Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

Fast Gradient Computation for Gromov-Wasserstein Distance

Wei Zhang [*] [1], Zihao Wang [†] [2], Jie Fan [‡] [1], Hao Wu [§] [1], and Yong Zhang [¶] [3]

[1]Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China
[2]Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong SAR, China
[3]BNRist, RIIT, Institute of Internet Industry, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract. The Gromov-Wasserstein distance is a notable extension of optimal transport. In contrast to the classic Wasserstein distance, it solves a quadratic assignment problem that minimizes the pair-wise distance distortion under the transportation of distributions and thus could apply to distributions in different spaces. These properties make Gromov-Wasserstein widely applicable to many fields, such as computer graphics and machine learning. However, the computation of the Gromov-Wasserstein distance and transport plan is expensive. The well-known Entropic Gromov-Wasserstein approach has a cubic complexity since the matrix multiplication operations need to be repeated in computing the gradient of Gromov-Wasserstein loss. This becomes a key bottleneck of the method. Currently, existing methods accelerate the computation focus on sampling and approximation, which leads to low accuracy or incomplete transport plans. In this work, we propose a novel method to accelerate accurate gradient computation by dynamic programming techniques, reducing the complexity from cubic to quadratic. In this way, the original computational bottleneck is broken and the new entropic solution can be obtained with total quadratic time, which is almost optimal complexity. Furthermore, it can be extended to some variants easily. Extensive experiments validate the efficiency and effectiveness of our method.

Keywords:
Optimal transport,
Gromov-Wasserstein distance,
Fast gradient computation algorithm,
Fast algorithm.

Article Info.:
Volume: 3
Number: 3
Pages: 282 - 299
Date: September/2024
doi.org/10.4208/jml.240416

Article History:
Received: 16/04/2024
Accepted: 25/06/2024

Communicated by:
Zhi-Qin John Xu

1 Introduction

The Gromov-Wasserstein (GW) distance [31], as an important member of optimal transport [36, 40], is a powerful tool for distribution comparison. It is related to the Gromov-Hausdorff (GH) distance [17], a fundamental distance in metric geometry that measures how far two metric spaces are from being isometric [11]. Specifically, it measures the minimal distortion of pair-wise geodesic distances under the transport plan between two probabilistic distributions, even defined on different underlying spaces. Inherited from GH distance, GW is invariant to translation, rotation, and reflection of metric space. In

[*] zhang-w20@mails.tsinghua.edu.cn
[†] Corresponding author. zwanggc@cse.ust.hk
[‡] fanj21@mails.tsinghua.edu.cn
[§] hwu@tsinghua.edu.cn
[¶] zhangyong05@tsinghua.edu.cn

this way, GW has particular advantages to applications that require preserving geometry structures including computer graphics [30,37,48], natural language processing [3], graph factorization and clustering [14,55], and machine learning [10,57]. Moreover, variants of GW distance have been proposed for wider applications. For example, unbalanced GW (UGW) extends the comparison from probabilistic distributions to positive measures [44]. Fused GW (FGW) combines the GW and Wasserstein distances by interpolating their objectives, which is shown to be particularly effective for networks [50,52] and cross-domain distributions [35].

The computation of the Gromov-Wasserstein distance boils down to solving a nonconvex quadratic assignment problem that is NP-hard [22]. For this, some numerical methods built on relaxations have been developed, including convex relaxations [20,46,47], eigenvalue relaxations [24], etc. Nevertheless, these methods often require a large number of relaxed variables (for instance, $N^2 \times N^2$ variables in [20] where N is the number of discrete points of two spaces), resulting in high computational complexity. And they frequently provide unsatisfying solutions, especially in the presence of a symmetric metric matrix [37]. Entropic GW is a seminal and now the most popular work to compute GW distance from another perspective, which minimizes GW objective with an entropy regularization term [37,48]. In contrast to the non-entropy-based method mentioned above, it exhibits global convergence without removing constraints and offers a more concise computation. Moreover, it can adapt to solve GW variants, such as FGW [52] and UGW [44]. In each iteration of it, one first computes the GW gradient with matrix multiplications in $\mathcal{O}(N^3)$ time, which dominates the total complexity, and then solves the subproblem by the Sinkhorn algorithm [15] in $\mathcal{O}(N^2)$ time.

The computational cost of entropic GW remains unsatisfactory in large-scale scenarios. There are various methods to accelerate it (see Table 1.1 for the comparison). Scalable Gromov-Wasserstein learning method (S-GWL) [56] assumes the hierarchical structure of

Table 1.1: The comparison of different methods for the computation of GW metric and its variants. For SaGrow, the parameter s serves as a sampling parameter that dictates the quantity of specific sampled matrices. For spar-GW, the parameter s' designates the number of elements sampled from the GW gradient matrix. For LR-GW, the parameters r and d represent the presumed ranks of the distance matrices and the coupling matrices, respectively.

Method	Complexity	Exact and full-sized plan
Entropic GW and its approximations		
Entropic GW [37]	$\mathcal{O}(N^3)$	✓
S-GWL [56]	$\mathcal{O}(N^2 \log N)$	not exact
SaGroW [19]	$\mathcal{O}(N^2 s)$	not full-sized
Spar-GW [25]	$\mathcal{O}(N^2 + s'^2)$	not full-sized
LR-GW [42]	$\mathcal{O}(N(r^2 + d^2 + rd))$	not exact
AE [41]	$\mathcal{O}(N^2 \log N)$	not exact
GW on special structures		
FlowAlign [23]	$\mathcal{O}(N^2)$	tree only
FGC-GW (This work)	$\mathcal{O}(N^2)$	✓