Learning Population-Level Diffusions with Generative Recurrent Networks

Tatsunori B. Hashimoto THASHIM@MIT.EDU
David K. Gifford DKG@MIT.EDU
Tommi S. Jaakkola TOMMI@CSAIL.MIT.EDU

Abstract

We estimate stochastic processes that govern the dynamics of evolving populations such as cell differentiation. The problem is challenging since longitudinal trajectory measurements of individuals in a population are rarely available due to experimental cost and/or privacy. We show that cross-sectional samples from an evolving population suffice for recovery within a class of processes even if samples are available only at a few distinct time points. We provide a stratified analysis of recoverability conditions, and establish that reversibility is sufficient for recoverability. For estimation, we derive a natural loss and regularization, and parameterize the processes as diffusive recurrent neural networks. We demonstrate the approach in the context of uncovering complex cellular dynamics known as the 'epigenetic landscape' from existing biological assays.

1. Motivation

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

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

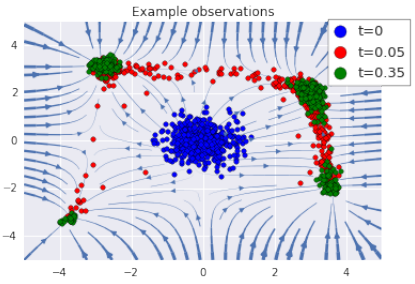


Figure 1. In population-level inference we observe samples (colored points) drawn from the process at different times. The goal is to infer the dynamics (blue vectors). In this toy dataset each point can be thought of as a single cell and the x and y axes as gene expression levels of two genes.

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

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

Analyzing the inference problem for population-level diffusions, we utilize the connection between partial differential equations, diffusion processes, and recurrent neural

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

networks (RNN) to derive a principled loss function and estimation procedure.

Our contributions are the following

- First, we rigorously study whether the dynamics of a diffusion can be recovered from cross-sectional observations, and establish the first identifiability results.
- Second, we show that a particular regularized recurrent neural network (RNN) with Wasserstein loss is a natural model for this problem and use this to construct a fast scalable initializer that exploits the connection between diffusions and RNNs.
- Finally, our method is verified to recover known dynamics from simulated data in the high-dimensional regime better than both parametric and local diffusion models, as well as predict the differentiation time-course on tens of genes for real RNA-seq data.

2. Prior work

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

2.1. Population level inference

Inferring dynamics from population appears in three areas: In home-range estimation, one estimates the support of a two-dimensional continuous time series from the stationary distribution (Fleming et al., 2015). Our work is distinguished by our focus on the high-dimensional $(d > 2)$ and non-stationary settings. The stationary case is discussed in section 4.1.

Inverse problems in parabolic differential equations identify continuous, low-dimensional dynamics given noisy but complete measurements (rather than samples) along a known boundary (Tarantola, 2005). One-dimensional methods using plug-in kernel density estimates exist (Lund et al., 2014) but do not generalize to greater than one dimension.

Finally, estimation of discrete Markov chains using 'macro' data is the discrete time and space equivalent of our problem. This is a classic problem in econometrics, and recovery properties (Van Der Plas, 1983), estimation algorithms (Kalbfleisch & Lawless, 1984), and the effect of noise (Bernstein & Sheldon, 2016) are all well-known. The discrete solutions above observe multiple populations stopped at the same time points, which allows for the more general solutions. Our problem cannot be solved trivially via discretization: discretizing the space scales exponentially with dimension, and discretizing time results in a solution which is conceptually equivalent to the time-derivative model in section 4.2 and does not capture the underlying geometry of the problem.

2.2. Diffusive RNNs

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

2.3. Computational biology

Pseudo-time analysis (Trapnell et al., 2014) models the differentiation of cells as measured by single-cell RNA-seq by assigning each cell to a differentiation path via bifurcations and a 'pseudo-time' indicating its level of differentiation. Such analysis is driven by the desire to identify the cell-states and relevant marker genes during differentiation. Recent sophisticated methods can recover such bifurcations quite effectively (Setty et al., 2016; Marco et al., 2014).

Our work complements such analyses by showing that it is possible to recover quantitative parameters such as the underlying epigenetic landscape from few population measurements. Our results on identifiability of the epigenetic landscape will become more valuable as the number of captured cells in a single-cell RNA-seq experiment grows from hundreds (Klein et al., 2015) to tens of thousands.

Systems biology models of the epigenetic landscape have focused on constructing landscapes which recapitulate the qualitative properties of differentiation systems (Qiu et al., 2012; Bhattacharya et al., 2011). Our work distinguished by a focus on data-driven identification of the epigenetic landscape. Existing data-driven models of the epigenetic landscape are for a single gene and either rely on longitudinal tracking (Sisan et al., 2012) or require assuming that a particular cell population is stationary (Luo et al., 2013).

3. Population-level behavior of diffusions

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

A d-dimensional diffusion process $X(t)$ represents the state (such as gene expression) of an individual at time t.

Formally we define $X(t)$ as a stochastic differential equation (SDE):

$$dX(t) = \mu(X(t))dt + \sqrt{2\sigma^2}dW(t). \quad (1)$$

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

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

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

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

Definition 1 (Population-level inference)**.** *Define the marginal distribution* $\rho(t,x) = P(X(t) = x)$.

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

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

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

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

Theorem 1 (The JKO theorem)**.** *Given a diffusion process defined by equation 1 with* $\mu(x) = -\nabla\Psi(x)$*, then the marginal distribution* $\rho(t,x) = P(X(t) = x)$ *is approximated by the solution to the following recurrence equation for* $\rho^{(t)}$ *with* $\rho^{(0)} = \rho(0,x)$.

$$\rho^{(t+\Delta t)} = \underset{\rho^{(t+\Delta t)}}{\textit{argmin}} \quad W_2(\rho^{(t+\Delta t)}, \rho^{(t)})^2 + \frac{\Delta t}{\sigma^2} D\left(\rho^{(t+\Delta t)} || \exp\left(\frac{-\Psi(x)}{\sigma^2}\right)\right). \quad (3)$$

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

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

4. Recoverability of the potential Ψ

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

We consider recoverability in three regimes of increasing difficulty. First, in section 4.1, we consider the stationary case of observing $\rho(\infty,x)$ which results in a closed-form estimator for Ψ, but requires unrealistic assumptions on our model. Next, in section 4.2 we consider a large number of observations across time, and show that exact identifiability is possible. However, this case requires a prohibitively large number of experiments to guarantee identifiability. Finally, in section 4.3 we will consider the most realistic case of observing a few observations across time, and discuss the conditions under which recovery of Ψ is possible.

4.1. Stationary observations

In the stationary observation model, we are given samples from a fully mixed process $\rho(\infty,x)$. In this case, one time observation is sufficient to exactly identify the potential. This follows from representing the stochastic process in Eq. 1 as a parabolic partial differential equation (PDE).

Theorem 2 (Fokker-Planck (Jordan et al., 1998))**.** *Given the SDE in equation 1, with drift* $\mu(x) = -\nabla\Psi(x)$*, the marginal distribution* $\rho(t,x)$ *fulfills:*

$$\frac{\partial\rho}{\partial t} = div(\rho(t,x)\nabla\Psi(x)) + \sigma^2\nabla^2\rho(t,x) \quad (4)$$

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

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

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

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

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

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

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

4.2. Many time observations

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

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

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

Theorem 3 (Uniqueness of Fokker-Planck like operators)**.** *Let $\Psi(x)$ be a continuously differentiable solution to the following elliptic PDE:*

$$f(x) = \nabla^2\Psi(x)\tau(x) + \nabla\Psi(x)\nabla\tau(x) + \sigma^2\nabla^2\tau(x) \quad (5)$$

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

Equation 5 is fulfilled in the short-time case with, $f = \frac{\partial\rho}{\partial t}$, $\tau = \rho$ and in the time-integral case, $f(x) = \rho(t_0,x) - \rho(t_n,x)$ and $\tau(x) = \int_0^T \rho(t,x)dt$.

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

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

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

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

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

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

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

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

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

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

The proof of Thm. 3 illustrates that the recoverability depends critically on $\tau(x) > 0$. Thus in the time-integral case, the regions which can be clearly recovered are those over which $\tau(x) = \int_0^T \rho(t,x)dt$ has large mass. Compared to the stationary situation, this is substantially better; we will get accurate estimates of Ψ over the entire timecourse of $\rho(0,x)\dots\rho(T,x)$.

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

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

4.3. Few time observations

In more realistic settings, we may get many samples, but very few time observations such that the time-integral uniqueness theorem does not hold. We analyze this case and establish two results: first, we establish exact identifiability in one dimension (Thm. 4) and give evidence for the conjecture in multiple dimensions (Cor. 1). Next, we establish that a sufficiently mixed final time observation is sufficient for uniqueness (Thm. 5) and derive a model constraint based on this theorem (Eq. 6).

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

Theorem 4 (1-D identifiability)**.** *Assume there exists some c such that $\sigma > c > 0$; boundaries a, b such that $\rho(t,a) = 0$ and $\rho(t,b) = 0$ for all t; and the marginal densities are Holder continuous with $\rho(t,x) \in H^{2+\lambda}$.*

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

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

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

Corollary 1 (Finite identifiability of Ψ)**.** *Let Ψ_0 and Ψ_1 be candidate potentials such that given $\rho_0(0,x) = \rho_1(0,x)$ and*

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

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

[3]The requirement of three marginal distributions is due to the more general nature of (Goldman, 2010, Problem 1). We believe only two marginals are necessary.

Proof. See Supp. section S.1. The statement reduces to short-time uniqueness studied in section 4.2. □

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

Theorem 5 (Relative fisher information constraint)**.** *Let $\rho(0,x)$ and $\rho(t_n,x)$ be marginal distributions associated with the potential Ψ. Then, if the final time $\rho(t_n,x)$ is sufficiently mixed:*

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

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

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

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

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

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

Theorem 5 implies that if we are willing to assume that $\rho(t_n,x)$ is close to mixed, and we can ensure that our estimated $\hat{\Psi}$ has a tight bound on $-\frac{\partial}{\partial t} D(\rho(t_n,x)||\exp(-\hat{\Psi}(x)/\sigma^2))$, then we can recover a good approximation to the true Ψ. In practice this assumption and constraint is straightforward to fulfill: experimental designs often track cell populations until they do not show substantial changes ($\rho(t_n,x)$ is close to mixed) and we can fit $\hat{\Psi}$ under the constraint that it is smooth with bounded gradient and

$$D(\rho(t_n,x)||\exp(-\hat{\Psi}(x)/\sigma^2)) \leq \eta. \tag{6}$$

Which implicitly bounds the mixedness in Thm. 5 by the JKO theorem (Thm. 1). Thus we have established a constraint (Eq. 6) and experimental condition (Thm. 5) under which we can reliably recover the underlying dynamics even with few timepoints.

5. Inference

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

5.1. Loss function and regularization

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

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

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

$$\lim_{\hat{t}\to 0} -\hat{t}\log(P(X(t+\hat{t}) = x(t+\hat{t})|X(t) = x(t))) = \frac{1}{4}\inf_{\sigma\in S_{m_n}} \sum_{i=1}^{m_n} |x(t+\hat{t})_i - x(t)_{\sigma(i)}|_2^2. \quad (7)$$

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

We regularize this loss function with an entropic regularizer. Thm. 5 states that if $\frac{\partial}{\partial t} D(\rho(t_n, x)||\exp(-\Psi(x)/\sigma^2))$ is small then we can recover any mixed potential. We fulfill this mixing constraint by controlling the relative entropy in Eq. 6, which we write as

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

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

$$\min_{\Psi\in K} \left[\sum_{i=1}^{n} W_2(\hat{\rho}(t_i, x), \rho_\Psi(t_i, x))^2\right] + \tau \sum_{j=1}^{m_n} \frac{\Psi(x_j)}{\sigma^2}. \quad (8)$$

The similarity of Eq. 8 to the JKO theorem (Thm. 1) is not coincidental. One interpretation of the JKO theorem is that W_2 is the natural metric over marginal distributions and likelihood is the natural measure of model fit over Ψ.

5.2. Diffusions as a recurrent network

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

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

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

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

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

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

Applying this to the first order time discretization in Eq. 2, a draw $\overline{y}_i^t$ of our stochastic process can be simulated as:

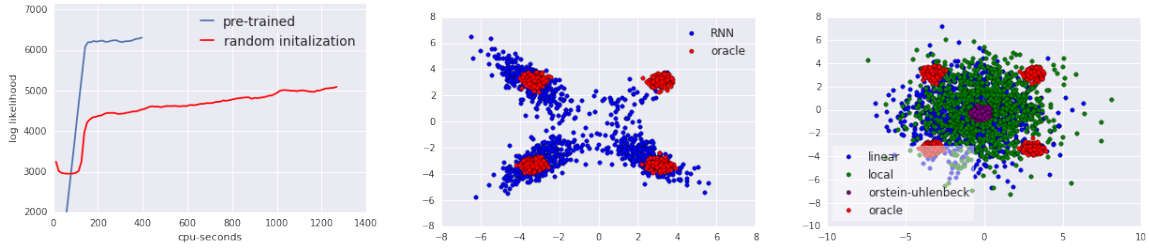
$$\overline{y}_i^{t+dt} = \overline{y}_i^t + \Delta t \sum_k h'(w_k \overline{y}_i^t + b_k)w_k g_k + \sqrt{\Delta t \sigma^2} z_{it} \quad (9)$$

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

5.3. Optimization

Optimizing the full objective function (Eq. 8) directly via backpropagation across time is slow and sensitive to the initialization. Exploiting the connection between RNNs and the diffusion, we can pre-train the model by optimizing the regularizer alone: $\sum_{j=1}^{m_n} \Psi(x_j)/\sigma^2$ under the constraint

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

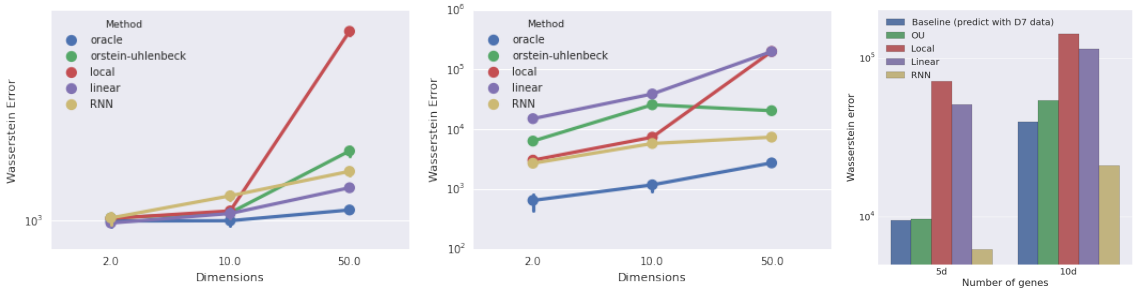


(a) Stationary pre-training improves both runtime and goodness of fit.

(b) RNN predictions are similar to the true dynamics on 50D data.

(c) Example prediction of baselines on same data

Figure 2. The pre-trained RNN captures the multimodal dynamics of the Styblinski flow even in 50-dimensions.



(a) Quadratic high dimensional flow

(b) Styblinski flow in high dimensions

(c) Gene expression (D4)

Figure 3. Held-out goodness of fit (lower is better), as measured by Wasserstein distance. 'Oracle' represents the error from Monte Carlo sampling for the true gradient flow. The RNN parametrization performs best across a wide range of tasks.

that $\int \exp(-\Psi(x)/\sigma^2)dx = 1$. We solve this optimization problem with contrastive divergence (Hinton, 2002) using the first-order Euler scheme in Eq. 9 to generate negative samples.

After this initialization, we perform backpropagation over time on our objective function, with ρ_Ψ approximated via Monte Carlo samples using Eq. 9 and the Wasserstein error approximated using Sinkhorn distances. These stochastic gradients are then used in Adagrad to optimize Ψ(Duchi et al., 2011). We implement the entire method in Theano, and code is available at `https://github.com/thashim/population-diffusions`.

6. Results

We now demonstrate the effectiveness of both the pre-training and RNN parametrization. [5]

6.1. Effectiveness of the stationary pre-training

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

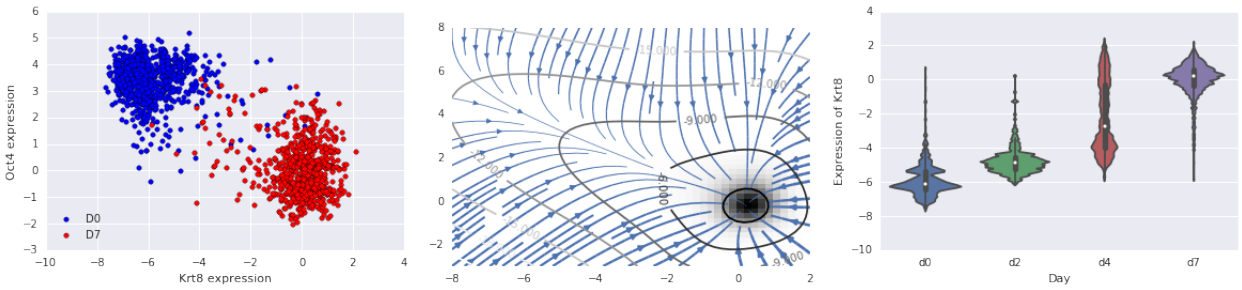
6.2. Learning high dimensional flows

One of the primary advantages of using recurrent networks and sums-of-ramps as a potential is that they behave well in high-dimensional estimation problems. We compare our RNN model against a linear $\Psi(x)$, the Orstein-Uhlenbeck process (quadratic $\Psi(x)$), and a local sum-of-gaussian potentials parametrization for $\Psi(x)$ (details in Sec. S.4).

In the first task (Fig. 3(a)), we have a population evolution in $\mathbb{R}^d$ for $d \in \{2, 10, 50\}$ according to a unit quadratic potential $\Psi(x) = |x|_2^2$. The initial measurement is 500 samples drawn from a normal distribution with $1/2$ scale centered at $(5, 0, 0 \ldots 0)$, and the final time measurement is 500 samples at $t = 1$ with $\sigma = 1.5$. This tests whether our model can recover a simple, high-dimensional potential function. In this case, the simple dynamics mean that the parametric models (Orstein-Uhlenbeck and Linear flows) perform quite well. The RNN parametrization is competitive with these models in as the dimensionality increases, and substantially outperforms the local model (Fig. 3(a)).

In the second task (Fig. 3(b)), we consider a population over $d \in \{2, 10, 50\}$ with two of the dimensions evolving according to the Styblinski flow ($\Psi(x) = ||3x^3 - 32x + 5||_2^2$), and the other dimensions set to zero. This tests

[5] Step-size is selected by grid search (see section S.3 for other parameter settings). σ is assumed known in the simulations, and fixed to the observed marginal variance for the RNA-seq data.



(a) D0 and D7 distributions of Oct4 (y-axis) and Krt8 (x-axis) (b) Learned differentiation dynamics (c) Distributions of true Krt8 expression

Figure 4. Observed data and learned model for single-cell RNA-seq data

whether our model can identify a complex low-dimensional potential embedded in a high-dimensional space. Example outputs in Fig. 2(b) and 2(c) demonstrate that our RNN model can model the multi-modal dynamics embedded within a high-dimensional space. The quantitative error in Fig. 3(b) shows that the local and RNN methods perform best at low (2-10) dimensions, but the local method rapidly degenerates in higher dimensions. In both cases, our RNN approach produces substantially lower Wasserstein loss compared both parametric and local approaches.

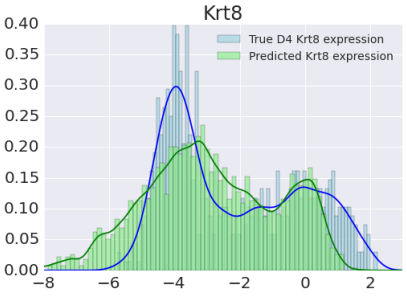


Figure 5. D4 predictions of Krt8 recapitulate bimodality

6.3. Analysis of Single-cell RNA-seq

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

Fitting our RNN model across the top five and ten most differential genes (as determined by the Wassertein distance between D0 and D7 distributions for each gene), our RNN method performs best compared to baselines (Fig3(c)), and is the only one to perform better than the trivial baseline of predicting the D4 gene expression using D7 data. We find that ten genes is the limit for accurate prediction with a few hundred cells; in higher dimensions the RNN begins to behave much like the linear model. As the number of captured cells in single-cell RNA-seq grows, our RNN model will be capable of modeling more complex multivariate potentials.

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

Fitting our RNN on this two dimensional problem shown in Fig. 4(a) we obtain a potential function with a single minimum (Fig. 4(b)) demonstrating that differentiation is concentrated around a linear path connecting the D0 and D7 distributions. Surprisingly, this simple unimodal potential predicts a bimodal distribution for the D4 Krt8 distribution shown in Fig. 5 despite the lack of bimodality in either the input data (Fig 4(a)) or the potential (Fig 4(b)). [6]

The bimodality arises from modeling the quantitative dynamics from D0 to D7, and provides evidence that even with as few as two time measurments, complex dynamics can be recovered from population level observations.

7. Discussion

Our work establishes the problem of recovering an underlying potential function using samples from the population distribution. Using a variational interpretation of diffusions, we derive natural and scalable losses and regularizers. Finally, we demonstrate through multiple synthetic datasets and a real single cell RNA-seq dataset that our model performs well in a high-dimensional setting.

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

Acknowledgements

We would like to thank the reviewers for their helpful comments in revising the paper.

This research was funded by the National Institute of Health under grants to D.G. and T.J. 1U01HG007037-01 and 1R01HG008363-01.

References

Adams, Stefan, Dirr, Nicolas, Peletier, Mark, and Zimmer, Johannes. Large deviations and gradient flows. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 371 (2005):20120341, 2013.

Bejan, Adrian and Merkx, Gilbert W. *Constructal theory of social dynamics*. Springer, 2007.

Bernstein, Garrett and Sheldon, Daniel. Consistently estimating markov chains with noisy aggregate data. In *Artificial Intelligence and Statistics*, pp. 1142–1150, 2016.

Bhattacharya, Sudin, Zhang, Qiang, and Andersen, Melvin E. A deterministic map of waddington's epigenetic landscape for cell fate specification. *BMC systems biology*, 5(1):85, 2011.

Cuturi, Marco. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

De Cezaro, Adriano and Johansson, B Tomas. A note on uniqueness in the identification of a spacewise dependent source and diffusion coefficient for the heat equation. *arXiv preprint arXiv:1210.7346*, 2012.

Duchi, John, Hazan, Elad, and Singer, Yoram. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

Fleming, Chris H, Fagan, William F, Mueller, Thomas, Olson, Kirk A, Leimgruber, Peter, and Calabrese, Justin M. Rigorous home range estimation with movement data: a new autocorrelated kernel density estimator. *Ecology*, 96 (5):1182–1188, 2015.

Fukunaga, Keinosuke and Hostetler, Larry D. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1):32–40, 1975.

Gers, Felix A, Schmidhuber, Jürgen, and Cummins, Fred. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

Goldman, NL. Uniqueness classes in inverse problems for parabolic equations with several unknown coefficients. In *Doklady Mathematics*, volume 82, pp. 573–577. Springer, 2010.

Hanna, Jacob, Saha, Krishanu, Pando, Bernardo, Van Zon, Jeroen, Lengner, Christopher J, Creyghton, Menno P, van Oudenaarden, Alexander, and Jaenisch, Rudolf. Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature*, 462(7273):595–601, 2009.

Hicks, Stephanie C, Teng, Mingxiang, and Irizarry, Rafael A. On the widespread and critical impact of systematic bias and batch effects in single-cell rna-seq data. *bioRxiv*, pp. 025528, 2015.

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

Jordan, Richard, Kinderlehrer, David, and Otto, Felix. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Kalbfleisch, John David and Lawless, Jerald F. Least-squares estimation of transition probabilities from aggregate data. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pp. 169–182, 1984.

Klein, Allon M, Mazutis, Linas, Akartuna, Ilke, Tallapragada, Naren, Veres, Adrian, Li, Victor, Peshkin, Leonid, Weitz, David A, and Kirschner, Marc W. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

Lund, Steven P, Hubbard, Joseph B, and Halter, Michael. Nonparametric estimates of drift and diffusion profiles via fokker–planck algebra. *The Journal of Physical Chemistry B*, 118(44):12743–12749, 2014.

Luo, Yang, Lim, Chea Lu, Nichols, Jennifer, Martinez-Arias, Alfonso, and Wernisch, Lorenz. Cell signalling regulates dynamics of nanog distribution in embryonic stem cell populations. *Journal of The Royal Society Interface*, 10(78):20120525, 2013.

Manton, Kenneth G, Gu, XiLiang, and Lowrimore, Gene R. Cohort changes in active life expectancy in the us elderly population: experience from the 1982–2004 national long-term care survey. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 63(5):S269–S281, 2008.

Marco, Eugenio, Karp, Robert L, Guo, Guoji, Robson, Paul, Hart, Adam H, Trippa, Lorenzo, and Yuan, Guo-Cheng. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings*

of the National Academy of Sciences, 111(52):E5643–E5650, 2014.

Markowich, Peter A and Villani, Cédric. On the trend to equilibrium for the fokker-planck equation: an interplay between physics and functional analysis. *Mat. Contemp*, 19:1–29, 2000.

Mineiro, Paul, Movellan, Javier, and Williams, Ruth J. Learning path distributions using nonequilibrium diffusion networks. *Advances in neural information processing systems*, pp. 598–604, 1998.

Morris, Rob, Sancho-Martinez, Ignacio, Sharpee, Tatyana O, and Belmonte, Juan Carlos Izpisua. Mathematical approaches to modeling development and reprogramming. *Proceedings of the National Academy of Sciences*, 111(14):5076–5082, 2014.

Pavliotis, Grigorios A. Stochastic processes and applications. *Diffusion Processes, the Fokker-Planck*, 2014.

Qiu, Xiaojie, Ding, Shanshan, and Shi, Tieliu. From understanding the development landscape of the canonical fate-switch pair to constructing a dynamic landscape for two-step neural differentiation. *PLoS One*, 7(12), 2012.

Robine, Jean-Marie and Michel, Jean-Pierre. Looking forward to a general theory on population aging. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 59(6):M590–M597, 2004.

Setty, Manu, Tadmor, Michelle D, Reich-Zeliger, Shlomit, Angel, Omer, Salame, Tomer Meir, Kathail, Pooja, Choi, Kristy, Bendall, Sean, Friedman, Nir, and Pe'er, Dana. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology*, 2016.

Sisan, Daniel R, Halter, Michael, Hubbard, Joseph B, and Plant, Anne L. Predicting rates of cell state change caused by stochastic fluctuations using a data-driven landscape model. *Proceedings of the National Academy of Sciences*, 109(47):19262–19267, 2012.

Sohl-Dickstein, Jascha, Weiss, Eric A, Maheswaranathan, Niru, and Ganguli, Surya. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015.

Tarantola, Albert. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, 2005.

Tereshko, Valery. Reaction-diffusion model of a honeybee colonys foraging behaviour. In *Parallel Problem Solving from Nature PPSN VI*, pp. 807–816. Springer, 2000.

Trapnell, Cole, Cacchiarelli, Davide, Grimsby, Jonna, Pokharel, Prapti, Li, Shuqiang, Morse, Michael, Lennon, Niall J, Livak, Kenneth J, Mikkelsen, Tarjei S, and Rinn, John L. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature biotechnology*, 32(4):381, 2014.

Van Der Plas, Adriaan P. On the estimation of the parameters of markov probability models using macro data. *The Annals of Statistics*, pp. 78–85, 1983.

Waddington, Conrad Hal et al. Organisers and genes. *Organisers and Genes.*, 1940.