Analyzing Biological Models with the Wasserstein Distance

Shane Lubold
shane.lubold@asu.edu
Arizona State University

Sebastien Motsch
smotsch@asu.edu
Arizona State University

May 24, 2016

Abstract

In recent years a number of mathematical models have been proposed that describe biological systems. One needs a way to compare the accuracy of these methods, in relation to each other and in relation to empirical data. The Wasserstein distance is a new metric that can be used for such a purpose. In one dimension, there exists a closed form expression for the Wasserstein distance, but in two or more dimensions, no such expression exists for most functions. In this paper we present a method for approximating the Wasserstein distance in two or more dimensions. We also consider several biological models and demonstrate how the Wasserstein distance can be used to analyze the effictiveness of such models.

1 Introduction

Modeling biological systems has increased in popularity in recent years. From bacteria to swarm bees, a number of models have been proposed to describe each system. But the increase in number of models does not necessary improve the comprehension of the underlying population. One needs to be able to compare there predictions and more important to measure their agreement with experimental data. One way to do this is to use the Wassertstein distance, which allows us to compare discrete and continuous dynamics (i.e.e, systems of ODEs and PDEs). But so far Wasserstein distance has been used mainly as a powerful tool to prove analytic results (i.e. existence/uniqueness of solutions). In this paper we use the Wasserstein distance to compare the accuracy of biological models in relation to empirical data.

The challenge is that the Wasserstein distance, defined as a minimization problem, can only be computed explicity in one dimension. In two or higher dimensions, we must rely on an approximation. In this paper, we develop a numerical method that approximates the Wasserstein distance through the Simplex method for continuous distributions, and the Hungarian algorithm for discrete distributions (sets of points). Once the numerical methods have been implemented and tested, we will use the Wasserstein distance to perform data-model comparison for several dynamics (e.g. pedestrian dynamics, swarming model). In many studies, data-model comparisons are limited to only visual or qualitative agreement. This framework aims at providing quantitative measurement for the agreement (or discrepancy) between a model and experimental data.

The rest of the paper is organized as follows. In Section 3, we present the general problem, outline the theory and notation, and provide some theoretical results regarding the Wasserstein Distance. In Section 4, we provide the theory and computations of the Wasserstein distance in one dimension. We also consider the case of noisy data and show how this affects the computation of the Wasserstein distance. Section 5 outlines the problem in two dimensions, and shows how the simplex method may be used to compute the Wasserstein distance between continuous and discrete distributions. In Section 6 we examine how models used to study biological systems compare to empirical data through the use of the Wasserstein distance. Section 7 provides concluding remarks and points to future work.

2 Literature Review

The optimal transport problem, the problem of moving a mass from one location to another in the most efficient way possible, has been studied for several centuries. One way to phrase the problem is this. Consider a number n of bakeries, each of which makes a certain amount of bread, and we know that there are m restaurants that will consume a fixed amount of bread. Since the amount of bread produced and consumed is fixed for each bakery and restaurant, we can model these quantities as probability measures and create a "density of production" and a "density of consumption" on some space, which we can consider to be a metric space with the distance between a bakery and a restaurant as the length of the straight line between them. The optimal transport problem then seeks to move the bread from the bakeries to the restaurants by minimizing the total cost of transportation. For a more complete description of this problem, see Section 3 of [1].

Recent work has shown a number of theoretical properties of the Wasserstein distance, including the existence of solutions to the optimal transport problem under certain assumptions, [1], the convexity of the solution space, [2], and the weak convergence of measures, [2].

In recent years there has been a proliferation of biological models concerning a vast array of biological systems.

3 Preliminaries

Given a Polish (complete, separable, metric) space (M, d), define $\mathcal{P}(M)$ as the set of probability measures on $(M, \mathcal{B}(M))$. If X, Y are two Polish spaces, T is a Borel map, and $\phi \in \mathcal{P}(X)$ is a measure, we can define $T_{\#}\phi \in \mathcal{P}(Y)$, which we call the *pushforward of ϕ through T*, as

$$T_{\#}\phi(A) = \phi(T^{-1}(A)), \qquad \forall A \subseteq Y,\ A \text{ is Borel.} \tag{1}$$

The pushforward of ϕ through T has the property that

$$\int_Y h\ dT_{\#}u = \int_X h \circ T d\mu, \tag{2}$$

for every Borel function $h : Y \to \mathbb{R} \cup \{\pm\infty\}$. Define a Borel cost function $c : X \times Y \to \mathbb{R} \cup \{\infty\}$. Following the notation of [2], we call the following problem formulation the *Monge optimal transport problem.*

Monge Optimal Transport Problem: Let $v \in \mathcal{P}(X)$, $\mu \in \mathcal{P}(Y)$. We define the jth Wasserstein distance between v and μ as

$$W_j(\mu, \nu) = \inf_T \sqrt[j]{\int_M c(x, T(x))^j\ d\mu(x)}, \tag{3}$$

where the inf is taken over all T satisfying (2). As noted in [2], Monge's definition can be ill-posed if no such admissible T exists (if μ is a Dirac delta and ν is not) or if the constraint that $T_{\#}\mu = \nu$ is not weakly sequentially closed, w.r.t any reasonable weak topology. To resolve these issues, we also present the formulation attributed to Kantorovich, [2].

Kantorovich Optimal Transport Problem: For each $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, let $\Pi(f, g)$ denote the collection of couplings between μ and ν. In other words,

$$\Pi(f, g) = \Big\{\pi : \pi(A \times M) = \mu(A),\ \text{and } \pi(A \times M) = \nu(A),\ \text{for each } A \in \mathcal{B}(M)\Big\}. \tag{4}$$

We then define the j-th Wasserstein distance between μ and ν as

$$W_j(f, g) = \inf_{\pi \in \Pi(f,g)} \sqrt[j]{\int_{M \times M} d(x, y)^j d\pi(x, y)}, \tag{5}$$

where the inf is taken over all $\pi \in \Pi(f, g)$ that satisfy (4). In some cases, $W_p(\mu, \nu) = \infty$, so we will restrict ourselves $W_j(f, g)$ to $\mathcal{P}_j(M) \subset \mathcal{P}(M)$, where $\mathcal{P}_j(M)$ denotes the set of probability measures on M with finite j-th moment. In other words,

$$\mathcal{P}_j(M) = \Big\{ f \in \mathcal{P}(M) : \int_M d(x, y)^j f(dy) < \infty, \text{ for some } x \in M \Big\}. \tag{6}$$

As shown in [2], $\Pi(f, g)$ is a convex set that contains $f \times g$. If c is lower semicontinuous and bounded below, then there exists a $\pi \in \Pi(f, g)$ that satisfies (5). [2] also proves that if c is continuous and μ is non atomic, then the values of (3) and (5) are equal. It is also possible to show that W_j defines a distance function on $\mathcal{P}_j(M)$. Clearly $W_j(\mu, \nu) = W_j(\nu, \mu)$ and $W_j(\nu, \mu) \geq 0$, with equality only holding in the case that $\mu = \nu$. To prove that $W_j(\mu, \nu) \leq W_j(\nu, \psi) + W_j(\psi, \nu)$ for some measure ψ, it's convenient to use the so-called gluing lemma, [1].

Gluing Lemma: Let $(X_i, \mu_i), i = 1, 2, 3$, be Polish probability spaces. If (X_1, X_2) is a coupling of (μ_1, μ_2) and (Y_2, Y_3) is a coupling of (μ_2, μ_3), then it is possible to construct a triple of random variables (Z_1, Z_2, Z_3) such that (Z_1, Z_2) has the same law as (X_1, X_2) and (Z_2, Z_3) has the same law as (Y_2, Y_3).

For the remainder of the paper, we restict ourselves to $(M, d) = (\mathbb{R}^n, d)$, with $n \geq 1$ and d being the standard metric on $\mathbb{R}^n$.

4 One Dimensional Case

In one dimension, there exists a closed form solution to computing the Wasserstein distance between two probability density functions. Let $f, g : \mathbb{R} \to \mathbb{R}$ such that

$$\int_{\mathbb{R}} f \, dx = \int_{\mathbb{R}} g \, dx. \tag{7}$$

If we define

$$F(x) = \int_{-\infty}^{x} f(s) \, ds \quad \text{and} \quad G(x) = \int_{-\infty}^{x} g(s) \, ds, \tag{8}$$

then the 2nd Wasserstein distance between f and g is

$$W_2^2(f, g) = \int_0^1 |F^{-1}(y) - G^{-1}(y)|^2 \, dy. \tag{9}$$

From now on, we will refer to the value of (9) as the Wasserstein distance between f and g, or simply the WD between f and g. One can think of (9) as moving the j-th percentile of f to the j-th percentile of g (Figure 1).

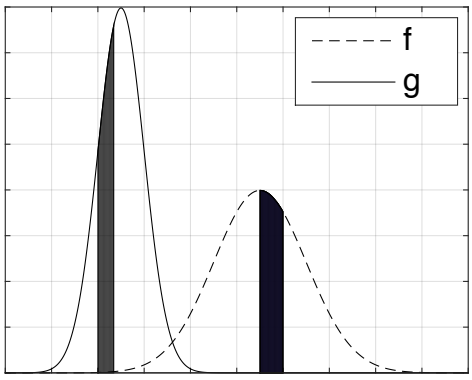
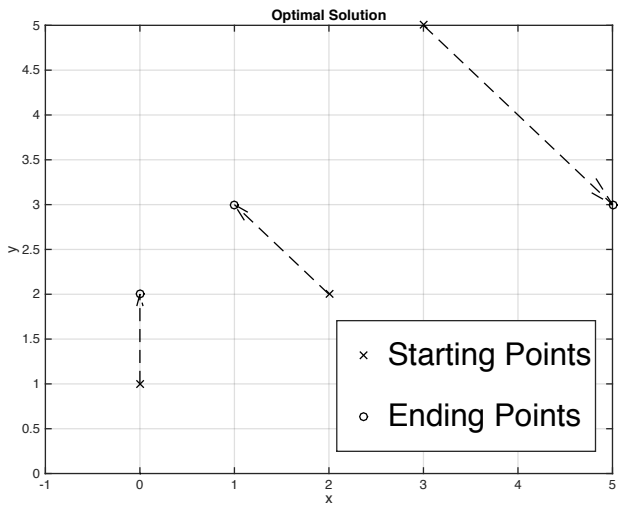
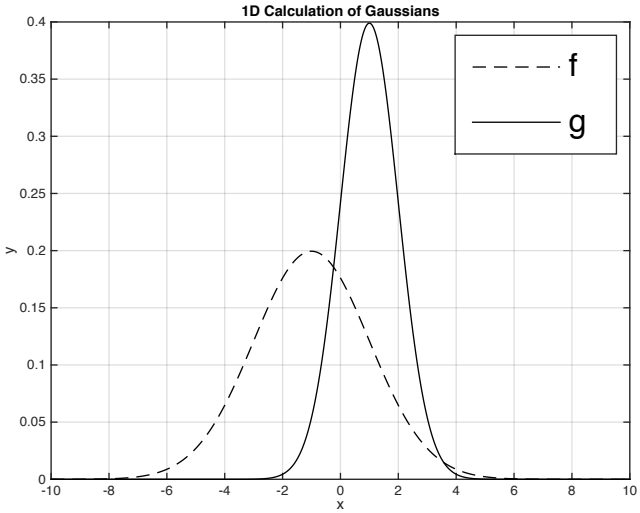


Figure 1: Moving j-th percentile (shaded) of f to j-th percentile of g.

The WD is used to quantify the distance between discrete and continuous functions, as shown in Figure (3). In Figure 2a the WD between three points in computed, and in Figure 2b the WD between two Gaussians is computed.



(a) Discrete Points, WD ≈ 1.7475



(b) Continuous Functions, WD ≈ 2.236

Figure 2: Computations of W_2 in one dimension

5 Two Dimensional Case

In two or more dimensions, there is no closed form expression for the Wasserstein distance. One prominent exception is in the case of two Gaussians. If

$$x \sim \mathcal{N}(\mu_x, \Sigma_x) \text{ and } y \sim \mathcal{N}(\mu_y, \Sigma_y), \tag{10}$$

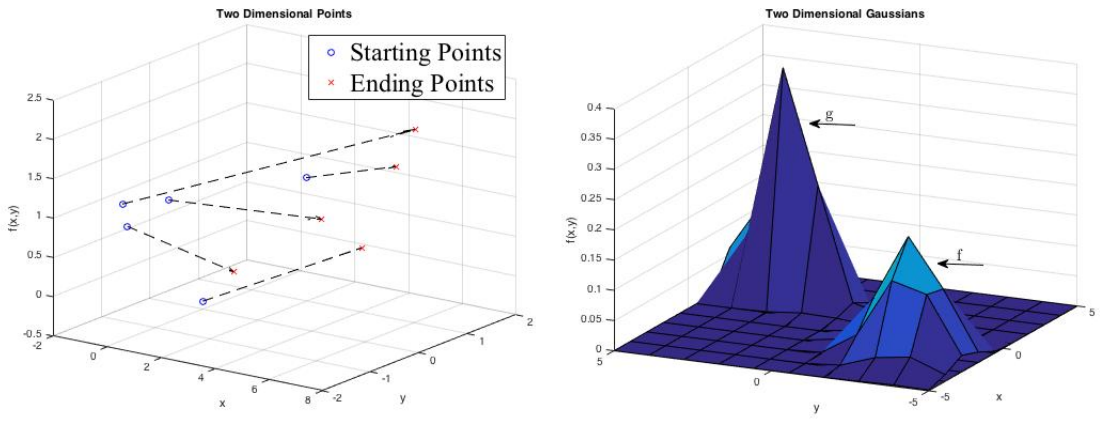
and f, g denote the PDFs of x, y respectively, then [3] shows

$$W_2(f, g) = |\mu_x - \mu_y| + \text{tr}(\Sigma_x + \Sigma_y - 2\sqrt{\Sigma_x \Sigma_y}). \tag{11}$$

In this section a method to approximate the Wasserstein distance between discrete or continuous distributions in $\mathbb{R}^2$ is presented. Consider a physical domain with points $x_i, 1 \leq i \leq n \times m$, where n, m are the number of points in the x, y dimensions respectively. Form the matrix X such that $(X)_{j,k}$ is the amount transported from x_j under f to x_k under g. Form the matrix $(C)_{j,k} = |x_j - x_k|$ as the distance between points x_j and x_k. With this framework, computing the WD is then equivalent to solving

$$\begin{aligned} \text{minimize} \quad & \sum_{ij} X_{ij} C_{ij} \\ \text{subject to} \quad & X\mathbf{1} = (a_1, \ldots, a_m)^T \\ & \mathbf{1}^T X = (b_1, \ldots, b_n) \\ & (X)_{ij} \geq 0, \end{aligned} \tag{12}$$

where $\mathbf{1}$ indicates a vector of ones of appropriate length. Because of the linearity of the objective function, (12) can be solved with the simplex method. In Figures 3a and 3b, the WD between discreet functions and continuous functions, respectively, in two dimensions is computed.



(a) Discrete Points, WD ≈ 5.69

(b) Continuous Functions, WD ≈ 7.12

Figure 3: Computations of WD in two dimensions

Remark: Although the continuous examples given were normal distributions, the WD can be computed between any two continuous functions.

6 Application to Biological Systems

This section applies the tools from previous sections to quantify the differences between biological models that use particles to study larger phenomenon ("micro" models), such as the growth of cancer, and those that study such phenomenon on a macroscopic scope. This question is of critical importance to researchers, since the micro-level models are computationally cheaper and thus providing measures of their accuracy can be very useful.

6.1 One Dimensional Case

For illustration, we begin by examining the growth of cancer cells. To model such a phenomenon on a macroscopic level, we study the model

$$\frac{\partial \rho}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2 \rho}{\partial^2 x} + \rho(1-\rho), \tag{13}$$

where σ is a scalar and $\rho(x,t)$ denotes the density of cancer cell at time t and a location x. For the micro-model, we consider the equation

$$dX_t = \sigma dB_t, \tag{14}$$

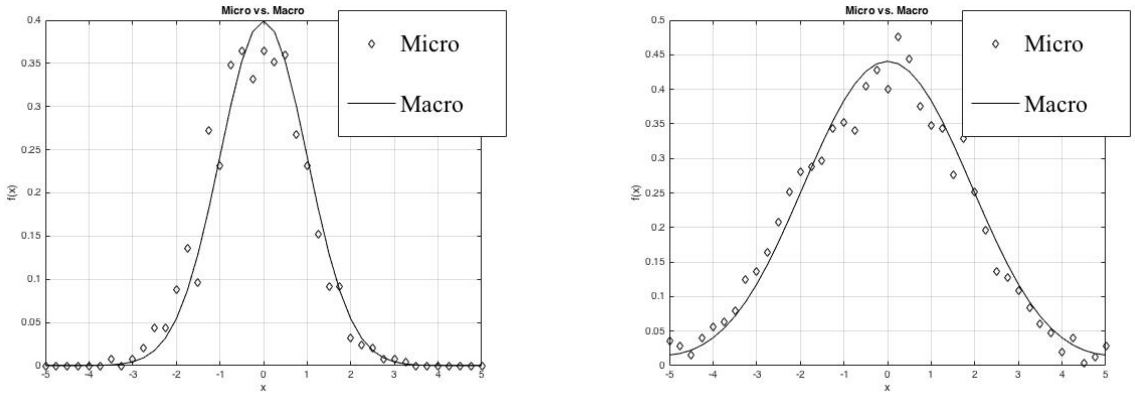
where B_t denotes a standard Brownian process, with $B(0) = 0$. For $\Delta t > 0$, we have

$$\int_0^{\Delta t} dX_t = \sigma \int_0^{\Delta t} dB_t \implies x_{\Delta t} - x_0 = \sigma \mathbf{Z} \tag{15}$$

where $\mathbf{Z} \sim \mathcal{N}(0, \Delta t)$. Using properties of variance,

$$x_{\Delta t} = x_0 + \sigma\sqrt{\Delta t}\mathbf{W} \tag{16}$$

where $\mathbf{W} \sim \mathcal{N}(0,1)$. To compute $Micro_\rho$, $N \in \mathbb{Z}^+$ i.i.d. realizations of (16) are computed, from which a histogram ($Micro_\rho$) is created over the physical domain. In Figures 4a and 4b we plot $Macro_\rho, Micro_\rho$ at the initial and final times, respectively.



(a) Solutions at initial time

(b) Solutions at final time

Figure 4: $N = 10^3,\ T_{final} = 1,\ \Delta t = .01.$

Using a standard normal PDF as the initial condition for $Macro_\rho$, and $\Delta x = .05$, we see in Figures 5a and 5b the decay of WD and log(WD) against number of particles, respectively. Figure 5c shows that WD decays linearly a Δx decreases. All WD reported are at $T_{final} = 1,\ \Delta t = .01$.

We conclude from Figure 5a that the error between Macro_ρ and Micro_ρ follows the relationship

$$\Big|W_2\Big(Macro_\rho, Micro_\rho\Big)\Big| \approx C\Big(\frac{1}{N}\Big)^\alpha,$$

where $C \approx 3.33,\ \alpha \approx 0.44$. When computing the fit $log(WD) = a \times log(N) + b$, we find $a \approx -.41, b \approx .53$. We thus conjecture that the Wasserstein distance between Micro_ρ and Macro_ρ decays at a rate of $1/\sqrt{N}$.

Theorem 1. $\Big|W_2\Big(Macro_\rho, Micro_\rho\Big)\Big| \approx C\sqrt{\frac{1}{N}}$, *for some* $C \in \mathbb{R}$.

Proof. Denote Macro_ρ by ρ_{mi}, Micro_ρ by ρ_{ma}. Then, for some $\phi \in C_b(\mathbb{R})$

$$\langle \rho_{mi}, \phi \rangle = \int_{\mathbb{R}} \rho_{mi}\phi(x)\ dx = \frac{1}{N}\sum_x \phi(x_i(t)).$$

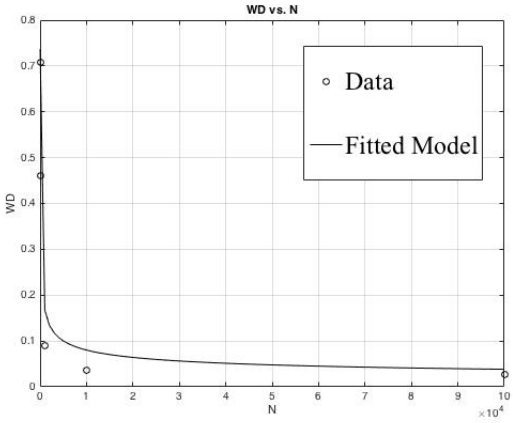
By the Law of Large Numbers,

$$\frac{1}{N}\sum_x \phi(x_i(t)) \xrightarrow{N\to\infty} \int_{\mathbb{R}} \phi(x)\rho_{mi}(x)\ dx + \frac{1}{\sqrt{N}}\text{std}(\phi(x))$$

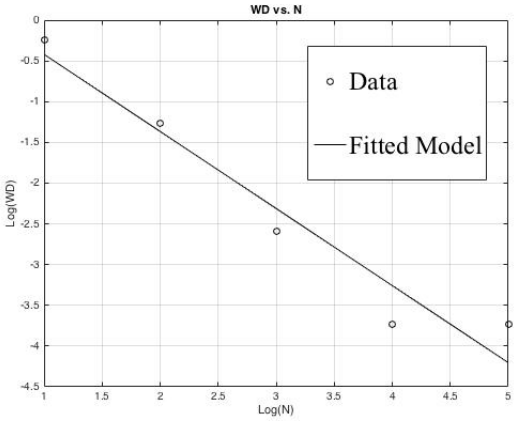
where $\text{std}(\phi(x))$ denotes the standard deviation of ϕ. □

Figures 6a and 6b plot the respective velocities of $Micro_\rho, Macro_\rho$. To find the velocity between functions f_1, f_2 at time t_j, we set a value of $c = .1$, then find $x_{j+1} = f_2^{-1}(c), x_j = f_1^{-1}(c)$. Then, the velocity at time j is $v(j) = \frac{x_{j+1} - x_j}{\Delta t}$. Clearly, the velocity of $Micro_\rho \approx 1.38 \times 10^2$ is a random variable, while the velocity of $Macro_\rho \approx 1.5 \times 10^2$ is deterministic.

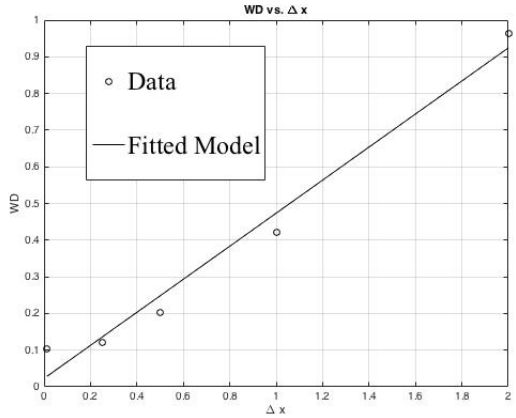
We now investigate how the ratio $\frac{N}{\Delta x}$ impacts the WD. We consider the cases when $(N,\ \Delta x) = (10, 1),\ (10^2, .5),\ (10^3, .125),\ (10^4, .0625)$. For each case, 1,000 i.i.id realizations of (14) are run. The WD between $Macro_\rho$ and $Micro_\rho$ at $T_{final=1},\ \Delta t = .01$ is computed, and then the max of these WD values is computed. We plot the corresponding max WD and pairings of $(N,\ \Delta x)$ in Figures 7a and 7b.



(a) Fitting WD to N

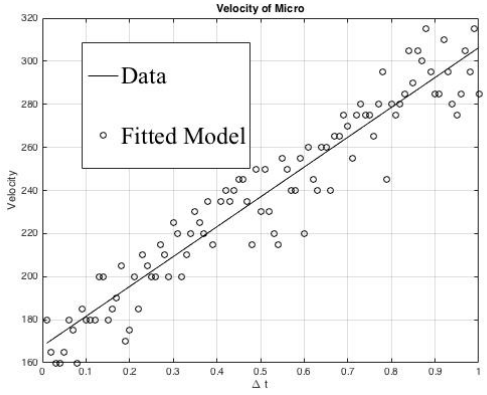


(b) Fitting Log(WD) to Log(N)

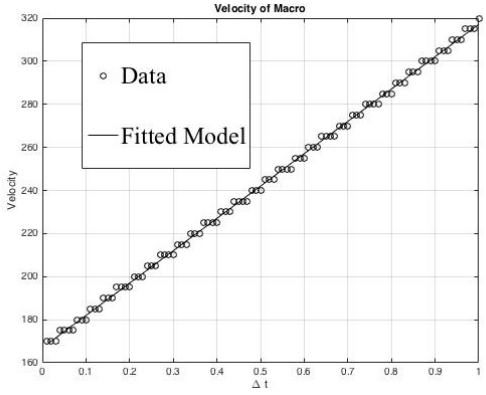


(c) Fitting WD to Δx

Figure 5: One Dimensional WD Computations

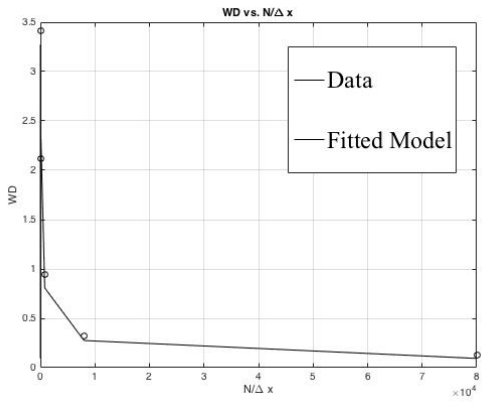


(a) Fitting Micro$_\rho$ velocity

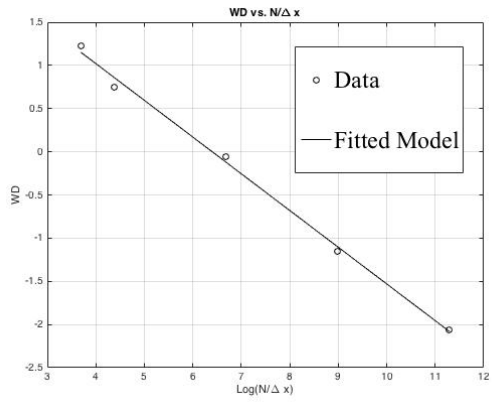


(b) Fitting Macro$_\rho$ velocity

Figure 6: Velocities, $T_{final} = 1, \Delta t = .01, N = 10^4, \sigma = \sqrt{2}$.



(a) Fitting WD to $\frac{N}{\Delta x}$



(b) Fitting Log(WD) to $Log(\frac{N}{\Delta x})$

Figure 7: One Dimensional WD Computations

6.2 Two Dimensional Case

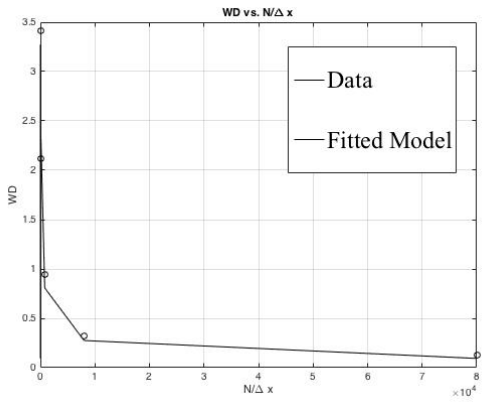
In two dimensions, the equation for the macro model becomes

$$\frac{\partial \rho}{\partial t} = D\Big(\frac{\partial^2 \rho}{\partial^2 x} + \frac{\partial^2 \rho}{\partial^2 y}\Big) + \rho(1-\rho), \tag{17}$$

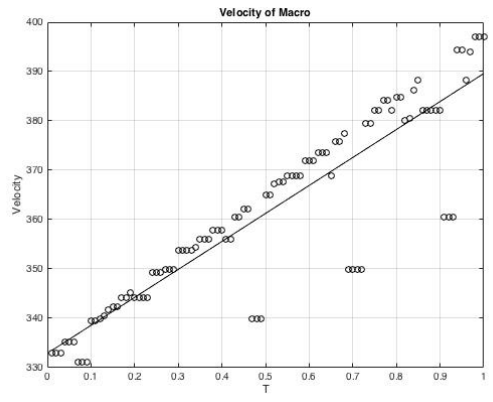
where $\rho(t, x, y)$ denotes the density of cancer cell at time t at location (x, y). Similarly, (16) becomes

$$dX_t = \sigma dB_t, \tag{18}$$

where B_t denotes a standard Brownian motion in 2 dimensions, with $B(0) = (0, 0)$. Figures 8a and 8b plot the respective velocities of $Micro_\rho, Macro_\rho$. Clearly, the velocity of $Micro_\rho \approx$ is a random variable, while the velocity of $Macro_\rho \approx 1.5 \times 10^2$ is deterministic.



(a) Fitting WD to $\frac{N}{\Delta x}$



(b) Fitting Log(WD) to $Log(\frac{N}{\Delta x})$

Figure 8: 2D Velocities, $T_{final} = 1, \Delta t = .01, N = 10^3, \sigma = \sqrt{2}$.

7 Conclusion

8 Definition: More for My Use

- Let X be a set. An algebra on X is a collection $\Sigma \subseteq 2^X$ that is closed under finite set operations (complement, union, and intersection). A $\sigma-$algebra on X is a collection $\Sigma \subseteq 2^X$ that is closed under countably many such operations. The intersection of a collection of $\sigma-$algebras is a $\sigma-$algebra.
- A **Borel set** in a topological space $(X, \mathscr{T})$ is a set that can be made from open sets $U \in \mathscr{T}$ by countable union, countable intersection, or relative complement. The collection of all Borel sets on X forms a $\sigma-$algebra, known as the Borel $\sigma-$algebra, and is denoted by $\mathcal{B}(X)$.
- Let X, Y be two topological spaces. A function $f : X \to Y$ is called a Borel map if 1) $f^{-1}(A)$, A open, is a Borel subset, 2) $f^{-1}(B)$, B closed, is a Borel subset, and 3) $f^{-1}(C)$, C Borel, is a Borel subset.
- Given a measurable space (X, Σ) and a measure μ on that space, a set $A \subset X$ in Σ is called an **atom** if $\mu(A) > 0$ and for any measurable subset $B \subset A$ with $\mu(B) < \mu(A)$, the set B has measure zero. A measure that has no atoms is called **non-atomic**.

Informal sources:

- (1) http://www.math.umd.edu/ yanir/OT/AmbrosioGigliDec2011.pdf
- (2) http://www.math.tohoku.ac.jp/ aida/workshop/H23-Daisympo/kuwada.pdf

References

[1] Villani, C., Optimal transport, old and new, June 13, 2008, http://cedricvillani.org/wp-content/uploads/2012/08/preprint-1.pdf

[2] Ambrosio, L., Gigli, N. A user's guide to optimal transport. `http://www.math.umd.edu/~yanir/OT/AmbrosioGigliDec2011.pdf`.

[3] Dowson, D.C., B.V Landau, The Frechet Distance between Multivariate Normal Distributions, J. Multivariate Analysis 12, 450-455 (1982).

[4] Prokhorov, Yu. V., Convergence of random processes and limit theorems in probability theory. (Russian) Teor. Veroyatnost. i Primenen. 1 (1956), 177-238.