Entropy-regularized Wasserstein Distances for Analyzing Environmental and Ecological Data

†, *Hidekazu Yoshioka [1,2], * Yumi Yoshioka [1], and Yuta Yaegashi[3]

[1] Graduate School of Natural Science and Technology, Shimane University, Japan
[1] Fisheries Ecosystem Project Center, Shimane University, Japan
[2] Independent Researcher, Dr. of Agr., Japan

*Presenting author: yoshih@life.shimane-u.ac.jp
†Corresponding author: yoshih@life.shimane-u.ac.jp

Abstract

We explore applicability of entropy-regularized Wasserstein (pseudo-)distances as new tools for analyzing environmental and ecological data. In this paper, the two specific examples are considered and are numerically analyzed using the Sinkhorn algorithm. The first example is the inflow and outflow discharges of a dam-reservoir system. The inflow and outflow discharges are described as discrete-time Markov chains, and their transition rates among the discharge regimes and the corresponding stationary probability distributions are identified. The optimal transport plan leading to the regularized Wasserstein distance between the two Markov chains is considered as the system optimization policy decided by the operator. The second example is the body weight distributions of a fish serving as a major inland fishery resource in Japan. We quantify differences of the collected body weight distributions among the different years focusing on the summer growing season. The obtained analysis results imply usefulness of the regularized Wasserstein distances for assessing probability distributions arising in environmental and ecological problems.

Keywords: Aquatic environment and ecology, Optimal transport, Entropy-regularized Wasserstein distance, Sinkhorn algorithm

Introduction

Environmental and ecological dynamics in our world are inherently uncertain. The probability density functions or equivalently probability distributions can effectively quantify uncertainties involved in the target phenomena. Quantifying and comparing the probability distributions play an essential part in understanding and managing environmental and ecological dynamics [1].

The Wasserstein distances [2] are the metrics to rigorously measure difference between probability distributions. They originate from an optimization problem of transportation plans of materials from a set of starting points to a set of terminal points. They have been applied to a wide variety of research areas in both science and engineering, such as image processing, machine learning, and mathematical analysis and discretization of partial differential equations [2]-[3]. However, their applications to problems of environment and ecology have been far less explored to the best of the authors' knowledge. This is the motivation of our research.

In this paper, we apply the robust entropy-regularized Wasserstein (pseudo-)distances to unique environmental and ecological data collected in a river environment. The first application is to the discrete-time Markov chains representing inflow and outflow discharge

processes of an existing dam-reservoir system. The optimal plan as a minimizer in a Wasserstein distance is computed as the system optimization policy of the operator. The second application is to the body weight distributions of a fish in different years. We quantify difference among the distributions. The entropic regularization allows us to efficient as well as robust numerical computation of the Wasserstein distances. Our results would advance understanding and assessment of environmental and ecological data from a new viewpoint based on the Wasserstein distances.

Wasserstein distances

Standard Wasserstein distances

Wasserstein distances are the distances that can measure the differences between probability distributions [3]. For finite discrete probability distributions, namely for the two normalized histograms $a=\{a_i\}_{1\le i\le n}$ and $b=\{b_i\}_{1\le i\le n}$ with some $n\in\mathbb{N}$, the p th-order Wasserstein distance $W_p=W_p(a,b)=W_p(b,a)$ between a and b is set as

$$W_p^p=\min_P\sum_{i,j=1}^n C_{ij}P_{ij} \text{ with } C_{ij}=|i-j|^p \tag{1}$$

subject to the constraints

$$\sum_{j=1}^n P_{ij}=a_i,\ \sum_{i=1}^n P_{ij}=b_j,\ P_{ij}\ge 0,\ (1\le i,j\le n). \tag{2}$$

Here, C_{ij} is the transportation cost quantifying the difference between the classes i,j and the matrix $P=\{P_{ij}\}_{1\le i,j\le n}$ is referred to as a plan. A minimizing plan of (1) is called an optimal plan. This is a linear programming problem subject to constraints, but the resulting optimal plans are possibly not robust against the uncertainties in a and b because they are often non-unique and are of the non-interior type [2].

Regularized Wasserstein distances

The above-explained formulation would not be appropriate for problems under uncertain environment, where the histograms are not always accurate. This is often the case in handling histograms of environmental and ecological data. In such a case, it is more reasonable to consider the penalized problem subject to the same constraint (2):

$$W_{\varepsilon,p}^p=\min_P\left\{\sum_{i,j=1}^n C_{ij}P_{ij}+\varepsilon\sum_{i,j=1}^n P_{ij}\left(\ln P_{ij}-1\right)\right\} \tag{3}$$

with a penalty parameter $\varepsilon>0$. This $W_{\varepsilon,p}$ is not a distance due to not satisfying the triangle inequality, but we call it a “distance” for the sake of brevity. The added term is understood as the penalization against model uncertainty to improve the robustness, which formally vanishes as $\varepsilon\to+0$ [4]. The problems (1) and (3) coincide under this limit. Furthermore, the minimizer $P=P_\varepsilon$ of (3) converges to an optimal plan of the problem as $\varepsilon\to+0$ [3].

Using the penalized formulation has the following two computational advantages. Firstly, the optimal plan $P=P_\varepsilon$ is unique in (3) because the regularized problem is ε-convex. Secondly,

there exists a simple, fast, and stable algorithm for numerically finding P_ε: the Sinkhorn algorithm [3]:

$$u_i^{(m+1)} = \frac{a_i}{\sum_{j=1}^{n} K_{ij} v_j^{(m)}} \text{ and } v_j^{(m+1)} = \frac{b_j}{\sum_{j=1}^{n} K_{ji} u_i^{(m+1)}} \ (1 \le i, j \le n,\ m \ge 0) \tag{4}$$

with positive initial guesses $u_i^{(0)}, v_i^{(0)}$ $(1 \le i \le n)$, from which P_ε is obtained as

$$P_{\varepsilon,ij} = \lim_{m \to +\infty} u_i^{(m)} K_{ij} v_j^{(m)} \ (1 \le i, j \le n). \tag{5}$$

Here, we have set $K_{ij} = \exp\left(-C_{ij} / \varepsilon\right)$. Computational efficiency of the algorithm has been demonstrated in Cuturi [5]. Our implementation of the algorithm is based on the logarithmic rewriting [3] to avoid computational instability with small ε. Notice that $W_{\varepsilon,p}^p$ is increasing with respect to ε. We terminate the algorithm if the differences $\left|\ln u_i^{(m+1)} - \ln u_i^{(m)}\right|$ and $\left|\ln v_i^{(m+1)} - \ln v_i^{(m)}\right|$ become smaller than a sufficiently small error threshold (10^{-7} in this paper).

Wasserstein distances

Regularized Wasserstein distances

The first application is the inflow and outflow discharges of a dam-reservoir system in H River in Japan. The system has been operated from 2011 for multiple purposes including water resources supply and flood mitigation. Hourly inflow and outflow discharges data of the dam-reservoir system are available from April 1 in 2016.

We identify hourly discrete-time and discrete-state Markov chains of the inflow and outflow discharges using the collected data from April 1, 2016 to September 31, 2019. Seasonality of the data is not considered in this paper for the sake of simplicity, but will be addressed elsewhere. The discharge regimes are classified as follows: $S_i = \left[s_i, s_{i+1}\right)$ $(1 \le i \le n)$ with $s_i = i - 1$ (m^3/s) $(1 \le i \le 11)$ and $s_i = 10 + 8\left(i - 11\right)$ (m^3/s) $(12 \le i \le n)$, $n = 41$, and $S_{42} = +\infty$. This non-uniform partition has been employed because the average discharges are around 5.5 (m^3/s) for both the inflow and outflow records. We remark that, in this case, the system operation policy depending on the cost C_{ij} and the penalty parameter ε can be identified as the probability matrix $\left\{a_i^{-1} P_{\varepsilon,ij}\right\}_{1 \le i, j \le n}$ if $a_i > 0$ for all i.

Fig. 1 shows the estimated transition matrices $q = \left\{q_{ij}\right\}$ of the Markov chains of the inflow and outflow discharges. The estimated results show that the Markov chain for the outflow is closer to diagonally-dominant, meaning that the regime transitions occurred less frequently than in the inflow. The stationary probability distributions of the inflow ($a = \left\{a_i\right\}_{1 \le i \le n}$) and outflow discharges ($b = \left\{b_i\right\}_{1 \le i \le n}$) are used to numerically compute the regularized Wasserstein distances. Although not presented, it has been found that they are indeed increasing with respect to ε as theoretically expected.

Figs. 2 and **3** show the optimal plans P_ε for $p=1,2$ with $\varepsilon=0.01$ and $\varepsilon=10$. The optimal plan is sparser for smaller ε, implying its less robustness against perturbation of the input data. The computational results clearly indicate that the optimal plans are such that the discharges are significantly decreased through the dam-reservoir system for the lower regimes $1\le i\le 11$ where the discharges are smaller than 10 (m^3/s). Considering the Markov chains estimated in **Fig. 1** and the optimal plans in **Figs. 2** and **3**, this dam-reservoir system is serving as a filter to lower the lower flow and to a less transient flow. It is important to see that this characteristic of the system is visible both for the cases $p=1$ and $p=2$, although they are somewhat different for relatively high flow regimes $i, j\ge 12$.

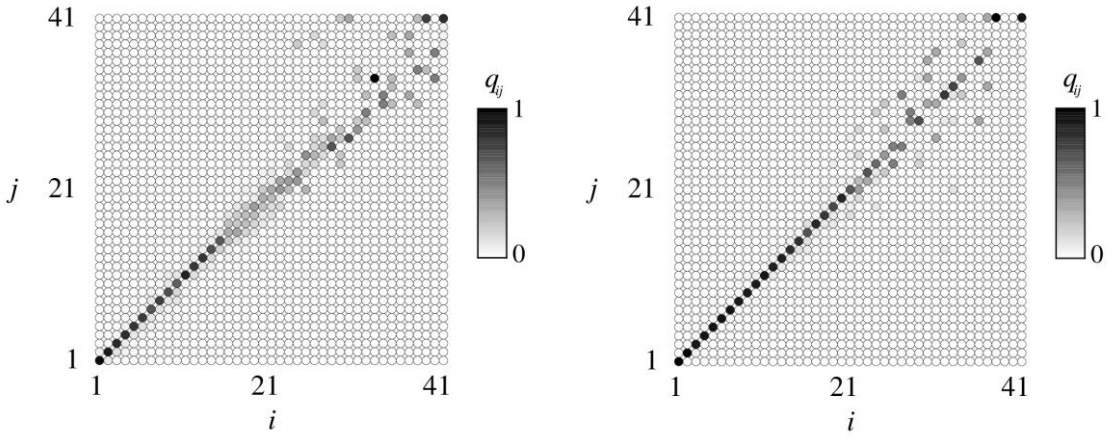

Figure 1. Transition probabilities of the Markov chains for the inflow discharge (Left) and outflow discharge (Right)

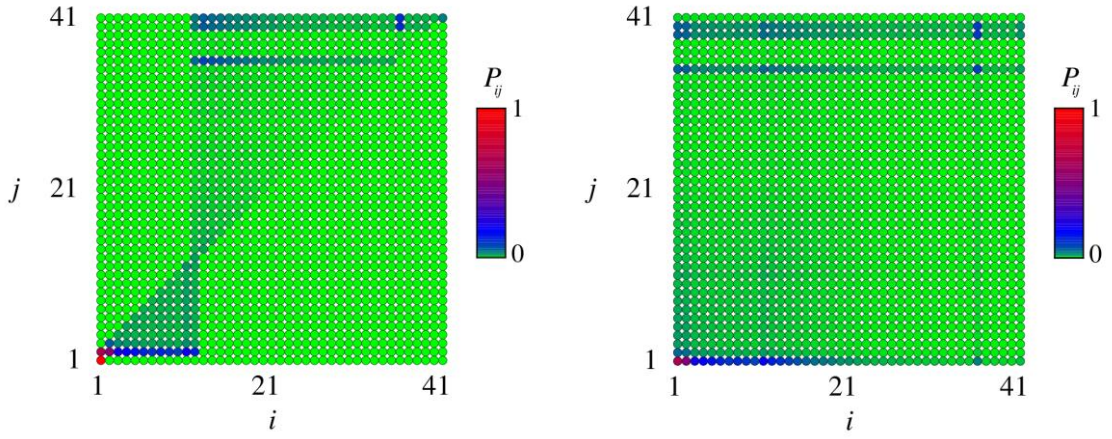

Figure 2. Optimal plans with $p=1$ for $\varepsilon=0.01$ (Left) and $\varepsilon=10$ (Right)

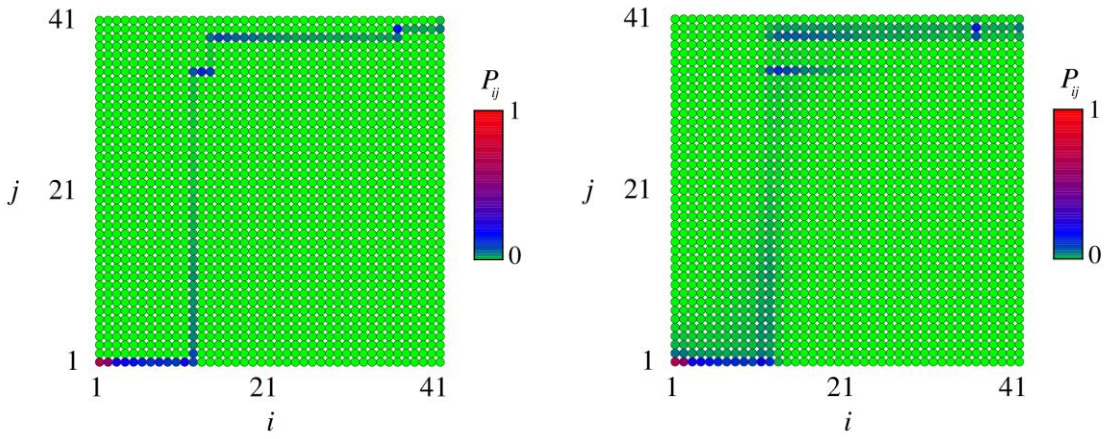

Figure 3. Optimal plans with $p=2$ for $\varepsilon=0.01$ (Left) and $\varepsilon=10$ (Right)

Body weights of a fish species

The second application focuses on the collected body weight distributions of the fish *Plecoglossus altivelis altivelis* as a major inland fishery resource in Japan [6]. The fish is one of the most important incomes for inland fishery cooperatives in the country. In addition, the fish is a key species in the aquatic ecosystems in and around river environment. Therefore, their growth dynamics are of critical importance. The life history of the fish is not explained here, but is found in Yoshioka et al. [7]. An important fact is that they have a one-year life history and grow significantly in summer, during which harvesting the fish is carried out.

We collected the body weight distributions of the fish at the beginning of August in 2017, 2018, 2019 in H River, and obtained the statistical estimates as demonstrated in **Table 1**. The data for 2017 and 2018 is found also in Yoshioka et al. [6, 8]. **Fig. 4** plots their distributions. The average values are around 56 to 57 (g) and the standard deviations around 18 to 19 (g). All the distributions have positive skewness values around 1. The collected data implies that the distributions are qualitatively the same.

An interest from a fisheries viewpoint is whether there exist significant quantitative differences among the three distributions. **Figs. 5** and **6** plot the computed regularized Wasserstein distances $W_{\varepsilon,p}$ for $p=1,2$ with respect to the different values of ε. There are at least two important findings from the figures. Firstly, the distance between 2017 and 2018 are the largest for both $p=1,2$. On the other hand, the relationship of the distances between 2018 and 2019 and that between 2017 and 2019 are opposite between $p=1,2$, especially when ε is small. In fact, $W_{0.01,1}$ are 0.159 and 0.193 between 2018 and 2019 and that between 2017 and 2019, respectively. On the other hand, $W_{0.01,2}$ are 0.222 and 0.201 between 2018 and 2019 and that between 2017 and 2019, respectively. This finding suggests that exploring a more biologically reasonable C_{ij} would be required. Nevertheless, the results suggest a significant difference between the data of 2017 and 2018.

Table 1. The collected data of the body weights of the fish

	2019	2018	2017
Total number of caught fishes	227	189	234
Average (g)	56.4	57.3	55.6
Standard deviation (g)	18.2	18.5	19.1
Skewness	0.95	1.16	0.78

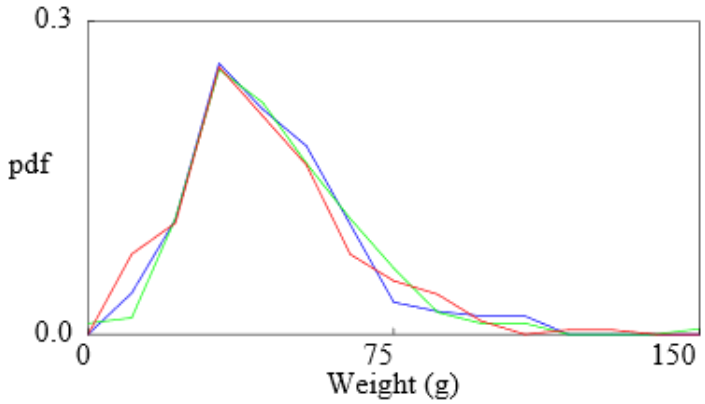

Figure 4. The body weight distributions in 2017 (Red), 2018 (Green), and 2019 (Blue)

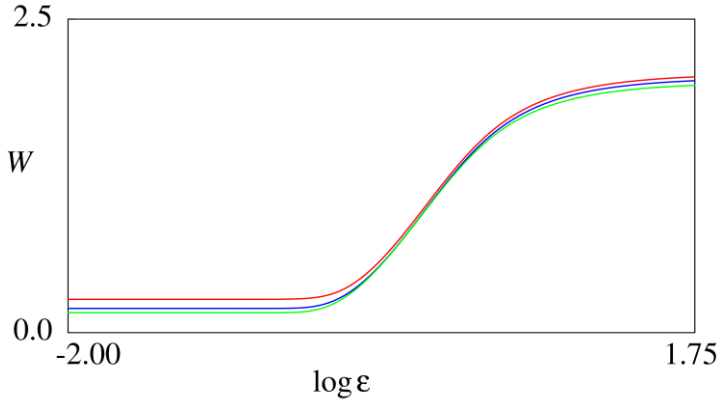

Figure 5. $W = W_{\varepsilon,p}$ **for** $p = 1$ **with respect to** ε**: The distance between 2017 and 2018 (Red), 2018 and 2019 (Green), and 2019 and 2017 (Blue)**

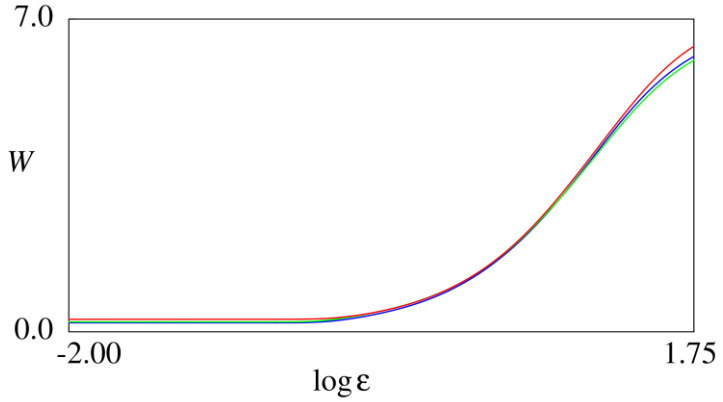

Figure 6. $W = W_{\varepsilon,p}$ **for** $p = 2$ **with respect to** ε **(the same legend with Fig. 5)**

Conclusions

The entropy-regularized (pseudo-)Wasserstein distances were applied to analyzing the unique environmental and ecological data. The application to the dam-reservoir system identified the optimal plan representing the system operation policy. Another application on the fish growth distributions quantified the differences among the collected distributions of the fish in different years.

Our results suggest that the regularized Wasserstein distances can serve as new tools for analyzing environmental and ecological data. A future reach topic would be computing the optimal plans of different dam-reservoir systems across the country, or across the globe, with which actual operational characteristics among them can be clarified. Analyzing applicability of the Wasserstein distances to other species, such as recently-found unique land-locked *P. altivelis* in Japan, is also an interesting topic.

Acknowledgements

JSPS Research Grant 19H03073, Kurita Water and Environment Foundation Grant 19B018, grants from MLIT Japan for ecological survey of a life history of the landlocked *Plecoglossus altivelis altivelis* and management of seaweed in Lake Shinji, and a research grant for young researchers in Shimane University support this research.

References

[1] Folke, C., Hahn, T., Olsson, P. and Norberg, J. (2005) Adaptive governance of social-ecological systems, *Annual Review of Environment and Resources* **30**, 441–473 (2005)
[2] Santambrogio, F. (2015) *Optimal Transport for Applied Mathematicians*, Birkäuser, New York.
[3] Peyré, G. and Cuturi, M. (2019) Computational optimal transport, *Foundations and Trends in Machine Learning* **11**, 355–607.
[4] Rigollet, P. and Weed, J. (2018) Entropic optimal transport is maximum-likelihood deconvolution, *Comptes Rendus Mathematique* **356**, 1228–1235.
[5] Cuturi, M. (2013) Sinkhorn distances: Lightspeed computation of optimal transport. In Advances in neural information processing systems. (pp. 2292–2300).
[6] Yoshioka, H., Yaegashi, Y., Yoshioka, Y. and Tsugihashi, K. (2019) A short note on analysis and application of a stochastic open-ended logistic growth model, *Letters in Biomathematics* **6**, 67–77.
[7] Yoshioka, H., Tanaka, T., Aranishi, F., Izumi, T. and Fujihara, M. (2019) Stochastic optimal switching model for migrating population dynamics, *Journal of Biological Dynamics* **13**, 706–732.
[8] Yoshioka, H., Yoshioka, Y., Yaegashi, Y. and Tsujimura, M. (2020) Chapter 2. Growth of the fish *Plecoglossus altivelis altivelis*. In Ayu and River Environment in Hii River, Japan -Research results from 2015 to 2020-. (pp. 13–26), Laboratory of Mathematical Sciences for Environment and Ecology, Shimane University, Matsue, Japan.