Population Dynamics for Discrete Wasserstein Gradient Flows over Networks

Gilberto Diaz-Garcia [1] **César A. Uribe** [2] **Nicanor Quijano** [1]

Abstract

We study the problem of minimizing a convex function over probability measures supported in a graph. We build upon the recent formulation of optimal transport over discrete domains to propose a method that generates a sequence that provably converges to a minimum of the objective function and smoothly transports mass over the edges of the graph. Moreover, we identify novel relation between Riemannian gradient flows and perturbed best response protocols that provide sufficient conditions for the convergence of the proposed algorithm. Numerical results show practical advantages over existing approaches with respect to the implementability and convergence rates.

1. Introduction

Optimal transport theory provides a mathematical formulation to describe how a probability distribution could be efficiently transported into another (Villani, 2008). This theoretical framework gives us geometrical tools to define distances over probability spaces, extending concepts from classical Euclidean geometries (Peyré et al., 2019). Such flexibility, alongside the geometric and statistical properties of probability distributions has led to a large number of applications, e.g., image morphing and image interpolation of natural images (Simon & Aberdam, 2020), averaging atmospheric gas concentration data (Barré et al., 2020), fairness in machine learning (Chzhen et al., 2020), Bayesian learning (Backhoff-Veraguas et al., 2018), among others.

One of the most remarkable properties of optimal transport is that it endows the spaces of distributions with a metric known as Wasserstein metrics (Peyré et al., 2019). This induced geometry allows to apply concepts that applies to metric spaces to probability distributions $\mathcal{M}(\mathcal{X})$. In particular, one well-studied applications is that gradient flows in Wasserstein metric, i.e., $\partial_t \rho = -\operatorname{grad} \mathcal{H}(\rho)$ for $\mathcal{H}(\rho) : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$, can be used to model partial differential equations (PDEs) (Otto, 2001). This significantly facilitates the study of PDEs by analyzing the behavior of the corresponding energy functions.

Although the optimal transport theory results are compelling, with generalizations also to infinite-dimensional spaces (Ambrosio et al., 2008), there is an increasing interest in discrete time optimal transport algorithms over discrete measures spaces and domains (Erbar et al., 2020; Lavenant et al., 2018). Modern data-driven machine learning problems have intrinsic communication constraints and structured domains, then the notions of discrete measures on graphs naturally emerge (Bécigneul et al., 2020; Dong & Sawin, 2020; Essid & Solomon, 2018). In spite of its flexibility, optimal transport formulations require significant computational efforts (Cuturi, 2013), particularly in discrete structure domains like graphs (Erbar et al., 2020; Lavenant et al., 2018).

In this paper, we focus on the optimization problem:

$$\min_{\rho \in \Delta^+} \mathcal{F}(\rho), \tag{1}$$

where Δ^+ denotes the interior of a probability simplex defined on the nodes $\mathcal{V} = [1, \ldots, n]$ of a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, and $\mathcal{F} : \Delta^+ \to \mathbb{R}$ is a convex function. Moreover, we study the case where we endow the domain Δ^+ with a structured discrete Wasserstein metric on the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. We assume a minimizer exists and denote it as ρ^*.

When focusing on discrete distributions over graphs, the advantage of optimal transport formulation does not rely on the analysis of higher dimensional PDEs. Instead, the induced metric contains information about the probability mass exchange between nodes. Therefore, the gradient of $\mathcal{F}(\rho)$ flows in the Wasserstein metric and generates curves that not only minimize $\mathcal{F}(\rho)$, but *smoothly transports the mass over the edges of the graph.*

Among the wide variety of algorithms for optimization over probability measures of the form (1), the most broadly discussed are those derived from Jordan-Kinderlehrer-Otto

[1] Department of Electronics Engineering at the University of Los Andes, Colombia. [2] Department of Electrical and Computer Engineering, Rice University, TX, USA.. Correspondence to: Gilberto Diaz-Garcia <gj.diaz10@uniandes.edu.co>, César A. Uribe <cauribe@rice.edu>, Nicanor Quijano <nquijano@uniandes.edu.co>.

Proceedings of the LatinX in AI (LXAI) Research workshop at ICML 2021. Copyright 2021 by the author(s).

(JKO) schemes (Jordan et al., 1998). The idea behind them is to minimize $\mathcal{F}(\rho)$ and impose the smoothness of the solution via regularization with Wasserstein distance $\mathcal{W}(\cdot,\cdot)$ (Peyré, 2015). While such algorithms have successfully optimize functions over a broad range of networks, there is immense computational cost.

Therefore, there is a need to propose simpler algorithms to implement. Furthermore, since current applications demand higher network complexity, making the optimization algorithm work in a distributed manner is desirable. We take advantage of the structure of $\operatorname{grad}\mathcal{F}(\rho)$ to develop an algorithm that generate a sequence $\{\rho_k\}_k$ that minimizes $\mathcal{F}(\rho)$ for the interaction constraints imposed by a graph. Using game theory concepts, we relate gradient flows with a special class of better response dynamics, giving us new tools to define minimizing discrete dynamics.

The main contribution of this paper can be summarized as follows. First, we formally establish an equivalence between Wasserstein gradient flows over discrete measures and population game dynamics. Second, we propose a discrete algorithm for Wasserstein gradient flows to minimize functions on the space of discrete functions supported on a graph. Third, we determine sufficient conditions for convergence of our proposed algorithm. Finally, we present numerical experiments that support our theoretical results and compare them with other proposed strategies.

2. Optimal Transport & Wasserstein Distance in Graphs

Optimal transport framework describes optimal transformations between arbitrary measures. Given two measures $\mu \in \mathcal{M}(X)$ and $\sigma \in \mathcal{M}(Y)$, and a transport cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ the optimal transport formulation searches the optimal coupling between μ and σ with respect to c. In the specific case when $\mathcal{X} = \mathcal{Y}$ and the transport cost is expressed as $c(x,y) = d_{\mathcal{X}}(x,y)^p$, with $d_{\mathcal{X}}$ a distance in $\mathcal{X}$, the optimal transport problem is equivalent to a minimal-length path problem (Benamou & Brenier, 2000). This provides an extension to metrics over measures spaces called $p-$Wasserstein distances. In (Chow et al., 2017), the authors tackle this geometrical extension for distributions that are supported by a weighted graphs proposing that the optimal transport formulation over graphs can be written as,

$$\inf_{\Phi(t)} \int_0^1 \Phi(t)^\top L(\rho(t))\Phi(t)dt \tag{2}$$
$$\text{s.t } \partial_t\rho(t) = L(\rho(t))\Phi(t),\ \rho(0) = \mu \text{ and } \rho(1) = \sigma,$$

where $\Phi(t)$ is a vector valued function and $L(\rho(t))$ is a mass dependent Laplacian matrix. Thus, if we define an inner product over the tangent space of Δ^+, $\mathcal{T}_\rho\Delta^+$ as $g_\rho(\mu,\sigma) = \mu^\top L(\rho)^+\sigma$ where A^+ is the pseudo-inverse of A, the formulation in Equation (2) can be restated as the definition of distance for the Riemannian manifold $(\mathcal{M}(\mathcal{G}), g_\rho)$. Therefore, we have access to the optimal transport induced geometry on Δ^+ via $g_\rho(\cdot,\cdot)$. With the metric tensor we are able to define the gradient of a function $\mathcal{F}(\rho)$ under this geometry.

Proposition 1. *The gradient of a function $\mathcal{F}(\rho) : \Delta^+ \to \mathbb{R}$ is* $\operatorname{grad} F(\rho) = L(\rho)\nabla\mathcal{F}(\rho)$.

The gradient described in Proposition 1 not only grants us the instant steepest direction for a function $\mathcal{F}(\rho)$. It also contains information about the graph topology. Thus, this gradient will help define algorithms to optimize functions $\mathcal{F}(\rho)$ over probability distributions.

3. Discrete Wasserstein Gradient Flows and Population Dynamics

Once defined the notion of gradient under the geometry induced by the Wasserstein distance, we can design algorithms to generate sequences $\{\rho^{(k)}\}_k$ that optimize functions over probability distributions. One common approach is to make a generalization of steepest descend algorithms (Udriste, 2013), where a sampled version of the descend direction $-\operatorname{grad}\mathcal{F}(\rho^{(k)}) \in \mathcal{T}_{\rho^{(k)}}\Delta^+$ moves towards ρ^* using a fixed step ϵ. Nevertheless, the convergence of this family of algorithms, e.g. Riemannian Gradient Descent (Bonnabel, 2013), entirely depends on the choice of ϵ. Consequently, we want to develop a more accessible way to define a proper value for ϵ.

In (Mertikopoulos & Sandholm, 2018) the authors generalize population dynamics using the notion of gain of motion $G(z;\rho)$ and cost of motion $C(z;\rho)$. If we want to choose a direction $z \in \mathcal{T}_\rho\Delta^+$ they propose to find it by maximizing the profit of moving in such direction. An interesting characteristic the resulting dynamics is that they are closely related to the Hopkins dynamics (Hopkins, 1999). Using this relationship and properties of graph Laplacians it is possible to formulate gradient flows as a population playing a potential game under the Riemannian dynamics.

Lemma 1. *The Riemannian gradient flow $\dot{\rho} = L(\rho)\nabla\mathcal{F}(\rho)$ is equivalent to the Riemannian dynamics with $G(z;\rho) = z^\top\nabla\mathcal{F}(\rho)$ and $C(z;\rho) = z^\top g(\rho)z$, where $g^{-1}(\rho) = L(\rho) + \frac{1}{n}\mathbf{1}\mathbf{1}^\top$.*

With this in mind, let us define the direction $z^{(k)} = \alpha - \rho^{(k)}$. This means that instead of choosing a direction $z \in \mathcal{T}_{\rho^{(k)}}\Delta^+$ we choose a distribution $\alpha \in \Delta^+$ to move towards. Thus, we can formulate the iterative protocol,

$$\rho^{(k+1)} = \arg\max_{\alpha\in\Delta^+} \{\alpha^\top\nabla\mathcal{F}(\rho^{(k)}) - \frac{1}{\gamma}(\alpha - \rho^{(k)})^\top g(\rho^{(k)})(\alpha - \rho^{(k)})\}. \tag{3}$$

Equation (3) describes a family of target protocols called *Perturbed Best Response* protocols (Hopkins, 1999). In them, the players of a game choose the mixed strategy $\alpha \in \Delta^+$ that maximizes their profit when a noisy payoff is perceived. Thus, we can relate our optimization problem with them to grant us several tools to analyze our discrete protocol for Wasserstein gradient flows shown in (3).

Theorem 1. *Problem* (3) *has a closed form solution:*

$$\rho^{(k+1)} = \rho^{(k)} + \frac{\gamma}{2} L(\rho^{(k)}) \nabla \mathcal{F}(\rho^{(k)}), \tag{4}$$

and describes a perturbed best response protocol.

Theorem 1 gives us an insight of the behavior of the sequence $\{\rho^{(k)}\}_k$ generated by Equation (3). Since perturbed best response dynamics are an approximation of best response dynamics (Sandholm, 2015), the sequence will pursue the Nash equilibrium of a noisy version of the game proposed in Lemma 1. Given that the fitness is generated by the gradient of the function $\mathcal{F}(\rho)$, such game describes a potential game (Monderer & Shapley, 1996). A property of these games is that their Nash equilibrium matches the optimizer of the potential function. To ensure convergence of our proposed dynamics, note that perturbed best response is a particular case of a sub-gradient algorithm with nonlinear projection (Beck & Teboulle, 2003). We adapt a particular result of them in Proposition 2 that gives us a criterion for γ to generate a convergent sequence $\{\rho^{(k)}\}_k$.

Proposition 2. *Let* $\mathcal{F}(\rho) : \Delta^+ \to \mathbb{R}$ *a convex Lipschitz continuous function with respect to a fixed given norm* $||\cdot||$. *If* $\lim_{k\to\infty} \gamma^{(k)} = 0$ *and* $\sum_k \gamma^{(k)} = \infty$, *then, the sequence of points* $\{\rho^{(k)}\}_k$ *generated by* (4) *converge to* ρ^*.

Using the results of Theorem 1 and Proposition 2, we design an algorithm to choose a proper sequence of $\gamma^{(k)}$ such as $\rho^{(k)} \to \rho^*$. This protocol is summarized in Algorithm 1.

Algorithm 1 Discrete Wasserstein Gradient Flow

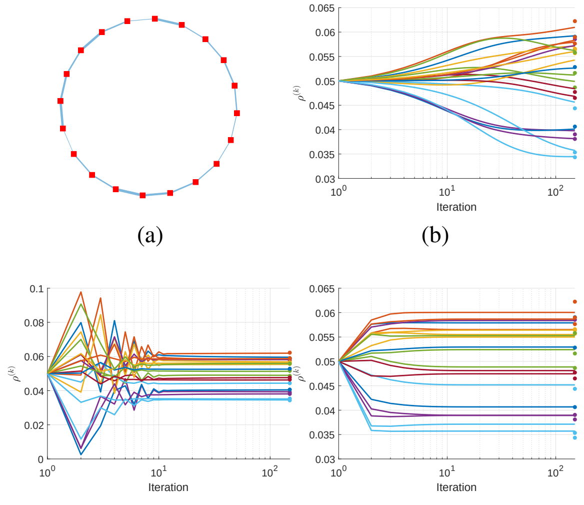
```
Read $\rho^{(0)}$, $\eta \in (0,1)$
$d \leftarrow \operatorname{grad} \mathcal{F}(\rho^{(0)})$, $\gamma^{(0)} \leftarrow \eta / \min_i d_i$
for $k \in \{1, \dots, K\}$ do
    $d \leftarrow \operatorname{grad} \mathcal{F}(\rho^{(k-1)})$
    for $i \in \mathcal{V}$ do
        $\gamma_i \leftarrow \min \left\{ \eta \rho_i^{(k-1)} / d_i, \gamma^{(k-1)} k/(k+1) \right\}$
        if $\gamma_i < 0$ then $\gamma_i \leftarrow \gamma^{(k-1)} k/(k+1)$ end
    end for
    $\rho^{(k)} \leftarrow \rho^{(k-1)} - (\min_i \gamma_i) d$
end for
```

4. Numerical Analysis

In this section, we present numerical simulations of our theoretical results. Therefore, we implement Algorithm 1 together with the Riemannian Gradient Descent (Bonnabel, 2013) and JKO flows (Peyré, 2015) to verify their performance. In these simulations we minimize the Kullback–Leibler divergence, that is $\mathcal{F}(\rho) = \sum_i \rho_i \log(\rho_i / q_i)$, with $q \in \Delta^+$. For all algorithms we set $\rho_i^{(0)} = 1/n$ for all $i \in \mathcal{V}$. Other additional parameters used in our simulations are $\epsilon = 0.01$ for Riemannian Gradient Descent and $\eta = 0.95$ and

$$\theta_{ij} = \begin{cases} \rho_i & \text{if } \partial_{\rho_i} \mathcal{F}(\rho) < \partial_{\rho_j} \mathcal{F}(\rho), \\ \rho_j & \text{if } \partial_{\rho_i} \mathcal{F}(\rho) > \partial_{\rho_j} \mathcal{F}(\rho), \\ \frac{\rho_i + \rho_j}{2} & \text{if } \partial_{\rho_i} \mathcal{F}(\rho) = \partial_{\rho_j} \mathcal{F}(\rho), \end{cases}$$

for Algorithm 1. For the Entropic Wasserstein Gradient Flows the parameters used are $\tau = 20$, $\bar{\gamma} = 2$ and $\varepsilon = 10^{-10}$. Additionally, since this algorithm needs the explicit definition of a distance over $\mathcal{V}$, we use the usual geodesic distance over graph.



(a) (b)

(c) (d)

Figure 1. Different implementations of Wasserstein gradient flows. (a) Interaction graph that supports the distribution. (b) Evolution of probability distribution using Riemannian Gradient Descent (Bonnabel, 2013). (c) Evolution of probability distribution using Algorithm 1. (d) Evolution of probability distribution using the implementation of JKO Flow presented in (Peyré, 2015). In Figure (a) the thickness of the edges is proportional its weight. In Figures (b)-(d) the filled markers at the end represents ρ^*.

We execute the algorithms in the graph shown in Fig. 1a. In Figures 1b-1d we present the evolution of the distributions through iteration for all three algorithms. It can be noticed in Fig. 1b that, at the expense of being a fully distributed algorithm, the Riemannian gradient descent is a slower algorithm. Instead, the evolution of Algorithm 1 is faster as shown in Fig. 1c but preserving the distributed property. Even it is comparable to the performance of a centralized algorithm as presented in Figure 1d.

5. Conclusions and Future Work

In this paper, we propose an alternative algorithm to calculate gradient flows under the Wasserstein-like metrics. By relating concepts from optimal transport and population game theory we not only present a new algorithm to minimize functions whose domain are probability distributions. In addition, we provide sufficient conditions to guarantee the convergence of the aforementioned algorithm. Moreover, we illustrate our theoretical results in an assortment of simulations. With them we can evidence that its performance is comparable to other presented strategies that tackle the same issue but with a lower computational cost. Future work should study accelerated gradient flows, capacity constraints in the network nodes and edges, and non-asymptotic convergence analysis.

References

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Backhoff-Veraguas, J., Fontbona, J., Rios, G., and Tobar, F. Bayesian learning with wasserstein barycenters. *arXiv:1805.10833*, 2018.

Barré, M., Giron, C., Mazzolini, M., and d'Aspremont, A. Averaging atmospheric gas concentration data using wasserstein barycenters. *arXiv:2010.02762*, 2020.

Bécigneul, G., Ganea, O.-E., Chen, B., Barzilay, R., and Jaakkola, T. Optimal transport graph neural networks. *arXiv preprint arXiv:2006.04804*, 2020.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Bonnabel, S. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58 (9):2217–2229, 2013.

Chow, S.-N., Li, W., and Zhou, H. Entropy dissipation of fokker-planck equations on graphs. *arXiv preprint arXiv:1701.04841*, 2017.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Fair regression with wasserstein barycenters. *arXiv:2006.07286*, 2020.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

Dong, Y. and Sawin, W. Copt: Coordinated optimal transport on graphs. *arXiv preprint arXiv:2003.03892*, 2020.

Erbar, M., Rumpf, M., Schmitzer, B., and Simon, S. Computation of optimal transport on discrete metric measure spaces. *Numerische Mathematik*, 144(1):157–200, 2020.

Essid, M. and Solomon, J. Quadratically regularized optimal transport on graphs. *SIAM Journal on Scientific Computing*, 40(4):A1961–A1986, 2018.

Hopkins, E. A note on best response dynamics. *Games and Economic Behavior*, 29(1-2):138–150, 1999.

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

Lavenant, H., Claici, S., Chien, E., and Solomon, J. Dynamical optimal transport on discrete surfaces. *ACM Transactions on Graphics (TOG)*, 37(6):1–16, 2018.

Mertikopoulos, P. and Sandholm, W. H. Riemannian game dynamics. *Journal of Economic Theory*, 177:315–364, 2018.

Monderer, D. and Shapley, L. S. Potential games. *Games and economic behavior*, 14(1):124–143, 1996.

Otto, F. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 2001.

Peyré, G. Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Sandholm, W. H. Population games and deterministic evolutionary dynamics. In *Handbook of game theory with economic applications*, volume 4, pp. 703–778. Elsevier, 2015.

Simon, D. and Aberdam, A. Barycenters of natural images constrained wasserstein barycenters for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7910–7919, 2020.

Udriste, C. *Convex functions and optimization methods on Riemannian manifolds*, volume 297. Springer Science & Business Media, 2013.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.