Graph Learning for Inverse Landscape Genetics

Prathamesh Dharangutte, Christopher Musco
Dept. of Computer Science and Engineering, New York University, Brooklyn, New York
prathamesh.d, cmusco@nyu.edu

Abstract

The problem of inferring unknown graph edges from numerical data at a graph's nodes appears in many forms across machine learning. We study a version of this problem that arises in the field of *landscape genetics*, where genetic similarity between organisms living in a heterogeneous landscape is explained by a weighted graph that encodes the ease of dispersal through that landscape. Our main contribution is an efficient algorithm for *inverse landscape genetics*, which is the task of inferring this graph from measurements of genetic similarity at different locations (graph nodes).

Inverse landscape genetics is important in discovering impediments to species dispersal that threaten biodiversity and long-term species survival. In particular, it is widely used to study the effects of climate change and human development. Drawing on influential work that models organism dispersal using graph *effective resistances* (McRae 2006), we reduce the inverse landscape genetics problem to that of inferring graph edges from noisy measurements of these resistances, which can be obtained from genetic similarity data.

Building on the NeurIPS 2018 work of Hoskins et al. (2018) on learning edges in social networks, we develop an efficient first-order optimization method for solving this problem. Despite its non-convex nature, experiments on synthetic and real genetic data establish that our method provides fast and reliable convergence, significantly outperforming existing heuristics used in the field. By providing researchers with a powerful, general purpose algorithmic tool, we hope our work will have a positive impact on accelerating work on landscape genetics.

Introduction

Many datasets can be modeled as a weighted, undirected graph: $G = (V, E)$ with nodes $V = \{v_1, \ldots, v_n\}$ and additional numerical data vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ at each node. For example, in social networks, each node is a user, each edge is a connection or interaction between users, and x_i might contain demographic information about user i like age, gender, or expressed political party.

Often, node data is correlated with G's *connectivity structure*: if v_i and v_j are strongly connected, x_i and x_j tend to be more similar than for poorly connected nodes (Kalofolias 2016; Ortega et al. 2018). Formally, connectivity between two nodes can be quantified in many of ways, from simple statistics like shortest path distance or number of common neighbors, to more advanced metrics like personalized PageRank (Page et al. 1999; Jeh and Widom 2003), SimRank (Jeh and Widom 2002), or DeepWalk distance (Perozzi, Al-Rfou, and Skiena 2014). While these measures depend solely on G's structure (i.e. edges and their weights), they often align with measured similarities between $x_1, \ldots, x_n$.

This observation leads to an interesting possibility: even when edges in G are *unknown*, node data can be useful in *inferring edges and weights*, or at least in inferring a graph *whose connectivity structure is consistent with the observed data*. This possibility has been explored across statistics, machine learning, and network science (Raskutti et al. 2009; Cai, Liu, and Luo 2011; Egilmez, Pavez, and Ortega 2017; Liben-Nowell and Kleinberg 2007; Hoskins et al. 2018). In many cases, pairwise measures of connectivity can reveal a striking amount of information about G, and by proxy, so can similarity information between $x_1, \ldots, x_n$ (Hoskins et al. 2018). Applications of graph inference from node data include understanding structured statistical correlation, link prediction, and phylogeny reconstruction.

In this work, we examine an application of graph inference in *landscape genetics*, a field at the intersection of landscape ecology, spatial statistics, and population genetics (Manel et al. 2003; Sanderson 2020). Landscape genetics seeks to explain genetic differences between populations of the same species that live at different geographic locations. The goal is to understand how ease of movement between these geographic locations (i.e., through the landscape) affects population genetics. Geographically isolated populations tend to differ genetically, whereas ease of travel and intermixing between populations leads to genetic similarity.

Early methods in landscape genetics correlate genetic similarity with simple measures of geographic isolation, like the Euclidean distance between populations (Wright 1943; Sokal and Oden 1978), or distance along an oriented direction or curve, leading to concepts like clines and ring species (Endler 1977; Huggett 2004). These tried-and-true approaches have been successfully applied to understanding genetic variation in a variety of species, including humans (Novembre et al. 2008). More recently, however, work in landscape genetics considers finer-grained measures of landscape-driven isolation, largely based on modeling the landscape as an undirected graph (the *landscape graph*). Each location (spatial

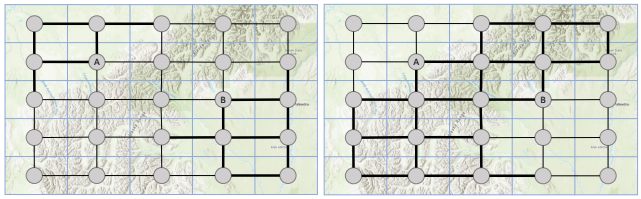


Figure 1: Example landscape graphs for two species, with edge thickness indicating edge weight. In the first graph, edges in low-altitude areas have higher weight, so this graph is apt for a species that prefers low-land habitat. The second graph is natural for a species that prefers high altitudes. We expect higher genetic similarity between populations at nodes A and B for the second graph, due to better connectivity.

cell) in the landscape is associated with a graph node, and each node is connected by a weighted edge to all geographically adjacent nodes (see Fig. 1). Edge weights are chosen to reflect the ease of organism dispersal between adjacent nodes: we follow the convention that high weight indicates ease of dispersal and low weight indicates inhibition to movement, although note that the opposite meaning is sometimes used (Coulon et al. 2004). Weights are tailored to specific species: e.g., an edge across a span of water would have low weight for a ground-dwelling species which cannot easily traverse the edge. For an organism that prefers low-land environments, edges crossing areas of high elevation might receive lower weight than those crossing low-land areas.

In addition to the landscape graph, we consider numerical genetic data about populations of organisms living at different nodes of the graph. Usually this data is sparse, meaning we only have information for a subset of nodes (Oyler-McCance, Fedy, and Landguth 2013). Regardless, the goal is to correlate pairwise genetic similarity between these nodes with pairwise connectivity in the underlying landscape graph. For example, the weight of the least cost path between two nodes is a common connectivity measure, and shown to correlate with genetic similarity, measured using e.g., the fixation index (Arnaud 2003; Coulon et al. 2004; Vignieri 2005). More recently, McRae's influential paper *Isolation by Resistance* popularized the use of *effective resistance distance* as a connectivity measure in landscape genetics (McRae 2006; Yen et al. 2007). Effective resistances better model organism dispersal, and thus correlate more closely with genetic differences across landscapes (McRae and Beier 2007).

Amongst many other applications, effective resistance-based landscape ecology has been important in understanding the effects of climate change on species dispersal and migration (McRae, Shah, and Edelman 2016). [1]

Our Contributions

So where does graph inference come in? Most studies that use landscape graphs to model species dispersal construct these graphs based on *expert knowledge* (McRae 2006; Shirk et al. 2010). Knowledge of a species' behavioral preferences (e.g, preferred elevation, vegetation cover, or climate) are used to determine edge weights, which are then used to compute pairwise connectivities like least cost paths or effective resistances. Multiple landscape graphs proposed by experts can be tested for fit (Vos et al. 2001; Lugon-Moulin and Hausser 2002; Vignieri 2005; Short Bull et al. 2011), but achieving high levels of correlation with genetic data requires significant background information on a species (which may be imperfect) and laborious hand-tuning of the landscape graph.

Inverse landscape genetics. To address this issue, there has been interest in moving beyond expert opinion, by *algorithmically* determining optimal edge weights (Zeller, McGarigal, and Whiteley 2012; Peterman et al. 2019). Specifically, the goal is to learn a function that maps measurable landscape parameters for each edge (e.g. what vegetation cover it goes through, or if there is human development along the edge) to edge weights. The resulting weighted graph should have connectivity structure that correlates as well as possible with genetic differences across the landscape.

We call this parameterized graph inference problem *inverse landscape genetics*. Not only does this exciting problem offer the possibility of refining expert-designed landscape graphs, but a solution would allow ecologist to infer information about species dispersal based *purely on collected genetic data* (Oyler-McCance, Fedy, and Landguth 2013), as opposed to the traditional perspective of explaining genetic data with known ecological knowledge. Genetic information could be used to understand species habitat preferences, find bottlenecks in migration, or understand how human development is impeding species movement (McRae, Shah, and Edelman 2016). As discussed in Zeller, McGarigal, and Whiteley (2012) and Graves, Beier, and Royle (2013), algorithms for learning landscape graphs from data could therefore be essential in future conservation and planning decisions involving e.g. wildlife corridor design.

However, despite interest in the inverse landscape genetics problem, few effective algorithms have been developed to solve it. Zeller, McGarigal, and Whiteley (2012) surveys of existing techniques. Most current approaches optimize landscape graphs (i.e. find a graph consistent with observed genetic data) using variants of brute force search. For example, a common approach is to rely on expert opinion to obtain an initial graph and then search over a small set of nearby weight functions to improve the fit (Shirk et al. 2010). There has been some work on more systematic algorithms. Peterman (2018) introduce a framework for optimizing landscape graphs using a genetic algorithm and compare their method with other approaches (Peterman et al. 2019). Graves, Beier, and Royle (2013) develop an approach based on local search heuristics, using Nelder-Mead and Newton line search algorithms to optimize landscape graphs.

A differentiable approach. Our main contribution is to show that one of the most common formalizations of the inverse landscape genetics problem can be solved efficiently and reliably using *gradient based* optimization methods. In particular, we consider a version of the problem which correlates the *effective resistance* between two nodes (a measure of graph connectivity) with the *fixation index* between genetic data at those nodes (a measure of genetic differentiation). We

[1] See https://circuitscape.org/pubs.html for further details.

build on recent work of Hoskins et al. (2018) that studies the problem of learning graph edges based on noisy measurements of effective resistances in the graph. As in that result, we show how to compute a gradient for an appropriately chosen graph-learning loss involving the effective resistances, and in our case, fixation index values. To do so, we need to differentiate through the effective resistances computation, which involves the pseudoinverse of a graph Laplacian. We implement this step efficiently using an iterative linear system solver for positive semidefinite matrices. Our approach is detailed in the Proposed Method section.

To the best of our knowledge, our method is the first for the inverse landscape genetics problem that uses a gradient based optimization method. In the Empirical Results section, we compare it against local search heuristics used in prior work (Graves, Beier, and Royle 2013), showing that it obtains much more reliable convergence on both synthetic and real-world data sets. As an application of our fast algorithm, we are able to explore questions of statistical complexity that have been raised in the landscape genetics literature (Oyler-McCance, Fedy, and Landguth 2013). In particular, there are concerns that algorithmic methods might overfit the landscape graph if learned using genetic data from an insufficient number of nodes. By varying the amount of data available in a sequence of large synthetic data experiments, we empirically explore the precise number of samples required to obtain a generalizing solution, showing that in some cases, as few as 25 populations are needed to reliably fit the parameters of a landscape graph involving 1000s of nodes.

Additional related work. Relevant work on landscape genetics is included in Our Contributions section . We discuss additional related work on graph learning in Appendix C of this paper's full version (Dharangutte and Musco 2020).

Proposed Method

We first describe notation needed to formalize the *inverse landscape genetics* problem from previous section.

Graph and genetic data notation: We denote the weighted, undirected landscape graph by $G = (V, E, w)$, where $V = \{v_1, \ldots, v_n\}$ is the vertex set, E is the edge set, and w is a vector of weights assigned to each edge. Let m denote $m = |E|$. Typically $m \ll \binom{n}{2}$ since for most landscapes G will be a grid graph with $m = O(n)$. We index both E and w by their terminal nodes: edges are $e_{i_1 j_1}, \ldots, e_{i_m j_m}$ and weights are $w_{i_1 j_1}, \ldots, w_{i_m j_m}$. It is often helpful to view graphs as electrical networks where e_{ij} represents an electrical connection with *conductance* w_{ij} between nodes v_i and v_j (Spielman and Srivastava 2011). Let $r_{ij} = 1/w_{ij}$ denote the resistance of the connection.

For a subset $S \subseteq V$ of nodes we have measured vectors of population genetic data $x_1, \ldots, x_{|S|} \in \mathbb{R}^d$. We only interact with this data through a black-box measure of genetic *dissimilarity*: the specific choice is not important. In keeping with prior work, our experiments use the fixation index, typically denoted F_{ST}. For two populations, i and j a high F_{ST} (close to 1) indicates greater difference between the measured genetic information in x_i and x_j. Let $F \in \mathbb{R}^{|S| \times |S|}$ contain pairwise F_{ST} (or another dissimilarity) for all nodes in S. Let $F_{i,i} = 0$ for all diagonal entries.

It has been established that the values in F will correlate well with the *effective resistances* of an appropriately chosen landscape graph G (McRae 2006). To define these measures, let $D \in \mathbb{R}_+^{n \times n}$ be the diagonal degree matrix with $D_{i,i} = \sum_{j : e_{ij} \in E} w_{ij}$. Let A be the adjacency matrix with $A_{ji} = A_{ij} = w_{ij}$ for all $e_{ij} \in E$, and 0 otherwise. Let L be the weighted graph Laplacian as $D - A$.

Definition 1 (Effective resistance)**.** *The effective resistance R_{ij} between two nodes i and j satisfies*

$$R_{ij} = b_{ij}^T L^+ b_{ij}$$

where L^+ is the Moore-Pensore pseudoinverse of the graph laplacian L and $b_{ij} \in \mathbb{R}^n$ is the vector with 1 at position i, -1 at position j and 0's elsewhere.

The effective resistance between two nodes v_i and v_j is lower when there exist more low-resistance paths (i.e., high weight paths) between v_i and v_j. It is known to be equal to the *commute time* between v_i and v_j for a random walk with steps taken proportional to edges weights (Chandra et al. 1996), which gives some intuition for why the measure effectively quantifies organism dispersal through a landscape. We refer the reader to McRae, Shah, and Edelman (2016) for further discussion of the important of effective resistances in landscape ecology.

Let R be the matrix of all pairwise effective resistances and note that $R_{ij} = R_{ji}$, which can be thought of as the resistance surface for the landscape. R is 0 along its diagonal. Let $R_S \in |S| \times |S|$ be the principal submatrix of R containing only the pairwise effective resistances between nodes in S. The main problem we study is as follows:

Problem 1 (Inverse Landscape Genetics)**.** *Given landscape graph nodes V and edges E we are given a vector of environmental parameters $C_{i_k j_k} \in \mathbb{R}^q$ for each $e_{i_k j_k} \in E$ and a function class $\mathcal{P}$ from $\mathbb{R}^q \to \mathbb{R}_+$ which maps these parameters to a weight for each edge. Assume $\mathcal{P}$ is parameterized by parameters θ and denote functions in the class by $p_\theta \in \mathcal{P}$. For p_θ, let $p_\theta(E) = [p_\theta(C_{i_1 j_1}), \ldots, p_\theta(C_{i_m j_m})]$. Our goal is to find $\hat{\theta}$ minimizing the loss:*

$$\hat{\theta} = \arg\min_\theta \mathcal{L}(\theta) = \arg\min_\theta \|R_S(p_\theta(E)) - F\|_F^2, \tag{1}$$

where $R_S(p_\theta(E))$ is the effective resistance matrix for the graph $G = (V, E, p_\theta(E))$ (restricted to nodes in S). $\|A\|_F^2 = \sum_i \sum_j A_{ij}^2$ denotes the standard Frobenius norm.

Note that both $R_S(p_\theta(E))$ and F have zeros on the diagonal, so the Frobenius norm above is equal to $2\times$ the standard squared loss between effective resistances and genetic dissimilarities. Other natural choices could be used instead of $\mathcal{L}$, e.g. the inverse of the Mantel correlation between $R_S(p_\theta(E))$ and F (Graves, Beier, and Royle 2013). In either case, the goal is to find edge weights such that the landscape graph G

induces effective resistances between nodes in S which are as close as possible to the genetic dissimilarities in F. Alternatively, under the assumption that genetic dissimilarities represent noisy measurements of the *true effective resistances for some unknown landscape graph* G^*, then Problem 1 can be viewed as the task of recovering that graph.

Example Functional Forms

The problem is stated under the constraint that weights in the learned graph are a function p_θ of q environmental parameters $C_{i_k j_k}$ about each edge. This function can take any form: we only require that it is differentiable with respect to its parameters. For example, prior work often considers $C_{i_k j_k}$ which is a single continuous scale parameter like edge elevation or temperature. A typical choice (see e.g. (Graves, Beier, and Royle 2013)) is to assume that $1/w_{i_k j_k} = r_{i_k j_k}$ follows an inverted Gaussian relation governed by parameters $\theta = [\beta, \beta_{opt} \text{ and } \beta_{SD}]$:

$$\frac{1}{w_{i_k j_k}} = r_{i_k j_k} = \beta + 1 - \beta \exp\left(\frac{-(C_{i_k j_k} - \beta_{\text{opt}})^2}{2\beta_{\text{SD}}^2}\right) \quad (2)$$

This form captures the fact that many species have for example a preferred "ideal" elevation β_{opt} and are more likely to travel along edges of similar elevation: the resistance to dispersal $r_{i_k j_k}$ increases as $C_{i_k j_k}$ moves further from β_{opt}. Other papers consider slightly different functions, but they typically have the same general structure as (2) (Peterman 2018).

Another common functional form is linear. We simply let:

$$\frac{1}{w_{i_k j_k}} = r_{i_k j_k} = \alpha^T C_{i_k j_k}. \quad (3)$$

For instance $C_{i_k j_k}$ might contain one-hot-encoded categorical data indicating what landcover an edge traverses (e.g. water, marshland, tundra). Each entry in α is a scalar associated with each category type that conveys how permeable the category is for movement. When continuous and discrete data at nodes is considered in unison, it is natural to add multiple functional forms linearly: e.g. we might have that $r_{i_k j_k} = r^{\text{E}}_{i_k j_k} + r^{\text{LC}}_{i_k j_k}$ where $r^{\text{E}}_{i_k j_k}$ is an elevation term in the form of (2) and $r^{\text{LC}}_{i_k j_k}$ is a landcover term in the form of (3).

Gradient Computation

Due to its non-convex nature, these is no closed form solution for (1). The cornerstone of our approach is to instead find approximate solution by using projected gradient descent to minimize $\mathcal{L}(\theta)$. To do so, we need an efficient method for computing the gradient of this loss.

Proposition 1. *Let n_θ denote the number of parameters in θ (typically a small constant) and let $J \in \mathbb{R}^{m \times n_\theta}$ denote the Jacobian with $J_{k,h} = \frac{\partial w_{i_k j_k}}{\partial \theta_h}$. Let $B \in \mathbb{R}^{m \times n}$ denote the edge-vertex incidence matrix of G with k^{th} row equal to $e_{i_k} - e_{j_k}$ where i_k and j_k are the terminal nodes of G's k^{th} edge.*

$$\nabla_\theta(\mathcal{L}) = \sum_{v_l, v_k \in S} \left(F_{lk} - b_{lk}^T L_\theta^+ b_{lk}\right) \cdot 2J^T \cdot (BL_\theta^+ b_{lk})^{\circ 2},$$

where $^{\circ 2}$ denote the Hadamard power (i.e. square every vector element entrywise) and L_θ denotes the Laplacian of the landscape graph with edge weights $p_\theta(E)$

Proof. For given parameters θ, let $w^\theta = p_\theta(E)$, where $p_\theta(E)$ is as defined in Problem 1. We have:

$$\nabla_{w^\theta} \mathcal{L} = -2 \sum_{v_l, v_k \in S} \left(F_{lk} - R(w^\theta)_{lk}\right) \cdot \nabla_{w^\theta} R(w^\theta)_{lk} \quad (4)$$

As in Problem 1, $R(w^\theta)$ is the matrix of all pairwise effective resistances for the graph $G = (V, E, w^\theta)$. From the definition for effective resistance, $R(w^\theta)_{lk} = b_{lk}^T L_\theta^+ b_{lk}$. As in (Hoskins et al. 2018), we can obtain a partial derivative for entries of L^+ with respect to w^θ via the Sherman-Morrison formula for rank one updates to the pseudoinverse. Specifically, we have $\frac{\partial L_\theta^+}{\partial w_{ij}^\theta} = -L_\theta^+ b_{ij} b_{ij}^T L_\theta^+$ and thus

$$\frac{\partial R(w^\theta)_{lk}}{\partial w_{ij}^\theta} = -b_{lk}^T \left(L_\theta^+ b_{ij} b_{ij}^T L_\theta^+\right) b_{lk} = -(b_{ij}^T L_\theta^+ b_{lk})^2$$

It follows that $\nabla_{w^\theta} R(w^\theta)_{lk} = -(BL_\theta^+ b_{lk})^{\circ 2}$. The proposition follows from plugging this equation into (4) and noting that $\nabla_\theta(\mathcal{L}) = J^T \cdot \nabla_{w^\theta} \mathcal{L}$. □

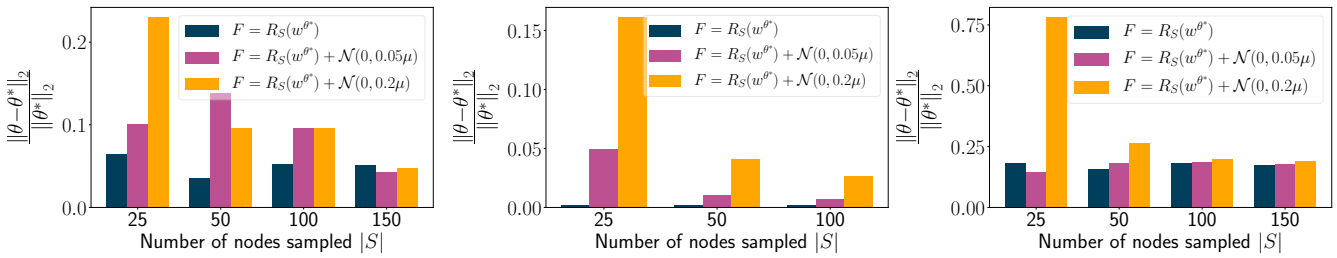
Efficient computation of the gradient: Proposition 1 yields an efficient algorithm for computing $\nabla_\theta \mathcal{L}$. In particular, since n_θ is typically a small constant computing the Jacobian J is efficient for any differentiable functional form p_θ. Then, ignoring the cost of computing $b_{lk}^T L_\theta^+ = L_\theta^+ b_{lk}$ for all $v_l, v_k \in S$, the gradient can be computed in $O(|S|^2 \cdot m \cdot n_\theta)$ time. Note that since every row in B is 2 sparse, $b_{lk}^T L^+ B$ can be computed in $O(m)$ time once $b_{lk}^T L^+$ is computed. Since $m = O(n)$ in most landscape genetics applications, the bottle neck is therefore computing each $L_\theta^+ b_{lk}$.

This would naively require inverting the $n \times n$ Laplacian L_θ, which would be computationally intensive and impractical for large graphs. We instead approximate the matrix-vector product $L_\theta^+ b_{lk}$ using an iterative solver for positive semidefinite linear systems (L is positive semidefinite). In our experiments we use the standard MINRES method. To optimize the approach further, we note that $L_\theta^+ b_{lk} = L_\theta^+ e_l - L_\theta^+ e_k$ where e_l and e_k are the l^{th} and k^{th} standard basis vectors. Accordingly, we only need to solve $|S|$ linear systems (either e_l as the right hand side for all $v_l \in S$), and can then recombine those solutions to return all $\binom{|S|}{2}$ vectors $L(w)^+ b_{lk}$ needed for the gradient computation.

Each linear system solve could be further optimized by constructing e.g., a multigrid or partial Choleksy preconditioner. However, we found that the MINRES converged quickly in experiments without preconditioning, so the possible improvement is relatively small.

Empirical Results

With an efficient gradient oracle in hand for the loss function in Problem 1, we test a gradient based optimization approach on both synthetic and real genetic data. Real genetic data is obtained for the North American wolverine (Gulo gulo) from Kyle and Strobeck (2001), which provides F_{ST} values for



(a) Discrete and combined case (b) Discrete case (landcover data) (c) Continuous case (elevation data)

Figure 2: Relative error between recovered parameters and true parameters for synthetic data experiments with different numbers of nodes sampled N and noise standard deviation $\tilde{\sigma}$. Parameter recovery improves with more samples (i.e., more locations with genetic similarity data), and generally with less noise (i.e., more highly correlated resistance and genetic data).

6 populations living across a region in Alaska. Our goal is to understand the interplay between genetic variation in this region and the underlying landscape. Specifically, we obtain elevation data [2] and land cover data (Homer et al. 2020), which will be used as the basis for selecting edge weights in a landscape graph.

The landscape graph graph is constructed by dividing the Alaska region into a grid of square cells. In previous landscape genetics studies of the North American wolverine, cell sizes of 5 km and 50 km have been used (McRae and Beier 2007). We choose a resolution of 15 km, which lead to a graph $G = (V, E)$ with $|V| = 24035$ and $|E| = 47746$. For each cell we create a node in the grid graph, and connect adjacent nodes with edges (as in Figure 1). Our landscape data comes as raster images, with each pixel corresponding to a region of 100×100 meters for elevation data and 30×30 meters for landcover data, so we have multiple pixels of information within each landscape cell. This data was resampled to cell resolution using standard GIS methods (see Appendix A in (Dharangutte and Musco 2020) for details).

Continuous and discrete environmental parameters are then collected for each edge in the graph. For edge k, edge elevation $C^{\text{E}}_{i_k j_k}$ is taken as the average elevation at cells i and j and scaled to lie within the range 0-10. For each edge we also construct a vector of one-hot-encoded landcover data $C^{\text{LC}}_{i_k j_k}$, which has 17 entries for landcover types like evergreen forest, barren land, or open water. Each entry in $C^{\text{LC}}_{i_k j_k}$ is given values as follows: 0 if the landcover type is absent at cell i and j, 0.5 if present at either cell i or j, or 1 if present at both cells i and j. We model edge weights as a function of these parameters by linearly combining equation (2) for elevation data and (3) for landcover data. So, the final parameter vector we hope to learn when solving Problem 1 is $\theta = \{\beta, \beta_{\text{opt}}, \beta_{\text{SD}}, \alpha \in \mathbb{R}^{17}\}$.

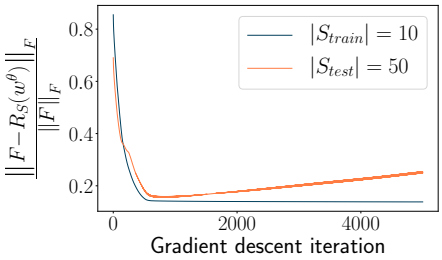
To minimize (1), we implement a projected gradient descent method with RMSProp step size adjustment, which adjusts learning rate by a decaying average of squared gradients (Tieleman and Hinton 2012). Since edge weights are constrained to be non-negative, and all edge data is non-negative, we project parameters to $\max(\epsilon, \theta)$ with $0 < \epsilon \leq 1$ at each gradient step. This ensures non-zero resistance value for all landcover types, which is a constraint often imposed in prior work. All experiments were run on server with 2vCPU @2.2GHz and 13 GB main memory.

Synthetic data: Our first set of data experiments uses the *real landscape data* from Alaska, but in conjunction with carefully *simulated genetic data*, which makes it possible to better assess the performance of our method. Specifically, we select a random 50×50 subgrid of our Alaska graph to obtain a grid graph with $|V| = 2500$ and $|E| = 4900$. We then construct a *ground truth* graph by randomly sampling a set of parameters, θ^*, and evaluating the weights for all edges in E. The goal in our synthetic experiments is to recover this ground truth, which is a common set up in testing algorithms for inverse landscape genetics as real ground truth data is never available (Graves, Beier, and Royle 2013).

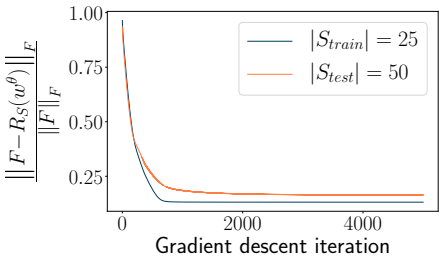
In particular, we construct the pairwise effective resistance matrix $R_S(w^{\theta^*})$ for a set of nodes S with $N = |S| \ll |V|$. For the nodes in S, we produce a simulated genetic similarity matrix F by setting $F_{lk} = \left[R_S(w^{\theta^*})\right]_{lk} + \tilde{z}$ where $\tilde{z} \sim \mathcal{N}(0, \tilde{\sigma})$. We run experiments with $\tilde{\sigma} = \{0, 0.05\mu, 0.2\mu\}$, where μ is the mean of the resistances in $R_S(w^{\theta^*})$. These cases (no, low, and high noise) range from perfect to poor alignment between genetic data and landscape resistance. For parameters θ obtained after optimization, we report the relative parameter error as $\|\theta - \theta^*\|_2 / \|\theta^*\|_2$. We ignore parameters for landcover types present at less than 1% of nodes as these parameters can't be determined with any level of accuracy (since they have essentially no impact on graph effective resistances). Results are shown in Fig. 2, with additional experiments in (Dharangutte and Musco 2020).

We conclude that, as N increases, our method obtains high quality approximations to the true parameters θ^*, even in the high noise regime. For example, $N = 150$ was sufficient for fitting the graph parameters in all cases. This is a pretty typical number of samples for a landscape genetics study (e.g. Shirk et al. (2010) obtain genetic data for mountain goats from $N = 149$ locations over a comparably sized area). Accordingly, even for a reasonably large number of landscape parameters, reliable learning of landscape data should be possible with existing data collection methods.

[2] National Atlas of the United States. (2012). 100-Meter Resolution Elevation of Alaska, Albers projection. National Atlas of the United States. Available at: http://purl.stanford.edu/sg962yb7367.

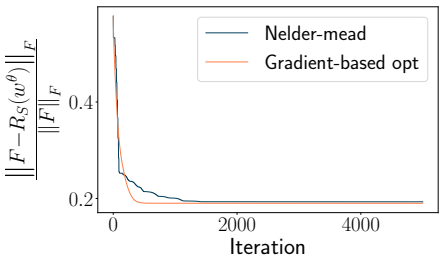


(a) Relative loss. vs iteration for $N = 10$

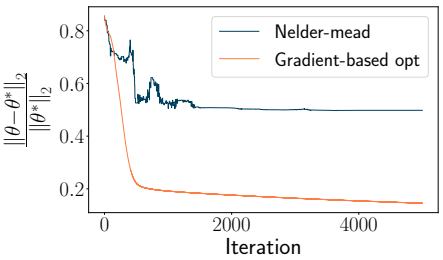


(b) Relative loss vs. iteration for $N = 25$

Figure 3: Train and test loss for different values of N on synthetic data with $\tilde{\sigma} = 0.2\mu$. Parameters are learnt for nodes belonging to S_{train} and used to infer pairwise effective resistance for nodes in S_{test}. We obtain good generalization for N as low as 25, but observe clear overfitting for $N = 10$.



(a) Relative loss vs. iteration for Nelder-Mead and gradient-based optimization. After 5000 iterations, the loss value is 0.193 for Nelder-Mead and 0.19 for gradient-based optimization.



(b) Relative parameter error between recovered parameters and true parameters with iteration. Gradient-based optimization is better at recovering true parameters.

Figure 4: Comparison of proposed method to a heuristic optimization technique. Gradient-based optimization is faster in convergence and better at recovering true parameters with enough data. Experiments are for synthetic data with high noise setting with $N = 150$ and $\tilde{\sigma} = 0.2\mu$, where μ is mean of entries in true resistance surface $R_S(w^{\theta^*})$ corresponding to nodes in S.

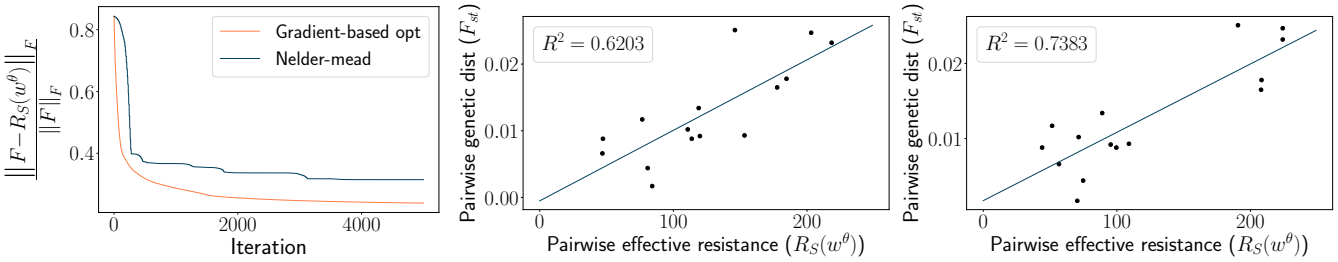
Addressing overfitting: It has been reported that a potential concern with optimizing landscape graphs is overfitting when N is small. I.e., the landscape graph fit to F does not generalize to new data (Oyler-McCance, Fedy, and Landguth 2013). To validate against overfitting, we randomly split nodes into sets S_{train} and S_{test}. We learn parameters θ for nodes in S_{train} and evaluate these parameters against pairwise effective resistances in S_{test}. Even in the high noise setting, with $\tilde{\sigma} = 0.2\mu$, test loss converges along with train loss when N is as low as 25, (Fig. 3). This implies good generalization and a lack of overfitting, *even though we do not accurately recover all parameters in* θ^*. This is not necessarily surprisingly: it indicates that, while the inverse landscape genetics problem may be poorly conditioned with respect to θ (as observed in Graves, Beier, and Royle (2013)) it is still possible to obtain reliable predictive models with little data.

Comparison with existing approaches : We compare gradient-based optimization to the Nelder-Mead method [3], which has been used in prior work on inverse landscape genetics (Graves, Beier, and Royle 2013). We observe that our method is faster in terms of convergence and also better at

[3]Note that Nelder-Mead is an unconstrained optimization method, so we add a projection step to ensure interpretable parameters are returned. This does not noticeably affect the behavior of convergence in our experiments.

Parameter	Nelder Mead	Gradient based optimization
β	0	0
β_{opt}	10	9
β_{SD}	0	0
Open water	227	502
Barren Land	151	5
Deciduous forest	0	0
Evergreen forest	0	12
Mixed forest	0	0
Dwarf Shrub	18	0
Shrub/Scrub	107	95
Sedge/Herbaceous	0	500
Woody Wetlands	25	26

Table 1: Final parameter values after optimization, rounded to nearest whole number. We do not report for landcover types which were present at less that 2% of nodes in the graph. The β parameters are for elevation data – see equation (2).



(a) Relative objective function value vs iteration. Gradient-based optimization obtains a better solution faster.

(b) F_{ST} vs effective resistance from learnt parameters using Nelder-Mead algorithm.

(c) F_{ST} vs effective resistance from learnt parameters using gradient-based optimization.

Figure 5: Relative objective function value and R^2 values computed for a linear fit between F_{ST} and effective resistances in the final learnt landscape graphs for real-world data. Gradient-based optimization obtains a slightly better fit.

recovering true parameters with enough data. Nelder-Mead eventually achieves comparable performance in terms of train loss but fails at recovering the true parameters (Figure 4). To ensure a fair comparison, we choose the same random initialization of parameters and non-negativity constraints.

North American wolverine (gulo gulo) : For experiments on real data, F_{ST} values range from 0 to 1 and we have access to genetic data at 15 nodes out of 24035 nodes. After fitting θ with our gradient based method, we compute the R^2 value for a linear fit between recovered resistances and F_{ST} values (Fig. 5), a metric used in prior work (McRae and Beier 2007). We obtain an R^2 value of 0.7383 using gradient-based optimization and 0.6203 using Nelder-mead, in comparison to 0.68 (5km resolution) and 0.71 (50km resolution) obtained by McRae and Beier (2007) using expert opinions. Note that McRae and Beier (2007) use a binary map as habitat/nonhabitat for underlying landscape with 12 populations whereas we use a multivariate surface with continuous and discrete data with 6 populations. We provide the final parameters θ in Table 1. The solutions for Nelder-Mead optimization and our gradient method largely agree: landcover types that allow for movement under cover (e.g., forests) are assigned low resistances values, and open water is assigned the highest resistance. There is a notable difference between learned parameters for barren land, and sedge/herbaceous landscape, which would be interesting to explore further.

Conclusion and Future Work

By formalizing the Inverse Landscape Genetics problem as a graph inference problem involving noisy measurements of effective resistances, we show how to apply powerful optimization methods from machine learning to this scientifically important problem. These methods already provide a promising alternative to existing heuristics, and will allow researchers to more efficiently and effectively solve real-world problems, or to explore synthetic problems at scale. This could facilitate, for example, more widespread investigations of the statistical complexity of inverse landscape genetics.

A major open research direction is to develop further theory around the problem formalized in this paper. For example, as discussed in (Hoskins et al. 2018), while non-convex gradient descent methods seem to perform well, it remains unclear if Problem 1 can be provably solved in polynomial time.

In terms of statistical complexity, our problem is related to that of inferring graphical models (Attias 2000; Mohan et al. 2012), which has been studied in different formulations across machine learning, statistics, and graph signal processing (Egilmez, Pavez, and Ortega 2017; Ortega et al. 2018). The common assumption is that the correlation matrix between data at graph nodes is related to the adjacency or Laplacian matrix of an unknown graph. Several works explore how many samples are needed to learn the structure of this graph, often under additional assumptions like graph sparsity (Raskutti et al. 2009; Cai, Liu, and Luo 2011).

Our work makes a structural assumption that the graph underlying our data has both a simple edge structure (i.e., its a grid graph) and that edges weights are functions of relatively low-dimensional edge data (i.e., landscape information). An interesting direction for future work is understanding if these natural assumptions can be used to formally bound the sample complexity of the inverse landscape genetics problem.

Doing so will likely require a better understanding of *how* samples should be collected for optimal inference. By choosing to collect organism samples in specific geographical locations, we often have control over exactly which graph nodes data is collected for. Empirically, sample design can have substantial impact on how much data is needed to solve the inverse landscape genetics problem (Oyler-McCance, Fedy, and Landguth 2013). Again, we hope that our work provides a starting point for further exploration of this important question. Progress would allow researchers to more efficiently study the dispersion of at-risk species, for which it is difficult to collect substantial genetic data.

Acknowledgments

The authors would like to thank Cameron Musco and Charalampos E. Tsourakakis for early discussions about this work, as well as Uthsav Chitra who provided assistance in formalizing the inverse landscape genetics problem studied. Funding in direct support of this work came solely from NYU's Tandon School of Engineering. There are no other relevant financial activities to disclose.

Ethics Statement

As discussed in the introduction, we believe our work has high potential for positive broader impacts related to environmental protection and conversation. We also hope this paper highlights an interesting applications of graph-inference that we believe is not well known to the machine learning community. Storfer et al. (2007) emphasize the need for forming bridges between research areas with different technical expertise to move landscape genetics forward. Already there has been successful cross-field collaboration between ecologists and those working in spatial statistics. We hope to bring the machine learning community into the fold.

With those benefits in mind, our work does have some potential for negative impact. In particular, the graph inference problem studied has potential applications to de-anonymizing edges in social networks (Hoskins et al. 2018; Liben-Nowell and Kleinberg 2007). Our contributions would probably have limited impact on this sort of application (since our methods are developed specifically for parameterized, planar graphs used in modeling landscapes). Nevertheless, continued work in the area could have negative privacy implications.

References

Arnaud, J.-F. 2003. Metapopulation genetic structure and migration pathways in the land snail Helix aspersa: influence of landscape heterogeneity. *Landscape Ecology* 18(3): 333–346.

Attias, H. 2000. A variational baysian framework for graphical models. In *Advances in neural information processing systems*, 209–215.

Cai, T.; Liu, W.; and Luo, X. 2011. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* 106(494): 594–607.

Chandra, A. K.; Raghavan, P.; Ruzzo, W. L.; Smolensky, R.; and Tiwari, P. 1996. The electrical resistance of a graph captures its commute and cover times. *Computational Complexity* 6(4): 312–340.

Coulon, A.; Cosson, J.; Angibault, J.; Cargnelutti, B.; Galan, M.; Morellet, N.; Petit, E.; Aulagnier, S.; and Hewison, A. 2004. Landscape connectivity influences gene flow in a roe deer population inhabiting a fragmented landscape: an individual–based approach. *Molecular ecology* 13(9): 2841–2850.

Dharangutte, P.; and Musco, C. 2020. Graph learning for inverse landscape genetics. *arXiv preprint arXiv:2006.12334* .

Egilmez, H. E.; Pavez, E.; and Ortega, A. 2017. Graph learning from data under Laplacian and structural constraints. *IEEE Journal of Selected Topics in Signal Processing* 11(6): 825–841.

Endler, J. A. 1977. *Geographic variation, speciation, and clines*. Princeton University Press.

Graves, T. A.; Beier, P.; and Royle, J. A. 2013. Current approaches using genetic distances produce poor estimates of landscape resistance to interindividual dispersal. *Molecular ecology* 22(15): 3888–3903.

Homer, C.; Dewitz, J.; Jin, S.; Xian, G.; Costello, C.; Danielson, P.; Gass, L.; Funk, M.; Wickham, J.; Stehman, S.; et al. 2020. Conterminous United States land cover change patterns 2001–2016 from the 2016 National Land Cover Database. *ISPRS Journal of Photogrammetry and Remote Sensing* 162: 184–199.

Hoskins, J.; Musco, C.; Musco, C.; and Tsourakakis, B. 2018. Inferring Networks From Random Walk-Based Node Similarities. In *Advances in Neural Information Processing Systems 31*, 3704–3715.

Huggett, R. J. 2004. *Fundamentals of biogeography*. Routledge.

Jeh, G.; and Widom, J. 2002. SimRank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 538–543.

Jeh, G.; and Widom, J. 2003. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, 271–279.

Kalofolias, V. 2016. How to learn a graph from smooth signals. In *Artificial Intelligence and Statistics*, 920–929.

Kyle, C.; and Strobeck, C. 2001. Genetic structure of North American wolverine (Gulo gulo) populations. *Molecular Ecology* 10(2): 337–347.

Liben-Nowell, D.; and Kleinberg, J. 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology* 58(7): 1019–1031.

Lugon-Moulin, N.; and Hausser, J. 2002. Phylogeographical structure, postglacial recolonization and barriers to gene flow in the distinctive Valais chromosome race of the common shrew (Sorex araneus). *Molecular ecology* 11(4): 785–794.

Manel, S.; Schwartz, M. K.; Luikart, G.; and Taberlet, P. 2003. Landscape genetics: combining landscape ecology and population genetics. *Trends in ecology & evolution* 18(4): 189–197.

McRae, B.; Shah, V.; and Edelman, A. 2016. Circuitscape: modeling landscape connectivity to promote conservation and human health. *The Nature Conservancy* 14.

McRae, B. H. 2006. Isolation by resistance. *Evolution* 60(8): 1551–1561.

McRae, B. H.; and Beier, P. 2007. Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences* 104(50): 19885–19890.

Mohan, K.; Chung, M.; Han, S.; Witten, D.; Lee, S.-I.; and Fazel, M. 2012. Structured learning of Gaussian graphical models. In *Advances in neural information processing systems*, 620–628.

Novembre, J.; Johnson, T.; Bryc, K.; Kutalik, Z.; Boyko, A. R.; Auton, A.; Indap, A.; King, K. S.; Bergmann, S.; Nelson, M. R.; et al. 2008. Genes mirror geography within Europe. *Nature* 456(7218): 98–101.

Ortega, A.; Frossard, P.; Kovačević, J.; Moura, J. M.; and Vandergheynst, P. 2018. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE* 106(5): 808–828.

Oyler-McCance, S. J.; Fedy, B. C.; and Landguth, E. L. 2013. Sample design effects in landscape genetics. *Conservation Genetics* 14(2): 275–285.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710.

Peterman, W. E. 2018. ResistanceGA: An R package for the optimization of resistance surfaces using genetic algorithms. *Methods in Ecology and Evolution* 9(6): 1638–1647.

Peterman, W. E.; Winiarski, K. J.; Moore, C. E.; da Silva Carvalho, C.; Gilbert, A. L.; and Spear, S. F. 2019. A comparison of popular approaches to optimize landscape resistance surfaces. *Landscape Ecology* 34(9): 2197–2208.

Raskutti, G.; Yu, B.; Wainwright, M. J.; and Ravikumar, P. K. 2009. Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of ℓ_1-regularized MLE. In *Advances in Neural Information Processing Systems*, 1329–1336.

Sanderson, J. 2020. *Landscape ecology: a top down approach*. CRC Press.

Shirk, A.; Wallin, D.; Cushman, S.; Rice, C.; and Warheit, K. 2010. Inferring landscape effects on gene flow: a new model selection framework. *Molecular ecology* 19(17): 3603–3619.

Short Bull, R.; Cushman, S.; Mace, R.; Chilton, T.; Kendall, K.; Landguth, E.; Schwartz, M.; McKelvey, K.; Allendorf, F. W.; and Luikart, G. 2011. Why replication is important in landscape genetics: American black bear in the Rocky Mountains. *Molecular ecology* 20(6): 1092–1107.

Sokal, R. R.; and Oden, N. L. 1978. Spatial autocorrelation in biology: 2. Some biological implications and four applications of evolutionary and ecological interest. *Biological Journal of the Linnean Society* 10(2): 229–249.

Spielman, D. A.; and Srivastava, N. 2011. Graph sparsification by effective resistances. *SIAM Journal on Computing* 40(6): 1913–1926.

Storfer, A.; Murphy, M.; Evans, J.; Goldberg, C.; Robinson, S.; Spear, S.; Dezzani, R.; Delmelle, E.; Vierling, L.; and Waits, L. 2007. Putting the 'landscape'in landscape genetics. *Heredity* 98(3): 128–142.

Tieleman, T.; and Hinton, G. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4(2): 26–31.

Vignieri, S. N. 2005. Streams over mountains: influence of riparian connectivity on gene flow in the Pacific jumping mouse (Zapus trinotatus). *Molecular Ecology* 14(7): 1925–1937.

Vos, C. C.; Antonisse-De Jong, A.; Goedhart, P.; and Smulders, M. 2001. Genetic similarity as a measure for connectivity between fragmented populations of the moor frog (Rana arvalis). *Heredity* 86(5): 598–608.

Wright, S. 1943. Isolation by distance. *Genetics* 28(2): 114.

Yen, L.; Fouss, F.; Decaestecker, C.; Francq, P.; and Saerens, M. 2007. Graph nodes clustering based on the commute-time kernel. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1037–1045. Springer.

Zeller, K. A.; McGarigal, K.; and Whiteley, A. R. 2012. Estimating landscape resistance to movement: a review. *Landscape ecology* 27(6): 777–797.