Graph Learning for Inverse Landscape Genetics

Prathamesh Dharangutte
Tandon School of Engineering
New York University
prathamesh.d@nyu.edu

Christopher Musco
Tandon School of Engineering
New York University
cmusco@nyu.edu

Abstract

Inferring unknown edges from data at a graph's nodes is a common problem across statistics and machine learning. We study a version that arises in the field of *landscape genetics*, where genetic similarity between organisms living in a heterogeneous landscape is explained by a graph that encodes the ease of dispersal through that landscape. Our main contribution is an efficient algorithm for *inverse landscape genetics*, the task of inferring edges in this graph based on the similarity of genetic data from populations at different nodes. This problem is important in discovering impediments to dispersal that threaten biodiversity and species survival. Drawing on influential work that models dispersal using graph *effective resistances* [21], we reduce the inverse landscape genetics problem to that of inferring graph edges from noisy measurements of these resistances. Then, building on edge-inference techniques for social networks [12], we develop an efficient first-order optimization method for solving the problem, which significantly outperforms existing techniques in experiments on synthetic and real genetic data.

1 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 $x_1, \ldots, x_n \in \mathbb{R}^d$ at each node. E.g., in a social network, nodes are users, edges are connections between users, and x_i might contain information like a user's age, or political leaning. Often, node data is correlated with G's *connectivity structure*, which can be quantified using simple statistics like shortest path distance, or advanced metrics like personalized PageRank or DeepWalk distance [27, 14, 28]. We often observe that, if v_i and v_j are strongly connected under such measures, x_i and x_j tend to be more similar [15, 25]. This fact leads to an interesting possibility that has been studied widely: even when edges in G are *unknown*, we can often infer a graph *whose connectivity structure is consistent with the observed node data* [31, 4, 7, 17, 12].

We study an application of this sort of graph inference problem to *landscape genetics*, a field at the intersection of landscape ecology and population genetics [19, 33]. 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 locations (through the landscape) affects genetics. Geographically isolated populations tend to have highly differentiated gene pools, whereas ease of travel and intermixing leads to genetic similarity. In contrast to simpler classical approaches [42, 36], modern methods in landscape genetics often measures isolation by modeling the landscape as an undirected graph (the *landscape graph*). Each location (spatial cell) in the landscape is associated with a graph node, which is connected by an edge to all geographically adjacent nodes (see Fig. 1). Edge weights are chosen to reflect the ease of organism dispersal between adjacent nodes, with high weight indicating ease of dispersal and low weight indicates inhibition. 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 traverse the edge. For an organism that prefers low-land environments, edges crossing areas of high elevation would receive lower weight than low-land areas.

AI for Earth Sciences Workshop at NeurIPS 2020.

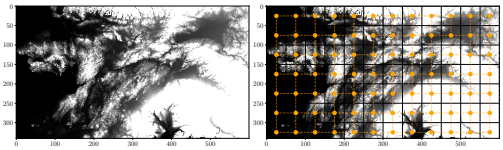


Figure 1: To model a landscape as a graph it is divided into cells (black grid) and each cell is associated with a node (orange markers). Adjacent nodes are connected by weighted edges (orange lines). In landscape genetics, each weights are a function of underlying landscape data, e.g. the average elevation of the edge.

It has been established that measures of pairwise connectivity in a well-constructed landscape graph correlate with pairwise genetic similarities between populations living at different nodes. For example, for many species, the weight of the least cost path is known to correlate with genetic similarity measured using e.g., the fixation index [1, 6, 40]. More recently, McRae's influential paper *Isolation by Resistance* popularized the use of *effective resistance distance* as a connectivity measure in landscape genetics [21, 43]. Effective resistances better model organism dispersal, and thus correlate more closely with genetic differences across landscapes [22].

Most studies that use landscape graphs to model species dispersal construct these graphs based on *expert knowledge* of a species' behavioral preferences (e.g, preferred elevation, vegetation cover, or climate) [21, 34]. Multiple expert proposed graphs can be tested for fit [41, 18, 40, 35], but achieving high levels of correlation with genetic data requires significant background information on a species (which may be imperfect) and laborious hand-tuning. To address this issue, there has been interest in moving beyond expert opinion, by *algorithmically* determining optimal edge weights [44, 30]. The aim is to learn a function that maps measurable landscape parameters (e.g. vegetation cover, or if there is human development along an edge) to edge weights. The resulting weighted graph should have connectivity structure that correlates well with genetic differences across the landscape. We call this parameterized graph inference problem *inverse landscape genetics*. Beyond refining expert-designed landscape graphs, solutions to this problem would allow ecologist to infer information about species dispersal based *purely on collected genetic data*, as opposed to simply explaining genetic data with known ecological knowledge. Genetic information can then be used to understand species habitat preferences, find bottlenecks in migration, or understand the impact of human development [20]. As discussed in [44] and [10], algorithms for learning landscape graphs from data could therefore be essential in future conservation and planning decisions, e.g. for wildlife corridor design.

Our Contributions. Despite interest in inverse landscape genetics, few effective algorithms exist for the problem. Most approaches use simple variants of brute force search, relying on expert opinion to obtain an initial graph and then searching over nearby weight functions to improve fit [34]. There has been some work on more systematic approaches: [29] uses a genetic algorithm and [10] uses local search heuristics like the Nelder-Mead and Newton line search. Our main contribution is to show that the inverse landscape genetics problem can be solved more effectively using *gradient based* optimization. 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 (a measure of genetic differentiation). Building on recent work on learning social network edges [12], we show how to compute a gradient for an appropriately chosen graph-learning loss, which requires differentiating through an effective resistances computation. We do so efficiently using iterative solvers for positive semidefinite linear systems. To the best of our knowledge, our method is the first for inverse landscape genetics that uses gradient based optimization and, while our underlying objective is non-convex, it obtains faster and more reliable convergence on both synthetic and real-world data than prior work. 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 [26].

2 Proposed Method

Notation. Denote the undirected landscape graph by $G = (V, E, w)$, where $V = \{v_1, \ldots, v_n\}$ is the vertex set, E is the edge set with size $m = |E|$, and w is a vector of non-negative edge weights. We index E and w by their terminal nodes: edges are $e_{i_1 j_1}, \ldots, e_{i_m j_m}$ and weights are $w_{i_1 j_1}, \ldots, w_{i_m j_m}$. We can view the landscape graph as an electrical network where e_{ij} represents a connection with *conductance* w_{ij} [38, 21]. Let $r_{ij} = 1/w_{ij}$ denote the resistance of the connection. For a subset $S \subseteq V$ of nodes we have 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 of the measure is not important. In keeping with prior work, our experiments use the fixation index, denoted by F_{ST}. A high F_{ST} (close to 1) indicates greater difference between the genetic data 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, with $F_{i,i} = 0$ for all i.

A central result in landscape genetics is that values in F will correlate well with the *effective resistances* of an appropriately chosen landscape graph G [21]. Letting L be the weighted, unnormalized graph Laplacian of G, the effective resistance between nodes i and j is equal to $R_{ij} = b_{ij}^T L^+ b_{ij}^T$, where L^+ is the pseudoinverse of L and $b_{ij} \in \mathbb{R}^n$ is a vector with value 1 at position i, -1 at position j, and 0 elsewhere. Effective resistance is lower when there are more low-resistance (i.e., high weight) paths between v_i and v_j. It is known to be equal to the edge-weighted *commute time* between v_i and v_j [5], which gives intuition for why it effectively quantifies organism dispersal through a landscape [20]. Let R be the matrix of all pairwise effective resistances and let $R_S \in |S| \times |S|$ be its principal submatrix containing only resistances between nodes in S. The 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 a vector θ with n_θ parameters 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 θ^* minimizing the loss:*

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

where $R_S(p_\theta(E))$ is the effective resistance matrix for the landscape graph $G = (V, E, p_\theta(E))$.

$R_S(p_\theta(E))$ and F have zero diagonals, so the Frobenius norm in (1) is just twice the standard squared error between effective resistances and genetic dissimilarities. Other loss functions could also be used, like the inverse Mantel correlation [10]. In any case, the goal is to find a weight function so that G's effective resistances are as close as possible to the measured genetic dissimilarities in F. Alternatively, if we can think of these dissimilarities as noisy measures of the *true effective resistances for some unknown landscape graph G^**, then Problem 1 can be viewed as the task of recovering G^*.

Example functional forms. In Problem 1, weights in the learned graph are a function of environmental parameters about each edge. This function can take any form: we only require that it is differentiable. A typical choice is to have $1/w_{i_k j_k} = r_{i_k j_k}$ follow an inverted Gaussian relation governed by parameters $\theta = [\beta, \beta_{opt}$ and $\beta_{SD}]$, which indicates a preferred parameter value/range for a species (e.g. preferred elevation) [10]. Another common functional form is linear, which is often used with for discrete (e.g., binary) data which indicates the presents of conditions that could impact dispersal, like forest cover or human development. It is also natural to add multiple functional forms for a combination of continuous and discrete data. E.g. we might have $r_{i_k j_k} = r^{\text{E}}_{i_k j_k} + r^{\text{LC}}_{i_k j_k}$ where $r^{\text{E}}_{i_k j_k}$ is an elevation term in the form of (2a) and $r^{\text{LC}}_{i_k j_k}$ is a landcover term in the form of (2b).

$$\frac{1}{w_{i_k j_k}} = \beta + 1 - \beta e^{\frac{-(C_{i_k j_k} - \beta_{\text{opt}})^2}{2\beta_{\text{SD}}^2}} \tag{2a}$$

and

$$\frac{1}{w_{i_k j_k}} = \alpha^T C_{i_k j_k}. \tag{2b}$$

Efficient gradient computation. Since there is no closed form solution for (1), the cornerstone of our approach is to find an approximate solution by using projected gradient descent. To do so, we need an efficient method for computing the gradient of $\mathcal{L}(\theta)$. The following is proven in Appendix A.

Proposition 1. *Let $J \in \mathbb{R}^{m \times n_\theta}$ be the Jacobian with $J_{k,h} = \frac{\partial w_{i_k j_k}}{\partial \theta_h}$. Let B denote G's edge-vertex incidence matrix, with k^{th} row equal to $b_{i_k j_k}$, where i_k and j_k are the terminal nodes of edge $e_{i_k j_k}$.*

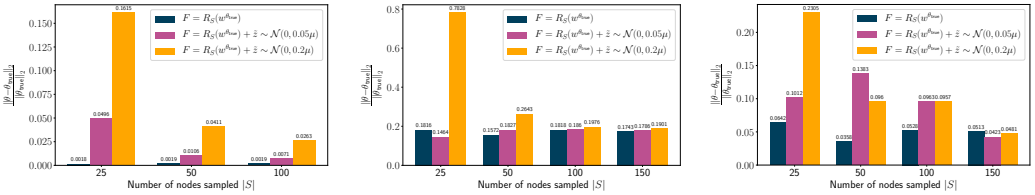
$$\nabla_\theta(\mathcal{L}) = \sum_{v_l, v_k \in S} \left(F_{lk} - b_{lk}^T L_\theta^+ b_{lk}\right) \cdot 2J \cdot (b_{lk}^T L_\theta^+ B)^{\circ 2},$$

where $^{\circ 2}$ is the Hadamard power and L_θ is the Laplacian of the graph with weights $p_\theta(E)$.

Since the number of parameters n_θ is typically small, computing J is efficient for any differentiable p_θ. So, if we have computed $b_{lk}^T L_\theta^+ B$ for all $v_l, v_k \in S$, the gradient can be computed in $O(|S|^2 \cdot m \cdot n_\theta)$ additional time. And, since every row in B is 2 sparse, $b_{lk}^T L^+ B$ can be computed in $O(m)$ time once $b_{lk}^T L^+ = L^+ b_{lk}$ is computed. Since $m = O(n)$ in most landscape genetics applications (e.g. $m \approx 4n$ for a grid), the bottle neck is therefore computing each $L_\theta^+ b_{lk}$. This naively requires inverting the $n \times n$ Laplacian L_θ, which is computationally impractical for large graphs. We instead approximate the $L_\theta^+ b_{lk}$ using an iterative linear system solver, specifically MINRES in our experiments. To optimize the approach further, we note that $L_\theta^+ b_{lk} = L(w)^+ e_l - L_\theta^+ e_k$ where e_l and e_k are standard basis vectors. Accordingly, we only need to solve $|S|$ linear systems and can then recombining those solutions to return all $\binom{|S|}{2}$ vectors $L(w)^+ b_{lk}$ needed for the gradient computation.

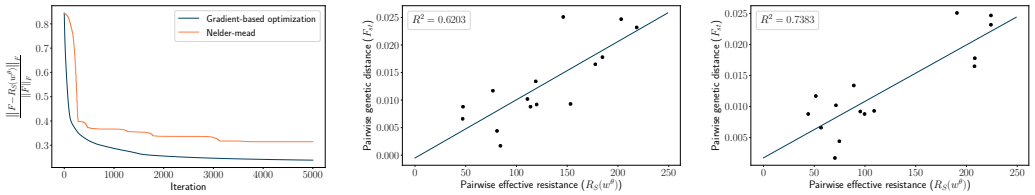
3 Empirical Results

We test our optimization method and compare it against local search heuristics used in prior work [10] on synthetic genetic data and real data for the North American wolverine (Gulo gulo) [16], for which we have F_{ST} values for 6 populations living across Alaska. We use elevation [24] and land cover data [11] to parameterize the edge weight function in the landscape graph. Details of the experimental setup are discussed in Appendix B, and additional results are provided in Appendix C.



(a) (b) (c)

Figure 2: Relative error between recovered and true parameters for synthetic data with increasing $|S|$ and noise $\tilde{\sigma}$ for discrete (2a), continuous (2b) and combined edge functions. Parameter recovery improves with more samples and descreased noise.



(a) (b) (c)

Figure 3: Relative objective function value (Fig. 3a) and R^2 values computed for a linear fit between F_{ST} and effective resistances in the learnt landscape graphs for real-world data. Gradient-based optimization (Fig. 3c) obtains a better fit compared to Nelder-Mead (Fig. 3b) and converges faster.

Synthetic data. We select a 50x50 subgrid of an Alaska landscape ($|V| = 2500$, $|E| = 4900$). We then construct a *ground truth* graph by randomly sampling parameters θ_{true} and evaluating the weight function for all edges. We construct the pairwise effective resistance matrix $R_S(w^{\theta_{\text{true}}})$ for a small subset of nodes S and produce a simulated genetic similarity matrix F by setting $F_{lk} = \left[R_S(w^{\theta_{\text{true}}})\right]_{lk} + \tilde{z}$ where $\tilde{z} \sim \mathcal{N}(0, \tilde{\sigma})$. The goal in our synthetic experiments is to recover the ground truth. We test with $\tilde{\sigma} = \{0, 0.05\mu, 0.2\mu\}$, where μ is the mean resistance in $R_S(w^{\theta_{\text{true}}})$, corresponding to (no, low and high noise) case. For parameters θ obtained after optimization, we report the relative error $\|\theta - \theta_{\text{true}}\|_2 \,/\, \|\theta_{\text{true}}\|_2$. We ignore parameters for landcover types present at less than 1% of nodes as they cannot be determined with any level of accuracy (since they have essentially no impact on graph effective resistances). As seen in Figure 2, as $|S|$ increases, we obtain high quality approximations of θ_{true}, even in the high noise regime. Overfitting for small $|S|$ has been raised as concern in the literature, i.e. the landscape graph fit to F may not generalize to new data [26]. We further validate our method against overfitting in Appendix C, showing that even when $|S|$ relatively is small (e.g., ≥ 25), test loss converges along with train loss, implying good generalization.

North American wolverine (Gulo gulo). For real-world data, the 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 (Figure 3), a metric used in prior work [22]. 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 [22] using expert opinions. Note that [22] 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.

Conclusion. By formalizing the Inverse Landscape Genetics problem as a graph inference problem involving effective resistances, we show how to apply powerful optimization methods to this scientifically important problem. Doing so already provides 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. A major open research direction is to develop further theory around the problem formalized in this paper. We discuss several specific questions in Appendix E.

References

[1] Jean-François Arnaud. Metapopulation genetic structure and migration pathways in the land snail helix aspersa: influence of landscape heterogeneity. *Landscape Ecology*, 18(3):333–346, 2003.

[2] Hagai Attias. A variational baysian framework for graphical models. In *Advances in neural information processing systems*, pages 209–215, 2000.

[3] Vladimir Batagelj, Tomaž Pisanski, and JMS Sim [otilde] es Pereira. An algorithm for tree-realizability of distance matrices. *International Journal of Computer Mathematics*, 34(3-4):171–176, 1990.

[4] Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[5] Ashok K Chandra, Prabhakar Raghavan, Walter L Ruzzo, Roman Smolensky, and Prasoon Tiwari. The electrical resistance of a graph captures its commute and cover times. *Computational Complexity*, 6(4):312–340, 1996.

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

[7] Hilmi E Egilmez, Eduardo Pavez, and Antonio Ortega. Graph learning from data under laplacian and structural constraints. *IEEE Journal of Selected Topics in Signal Processing*, 11(6):825–841, 2017.

[8] Joseph Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *evolution*, 39(4):783–791, 1985.

[9] Joseph Felsenstein and Joseph Felenstein. *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA, 2004.

[10] Tabitha A Graves, Paul Beier, and J Andrew Royle. Current approaches using genetic distances produce poor estimates of landscape resistance to interindividual dispersal. *Molecular ecology*, 22(15):3888–3903, 2013.

[11] Collin Homer, Jon Dewitz, Suming Jin, George Xian, Catherine Costello, Patrick Danielson, Leila Gass, Michelle Funk, James Wickham, Stephen Stehman, et al. 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, 2020.

[12] Jeremy Hoskins, Cameron Musco, Christopher Musco, and Babis Tsourakakis. Inferring networks from random walk-based node similarities. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3704–3715. Curran Associates, Inc., 2018.

[13] Arun Jambulapati, Kirankumar Shiragur, and Aaron Sidford. Efficient structured matrix recovery and nearly-linear time algorithms for solving inverse symmetric m-matrices. *arXiv preprint arXiv:1812.06295*, 2018.

[14] Glen Jeh and Jennifer Widom. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, pages 271–279, 2003.

[15] Vassilis Kalofolias. How to learn a graph from smooth signals. In *Artificial Intelligence and Statistics*, pages 920–929, 2016.

[16] CJ Kyle and C Strobeck. Genetic structure of north american wolverine (gulo gulo) populations. *Molecular Ecology*, 10(2):337–347, 2001.

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

[18] N Lugon-Moulin and J Hausser. 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, 2002.

[19] Stéphanie Manel, Michael K Schwartz, Gordon Luikart, and Pierre Taberlet. Landscape genetics: combining landscape ecology and population genetics. *Trends in ecology & evolution*, 18(4):189–197, 2003.

[20] BH McRae, Viral Shah, and Alan Edelman. Circuitscape: modeling landscape connectivity to promote conservation and human health. *The Nature Conservancy*, 14, 2016.

[21] Brad H McRae. Isolation by resistance. *Evolution*, 60(8):1551–1561, 2006.

[22] Brad H McRae and Paul Beier. Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences*, 104(50):19885–19890, 2007.

[23] Karthik Mohan, Mike Chung, Seungyeop Han, Daniela Witten, Su-In Lee, and Maryam Fazel. Structured learning of gaussian graphical models. In *Advances in neural information processing systems*, pages 620–628, 2012.

[24] National Atlas of the United States. 100-meter resolution elevation of alaska, albers projection. `http://purl.stanford.edu/sg962yb7367`, 2012.

[25] Antonio Ortega, Pascal Frossard, Jelena Kovačević, José MF Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018.

[26] Sara J Oyler-McCance, Bradley C Fedy, and Erin L Landguth. Sample design effects in landscape genetics. *Conservation Genetics*, 14(2):275–285, 2013.

[27] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[28] 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, 2014.

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

[30] William E Peterman, Kristopher J Winiarski, Chloe E Moore, Carolina da Silva Carvalho, Anthony L Gilbert, and Stephen F Spear. A comparison of popular approaches to optimize landscape resistance surfaces. *Landscape Ecology*, 34(9):2197–2208, 2019.

[31] Garvesh Raskutti, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. In *Advances in Neural Information Processing Systems*, pages 1329–1336, 2009.

[32] Lev Reyzin and Nikhil Srivastava. Learning and verifying graphs using queries with a focus on edge counting. In *International Conference on Algorithmic Learning Theory*, pages 285–297. Springer, 2007.

[33] James Sanderson. *Landscape ecology: a top down approach*. CRC Press, 2020.

[34] AJ Shirk, DO Wallin, SA Cushman, CG Rice, and KI Warheit. Inferring landscape effects on gene flow: a new model selection framework. *Molecular ecology*, 19(17):3603–3619, 2010.

[35] RA Short Bull, SA Cushman, R Mace, T Chilton, KC Kendall, EL Landguth, MK Schwartz, Kevin McKelvey, Fred W Allendorf, and G Luikart. Why replication is important in landscape genetics: American black bear in the rocky mountains. *Molecular ecology*, 20(6):1092–1107, 2011.

[36] Robert R Sokal and Neal L Oden. 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, 1978.

[37] Daniel A Spielman. Trees and distances. *University Lecture*, 2012.

[38] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

[39] Tijmen Tieleman and Geoffrey Hinton. 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, 2012.

[40] Sacha N Vignieri. Streams over mountains: influence of riparian connectivity on gene flow in the pacific jumping mouse (zapus trinotatus). *Molecular Ecology*, 14(7):1925–1937, 2005.

[41] Claire C Vos, AG Antonisse-De Jong, PW Goedhart, and MJM Smulders. Genetic similarity as a measure for connectivity between fragmented populations of the moor frog (rana arvalis). *Heredity*, 86(5):598–608, 2001.

[42] Sewall Wright. Isolation by distance. *Genetics*, 28(2):114, 1943.

[43] Luh Yen, Francois Fouss, Christine Decaestecker, Pascal Francq, and Marco Saerens. Graph nodes clustering based on the commute-time kernel. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 1037–1045. Springer, 2007.

[44] Katherine A Zeller, Kevin McGarigal, and Andrew R Whiteley. Estimating landscape resistance to movement: a review. *Landscape ecology*, 27(6):777–797, 2012.

A Details of Proposed Method

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

$$\nabla_{w^\theta}\mathcal{L} = -2 \sum_{v_l, v_k \in S} \left(F_{lk} - R(w^\theta)_{lk}\right) \cdot \nabla_{w^\theta} R(w^\theta)_{lk} \tag{3}$$

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 [12], 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 $\nabla_{w^\theta} R(w^\theta)_{lk} = -(b_{lk}^T L_\theta^+ B)^{\circ 2}$. The proposition follows from plugging this equation into (3) and noting that $\nabla_\theta(\mathcal{L}) = J \cdot \nabla_{w^\theta}\mathcal{L}$. □

B Details of Experiments

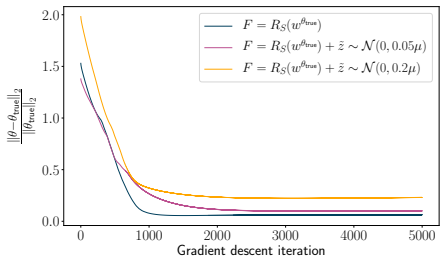
The landscape graph graph is constructed by dividing the Alaska region into a grid of square cells. We choose a resolution of 15 km, which lead to a graph $G = (V, E)$ with $|V| = 24035$ and $|E| = 47746$. Our landscape data comes as raster images, with each pixel corresponding to a region of 100×100 meters for elevation data and 30×30 meters for landcover data, so we have multiple pixels of information within each landscape cell. We choose resolution of 15km to have a reasonably sized graph, compared to previously used 5km and 50km [22]. We use the re-sampling tool from ArcMap with the 'nearest' technique for land cover data and 'bilinear' technique for elevation data, to change the resolution[1].

Continuous and discrete environmental parameters are then collected for each edge in the graph from the raster. 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 (2a) for elevation data and (2b) for landcover data.

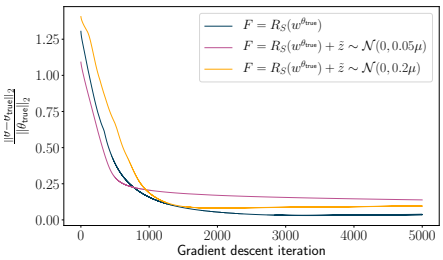
To minimize (1), we implement a projected gradient descent method with RMSProp step size adjustment, which adjusts learning rate by a decaying average of squared gradients [39]. Since edge weights are constrained to be non-negative, and all edge data is non-negative, we project parameters to $\max(\epsilon, \theta)$ 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.

For experiments with real-world genetic data, we set the edge weight to 0 if any node for the edge had landcover type as 'unclassified' whereas for synthetic data we set corresponding α parameter to high value which results in minuscule edge weight. For optimization, we initialize β as 1 and all other parameters (θ_{true}) randomly from a discrete uniform distribution, as a small initial value for β provided us with most consistent results. For setting the true parameters θ_{true} and initialization, we sample β_{opt} and β_{SD} from range 0-10 and α from range 0-100 for synthetic data experiments. For experiments with real-world data, we initialize β_{opt} and β_{SD} similarly but for alpha we use the range 1-10. We run all experiments with RMSProp for 5000 iterations using learning rate of 0.1 and 0.9 for γ parameter. For projecting parameters as $max(\epsilon, \theta)$, we use $\epsilon = 1$ for synthetic data experiments and $\epsilon = 10^{-20}$ for experiments with genetic data for North American wolverine for the parameters $\theta = \{\beta, \alpha\ \in \mathbb{R}^{17}\}$. For β_{SD} we use $\epsilon = 10^{-3}$. Projecting to a small $\epsilon > 0$ instead of exactly 0 helps

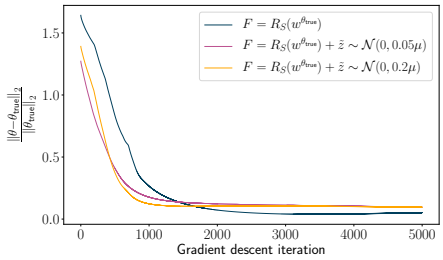
[1]These are the recommended techniques from the software documentation. We refer the reader to `https://desktop.arcgis.com/en/arcmap/10.3/tools/data-management-toolbox/resample.htm`



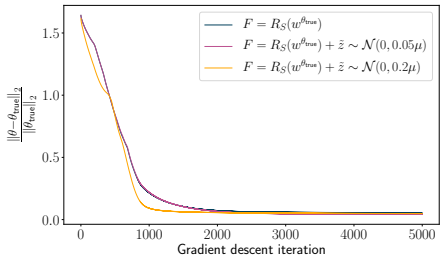
(a) Landscape with continuous and discrete data with $N = 25$.



(b) Landscape with continuous and discrete data with $N = 50$.



(c) Landscape with continuous and discrete data with $N = 100$.



(d) Landscape with continuous and discrete data with $N = 150$.

Figure 4: Relative error between recovered parameters and true parameters with gradient descent iteration, for various values of noise standard deviation $\tilde{\sigma}$, for landscape graph with continuous and discrete data. Higher value of N provides better recovery.

avoid numerical issues when computing the gradient and prevented the algorithm from getting stuck in local minima. We scale the F_{st} values by a factor of 10^5 to ensure stable gradient computations.

Note that for reporting R^2 values for linear fit between pairwise effective resistances of learnt landscape graph and F_{ST} values, [22] use $\frac{F_{\text{ST}}}{1-F_{\text{ST}}}$ instead of F_{ST} values whereas we use F_{ST}, but as $F_{\text{ST}} << 1$ for the data under consideration, the R^2 values are approximately equal.

C Additional Experiments

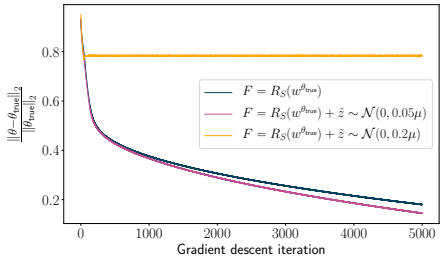
C.1 Synthetic data experiments

We present results for synthetic data experiments. We note that when only elevation data is considered, we consistently obtain good recovery for β_{opt} and β_{SD} but not for β. Although the recovery is not consistent across all parameters, β_{opt} and β_{SD} are typically more meaningful to researchers, indicating the preferred elevation range and range of elevation of the species. We conclude that as the number of nodes sampled(N) increases, we obtain good approximations to the true parameters θ_{true}.

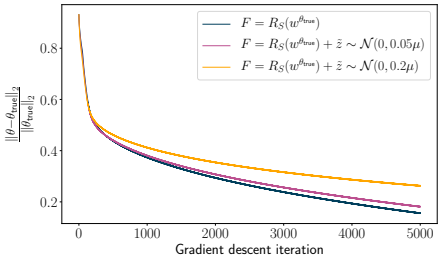
For parameter recovery with different settings of N, we observe that $N = 150$ is sufficient for reliable recovery of parameters. Appendix Figures 4, 5 and 6 show the relative parameter approximation error with gradient descent iteration. For landscape graph with only discrete data with $N = 25$ and $\tilde{\sigma} = 0.2\mu$, we observe overfitting where the relative parameter error increases with iteration in Appendix Figure 6a.

C.2 Addressing overfitting

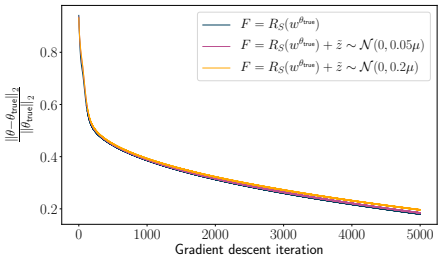
It has been noted in the literature that a potential concern with optimizing landscape graphs is overfitting when N is small. i.e., the landscape graph fit to F does not generalize to new data [26]. To validate our method against overfitting, we randomly split nodes into sets S_{train} and S_{test}. 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



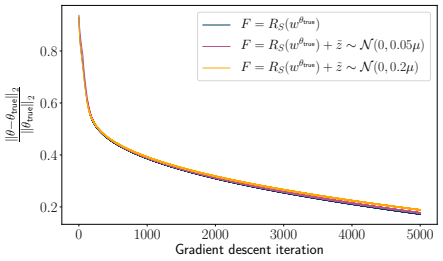
(a) Landscape with continuous data and $N = 25$.



(b) Landscape with continuous data and $N = 50$.

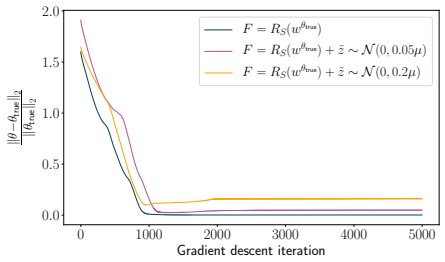


(c) Landscape with continuous data and $N = 100$.

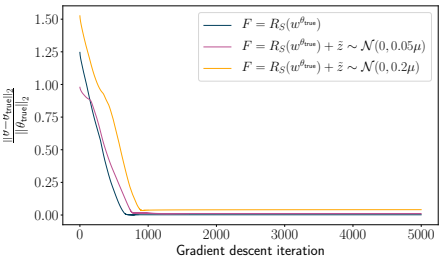


(d) Landscape with continuous data and $N = 150$.

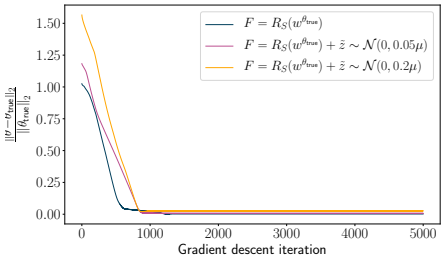
Figure 5: Relative error between recovered parameters and true parameters for continuous data with gradient descent iteration, for various values of noise standard deviation $\tilde{\sigma}$, for landscape graph with continuous data. Higher value of N provides better recovery.



(a) Landscape with discrete data and $N = 25$.

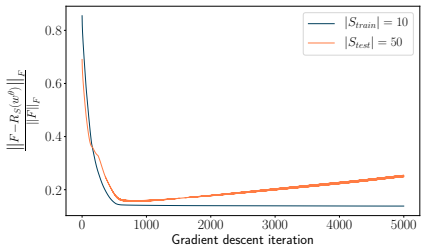


(b) Landscape with discrete data and $N = 50$.

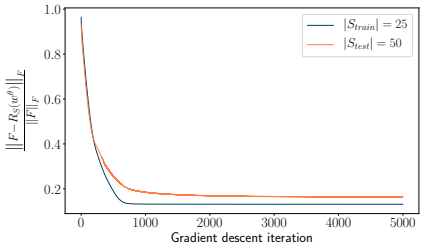


(c) Landscape with discrete data and $N = 100$.

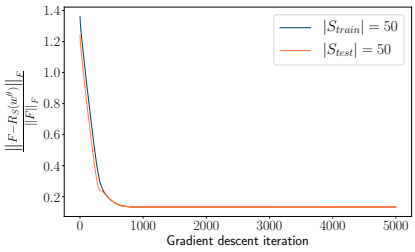
Figure 6: Relative error between recovered parameters and true parameters for discrete data with gradient descent iteration, for various values of noise standard deviation $\tilde{\sigma}$, for landscape graph with discrete data. Higher value of N provides better recovery and we observe overfitting for high-noise with $N = 25$.



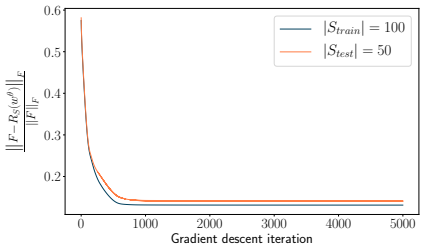
(a) Relative train and test loss for $N_{\text{train}} = 10$.



(b) Relative train and test loss for $N_{\text{train}} = 25$.



(c) Relative train and test loss for $N_{\text{train}} = 50$.



(d) Relative train and test loss for $N_{\text{train}} = 100$.

Figure 7: Relative train and test loss for different values of sampled train nodes with $\tilde{\sigma} = 0.2\mu$. We obtain good generalization, for $N \geq 25$, mitigating the concern for overfitting with low data.

is as low as 25, (appendix figure 7). This implies good generalization and a lack of overfitting, *even though we do not accurately recover all parameters in θ_{true}*. This is not necessarily surprisingly: it indicates that, while the inverse landscape genetics problem may be poorly conditioned with respect to θ (a fact observed in [10]) it is still possible to obtain reliable predictive models with little data.

C.3 Comparision with existing approaches

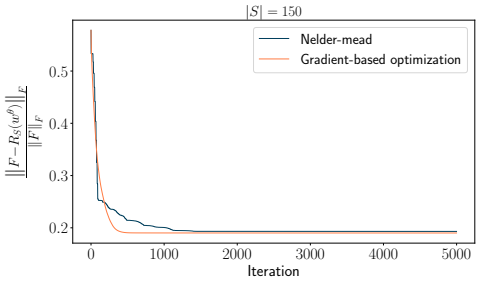
We compare gradient-based optimization to the Nelder-Mead method, which has been used in prior work on inverse landscape genetics [10]. We observe that our method is faster in terms of convergence and also better at recovering true parameters with enough data. Nelder-Mead eventually achieves comparable performance in terms of train loss but fails at recovering the true parameters (Figure 8). To ensure a fair comparison, we choose the same random initialization of parameters and non-negativity constraints[2].

D Additional Related Work

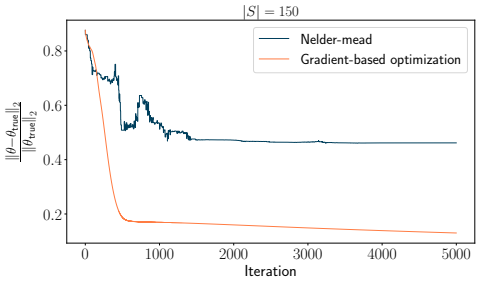
Relevant related work on landscape genetics is discussed in the introduction. Here we add important comments on additionally related work in graph learning. In particular, in Section 2, we frame the inverse landscape genetics problem as a problem of learning edge weights in a graph from (noisy) measurements of the effective resistances. This problem was directly addressed in [12], which our paper builds on. It has also been studied elsewhere. For example, it is well known that you can recover a graph exactly if you know effective resistances between *all pairs of nodes*.

This can be done in polynomial time [37]: access to all effective resistances allows you to reconstruct the pseudoinverse of the graph Laplacian, which can then be inverted using a generic $O(n^3)$ time method, or more efficient algorithms [13]. Unfortunately, when only a subset of effective resistances are known, no polynomial time algorithm is known for recovering a graph consistent with those measurements. However, as observed in [12] and this paper (where we study a somewhat different parameterized problem) graph recovery can be framed as an optimization problem and solved to a global optimal with first order methods, despite inherent non-convexity. Recovering edge information from effective resistances has also been studied for the special case of tree graphs. In a tree, the effective resistance is the inverse of the shortest path distance between nodes i and j. There has been

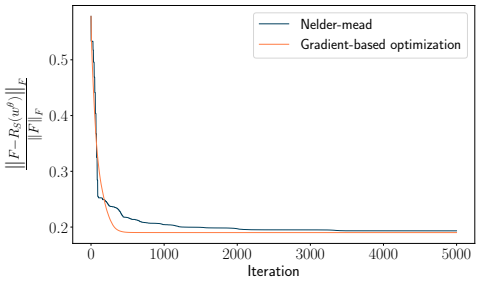
[2]Note that Nelder-Mead is an unconstrained optimization method, so we add a projection step to ensure interpretable parameters are found. This does not noticably effect the behavior of convergence in our experiments.



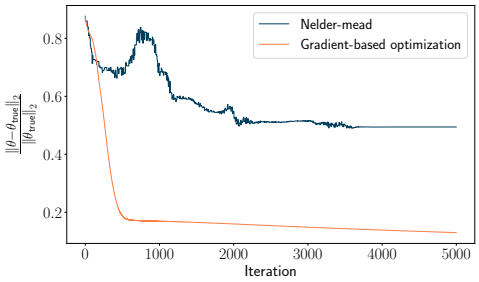
(a) Comparison of relative loss with iteration for Nelder-mead with projection and gradient-based approach.



(b) Relative parameter error between recovered parameters and true parameters with iteration(with projection for Nelder-Mead).



(c) Comparison of relative loss with iteration for Nelder-mead without projection and gradient-based approach.



(d) Relative parameter error between recovered parameters and true parameters with iteration(without projection for Nelder-Mead).

Figure 8: 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 $\tilde{\sigma} = 0.2\mu$.

a lot of interest in reconstructing trees from actual and partial measurements of these distances [32, 3]. Applications include the reconstruction of phylogenetic trees in genetics [8, 9].

Finally, we note that our problem is related to that of inferring graphical models [2, 23], which has been studied in different formulations across machine learning, statistics, and graph signal processing [7, 25]. The common assumption is that the correlation matrix between data at each node is related to the adjacency or Laplacian matrix of an unknown graph. Several work explore how many samples are needed to learn the structure of this graph, often under additional assumptions like graph sparsity [31, 4]. 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 theoretically exploring the implications of these strong assumptions on bounding the sample complexity of the inverse landscape genetics problem.

E Future Directions

As mentioned, an important research direction is to develop further theory around the problem formalized in this paper. For example, as discussed in [12], 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 [2, 23], which has been studied in different formulations across machine learning, statistics, and graph signal processing [7, 25]. 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 [31, 4].

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 [26]. 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.