycle. Ordering the candidates before the anchors gives

$$L_G = \begin{pmatrix} I_m & -I_m \\ -I_m & I_m + L_{C_m} \end{pmatrix}.$$

and hence its Schur complement onto the candidate set S is

$$\widetilde{L} = I_m - (I_m + L_{C_m})^{-1} = L_{C_m}(I_m + L_{C_m})^{-1},$$

whose nonzero eigenvalues are

$$\frac{\lambda_k}{1+\lambda_k}, \qquad \lambda_k = 4\sin^2(\pi k/m).$$

After pinning $X_0 = x$, Proposition 8 gives

$$X_{1:J} \mid X_0 = x \sim \mathcal{N}\Big(\mathbf{1}_J \otimes x,\ t(\widetilde{L}^{(0)})^{-1} \otimes \mathbf{I}_d\Big).$$

Equivalently,

$$(\widetilde{L}^{(0)})^{-1} = (L^{(0)}_{C_m})^{-1} + I_J + \mathbf{1}\mathbf{1}^\top,$$

so the sampler can be realized by first drawing $H_0 \sim r_t(\cdot \mid x)$, then drawing the remaining anchors as a cyclic Gaussian bridge pinned at H_0, and finally sampling each $X_j \sim r_t(\cdot \mid H_j)$ independently.

Two-rank anchoring. Let each candidate X_j be attached to a rank-one anchor $H_{1,j}$, and connect all rank-one anchors to a common rank-two anchor H_2. With vertices ordered as (S, H_1, H_2),

$$L_G = \begin{pmatrix} I_m & -I_m & 0 \\ -I_m & 2I_m & -\mathbf{1} \\ 0 & -\mathbf{1}^\top & m \end{pmatrix}.$$

Eliminating the rank-one anchors first gives

$$L' = \frac{1}{2}\begin{pmatrix} I_m & -\mathbf{1} \\ -\mathbf{1}^\top & m \end{pmatrix},$$

and eliminating H_2 then yields

$$\widetilde{L} = \frac{1}{2}\left(I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^\top\right) = \frac{1}{2m}L_{K_m}.$$

Hence

$$(\widetilde{L}^{(0)})^{-1} = 2(I_J + \mathbf{1}\mathbf{1}^\top),$$

and

$$X_{1:J} \mid X_0 = x \sim \mathcal{N}\big(\mathbf{1}_J \otimes x,\ 2t(I_J + \mathbf{1}\mathbf{1}^\top) \otimes \mathbf{I}_d\big).$$

Thus two-rank anchoring induces the same complete-graph candidate geometry as ordinary anchoring, with twice the effective scale.

C Graph sampling procedure induces a reversible CD kernel

Proposition 9 (Symmetric sampling kernel induces a reversible CD kernel)**.** *Let (G, S) be a symmetric generation graph, and let $Q_G(X_1, \dots, X_J \mid X_0)$ be its induced sampling kernel. For each $k \in \{0, \dots, J\}$, let σ_k be a natural permutation of Q_G.*

Given a current state $X_0 = x$, draw

$$X_1, \dots, X_J \sim Q_G(\cdot \mid x),$$

and with $w_{\theta,k}(X_0, \dots, X_J) = p_\theta(X_k)/\sum_{\ell=0}^{J} p_\theta(X_\ell)$, sample

$$K \sim \mathrm{Cat}(w_{\theta,0}(X_0, \dots, X_J), \dots, w_{\theta,J}(X_0, \dots, X_J)).$$

Then the transition $x \mapsto X_K$ defines a Markov kernel $K_{\theta,G}$ that is reversible with respect to p_θ. Moreover,

$$\nabla_\theta \ell_G(\theta; X_0, \dots, X_J) = \mathbb{E}[\nabla_\theta U_\theta(X_0) - \nabla_\theta U_\theta(X_K) \,|\, X_0, \dots, X_J],$$

so the symmetric NCE gradient is the Rao–Blackwellized CD-1 update associated with $K_{\theta,G}$.

To prove Proposition 9, we first define a *balanced natural permutation set* and establish its existence.

Lemma 10 (Balanced natural permutation set)**.** *Let (G, S) be a symmetric generation graph with $S = \{0, \dots, J\}$. A natural permutation set $\Sigma^* := \{\sigma_0^*, \sigma_1^*, \dots, \sigma_J^*\}$ is called* balanced *if*

$$\{\sigma_j^{*-1}(0) : j = 0, \dots, J\} = S.$$

For every symmetric generation graph (G, S), there exists a balanced natural permutation set Σ^*.

Proof of Lemma. Let Γ be the set of all natural permutations of (G, S). For each $j \in S$, define $L_j := \{\sigma \in \Gamma : \sigma(0) = j\}$ and $R_j := \{\sigma \in \Gamma : \sigma^{-1}(0) = j\}$.

We first show that all L_j's have the same size. Since (G, S) is symmetric, for every $j \in S$ there exists a natural permutation τ_j such that $\tau_j(0) = j$. Then $\sigma \mapsto \tau_j \circ \sigma$ is a bijection from L_0 to L_j, because the composition and inverse of natural permutations are again natural permutations. Hence $|L_0| = |L_1| = \cdots = |L_J|$. Similarly, $\sigma \mapsto \sigma \circ \tau_j^{-1}$ is a bijection from R_0 to R_j, and therefore $|R_0| = |R_1| = \cdots = |R_J|$. Moreover, $L_0 = R_0$, since $\sigma(0) = 0$ if and only if $\sigma^{-1}(0) = 0$. Thus there is some $n > 0$ such that $|L_j| = |R_j| = n$ for every $j \in S$.

Now construct a bipartite multigraph with left vertices $0, \dots, J$ and right vertices $0, \dots, J$. For each natural permutation $\sigma \in \Gamma$, draw an edge from the left vertex $\sigma(0)$ to the right vertex $\sigma^{-1}(0)$.

By the definitions of L_j and R_j, every left vertex and every right vertex has exactly n incident edges. Consider any set A of left vertices, and let $N(A)$ be the set of right vertices connected to A. There are exactly $n|A|$ edges leaving A. Since each vertex in $N(A)$ is incident to at most n of these edges, $n|A| \le n|N(A)|$. Therefore, $|A| \le |N(A)|$.

By Hall's marriage theorem, the bipartite graph has a perfect matching. For each left vertex j, let σ_j^* be the natural permutation corresponding to its matched edge. Then $\sigma_j^*(0) = j$, and the right endpoints of the matched edges are all distinct and cover S. Hence $\{\sigma_j^{*-1}(0) : j = 0, \dots, J\} = S$. Therefore, $\Sigma^* = \{\sigma_0^*, \dots, \sigma_J^*\}$ is balanced. □

Proof of Proposition 9. We start by writing $p_\theta(X_0) K_{G,\theta}(X'|X_0)$ with $X' \neq X_0$ as

$$\begin{aligned}
= p_\theta(X_0) \Bigg[& \int Q_G(X', X_2, \dots, X_J | X_0) \frac{p_\theta(X')}{p_\theta(X') + p_\theta(X_0) + \sum_{\ell \neq 0,1} p_\theta(X_\ell)} dX_2 dX_3 \dots dX_J \\
& + \int Q_G(X_1, X', \dots, X_J | X_0) \frac{p_\theta(X')}{p_\theta(X') + p_\theta(X_0) + \sum_{\ell \neq 0,2} p_\theta(X_\ell)} dX_1 dX_3 \dots dX_J \\
& + \dots \\
& + \int Q_G(X_1, X_2, \dots, X' | X_0) \frac{p_\theta(X')}{p_\theta(X') + p_\theta(X_0) + \sum_{\ell \neq 0,J} p_\theta(X_\ell)} dX_1 dX_2 \dots dX_{J-1} \Bigg]
\end{aligned}$$

By Lemma 10, choose a balanced natural permutation set $\Sigma^* = \{\sigma_0^*, \dots, \sigma_J^*\}$, and define $r_j := (\sigma_j^*)^{-1}(0)$. Since $\sigma_0^*(0) = 0$, we have $r_0 = 0$. Hence $\{r_1, \dots, r_J\} = \{1, \dots, J\}$.

Apply σ_j^* to the j-th term of the preceding sum. Since $\sigma_j^*(0) = j$, the conditioning variable becomes $X_j = X'$, while the original root X_0 appears in position r_j. The remaining coordinates are only permuted. A permutation change of variables preserves Lebesgue measure, and $\sum_{\ell=0}^{J} p_\theta(X_\ell)$ is invariant under the same permutation. Therefore, the j-th forward term is equal to the r_j-th term of the reverse transition.

Writing the j-th contribution as F_j, we therefore have

$$F_j(X_0, X') = F_{r_j}(X', X_0).$$

Consequently,

$$\begin{aligned} p_\theta(X_0)K_{G,\theta}(X'|X_0) &= \sum_{j=1}^{J} F_j(X_0, X') \\ &= \sum_{j=1}^{J} F_{r_j}(X', X_0) \\ &= \sum_{r=1}^{J} F_r(X', X_0) = p_\theta(X')K_{G,\theta}(X_0|X'). \end{aligned}$$

Thus detailed balance holds for $X' \neq X_0$. The self-transition satisfies detailed balance trivially, and hence $K_{G,\theta}$ is reversible with respect to p_θ.

For the second statement, the energy-only objective is

$$\ell_G(\theta; \mathbf{X}) = U_\theta(X_0) + \log \sum_{k=0}^{J} e^{-U_\theta(X_k)}.$$

Its gradient is

$$\nabla_\theta \ell_G(\theta; \mathbf{X}) = \nabla_\theta U_\theta(X_0) - \sum_{k=0}^{J} \pi_{\theta,k}(\mathbf{X}) \nabla_\theta U_\theta(X_k).$$

Since

$$K \mid \mathbf{X} \sim \mathrm{Cat}\big(\pi_{\theta,0}(\mathbf{X}), \ldots, \pi_{\theta,J}(\mathbf{X})\big),$$

we obtain

$$\nabla_\theta \ell_G(\theta; \mathbf{X}) = \mathbb{E}[\nabla_\theta U_\theta(X_0) - \nabla_\theta U_\theta(X_K) \mid \mathbf{X}].$$

Averaging over $X_{1:J} \sim Q_G(\cdot \mid X_0 = x)$ gives

$$\mathbb{E}_{Q_G(\cdot|x)} [\nabla_\theta \ell_G(\theta; x, X_{1:J})] = \nabla_\theta U_\theta(x) - \mathbb{E}_{X' \sim K_{\theta,G}(\cdot|x)} [\nabla_\theta U_\theta(X')].$$

Thus the symmetric NCE gradient is the Rao–Blackwellized CD-1 update associated with $K_{\theta,G}$. □

D EXPECTED NEGATIVES WITHIN A R-RADIUS BALL

The proposition below shows the local sampling of a Gaussian bridge relative to anchor sampling.

Proposition 11 (The expected fraction of negatives within a radius)**.** *Let $N(a)$ denote the expected fraction of negatives whose per-coordinate distance from X_0 is at most av:*

$$N(a) := \frac{1}{J} \sum_{r=1}^{J} \Pr\left(\frac{\|X_r - X_0\|}{\sqrt{d}} \leq av \right).$$

Then:

1. *For anchoring,*

$$N_{\mathrm{anc}}(a) = F_{\chi^2_d}(da^2).$$

2. *For the cyclic Gaussian bridge,*

$$N_{\mathrm{cyc}}(a) = \frac{1}{J} \sum_{r=1}^{J} F_{\chi^2_d}\left(\frac{da^2}{\beta_r} \right),$$

where

$$\beta_r := \frac{6r(J+1-r)}{(J+1)(J+2)}.$$

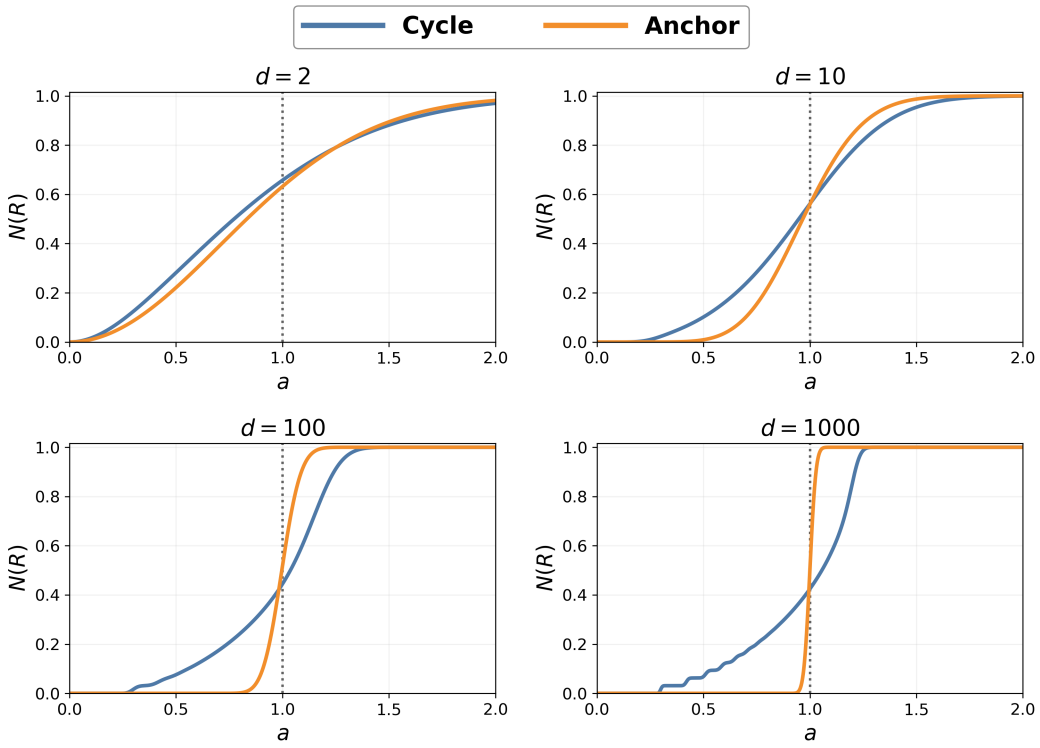


Figure 3: The expected fraction of negatives inside per coordinate $R = av$-radius ball in the case of negative sample size $J = 64$, with dimension $d = 2, 10, 100, 1000$.

Proof. Let $m = J + 1$. We measure locality using the per-coordinate normalized distance, so that a negative X_r lies within radius av whenever $\|X_r - X_0\|^2/d \le a^2v^2$.

For anchoring, marginalizing the anchor gives

$$X_r - X_0 \sim \mathcal{N}(0, v^2 I_d),$$

where $v^2 = 2t$. Hence

$$\frac{\|X_r - X_0\|^2}{v^2} \sim \chi^2_d,$$

and therefore

$$N_{\text{anc}}(a) = \Pr\left(\frac{\|X_r - X_0\|^2}{v^2} \le da^2\right) = F_{\chi^2_d}(da^2).$$

For the cyclic Gaussian bridge, recall that the marginal displacement of the r-th candidate is

$$X_r - X_0 \sim \mathcal{N}\left(0,\, t\frac{r(m-r)}{m} I_d\right).$$

As we derived in section A.3, the standard Gaussian-bridge variance is given as $v^2 = t(J+2)/6$. Thus

$$\text{Var}(X_r - X_0) = \beta_r v^2 I_d$$

where

$$\beta_r = \frac{6r(J+1-r)}{(J+1)(J+2)}.$$

Consequently,

$$\Pr\left(\frac{\|X_r - X_0\|^2}{d} \le a^2v^2\right) = F_{\chi^2_d}\left(\frac{da^2}{\beta_r}\right).$$

□

We plot the expected fraction as a function of the radius ratio a in Figure 3. The anchoring kernel suffers from thin-shell concentration in high dimensions, whereas the cyclic bridge distributes negatives across a wider range of radii and retains a larger fraction of local samples.

E EXPERIMENTAL DETAILS

E.1 EXPERIMENTAL SETTING: DENSITY ESTIMATION ON SYNTHETIC DATASETS

We provide experimental details for Figure 2, which compares Direct star, Anchor (ED), and Cycle (GORI) on three synthetic Gaussian-mixture benchmarks.

Datasets. All three benchmarks have uniform mixture weights and analytically available target densities $p_\star$. The first benchmark, Uniform 25-Gaussians, is a two-dimensional mixture with component means $(2i, 2j)/1.414$, where $i, j \in \{-2, -1, 0, 1, 2\}$, and a common isotropic standard deviation $0.2/1.414$. The second benchmark, Tiered-width 25-Gaussians, has means $(4i, 4j)$ on the same index grid. Its isotropic standard deviations are 0.6 for the central component, 0.4 for the eight components with $\max(|i|, |j|) = 1$, and 0.2 for the sixteen components with $\max(|i|, |j|) = 2$. Both mixtures assign weight $1/25$ to each component. The third benchmark, Uniform hypercube, is a six-dimensional mixture with 64 component means at the vertices of $\{-1.25, +1.25\}^6$, isotropic standard deviation 0.18, and component weights $1/64$. Training minibatches are sampled online from the corresponding target distribution.

Model architecture. We parameterize the scalar energy $E_\theta : \mathbb{R}^d \to \mathbb{R}$ using a fully connected network with four hidden layers, each containing 128 units and a Softplus activation, followed by a linear scalar output. Here $d = 2$ for the two Gaussian-grid benchmarks and $d = 6$ for the hypercube benchmark. The unnormalized model density is $\exp(-E_\theta(x))$.

Training details. We use Adam with learning rate 10^{-3}, batch size 128, and $50{,}001$ parameter updates. We maintain an exponential moving average (EMA) of the parameters with decay 0.999 and use the EMA model for evaluation. Each method is run with seeds 0–9. All three panels use $J = 16$ negative candidates in addition to the positive sample, and the same per-coordinate RMS displacement $v = 1$. We use the objective in Equation 3, including its geometric correction, for Direct star sampling, and the objective in Equation 4 for Anchor and Cycle sampling. Sampler variances and the corresponding geometric corrections are converted consistently to enforce this common RMS scale; equal v does not mean equal internal noise variances across the samplers.

Metric and evaluation. We evaluate the EMA model every 100 training iterations using the mean squared error between its estimated normalized log-density and the analytic target log-density. At each evaluation, we estimate the normalizing constant using $N_Z = 100{,}000$ independent target-distributed importance samples:

$$Y_n \sim p_\star, \qquad \widehat{Z}_\theta = \frac{1}{N_Z} \sum_{n=1}^{N_Z} \frac{\exp(-E_\theta(Y_n))}{p_\star(Y_n)}, \qquad \widehat{\ell}_\theta(x) = -E_\theta(x) - \log \widehat{Z}_\theta.$$

These samples are drawn from a separate random-number stream, and processed in chunks of $5{,}000$ using log-space accumulation. The reported error is

$$\widehat{\mathrm{MSE}}_{\log p} = \frac{1}{N_{\mathrm{eval}}} \sum_{i=1}^{N_{\mathrm{eval}}} \left[\widehat{\ell}_\theta(X_i) - \log p_\star(X_i)\right]^2, \qquad N_{\mathrm{eval}} = 5{,}000.$$

Curves show arithmetic means over ten seeds, and shaded regions show pointwise 95% nonparametric bootstrap confidence intervals from $10{,}000$ resamples of the seeds. The error axis is logarithmic.

E.2 EXPERIMENTAL SETTING: DENSITY ESTIMATION ON REAL DATASETS

We describe the full-vector joint energy-based density-estimation experiments on UCI benchmarks, as detailed below.

Datasets. We evaluate on POWER, GAS, HEPMASS, MINIBOONE, and BSDS300, using the released data splits and the preprocessing code of Papamakarios et al. (2017), verified against the checksum of the official archive. For BSDS300 we apply one additional step: the released 63-dimensional patches have a per-coordinate standard deviation of about 0.09, so we standardize each coordinate, $y_j = (x_j - \mu_j)/s_j$, with μ_j and s_j computed on the training and validation splits, exactly as Papamakarios et al. (2017) do for the other four datasets.

Model architecture. We use a joint residual network that maps the complete input vector to a scalar $g_\theta(x)$. An input linear projection to width 128 with ReLU activation is followed by four width-128 residual blocks and a biased scalar linear output layer. The dropout probability is 0.1, except on MINIBOONE, where it is 0.5. The unnormalized log-density is $f_\theta(x) = -E_\theta(x) = g_\theta(x) - \|x\|_2^2/(2\sigma^2)$, i.e. the network output is combined with a fixed isotropic Gaussian base density.

The same model, without spectral normalization or output scaling, is used for all five datasets (POWER, GAS, HEPMASS, MINIBOONE, and BSDS300). Because a ReLU network grows at most linearly, the Gaussian base term guarantees a finite normalizing constant and prevents probability mass far from the data, which otherwise made the SMC estimate of $\log Z$ depend on the sampler budget. We set $\sigma = 2$ for every method and dataset; it was selected from $\{1, 2, 3\}$ on a MINIBOONE seed-0 pilot by the stability of $\widehat{\log Z}$ across MALA budgets and the mean validation NLL, never using test data.

Training details. We compare Direct star, Anchor (ED), Cycle (GORI), Cyclic anchoring (CA) 2×10 (the two anchor vertices are connected by a single edge, and ten candidate vertices are attached to each anchor in a star configuration), and binary CNCE, using seeds 0–6. POWER 1M, GAS 600k, HEPMASS 200k, MINIBOONE 300k, and BSDS300 400k updates, validation every 10k. Every method uses $J = 19$ negative candidates, giving 20 observed candidates including the positive sample. All methods use per-coordinate RMS negative displacement $v = 1$. The internal variance is converted separately for each sampler, and the same conversion is used in its geometric correction. Direct star uses Equation 3; Anchor, Cycle, and the symmetric grouped sampler use Equation 4. Binary CNCE averages its pairwise logistic loss over the negative candidates.

Training uses Adam with learning rate 5×10^{-4}, $(\beta_1, \beta_2) = (0, 0.999)$, cosine learning-rate decay, batch size 512, and gradient-norm clipping at 5. We use no EMA, weight decay, or gradient penalty. Training runs to a fixed dataset-specific horizon; validation selects the checkpoint but does not terminate training early.

Metric and evaluation. We report the held-out test log-likelihood in nats per example, not per-dimension, selecting the checkpoint with the highest validation log-likelihood (evaluated every $10{,}000$ updates). Since the model specifies an unnormalized joint density, we estimate its normalizing constant and compute

$$\mathrm{Log}-\widehat{\mathrm{Likelihood}}_{\mathrm{test}} = -\log \widehat{Z}_\theta + \frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} f_\theta(x_i),$$

where f_θ includes the Gaussian base term. For BSDS300, we report log-likelihoods in the original coordinates through the exact change of variables $\log p_X(x) = \log p_Y(y) - \sum_{j=1}^{63} \log s_j$, where $\sum_j \log s_j = -153.33$. The normalizer is estimated by sequential Monte Carlo (SMC) along the geometric path from a 50-component Gaussian-mixture proposal, fitted only to training data, to the model: adaptive tempering with target relative effective sample size 0.5 (at most $10{,}000$ stages), multinomial resampling at every stage, and 50 Metropolis-adjusted Langevin rejuvenation steps per stage with an adapted step size (20 steps for the validation estimates on MINIBOONE). The reported estimate uses $500{,}000$ particles, 50 Langevin steps per stage, and three independent replicates. To check that the estimate has converged, every selected checkpoint is re-evaluated with 10^6 particles and with 100 Langevin steps per stage (three replicates each), and with 200 Langevin steps per stage (one replicate).

E.3 Experimental Setting: image generation

Datasets. We evaluate on SVHN, CIFAR-10, and CelebA. We use the official training and test splits of SVHN (73,257/26,032 images) and CIFAR-10 (50,000/10,000 images), retaining their native 32×32 resolution. For CelebA, we use the first 40,000 filenames in sorted order from the official training split and all 19,962 test images. Following Pang et al. (2020), we resize CelebA images to 64×64. We apply the same preprocessing to training and FID reference images.

Model architecture. We use a latent energy-based prior coupled to a convolutional generator:

$$p_\theta(z) = \frac{1}{Z_\theta} \exp\left(-E_\theta(z) - \frac{1}{2}\|z\|_2^2\right), \qquad p_\phi(x \mid z) = \mathcal{N}\big(g_\phi(z), 0.3^2 I\big) .$$

The latent dimension is 100 for SVHN and CelebA and 128 for CIFAR-10. The energy network has two hidden layers of 200 units with Leaky ReLU activations of slope 0.1 and a scalar linear output. The generator uses transposed convolutions, Leaky ReLU activations of slope 0.2 between layers, and a final Tanh activation. We use no encoder, batch normalization, spectral normalization, or dropout. Linear and transposed-convolution weights are initialized with Xavier normal initialization, and their biases are initialized to zero.

Training details. We compare Direct star, Anchor (ED), Cycle (GORI), Cyclic anchoring (CA) 2×10, and binary CNCE using seeds 0–2. All methods use $J = 19$ negatives plus one positive sample and per-coordinate RMS negative displacement $v = 0.25$. The internal noise variance and geometric correction are converted consistently for each sampler. Prior-learning objectives are evaluated using the total energy $U_\theta(z) = E_\theta(z) + \|z\|_2^2/2$, including the Gaussian base contribution. Direct star uses Equation 3, while Anchor, Cycle, and the symmetric grouped sampler use Equation 4. Binary CNCE averages its pairwise logistic loss over the negative candidates.

We jointly train the generator and prior for 200 epochs with an effective batch size of 512, dropping the final incomplete training batch. CIFAR-10 and CelebA use gradient accumulation over micro-batches of 64. We use separate Adam optimizers and fixed learning rates 10^{-4} for the generator and 5×10^{-5} for the energy network. No learning-rate schedule or EMA is used. The generator minimizes the pixel-summed squared reconstruction error averaged over the minibatch. Posterior samples are detached before the parameter updates. For each training minibatch, we initialize latent variables independently from $\mathcal{N}(0, I)$ and apply unadjusted Langevin dynamics to the posterior. We use 20 steps for SVHN and CelebA and 40 steps for CIFAR-10, with step-size parameter $s = 0.1$.

Metric and evaluation. We report Fréchet Inception Distance (FID). For generation, we initialize latent variables from $\mathcal{N}(0, I)$, run 500 prior Langevin steps with $s = 0.2$, and decode the resulting latent variables. Each FID evaluation uses 50,000 generated images and the full training-image reference set for that dataset (73,257 for SVHN, 50,000 for CIFAR-10, and 40,000 for CelebA). Images are mapped to $[0, 1]$ and clipped before conversion to the Inception input range. We report TensorFlow Inception FID using the final checkpoint after 200 epochs. Results are summarized as arithmetic means and standard deviations across the three training seeds.

Generated samples. Figure 4 shows representative SVHN, CIFAR-10, and CelebA samples generated from the seed-0 models for Anchor, Cycle, and Cyclic anchoring (CA) 2×10.

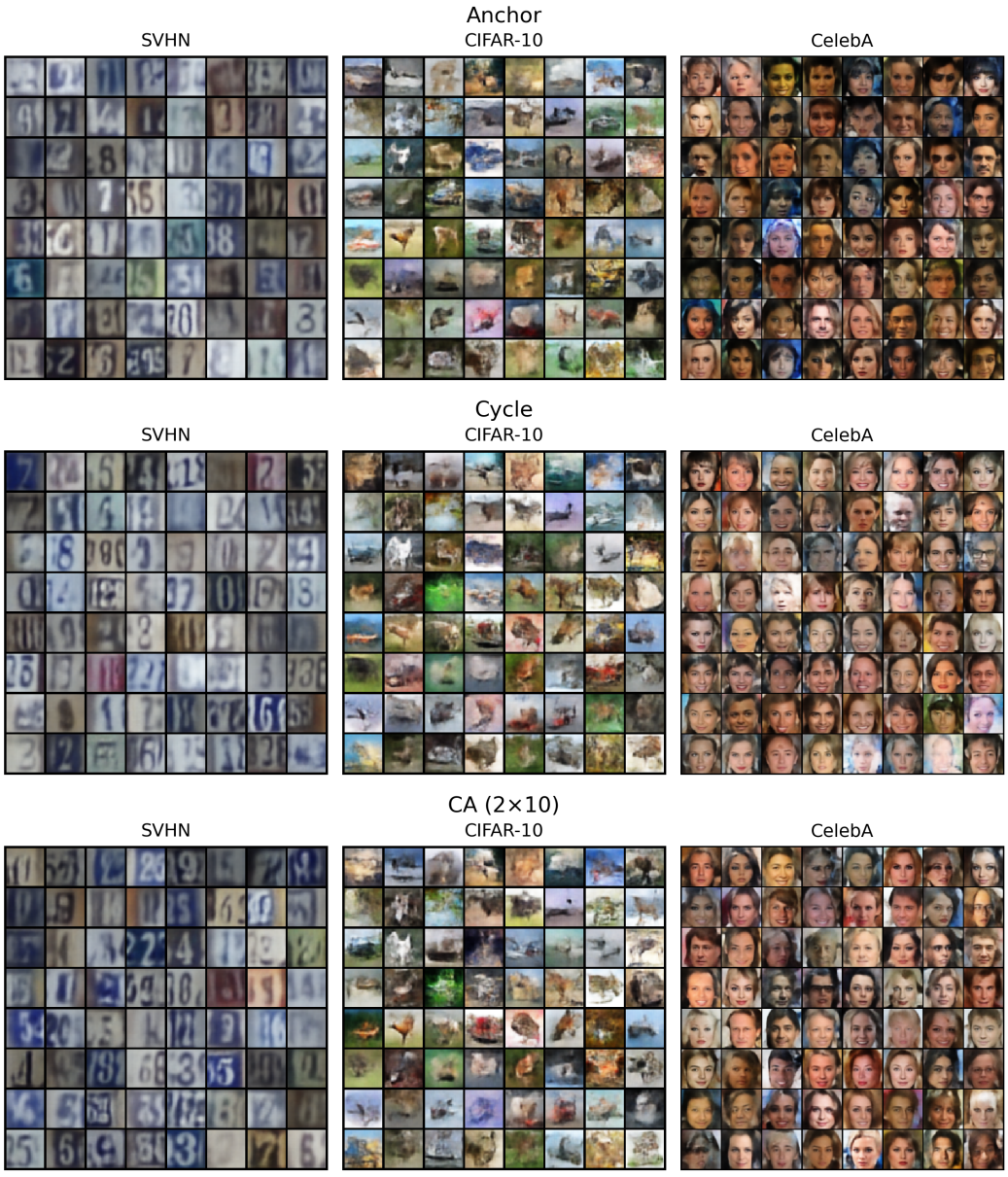


Figure 4: Representative SVHN, CIFAR-10, and CelebA samples generated from the seed-0 models. From top to bottom: Anchor (ED), Cycle (GORI), and Cyclic anchoring (CA) 2×10.