Gromov-Monge distance helps understand the Gromov-Wasserstein distance

Quang Huy Tran
Univ. Bretagne-Sud, CNRS, IRISA
F-56000 Vannes
quang-huy.tran@univ-ubs.fr

Introduction. Let C be a 4-D tensor, where $C_{ijkl} = |C^x_{ik} - C^y_{jl}|^p$, for $p \geq 1$. The product measure $\mu^{\otimes 2} := \mu \otimes \mu$, where $(\mu \otimes \nu)_{ij} = \mu_i \otimes \nu_j$.

For $\mathcal{X} = (C_x, \mu_x)$ and $\mathcal{Y} = (C_y, \mu_y)$, where $C_x \in \mathbb{R}^{m \times m}$, $C_y \in \mathbb{R}^{n \times n}$, and $\mu_x \in \mathbb{R}^m_{\geq 0}$, $\mu_y \in \mathbb{R}^n_{\geq 0}$, define the UCOOT's objective function: for $\rho_1, \rho_2 \geq 0$ and $P, Q \geq 0$,

$$G_{C,\rho_{12}}(P, Q) = \langle C, P \otimes Q \rangle + \rho_1 \text{KL}(P_{\#1} \otimes Q_{\#1} | \mu_x \otimes \mu_x) + \rho_2 \text{KL}(P_{\#2} \otimes Q_{\#2} | \mu_y \otimes \mu_y) \quad (1)$$

The UGW reads

$$\begin{aligned}
\text{UGW}_{\rho_{12}}(\mathcal{X}, \mathcal{Y}) &= \inf_{P \geq 0} G_{C,\rho_{12}}(P, P) = \inf_{\substack{P,Q \geq 0 \\ P=Q}} G_{C,\rho_{12}}(P, Q) \\
&\geq \inf_{P,Q \geq 0} G_{C,\rho_{12}}(P, Q) = \inf_{\substack{P,Q \geq 0 \\ m(P)=m(Q)}} G_{C,\rho_{12}}(P, Q) = \text{LB-UGW}_{\rho_{12}}(\mathcal{X}, \mathcal{Y})
\end{aligned}$$

Let D be a $m \times n$ matrix whose coordinates are distances between features. Define the unbalanced OT's objective function: for $P \geq 0$,

$$F_{D,\rho_{34}}(P) = \langle D, P \rangle + \rho_3 \text{KL}(P_{\#1} | \mu_x) + \rho_4 \text{KL}(P_{\#2} | \mu_y) \quad (2)$$

and the UOT reads

$$\text{UOT}_{\rho_{34}}(\mu_x, \mu_y) = \inf_{P \geq 0} F_{D,\rho_{34}}(P)$$

The FGW reads: for $\lambda \in [0, 1]$,

$$\text{FGW}_\lambda(\mathcal{X}, \mathcal{Y}) = \inf_{P \in U(\mu_x, \mu_y)} \lambda \langle C, P \otimes P \rangle + (1 - \lambda) \langle D, P \rangle$$

Formulation. Now, fused UGW reads

$$\begin{aligned}
\text{FUGW}_{\rho,\lambda}(\mathcal{X}, \mathcal{Y}) &= \inf_{P \geq 0} \lambda G_{C,\rho_{12}}(P, P) + (1 - \lambda) F_{D,\rho_{34}}(P) \\
&= \inf_{\substack{P,Q \geq 0 \\ P=Q}} \lambda G_{C,\rho_{12}}(P, Q) + \frac{1 - \lambda}{2} \big[F_{D,\rho_{34}}(P) + F_{D,\rho_{34}}(Q) \big]
\end{aligned} \quad (3)$$

Remark 0.1. When $\rho_1, \rho_2, \rho_3, \rho_4 \to \infty$, then we recover FGW. When $\rho_1 = \rho_3 = \rho_4 = 0$, and either $\rho_2 = \infty$, then we recover semi-relaxed FGW.

Estimating UGW is numerically difficult, let alone FUGW. Thus, we study its lower bound:

$$\text{LB-FUGW}_{\rho,\lambda}(\mathcal{X}, \mathcal{Y}) = \inf_{\substack{P,Q \geq 0 \\ m(P)=m(Q)}} \lambda G_{C,\rho_{12}}(P, Q) + \frac{1 - \lambda}{2} \big[F_{D,\rho_{34}}(P) + F_{D,\rho_{34}}(Q) \big] \quad (4)$$

The additional mass constraint $m(P) = m(Q)$ may be advantageous because it may help BCD algo more numerically stable, similar to the UGW.

Preprint. Under review.

Proposition 0.2. *The problem 4 admits a minimiser under which condition? Should be similar to UGW and UOT.*

Proposition 0.3. *(Interpolation property)*

- *Intuitively, $FUGW_{\rho,\lambda}(\mathcal{X},\mathcal{Y})$ converges to $UGW_{\rho_{12}}(\mathcal{X},\mathcal{Y})$, when $\lambda \to 1$, and to $UOT_{\rho_{34}}(\mu_x,\mu_y)$ when $\lambda \to 0$.*
- *LB-$FUGW_{\rho,\lambda}(\mathcal{X},\mathcal{Y})$ converges to LB-$UGW_{\rho_{12}}(\mathcal{X},\mathcal{Y})$ when $\lambda \to 1$, and to $UOT_{\rho_{34}}(\mu_x,\mu_y)$ when $\lambda \to 0$.*

Proof. When $\lambda \to 1$, then same proof as the FUGW. When $\lambda \to 0$, consider the following lower bound

$$\widetilde{\text{LB-FUGW}}_{\rho,\lambda}(\mathcal{X},\mathcal{Y}) = \inf_{P,Q\geq 0} \lambda G_{C,\rho_{12}}(P,Q) + \frac{1-\lambda}{2}\big[F_{D,\rho_{34}}(P) + F_{D,\rho_{34}}(Q)\big]$$

Clearly, FUGW $\geq$ LB-FUGW $\geq \widetilde{\text{LB-FUGW}}$. When $\lambda \to 0$, show that $\widetilde{\text{LB-FUGW}} \to \frac{1}{2}(\text{UOT} + \text{UOT}) = \text{UOT}$ (intuitively, this should be true). By sandwich theorem and proposition 0.3, we conclude that LB-FUGW $\to$ UOT when $\lambda \to 0$. This is interesting because despite the mass constraint in the LB-FUGW, it has virtually no impact on the two UOT terms, for small λ. ■

Proposition 0.4. *For fixed $\lambda \in [0,1]$, for every $\rho_1,\rho_2,\rho_3,\rho_4 > 0$, we have*

- *$FUGW_{\rho,\lambda}(\mathcal{X},\mathcal{Y}) \leq FGW_{\lambda}(\mathcal{X},\mathcal{Y})$. Furthermore, $FUGW_{\rho,\lambda}(\mathcal{X},\mathcal{Y}) = 0$ iff $FGW_{\lambda}(\mathcal{X},\mathcal{Y}) = 0$.*
- *LB-$FUGW_{\rho,\lambda}(\mathcal{X},\mathcal{Y}) \leq$ LB-$FGW_{\lambda}(\mathcal{X},\mathcal{Y})$. Furthermore, LB-$FUGW_{\rho,\lambda}(\mathcal{X},\mathcal{Y}) = 0$ iff LB-$FGW_{\lambda}(\mathcal{X},\mathcal{Y}) = 0$.*

Entropic LB-UGW setting. Two possible entropic regularisation versions

1. Following UGW (corresponding to `reg_mode = "joint"`).
$$\text{LB-FUGW}_{\varepsilon,\rho,\lambda}(\mathcal{X},\mathcal{Y}) = \inf_{\substack{P,Q\geq 0\\ m(P)=m(Q)}} H_{\rho,\lambda}(P,Q) + \varepsilon\text{KL}\big(P\otimes Q|(\mu_x\otimes\mu_y)^{\otimes 2}\big) \tag{5}$$
2. Following COOT (corresponding to `reg_mode = "independent"`)
$$\text{LB-FUGW}_{\varepsilon,\rho,\lambda}(\mathcal{X},\mathcal{Y}) = \inf_{\substack{P,Q\geq 0\\ m(P)=m(Q)}} H_{\rho,\lambda}(P,Q) + \varepsilon\text{KL}(P|\mu_x\otimes\mu_y) + \varepsilon\text{KL}(Q|\mu_x\otimes\mu_y) \tag{6}$$

Proposition 0.5. *In both versions, we have LB-$FUGW_{\varepsilon,\rho,\lambda}(\mathcal{X},\mathcal{Y}) \to$ LB-$FUGW_{\rho,\lambda}(\mathcal{X},\mathcal{Y})$, when $\varepsilon \to 0$.*

The previous formulation is nice in terms of theoritical properties but bad in terms of implementation because it introduces too many hyperparameters (coming from the UOT term). It may be enough to relax the mass via the UGW term, no need to further introduce in the UOT term. Only linear terms are kept.

$$\text{FUGW}_{\rho}(\mathcal{X},\mathcal{Y}) = \inf_{\substack{P,Q\geq 0\\ m(P)=m(Q)}} G_\rho(P,Q) + \lambda\langle D, P+Q\rangle$$

Few observations:

1. If $\rho_1 = \rho_2 = \infty$, then recover fused GW.
2. FUGW is robust to outliers (the proof should be similar to that of UGW).
3. With the usual choice of cost C, we have
$$\begin{aligned}\text{FUGW}_{\rho}(\mathcal{X},\mathcal{Y}) &= \inf_{P,Q\geq 0} G_\rho(P,Q) + \lambda\langle D, P+Q\rangle \\ &= \inf_{P\geq 0} G_\rho(P,P) + 2\lambda\langle D, P\rangle\end{aligned}$$

In practice, we consider

$$\text{FUCOOT}_{\rho,\lambda}(\mathcal{X},\mathcal{Y}) = \inf_{\substack{P,Q\geq 0\\ m(P)=m(Q)}} G_{C,\rho_{xy}}(P,Q) + \lambda_s F_{D_s,\rho_s}(P) + \lambda_f F_{D_f,\rho_f}(Q) \tag{7}$$

Under the constraint $m(P) = m(Q) = m$, the complete objective function of FUCOOT reads

$$\begin{aligned}
H_{\rho,\lambda}(P,Q) &= G_{C,\rho_{xy}}(P,Q) + \lambda_s F_{D_s,\rho_s}(P) + \lambda_f F_{D_f,\rho_f}(Q) \\
&= \langle C, P\otimes Q\rangle + \rho_x \text{KL}(P_{\#1}\otimes Q_{\#1}|\mu_{nx}\otimes\mu_{dx}) + \rho_y \text{KL}(P_{\#2}\otimes Q_{\#2}|\mu_{ny}\otimes\mu_{dy}) \\
&+ \lambda_s\Big(\langle D_s, P\rangle + \rho_1^{(s)}\text{KL}(P_{\#1}|\mu_{nx}) + \rho_2^{(s)}\text{KL}(P_{\#2}|\mu_{ny})\Big) \\
&+ \lambda_f\Big(\langle D_f, Q\rangle + \rho_1^{(f)}\text{KL}(Q_{\#1}|\mu_{dx}) + \rho_2^{(f)}\text{KL}(Q_{\#2}|\mu_{dy})\Big)
\end{aligned}$$

Two possible entropic regularisation versions, define $\mu_n := \mu_{nx}\otimes\mu_{ny}$ and $\mu_d := \mu_{dx}\otimes\mu_{dy}$.

1. Following UGW (corresponding to `reg_mode = "joint"`).

$$\text{LB-FUGW}_{\varepsilon,\rho,\lambda}(\mathcal{X},\mathcal{Y}) = \inf_{\substack{P,Q\geq 0\\ m(P)=m(Q)}} H_{\rho,\lambda}(P,Q) + \varepsilon\text{KL}(P\otimes Q|\mu_n\otimes\mu_d) \tag{8}$$

2. Following COOT (corresponding to `reg_mode = "independent"`)

$$\text{LB-FUGW}_{\varepsilon,\rho,\lambda}(\mathcal{X},\mathcal{Y}) = \inf_{\substack{P,Q\geq 0\\ m(P)=m(Q)}} H_{\rho,\lambda}(P,Q) + \varepsilon_s\text{KL}(P|\mu_n) + \varepsilon_f\text{KL}(Q|\mu_d) \tag{9}$$

Using the relation

$$\text{KL}(P|\mu) = \int \log\big(\frac{dP}{d\mu}\big)dP - m(P) + m(\mu)$$

and for fixed P, denote $m = m(P)$.

$$\begin{aligned}
&\text{KL}(P_{\#1}\otimes Q_{\#1}|\mu\otimes\nu) \\
&= m(Q)\text{KL}(P_{\#1}|\mu) + m(P)\text{KL}(Q_{\#1}|\nu) + \big[m(P)-m(\mu)\big]\big[m(Q)-m(\nu)\big] \\
&= \int\Big(\int \log\big(\frac{dP_{\#1}}{d\mu}\big)dP_{\#1}\Big)dQ + m(P)\text{KL}(Q_{\#1}|\nu) - m(\nu)\big[m(P)-m(\mu)\big] \\
&= \int\Big(\int \log\big(\frac{dP_{\#1}}{d\mu}\big)dP_{\#1}\Big)dQ + m\text{KL}(Q_{\#1}|\nu) + \text{ constant}
\end{aligned}$$

Then, solving the problem 8 is equivalent to solving

$$\inf_{Q\geq 0}\langle L, Q\rangle + \Big(m\rho_x + \lambda_f\rho_1^{(f)}\Big)\text{KL}(Q_{\#1}|\mu_{dx}) + \Big(m\rho_y + \lambda_f\rho_2^{(f)}\Big)\text{KL}(Q_{\#2}|\mu_{dy}) + \varepsilon m\text{KL}(Q|\mu_d)$$

where

$$L = C\otimes P + \lambda_f D_f + \rho_x\langle\log\big(\frac{P_{\#1}}{\mu_{nx}}\big), P_{\#1}\rangle + \rho_y\langle\log\big(\frac{P_{\#2}}{\mu_{ny}}\big), P_{\#2}\rangle + \varepsilon\langle\log\big(\frac{P}{\mu_n}\big), P\rangle$$

and olving the problem 9 is equivalent to solving

$$\inf_{Q\geq 0}\langle L, Q\rangle + \Big(m\rho_x + \lambda_f\rho_1^{(f)}\Big)\text{KL}(Q_{\#1}|\mu_{dx}) + \Big(m\rho_y + \lambda_f\rho_2^{(f)}\Big)\text{KL}(Q_{\#2}|\mu_{dy}) + \varepsilon_f\text{KL}(Q|\mu_d)$$

where

$$L = C\otimes P + \lambda_f D_f + \rho_x\langle\log\big(\frac{P_{\#1}}{\mu_{nx}}\big), P_{\#1}\rangle + \rho_y\langle\log\big(\frac{P_{\#2}}{\mu_{ny}}\big), P_{\#2}\rangle$$

Algorithm 1 Generic scaling algorithm

Input. Solving

$$\min_{P\geq 0} \langle C, P\rangle + \rho_1 \mathrm{KL}(P_{\#1}|\mu) + \rho_2 \mathrm{KL}(P_{\#2}|\nu) + \varepsilon \mathrm{KL}(P|\mu\otimes\nu)$$

.

Output. Pair of optimal dual vectors (f, g) and coupling P.

1. While not converge, update
$$\begin{cases} f = -\frac{\rho_1}{\rho_1+\varepsilon}\log\sum_j \exp\left(g_j + \log\nu_j - \frac{C_{\cdot,j}}{\varepsilon}\right) \\ g = -\frac{\rho_2}{\rho_2+\varepsilon}\log\sum_i \exp\left(f_i + \log\mu_i - \frac{C_{i,\cdot}}{\varepsilon}\right) \end{cases}$$
2. Calculate $P = (\mu\otimes\nu)\exp\left(f\oplus g - \frac{C}{\varepsilon}\right)$.

Here $\otimes$ and $\oplus$ are the tensor product and sum, respectively. Some tricks: for any matrix M, we write $M^{\odot 2} := M\odot M$, where $\odot$ is the element-wise multiplication.

1. Suppose $A\in\mathbb{R}^{n_1\times d_1}$ and $B\in\mathbb{R}^{n_2\times d_2}$. For $P\in\mathbb{R}^{d_1\times d_2}$, we have $|A-B|^2\otimes P\in\mathbb{R}^{n_1\times n_2}$, where
$$|A-B|^2\otimes P = A^{\odot 2}P_{\#1}\oplus B^{\odot 2}P_{\#2} - 2APB^T.$$
2. If $A = (a_1, ..., a_m)\in\mathbb{R}^{m\times d}$ and $B = (b_1, ..., b_n)\in\mathbb{R}^{n\times d}$, then the matrix $D\in\mathbb{R}^{m\times n}$ defined by $D_{ij} = ||a_i - b_j||_2^2$ can be decomposed as $D = D_a D_b^T$, where $D_a = (A^{\odot 2}1_d, 1_m, -\sqrt{2}A)\in\mathbb{R}^{m\times(d+2)}$ and $D_b = (1_n, B^{\odot 2}1_d, \sqrt{2}B)\in\mathbb{R}^{n\times(d+2)}$. So, instead of storing D, we store D_a and D_b, so that we can scale up easily when the dimension d is small.

So, for $C = |C_x - C_y|^2$, with $(C_x)_{ij} = ||x_i - x_j||_2^2$ and $(C_y)_{kl} = ||y_k - y_l||_2^2$ and $D_{ij} = ||a_i^{(x)} - a_j^{(y)}||_2^2$, we have $C_x P C_y^T = A_1 A_2^T P B_2 B_1^T$. Denote $M = A_2^T P B_2\in\mathbb{R}^{(d_1+2)\times(d_2+2)}$, then $C_x P C_y^T = A_1 M B_1^T$.

Algorithm 2 Approximation algorithm for FUGW

Input. Graphs $X = (C^x, \mu_x), Y = (C^y, \mu_y)$, with distance matrix D between features, parameters $\rho_1, \rho_2 > 0$, interpolation parameter $\lambda\in[0,1]$, the regularisation parameter $\varepsilon > 0$ and initialisation P_0.

Output. Pair of optimal couplings (P, Q).

- **While** P_k has not converged **do**
  1. Q_{k+1} is the solution for fixed P_k.
  2. Rescale $Q_{k+1} = \sqrt{\frac{m(P_k)}{m(Q_{k+1})}}Q_{k+1}$.
  3. P_{k+1} is the solution for fixed Q_{k+1}.
  4. Rescale $P_{k+1} = \sqrt{\frac{m(Q_{k+1})}{m(P_{k+1})}}P_{k+1}$.

The regularised and unregularised UOT can be solved with MM algorithm: the iteration reads

$$\begin{aligned} P &= \left[\left(\frac{\mu}{P_{\#1}}\right)^{\lambda_1}\otimes\left(\frac{\nu}{P_{\#2}}\right)^{\lambda_2}\right]\odot P^{\lambda_1+\lambda_2}\odot(\mu\otimes v)^r\odot\exp\left(-\frac{C}{\lambda}\right) \\ &= \frac{P^{\lambda_1+\lambda_2}}{P_{\#1}^{\lambda_1}\otimes P_{\#2}^{\lambda_2}}\odot\left(\mu^{\lambda_1+r}\otimes\nu^{\lambda_2+r}\right)\odot\exp\left(-\frac{C}{\lambda}\right) \end{aligned} \tag{10}$$

where $\lambda = \rho_1+\rho_2+\varepsilon$ and $\lambda_i = \frac{\rho_i}{\lambda}$ and $r = \frac{\varepsilon}{\lambda}$. Or for more stability,

$$\begin{aligned} \log P &= (\lambda_1+\lambda_2)\log P - (\lambda_1\log P_{\#1}\oplus\lambda_2\log P_{\#2}) \\ &\quad + [(\lambda_1+r)\log\mu\oplus(\lambda_2+r)\log\nu] - \frac{C}{\lambda} \end{aligned} \tag{11}$$

In the example of neuro-image: source and target data

- Functional data: $F_s \in \mathbb{R}^{160k \times 300}, F_t \in \mathbb{R}^{60k \times 300}$.
- Anatomy data: $A_s \in \mathbb{R}^{160k \times 6}, A_t \in \mathbb{R}^{60k \times 6}$.

Input of FUGW: distance matrix $K \in \mathbb{R}^{160k \times 60k}$ between A_s and A_t for the fused part. For GW part: distance matrix $D_s \in \mathbb{R}^{160k \times 160k}$ and $D_t \in \mathbb{R}^{60k \times 60k}$.

Formulation used in fugw full

$$
\begin{aligned}
\text{FUGW}_{\rho,\alpha}(X, Y) = \min_{P,Q \geq 0} & \langle \text{cost}, P \otimes Q \rangle \\
& + \rho_1 \text{KL}(P_{\#1} \otimes Q_{\#1} | \mu \otimes \mu) + \rho_2 \text{KL}(P_{\#2} \otimes Q_{\#2} | \nu \otimes \nu) \\
& + \alpha \left[\langle K, P \rangle + \rho_3 \text{KL}(P_{\#1} | \mu) + \rho_4 \text{KL}(P_{\#2} | \nu) \right] \\
& + \alpha \left[\langle K, Q \rangle + \rho_3 \text{KL}(Q_{\#1} | \mu) + \rho_4 \text{KL}(Q_{\#2} | \nu) \right]
\end{aligned}
\tag{12}
$$

References