Mathematical background

for

Optimal-transport analysis
of single-cell gene expression
identifies developmental trajectories in
reprogramming.

Contents

I. Modeling developmental processes with optimal transport

We developed a method to model development based on optimal transport. Section 1 reviews the concept of gene expression space and introduces our probabilistic framework for time series of expression profiles. Section 2 introduces our key modeling assumption to infer temporal couplings over short time scales. Section 3 shows how we can compute an optimal coupling between adjacent time points by solving a convex optimization problem, and how we can leverage an assumption of Markovity to compose adjacent time points and estimate temporal couplings over longer intervals. Section 4 describes how to interpret transport maps. Specifically, Section 4.1 shows how to compute ancestors and descendants of cells, Section 4.2 establishes a connection between entropic OT and Brownian motion of indistinguishable particles, and Section 4.3 shows how OT generalizes Waddington's classical metaphor of development in terms of a potential energy landscape. While Waddington's picture can only describe cell autonomous processes, which are akin to gradient flows in gene expression space, OT can describe gradient flows in the space of probability distributions on gene expression space. These can involve interactions between particles.

1. Developmental processes in gene expression space

A collection of mRNA levels for a single cell is called an *expression profile* and is often represented mathematically by a vector in *gene expression space*. This is a vector space that has dimension equal to the number of genes, with the value of the ith coordinate of an expression profile vector representing the number of copies of mRNA for the ith gene. Note that real cells only occupy an integer lattice in gene expression space (because the number of copies of mRNA is an integer), but we pretend that cells can move continuously through a real-valued G dimensional vector space.

As an individual cell changes the genes it expresses over time, it moves in gene expression space and describes a trajectory. As a population of cells develops and grows, a *distribution* on gene expression space evolves over time. When a single cell from such a population is measured with single cell RNA-seq, we obtain a noisy estimate of the number of molecules of mRNA for each gene. We represent the measured expression profile of this single cell as a sample from a probability distribution on gene expression space. This sampling captures both (a) the randomness in the measurement process (due to subsampling reads, technical issues, etc.) and (b) the random selection of a cell from the population. We treat this probability distribution as *nonparametric* in the sense that it is not specified by any finite list of parameters.

In the remainder of this section we introduce a precise mathematical notion for a *developmental process* as a special type of stochastic process (with a modified notion of coupling to accommodate cellular growth and death). Our primary goal is to infer the ancestors and descendants of subpopulations evolving according to an unknown developmental process. This information is encoded in the *temporal coupling* of the process, which is lost because we kill the cells when we perform scRNA-Seq. We claim it is possible to recover the temporal coupling over short time scales provided that cells don't change too much. We show in the remainder of this appendix how to do this with *optimal transport.*

1.1. A mathematical model of developmental processes

We begin by formally defining a precise notion of the developmental trajectory of an individual cell and its descendants. Intuitively, it is a continuous path in gene expression space that bifurcates with every cell division. Formally, we define it as follows:

Definition 1 (single-cell developmental trajectory)**.** *Consider a cell $x(0) \in \mathbb{R}^G$. Let $k(t) \geq 0$ specify the number of descendants at time t, where $k(0) = 1$. A single-cell developmental trajectory is a continuous function*

$$x : [0, T) \to \underbrace{\mathbb{R}^G \times \mathbb{R}^G \times \ldots \times \mathbb{R}^G}_{k(t) \text{ times}}.$$

This means that $x(t)$ is a $k(t)$-tuple of cells, each represented by a vector in $\mathbb{R}^G$:

$$x(t) = \left(x_1(t), \ldots, x_{k(t)}(t)\right).$$

We refer to the cells $x_1(t), \ldots, x_{k(t)}(t)$ as the descendants of $x(0)$.

Note that we cannot directly measure the temporal dynamics of an individual cell because scRNA-Seq is a destructive measurement process: scRNA-Seq lyses cells so it is only possible to measure the expression profile of a cell at a single point in time. As a result, it is not possible to directly measure the descendants of that cell, and the full trajectory is unobservable. However, one can hope to learn something about the probable trajectories of individual cells by measuring snapshots from an evolving population.

Published methods typically represent the aggregate trajectory of a population of cells by means of a graph structure. While this recapitulates the branching path traveled by the descendants of an individual cell, it may over-simplify the stochastic nature of developmental processes. Individual cells have the potential to travel through different paths, but any given cell travels one and only one such path. Our goal is to assign a likelihood to the set of possible paths, which in general are not finite and therefore cannot be a represented by a graph.

We define a developmental process to be a time-varying probability distribution on gene expression space. One simple example of a distribution of cells is that we can represent a set of cells $x_1, \ldots, x_n$ by the distribution

$$\mathbb{P} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i},$$

where δ_x denote the Dirac delta (a distribution placing unit mass on x). Similarly, we can represent a set of single-cell trajectories $x_1(t), \ldots, x_n(t)$ with a distribution over trajectories. This is a special case of a developmental process, which we define as follows:

Definition 2 (developmental process)**.** *A developmental process $\mathbb{P}_t$ is a time-varying distribution (i.e. stochastic process) on gene expression space.*

Recall that a stochastic process is determined by its temporal dependence structure. This is specified by the coupling (i.e. joint distribution) between random variables at different time points. Given that a cell has a particular expression profile y at time t_2, where did it come from at time t_1? This is precisely the information lost by not tracking individual cells over time.

Definition 3 (temporal coupling)**.** *Let $\mathbb{P}_t$ be a developmental process and consider two time points $s < t$. Let $X_t \sim \mathbb{P}_t$ denote the expression profile of a random cell at time t and let X_s denote the expression profile of its cell of origin at time s.*

The temporal coupling $\gamma_{s,t}$ is defined as the law of the joint distribution:

$$\gamma_{s,t} = \mathcal{L}(X_s, X_t).$$

Equivalently,

$$\int_{x \in A} \int_{y \in B} \gamma_{s,t}(x, y) dx dy = \Pr\{X_s \in A, X_t \in B\}$$

for any sets $A, B \subset \mathbb{R}^G$.

The temporal coupling $\gamma_{s,t}$ is not technically a coupling of $\mathbb{P}_s$ and $\mathbb{P}_t$ in the standard sense because it does not necessarily have marginals $\mathbb{P}_s$ and $\mathbb{P}_t$:

$$\int \gamma_{s,t}(x, y) dx = \mathbb{P}_t(y), \quad \text{but} \quad \int \gamma_{s,t}(x, y) dy \neq \mathbb{P}_s(x).$$

Biologically, this is the case when cells grow at different rates. Then proliferative cells from the earlier time point will be over-represented when we look for the origin of cells at the later time point. In the following definition, we introduce a relative growth rate function to describe the relationship between the expression profile of a cell and the average number of living descendants it gives rise to after certain amount of time.

Definition 4. *A relative growth rate function associated with a temporal coupling is a function $g(x)$ satisfying*

$$\int \gamma_{s,t}(x, y) dy = \mathbb{P}_s(x) \frac{g(x)^{t-s}}{\int g(x)^{t-s} d\mathbb{P}_s(x)}.$$

The integral on the left-hand side represents the amount of mass coming out of x and going to any y. The term $\mathbb{P}(x)$ on the right hand side accounts for the abundance of cells with expression profile x, and the function $g(x)$ represents the exponential increase in mass per unit time.

Having defined the notion of developmental processes and temporal couplings, we now turn to estimating these from data.

2. The optimal transport principle for developmental processes

ScRNA-Seq allows us to sample cells from a developmental process at various time points, but it does not give any information about the coupling between successive time points. Without making any assumptions, it is impossible to recover the temporal coupling even given infinite data in the form of the full distributions $\mathbb{P}_s$ and $\mathbb{P}_t$. However, we claim that it is reasonable to assume that cells don't change expression by large amounts over short time scales. This assumption allows us to estimate the coupling and infer which cells go where.

We begin with a simple one-dimensional example to build intuition.

Example 1. *Let $X_0 \sim \mathcal{N}(0, \sigma^2)$ and $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ be one dimensional Gaussian variables representing the location of a particle at time 0 and at time 1. If we believe that the particle cannot move very far over a short amount of time, then how can we infer the coupling γ specifying the joint distribution of the pair (X_0, X_1)? One simple heuristic to estimate $\hat{\gamma}$ is to minimize the squared distance that the particle moves from time 0 to time 1:*

$$\hat{\gamma} \leftarrow \arg\min_{\pi} \ \mathbb{E}_\pi \|X_0 - X_1\|^2.$$

We minimize over all couplings π with marginals $\mathcal{N}(0, \sigma^2)$ and $\mathcal{N}(\mu, \sigma^2)$. One can check that the optimal joint distribution is a two dimensional Gaussian with the following dependence structure:

$$X_1 = X_0 + \mu.$$

This heuristic to couple marginals is called *optimal transport* (OT) (Villani, 2008). If $c(x, y)$ denotes the cost of transporting a unit mass from x to y, and the amount we transfer from x to y is $\pi(x, y)$, then the total cost of transporting mass according to such a transport plan π is given by

$$\iint c(x,y)\pi(x,y)dxdy.$$

In this paper we focus exclusively on the cost defined by the squared-Euclidean distance

$$c(x,y) = \|x-y\|^2,$$

on an appropriate input space (see Chapter II for details). We make this choice to focus on this cost function because of the many well-known attractive theoretical properties it enjoys over other cost functions (Villani, 2008).

The *optimal* transport plan minimizes the expected cost subject to marginal constraints:

$$\begin{aligned} \pi(\mathbb{P},\mathbb{Q}) = \underset{\pi}{\text{minimize}} \quad & \iint c(x,y)\pi(x,y)dxdy \\ \text{subject to} \quad & \int \pi(x,\cdot)dx = \mathbb{Q} \\ & \int \pi(\cdot,y)dy = \mathbb{P}. \end{aligned} \tag{1}$$

Note that this is a linear program in the variable π because the objective and constraints are both linear in π. The optimal objective value defines the *transport distance* between $\mathbb{P}$ and $\mathbb{Q}$ (it is also called the Earthmover's distance or Wasserstein distance). Unlike many other ways to compare distributions (such as KL-divergence or total variation), optimal transport takes the geometry of the underlying space into account. For example, the KL-Divergence is infinite for any two distributions with disjoint support, but the transport distance depends on the separation of the support. For a comprehensive treatment of the rich mathematical theory of optimal transport, we refer the reader to (Villani, 2008).

2.1. The optimal transport principle

We propose to use optimal transport to estimate the temporal coupling of a developmental process. We make two modifications to classical optimal transport to adapt it to our biological setting.

1. Classical optimal transport has conservation of mass built into the constraints (1). We account for growth by rescaling the distribution $\mathbb{P}_t$ before applying OT.

2. The coupling identified by classical optimal transport is purely deterministic in the sense that each point is transported to a single point[1]. However, for cells whose fates are not completely determined, the true coupling should have a degree of entropy to it. We therefore add a term to the objective to promote entropy in the transport coupling.

   Injecting a small amount of entropy also makes sense even for a population of cells with truly deterministic descendant distribution. When we sample finitely many cells at time t_2, the true descendants of any given t_1 cell are not captured. Therefore entropy in the transport map can be used to represent our statistical uncertainty in the inferred descendant distribution.

[1] There may be non-deterministic plans achieving the same cost (e.g. if all points are equidistant), but there is always an optimal plan that is deterministic.

In order to state the optimal transport principle, we first introduce some notation. Let $\mathbb{P}_t$ denote a developmental process with temporal coupling $\gamma_{s,t}$ and with relative growth function $g(x)$. Let $\mathbb{Q}_s$ denote the distribution obtained by rescaling $\mathbb{P}_s$ by the relative growth rate:

$$\mathbb{Q}_s(x) = \mathbb{P}_s(x) \frac{g^{t-s}(x)}{\int g^{t-s}(z) d\mathbb{P}_s(z)}.$$

Finally, let $\pi_{s,t}(\epsilon)$ denote the entropy-regularized optimal transport coupling of $\mathbb{Q}_s$ and $\mathbb{P}_t$, defined as the solution to the following optimization problem:

$$\begin{aligned} \pi_{s,t}(\epsilon) = \underset{\pi}{\text{minimize}} \quad & \iint c(x,y)\pi(x,y)dxdy - \epsilon \iint \pi(x,y) \log \pi(x,y) dxdy \\ \text{subject to} \quad & \int \pi(x,\cdot)dx = \mathbb{Q}_s \\ & \int \pi(\cdot,y)dy = \mathbb{P}_t. \end{aligned} \tag{2}$$

We now state the optimal transport principle for developmental processes:

$$s \approx t \implies \pi_{s,t}(\epsilon) \approx \gamma_{s,t}.$$

In words, over short time scales, the true coupling is well approximated by the OT coupling. In section 3, we show how to estimate $\pi_{s,t}(\epsilon)$ from data (we occasionally omit the dependence on ϵ and write $\pi_{s,t}$). This in turn gives us an estimate of $\gamma_{s,t}$.

3. Inferring temporal couplings from empirical data

In this section we show how to estimate the temporal couplings of a developmental process from data.

Definition 5 (developmental time series)**.** *A developmental time series is a sequence of samples from a developmental process $\mathbb{P}_t$ on $\mathbb{R}^G$. This is a sequence of sets $S_1, \dots, S_T \subset \mathbb{R}^G$ collected at times $t_1, \dots, t_T \in \mathbb{R}$. Each S_i is a set of expression profiles in $\mathbb{R}^G$ drawn independently from $\mathbb{P}_{t_i}$.*

From this input data, we form an empirical version of the developmental process. Specifically, at each time point t_i we form the empirical probability distribution supported on the data $x \in S_i$. We summarize this in the following definition:

Definition 6 (Empirical developmental process)**.** *An empirical developmental process $\hat{\mathbb{P}}_t$ is a time varying distribution constructed from a developmental time course $S_1, \dots, S_T$:*

$$\hat{\mathbb{P}}_{t_i} = \frac{1}{|S_i|} \sum_{x \in S_i} \delta_x. \tag{3}$$

The empirical developmental process is undefined for t $\notin \{t_1, \dots, t_T\}$.

In order to estimate the coupling from time t_1 to time t_2, we first construct an initial estimate of the growth rate function $g(x)$. In practice, we form an initial estimate $\hat{g}(x)$ as the expectation of a birth-death process on gene expression space with birth-rate $\beta(x)$ and death rate $\delta(x)$ defined in terms of expression levels of genes involved in cell proliferation and apoptosis (see **Estimating birth and death rates and computing transport maps** in STAR Methods). We ultimately leverage techniques from *unbalanced transport* (Chizat et al., 2018) to refine this initial estimate to learn cellular growth and death rates automatically from data (see Chapter II).

We then form the rescaled empirical distribution

$$\hat{\mathbb{Q}}_{t_1}(x) = \hat{\mathbb{P}}_{t_1}(x) \frac{\hat{g}(x)^{t_1 - t_2}}{\int \hat{g}(z)^{t_1 - t_2} d\hat{\mathbb{P}}_{t_i}(z)},$$

and compute the optimal transport map $\hat{\pi}_{t_1,t_2}$ between $\hat{\mathbb{Q}}_{t_1}$ and $\hat{\mathbb{P}}_{t_2}$.

3.1. Estimating couplings between adjacent time points

In order to identify an optimal transport plan connecting $\hat{\mathbb{Q}}_{t_1}$ and $\hat{\mathbb{P}}_{t_2}$, we must solve an optimization problem with a matrix-valued optimization variable. In the classical zero-entropy setting, the optimization problem (2) is a linear program (when $\epsilon = 0$). While the classical optimal transport linear program can be difficult to solve for large numbers of points, fast algorithms have been recently developed (Cuturi, 2013) to solve the entropically regularized version of the transport program. Entropic regularization speeds up the computations because it makes the optimization problem strongly convex, and gradient ascent on the dual can be realized by successive diagonal matrix scalings called Sinkhorn iterations (Cuturi, 2013). These are very fast operations.

The scaling algorithm for entropically regularized transport has also been extended to work in the setting of **unbalanced transport** (Chizat et al., 2018), where the equality constraints are relaxed to bounds on the marginals of the transport plan (in terms of KL-divergence or total variation or a general f-divergence). In our application this is very attractive from a modeling perspective for the following reasons:

1. We may have misspecified the growth rate function $\hat{g}(x)$. Unbalanced transport adjusts the input growth rate in order to reduce the transport cost. This allows us to automatically learn growth rates from scratch (see Chapter II).
2. Even if the growth rates are completely uniform, the random sampling can introduce what looks like growth. For example, suppose there is a rare subpopulation of cells consisting of 5% of the total. If at one time point, we randomly sample fewer of these cells so that they comprise 4% of the total, and at the next time point we sample 6%, then it will look like this population has increased by 50%. Unbalanced transport can automatically adjust for this apparent growth.

We use both entropic regularization and unbalanced transport. To compute the transport map between the empirical distributions of expression profiles observed at time t_i and t_{i+1}, we solve the following optimization problem:

$$\begin{aligned} \hat{\pi}_{t_i,t_{i+1}} &= \arg\min_{\pi} \sum_{x \in S_i} \sum_{y \in S_{i+1}} c(x,y)\pi(x,y) - \epsilon \iint \pi(x,y) \log \pi(x,y) dx dy \\ &\qquad + \lambda_1 \mathrm{KL}\left[\sum_{x \in S_i} \pi(x,y) \middle\| d\hat{\mathbb{P}}_{t_{i+1}}(y) \right] + \lambda_2 \mathrm{KL}\left[\sum_{y \in S_{i+1}} \pi(x,y) \middle\| d\hat{\mathbb{Q}}_{t_i}(x) \right] \end{aligned} \tag{4}$$

where ϵ, λ_1 and λ_2 are regularization parameters. We provide guidelines for tuning these parameters in Chapter II.

This is a convex optimization problem in the matrix variable $\pi \in \mathbb{R}^{N_i \times N_{i+1}}$, where $N_i = |S_i|$ is the number of cells profiled at time t_i. It takes about 5 seconds to solve this unbalanced transport problem using the scaling algorithm of (Chizat et al., 2018) on a standard laptop with $N_i \approx 5000$.

Note that by default the densities (on the discrete set S_i) of the empirical distributions specified in equation (3) are simply $d\hat{\mathbb{P}}_{t_i}(x) = \frac{1}{N_i}$. However, in principle one could use nonuniform empirical distributions (e.g. if one wanted to include information about cell quality).

To summarize: given a sequence of expression profiles $S_1, \ldots, S_T$, we solve the optimization problem (4) for each successive pair of time points S_i, S_{i+1}. For the pair of time-points (t_i, t_{i+1}), this gives us a transport map $\hat{\pi}_{t_i,t_{i+1}}$. When we have enough data, this is a good estimate of $\pi_{t_i,t_{i+1}}$ because it is well known that transport maps are consistent in the sense that

$$\lim_{N_i, N_{i+1} \to \infty} \hat{\pi}_{t_i,t_{i+1}} = \pi_{t_i,t_{i+1}}.$$

Taken together with the optimal transport principle:

$$\pi_{t_i,t_{i+1}} \approx \gamma_{t_i,t_{i+1}},$$

we therefore can estimate $\gamma_{t_i,t_{i+1}}$ from $\hat{\pi}_{t_i,t_{i+1}}$ when N_i is large enough.

3.2. Estimating long-range couplings

We rely on an assumption of Markovity (or memorylessness) in order to estimate couplings over longer time intervals. Recall that a stochastic process is Markov if the future is independent of the past, given the present. Equivalently, it is fully specified by the couplings between pairs of time points. We define Markov developmental processes in a similar spirit:

Definition 7 (Markov developmental process)**.** *A Markov developmental process $\mathbb{P}_t$ is a time-varying distribution on $\mathbb{R}^G$ that is completely specified by couplings between pairs of time points in the following sense. For any three time points $s < t < \tau$, the long-range coupling $\gamma_{s,\tau}$ is equal to the composition of short-range couplings:*

$$\gamma_{t,\tau} \circ \gamma_{s,t} = \gamma_{s,\tau}.$$

Note that the optimal transport maps $\hat{\pi}_{s,t}$ do **not** necessarily have this compositional property! Composing the OT coupling from time s to t and then from t to τ is not the same as optimally transporting from s directly to τ. In general, we do not recommend computing OT maps directly between distant time points.

We leverage the Markovity assumption to estimate couplings over long time intervals by composing estimates over shorter intervals. Formally, for any pair of time points t_i, t_{i+k}, we estimate the coupling $\hat{\gamma}_{t_i,t_{i+k}}$ by composing as follows:

$$\hat{\gamma}_{t_i,t_{i+k}} = \hat{\pi}_{t_i,t_{i+1}} \circ \hat{\pi}_{t_{i+1},t_{i+2}} \circ \ldots \circ \hat{\pi}_{t_{i+k-1},t_{i+k}}.$$

These compositions are computed via ordinary matrix multiplication.

It is an interesting question to what extent developmental processes are Markov. On gene expression space, they are likely not strictly Markov because, for example, the history of gene expression can influence chromatin modifications, which may not themselves be fully reflected in the observed expression profile but could still influence the subsequent evolution of the process. However, it is possible that developmental processes could be considered Markov on some augmented space.

4. Interpreting transport maps

In the previous section we introduced the principle of optimal transport for time series of gene expression profiles. Given a time series of expression profiles $S_1, \ldots, S_T$, we use this principle to compute a sequence of transport maps between subsequent time slices. In this section we define the *ancestors* and *descendants* of any subset of cells from this sequence of transport maps in Section 4.1. Then, in Section 4.2 we explain an intuitive physical interpretation of entropy-regularization. Finally, in Section 4.3 we describe a connection between optimal transport, gradient flows, and Waddington's landscape.

4.1. Defining ancestors, descendants and trajectories

We now define the descendants and ancestors of subgroups of cells evolving according to a Markov (i.e. memoryless) developmental process.

Our definition of ancestors and descendants relies on a notion of *pushing* sets of cells through a transport map. Before defining ancestors and descendants, we introduce this terminology. As a distribution on the product space $\mathbb{R}^G \times \mathbb{R}^G$, a coupling γ assigns a number $\gamma(A, B)$ to any pair of sets $A, B \subset \mathbb{R}^G$

$$\gamma(A, B) = \int_{x \in A} \int_{y \in B} \gamma(x, y) dx dy.$$

This number $\gamma(A, B)$ represents the amount of mass coming from A and going to B. When we don't specify a particular destination, the quantity $\gamma(A, \cdot)$ specifies the full distribution of mass coming from A. We refer to this action as *pushing* A through the transport plan γ. More generally, we can also push a *distribution* μ forward through the transport plan γ via integration

$$\mu \mapsto \int \gamma(x, \cdot) d\mu(x).$$

We refer to the reverse operation as pulling a set B back through γ. The resulting distribution $\gamma(\cdot, B)$ encodes the mass ending up at B. We can also pull distributions μ back through γ in a similar way:

$$\mu \mapsto \int \gamma(\cdot, y) d\mu(y).$$

We sometimes refer to this as *back-propagating* the distribution μ (and to pushing μ forward as *forward propagation*).

Equipped with this terminology, we define ancestors and descendants as follows:

Definition 8 (descendants in a Markov developmental process)**.** *Consider a set of cells $C \subset \mathbb{R}^G$, which live at time t_1 are part of a population of cells evolving according to a Markov developmental process $\mathbb{P}_t$. Let γ_{t_1,t_2} denote the coupling from time t_1 to time t_2. The descendants of C at time t_2 are obtained by pushing C through γ.*

Definition 9 (ancestors in a Markov developmental process)**.** *Consider a set of cells $C \subset \mathbb{R}^G$, which live at time t_2 and are part of a population of cells evolving according to a Markov developmental process $\mathbb{P}_t$. Let π denote the transport map for $\mathbb{P}_t$ from time t_2 to time t_1. The ancestors of C at time t_1 are obtained by pulling C back through γ.*

Trajectories: We define to the *ancestor trajectory* to a set C as the sequence of ancestor distributions at earlier time points. Similarly, we refer to the *descendant trajectory* from a set C as the sequence of descendant distributions at later time points.

4.2. Interpreting the entropy regularization parameter

In this section we explain a physical interpretation of entropy-regularized optimal transport.

Consider a collection of N indistinguishable particles undergoing Brownian motion with diffusion coefficient ϵ. Suppose we observe the positions of N particles at times 0 and 1. But because the particles are indistinguishable, we don't know which particle at time 0 corresponds to each particle at time 1. If $N = 1$, this is of course not an issue, and the distribution on paths connecting the starting and ending point is called a *Brownian bridge*.

For $N > 1$, the distribution over possible paths connecting the starting and ending points involves two components:

1. A coupling of the particles specifying which particle goes where (because the particles are indistinguishable, this is not uniquely specified by the observations).

2. Given a matching, the distribution on paths for each matched pair is a Brownian bridge.

The coupling is a random permutation that matches points at time 0 to points at time 1. The distribution of this random permutation depends on the variance (or diffusion coefficient) of the Brownian motion. If the diffusion coefficient is larger, then it is more likely that particles will swap positions over larger distances. It turns out that the expected (i.e. average) coupling can be computed by maximum entropy optimal transport. These ideas can be traced back to Schrodinger's 1932 work in statistical electrodynamics (Schrodinger, 1932), but the connection to optimal transport was not made explicit until recently (Cuturi, 2013; Léonard, 2014). We summarize this in the following theorem:

Theorem 1. *Entropy regularized optimal transport gives the expectation of the distribution over couplings induced by Brownian motion, when the diffusion coefficient of the Brownian motion is equal to the entropy regularization parameter.*

4.3. Gradient flow and Waddington's landscape

In this section we show how optimal transport can be interpreted as a gradient flow in gene expression space (capturing cell-autonomous processes) or in the space of distributions (capturing cell-nonautonomous processes). For a full treatment of the rich OT theory of gradient flows, we refer the reader to (Ambrosio et al., 2005; Santambrogio, 2015).

We begin by considering the simple setting described by Waddington's landscape, which describes a gradient flow in gene expression space and is a special case of what we can capture with optimal transport. Mathematically, Waddington's landscape defines a potential function Φ assigning potential energy $\Phi(x)$ to a cell with expression profile x. The cells roll downhill according to the gradient of Φ to describe a trajectory $x(t)$ satisfying the differential equation

$$\frac{dx}{dt} = -\nabla\Phi(x). \tag{5}$$

This equation governing the trajectory of individual cells induces a flow in the distribution of the population of cells:

$$\frac{d\mathbb{P}_t}{dt} = \text{div}[\nabla\Phi(x)\mathbb{P}_t]. \tag{6}$$

Intuitively, this equation states that the change in mass for each small volume of space (on the left-hand side) is equal to the flux of mass in and out (given by the divergence on the right hand side).

Optimal transport can capture this type of potential driven dynamics: the true coupling specified by (5) is close to the optimal transport coupling over short time scales. To motivate this, we appeal to a classical theorem establishing a dynamical formulation of optimal transport.

Theorem 2 (Benamou and Brenier, 2001)**.** *The optimal objective value of the transport problem* (1) *is equal to the optimal objective value of the following optimization problem:*

$$\begin{array}{ll} \underset{\rho,v}{\text{minimize}} & \displaystyle\int_0^1 \int_{\mathbb{R}^G} \|v(t,x)\|^2 \rho(t,x) dt dx \\ \text{subject to} & \rho(0,\cdot) = \mathbb{P}, \quad \rho(1,\cdot) = \mathbb{Q} \qquad \cdot \\ & \nabla \cdot (\rho v) = \dfrac{\partial \rho}{\partial t} \end{array} \tag{7}$$

In this theorem, v is a vector-valued velocity field that advects[2] the distribution ρ from $\mathbb{P}$ to $\mathbb{Q}$, and the objective value to be minimized is the kinetic energy of the flow (mass × squared velocity). In our setting, the two distributions are snapshots $\mathbb{P}_s$ and $\mathbb{P}_t$ of a developmental process at two time points, and the theorem shows that the transport map $\pi_{s,t}$ can be seen as a point-to-point summary of a least-action continuous time flow, according to an unknown velocity field. In the special case when the velocity field is the gradient of a potential Φ (i.e. Waddington landscape), the theorem implies that the coupling (5) achieves the optimal transport cost. In other words, OT can capture potential driven dynamics. In addition, optimal transport can also describe much more general settings. This velocity field could change over time and also depend on the entire distribution of cells, so optimal transport can describe very general developmental processes including those with cell-cell interactions, as we describe below.

We will show that the evolution (6) is a special case of a *Wasserstein gradient flow* to minimize the linear energy functional

$$E(\mathbb{P}) = \int \Phi(x) d\mathbb{P}(x).$$

We will then describe non-linear gradient flows, which can capture cell-cell interactions.

To understand gradient flows, let's start with the familiar notion of gradient descent:

$$x_{k+1} = -\eta \nabla E(x_k) + x_k.$$

This can be rewritten as a *proximal procedure*, where one seeks to minimize E over all x in the proximity of x_k:

$$x_{k+1} = \quad \arg\min_x \qquad E(x) + \frac{1}{2\eta}\|x - x_k\|^2. \tag{8}$$

We can perform a similar proximal procedure in the space of distributions, replacing the Euclidean norm $\|\cdot\|^2$ with the Wasserstein distance:

$$\mathbb{P}_{k+1} = \quad \arg\min_\rho E(\rho) + \frac{1}{2\eta} W_2^2(\rho, \mathbb{P}_k). \tag{9}$$

This produces a sequence of iterates $\mathbb{P}_0, \mathbb{P}_1, \ldots, \mathbb{P}_k$. The gradient flow is the limit obtained as we shrink the step-size $\eta \downarrow 0$. In (Jordan et al., 1998), it's proven that for the linear energy functional

$$E(\mathbb{P}) = \int \Phi(x) d\mathbb{P}(x),$$

the limiting gradient flow converges to a solution of (6).

Going beyond the linear energy functional associated with Waddington's landscape, one could describe cell-cell interactions with an interaction energy of the form

$$E(\mathbb{P}) = \iint I(x, y) d\mathbb{P}(x) d\mathbb{P}(y).$$

Gradient flows for interaction potentials are discussed in chapter 7 of (Santambrogio, 2015).

Learning models of gene regulation Motivated by this interpretation of optimal transport as a gradient flow according to an unknown vector field, we describe a strategy to estimate such a vector field from data in Chapter II. We interpret the vector field as a model of gene regulation – it predicts gene expression at later time points as a function of transcription factor expression at current time points. We assume that the vector field does not change over time, and describes a cell-autonomous flow, but we do not assume that it comes from a potential function.

[2] *Advection*, a term borrowed from fluid mechanics, refers to the transport of a substance by bulk motion. The constraint that the divergence of the flow is equal to the rate of change of ρ means that ρ flows according to the velocity field v, without gaining or losing mass.

II. WADDINGTON-OT : Concepts and Implementation

Building on the theoretical foundations developed in Modeling developmental processes with optimal transport, we developed WADDINGTON-OT: our method for computing ancestor and descendant trajectories, interpolating developmental processes, inferring gene regulatory models, and visualizing developmental landscapes. We begin with an overview in Section 1, and we then describe the specific details in Sections 2 - 8.

1. Overview

To apply WADDINGTON-OT to a dataset, we pursue the following steps. The code is available on GitHub:

`https://github.com/broadinstitute/wot/`

Specifically, in the sections below we describe our procedures for

- computing transport maps
- computing trajectories to cell sets
- fitting local and global regulatory models
- interpolating the distribution of cells at held-out time points.

To keep the focus here general-purpose, we defer all reprogramming-specific details to the subsequent sections of STAR Methods.

Input data: The input to our suite of methods is a temporal sequence of single cell gene expression matrices, prepared as described in STAR Methods: **Preparation of expression matrices**.

Computing transport maps: Waddington-OT calculates transport maps between consecutive time points and automatically estimates cellular growth and death rates. In Section 2 below we provide guidelines for defining the cost function, selecting regularization parameters and (optionally) providing an initial estimate of growth and death rates.

Ancestors, descendants, and trajectories: We describe in Section 3 how we compute trajectories plot trends in gene expression. Briefly, the *developmental trajectory* of a subpopulation of cells refers to the sequence of ancestors coming before it and descendants coming after it. Using the transport maps, we can calculate the forward or backward transport probabilities between any two classes of cells at any time points. For example, we can take successfully reprogrammed cells at day 18 and use back-propagation to infer the distribution over their precursors at day 17.5. We can then propagate this back to day 17, and so on to obtain the ancestor distributions at each previous time point. This is the developmental trajectory to iPS cells. We can then readily plot trends in gene expression over time.

Fitting regulatory models: We describe our method to fit a regulatory model to the transport maps in Section 4. Transcription factors (TFs) that appear to play important roles along trajectories to key destinations are identified by two approaches. The first approach involves constructing a global regulatory model, related to the framework we describe above in Section I.4.3. Pairs of cells at consecutive time points are sampled according to their transport probabilities; expression levels of TFs in the cell at time t are used to predict expression levels of all non-TFs in the paired cell at time $t+1$, under the assumption that the regulatory rules are constant across cells and time points. (TFs are excluded from the predicted set to avoid cases of spurious self-regulation). The second approach involves local enrichment analysis. TFs are identified based on enrichment in cells at an earlier time point with a high probability ($> 80\%$) of transitioning to a given fate vs. those with a low probability ($< 20\%$).

Geodesic interpolation: To validate the temporal couplings, Waddington-OT can interpolate the distribution of cells at a held-out time point. The method is performing well if the interpolated distribution is close to the true held-out distribution (compared to the distance between different batches of the held-out distribution). Otherwise, it is possible that the method requires more data or finer temporal resolution.

Section 5 describes our method to interpolate the distribution of cells at a held-out time point. The specific application for validation of our method on iPS reprogramming data is presented in STAR Methods: **Validation by geodesic interpolation**. We performed extensive sensitivity analysis to show that our temporal couplings produce valid interpolations over a wide range of parameter settings perturbations to the data (downsampling cells or reads). See STAR Methods: **QUANTIFICATION AND STATISTICAL ANALYSIS** for this sensitivity analysis.

2. Computing transport maps

Recall that for any pair of time points we compute a transport plan that minimizes the expected cost of redistributing mass, subject to constraints involving the relative growth rate (see Chapter I for a precise statement of the optimization problem).

The transport map $\hat{\pi}_{t_1,t_2}$ connecting cells from time t_1 to cells from time t_2 has a row for each cell x at time t_1 and a column for each cell y at time t_2. Each row specifies the *descendant distribution* of a single cell x from time t_1. The descendant mass is the sum of all the entries across a row. This row-sum is proportional to the number of descendants that x will contribute to the next time point. Intuitively, the descendant distribution specifies which cells at time t_2 are likely to be descendants of x (see Section 4.1 of Chapter I for the formal definition of descendants in a developmental process).

Similarly, each column specifies the ancestor distribution of a cell y from time t_2. The ancestor mass is usually the same for each cell y. The ancestor distribution tells us which cells at time t_1 are likely to give rise to the cell y.

To compute these transport matrices, we need to specify a cost function, numerical values for the regularization parameters, and (optionally) an initial estimate for the relative growth rate.

2.1. Cost function

To compute the cost of transporting each individual point x from time t_1 to position y at time t_2, we first perform principal components analysis (PCA) on the data from this pair of time points. This dimensionality reduction is performed separately for each pair of adjacent time points. We define the cost function to be squared Euclidean distance in this 'local-PCA space'.

Finally, we normalize the cost matrix by dividing each entry by the median cost for that time interval. Here the cost matrix is the matrix with entries $C_{i,j} = c(x_i, y_j)$ for each x_i form time t_1 and y_j at time t_2.

This rescaling of the cost allows us to refer to specific numerical values of the regularization parameters, without worrying about the global scale of distances.

2.2. Regularization parameters

The optimization problem (4) involves three regularization parameters:

- The *entropy* parameter ϵ controls the entropy of the transport map. An extremely large entropy parameter will give a maximally entropic transport map, and an extremely small entropy parameter will give a nearly deterministic transport map. The default value is 0.05.
- λ_1 controls the degree to which transport is unbalanced along the rows. Large values of λ_1 impose stringent constraints related to relative growth rates. Small values of λ_1 give the algorithm more flexibility to change the relative growth rates in order to improve the transport objective. The default value is 1. To visually inspect the degree of unbalancedness, we recommend plotting the input row-sums vs the output row-sums of the transport map **(Figure S1D-F)**.
- λ_2 controls the degree to which transport is unbalanced along the columns. The default value is $\lambda_2 = 50$. This large value essentially imposes equality constraints for the column marginals. A smaller value of λ_2 would allow different amounts of mass to transport to some cells at time t_2. We strongly recommend keeping a large value for λ_2 so that the results are balanced along the columns. To visually inspect the degree of unbalancedness, one can plot the input column-sums vs the output column-sums of the transport map.

As we demonstrate in **QUANTIFICATION AND STATISTICAL ANALYSIS** in STAR Methods, our validation results are stable over a wide range of values for ϵ and λ_1.

2.3. Estimating relative growth rates

Our method solves the optimization problem (4) several times, using the output row-sums of the optimal transport map $\hat{\pi}_{t_1,t_2}$ as a new estimate for the relative growth rate function $\hat{g}(x)$. By default, we initialize with

$$\hat{g}(x) = 1,$$

so that all cells grow at the same rate. If one has some prior knowledge of growth rates (e.g. based on gene signatures of proliferation and apoptosis), this can be incorporated in the initial estimate for $\hat{g}(x)$. For our reprogramming data, we show how we formed an initial estimate for relative growth rates in **Estimating growth and death rates and computing transport maps**.

3. Ancestors, descendants, and trajectories

Given a set of cells C, we can compute the descendant distribution of the entire set by adding the descendant distributions of each cell in the set. This can be computed efficiently via matrix multiplication as follows: Let S_1 denote all the cells from time point t_1, and let

$$p(x) = \begin{cases} 1 & x \in C \\ 0 & \text{otherwise} \end{cases}$$

denote the uniform distribution on $C \subset S$. The descendant distribution of C is given by $\hat{\pi}_{t_1,t_2} p$. We compute ancestor distributions in a similar way, except instead of taking the sum we compute an average.

In particular, we define a function $p(x)$ as above, then normalize it to sum to 1 and then form the matrix-vector product

$$p^T \hat{\pi}_{t_0,t_1}$$

to obtain the ancestor distribution on time t_0.

After computing the trajectory to or from a cell set C (in the form of a sequence of ancestor and descendant distributions), we compute trends in expression for any gene or gene signature of interest along the trajectory. For each time point, we compute the mean expression weighting each cell according to the probability distribution defined by the ancestor or descendant distribution.

4. Learning gene regulatory models

We employ two strategies to summarize the transport maps by learning models of gene regulation. The first model uses local enrichment analysis to identify transcription factors (TFs) enriched in ancestors of a set of cells. The second model is motivated by the dynamical systems formulation of optimal transport, as described above in Section 4.3 of Chapter I.

4.1. Local model: TF enrichment analysis of top ancestors

We perform local enrichment analysis as follows. Given a set of cells C at time t_2, we first compute the ancestor distribution of C at an earlier time t_1, as described in Section 3 above. We then select cells contributing the most mass to the ancestor distribution, until a certain amount of mass is accounted for (e.g. 30% of the ancestor mass). We refer to these as the *top ancestors* at time t_1 of the cell set C. Finally, we compare the top ancestors to a null set of cells from the same time point. For example, this null cell set could be:

- all cells except for the top ancestors,
- the *bottom ancestors* (defined to be all cells except for the top ancestors of a less-strict cut-off),
- the bottom ancestors restricted to a specialized subset (e.g. all other trophoblasts when C is a specific subset of trophoblasts like spongiotrophoblasts).

4.2. Global model: learning a cell-autonomous gradient flow

To learn a simple description of the temporal flow, we assume that a cell's trajectory is cell-autonomous and, in fact, depends only on its own internal gene expression. We know this is wrong as it ignores paracrine signaling between cells, and we discuss models that include cell-cell communication below. However, this assumption is powerful because it exposes the time-dependence of the stochastic process $\mathbb{P}_t$ as arising from pushing an initial measure through a differential equation:

$$\dot{x} = f(x). \tag{10}$$

Here f is a vector field that prescribes the flow of a particle x. Our biological motivation for estimating such a function f is that it encodes information about the cell-autonomous regulatory networks that create the equations of motion in gene-expression space.

We propose to set up a regression to learn a regulatory function f that models the fate of a cell at time t_{i+1} as a function of its expression profile at time t_i. Our approach involves sampling pairs of points using the couplings from optimal transport:

- For each pair of time points t_i, t_{i+1}, we sample pairs of cells $\left(X_{t_i}, X_{t_{i+1}}\right)$ from the joint distribution specified by the transport map $\hat{\pi}_{t_i,t_{i+1}}$.
- Using the training data generated in the first step, we set up the following regression:

$$\min_{f \in \mathcal{F}} \quad \mathbb{E}_{\hat{\pi}_{t_i,t_{i+1}}} \left\| X_{t_{i+1}} - f(X_{t_i}) \right\|^2,$$

where $\mathcal{F}$ is a rectified-linear function class defined in terms of a specific generalized logistic function $\ell : \mathbb{R} \mapsto \mathbb{R}$:

$$\ell(x; k, b, y_0, x_0) = \frac{k y_0}{y_0 + (k - y_0) e^{-b(x - x_0)}},$$

where $k, b, y_0, x_0 \in \mathbb{R}$ are parameters of the generalized logistic function $\ell(x)$.

We define a function class $\mathcal{F}$ consisting of functions $f : \mathbb{R}^G \to \mathbb{R}^G$ of the form

$$f(x) = U \ell(WTx),$$

where ℓ is applied entry-wise to the vector $WTx \in \mathbb{R}^M$ to obtain a vector that we multiply against $U \in \mathbb{R}^{G \times M}$. Here $T \in \mathbb{R}^{G_{\text{TF}} \times G}$ denotes a projection operator that selects only the coordinates of x that are transcription factors, and G_{TF} is the number of transcription factors. Intuitively, this gives a set of low-rank, linear functions with sparse factors. Each rank-1 component can be interpreted as a regulatory module of transcription factors acting on a module of regulated genes.

We set up the following optimization over matrices $U \in \mathbb{R}^{G \times M}$ and $W \in \mathbb{R}^{M \times G_{\text{TF}}}$:

$$\begin{aligned} \min_{U,W} \quad & \mathbb{E}_r \left\| X_{t_{i+1}} - U \ell(WTX_{t_i}) \right\|^2 + \eta_1 \|U\|_1 + \eta_2 \|W\|_1, + \eta_3 \|W\|_2^2 \\ \text{s.t.} \quad & U \geq 0. \end{aligned} \tag{11}$$

where $(X_{t_i}, X_{t_{i+1}})$ is a pair of random variables distributed according to the normalized transport map r, and $\|U\|_1$ denotes the sparsity-promoting ℓ_1 norm of U, viewed as a vector (that is, the sum of the absolute value of the entries of U). Each rank one component (row of U or column of W) gives us a group of genes controlled by a set of transcription factors. The regularization parameters η_1 and η_2 control the sparsity level (i.e. number of genes in these groups).

Implementation: We designed a stochastic gradient descent algorithm to solve (11). Over a sequence of epochs, the algorithm samples batches of points $(X_{t_i}, X_{t_{i+1}})$ from the transport maps, computes the gradient of the loss, and updates the optimization variables U and W. The batch sizes are determined by the Shannon diversity of the transport maps: for each pair of consecutive time points, we compute the Shannon diversity S of the transport map, then randomly sample $\max(S \times 10^{-5}, 10)$ pairs of points to add to the batch. We run for a total of $10,000$ epochs.

Cell non-autonomous processes: The gradient flow (10) addresses cell-autonomous processes. Otherwise, the rate of change in expression $\dot{x}$ is not just a function of a cell's own expression vector $x(t)$, but also of other expression vectors from other cells. We can accommodate cell non-autonomous processes by allowing f to also depend on the full distribution $\mathbb{P}_t$:

$$\frac{dx}{dt} = f(x, \mathbb{P}_t). \tag{12}$$

Concretely, we could allow f to depend on the mean expression levels of specific genes (expressed by any cell) encoding, for example, secreted factors or direct protein measurements of the factors themselves. For a theoretical description of gradient flows with interactions, see Section 4.3 of Chapter I.

5. Geodesic interpolation for validation

Optimal transport provides an elegant way to interpolate distribution-valued data, analogous to how linear regression can be used to interpolate numerical or vector-valued data. Given two numerical datapoints, the simplest way to interpolate is to connect them with a line; this is the shortest path connecting the observed data. Given two distributions, we interpolate by finding the shortest path in the space of distributions. To do this we need a notion of distance between distributions, and for this we use the metric induced by optimal transport. This metric space is called Wasserstein space, and this form of interpolation is called geodesic interpolation (Villani, 2008).

We derive a modified version of geodesic interpolation that takes into account cell growth. Ordinarily, an interpolating distribution is computed by first computing a transport map between the distributions, and then connecting each point in the first distribution to points in the second according to the transport map. Finally, an interpolating point cloud is produced by from the midpoints of those line segments. (More generally, instead of taking just midpoints, one one can also construct a family of interpolations that sweep from the first distribution to the second). We extend this framework to accommodate growth by changing the mass of the point we place at the midpoint (to account for the fact that cells will have a different number of descendants at time t_1 than they will at time t_2).

Specifically, to interpolate at time $s \in (t_1, t_2)$, we first renormalize the rows of the transport map so they sum to roughly $\frac{\hat{g}(x)^{s-t_1}}{\int \hat{g}(x)^{s-t_1} d\hat{\mathbb{P}}_{t_1}}$ instead of $\frac{\hat{g}(x)^{t_2-t_1}}{\int \hat{g}(x)^{t_2-t_1} d\hat{\mathbb{P}}_{t_1}(x)}$. This takes into account the descendant mass each cell will have by time s instead of by time t_2. We then sample points $z_1, \ldots, z_N$ as follows:

1. Sample a pair of points (x, y) from the joint distribution specified by the transport map.
2. Identify the point
$$z = \alpha x + (1 - \alpha) y$$
along the line segment connecting x and y. Here α is given by $s = \alpha t_1 + (1 - \alpha) t_2$.

By repeating the steps above, we accumulate a point-cloud of points $z_1, \ldots, z_N$. Finally, we define the interpolating distribution as

$$\hat{\mathbb{P}}(s) = \frac{1}{N} \sum_{i=1}^{N} \delta_{z_i}.$$

Equipped with this notion of interpolation, we can test the performance of optimal transport by comparing the interpolated distribution to held-out time points. Using the data from time t_i and t_{i+2}, we interpolate to estimate the distribution $\mathbb{P}_{t_{i+1}}$. We then compute the Wasserstein distance between the interpolated distribution and the observed distribution. We compare this distance to a null model generated from the independent coupling where we sample pairs (x, y) independently $x \sim \hat{\mathbb{P}}_{t_i}$ and $y \sim \hat{\mathbb{P}}_{t_{i+2}}$ in step 1 above. We also compare the interpolated distance to distance between batches of $\mathbb{P}_{t_{i+1}}$. Optimal transport is performing well if the interpolated point cloud is as close to the batches of the held out time point as the batches are to each other, and the null-interpolated point cloud is farther away.

We present our application for validation in the case of IPS reprogramming in the STAR Methods (**Validation by geodesic interpolation**).