Improved Complexity Bounds in Wasserstein Barycenter Problem

Darina Dvinskikh, Daniil Tiapkin∗
Weierstrass Institute, HSE

2021. 3. 19

Outlines

OT: Arithmetic calculations problem

- Simplex method or interior point method: $\widetilde{O}(n^3)$
- Sinkhorn algorithm: $\widetilde{O}\left(n^2\|C\|_\infty^2/\varepsilon^2\right)$, with ε-precision. C is cost matrix, the regularization parameter before negative entropy is of order ε.
- Accelerated Sinkhorn algorithm: $\widetilde{O}\left(n^{2.5}\|C\|_\infty/\varepsilon\right)$. In practice, it has better dependence on ε but not on n.
- All entropy-regularized based approaches are numerically unstable when the regularizer parameter γ before negative entropy is small (this also means that precision ε is high as γ must be selected proportional to ε.

The recent work(Jambulapati et al. (2019)) privides an optimal method: $\widetilde{O}\left(n^2\|C\|_\infty/\varepsilon\right)$

- Based on dual extrapolation and area-convexity.
- Without additional penalization.

Wasserstein Barycenter problem

- Iterative Bregman projections (IBP) algorithm: The IBP is an extension of the Sinkhorn' s algorithm for m measures, and hence, its complexity is m times more than the Sinkhorn complexity: $\widetilde{O}\left(mn^2\|C\|_\infty^2/\varepsilon^2\right)$.
- Accelerated IBP algorithm: $\widetilde{O}\left(mn^{2.5}\|C\|_\infty/\varepsilon\right)$.
- Another fast version of IBP, FastIBP: $\widetilde{O}\left(mn^{\frac{7}{3}}\|C\|_\infty^{\frac{4}{3}}/\varepsilon^{\frac{4}{3}}\right)$.

Contribution

The first contribution:

- Propose an algorithm which does not suffer from a small value of the regularization parameter.
- Convergence rate: $\widetilde{O}\left(mn^{2.5}\|C\|_\infty/\varepsilon\right)$, not worse than the celebrated accelerated IBP.
- Based on mirror prox with specific prox-function.

The second contribution:

- Propose an algorithm that has better complexity than the accelerated IBP.
- Convergence rate: $\widetilde{O}\left(mn^{2}\|C\|_\infty/\varepsilon\right)$.
- Based on rewriting the WB problem as a saddle-point problem and further application of the dual extrapolation scheme under the weaker convergence requirements of area-convexity.

In some sense, the first algorithm can be seen as a simplified version of the second algorithm.

Contribution

Table 1: Algorithms and their rates of convergence for the Wasserstein barycenter problem

Approach	**Paper**	**Complexity**
IBP	(Kroshnin et al., 2019)	$\widetilde{O}\left(\frac{mn^2\|C\|_\infty^2}{\varepsilon^2}\right)$
Accelerated IBP	(Guminov et al., 2019)	$\widetilde{O}\left(\frac{mn^2\sqrt{n}\|C\|_\infty}{\varepsilon}\right)$
FastIBP	(Lin et al., 2020)	$\widetilde{O}\left(\frac{mn^2\sqrt[3]{n}\|C\|_\infty^{4/3}}{\varepsilon\sqrt[3]{\varepsilon}}\right)$
Mirror prox with specific norm	This work	$\widetilde{O}\left(\frac{mn^2\sqrt{n}\|C\|_\infty}{\varepsilon}\right)$
Dual extrapolation with area-convexity	This work	$\widetilde{O}\left(\frac{mn^2\|C\|_\infty}{\varepsilon}\right)$

Problem Statement

- Let $\Delta_n = \left\{p \in \mathbb{R}_+^n : \sum_{i=1}^n p_i = 1\right\}$ be the probability simplex. Given two histograms $p, q \in \Delta_n$ and ground cost $C \in \mathbb{R}_+^{n*n}$, the OT problem is formulated as follows

$$W(p, q) = \min_{X \in \mathcal{U}(p,q)} \langle C, X \rangle \tag{1}$$

- where X is a transport plan.$\mathcal{U}$ is the transport polytope.
- Let d be vectorized cost matrix of C, x be vectorized transport plan of X, $b = \begin{pmatrix} p \\ q \end{pmatrix}$, and $A = \{0, 1\}^{2n \times n^2}$ be an incidence matrix.
- As $\sum_{i,j=1}^n X_{ij} = 1$, (Jambulapati et al. (2019)) rewrite (1) as

$$W(p, q) = \min_{x \in \Delta_n 2} \max_{y \in [-1,1]^{2n}} \left\{d^\top x + 2\|d\|_\infty \left(y^\top Ax - b^\top y\right)\right\} \tag{2}$$

Problem Statement

- Given histograms $q_1, q_2, ..., q_m \in \Delta_n$, a WB of those measures is a solution of the following problem:

$$p^* = \arg\min_{p \in \Delta_n} \frac{1}{m} \sum_{i=1}^{m} W(p, q_i) \tag{3}$$

- Rewrite (3) using the reformulation (2) of OT as follows:

$$\min_{p \in \Delta_n} \frac{1}{m} \sum_{i=1}^{m} \min_{x_i \in \Delta_n 2} \max_{y_i \in [-1,1]^{2n}} \left\{ d^\top x_i + 2\|d\|_\infty \left(y_i^\top A x_i - b_i^\top y_i \right) \right\} \tag{4}$$

Problem Statement

- Define spaces $\mathcal{X} \triangleq \prod^m \Delta_{n^2} \times \Delta_n$ and $\mathcal{Y} \triangleq [-1,1]^{2mn}$, where $\prod^m \Delta_{n^2} \times \Delta_n = \underbrace{\Delta_{n^2} \times \ldots \times \Delta_{n^2}}_{m} \times \Delta_n$. Reweite (4) for column vectors $\mathbf{x} = \left(x_1^\top, \ldots, x_m^\top, p^\top\right)^\top \in \mathcal{X}$ and $\mathbf{y} = \left(y_1^\top, \ldots, y_m^\top\right)^\top \in \mathcal{Y}$ as follows:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{m}\left\{\boldsymbol{d}^\top \mathbf{x} + 2\|d\|_\infty \left(\mathbf{y}^\top \boldsymbol{A}\mathbf{x} - \boldsymbol{c}^\top \mathbf{y}\right)\right\} \tag{5}$$

- As objective $F(\mathbf{x},\mathbf{y})$ in (5) is convex in x and concave in y, problem (5) is a saddle-point problem.
- where $\mathbf{d} = (d^\top, \ldots, d^\top, 0_n^\top)^\top$, $\mathbf{c} = \left(0_n^\top, q_1^\top, \ldots, 0_n^\top, q_m^\top\right)^\top$ and $\boldsymbol{A} = \begin{pmatrix} \hat{A} & \mathcal{E} \end{pmatrix} \in \{-1,0,1\}^{2mn\times(mn^2+n)}$ with block-diagonal matrix $\boldsymbol{A}$ of m blocks $\hat{A} = \begin{pmatrix} A & 0_{2n\times n^2} & \cdots & 0_{2n\times n^2} \\ 0_{2n\times n^2} & A & \cdots & 0_{2n\times n^2} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{2n\times n^2} & 0_{2n\times n^2} & \cdots & A \end{pmatrix}$ and matrix $\mathcal{E}^\top = (\underbrace{(-I_n \quad 0_{n\times n})}_{-B_2^\top} \underbrace{(-I_n \quad 0_{n\times n})}_{-B_2^\top} \cdots \underbrace{(-I_n \quad 0_{n\times n})}_{-B_2^\top})$.

Setup

- Notation: For a prox-function $d(x)$, define the corresponding Bregman divergence $B(x, y) = d(x) - d(y) - \langle \nabla d(y), x - y\rangle$. For example, the Euclidean $\mathbf{l}_2$-norm $\|\mathbf{y}\|_2$, prox-function $d_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|_2^2$, and the $B_{\mathcal{Y}}(\mathbf{y}, \breve{\mathbf{y}}) = \frac{1}{2}\|\mathbf{y} - \breve{\mathbf{y}}\|_2^2$
- For space $\mathcal{X} \triangleq \prod^m \Delta_{n^2} \times \Delta_n$, we choose the following specific norm $\|\mathbf{x}\|_{\mathcal{X}} = \sqrt{\sum_{i=1}^m \|x_i\|_1^2 + m\|p\|_1^2}$ for $\mathbf{x} = (x_1, \ldots, x_m, p)^T$. Given $\mathcal{X}$ with prox-function $d_{\mathcal{X}}(\mathbf{x}) = \sum_{i=1}^m \langle x_i, \ln x_i\rangle + m\langle p, \ln p\rangle$ and $B_{\mathcal{X}}(\mathbf{x}, \breve{\mathbf{x}}) = \sum_{i=1}^m \langle x_i, \ln(x_i/\breve{x}_i)\rangle - \sum_{i=1}^m \mathbf{1}^\top (x_i - \breve{x}_i) + m\langle p, \ln(p/\breve{p})\rangle - m\mathbf{1}^\top(p - \breve{p})$.
- Define $R_{\mathcal{X}}^2 = \sup_{\mathbf{x}\in\mathcal{X}} d_{\mathcal{X}}(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}} d_{\mathcal{X}}(\mathbf{x})$ and $R_{\mathcal{Y}}^2 = \sup_{\mathbf{y}\in\mathcal{Y}} d_{\mathcal{Y}}(\mathbf{y}) - \min_{\mathbf{y}\in\mathcal{Y}} d_{\mathcal{Y}}(\mathbf{y})$.
- Definition: $f(\widetilde{\mathbf{x}}, \mathbf{y})$ is $(L_{xx}, L_{xy}, L_{yx}, L_{yy})$-smooth if for any $x, x^{'} \in \mathcal{X}$ and $y, y^{'} \in \mathcal{Y}$

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}', \mathbf{y})\|_{\mathcal{X}^*} \le L_{\mathbf{xx}} \|\mathbf{x} - \mathbf{x}'\|_{\mathcal{X}},$$
$$\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}')\|_{\mathcal{X}^*} \le L_{\mathbf{xy}} \|\mathbf{y} - \mathbf{y}'\|_{\mathcal{Y}},$$
$$\left\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}')\right\|_{\mathcal{Y}^*} \le L_{\mathbf{yy}} \|\mathbf{y} - \mathbf{y}'\|_{\mathcal{Y}},$$
$$\left\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}', \mathbf{y})\right\|_{\mathcal{Y}^*} \le L_{\mathbf{yx}} \|\mathbf{x} - \mathbf{x}'\|_{\mathcal{X}}.$$

Implementation

- As problem (5) is a saddle-point problem, we will evaluate the quality of an algorithm that outputs a pair of solutions $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) \in (\mathcal{X}, \mathcal{Y})$ through the so-called duality gap

$$\max_{\mathbf{y}\in\mathcal{Y}} F(\widetilde{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x}, \widetilde{\mathbf{y}}) \leq \varepsilon \tag{6}$$

- The first algorithm is based on mirror prox (MP) algorithm (Nemirovski, 2004) on space $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ with prox-function $d_{\mathcal{Z}}(\mathbf{z}) = a_1 d_{\mathcal{X}}(\mathbf{x}) + a_2 d_{\mathcal{Y}}(\mathbf{y})$ and $B_{\mathcal{Z}}(\mathbf{z}, \breve{\mathbf{z}}) = a_1 B_{\mathcal{X}}(\mathbf{x}, \breve{\mathbf{x}}) + a_2 B_{\mathcal{Y}}(\mathbf{y}, \breve{\mathbf{y}})$, where $a_1 = \frac{1}{R_{\mathcal{X}}^2}, a_2 = \frac{1}{R_{\mathcal{Y}}^2}$,

$$\begin{pmatrix} \mathbf{u}^{k+1} \\ \mathbf{v}^{k+1} \end{pmatrix} = \arg\min_{\mathbf{z}\in\mathcal{Z}} \left\{\eta G\left(\mathbf{x}^k, \mathbf{y}^k\right)^\top \mathbf{z} + B_{\mathcal{Z}}\left(\mathbf{z}, \mathbf{z}^k\right)\right\},$$

$$\mathrm{z}^{k+1} = \arg\min_{\mathbf{z}\in\mathcal{Z}} \left\{\eta G\left(\mathbf{u}^{k+1}, \mathbf{v}^{k+1}\right)^\top \mathbf{z} + B_{\mathcal{Z}}\left(\mathbf{z}, \mathbf{z}^k\right)\right\}.$$

Implementation

- Here η is learning rate, $\mathbf{z}^1 = \arg\min_{\mathbf{z}\in\mathcal{Z}} d_{\mathcal{Z}}(\mathbf{z})$ and $G(\mathbf{x},\mathbf{y})$ is a gradient operator defined as follows

$$G(\mathbf{x},\mathbf{y}) = \begin{pmatrix} \nabla_{\mathbf{x}} F(\mathbf{x},\mathbf{y}) \\ -\nabla_{\mathbf{y}} F(\mathbf{x},\mathbf{y}) \end{pmatrix} = \frac{1}{m}\begin{pmatrix} \boldsymbol{d} + 2\|d\|_\infty \boldsymbol{A}^\top \mathbf{y} \\ 2\|d\|_\infty(\boldsymbol{c} - \boldsymbol{A}\mathbf{x}) \end{pmatrix} \tag{7}$$

- If $F(\mathbf{x},\mathbf{y})$ is $(L_{xx}, L_{xy}, L_{yx}, L_{yy})$-smooth, then to satify (6) with $\widetilde{\mathbf{x}} = \frac{1}{N}\sum_{k=1}^N \mathbf{u}^k, \tilde{\mathbf{y}} = \frac{1}{N}\sum_{k=1}^N \mathbf{v}^k$ one needs to perform

$$N = \frac{4}{\varepsilon}\max\left\{L_{\mathbf{xx}}R_{\mathcal{X}}^2, L_{\mathrm{xy}}R_{\mathcal{X}}R_{\mathcal{Y}}, L_{\mathbf{yx}}R_{\mathcal{Y}}R_{\mathcal{X}}, L_{\mathbf{yy}}R_{\mathcal{Y}}^2\right\}\Bigg) \tag{8}$$

- iterations of MP(Bubeck, 2014) with

$$\eta = 1/\left(2\max\left\{L_{\mathbf{xx}}R_{\mathcal{X}}^2, L_{\mathrm{xy}}R_{\mathcal{X}}R_{\mathcal{Y}}, L_{\mathbf{yx}}R_{y}R_{\mathcal{X}}, L_{\mathbf{yy}}R_{\mathcal{Y}}^2\right\}\right) \tag{9}$$

Complexity Bound

- Lemma: Objective $F(\mathbf{x}, \mathbf{y})$ in (5) is $(L_{xx}, L_{xy}, L_{yx}, L_{yy})$-smooth with $L_{xx} = L_{yy} = 0$ and $L_{\mathrm{xy}} = L_{\mathbf{yx}} = 2\sqrt{2}\|d\|_\infty/m$.
- Theorem: Assume that $F(\mathbf{x}, \mathbf{y})$ in (5) is $\left(0, 2\sqrt{2}\|d\|_\infty/m, 2\sqrt{2}\|d\|_\infty/m, 0\right)$-smooth and $R_{\mathcal{X}} = \sqrt{3m\ln n}$, $R_{\mathcal{Y}} = \sqrt{mn}$. Then after $N = 8\|d\|_\infty\sqrt{6n\ln n}/\varepsilon$ iterations, Algorithm 1 with $\eta = \frac{1}{4\|d\|_\infty\sqrt{6n\ln n}}$ outputs a pair $(\widetilde{\mathbf{u}}, \widetilde{\mathbf{v}}) \in (\mathcal{X}, \mathcal{Y})$ such that $\max_{\mathbf{y}\in\mathcal{Y}} F(\widetilde{\mathbf{u}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x}, \widetilde{\mathbf{v}}) \le \varepsilon$.
- The total complexity of Algorithm 1 is $O\left(mn^2\sqrt{n\ln n}\|d\|_\infty\varepsilon^{-1}\right)$. The complexity of one iteration of Algorithm 1 is $O\left(mn^2\right)$ as the number of non-zero elements in matrix A is $2n^2$, and m is the number of vector-components in $\mathbf{y}$ and $\mathbf{x}$. Multiplying this by the number of iterations N, we get the result.
- As d is the vectorized cost matrix of C, we may reformulate the complexity results of Theorem with respect to C as $O\left(mn^2\sqrt{n\ln n}\|C\|_\infty\varepsilon^{-1}\right)$.

Algorithm 1 Mirror Prox for Wasserstein Barycenters

Input: measures $q_1, ..., q_m$, linearized cost matrix d, incidence matrix A, step η, starting points $p^1 = \frac{1}{n}\mathbf{1}_n$, $x_1^1 = ... = x_m^1 = \frac{1}{n^2}\mathbf{1}_{n^2}$, $y_1^1 = ... = y_m^1 = \mathbf{0}_{2n}$

$\alpha = 2\|d\|_\infty \eta n$, $\beta = 6\|d\|_\infty \eta \ln n$, $\gamma = 3m\eta \ln n$.

for $k = 1, 2, \cdots, N-1$ **do**

  for $i = 1, 2, \cdots, m$ **do**

    $v_i^{k+1} = y_i^k + \alpha \left(A x_i^k - \begin{pmatrix} p^k \\ q_i \end{pmatrix} \right)$,

    Project v_i^{k+1} onto $[-1,1]^{2n}$

$$u_i^{k+1} = \frac{x_i^k \odot \exp\left\{-\gamma\left(d + 2\|d\|_\infty A^\top y_i^k\right)\right\}}{\sum\limits_{l=1}^{n^2} [x_i^k]_l \exp\left\{-\gamma\left([d]_l + 2\|d\|_\infty [A^\top y_i^k]_l\right)\right\}}$$

  end for

$$s^{k+1} = \frac{p^k \odot \exp\left\{\beta \sum_{i=1}^m [y_i^k]_{1...n}\right\}}{\sum_{l=1}^n [p^k]_l \exp\left\{\beta \sum_{i=1}^m [y_i^k]_l\right\}}$$

  for $i = 1, 2, \cdots, m$ **do**

    $y_i^{k+1} = y_i^k + \alpha \left(A u_i^{k+1} - \begin{pmatrix} s^{k+1} \\ q_i \end{pmatrix} \right)$

    Project y_i^{k+1} onto $[-1,1]^{2n}$

$$x_i^{k+1} = \frac{x_i^k \odot \exp\left\{-\gamma\left(d + 2\|d\|_\infty A^\top v_i^{k+1}\right)\right\}}{\sum\limits_{l=1}^{n^2} [x_i^k]_l \exp\left\{-\gamma\left([d]_l + 2\|d\|_\infty [A^\top v_i^{k+1}]_l\right)\right\}}$$

  end for

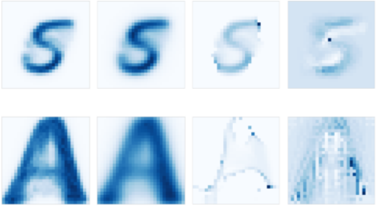
$$p^{k+1} = \frac{p^k \odot \exp\left\{\beta \sum_{i=1}^m [v_i^{k+1}]_{1...n}\right\}}{\sum_{l=1}^n [p^k]_l \exp\left\{\beta \sum_{i=1}^m [v_i^{k+1}]_l\right\}}$$

end for

Output: $\tilde{\mathbf{u}} = \sum\limits_{k=1}^N \begin{pmatrix} u_1^k \\ \vdots \\ u_m^k \\ s^k \end{pmatrix}$, $\tilde{\mathbf{v}} = \sum\limits_{k=1}^N \begin{pmatrix} v_1^k \\ \vdots \\ v_m^k \end{pmatrix}$

MNIST and notMNIST

- The result of IBP with regularizing parameter γ is numberically unstable, as γ must be selected proportional to ε.



Mirror Prox for WB | Dual Extra-polation | IBP, $\gamma = 10^{-3}$ | IBP, $\gamma = 10^{-5}$

Figure 1: WBs of hand-written digit '5' (first row) and of letters 'A' (second row) computed by Algorithm 1 (Mirror Prox for WB), Algorithm 4 (Dual Extrapolation for WB) and the IBP with small values of the regularizing parameter.

Gaussian measures

- To compare the convergence of the proposed algorithms, we randomly generated 10 Gaussian measures with equally spaced support of 100 points in [−10,10], mean from [−5,5] and variance from [0.8,1.8].
- Figure 2 presents the convergence with respect to the function optimality gap $\frac{1}{m}\sum_{i=1}^{m}\mathcal{W}(p, q_i) - \frac{1}{m}\sum_{i=1}^{m}\mathcal{W}(p^*, q_i)$. Here p^* is the true bartcenter.

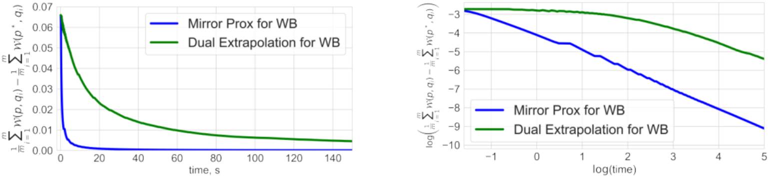


Figure 2: Convergence of Algorithm 1 (Mirror Prox for WB) and Algorithm 4 (Dual Extrapolation for WB) to the true barycenter of Gaussian measures w.r.t the function optimality gap $\frac{1}{m}\sum_{i=1}^{m}\mathcal{W}(p, q_i) - \frac{1}{m}\sum_{i=1}^{m}\mathcal{W}(p^*, q_i)$. Here p^* is the true barycenter.

Gaussian measures

- Algorithm 4 has better complexity bound, Algorithm 1 has better convergence in practice. The slope ration -1 for the convergence of Algorithm 1 in log-scale perfectly fits the theoretical dependence of working time (iteration number N) on the desired accuracy ε ($N \sim \varepsilon^{-1}$ from Theorem).
- For Algorithm 4, this slope ratio -1 is achieved only after a number of iterations but this is due to the need of solving practically computationally costly subproblems.

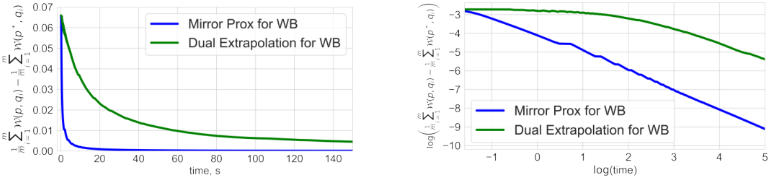
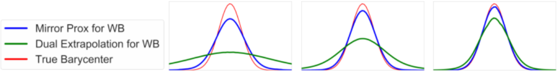


Figure 2: Convergence of Algorithm 1 (Mirror Prox for WB) and Algorithm 4 (Dual Extrapolation for WB) to the true barycenter of Gaussian measures w.r.t the function optimality gap $\frac{1}{m}\sum_{i=1}^{m}\mathcal{W}(p,q_i) - \frac{1}{m}\sum_{i=1}^{m}\mathcal{W}(p^*,q_i)$. Here p^* is the true barycenter.

Gaussian measures

- Figure 3 illustrates the convergence of the barycenters obtained by Algorithms 1 and 4 to the true barycenter.



After 20 seconds | After 100 seconds | After 500 seconds

Figure 3: Convergence of the barycenters obtained by Algorithm 1 (Mirror Prox for WB) and Algorithm 4 (Dual Extrapolation for WB) to the true barycenter of Gaussian measures.

- Figure 4 illustrates better approximations of the true Gaussian barycenter by Algorithms 1 and 4 compared to the *gamma*-regularized IBP barycenter.

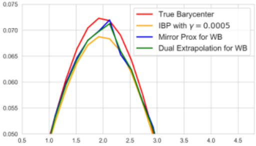


Figure 4: Convergence of the barycenters obtained by Algorithm 1 (Mirror Prox for WB), Algorithm 4 (Dual Extrapolation for WB), and the IBP to the true barycenter of Gaussian measures.