Continuous representations and models from random walk diffusion limits

by

Tatsunori B. Hashimoto

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Signature redacted

Author ....
Department of Electrical Engineering and Computer Science
Signature redacted May 3, 2016

Certified by ....
Tommi S. Jaakkola
Professor
Signature redacted Thesis Supervisor

Certified by ....
David K. Gifford
Professor
Thesis Supervisor

Accepted by .... Signature redacted ....
Leslie A. Kolodziejski
Chairman, Department Committee on Graduate Theses

Continuous representations and models from random walk diffusion limits

by

Tatsunori B. Hashimoto

Submitted to the Department of Electrical Engineering and Computer Science
on May 3, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Structured data such as sequences and networks pose substantial difficulty for traditional statistical theory which has focused on data drawn independently from a vector space. A popular and empirically effective technique for dealing with such data is to map elements of the data to a vector space and to operate over the embedding as a summary statistic. Such a vector representation of discrete objects is known as a 'continuous representation'. Continuous space models of words, objects, and signals have become ubiquitous tools for learning rich representations of data, from natural language processing to computer vision. Even in cases that the embedding is not explicit, many algorithms operate over similarity measures which implicitly embed the original dataset. In this thesis, we attempt to understand the intuition behind continuous representations. Can we construct a general theory of continuous representations? Are there general principles for semantically meaningful representations?

In order to answer these questions, we develop a framework for analyzing continuous representations through diffusion limits of random walks. We show that measureable quantities of discrete random walks with a latent metric structure have closed form diffusion limits. These diffusion limits allow us to approximate attributes of the discrete random walk such as the stationary distribution, hitting time, or co-occurrence using closed-form expressions from diffusions. We establish limits which guarantee asymptotic consistency of such estimators, and show they work well in practice.

Using this new approach, we solve three classes of problems: first, we derive principled network algorithms which connect statistical estimation tasks such as density estimation to network algorithms such as PageRank. Next, we demonstrate that continuous representations of words are a type of random walk metric estimator with close connections to manifold learning. Finally, we apply our theory to single-cell RNA seq data, and derive a way to learn time-series models without trajectories by using stochastic recurrent neural networks.

Thesis Supervisor: Tommi S. Jaakkola
Title: Professor

Thesis Supervisor: David K. Gifford
Title: Professor

Acknowledgments

This thesis and my PhD were made possible only through the contributions of many mentors, colleagues and friends.

I am fortunate to have had two incredible advisors: Tommi Jaakkola and David Gifford who pushed me to be rigorous and productive in my research. Dave and Tommi's patience with my ever-changing research directions allowed me to explore a wide-variety of research problems related to machine learning and biology, only a small part of which is included in this thesis. I'm also thankful for their role in my development as a researcher: Dave, for pushing me to be more productive and work on biologically relevant problems, and Tommi for demanding both rigor and real-world impact in my work.

I'm also thankful for my third thesis commitee member and undergraduate advisor, Edoardo Airoldi, who provided substantial encouragement for going into a PhD program, as well as valuable feedback on the thesis document.

Most of the content in the thesis would not have been possible without my co-authors and friends: Yi Sun and David Alvarez-Melis. Yi has been a great friend and roommate for many years and his combination of mathematical rigor and understanding of probability theory was invaluable to much of the work in this thesis. Without him, I would not have known about several key results such as the Stroock-Varadhan lemma. David was invaluable to the work in chapter 4, which required combining graph limits with natural language processing (NLP). David's knowledge of the NLP and psychometric literature elevated the work from a simple theoretical exercise to a genuinely useful paper on word embeddings.

My co-authors for several computational biology papers not included in this thesis were also invaluable to my development as a PhD student: Richard Sherwood, who was the biological mastermind behind assay development, has been a great collaborator who is patient enough to explain the basic biology and cares about the underlying statistics. Rich was the best biological collaborator I could have hoped for. Matthew Edwards, who always seriously discussed my half-baked ideas, has been a great friend and made the lab a fun place to be, especially with his barbeque. Daniel Kang, who was the first undergraduate I supervised, surprised me by learning the material so quickly and was the best programmer I've had the pleasure of working with.

The entire Gifford and Jaakkola labs made PhD an enjoyable time, and I'm especially thankful for those members who I've pestered (Jonas, Logan), who organized lab activities (Matt, Haoyang), and those who would come play boardgames or go to the muddy charles when research progress was slow (pretty much everyone).

Finally, I'd like to thank my parents without whom I (literally) would not be here, and my fiancee Victoria Hung. My parents for encouraging me to go into PhD rather than industry, and making my time at MIT as stress-free as can be. Vicky has been an essential part of my life these five years, complementing and correcting for my many faults. I'm so lucky to have found such a kind, skilled scientist for a partner. PhD wasn't so hard when it was with you.

Contents

List of Figures

List of Tables

Nomenclature

Graphs

G_n A unweighted, directed graph with n vertices

$\mathsf{NB}_n(x)$ Out-neighbors of vertex x

$\mathsf{NB}^{\mathsf{in}}(x)$ In-neighbors of vertex x

$\overline{\varepsilon}(x)$ Limiting neighborhood radius of x (e.g. in the k-nn graph, this is $(p(x)V_dn/k)^{1/d}$)

$\varepsilon_n(x)$ Neighborhood radius of x (e.g. in the k-nn graph, this is thedistance to the k-th nearest neighbor)

g_n Scaling limit, the rate at which $\varepsilon_n(x) \to 0$ as $n \to \infty$.

$\widehat{\varepsilon}$ Estimatior for $\varepsilon_n(x)$

$\widehat{p}$ Estimatior for the density $p(x)$ based upon G_n

Other

Tr Trace of a matrix

Var, Cov Variance / covariancex of a random variable

Random walks

$W(t)$ or W_t Brownian motion at time t

$X_n(t)$ A discrete random walk over n objects at time t

$Y(t)$ A diffusion process at time t

$\overline{T}^x_E$ Hitting time of a Brownian motion started at x stopped at E

$\pi_Y(x)$ Density of the stationary distribution of a random walk Y at position x

$T^{x_i}_{x_j,n}$ Discrete hitting time on a n vertex graph started at vertex x_j and stopped at x_i

T^x_E Hitting time started at x stopped at a domain E

$\widehat{t}$ Rescaled time. As the graph grows large, neigbhorhoods shrink and so a single step in the random-walk moves ever smaller distances. $\widehat{t}$ is re-scaled by g_n^2 to prevent this from happenig (See entry for g_n^2).

Spaces

$\mathbb{R}^d$ d-dimensional Euclidean space

$\overline{D}$ Closure of the domain

∂D Boundary of the domain

$D \subset \mathbb{R}^d$ Compact, path-connected domain

D Skorokhod space (the space of right-continuous functions)

$\mathcal{X}$ Set of discrete objects (vertices or words)

$\mathbb{Z}$ Integers

C Continuous functions

$\mathsf{C}([0,T],\mathbb{R}^d)$ Continuous functions mapping $[0,T] \to \mathbb{R}^d$

$W_2(x,y)$ 2-Wasserstein distance between x and y

Chapter 1

Introduction and background

Discrete-time random walks lie at the heart of many domains in machine learning, ranging from graph algorithms to biological processes. This thesis unifies many algorithmic approaches for inference on random walks by approximating discrete-time and space random walks using diffusion processes. Much like how a random walk over a lattice converges to a Brownian motion, we exploit the fact that random walks over large domains behave as diffusions.

Many empirically effective random-walk based algorithms over discrete structures, such as the PageRank algorithm for graph vertex importance and word2vec for word embeddings are poorly understood from a statistical point of view. The diffusion process based techniques in this thesis develops connections between classical statistical ideas such as density estimation or manifold learning, and discrete algorithms such as PageRank and word embeddings. These connections allow us to understand new, empirically effective algorithms through the lens of parameter estimation for a well-specified metric model.

In this chapter, we will first introduce the ideas behind diffusion based random walk inference as well as the basic intuition for the diffusion limit. For now we will focus on intuition and defer precise technical discussions to Chapter 2.

1.1 Introduction

Structured non-metric inputs such as DNA sequences, social networks and natural language texts pose difficulties for traditional statistical and machine learning techniques which are designed to process real valued vectors. One approach for rigorously analyzing such structured data has been to infer latent embeddings which are assumed to be independent and identically distributed.

For example, if we are asked to determine the most important vertices of a social network, we may first assign a vector to each vertex in the graph such that connected vertices are close to each other, and then return the modes of the density of these embedded vertices as important vertices.

However, these embedding methods have thus far been ad-hoc and domain specific. For example, the techniques for embedding networks [Alamgir and von Luxburg, 2011] and English words [Mikolov et al., 2013a] rely upon completely different justifications. We show that there is a straightforward, unified approach to understanding and estimating latent metric structures by analyzing embeddings that exploit random walk structure. These techniques are simple, widely applicable and answer several open questions in nonparametric statistics and machine learning.

Conceptually, our work consists of three main ideas:

1. Random walks over discrete objects encode metric structure by frequently transitioning between more similar words or graph vertices.

2. Metric random walks over large spaces behave similarly to diffusion processes which are simple to analyze

3. Many algorithms such as PageRank, or word2vec for word embeddings, can be reduced to methods to infer properties of the underlying metric random walk.

We begin by analyzing the problem of performing density estimation given only the undirected connectivity structure of a k-nearest neighbor graph, and use this to develop the theory of diffusion process limits in Chapter 2. Having established a basic framework for connecting random walks and diffusions, we solve the problem of fully recovering a latent embedding from a geometric graph using weighted hitting times in Chapter 3. Applying these methods to a practical machine-learning problem, we show that word embeddings are a special type of scalable weighted graph embedding in Chapter 4.

Finally, having established results for discrete random walks and their relationship to diffusions, we show that a particular type of recurrent neural network naturally generates diffusion processes. Using this connection we analyze a problem of modeling the differentiation of embryonic stem cells using single-cell RNA-seq (Chapter 5).

1.1.1 Graph vertex embeddings

Many graph algorithms rely upon an implicit assumption of a **metric graph**, which is a graph where latent attributes of a vertex determine its connectivity. For example, a k-nearest neighbor graph is a type of metric graph, and finding similar vertices using the number of common neighbors attemps to recover latent vertex similarities.

One instance of such a metric assumption is the assumption that a simple random walk over the graph will frequently transition to similar vertices. For example, if our graph is the integer lattice, then a random walk will remain 'close' over short timespans. Our goal is to formalize the implicit metric assumption, and then to derive results under this assumption.

Informally, our metric assumption will be the following:

1. Each graph vertex is assigned a coordinate in a metric space

2. Graph vertices connects to neighbors that are sufficiently close

3. The graph is sufficiently large

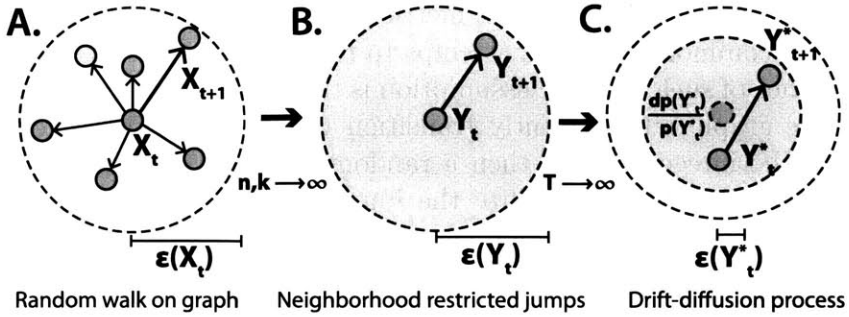
The simplest example of such a metric graph is the integer lattice, where a vertex has a multidimensional integer coordinate, and is connected to neighbors within a distance of one. We will prove a surprising property of metric graphs: simple random walks on metric graphs converge to diffusion processes.

The convergence of a simple random walk over a lattice to a Brownian motion is a classic diffusion limit. For example, if $X_i \sim \text{Bernoulli}(0.5)$ then $S_n = \sum_i X_i$ is the one-dimensional simple random walk after n steps, and as n grows this simple random walk behaves as a Brownian motion, $n^{-1/2}S_{tn} \rightarrow W_t$ [Lawler and Limic, 2010, Chapter 3.1].

In the case of a general metric graph, a similar phenomenon holds. Figure 1-1 is an informal sketch of this convergence to a diffusion process for a k-nearest neighbor graph. As the graph grows large, the discrete-space and discrete-time random walk over a graph (X_t) resembles a continuous-space discrete time process Y_t (Figure 1-1, Panels A and B). After a time-rescaling which considers many discrete steps per unit time we obtain a Gaussian limit which turns our random walk into a diffusion process (Figure 1-1, Panels C and D).

A surprising result is that not only does this diffusion process limit exist under very general conditions, we show that the limiting process can be written down in closed form and does not depend on the precise connectivity rules used to construct the graph (Theorem 2.3.3).

The main idea of our graph diffusion limit is that even though we do not observe the latent metric, the diffusion limit can reveal the underlying metric properties of the graph. For example, in Chapter 2 we show that the stationary distribution of a discrete random walk can be interpreted as a type of density estimation procedure over the latent metric of a graph, and in Chapter 3 we show that a exponentially



Random walk on graph Neighborhood restricted jumps Drift-diffusion process

Figure 1-1: An informal description of the convergence of random walks to diffusions. In panel A) we have a simple random walk over a graph. In B), as the size of the graph (n) and the number of points in each neighborhood (k) grows we can jump anywhere within the ball. In C), if we take many steps the sum of many independent jumps begins to resemble a Gaussian limit.

weighted hitting time (termed the log-LTHT) can be used to robustly recover the latent metric structure of a graph.

1.1.2 Word embeddings

Word embeddings are a class of techniques which use the co-occurrence of words across a large text corpus in order to assign a vector representation to each word. Generally, word embedding methods ensure that frequently co-occurring words are close in the embedded space.

These simple un-supervised representations were found to capture surprising semantic structures such as analogical reasoning [Mikolov et al., 2013b]. For example, if we define the vector for the word 'man' as vec(man) and apply the following operations: vec(king)-vec(man)+vec(woman)≈vec(queen).

Word embeddings have been remarkably useful across NLP tasks but remain poorly understood. The main approach to understanding the semantic properties of word embeddings have been to examine the properties of text corpora, which has lead to a very narrow view of semantic embeddings as being driven by text co-occurrences. Our goal in this thesis is to develop a theory of word embeddings that can be used to generalize semantically meaningful vector representations beyond natural language processing.

In order to achieve this goal, we ground word embeddings in semantic spaces studied in the cognitive-psychometric literature, taking these spaces as the primary objects to recover. To this end, we relate log co-occurrences of words in large corpora to semantic similarity assessments and show that co-occurrences are indeed consistent with an Euclidean semantic space hypothesis.

Fundamentally, we take *metric recovery* as the key theoretical goal. This perspective unifies existing word embedding algorithms, ties them to manifold learning, and demonstrates that existing algorithms are consistent metric recovery methods of co-occurrence counts from random walks.

To formalize this, we use the diffusion limits developed in Chapter 2 as a way to analyze the observed co-occurrence of words in a corpus. We show that log co-occurrences between vertices on a graph with a latent metric converge to the squared distance in the latent metric.

Further, we propose a simple, principled direct metric recovery algorithm that is comparable to the state-of-art in both word embedding and manifold learning. Finally, we complement recent focus on analogies by constructing two new inductive reasoning datasets – series completion and classification – and demonstrate that word embeddings can be used to solve them as well.

1.1.3 Recurrent networks as diffusions

Using many of the diffusion process techniques, we analyze a central problem in developmental biology: understanding the forces which control gene expression changes during differentiation.

We model the gene expression of a single cell over differentiation as a high-dimensional diffusion process which is stopped and observed at a given time point through single-cell RNA-seq. This problem is particularly challenging since RNA-seq measurement kills the cell, preventing longitudinal trajectory measurements of individuals.

We show that cross-sectional samples from an evolving population suffice for recovery within a class of processes even if samples are available only at a few distinct time points. We provide a stratified analysis of recoverability conditions, and establish that reversibility is sufficient for recoverability. This reversibility condition is closely related to a concept known as the "epigenetic landscape" in developmental biology, and we show that the landscape can be recovered from existing single-cell RNA-seq data.

For estimation, we derive a natural loss and regularization, and parameterize the processes as diffusive recurrent neural networks. Using our diffusion limit, we show that a particular recurrent neural network architecture naturally models reversible diffusions, and use this as a way to develop fast, scalable inference for single-cell RNA-seq data.

1.1.4 Prior publication

This thesis is derived from several prior publications with collaborators who have been invaluable to developing the results in this thesis.

The graph diffusion process limits in Chapters 2, 3 are expanded versions of Hashimoto et al. [2015c,a] with corrections for typographical errors and is based on joint work with Yi Sun and Tommi Jaakkola. Chapter 4 specializes these limits for word embeddings and is based upon work in submission done jointly with David Alvarez-Melis and Tommi Jaakkola. Chapter 5 applies diffusion limits to developmental biology and is joint work with David Gifford and Tommi Jaakkola.

Chapter 2

Density estimation on metric graphs

Data for unsupervised learning is increasingly available in the form of graphs or networks. For example, we may analyze gene networks, social networks, or general co-occurrence graphs (*e.g.*, built from purchasing patterns). While classical unsupervised tasks such as density estimation or clustering are naturally formulated for data in vector spaces, these tasks have analogous problems over graphs such as centrality and community detection. We provide a step towards unifying unsupervised learning by recovering the underlying density and metric directly from graphs.

We consider "unweighted directed geometric graphs" that are assumed to have been built from underlying (unobserved) points x_i, $i = 1, \ldots, n$. In particular, we assume that graphs are formed by drawing an arc from each vertex i to its neighbors within distance $\varepsilon_n(x_i)$. Note that the graphs are typically not symmetric since the distance (the ε_n-ball) may vary from point to point. By allowing $\varepsilon_n(x_i)$ to be stochastic, *e.g.*, depend on the set of points, the construction subsumes also typical k-nearest neighbor graphs. Arguably, graphs from top k friends/products, or co-association graphs may also be approximated in this manner.

The key property of our family of geometric graphs is that their structure is completely characterized by two functions over the latent space: the local density $p(x)$ and the local scale $\varepsilon(x)$. Indeed, global properties such as the distances between points can be recovered by integrating these quantities. We show that asymptotic behavior of random walks on the directed graphs relate to the density and metric. In particular, we show that random walks on such graphs with minimal degree at least $\omega(n^{2/(2+d)} \log(n)^{\frac{d}{d+2}})$ can be completely characterized in terms of p and ε using drift-diffusion processes. This enables us to recover both the density and distance given only the observed graph and the (hypothesized) underlying dimension d.

The fact that we may recover the density (up to constant scale) is surprising. For example, in k-nearest neighbor graphs, each vertex has degree exactly k. There is no immediate local information about the density, *i.e.*, whether the corresponding point lies in a high-density region with small ball radii, or in a low-density region with large ball radii. The key insight of this paper is that random walks over such graphs naturally drift toward higher density regions, allowing for density recovery.

While the paper is primarily focused on the theoretical aspects of recovering the metric and density, we believe our results offer useful strategies for analyzing real-world networks. For example, we analyzed the Amazon co-purchasing graph where an edge is drawn from an item i to j if j is among the top k co-purchased items with i. These Amazon products may be co-purchased if they are similar enough to be complementary, but not so similar that they are redundant. We extend our model to deal with connectivity rules shaped like an annulus, and demonstrate that our estimator can simultaneously recover product similarities, product categories, and central products by metric embedding.

2.1 Relation to prior work

The density estimation problem addressed by this paper was proposed and partially solved by von Luxburg and Alamgir [2013] using integration of local density gradients over shortest paths. This estimator has since been used for drawing graphs with ordinal constraints in von Luxburg and Alamgir [2013] and graph down-sampling in Alamgir et al. [2014]. However, the recovery algorithm is restricted to 1-dimensional k-nearest neighbor graphs under the constraint $k = \omega(n^{2/3}\log(n)^{\frac{1}{3}})$. Our paper provides an estimator that works in all dimensions, applies to a more general class of graphs, and strongly outperforms that of von Luxburg-Alamgir in practice.

On a technical level, our work has similarities to the analysis of convergence of graph Laplacians and random walks on manifolds Woess [1994], Hein et al. [2006]. For example, Ting et al. [2010b] used infinitesimal generators to capture the convergence of a discrete Laplacian to its continuous equivalent on k-nearest neighbor graphs. However, their analysis was restricted to the Laplacian and did not consider the latent recovery problem. In addition, our approach proves convergence of the entire random walk trajectory and allows us to analyze the stationary distribution function directly.

2.2 Main results and proof outline

2.2.1 Problem setup

Let $\mathcal{X} = \{x_1, x_2, \ldots\}$ be an infinite sequence of latent coordinate points drawn independently from a distribution with probability density $p(x)$ in $\mathbb{R}^d$. Let $\varepsilon_n(x_i)$ be a radius function which may depend on the draw of $\mathcal{X}$. In this paper, we analyze a single draw of $\mathcal{X}$. Let $G_n = (\mathcal{X}_n, E_n)$ be the unweighted directed neighborhood graph with vertex set $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ and with a directed edge from i to j if and only if $|x_i - x_j| < \varepsilon_n(x_i)$.

Fix now a large n. We consider the random directed graph model given by observing the single graph G_n. The model is completely specified by the latent function $p(x)$ and the possibly stochastic $\varepsilon_n(x)$. Under the conditions $(\star)$ to be specified below, we solve the following problem:

> Given only G_n and d, form a consistent estimate of $p(x_i)$ and $|x_i - x_j|$ up to proportionality constants.

In the case that the graph is disconnected, we will recover the corresponding quantity up to scaling for each separate connected component.

The conditions we impose on $p(x)$, $\varepsilon_n(x)$, and the stationary density function $\pi_{X_n}(x)$ of the simple random walk $X_n(t)$ on G_n are the following, which we refer to as $(\star)$. We assume $(\star)$ holds throughout the paper.

- The density $p(x)$ is differentiable with bounded $\nabla \log(p(x))$ on a path-connected compact domain $D \subset \mathbb{R}^d$ with smooth boundary ∂D.

- There is a deterministic continuous function $\overline{\varepsilon}(x) > 0$ on $\overline{D}$ and scaling constants g_n satisfying

$$g_n \to 0 \text{ and } g_n n^{\frac{1}{d+2}} \log(n)^{-\frac{1}{d+2}} \to \infty$$

  so that, a.s. in the draw of $\mathcal{X}$, $g_n^{-1}\varepsilon_n(x)$ converges uniformly to $\overline{\varepsilon}(x)$.

- The rescaled density functions $n\pi_{X_n}(x)$ are a.s. eventually uniformly equicontinuous.

Remark. We conjecture that the last condition in $(\star)$ holds for any p and $\overline{\varepsilon}$ satisfying the other conditions in $(\star)$ (see Conjecture A.1.1).

Let $\mathsf{NB}_n(x)$ denote the set of out-neighbors of x so that y is in $\mathsf{NB}_n(x)$ if there is a directed edge from x to y. The second condition in $(\star)$ implies for all $x \in \mathcal{X}_n$ that

$$|\mathsf{NB}_n(x)| = \omega(n^{\frac{2}{d+2}} \log(n)^{\frac{d}{d+2}}). \tag{2.1}$$

2.2.2 Statement of results

Our approach is based on the simple random walk $X_n(t)$ on the graph G_n. Let $\pi_{X_n}(x)$ denote the stationary density of $X_n(t)$. We first show that when appropriately renormalized, $\pi_{X_n}(x)$ converges to an explicit function of $p(x)$ and $\overline{\varepsilon}(x)$ up to a scaling constant c.

Theorem 2.2.1. *Given $(\star)$, a.s. in $\mathcal{X}$, we have*

$$n\pi_{X_n}(x) \to c\frac{p(x)}{\overline{\varepsilon}(x)^2}, \tag{2.2}$$

for the normalization constant $c^{-1} = \int p(x)^2 \overline{\varepsilon}(x)^{-2} dx$.

Combining this result with an estimate on the out-degree of points in G_n gives our general result on recovery of density and scale. Let V_d be the volume of the unit d-ball.

Corollary 2.2.2 (Density and metric estimator). *Assuming $(\star)$, the estimators*

$$c_1 |\mathsf{NB}_n(x)|^{\frac{2}{d+2}} \pi_{X_n}(x)^{\frac{d}{d+2}} \to p(x) \text{ and}$$
$$c_2 |\mathsf{NB}_n(x)|^{\frac{1}{d+2}} \pi_{X_n}(x)^{-\frac{1}{d+2}} \to \overline{\varepsilon}(x) \text{ with}$$
$$c_1 = \left(\frac{n^{\frac{d-2}{d}}}{cV_d^{2/d} g_n^2}\right)^{\frac{d}{d+2}} \qquad c_2 = \left(\frac{1}{c^{d/2} V_d n^2 g_n^d}\right)^{\frac{1}{d+2}}$$

consistently recover the underlying density a.s. in $\mathcal{X}$ up to a scale depending on the normalizer c defined in (2.2) which cannot be identified from the graph alone.

Proof. Immediate from the out-degree estimate $p(x)\varepsilon_n(x)^d V_d = |\mathsf{NB}_n(x)|/n$ and Theorem 2.2.1. □

Remark. If $\varepsilon_n(x)$ is constant, every edge is bidirectional, so $\pi_{X_n}(x)$ is proportional to the degree of x, and we recover the standard ε-ball density estimator.

Our estimator for the density $p(x)$ closely resembles the PageRank algorithm without damping Page et al. [1999]. For the k-nearest neighbor graph, it gives the same rank ordering as PageRank and reduces to PageRank as $d \to \infty$.

For the k-nearest neighbor density estimation problem posed in von Luxburg and Alamgir [2013], we obtain the following.

Corollary 2.2.3. *If $\varepsilon_n(x)$ is selected via the k-nearest neighbors procedure with $k = \omega(n^{\frac{2}{d+2}} \log(n)^{\frac{d}{d+2}})$ and satisfies the first and last conditions in $(\star)$, for c_1 and c_2 depending on c as in Corollary 2.2.2 we have a.s in $\mathcal{X}$ that*

$$c_1 \pi_{X_n}(x)^{\frac{d}{d+2}} \to p(x) \text{ and}$$
$$c_2 \pi_{X_n}(x)^{-\frac{1}{d+2}} \to \bar{\varepsilon}(x).$$

Proof. By Devroye and Wagner [1977], the empirical $\varepsilon_n(x)$ induced by the k-nearest neighbors procedure satisfies the second condition of $(\star)$ with

$$\bar{\varepsilon}(x) = \frac{1}{V_d^{1/d} p(x)^{1/d}} \text{ and } g_n = (k/n)^{1/d}. \qquad \square$$

2.2.3 Outline of approach

Our proof proceeds via the following steps.

1. As $n \to \infty$, the simple random walk $X_n(t)$ on G_n converges weakly to an Itô process $Y(t)$, yielding weak convergence of stationary measures. (Theorem 2.3.4)
2. The stationary density $\pi_Y(x)$ is explicitly determined via Fokker-Planck equation. (Lemma 2.4.1)
3. Uniform equicontinuity of $n\pi_{X_n}(x)$ yields convergence in density after rescaling. (Theorem 2.2.1)

An intuitive explanation for our results is as follows. For large n, the simple random walk on G_n, when considered with its original metric embedding, closely approximates the behavior of a drift-diffusion process. Both the process and the approximating walk move preferentially toward regions where $p(x)$ is large and diffuse more slowly out of regions where $\bar{\varepsilon}(x)$ is small. Occupation times therefore give us information about $p(x)$ and $\bar{\varepsilon}(x)$ which allow us to recover them.

Formally, the convergence of $X_n(t)$ to $Y(t)$ follows by verifying the conditions of the Stroock-Varadhan criterion (Theorem A.2.3) for convergence of discrete time

Markov processes to Itô processes Stroock and Varadhan [1971a]. This criterion states that if a process reflects at the boundary and the variance a_n, expected value b_n, and higher order moments $\Delta_{n,\alpha}$ of a jump are continuous and well-controlled in the limit, the process converges to an Itô process. Via the Fokker-Planck equation, we can express the stationary density of this process solely in terms of $p(x)$ and $|\mathsf{NB}_n(x)|$. This allows us to estimate the density using only the unweighted graph.

Let $\overline{D}$ and ∂D be the closure and boundary of the support D of $p(x)$. Let $B(x, \varepsilon)$ be the ball of radius ε centered at x. Let $h_n = g_n^2$ be the time rescaling necessary for $X_n(t)$ to have equal timescale to $Y(t)$.

2.3 Convergence of the simple random walk to an Itô process

We will verify the regularity conditions of the Stroock-Varadhan criterion (see Stroock and Varadhan [1971a, Section 6]).

Theorem 2.3.1 (Stroock-Varadhan). *Let $X_n(t)$ be discrete-time Markov processes defined over a domain D with boundary ∂D. Define the discrete time drift and diffusion coefficients by*

$$a_n^{ij}(s,x) = \frac{1}{h_n} \sum_{y \in \mathsf{NB}_n(x)} \frac{1}{|\mathsf{NB}_n(x)|}(y_i - x_i)(y_j - x_j)$$

$$b_n^i(s,x) = \frac{1}{h_n} \sum_{y \in \mathsf{NB}_n(x)} \frac{1}{|\mathsf{NB}_n(x)|}(y_i - x_i)$$

$$\Delta_{n,\alpha}(s,x) = \frac{1}{h_n} \sum_{y \in \mathsf{NB}_n(x)} \frac{1}{|\mathsf{NB}_n(x)|}|y - x|^{2+\alpha}.$$

If we have $a_n^{ij}(s,x) \xrightarrow{a.s} a^{ij}(s,x)$, $b_n^i(s,x) \xrightarrow{a.s} b^i(s,x)$, $\Delta_{n,1}(s,x) \xrightarrow{a.s} 0$, and regularity conditions to ensure reflection at ∂D (Theorem A.2.2 and Theorem A.2.3), the time-rescaled stochastic processes $X_n(\lfloor t/h_n \rfloor)$ converge weakly in Skorokhod space $\mathsf{D}([0,\infty), \overline{D})$ to an Itô process with reflecting boundary condition

$$dY(t) = \sigma(t, Y(t))dW_t + b(t, Y(t))dt,$$

with W_t a standard d-dimensional Brownian motion and $\sigma(t, Y(t))\sigma(t, Y(t))^T = a(t, Y(t))$.

Remark. The original result of Stroock-Varadhan was stated for $\mathsf{D}([0,T],\overline{D})$ for all finite T; our version for $\mathsf{D}([0,\infty),\overline{D})$ is equivalent by Whitt [1980, Theorem 2.8].

The technical conditions of Theorem 2.3.1 enforcing reflecting boundary conditions are checked in Theorem A.2.8 to Theorem A.2.12. We focus on convergence of the drift and diffusion coefficients.

Lemma 2.3.2 (Strong LLN for local moments). *For a function $f(x)$ such that $\sup_{x \in B(0,\varepsilon)} |f(x)| < \varepsilon$, given $(\star)$ we have uniformly on $x \in \mathcal{X}_n$ that*

$$\frac{1}{h_n} \sum_{y \in \mathsf{NB}_n(x)} \frac{1}{|\mathsf{NB}_n(x)|} f(y-x) \\ \xrightarrow{a.s.} \frac{1}{h_n} \int_{y \in B(x,\varepsilon_n(x))} f(y-x) \frac{p(y)}{p_{\varepsilon_n(x)}(x)} dy.$$

Proof. Denote the claimed value of the limit by $\mu(x)$. For convergence in expectation, we condition on $|\mathsf{NB}_n(x)|$ and apply iterated expectation to get

$$E\left[\frac{1}{h_n} \sum_{y \in \mathsf{NB}_n(x)} \frac{1}{|\mathsf{NB}_n(x)|} f(y-x)\right] \\ = E\left[\frac{1}{h_n} E\left[f(y-x) \middle| |\mathsf{NB}_n(x)|\right]\right] = \mu(x).$$

For $y \in B(x, \varepsilon_n(x))$, we have $|f(y-x)| \leq \varepsilon_n(x)$, so Hoeffding's inequality yields

$$\begin{aligned} P\left(\left|\frac{1}{h_n} \sum_{y \in \mathsf{NB}_n(x)} \frac{1}{|\mathsf{NB}_n(x)|} f(y-x) - \mu(x)\right| \geq t\right) \\ &\leq 2\exp\left(-\frac{2h_n^2 |\mathsf{NB}_n(x)|^2 t^2}{|\mathsf{NB}_n(x)| \varepsilon_n(x)^2}\right) \\ &= \Theta\left(\exp\left(-2g_n^2 \overline{\varepsilon}(x)^{-2} |\mathsf{NB}_n(x)| t^2\right)\right) \qquad (2.3) \\ &= o(n^{-2t^2\omega(1)}) \end{aligned}$$

for $|\mathsf{NB}_n(x)| = \omega\left(n^{2/(d+2)} \log(n)^{d/(d+2)}\right)$ by (2.1). Borel-Cantelli then yields a.s. convergence. □

Remark. This limit holds for stochastic $\varepsilon_n(x)$ if $g_n^{-1}\varepsilon_n(x)$ a.s. converges uniformly to a deterministic continuous $\overline{\varepsilon}(x)$. An example of such a graph is the k-nearest neighbors graph.

Theorem 2.3.3 (Drift diffusion coefficients). *Almost surely on the draw of $\mathcal{X}$, as $n \to \infty$, we have*

$$\begin{aligned} \lim_{n\to\infty} a_n^{ij}(s,x) &= \delta_{ij}\frac{1}{3}\overline{\varepsilon}(x)^2 \\ \lim_{n\to\infty} b_n^{i}(s,x) &= \frac{\partial_i p(x)}{3p(x)}\overline{\varepsilon}(x)^2 \\ \lim_{n\to\infty} \Delta_{n,1}(s,x) &= 0, \end{aligned}$$

where δ_{ij} is the Kronecker delta function.

Proof. By Lemma 2.3.2, a_n, b_n, and $\Delta_{n,1}$ converge a.s. to their expectations, so it suffices to verify that the integrals in Lemma 2.3.2 have the claimed limits. Because p is differentiable on D, for any $x \in D$ we have the Taylor expansion

$$p(x+y) = p(x) + y \cdot \nabla p(x) + o(|y|^2)$$

of p at x, where the convergence is uniform on compact sets. For n large so that $B(x, \varepsilon_n(x))$ lies completely inside D, substituting this expansion into the definitions of a_n, b_n, and $\Delta_{n,1}$ and integrating over spheres yields the result. Full details are in Theorem 2.3.3. □

Theorem 2.3.4. *Under ($\star$), as $n \to \infty$ a.s. in the draw of $\mathcal{X}$ the process $X_n(\lfloor t/h_n \rfloor)$ converges in $\mathsf{D}([0,\infty), \overline{D})$ to the isotropic $\overline{D}$-valued Itô process $Y(t)$ with reflecting boundary condition defined by*

$$dY(t) = \frac{\nabla p(Y(t))}{3p(Y(t))}\overline{\varepsilon}(Y(t))^2 dt + \frac{\overline{\varepsilon}(Y(t))}{\sqrt{3}} dW(t). \tag{2.4}$$

Proof. Lemma 2.3.2 and Theorem 2.3.3 show that $X_n(\lfloor t/h_n \rfloor)$ fulfills the conditions of Theorem A.2.3. The result follows from the Stroock-Varadhan criterion using the drift and diffusion terms from Theorem 2.3.3. □

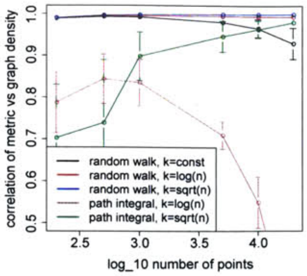


Figure 2-1: Accuracy vs sample and neighborhood size. Path integral (green, maroon) is from von Luxburg and Alamgir [2013]. Our estimator (red, blue, black) is nearly perfect at all sample sizes and neighborhood sizes.

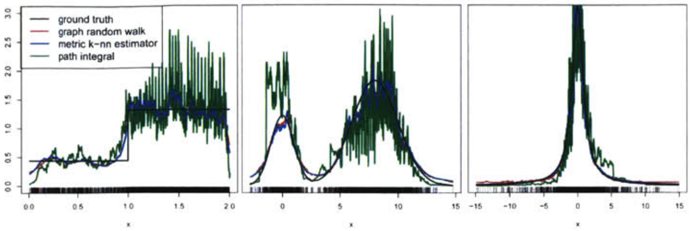


Figure 2-2: Examples of four density estimates: our method (red) using no metric information is indistinguishable from metric k-nearest neighbor (blue) and close to ground truth (black). Path integral estimator of von Luxburg and Alamgir [2013] (green) shows higher error in all cases.

2.4 Convergence and computation of the stationary distribution

2.4.1 Graphs satisfying condition $(\star)$

The Itô process $Y(t)$ is an isotropic drift-diffusion process, so the Fokker-Planck equation [Risken, 1984] implies its density $f(t, x)$ at time t satisfies

$$\partial_t f(t,x) \quad = \quad \sum_i \Big(\; - \; \partial_{x_i}[b^i(t,x)f(t,x)] \; + \; \frac{1}{2}\partial_{x_i^2}[a^{ii}(t,x)f(t,x)]\Big), \quad (2.5)$$

where $b^i(t, x)$ and $a^{ii}(t, x)$ are given by

$$b(t,x) = \frac{\nabla p(x)}{3p(x)}\bar{\varepsilon}(x)^2 \text{ and } a^{ii}(t,x) = \frac{1}{3}\bar{\varepsilon}(x)^2.$$

Lemma 2.4.1. *The process $Y(t)$ defined by (2.4) has absolutely continuous stationary measure with density*

$$\pi_Y(x) = cp(x)^2\bar{\varepsilon}(x)^{-2},$$

where c was defined in (2.2).

Proof. By (2.5), to check that $\pi_Y(x) = cp(x)^2\bar{\varepsilon}(x)^{-2}$, it suffices to show

$$\sum_i \left(\partial_{x_i} p(x) \left(p(x)^{-1}\bar{\varepsilon}(x)^2 c\frac{p(x)^2}{\bar{\varepsilon}(x)^2} \right) - \frac{1}{2}\partial_{x_i} \left(\bar{\varepsilon}(x)^2 c\frac{p(x)^2}{\bar{\varepsilon}(x)^2} \right) \right) = 0. \quad \square$$

We now prove Theorem 2.2.1 by showing that a rescaling of $\pi_{X_n}(x)$ converges to $\pi_Y(x)$.

Proof of Theorem 2.2.1. The a.s. convergence of processes of Theorem 2.3.4 implies by Ethier and Kurtz [1986, Theorem 4.9.12] that the empirical stationary measures

$$d\mu_n = \sum_{i=1}^{n} \pi_{X_n}(x_i)\delta_{x_i}$$

converge weakly to the stationary measure $d\mu = \pi_Y(x)dx$ for $Y(t)$. For any $x \in \mathcal{X}$ and $\delta > 0$, weak convergence against $1_{B(x,\delta)}$ yields

$$\sum_{y \in \mathcal{X}_n, |y-x|<\delta} \pi_{X_n}(y) \to \int_{|y-x|<\delta} \pi_Y(y)dy.$$

By eventual uniform equicontinuity of $n\pi_{X_n}(x)$, for any $\varepsilon > 0$ there is small enough $\delta > 0$ so that for all n we have

$$\left| \sum_{y \in \mathcal{X}_n, |y-x|<\delta} \pi_{X_n}(y) - |\mathcal{X}_n \cap B(x,\delta)|\pi_{X_n}(x) \right| \leq n^{-1}|\mathcal{X}_n \cap B(x,\delta)|\varepsilon,$$

which implies that

$$\begin{aligned}
&\lim_{n\to\infty} \pi_{X_n}(x)p(x)n \\
&\quad = \lim_{\delta\to 0}\lim_{n\to\infty} V_d^{-1}\delta^{-d} n\pi_{X_n}(x) \int_{|y-x|<\delta} p(y)dy \\
&\quad = \lim_{\delta\to 0}\lim_{n\to\infty} V_d^{-1}\delta^{-d} |\mathcal{X}_n \cap B(x,\delta)|\pi_{X_n}(x) \\
&\quad = \lim_{\delta\to 0} V_d^{-1}\delta^{-d} \int_{|y-x|<\delta} \pi_Y(y)dy = \pi_Y(x).
\end{aligned}$$

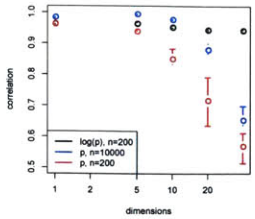


Figure 2-3: Estimate performance degrades in high dimensions due to over-smoothing (blue and red), but the estimator is still highly accurate up to log concentration parameter (black).

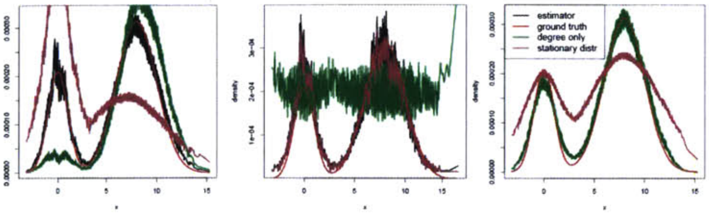


Figure 2-4: Example isotropic metric graphs. Our estimator (black) agrees with the true density (red) in all cases. Degree and stationary distribution (green and maroon) based density estimates work for some cases (right two panels) but cannot work if the degree is tied to spatial location (left).

Combining with Lemma 2.4.1 yields the desired

$$\lim_{n\to\infty} n\pi_{X_n}(x) = \frac{\pi_Y(x)}{p(x)} = c\frac{p(x)}{\overline{\varepsilon}(x)^2}. \qquad \square$$

2.4.2 Extension to general metric graphs

To obtain our stationary distribution in Theorem 2.2.1 we require only convergence to some Itô process via the Stroock-Varadhan criterion. We can achieve this under substantially more general conditions. We define a class of metric graphs on $\mathcal{X}_n$ termed *isotropic* over which we have consistent metric recovery without knowledge of the graph construction method.

Definition 1 (Isotropic metric graph). A graph edge connection procedure on $\mathcal{X}_n$ is isotropic if it satisfies:

Distance kernel: The probability of placing a directed edge from i to j is defined by a kernel function $h(r_{ij})$ mapping locally scaled distances

$$r_{ij} = |x_i - x_j|\varepsilon_n(x_i)^{-1}.$$

with $\varepsilon_n(x)$ obeying ($\star$).

Nonzero mass: The kernel function $h(r)$ has nonzero integral $\int_0^1 h(r)r^{d-1}dr > 0$.

Bounded tails: For all $r > 1$, $h(r) = 0$.

Continuity: The scaling $n\pi_{X_n}(x)$ of the stationary distribution is eventually uniformly equicontinuous.

This class of graph preserves the property that the random graph is entirely determined by the underlying density $p(x)$ and local scale $\overline{\varepsilon}(x)$; this allows us to have the same tractable form for the stationary distribution.

Both constant ε and k-nearest neighbor graphs are isotropic upon assumption of eventual uniform equicontinuity. Another interesting class of graphs allowed by this generalization is truncated Gaussian kernels, where connectivity probability decreases exponentially. Note that $h(r)$ might not be monotonic or continuous in r; one surprising example is $h(r) = 1_{[0.5,1]}(r)$, which deterministically connects points in an annulus.

Corollary 2.4.2 (Generalization). *If a neighborhood graph is isotropic, then the limiting stationary distribution follows Theorem 2.2.1, and the density and distances can be estimated by Corollary 2.2.2.*

Proof. We check the Stroock-Varadhan condition stated in Theorem A.2.3. For this, we use a version of Lemma 2.3.2 for isotropic graphs, which requires that the ball radius vanishes and that the neighborhood size scales as $\omega(n^{\frac{2}{d+2}} \log(n)^{\frac{d}{d+2}})$.

Vanishing neighborhood radius follows because bounded tails and the fact that the kernel is evaluated on $|x_i - x_j| \varepsilon_n(x_i)^{-1}$ ensure the isotropic graph is a subgraph of the $\varepsilon_n(x)$-ball graph. Kolmogorov's strong law implies that the stochastic out-degree concentrates around its expectation. It has the correct scaling because the argument of $h(r)$ is scaled by $\varepsilon_n(x)$. See Theorem A.3.2 for details. Thus the analogue of Lemma 2.3.2 holds.

We then check that the limiting local moments for isotropic graphs are proportional to those of $\varepsilon_n(x)$-ball graphs in Lemma A.3.3. All but one of the conditions for the Stroock-Varadhan criterion follow from this; the last Theorem A.2.11 follows from the bounded ball structure of the connectivity kernel.

To check that we obtain the same limiting process and stationary measure, note the ratios of integrals in Theorem 2.3.3 are unchanged in the isotropic setting. See Lemma A.3.3 for details. Recovering the stationary distribution, density, and local scale is then done in the same manner as in the ε-ball setting. □

2.5 Distance recovery via shortest paths

Our results in Theorem 2.2.1 give a consistent estimator for the density $p(x)$ and the local scale $\overline{\varepsilon}(x)$. These two quantities specify up to isometry and scaling the latent metric embedding of $\mathcal{X}$.

In order to reconstruct distances between non-neighbor points, we weight the edges of G_n by $w_{ij} = \varepsilon_n(x_i)$ and find the shortest paths over this graph, which we call $\overline{G}_n$. The results in Alamgir and von Luxburg [2012a, Section 4.1] show that in the k-nearest neighbor case, setting $w_{ij} = \widehat{\varepsilon}_n(x_i)$ for the estimator $\widehat{\varepsilon}_n$ of ε_n results in consistent recovery of pairwise distances.

In Theorem A.4.5, we give a straightforward extension of this approach to show that given any uniformly convergent estimator of $\varepsilon_n(x)$, the shortest path on the weighted graph $\overline{G}_n$ converges to the geodesic distance. Applying standard metric multidimensional scaling then allows us to embed these distances and recover the latent space up to isometry and scaling.

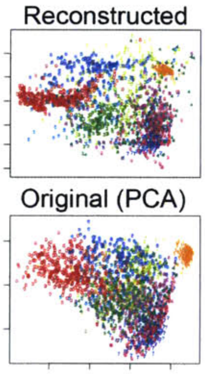


Figure 2-5: Reconstruction closely matches projection of the true metric.

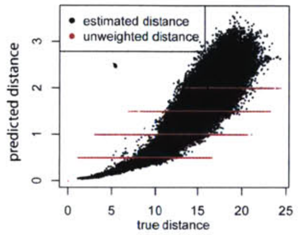


Figure 2-6: Distances estimated by our method are globally close to the true metric.

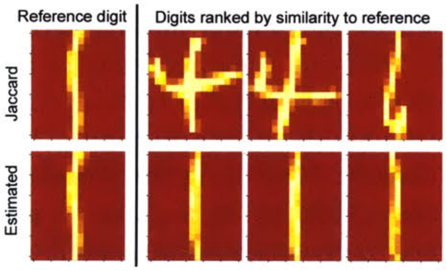


Figure 2-7: Items close in our weighted graph (bottom) are more similar than those under the Jaccard index (top).

2.6 Empirical results

We demonstrate extremely good finite sample performance of our estimator in simulated density reconstruction problems and two real-world datasets. Some details such as exact graph degrees and distribution parameters are in the supplementary code which reproduces all figures in this paper. Standard graph statistics such as centrality and Jaccard index are calculated via the `igraph` package Csardi and Nepusz [2006].

k-nearest neighbor graphs We compared our random-walk based estimator and the path-integral based estimator in von Luxburg and Alamgir [2013] to the metric k-nearest neighbor density estimator. The number of samples n was varied from 100 to 20000 along with the sparsity level k (Figure 2-1).

While our theoretical results suggest that both our algorithm and the path-integral estimator of von Luxburg and Alamgir [2013] might fail to converge at $\sqrt{n}$ and $\log(n)$ sparsity levels, in practice our estimator performs nearly perfectly at both low sparsity levels.

For constant degree $k = 50$ we achieve near-perfect performance for all choices of n, while the path-integral estimator fails to converge in the $k = \log(n)$ regime.

Some specific examples of our density estimator with $n = 2000, k = 100$ are shown in Figure 2-2. The examples are mixture of uniforms (left), mixture of Gaussians (center), and t-distribution (right). As predicted, our estimator tracks extremely closely with the metric k-nearest neighbor estimator (red and blue), as well as the true density (black). The path integral estimator has high estimate variance at points with large density and fails to cope with the two mixture densities.

Varying the dimension for an isotropic multivariate normal with $k = \sqrt{n}$, we find that a large number of points are required to maintain high accuracy as d grows large (red and blue lines in Figure 2-3). However, this is due to a global 'flattening' of the density. Measuring the correlation between the true and estimated log probabilities show that up to a global concentration parameter, the estimator maintains high accuracy across a large number of dimensions (black lines).

Kernel graphs We validate the nonparametric estimator in Corollary 2.4.2 for kernel graphs by constructing three different kernel graphs. In all cases, we sample 5000 points with connection probability following $p_{i,j} = \exp(-\varepsilon(x_i)^{-1}|x_i - x_j|)$. We vary the neighborhood structure ε in three ways: a constant kernel, $\varepsilon(x_i) \propto 1$; k-nearest neighbor kernel: $\varepsilon(x) \propto 1/\varepsilon_{k=100}$; and spatially varying kernel $\varepsilon(x) \propto |x|$.

In Figure 2-4, we find that our nonparametric estimator (black) always matches the ground truth (red). This example also shows that both degree and stationary distribution can be valid density estimators under certain assumptions, but only our estimator can deal with arbitrary isotropic graph construction methods without knowledge of the graph construction technique.

Metric recovery on real data As an example of metric reconstruction, we take the first 2000 examples in the U.S. postal service (USPS) digits dataset Hull [1994] and construct an unweighted k-nearest neighbor graph. We use our method to reconstruct the metric and perform similarity queries, and the Jaccard index was used to tie-break direct neighbors.

The USPS digits dataset is known to have a high-density cluster of ones digits (orange). Results in Figure 2-5 show that we are able to successfully recover the density structure of the data (top). Inter-point distances estimated by our method (Figure 2-6, y-axis) show nearly linear agreement to the true metric (x-axis) at short distances and high similarity globally.

Performing a similarity query on the data (Figure 2-7) shows that the our reconstructed distances (bottom row) have a more coherent set of similar digits when compared to the Jaccard index (top row) Jaccard [1901]. The behavior of the unweighted Jaccard similarity is due to a known problem with shortest paths in k-nearest neighbor graphs preferring low density regions von Luxburg and Alamgir [2013].

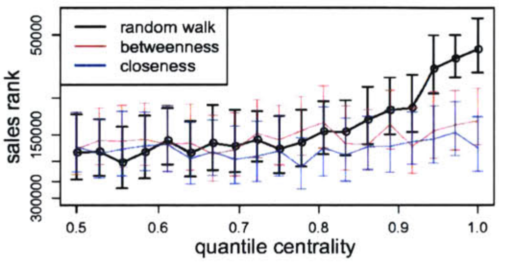


Figure 2-8: Density estimates in the graph correlate well with sales rank, unlike the other measures of centrality.

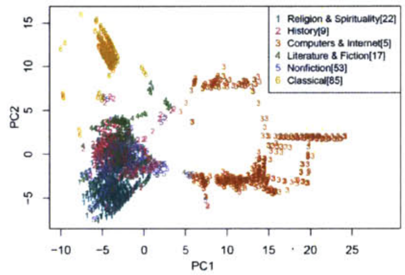


Figure 2-9: Embeddings from estimated distances recover the separation between different product categories.

Classics	Literature	Classical music	Philosophy
The Prince	The Stranger	Beethoven: Symphonien Nos. 5 &	The Practice of Everyday Life
The Communist Manifesto	The Myth of Sisyphus	Mozart: Symphonies Nos. 35-41	The Society of the Spectacle
The Republic	The Metamorphosis	Mozart: Violin Concertos	The Production of Space
Wealth of Nations	Heart of Darkness	Tchaikovsky: Concerto No. 1/Rac	Illuminations
On War	The Fall	Beethoven: Symphonies Nos. 3 &	Space and Place: The Perspectiv

Table 2.1: Top 4 clusters formed by mapping each item to its mode (first row). Each group is a coherent genre.

Amazon co-purchasing data Finally, we recover density and metric on a real network dataset with no ground truth. We analyzed the largest connected component of the Amazon co-purchasing network dataset ($n = 7175$, $k = 21804$) Leskovec et al. [2007]. Each vertex is a product on `amazon.com` along with its category and sales rank, and each directed edge represents a co-purchasing recommendation of the form "person who bought x also bought y." This dataset naturally fulfills our assumptions of having edges that are asymmetric, where edges represent a notion of similarity in some space.

The items that lie in regions of highest density should be archetypal products for a category, and therefore be more popular. We show that the density estimates using our method with $d = 10$ show a strong positive association between density and sales (Figure 2-8). We found that this effect persisted regardless of choice of d. Other popular measures of network centrality such as betweenness and closeness fail to display this effect.

We then attempted metric recovery using our random walk based reconstruction (Figure 2-9). For visualization purposes, we used classical multidimensional scaling on the recovered metric to embed points belonging to categories with at least two hundred items. The embedding shows that our method captures separation across different product categories. Notably, nonfiction and history have substantial overlap as expected, while classical music CD's and computer science books have little overlap with the other clusters.

Analyzing the modes of the density estimate by clustering each point to its local mode, we find coherent clusters where top items serve as archetypes for the cluster (Table 2.1). This suggests that there may be a close connection between clustering in a metric space and community detection in network data. The overall performance of our method on density estimation and metric recovery for the Amazon dataset suggests that when a metric assumption is appropriate, our random walk based metric quantities can be used directly for centrality and cluster estimates on a network.

2.7 Conclusions

We have presented a simple explicit identity linking the stationary distribution of a random walk on a neighborhood graph to the density and neighborhood size.

The density estimator constructed by inverting this identity matches the metric k-nn density estimate with $r > 0.95$ at $\log(n)$ degree with as few as a hundred points (Figures 2-1,2-2). We also generalized the theorem to a large class of graph construction techniques and demonstrated that the choice of construction technique matters little for accuracy (Figures 2-4).

Our estimator performed well on real-world data, recovering underlying metric information in test data (Figures 2-6,2-7) and predicting popular Amazon products through density estimates (Figure 2-8).

There are several open questions left unanswered by our work. Our results required that the graphs be of degree $k = \omega(n^{2/(d+2)} \log(n)^{d/(d+2)})$ rather than the

$\log(n)$ required for connectivity. Our simulation results suggest that even near the $\log(n)$ regime our estimator performs similarly to the dense case, suggesting that the true degree lower bound may be much lower.

The close connection of our density estimate to PageRank suggests that combining the latent spatial map with vector space estimates may lead to highly effective and theoretically principled network algorithms.

Chapter 3

Metric recovery on graphs

In chapter 2 we considered the problem of estimating an unknown, latent density function from unweighted graphs. As a consequence of this estimator, we constructed consistent, and empirically effective estimators of two quantities: the neighborhood radius $\varepsilon_n(x)$ and the density $p(x)$ at each vertex. While this estimate can be used to identify the latent embedding of a noiseless graph (Section 2.5), it does not serve as an robust, empirically effective metric over vertices on a graph. In this section, we will formalize the problem of estimating vertex similarity as recovering the latent pairwise distances in a metric graph and derive a new graph embedding algorithm.

Many network metrics have been introduced to measure the similarity between any two vertices. Such metrics can be used for a variety of purposes, including uncovering missing edges or pruning erroneous ones. Since the metrics tacitly assume that vertices lie in a latent (metric) space, one could expect that they also recover the underlying metric in some well-defined limit. Surprisingly, there are nearly no known results on this type of consistency. Indeed, it was recently shown by von Luxburg et al. [2014] that the expected hitting time degenerates and does not measure any notion of distance.

We analyze an improved hitting-time metric – Laplace transformed hitting time (LTHT) – and rigorously evaluate its consistency, cluster-preservation, and robustness under a general network model which encapsulates the latent space assumption. This network model, specified in Section 3.1, posits that vertices lie in a latent metric space, and edges are drawn between nearby vertices in that space. To analyze the LTHT, we develop two key technical tools. We establish a correspondence between functionals of hitting time for random walks on graphs, on the one hand, and limiting Itô processes (Corollary 3.3.3) on the other. Moreover, we construct a weighted random walk on the graph whose limit is a Brownian motion (Corollary 3.3.1). We apply these tools to obtain three main results.

First, our Theorem 3.2.5 recapitulates and generalizes the result of von Luxburg et al. [2014] pertaining to degeneration of expected hitting time in the limit. Our proof is direct and demonstrates the broader applicability of the techniques to general random walk based algorithms. Second, we analyze the Laplace transformed hitting time as a one-parameter family of improved distance estimators based on random walks on the graph. We prove that there exists a scaling limit for the parameter β such that the LTHT can become the shortest path distance (Theorem B.5.1) or a consistent metric estimator averaging over many paths (Theorem 3.3.4). Finally, we prove that the LTHT captures the advantages of random-walk based metrics by respecting the cluster structure (Theorem 3.3.5) and robustly recovering similarity queries when the majority of edges carry no geometric information (Theorem 3.3.7). We now discuss the relation of our work to prior work on similarity estimation.

Quasi-walk metrics: There is a growing literature on graph metrics that attempts to correct the degeneracy of expected hitting time [von Luxburg et al., 2014] by interpolating between expected hitting time and shortest path distance. The work closest to ours is the analysis of the phase transition of the p-resistance metric in Alamgir and von Luxburg [2011] which proves that p-resistances depend on distances for some parameters of p; however, their work did not address consistency or bias of p-resistances. Other approaches to quasi-walk metrics such as logarithmic-forest [Chebotarev, 2011], distributed routing distances [Tahbaz-Salehi and Jadbabaie, 2006], truncated hitting times [Sarkar and Moore, 2007], and randomized shortest paths [Kivimäki et al., 2014, Yen et al., 2008] exist but their statistical properties are unknown. Our paper is the first to prove consistency properties of a quasi-walk metric and our techniques could be applied to other quasi-walk metrics with appropriate scaling limits to derive consistency properties.

Nonparametric statistics: In the nonparametric statistics literature, the behavior of k-nearest neighbor and ε-ball graphs has been the focus of extensive study. For undirected graphs, Laplacian-based techniques have yielded consistency for clusters [von Luxburg et al., 2008] and shortest paths [Alamgir and von Luxburg, 2012b] as well as the degeneracy of expected hitting time [von Luxburg et al., 2014]. Algorithms for exactly embedding k-nearest neighbor graphs are similar and generate metric estimates, but require knowledge of the graph construction method and their consistency properties are unknown [Shaw and Jebara, 2009]. Stochastic differential equation techniques similar to ours were applied to prove Laplacian convergence results in Ting et al. [2010a], while the process-level convergence was exploited in Hashimoto et al. [2015c]. Our work advances the techniques of Hashimoto et al. [2015c] by extracting more robust estimators from process-level information.

Network analysis: The task of predicting missing links in a graph, known as link prediction, is one of the most popular uses of similarity estimation. The survey by Lü and Zhou [2011] compares several common link prediction methods on synthetic benchmarks. The consistency of some local similarity metrics was analyzed under a specific model which required strong assumptions on the graph class and similarity estimator [Sarkar et al., 2011]. The LTHT is the first global metric to achieve such a consistency property.

3.1 Continuum limits of random walks on networks

3.1.1 Definition of a metric graph

We take a generative approach to defining similarity between vertices. We suppose that each vertex i of a graph is associated with a latent coordinate $x_i \in \mathbb{R}^d$ and that the probability of finding an edge between two vertices depends solely on their latent coordinates. In this model, given only the un-weighted edge connectivity of a graph, we define natural distances between vertices as the distances between the latent coordinates x_i. Formally, let $\mathcal{X} = \{x_1, x_2, \dots\} \subset \mathbb{R}^d$ be an infinite sequence of points drawn i.i.d. from a differentiable density with bounded log gradient $p(x)$ with compact support D.

As defined before in Chapter 2, a metric graph is defined by the following:

Definition 2 (Metric graph). Let $\varepsilon_n : \mathcal{X}_n \to \mathbb{R}_{>0}$ be a local scale function and $h : \mathbb{R}_{\geq 0} \to [0, 1]$ a piecewise continuous function with $h(x) = 0$ for $x > 1$, $h(1) > 0$, and h left-continuous at 1. The *metric graph* G_n corresponding to ε_n and h is the random graph with vertex set $\mathcal{X}_n$ and a directed edge from x_i to x_j with probability $p_{ij} = h(|x_i - x_j|\varepsilon_n(x_i)^{-1})$.

This graph was proposed in Definition 1 as the generalization of k-nearest neighbors to isotropic kernels. To make inference tractable, we focus on the large-graph, small-neighborhood limit as $n \to \infty$ and $\varepsilon_n(x) \to 0$. In particular, we will suppose that there exist scaling constants g_n and a deterministic continuous function $\overline{\varepsilon} : D \to \mathbb{R}_{>0}$ so that

$$g_n \to 0, \qquad g_n n^{\frac{1}{d+2}} \log(n)^{-\frac{1}{d+2}} \to \infty, \qquad \varepsilon_n(x) g_n^{-1} \to \overline{\varepsilon}(x) \text{ for } x \in \mathcal{X}_n,$$

where the final convergence is uniform in x and a.s. in the draw of $\mathcal{X}$. The scaling constant g_n represents a bound on the asymptotic sparsity of the graph.

We give a few concrete examples to make the quantities h, g_n, and ε_n clear.

1. The directed k-nearest neighbor graph is defined by setting $h(x) = 1_{x \in [0,1]}$, the indicator function of the unit interval, $\varepsilon_n(x)$ the distance to the k^{th} nearest neighbor, and $g_n = (k/n)^{1/d}$ the rate at which $\varepsilon_n(x)$ approaches zero.

2. A Gaussian kernel graph is approximated by setting $h(x) = \exp(-x^2/\sigma^2)1_{x \in [0,1]}$. The truncation of the Gaussian tails at σ is an analytic convenience rather than a fundamental limitation, and the bandwidth can be varied by rescaling $\varepsilon_n(x)$.

3.1.2 Continuum limit of the random walk

Our techniques rely on analysis of the limiting behavior of the simple random walk X_t^n on a spatial graph G_n, viewed as a discrete-time Markov process with domain D. The increment at step t of X_t^n is a jump to a random point in $\mathcal{X}_n$ which lies within the ball of radius $\varepsilon_n(X_t^n)$ around X_t^n. We observe three effects: (A) the random walk jumps more frequently towards regions of high density; (B) the random walk moves more quickly whenever $\varepsilon_n(X_t^n)$ is large; (C) for ε_n small and a large step count t, the random variable $X_t^n - X_0^n$ is the sum of many small independent (but not necessarily identically distributed) increments. In the $n \to \infty$ limit, we may identify X_t^n with a continuous-time stochastic process satisfying (A), (B), and (C) via the following result, which is a slight strengthening of Theorem 2.3.4 obtained by applying Stroock and Varadhan [1979, Theorem 11.2.3] in place of the original result of Stroock-Varadhan.

Theorem 3.1.1. *The simple random walk X_t^n converges uniformly in Skorokhod space $\mathsf{D}([0,\infty),\overline{D})$ after a time scaling $\widehat{t} = tg_n^2$ to the Itô process $Y_{\widehat{t}}$ valued in the space of continuous functions $\mathsf{C}([0,\infty),\overline{D})$ defined by*

$$dY_{\widehat{t}} = \frac{\nabla \log(p(Y_{\widehat{t}}))}{3}\overline{\varepsilon}(Y_{\widehat{t}})^2 d\widehat{t} + \frac{\overline{\varepsilon}(Y_{\widehat{t}})}{\sqrt{3}} dW_{\widehat{t}} \tag{3.1}$$

with reflecting boundary conditions on D.

Effects (A), (B), and (C) may be seen in the stochastic differential equation (C.1) as follows. The direction of the drift is controlled by $\nabla \log(p(Y_{\widehat{t}}))$, the rate of drift is controlled by $\overline{\varepsilon}(Y_{\widehat{t}})^2$, and the noise is driven by a Brownian motion $W_{\widehat{t}}$ with location-dependent scaling $\frac{\overline{\varepsilon}(Y_{\widehat{t}})}{\sqrt{3}}$.[1]

[1] Both the variance $\Theta(\varepsilon_n(x)^2)$ and expected value $\Theta(\nabla \log(p(x))\varepsilon_n(x)^2)$ of a single step in the simple random walk are $\Theta(g_n^2)$. The time scaling $\widehat{t} = tg_n^2$ in Theorem 3.1.1 was chosen so that as $n \to \infty$ there are g_n^{-2} discrete steps taken per unit time, meaning the total drift and variance per unit time tend to a non-trivial limit.

We view Theorem 3.1.1 as a method to understand the simple random walk X_t^n through the continuous walk $Y_{\hat{t}}$. Attributes of stochastic processes such as stationary distribution or hitting time may be defined for both $Y_{\hat{t}}$ and X_t^n, and in many cases Theorem 3.1.1 implies that an appropriately-rescaled version of the discrete attribute will converge to the continuous one. Because attributes of the continuous process $Y_{\hat{t}}$ can reveal information about proximity between points, this provides a general framework for inference in spatial graphs. We use hitting times of the continuous process to a domain $E \subset D$ to prove properties of the hitting time of a simple random walk on a graph via the limit arguments of Theorem 3.1.1.

3.2 Degeneracy of expected hitting times in networks

The hitting time, commute time, and resistance distance are popular measures of distance based upon the random walk which are believed to be robust and capture the cluster structure of the network. However, it was shown in a surprising result that on undirected geometric graphs the scaled expected hitting time from x_i to x_j converges to inverse of the degree of x_j [von Luxburg et al., 2014].

In Theorem 3.2.5, we give an intuitive explanation and generalization of this result by showing that if the random walk on a graph converges to any limiting Itô process in dimension $d \geq 2$, the scaled expected hitting time to any point converges to the inverse of the stationary distribution. This answers the open problem in von Luxburg et al. [2014] on the degeneracy of hitting times for directed graphs and graphs with general degree distributions such as directed k-nearest neighbor graphs, lattices, and power-law graphs with convergent random walks. Our proof can be understood as first extending the transience or neighborhood recurrence of Brownian motion for $d \geq 2$ to more general Itô processes and then connecting hitting times on graphs to their Itô process equivalents.

3.2.1 Hitting times of an Itô process

For a domain $E \subset D$, let T_E^x be the hitting time of $Y_{\hat{t}}$ started at x to E and $T_{E,n}^{x_i}$ the hitting time of X_t^n started at x_i to E. We will give a lower bound for $\mathbb{P}(T_{x_j,n}^{x_i} > cg_n^{-2})$ for any constant c using a similar bound on $\mathbb{P}(T_E^{x_i} > c)$. Such bounds arise naturally from the Feynman-Kac theorem, which shows that functionals of hitting times are solutions to partial differential equations. We apply it to the Itô process in (C.1) with drift and diffusion functions $\mu(x) = \frac{\nabla \log(p(x))}{3}\overline{\varepsilon}(x)^2$ and $\sigma(x) = \frac{\overline{\varepsilon}(x)}{\sqrt{3}}$.

Theorem 3.2.1 ([Øksendal, 2003, Exercise 9.12] Feynman-Kac for the Laplace transform)**.** *The Laplace transform of the hitting time (LTHT) $u(x) = \mathbb{E}[\exp(-\beta T^x_E)]$ is the solution to the boundary value problem with boundary condition $u|_{\partial E} = 1$:*

$$\frac{1}{2} Tr[\sigma^T H(u)\sigma] + \mu(x) \cdot \nabla u - \beta u = 0.$$

Let $B(x, s)$ be the d-dimensional ball of radius s centered at x.

Lemma 3.2.2. *For $x, y \in D$, $d \geq 2$, and any $\delta > 0$, there exists $s > 0$ such that $\mathbb{E}[e^{-T^x_{B(y,s)}}] < \delta$.*

Proof. We compare the Laplace transformed hitting time of the general Itô process to that of Brownian motion via Feynman-Kac and handle the latter case directly. Details are in Section B.2.1. □

Corollary 3.2.3 (Typical hitting times are large)**.** *For any $d \geq 2$, $c > 0$, and $\delta > 0$, for large enough n we have $\mathbb{P}(T^{x_i}_{x_j,n} > cg_n^{-2}) > 1 - \delta$.*

Proof. Because $T^{x_i}_{x_j,n} \geq T^{x_i}_{B(x_j,s),n}$, by Theorem 3.1.1 we have

$$\lim_{n\to\infty} \mathbb{E}[e^{-T^{x_i}_{x_j,n} g_n^2}] \leq \mathbb{E}[e^{-T^{x_i}_{B(x_j,s)}}] \text{ for any } s > 0. \tag{3.2}$$

Applying Lemma 3.2.2, we have $\mathbb{E}[e^{-T^{x_i}_{B(x_j,s)}}] < \frac{1}{2}\delta e^{-c}$ for some $s > 0$. For large enough n, this combined with (3.2) implies $\mathbb{P}(T^{x_i}_{x_j,n} \leq cg_n^{-2})e^{-c} < \delta e^{-c}$ and hence $\mathbb{P}(T^{x_i}_{x_j,n} \leq cg_n^{-2}) < \delta$. □

3.2.2 Expected hitting times degenerate to the stationary distribution

To translate results from Itô processes to directed graphs, we require a regularity condition. Let $q_t(x_j, x_i)$ denote the probability that $X^n_t = x_j$ conditioned on $X^n_0 = x_i$. We make the following technical conjecture which we assume holds for all spatial graphs.

($\star$) For $t = \Theta(g_n^{-2})$, the rescaled marginal $nq_t(x, x_i)$ is a.s. eventually uniformly equicontinuous.

Assumption ($\star$) is related to smoothing properties of the graph Laplacian and is known to hold for undirected graphs [Croydon and Hambly, 2008a]. No directed analogue is known, and Conjecture A.1.1 in Section 2 is a weaker property conjectured to hold for all spatial graphs. See Section B.1 for further details. Let $\pi_{X^n}(x)$ denote the stationary distribution of X_t^n. The following was shown in Theorem 2.2.1 under conditions implied by our condition $(\star)$ (Corollary B.2.6).

Theorem 3.2.4. *Assuming $(\star)$, for $a^{-1} = \int p(x)^2 \overline{\varepsilon}(x)^{-2} dx$, we have the a.s. limit*

$$\widehat{\pi}(x) := \lim_{n\to\infty} n\pi_{X^n}(x) = a\frac{p(x)}{\overline{\varepsilon}(x)^2}.$$

We may now express the limit of expected hitting time in terms of this result.

Theorem 3.2.5. *For $d \geq 2$ and any i, j, we have*

$$\frac{\mathbb{E}[T^{x_i}_{x_j,n}]}{n} \overset{a.s.}{\to} \frac{1}{\widehat{\pi}(x_j)}.$$

Proof. By definition, we have

$$\mathbb{E}[T^{x_i}_{x_j,n} \mid T^{x_i}_{x_j,n} > \widehat{t}g_n^{-2}] \geq \mathbb{E}[T^{x_i}_{x_j,n}] \geq \mathbb{P}(T^{x_i}_{x_j,n} > \widehat{t}g_n^{-2})\mathbb{E}[T^{x_i}_{x_j,n} \mid T^{x_i}_{x_j,n} > \widehat{t}g_n^{-2}]. \quad (3.3)$$

By Corollary 3.2.3, for any $\delta > 0$ and $\widehat{t}_0 > 0$, there is some n_1 so that for $n > n_1$ and $\widehat{t} > \widehat{t}_0$ we have $\mathbb{P}(T^{x_i}_{x_j,n} > \widehat{t}g_n^{-2}) > (1-\delta)$. Define now $p_t = \mathbb{P}(T^{x_i}_{x_j,n} = t \mid T^{x_i}_{x_j,n} \geq t)$; by definition we have

$$\mathbb{E}[T^{x_i}_{x_j,n} \mid T^{x_i}_{x_j,n} > \widehat{t}g_n^{-2}] = \sum_{t=\lceil \widehat{t}g_n^{-2}\rceil}^{\infty} tp_t \prod_{r=\lceil \widehat{t}g_n^{-2}\rceil}^{t-1} (1-p_r).$$

By Theorem B.2.5 obtained as a consequence of Corollary 3.2.3, the simple random walk X_t^n mixes at exponential rate, implying in Lemma B.2.8 that the probability of first hitting at step $t > cg_n^{-2}$ is approximately the stationary distribution at x_j (See Section B.2 for a full proof).

By Lemma B.2.8 , we have for some n_2 that for $n > n_2$ and $t > \widehat{t}_0 g_n^{-2}$ that

$$|p_t - \theta_n(x_j)| < \frac{C\exp(-\beta t g_n^2)}{n}$$

so in particular for $\delta = \frac{1}{2}\min_{x\in D}\widehat{\pi}(x)$ and $\tau = 2\max_{x\in D}\widehat{\pi}(x)$, we have for some n_3 that for $n > n_3$ we have

$$\delta < np_t < \tau \text{ and } \delta < n\theta_n(x_j) < \tau.$$

For n_4 large enough that $1 - \tau/n_4 > \delta/n_4$, for $n > n_4$ we have

$$\begin{aligned}\left| p_t \prod_{r=\lceil \hat{t} g_n^{-2} \rceil}^{t-1} (1 - p_r) - \theta_n(x_j)(1 - \theta_n(x_j))^{t - \lceil \hat{t} g_n^{-2} \rceil} \right| &< \sum_{r=\lceil \hat{t} g_n^{-2} \rceil}^{t-1} \frac{C \exp(-\beta r g_n^2)}{n} (1 - \tau/n)^{t - \lceil \hat{t} g_n^{-2} \rceil - 1} \\ &< \frac{C}{n} \frac{e^{-\beta \hat{t}}}{1 - e^{-\beta g_n^2}} (1 - \tau/n)^{t - \lceil \hat{t} g_n^{-2} \rceil - 1}.\end{aligned}$$

This implies that

$$\begin{aligned}&\frac{1}{n} \left| \mathbb{E}[T_{x_j,n}^{x_i} \mid T_{x_j,n}^{x_i} > \hat{t} g_n^{-2}] - \sum_{t=\lceil \hat{t} g_n^{-2} \rceil}^{\infty} t \theta_n(x_j)(1 - \theta_n(x_j))^{t - \lceil \hat{t} g_n^{-2} \rceil} \right| \\ &< \sum_{t=\lceil \hat{t} g_n^{-2} \rceil}^{\infty} \frac{C}{n^2} \frac{e^{-\beta \hat{t}}}{1 - e^{-\beta g_n^2}} (1 - \tau/n)^{t - \lceil \hat{t} g_n^{-2} \rceil - 1} \\ &< \frac{C}{\tau(n - \tau)} \frac{e^{-\beta \hat{t}}}{1 - e^{-\beta g_n^2}},\end{aligned}$$

where we note that for $n > 2\tau$, we have

$$\frac{C}{\tau(n - \tau)} \frac{e^{-\beta \hat{t}}}{1 - e^{-\beta g_n^2}} < \frac{2 C e^{-\beta \hat{t}}}{\tau} n^{-1} (g_n^{-2} + \frac{1}{2} + \frac{1}{12} g_n^2).$$

Because $\lim_{n \to \infty} n^{-1}(g_n^{-2} + \frac{1}{2} + \frac{1}{12} g_n^2) = 0$, considering $n > \max\{n_1, n_2, n_3, n_4\}$, we conclude that

$$\begin{aligned}\lim_{n \to \infty} \frac{1}{n} \mathbb{E}[T_{x_j,n}^{x_i} \mid T_{x_j,n}^{x_i} > \hat{t} g_n^{-2}] &= \lim_{n \to \infty} \frac{1}{n} \sum_{t=\lceil \hat{t} g_n^{-2} \rceil}^{\infty} t \theta_n(x_j)(1 - \theta_n(x_j))^{t - \lceil \hat{t} g_n^{-2} \rceil} \\ &= \lim_{n \to \infty} \frac{1}{n} \frac{1 - \theta_n(x_j) + \theta_n(x_j) \lceil \hat{t} g_n^{-2} \rceil}{\theta_n(x_j)} \\ &= \lim_{n \to \infty} \frac{1}{n \theta_n(x_j)} = \frac{1}{\hat{\pi}(x_j)},\end{aligned}$$

where the last equality follows from Lemma B.2.10. Now by (3.3), we conclude that

$$\lim_{n \to \infty} \frac{1}{n} \mathbb{E}[T_{x_j,n}^{x_i}] = \frac{1}{\hat{\pi}(x_j)}. \qquad \square$$

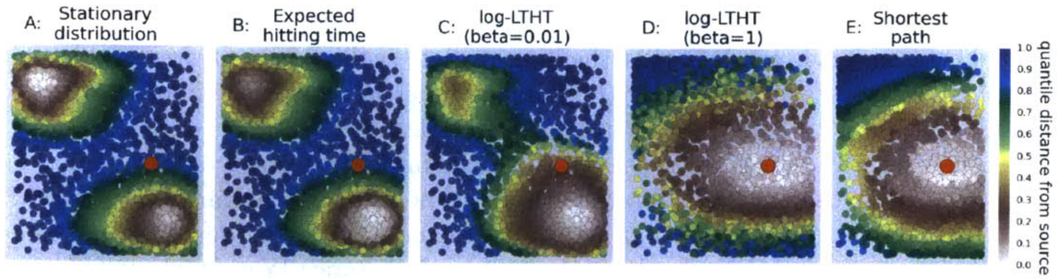


Figure 3-1: Estimated distance from orange starting point on a k-nearest neighbor graph constructed on two clusters. A and B show degeneracy of hitting times (Theorem 3.2.5). C, D, and E show that log-LTHT interpolate between hitting time and shortest path.

Theorem 3.2.5 is illustrated in Figures 3-1A and 3-1B, which show that on as few as 3000 points, expected hitting times on a k-nearest neighbor graph converge to the stationary distribution rather than any measure of distance. [2]

3.3 The Laplace transformed hitting time (LTHT)

In Theorem 3.2.5 we showed that expected hitting time is degenerate because a simple random walk mixes before hitting its target. To correct this we penalize longer paths. More precisely, consider for $\widehat{\beta} > 0$ and $\beta_n = \widehat{\beta} g_n^2$ the *Laplace transforms* $\mathbb{E}[e^{-\widehat{\beta} T_E^x}]$ and $\mathbb{E}[e^{-\beta_n T_{E,n}^x}]$ of T_E^x and $T_{E,n}^x$.

These Laplace transformed hitting times (LTHT's) have three advantages. First, while the expected hitting time of a Brownian motion to a domain is dominated by long paths, the LTHT is dominated by direct paths. Second, the LTHT for the Itô process can be derived in closed form via the Feynman-Kac theorem, allowing us to make use of techniques from continuous stochastic processes to control the continuum LTHT. Lastly, the LTHT can be computed both by sampling and in closed form as a matrix inversion (Section B.3). Now define the scaled log-LTHT as

$$-\log(\mathbb{E}[e^{-\beta_n T_{x_j,n}^{x_i}}])/\sqrt{2\beta_n} g_n.$$

Taking different scalings for β_n with n interpolates between expected hitting time

[2] Surprisingly, von Luxburg et al. [2014] proved that 1-D hitting times diverge despite convergence of the continuous equivalent. This occurs because the discrete walk can jump past the target point. In Section B.2.4, we consider 1-D hitting times to small out neighbors which corrects this problem and derive closed form solutions (Theorem B.2.13). This hitting time is non-degenerate but highly biased due to boundary terms (Corollary B.2.15).

($\beta_n \to 0$ on a fixed graph) and shortest path distance ($\beta_n \to \infty$) (Figures 3-1C, D, and E). In Theorem 3.3.4, we show that the intermediate scaling $\beta_n = \Theta(\widehat{\beta} g_n^2)$ yields a consistent distance measure retaining the unique properties of hitting times. Most of our results on the LTHT are novel for any quasi-walk metric.

While considering the Laplace transform of the hitting time is novel to our work, this metric has been used in the literature in an ad-hoc manner in various forms as a similarity metric for collaboration networks [Yazdani, 2013], hidden subgraph detection [Smith et al., 2014], and robust shortest path distance [Yen et al., 2008]. However, these papers only considered the elementary properties of the limits $\beta_n \to 0$ and $\beta_n \to \infty$ and simple triangle inequalities. Our consistency proof demonstrates the advantage of the stochastic process approach over existing combinatorial ones.

3.3.1 Consistency

We now consider metric recovery. It was shown previously that for n fixed and $\beta_n \to \infty$, $-\log(\mathbb{E}[-\beta_n T^{x_i}_{x_j,n}])/\beta_n g_n$ converges to shortest path distance from x_i to x_j. We investigate more precise behavior in terms of the scaling of β_n. There are two regimes: if $\beta_n = \omega(\log(g_n^d n))$, then the shortest path dominates and the LTHT converges to shortest path distance (See Theorem B.5.1). If $\beta_n = \Theta(\widehat{\beta} g_n^2)$, the graph log-LTHT converges to its continuous equivalent, which for large $\widehat{\beta}$ averages over random walks concentrated around the geodesic. To our knowledge, this is the first method with any consistency property without appeal to shortest paths.

To show consistency for $\beta_n = \Theta(\widehat{\beta} g_n^2)$, we proceed in three steps: (1) we reweight the random walk on the graph so the limiting process is Brownian motion; (2) we show that log-LTHT for Brownian motion recovers latent distance; (3) we show that log-LTHT for the reweighted walk converges to its continuous limit; (4) we conclude that log-LTHT of the reweighted walk recovers latent distance.

(1) Reweighting the random walk to converge to Brownian motion: We define weights using the estimators $\widehat{p}$ and $\widehat{\varepsilon}$ for $p(x)$ and $\overline{\varepsilon}(x)$ from Corollary 2.2.2.

Theorem 3.3.1. *Let $\widehat{p}$ and $\widehat{\varepsilon}$ be consistent estimators of the density and local scale and A be the adjacency matrix. Then the random walk $\widehat{X}_t^n$ defined below converges to a Brownian motion.*

$$\mathbb{P}(\widehat{X}_{t+1}^n = x_j \mid \widehat{X}_t^n = x_i) = \begin{cases} \frac{A_{i,j}\widehat{p}(x_j)^{-1}}{\sum_k A_{i,k}\widehat{p}(x_k)^{-1}} \widehat{\varepsilon}(x_i)^{-2} & i \neq j \\ 1 - \widehat{\varepsilon}(x_i)^{-2} & i = j \end{cases}$$

Proof. Reweighting by $\widehat{p}$ and $\widehat{\varepsilon}$ is designed to cancel the drift and diffusion terms in Theorem 3.1.1 by ensuring that as n grows large, jumps have means approaching 0 and variances which are asymptotically equal (but decaying with n).

Set $a_n(x) = \widehat{p}(x)^{-1}$ and $b_n(x) = \widehat{\varepsilon}(x)^{-2}$ as estimated by Corollary 2.2.2 so that $\lim_{n\to\infty} a_n(x) = p(x)^{-1}$ and $\lim_{n\to\infty} b_n(x) = \overline{\varepsilon}(x)^{-2}$. Verifying the limiting drift and diffusion coefficients using the Stroock-Varadhan criterion (Theorem B.4.1) shows that the limiting process is a Brownian motion.[3] □

(2) Log-LTHT for a Brownian motion: Let W_t be a Brownian motion with $W_0 = x_i$, and let $\overline{T}^{x_i}_{B(x_j,s)}$ be the hitting time of W_t to $B(x_j, s)$. We show that log-LTHT converges to distance.

Lemma 3.3.2. *For any $\alpha < 0$, if $\widehat{\beta} = s^\alpha$, as $s \to 0$ we have*

$$-\log(\mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)})])/\sqrt{2\widehat{\beta}} \to |x_i - x_j|.$$

Proof. We consider hitting time of Brownian motion started at distance $|x_i - x_j|$ from the origin to distance s of the origin, which is controlled by a Bessel process.

Let $B_t = |W_t|$ be the order $\nu = d/2 - 1$ Bessel process. The LTHT of B_t to hit $x_j \pm s$ is equivalent to the LTHT of W_t to hit $B(x_j, s)$. Defining $w = |x_i - x_j|$, by Borodin and Salminen [2002, Eq 4.2.0.1], this is:

$$\mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)})] = \frac{K(\nu, w\sqrt{2\widehat{\beta}})w^{-\nu}}{K(\nu, s\sqrt{2\widehat{\beta}})s^{-\nu}},$$

where $K(\nu, w)$ is a modified Bessel function of the second kind.

Write $-\log(\mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)})])/\sqrt{2\widehat{\beta}} = c_1 + c_2$ for

$$\begin{aligned} c_1 &= -\log(K(\nu, w\sqrt{2\widehat{\beta}})w^{-\nu})/\sqrt{2\widehat{\beta}} \\ c_2 &= -\log(K(\nu, s\sqrt{2\widehat{\beta}})s^{-\nu})/\sqrt{2\widehat{\beta}}. \end{aligned}$$

Taylor expansion of c_1 at $\widehat{\beta}^{-1} = 0$ yields

$$c_1 = w - \frac{\log(\pi^2/(8\widehat{\beta})) + 4\log(w^{-1/2-\nu})}{4\sqrt{2\widehat{\beta}}} + o\left(\frac{\nu^2}{w\widehat{\beta}}\right),$$

[3] This is a special case of a more general theorem for transforming limits of graph random walks (Theorem 3.3.1). Figure B-4 shows that this modification is highly effective in practice.

hence $c_1 \to w$. For c_2, note that $\nu \log(s)/\sqrt{2\widehat{\beta}} \to 0$ and for s small,

$$K(\nu, s\sqrt{2\widehat{\beta}}) \sim \begin{cases} -\log(s\sqrt{2\widehat{\beta}}) & d = 2 \\ \frac{1}{2}\Gamma(s\sqrt{2\widehat{\beta}})(\frac{1}{2}s\sqrt{2\widehat{\beta}})^{-\nu} & d > 2 \end{cases}.$$

by Abramowitz and Stegun [1972, p375]. Checking that $-\log(K(\nu, s\sqrt{2\widehat{\beta}}))/\sqrt{2\widehat{\beta}} \to 0$ and combining estimates gives $-\log(\mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)})])/\sqrt{2\widehat{\beta}} = c_1 + c_2 \to w$. □

(3) Convergence of LTHT on graphs with $\beta_n = \Theta(\widehat{\beta} g_n^2)$: To compare continuous and discrete log-LTHT's, we will first define the s-neighborhood of a vertex x_i on G_n as the graph equivalent of the ball $B(x_i, s)$.

Definition 3 (s-neighborhood). Let $\widehat{\varepsilon}(x)$ be the consistent estimate of the local scale from Corollary 2.2.2 so that $\widehat{\varepsilon}(x) \to \overline{\varepsilon}(x)$ uniformly a.s. as $n \to \infty$. The $\widehat{\varepsilon}$-weight of a path $x_{i_1} \to \cdots \to x_{i_l}$ is the sum $\sum_{m=1}^{l-1} \widehat{\varepsilon}(x_{i_m})$ of vertex weights $\widehat{\varepsilon}(x_i)$. For $s > 0$ and $x \in G_n$, the s-neighborhood of x is

$$\mathsf{NB}_n^s(x) := \{y \mid \text{there is a path } x \to y \text{ of } \widehat{\varepsilon}\text{-weight} \leq g_n^{-1}s\}.$$

For $x_i, x_j \in G_n$, let $\widehat{T}^{x_i}_{B(x_j,s)}$ be the hitting time of the transformed walk on G_n from x_i to $\mathsf{NB}_n^s(x_j)$. We now verify that hitting times to the s-neighborhood on graphs and the s-radius ball coincide.

Corollary 3.3.3. *For $s > 0$, we have $g_n^2 \widehat{T}^{x_i}_{\mathsf{NB}_n^s(x_j),n} \xrightarrow{d} \overline{T}^{x_i}_{B(x_j,s)}$.*

Proof. Consistency of the $\widehat{\varepsilon}$-balls imply that the ball and the neighborhood will have nearly identical sets of points.

After verifying that $\widehat{\varepsilon}$-balls are close to ε balls, we can directly apply Theorem 3.1.1. See Section B.6.2 for details. □

The requirement that we consider hitting times to s-neighborhoods are likely a technical condition, similar to that of requiring equicontinuity. In practice we consider hitting times to single vertices and find that these quantities are well-behaved. Proving that such single-vertex LTHTs are well behaved will require more advanced techniques which are outside the scope of this thesis.

(4) Proving consistency of log-LTHT: Properly accounting for boundary effects, we obtain a consistency result for the log-LTHT for small neighborhood hitting times.

Theorem 3.3.4. *Let $x_i, x_j \in G_n$ be connected by a geodesic not intersecting ∂D. For any $\delta > 0$, there exists a choice of $\widehat{\beta}$ and $s > 0$ so that if $\beta_n = \widehat{\beta} g_n^2$, for large n we have with high probability*

$$\left| - \log(\mathbb{E}[\exp(-\beta_n \widehat{T}^{x_i}_{\mathsf{NB}^s_n(x_j),n})]) / \sqrt{2\widehat{\beta}} - |x_i - x_j| \right| < \delta.$$

Proof of Theorem 3.3.4. The proof has three steps. First, we convert to the continuous setting via Corollary 3.3.3. Second, we show the contribution of the boundary is negligible. The conclusion follows from the explicit computation of Lemma 3.3.2. Full details are in Section B.6. □

The stochastic process limit proof of Theorem 3.3.4 has two unique implications. First, small perturbations to the graph such as removing g_n^2/n edges do not affect the limit; this shows that the log-LTHT does not rely heavily on any one path in the limit (Section B.8). Second, while we explicitly removed the effect of cluster structure by reweighting, we can construct a cluster preserving metric by applying the log-LTHT to the unweighted simple random walk.

3.3.2 Bias

Random walk based metrics are often motivated as recovering a cluster preserving metric. We now show that the log-LTHT of the un-weighted simple random walk preserves the underlying cluster structure. In the 1-D case, we provide a complete characterization.

Theorem 3.3.5. *Suppose the spatial graph has $d = 1$ and $h(x) = 1_{x \in [0,1]}$. Let $T^{x_i}_{\mathsf{NB}_n^{\widehat{\varepsilon}(x_j) g_n}(x_j),n}$ be the hitting time of a simple random walk from x_i to the out-neighborhood of x_j. It converges to*

$$- \log(\mathbb{E}[-\beta T^{x_i}_{\mathsf{NB}_n^{\widehat{\varepsilon}(x_j) g_n}(x_j),n}]) / \sqrt{8\beta} \to \int_{x_i}^{x_j} \sqrt{m(x)} dx + o\left(\log(1 + e^{-\sqrt{2\beta}}) / \sqrt{2\beta}\right),$$

where $m(x) = \frac{2}{\overline{\varepsilon}(x)^2} + \frac{1}{\beta} \frac{\partial \log(p(x))}{\partial x^2} + \frac{1}{\beta} \left(\frac{\partial \log(p(x))}{\partial x} \right)^2$ defines a density-sensitive metric.

Proof. We first prove an approximation for the LTHT of the continuous stochatsic process:

Let $\mathbb{E}[\exp(-\beta T_{x_j}^{x_i})] = u(x_i)$, where $u(x)$ is the hitting time to x_j from point x. By Feynman-Kac, this is

$$\frac{\partial^2 u}{\partial x^2} + 2\frac{\partial \log(p(x))}{\partial x}\frac{\partial u}{\partial x} + q(x)u = 0,$$

where $q(x) = -2\beta\bar{\varepsilon}(x)^{-2}$. Rewrite this as a perturbation of a second order ODE via the change of variables to obtain

$$\begin{aligned} y(x) &= u(x)\exp\left(\int_\gamma^x \frac{\partial \log(p(y))}{\partial y}dy\right) = u(x)p(x)p(\gamma)^{-1} \\ f(x) &= \frac{2}{\bar{\varepsilon}(x)^2} + \frac{1}{\beta}\left(\frac{\partial \log(p(x))}{\partial x^2} + \left(\frac{\partial \log(p(x))}{\partial x}\right)^2\right) \\ \frac{1}{\beta}\frac{\partial^2 y}{\partial x} &= f(x)y(x). \end{aligned}$$

Since this is a type of Schrodinger's equation with $f(x) \neq 0$ everywhere we can apply the WKBJ asymptotic expansion [Bender and Orszag, 1999, section 10.1] to obtain

$$y(x) = \frac{c_1}{f(x)^{1/4}}\exp\left(-\sqrt{\beta}\int_{x_0}^x \sqrt{f(s)}ds\right) + \frac{c_2}{f(x)^{1/4}}\exp\left(\sqrt{\beta}\int_{x_0}^x \sqrt{f(s)}ds\right) + o(\exp(-\beta)).$$

Since we assumed $x_i < x_j$ and by the boundary condition $u(x_j) = 1$ we have

$$u(x) = \frac{c_2 p(\gamma)}{f(x)^{1/4}p(x)}\exp\left(-\sqrt{\beta}\int_x^{x_j} \sqrt{f(s)}ds\right) + \frac{c_1 p(\gamma)}{f(x)^{1/4}p(x)}\exp\left(\sqrt{\beta}\int_x^{x_j} \sqrt{f(s)}ds\right) + o(\exp(-\beta$$

To obtain the boundary conditions, note that $u'(\gamma) = 0$. Taking the derivative for $y(x)p(x)$, setting to zero and solving for c_2 results in

$$c_2 = c_1\frac{\exp(-2\sqrt{\beta}\int_\gamma^{x_j}\sqrt{f(s)}ds)(p(\gamma)4\sqrt{\beta}f(\gamma)^{3/2} + f'(\gamma)) - f(\gamma)p'(\gamma)}{4\sqrt{\beta}f(\gamma)^{3/2}p(\gamma) - p(\gamma)f'(\gamma) + 4f(\gamma)p'(\gamma)} + o(\exp(-\beta)),$$

from which we obtain

$$c_2 = c_1\exp\left(-2\sqrt{\beta}\int_\gamma^{x_j}\sqrt{f(s)}ds\left(1 + o\left(\sqrt{\frac{1}{\beta}}\right)\right)\right).$$

Pulling out the $-\sqrt{\beta}$ term, we get

$$u(x_i) = \mathbb{E}[\exp(-\beta T_{x_j}^{x_i})] = \frac{c_1 p(\gamma)}{f(x_i)^{1/4} p(x_i)} \exp\left(-\sqrt{\beta} \int_{x_i}^{x_j} \sqrt{f(s)} ds\right)$$
$$\left(1 + \left(1 + o\left(\frac{1}{\sqrt{\beta}}\right)\right) \exp\left(-2\sqrt{\beta} \int_{\gamma}^{x_i} \sqrt{f(x)} dx\right) + o(\exp(-\beta))\right). \quad \square$$

This expansion give us the first-order terms of the LTHT in a continuous setting. Taking the taylor expansion of the log (Corollary B.7.2) and using Stroock-Varadhan to convert this result to the graph setting (Corollary B.2.14) gives the desired result.

The leading order terms of the density-sensitive metric appropriately penalize crossing regions of large changes to the log density. This is not the case for the expected hitting time where the boundary effects dominate density effects (Theorem B.2.13). In $d > 1$, no approximation analogous to WKBJ is known, but for linear drift, an exact solution exists and is similar to our 1-D result [Yin, 1999].

3.3.3 Robustness

While shortest path distance is a consistent measure of the underlying metric, it breaks down catastrophically with the addition of a single non-geometric edge and does not meaningfully rank vertices that share an edge. In contrast, we show that LTHT breaks ties between vertices via the resource allocation (RA) index, a robust local similarity metric under Erdős-Rényi-type noise. [4]

Definition 4. The noisy spatial graph G_n over $\mathcal{X}_n$ with noise terms $q_1(n), \ldots, q_n(n)$ is constructed by drawing an edge from x_i to x_j with probability

$$p_{ij} = h(|x_i - x_j|\varepsilon_n(x_i)^{-1})(1 - q_j(n)) + q_j(n).$$

Define the directed RA index in terms of the out-neighborhood set $\mathsf{NB}_n(x_i)$ and the in-neighborhood set $\mathsf{NB}_n^{\mathsf{in}}(x_i)$ as $R_{ij} := \sum_{x_k \in \mathsf{NB}_n(x_i) \cap \mathsf{NB}_n^{\mathsf{in}}(x_j)} |\mathsf{NB}_n(x_k)|^{-1}$ and two step log-LTHT by $M_{ij}^{\mathrm{ts}} := -\log(\mathbb{E}[\exp(-\beta T_{x_j,n}^{x_i}) \mid T_{x_j,n}^{x_i} > 1])$. [5] We show two step

[4] Modifying the graph by changing fewer than g_n^2/n edges does not affect the continuum limit of the random graph, and therefore preserve the LTHT with parameter $\beta = \Theta(g_n^2)$. While this weak bound allows on average $o(1)$ noise edges per vertex, it does show that the LTHT is substantially more robust than shortest paths without modification. See Section B.8 for proofs.

[5] The conditioning $T_{x_j,n}^{x_i} > 1$ is natural in link-prediction tasks where only pairs of disconnected vertices are queried. Empirically, we observe it is critical to performance (Figure 3-3).

log-LTHT and RA index give equivalent methods for testing if vertices are within distance $\varepsilon_n(x)$.

Theorem 3.3.6. *If $\beta = \omega(\log(g_n^d n))$ and x_i and x_j have at least one common neighbor, then*

$$M_{ij}^{ts} - 2\beta \to -\log(R_{ij}) + \log(|\mathsf{NB}_n(x_i)|).$$

Proof. Let $P_{ij}(t)$ be the probability of going from x_i to x_j in t steps, and $H_{ij}(t)$ the probability of not hitting before time t. Factoring the two-step hitting time yields

$$M_{ij}^{\text{ts}} = 2\beta - \log(P_{ij}(2)) - \log\left(1 + \sum_{t=3}^{\infty} \frac{P_{ij}(t)}{P_{ij}(2)} H_{ij}(t) e^{-\beta(t-2)}\right).$$

Let $k_{\max}$ be the maximal out-degree in G_n. The contribution of paths of length greater than 2 vanishes because $H_{ij}(t) \leq 1$ and $P_{ij}(t)/P_{ij}(2) \leq k_{\max}^2$, which is dominated by $e^{-\beta}$ for $\beta = \omega(\log(g^n n))$. Noting that $P_{ij}(2) = \frac{R_{ij}}{|\mathsf{NB}_n(x_i)|}$ concludes. For full details see Theorem B.9.1. □

For edge identification within distance $\varepsilon_n(x)$, the RA index is robust even at noise level $q = o(g_n^{d/2})$.

Theorem 3.3.7. *If $q_i = q = o(g_n^{d/2})$ for all i, for any $\delta > 0$ there are c_1, c_2 and h_n so that for any i, j, with probability at least $1 - \delta$ we have*

- $|x_i - x_j| < \min\{\varepsilon_n(x_i), \varepsilon_n(x_j)\}$ *if $R_{ij} h_n < c_1$;*
- $|x_i - x_j| > 2\max\{\varepsilon_n(x_i), \varepsilon_n(x_j)\}$ *if $R_{ij} h_n > c_2$.*

Proof. The minimal RA index for overlapping balls is bounded using Chebyshev's inequality. Since much of the calculations are routine, we provide a short summary of the proof, and defer details to (Theorem B.9.2).

First note that the out-degree of x_i can be decomposed into expectation and noise terms:

$$|\mathsf{NB}_n(x_i)| = nq + k_i + z_i.$$

Applying a Taylor expansion to $1/|\mathsf{NB}_n(x_i)|$ and applying the definition of the RA index gives:

$$R_{ij} = \sum_{x_k \in \mathsf{NB}_n(x_i) \cap \mathsf{NB}_n^{\text{in}}(x_j)} \left(\frac{1}{nq + k_i} + O(\delta_1^{-1/2}(nq + k_i)^{-3/2})\right).$$

We can then apply this bound to the two cases above, where x_i and x_j are less than $\varepsilon_n(x_i)$ or greater than $2\varepsilon_n(x_i)$. □

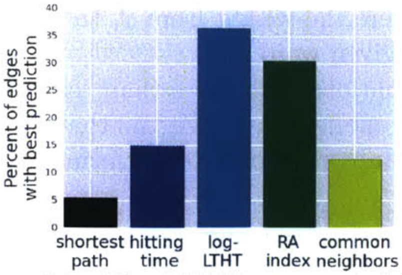


Figure 3-2: The LTHT recovered deleted edges most consistently on a citation network

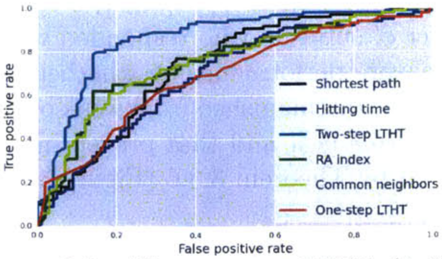


Figure 3-3: The two-step LTHT (defined above Theorem 3.3.6) performs best at word similarity estimation, outperforming even the basic log-LTHT (one-step).

3.4 Link prediction tasks

We compare the LTHT against other baseline measures of vertex similarity: shortest path distance, expected hitting time, number of common neighbors, and the RA index. A comprehensive evaluation of these quasi-walk metrics was performed in Kivimäki et al. [2014] who showed that a metric equivalent to the LTHT performed best. We consider two separate link prediction tasks on the largest connected component of vertices of degree at least five, fixing $\beta = 0.2$.[6] The degree constraint is to ensure that local methods using number of common neighbors such as the resource allocation index do not have an excessive number of ties. All code to generate figures in this paper are contained in an iPython-notebook in the supplement.

Citation network: The KDD 2003 challenge dataset [Gehrke et al., 2003] includes a directed, unweighted network of e-print arXiv citations whose dense connected component has 11,042 vertices and 222,027 edges. We use the same benchmark method as Lü and Zhou [2011] where we delete a single edge and compare the similarity of the deleted edge against the set of control pair of vertices i, j which do not share an edge. We count the fraction of pairs on which each method rank the deleted edge higher than all other methods. We find that LTHT is consistently best at this task (Figure 3-2). [7]

Associative Thesaurus network: The Edinburgh associative thesaurus [Kiss et al., 1973] is a network with a dense connected component of 7754 vertices and 246,609 edges in which subjects were shown a set of ten words and for each word was asked to respond with the first word to occur to them. Each vertex represents

[6] Results are qualitatively identical when varying β from 0.1 to 1; see supplement for details.

[7] The two-step LTHT is not shown, since the LTHT is equivalent to the two-step LTHT for missing link prediction.

a word and each edge is a weighted, directed edge where the weight from x_i to x_j is the number of subjects who responded with word x_j given word x_i.

We measure performance by whether strong associations with more than ten responses can be distinguished from weak ones with only one response. We find that the LTHT performs best and that preventing one-step jumps is critical to performance as predicted by Theorem 3.3.6 (Figure 3-3).

3.5 Conclusion

Our work has developed an asymptotic equivalence between hitting times for random walks on graphs and the Feynman-Kac theorem for diffusion processes. Using this, we have provided a short extension of the proof for the divergence of expected hitting times, and derived a new consistent graph metric that is theoretically principled, computationally tractable, and empirically successful at well-established link prediction benchmarks. These results open the way for the development of other principled quasi-walk metrics that can provably recover underlying latent similarities for spatial graphs.

Chapter 4

Word embeddings as metric recovery

Word embeddings, which seek to represent words as vectors whose similarities capture co-occurrence, are another class of embeddings which are closely related to the graph embeddings. Words can be viewed as vertices on a graph, and sentences can be viewed as a random walk over this graph. The major distinction between graph and word embeddings is that in word embeddings both the random-walk and measurable quantities are outside of our control. While in the graph setting we could measure sophisticated statistics such as hitting times, here we are given co-occurrences arising from a random walk and asked to embed it.

Continuous word representations have been remarkably useful across NLP tasks but remain poorly understood. We ground word embeddings in semantic spaces studied in the cognitive-psychometric literature, taking these spaces as the primary objects to recover. To this end, we relate log co-occurrences of words in large corpora to semantic similarity assessments and show that co-occurrences are indeed consistent with an Euclidean semantic space hypothesis. Fundamentally, we take *metric recovery* as the key theoretical goal. This perspective unifies existing word embedding algorithms, ties them to manifold learning, and demonstrates that existing algorithms are consistent metric recovery methods of co-occurrence counts from random walks. Further, we propose a simple, principled direct metric recovery algorithm that is comparable to the state-of-art in both word embedding and manifold learning. Finally, we complement recent focus on analogies by constructing two new inductive reasoning datasets – series completion and classification – and demonstrate that word embeddings can be used to solve them as well.

4.1 Introduction

Continuous space models of words, objects, and signals have become ubiquitous tools for learning rich representations of data, from natural language processing to computer vision. Specifically, there has been particular interest in word embeddings, largely due to their intriguing semantic properties [Mikolov et al., 2013b] and their success as features for downstream natural language processing tasks, including named entity recognition [Turian et al., 2010], parsing [Socher et al., 2013], and many others.

The empirical success of word embeddings has led some to seek a better understanding of their properties, associated estimation algorithms, and explore possible revisions. Recently, Levy and Goldberg [2014a] showed that linear linguistic regularities first observed with `word2vec` extend to other embedding methods. In particular, *explicit* representations of words in terms of co-occurrence counts can be used to solve analogies in the same way. In terms of algorithms, Levy and Goldberg [2014b] demonstrated that the global minimum of the skip-gram method with negative sampling in Mikolov et al. [2013b] implicitly factorizes a shifted version of the pointwise mutual information (PMI) matrix of word-context pairs. Arora et al. [2015] explored links between random walks and word embeddings, relating them to contextual (probability ratio) analogies, under specific (isotropic) assumptions about word vectors.

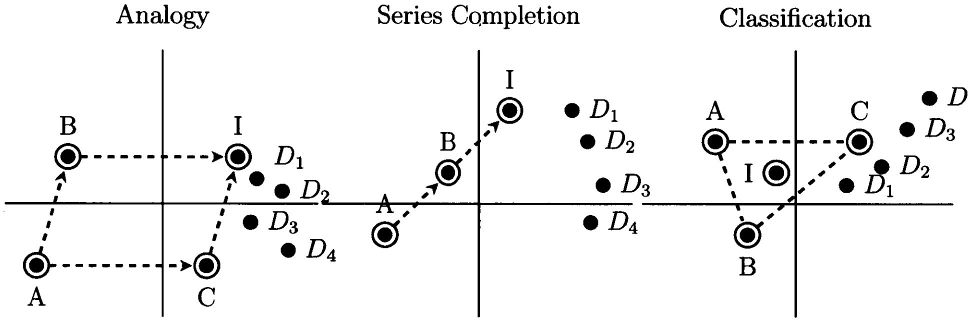


Figure 4-1: Inductive reasoning in semantic space proposed in Sternberg and Gardner [1983]. A, B, C are given, I is the ideal point and D are the choices. The correct answer is shaded green.

In this work, we take *semantic spaces* studied in the cognitive-psychometric literature as the prototypical objects that word embedding algorithms estimate. Semantic spaces are vector spaces over concepts where Euclidean distances between points are assumed to indicate semantic similarities. We link such semantic spaces to word co-occurrences through semantic similarity assessments, and demonstrate that the observed co-occurrence counts indeed possess statistical properties that are consistent with an underlying Euclidean space where distances are linked to semantic similarity.

Formally, we view word embedding methods as performing *metric recovery*. This perspective is significantly different from current approaches. Instead of aiming to find representations that exhibit desirable semantic properties, we seek methods that recover the underlying metric of the hypothesized semantic space. The clearer foundation afforded by this perspective enables us to study whether embedding algorithms indeed succeed. In particular, we ask whether word embedding algorithms are able to recover the metric under specific scenarios. To this end, we unify existing word embedding algorithms as statistically consistent metric recovery methods under the theoretical assumption that co-occurrences arise from (metric) random walks over semantic spaces. The new setting also suggests a simple and direct recovery algorithm which we evaluate and compare against other embedding algorithms.

The main contributions of this work are the following:

- We ground word embeddings in *semantic spaces* via log co-occurrence counts. We show that PMI (point-wise mutual information) relates linearly to human similarity assessments, and that nearest-neighbor statistics (centrality and reciprocity) are consistent with an Euclidean space hypothesis.

- In contrast to prior work Arora et al. [2015], we take *metric recovery* as the key object of study, unifying existing algorithms as consistent metric recovery methods based on co-occurrence counts from simple Markov random walks over graphs and manifolds. The strong link to manifold estimation promises fruitful extensions of word embeddings.

- We propose and evaluate a new principled direct metric recovery algorithm that performs comparably to the existing state of the art on both word embedding and manifold learning tasks, and show that the `GloVe` technique is closely related to the second-order Taylor expansion of our objective.

- We construct and make available two new inductive reasoning datasets beyond analogies – series completion and classification – and demonstrate that word embeddings can be used to solve these as well. For example, in the series completion task, given "body, arm, hand" we find the answer predicted by vector operations on word embeddings to be "fingers".

Task	Prompt	Answer
Analogy	king:man::queen:?	woman
Series	penny:nickel:dime:?	quarter
Classification	horse:zebra:{deer, dog, fish}	deer

Table 4.1: Examples of new inductive reasoning tasks from Sternberg and Gardner [1983].

4.2 Word vectors and semantic spaces

Most current word embedding algorithms build on the *distributional hypothesis* [Harris, 1954] (similar contexts imply similar meanings) so as to tie co-occurrences of words to their underlying meanings. The relation between semantics and co-occurrences has also been studied in psychometrics and cognitive science[Rumelhart and Abrahamson, 1973, Sternberg and Gardner, 1983], often by means of free word association tasks and *semantic spaces.* The semantic spaces, in particular, provide a natural conceptual framework for continuous word representations as vector spaces where semantically related words are close to each other. For example, the observation that word embeddings can solve analogies was already shown in Rumelhart and Abrahamson [1973] using vector representations of words derived from semantic similarity surveys, i.e., similarity judgments between pairs of words.

A fundamental question regarding vector space models of words is whether an Euclidean vector space is a valid representation of semantic concepts. There is substantial empirical evidence in favor of this hypothesis. For example, Rumelhart and Abrahamson [1973] showed experimentally that analogical problem solving with fictitious words and human mistake rates were consistent with an Euclidean space. Sternberg and Gardner [1983] provided further evidence supporting this hypothesis, proposing that general inductive reasoning was based upon operations in metric embeddings. Using the analogy, series completion and classification tasks shown in Table 4.1 as testbeds, they proposed that subjects solve these problems by finding the word closest (in semantic space) to an ideal point: the vertex of a parallelogram for analogies, a displacement from the last word in series completion, and the centroid in the case of classification (Figure 4-1).

We use semantic spaces as the prototypical structures that word embedding methods attempt to uncover, and we investigate the suitability of word co-occurrence counts to recover their metric structure. In the next section, we show that co-occurrences from large corpora indeed relate to semantic similarity assessments, and that the resulting metric is consistent with an Euclidean semantic space hypothesis.

4.3 The semantic space of log co-occurrences

Most word embedding algorithms are based on word co-occurrence counts. In order for such methods to uncover an underlying Euclidean semantic space, we must demonstrate that co-occurrences themselves are indeed consistent with some semantic space. We must relate co-occurrences to semantic similarity assessments, on one hand, and show that they can be embedded into a Euclidean metric space, on the other. We provide here empirical evidence for both of these properties.

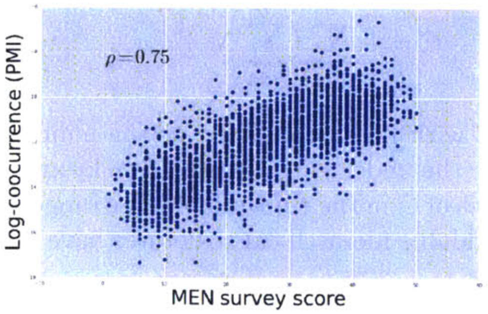


Figure 4-2: Normalized log co-occurrence (pointwise mutual information) linearly correlates with human semantic similarity judgements (MEN survey).

We commence by demonstrating in Figure 4-2 that Pointwise Mutual Information (PMI) evaluated from co-occurrence counts has a strong linear relationship with semantic similarity judgements from survey data (Pearson's r=0.75).[1] This suggestive linear relationship does not, however, by itself demonstrate that log co-occurrences (with normalization) can be used to define an Euclidean metric space.

Earlier psychometric studies employed two nearest neighbor statistics to gauge whether similarity evaluations are embeddable. Specifically, they viewed words as points in an Euclidean space, sampled from some unknown distribution, demonstrating that not all types of similarities are embeddable under this *GS model* [Tversky and Hutchinson, 1986, Schwarz and Tversky, 1980]. We extend this analysis to log co-occurrences and show that semantic similarity estimates from log co-occurrences are actually consistent with an Euclidean semantic space hypothesis.

The first statistic, *centrality* C, is defined in terms of nearest neighbor indicators

[1] Normalizing the log co-occurrence with the unigram frequency taken to the 3/4th power maximizes the linear correlation in Figure 4-2, explaining this choice of normalization in prior work [Levy and Goldberg, 2014a, Mikolov et al., 2013b].

Corpus	C	R_f
Free association	1.51	0.48
Wikipedia corpus	2.21	0.63
Word2vec corpus	2.24	0.73
GloVe corpus	2.66	0.62

Table 4.2: Semantic similarity data derived from multiple sources on multiple tests show evidence for latent embeddings consistent with the *GS model*

$N_{ij} = \mathbb{1}\{\text{i is j's nearest neighbor}\}$ by

$$C = \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{n} N_{ij})^2$$

Under the *GS model*, $C \leq 2$ with high probability as the number of words $n \to \infty$ regardless of the dimension or the underlying density [Tversky and Hutchinson, 1986]. Typical non-asymptotic values of C in the embeddable case range approximately from 1 to 3, while, e.g., non-embeddable hierarchical structures have $C > 10$ [Tversky and Hutchinson, 1986].

The second statistic, *reciprocity fraction* R_f [Schwarz and Tversky, 1980, Tversky and Hutchinson, 1986], is defined as

$$R_f = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} N_{ij}N_{ji}$$

and measures the fraction of words that are their nearest neighbor's nearest neighbor. Under the *GS model*, this fraction should be greater than 0.5 [2]

Table 4.2 shows the two statistics from three popular large corpora and a free word association dataset (see Section 4.6 for details). The nearest neighbor calculations are based on PMI. The results show a surprisingly high agreement on all three statistics for all corpora, with C and R_f contained in small intervals: $C \in [2.21, 2.66]$ and $R_f \in [0.62, 0.73]$. These results are consistent with Euclidean semantic spaces and the *GS model* in particular. The largest violators of C and R_f are consistent with Tversky's analysis: the word with the largest centrality in the non-stopword Wikipedia corpus is 'the', whose inclusion would increase C to 3.46 compared to 2.21 without it. In contrast to Tversky's original (small scale) analysis of semantic similarities, we find

[2] Both R and C are asymptotically dimension independent, due to the fact that they rely on only the single nearest neighbor, and represent very basic properties of embeddability. Understanding the number of necessary dimensions requires other more dimension sensitive measures such as ball-expansion rates, but these measures are often not stable beyond tens of dimensions.

that word co-occurrence and free association data are in fact largely embeddable into a metric space.

The results of this section are an important step towards justifying the use of word co-occurrence counts as the central object of interest for semantic vector representations of words. We have shown that they are *empirically* related to a human notion of semantic similarity and that they are metrically embeddable, a desirable condition if we expect word vectors derived from them to behave as true elements of a metric space. This, however, does not yet fully justify their use to derive semantic representations. The missing piece is to formalize the connection between these co-occurrence counts and some intrinsic notion of semantics, such as the semantic spaces described in Section 4.2. In the next two sections, we establish this connection by framing word embedding algorithms that operate on co-occurrences as metric recovery methods.

4.4 Semantic spaces and manifolds

The notion of semantic spaces and the associated parallelogram rule for analogical reasoning extend naturally to objects other than words. For example, images could be approximately viewed as points in an Euclidean space by representing them in terms of their underlying degrees of freedom (e.g., orientation, illumination). Manifold learning methods such as Isomap [Tenenbaum et al., 2000] aim precisely to uncover such an underlying Euclidean coordinate system, whenever possible. If we view geodesic distances on the manifold (represented as a graph) as semantic distances, then it suffices to isometrically embed these distances in an Euclidean space. Tenenbaum [1998] showed that such isometric embeddings of geodesic distances could solve analogies via the parallelogram rule.

Semantic spaces provide a common foundation for word embeddings and manifold learning. Both approximate semantic distances as Euclidean metrics. The difference is how semantically meaningful distances are extracted between the objects. We have argued that for word embeddings, negative log co-occurrences between words already provide approximate semantic distances as shown in Figure 4-2. For manifold learning, semantic distances (geodesics) may be obtained via neighborhood graphs built from the original, high-dimensional points.

We can now demonstrate that a simple random-walk model of sentences allows us to view word embeddings as a type of manifold learning problem. In Figure 4-2 we argued that co-occurrences between words correlated with semantic similarities. A simple toy model capturing this relationship would be to define a sentence as a random walk X_t where the probability of transitioning from word i to j is

$$P(X_t = j | X_{t-1} = i) = h(||x_i - x_j||_2^2/\sigma) \tag{4.1}$$

where $||x_i - x_j||_2^2$ is distance between words in semantic space and h is an unknown, subgaussian function linking semantic similarity to co-occurrence. [3]

Given a corpus over such sentences with m words and a vocabulary size of n, as m grows large, the co-occurrence converges to the joint probability $P(X_t = j, X_{t-1} = i)$ by the Markov chain law of large numbers. Consider $C_{ij}^{m,n}$ to be the co-occurrence matrix over this corpus. We can show that there exists unigram normalizers $a_i^{m,n}$, $b_i^{m,n}$ such that

Lemma 4.4.1. *Given a corpus generated by Equation 4.1 there exists a_i and b_j such that simultaneously over all i, j:*

$$\lim_{m,n\to\infty} -\log(C_{ij}^{m,n}) - a_i^{m,n} - b_j^{m,n} \to ||x_i - x_j||_2^2$$

if the window size for the co-occurrence is chosen appropriately large (See Corollary C.1.4 for precise statement and conditions). Intuitively, this result is an analog of the central limit theorem for random walks, although we do not know the true link h between semantic similarity and co-occurrence, the random walk structure allows us to approximate it with a squared-exponential form. The practical details of symmetrizing and considering windowed co-occurences only modify constant terms of 4.4.1 (see section C.3 for details).

We can make this connection between metric recovery and word embeddings more precise by providing a reduction from a manifold learning problem to a word embedding problem. Suppose that we are given a manifold learning problem which consists of n points $\{x_1 \dots x_n\} \in \mathbb{R}^d$ drawn i.i.d from a density p lying on a manifold isometrically embeddable into $D < d$ dimensions, and we are asked to find an embedding of $x_1 \dots x_n$ into D dimensions.

Such nonlinear dimensionality reduction problems arising from manifold learning can be solved exactly as word embedding problems in the following way:

1. Construct a neighborhood graph (such as k-nearest neighbors) over $\{x_1 \dots x_n\}$.

2. Record the vertex sequence of a simple random walk over these graphs as a sentence.

[3] This toy model ignores the role of semantics and function words, but the framework is robust to adding such complications as long as the moment bounds originally derive in Hashimoto et al. [2015c] are fulfilled. For examples of such constraints see Theorem A.2.2 and Theorem A.2.3

3. Co-occurrences over a corpus of such sentences fulfill Lemma 4.4.1, replacing the L_2 norm with the geodesic distance.

4. Apply word embedding to co-occurrences over the vertex sequence corpus to generate D-dimensional vectors.

The same argument for the proof of Lemma 4.4.1 demonstrates that with the appropriate graph construction, log co-occurrences will capture the squared geodesic distances[4]. A surprising result is that assuming that Lemma 4.4.1 holds, existing word embedding algorithms are nearly equivalent to manifold learning algorithms, allowing this reduction to solve manifold learning in a formal sense.

4.5 Recovering semantic distances with word embeddings

We can now take the random walk in section 4.4 as a simplified model of how log co-occurrences relate to the semantic space. Our goal is to understand how popular word embedding algorithms behave under this model. We ask the following question: given a corpus generated as a semantic random walk, can word embedding algorithms recover the latent semantic space? We show that, surprisingly, many popular word embedding algorithms consistently recover the latent, unknown embedding.

4.5.1 Word embeddings as metric recovery

Using Lemma 4.4.1, we show that three popular word embedding methods can be viewed as metric recovery algorithms from co-occurrences.

GloVE The Global Vectors (`GloVe`) [Pennington et al., 2014] method for word embedding optimizes the objective function

$$\min_{\widehat{x},\widehat{c},a,b} \sum_{i,j} f(C_{ij})(2\langle \widehat{x}_i, \widehat{c}_j \rangle + a_i + b_j - \log(C_{ij}))^2$$

[4] This approach of applying random walks and word embeddings to general graphs has already been shown to be surprisingly effective for social networks [Perozzi et al., 2014], and demonstrates that word embeddings serve as a general way to connect metric random walks to embeddings.

with $f(C_{ij}) = \min(C_{ij}, 100)^{3/4}$. If we rewrite the bias terms as $a_i = \widehat{a}_i - ||\widehat{x}_i||_2^2$ and $b_j = \widehat{b}_j - ||\widehat{c}_j||_2^2$, we obtain the equivalent representation:

$$\min_{\widehat{x},\widehat{c},\widehat{a},\widehat{b}} \sum_{i,j} f(C_{ij})(-\log(C_{ij}) - ||\widehat{x}_i - \widehat{c}_j||_2^2 + \widehat{a}_i + \widehat{b}_j))^2.$$

Together with Lemma 4.4.1, we recognize this as a weighted multidimensional scaling objective with weights $f(C_{ij})$. Splitting the word vector $\widehat{x}_i$ and context vector $\widehat{c}_i$ is helpful in practice but not necessary under the assumptions of Lemma 4.4.1 since the true embedding $\widehat{x}_i = \widehat{c}_i = x_i/\sigma$ and $\widehat{a}_i, \widehat{b}_i = 0$ is a global minimum whenever $\dim(\widehat{x}) = d$. In other words, GloVE can recover the true metric provided that we set d correctly.

word2vec The embedding algorithm `word2vec` approximates a softmax objective:

$$\min_{\widehat{x},\widehat{c}} \sum_{i,j} C_{ij} \log \left(\frac{\exp(\langle \widehat{x}_i, \widehat{c}_j \rangle)}{\sum_{k=1}^{n} \exp(\langle \widehat{x}_i, \widehat{c}_k \rangle)} \right).$$

Without loss of generality, we can rewrite the above with a bias term b_j by making $\dim(\widehat{x}) = d + 1$ and setting one of the dimensions of $\widehat{x}$ to 1. By redefining the bias $\widehat{b}_j = b_j - ||\widehat{c}_j||_2^2/2$, we see that `word2vec` solves

$$\min_{\widehat{x},\widehat{c},\widehat{b}} \sum_{i,j} C_{ij} \log \left(\frac{\exp(-||\widehat{x}_i - \widehat{c}_j||_2^2/2 + \widehat{b}_j)}{\sum_{k=1}^{n} \exp(-||\widehat{x}_i - \widehat{c}_k||_2^2/2 + \widehat{b}_k)} \right).$$

Since according to lemma 1 $C_{ij}/\sum_{k=1}^{n} C_{ik}$ approaches $\frac{\exp(-|||x_i - x_j||_2^2/\sigma^2)}{\sum_{k=1}^{n} \exp(-|||x_i - x_k||_2^2/\sigma^2)}$, this is the stochastic neighbor embedding (SNE) [Hinton and Roweis, 2002] objective weighted by $\sum_{k=1}^{n} C_{ik}$. Global optimum is achieved by $\widehat{x}_i = \widehat{c}_i = x_i(\sqrt{2}/\sigma)$ and $\widehat{b}_j = 0$. The negative sampling approximation used in practice behaves much like the SVD approach [Levy and Goldberg, 2014b]. In the absence of a bias term, by applying the same stationary point analysis as [Levy and Goldberg, 2014b], the true embedding is a global minimum under the additional assumption that $||x_i||_2^2(2/\sigma^2) = \log(\sum_j C_{ij}/\sqrt{\sum_{ij} C_{ij}})$ [Hinton and Roweis, 2002].

SVD The SVD method of Levy and Goldberg [2014b] uses the log pointwise mutual information matrix defined in terms of the unigram frequency C_i as:

$$M_{ij} = \log(C_{ij}) - \log(C_i) - \log(C_j) + \log\left(\sum_j C_j\right)$$

and applies the SVD to the shifted and truncated matrix : $(M_{ij} + \tau)_+$. This shift and truncation is done for computational reasons and to prevent M_{ij} from diverging. Assuming that the limit of Lemma 4.4.1 holds, and the corpus is sufficiently large that no truncation is necessary (i.e. $\tau = -\min(M_{ij}) < \infty$) we will recover the underlying embedding assuming $\frac{1}{\sigma}||x_i||_2^2 = \log \frac{C_i}{\sqrt{\sum_j C_j}}$ via the law of large numbers since $M_{ij} \to \langle x_i, x_j \rangle$. Centering the matrix M_{ij} prior to SVD would relax the norm assumption, resulting in exactly classical multidimensional scaling [Sibson, 1979].

4.5.2 Metric regression from log co-occurrences

We have shown above that with few additional assumptions and reparameterization, existing embedding algorithms can be seen as consistent metric recovery methods under Lemma 4.4.1. However, Lemma 4.4.1 suggests a more direct regression method for recovering the latent coordinates which we propose here. The new embedding algorithm serves as a litmus test for our metric recovery paradigm.

Lemma 4.4.1 describes a log-linear relationship between distance and co-occurrences. The canonical way to fit such a relationship would be to use a generalized linear model, where the co-occurrences C_{ij} follow a negative binomial distribution $C_{ij} \sim \text{NegBin}(\theta, p)$, where $p = \frac{\theta}{\theta + \exp(-\frac{1}{2}||x_i - x_j||_2^2 + a_i + b_j)}$

Under this overdispersed log linear model

$$\mathbb{E}[C_{ij}] = \exp(-||x_i - x_j||_2^2/2 + a_i + b_j)$$
$$\text{Var}(C_{ij}) = \mathbb{E}[C_{ij}]^2/\theta + \mathbb{E}[C_{ij}]$$

Here, the parameter θ controls the contribution of large C_{ij} and acts similarly to `GloVe`'s $f(C_{ij})$ weight function. Fitting this model is straightforward, as we can define the log-likelihood in terms of the expected rate $\lambda_{ij} = \exp(-||x_i - x_j||_2^2/2 + a_i + b_j)$ as follows

	Google Semantic		Google Synatactic		Google Total		SAT		Classification		Sequence	
Method	L_2	Cos	L_2	Cos	L_2	Cos	L_2	Cos	L_2	Cos	L_2	Cos
Regression	**75.5**	**78.4**	**70.9**	70.8	**72.6**	**73.7**	39.2	37.8	**87.6**	**84.6**	**58.3**	**59.0**
GloVE	71.1	76.4	68.6	71.9	69.6	**73.7**	36.9	35.5	74.6	80.1	53.0	58.9
SVD	50.9	58.1	51.4	52.0	51.2	54.3	32.7	24.0	71.6	74.1	49.4	47.6
Word2vec	71.4	73.4	**70.9**	**73.3**	71.1	73.3	**42.0**	**42.0**	76.4	**84.6**	54.4	56.2

Table 4.3: Accuracies on Google, SAT analogies and on two new inductive reasoning tasks.

	Manifold Learning	Word Embedding
Type	Isomap	Regression
Semantic	**83.3**	70.7
Syntactic	8.2	**76.9**
Total	51.4	**73.4**

Table 4.4: Semantic similarity alone can solve the Google analogy tasks

$$\text{LL}(x, a, b, \theta) = \sum_{i,j} \theta \log(\theta) - \theta \log(\lambda_{ij} + \theta) +$$

$$C_{ij} \log\left(1 - \frac{\theta}{\lambda_{ij} + \theta}\right) + \log\left(\frac{\Gamma(C_{ij} + \theta)}{\Gamma(\theta)\Gamma(C_{ij} + 1)}\right)$$

Optimizing this objective using stochastic gradient descent will randomly select word pairs i, j and attract or repulse the vectors $\widehat{x}$ and $\widehat{c}$ in order to achieve the relationship in Lemma 4.4.1. Our implementation uses the GloVe codebase (Appendix C.4.1).

Relationship to GloVE The overdispersion parameter θ sheds light on the role of GloVe's weight function $f(C_{ij})$. Taking the Taylor expansion of the log-likelihood at $\log(\lambda_{ij}) \approx -\log(C_{ij})$ we find that for a constant k_{ij},

$$\text{LL}(x, a, b, \theta) = \sum_{ij} k_{ij} - \frac{C_{ij}\theta}{2(C_{ij} + \theta)}(\log(\lambda_{ij}) - \log(C_{ij}))^2 + o((\log(\lambda_{ij}) - \log(C_{ij}))^3).$$

Note the similarity of the second order term with the GloVe objective. Both weight functions $\frac{C_{ij}\theta}{2(C_{ij}+\theta)}$ and $f(C_{ij}) = max(C_{ij}, x_{max})^{3/4}$ smoothly converge to $\theta/2$ and x_{max} for large C_{ij}, down-weighting large co-occurrences.

4.6 Empirical validation

We experimentally validate two aspects of our theory: the semantic space hypothesis, and the correspondence between word embedding and manifold learning. Our goal is not to find the absolute best method and evaluation metric for word embeddings, which has been studied in Levy et al. [2015]. Instead we will demonstrate evidence for the semantic space hypothesis, and show that our simple algorithm for metric recovery is competitive with state-of-the-art on both semantic induction and manifold learning.

4.6.1 Datasets

Corpus and training: We trained word embeddings on three different corpora: a Wikipedia snapshot of 03/2015 (2.4B tokens), the corpus used in the original `word2vec` paper Mikolov et al. [2013a] (6.4B tokens), and a combination of Wikipedia with Gigaword5 emulating `GloVe`'s corpus [Pennington et al., 2014] (5.8B tokens). We preprocessed all the corpora by removing punctuation, numbers and lower-casing all the text. We limited the vocabulary to the 100K most frequent words in each corpus, and trained embeddings using four methods: `word2vec`, `GloVe`, randomized SVD (referred to as SVD), and metric regression (referred to as regression). Full implementation details are provided in the Appendix.[5]

For fairness we fix all hyperparameters, develop and test the code for metric regression exclusively on the first 1GB subset of the wiki dataset. For open-vocabulary tasks, we restrict the set of answers to the top 30 thousand words, since this improves performance while covering the majority of the questions.

In the following, we show performance for the `GloVe` corpus throughout but include results for all corpora along with our code package.

Evaluation tasks: We test the quality of the word embeddings on three inductive tasks: analogies, sequence completion and classification (Figure 4-1). For the analogies, we used the standard open-vocabulary analogy task by Google [Mikolov et al., 2013a] (19,544 semantic and syntactic questions). In addition, we use the more difficult SAT analogy dataset (version 3) [Turney and Littman, 2005], which contains 374 questions from actual exams and guidebooks. Each question consists of 5 exemplar pairs of words *word1:word2*, where all the pairs hold the same relation. The task

[5] We used randomized, rather than full SVD due to the difficulty of scaling SVD to this problem size. For performance of full SVD factorizations see Levy et al. [2015].

is to pick from among another five pairs of words the one that best represents the relation represented by the exemplars.

Inspired by Sternberg and Gardner [1983], we propose two new difficult inductive reasoning tasks beyond analogies to verify the semantic field hypothesis: sequence completion and classification. As described in Section 4.2, the former involves choosing the next step in a semantically coherent sequence of words (e.g. *hour*, *minute*, ...), and the latter consists of selecting an element within the same category out of five possible choices. Given the lack of publicly available datasets with questions of this type, we generated our own questions using WordNet [Miller and Fellbaum, 1998] relations in combination with word-word PMI values. These datasets were constructed before training any embeddings to avoid biasing them towards any one method.

For the classification data, we created the in-category words by selecting words from various WordNet relations associated to some root words, after which we pruned down to four words based on PMI-similarity to the root word and the other words in the class. The additional options for the multiple choice question were created searching over words related to the root by a different relation type, and selecting those most similar to the root.

For the sequence data, we obtained WordNet trees of various relation types, and then pruned based on similarity to the root word. For the multiple-choice version of the data, we selected additional (incorrect) options by searching over other words related to the root word, and pruning, as for sequences, based on PMI similarity.

After pruning, we ended up with 215 classification questions and 220 sequence completion questions, of which 51 are open-vocabulary and 169 are multiple choice. These two datasets will be released along with the code used to generate all embeddings.

4.6.2 Results on inductive reasoning tasks

Solving analogies using survey data alone: We demonstrate that word embeddings trained directly on semantic similarity derived from survey data can solve analogy tasks. Extending a study by Rumelhart and Abrahamson [1973], we use a free-association dataset [Nelson et al., 2004] to construct a similarity graph, where vertices correspond to words and the weights w_{ij} are given by the number of times word j was considered most similar to word i in the survey. We take the largest connected component of this graph (consisting of 4845 words and 61570 weights) and embed it using Isomap for which squared edge distances are defined as $-\log(w_{ij}/\max_{kl}(w_{kl}))$. We use the resulting vectors as word embeddings to solve the Google analogy questions [Mikolov et al., 2013a]. The results in Table 4.4 show that embeddings obtained with Isomap on survey data can outperform the corpus based metric regression vectors on semantic, but not syntactic tasks; this is due to the fact that free-association surveys capture semantic, but not syntactic similarity between words.

Analogies: The results on the Google analogies shown in Table 4.3 demonstrate that our proposed framework of metric regression and L_2 distance is competitive with the baseline of `word2vec` with cosine distance. The performance gap across methods is small and fluctuates across corpora, but metric regression consistently outperforms `GloVe` on most tasks and outperforms all methods on semantic analogies, while `word2vec` does better on syntactic categories. For the SAT dataset, the L_2 distance performs better than the cosine similarity, and we find `word2vec` to perform best, followed by metric regression. The results on these two analogy datasets show that directly embedding the log co-occurrence metric and taking L_2 distances between vectors is competitive with current approaches to analogical reasoning.

Sequence and classification tasks: As predicted by the semantic field hypothesis, word embeddings perform well on the two novel inductive reasoning tasks (Examples in Table 4.1) . Again, we observe that the metric recovery approach of metric regression and L_2 distance consistently performs as well as and often better than the current state-of-the-art word embedding methods on these two additional semantic datasets.

4.6.3 Word embeddings can embed manifolds

In section 4.4 we argued that asymptotically, word embedding algorithms could be used to solve manifold learning problems. Surprisingly, this simple reduction is highly effective in practice, making word embeddings solve standard manifold learning benchmarks as well as manifold learning algorithms. The benchmark task we consider is nonlinear dimensionality reduction on the MNIST digits dataset from 256 to two dimensions. Using a four-thousand image subset, we generated a k-nearest neighbor graph ($k = 20$) and generated 10 simple random walks of length 200 from each point resulting in 40,000 sentences each of length 200. We compared the four word embedding methods against standard dimensionality reduction methods: PCA, Isomap, SNE and, t-SNE which use the full image. The cluster purity of an embedding was measured using the percentage of 5-nearest neighbors having the same digit label. The four embeddings shown in Fig. 4-3 demonstrate that metric regression is highly effective at this task, outperforming metric SNE and beaten only by t-SNE

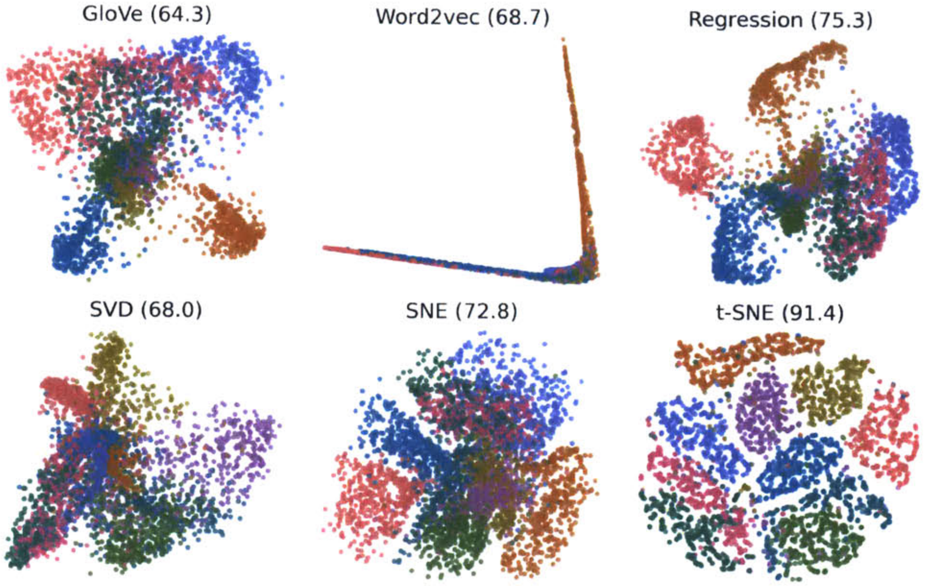


Figure 4-3: Dimensionality reduction using word embedding and manifold learning. Performance is quantified by percentage of 5-nearest neighbors sharing the same digit label.

(91% cluster purity), which is a visualization method designed to preserve cluster separation. All word embedding methods including SVD (68%) embed the MNIST digits well and outperform baselines of PCA (48%) and Isomap (49%) (Other figures included with code package).

4.7 Discussion

Our work recasts word embedding as a metric recovery problem pertaining to the underlying semantic space. We use co-occurrence counts from random walks as a theoretical tool to demonstrate that existing word embedding algorithms are consistent metric recovery methods. Our direct regression method is competitive with the state of the art on various semantics tasks, including on two new benchmark problems of series completion and classification

Our framework highlights the strong interplay and common foundation between word embedding methods and manifold learning, suggesting several avenues for recovering vector representations of phrases and sentences via properly defined Markov processes and their generalizations.

Chapter 5

Population-level estimation of diffusion processes

Thus far we have considered the embeddings of graphs and networks (Chapters 2 and 3) and words (Chapter 4) by applying the continuum limit of a random walk. In this chapter we consider a different continuum limit: we consider the problem of estimating a time-series, and use the continuum limit of a recurrent neural network (RNN) as a way to solve this estimation problem. Many of the techniques used in earlier chapters will be used to characterize the recoverability of the underlying inference problem.

5.1 Motivation

Understanding the population dynamics of individuals over time is a fundamental problem in a variety of areas, from biology (gene expression of a cell population [Waddington et al., 1940]), ecology (spatial distribution of animals [Tereshko, 2000]), to census data (life expectancy [Manton et al., 2008] and racially segregated housing [Bejan and Merkx, 2007]). In such areas, experimental cost or privacy concerns often prevent measurements of complete trajectories of individuals over time, and instead we observe samples from an evolving population over time (Fig. 5-1).

For example, modeling the active life expectancy and disabilities of an individual over time is an area of substantial interest for healthcare statistics [Manton et al., 2008], but the expense and difficulty of collecting longitudinal health data has meant that much of the data is cross-sectional [Robine and Michel, 2004]. Our technique replaces longitudinal data with cross-sectional data for inferring the underlying dynamics behind continuous-time time-series.

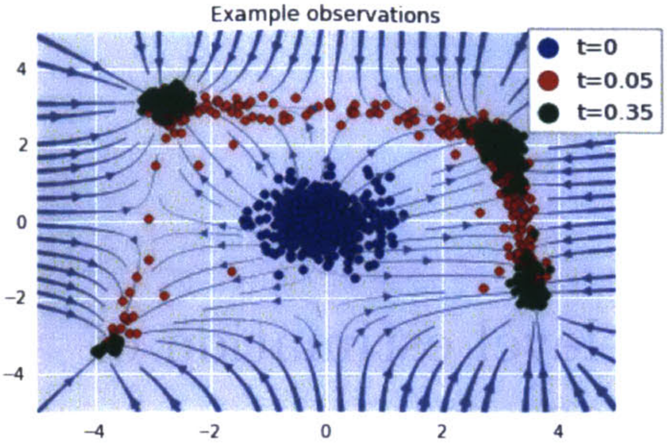


Figure 5-1: In population-level inference we observe samples (colored points) drawn from the process at different times. The goal is to infer the dynamics (blue vectors). In this toy dataset each point can be thought of as a single cell and the x and y axes as gene expression levels of two genes.

The framework we develop will be applicable to the general cross-sectional population inference problem, but in order to ground our discussion we will focus on a specific application in computational biology, where we seek to understand the process by which embryonic stem cells differentiate into mature cells. An individual cell's tendency to differentiate into a mature cell is thought to follow a 'epigenetic landscape' much like a ball rolling down a hill. The local minima of this landscape represents cell states and the slope represents the rate of differentiation [Waddington et al., 1940]. While more recent work has established the validity of modeling differentiation as a diffusion process [Hanna et al., 2009, Morris et al., 2014], direct inference of the epigenetic landscape has been limited to the dynamics of single genes [Sisan et al., 2012] due to the difficulty of longitudinally tracking single cells.

Our work establishes that no longitudinal tracking is necessary and population data alone can be used to recover the latent dynamics driving diffusions. This result allows cheap, high-throughput assays such as single cell RNA-seq to be used to infer the latent dynamics of tens to hundreds of genes.

Analyzing the inference problem for population-level diffusions, we utilize the connection between partial differential equations, diffusion processes, and recurrent neural networks (RNN) to derive a principled loss function and estimation procedure that performs well in practice.

Our contributions are the following

- First, we rigorously study whether the dynamics of a diffusion can be recovered from cross-sectional observations, and establish the first identifiability results.

- Second, we show that a particular regularized recurrent neural network (RNN) with Wasserstein loss is a natural model for this problem and use this to construct a fast scalable initializer that exploits the connection between diffusions and RNNs.

- Finally, our method is verified to recover known dynamics from simulated data in the high-dimensional regime better than both parametric and local diffusion models, as well as predict the differentiation time-course on tens of genes for real RNA-seq data.

5.2 Prior work

Population level inference of dynamics consists of observing samples drawn from a diffusion stopped at various times and inferring the forces driving the changes in the population (Fig. 5-1) which contrasts with inferring dynamics with trajectory data which tracks individuals longitudinally. Our work is distinct from existing approaches in that it considers sampled, multivariate, and non-stationary ($t < \infty$) observations.

5.2.1 Population level inference

Inferring dynamics from population appears in two areas: In home-range estimation, one estimates the support of a two-dimensional time series from the stationary distribution in order to find the grazing area of animals [Fleming et al., 2015]. Our work is distinguished by our focus on the high-dimensional ($d > 2$) and non-stationary settings. Identifiability and limitations of a stationary assumption are discussed in section 5.4.1.

Inverse problems in parabolic differential equations identify dynamics in one to three dimensions given noisy but complete measurements along the boundary [Tarantola, 2005]. These methods require complete observation (rather than samples) and additional boundary conditions which do not hold for our problem. One-dimensional methods which use plug-in kernel density estimates to translate samples to complete observations exist [Lund et al., 2014] but do not generalize to greater than one dimension.

5.2.2 Diffusive RNNs

Diffusive networks [Mineiro et al., 1998] connect diffusion processes and RNNs much like our work. Our work focuses on the more difficult problem of recoverable population-level diffusions (rather than full trajectory observations) and derives new pre-training schemes based on contrastive divergence. Our work shows that the connection between recurrent network and diffusions such as those in Mineiro et al. [1998] can be used to develop powerful inference techniques for general diffusions.

Ideas from diffusions have been used for unsupervised density estimates [Sohl-Dickstein et al., 2015] and our RNN architecture is reminiscent of memory models such as LSTMs [Gers et al., 2000], but both of these models do not make a formal connection to stochastic differential equations which is the focus of this paper.

5.2.3 Computational biology

Pseudo-time analysis [Trapnell et al., 2014] models differentiation of individual cells measured by single-cell RNA-seq by assigning each cell a 'pseudo-time' indicating its level of differentiation. Our approach of modeling the epigenetic landscape is both more general and fully generative. While Pseudo-time describes cells as having one dimension (pseudo-time) with bifurcations, the epigenetic landscape can describe multiple continuous dimensions (such as the toy example in Fig. 5-1) as well as provide predictions of future population distributions.

Although identifying the epigenetic landscape in a multivariate setting requires substantially more cells than identifying a pseudo-time ordering, our results on the identifiability of the epigenetic landscape will become increasingly more valuable as the number of captured cells in a single-cell RNA-seq experiment grows from hundreds [Klein et al., 2015] to tens of thousands.

Systems biology models of the epigenetic landscape have focused on understanding and constructing landscapes which recapitulate the qualitative properties of differentiation systems [Qiu et al., 2012, Bhattacharya et al., 2011]. Our work is instead focused on data-driven identification of the epigenetic landscape. Our results show that our system captures complex qualitative properties of differentiating gene expression.

Existing data-driven models of epigenetic landscape are for a single gene and either rely on longitudinal tracking [Sisan et al., 2012] or require assuming that a particular cell population is stationary [Luo et al., 2013]. Our approach requires neither, and we discuss the limitations of the stationarity assumption in section 5.4.1.

5.3 Population-level behavior of diffusions

We will begin with a short overview of our notation, observation model, and mathematical background.

A d-dimensional diffusion process $X(t)$ represents the state (such as gene expression) of an individual at time t. Formally we define $X(t)$ as a stochastic differential equation (SDE):

$$dX(t) = \mu(X(t))dt + \sqrt{2\sigma^2}dW(t). \tag{5.1}$$

Where $W(t)$ is the unit Brownian motion. This can be thought of as the continuous-time limit of the discrete stochastic process $Y(t)$ as $\Delta t \to 0$:

$$Y(t + \Delta t) = Y(t) + \mu(Y(t))\Delta t + \sqrt{2\sigma^2\Delta t}Z(t) \tag{5.2}$$

where $Z(t)$ are i.i.d standard Gaussians. The function $\mu(x)$ is called the **drift** and represents the force acting on an individual at a particular state x. In Fig. 5-1, the blue curves are $\mu(x)$ which result in $X(t)$ converging to one of four terminal states. The probability of observing $X(t)$ at any point x at time t is called the **marginal distribution** and corresponds to the colored points in Fig. 5-1.

We define the population-level inference task as finding the drift function μ given distributions over the marginals.

Definition 5 (Population-level inference)**.** Define the marginal distribution $\rho(t, x) = P(X(t) = x)$.

A population-level inference problem on $X(t)$ given diffusion constant σ, time points $\mathcal{T} = \{0, t_1 \dots t_n\}$, and samples $\mathcal{M} = \{m_0 \dots m_n\}$ consists of identifying $\mu(x)$ from samples $\{x(t)_i \sim \rho(t, x) \mid i \in \{1 \dots m_t\}, t \in \mathcal{T}\}$.

Fully general population level inference is impossible. Consider a process with the unit disk in $\mathbb{R}^2$ as $\rho(0, x)$, and the drift μ is a clockwise rotation. From a population standpoint, this would look identical to no drift at all.

This raises the question: what restrictions on $\mu(x)$ are natural, and allow for the recovery of the underlying drift? Our paper considers **gradient flows** which are stochastic processes with drift defined as $\mu(x) = -\nabla\Psi(x)$ [1]. The **potential function** $\Psi(x)$ corresponds to the 'epigenetic landscape' of our stochastic process. The force $\mu(x) = -\nabla\Psi(x)$ drives the process $X(t)$ toward regions of low $\Psi(x)$ much like a noisy gradient descent.

[1] For diffusion processes, the gradient flow condition is equivalent to reversibility [Pavliotis, 2014, Section 4.6].

A remarkable result on these gradient flows is that the marginal distribution $\rho(t,x)$ evolves by performing steepest descent on the relative entropy $D(\rho(t,x)||\exp(-\Psi(x)/\sigma^2))$ with respect to the 2-Wasserstein metric W_2. Formally, this is described by the Jordan-Kinderlehrer-Otto theorem [Jordan et al., 1998]:

Theorem 5.3.1 (The JKO theorem). *Given a diffusion process defined by equation 5.1 with $\mu(x) = -\nabla\Psi(x)$, then the marginal distribution $\rho(t,x) = P(X(t) = x)$ is approximated by the solution to the following recurrence equation for $\rho^{(t)}$ with $\rho^{(0)} = \rho(0,x)$.*

$$\rho^{(t+\Delta t)} = \underset{\rho^{(t+\Delta t)}}{argmin} \quad W_2(\rho^{(t+\Delta t)}, \rho^{(t)})^2 \\ + \frac{\Delta t}{\sigma^2} D\left(\rho^{(t+\Delta t)} || \exp\left(\frac{-\Psi(x)}{\sigma^2}\right)\right). \quad (5.3)$$

in the sense that $\lim_{\Delta t \to 0} \rho^{(t)}(x) \to \rho(t,x)$

This theorem is the conceptual core of our approach: the Wasserstein metric, which represents the probability of transforming one distribution to another via purely Brownian motion, will be our empirical loss [Adams et al., 2013]; and the relative entropy $D(\rho||\exp(-\Psi(x)/\sigma^2))$ describing the tendency of the system to maximize entropy, will be our regularizer.

5.4 Recoverability of the potential Ψ

Before we discuss our model, we must first establish that it is possible to asymptotically identify the true potential $\Psi(x)$ from sampled data. Otherwise the estimated $\Psi(x)$ will have limited value as a scientific and predictive tool.

We consider recoverability in three regimes of increasing difficulty. First, in section 5.4.1, we consider the stationary case of observing $\rho(\infty, x)$ which results in a closed-form estimator for Ψ, but requires unrealistic assumptions on our model. Next, in section 5.4.2 we consider a large number of observations across time, and show that exact identifiability is possible. However, this case requires a prohibitively large number of experiments to guarantee identifiability. Finally, in section 5.4.3 we will consider the most realistic case of observing a few observations across time, and discuss the conditions under which recovery of Ψ is possible.

5.4.1 Stationary observations

In the stationary observation model, we are given samples from a fully mixed process $\rho(\infty, x)$. In this case, one time observation is sufficient to exactly identify the potential. This follows from representing the stochastic process in Eq. 5.1 as a parabolic partial differential equation (PDE).

Theorem 5.4.1 (Fokker-Planck [Jordan et al., 1998]). *Given the SDE in equation 5.1, with drift $\mu(x) = -\nabla\Psi(x)$, the marginal distribution $\rho(t, x)$ fulfills:*

$$\frac{\partial \rho}{\partial t} = div(\rho(t,x)\nabla\Psi(x)) + \sigma^2\nabla^2\rho(t,x) \tag{5.4}$$

with given initial condition $\rho(0, x)$.

Now in the stationary case, we can note that the ansatz $\rho(\infty, x) = \exp(-\Psi(x)/\sigma^2)$ gives:

$$0 = \mathrm{div}(\nabla\Psi(x)\rho(\infty,x))/\sigma^2 + \nabla^2\rho(\infty,x)$$

implying that $\exp(-\Psi(x)/\sigma^2)$ is the stationary distribution, and we can estimate the underlying drift as $\nabla\Psi(x) = -\nabla\log(\rho(\infty,x))\sigma^2$. The quantity $-\nabla\log(\rho(\infty,x))\sigma^2$ can be estimated from samples via one step of the mean-shift algorithm [Fukunaga and Hostetler, 1975, Eq. 41].

Although estimation of $\nabla\Psi(x)$ from the stationary distribution is tractable, it has two substantial drawbacks. First, it is difficult to collect samples from the exact stationary distribution $\rho(\infty, x)$; we often collect marginal distributions that are close, but not exactly equal to, the stationary distribution. Second, our estimator $-\nabla\log(\rho(\infty,x))$ is only accurate over regions of high density in $\rho(\infty, x)$ which may be distinct from our region of interest. For differentiation systems, this means we will only know the behavior of $\nabla\Psi(x)$ near the fully differentiated state, rather than over the entire differentiation timecourse.

To make this drawback clear, consider the case where σ^2 is small. The stationary observations from $\exp(-\Psi(x)/\sigma^2)$ will concentrate around the global minimums of $\Psi(x)$ and will therefore only tell us about the local behavior of $\Psi(x)$ around the minima. On the other hand, observing a non-stationary sequence of distributions $\rho(0,x), \rho(t_1,x)\ldots$ does not have this drawback, as $\rho(0, x)$ may be initialized far from the minima of $\Psi(x)$ allowing us to observe how the distribution $\rho(0, x)$ converges to the minima of $\Psi(x)$.

5.4.2 Many time observations

We show that sampling multiple nonstationary timepoints is identifiable, and avoids the drawbacks of a single stationary observation. Consider a observation scheme where we obtain $\rho(0,x), \rho(t_1,x)\ldots$ up to some time $t_n = T$ such that we can estimate one of two quantities reliably:

- **Short-time:** $\frac{\partial \rho}{\partial t}\Big|_T \approx \sum_{i=1}^n \frac{\rho(t_i,x)-\rho(t_0,x)}{t_i-t_0}$
- **Time-integral:** $\int_0^T \rho(t,x)dt \approx \sum_{i=1}^n \rho(t_i,x)/n$

In both of these cases, we can show that the underlying potential $\Psi(x)$ is identifiable via direct inversion of the Fokker-Planck operator. The time-integral model is particularly interesting, as it can be implemented in practice for single cell RNA-seq by collecting cells at uniform times across development [Klein et al., 2015].

Theorem 5.4.2 (Uniqueness of Fokker-Planck like operators). *Let $\Psi(x)$ be a continuously differentiable solution to the following elliptic PDE:*

$$f(x) = \nabla^2\Psi(x)\tau(x) + \nabla\Psi(x)\nabla\tau(x) + \sigma^2\nabla^2\tau(x) \tag{5.5}$$

subject to the constraint $\int \exp(-\Psi(x)/\sigma^2)dx = 1$.

Equation 5.5 is fulfilled in the short-time case with, $f = \frac{\partial \rho}{\partial t}$, $\tau = \rho$ and in the time-integral case, $f(x) = \rho(t_0,x) - \rho(t_n,x)$ and $\tau(x) = \int_0^T \rho(t,x)dt$.

Additionally, the Fokker-Planck equation associated with $\rho(t,x)$ is constrained to domain Ω via a reflecting boundary. Formally, there exists a compact domain Ω with $\langle\nabla\Psi(x)\tau(x) + \sigma^2\nabla\tau(x), n_x\rangle = 0$ for any boundary normal vector n_x with $x \in \partial\Omega$. [2]

Then $\Psi(x)$ is unique up to sets of measure zero in $\tau(x)$.

Proof. Consider any $\Psi_1(x)$ and $\Psi_2(x)$, then by linearity of the PDE, $\Psi'(x) = \Psi_1(x) - \Psi_2(x)$ must be a solution to the homogeneous elliptic PDE

$$0 = \text{div}(\nabla\Psi'(x)\tau(x)) = \nabla^2\Psi'(x)\tau(x) + \nabla\Psi'(x)\nabla\tau(x).$$

Consider the set $R_\varepsilon = \{x : x \in \Omega, \Psi'(x) \leq \min_y \Psi'(y) + \varepsilon\}$. By smoothness of Ψ' and compactness of Ω, for all $\varepsilon > \varepsilon_{min} = \min_y \Psi'(y)$ the region R_ε is compact.

[2] This boundary condition is only necessary to keep the proof simple. We prove a relaxation in theorem D.0.2.

By construction, ∂R_ε can be decomposed into two parts: the boundary of the level set $\Psi'(x) = \min_y \Psi'(y) + \varepsilon$ which we define as $\partial R_\varepsilon^\circ$ and a possibly empty subset of the domain boundary $\partial\Omega$ defined as $\partial\Omega^\circ$.

By the divergence theorem we can integrate the elliptic PDE over any R_ε:

$$\int_{x\in R_\varepsilon} \mathrm{div}(\nabla\Psi'(x)\tau(x))dx = \int_{x\in\partial\Omega\mathrm{o}} \langle\nabla\Psi'(x)\tau(x), n_x\rangle dx + \int_{x\in\partial R_\varepsilon^\circ} |\nabla\Psi'(x)|_2\tau(x)dx = 0$$

By the boundary condition, for any n_x with $x \in \partial\Omega$, $\langle\nabla\Psi_1(x)\tau + \sigma^2\nabla\tau, n_x\rangle = 0$ which implies that $\langle\nabla\Psi'(x)\tau, n_x\rangle = 0$ and therefore $\int_{x\in\partial R_\varepsilon^\circ} |\nabla\Psi'(x)|_2\tau(x)dx = 0$.

By construction, $\tau(x) > 0$ over Ω and therefore $|\nabla\Psi'(x)| = 0$ for all $x \in \partial R_\varepsilon^\circ$. The union of sets $\partial R_\varepsilon^\circ$ contains all of Ω by construction, and therefore for $x \in \Omega$, $|\nabla\Psi'(x)| = |\nabla\Psi_1(x) - \nabla\Psi_2(x)| = 0$. Combined with the normalization constraint, $\int \exp(-\Psi(x)/\sigma^2)dx = 1$, this implies $\Psi_1(x) = \Psi_2(x)$. □

The proof of Thm. 5.4.2 illustrates that the recoverability depends critically on $\tau(x) > 0$. Thus in the time-integral case, the regions which can be clearly recovered are those over which $\tau(x) = \int_0^T \rho(t,x)dt$ has large mass. Compared to the stationary situation, this is substantially better; we will get accurate estimates of Ψ over the entire timecourse of $\rho(0,x)\dots\rho(T,x)$.

Finally, we ask whether Ψ is recoverable when the time observations $\rho(0,x), \rho(t_1,x)\dots$ are sufficiently few and separated in time such that both the short-time and time-integral assumptions are not valid.

5.4.3 Few time observations

In more realistic settings, we may get many samples, but very few time observations such that the time-integral uniqueness theorem does not hold. We analyze this case and establish two results: first, we establish exact identifiability in one dimension (Thm. 5.4.3) and give evidence for the conjecture in multiple dimensions (Cor. 5.4.4). Next, we establish that a sufficiently mixed final time observation is sufficient for uniqueness (Thm. 5.4.5) and derive a model constraint based on this theorem (Eq. 5.6).

In one dimension, three time points are sufficient to recover the underlying potential function[3]:

[3] The requirement of three marginal distributions is arises due to the more general nature of

Theorem 5.4.3 (1-D identifiability). *Assume there exists some c such that $\sigma > c > 0$; boundaries a, b such that $\rho(t, a) = 0$ and $\rho(t, b) = 0$ for all t; and the marginal densities are Holder continuous with $\rho(t, x) \in H^{2+\lambda}$.*

Given $\rho(0, x), \rho(t_1, x), \rho(t_2, x)$ with $0 \neq t_1 \neq t_2 < \infty$, there exists a unique continuous potential $\Psi(x) \in C^1$ fulfilling the Fokker-Planck equation.

Proof. This is a special case of problem 1 considered in GolâĂŽdman [2010] once we set $c(x, t, u) = 1$, $f(x, t) = 0$, $d(x, t, u) = 0$, $b_1(x, t, u) = 0$, $p(x) = d_1(x, t, u) = 0$. The result follows from GolâĂŽdman [2010, Theorem 1]. □

In the multivariate case, the adjoint technique used in GolâĂŽdman [2010] no longer applies, and the equivalent result is an open problem conjectured to be true [De Cezaro and Johansson, 2012]. We believe this conjecture is true and show that for any finite number of candidate Ψ which agrees at two marginals $\rho(0, x)$ and $\rho(t, x)$ we can identify the true potential using a third measurement.

Corollary 5.4.4 (Finite identifiability of Ψ). *Let Ψ_0 and Ψ_1 be candidate potentials such that given $\rho_0(0, x) = \rho_1(0, x)$ and*

$$\frac{\partial \rho_i}{\partial t} = div(\nabla \Psi_i(x) \rho_i(t, x)) + \sigma^2 \nabla^2 \rho_i(t, x)$$

such that $\rho_0(t, x) = \rho_1(t, x)$. Define $\rho_i(t_3, x)$ where $t_3 \sim T$ is a draw from T defined as a random variable absolutely continuous with respect to the Lebesgue measure, then $\rho_1(t_3, x) = \rho_0(t_3, x)$ with probability one if and only if $\forall x$, $\Psi_1(x) = \Psi_0(x)$.

Proof. See Supp. section D.0.1. The statement reduces to short-time uniqueness studied in section 5.4.2. □

In the case that the final marginal distribution $\rho(t_n, x)$ is sufficiently mixed, stationary identifiability allows us to derive an identifiability result regardless of the conjecture.

Theorem 5.4.5 (Relative fisher information constraint). *Let $\rho(0, x)$ and $\rho(t_n, x)$ be marginal distributions associated with the potential Ψ. Then, if the final time $\rho(t_n, x)$ is sufficiently mixed:*

$$-\frac{\partial}{\partial t} D(\rho(t_n, x) || \exp(-\Psi(x)/\sigma^2)) \leq \varepsilon,$$

the problem considered in GolâĂŽdman [2010, Problem 1]. We believe only two marginals are necessary.

all $\widehat{\Psi}$ which are consistent with $\rho(0,x)$ and $\rho(t_n,x)$ with similar mixing constraints: $-\frac{\partial}{\partial t}D(\rho(t_n,x)||\exp(-\widehat{\Psi}(x)/\sigma^2)) \leq \varepsilon$ must imply similar drifts:

$$\int |\nabla\Psi(x) - \nabla\widehat{\Psi}(x)|^2\rho(t_n,x)dx \leq 4\varepsilon.$$

Proof. This follows from a relative fisher information identity in Markowich and Villani [2000, Lemma 4.1]. We reproduce an abbreviated proof for completeness. Since ρ is the solution to the Fokker-Planck equation evolving according to Ψ, we can write $h_t(x) = \rho(t_n,x)/\exp(-\Psi(x)/\sigma^2)$, leading to

$$\begin{aligned}&-\frac{\partial D(\rho(t_n,x)||\exp(-\Psi(x)/\sigma^2))}{\partial t}\\ &=\int \frac{\exp(-\Psi(x)/\sigma^2)}{h_t(x)}|\nabla h_t(x)|^2dx\\ &=\int |\nabla\Psi(x) - \nabla\rho(t_n,x)|^2\rho(t_n,x)dx \leq \varepsilon.\end{aligned}$$

Where the second equality follows via integration by parts on the Fokker-Planck equation. Applying the Minkowski inequality to the last line gives the desired identity. □

Theorem 5.4.5 implies that if we are willing to assume that $\rho(t_n,x)$ is closed to mixed, and we can ensure that our estimated $\widehat{\Psi}$ has a tight bound on $-\frac{\partial}{\partial t}D(\rho(t_n,x)||\exp(-\widehat{\Psi}(x)/\sigma^2))$, then we can recover a good approximation to the true Ψ. In practice this assumption and constraint is straightforward to fulfill: experimental designs often track cell populations until they do not show substantial changes ($\rho(t_n,x)$ is close to mixed) and we can fit $\widehat{\Psi}$ under the constraint that it is smooth with bounded gradient and

$$D(\rho(t_n,x)||\exp(-\widehat{\Psi}(x)/\sigma^2)) \leq \eta. \tag{5.6}$$

Which implicitly bounds the mixedness in Thm. 5.4.5 by the JKO theorem (Thm. 5.3.1). Thus we have established a constraint (Eq. 5.6) and experimental condition (Thm. 5.4.5) under which we can reliably recover the underlying dynamics even with few timepoints.

5.5 Inference

We will show that a Wasserstein loss with an entropic regularization on a noisy RNN is natural for this model.

5.5.1 Loss function and regularization

To motivate the Wasserstein loss, consider the case where we observe full trajectories of a single stochastic process $X(t)$. Then one natural loss function is to consider the expected squared loss between the observed value x_t and the predicted distribution of $X(t)$ under the model.

The Wasserstein distance is exactly the analogous quantity to the L_2 distance when we switch from fully observed trajectories to populations of indistinguishable particles in a diffusion [Adams et al., 2013, Section 3]. We outline the intuition for this argument here: the squared loss for a diffusion arises from the fact that given m_t trajectories with $x(t) = \{x(t)_0, x(t)_1 \dots x(t)_{m_t}\}$, then $\lim_{\widehat{t}\to 0} -\widehat{t}\log(P(X(\widehat{t}+t) = x(\widehat{t}+t)|X(t) = x(t))) = \frac{1}{4}\sum_{i=1}^{m_t} |x(t+\widehat{t})_i - x(t)_i|_2^2$. The usual squared loss is then derived as the log-probability that Brownian motion transforms the predicted value $X(t)$ into the true value $x(t)$ in an infinitesimal time $\widehat{t}$.

If we make the particles indistinguishable via a random permutation $\sigma \in S_{m_0}$, the above limit becomes:

$$\lim_{\widehat{t}\to 0} -\widehat{t}\log(P(X(t+\widehat{t}) = x(t+\widehat{t})|X(t) = x(t))) = \frac{1}{4}\inf_{\sigma\in S_{m_n}} \sum_{i=1}^{m_n} |x(t+\widehat{t})_i - x(t)_{\sigma(i)}|_2^2. \quad (5.7)$$

This is a special case of the Wasserstein metric, implying that for population inference, the natural analog to empirical squared loss minimization is empirical Wasserstein loss minimization. Thus at time t_i we penalize $W_2(\widehat{\rho}(t_i, x), \rho_\Psi(t_i, x))^2$ which is the Wasserstein distance between the empirical distribution $\widehat{\rho}$ and the marginal distribution predicted by Ψ, ρ_Ψ. This loss is approximated via sampling and the Sinkhorn distance [Cuturi, 2013].

We regularize this loss function with an entropic regularizer. Thm. 5.4.5 states that if $\frac{\partial}{\partial t} D(\rho(t_n, x)|| \exp(-\Psi(x)/\sigma^2))$ is small then we can recover any mixed potential. We fulfill this mixing constraint by controlling the relative entropy in Eq. 5.6, which we write as

$$E_{X\sim\rho(t_n,x)}[\log(\rho(t_n, X))] + E_{X\sim\rho(t_i,x)}[\Psi(X)/\sigma^2] \leq \eta,$$

where $\rho(t_n, x)$ is the unknown, true marginal distribution at time t_n. Removing constant terms not involving $\Psi(x)$ and replacing $\rho(t_n, x)$ with samples $x_j \sim \rho(t_n, x)$ gives us the regularizer: $\sum_{j=1}^{m_n} \Psi(x_j)/\sigma^2$. Converting this constraint into a regularization term with parameter τ and assuming that Ψ is contained in a family of models K, our objective function is:

$$\min_{\Psi \in K} \left[\sum_{i=1}^{n} W_2(\widehat{\rho}(t_i, x), \rho_\Psi(t_i, x))^2 \right] + \tau \sum_{j=1}^{m_n} \frac{\Psi(x_j)}{\sigma^2}. \tag{5.8}$$

The similarity of Eq. 5.8 to the JKO theorem (Thm. 5.3.1) is not coincidental. One interpretation of the JKO theorem is that W_2 is the natural metric over marginal distributions and likelihood is the natural measure of model fit over Ψ.

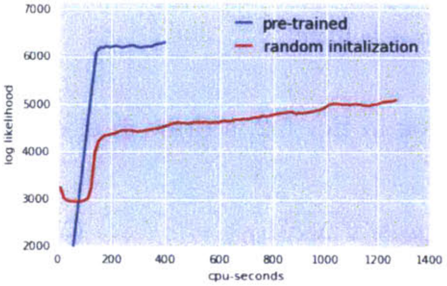


Figure 5-2: Stationary pre-training improves both runtime and goodness of fit

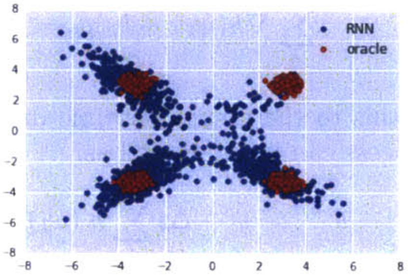


Figure 5-3: RNN predictions are similar to the true dynamics on 50D data

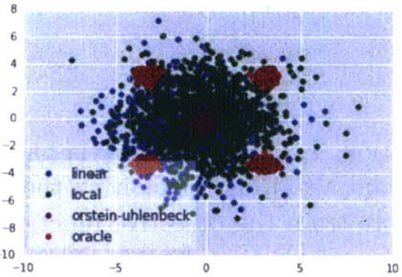


Figure 5-4: Example prediction of baselines on same data

5.5.2 Diffusions as a recurrent network

Thus far we have abstractly considered all stochastic processes of the form: $dX(t) = -\nabla\Psi(x)dt + \sqrt{2\sigma^2}dW(t)$.

A natural way to parametrize Ψ is to consider linearly separable potential functions, which we may write as:

$$\Psi(x) = \sum_k h(w_k x + b_k)g_k,$$

such that h is some strictly increasing function. This represents Ψ as the sum of energy barriers h in the direction of vectors w_k, allowing us to fit our model via gradient descent, while maintaining interpretability of the parameters.

Setting $h(x) = \log(1 + \exp(x))$ parametrizes $\Psi(x)$ as the sum of nearly linear ramps and we obtain that the drift $\nabla\Psi$ is a one layer of a sigmoid neural network, where the linear terms are tied together much like an autoencoder:

$$\sum_k \nabla h(w_k x + b_k)g_k = \sum_k h'(w_k x + b_k)g_k w_k^T$$

Applying this to the first order time discretization in Eq. 5.2, a draw $\overline{y}_i^t$ of our stochastic process can be simulated as:

$$\overline{y}_i^{t+dt} = \overline{y}_i^t + \Delta t \sum_k h'(w_k \overline{y}_i^t + b_k) w_k g_k + \sqrt{\Delta t \sigma^2} z_{it} \tag{5.9}$$

This can be interpreted as a type of RNN with noise based regularization. The network is generative and as $\Delta t \to 0$ the draws from this recurrent net converge to trajectories of the diffusion process X above. [4]

5.5.3 Optimization

Optimizing the full objective function (Eq. 5.8) directly via backpropagation across time is slow and sensitive to the initialization. Exploiting the connection between RNNs and the diffusion, we can pre-train the objective function on the regularizer alone: $\sum_{j=1}^{m_n} \Psi(x_j)/\sigma^2$ under the constraint that $\int \exp(-\Psi(x)/\sigma^2)dx = 1$. We solve this with contrastive divergence [Hinton, 2002] using the first-order Euler scheme in

[4] In practice, we set Δt to be 0.1 which gives at least a ten time-steps between observations in our experiments and find anywhere from five to hundred time-steps between observations to be sufficient.

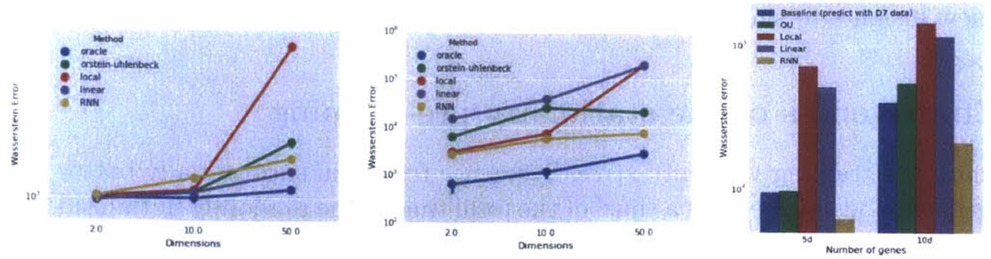


Figure 5-5: Quadratic high dimensional flow

Figure 5-6: Styblinski flow in high dimensions

Figure 5-7: Gene expression (D4)

Figure 5-8: Held-out goodness of fit (lower is better), as measured by Wasserstein distance. 'Oracle' represents the error from Monte Carlo sampling for the true gradient flow. The RNN parametrization performs best across a wide range of tasks.

Eq. 5.9 to generate our negative samples.

After this initialization, we perform backpropagation over time on our objective function, with ρ_Ψ approximated via Monte Carlo samples using Eq. 5.9 and the Wasserstein error approximated using Sinkhorn distances. These stochastic gradients are then used in Adagrad to optimize Ψ[Duchi et al., 2011]. We implement the entire method in Theano, and full code is available in the supplement as an ipython notebook.

5.6 Results

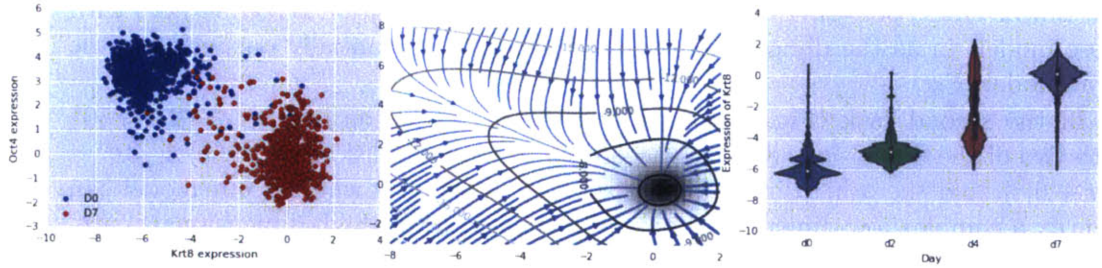


Figure 5-9: D0 and D7 distributions of Oct4 (y-axis) and Krt8 (x-axis)

Figure 5-10: Learned differentiation dynamics

Figure 5-11: Distributions of true Krt8 expression

Figure 5-12: Observed data and learned model for single-cell RNA-seq data

We now demonstrate that the pre-training and RNN parametrization are effective, and bring new insights to single-cell RNA-seq data. [5]

5.6.1 Effectiveness of the stationary pre-training

The stationary pre-training via contrastive divergence results in substantially better training log-likelihoods in less than a third of the total time of the randomly initialized case (Fig. 5-2) for the Himmelblau flow (Fig. 5-1). We control for initialization and runtime of both procedures by ensuring that the initial parameters of the pre-training matches that of the random initialization, and applying shared code for both pre-training and backpropagation.

5.6.2 Learning high dimensional flows

One of the primary advantages of using recurrent networks and sums-of-ramps as a potential is that they behave well in high-dimensional estimation problems. We compare our RNN model against a linear $\Psi(x)$, the Orstein-Uhlenbeck process (quadratic $\Psi(x)$), and a local sum-of-gaussian potentials parametrization for $\Psi(x)$ (details in Sec. D.0.4).

In the first task (Fig. 5-5), we have a population evolution in $\mathbb{R}^d$ for $d \in \{2, 10, 50\}$ according to a unit quadratic potential $\Psi(x) = |x|_2^2$. The initial measurement is 500 samples drawn from a normal distribution with 1/2 scale centered at $(5, 0, 0 \ldots 0)$, and the final time measurement is 500 samples at $t = 1$ with $\sigma = 1.5$. This tests whether our model can recover a simple, high-dimensional potential function. In this case, the simple dynamics mean that the parametric models (Orstein-Uhlenbeck and Linear flows) perform quite well. The RNN parametrization is competitive with these models in as the dimensionality increases, and substantially outperforms the local model (Fig. 5-5).

In the second task (Fig. 5-6), we consider a population over $d \in \{2, 10, 50\}$ with two of the dimensions evolving according to the Styblinski flow ($\Psi(x) = |3x^3 - 32x + 5|^2$), and the other dimensions set to zero. This tests whether our model can identify a complex low-dimensional potential embedded in a high-dimensional space. Example outputs in Fig. 5-3 and 5-4 demonstrate that our RNN model can model the multi-modal dynamics embedded within a high-dimensional space. The quantitative error in Fig. 5-6 shows that the local and RNN methods perform best at low (2-10) dimensions, but the local method rapidly degenerates in higher dimensions. In both

[5] Step-size is selected by grid search (see section D.0.3 for other parameter settings). σ is assumed known in the simulations, and fixed to the observed marginal variance for the RNA-seq data.

cases, our RNN approach produces substantially lower Wasserstein loss compared both parametric and local approaches.

5.6.3 Analysis of Single-cell RNA-seq

In Klein et al. [2015] an initially stable embryonic stem cell population (termed 'D0' for day 0) begins to differentiate after removal of LIF (leukemia inhibitory factor) and single-cell RNA-seq measurements are made at two, four, and seven days after LIF removal. At each time point, the expression of 24175 genes for several hundred cells (933 cells at D0, 303 at D2, 683 at D4, and 798 at D7) are measured. We apply standard normalization procedures [Hicks et al., 2015] to correct for batch effects, and impute dropout expression levels using nonnegative matrix factorization. Our tasks to predict the gene expression at D4 given only the D0 and D7 expression values.

Fitting our RNN model on the top five and ten most differential genes as determined by the Wassertein distance between D0 and D7 distributions, our RNN method performs best compared to baselines (Fig5-7), and is the only one to perform better than predicting the D4 gene expression with D7 data. We find that ten genes is the limit for with a few hundred cells and the RNN begins to behave much like the linear model with more genes. As the number of captured cells in single-cell RNA-seq grows, our RNN model will be capable of modeling more complex multivariate potentials.

We now focus on whether our model captures the qualitative dynamics of differentiation for the two main differentiation markers studied in Klein et al. [2015]: Keratin 8 (Krt8) which is an epithelial marker and Oct 4 (Pou5f1) which is a embryonic marker. Krt8 in particular shows two sub-populations at day 4 (Fig. 5-11) suggesting that epigenetic landscape may have multiple minima.

Fitting our RNN on this two dimensional problem shown in Fig. 5-9 we obtain a potential function with a single minimum (Fig. 5-10) demonstrating that differentiation is concentrated around a linear path connecting the D0 and D7 distributions. Surprisingly, this simple unimodal potential predicts a bimodal distribution for the D4 Krt8 distribution shown in Fig. 5-13 despite the lack of bimodality in either the input data (Fig 5-9) or the potential (Fig 5-10). [6]

The bimodality arises from modeling the quantitative dynamics from D0 to D7, and would not be possible to predict from either cluster or pseudo-time based analysis which would predict a unimodal distribution for D4. Estimation of the epigenetic

[6] Similar qualitative results hold for D4 Krt8 expression under five and ten-dimensional versions (Supp. Fig. D-1, D-2, D-3, D-4).

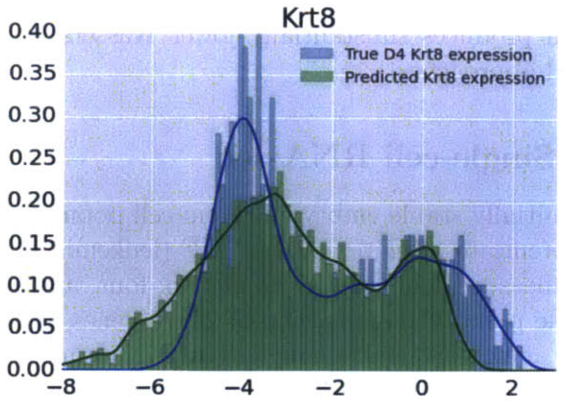


Figure 5-13: D4 predictions of Krt8 recapitulate bimodality

landscape offers an alternative and expressive representation of the cellular differentiation process.

5.7 Discussion

Our work establishes the problem of recovering an underlying potential function using samples from the population distribution. Using a variational interpretation of diffusions, we derive both loss and regularizers for the problem which are natural and amenable to scalable gradient based methods. Finally, we demonstrate through multiple synthetic datasets that our model performs well in a high-dimensional setting and captures unique dynamics of cellular differentiation in single-cell RNA seq data.

Chapter 6

Open problems and future work

This thesis represents the first, general attempts to unify machine learning over graphs and words with traditional unsupervised learning. We show that existing network analysis and word embedding algorithms are variants of well-understood metric algorithms in machine learning. The techniques developed in this thesis are general to any discrete structures which admit a natural random walk and allow for a tight link between the theory of embeddings and similarities.

Specifically, we applied techniques from diffusion processes to machine learning, and used this to substantially simplify the analysis of algorithms when there exists a latent, metric embedding. The method we develop uses the Stroock-Varadhan lemma (Theorem A.2.3) which is a very general way for mapping discrete-time random walks to diffusions. We believe that this technique can be applied to a variety of iterative algorithms in machine learning due to its simplicity.

However, the thesis also leaves open several major questions on the applicability of metric embeddings, as well as the convergence rates of estimators relying upon diffusion approximations. In this section, we briefly discuss the appropriateness of our assumptions as well as review conjectures made in the prior sections.

6.1 Robustness of embeddings to non-metric data

The assumption of a latent metric space associated with discrete objects is a common and convenient *implicit* assumption as discussed in the motivation of Chapter 3. Formally, these types of metric assumptions have been studied in the psychometric literature as part of understanding multidimensional scaling algorithms. For example, the *GS* model proposed by [Tversky and Hutchinson, 1986] formalizes the idea of a metric space by proposing that similarities between objects arise from a latent metric space where the coordinates are drawn independently from a common, smooth distribution. This is a minor weighted generalization of our metric graph model (Definition 1) used throughout the thesis.

How appropriate is the *GS* model as an assumption? For word embeddings we demonstrate that this assumption is a reasonable one based on several test statistics proposed in the literature (Section 4.3). For un-weighted graphs, it is unclear if the *GS* model can be directly tested, as we need to define an appropriate similarity metric over vertices. In future work, we hope to see if there is a single, universal distance function between vertices (such as the log-LTHT) that can be used to determine if a social network is compatible with a latent embedding assumption.

In the case that the metric assumption such as the *GS* model is violated, the accuracy of the estimators presented in this thesis would depend heavily on the type of non-metric violation. For example, if a part of the graph is metric and the non-metric component arises randomly, then our robustness results such as (Theorem 3.3.7) imply that we would still be able to consistently recover the underlying metric component of the graph. In the case that the non-metric violation is systematic, for example because the graph is bipartite, or because the graph is actually a block-model, then our estimators have no theoretical guarantees on performance. Optimization based estimators would embed these graphs on a best-effort basis, but it is still not clear if that is sufficient for obtaining reasonable embeddings.

A important direction for this line of work is to consider the implications of non-metric, standard network models such as block-models or preferential attachment type models. In these cases, it is possible to define a notion of similarity (for example, whether two vertices belong to the same cluster, or attached to the same hub) but it is clear that no exact embedding exists. The hitting time proofs used for the continuum limit are no longer applicable, but short-time asymptotic expansions used in Theorem B.5.1 could still be used to show that embeddings capture the underlying graph structure.

It is an interesting conjecture whether metric embedding type methods can recover the phase-transition of block-model recovery. Such a result would substantially increase the importance of network embedding methods in the analysis of network algorithms and properties.

6.2 Equicontinuity as an open conjecture

The other large assumption made in this work is that of equicontinuity of various properties of the random walk. To review, the results on density estimation require equicontinuity of the stationary distribution (Conjecture A.1.1):

Conjecture 6.2.1. *Given the other continuity and scaling conditions on $p(x)$ and $\varepsilon_n(x)$ in $(\star)$, $n\pi_{X_n}(x)$ is a.s. uniformly equicontinuous.*

For hitting times, we require a slightly stronger version of this condition in terms of equicontinuity of the marginal distribution:

Definition 6. If $t = \Theta(g_n^{-2})$, with probability 1, for any $\delta > 0$, there exist $\gamma > 0$ and n_0 so that for $n > n_0$, any $x_k \in \mathcal{X}_n$, and any $x_i, x_j \in \mathcal{X}_n$ with $|x_i - x_j| < \gamma$, we have

$$|nq_t(x_i, x_k) - nq_t(x_j, x_k)| < \delta.$$

The reason why these conditions are necessary is that the Stroock-Varadhan criterion (Theorem A.2.3) guarantees that two stochastic processes converge weakly in measure. This means that integrating any fixed continuous function against both the discrete random walk and the diffusion produce identical results. This is close, but not exactly the result we need, which is to show that the stationary density function (or the transition density in the case of the second version of the conjecture) between the two processes converge.

In order to go from weak convergence of measures to convergence of densities, it is sufficient to assume that the stationary distribution $\pi_{X_n}(x)$ is smooth. However, this result turns out to be extremely difficult to obtain. As discussed in Section B.1, this result is trivial if the graph is undirected, since a random walk can be defined to increase smoothness over time using standard spectral graph theory arguments. However, as soon as the graph becomes directed, the same argument breaks down. Cycles and other adversarial structures can be very badly behaved, and prevent the stationary distribution from being smooth.

Somehow, metric graphs are free of this phenomenon since they are 'almost reversible' graphs. That is to say, their limiting diffusions are reversible stochastic processes, and we can show in a precise sense that adversarial cycle like behavior is unlikely. However these techniques all fall short of proving the conjecture above. While standard spectral graph based approaches do not seem to be able to solve the conjecture, we are hopeful that more advanced techniques for analyzing graphs, such as directly characterizing the entropy of a random walk will do so.

6.3 Finding further generalizations of continuous embeddings as relaxations

Finally, while we have emphasized that the continuous diffusion limit is general, it is not clear whether this approach can be generalized beyond the obvious random-walk and time-series applications. Chapter 5 provides a new way forward by analyzing iterative algorithms such as recurrent neural nets as a type of diffusion process, but it is an open question whether there exists an even more general framework for understanding a diffusion limit as a type of continuous relaxation, much like how a convex relaxation is viewed as a general approach to difficult optimization problems.

The technical complexity of the arguments is one of the main roadblocks to widespread application of diffusion limits. This thesis provided first steps towards providing a simpler way of applying diffusion limits by presenting a simple, general scheme for applying diffusion limits via calculation of the first three moments of an increment of a stochastic process (Theorem A.2.3). Future application of diffusion limits to stochastic gradient methods and neural networks could help popularize such diffusion methods.

Finally, we have not fully resolved the question of generating semantically meaningful representations in an un-supervised manner. For example, we may ask whether the same strategy as word embeddings can be generalized to un-ordered collections of images. The random-walk analysis used in this thesis no longer applies in this case, and a new theory would have to be developed which links semantics and un-ordered collections. Clarifying the connection between semantic spaces and manifold learning could lead to a general theory of semantically meaningful continuous representations.

Appendix A

Density estimation

A.1 Conjecture on eventual uniform equicontinuity of the rescaled stationary distribution

In the conditions ($\star$) we imposed, we required the eventual uniform equicontinuity of $n\pi_{X_n}$. Without this condition, our proof technique implies the weak convergence

$$\sum_{x \in \mathcal{X}_n} \pi_{X_n}(x)\delta_x \to \pi_Y(x)dx$$

of the empirical stationary measures of $X_n(t)$ to the stationary measure of $Y(t)$. The additional imposition of eventual uniform equicontinuity was required solely to upgrade this convergence to a convergence of the rescaled discrete density functions to the continuous density function. We conjecture that this continuity is true in general.

Conjecture A.1.1. *Given the other continuity and scaling conditions on $p(x)$ and $\varepsilon_n(x)$ in ($\star$), $n\pi_{X_n}(x)$ is a.s. uniformly equicontinuous.*

We discuss a few reasons why we might believe this conjecture to hold.

- In the case of constant $\varepsilon_n(x)$, $n\pi_{X_n}(x)$ is proportional to $|\mathsf{NB}_n(x)|$, hence converges to $p(x)$ uniformly. The conjecture therefore holds in this case.

- Our empirical results produce robust results across a broad range of n, $\overline{\varepsilon}(x)$, and $p(x)$. One possible explanation would be that Conjecture A.1.1 holds for all datasets constructed according to ($\star$).

- For $x, y \in \mathcal{X}_n$, let $r_n(x)$ denote the expected first return time to x and $c_n(x,y)$ denote the expected commute time from x to y. It is known that

  $$\pi_{X_n}(x) = \frac{1}{r_n(x)},$$

  so to show that $n\pi_{X_n}(x)$ is uniformly equicontinuous, it suffices to show that $\frac{n}{r_n(x)}$ is uniformly equicontinuous. Notice that

  $$r_n(x) \leq c_n(x,y) + r_n(y) + c_n(y,x)$$

  and that

  $$r_n(y) \leq c_n(x,y) + r_n(x) + c_n(y,x),$$

  which together imply that

  $$|r_n(x) - r_n(y)| \leq |c_n(x,y) + c_n(y,x)|.$$

  This relates continuity of $r_n(x)$ and hence $\pi_{X_n}(x)$ to the commute time $c_n(x,y)$. On the other hand, our techniques using the Stroock-Varadhan criterion yield convergence of the simple random walk $X_n(t)$ to the Itô process $Y(t)$ in $\mathsf{D}([0,\infty), \overline{D})$ without assumption of eventual uniform equicontinuity. In a scaling limit, this should lead to a relation between $c_n(x,y)$ and a rescaling of the commute time of the corresponding Itô process. In future work, we intend to use this result to relate a scaling of $c_n(x,y)$ to $|x-y|$ and approach Conjecture A.1.1 in conjunction with new methods for metric estimation.

A.2 Full proof of Theorem 2.2.1

The goal of this section will be to give a fully rigorous proof of Theorem 2.3.4 from the main text. We first restate the theorem as Theorem A.2.1.

Theorem A.2.1. *Under $(\star)$, if $h_n \to g_n^2$ as $n \to \infty$, then a.s. in $\mathcal{X}$, the process $X_n(\lfloor t/h_n \rfloor)$ converges in $\mathsf{D}([0,\infty), \overline{D})$ to the isotropic $\overline{D}$-valued Itô process $Y(t)$ with reflecting boundary condition defined by*

$$dY(t) = \frac{\nabla p(Y(t))}{3p(Y(t))}\bar{\varepsilon}(Y(t))^2 dt + \frac{\bar{\varepsilon}(Y(t))}{\sqrt{3}} dW(t), \tag{A.1}$$

where the precise meaning of the reflecting boundary condition is given in Subsec-

tion A.2.1.

Our technique is an application of the Stroock-Varadhan criterion (see [Stroock and Varadhan, 1971a, Theorem 6.3]) for convergence of discrete time Markov processes in a bounded domain to drift-diffusion processes with reflecting boundary conditions in that domain. In what follows, we preserve the notation used by Stroock and Varadhan [1971a] whenever possible.

A.2.1 Definition of the objects

In this subsection, we recall in detail the problem setup. We are given an infinite sequence $\mathcal{X} = \{x_1, x_2, \dots\}$ of latent coordinate points drawn independently from a distribution with probability density $p(x)$ in $\mathbb{R}^d$ supported on a compact domain $D \subset \mathbb{R}^d$ with smooth boundary ∂D. We may then find a bounded C^2 function $\phi(x)$ on $\mathbb{R}^d$ so that

$$D = \{x \mid \phi(x) > 0\}, \qquad \partial D = \{x \mid \phi(x) = 0\}, \text{and} \qquad |\nabla \phi(x)| \geq 1 \text{ on } \partial D.$$

We fix a single random draw of $\mathcal{X}$ and analyze the quenched setting.

We are then given a radius function $\varepsilon_n(x_i)$ which may depend on the draw of $\mathcal{X}$ and a scaling factor g_n so that

$$\lim_{n\to\infty} g_n^{-1} \varepsilon_n(x) = \overline{\varepsilon}(x)$$

for some deterministic $\overline{\varepsilon}(x)$ on $\overline{D}$. Let $G_n = (\mathcal{X}_n, E_n)$ be the unweighted directed neighborhood graph with vertex set $\mathcal{X}_n = \{x_1, \dots, x_n\}$ and with a directed edge from i to j if and only if

$$|x_i - x_j| < \varepsilon_n(x_i).$$

Note that G_n is stochastic and depends on the specific realization of $\mathcal{X}_n$ which is drawn.

Let $X_n(t)$ be the simple random walk on the directed graph G_n so that $X_n(t)$ is a discrete-time Markov process with state space $\mathcal{X}_n$. We normalize the timestep of $X_n(t)$ to be $h_n = g_n^2$ and identify $X_n(t)$ with the continuous time process given by $t \mapsto X_n(\lfloor t/h_n \rfloor)$. From now on, we refer to these two processes interchangeably.

In Theorem A.2.1, we wish to show that $X_n(t)$ converges weakly in $\mathsf{D}([0, \infty), \overline{D})$ to the continuous-time continuous-space Itô process $Y(t)$ defined by (A.1) with reflecting boundary conditions. We interpret the boundary conditions in terms of the submartingale condition of Stroock and Varadhan [1971a]. That is, we define the vector function $\gamma(s,x)$ to be the normal vector to ∂D at x whose length is normalized so that

$$\langle \gamma(s,x), \nabla\phi(x)\rangle = 1.$$

Take also the scalar function $\rho(s,x) = 0$. Together, γ and ρ specify the boundary conditions in the following sense.

We say that a process $Y(t)$ solves the submartingale problem for a, b, ρ, and γ if for any function $f \in \mathsf{C}_0^{1,2}([0,\infty) \times \overline{D})$ satisfying

$$\rho(\partial f/\partial t) + \langle \gamma, \nabla f\rangle \geq 0$$

on $[0,\infty) \times \partial D$, the random variable

$$f(t, Y(t)) - \int_0^t (f_s + L_s f)(s, Y(s))\, 1_D(Y(s))ds$$

is a submartingale, where

$$L_s f = \frac{1}{2}\sum_{i,j=1}^{d} a^{ij}\frac{\partial^2 f}{\partial x_i \partial x_j} + \sum_{j=1}^{d} b^j \frac{\partial f}{\partial x_j}.$$

As explained in Stroock and Varadhan [1971a], when $\rho = 0$, this formulation is equivalent to specifying that $Y(t)$ satisfies (A.1) on the interior of D and has reflecting boundary conditions on ∂D.

A.2.2 Quantities used in the Stroock-Varadhan criterion

We now define the moment and boundary quantities which are used in the Stroock-Varadhan criterion. We follow the notations of Stroock and Varadhan [1971a]. Our discrete time Markov process $X_n(t)$ has time increment $h_n = g_n^2$ and transition kernel

$$\Pi_n(x, A) = p(X_n(t+1) \in A | X_n(t) = x) = \frac{|\mathcal{X}_n \cap A \cap B(x, \varepsilon_n(x))|}{|\mathcal{X}_n \cap B(x, \varepsilon_n(x))|}$$

for $x \in \mathcal{X}_n$, where we recall that $\mathcal{X}_n \cap B(x, \varepsilon_n(x)) = \mathsf{NB}_n(x)$.

The moment quantities in Stroock and Varadhan [1971a] are the discrete time drift b_n, diffusion a_n, and tail $\Delta_{n,\alpha}$ coefficients, defined for $x \in \mathcal{X}_n$ by

$$\begin{aligned} a_n^{ij}(s,x) &= \frac{1}{h_n}\int (y_i - x_i)(y_j - x_j)\Pi_n(x,dy) = \frac{1}{h_n}\sum_{y \in \mathsf{NB}_n(x)} \frac{(y_i - x_i)(y_j - x_j)}{|\mathsf{NB}_n(x)|} \\ b_n^i(s,x) &= \frac{1}{h_n}\int (y_i - x_i)\Pi_n(x,dy) = \frac{1}{h_n}\sum_{y \in \mathsf{NB}_n(x)} \frac{y_i - x_i}{|\mathsf{NB}_n(x)|} \\ \Delta_{n,\alpha}(s,x) &= \frac{1}{h_n}\int |y - x|^{2+\alpha}\Pi_n(x,dy) = \frac{1}{h_n}\sum_{y \in \mathsf{NB}_n(x)} \frac{|y - x|^{2+\alpha}}{|\mathsf{NB}_n(x)|}. \end{aligned}$$

The boundary conditions are specified by γ and ρ, where we recall that $\rho \equiv 0$. We note that γ has the alternate expression

$$\gamma(s,x) = C_\gamma(x) \lim_{n\to\infty} \varepsilon_n(x)^{-1} \int_{|y|<\varepsilon_n(x)} y \frac{p(x+y)}{p_{\varepsilon_n(x)}(x)} dy,$$

where $p_r(x) = \int_{|y|<r} p(x+y)dy$ for a normalization factor $C_\gamma(x)$. Define $J_0 = \{(t,y) : \rho(t,y) = 0\}$ and J_1 as its complement. In our setting, $J_0 = \partial D$ and J_1 is empty.

Remark. In the definitions above, we have included possible time dependence in all quantities to be consistent with the notation of Stroock and Varadhan [1971a]. However, all processes we consider are time-independent, so this dependence will not exist in our case.

A.2.3 Statement of the Stroock-Varadhan criterion

We now state two theorems of Stroock-Varadhan which together imply the convergence of $X_n(t)$ to $Y(t)$ in $\mathsf{D}([0,T],\overline{D})$ for any $T > 0$. These theorems will depend on several conditions which we label A-E and F1-4 and check in the next subsection.

Remark. By Whitt [1980, Theorem 2.8], convergence in $\mathsf{D}([0,T],\overline{D})$ for all $T > 0$ implies convergence $\mathsf{D}([0,\infty),\overline{D})$. Further, by [Ethier and Kurtz, 1986, Theorem 4.9.12], this implies weak convergence of the stationary measures of $X_n(t)$ to the stationary measure of $Y(t)$.

The first theorem yields tightness of measures of $X_n(t)$ on Skorokhod space.

Theorem A.2.2 ([Stroock and Varadhan, 1971a, Theorem 6.1]). *Suppose a discrete time Markov process $X_n(t)$ satisfies the following conditions.*

A. (bounded tail mass): For some $\alpha > 0$, as $n \to \infty$, we have

$$\sup_{0 \le t \le T} \sup_{x \in G} \Delta_{n,\alpha}(t, x) \to 0.$$

B. (all large drifts are reflections): There exists M and c such that for all $n > n_0$, $|b_n(t, x)| > M$ implies $\frac{\langle \nabla\phi(x), b_n(t,x)\rangle}{|b_n(t,x)|} \ge c$.

C. (bounded drift outside boundary): For every $\delta > 0$ there exists some $M_\delta < \infty$ such that for all $n > n_0$, $|b_n(t, x)| > M_\delta$ implies $\phi(x) < \delta$.

D. (bounded diffusion): There exists $M < \infty$ such that for all $n > n_0$,

$$\sup_{0 \le t \le T} \sup_{x \in G} ||a_n(t, x)|| \le M,$$

where $|| \cdot ||$ denotes the Frobenius norm.

Then, the family of distributions P_x^n induced by $X_n^x(t)$ over trajectories is conditionally compact in $\mathsf{D}([0, T], \overline{D})$. Moreover, any weak limit of these is concentrated on the subset $\mathsf{C}([0, T], \overline{D}) \subset \mathsf{D}([0, T], \overline{D})$.

The next theorem yields convergence of $X_n(t)$ under convergence of the moment quantities and some regularity conditions on the boundary.

Theorem A.2.3 ([Stroock and Varadhan, 1971a, Theorem 6.3]). *Suppose $X_n(t)$ satisfies the following.*

E. (convergence of coefficients): Drift and diffusion coefficients a_n and b_n converge uniformly on compact subsets $K \subset [0, T] \times D$ to some a and b.

F1. (reflectivity at absorbing boundary): Given $(t, y) \in J_1$ and $\varepsilon > 0$, there exists $n_0 < \infty$, $\delta_0 > 0$ such that if $|t - s| < \delta_0$, $|x - y| < \delta_0$, $n > n_0$ and $\langle \nabla\phi(x), a_n(s, x)\nabla\phi(x)\rangle < \delta_0$ the following hold:

$$|a_n(s, x)| < \varepsilon \qquad \text{and} \qquad |b_n(s, x) - \rho^{-1}(t, y)\gamma(t, y)| < \varepsilon.$$

F2. (bounded drift under absorption): Given $(t, y) \in J_1$ there exist $M_0 < \infty$ and $\delta_0 > 0$ such that if $|s - t| < \delta_0$ and $|y - x| < \delta_0$, then

$$|b_n(s, x)| \le M_0 \quad \text{for all} \quad n.$$

F3. (drift dominates diffusion on reflection): Given $(t, y) \in J_0$ and $M < \infty$ there exist $\delta_0 > 0$ and $n_0 < \infty$ such that if $|t - s| < \delta_0, |x - y| < \delta_0, n > n_0$, and $\langle \nabla\phi(x), a_n(s, x)\nabla\phi(x)\rangle < \delta_0$, we have

$$|b_n(s, x)| \geq M.$$

F4. (drifts at boundary simulate reflection): Given $(t, y) \in J_0$ and $\varepsilon > 0$ there exist $\delta_0 > 0, n < \infty$ and $M < \infty$ such that if $|s - t| < \delta_0, |x - y| < \delta_0, n > n_0$, and $|b_n(s, x)| > M$, then

$$\left| \frac{b_n(s, x)}{\langle b_n(s, x), \nabla\phi(x)\rangle} - \gamma(t, y) \right| < \varepsilon.$$

Then any weak limit $Y(t)$ of $X_n(t)$ in $\mathsf{D}([0, T], \overline{D})$ solves the submartingale problem for a, b, ρ, and γ.

Finally, we state a criterion for uniqueness of solution to the submartingale problem for a, b, ρ, and γ.

Theorem A.2.4 ([Stroock and Varadhan, 1971a, Theorem 5.8]). *Suppose a, b, ρ, and γ are time independent and satisfy the following conditions.*

1. *a is continuous, symmetric, and positive definite on $\overline{D}$;*

2. *b is bounded and measurable;*

3. *γ is bounded, locally Lipschitz, and on ∂D satisfies*

$$\langle \gamma(x), \nabla\phi(x)\rangle \geq \beta > 0;$$

4. *$\rho(x)$ is bounded, continuous, and non-negative.*

Then the solution to the submartingale problem for a, b, ρ, and γ is unique.

Combining Theorem A.2.2, Theorem A.2.3, and Theorem A.2.4 yields the following conclusion.

Corollary A.2.5. *Suppose that $X_n(t)$ satisfies the conditions of Theorem A.2.2, Theorem A.2.3, and Theorem A.2.4. Then $X_n(t)$ converges to $Y(t)$ in $\mathsf{D}([0, T], \overline{D})$.*

Proof. By Theorem A.2.2, some subsequential limit of $X_n(t)$ exists. Theorem A.2.3 implies that any such limit is a solution to the submartingale problem for a, b, ρ, and γ, so the uniqueness of Theorem A.2.4 yields the desired result. □

A.2.4 Verification of the Stroock-Varadhan conditions

We now verify each of the nine conditions necessary for weak convergence. Conditions F1 and F2 are vacuous because J_1 is empty for us. We now verify each of the remaining conditions.

Moment conditions

Theorem A.2.6 (Condition A). *As $n \to 0$, we have*

$$\sup_{0 \leq t \leq T} \sup_{x \in D} \Delta_{n,1}(t,x) \to 0.$$

Specifically, we have

$$\Delta_{n,1}(s,x) \to \lim_{n\to\infty} \frac{1}{h_n} \int_{|y|<\varepsilon_n(x)} |y|^3 \frac{p(x+y)}{p_{\varepsilon_n(x)}(x)} dy = 0.$$

Proof. From Lemma 2.3.2 with $f(x) = |x|^3$. □

Theorem A.2.7 (Condition E). *The sequences of drift and diffusion coefficients $a_n \to a$ and $b_n \to b$ converge uniformly on compact subsets $K \subset [0,T] \times G$. More specifically, the limiting quantities are*

$$a_n^{ij}(s,x) \to \lim_{n\to\infty} \frac{1}{h_n} \int_{|y|<\varepsilon_n(x)} y_i y_j \frac{p(x+y)}{p_{\varepsilon_n(x)}(x)} dy$$
$$b_n^{i}(s,x) \to \lim_{n\to\infty} \frac{1}{h_n} \int_{|y|<\varepsilon_n(x)} y_i \frac{p(x+y)}{p_{\varepsilon_n(x)}(x)} dy.$$

Proof. From Lemma 2.3.2 with $f(x) = x$ and $f^{ij}(x) = x_i x_j$. □

Boundary conditions

Theorem A.2.8 (Condition C). *For $\delta > 0$, there exists $M_\delta < \infty$ and n_0 so that for $n > n_0$, $|b_n(t,x)| > M_\delta$ implies $\phi(x) < \delta$.*

Proof. On the compact set $\{\phi(x) \geq \delta\}$, $b_n(t,x)$ converges uniformly by Theorem A.2.7 and Theorem 2.3.3 to $\frac{1}{3}\frac{\nabla p(x)}{p(x)}\bar{\varepsilon}(x)^2$, hence is uniformly bounded on this set. □

Theorem A.2.9 (Condition D). *The diffusion term a_n is uniformly bounded by some $M < \infty$ so that*

$$\sup_{s,x,n} ||a_n(s,x)|| \leq M.$$

Proof. By definition the diffusion term

$$a_n^{ij}(s,x) = \frac{1}{h_n} \sum_{y \in \mathsf{NB}_n(x)} \frac{1}{|\mathsf{NB}_n(x)|}(y_i - x_i)(y_j - x_j)$$

is an average of numbers bounded by $\frac{\varepsilon_n(x)^2}{h_n}$. This quantity converges to the bounded function $\overline{\varepsilon}(x)$ as $n \to \infty$, yielding the result. □

Theorem A.2.10 (Condition F3). *Given $(t,y) \in J_0$ and $M < \infty$, there exist $\delta_0 > 0$ and $n_0 < \infty$ so that if $|t-s| < \delta_0$, $|x-y| < \delta_0$, $n > n_0$, and $\langle \nabla\phi(x), a_n(s,x)\nabla\phi(x)\rangle < \delta_0$, then $|b_n(s,x)| \geq M$.*

Proof. For any $\delta_1 > 0$, by Lemma 2.3.2, we may choose n_0 large enough so that for all $n > n_0$ and $x \in \mathcal{X}_n$, we have

$$||a_n(s,x) - a(s,x)|| < \delta_1,$$

which implies that

$$\langle \nabla\phi(x), a_n(s,x)\nabla\phi(x)\rangle \geq \Big(\frac{1}{3}\overline{\varepsilon}(x)^2 - \delta_1^2)\Big)|\nabla\phi(x)|^2.$$

Because $\overline{\varepsilon}(y) > 0$, we can choose $\delta_0 > 0$ so that $\overline{\varepsilon}(x)^2$ is uniformly bounded away from 0 on $|x-y| < \delta_0$, hence choosing δ_1 small makes the condition vacuous. □

Theorem A.2.11 (Condition F4). *Given $(t,y) \in J_0$ and $\varepsilon > 0$, there exist $\delta_0 > 0$, $n_0 < \infty$, and $M < \infty$ so that if $|t-s| < \delta_0$, $|x-y| < \delta_0$, $n > n_0$, and $|b_n(s,x)| > M$, then*

$$\left| \frac{b_n(s,x)}{\langle b_n(s,x), \nabla\phi(x)\rangle} - \gamma(t,y) \right| < \varepsilon.$$

Proof. For any $\varepsilon > 0$, fix $M > 0$ to be chosen later. Choose δ_0 small enough so that if $|x-y| < 2\delta_0$, we have

$$\left| p(x) - p(y) - \frac{(y-x)\cdot\nabla p(y)}{p(y)} \right| < C_1$$

for some uniform C_1. By Lemma 2.3.2 and the fact that $|\nabla\phi(x)| \geq 1$ on ∂D and is continuous, we may choose n_0 large enough so that for all $n > n_0$ and $|x-y| < \delta_0$, we have

- $\varepsilon_n(x) < \delta_0$;

- $|\varepsilon_n(x)^2 h_n^{-1} - \overline{\varepsilon}(x)^2| < C_2$ for a uniform $C_2 > 0$;

- $\left|b_n(s,x) - E[b_n(s,x)]\right| < M/2$ for $x \in \mathcal{X}_n$;

- $\left|\frac{b_n(s,x)}{\langle b_n(s,x), \nabla\phi(x)\rangle} - \frac{E[b_n(s,x)]}{\langle E[b_n(s,x)], \nabla\phi(x)\rangle}\right| < \varepsilon/2$ for $x \in \mathcal{X}_n$.

If $|b_n(s,x)| > M$ for $n > n_0$, then

$$\left|E[b_n(s,x)]\right| > M/2.$$

Now, orient the coordinate axes so that the first coordinate axis lies on the normal vector from x to ∂D, and let τ be the distance from x to ∂D. In this case, we compute

$$\begin{aligned} E[b_n^1(s,x)] &= h_n^{-1} \int_{z \in B(x,\varepsilon_n(x)) \cap D} (z_1 - x_1) \frac{p(z)}{p_{\varepsilon_n(x)}(x)} dy \\ &= \frac{\varepsilon_n(x) - \min\{\tau, \varepsilon_n(x)\}}{h_n} + \frac{1}{6}\frac{\partial_1 p(x)}{p(x)} \frac{\varepsilon_n(x)^2 + \tau^2}{h_n} + C_3 \end{aligned}$$

and for $i > 1$ that

$$E[b_n^i(s,x)] = \frac{1}{6}\frac{\partial_i p(x)}{p(x)} \frac{\varepsilon_n(x)^2}{h_n} + C_4 \tag{A.2}$$

for error terms C_3 and C_4 independent of n. Choosing M large enough, we find

$$\tau < (1 - C_5(M))\varepsilon_n(x)$$

for a constant $C_5(M) > 0$ independent of n, which implies that

$$E[b_n^1(s,x)] \geq C_5(M)\frac{\varepsilon_n(x)}{h_n} + \frac{1}{6}\frac{\partial_1 p(x)}{p(x)} \frac{\varepsilon_n(x)^2 + \tau^2}{h_n} + C_3. \tag{A.3}$$

Now, notice that $\gamma(s,y)$ is a vector purely in the normal direction to ∂D at y normalized so that $\langle \gamma(s,y), \nabla\phi(y)\rangle = 1$. Because the constants $C_3, C_4, C_5(M)$ in (A.3) and (A.2) are independent of n, all terms in these equations aside from $C_5(M)\frac{\varepsilon_n(x)}{h_n}$ scale to constants as we take n_0 and M large, so $\frac{E[b_n(s,x)]}{\langle E[b_n(s,x)], \nabla\phi(x)\rangle}$ becomes arbitrarily close to a vector purely in the normal direction to ∂D from x. Choosing δ_0 small enough makes these vectors coincide up to error $\varepsilon/2$, which gives the result when combined with the bound

$$\left|\frac{b_n(s,x)}{\langle b_n(s,x), \nabla\phi(x)\rangle} - \frac{E[b_n(s,x)]}{\langle E[b_n(s,x)], \nabla\phi(x)\rangle}\right| < \varepsilon/2$$

we obtained by taking n_0 large. □

Theorem A.2.12 (Condition B). *There exist M, c, and n_0 so that for all $n > n_0$, $|b_n(t,x)| > M$ implies*

$$\frac{\langle \nabla\phi(x), b_n(t,x)\rangle}{|b_n(t,x)|} \geq c.$$

Proof. By definition, $\gamma(t,x)$ is uniformly bounded above by some C_0. Now, by compactness of $\partial D = J_0$, there exists some $\delta > 0$ so that each $x \in \{\phi(y) < \delta\}$ has a corresponding $x' \in \delta D$ so that the conclusion of Theorem A.2.11 applies with $\varepsilon = C_0/2$. Taking $M = M_\delta$ and n_0 from Theorem A.2.8 for this δ and applying Theorem A.2.11 yields that

$$\frac{\langle \nabla\phi(x), b_n(t,x)\rangle}{|b_n(t,x)|} \geq \frac{2}{C_0}.$$ □

A.2.5 Completing the proof of Theorem A.2.1

By Corollary A.2.5, to complete the proof of Theorem A.2.1, it suffices for us to compute the limiting terms a and b and to verify the conditions of Theorem A.2.4 for uniqueness of the submartingale problem. We begin by computing the limiting a and b, for which we will need the following lemma.

Lemma A.2.13. *For $d \geq 2$, let $B_d(r)$ be the d-dimensional ball of radius r and $V_d(r) = V_d r^d$ be its volume. As $r \to 0$, we have*

$$\int_{B_d(r)} x_i^n dx = \begin{cases} 0 & n \text{ odd} \\ \frac{2V_{d-1}}{n+1} r^{n+d} + o(r^{n+d}) & n \text{ even} \end{cases}$$

and

$$\int_{B_d(r)} x_i^n x_j^m dx = 0 \text{ if } n \text{ odd.}$$

Proof. If n is odd, both claims follow because the integrands are odd functions integrated over symmetric domains. If n is even, for the first claim we compute

$$\int_{B_d(r)} x_i^n dx = \int_{-r}^{r} V_{d-1}(\sqrt{r^2 - x^2}) x^n dx = \frac{2V_{d-1}}{n+1} r^{n+d} + o(r^{n+d}).$$ □

Theorem A.2.14 (Drift diffusion coefficients). *The limiting integrals for drift and diffusion are*

$$
\begin{aligned}
a_n^{ii}(s,x) &= \frac{1}{h_n}\left(\frac{1}{3}\varepsilon_n(x)^2 + o(\varepsilon_n(x)^2)\right) \to \frac{1}{3}\overline{\varepsilon}(x)^2 \\
a_n^{ij}(s,x) &= \frac{1}{h_n}\frac{o(\varepsilon_n(x)^{d+2})}{2V_{d-1}p(x)\varepsilon_n(x)^d + o(\varepsilon_n(x)^d)} \to 0 \\
b_n^{i}(s,x) &= \frac{1}{h_n}\Big(\frac{1}{3}\frac{\partial_i p(x)}{p(x)}\varepsilon_n(x)^2 + o(\varepsilon_n(x)^2)\Big) \to \frac{\partial_i p(x)}{3p(x)}\overline{\varepsilon}(x)^2 \\
\Delta_{n,1}(x,s) &= \frac{1}{h_n}\Big(\frac{\varepsilon_n(x)^{d+4}p(x)V_{d-1} + o(\varepsilon_n(x)^{d+4})}{2V_{d-1}p(x)\varepsilon_n(x)^d + o(\varepsilon_n(x)^d)}\Big) \to 0.
\end{aligned}
$$

Proof. Because p is differentiable on D, for any $x \in D$ we have the Taylor expansion

$$p(x+y) = p(x) + y \cdot \nabla p(x) + o(|y|^2)$$

of p at x, where the convergence is uniform on compact sets. For n large enough so that the ball of radius $\varepsilon_n(x)$ centered at x lies completely inside D, we can substitute this expansion into the definitions of a_n and b_n. Using Lemma A.2.13 to estimate the resulting expressions yields

$$
\begin{aligned}
a_n^{ii}(s,x) &= \frac{1}{h_n}\frac{\int_{|y|<\varepsilon_n(x)} y_i^2 p(x) + y_i^2 y \cdot \nabla p(x) + y_i^2 o(|y|^2)dy}{\int_{|y|<\varepsilon_n(x)} p(x) + y \cdot \nabla p(x) + o(|y|^2)dy} \\
&= \frac{1}{h_n}\frac{\frac{2}{3}V_{d-1}p(x)\varepsilon_n(x)^{d+2} + o(\varepsilon_n(x)^{d+2})}{2V_{d-1}p(x)\varepsilon_n(x)^d + o(\varepsilon_n(x)^d)} \\
&= \frac{1}{h_n}\Big(\frac{1}{3}\varepsilon_n(x)^2 + o(\varepsilon_n(x)^2)\Big)
\end{aligned}
$$

,

$$
\begin{aligned}
a_n^{ij}(s,x) &= \frac{1}{h_n}\frac{\int_{|y|<\varepsilon_n(x)} y_i y_j p(x) + y_i y_j y \cdot \nabla p(x) + y_i y_j o(|y|^2)dy}{\int_{|y|<\varepsilon_n(x)} p(x) + y \cdot \nabla p(x) + o(|y|^2)dy} \\
&= \frac{1}{h_n}\frac{o(\varepsilon_n(x)^{d+2})}{2V_{d-1}p(x)\varepsilon_n(x)^d + o(\varepsilon_n(x)^d)}
\end{aligned}
$$

, and

$$\begin{aligned}
b_n^i(s,x) &= \frac{1}{h_n}\frac{\int_{|y|<\varepsilon_n(x)} y_i p(x) + y_i y\cdot\nabla p(x) + y_i o(|y|^2)dy}{\int_{|y|<\varepsilon_n(x)} p(x) + y\cdot\nabla p(x) + o(|y|^2)dy}\\
&= \frac{1}{h_n}\frac{\frac{2}{3}V_{d-1}\frac{\partial_i p(x)}{p(x)}\varepsilon_n(x)^{d+2} + o(\varepsilon_n(x)^{d+2})}{2V_{d-1}p(x)\varepsilon_n(x)^d + o(\varepsilon_n(x)^d)}\\
&= \frac{1}{h_n}\Big(\frac{1}{3}\frac{\partial_i p(x)}{p(x)}\varepsilon_n(x)^2 + o(\varepsilon_n(x)^2)\Big).
\end{aligned}$$

Defining $S_d(r)$ to be the surface area of a radius r ball in d dimensions, we find

$$\begin{aligned}
\Delta_{n,1}(s,x) &= \frac{1}{h_n}\frac{\int_{|y|<\varepsilon_n(x)} |y|^3 p(x) + |y|^3 p(x) + |y|^3 o(|y|^3)dy}{\int_{|y|<\varepsilon_n(x)} p(x) + y\cdot\nabla p(x) + o(|y|^2)dy}\\
&= \frac{1}{h_n}\frac{\int_0^{\varepsilon_n(x)} r^3 S_d(r)p(x) + o(\varepsilon_n(x)^{d+4})}{2V_{d-1}p(x)\varepsilon_n(x)^d + o(\varepsilon_n(x)^d)}\\
&= \frac{1}{h_n}\frac{o(\varepsilon_n(x)^{d+4})}{2V_{d-1}p(x)\varepsilon_n(x)^d + o(\varepsilon_n(x)^d)}.
\end{aligned}$$

The result follows by taking the $n \to \infty$ limit in each estimate and recalling that h_n was chosen so that $h_n^{-1}\varepsilon_n(x)^2 \to \bar{\varepsilon}_n(x)^2$ and $h_n^{-1}\varepsilon_n(x)^{2+\alpha} \to 0$. The final convergence is uniform on compact sets because the convergence of the initial Taylor expansion was, each integration estimate preserves uniformity, and the limit $h_n^{-1}\varepsilon_n(x)^2 \to \bar{\varepsilon}(x)^2$ is uniform over all of D. □

Proof of Theorem A.2.1. To prove Theorem A.2.1, it remains only to check the conditions of Theorem A.2.4. Condition (1) follows because $a(x) = \frac{1}{3}\bar{\varepsilon}(x)^2 \cdot I$ is a continuous multiple of the identity. Condition (2) follows because $b(x) = \frac{1}{3}\frac{\nabla p(x)}{p(x)}\bar{\varepsilon}(x)^2$ is evidently bounded and measurable. For Condition (3), γ is evidently bounded, locally Lipschitz because it is a normalized vector normal to the smooth ∂D, and $\langle\gamma(x), \nabla\phi(x)\rangle = 1$ by definition. Finally, Condition (4) is evident because $\rho \equiv 0$. □

A.3 Generalizing to metric graphs

In this section, we give details on how to generalize our results for $\varepsilon_n(x)$-ball graphs to isotropic metric graphs. The approach is exactly parallel; we verify the conditions of the Stroock-Varadhan criterion and consider the limiting rescaled stationary distribution. We give in this section the necessary estimates of the minimal degree and the drift and diffusion terms. We first present a technical lemma.

Lemma A.3.1. *For $d \geq 2$, Let $S_d(r)$ be the d-dimensional shell of radius r and $V_d(r) = C_d r^d$ be its volume. As $r \to 0$, we have*

$$\int_{S_d(r)} x_i^n dx = \begin{cases} 0 & n \text{ odd} \\ \frac{2C_{d-1}}{n+1}(n+d)r^{n+d-1} + o(r^{n+d-1}) & n \text{ even} \end{cases}$$

and

$$\int_{S_d(r)} x_i^n x_j^m dx = 0 \text{ if } n \text{ odd.}$$

Proof. This follows by differentiating Lemma A.2.13. □

Let us now consider an isotropic metric graph model with kernel function $h(r)$. In particular, this implies that there is an edge from x_i to x_j with probability $h(|x_i - x_j|\varepsilon_n(x_i)^{-1})$ and that

$$\int_0^1 h(r)r^{d-1}dr > 0.$$

We characterize the minimal out-degree in this setting.

Theorem A.3.2 (Minimal out-degree). *For an isotropic metric graph with kernel $h(r)$ satisfying ($\star$), we have the almost sure convergence*

$$\frac{|\mathsf{NB}_n(x)|}{|\mathcal{X}_n \cap B(x, \varepsilon_n(x))|} \to C(h)p(x)$$

for a constant $C(h)$ independent of x and n, which implies that the minimal degree $|\mathsf{NB}_n(x)| = \omega(n^{2/(d+2)} \log(n)^{d/(d+2)})$.

Proof. The out-degree of a vertex is the independent sum of binary variables, each with probability $h(|x_i - x_j|\varepsilon_n(x_i)^{-1})$, so Kolmolgorov's strong law yields

$$\varepsilon_n(x)^{-d} \frac{|\mathsf{NB}_n(x)|}{|\mathcal{X}_n \cap B(x, \varepsilon_n(x))|} \to E\left[\varepsilon_n(x)^{-d} \frac{|\mathsf{NB}_n(x)|}{|\mathcal{X}_n \cap B(x, \varepsilon_n(x))|}\right].$$

Let $y(r, \theta)$ be the radial representation of y and let $C = \frac{2C_{d+1}(n+d)}{n+1}$ be the constants in Lemma A.3.1. The desired expected value is the integral

$$\begin{aligned}
E\left[\frac{\varepsilon_n(x)^{-d}|\mathsf{NB}_n(x)|}{|\mathcal{X}_n \cap B(x,\varepsilon_n(x))|}\right] &= \int_{y\in B(x,\varepsilon_n(x))} p(x+y)h(|y|\varepsilon_n^{-1}(x))dy \\
&\sim \varepsilon_n(x)^{-d}\int_{y\in B(x,\varepsilon_n(x))} (p(x)+\nabla p(x)\cdot y)h(|y|\varepsilon_n^{-1}(x))dy \\
&= \varepsilon_n(x)^{-d}\int_0^{\varepsilon_n(x)}\int_{\theta\in S_d(r)} (p(x)+\nabla p(x)\cdot y(r,\theta))h(r)dyd\theta \\
&= Cp(x)\varepsilon_n(x)^{-d}\int_0^{\varepsilon_n(x)} h(r)r^{d-1}dr \\
&\quad + \varepsilon_n(x)^{-d}\int_0^{\varepsilon_n(x)} h(r)r^{d-1}\int_{\theta\in S_d(1)} \nabla p(x)\cdot y(1,\theta)drd\theta.
\end{aligned}$$

The latter term is zero by Lemma A.3.1 since it is the integral of the odd function $y(1,\theta)$ over a symmetric domain. Now take the substitution $s = r/\varepsilon_n(x)$ to obtain

$$E\left[\frac{\varepsilon_n(x)^{-d}|\mathsf{NB}_n(x)|}{|\mathcal{X}_n \cap B(x,\varepsilon_n(x))|}\right] = Cp(x)\int_0^1 h(s)s^{d-1}ds.$$

The Kolmogorov strong law provides concentration around this value. Noting that $\varepsilon_n(x)^d = \omega(n^{2/(d+2)}\log(n)^{d/(d+2)})$ gives the asymptotic claim. □

Since Theorem A.3.2 guarantees that asymptotically we achieve the necessary minimal number of points, and $h(x)$ is zero for $x > 1$, Lemma 2.3.2 applies to show the moment conditions in the Stroock-Varadhan criterion. For the boundary conditions, note that C, D, and F3 only require convergence of coefficients in Lemma A.3.3 to those in Theorem A.2.14. Conditions F4 and B rely on two facts, the uniform convergence of coefficients given by Lemma A.3.3, and the asymmetry induced by the boundary (A.3), the proof of which is parallel to the one given for ε-ball graphs. Therefore, to complete the proof the generalization, it remains only to compute the limiting drift and diffusion coefficients.

Lemma A.3.3 (Polynomial integrals with respect to kernel). *Under the same conditions as Theorem A.3.2, for any positive integer α we have*

$$\int_{y\in B(x,\varepsilon_n(x))} y_i^\alpha p(x+y)h(|y|\varepsilon_n^{-1}(x))dy \sim V(h,\alpha)\int_{y\in B(x,\varepsilon_n(x))} y_i^\alpha p(x+y)dy$$

as $n \to \infty$ for a constant $V(h,\alpha)$ independent of n with $V(h,1) = V(h,2)$.

Proof. Perform the same Taylor approximation and radial decomposition as in Theorem A.3.2 to obtain

$$\int_{y\in B(x,\varepsilon_n(x))} y_i^\alpha p(x+y)h(|y|\varepsilon_n^{-1}(x))dy$$
$$\sim \int_0^{\varepsilon_n(x)} \int_{\theta\in S_d(r)} y_i(r,\theta)^\alpha(p(x)+\nabla p(x)\cdot y(r,\theta))h(r\varepsilon_n^{-1}(x))drd\theta.$$

For α an odd integer, by Lemma A.3.1 we have

$$\begin{aligned}
&\int_{y\in B(x,\varepsilon_n(x))} y_i^\alpha p(x+y)h(|y|\varepsilon_n^{-1}(x))dy \\
&\sim \int_0^{\varepsilon_n(x)} h(r\varepsilon_n^{-1}(x))r^{\alpha+d}\int_{\theta\in S_d(1)} y_i(1,\theta)^\alpha\nabla p(x)\cdot y(1,\theta)drd\theta \\
&= \partial p_i(x)\int_0^1 h(r)r^{\alpha+d}dr\varepsilon_n(x)^{\alpha+d}\int_{\theta\in S_d(1)} y_i(1,\theta)^{\alpha+1}d\theta \\
&\sim V(h,\alpha)\int_{y\in B(x,\varepsilon_n(x))} y_i^\alpha p(x+y)dy
\end{aligned}$$

for

$$V(h,\alpha)=(\alpha+d+1)\int_0^1 h(r)r^{\alpha+d}dr.$$

If α is an even integer, we have

$$\begin{aligned}
&\int_{y\in B(x,\varepsilon_n(x))} y_i^\alpha p(x+y)h(|y|\varepsilon_n^{-1}(x))dy \\
&\sim \int_0^{\varepsilon_n(x)} h(r\varepsilon_n^{-1}(x))r^{\alpha+d-1}\int_{\theta\in S_d(1)} y_i(1,\theta)^\alpha p(x)drd\theta \\
&= p(x)\int_0^{\varepsilon_n(x)} h(r\varepsilon_n^{-1}(x))r^{\alpha+d-1}dr\int_{\theta\in S_d(1)} y_i(1,\theta)^\alpha d\theta \\
&= p(x)\varepsilon_n(x)^{\alpha+d}\int_0^1 h(r)r^{\alpha+d-1}dr\int_{\theta\in S_d(1)} y_i(1,\theta)^\alpha d\theta \\
&\sim V(h,\alpha)\int_{y\in B(x,\varepsilon_n(x))} y_i^\alpha p(x+y)dy
\end{aligned}$$

for

$$V(h,\alpha) = (\alpha + d) \int_0^1 h(r) r^{\alpha+d-1} dr. \qquad \square$$

The limits of drift and diffusion terms in Theorem A.2.14 depend only on ratios of these integrals for $\alpha = 1, 2$, so applying Lemma A.3.3 shows that the limits for isotropic metric graphs are identical to the ones for ε-ball graphs. The remainder of the analysis proceeds unchanged.

A.4 Recovery of distances via ball-radii

We will prove that given the ball radii $\varepsilon_n(x_i)$, we can recover point-to-point distances if x_i are located in a convex domain. Otherwise, we recover the geodesic distances. Our goal is to show that for any points x_i and x_j, the weighted shortest path distance d_{ij} between the points on the graph $\overline{G}_n$ where outgoing edges are weighted by $\varepsilon_n(x_i)$ converges to the distance $|x_i - x_j|$.

A.4.1 Outline of proof approach

We proceed in two steps. First, we consider the case when $\varepsilon_n(x_i)$ is known exactly. In this case, the weighted shortest path is an upper bound on the true distance. We bound its weighted distance d_{ij} by constructing a path whose weighted distance is close to the geodesic distance.

To control the upper bound, we show that there exists a δ that converges to zero faster than $\min_{x_i} \varepsilon_n(x_i)$ while still guaranteeing that every ball of size δ in the domain contains at least one point. Once we find such a δ, the upper bound will follow. Indeed, if we are at some x, we can always find a point that whose distance from our target x_j is smaller by at least $\varepsilon_n(x) - \delta$. This gives an upper bound on the number of steps in our path and therefore the total error.

Second, we assume that we are given noisy estimates of $\overline{\varepsilon}(x)$ from our algorithm via the stationary distribution. We use uniform convergence of $\overline{\varepsilon}(x)$ to control the overall pathwise error.

We give a detailed analysis of each step in separate subsections below.

A.4.2 The case of exact knowledge of ε_n.

We begin with two lemmas allowing us to construct for each pair of points i, j a point k along which to start a path from i to j.

Lemma A.4.1. *Let $\delta_n = \Omega(n^{-\frac{1}{d+1}})$. For any set of n^2 balls with radius δ_n, all n^2 balls will have at least one point of $\mathcal{X}_n$ with high probability.*

Proof. The number of points $N(x)$ in a ball of radius δ_n follows a binomial distribution with n draws and success probability

$$p_{\delta_n}(x) = \int_{|y-x|<\delta(n)} p(y)dy \sim V_d p(x)\delta_n^d.$$

Therefore, the probability that $N(x) = 0$ is

$$P(N(x) = 0) = (1 - p_{\delta_n}(x))^n = \left((1 - p_{\delta_n}(x))^{p_{\delta_n}(x)^{-1}}\right)^{np_{\delta_n}(x)} \to e^{-np_{\delta_n}(x)}$$

if $n\delta_n^d \to \infty$. Recalling that $\delta_n = \Omega(n^{-\frac{1}{d+1}})$, this implies that

$$np_{\delta_n}(x) \sim n^{\frac{1}{d+1}}$$

and in particular that $P(N(x) = 0) = o(n^{-2})$, so taking the union bound over all n^2 balls yields the result. □

Lemma A.4.2. *Let $\delta_n = \Omega(n^{-\frac{1}{d+1}})$. For all i, j, there exists $x_k \in B(x_i, \varepsilon_n(x_i))$ such that*

$$\left|\left(|x_i - x_j| - |x_k - x_j|\right) - |x_i - x_k|\right| \leq 2\delta_n \text{ and } \left||x_k - x_i| - \varepsilon_n(x_i)\right| \leq 2\delta_n.$$

Proof. Let $v = \frac{x_j - x_i}{|x_j - x_i|}$ and consider the n^2 balls

$$B_{ij} = B(x_i + v(\varepsilon_n(x_i) - \delta_n), \delta_n).$$

By Lemma A.4.1, there must exist with high probability at least one point of $\mathcal{X}_n$ in each B_{ij}. Any such $x_k \in B_{ij}$ verifies the desired conditions. □

Theorem A.4.3. *Let $x_i, x_j \in \mathcal{X}_n$ and d_{ij} be the weighted shortest path distance over the weighted graph $\overline{G}_n$ constructed from G_n by assigning weight $\varepsilon_n(x_i)$ to all outgoing edges from x_i. For any $\varepsilon > 0$, there exists an n such that*

$$\left||x_i - x_j| - d_{ij}\right| < \varepsilon.$$

Proof. Take $\delta_n = \Theta(n^{-\frac{1}{d+1}})$. We show that with high probability, there exists a path with M steps whose weighted path distance d satisfies

$$|x_i - x_j| \leq d \leq |x_i - x_j| + 2M\delta_n + \max_{x \in \mathcal{X}_n} \varepsilon_n(x)$$

and so that $\lim_{n\to\infty} M\delta_n = 0$. The result then follows because $d_{ij} \leq d$.

To construct such a path from x_i to x_j, we apply the following procedure. Start at the point x_i. If the current point is x_k and $x_j \in B(x_k, \varepsilon_n(x_k))$, move to it and terminate. Otherwise, pick a point $x_l \in B_{kj}$ and repeat until x_j is reached.

The lower bound holds because each edge weight is at least its length. For the upper bound, by Lemma A.4.2, moving to x_l reduces the geodesic distance to x_j by at least $|x_k - x_l| - 2\delta_n$ and moves a weighted distance of $\varepsilon_n(x_k) < |x_k - x_l| + 2\delta_n$. Thus, if our path has M steps, the difference between our weighted distance and the geodesic distance is at most $4M\delta_n + \max_x \varepsilon_n(x)$, where we add the weighted distance of the last step. This gives the upper bound.

It remains now to bound M. For this, notice that the geodesic distance to x_j decreases by at least $\min_{x\in\mathcal{X}_n} \varepsilon_n(x) - 2\delta_n$ at each step, leading to the bound

$$M \leq \frac{|x_i - x_j|}{\min_{x\in\mathcal{X}_n} \varepsilon_n(x) - 2\delta_n}.$$

Recall now that $\delta_n = \Theta(n^{-\frac{1}{d+1}})$ so that $\varepsilon_n(x) = \omega(\delta_n)$ and hence

$$M\delta_n = \frac{|x_i - x_j|}{\min_{x\in\mathcal{X}_n} \frac{\varepsilon_n(x)}{\delta_n} - 1} \to 0. \qquad \square$$

A.4.3 The case of stochastic estimates of ε_n

We now consider the case where we are given only an estimate $\widehat{\varepsilon}_n(x)$ of ε_n, obtained by first estimating $\overline{\varepsilon}(x)$ via the stationary distribution and then applying a normalization to obtain $\widehat{\varepsilon}_n(x)$ on $\mathcal{X}_n$. We first control the error in $\widehat{\varepsilon}_n(x)$ along a single path.

Lemma A.4.4. *For $k_1 = i$ and $k_{l_n} = j$, let $x_{k_1}, \ldots, x_{k_{l_n}}$ be a path between i and j in $\overline{G}_n$. If $l_n = O(g_n^{-1})$, we have*

$$\sum_{i=1}^{l_n} |\widehat{\varepsilon}_n(x_{k_i}) - \varepsilon_n(x_{k_i})| \to 0$$

in probability.

Proof. By uniform convergence of the stationary distribution and continuity of the out degree estimate $p(x)\varepsilon_n(x)^d V_d = k/n$, for all γ and δ, we have

$$P\left(\sup_{x\in\mathcal{X}_n}\left|\frac{\widehat{\varepsilon}_n(x)}{g_n} - \overline{\varepsilon}(x)\right| > \gamma\right) < \delta$$

for large enough n. This implies that

$$P\left(\sup_{x\in\mathcal{X}_n}|\widehat{\varepsilon}_n(x) - \varepsilon_n(x)| > \gamma g_n\right) < \delta.$$

Now notice that

$$P\left(\sum_{i=1}^{l_n}|\widehat{\varepsilon}_n(x_{k_i}) - \varepsilon_n(x_{k_i})| > \gamma\right) < P\left(l_n \sup_x |\widehat{\varepsilon}_n(x_{k_i}) - \varepsilon_n(x_{k_i})| > \gamma\right).$$

By assumption, the number of steps in the path is $l_n = O(g_n^{-1})$. Therefore, there exists a constant $M > 0$ such that

$$P\left(\sum_{i=1}^{l_n}|\widehat{\varepsilon}_n(x_{k_i}) - \varepsilon_n(x_{k_i})| > \gamma\right) < P\left(\sup_x |\widehat{\varepsilon}_n(x) - \varepsilon_n(x)| > M\gamma g_n\right) < \delta,$$

from which the claim follows by choosing n large enough. □

We now show that the shortest weighted distance path recovers the geodesic distance with stochastic estimates $\widehat{\varepsilon}_n(x)$ instead of the true values. Our approach is the same as in the deterministic case; we will construct a weighted path and show that its weighted distance converges to the geodesic distance and is close to the weighted distance of the shortest weighted path. Let $\widehat{d}_{ij}$ denote the weighted distance of the shortest weighted distance path from x_i to x_j.

Theorem A.4.5. *For any $\varepsilon > 0$, there exists n such that*

$$\left||x_i - x_j| - \widehat{d}_{ij}\right| < \varepsilon$$

with high probability.

Proof. Let $\delta_n = \Theta(n^{-\frac{1}{d+2}})$. For any $\gamma > 0$, we show that for large enough n, with high probability there exists a path from x_i to x_j with M steps whose weighted path distance $\widehat{d}$ satisfies

$$\widehat{d} \leq |x_i - x_j| + 4M\delta_n + \gamma + \max_{x \in \mathcal{X}_n} \widehat{\varepsilon}_n(x). \tag{A.4}$$

Construct the path as in Theorem A.4.3 with $\varepsilon_n(x)$ replaced by $\widehat{\varepsilon}_n(x)$.

We now analyze its weighted distance. Arguing as in Lemma A.4.2, a step from x_k to x_l which is not the last step in this path reduces the geodesic distance to x_j by between $|x_k - x_l| - 2\delta_n$ and $|x_k - x_l|$. On the other hand, this step has a weighted distance of $\widehat{\varepsilon}_n(x_k)$, which satisfies

$$|x_k - x_l| - 2\delta_n - |\widehat{\varepsilon}_n(x_k) - \varepsilon_n(x_k)| \leq \widehat{\varepsilon}_n(x_k) \leq |x_k - x_l| + |\widehat{\varepsilon}_n(x_k) - \varepsilon_n(x_k)|.$$

Therefore, the geodesic distance traveled and weighted distance $\widehat{d}$ along our constructed path differ by at most

$$\sum_{i=1}^{M-1} |\widehat{\varepsilon}_n(x_{k_i}) - \varepsilon_n(x_{k_{i+1}})| + 4M\delta_n$$

in the first $M - 1$ steps. By arguing as in the proof of Theorem A.4.3 with $\varepsilon_n(x)$ replaced by $\widehat{\varepsilon}_n(x)$ and noting that $\widehat{\varepsilon}_n(x)$ converges uniformly to $\varepsilon_n(x)$, the number of steps in the constructed path satisfies

$$\frac{|x_i - x_j|}{\max_x \varepsilon_n(x)} \leq M \leq \frac{|x_i - x_j|}{\min_x \varepsilon_n(x) - 2\delta_n}. \tag{A.5}$$

In particular, we note that $M = O(g_n^{-1})$. Applying Lemma A.4.4 to choose n large enough so that

$$\sum_{i=1}^{M-1} |\widehat{\varepsilon}_n(x_{k_i}) - \varepsilon_n(x_{k_{i+1}})| < \gamma$$

and adding $\max_{x \in \mathcal{X}_n} \widehat{\varepsilon}_n(x)$ for the last step yields (A.4). Noting by (A.5) that $M\delta_n \to 0$, taking large enough n in (A.4) shows that $\widehat{d}_{ij} \leq \widehat{d} \leq |x_i - x_j|$.

We now show that $\widehat{d}_{ij} \geq |x_i - x_j|$. It suffices to show that the length L of the shortest weighted distance path must be $L = O(g_n^{-1})$, as Lemma A.4.4 would then imply that its weighted distance with respect to $\widehat{\varepsilon}_n(x)$ converges to its weighted distance with respect to $\varepsilon_n(x)$, which is bounded below by $|x_i - x_j|$.

To bound L, note that the minimum weighted distance at each step is $\min_{x \in \mathcal{X}_n} \widehat{\varepsilon}_n(x)$, while the total weighed distance is at most $\widehat{d}$. Therefore, by (A.4), we obtain that for any $\gamma > 0$ we have

$$L \min_{x \in \mathcal{X}_n} \widehat{\varepsilon}_n(x) \leq |x_i - x_j| + 4M\delta_n + \gamma + \max_{x \in \mathcal{X}_n} \widehat{\varepsilon}_n(x)$$

for large enough n. By uniform convergence of $\widehat{\varepsilon}_n(x)$ to $\varepsilon_n(x)$, this shows that for any $\gamma > 0$ we have

$$L \leq \frac{|x_i - x_j| + 4M\delta_n + \gamma + \max_{x \in \mathcal{X}_n} \widehat{\varepsilon}_n(x)}{\min_{x \in \mathcal{X}_n} \varepsilon_n(x)} = O(g_n^{-1})$$

for large enough n, yielding the desired. □

Appendix B

Metric estimation

B.1 Uniform equicontinuity of the marginals

We discuss condition $(\star)$ stated in the main text. We conjecture and assume that the following technical condition holds on all metric graphs. Note that as shown in Corollary B.2.6 this is a strictly stronger condition than Conjecture A.1.1 in Appendix A, which is necessary due to the fact that we are looking at individual trajectories of the stochastic process.

$(\star)$ For $t = \Theta(g_n^{-2})$, the rescaled marginal density $nq_t(x, x_i)$ is a.s. eventually uniformly equicontinuous.

To make the terminology we use in $(\star)$ clear, we rephrase it as follows.

Definition 7 (Condition $\star$). If $t = \Theta(g_n^{-2})$, with probability 1, for any $\delta > 0$, there exist $\gamma > 0$ and n_0 so that for $n > n_0$, any $x_k \in \mathcal{X}_n$, and any $x_i, x_j \in \mathcal{X}_n$ with $|x_i - x_j| < \gamma$, we have

$$|nq_t(x_i, x_k) - nq_t(x_j, x_k)| < \delta.$$

This statement allows us to convert the weak convergence in distribution ensured by Theorem C.1.1 and the results of Stroock and Varadhan [1971b] to the convergence in density required by Corollary B.2.6. Such a statement rules out the possibility that the density function $q_t(x, x_i)$ oscillates at frequencies increasing with n as $n \to \infty$. Controlling regularity of $q_t(x, x_i)$ seems to be a critical ingredient for approaching $(\star)$.

In the case of an undirected graph, $(\star)$ follows from results in the literature. The strong local limit law for simple random walks shown in Croydon and Hambly [2008a] yields an even stronger result than $(\star)$. In addition, in the same setting, the convergence result for spectral clustering of [von Luxburg et al., 2008] yields an equicontinuity result for eigenvectors of the Laplacian which implies $(\star)$.

However, for directed graphs no such result exists, and non-reversibility of the Markov chain seems to be an obstacle to proving such a result. Some results on utilizing the directed Laplacian as a smoothing operator Zhou et al. [2005] exist in this direction. We believe these techniques could lead to an approach to $(\star)$ but thus far they have not yielded a sufficiently strong equicontinuity result.

One consequence of equicontinuity is that convergence in distribution implies convergence in density. We prove this for the marginal distribution $\widehat{q_t}(x_k, x_i)$ of $Y_{\widehat{t}}$ for the purpose of Theorem 3.2.5, following the original strategy in Appendix A.

Lemma B.1.1 (Convergence of marginal densities). *If* $t_n g_n^2 = \widehat{t} = \Theta(1)$, *then under* $(\star)$ *we have*

$$\lim_{n\to\infty} n q_{t_n}(x, x_i) = \frac{\widehat{q_t}(x, x_i)}{p(x)},$$

where the convergence is uniform in x *and* x_i.

Proof. The a.s. weak convergence of processes of Theorem C.1.1 (which is uniform in x_i) implies by Ethier and Kurtz [1986, Theorem 4.9.12] that the empirical marginal distribution

$$d\mu_n = \sum_{j=1}^{n} q_{t_n}(x_j, x_i)\delta_{x_j}$$

converges weakly to the marginal distribution $d\mu = \widehat{q_t}(x, x_i)dx$ for $Y_{\widehat{t}}$. For any $x \in \mathcal{X}$ and $\delta > 0$, weak convergence against the test function $1_{B(x,\delta)}$ yields

$$\sum_{y\in\mathcal{X}_n, |y-x|<\delta} q_{t_n}(y, x_i) \to \int_{|y-x|<\delta} \widehat{q_t}(y, x_i)dy.$$

By eventual uniform equicontinuity of $nq_t(x, x_i)$, for any $\varepsilon > 0$ there is small enough $\delta > 0$ so that for all n we have

$$\left| \sum_{y\in\mathcal{X}_n, |y-x|<\delta} q_{t_n}(y, x_i) - |\mathcal{X}_n \cap B(x, \delta)| q_{t_n}(x, x_i) \right| \leq n^{-1} |\mathcal{X}_n \cap B(x, \delta)| \varepsilon,$$

which implies that

$$\lim_{n\to\infty} q_{t_n}(x, x_i)p(x)n = \lim_{\delta\to 0}\lim_{n\to\infty} V_d^{-1}\delta^{-d} n q_{t_n}(x, x_i)\int_{|y-x|<\delta} p(y)dy$$

$$= \lim_{\delta\to 0}\lim_{n\to\infty} V_d^{-1}\delta^{-d}|\mathcal{X}_n \cap B(x, \delta)| q_{t_n}(x, x_i) = \lim_{\delta\to 0} V_d^{-1}\delta^{-d}\int_{|y-x|<\delta} \widehat{q_t}(y, x_i)dy = \widehat{q_t}(x, x_i).$$

We conclude the desired

$$\lim_{n\to\infty} nq_t(x, x_i) = \frac{\widehat{q_t}(x, x_i)}{p(x)},$$

where uniformity in x comes from ($\star$) and uniformity in x_i comes from uniformity of Theorem C.1.1. □

B.2 Hitting times

In this section, we prove Theorem 3.2.5 (restated as Theorem B.2.11) to generalize the result of von Luxburg et al. [2014] on degenerate behavior of hitting times via Lemma 3.2.2. Our proof consists of two parts. First, we show that by Lemma 3.2.2 we can make the random walk mix before hitting any point. Next, we use this to show that if the chain is sufficiently mixed, then the expected hitting time is degenerate.

B.2.1 Typical hitting times are large

In this subsection, we give a complete proof of Lemma 3.2.2, reproduced here as Lemma B.2.2. Recall that $T_E^{x_i}$ is the hitting time of Y_t from x_i to a domain $E \subset D$. We will require a more general version of the Feynman-Kac theorem.

Theorem B.2.1 ([Øksendal, 2003, Exercise 9.12] Feynman-Kac). *Let Z_t be an Itô process in $\mathbb{R}^d$ defined by*

$$dZ_t = \mu(Z_t)dt + \sigma(Z_t)dB_t.$$

For a function $V(x)$ and T_E^x the hitting time to a domain $E \subset D$, the function

$$u(x) = \mathbb{E}\left[e^{-\int_0^{T_E^x} V(Z_s)ds}\right]$$

is the solution to the boundary value problem

$$\frac{1}{2}Tr[\sigma^T H u \sigma] + \mu(x)\cdot\nabla u - V(x)u = 0$$

with boundary condition $u|_{\partial E} = 1$.

Lemma B.2.2. *For* $x, y \in D$, $d \geq 2$, *and any* $\delta > 0$, *there exists* $s > 0$ *such that* $\mathbb{E}[e^{-T^x_{B(y,s)}}] < \delta$.

Proof. We use Feynman-Kac to compare $\mathbb{E}[e^{-T^x_{B(y,s)}}]$ for the general process to that of Brownian motion. By Theorem 3.2.1, $u_s(x) = \mathbb{E}[e^{-T^x_{B(y,s)}}]$ satisfies the boundary value problem

$$\Delta u_s + 2\nabla p(x) \cdot \nabla u_s - 2u_s\overline{\varepsilon}(x)^{-2} = 0$$

with $u_s|_{B(y,s)} \equiv 1$. This is equivalent to

$$\sum_i \Big(\partial_i[p(x)^2\partial_i u_s] - \frac{2}{d}u_s\frac{p(x)^2}{\overline{\varepsilon}(x)^2}\Big) = 0.$$

Set $v_s(x) = p(x)u_s(x)$ and change variables to obtain

$$\sum_i \Big(\partial_i[p(x)\partial_i v_s - v_s(x)\partial_i p(x)] - \frac{2}{d}v_s\frac{p(x)}{\overline{\varepsilon}(x)^2}\Big) = p(x)\Delta v_s - \Delta p(x)v_s - v_s\frac{2p(x)}{\overline{\varepsilon}(x)^2} = 0,$$

which is equivalent to

$$\Delta v_s - \Big(\frac{\Delta p(x)}{p(x)} + 2\overline{\varepsilon}(x)^{-2}\Big)v_s = 0$$

with boundary condition $v_s|_{\partial B(y,s)} = 1$. Theorem B.2.1 for $V(x) = \frac{\Delta p(x)}{p(x)} + 2\overline{\varepsilon}(x)^{-2}$ implies

$$v_s(x) = \mathbb{E}\left[e^{-\int_0^{\overline{T}^x_{B(y,s)}} \frac{\Delta p(B_r)}{p(B_r)} + 2\overline{\varepsilon}(B_r)^{-2}dr}\right]$$

for B_t Brownian motion started at x and $\overline{T}^x_{B(y,s)}$ the hitting time of B_t to $B(y,s)$. For a constant C depending on p and $\overline{\varepsilon}$, we have

$$u_s(x) = \frac{v_s(x)}{p(x)} \leq \mathbb{E}[e^{-C\overline{T}^x_{B(y,s)}}].$$

Applying Lemma B.2.3 with this C implies $u_s(x) < \delta e^{-c}$, as needed. □

Lemma B.2.3. *For* $x, y \in D$, *let* B_t *be a Brownian motion with reflecting boundary condition in* D *started at* x *and* $T^x_{B(y,s)}$ *its hitting time to* $B(y,s)$. *Then for any* *sufficiently small $C, c, \delta > 0$, there exists some $s > 0$ so that*

$$\mathbb{E}[e^{-CT^x_{B(y,s)}}] < \delta e^{-c}.$$

Proof. Fix $c > 0$ and $\delta > 0$ small enough so that $e^{-c}\delta < 1/10$. If $|x - y| = p$, by Theorem 3 of Byczkowski et al. [2013], the probability density of $T^x_{B(y,s)}$ started at x if there were no outer boundary is bounded by

$$p(t) < C_1 \frac{s^3(p-s)e^{-\frac{(p-s)^2}{2t}}}{pt^{3/2}} \begin{cases} ((t/s^2)^{\frac{d-3}{2}} + (p/s)^{\frac{d-3}{2}})^{-1} & d > 2 \\ \frac{(p/s+t/s^2)^{1/2}(1+\log(p/s))}{(1+\log(1+t/ps))(1+\log(p/s+t/s^2))} & d = 2 \end{cases}$$

for some constant C_1.
Choosing constants: We claim that we can choose s, p, and r with $r > p > s > 0$ so that:

(a) $B(y, r)$ is contained entirely in the domain D;

(b1) $\frac{r^{2-d}-p^{2-d}}{r^{2-d}-s^{2-d}} > \max\{1/2, \frac{1-e^{-c}\delta}{1-\frac{1}{2}e^{-c}\delta}\}$ if $d > 2$;

(b2) $\frac{\log r - \log p}{\log r - \log s} > \max\{1/2, \frac{1-e^{-c}\delta}{1-\frac{1}{2}e^{-c}\delta}\}$ if $d = 2$;

(c1) $C_1 s^d \left(C^{-1} + \int_1^\infty u^{d/2-2}e^{-(p-s)^2u}du\right) < \frac{1}{4}\delta e^{-c}$ if $d > 2$;

(c2a) $C_1 s^2 \int_1^\infty e^{-Ct}(ps+t)^{1/2}dt < \frac{1}{8}\delta e^{-c}$ if $d = 2$;

(c2b) $C_1 s^2(ps+1)^{1/2} \int_0^1 t^{-3/2}e^{-(p-s)^2/2t}dt < \frac{1}{8}\delta e^{-c}$ if $d = 2$.

For $d > 2$, we have that

$$\frac{\Gamma(d/2-1)}{(p-s)^{d-2}} = (p-s)^{2-d}\int_0^\infty x^{d/2-2}e^{-x}dx = \int_0^\infty u^{d/2-2}e^{-(p-s)^2u}du > \int_1^\infty u^{d/2-2}e^{-(p-s)^2u}du.$$

Then for $p = 2qr$ and $s = qr$, we have that

$$\frac{r^{2-d} - p^{2-d}}{r^{2-d} - s^{2-d}} = \frac{1 - 2^{d-2}q^{d-2}}{1 - q^{d-2}} > 1 - 2^{d-2}q^{d-2}$$

and that

$$s^d \frac{\Gamma(d/2-1)}{(p-s)^{d-2}} < q^2r^2\Gamma(d/2-1)$$

Therefore, sending $r \to 0$ and $q \to 0$ gives a choice of $r > p > s$ satisfying (a), (b1), and (c1) as needed.

For $d = 2$, notice that for $t = u^{-1}$, we have

$$\int_0^1 t^{-3/2} e^{-(p-s)^2/2t} dt = \int_1^{\infty} u^{-1/2} e^{-(p-s)^2 u/2} du.$$

Observe now that

$$(p-s)^{-1}\Gamma(1/2) = (p-s)^{-1} \int_0^{\infty} t^{-1/2} e^{-t} dt = \int_0^{\infty} u^{-1/2} e^{-(p-s)^2 u} du > \int_1^{\infty} u^{-1/2} e^{-(p-s)^2 u} du,$$

whence we conclude that

$$C_1 \frac{s^2\sqrt{ps+1}}{p-s}\Gamma(1/2) > C_1 s^2 (ps+1)^{1/2} \int_0^1 t^{-3/2} e^{-(p-s)^2/2t} dt.$$

Again choose $p = 2qr$ and $s = qr$, for which we obtain

$$C_1 s^2 (ps+1)^{1/2} \int_0^1 t^{-3/2} e^{-(p-s)^2/2t} dt < C_1 \Gamma(1/2) qr \sqrt{4q^2r^2+1}.$$

Sending q and r to 0 then yields $r > p > s$ satisfying (a), (b2) because $q \to 0$, (c2a) because $s \to 0$ and $(ps+t)^{1/2} < (1+t)^{1/2}$, and (c2b) by the estimate above.

Bounding the Laplace transform: Having chosen $r > p > s > 0$ with the desired properties, we have for any $z \in D$ that

$$\mathbb{E}[e^{-CT^z_{B(y,s)}}] \leq \max_{|x-y|=p} \mathbb{E}[e^{-CT^x_{B(y,s)}}].$$

Our strategy will be to bound $\mathbb{E}[e^{-CT^x_{B(y,s)}}]$ for any x with $|x-y| = p$. Fix such an x. Let E be the event that the walk hits $B(y,s)$ before $B(y,r)$. By Theorem 3.17 of Mörters and Peres [2010], the probability of E is $\frac{r^{2-d}-p^{2-d}}{r^{2-d}-s^{2-d}}$ if $d > 2$ and $\frac{\log r - \log p}{\log r - \log s}$ if $d = 2$. By our choice of parameters, this probability is at least $\mathbb{P}(E) > \max\{1/2, \frac{1-e^{-c\delta}}{1-\frac{1}{2}e^{-c\delta}}\}$.

Let $\mathbb{E}'[e^{-CT^x_{B(y,s)}}]$ denote the case where there is no outside boundary. For $d > 2$, we have

$$\begin{aligned}\mathbb{E}'[e^{-CT^x_{B(y,s)}}] &< C_1 s^d \int_0^\infty e^{-Ct} t^{-d/2} e^{-(p-s)^2/2t} dt \\ &< C_1 s^d \left(C^{-1} + \int_0^1 t^{-d/2} e^{-(p-s)^2/2t} dt\right) \\ &< C_1 s^d \left(C^{-1} + \int_1^\infty u^{d/2-2} e^{-(p-s)^2 u} du\right) \\ &< \frac{1}{4}\delta e^{-c}\end{aligned}$$

by the choice of s and p. For $d = 2$, we have

$$\mathbb{E}'[e^{-CT^x_{B(y,s)}}] < C_1 s^2 \int_0^\infty e^{-Ct} t^{-3/2} e^{-(p-s)^2/2t} (ps + t)^{1/2} dt < \frac{1}{4}\delta e^{-c}$$

again by our choice of s and p. Conditioning on E, we find that

$$\mathbb{E}[e^{-CT^x_{B(y,s)}} | E] \leq \mathbb{P}(E)^{-1} \mathbb{E}'[e^{-CT^x_{B(y,s)}}] < \frac{1}{2}\delta e^{-c}.$$

This implies the desired

$$\mathbb{E}[e^{-CT^x_{B(y,s)}}] \leq \mathbb{P}(E)\mathbb{E}[e^{-CT^x_{B(y,s)}}] + (1 - \mathbb{P}(E)) < \delta e^{-c}. \qquad \square$$

B.2.2 Exponential mixing on metric graphs

In this subsection, we show that mixing rates are exponential on metric graphs as assuming $(\star)$.

Lemma B.2.4 (Uniform Doeblin condition). *Assuming $(\star)$, there exist $\alpha > 0$ and $K < \infty$ so that for some $n_0 > 0$ and $\widehat{t}_0 > 0$, we have for $n > n_0$ and $\widehat{t} > \widehat{t}_0$ that*

1. $\min_{x, x_i \in \mathcal{X}_n} q^n_{\lceil \widehat{t} g_n^{-2} \rceil}(x, x_i) > \frac{\alpha}{n}$;

2. $\max_{x, x_i \in \mathcal{X}_n} q^n_{\lceil \widehat{t} g_n^{-2} \rceil}(x, x_i) < \frac{K}{n}$.

Proof. By Lemma B.1.1, assuming $(\star)$ we have $nq^n_{\lceil \widehat{t} g_n^{-2} \rceil}(x, x_i) \to \widehat{q_t}(x, x_i)/p(x)$, where convergence is uniform in x and x_i. Therefore, we may choose $n_0 > 0$ and $\widehat{t}_0 > 0$ so that for $n > n_0$ and $\widehat{t} > \widehat{t}_0$, we have for any $x, x_i \in \mathcal{X}_n$ that

$$\min_{x, x_i \in D} \widehat{q_t}(x, x_i)/p(x) \leq nq^n_{\lceil \widehat{t} g_n^{-2} \rceil}(x, x_i) \leq \max_{x, x_i \in D} \widehat{q_t}(x, x_i)/p(x),$$

where the first and last quantities are well-defined by compactness of D. Taking

$$\alpha = \frac{1}{2}\min_{x,y\in D} \widehat{q_t}(x,y)/p(x) \qquad \text{and} \qquad K = 2\max_{x,y\in D} \widehat{q_t}(x,y)/p(x)$$

thus fulfills the desired conditions. □

Theorem B.2.5. *Then we may choose $\widehat{t}_0, n_0 > 0$ and $C, \beta > 0$ so that for $\widehat{t} > \widehat{t}_0$, $n > n_0$ and $x_i, x \in \mathcal{X}_n$, we have*

$$|q^n_{\lceil \widehat{t} g_n^{-2}\rceil}(x, x_i) - \pi_{X_n}(x)| < C\exp(-\beta\widehat{t})\pi_{X_n}(x).$$

Proof. By Lemma B.2.4, the family of processes X^n_t satisfies the uniform Doeblin condition of Eloranta [1990, Section 2.8]. The claim follows by the consequences for exponential mixing given in the analogue of [Eloranta, 1990, Theorem 2.7]. □

Corollary B.2.6. *Assuming ($\star$), the rescaled stationary distribution $n\pi_{X^n}(x)$ is a.s. eventually uniformly equicontinuous.*

Proof. Choose $\alpha, K, \widehat{t}_1, n_1$ by Lemma B.2.4 so that for $n > n_1, \widehat{t} > \widehat{t}_1$, we have

$$\frac{\alpha}{n} < q^n_{\lceil \widehat{t} g_n^{-2}\rceil}(x, x_i) < \frac{K}{n}.$$

Choose $C, \beta, \widehat{t}_2, n_2$ by Theorem B.2.5 so that for $n > n_2, \widehat{t} > \widehat{t}_2$, we have

$$|q^n_{\lceil \widehat{t} g_n^{-2}\rceil}(x, x_i) - \pi_{X_n}(x)| < C\exp(-\beta\widehat{t})\pi_{X_n}(x).$$

Thus, for $n > \max\{n_1, n_2\}$ and $\widehat{t} > \max\{\widehat{t}_1, \widehat{t}_2\}$, we have

$$|q^n_{\lceil \widehat{t} g_n^{-2}\rceil}(x, x_i) - \pi_{X_n}(x)| < C\exp(-\beta\widehat{t})\pi_{X_n}(x) < \frac{CK}{n}\exp(-\beta\widehat{t}). \tag{B.1}$$

Now, for any $\gamma > 0$, choose $n_0 > \max\{n_1, n_2\}$ and $\widehat{t}_0 > \max\{\widehat{t}_1, \widehat{t}_2\}$ large enough and $\delta > 0$ so that

- $\frac{CK}{n_0}\exp(-\beta\widehat{t}_0) < \gamma/3$;
- by eventual uniform equicontinuity of $nq^n_{\lceil \widehat{t}_0 g_n^{-2}\rceil}(x, x_i)$, for $n > n_0$, if $|x - y| < \delta$, then
$$|nq^n_{\lceil \widehat{t}_0 g_n^{-2}\rceil}(x, x_i) - nq^n_{\lceil \widehat{t}_0 g_n^{-2}\rceil}(y, x_i)| < \gamma/3.$$

Now, for $n > n_0$ and $|x - y| < \delta$, we find that

$$\begin{aligned} |n\pi_{X^n}(x) - n\pi_{X^n}(y)| &\leq |nq^n_{\lceil \widehat{t}_0 g_n^{-2} \rceil}(x, x_i) - nq^n_{\lceil \widehat{t}_0 g_n^{-2} \rceil}(y, x_i)| + |nq^n_{\lceil \widehat{t}_0 g_n^{-2} \rceil}(x, x_i) - n\pi_{X^n}(x)| \\ &\quad + |nq^n_{\lceil \widehat{t}_0 g_n^{-2} \rceil}(y, x_i) - n\pi_{X^n}(y)| \\ &< \gamma/3 + \frac{2CK}{n} \exp(-\beta \widehat{t}_0) \\ &< \gamma, \end{aligned}$$

where we apply (B.1). This implies that $n\pi_{X^n}(x)$ is eventually uniformly equicontinuous, as needed. □

B.2.3 Expected hitting times degenerate to the stationary distribution

For any x_j, let π'_{X^n} be the stationary distribution of the simple random walk on the graph G'_n formed from G_n by removing x_j and all edges incident to it.

Lemma B.2.7. *Assuming ($\star$), the rescaled stationary density $n\pi'_{X^n}(x)$ is a.s. eventually uniformly equicontinuous and satisfies*

$$\lim_{n \to \infty} n\pi'_{X^n}(x) = \widehat{\pi}(x).$$

Proof. Let $q'_t(x, x_i)$ be the marginal distribution of the simple random walk on the modified graph G'_n. Because G'_n is also a metric graph, by Theorem 2.3.4, the time-rescaled simple random walks on G'_n and G_n converge to the same continuous-time Itô process, and we have under ($\star$) that

$$\lim_{n \to \infty} n\pi'_{X^n}(x_i) = \widehat{\pi}(x_i),$$

where the convergence is uniform in x_i. □

Lemma B.2.8. *There exist $t_0 > 0$, $n_0 > 0$, and $C, \beta > 0$ so that for all $\widehat{t} > t_0$ and $n > n_0$ and any integer $t > \widehat{t} g_n^{-2}$, we have*

$$\left| n\mathbb{P}\left(T^{x_i}_{x_j,n} = t \mid T^{x_i}_{x_j,n} \geq t\right) - n \sum_{x \in \mathsf{NB}^{\mathsf{in}}_n(x_j)} \frac{\pi'_{X^n}(x)}{|\mathsf{NB}_n(x)|} \right| < C \exp(-\beta t g_n^2).$$

Proof. By Theorem B.2.5, we may choose $\widehat{t}_0 > 0$, $n_0 > 0$, $C_1 > 0$, and $\beta > 0$ so that for $\widehat{t} > \widehat{t}_0$ and $n > n_1$, we have

$$|q'_{\lceil \widehat{t} g_n^{-2} \rceil - 1}(x, x_i) - \pi'_{X^n}(x)| < C_1 \exp(-\beta \widehat{t}) \pi'_{X^n}(x_j).$$

We claim that the desired result will hold for $\widehat{t}_0$ and this n_0.

By definition, we have

$$\mathbb{P}\left(T^{x_i}_{x_j,n} = t \mid T^{x_i}_{x_j,n} \geq t\right) = \sum_{x \in \mathsf{NB}^{\mathsf{in}}_n(x_j)} \frac{q'_{t-1}(x, x_i)}{|\mathsf{NB}_n(x)|}$$

from which we conclude that for $t > \widehat{t}_0 g_n^{-2}$ we have

$$\begin{aligned}\left|\mathbb{P}\left(T^{x_i}_{x_j,n} = t \mid T^{x_i}_{x_j,n} \geq t\right) - \sum_{x \in \mathsf{NB}^{\mathsf{in}}_n(x_j)} \frac{\pi'_{X^n}(x)}{|\mathsf{NB}_n(x)|}\right| &\leq \left| \sum_{x \in \mathsf{NB}^{\mathsf{in}}_n(x_j)} \frac{q'_{t-1}(x, x_i) - \pi'_{X^n}(x)}{|\mathsf{NB}_n(x)|} \right| \\ &< C_1 \exp(-\beta t g_n^2) \sum_{x \in \mathsf{NB}^{\mathsf{in}}_n(x_j)} \frac{\pi'_{X^n}(x)}{|\mathsf{NB}_n(x)|} \\ &< C_1 \exp(-\beta t g_n^2) \frac{|\mathsf{NB}^{\mathsf{in}}_n(x_j)|}{\min_x |\mathsf{NB}_n(x)|} \max_x \pi'_{X^n}(x).\end{aligned}$$

We now show there exists $C_2 > 0$ such that $\frac{|\mathsf{NB}^{\mathsf{in}}_n(x_j)|}{\min_x |\mathsf{NB}_n(x)|} < C_2$ almost surely due to the construction of g_n and $\overline{\varepsilon}$. By the out-degree estimate of an isotropic graph Theorem A.3.2, we have

$$\frac{|\mathsf{NB}_n(x)|}{|\mathcal{X}_n \cap B(x, \varepsilon_n(x))|} \to C(h)p(x)$$

for some constant $C(h)$ independent of x and n. Further, since the number of points in $|\mathcal{X}_n \cap B(x, \varepsilon_n(x))| \sim \mathrm{Pois}(\varepsilon_n(x)^d V_d p(x))$, we obtain for a constant $0 < C_x < \infty$ dependent on p, V_d and $C(h)$ that

$$\frac{|\mathsf{NB}_n(x)|}{\varepsilon_n(x)^d n} \to C_x.$$

For the denominator $\min_x |\mathsf{NB}_n(x)|$, the above limit immediately implies that $\min_x |\mathsf{NB}_n(x)| \varepsilon_n^{-d} n^{-1}$ - $\min_x C_x > 0$ by the lower bounds on $p(x)$ and $\varepsilon_n(x)$.

For the numerator, note that by construction of $\overline{\varepsilon}$, for any $\delta > 0$ there exists a n such that $\varepsilon_n(x)^{-d} < (1 + \delta) \max_x \overline{\varepsilon}(x)^{-d} g_n^{-d}$ almost surely. By the expectation in Theorem A.3.2, the out-neighborhood of a graph constructed with uniform scale $\max_x \bar{\varepsilon}(x) g_n$ asymptotically dominate the in-neighborhood of the original graph. Therefore,

$$\max_x |\mathsf{NB}_n(x)^{\mathsf{in}}| \bar{\varepsilon}(x)^{-d} g_n^{-d} n^{-1} < \max_x C_x(1+\delta) < \infty.$$

Combining the two bounds gives that

$$\frac{|\mathsf{NB}_n^{\mathsf{in}}(x_j)|}{\min_x |\mathsf{NB}_n(x)|} < (1+\delta)\frac{\max_x C_x \bar{\varepsilon}(x)^d}{\min_x C_x \bar{\varepsilon}(x)^d}.$$

The ratio $\frac{\max_x C_x}{\min_x C_x}$ is bounded by definition of $p(x)$, and therefore there exists $C_2 > 0$ such that $\frac{|\mathsf{NB}_n^{\mathsf{in}}(x_j)|}{\min_x |\mathsf{NB}_n(x)|} < C_2$ almost surely. Finally, by Lemma B.2.7, there exists $C_3 > 0$ such that $\pi'_{\mathcal{X}^n}(x) \leq C_3/n$ for large enough n. The original statement follows by setting $C = C_1 C_2 C_3$. □

Lemma B.2.9. *We have the limit*

$$\lim_{n\to\infty} \sum_{x \in \mathsf{NB}_n^{\mathsf{in}}(x_j)} \frac{1}{|\mathsf{NB}_n(x)|} = 1.$$

Proof. We will proceed through three estimates.

Estimating $\bar{\varepsilon}(x)$ for $x \in \mathsf{NB}_n^{\mathsf{in}}(x_j)$: For $\sigma > 0$, define $\gamma = \sigma \min_x \bar{\varepsilon}(x) > 0$. We may choose $\delta > 0$ so that if $|x - y| < \delta$, then $|\bar{\varepsilon}(x) - \bar{\varepsilon}(y)| < \gamma$. Choose n_0 so that if $n > n_0$ then $g_n \max_x \bar{\varepsilon}(x) < \delta/2$. For $n > n_0$, we find that for $x \in \mathsf{NB}_n^{\mathsf{in}}(x_j)$, we have

$$|x - x_j| \leq \varepsilon_n(x) \leq g_n \max_x \bar{\varepsilon}(x) < \delta$$

and therefore that

$$|\bar{\varepsilon}(x) - \bar{\varepsilon}(x_j)| < \sigma \min_x \bar{\varepsilon}(x) \leq \sigma \bar{\varepsilon}(x_j).$$

This implies that for $n > n_0$ we have

$$(1-\sigma)\bar{\varepsilon}(x_j) < \bar{\varepsilon}(x) < (1+\sigma)\bar{\varepsilon}(x_j). \tag{B.2}$$

Estimating $|\mathsf{NB}_n(x)|$ for $x \in \mathsf{NB}_n^{\mathsf{in}}(x_j)$: By Theorem A.3.2, we have

$$\frac{|\mathsf{NB}_n(x)|}{|\mathcal{X}_n \cap B(x, \varepsilon_n(x))|} \to C(h)p(x)$$

for some constant $C(h)$ independent of x and n. For any $\tau > 0$, we may therefore find some n_1 so that for $n > n_1$ we have

$$(1-\tau)C(h)p(x)|\mathcal{X}_n \cap B(x, \varepsilon_n(x))| < |\mathsf{NB}_n(x)| < (1+\tau)C(h)p(x)|\mathcal{X}_n \cap B(x, \varepsilon_n(x))|.$$

On the other hand, by (B.2), for $x \in \mathsf{NB}_n^{\mathsf{in}}(x_j)$ and any $\sigma > 0$ there is some n_0 so that for $n > n_0$ we have

$$|\mathcal{X}_n \cap B(x, (1-\sigma)\varepsilon_n(x_j))| < |\mathcal{X}_n \cap B(x, \varepsilon_n(x))| < |\mathcal{X}_n \cap B(x, (1+\sigma)\varepsilon_n(x_j))|. \quad \text{(B.3)}$$

Estimating $|\mathsf{NB}_n^{\mathsf{in}}(x_j)|$: By (B.2) and an analogue of the proof of Theorem A.3.2, we have for $x \in \mathsf{NB}_n^{\mathsf{in}}(x_j)$ that for any $\rho > 0$, there is $n_2 > 0$ so that if $n > n_2$ then

$$(1-\rho)C(h)p(x)|\mathcal{X}_n \cap B(x, (1-\sigma)\varepsilon_n(x_j))| < |\mathsf{NB}_n^{\mathsf{in}}(x_j)| < (1+\rho)C(h)p(x)|\mathcal{X}_n \cap B(x, (1+\sigma)\varepsilon_n(x_j))|. \quad \text{(B.4)}$$

Completing the proof: The conclusion follows by taking $\tau, \sigma, \rho \to 0$, choosing n large, and combining (B.3) and (B.4). □

Lemma B.2.10. *The quantity $\theta_n(x_j) = \sum_{x \in \mathsf{NB}_n^{\mathsf{in}}(x_j)} \frac{\pi'_{X^n}(x)}{|\mathsf{NB}_n(x)|}$ satisfies*

$$\lim_{n\to\infty} n\theta_n(x_j) = \widehat{\pi}(x_j).$$

Proof. Fix a sequence of points $y_1, y_2, \dots$ in $\mathcal{X}$ with $y_k \in G'_k$ so that $\lim_{k\to\infty} y_k = x_j$. Fix any $\delta > 0$. By Lemma B.2.7, we may find some n_0 so that for $n > n_0$, for each $x \in \mathsf{NB}_n^{\mathsf{in}}(x_j)$ we have

$$|\pi'_{X^n}(x) - \pi'_{X^n}(y_n)| < \delta/2.$$

This implies that for $n > n_0$ we have

$$\left| n\theta_n(x_j) - n\pi'_{X^n}(y_n) \sum_{x \in \mathsf{NB}_n^{\mathsf{in}}(x_j)} \frac{1}{|\mathsf{NB}_n(x)|} \right| < \frac{\delta}{2} \sum_{x \in \mathsf{NB}_n^{\mathsf{in}}(x_j)} \frac{1}{|\mathsf{NB}_n(x)|}.$$

The result then follows by Lemma B.2.9 and Lemma B.2.7. □

Theorem B.2.11. *For any x_i and x_j, we have*

$$\frac{\mathbb{E}[T^{x_i}_{x_j,n}]}{n} \to \frac{1}{\widehat{\pi}(x_j)},$$

where the convergence is a.s. in the draw of $\mathcal{X}$.

Proof. By definition, we have

$$\mathbb{E}[T^{x_i}_{x_j,n} \mid T^{x_i}_{x_j,n} > \widehat{t}g_n^{-2}] \geq \mathbb{E}[T^{x_i}_{x_j,n}] \geq \mathbb{P}(T^{x_i}_{x_j,n} > \widehat{t}g_n^{-2})\mathbb{E}[T^{x_i}_{x_j,n} \mid T^{x_i}_{x_j,n} > \widehat{t}g_n^{-2}]. \tag{B.5}$$

By Corollary 3.2.3, for any $\delta > 0$ and $\widehat{t}_0 > 0$, there is some n_1 so that for $n > n_1$ and $\widehat{t} > \widehat{t}_0$ we have $\mathbb{P}(T^{x_i}_{x_j,n} > \widehat{t}g_n^{-2}) > (1 - \delta)$. Define now $p_t = \mathbb{P}(T^{x_i}_{x_j,n} = t \mid T^{x_i}_{x_j,n} \geq t)$; by definition we have

$$\mathbb{E}[T^{x_i}_{x_j,n} \mid T^{x_i}_{x_j,n} > \widehat{t}g_n^{-2}] = \sum_{t=\lceil \widehat{t}g_n^{-2} \rceil}^{\infty} tp_t \prod_{r=\lceil \widehat{t}g_n^{-2} \rceil}^{t-1} (1 - p_r).$$

By Lemma B.2.8, we have for some n_2 that for $n > n_2$ and $t > \widehat{t}_0 g_n^{-2}$ that

$$|p_t - \theta_n(x_j)| < \frac{C \exp(-\beta t g_n^2)}{n}$$

so in particular for $\delta = \frac{1}{2} \min_{x \in D} \widehat{\pi}(x)$ and $\tau = 2 \max_{x \in D} \widehat{\pi}(x)$, we have for some n_3 that for $n > n_3$ we have

$$\delta < np_t < \tau \text{ and } \delta < n\theta_n(x_j) < \tau.$$

For n_4 large enough that $1 - \tau/n_4 > \delta/n_4$, for $n > n_4$ we have

$$\begin{aligned} \left| p_t \prod_{r=\lceil \widehat{t}g_n^{-2} \rceil}^{t-1} (1 - p_r) - \theta_n(x_j)(1 - \theta_n(x_j))^{t - \lceil \widehat{t}g_n^{-2} \rceil} \right| &< \sum_{r=\lceil \widehat{t}g_n^{-2} \rceil}^{t-1} \frac{C \exp(-\beta r g_n^2)}{n} (1 - \tau/n)^{t - \lceil \widehat{t}g_n^{-2} \rceil - 1} \\ &< \frac{C}{n} \frac{e^{-\beta \widehat{t}}}{1 - e^{\neg \beta g_n^2}} (1 - \tau/n)^{t - \lceil \widehat{t}g_n^{-2} \rceil - 1}. \end{aligned}$$

This implies that

$$\begin{aligned}
&\frac{1}{n}\left|\mathbb{E}[T^{x_i}_{x_j,n} \mid T^{x_i}_{x_j,n} > \widehat{t}g_n^{-2}] - \sum_{t=\lceil \widehat{t}g_n^{-2}\rceil}^{\infty} t\theta_n(x_j)(1-\theta_n(x_j))^{t-\lceil \widehat{t}g_n^{-2}\rceil}\right| \\
&< \sum_{t=\lceil \widehat{t}g_n^{-2}\rceil}^{\infty} \frac{C}{n^2}\frac{e^{-\beta\widehat{t}}}{1-e^{-\beta g_n^2}}(1-\tau/n)^{t-\lceil \widehat{t}g_n^{-2}\rceil-1} \\
&< \frac{C}{\tau(n-\tau)}\frac{e^{-\beta\widehat{t}}}{1-e^{-\beta g_n^2}},
\end{aligned}$$

where we note that for $n > 2\tau$, we have

$$\frac{C}{\tau(n-\tau)}\frac{e^{-\beta\widehat{t}}}{1-e^{-\beta g_n^2}} < \frac{2Ce^{-\beta\widehat{t}}}{\tau}n^{-1}(g_n^{-2} + \frac{1}{2} + \frac{1}{12}g_n^2).$$

Because $\lim_{n\to\infty} n^{-1}(g_n^{-2} + \frac{1}{2} + \frac{1}{12}g_n^2) = 0$, considering $n > \max\{n_1, n_2, n_3, n_4\}$, we conclude that

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\mathbb{E}[T^{x_i}_{x_j,n} \mid T^{x_i}_{x_j,n} > \widehat{t}g_n^{-2}] &= \lim_{n\to\infty}\frac{1}{n}\sum_{t=\lceil \widehat{t}g_n^{-2}\rceil}^{\infty} t\theta_n(x_j)(1-\theta_n(x_j))^{t-\lceil \widehat{t}g_n^{-2}\rceil} \\
&= \lim_{n\to\infty}\frac{1}{n}\frac{1-\theta_n(x_j)+\theta_n(x_j)\lceil \widehat{t}g_n^{-2}\rceil}{\theta_n(x_j)} \\
&= \lim_{n\to\infty}\frac{1}{n\theta_n(x_j)} = \frac{1}{\widehat{\pi}(x_j)},
\end{aligned}$$

where the last equality follows from Lemma B.2.10. Now by (B.5), we conclude that

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}[T^{x_i}_{x_j,n}] = \frac{1}{\widehat{\pi}(x_j)}. \qquad \square$$

B.2.4 The case of one dimension

The Laplacian-based bounds in von Luxburg et al. [2014] suggest that the hitting time should diverge even when the dimension of the underlying geometric graph is 1. This is a very surprising result, since the continuous random walk in one dimension converges to a non-trivial limit. We provide another explanation of this result in our framework.

Intuitively, this happens since we are concerned with the hitting time to a single point, and the discrete random walk may jump over the point, while the continuous walk cannot. To demonstrate this, we show that considering the hitting time to a sufficiently large out-neighborhood of a vertex instead of the vertex itself fixes this problem.

Pick $x_i, x_j \in G_n$, and let X_t^n be the simple random walk on G_n. Suppose without loss of generality that $x_i < x_j$ and define

$$\gamma = \inf_n \min_{x_i \in \mathcal{X}_n} x_i$$

to be the left boundary of D. Pick a sequence of sets of vertices $S_n \subset \mathcal{X}_n$ so that every element in S_n is reachable from x_j in $o(g_n^{-1})$ steps and the removal of S_n from G_n disconnects G_n. Let $T_{S_n}^{x_i}$ be the hitting time to any point in S_n. We will use the Feynman-Kac theorem for functionals of hitting time.

Theorem B.2.12 ([Øksendal, 2003, Exercise 9.12] Feynman-Kac). *Let Z_t be an Itô process in $\mathbb{R}^d$ defined by*

$$dZ_t = \mu(Z_t)dt + \sigma(Z_t)dB_t.$$

For a function $f(x)$ and T_E^x the hitting time to a domain $E \subset D$, the function

$$u(x) = \mathbb{E}\left[\int_0^{T_E^x} f(Z_s)ds\right]$$

is the solution to the boundary value problem

$$\frac{1}{2}Tr[\sigma^T H u \sigma] + \mu(x) \cdot \nabla u + f(x) = 0$$

with boundary condition $u|_{\partial E} = 1$.

Theorem B.2.13. *Such a sequence of vertex sets S_n always exists and the expected hitting time $\mathbb{E}[T_{S_n,n}^{x_i}]$ converges to a non-degenerate continuum limit defined by*

$$\mathbb{E}[T_{S_n,n}^{x_i} g_n^2] \to \int_{x_i}^{x_j} \frac{1}{p(y)^2} \int_\gamma^y \frac{2p(z)^2}{\bar{\varepsilon}(z)^2} dz dy.$$

Proof. First we prove a sequence S_n exists. Take the set of points $\widehat{S}_n = \{x_k : |x_k - x_j| < c_n\}$ for a sequence c_n with $c_n \to 0$ and $c_n g_n \to \infty$. Let s be the maximum shortest path distance to any element in $\widehat{S}_n$. Then we have $s = o(g_n^{-1})$ since $c_n \to 0$ and the length of the shortest path between any two points scales as

$\Theta(g_n^{-1})$. Therefore the set S_n defined by all points whose shortest path distance to x_j is at most s fulfills the requirements.

Let $\widehat{T}_{ij}$ be the hitting time to x_j of $Y_{\widehat{t}}$ started at x_i. Note that it is not infinite because we have $d = 1$. By Corollary 3.3.3 and the fact that $sg_n^{-1} \to 0$, we have

$$T_{S_n,n}^{x_i} g_n^2 \xrightarrow{d} \widehat{T}_{ij}.$$

Finally, by Theorem B.2.12 with $f(x) \equiv 1$, the expected hitting time $u(x)$ to x_j under the continuous process $Y_{\widehat{t}}$ started at x is the solution to the boundary value problem

$$\frac{1}{2}\overline{\varepsilon}(x)^2 u''(x) + \frac{p'(x)}{p(x)}\overline{\varepsilon}(x)^2 u'(x) + 1 = 0$$

We may rewrite this as

$$p(x)^2 u''(x) + 2p(x)p'(x)u'(x) = -\frac{2p(x)^2}{\overline{\varepsilon}(x)^2},$$

after which integration of both sides and application of $u'(\gamma) = 0$ implies that

$$p(x)^2 u'(x) = -\int_\gamma^x \frac{2p(z)^2}{\overline{\varepsilon}(z)^2} dz.$$

Another integration and application of $u(x_j) = 0$ implies that

$$u(x) = -\int_{x_j}^x \frac{1}{p(y)^2} \int_\gamma^y \frac{2p(z)^2}{\overline{\varepsilon}(z)^2} dz dy.$$

Setting $x = x_i$ then implies that

$$\lim_{n\to\infty} \mathbb{E}[T_{S_n,n}^{x_i} g_n^2] = \mathbb{E}[\widehat{T}_{ij}] = \int_{x_i}^{x_j} \frac{1}{p(y)^2} \int_\gamma^y \frac{2p(z)^2}{\overline{\varepsilon}(z)^2} dz dy. \qquad \square$$

For cases where the kernel function takes values in $\{0, 1\}$, such as the k-nearest neighbor graph, the following corollary is useful.

Corollary B.2.14. *Suppose that G_n is constructed by the kernel $h(x) = 1_{[0,1]}$. Then the expected hitting time of X_t^n started at x_i to the out-neighbors of x_j converges to the limit of Theorem B.2.13*

Proof. From the fact that the out-neighborhood of x_j satisfies the conditions for S_n in Theorem B.2.13. □

Although this metric is nontrivial in the sense that it retains some information about the latent space metric, it is still highly distorted. We examine this phenomenon in the case of $\overline{\varepsilon}(x) = 1$ and $p(x) = 1$ in the following Corollary.

Corollary B.2.15. *If $\overline{\varepsilon}(x) = 1$ and $p(x) = 1$ in Corollary B.2.14, for any x_i and x_j the rescaled expectation of the hitting time $T^{x_i}_{\mathsf{NB}_n(x_j),n}$ of X^n_t started at x_i to the out-neighborhood of x_j has the limit*

$$\mathbb{E}[T^{x_i}_{\mathsf{NB}_n(x_j),n} g_n^2] \to |x_j - x_i| \cdot |x_j + x_i - 2\gamma|.$$

Proof. This follows by applying Theorem B.2.13 with our $\overline{\varepsilon}(x)$ and $p(x)$. □

Remark. Note that the boundary condition in Corollary B.2.15 induces a large non-uniform multiplicative error. Because of this, the expected hitting time is not consistent even in the ideal situation of a one-dimensional latent space with random walk converging to Brownian motion. Compare this result with Theorem 3.3.4, which shows a much stronger consistency property.

B.3 Computing the LTHT

Algorithmically, computing the LTHT can be done in two major ways: matrix inversion, or sampling. For the results in the paper, we use the direct sampling method of drawing a simple random walk and calculating the exponentially discounted hitting time. This same computation can be performed using a truncated power method Yazdani [2013, Algorithm 1].

Alternative approaches for computing the LTHT involve the following matrix inversion method. Let P be the transition matrix for some random walk. Then the LTHT $\mathbb{E}[\exp(-\beta T^{x_i}_{x_j,n})]$ is given by

$$\mathbb{E}[\exp(-\beta T^{x_i}_{x_j,n})] = (I - W\exp(-\beta))^{-1}_{ji}.$$

Note that this expression is a close discrete analog of Feynman-Kac (Theorem B.2.1). This relationship was used in prior work [Smith et al., 2014, Eq. 22] to calculate the LTHT in a different setting and formulation. Correctness of this expression can be seen via the series expansion which was computed as a normalizer for randomized shortest paths [Françoisse et al., 2013, Algorithm 2]. This method has been used to calculate the LTHT in in prior work [Kivimäki et al., 2014].

B.4 Reweighting the random walk

Recall that A^n_{ij} was the adjacency matrix of G_n. In Corollary B.4.2, we give a complete proof of Theorem 3.3.1 from the maintext.

B.4.1 General construction and application to Brownian motion

Let $a_n(x)$ and $b_n(x)$ be scalar functions on $\mathcal{X}_n$ with possibly stochastic dependence on $\mathcal{X}_n$ so that

$$\lim_{n\to\infty} a_n(x) = \overline{a}(x) \qquad \text{and} \qquad \lim_{n\to\infty} b_n(x) = \overline{b}(x)$$

uniformly in x for some deterministic $\overline{a}(x)$ and $\overline{b}(x)$.

Theorem B.4.1. *If $a_n(x)$ is a.s. eventually equicontinuous, $\overline{a}(x)$ is smooth with bounded gradient, and $\overline{b}(x)$ is continuous and bounded in $(0,1]$, the weighted random walk Z_t defined by the transition matrix*

$$\mathbb{P}(Z_{t+1} = x_j \mid Z_t = x_i) = \begin{cases} A^n_{i,j} \frac{a_n(x_j)}{\sum_{x_k \in \mathsf{NB}_n(x_i)} a_n(x_k)} b_n(x_i) & i \neq j \\ 1 - b_n(x_i) & i = j \end{cases}$$

converges to the Itô process with drift $\nabla \log(p(x)\overline{a}(x))/3$ and diffusion $\overline{\varepsilon}(x)^2\overline{b}(x)/3$.

Proof. To show convergence to an Itô process, it suffices to check the Stroock-Varadhan criterion [Stroock and Varadhan, 1971b]. Since the boundary for both the original and modified walk are the same, we only need check that

$$\mathbb{E}[Z_{t+1} - x_i \mid Z_t = x_i] \xrightarrow{P} \frac{1}{3}\frac{\nabla[p(x_i)\overline{a}(x_i)]}{p(x_i)\overline{a}(x_i)}\overline{\varepsilon}(x_i)^2\overline{b}(x_i), \text{ and}$$
$$\mathbb{E}[(Z_{t+1} - x_i)^2 \mid Z_t = x_i] \xrightarrow{P} \frac{1}{3}\overline{\varepsilon}(x_i)^2\overline{b}(x_i).$$

For this, by definition we have that

$$\mathbb{E}[Z_{t+1} - x_i \mid Z_t = x_i] = \mathbb{P}(Z_{t+1} \neq x_i)\frac{1}{\sum_{x_k \in \mathsf{NB}_n(x_i)} a_n(x_k)} \sum_{x_k \in \mathsf{NB}_n(x_i)} (x_k - x_i)a_n(x_k), \text{ and}$$
$$\mathbb{E}[(Z_{t+1} - x_i)^2 \mid Z_t = x_i] = \mathbb{P}(Z_{t+1} \neq x_i)\frac{1}{\sum_{x_k \in \mathsf{NB}_n(x_i)} a_n(x_k)} \sum_{x_k \in \mathsf{NB}_n(x_i)} (x_k - x_i)^2 a_n(x_k),$$

from which the desired estimates follow by using $\mathbb{P}(Z_{t+1} \neq x_i) = b_n(x_i)$ and the values and concentration of conditional moments $\mathbb{E}[f(Z_t - Z_{t-1}) \mid Z_{t-1}, Z_t \neq Z_{t-1}]$ given by applying Lemma B.4.3 and Lemma 2.3.2 to $f(x) = x$ and $f(x) = x^2$. □

Corollary B.4.2. *Let $\widehat{p}$ and $\widehat{\varepsilon}$ be consistent estimators of the density and local scale and A be the adjacency matrix. Then the random walk $\widehat{X}_t^n$ defined by the following transition*

$$\mathbb{P}(\widehat{X}_{t+1}^n = x_j \mid \widehat{X}_t^n = x_i) = \begin{cases} \frac{A_{i,j}\widehat{p}(x_j)^{-1}}{\sum_k A_{i,k}\widehat{p}(x_k)^{-1}}\widehat{\varepsilon}(x_i)^{-2} & i \neq j \\ 1 - \widehat{\varepsilon}(x_i)^{-2} & i = j \end{cases}$$

converges to a Brownian motion.

Proof. Set $a_n(x) = \widehat{p}(x)^{-1}$ and $b_n(x) = \widehat{\varepsilon}(x)^{-2}$ as estimated by Corollary 2.2.2 so that $\lim_{n\to\infty} a_n(x) = p(x)^{-1}$ and $\lim_{n\to\infty} b_n(x) = \overline{\varepsilon}(x)^{-2}$. These satisfy the conditions of Theorem B.4.1 and yield limiting drift and diffusion coefficients for Brownian motion. □

B.4.2 Technical moment estimates

In this subsection, we give the moment estimates necessary in the proof of Theorem B.4.1. We first derive the expected values of each moment quantity averaged over draws of $\mathcal{X}_n$.

Lemma B.4.3 (Expected values of reweighting). *Let $x = X_t^n$ and $y = X_{t+1}^n$. Then the conditional expectation after weighting by $a_n(x)$ converges to the weighted draw over $p(x)a_n(x)$; that is, we have a.s. that*

$$\lim_{n\to\infty} \left| \frac{1}{h_n} |\mathsf{NB}_n(x)| \, \mathbb{E}\left[\frac{a_n(y)}{\sum_{z \in \mathsf{NB}_n(x)} a_n(z)} f(y - x) \mid y \neq x \right] \right. \\ \left. - \frac{1}{h_n} \int_{y \in B(x, \varepsilon_n(x))} f(y - x) \frac{p(y)\overline{a}(y)}{\int_{z \in B(x,\varepsilon_n(x))} p(z)\overline{a}(z) dz} dy \right| = 0.$$

Proof. By the continuity of p and a.s. eventual equicontinuity of $a_n(y)$, we have $\sup_{y \in B(x,\varepsilon_n(x))} |a_n(y)p(y) - a_n(x)p(x)| \to 0$ and $\sup_{y \in B(x,\varepsilon_n(x))} |p(y) - p(x)| \to 0$. These together imply

$$\frac{\int_{y \in B(x,\varepsilon_n(x))} a_n(y)p(y)dy}{\int_{y \in B(x,\varepsilon_n(x))} p(y)dy} \xrightarrow{a.s.} \overline{a}(x). \tag{B.6}$$

Because $a_n(x) \to \overline{a}(x)$ uniformly in x, for any $\delta > 0$, we may choose n_0 so that for $n > n_0$, we have $|a_n(x) - \overline{a}(x)| < \delta/2$ and $\varepsilon_n(x)$ is small enough so that if $|y - x| < \varepsilon_n(x)$, then $|\overline{a}(y) - \overline{a}(x)| < \delta/2$. For $n > n_0$, we then have

$$\sup_{z \in \mathsf{NB}_n(x)} |a_n(z) - \overline{a}(x)| \leq \sup_{z \in \mathsf{NB}_n(x)} |a_n(z) - \overline{a}(z)| + |\overline{a}(z) - \overline{a}(x)| < \delta.$$

This shows that $\sup_{z \in \mathsf{NB}_n(x)} |a_n(z) - \overline{a}(x)| \to 0$ and therefore

$$\frac{\sum_{z \in \mathsf{NB}_n(x)} a_n(z)}{|\mathsf{NB}_n(x)|} \xrightarrow{a.s.} \overline{a}(x). \tag{B.7}$$

Applying (B.6) and (B.7), we find that

$$\lim_{n\to\infty} \left| \frac{1}{h_n} |\mathsf{NB}_n(x)| \, \mathbb{E}\left[\frac{a_n(y)}{\sum_{z \in \mathsf{NB}_n(x)} a_n(z)} f(y-x) \mid x \neq y \right] \right. \\ \left. - \frac{1}{h_n} \mathbb{E}\left[a_n(y) f(y-x) \mid x \neq y \right] \frac{\int_{y \in B(x,\varepsilon_n(x))} p(y)dy}{\int_{y \in B(x,\varepsilon_n(x))} a_n(y)p(y)dy} \right| \to 0.$$

We apply the argument of Lemma 2.3.2 to this iterated expectation to obtain

$$\lim_{n\to\infty} \left| \frac{1}{h_n} \mathbb{E}\left[a_n(y) f(y-x) \mid x \neq y \right] \frac{\int_{y \in B(x,\varepsilon_n(x))} p(y)dy}{\int_{y \in B(x,\varepsilon_n(x))} a_n(y)p(y)dy} \right. \\ \left. - \frac{1}{h_n} \int_{y \in B(x,\varepsilon_n(x))} f(y-x) \frac{p(y)\overline{a}(y)}{\int_{z \in B(x,\varepsilon_n(x))} p(z)\overline{a}(z)dz} dy \right| \to 0. \quad \square$$

Evaluating the integrals for $f(x) = x$ and $f(x) = x^2$ in Lemma B.4.3 implies that the expected value of an increment of the reweighted walk across all draws of $\mathcal{X}_n$ limits to $\nabla \log[p(x)\overline{a}(x)]/3$ and the expected variance of the increment limits to $\overline{\varepsilon}(x)^2 \overline{b}(x)/3$. However, in order to apply the Stroock-Varadhan criteria we require that this hold with high probability over all draws of $\mathcal{X}_n$.

Lemma B.4.4 (Strong LLN for local moments). *For a function $f(x)$ such that $\sup_{x \in B(0,\varepsilon)} |f(x)| < \varepsilon$ for small $\varepsilon > 0$, we have a.s. that*

$$\lim_{n\to\infty} \left| \frac{1}{h_n} \sum_{y \in \mathsf{NB}_n(x)} \frac{a_n(y)}{\sum_{z \in \mathsf{NB}_n(x)} a_n(z)} f(y-x) - \frac{1}{h_n} \int_{y \in B(x,\varepsilon_n(x))} f(y-x) \frac{p(y)\overline{a}(y)}{\int_{z \in B(x,\varepsilon_n(x))} p(z)\overline{a}(z)dz} dy \right| = 0.$$

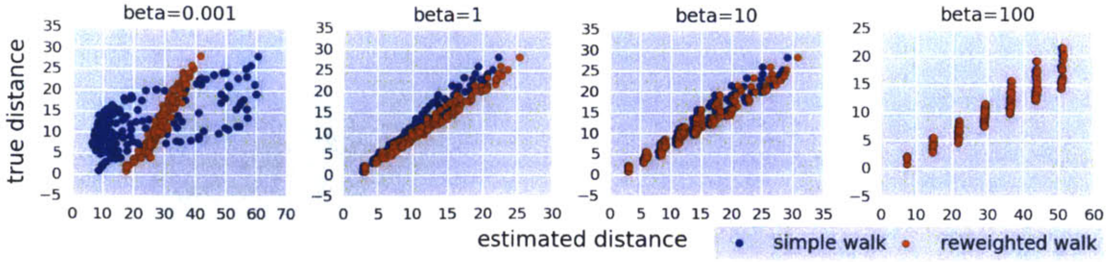


Figure B-1: Distance estimates for various values of β on re-weighted walks on a simulated dataset

Proof. Define the quantity

$$\mu(x) = \frac{1}{h_n} \int_{y \in B(x, \varepsilon_n(x))} f(y - x) \frac{p(y)\overline{a}(y)}{\int_{z \in B(x, \varepsilon_n(x))} p(z)\overline{a}(z) dz} dy.$$

We wish to bound

$$p_n(t) = \mathbb{P}\left(\left| \frac{1}{h_n} \sum_{y \in \mathsf{NB}_n(x)} \frac{a_n(y)}{\sum_{z \in \mathsf{NB}_n(x)} a_n(z)} f(y - x) - \mu(x) \right| \geq t \right). \tag{B.8}$$

By a.s. eventual equicontinuity of $a_n(y)$, we have for some $c > 0$ and large enough n that

$$\frac{a_n(y)}{\sum_{z \in \mathsf{NB}_n(x)} a_n(z)} \leq c \frac{1}{|\mathsf{NB}_n(x)|}.$$

By the construction of $\varepsilon_n(x)$, if $|y - x| < \varepsilon_n(x)$, then $|f(y - x)| \leq \varepsilon_n(x)$. Combining these two we apply Hoeffding's inequality to obtain that

$$p_n(t) \leq 2 \exp\left(-\frac{2h_n^2 |\mathsf{NB}_n(x)|^2 c^2 t^2}{|\mathsf{NB}_n(x)| \varepsilon_n(x)^2} \right) = o(n^{-2t^2 \omega(1)}), \tag{B.9}$$

where we use that $|\mathsf{NB}_n(x)| = \omega\left(n^{2/(d+2)} \log(n)^{d/(d+2)}\right)$. This completes the proof by Borel-Cantelli. □

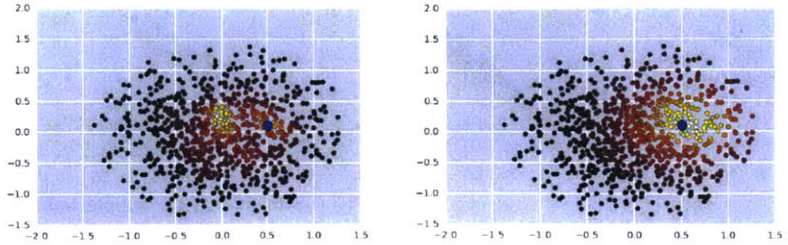


Figure B-2: Simple random walk is biased toward region with high density

Figure B-3: Re-weighted walk diffuses evenly on the true metric

Figure B-4: Visualization of the marginal distribution $P_{ij}(t)$ of a random walk over a k-nn graph on a Gaussian restricted to a disk, starting at the blue initial point and run for 40 steps. The re-weighted walk diffuses evenly from the starting point, ignoring biases due to density p and neighborhood size ε.

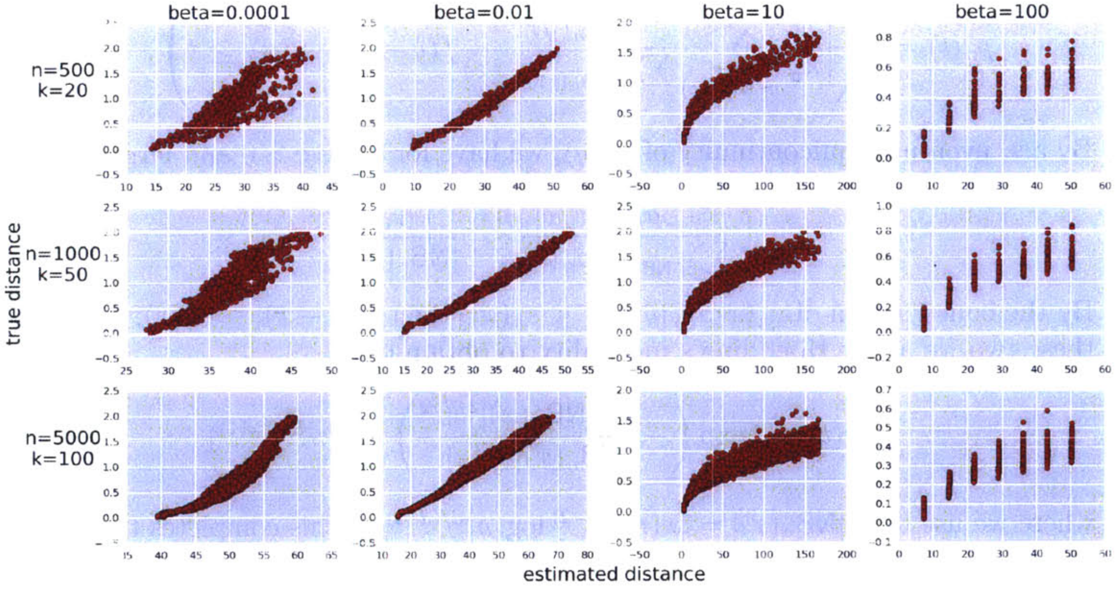


Figure B-5: Distance estimates for various values of β on re-weighted walks on a simulated dataset

B.5 Consistency at $\beta = \omega(\log(g_n^d n))$ via shortest paths

Definition 8. Define the f-length of any path $\gamma \subset D$ as given in Alamgir and von Luxburg [2012b] as

$$D_{f,\gamma} = \int_\gamma f(\gamma(t))|\gamma'(t)|dt.$$

Let the f-distance from x to y be the minimum path length between two points

$$D_f(x,y) = \min_{\gamma \in C^1, \gamma(0)=x, \gamma(1)=y} D_{f,\gamma}.$$

Theorem B.5.1. *Let $\beta = \omega(\log(g_n^d n))$, then for $f(x) = \overline{\varepsilon}(x)^{-1}$ we have*

$$-\log(\mathbb{E}[\exp(-\beta T^{x_i}_{x_j,n})])/\beta g_n \to D_f(x_i, x_j).$$

Proof. Define $H_{ij}(t)$ to be the probability of not hitting x_j by step t, and $P_{ij}(t)$ to be the probability of going from x_i to x_j in exactly t steps. The expected value is the series

$$-\log(\mathbb{E}[\exp(-\beta T^{x_i}_{x_j,n})])/\beta g_n = -\beta^{-1} g_n \log\left(\sum_{t=0}^{\infty} P_{ij}(t) H_{ij}(t) \exp(-\beta t)\right).$$

Now, let D_{ij} be the length of the shortest path from i to j. By definition $H_{ij}(D_{ij}) = 1$ and

$$\begin{aligned} -\log(\mathbb{E}[\exp(-\beta T^{x_i}_{x_j,n})])/\beta g_n &= D_{ij} g_n - \log(P_{ij}(D_{ij}))\frac{g_n}{\beta} \\ &\quad - \log\left(1 + \sum_{t=D_{ij}+1}^{\infty} \frac{P_{ij}(t)}{P_{ij}(D_{ij})} H_{ij}(t) \exp(-\beta(t - D_{ij}))\right) \frac{g_n}{\beta}. \end{aligned}$$

This forms the upper bound

$$-\log(\mathbb{E}[\exp(-\beta T^{x_i}_{x_j,n})])/\beta g_n \leq D_{ij} g_n - \log(P_{ij}(D_{ij}))\frac{g_n}{\beta}.$$

The probability $P_{ij}(D_{ij})$ of hitting x_j in exactly D_{ij} steps is lower bounded by $(g_n^d n)^{-D_{ij}}$ since by definition at least one path exists. This implies that $\log(P_{ij}(D_{ij})) =$

$o(g_n^{-1}\log(g_n^d n))$ and therefore

$$D_{ij}g_n \leq -\log(\mathbb{E}[\exp(-\beta T^{x_i}_{x_j,n})])/\beta g_n \leq D_{ij}g_n - o(1),$$

where the lower bound follows because it is impossible to reach vertex x_j in less than D_{ij} steps. By Alamgir and von Luxburg [2012b] for the k-nearest neighbor case and Theorem A.4.5 with Lemma B.6.3 for the other cases of a metric graph, $D_{ij}g_n$ converges to the f-distance defined by $\overline{\varepsilon}(x)^{-1}$, completing the proof. □

B.6 Consistency of LTHT

In this section, we prove some results needed in the proof of Theorem 3.3.4 (restated as Theorem B.6.6).

B.6.1 LTHT of the Brownian motion

Lemma B.6.1. *Let W_t be a Brownian motion with $W_0 = x_i$. Let $\overline{T}^{x_i}_{B(x_j,s)}$ be the hitting time of W_t to $B(x_j, s)$. For any $\alpha < 0$, if $\widehat{\beta} = s^\alpha$, as $s \to 0$ we have*

$$-\log(\mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)})])/\sqrt{2\widehat{\beta}} \to |x_i - x_j|.$$

Proof of Lemma B.6.1. Let $B_t = |W_t|$ be the order $\nu = d/2 - 1$ Bessel process. The LTHT of B_t to hit $x_j \pm s$ is equivalent to the LTHT of W_t to hit $B(x_j, s)$. Defining $w = |x_i - x_j|$, by Borodin and Salminen [2002, Eq 4.2.0.1], this is:

$$\mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)})] = \frac{K(\nu, w\sqrt{2\widehat{\beta}})w^{-\nu}}{K(\nu, s\sqrt{2\widehat{\beta}})s^{-\nu}},$$

where $K(\nu, w)$ is a modified Bessel function of the second kind.

Write $-\log(\mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)})])/\sqrt{2\widehat{\beta}} = c_1 + c_2$ for

$$c_1 = -\log(K(\nu, w\sqrt{2\widehat{\beta}})w^{-\nu})/\sqrt{2\widehat{\beta}}$$
$$c_2 = -\log(K(\nu, s\sqrt{2\widehat{\beta}})s^{-\nu})/\sqrt{2\widehat{\beta}}.$$

Taylor expansion of c_1 at $\widehat{\beta}^{-1} = 0$ yields

$$c_1 = w - \frac{\log(\pi^2/(8\widehat{\beta})) + 4\log(w^{-1/2-\nu})}{4\sqrt{2\widehat{\beta}}} + o\left(\frac{\nu^2}{w\widehat{\beta}}\right),$$

hence $c_1 \to w$. For c_2, note that $\nu \log(s)/\sqrt{2\widehat{\beta}} \to 0$ and for s small,

$$K(\nu, s\sqrt{2\widehat{\beta}}) \sim \begin{cases} -\log(s\sqrt{2\widehat{\beta}}) & d = 2 \\ \frac{1}{2}\Gamma(s\sqrt{2\widehat{\beta}})(\frac{1}{2}s\sqrt{2\widehat{\beta}})^{-\nu} & d > 2 \end{cases}.$$

by Abramowitz and Stegun [1972, p375]. Checking that $-\log(K(\nu, s\sqrt{2\widehat{\beta}}))/\sqrt{2\widehat{\beta}} \to 0$ and combining estimates gives $-\log(\mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)})])/\sqrt{2\widehat{\beta}} = c_1 + c_2 \to w$. □

B.6.2 Proof of Corollary 3.3.3

We prove here Corollary 3.3.3 (restated below as Corollary B.6.2). We recall the setup. For points $x_i, x_j \in G_n$ and $s > 0$, $\widehat{T}^{x_i}_{B(x_j,s)}$ is the hitting time of the de-biased walk on G_n from x_i to $\mathsf{NB}^s_n(x_j)$. In the continuous setting, $\overline{T}^{x_i}_{B(x_j,s)}$ is the hitting time of Brownian motion with reflecting boundary conditions in D from x_i to $B(x_j, s)$. We would like to show the following.

Corollary B.6.2. *For $s > 0$, we have $g_n^2 \widehat{T}^{x_i}_{B(x_j,s)} \xrightarrow{d} \overline{T}^{x_i}_{B(x_j,s)}$.*

Our proof consists of two steps. First, we show that hitting $\mathsf{NB}^s_n(x_j)$ is equivalent to hitting $B(x_j, s)$ with the discrete walk. Second, we use Corollary B.4.2 to show convergence in distribution of this second hitting time. We require a few lemmas.

Lemma B.6.3. *For any $\delta > 0$ and $s > 0$ so that $B(x_j, s + \delta) \subset D$, we have with high probability that*

$$\mathcal{X}_n \cap B(x_j, s - \delta) \subset \mathsf{NB}^s_n(x_j) \subset B(x_j, s + \delta).$$

Proof. Recall that $\mathsf{NB}^s_n(x_j)$ is defined as

$$\mathsf{NB}^s_n(x) := \{y \mid \text{there is a path } x \to y \text{ of } \widehat{\varepsilon}\text{-weight} \le s\}.$$

The estimator $\widehat{\varepsilon}(x)$ is appropriately scaled such that $\widehat{\varepsilon}(x) \to \overline{\varepsilon}(x)$ uniformly and almost surely. Thus, we need to show that $\widehat{\varepsilon}$-weighted shortest path distance converges to true shortest path distance up to error $\Theta(g_n)$.

We first present the simpler case of a constant kernel $h(x) \equiv 1$ over $[0,1]$; this includes the k-nearest neighbor and ε-ball cases. Let D_{ij} be the minimum $\widehat{\varepsilon}$-weight of a path from x_i to x_j. The proof of Theorem A.4.5 shows that in this case

$$\Big| |x_i - x_j| - D_{ij} g_n \Big| \leq \varepsilon_n(x_j). \tag{B.10}$$

If $x_k \in \mathcal{X}_n \cap B(x_j, s - \delta)$, this implies that $D_{jk} g_n \to |x_k - x_j| \leq s - \delta$. Therefore $D_{jk} g_n \leq s$ with high probability and $x_k \in \mathsf{NB}_n^s(x_j)$. If $x_k \in \mathsf{NB}_n^s(x_j)$, this implies that $D_{ik} \leq s$. By Equation (B.10), we have $s \geq D_{ij} g_n \to |x_i - x_j|$. Therefore $x_k \in B(x_j, s + \delta)$ with high probability.

The proof for the case of generic $h(x)$ is closely analogous. The same proof as used for Theorem A.4.5 shows that there exists some k such that $|x_k - x_j| \leq \varepsilon_n(x_k)$ such that

$$\Big| |x_i - x_j| - D_{ik} g_n \Big| \leq \varepsilon_n(x).$$

At this stage, a difference arises. The proof of Theorem A.4.5 bounds the number of steps necessary to reach distance $\varepsilon_n(x_j)$ to the target, but for a general choice of $h(x)$ this does not guarantee that we can reach x_j.

For general $h(x)$, we instead show that two extra jumps are sufficient. Because $h(1) > 0$ and h is continuous at 1, there exists some interval $(c_1, 1)$ and some $c_2 > 0$ such that

$$\inf_{x \in (c_1, 1)} h(x) > c_2.$$

This annulus will yield a lower bound on the true connectivity. If $|x_i - x_j| \leq \varepsilon_n(x_i)$, then the probability that there is some point x_k such that the path $x_i \to x_k \to x_j$ exists in G_n is governed by

$$\mathbb{P}(D_{ij} > 2) = (1 - c_2)^{2|\mathsf{NB}_n(x_i) \cap \mathsf{NB}_n^{in}(x_j)|}$$

where

$$|\mathsf{NB}_n(x_i) \cap \mathsf{NB}_n^{in}(x_j)| \sim \mathrm{Pois}(g_n^d n \tau(x_i - x_j))$$

and $\tau(z)$ is the total overlapping density between the connectivity kernel of x_i and x_j. This is lower bounded by the annulus; for any $d > 2$ the annuli have nonzero overlap volume and

$$\tau(z) \geq c_2^2 \int_{x \in B(0,1)} 1_{1>|x|>c_1} 1_{1>|1-x|>c_1} dx \geq 0.$$

This implies that $|\mathsf{NB}_n(x_i) \cap \mathsf{NB}_n^{in}(x_j)| = \Theta(k)$ with high probability and therefore

$$\mathbb{P}(D_{ij} > 2) = (1 - c_2)^{2|\mathsf{NB}_n(x_i) \cap \mathsf{NB}_n^{in}(x_j)|} \to 0.$$

Thus, there exists a two step path from x_i to x_j whenever $|x_i - x_j| < \varepsilon_n(x_i)$. Combined with the analogue of Theorem A.4.5, this shows that with high probability there is a walk of $\widehat{\varepsilon}$-weight at most $|x_i - x_k| + 2\varepsilon_n(x_k)$ from x_i to x_k. We conclude in the same way as in the constant kernel case. □

We now require a lemma on the continuity of functions on Skorokhod space. For this, we recall the metric which induces the relevant topology on Skorokhod space. Let Λ be the set of strictly increasing continuous bijections $[0, \infty) \to [0, \infty)$. The Skorokhod metric on $\mathsf{D}([0, \infty), \overline{D})$ and $\mathsf{D}([0, \infty), \mathbb{R}_{\geq 0})$ is given by

$$\sigma(f, g) = \inf_{\lambda \in \Lambda} \max\{||\lambda - \mathrm{id}||_\infty, ||f - g \circ \lambda||_\infty\},$$

where $|| \cdot ||_\infty$ denotes the sup-norm on the relevant space.

Lemma B.6.4. *Let $B \subset D$ be any ball and $\overline{T}_B^x$ the hitting time from x to B of Brownian motion with reflecting boundary condition in D. As a map $\mathsf{D}([0, \infty), \overline{D}) \to \mathbb{R}_{\geq 0}$, the hitting time $\overline{T}_B^x$ is continuous on the subset of $\mathsf{C}([0, \infty), \overline{D})$ of paths whose hitting time to B is finite.*

Proof. Denote by C_B the subset of $\mathsf{C}([0, \infty), \overline{D})$ of paths whose hitting time to B is finite. We first claim that the function

$$d_B : \mathsf{D}([0, \infty), \overline{D}) \to \mathsf{D}([0, \infty), \mathbb{R}_{\geq 0})$$

given by composition with the function $\overline{d}_B : \overline{D} \to B$ giving the distance to B is continuous. For any $\varepsilon > 0$, pick δ by uniform continuity of $\overline{d}_B$ so that $\delta < \varepsilon$ and if $|x - y| < \delta$, then $|\overline{d}_B(x) - \overline{d}_B(y)| < \varepsilon$. If $\sigma(f, g) < \delta$, we have

$$\sigma(d_B(f), d_B(g)) = \inf_{\lambda \in \Lambda} \max\{||\lambda - \mathrm{id}||_\infty, ||\overline{d}_B \circ f - \overline{d}_B \circ g \circ \lambda||_\infty\}.$$

Because $\sigma(f, g) < \delta$, we may find $\lambda \in \Lambda$ so that $||f - g \circ \lambda||_\infty < \delta$ and $||\lambda - \mathrm{id}||_\infty < \delta$.

By our choice of δ, this implies that

$$\max\{||\lambda - \mathrm{id}||_\infty, ||\overline{d}_B \circ f - \overline{d}_B \circ g \circ \lambda||_\infty\} < \varepsilon$$

and therefore that $\sigma(d_B(f), d_B(g)) < \varepsilon$, establishing continuity.

Now, the image of C_B under d_B is the subset C_0 of $\mathsf{C}([0,\infty), \mathbb{R}_{\geq 0})$ of paths whose hitting time to 0 is finite. By Whitt [1980, Theorem 7.1], the first passage time to 0 is continuous on C_0. The hitting time $\overline{T}^x_B$ is the composition of the first passage time and d_B, hence is continuous on C_B as claimed. □

Lemma B.6.5. *Let $B \subset D$ be any ball containing at least one point of G_n. For $x_i \in G_n$, let $T^{x_i}_{B,n}$ be the hitting time from x_i to B of the de-biased random walk on G. Then $g_n^2 \widehat{T}^{x_i}_{B,n} \xrightarrow{d} \overline{T}^x_B$.*

Proof. First, note that both the de-biased random walk and Brownian motion with reflecting boundary condition started at x_i have a.s. finite hitting time to B. By Lemma B.6.4, the hitting time to B is a.s. continuous on the subset of $\mathsf{D}([0,\infty), \overline{D})$ containing their trajectories. The desired convergence in distribution then follows from Corollary B.4.2, the continuous mapping theorem (see [Whitt, 1980, Section 1]), and noting the time-rescaling used in Corollary B.4.2. □

Proof of Corollary B.6.2. Recall that $T^{x_i}_{B(x_j,p),n}$ is the hitting time of the simple random walk on G_n to $B(x_j, p)$. By Lemma B.6.3, for any $\delta > 0$, we have with high probability that

$$T^{x_i}_{B(x_j,s+\delta),n} \leq \widehat{T}^{x_i}_{B(x_j,s),n} \leq T^{x_i}_{B(x_j,s-\delta),n}.$$

Applying Lemma B.6.5 to $B(x_j, s \pm \delta)$, we see that

$$g_n^2 T^{x_i}_{B(x_j,s\pm\delta),n} \xrightarrow{d} \overline{T}^{x_i}_{B(x_j,s\pm\delta)},$$

which shows that

$$\overline{T}^{x_i}_{B(x_j,s-\delta)} \leq \lim_{n\to\infty} g_n^2 \widehat{T}^{x_i}_{B(x_j,s),n} \leq \overline{T}^{x_i}_{B(x_j,s+\delta)}$$

for all $\delta > 0$. Sending $\delta \to 0$ yields the result. □

B.6.3 Proof of Theorem 3.3.4

We prove here Theorem 3.3.4. Recall we chose an estimator $\widehat{\varepsilon}(x) \to \overline{\varepsilon}(x)$.

Theorem B.6.6. *Let x_i and x_j be points in G_n connected by a geodesic not intersecting ∂D. For any $\delta > 0$, there exists a choice of $\widehat{\beta}$ and $s > 0$ so that if $\beta = \widehat{\beta} g_n^2$, we have for large n with high probability that*

$$\left| -\log(\mathbb{E}[\exp(-\beta \widehat{T}^{x_i}_{B(x_j,s),n})])/\sqrt{2\widehat{\beta}} - |x_i - x_j| \right| < \delta.$$

Our proof will proceed by converting to the continuous setting by Corollary B.6.2 and then reducing to the case of Brownian motion without boundary which was analyzed in Lemma B.6.1. Because we are in the setting of Brownian motion with reflecting boundary conditions, we must apply the "principle of not feeling the boundary" to show that our results are unaffected by it. For this, we define some events to condition on.

Let $\mathcal{G}$ be the geodesic from x_i to x_j, and for a distance scale ρ, let $\mathcal{G}(\rho)$ be the set of all points of distance less than ρ from $\mathcal{G}$. Choose ρ small enough so that $\mathcal{G}(\rho) \subset D$. For a distance $s > 0$, let B_t be a Brownian motion without boundary started at x_i, and let $\overline{T}^{x_i}_{B(x_j,s)}$ be its hitting time to $B(x_j, s)$. For a time $t^\star > 0$, define the following events:

- let E_1 be the event that $\overline{T}^{x_i}_{B(x_j,s)} < t^\star$;
- let E_2 be the event that E_1 holds and B_t hits $B(x_j, s)$ before $\mathcal{G}(\rho)$;
- let E_3 and E_4 denote the analogous events for Brownian motion with boundary.

Notice that $\mathbb{P}(E_2) = \mathbb{P}(E_4)$. In the rest of this section, we will consider the scalings $t^\star = s^\gamma$ and $\widehat{\beta} = s^\alpha$ for some $\gamma > 0$ and $\alpha < 0$ so that $\alpha + \gamma > 0$, so that $\widehat{\beta} t^\star \to \infty$ as $s \to 0$.

Let $p_t^R(x, y)$, $p_t^K(x, y)$, $p_t^G(x, y)$, and $p_t^F(x, y)$ be the transition density of Brownian motion started at x and run for time t with reflecting boundary condition, killed at ∂D, killed at $\partial G(\rho)$, and no boundary condition, respectively. For $\star \in \{R, K, G, F\}$, let $h^\star(T)$ be the probability that the respective process hits $B(x_j, s)$ before time T, and let $h^\star(t, x)$ be the density of hitting at $x \in B(x_j, s)$ at time t. Note that $p_t^K(x, y) \leq p_t^R(x, y)$, $p_t^K(x, y) \leq p_t^F(x, y)$, and $p_t^G(x, y) \leq p_t^F(x, y)$. We have the following three lemmas, which are instances of "the principle of not feeling the boundary."

Lemma B.6.7. *For x, y a distance at least $\rho' > 0$ to $\partial G(\rho)$, there are constants $t_0 > 0$ and $\lambda > 0$ dependent only on ρ so that for $t < t_0$, we have*

$$\frac{p_t^G(x, y)}{p_t^F(x, y)} \geq 1 - e^{-\lambda t^{-1}}.$$

Proof. This follows from Hsu [1995, Theorem 1.2] and the results of Varadhan [1967a]. □

Lemma B.6.8. *For x, y a distance at least $\rho' > 0$ to ∂D, there are constants $t_0 > 0$ and $\lambda > 0$ dependent only on ρ so that for $t < t_0$, we have*

$$\frac{p_t^K(x,y)}{p_t^F(x,y)} \geq 1 - e^{-\lambda t^{-1}}.$$

Proof. This follows from Hsu [1995, Theorem 1.2] and the results of Varadhan [1967a]. □

Lemma B.6.9. *For x, y a distance at least $\rho' > 0$ to ∂D, there are constants $t_0 > 0$ and $\lambda > 0$ dependent only on ρ so that for $t < t_0$, we have*

$$\frac{p_t^K(x,y)}{p_t^R(x,y)} \geq 1 - e^{-\lambda t^{-1}}.$$

Proof. Note that our domain D is a Lipschitz domain in the sense of Bass and Hsu [1991, Section 3]. Therefore, by Bass and Hsu [1991, Theorem 3.1, Theorem 3.4, and Remark 3.11], the reflecting Brownian motion in D has transition density $p_t^R(x,y)$ satisfying

$$C_1 t^{-d/2} e^{-c_1 \frac{|x-y|^2}{t}} \leq p_t^R(x,y) \leq C_2 t^{-d/2} e^{-c_2 \frac{|x-y|^2}{t}} \tag{B.11}$$

for constants c_1, c_2, C_1, C_2 and small enough t. This verifies the conditions of Hsu [1995, Theorem 1.2], yielding the conclusion. □

We now prove a general lemma on when the probability of hitting $B(x_j, s)$ before a time $t^\star$ is asymptotically equal for two processes.

Lemma B.6.10. *Let Q be a diffusion process with transition densities $p_t^Q(x,y)$, and let $p_t^K(x,y)$ of Q killed at some boundary. If for some $\lambda > 0$ and small enough s, we have for all $t < t^\star$ and $x, y \in B$ that*

$$1 \geq \frac{p_t^K(x_i,x)}{p_t^Q(x_i,x)} \geq 1 - e^{-\lambda t^{-1}} \text{ and } 1 \geq \frac{p_t^K(x,y)}{p_t^Q(x,y)} \geq 1 - e^{-\lambda t^{-1}},$$

then the probabilities $h^K(t^\star)$ and $h^Q(t^\star)$ that K and Q hit $B(x_j, s)$ before $t^\star$ are asymptotically equal.

Proof. Let $B = B(x_j, s)$, and consider s small enough so that $B(x_j, s) \subset D$. For $x \in B$ and $t > 0$, let $h^K(t,x)$ and $h^Q(t,x)$ be the densities of the first passage time to B, and let $h^K(T)$ and $h^Q(T)$ be the probabilities that the respective first passage times are at most T. Note that $h^K(t,x) \le h^Q(t,x)$. For $\star \in \{K,Q\}$, we have

$$h^\star(t,x) = p_t^\star(x_i,x) - \int_0^t \int_{y\in B} p_{t-\tau}^\star(y,x)h^\star(\tau,y)dyd\tau$$

so we may integrate to obtain

$$h^\star(T) = \int_0^T \int_{x\in B} p_t^\star(x_i,x)dtdx - \int_0^T \int_0^t \int_{x,y\in B} p_{t-\tau}^\star(y,x)h^\star(\tau,y)\, dydxd\tau dt. \quad \text{(B.12)}$$

Define the differences $d(T) := h^Q(T) - h^K(T)$, $d(t,x) = h^Q(t,x) - h^K(t,x)$, and $e_t(x,y) := p_t^Q(x,y) - p_t^K(x,y)$. By assumption, if $x = x_i$ or $x, y \in B$, we have

$$e_t(x,y) \le e^{-\lambda t^{-1}} p_t^Q(x,y).$$

Subtracting (B.12) for $\star \in \{K,Q\}$, we obtain

$$\begin{aligned}
d(T) &= \int_0^T \int_{x\in B} e_t(x_i,x)dtdx + \int_0^T \int_0^t \int_{x,y\in B} e_{t-\tau}(y,x)h^K(\tau,y)\, dydxd\tau dt \\
&\quad + \int_0^T \int_0^t \int_{x,y\in B} p_{t-\tau}^K(y,x)d(\tau,y)\, dydxd\tau dt \\
&\le \int_0^T \int_{x\in B} e^{-\lambda t^{-1}} p_t^Q(x_i,x)dtdx + \int_0^T \int_0^t \int_{x,y\in B} e^{-\lambda(t-\tau)^{-1}} p_{t-\tau}^Q(y,x)h^K(\tau,y)\, dydxd\tau dt \\
&\quad + \int_0^T \int_0^t \int_{y\in B} d(\tau,y)d\tau dtdy \\
&\le \int_0^T e^{-\lambda t^{-1}}dt + \int_0^T \int_0^t \int_{y\in B} e^{-\lambda(t-\tau)^{-1}} h^K(\tau,y)\, dyd\tau dt + \int_0^T d(\tau)d\tau \\
&\le 2\int_0^T e^{-\lambda t^{-1}}dt + \int_0^T d(\tau)d\tau \\
&\le 2Te^{-\lambda T^{-1}} + \int_0^T d(\tau)d\tau.
\end{aligned}$$

By Gronwall's inequality, this implies that

$$d(T) \le 2Te^{-\lambda T^{-1}} + 2\int_0^T \tau e^{-\lambda \tau^{-1}}(T-\tau)d\tau \le 2(T+T^3)e^{-\lambda T^{-1}}.$$

We conclude that

$$\lim_{s\to\infty} d(t^\star) = \lim_{s\to\infty} h^Q(t^\star) - h^K(t^\star) = 0. \qquad \square$$

Lemma B.6.11. *As $s \to 0$, we have $\mathbb{P}(E_2 \mid E_1) \to 1$.*

Proof. Let $B = B(x_j, s)$. By Lemma B.6.7 with ρ' small enough, we have for some $\lambda > 0$ and small enough s that for all $t < t^\star$ and $x, y \in B$ that

$$\frac{p_t^G(x_i, x)}{p_t^F(x_i, x)} \geq 1 - e^{-\lambda t^{-1}} \text{ and } \frac{p_t^G(x, y)}{p_t^F(x, y)} \geq 1 - e^{-\lambda t^{-1}}.$$

Notice that $\mathbb{P}(E_2)$ is the probability that the Brownian motion killed at $G(\rho)$ hits B before $t^\star$ and $\mathbb{P}(E_1)$ is the probability that the free Brownian motion hits B before $t^\star$. Therefore, Lemma B.6.10 implies that

$$\lim_{s\to 0} \mathbb{P}(E_1) = \lim_{s\to 0} \mathbb{P}(E_2),$$

from which we conclude that

$$\lim_{s\to 0} \mathbb{P}(E_2 \mid E_1) = \lim_{s\to 0} \frac{\mathbb{P}(E_2)}{\mathbb{P}(E_1)} = 1. \qquad \square$$

Lemma B.6.12. *As $s \to 0$, we have $\mathbb{P}(E_4 \mid E_3) \to 1$.*

Proof. Applying Lemma B.6.10 twice using Lemmas B.6.9 and B.6.8 implies that

$$\lim_{s\to\infty} \mathbb{P}(E_1) = \lim_{s\to\infty} h^F(t^\star) = \lim_{s\to\infty} h^K(t^\star) = \lim_{s\to\infty} h^R(t^\star) = \lim_{s\to\infty} \mathbb{P}(E_3).$$

We conclude from Lemma B.6.11 that

$$\lim_{s\to\infty} \mathbb{P}(E_4 \mid E_3) = \lim_{s\to\infty} \frac{\mathbb{P}(E_4)}{\mathbb{P}(E_3)} = \lim_{s\to\infty} \frac{\mathbb{P}(E_2)}{\mathbb{P}(E_1)} = \lim_{s\to\infty} \mathbb{P}(E_2 \mid E_1) = 1. \qquad \square$$

Proof of Theorem B.6.6. Throughout this proof, we will take $t^\star = s^\gamma$ and $\widehat{\beta} = s^\alpha$ for some fixed $\alpha < 0$ and $\gamma > 0$ so that $\alpha + \gamma > 0$. We will pick a small $s > 0$ at the end of the proof.

Bounding the effect of conditioning on E_4 on the process with boundary: By Corollary B.6.2, for any $\widehat{\beta}$ we have that

$$-\log(\mathbb{E}[\exp(-\widehat{\beta} g_n^2 \widehat{T}^{x_i}_{B(x_j,s),n})])/\sqrt{2\widehat{\beta}} \xrightarrow{d} -\log(\mathbb{E}[\exp(-\widehat{\beta\overline{T}}{}^{x_i}_{B(x_j,s)})])/\sqrt{2\widehat{\beta}}.$$

Conditioning on E_3 and E_4, we see that

$$\begin{aligned}\mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)})] &= \mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)}) \mid E_3^c]\,\mathbb{P}(E_3^c) \\ &\quad + \mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)}) \mid E_4]\,\mathbb{P}(E_4) \\ &\quad + \mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)}) \mid E_3 \cap E_4^c]\,\mathbb{P}(E_4^c \mid E_3)\,\mathbb{P}(E_3).\end{aligned}$$

By definition of E_3, we have $0 \leq \mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)}) \mid E_3^c]\,\mathbb{P}(E_3^c) \leq e^{-\widehat{\beta}t^\star}$. By the trivial bound $\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)}) \leq 1$, we find that

$$0 \leq \mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)}) \mid E_3 \cap E_4^c]\,\mathbb{P}(E_4^c \mid E_3)\,\mathbb{P}(E_3) \leq 1 - \mathbb{P}(E_4 \mid E_3).$$

By Lemma B.6.12, for any $\tau > 0$, for small enough $s > 0$ we have $e^{-\widehat{\beta}t^\star} < \tau$ and $1 - \mathbb{P}(E_4 \mid E_3) < \tau$. Noting also that $\mathbb{P}(E_4) = \mathbb{P}(E_2)$ and $\mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)}) \mid E_4] = \mathbb{E}[\exp(-\widehat{\beta}\widehat{T}^{x_i}_{B(x_j,s)}) \mid E_2]$, we conclude for small enough s that

$$\left|\mathbb{E}[\exp(-\widehat{\beta}\overline{T}^{x_i}_{B(x_j,s)})] - \mathbb{E}[\exp(-\widehat{\beta}\widehat{T}^{x_i}_{B(x_j,s)}) \mid E_2]\,\mathbb{P}(E_2)\right| < 2\tau. \tag{B.13}$$

Bounding the effect of conditioning on E_2 on the process without boundary: We now compare to the computations for Brownian motion without boundary. By conditioning on E_1 and E_2, we have that

$$\begin{aligned}\mathbb{E}[\exp(-\widehat{\beta}\widehat{T}^{x_i}_{B(x_j,s)})] &= \mathbb{E}[\exp(-\widehat{\beta}\widehat{T}^{x_i}_{B(x_j,s)}) \mid E_1^c]\,\mathbb{P}(E_1^c) \\ &\quad + \mathbb{E}[\exp(-\widehat{\beta}\widehat{T}^{x_i}_{B(x_j,s)}) \mid E_2]\,\mathbb{P}(E_2) \\ &\quad + \mathbb{E}[\exp(-\widehat{\beta}\widehat{T}^{x_i}_{B(x_j,s)}) \mid E_1 \cap E_2^c]\,(1 - \mathbb{P}(E_2 \mid E_1))\,\mathbb{P}(E_1).\end{aligned}$$

We again note that

$$0 \leq \mathbb{E}[\exp(-\widehat{\beta}\widehat{T}^{x_i}_{B(x_j,s)}) \mid E_1^c]\,\mathbb{P}(E_1^c) \leq e^{-\widehat{\beta}t^\star}$$

and

$$0 \leq \mathbb{E}[\exp(-\widehat{\beta}\widehat{T}^{x_i}_{B(x_j,s)}) \mid E_1 \cap E_2^c]\,(1 - \mathbb{P}(E_2 \mid E_1))\,\mathbb{P}(E_1) \leq 1 - \mathbb{P}(E_2 \mid E_1).$$

These together with Lemma B.6.11 imply that for any $\tau > 0$, for small enough $s > 0$ we have that $e^{-\widehat{\beta}t^\star} < \tau$ and $1 - \mathbb{P}(E_2 \mid E_1) < \tau$. We conclude for small enough s that

$$\left|\mathbb{E}[\exp(-\widehat{\beta}\widehat{T}^{x_i}_{B(x_j,s)})] - \mathbb{E}[\exp(-\widehat{\beta}\widehat{T}^{x_i}_{B(x_j,s)}) \mid E_2]\,\mathbb{P}(E_2)\right| < 2\tau. \tag{B.14}$$

Combining (B.13) and (B.14), we conclude for small enough s that

$$\left|\mathbb{E}[\exp(-\widehat{\beta\overline{T}}{}^{x_i}_{B(x_j,s)})] - \mathbb{E}[\exp(-\widehat{\beta}\widehat{T}^{x_i}_{B(x_j,s)})]\right| < 4\tau. \tag{B.15}$$

Aggregating the estimates: To conclude, for any $\delta > 0$ and $\widehat{\beta}_0 > 0$, choose $\tau > 0$ small enough so that if $|x-y| < 4\tau$, then for all $\widehat{\beta} > \widehat{\beta}_0$, we have

$$\left|\log(x)/\sqrt{2\widehat{\beta}} - \log(y)/\sqrt{2\widehat{\beta}}\right| < \delta/3.$$

Now, choose $s > 0$ small enough and n large enough so that $\widehat{\beta} > \widehat{\beta}_0$, and for this τ, we have:

- by our previous discussion, (B.15) holds;

- by Lemma B.6.1, we have

$$\left|-\log(\mathbb{E}[\exp(-\widehat{\beta}\widehat{T}^{x_i}_{B(x_j,s)})])/\sqrt{2\widehat{\beta}} - |x_i - x_j|\right| < \delta/3;$$

- by Corollary B.6.2, we have

$$\left|-\log(\mathbb{E}[\exp(-\widehat{\beta\overline{T}}{}^{x_i}_{B(x_j,s)})])/\sqrt{2\widehat{\beta}} + \log(\mathbb{E}[\exp(-\widehat{\beta} g_n^2 \widehat{T}^{x_i}_{B(x_j,s),n})])/\sqrt{2\widehat{\beta}}\right| < \delta/3.$$

For these choices of τ, s, and n, we have by (B.15) that

$$\left|\log(\mathbb{E}[\exp(-\widehat{\beta\overline{T}}{}^{x_i}_{B(x_j,s)})])/\sqrt{2\widehat{\beta}} - \log(\mathbb{E}[\exp(-\widehat{\beta}\widehat{T}^{x_i}_{B(x_j,s)})])/\sqrt{2\widehat{\beta}}\right| < \delta/3.$$

Combining the last three inequalities yields the desired

$$\left|\log(\mathbb{E}[\exp(-\widehat{\beta} g_n^2 \overline{T}^{x_i}_{B(x_j,s),n})])/\sqrt{2\widehat{\beta}} - |x_i - x_j|\right| < \delta. \qquad \square$$

B.7 1-D bias calculation

We repeat the full theorem statement and proof for the bias characterization.

Theorem B.7.1. *Let $T_{x_j}^{x_i}$ be the hitting time to x_j of a 1-dimensional Itô process with drift $\mu(x) = \frac{\partial \log(p(x))}{\partial x}\overline{\varepsilon}^2(x)$ and diffusion $\overline{\varepsilon}^2(x)$ started at x_i with reflecting boundary γ for $\gamma < x_i < x_j$. The Laplace transform of $T_{x_j}^{x_i}$ admits the asymptotic expansion*

$$\mathbb{E}[-\exp(\beta T_{x_j}^{x_i})] = \frac{c_1}{f(x_i)^{1/4}p(x_i)}\exp\left(-\sqrt{\beta}\int_{x_i}^{x_j}\sqrt{f(s)}ds\right)$$
$$\left(1+\left(1+o\left(\frac{1}{\sqrt{\beta}}\right)\right)\exp\left(-2\sqrt{\beta}\int_{\gamma}^{x_i}\sqrt{f(x)}dx\right)+o(\exp(-\beta))\right),$$

where $f(x) = \frac{2}{\overline{\varepsilon}(x)^2} + \frac{1}{\beta}\frac{\partial \log(p(x))}{\partial x^2} + \frac{1}{\beta}\left(\frac{\partial \log(p(x))}{\partial x}\right)^2$, and c_1 is a normalization constant depending on p, $\overline{\varepsilon}$,and j to make $E[-\beta T_{x_j}^{x_i}] = 1$.

Proof. Let $\mathbb{E}[\exp(-\beta T_{x_j}^{x_i})] = u(x_i)$, where $u(x)$ is the hitting time to x_j from point x. By Feynman-Kac, this is

$$\frac{\partial^2 u}{\partial x^2} + 2\frac{\partial \log(p(x))}{\partial x}\frac{\partial u}{\partial x} + q(x)u = 0,$$

where $q(x) = -2\beta\overline{\varepsilon}(x)^{-2}$. Rewrite this as a perturbation of a second order ODE via the change of variables to obtain

$$y(x) = u(x)\exp\left(\int_{\gamma}^{x}\frac{\partial \log(p(y))}{\partial y}dy\right) = u(x)p(x)p(\gamma)^{-1}$$
$$f(x) = \frac{2}{\overline{\varepsilon}(x)^2} + \frac{1}{\beta}\left(\frac{\partial \log(p(x))}{\partial x^2} + \left(\frac{\partial \log(p(x))}{\partial x}\right)^2\right)$$
$$\frac{1}{\beta}\frac{\partial^2 y}{\partial x} = f(x)y(x).$$

Since this is a type of Schrodinger's equation with $f(x) \neq 0$ everywhere we can apply the WKBJ asymptotic expansion [Bender and Orszag, 1999, section 10.1] to obtain

$$y(x) = \frac{c_1}{f(x)^{1/4}}\exp\left(-\sqrt{\beta}\int_{x_0}^{x}\sqrt{f(s)}ds\right)+\frac{c_2}{f(x)^{1/4}}\exp\left(\sqrt{\beta}\int_{x_0}^{x}\sqrt{f(s)}ds\right)+o(\exp(-\beta)).$$

Since we assumed $x_i < x_j$ and by the boundary condition $u(x_j) = 1$ we have

$$u(x) = \frac{c_2 p(\gamma)}{f(x)^{1/4} p(x)} \exp\left(-\sqrt{\beta}\int_x^{x_j} \sqrt{f(s)}ds\right) + \frac{c_1 p(\gamma)}{f(x)^{1/4} p(x)} \exp\left(\sqrt{\beta}\int_x^{x_j} \sqrt{f(s)}ds\right) + o(\exp(-\beta$$

To obtain the boundary conditions, note that $u'(\gamma) = 0$. Taking the derivative for $y(x)p(x)$, setting to zero and solving for c_2 results in

$$c_2 = c_1 \frac{\exp(-2\sqrt{\beta}\int_\gamma^{x_j} \sqrt{f(s)}ds)(p(\gamma)4\sqrt{\beta}f(\gamma)^{3/2} + f'(\gamma)) - f(\gamma)p'(\gamma)}{4\sqrt{\beta}f(\gamma)^{3/2}p(\gamma) - p(\gamma)f'(\gamma) + 4f(\gamma)p'(\gamma)} + o(\exp(-\beta)),$$

from which we obtain

$$c_2 = c_1 \exp\left(-2\sqrt{\beta}\int_\gamma^{x_j} \sqrt{f(s)}ds\left(1 + o\left(\sqrt{\frac{1}{\beta}}\right)\right)\right).$$

Pulling out the $-\sqrt{\beta}$ term, we get

$$u(x_i) = \mathbb{E}[\exp(-\beta T_{x_j}^{x_i})] = \frac{c_1 p(\gamma)}{f(x_i)^{1/4} p(x_i)} \exp\left(-\sqrt{\beta}\int_{x_i}^{x_j} \sqrt{f(s)}ds\right)$$
$$\left(1 + \left(1 + o\left(\frac{1}{\sqrt{\beta}}\right)\right) \exp\left(-2\sqrt{\beta}\int_\gamma^{x_i} \sqrt{f(x)}dx\right) + o(\exp(-\beta))\right). \quad \square$$

We now connect this statement to the discrete walk.

Corollary B.7.2. *Let $T_{B(x_j,s),n}^{x_i}$ be the discrete hitting time to a s ball around x_j where s is selected as given in Theorem B.2.13. Then the simple random walk over a graph constructed on density $p(x)$ and scale $\overline{\varepsilon}(x)$ has the following log-LTHT under the boundary conditions of Theorem 3.3.5*

$$-\log(\mathbb{E}[\exp(-\beta T_{B(x_j,s),n}^{x_i} g_n^2)])/\sqrt{2\beta} \to \int_{x_i}^{x_j} \sqrt{\frac{1}{\overline{\varepsilon}(x)} + \frac{1}{\beta}\left(\frac{\partial \log(p(x))}{\partial x^2} + \left(\frac{\partial \log(p(x))}{\partial x}\right)^2\right)}dx$$
$$+ \frac{\log(p(x_i)/p(x_j)) + \log(f(x_i)/f(x_j))/4}{\sqrt{2\beta}} + o(\log(1 + e^{-\sqrt{2\beta}})/\sqrt{2\beta}).$$

Proof. Taking the logarithm of the result of Theorem B.7.1 and noting the initial condition $u(x_j) = 1$ implies that asymptotically we have

$$c_1 \propto \left(\frac{1}{f(x_j)^{1/4}p(x_j)}(1 + o(e^{-2\sqrt{\beta}}))\right)^{-1} \to f(x_j)^{1/4}p(x_j),$$

which completes the continuous statement. The convergence of the hitting time to its discrete counterpart follows from Theorem B.2.13. □

B.8 Basic noise resistance

We give details for the basic noise bound from the main text footnote. Our goal is to prove the following statement about random walks.

Theorem B.8.1. *Let G_n be generated by the noise model of Definition 4 with $\sum_j q_j = o(g_n^2)$. Then the simple random walk over G_n converges to the same limit as the noiseless case in Theorem C.1.1.*

Proof. Since the boundaries of both noisy and noiseless graphs are identical, we need only verify the moment conditions in the proof of Theorem C.1.1. In particular we require that under any noise q, we have

$$\lim_{n\to\infty} g_n^{-2}\mathbb{E}[X_{t+1}^n - X_t^n | X_t^n] = \nabla \log(p(X_t^n))\bar{\varepsilon}(X_t^n)^2$$
$$\lim_{n\to\infty} g_n^{-2}\mathsf{Cov}[X_{t+1}^n | X_t^n] = \bar{\varepsilon}(X_t^n)^2 \cdot I_n$$
$$\lim_{n\to\infty} g_n^{-2}\mathbb{E}[|X_{t+1}^n - X_t^n|^{2+\alpha} \mid X_t^n] = 0,$$

which we show in the Lemma B.8.2 and Lemma B.8.3 below. By the Stroock-Varadhan criterion, this implies convergence to Theorem C.1.1, as well as any macroscopic quantities such as hitting times, or LTHTs with $\beta = \Theta(g_n^2)$. □

We now prove the moment bounds required for convergence of the noisy graph.

Lemma B.8.2 (Noisy moments). *If the noisy graph G_n is generated by the noise model of Definition 4, for any choice of latent noise parameters q_j such that $\sum_j q_j = o(g_n^2)$ then we have for $\alpha > 0$ that*

$$\lim_{n\to\infty} g_n^{-2}\mathbb{E}[X_{t+1}^n - X_t^n | X_t^n] = \nabla \log(p(X_t^n))\bar{\varepsilon}(X_t^n)^2$$
$$\lim_{n\to\infty} g_n^{-2}\mathsf{Cov}[X_{t+1}^n | X_t^n] = \bar{\varepsilon}(X_t^n)^2 \cdot I_n$$
$$\lim_{n\to\infty} g_n^{-2}\mathbb{E}[|X_{t+1}^n - X_t^n|^{2+\alpha} \mid X_t^n] = 0.$$

Proof. Let $\overline{X}$ denote quantities in the noise-free graph. We recall from Theorem 2.3.3 that

$$\lim_{n\to\infty} g_n^{-2}\mathbb{E}[\overline{X}_{t+1}^n - \overline{X}_t^n | \overline{X}_t^n = x] = \nabla \log(p(x))\overline{\varepsilon}(x)^2$$
$$\lim_{n\to\infty} g_n^{-2}\mathsf{Cov}[\overline{X}_{t+1}^n | \overline{X}_t^n = x] = \overline{\varepsilon}(x)^2 \cdot I_n$$
$$\lim_{n\to\infty} g_n^{-2}\mathbb{E}[|\overline{X}_{t+1}^n - \overline{X}_t^n|^{2+\alpha} \mid \overline{X}_t^n = x] = 0.$$

Let $\widehat{q} = \sum_i q_i$ so that $\widehat{q} = o(g_n^2)$. In the noisy graph, we first check the expectation via

$$\begin{aligned}\lim_{n\to\infty} g_n^{-2}\mathbb{E}[X_{t+1}^n - X_t^n \mid X_t^n = x] &= \lim_{n\to\infty} (1-\widehat{q}) g_n^{-2}\mathbb{E}[\overline{X}_{t+1}^n - \overline{X}_t^n \mid \overline{X}_t^n = x] + g_n^{-2}\sum_i q_i(x_i - x)\\ &= \lim_{n\to\infty} g_n^{-2}\mathbb{E}[\overline{X}_{t+1}^n - \overline{X}_t^n | \overline{X}_t^n = x]\\ &= \nabla \log(p(x))\overline{\varepsilon}(x)^2.\end{aligned}$$

The covariance follows because for all indices i and j we have

$$\begin{aligned}&\lim_{n\to\infty} g_n^{-2}\mathbb{E}[(X_{t+1}^n - X_t^n)_i(X_{t+1}^n - X_t^n)_j \mid X_t^n = x]\\ &= \lim_{n\to\infty} (1-\widehat{q}) g_n^{-2}\mathbb{E}[(X_{t+1}^n - X_t^n)_i(X_{t+1}^n - X_t^n)_j \mid \overline{X}_t^n = x] + g_n^{-2}\sum_k q_k(x_k - x)_i(x_k - x)_j\\ &= \lim_{n\to\infty} g_n^{-2}\mathbb{E}[(\overline{X}_{t+1}^n - \overline{X}_t^n)_i(\overline{X}_{t+1}^n - \overline{X}_t^n)_j | \overline{X}_t^n = x]\\ &= \delta_{ij}\overline{\varepsilon}(x)^2.\end{aligned}$$

Finally, the higher moments follow because we have

$$\begin{aligned}&\lim_{n\to\infty} g_n^{-2}\mathbb{E}[|X_{t+1}^n - X_t^n|^{2+\alpha} \mid X_t^n = x]\\ &= \lim_{n\to\infty} g_n^{-2}(1-\widehat{q}) g_n^{-2}\mathbb{E}[|\overline{X}_{t+1}^n - \overline{X}_t^n|^{2+\alpha} \mid \overline{X}_t^n = x] + g_n^{-2}\sum_i q_i|x_i - x|^{2+\alpha}\\ &= \lim_{n\to\infty} g_n^{-2}(1-\widehat{q}) g_n^{-2}\mathbb{E}[|\overline{X}_{t+1}^n - \overline{X}_t^n|^{2+\alpha} \mid \overline{X}_t^n = x]\\ &= 0,\end{aligned}$$

where we use that $|x_i - x|^{2+\alpha} = O(1)$. □

Lemma B.8.3 (Strong LLN for noisy moments). *For a function $f(x)$ such that $\sup_{x\in B(0,\varepsilon)} |f(x)| < \varepsilon$ and $\sup_{x\in D} |f(x)| < C$ for some constant C, given $(\star)$ we have* *uniformly in $x \in \mathcal{X}_n$ that*

$$g_n^{-2} \sum_{y \in \mathsf{NB}_n(x)} \frac{1}{|\mathsf{NB}_n(x)|} f(y-x) \overset{a.s.}{\to} g_n^{-2} \int_{y \in B(x, \varepsilon_n(x))} f(y-x) \frac{p(y)}{p_{\varepsilon_n(x)}(x)} dy.$$

Proof. Denote the claimed value of the limit by $\mu(x)$. Let the set of non-noise out-neighbors of x be $\overline{\mathsf{NB}}_n(x)$ and the set of noise out-neighbors of x be $\widetilde{\mathsf{NB}}_n(x)$, where we consider noise edges to be strictly non-geometric edges. We have uniformly in $x \in cX$ that

$$g_n^{-2} \frac{\sum_{y \in \overline{\mathsf{NB}}_n(x)} f(y-x) + \sum_{y \in \widetilde{\mathsf{NB}}_n(x)} f(y-x)}{|\overline{\mathsf{NB}}_n(x)| + |\widetilde{\mathsf{NB}}_n(x)|} = g_n^{-2} \frac{\sum_{y \in \overline{\mathsf{NB}}_n(x)} f(y-x) + o(Cg_n^2)}{|\overline{\mathsf{NB}}_n(x)| + o(g_n^2)}$$

$$\overset{\text{a.s.}}{\to} g_n^{-2} \sum_{y \in \overline{\mathsf{NB}}_n(x)} \frac{f(y-x)}{|\overline{\mathsf{NB}}_n(x)|},$$

so the result follows by the noise-less result in Lemma B.4.4. □

Now the behavior of noisy hitting times can be recovered by combining Lemma B.8.1 with the convergence result of Corollary B.6.2.

Theorem B.8.4. *Let G_n be a noisy geometric graph with noise $\sum_j q_j = o(g_n^2)$. For any δ, there exists some $\beta = \widehat{\beta} g_n^2$, s, c such that*

$$\left| -\frac{\log(\mathbb{E}[\exp(-\beta T^{x_i}_{B(x_j,s),n})])}{\sqrt{2\beta}} g_n - c|x_i - x_j| \right| \leq \delta$$

with high probability as $n \to \infty$.

Proof. By Lemma B.8.1, the noisy and noise-free walks converge to the same continuum limit, and this guarantees that by Corollary B.6.2 that their hitting times converge in distribution. Applying Theorem 3.3.4 gives the desired result. □

This is a basic, but useful result for robustness of hitting times. Up to $o(1)$ noise edges can be allowed for each vertex without disrupting the global convergence of hitting times.

B.9 Resource allocation index

Recall that the directed RA index was defined by

$$R_{ij} = \sum_{x_k \in \mathsf{NB}_n(x_i) \cap \mathsf{NB}_n^{\mathsf{in}}(x_j)} \frac{1}{|\mathsf{NB}_n(x_k)|}$$

and the modified log-LTHT was defined by

$$M_{ij}^{\text{mod}} = -\log(\mathbb{E}[\exp(-\beta T_{x_j,n}^{x_i}) \mid T_{x_j,n}^{x_i} > 1]).$$

B.9.1 RA index reduction

Theorem B.9.1. *If $\beta = \omega(\log(g_n^d n))$ and x_i and x_j have at least one common neighbor, then*

$$M_{ij}^{mod} - 2\beta \to -\log(R_{ij}) + \log(|\mathsf{NB}_n(x_i)|).$$

Proof. Let $P_{ij}(t)$ be the probability of going from x_i to x_j in t steps, and $H_{ij}(t)$ the probability of not hitting before time t. Factoring the two-step hitting time yields

$$M_{ij}^{\text{mod}} = 2\beta - \log(P_{ij}(2)) - \log\Big(1 + \sum_{t=3}^{\infty} \frac{P_{ij}(t)}{P_{ij}(2)} H_{ij}(t) e^{-\beta(t-2)}\Big).$$

Let $k_{\max}$ be the maximal out-degree which occurs in G_n. By assumption, at least one of the at most $k_{\max}^2$ two-step paths from x_i goes to x_j, we have the bound $\frac{P_{ij}(t)}{P_{ij}(2)} \le k_{\max}^2$. For $\beta = \omega(\log(g_n^d n))$, we see that $\beta = \omega(2\log(k_{\max}))$ with high probability. Applying the bounds $H_{ij}(t) \le 1$ and $\frac{P_{ij}(t)}{P_{ij}(2)} \le k_{\max}^2$, we obtain

$$\sum_{t=3}^{\infty} \frac{P_{ij}(t)}{P_{ij}(2)} H_{ij}(t) e^{-\beta(t-2)} \le \frac{k_{\max}^2}{e^{\beta} - 1} = o(k_{\max}^{-1}).$$

We conclude that $M_{ij}^{\text{mod}} \to 2\beta - \log(P_{ij}(2))$. It remains to verify that $\log(P_{ij}(2))$ is related to the resource allocation index by

$$\log(P_{ij}(2)) = \log\left(\frac{1}{|\mathsf{NB}_n(x_i)|} \sum_{k \in \mathsf{NB}_n(x_i) \cap \mathsf{NB}_n^{in}(x_j)} \frac{1}{\mathsf{NB}_n(x_k)}\right) = \log(R_{ij}) - \log(|\mathsf{NB}_n(x_i)|).$$

□

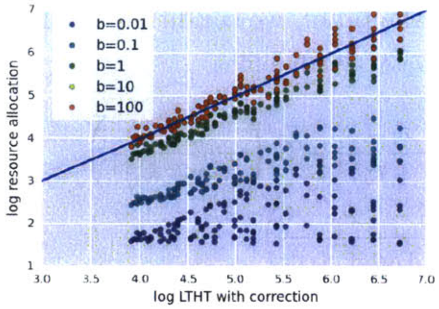


Figure B-6: Modified LTHT rapidly converges to RA index.

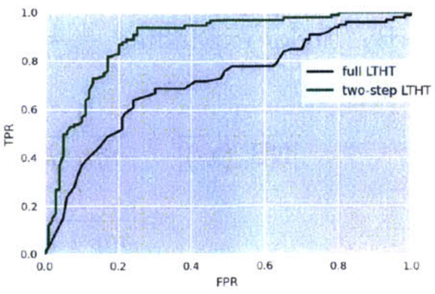


Figure B-7: Conditioning on $t > 1$ substantially outperforms naive LTHT.

B.9.2 RA index robustness

We verify the robustness of the RA index by directly bounding the statistics involved.

Theorem B.9.2. *If $q_i = q = o(g_n^{d/2})$ for all i, then for any $\delta > 0$ there exist cutoffs c_1, c_2 and scaling h_n so that with probability at least $1 - \delta$, for any i, j we have*

- $|x_i - x_j| < \min\{\varepsilon_n(x_i), \varepsilon_n(x_j)\}$ *if* $R_{ij}h_n < c_1$*;*
- $|x_i - x_j| > 2\max\{\varepsilon_n(x_i), \varepsilon_n(x_j)\}$ *if* $R_{ij}h_n > c_2$*.*

Proof. Decompose the out-degree of x_i into expectation and noise terms by

$$|\mathsf{NB}_n(x_i)| = nq + k_i + z_i,$$

where $k_i = \varepsilon_n(x_i)^d p(x_i) V_d n$, V_d is the volume of the d-unit ball, and z_i is a random variable giving the remaining error. The number of noise edges has a binomial distribution with n draws and success probability q, and the number of geometric edges has a Poisson distribution with rate k_i. Therefore, the Chebyshev inequality implies

$$\mathbb{P}(|z_i| > c) \leq \frac{k_i + nq(1-q)}{c^2} < \frac{k_i + nq}{c^2}. \tag{B.16}$$

Let $\delta_1 = \delta/4$ and define c by the equality

$$\delta_1 = \frac{k_i + nq(1-q)}{c^2} \tag{B.17}$$

so that $c = \delta_1^{-1/2}\sqrt{k_i + nq}$. For the rest of the proof, we condition on the event that $|z_i| < c$. By Taylor expanding $\frac{1}{|\mathsf{NB}_n(x_i)|}$ in z_i, we have that

$$\begin{aligned}\frac{1}{|\mathsf{NB}_n(x_i)|} &= \frac{1}{nq+k_i} - \frac{z_i}{(nq+k_i)^2} + O\left(\frac{z_i^2}{(nq+k_i)^3}\right)\\ &= \frac{1}{nq+k_i} - O\left(\frac{c}{(nq+k_i)^2}\right).\end{aligned} \tag{B.18}$$

By (B.17), we see that $|z_i| < c$ with probability at least $1 - \delta_1$, which implies that

$$\frac{c}{(nq+k_i)^2} < \delta_1^{-1/2}(nq+k_i)^{-3/2}.$$

By the definition of the RA index, we obtain

$$R_{ij} = \sum_{x_k \in \mathsf{NB}_n(x_i) \cap \mathsf{NB}_n^{\mathsf{in}}(x_j)} \left(\frac{1}{nq+k_i} + O(\delta_1^{-1/2}(nq+k_i)^{-3/2})\right).$$

Since our domain is compact, we may define

$$k_n^+ = \sup_x \varepsilon_n(x)^d p(x) V_d n \qquad \text{and} \qquad k_n^- = \inf_x \varepsilon_n(x)^d p(x) V_d n.$$

By construction, $k_n^+ > k_i > k_n^-$ for all i. Let $C_{ij} := |\mathsf{NB}_n(x_i) \cap \mathsf{NB}_n^{in}(x_j)|$. Then we have

$$\frac{C_{ij}}{nq+k_n^+} - O\left(\frac{C_{ij}}{\delta_1^{1/2}(nq+k_n^+)^{3/2}}\right) \leq R_{ij} \leq \frac{C_{ij}}{nq+k_n^-} + O\left(\frac{C_{ij}}{\delta_1^{1/2}(nq+k_n^-)^{3/2}}\right). \tag{B.19}$$

Choose the scaling

$$h_n = \frac{nq+k_n^+}{k_n^+}.$$

We will now bound C_{ij} to control $h_n R_{ij}$. To do this, decompose C_{ij} as

$$C_{ij} = C_{ij}^g + C_{ij}^{n1} + C_{ij}^{n2},$$

where C_{ij}^g, C_{ij}^{n1}, and C_{ij}^{n2} are defined as follows.

1. Geometric edges (C_{ij}^g): If $|x_i - x_j| < \min\{\varepsilon_n(x_i), \varepsilon_n(x_j)\}$ then they share common neighbors due to the geometric graph. Specifically their number of com-

mon neighbors has Poisson distribution with mean at least $\tau_d k_i(1-q)$, where τ_d is a constant independent of n defined as the overlapping density of two kernels at a unit distance.

2. One noise edge (C_{ij}^{n1}): The edge $x_i \to x_k$ occurs by noise but $x_k \to x_j$ is geometric. There are at most k_n^+ such vertices with in-edges to x_j and so this is at most a binomial random variable with k_n^+ draws and success probability q.

3. Two noise edges (C_{ij}^{n2}): Both $x_i \to x_k$ and $x_k \to x_j$ may occur by noise, this is at most a binomial random variable with $n - k_n^-$ draws and success probability q^2.

The case of $|x_i - x_j| < \min\{\varepsilon_n(x_i), \varepsilon_n(x_j)\}$: All types of edges may occur, so we obtain the moment bounds

$$\begin{aligned}
\mathbb{E}\left[\frac{C_{ij}}{k_n^+}\right] &\geq \tau_d(1-q)\frac{k_n^-}{k_n^+} \\
\mathsf{Var}\left[\frac{C_{ij}}{k_n^+}\right] &\leq \frac{\tau_d(1-q)k_n^-}{(k_n^+)^2} + \frac{(n-k_n^-)q^2(1-q^2)+k_n^+q(1-q)}{(k_n^+)^2} < \frac{\tau_d(1-q)k_n^-}{(k_n^+)^2} + \frac{nq^2+k_n^+q}{(k_n^+)^2}.
\end{aligned}$$

Notice that $\frac{k_n^+}{k_n^-}$ is bounded between the minimum and maximum of $\frac{\varepsilon_n(x)p(x)}{\varepsilon_n(y)p(y)}$ for $x, y \in D$, so $\lim_{n\to\infty} \mathbb{E}\left[\frac{C_{ij}}{k_n^+}\right] \geq c_{ij}$ for some $c_{ij} > 0$. Further, we find that $\mathsf{Var}\left[\frac{C_{ij}}{k_n^+}\right] \to 0$. These imply that for large enough n, we have $\frac{C_{ij}}{k_n^+} > c_{ij}$ with probability at least $1-\delta_1$. Therefore, if $|x_i - x_j| < \min\{\varepsilon_n(x_i), \varepsilon_n(x_j)\}$, we have

$$\lim_{n\to\infty} h_n R_{ij} \geq \lim_{n\to\infty} \frac{C_{ij}}{k_n^+} - O\left(\frac{C_{ij}}{\delta^{1/2}(nq+k_n^+)^{1/2}k_n^+}\right) \geq c_{ij}. \tag{B.20}$$

The case of $|x_i - x_j| > 2\max\{\varepsilon_n(x_i), \varepsilon_n(x_j)\}$: Only noise cases occur, hence we have the moment bounds

$$\begin{aligned}
\mathbb{E}\left[\frac{C_{ij}}{k_n^-}\right] &\leq \frac{q^2(n-k_n^-)+qk_n^+}{k_n^-} < \frac{q^2 n}{k_n^-} + \frac{qk_n^+}{k_n^-} \\
\mathsf{Var}\left[\frac{C_{ij}}{k_n^-}\right] &\leq \frac{(n-k_n^-)q^2(1-q^2)+k_n^+q(1-q)}{(k_n^-)^2} < \frac{qn}{(k_n^-)^2} + \frac{qk_n^+}{(k_n^-)^2}.
\end{aligned}$$

Because $q = o(g_n^{d/2})$, both the expectation and variance converge to zero and for large enough n, we have $C_{ij}/k_n^- \to 0$ probability at least $1-\delta_1$. Therefore, if

$|x_i - x_j| > 2\max\{\varepsilon_n(x_i), \varepsilon_n(x_j)\}$, for we have

$$\lim_{n\to\infty} h_n R_{ij} \leq \lim_{n\to\infty} \frac{C_{ij}(nq + k_n^+)}{k_n^+(nq + k_n^-)} + O\left(\frac{C_{ij}(nq + k_n^+)}{\delta^{1/2} k_n^+(nq + k_n^-)^{3/2}}\right) \to 0. \tag{B.21}$$

Combining the cases: Taking $h_n = \frac{nq+k_n^+}{k_n^+}$, we combine (B.20) and (B.21) to conclude that the desired holds with probability at least $1 - 3\delta_1 > 1 - \delta$ for any $c_1 \leq c_{ij}$ and $c_2 > 0$. □

Appendix C

Word embeddings

C.1 Metric recovery from Markov processes on graphs and manifolds

Consider an infinite sequence of points $\mathcal{X}_n = \{x_1, \dots, x_n\}$, where x_i are sampled i.i.d. from a density $p(x)$ over a compact Riemannian manifold equipped with a geodesic metric ρ. For our purposes, $p(x)$ should have a bounded log-gradient and a strict lower bound p_0 over the manifold. The random walks we consider are over *unweighted spatial graphs* defined as

Definition 9 (Spatial graph). Let $\sigma_n : \mathcal{X}_n \to \mathbb{R}_{>0}$ be a local scale function and $h : \mathbb{R}_{\geq 0} \to [0, 1]$ a piecewise continuous function with sub-Gaussian tails. A *spatial graph* G_n corresponding to σ_n and h is a random graph with vertex set $\mathcal{X}_n$ and a directed edge from x_i to x_j with probability $p_{ij} = h(\rho(x_i, x_j)^2 / \sigma_n(x_i)^2)$.

Simple examples of spatial graphs where the connectivity is not random include the ε ball graph ($\sigma_n(x) = \varepsilon$) and the k-nearest neighbor graph ($\sigma_n(x)$ =distance to k-th neighbor).

Log co-occurrences and the geodesic will be connected in two steps. (1) we use known results to show that a simple random walk over the spatial graph, properly scaled, behaves similarly to a diffusion process; (2) the log-transition probability of a diffusion process will be related to the geodesic metric on a manifold.

(1) The limiting random walk on a graph: Just as the simple random walk over the integers converges to a Brownian motion, we may expect that under specific constraints the simple random walk X_t^n over the graph G_n will converge to some well-defined continuous process. We require that the scale functions converge to a continuous function $\bar{\sigma}$ ($\sigma_n(x) g_n^{-1} \xrightarrow{a.s.} \bar{\sigma}(x)$); the size of a single step vanish ($g_n \to 0$) but contain at least a polynomial number of points within $\sigma_n(x)$ ($g_n n^{\frac{1}{d+2}} \log(n)^{-\frac{1}{d+2}} \to \infty$). Under this limit, our assumptions about the density $p(x)$, and regularity of the transitions, [1]

Theorem C.1.1 ([Hashimoto et al., 2015c, Ting et al., 2011]). *The simple random walk X^n_t on G_n converges in Skorokhod space $\mathsf{D}([0,\infty),\overline{D})$ after a time scaling $\widehat{t} = tg_n^2$ to the Itô process $Y_{\widehat{t}}$ valued in $\mathsf{C}([0,\infty),\overline{D})$ as $X^n_{\widehat{t}g_n^{-2}} \to Y_{\widehat{t}}$. The process $Y_{\widehat{t}}$ is defined over the normal coordinates of the manifold (D,g) with reflecting boundary conditions on D as*

$$dY_{\widehat{t}} = \nabla \log(p(Y_{\widehat{t}}))\overline{\sigma}(Y_{\widehat{t}})^2 d\widehat{t} + \overline{\sigma}(Y_{\widehat{t}})dW_{\widehat{t}} \tag{C.1}$$

The equicontinuity constraint on the marginal densities of the random walk implies that the transition density for the random walk converges to its continuum limit.

Lemma C.1.2 (Convergence of marginal densities). *[Hashimoto et al., 2015b] Let x_0 be some point in our domain $\mathcal{X}_n$ and define the marginal densities $\widehat{q}_t(x) = \mathbb{P}(Y_t = x|Y_0 = x_0)$ and $q_{t_n}(x) = \mathbb{P}(X^n_t = x|X^n_0 = x_0)$. If $t_n g_n^2 = \widehat{t} = \Theta(1)$, then under condition ($\star$) and the results of Theorem C.1.1 such that $X^n_t \to Y^n_t$ weakly, we have*

$$\lim_{n\to\infty} nq_{t_n}(x) = \widehat{q}_{\widehat{t}}(x)p(x)^{-1}.$$

(2) Log transition probability as a metric We may now use the stochastic process $Y_{\widehat{t}}$ to connect the log transition probability to the geodesic distance using Varadhan's large deviation formula.

Theorem C.1.3 ([Varadhan, 1967a, Molchanov, 1975]). *Let Y_t be a Itô process defined over a complete Riemann manifold (D,g) with geodesic distance $\rho(x_i,x_j)$ then*

$$\lim_{t\to 0} -t\log(\mathbb{P}(Y_t = x_j|Y_0 = x_i)) \to \rho(x_i,x_j)^2.$$

This estimate holds more generally for any space admitting a diffusive stochastic process [Saloff-Coste, 2010]. Taken together, we finally obtain:

Corollary C.1.4 (Varadhan's formula on graphs). *For any δ,γ,n_0 there exists some $\widehat{t}$, $n > n_0$, and sequence b^n_j such that the following holds for the simple random walk*

[1] For $t = \Theta(g_n^{-2})$, the marginal distribution $n\mathbb{P}(X_t|X_0)$ must be a.s. uniformly equicontinuous. For undirected spatial graphs, this is always true[Croydon and Hambly, 2008b] , but for directed graphs this is an open conjecture from [Hashimoto et al., 2015c]

X_t^n:

$$\mathbb{P}\Big(\sup_{x_i, x_j \in \mathcal{X}_{n_0}} \Big|\widehat{t}\log(\mathbb{P}(X^n_{\widehat{t}g_n^{-2}} = x_j \mid X^n_0 = x_i)) - \widehat{t}b^n_j - \rho_{\overline{\sigma}(x)}(x_i, x_j)^2\Big| > \delta\Big) < \gamma$$

Where $\rho_{\overline{\sigma}(x)}$ is the geodesic defined as

$$\rho_{\overline{\sigma}(x)}(x_i, x_j) = \min_{f \in C^1 : f(0)=x_i, f(1)=x_j} \int_0^1 \overline{\sigma}(f(t))dt$$

Proof. The proof is in two parts. First, by Varadhan's formula (Theorem C.1.3, [Molchanov, 1975, Eq. 1.7]) for any $\delta_1 > 0$ there exists some $\widehat{t}$ such that:

$$\sup_{y, y' \in D} |-\widehat{t}\log(\mathbb{P}(Y_{\widehat{t}} = y'|Y_0 = y)) - \rho_{\overline{\sigma}(x)}(y', y)^2| < \delta_1$$

The uniform equicontinuity of the marginals implies their uniform convergence (Lemma B.1.1), so for any $\delta_2 > 0$ and γ_0, there exists a n such that

$$\mathbb{P}(\sup_{x_j, x_i \in \mathcal{X}_{n_0}} |\mathbb{P}(Y_{\widehat{t}} = x_j|Y_0 = x_i) - np(x_j)\mathbb{P}(X^n_{g_n^{-2}\widehat{t}} = x_j|X^n_0 = x_i)| > \delta_2) < \gamma_0$$

By the lower bound on p and compactness of D, $\mathbb{P}(Y_{\widehat{t}}|Y_0)$ is lower bounded by some strictly positive constant c and we can apply uniform continuity of $\log(x)$ over (c, ∞) to get that for some δ_3 and γ,

$$\begin{aligned}\mathbb{P}(\sup_{x_j, x_i \in \mathcal{X}_{n_0}} &|\log(\mathbb{P}(Y_{\widehat{t}} = x_j|Y_0 = x_i)) - \log(np(x_j)) \\ &- \log(\mathbb{P}(X^n_{g_n^{-2}\widehat{t}} = x_j|X^n_0 = x_i))| > \delta_3) < \gamma. \quad \text{(C.2)}\end{aligned}$$

Finally we have the bound,

$$\begin{aligned}\mathbb{P}(\sup_{x_i, x_j \in \mathcal{X}_{n_0}} \Big| &-\widehat{t}\log(\mathbb{P}(X^n_{g_n^{-2}\widehat{t}} = x_j|X^n_0 = x_i)) \\ &- \widehat{t}\log(np(x_j)) - \rho_{\overline{\sigma}(x)}(x_i, x_j)^2\Big| > \delta_1 + \widehat{t}\delta_3) < \gamma\end{aligned}$$

To combine the bounds, given some δ and γ, set $b^n_j = \log(np(x_j))$, pick $\widehat{t}$ such that $\delta_1 < \delta/2$, then pick n such that the bound in Eq. C.2 holds with probability γ and error $\delta_3 < \delta/(2\widehat{t})$. □

C.2 Consistency proofs for word embedding

Lemma C.2.1 (Law of large numbers for log coocurrences). *Let X_t be a Markov chain defined by the transition*

$$\mathbb{P}(X_t = x_j | X_{t-1} = x_i) = \frac{\exp(-||x_i - x_j||_2^2/\sigma^2)}{\sum_{k=1}^n \exp(-||x_i - x_k||_2^2/\sigma^2)} \tag{C.3}$$

and C_{ij} be the number of times that $X_t = x_j$ and $X_{t-1} = x_i$ over m steps of this chain. Then for any $\delta > 0$ and $\varepsilon > 0$ there exist some m and constants a_i^m and b_j^m such that

$$\mathbb{P}\left(\sup_{i,j}\left| - \log(C_{ij}) - ||x_i - x_j||_2^2/\sigma^2 + a_i^m + b_j^m\right| > \delta\right) < \varepsilon$$

Proof. By detailed balance we observe that the stationary distribution $\pi_X(x_i)$ exists and is the normalization constant of the transition

$$\begin{aligned}\mathbb{P}(X_t = x_j | X_{t-1} = x_i)\pi_X(x_i) &= \frac{\exp(-||x_i - x_j||_2^2/\sigma^2)}{\sum_{k=1}^n \exp(-||x_i - x_k||_2^2/\sigma^2)} \sum_{k=1}^n \exp(-||x_i - x_k||_2^2/\sigma^2)\\ &= \mathbb{P}(X_t = x_i | X_{t-1} = x_j)\pi_X(x_j).\end{aligned}$$

Define m_i as the number of times that $X_t = x_i$ in a m word corpus. Applying the Markov chain law of large numbers, we obtain that for any $\varepsilon_0 > 0$ and $\delta_0 > 0$ there exists some m such that

$$P\Big(\sup_i \Big|\pi_X(x_i) - m_i/m\Big| > \delta_0\Big) < \varepsilon_0.$$

Therefore with probability ε_0, $m_i > m(\pi_X(x_i) - \delta_0)$.

Now given m_i, $C_{ij} \sim \text{Binom}(\mathbb{P}(X_t = x_j | X_{t-1} = X_i), m_i)$ applying Hoeffding's inequality and union bounding for any $\delta_1 > 0$ and $\varepsilon_1 > 0$ there exists some set of m_i such that

$$\mathbb{P}\left(\sup_{i,j}\Big|C_{ij}/m_i - \mathbb{P}\Big(X_t = x_j \mid X_{t-1} = x_i\Big)\Big| < \delta_1\right) \geq (1 - 2\exp(-2\delta_1^2 m_i))^{n^2} = \varepsilon_1.$$

Since $||x_i - x_j||_2 < \infty$, $\mathbb{P}(X_t = x_j | X_{t-1} = x_i)$ is lower bounded by some strictly positive constant c and we may apply the continuous mapping theorem on $\log(c)$ uniformly continuous over (c, ∞) to obtain that for all δ_2 and ε_2 there exists some set of m_i such that

$$\mathbb{P}\Big(\sup_{i,j}\Big|\log(C_{ij}) - \log(m_i) - \log(\mathbb{P}(X_t = x_j|X_{t-1} = x_i))\Big| < \delta_2\Big) \geq \varepsilon_2.$$

Therefore given any δ and ε for the theorem statement, set $\delta_2 = \delta$ and $\varepsilon_2 = \sqrt{\varepsilon}$ and define m' as the smallest m_i required. Since $\sup_{ij} ||x_i - x_j|| < \infty$, the Markov chain law of large numbers implies we can always find some m such that $\inf_i m_i > m'$ with probability at least $\sqrt{\varepsilon}$ which completes the original statement. □

Theorem C.2.2 (Consistency of GloVE). *Define the GloVe objective function as*

$$g(\widehat{x}, \widehat{c}, \widehat{a}, \widehat{b}) = \sum_{i,j} f(C_{ij})(2\widehat{x}_i\widehat{c}_j + \widehat{a} + \widehat{b} - \log(C_{ij}))^2$$

Define $\overline{x}_m, \overline{c}_m, \overline{a}_m, \overline{b}_m$ as the global minima of the above objective function for a corpus of size m.

Then the parameters derived from the true embedding in Lemma C.2.1, $x' = x/\sigma$, $a_i' = a_i^m - ||x_i||_2^2/\sigma^2$, $b_i' = b_i^m - ||x_i||_2^2/\sigma^2$ is arbitrarily close to the global minima in the sense that for any $\varepsilon > 0$ and $\delta > 0$ there exists some m such that

$$\mathbb{P}(|g(x', x', a', b') - g(\overline{x}_m, \overline{c}_m, \overline{a}_m, \overline{b}_m)| > \delta) < \varepsilon$$

Proof. Using Lemma C.2.1 with error δ_0 and probability ε_0 there exists some m such that uniformly over i and j,

$$(-||x_i - x_j||_2^2/\sigma^2 + a_i^m + b_i^m + \log(C_{ij}))^2 \leq \delta_0^2.$$

Now recall that $f(C_{ij}) \leq 10^{3/4} = c$ therefore

$$\mathbb{P}(g(x', x', a', b') > cn^2\delta_0^2) < \varepsilon_0.$$

Now the global minima $g(\overline{x}_m, \overline{c}_m, \overline{a}_m, \overline{b}_m)$ must be less than $g(x', x', a', b')$ and we have $0 < g(\overline{x}_m, \overline{c}_m, \overline{a}_m, \overline{b}_m) < g(x', x', a', b')$.

Therefore,

$$\mathbb{P}(|g(x', x', a', b') - g(\overline{x}_m, \overline{c}_m, \overline{a}_m, \overline{b}_m)| > cn^2\delta_0/2) < \varepsilon_0.$$

Picking a m such that $\delta_0 = 2\delta/(cn^2)$ and $\varepsilon_0 = \varepsilon$ concludes the proof. □

Lemma C.2.3 (Consistency of SVD). *Assume the norm of the latent embedding is proportional to the unigram frequency*

$$||x_i||/\sigma^2 = \frac{C_i}{\sqrt{\sum_j C_j}}.$$

Under these conditions, Let $\widehat{X}$ be the embedding derived from the SVD of M_{ij} as

$$2\widehat{X}\widehat{X}^T = M_{ij} = \log(C_{ij}) - \log\left(C_i\right) - \log\left(C_j\right) + \log\left(\sum_i C_i\right) + \tau.$$

Then there exists a τ such that this embedding is close to the true embedding under the same equivalence class as Lemma C.2.3

$$\mathbb{P}\Big(\sum_i ||A\widehat{x}_i/\sigma^2 - x_j||_2^2 > \delta\Big) < \varepsilon.$$

Proof. By Corollary C.1.4 for any $\delta_1 > 0$ and $\varepsilon_1 > 0$ there exists a m such that

$$P\Big(\sup_{i,j}\Big| - \log(C_{ij}) - \left(||x_i - x_j||_2^2/\sigma^2\right) - \log(mc)\Big| > \delta_1\Big) < \varepsilon_1.$$

Now additionally, if $C_i/\sqrt{\sum_j C_j} = ||x_i||^2/\sigma^2$ then we can rewrite the above bound as

$$P\Big(\sup_{i,j}\Big|\log(C_{ij}) - \log(C_i) - \log(C_j) + \log\left(\sum_i C_i\right) \\ - 2\langle x_i, x_j\rangle/\sigma^2 - \log(mc)\Big| > \delta_1\Big) < \varepsilon_1.$$

and therefore,

$$P\left(\sup_{i,j}\Big|M_{ij} - 2\langle x_i, x_j\rangle/\sigma^2 - \log(mc)\Big| > \delta_1\right) < \varepsilon_1.$$

Given that the dot product matrix has error at most δ_1, the resulting embedding it known to have at most $\sqrt{\delta_1}$ error [Sibson, 1979].

This completes the proof, since we can pick $\tau = -\log(mc)$, $\delta_1 = \delta^2$ and $\varepsilon_1 = \varepsilon$. □

Theorem C.2.4 (Consistency of SVD-MDS). *Let C_{ij} be defined as above and $M_{ij} = \log(C_{ij})$ and the centering matrix $V = I - 11^T/n$. Define the SVD based embedding $\widehat{X}$ as*

$$\widehat{X}\widehat{X}^T = \widehat{M} = VMV/2.$$

Without loss of generality, also assume that the latent vectors x have zero mean, then for any $\varepsilon > 0$ and $\delta > 0$, there exists some m, scaling constant σ, and an orthogonal matrix A such that

$$\mathbb{P}(\sum_i ||A\widehat{x}_i/\sigma^2 - x_j||_2^2 > \delta) < \varepsilon$$

Proof. By Lemma C.2.1 we have that

$$P\left(\sup_{i,j}\left| -\log(C_{ij}) - ||x_i - x_j||_2^2/\sigma^2 + a_i^m + b_j^m \right| > \widehat{\delta}\right) < \widehat{\varepsilon}$$

Since mean error cannot exceed entrywise error we can bound the row averages of $\log(C_{ij})$, where the dot product term is zero since x is zero mean.

$$P\left(\sup_i\left| -\frac{\sum_j \log(C_{ij})}{n} - a_i^m - \frac{\sum_j b_j^m}{n} - \frac{||x_i||_2^2}{\sigma} - \frac{\sum_j ||x_j||_2^2}{\sigma^2 n} + 2\left\langle x_i, \frac{\sum_j x_j}{n}\right\rangle \right| > \widehat{\delta}\right) < \widehat{\varepsilon}$$

Or in other words, $-\frac{\sum_j \log(C_{ij})}{n} \approx a_i^m + \frac{\sum_j b_j^m}{n} - \frac{||x_i||_2^2}{\sigma} - \frac{\sum_j ||x_j||_2^2}{\sigma^2 n}$

Define $M'_{ij} = -\log(C)_{ij} - \frac{\sum_j - \log(C)_{ij}}{n}$; applying the triangle inequality and combining both bounds gives

$$P\left(\sup_j\left|\frac{\sum_i M'_{ij}}{n} - \left(b_j - \frac{\sum_k b_k^m}{n} + ||x_j|| - \frac{\sum_k ||x_k||_2^2}{\sigma^2 n}\right)\right| > 2\widehat{\delta}\right) < 1 - (1 - \widehat{\varepsilon})^2.$$

Note that $M'_{ij} - \sum_i M'_{ij}/n = 2\widehat{M}_{ij}$ is the doubly centered matrix as defined above and combining all above bounds we have,

$$P\left(\sup_{ij}\left|\widehat{M}_{ij} - \langle x_i, x_j\rangle\right| > 4\widehat{\delta}\right) < 1 - (1 - \widehat{\varepsilon})^4.$$

Given that the dot product matrix has error at most 4δ the resulting embedding it known to have at most $\sqrt{4\widehat{\delta}}$ error [Sibson, 1979].

This completes the proof, since we can pick $\widehat{\delta} = \delta^2/4$ and $\widehat{\varepsilon} = 1 - (1 - \varepsilon)^{1/4}$ □

Theorem C.2.5 (Consistency of softmax/word2vec). *Define the softmax objective function with bias as*

$$g(\widehat{x}, \widehat{c}, \widehat{b}) = \sum_{ij} C_{ij} \log \left(\frac{\exp(-||\widehat{x}_i - \widehat{c}_j||_2^2 + \widehat{b}_j)}{\sum_{k=1}^n \exp(-||\widehat{x}_i - \widehat{c}_k||_2^2 + \widehat{b}_k)} \right)$$

Define $\overline{x}_m, \overline{c}_m, \overline{b}_m$ as the global minima of the above objective function for a co-occurence C_{ij} over a corpus of size m. For any $\varepsilon > 0$ and $\delta > 0$ there exists some m such that

$$\mathbb{P}(|g(x/\sigma, x/\sigma, 0) - g(\overline{x}_m, \overline{c}_m, \overline{b}_m)| > \delta) < \varepsilon$$

Proof. By differentiation, any objective of the form

$$\min_{\lambda_{ij}} C_{ij} \log \left(\frac{\exp(-\lambda_{ij})}{\sum_k \exp(-\lambda_{ik})} \right)$$

has the minima $\lambda_{ij} = -\log(C_{ij}) + a_i$ up to un-identifiable a_i with objective function value $C_{ij} \log(C_{ij}/\sum_k C_{ik})$. This gives a global function lower bound

$$g(\overline{x}_m, \overline{c}_m, \overline{b}_m) \geq \sum_{ij} C_{ij} \log \left(\frac{C_{ij}}{\sum_k C_{ik}} \right)$$

Now consider the function value of the true embedding x/σ;

$$\begin{aligned} &g(x/\sigma, x/\sigma, 0) \\ &= \sum_{ij} C_{ij} \log \left(\frac{\exp(-||x_i - x_j||_2^2/\sigma^2)}{\sum_k \exp(-||x_i - x_k||_2^2/\sigma^2)} \right) \\ &= \sum_{ij} C_{ij} \log \left(\frac{\exp(\log(C_{ij}) + \delta_{ij} + a_i)}{\sum_k \exp(\log(C_{ik}) + \delta_{ik} + a_i)} \right). \end{aligned}$$

We can bound the error variables δ_{ij} using Corollary C.1.4 as $\sup_{ij} |\delta_{ij}| < \delta_0$ with probability ε_0 for sufficiently large m with $a_i = \log(m_i) - \log(\sum_{k=1}^n \exp(-||x_i - x_k||_2^2/\sigma^2))$.

Taking the Taylor expansion at $\delta_{ij} = 0$, we have

$$g(x/\sigma, x/\sigma, 0) = \sum_{ij} C_{ij} \log\left(\frac{C_{ij}}{\sum_k C_{ik}}\right) + \sum_{l=1}^{n} \frac{C_{il}}{\sum_k C_{ik}} \delta_{il} + o(||\delta||_2^2)$$

By the law of large number of C_{ij},

$$\mathbb{P}\left(\left|g(x/\sigma, x/\sigma, 0) - \sum_{ij} C_{ij} \log\left(\frac{C_{ij}}{\sum_k C_{ik}}\right)\right| > n\delta_0\right) < \varepsilon_0$$

Combining with the global function lower bound we have that

$$\mathbb{P}\Big(\Big|g(x/\sigma, x/\sigma, 0) - g(\overline{x}, \overline{c}, \overline{b})\Big| > n\delta_0\Big) < \varepsilon_0.$$

To obtain the original theorem statement, take m to fulfil $\delta_0 = \delta/n$ and $\varepsilon_0 = \varepsilon$. □

Note that for negative-sampling based word2vec, applying the stationary point analysis of Levy and Goldberg [2014b] combined with the analysis in Lemma C.2.3 shows that the true embedding is a global minima.

C.3 Symmetry and windowing co-occurences

Existing word embedding algorithms utilize weighted, windowed, symmetrized word counts. Let C_{ij}^t define the t-step co-occurence which counts the number of times $X_{t+t'} = x_j$ and $X_{t'} = x_i$.

Then for some weight function $w(t)$ such that $\sum_{t=1}^{\infty} w(t) = 1$, we define

$$\widehat{C}_{ij} = \sum_{t=1}^{\infty} w(t)(C_{ij}^t + C_{ji}^t).$$

This is distinct from our stochastic process approach in two ways: first, there is symmetrization by counting both forward and backward transitions of the Markov chain. second, all words within a window of the center word $X_{t'}$ are used to form the co-occurences.

Symmetry: We begin by considering asymmetry of the random walk. If the Markov chain is reversible as in the cases of the Gaussian random walk, un-directed graphs, and the topic model, we can apply detailed balance to show that the joint distributions are symmetric

$$\mathbb{P}(X_{t+1} = x_j | X_t = x_i)\pi_X(x_i) = \mathbb{P}(X_{t+1} = x_i | X_t = x_j)\pi_X(x_j)$$

Therefore the empirical sum converges to

$$C^t_{ij}+C^t_{ji} \rightarrow \mathbb{P}(X_{t+t'} = x_j, X_{t'} = x_i)+\mathbb{P}(X_{t+t'} = x_i, X_{t'} = x_j) = 2\mathbb{P}(X_{t+t'} = x_j, X_{t'} = x_i)$$

In the cases where the random walk is non-reversible, such as a k-nearest neighbor graph then the two terms are not exactly equal, however note that if the non-symmetrized transition matricies C_{ij} fulfill Varadhan's formula both ways:

$$-t\log(C_{ij}) - a^m_i \rightarrow ||x_i - x_j||_2^2 + b^m_j \qquad \text{and} \qquad -t\log(C_{ji}) - a^m_j \rightarrow ||x_j - x_i||_2^2 + b^m_i$$

The sum $\widehat{C}_{ij}$ will fulfil

$$(C^t_{ij} + C^t_{ji}) = \exp(-||x_i - x_j||_2^2/t + o(1/t))\,(\exp(a_i/t + b_j/t) + \exp(b_i/t + a_j/t))$$

and

$$-t\log(C^t_{ij} + C^t_{ji}) = ||x_i - x_j||_2^2 + \log\left(\exp(a_i/t + b_j/t) + \exp(b_i/t + a_j/t)\right)t + o(1)$$

More specifically, for the manifold case, $a_i = \log(\pi_{X_n}) \rightarrow \log(np(x)/\overline{\sigma}(x_i)^2)$ and $b_j = -\log(np(x))$, and so the above term reduces to

$$-t\log(C^t_{ij} + C^t_{ji}) = ||x_i - x_j||_2^2 + \log\left(\overline{\sigma}^{-2}(x_i) + \overline{\sigma}^{-2}(x_j)\right)t + o(1)$$

Since the $\overline{\sigma}$ is independent of t, as $t \rightarrow 0$, we are once again left with Varadhan's formula in the symmetrized case.

In practice, this does not seem to affect the manifold embedding approaches much; in the results section we attempt embedding the MNIST digits dataset using the k-nearest neighbor simple random walk which is nonreversible.

Windowing: Now we consider the effect of windowing. We focus on the manifold case for analytic simplicity, but the same limits apply to the other two examples of Gaussian random walks and topic models.

Let $q_t(x, x') = \mathbb{P}(Y_t = x | Y_0 = x')$ and where Y_t fulfills Varadhan's formula such that there exists a metric function ρ,

$$\lim_{t\rightarrow 0} -t\log(q_t(x, x')) \rightarrow \rho(x, x')^2$$

Under these conditions, let $\widehat{q}_t(x,x') = \int_0^t q_{t'}(x,x')/tdt'$ define the windowed marginal distribution. We show this follows a windowed Varadhan's formula.

$$\lim_{t\to 0} t\widehat{q}_t(x,x') \to \rho(x,x')^2$$

This can be done via a direct argument. Varadhan's formula implies that,

$$q_t(x,x') = \exp\left(-\frac{\rho(x,x')^2}{t} + o\left(\frac{1}{t}\right)\right).$$

Thus we can find some bounding constants $0 < c = o(1)$ such that

$$\int_0^t \frac{1}{t}\exp\left(-\frac{\rho(x,x')^2}{t'} - \frac{c}{t'}\right) dt \leq \widehat{q}_t(x,x') \leq \int_0^t \frac{1}{t}\exp\left(-\frac{\rho(x,x')^2}{t'} + \frac{c}{t'}\right) dt'$$

Performing the bounding integral for general $c \in \mathbb{R}$,

$$\begin{aligned}\int_0^t \frac{1}{t}\exp\left(-\frac{\rho(x,x')^2}{t'} + \frac{c}{t'}\right) dt' &= \frac{1}{t}\left(\exp\left(-\frac{\rho(x,x')^2 - 2c}{2t}\right) t + (c - \rho(x,x')^2/2)\Gamma\left(\frac{\rho(x,x')^2 - 2c}{2t}\right)\right) \\ &= \frac{1}{t}\left(\exp\left(-\frac{c}{t} - \frac{\rho(x,x')^2}{t}\right)\left(-\frac{2t}{2c - \rho(x,x')^2} + t^2\right)\right)\end{aligned}$$

Therefore we have that for any c,

$$\lim_{t\to 0} -t\log\left(\int_0^t \frac{1}{t}\exp\left(-\frac{\rho(x,x')^2}{t'} + \frac{c}{t'}\right) dt'\right) \to \rho(x,x')^2 - c$$

By the two-sided bound and $c = o(1)$,

$$\lim_{t\to 0} t\widehat{q}_t(x,x') \to \rho(x',x)^2.$$

as desired.

C.4 Empirical evaluation details

C.4.1 Implementation details

We used off-the-shelf implementations of `word2vec`[2] and `GloVe`[3]. The two other methods (randomized) SVD and regression embedding are both implemented on top of the `GloVe` codebase. We used 300-dimensional vectors and window size 5 in all models. Further details are provided below.

`word2vec`. We used the skip-gram version with 5 negative samples, 10 iterations, $\alpha = 0.025$ and frequent word sub-sampling with a parameter of 10^{-3}.

`GloVe`. We disabled GloVe's corpus weighting, since this generally produced superior results. The default step-sizes results in NaN-valued embeddings, so we reduced them. We used $X_{\text{MAX}} = 100$, $\eta = 0.01$ and 10 iterations.

SVD For the SVD algorithm of Levy and Goldberg [2014b], we use the GloVe co-occurrence counter combined with a parallel randomized projection SVD factorizer, based upon the `redsvd` library due to memory and runtime constraints[4]. Following Levy et al. [2015], we used the square root factorization, no negative shifts ($\tau = 0$ in our notation), and 50,000 random projections.

Regression Embedding We do standard stochastic gradient descent with two differences. First, we drop co-occurrences values C_{ij} smaller than 10 with probability proportional to $1 - C_{ij}/10$ and scale the gradient, which resulted in training time speedups with no loss in accuracy. Second, we use an initial line search step combined with a linear stepsize decay by epoch. We use $\theta = 50$ and η is line-searched starting at $\eta = 10$.

C.4.2 Word embedding corpora

We used three corpora to train the word embeddings: the full Wikipedia dump of 03/2015 (about 2.4B tokens), a larger corpus similar to that used by GloVe [Pennington et al., 2014]: Wikipedia2015 + Gigaword5 (5.8B tokens in total) and the one used word2vec [Mikolov et al., 2013b], which consists of a mixture of several corpora

[2] http://code.google.com/p/word2vec

[3] http://nlp.stanford.edu/projects/glove

[4] https://github.com/ntessore/redsvd-h

from different sources (6.4B tokens in total). We preprocessed all the corpora by removing punctuation, numbers and lower-casing all the text. Finally we ran two passes of word2vec's tokenizer word2phrase. As a final step, we removed function words from the vocabulary and kept only the 100K most common words for all our experiments.

C.4.3 Datasets for semantic tasks

Our first set of experiments is on two standard open-vocabulary analogy tasks: Google [Mikolov et al., 2013a] and MSR [Mikolov et al., 2013c]. Google consists of 19,544 semantic and syntactic analogy questions, while MSR's 8,000 questions are all syntactic. As an additional analogy task, we use the SAT analogy questions (version 3) of Turney [Turney and Littman, 2005]. The dataset contains 374 questions from actual SAT exams, guidebooks, from the ETS web site and other sources. Each question consists of 5 exemplar pairs of words *word1:word2*, where all the pairs hold the same relation. The task is to pick from among another five pairs of words the one that best represents the relation represented by the exemplars. To the best of our knowledge, this is the first time word embeddings are used to solve this task.

Given the current lack of freely available datasets with category and sequence questions, as described in Section 2, we decided to create them. We used `nltk`'s[5] interface to WordNet [Miller and Fellbaum, 1998] in combination with word-word PMI values computed on the Wiki corpus to create the sequences and classes.

As a first step, we collected a set of root words from other semantic tasks to initialize the methods. For the classification data, we created the in-category words by selecting words from various WordNet relations associated to the root words, after which we pruned down to four words based on PMI-similarity to the root word and the other words in the class. The additional options for the multiple choice question were created searching over words related to the root by a different relation type, and selecting those most similar to the root.

For the sequence data, we obtained from WordNet trees of words given by various relation types, and then pruned based on similarity to the root word. For the multiple-choice version of the data, we selected additional (incorrect) options by searching over other words related to the root word, and pruning, as for sequences, based on PMI similarity. Finally, we manually pruned all three sets of questions, keeping only the most coherent questions, in order to increase the quality of the datasets. After pruning, the category dataset was left with 215 questions and the

[5] http://www.nltk.org/

sequence dataset with 51 questions in its open-vocabulary version and 169 in its multiple choice version.

The two datasets will be made publicly available, in the hopes of broadening the type of tasks used to evaluate semantic content of word embeddings.

C.4.4 Solving inductive reasoning tasks

The ideal point for a task is defined below:

- **Analogies**: Given A:B::C, the ideal point is given by $B-A+C$ (parallelogram rule).
- **Analogies (SAT)**: Given prototype A:B and candidates $C_1 : D_1 \ldots C_n : D_n$, we compare $D_i - C_i$ to the ideal point $B - A$.
- **Categories**: Given a category implied by $w_1, \ldots, w_n$, the ideal point is $I = \frac{1}{n}\sum_{i=1}^{n} w_i$.
- **Sequence**: Given sequence $w_1 : \cdots : w_n$ we compute the ideal as $I = w_n + \frac{1}{n}(w_n - w_1)$.

Once we have the ideal point I, we pick the answer as the word closest to I among the options, using L_2 or cosine distance. For the latter, we normalize I to unit norm before taking the cosine distance. For L_2 we do not apply any normalization.

C.5 Full table of analogy results

C.5.1 Top-30k vocabulary resctriction

	Google Analogies			SAT	MSR Analogies
	Semantic	Syntactic	Total		
Covered	5022	8195	13217	217	4358
Total	8869	10675	19544	374	8000

Table C.1: glove corpus question coverage

	Google Analogies			SAT	MSR Analogies
	Semantic	Syntactic	Total		
Covered	4746	7679	12425	199	4340
Total	8869	10675	19544	374	8000

Table C.2: wiki corpus question coverage

	Google Analogies			SAT	MSR Analogies
	Semantic	Syntactic	Total		
Covered	3965	8447	12412	257	4554
Total	8869	10675	19544	374	8000

Table C.3: w2v corpus question coverage

	Google Analogies (cosine)			Google Analogies (l2)		
Method	Semantic	Syntactic	Total	Semantic	Syntactic	Total
regression	**78.4**	70.5	**73.5**	**74.1**	70.0	**71.2**
GloVE	70.2	70.9	70.6	59.2	67.7	64.5
SVD	55.8	46.4	50.0	49.1	41.2	44.3
word2vec	68.0	**73.8**	71.6	66.6	**71.2**	69.4

Table C.4: wiki corpus analogy accuracy

	SAT			MSR Analogies	
Method	L2	diff-cosine	cosine	cosine	L_2
regression	38.7	41.7	33.7	**67.2**	**64.0**
GloVE	37.2	40.7	35.7	61.2	53.5
SVD	32.7	32.2	28.1	33.5	30.3
word2vec	**41.9**	**42.9**	**41.4**	65.0	63.4

Table C.5: wiki corpus analogy accuracy for MSR and SAT datasets

	Classification		Sequence		Sequence (open vocab)		Sequence (open vocab, top5)	
Method	Cosine	L_2	Cosine	L_2	Cosine	L_2	Cosine	L_2
regression	**86.1**	**85.6**	58.0	55.6	**7.8**	**5.9**	**72.5**	**60.8**
GloVE	80.9	76.7	**59.2**	51.5	2.0	2.0	51.0	37.3
SVD	74.9	64.7	46.2	46.2	2.0	2.0	21.6	25.5
word2vec	85.1	71.6	57.4	**59.2**	2.0	**5.9**	49.0	51.0

Table C.6: wiki corpus for classification and sequence

	Google Analogies (cosine)			Google Analogies (l2)		
Method	Semantic	Syntactic	Total	Semantic	Syntactic	Total
regression	**78.4**	70.8	**73.7**	**75.5**	**70.9**	**72.6**
GloVE	72.6	71.2	71.7	65.6	66.6	67.2
SVD	57.4	50.8	53.4	53.7	48.2	50.3
word2vec	73.4	**73.3**	73.3	71.4	**70.9**	71.1

Table C.7: glove corpus analogy accuracy

	SAT			MSR Analogies	
Method	L2	diff-cosine	cosine	cosine	L_2
regression	39.2	40.6	37.8	65.6	63.9
GloVE	36.9	42.8	33.6	62.0	55.6
SVD	27.1	32.2	25.8	32.0	30.6
word2vec	**42.0**	**49.2**	**42.0**	**67.9**	**66.5**

Table C.8: glove corpus analogy accuracy for MSR and SAT datasets

	Classification		Sequence		Sequence (open vocab)		Sequence (open vocab, top5)	
Method	Cosine	L_2	Cosine	L_2	Cosine	L_2	Cosine	L_2
regression	**84.6**	**87.6**	**58.9**	**58.3**	0.0	0.0	23.5	21.6
GloVE	80.1	73.1	**58.9**	48.8	0.0	0.0	27.5	23.5
SVD	74.6	65.2	53.0	52.4	0.0	2.0	19.6	15.7
word2vec	**84.6**	76.4	56.2	54.4	0.0	**3.9**	**53.0**	**58.8**

Table C.9: glove corpus for classification and sequence

	Google Analogies (cosine)			Google Analogies (l2)		
Method	Semantic	Syntactic	Total	Semantic	Syntactic	Total
regression	**80.1**	73.0	**75.2**	**77.3**	**73.1**	**74.4**
GloVE	70.4	73.0	72.2	61.9	70.0	67.2
SVD	55.2	43.6	54.1	52.8	50.6	51.3
word2vec	66.8	**73.4**	71.3	67.2	72.2	70.6

Table C.10: w2v corpus analogy accuracy

C.5.2 Top-100k vocabulary

	SAT			MSR Analogies	
Method	L2	diff-cosine	cosine	cosine	L_2
regression	38.1	43.0	36.9	69.4	68.4
GloVE	27.9	37.3	29.1	35.6	35.4
SVD	27.1	32.2	25.8	32.0	30.6
word2vec	**39.0**	**46.4**	**42.3**	**75.3**	**75.6**

Table C.11: w2v corpus analogy accuracy for MSR and SAT datasets

	Classification		Sequence		Sequence (open vocab)		Sequence (open vocab, top5)	
Method	Cosine	L_2	Cosine	L_2	Cosine	L_2	Cosine	L_2
regression	81.4	**85.5**	57.1	**55.4**	0.0	0.0	25.5	21.6
GloVE	78.2	70.0	**57.7**	50.6	2.0	0.0	31.4	31.4
SVD	74.1	61.1	47.0	48.2	0.0	0.0	35.3	21.6
word2vec	**87.0**	75.0	52.7	50.9	**3.9**	**5.9**	**49.0**	**45.1**

Table C.12: w2v corpus for classification and sequence

	Google Analogies			SAT	MSR Analogies
	Semantic	Syntactic	Total		
Covered	7829	10411	18240	217	5612
Total	8869	10675	19544	374	8000

Table C.13: glove corpus question coverage

	Google Analogies			SAT	MSR Analogies
	Semantic	Syntactic	Total		
Covered	7667	10231	17898	199	5186
Total	8869	10675	19544	374	8000

Table C.14: wiki corpus question coverage

	Google Analogies			SAT	MSR Analogies
	Semantic	Syntactic	Total		
Covered	7213	10405	17618	244	5462
Total	8869	10675	19544	374	8000

Table C.15: w2v corpus question coverage

	Google Analogies (cosine)			Google Analogies (l2)		
Method	Semantic	Syntactic	Total	Semantic	Syntactic	Total
regression	**76.9**	64.6	**69.9**	64.9	62.5	63.5
GloVE	69.0	66.0	67.3	53.5	62.1	58.4
SVD	53.8	40.2	46.1	40.2	34.2	36.8
word2vec	67.9	**70.4**	69.3	**67.4**	**67.2**	**67.3**

Table C.16: wiki corpus analogy accuracy

	SAT			MSR Analogies	
Method	L2	diff-cosine	cosine	cosine	L_2
regression	38.7	41.7	33.7	**62.6**	57.4
GloVE	37.2	40.7	35.7	58.6	50.2
SVD	32.7	32.1	28.1	31.3	26.7
word2vec	**41.7**	**43.2**	**41.2**	62.4	**61.5**

Table C.17: wiki corpus analogy accuracy for MSR and SAT datasets

	SAT			MSR Analogies	
Method	L2	diff-cosine	cosine	cosine	L_2
regression	39.2	40.6	37.8	65.6	63.9
GloVE	36.9	42.8	33.6	62.0	55.6
SVD	27.1	32.2	25.8	32.0	30.6
word2vec	**42.0**	**49.2**	**42.0**	**67.9**	**66.5**

Table C.18: wiki corpus analogy accuracy for MSR and SAT datasets

	Classification		Sequence		Sequence (open vocab, top 5)		Sequence (open vocab)	
Method	Cosine	L_2	Cosine	L_2	Cosine	L_2	Cosine	L_2
regression	**86.0**	**85.6**	**58.0**	**55.6**	**62.7**	37.3	**11.8**	**15.7**
GloVE	80.9	76.7	59.2	51.5	51.0	37.3	3.9	3.9
SVD	74.9	64.7	46.2	46.2	21.6	25.5	3.9	3.9
word2vec	85.1	71.6	45.1	43.1	43.1	**45.1**	3.9	11.8

Table C.19: wiki corpus for classification and sequence

	Google Analogies (cosine)			Google Analogies (l2)		
Method	Semantic	Syntactic	Total	Semantic	Syntactic	Total
regression	**75.0**	66.4	70.1	**70.0**	66.1	67.7
GloVE	70.7	67.5	68.8	62.5	62.4	62.5
SVD	57.0	44.2	50.3	47.9	42.0	44.5
word2vec	71.7	**71.5**	**71.5**	**70.0**	**68.7**	**69.5**

Table C.20: glove corpus analogy accuracy

	SAT			MSR Analogies	
Method	L2	diff-cosine	cosine	cosine	L_2
regression	39.2	40.6	37.8	62.2	58.9
GloVE	36.9	42.9	33.6	61.0	53.0
SVD	27.2	32.3	25.8	30.6	27.4
word2vec	**42.1**	**48.2**	**41.7**	**67.0**	**66.8**

Table C.21: glove corpus analogy accuracy for MSR and SAT datasets

	Classification		Sequence		Sequence (open vocab top 5)		Sequence (top 1)	
Method	Cosine	L_2	Cosine	L_2	Cosine	L_2	Cosine	L_2
regression	**84.6**	**87.6**	**58.9**	**58.3**	23.5	17.6	0.0	0.0
GloVE	80.1	73.1	58.3	48.8	27.5	23.5	0.0	0.0
SVD	74.6	65.1	55.6	54.4	19.6	11.8	0.0	3.9
word2vec	**84.6**	76.4	55.6	54.4	**49.0**	**54.9**	0.0	**7.8**

Table C.22: glove corpus for classification and sequence

	Google Analogies (cosine)			Google Analogies (l2)		
Method	Semantic	Syntactic	Total	Semantic	Syntactic	Total
regression	**78.2**	68.9	**72.7**	**72.0**	**68.6**	**70.0**
GloVE	70.6	**69.8**	70.1	61.2	65.7	63.9
SVD	55.9	47.8	51.1	45.4	44.7	45.0
word2vec	67.1	71.6	69.8	68.0	70.4	69.4

Table C.23: w2v corpus analogy accuracy

	SAT			MSR Analogies	
Method	L2	diff-cosine	cosine	cosine	L_2
regression	38.1	43.0	36.9	66.1	63.1
GloVE	36.9	**53.9**	34.0	65.3	59.0
SVD	27.9	37.2	29.1	33.6	31.1
word2vec	**39.2**	47.1	**42.8**	**73.8**	**74.6**

Table C.24: w2v corpus analogy accuracy for MSR and SAT datasets

	Classification		Sequence		Sequence (top 5)		Sequence (top 1)	
Method	Cosine	L_2	Cosine	L_2	Cosine	L_2	Cosine	L_2
regression	81.3	**85.5**	57.1	**55.4**	24.5	21.6	0.0	0.0
GloVE	78.2	70.0	**58.3**	50.6	31.4	31.4	3.9	0.0
SVD	74.1	61.1	45.8	48.2	31.4	21.6	0.0	0.0
word2vec	**87.0**	75.0	53.3	50.9	**43.1**	**35.3**	**7.84**	**11.8**

Table C.25: w2v corpus for classification and sequence

Appendix D

Population level diffusion

D.0.1 Hypothesis test proof

Corollary D.0.1 (Hypothesis test for Ψ). *Let Ψ_0 and Ψ_1 be candidate potentials such that given $\rho_0(0,x) = \rho_1(0,x)$ and*

$$\frac{\partial \rho_i}{\partial t} = div(\nabla \Psi_i(x) \rho_i(t,x)) + \sigma^2 \nabla^2 \rho_i(t,x)$$

fulfill $\rho_0(t,x) = \rho_1(t,x)$. Define $\rho_i(t_3,x)$ where $t_3 \sim T$ is a draw from T defined as a random variable absolutely continuous with respect to the Lebesgue measure, then either

$$P(\rho_1(t_3,x) = \rho_0(t_3,x)) = 1$$

if $\forall x$, $\Psi_1(x) = \Psi_0(x)$, or

$$P(\rho_1(t_3,x) = \rho_0(t_3,x)) = 0$$

otherwise.

Proof. By theorem 5.4.2, we know that if both $\frac{\partial \rho_1}{\partial t} = \frac{\partial \rho_0}{\partial t}$ and $\rho_1(t,x) = \rho_0(t,x)$ for any t, then $\Psi_1(x) = \Psi_0(x)$. Therefore if $\Psi_1(x) \neq \Psi_0(x)$, any t such that $\rho_1(t,x) = \rho_0(t,x)$ must have distinct time derivatives.

Now by Bolzano Weierstrass, if $\rho_1(t,x) = \rho_0(t,x)$ an infinite times over any finite time interval $[0,T]$, then there must be some accumulation point such that $\rho_1(t,x) = \rho_0(t,x)$ has a convergent subsequence. By differentiability of ρ with respect to time, this implies $\frac{\partial \rho_0}{\partial t}$ at some $\rho_1(t,x) = \rho_0(t,x)$. Therefore, if $\Psi_1(x) \neq \Psi_0(x)$ there can only be a finite number of times such that $\rho_1(t,x) = \rho_0(t,x)$. This has measure zero over with respect to the Lebesgue measure, thus any random stopping time t_3

implies

$$P(\rho_1(t_3, x) = \rho_0(t_3, x)) = 0.$$

The other direction occurs by uniqueness of the solution to the Fokker Planck equation.

□

D.0.2 Boundary conditions for identifiability

We prove the non-compact boundary condition, which replaces the boundary with some sequence of compact sets such that the probability of leaving the set limits to zero.

Theorem D.0.2 (Uniqueness of Fokker-Planck like operators). *Let $\Psi(x)$ be a C^1 solution to the following elliptic PDE:*

$$f(x) = \nabla^2\Psi(x)\tau(x) + \nabla\Psi(x)\nabla\tau(x) + \sigma^2\nabla^2\tau(x) \tag{D.1}$$

subject to the constraint $\int \exp(-\Psi(x)/\sigma^2)dx = 1$, $\int \tau(x)dx < \infty$.

Equation D.1 is fulfilled in the short-time case with, $f = \frac{\partial\rho}{\partial t}$, $\tau = \rho$ and in the time-integral case, $f(x) = \rho(t_0, x) - \rho(t_n, x)$ and $\tau(x) = \int_0^T \rho(t, x)dt$.

In both cases, assume that the underlying Fokker-Planck boundary condition allows us to construct a sequence of compact sets Ω_n such that $\lim_{n\to\infty} \int_{x\in\Omega_n} \tau(x)dx = \int_{x\in\mathbb{R}^d} \tau(x)dx < \infty$ and $\lim_{n\to\infty} \int_{x\in\omega} f(x) \to 0$.

Then $\Psi(x)$ is unique up to sets of measure zero of $\tau(x)$.

Proof. Consider any $\Psi_1(x)$ and $\Psi_2(x)$, then by linearity of the PDE $\Psi'(x) = \Psi_1(x) - \Psi_2(x)$ must be a solution to the homogeneous elliptic PDE

$$0 = \operatorname{div}(\nabla\Psi'(x)\tau(x)) = \nabla^2\Psi'(x)\tau(x) + \nabla\Psi'(x)\nabla\tau(x)$$

Construct $R_{\varepsilon,n} = \{x : x \in \Omega_n, \Psi'(x) \leq \varepsilon\}$, which is the intersection of the level set of Ψ' with Ω_n.

Expanding the limit boundary constraint on f and taking the difference we obtain:

$$\lim_{n\to\infty} \int_{x\in\partial\Omega_n} \langle\nabla\Psi'(x)\tau(x), n_x\rangle dx = 0.$$

Analogously to the reflecting boundary condition, define $\partial R^{\circ}_{\varepsilon,n}$ as the boundary of the sublevel set and $\partial\Omega^{\circ}_{\varepsilon,n}$ as the boundary of Ω_n such that the union of the two sets forms the boundary of $R_{\epsilon,n}$.

Applying the divergence theorem over the decomposition of the boundary analogously to the other boundary condition:

$$\lim_{n\to\infty}\int_{x\in R_{\varepsilon,n}} \operatorname{div}(\nabla\Psi'(x)\tau(x))dx \tag{D.2}$$

$$= \lim_{n\to\infty}\int_{x\in\partial\Omega_n^\circ} \langle\nabla\Psi'(x)\tau(x), n_x\rangle dx \tag{D.3}$$

$$+ \lim_{n\to\infty}\int_{x\in\partial R_{\varepsilon,n}^\circ} |\nabla\Psi'(x)|_2\tau(x)dx = 0. \tag{D.4}$$

which implies via our boundary constraint

$$\lim_{n\to\infty}\int_{x\in\partial R_{\varepsilon,n}^\circ} |\nabla\Psi'(x)|_2\tau(x)dx = 0.$$

This limit occurs uniformly in ε, since the first line of Eq D.2 is exactly zero and Eq D.3 is uniformly bounded as

$$\lim_{n\to\infty}\int_{x\in\partial\Omega_n^\circ} \langle\nabla\Psi'(x)\tau(x), n_x\rangle dx$$
$$\leq \lim_{n\to\infty}\int_{x\in\partial\Omega_n} \langle\nabla\Psi'(x)\tau(x), n_x\rangle dx.$$

Now assume that there exists some compact set S of nonzero measure such that for all $x \in S$, $|\nabla\Psi(x)| \neq 0$. Since Ψ is continuous the extreme value theorem implies the existence of some $\varepsilon_{\min} = \min_{x\in S}\Psi(x)$ and $\varepsilon_{\max} = \max_{x\in S}\Psi(x)$. Using the fact that any x with $|\nabla\Psi'(x)| \neq 0$ must be a part of $\partial R_{\varepsilon,n}^\circ$ for sufficient large n and uniformity of our limit with respect to ε we obtain:

$$\lim_{n\to\infty}\int_{x\in S} |\nabla\Psi'(x)|_2\tau(x)dx$$
$$= \lim_{n\to\infty}\int_{\varepsilon_{\min}}^{\varepsilon_{\max}}\int_{x\in\{\partial R_{\varepsilon,n}^\circ\cap S\}} |\nabla\Psi'(x)|\tau(x)dxd\varepsilon$$
$$\leq \lim_{n\to\infty}\int_{\varepsilon_{\min}}^{\varepsilon_{\max}}\int_{x\in\partial R_{\varepsilon,n}^\circ} |\nabla\Psi'(x)|\tau(x)dxd\varepsilon = 0.$$

Which is a contradiction, as this implies $\lim_{n\to\infty}|\nabla\Psi(x)| = 0$ from the fact that $\tau(x)$ has a lower bound strictly greater than zero over S.

Equicontinuity of $\nabla\Psi'(x)$ then implies $|\nabla\Psi'(x)| = 0$ for all x, and therefore

$$|\nabla\Psi'(x)| = |\nabla\Psi_1(x) - \nabla\Psi_2(x)| = 0.$$

Combined with the normalization constraint, $\int \exp(-\Psi(x)/\sigma^2)dx = 1$, this implies $\Psi_1(x) = \Psi_2(x)$. □

D.0.3 Details on parameters and methods

The following are the 'free' hyperparameters of the model:

- K: The number of hidden layers (200 for simulated data, 500 for RNA-seq data)
- Δt: simulation timestep (0.01 for simulations, 0.1 for RNA-seq)
- τ: regularization constant (0.7 for all data)
- ε: step size of adagrad (Grid searched from starting with 0.1 for 10 steps with decaying powers of 2)
- γ: adagrad squared gradient decay rate (0.01, all experiments)
- NS: number of samples to draw from simulations (Fixed to be the same as the number of points at the first time point)
- burnin: number of steps of the first-order Euler scheme to burn-in for contrastive divergence (set to 50)

For initializing the contrastive divergence, W is set to be i.i.d unit Gaussians, b to draws from the $[-1, 1]$ uniform, and g to zero.

D.0.4 Alternative methods

We fit the following baseline models:

- **Orstein-Uhlenbeck:** Quadratic potential with one parameter μ, $\Psi(x) = (x - \mu)^2$
- **Linear:** Linear potential with one parameter w, $\Psi(x) = xw^T$.
- **Local:** Sum of Gaussian potentials with three parameters μ, g and b, $\Psi(x) = g\exp(-(x-\mu)^2/b^2)$.

D.0.5 High-dim gene expression

Applying our RNN model to the top 5 or 10 differentiating genes as measured by the Wasserstein distance between the marginal day 0 and 7 distributions results in qualitatively similar results. In order to fit the higher-complexity multivariate model, we modified a few hyperparameters ($K = 2000$, initialization of b as $b_i = |w_i x_i|^2$, increasing NS to 1000, $\sigma = \sqrt{2}$ and using continuous contrastive divergence) and included all (non-heldout) data for pre-training. The parameter changes result in producing a similar goodness-of-fit to the higher dimensional versions of the problem with only a few hundred points.

For example, the D4 nonstationary dynamics of Krt8 are re-capitulated

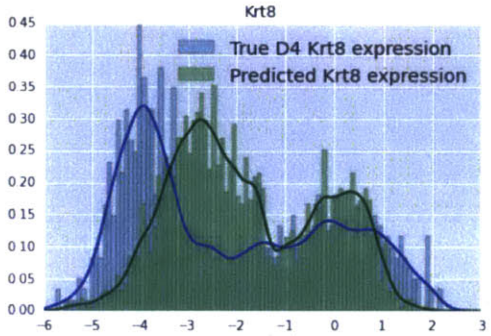


Figure D-1: 5-gene model prediction of Krt8 also reproduces the underlying bimodality of the data

Plotting the predicted marginal distribution for all 5 genes, we find that the RNN based model substantially outperforms other, parametric approaches to the same problem:

This same trend holds as we increase the number of genes from 5 to 10 where the RNN performs best compared to alternatives. We find that as we increase the dimensionality, the learned dynamics begin to become unimodal, as all models struggle to identify the true dynamics from sparse, high-dimensional data.

Even in this setting where we have a few hundred examples in 10 dimensions, we can still effectively identify correlations and other relationships between genes at this non-equilibrium state.

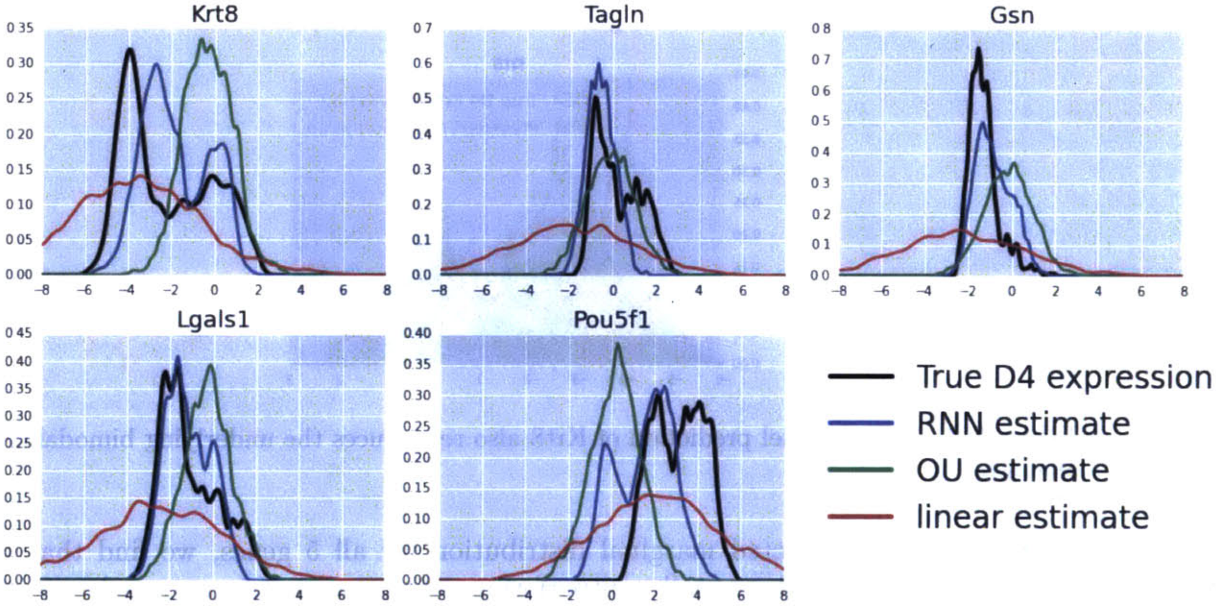


Figure D-2: Predicted marginal distributions of the top 5 differentiating genes at day 4

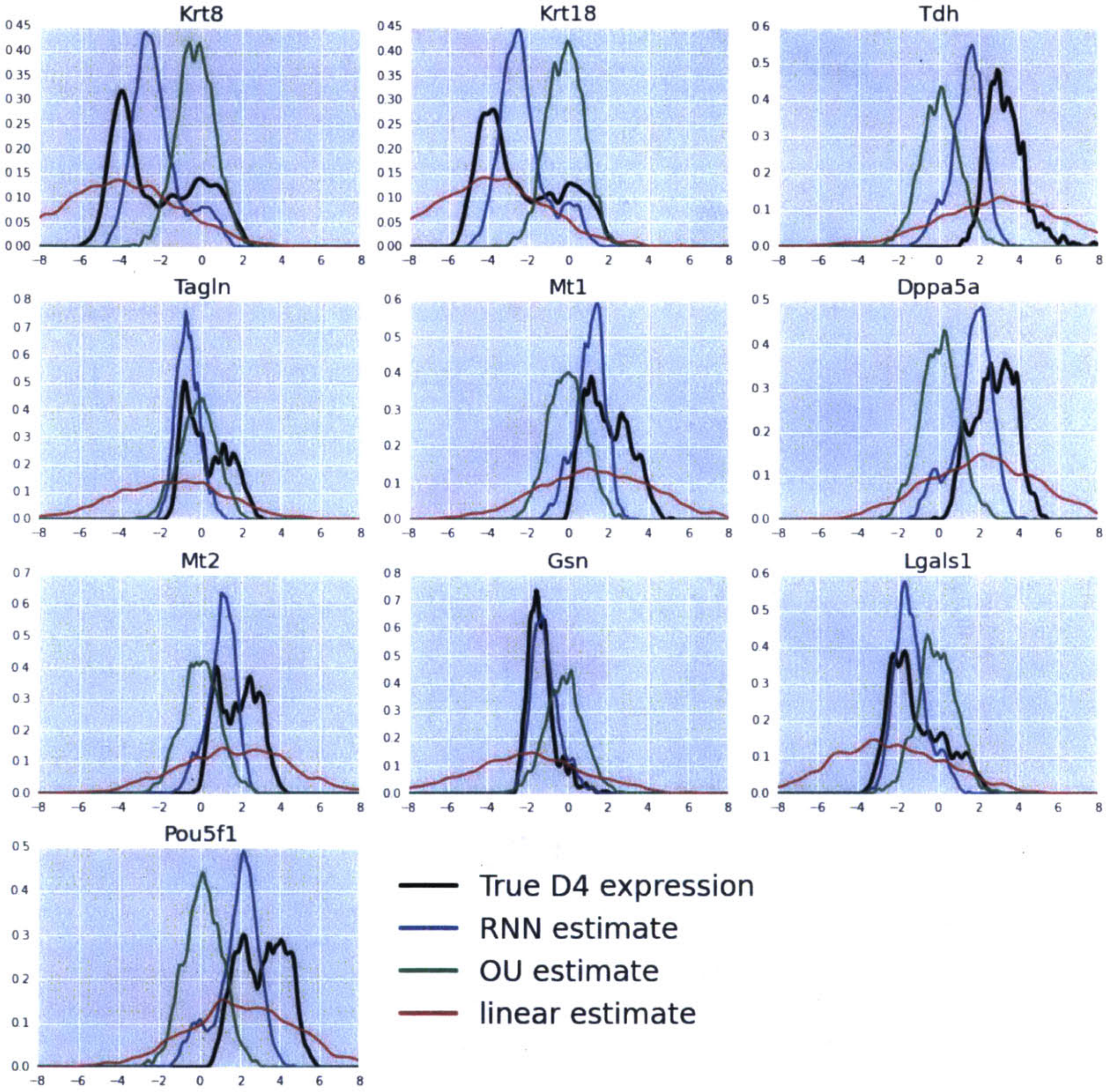


Figure D-3: Predicted marginal distributions of the top 10 differentiating genes at day 4

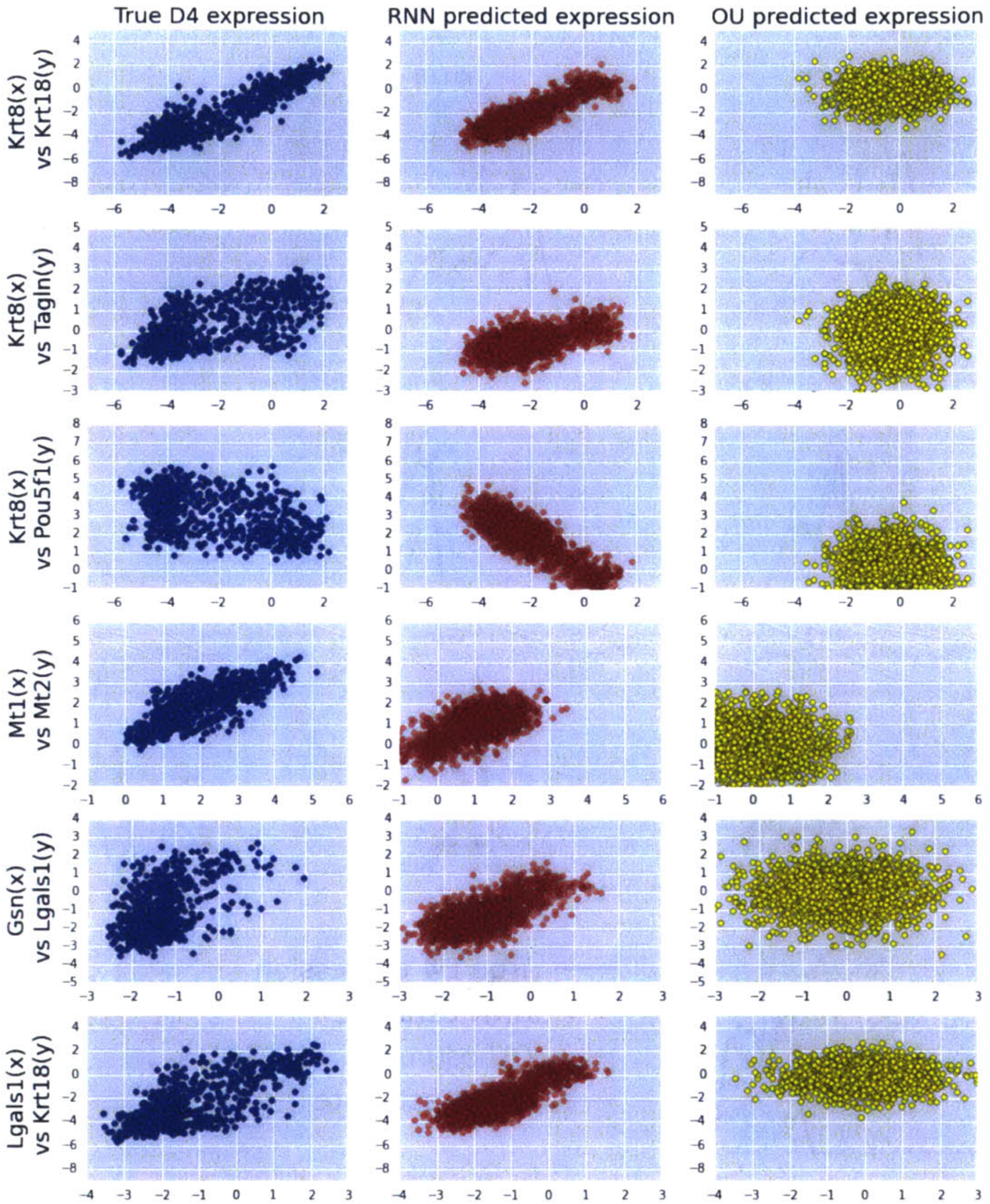


Figure D-4: Predicted against actual pairwise gene expression distributions at the D4 timepoint. The RNN models the correlational structure of the true dynamics.

Bibliography

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables.* Number 55. Courier Dover Publications, 1972.

Stefan Adams, Nicolas Dirr, Mark Peletier, and Johannes Zimmer. Large deviations and gradient flows. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 371(2005):20120341, 2013.

Morteza Alamgir and Ulrike von Luxburg. Phase transition in the family of p-resistances. In *Advances in Neural Information Processing Systems*, pages 379–387, 2011.

Morteza Alamgir and Ulrike von Luxburg. Shortest path distance in random k-nearest neighbor graphs. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1031–1038, 2012a.

Morteza Alamgir and Ulrike von Luxburg. Shortest path distance in random k-nearest neighbor graphs. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1031–1038, 2012b.

Morteza Alamgir, GÃąbor Lugosi, and Ulrike von Luxburg. Density-preserving quantization with application to graph downsampling. In *COLT*, 2014.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *arXiv preprint arXiv:1502.03520*, 2015.

Richard F. Bass and Pei Hsu. Some potential theory for reflecting Brownian motion in Hölder and Lipschitz domains. *Ann. Probab.*, 19(2):486–508, 1991.

Adrian Bejan and Gilbert W Merkx. *Constructal theory of social dynamics.* Springer, 2007.

Carl M Bender and Steven A Orszag. *Advanced Mathematical Methods for Scientists and Engineers I*. Springer Science & Business Media, 1999.

Sudin Bhattacharya, Qiang Zhang, and Melvin E Andersen. A deterministic map of waddington's epigenetic landscape for cell fate specification. *BMC systems biology*, 5(1):85, 2011.

Andrei N Borodin and Paavo Salminen. *Handbook of Brownian motion: facts and formulae*. Springer, 2002.

T. Byczkowski, J. Małecki, and M. Ryznar. Hitting times of Bessel processes. *Potential Anal.*, 38(3):753–786, 2013.

Pavel Chebotarev. A class of graph-geodetic distances generalizing the shortest-path and the resistance distances. *Discrete Applied Mathematics*, 159(5):295–302, 2011.

David A Croydon and Ben M Hambly. Local limit theorems for sequences of simple random walks on graphs. *Potential Analysis*, 29(4):351–389, 2008a.

David A Croydon and Ben M Hambly. Local limit theorems for sequences of simple random walks on graphs. *Potential Analysis*, 29(4):351–389, 2008b.

Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `http://igraph.org`.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

Adriano De Cezaro and B Tomas Johansson. A note on uniqueness in the identification of a spacewise dependent source and diffusion coefficient for the heat equation. *arXiv preprint arXiv:1210.7346*, 2012.

Luc P Devroye and TJ Wagner. The strong uniform consistency of nearest neighbor density estimates. *The Annals of Statistics*, pages 536–540, 1977.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

Kari Eloranta. α-congruence for markov processes. *The Annals of Probability*, pages 1583–1601, 1990.

Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, 1986.

Chris H Fleming, William F Fagan, Thomas Mueller, Kirk A Olson, Peter Leimgruber, and Justin M Calabrese. Rigorous home range estimation with movement data: a new autocorrelated kernel density estimator. *Ecology*, 96(5):1182–1188, 2015.

Kevin Françoisse, Ilkka Kivimäki, Amin Mantrach, Fabrice Rossi, and Marco Saerens. A bag-of-paths framework for network data analysis. *arXiv preprint arXiv:1302.6766*, 2013.

Keinosuke Fukunaga and Larry D Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1):32–40, 1975.

Johannes Gehrke, Paul Ginsparg, and Jon Kleinberg. Overview of the 2003 KDD Cup. *ACM SIGKDD Explorations Newsletter*, 5(2):149–151, 2003.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

NL GolâĂŹdman. Uniqueness classes in inverse problems for parabolic equations with several unknown coefficients. In *Doklady Mathematics*, volume 82, pages 573–577. Springer, 2010.

Jacob Hanna, Krishanu Saha, Bernardo Pando, Jeroen Van Zon, Christopher J Lengner, Menno P Creyghton, Alexander van Oudenaarden, and Rudolf Jaenisch. Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature*, 462(7273):595–601, 2009.

Zellig S Harris. Distributional structure. *Word*, 1954.

Tatsunori Hashimoto, Yi Sun, and Tommi Jaakkola. From random walks to distances on unweighted graphs. In *Advances in Neural Information Processing Systems*, pages 3411–3419, 2015a.

Tatsunori Hashimoto, Yi Sun, and Tommi Jaakkola. From random walks to distances on unweighted graphs. In *Advances in neural information processing systems*, 2015b.

Tatsunori B Hashimoto, Yi Sun, and Tommi S Jaakkola. Metric recovery from directed unweighted graphs. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 342–350, 2015c.

Matthias Hein, Jean-Yves Audibert, Ulrike von Luxburg, and Sanjoy Dasgupta. Graph Laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, page 2007, 2006.

Stephanie C Hicks, Mingxiang Teng, and Rafael A Irizarry. On the widespread and critical impact of systematic bias and batch effects in single-cell rna-seq data. *bioRxiv*, page 025528, 2015.

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

Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 833–840, 2002.

Elton P. Hsu. On the principle of not feeling the boundary for diffusion processes. *J. London Math. Soc. (2)*, 51(2):373–382, 1995.

Jonathan J. Hull. A database for handwritten text recognition research. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(5):550–554, 1994.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Sociëté Vaudoise des Sciences Naturelles*, 37: 547–579, 1901.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

George R Kiss, Christine Armstrong, Robert Milroy, and James Piper. An associative thesaurus of English and its computer analysis. *The Computer and Literary Studies*, pages 153–165, 1973.

Ilkka Kivimäki, Masashi Shimbo, and Marco Saerens. Developments in the theory of randomized shortest paths with a comparison of graph node distances. *Physica A: Statistical Mechanics and its Applications*, 393:600–616, 2014.

Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161 (5):1187–1201, 2015.

Gregory F Lawler and Vlada Limic. *Random walk: a modern introduction*, volume 123. Cambridge University Press, 2010.

Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

Omer Levy and Yoav Goldberg. Linguistic Regularities in Sparse and Explicit Word Representations. *Proc. 18th Conf. Comput. Nat. Lang. Learn. (CoNLL 2014)*, 2014a.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185, 2014b.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

Steven P Lund, Joseph B Hubbard, and Michael Halter. Nonparametric estimates of drift and diffusion profiles via fokker–planck algebra. *The Journal of Physical Chemistry B*, 118(44):12743–12749, 2014.

Yang Luo, Chea Lu Lim, Jennifer Nichols, Alfonso Martinez-Arias, and Lorenz Wernisch. Cell signalling regulates dynamics of nanog distribution in embryonic stem cell populations. *Journal of The Royal Society Interface*, 10(78):20120525, 2013.

Kenneth G Manton, XiLiang Gu, and Gene R Lowrimore. Cohort changes in active life expectancy in the us elderly population: experience from the 1982–2004 national long-term care survey. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 63(5):S269–S281, 2008.

Peter A Markowich and Cédric Villani. On the trend to equilibrium for the fokker-planck equation: an interplay between physics and functional analysis. *Mat. Contemp*, 19:1–29, 2000.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013b.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751, 2013c.

George Miller and Christiane Fellbaum. Wordnet: An electronic lexical database, 1998.

Paul Mineiro, Javier Movellan, and Ruth J Williams. Learning path distributions using nonequilibrium diffusion networks. *Advances in neural information processing systems*, pages 598–604, 1998.

SA Molchanov. Diffusion processes and riemannian geometry. *Russian Mathematical Surveys*, 30(1):1, 1975.

Rob Morris, Ignacio Sancho-Martinez, Tatyana O Sharpee, and Juan Carlos Izpisua Belmonte. Mathematical approaches to modeling development and reprogramming. *Proceedings of the National Academy of Sciences*, 111(14):5076–5082, 2014.

Peter Mörters and Yuval Peres. *Brownian motion.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2010. With an appendix by Oded Schramm and Wendelin Werner.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407, 2004.

Bernt Øksendal. *Stochastic differential equations: An introduction with applications.* Universitext. Springer-Verlag, Berlin, sixth edition, 2003.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. 1999.

Grigorios A Pavliotis. Stochastic processes and applications. *Diffusion Processes, the Fokker-Planck*, 2014.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12, 2014.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

Xiaojie Qiu, Shanshan Ding, and Tieliu Shi. From understanding the development landscape of the canonical fate-switch pair to constructing a dynamic landscape for two-step neural differentiation. *PLoS One*, 7(12), 2012.

Hannes Risken. *Fokker-Planck Equation.* Springer, 1984.

Jean-Marie Robine and Jean-Pierre Michel. Looking forward to a general theory on population aging. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 59(6):M590–M597, 2004.

David E Rumelhart and Adele A Abrahamson. A model for analogical reasoning. *Cognitive Psychology*, 5(1):1–28, 1973.

Laurent Saloff-Coste. The heat kernel and its estimates. *Probabilistic approach to geometry*, 57:405–436, 2010.

Purnamrita Sarkar and Andrew W Moore. A tractable approach to finding closest truncated-commute-time neighbors in large graphs. In *In Proc. UAI*, 2007.

Purnamrita Sarkar, Deepayan Chakrabarti, and Andrew W Moore. Theoretical justification of popular link prediction heuristics. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 2722, 2011.

Gideon Schwarz and Amos Tversky. On the reciprocity of proximity relations. *Journal of Mathematical Psychology*, 22(3):157–175, 1980.

Blake Shaw and Tony Jebara. Structure preserving embedding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 937–944. ACM, 2009.

Robin Sibson. Studies in the robustness of multidimensional scaling: Perturbational analysis of classical scaling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 217–229, 1979.

Daniel R Sisan, Michael Halter, Joseph B Hubbard, and Anne L Plant. Predicting rates of cell state change caused by stochastic fluctuations using a data-driven landscape model. *Proceedings of the National Academy of Sciences*, 109(47):19262–19267, 2012.

Steven T Smith, Edward K Kao, Kenneth D Senne, Garrett Bernstein, and Scott Philips. Bayesian discovery of threat networks. *IEEE Transactions on Signal Processing*, 62:5324–5338, 2014.

Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. Parsing with compositional vector grammars. In *In Proceedings of the ACL conference*. Citeseer, 2013.

Jascha Sohl-Dickstein, Eric A Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015.

Robert J Sternberg and Michael K Gardner. Unities in inductive reasoning. *Journal of Experimental Psychology: General*, 112(1):80, 1983.

Daniel Stroock and S.R.S. Varadhan. Diffusion processes with boundary conditions. *Communications on Pure and Applied Mathematics*, 24:147–225, 1971a.

Daniel W Stroock and SR Srinivasa Varadhan. Diffusion processes with boundary conditions. *Communications on Pure and Applied Mathematics*, 24(2):147–225, 1971b.

Daniel W Stroock and SR Srinivasa Varadhan. *Multidimensional diffussion processes*, volume 233. Springer Science & Business Media, 1979.

Alireza Tahbaz-Salehi and Ali Jadbabaie. A one-parameter family of distributed consensus algorithms with boundary: From shortest paths to mean hitting times. In *Decision and Control, 2006 45th IEEE Conference on*, pages 4664–4669. IEEE, 2006.

Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, 2005.

Joshua B Tenenbaum. Mapping a manifold of perceptual observations. *Advances in neural information processing systems*, pages 682–688, 1998.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Valery Tereshko. Reaction-diffusion model of a honeybee colonyâĂŹs foraging behaviour. In *Parallel Problem Solving from Nature PPSN VI*, pages 807–816. Springer, 2000.

Daniel Ting, Ling Huang, and Michael I Jordan. An analysis of the convergence of graph Laplacians. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1079–1086, 2010a.

Daniel Ting, Ling Huang, and Michael I Jordan. An analysis of the convergence of graph Laplacians. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1079–1086, 2010b.

Daniel Ting, Ling Huang, and Michael Jordan. An analysis of the convergence of graph laplacians. *arXiv preprint arXiv:1101.5435*, 2011.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature biotechnology*, 32(4):381, 2014.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

Peter D Turney and Michael L Littman. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1-3):251–278, 2005.

Amos Tversky and J Hutchinson. Nearest neighbor analysis of psychological spaces. *Psychological Review*, 93(1):3, 1986.

S. R. S. Varadhan. Diffusion processes in a small time interval. *Comm. Pure Appl. Math.*, 20:659–685, 1967a.

Srinivasa RS Varadhan. Diffusion processes in a small time interval. *Communications on Pure and Applied Mathematics*, 20(4):659–685, 1967b.

Ulrike von Luxburg and Morteza Alamgir. Density estimation from unweighted k-nearest neighbor graphs: a roadmap. In *Advances in Neural Information Processing Systems*, pages 225–233. Springer, 2013.

Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008.

Ulrike von Luxburg, Agnes Radl, and Matthias Hein. Hitting and commute times in large random neighborhood graphs. *Journal of Machine Learning Research*, 15: 1751–1798, 2014.

Conrad Hal Waddington et al. Organisers and genes. *Organisers and Genes.*, 1940.

Ward Whitt. Some useful functions for functional limit theorems. *Math. Oper. Res.*, 5(1):67–85, 1980. ISSN 0364-765X. doi: 10.1287/moor.5.1.67. URL `http://dx.doi.org/10.1287/moor.5.1.67`.

Wolfgang Woess. Random walks on infinite graphs and groups - a survey on selected topics. *Bull. London Math. Soc*, 26:1âĂŞ60, 1994.

Majid Yazdani. *Similarity Learning Over Large Collaborative Networks.* PhD thesis, École Polytechnique Fédérale de Lausanne, 2013.

Luh Yen, Marco Saerens, Amin Mantrach, and Masashi Shimbo. A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–793. ACM, 2008.

Chuancun Yin. The joint distribution of the hitting time and place to a sphere or spherical shell for brownian motion with drift. *Statistics & Probability Letters*, 42 (4):367–373, 1999.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning from labeled and unlabeled data on a directed graph. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 1036–1043. ACM, 2005.