Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Worm Draft

Marina Meilă MMP@STAT.WASHINGTON.EDU
Department of Statistics
University of Washington
Seattle, WA 98195-4322, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Division of Computer Science and Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Editor: Kevin Murphy and Bernhard Schölkopf

1. Introduction

Table 1: Notations

Notations	Definitions
$\mathbb{R}^d$	d-dimensional Euclidean space;
$[n]$	Denotes the set $\{1, 2, \ldots, n\}$;
$[n, +\infty)$	Real numbers greater than or equal to n;
$SE(d)$	Special Euclidean group over $\mathbb{R}^d$;

To simplify expressions, we define "Pattern" as nonnegative weighted nonempty finite point set with total weights equal to 1???

Definition 1 (Wasserstein Distance (Cuturi and Doucet (2014))) *Let (Ω, d) be a metric space and $P(\Omega)$ be the set of Borel probability measure on Ω. For any real number $s \geq 1$ and probability measures $\mu, v \in P(\Omega)$, their s^{th}-Wasserstein distance is*

$$W_s(\mu, v) := \left(\inf_{\pi \in \Pi(\mu, v)} \int_{\Omega^2} d(x, y)^s \mathrm{d}\pi(x, y) \right)^{1/s}, \tag{1}$$

where $\Pi(\mu, v)$ is the set of all probability measures on Ω^2 with marginals μ and v.

Definition 2 (Discrete Wasserstein Distance (DWD) (Rubner et al. (1998))) *Let $P = \{p_1, p_2, \ldots, p_{n_P}\}$ and $Q = \{q_1, q_2, \ldots, q_{n_Q}\}$ be two sets of weighted points in $\mathbb{R}^d$ with nonnegative weights α_i and β_j for each $p_i \in P$ and $q_j \in Q$, and $\sum_{i=1}^{n_P} \alpha_i = \sum_{j=1}^{n_Q} \beta_j = 1$. For any real number $s \geq 1$, their s^{th}-DWD is*

$$\mathcal{W}_s(P, Q) = \left(\min_F \sum_{i=1}^{n_P} \sum_{j=1}^{n_Q} f_{i_j} d(p_i - q_j)^s \right)^{1/s}, \tag{2}$$

where $d(\cdot,\cdot)$ is a distance function on $\mathbb{R}^d$ and $F = \{f_{i_j} \mid i \in [n_P], j \in [n_Q]\}$ is a feasible flow from P to Q, i.e., each $f_{i_j} \geq 0$, $\sum_{i=1}^{n_P} f_{i_j} = \beta_j$, and $\sum_{j=1}^{n_Q} f_{i_j} = \alpha_i$.

Definition 3 (Discrete Wasserstein Barycenter(DWB)) *Given a set of nonnegative weighted nonempty finite point sets $\{P_i \mid P_i \subset \mathbb{R}^d, i \in [m]\}$, where each P_i has the total weights equal to 1. The DWB of $\{P_i \mid P_i \subset \mathbb{R}^d, i \in [m]\}$ is a new point set $\tilde{P}$ such that*

$$\sum_{i=1}^{m} \mathcal{W}_s^s(P_i, \tilde{P}) \tag{3}$$

is minimized, where $\tilde{P}$ is a nonnegative weighted nonempty finite point sets with total weights equal to 1.

Since a rigid transformation is a geometric transformation for a Euclidean space (Wikipedia contributors (2021)), we just consider l_2-norm instead of arbitrary distance function $d(\cdot,\cdot)$ for DWD under rigid transformation.

Definition 4 (DWD under Rigid Transformation (DWD-RT) (Ding and Liu (2018))) *Let $P = \{p_1, p_2, \ldots, p_{n_P}\}$ and $Q = \{q_1, q_2, \ldots, q_{n_Q}\}$ be two sets of weighted points in $\mathbb{R}^d$ with nonnegative weights α_i and β_j for each $p_i \in P$ and $q_j \in Q$, and $\sum_{i=1}^{n_P} \alpha_i = \sum_{j=1}^{n_Q} \beta_j = 1$. For any real number $s \geq 1$, their s^{th}-DWD-RT is*

$$\mathcal{WRT}_s(P, Q) = \left(\min_{\substack{F,\\ e \in SE(d)}} \sum_{i=1}^{n_P} \sum_{j=1}^{n_Q} f_{i_j} \|p_i - e(q_j)\|_2^s \right)^{1/s}, \tag{4}$$

where $\|\cdot\|_2$ is l_2-norm on $\mathbb{R}^d$ and $F = \{f_{i_j} \mid i \in [n_P], j \in [n_Q]\}$ is a feasible flow from P to Q, i.e., each $f_{i_j} \geq 0$, $\sum_{i=1}^{n_P} f_{i_j} = \beta_j$, and $\sum_{j=1}^{n_Q} f_{i_j} = \alpha_i$.

Definition 5 (DWB under Rigid Transformation (DWB-RT)) *Given a set of nonnegative weighted nonempty finite point sets $\{P_i \mid P_i \subset \mathbb{R}^d, i \in [m]\}$, where each P_i has the total weights equal to 1. The DWB-RT of $\{P_i \mid P_i \subset \mathbb{R}^d, i \in [m]\}$ is a new point set $\tilde{P}$ such that*

$$\sum_{i=1}^{m} \mathcal{WRT}_s^s(P_i, \tilde{P}) \tag{5}$$

is minimized, where $\tilde{P}$ is a nonnegative weighted nonempty finite point sets with total weights equal to 1.

In this paper, we focus on the 2^{nd}-DWD

$$\mathcal{W}_2(P, Q) = \left(\min_{F} \sum_{i=1}^{n_P} \sum_{j=1}^{n_Q} f_{i_j} d\left(p_i - q_j\right)^2 \right)^{1/2}, \tag{6}$$

and the 2^{nd}-DWD-RT

$$\mathcal{WRT}_2(P,Q) = \left(\min_{\substack{F,\\ e\in SE(d)}} \sum_{i=1}^{n_P} \sum_{j=1}^{n_Q} f_{i_j} \|p_i - e(q_j)\|_2^2 \right)^{1/2}, \tag{7}$$

where s in Eq.(2) and (4) are replaced by constant 2.

Lemma 6 *Let x and $y_1, \ldots, y_r$ be non-negative real number, and real numbers $\epsilon > 0, p \geq 1$. Then*

$$\left(x + \sum_{i=1}^{r} y_i\right)^s \leq (1+\epsilon)^{s-1} x^s + \left(\frac{(1+\epsilon)r}{\epsilon}\right)^{s-1} \sum_{i=1}^{r} y_i^s.$$

Lemma 7 (Lemma 3) *Given three nonnegative weighted nonempty finite point sets $P_1, P_2, P_3 \subset \mathbb{R}^d$, where each point set has the total weights equal to 1. Then*

$$\mathcal{W}_2^2(P_1,P_2) \leq 2\mathcal{W}_2^2(P_1,P_3) + 2\mathcal{W}_2^2(P_3,P_2). \tag{8}$$

Proof Since $\mathcal{W}_2(\cdot,\cdot)$ is a distance function, we have

$$\begin{aligned}
\mathcal{W}_2^2(P_1,P_2) &= (\mathcal{W}_2(P_1,P_2))^2 \\
&\leq (\mathcal{W}_2(P_1,P_3) + \mathcal{W}_2(P_3,P_2))^2 \\
&= \mathcal{W}_2^2(P_1,P_3) + \mathcal{W}_2^2(P_3,P_2) + 2\mathcal{W}_2(P_1,P_3)\mathcal{W}_2(P_3,P_2) \\
&\leq \mathcal{W}_2^2(P_1,P_3) + \mathcal{W}_2^2(P_3,P_2) + (\mathcal{W}_2(P_1,P_3))^2 + (\mathcal{W}_2(P_3,P_2))^2 \\
&= 2\mathcal{W}_2^2(P_1,P_3) + 2\mathcal{W}_2^2(P_3,P_2).
\end{aligned}$$

■

Lemma 8 (Lemma 6) *Let $P_1, P_2, P_3 \subset \mathbb{R}^d$ be three nonnegative weighted nonempty finite point sets, where each point set has the total weights equal to 1. Then for any real numbers $\epsilon > 0$,*

$$|\mathcal{W}_2^2(P_1,P_2) - \mathcal{W}_2^2(P_1,P_3)| \leq \left(1 + \frac{1}{\epsilon}\right) \mathcal{W}_2^2(P_2,P_3) + \epsilon \mathcal{W}_2^2(P_1,P_2). \tag{9}$$

Proof First, we consider the case $\mathcal{W}_2^2(P_1,P_2) \geq \mathcal{W}_2^2(P_1,P_3)$. Then we have

$$\begin{aligned}
&|\mathcal{W}_2^2(P_1,P_2)-\mathcal{W}_2^2(P_1,P_3)|\\
=&\mathcal{W}_2^2(P_1,P_2)-\mathcal{W}_2^2(P_1,P_3)\\
=&\left(\mathcal{W}_2(P_1,P_2)\right)^2-\mathcal{W}_2^2(P_1,P_3)\\
\leq&\left(\mathcal{W}_2(P_1,P_3)+\mathcal{W}_2(P_3,P_2)\right)^2-\mathcal{W}_2^2(P_1,P_3)\\
\leq&\left(1+\frac{1}{\epsilon}\right)\left(\mathcal{W}_2(P_2,P_3)\right)^2+(1+\epsilon)\left(\mathcal{W}_2(P_1,P_3)\right)^2-\mathcal{W}_2^2(P_1,P_3) \qquad (10a)\\
=&\left(1+\frac{1}{\epsilon}\right)\mathcal{W}_2^2(P_2,P_3)+\epsilon\mathcal{W}_2^2(P_1,P_3)\\
\leq&\left(1+\frac{1}{\epsilon}\right)\mathcal{W}_2^2(P_2,P_3)+\epsilon\mathcal{W}_2^2(P_1,P_2), \qquad (10b)
\end{aligned}$$

where (10a) comes from Lemma 6. For this case, we assume $\mathcal{W}_2^2(P_1,P_2)\geq\mathcal{W}_2^2(P_1,P_3)$, thus (10b) holds.

For the other case $\mathcal{W}_2^2(P_1,P_2)<\mathcal{W}_2^2(P_1,P_3)$, we directly have

$$\begin{aligned}
&|\mathcal{W}_2^2(P_1,P_2)-\mathcal{W}_2^2(P_1,P_3)|\\
=&\mathcal{W}_2^2(P_1,P_3)-\mathcal{W}_2^2(P_1,P_2)\\
\leq&\left(1+\frac{1}{\epsilon}\right)\mathcal{W}_2^2(P_2,P_3)+\epsilon\mathcal{W}_2^2(P_1,P_2)
\end{aligned} \qquad (11)$$

by exchanging the roles of P_2 and P_3 in Eq.(10). ■

2. Dynamic

Lemma 9 (Lemma 3) *Given three nonnegative weighted nonempty finite point sets $P_1,P_2,P_3\subset\mathbb{R}^d$, where each point set has the total weights equal to 1. Then*

$$\mathcal{WRT}_2^2(P_1,P_2)\leq 2\mathcal{WRT}_2^2(P_1,P_3)+2\mathcal{WRT}_2^2(P_3,P_2). \qquad (12)$$

Proof W.l.o.g., we assume that the induced rigid transformations of $\mathcal{WRT}_2^2(P_1,P_3)$, $\mathcal{WRT}_2^2(P_3,P_2)$ are $e_1,e_2\in SE(d)$, respectively. That is,

$$\begin{aligned}
\mathcal{WRT}_2^2(P_1,P_3)&=\mathcal{W}_2^2(e_1(P_1),P_3),\\
\mathcal{WRT}_2^2(P_2,P_3)&=\mathcal{W}_2^2(e_2(P_2),P_3).
\end{aligned} \qquad (13)$$

Thus,

$$
\begin{aligned}
&\mathcal{WRT}_2^2(P_1,P_2)\\
\leq&\mathcal{W}_2^2\left(e_1(P_1),e_2(P_2)\right)\\
=&\left(\mathcal{W}_2\left(e_1(P_1),e_2(P_2)\right)\right)^2\\
\leq&\left(\mathcal{W}_2\left(e_1(P_1),P_3\right)+\mathcal{W}_2\left(P_3,e_2(P_2)\right)\right)^2\\
=&\mathcal{W}_2^2\left(e_1(P_1),P_3\right)+\mathcal{W}_2^2\left(P_3,e_2(P_2)\right)+2\mathcal{W}_2\left(e_1(P_1),P_3\right)\mathcal{W}_2\left(P_3,e_2(P_2)\right)\\
\leq&\mathcal{W}_2^2\left(e_1(P_1),P_3\right)+\mathcal{W}_2^2\left(P_3,e_2(P_2)\right)+\left(\mathcal{W}_2\left(e_1(P_1),P_3\right)\right)^2+\left(\mathcal{W}_2\left(P_3,e_2(P_2)\right)\right)^2\\
=&2\mathcal{W}_2^2\left(e_1(P_1),P_3\right)+2\mathcal{W}_2^2\left(P_3,e_2(P_2)\right)\\
=&2\mathcal{WRT}_2^2(P_1,P_3)+2\mathcal{WRT}_2^2(P_3,P_2).
\end{aligned}
\tag{14}
$$

■

Lemma 10 (Lemma 6) *Let $P_1,P_2,P_3\subset\mathbb{R}^d$ be three nonnegative weighted nonempty finite point sets, where each point set has the total weights equal to 1. Then for any real numbers $\epsilon>0$,*

$$
|\mathcal{WRT}_2^2(P_1,P_2)-\mathcal{WRT}_2^2(P_1,P_3)|\leq\left(1+\frac{1}{\epsilon}\right)\mathcal{WRT}_2^2(P_2,P_3)+\epsilon\mathcal{WRT}_2^2(P_1,P_2). \tag{15}
$$

Proof W.l.o.g., we assume that the induced rigid transformations of $\mathcal{WRT}_2^2(P_1,P_3)$, $\mathcal{WRT}_2^2(P_3,P_2)$ are $e_1,e_2\in SE(d)$, respectively. That is,

$$
\begin{aligned}
\mathcal{WRT}_2^2(P_1,P_3)=\mathcal{W}_2^2(e_1(P_1),P_3),\\
\mathcal{WRT}_2^2(P_2,P_3)=\mathcal{W}_2^2(e_2(P_2),P_3).
\end{aligned}
\tag{16}
$$

First, we consider the case $\mathcal{WRT}_2^2(P_1,P_2)\geq\mathcal{WRT}_2^2(P_1,P_3)$. Then we have

$$
\begin{aligned}
&|\mathcal{WRT}_2^2(P_1,P_2)-\mathcal{WRT}_2^2(P_1,P_3)|\\
=&\mathcal{WRT}_2^2(P_1,P_2)-\mathcal{WRT}_2^2(P_1,P_3)\\
\leq&\mathcal{W}_2^2\left(e_1(P_1),e_2(P_2)\right)-\mathcal{WRT}_2^2(P_1,P_3)\\
\leq&\left(\mathcal{W}_2\left(e_1(P_1),P_3\right)+\mathcal{W}_2\left(P_3,e_2(P_2)\right)\right)^2-\mathcal{WRT}_2^2(P_1,P_3)\\
=&\left(\mathcal{WRT}_2(P_1,P_3)+\mathcal{WRT}_2(P_3,P_2)\right)^2-\mathcal{WRT}_2^2(P_1,P_3)\\
\leq&\left(1+\frac{1}{\epsilon}\right)\left(\mathcal{WRT}_2(P_3,P_2)\right)^2+(1+\epsilon)\left(\mathcal{WRT}_2(P_1,P_3)\right)^2-\mathcal{WRT}_2^2(P_1,P_3) && \text{(17a)}\\
=&\left(1+\frac{1}{\epsilon}\right)\mathcal{WRT}_2^2(P_2,P_3)+\epsilon\mathcal{WRT}_2^2(P_1,P_3)\\
\leq&\left(1+\frac{1}{\epsilon}\right)\mathcal{WRT}_2^2(P_2,P_3)+\epsilon\mathcal{WRT}_2^2(P_1,P_2), && \text{(17b)}
\end{aligned}
$$

where (17a) comes from Lemma 6. For this case, we assume $\mathcal{WRT}_2^2(P_1,P_2) \geq \mathcal{WRT}_2^2(P_1,P_3)$, thus (17b) holds.

For the other case $\mathcal{WRT}_2^2(P_1,P_2) < \mathcal{WRT}_2^2(P_1,P_3)$, we directly have

$$\begin{aligned}
&|\mathcal{WRT}_2^2(P_1,P_2) - \mathcal{WRT}_2^2(P_1,P_3)|\\
=&\mathcal{WRT}_2^2(P_1,P_3) - \mathcal{WRT}_2^2(P_1,P_2)\\
\leq&\left(1+\frac{1}{\epsilon}\right)\mathcal{WRT}_2^2(P_2,P_3) + \epsilon\mathcal{WRT}_2^2(P_1,P_2)
\end{aligned} \tag{18}$$

by exchanging the roles of P_2 and P_3 in Eq.(17). ■

3. Proof for Uniform Sampling

For convenience, we represent both $\mathcal{W}_2^2(\cdot,\cdot)$ and $\mathcal{WRT}_2^2(\cdot,\cdot)$ as $\mathcal{M}(\cdot,\cdot)$.

We follow the notations in (Ding and Liu (2018)), we know

$$T = \sum_{P_l \in \mathbb{P}} t_{\mathbb{P}}(P_l) \leq 8(\alpha+1) + 4\alpha + 16.$$

Theorem 11 (Uniform Sampling (Langberg and Schulman (2010); Varadarajan and Xiao (2012) *Let $Q \subset \mathbb{R}^d$ be nonnegative weighted nonempty finite point set, and the total weights of Q equal to 1.*

- *If we take a sample P_i from $\mathbb{P}$ uniformly at random,*

$$E\left[\mathcal{M}(P_i,Q)\right] = \frac{1}{m}\sum_{P_l \in \mathbb{P}} \mathcal{M}(P_l,Q). \tag{19}$$

- *Take a sample $\mathbb{S}$ of size of r from $\mathbb{P}$ uiformunly at random, and $P_i \in \mathbb{S}$ are selected independently. We assume that*

$$\frac{\max_{P_i \in \mathbb{P}} \mathcal{M}(P_i,Q)}{\frac{1}{m}\sum_{P_l \in \mathbb{P}} \mathcal{M}(P_l,Q)} \leq C. \tag{20}$$

Then for any real number $\epsilon > 0$,

$$\Pr\left[|\frac{1}{m}\sum_{P_l \in \mathbb{P}} \mathcal{M}(P_l,Q) - \frac{1}{r}\sum_{P_l \in \mathbb{S}} \mathcal{M}(P_l,Q)| \leq \epsilon\frac{1}{m}\sum_{P_l \in \mathbb{P}} \mathcal{M}(P_l,Q)\right] \geq 1-2e^{-\frac{2r\epsilon^2}{C^2}}. \tag{21}$$

Proof Hoeffding's inequality indicates that

$$\begin{aligned}
&\Pr\left[|\frac{1}{m}\sum_{P_l\in\mathbb{P}}\mathcal{M}(P_l,Q)-\frac{1}{r}\sum_{P_l\in\mathbb{S}}\mathcal{M}(P_l,Q)|\le\epsilon\frac{1}{m}\sum_{P_l\in\mathbb{P}}\mathcal{M}(P_l,Q)\right]\\
\ge&1-2e^{-\frac{2r\epsilon^2\cdot\left[\frac{1}{m}\sum_{P_l\in\mathbb{P}}\mathcal{M}(P_l,Q)\right]^2}{[\max_{P_i\in\mathbb{P}}\mathcal{M}(P_i,Q)]^2}}\\
\ge&1-2e^{-\frac{2r\epsilon^2}{C^2}}, \qquad (22a)
\end{aligned}$$

where (22a) comes from inequality (20). And inequality (20) means that

$$\frac{\text{The maximum of }\mathcal{M}(P_i,Q)}{\text{The average of }\mathcal{M}(P_i,Q)}\le C.$$

■

Remark 12 *Let* $x=\frac{1}{m}\sum_{P_l\in\mathbb{P}}\mathcal{M}(P_l,Q)$ *and* $X=\max_{P_l\in\mathbb{P}}\mathcal{M}(P_l,Q)$. *For example, let* $m=10000$, $\mathcal{M}(P_1,Q)=10000$ *and* $\mathcal{M}(P_l,Q)=0.0001$ *for all* $P_l\in\mathbb{P}-\{P_1\}$.

$$x=\frac{1}{m}\sum_{P_l\in\mathbb{P}}\mathcal{M}(P_l,Q)=\frac{1}{10000}\times\left(10000+\overbrace{0.0001+\ldots+0.0001}^{9999}\right)\approx1.0001,$$

$$X=\max_{P_l\in\mathbb{P}}\mathcal{M}(P_l,Q)=10000.$$

The smaller the value of $C=\frac{X}{x}\le m$, *the better the result. However, from the above example, we know the result can be very bad. While* $T\le8(\alpha+1)+4\alpha+16$ *in sensitivity-based sampling may be much smaller than* C *in uniform sampling when* m *is big.*

$T\le8(\alpha+1)+4\alpha+16$ is a constant, while C can be close to the dataset size m in some extreme situations. Sensitivity-based sampling has a better bound than uniform sampling when $T<C$. So when m is big, sensitivity-based sampling can offer a better bound in theory.

4. Extensions of static cases

Lemma 13 (Lemma 3 (with s^{th} norm)) *Given three nonnegative weighted nonempty finite point sets* $P_1,P_2,P_3\subset\mathbb{R}^d$, *where each point set has the total weights equal to* 1. *Then for any real number* $s\in[1,+\infty)$,

$$\mathcal{W}_s^s(P_1,P_2)\le2^{s-1}\mathcal{W}_s^s(P_1,P_3)+2^{s-1}\mathcal{W}_s^s(P_3,P_2). \qquad (23)$$

Proof

$$
\begin{aligned}
\mathcal{W}_s^s(P_1,P_2) &= (\mathcal{W}_s(P_1,P_2))^s \\
&\le (\mathcal{W}_s(P_1,P_3)+\mathcal{W}_s(P_2,P_3))^s \\
&= \left(\frac{2\mathcal{W}_s(P_1,P_3)+2\mathcal{W}_s(P_2,P_3)}{2}\right)^s \\
&\le \frac{2^s\mathcal{W}_s^s(P_1,P_3)+2^s\mathcal{W}_s^s(P_2,P_3)}{2} \qquad (24a)\\
&= 2^{s-1}\mathcal{W}_s^s(P_1,P_3)+2^{s-1}\mathcal{W}_s^s(P_3,P_2),
\end{aligned}
$$

where (24a) comes from the property of Jensen's inequality for convex function; that is, $f(x)=x^s, x\in[0,+\infty)$ is a convex function when $s\in[1,+\infty)$. ■

Lemma 14 (Lemma 6 with s^{th} norm) *Let $P_1,P_2,P_3\subset\mathbb{R}^d$ be three nonnegative weighted nonempty finite point sets, where each point set has the total weights equal to 1. Then for any real numbers $\epsilon>0$, $s\in[2,+\infty)\cup\{1\}$,*

$$
|\mathcal{W}_s^s(P_1,P_2)-\mathcal{W}_s^s(P_1,P_3)| \le (1+\frac{1}{\epsilon})^{s-1}\mathcal{W}_s^s(P_2,P_3)+\left(\frac{2^{s-1}+(2\epsilon)^{s-1}}{2}-1\right)\mathcal{W}_s^s(P_1,P_2). \qquad (25)
$$

Proof

First, we consider the case $\mathcal{W}_s^s(P_1,P_2)\ge\mathcal{W}_s^s(P_1,P_3)$. For any $\epsilon>0$ and $s\in[2,+\infty)\cup\{1\}$, we have

$$
\begin{aligned}
&|\mathcal{W}_s^s(P_1,P_2)-\mathcal{W}_s^s(P_1,P_3)| \\
=&\mathcal{W}_s^s(P_1,P_2)-\mathcal{W}_s^s(P_1,P_3) \\
=&(\mathcal{W}_s(P_1,P_2))^s-\mathcal{W}_s^s(P_1,P_3) \\
\le&(1+\epsilon)^{s-1}(\mathcal{W}_s(P_1,P_3))^s+\left(1+\frac{1}{\epsilon}\right)^{s-1}(\mathcal{W}_s(P_1,P_2)-\mathcal{W}_s(P_1,P_3))^s-\mathcal{W}_s^s(P_1,P_3) \qquad (26a)\\
\le&(1+\epsilon)^{s-1}\mathcal{W}_s^s(P_1,P_3)+\left(1+\frac{1}{\epsilon}\right)^{s-1}\mathcal{W}_s^s(P_2,P_3)-\mathcal{W}_s^s(P_1,P_3) \\
\le&\left(\frac{2^{s-1}+(2\epsilon)^{s-1}}{2}\right)\mathcal{W}_s^s(P_1,P_3)+\left(1+\frac{1}{\epsilon}\right)^{s-1}\mathcal{W}_s^s(P_2,P_3)-\mathcal{W}_s^s(P_1,P_3) \qquad (26b)\\
=&\left(\frac{2^{s-1}+(2\epsilon)^{s-1}}{2}-1\right)\mathcal{W}_s^s(P_1,P_3)+\left(1+\frac{1}{\epsilon}\right)^{s-1}\mathcal{W}_s^s(P_2,P_3) \\
\le&(1+\frac{1}{\epsilon})^{s-1}\mathcal{W}_s^s(P_2,P_3)+\left(\frac{2^{s-1}+(2\epsilon)^{s-1}}{2}-1\right)\mathcal{W}_s^s(P_1,P_2), \qquad (26c)
\end{aligned}
$$

where (26a) comes from Lemma 6, and (26b) comes from the property of Jensen's inequality for convex function; that is, $f(x)=x^{s-1}, x\in[0,+\infty)$ is a convex function when $s\in[2,+\infty)\cup\{1\}$. For this case, we assume $\mathcal{W}_s^s(P_1,P_2)\ge\mathcal{W}_s^s(P_1,P_3)$, thus (26c) holds.

For the other case $\mathcal{W}_s^s(P_1,P_2) < \mathcal{W}_s^s(P_1,P_3)$, we directly have

$$\begin{aligned}
&|\mathcal{W}_s^s(P_1,P_2) - \mathcal{W}_s^s(P_1,P_3)| \\
=&\mathcal{W}_s^s(P_1,P_3) - \mathcal{W}_s^s(P_1,P_2) \\
\le&(1+\frac{1}{\epsilon})^{s-1}\mathcal{W}_s^s(P_2,P_3) + \left(\frac{2^{s-1}+(2\epsilon)^{s-1}}{2} - 1\right)\mathcal{W}_s^s(P_1,P_2)
\end{aligned} \tag{27}$$

by exchanging the roles of P_2 and P_3 in Eq.(26). ■

5. Extensions of dynamic cases

Lemma 15 (Lemma 3) *Given three nonnegative weighted nonempty finite point sets $P_1, P_2, P_3 \subset \mathbb{R}^d$, where each point set has the total weights equal to 1. Then for any real number $s \in [1,+\infty)$*

$$\mathcal{WRT}_s^s(P_1,P_2) \le 2^{s-1}\mathcal{WRT}_s^s(P_1,P_3) + 2^{s-1}\mathcal{WRT}_s^s(P_3,P_2). \tag{28}$$

Proof W.l.o.g., we assume that the induced rigid transformations of $\mathcal{WRT}_s^s(P_1,P_3)$, $\mathcal{WRT}_s^s(P_3,P_2)$ are $e_1, e_2 \in SE(d)$, respectively. That is,

$$\begin{aligned}
\mathcal{WRT}_s^s(P_1,P_3) = \mathcal{W}_s^s(e_1(P_1),P_3), \\
\mathcal{WRT}_s^s(P_2,P_3) = \mathcal{W}_s^s(e_2(P_2),P_3).
\end{aligned} \tag{29}$$

Thus,

$$\begin{aligned}
&\mathcal{WRT}_s^s(P_1,P_2) \\
\le&\mathcal{W}_s^s\left(e_1(P_1),e_2(P_2)\right) \\
=&\left(\mathcal{W}_s\left(e_1(P_1),e_2(P_2)\right)\right)^s \\
\le&\left(\mathcal{W}_s\left(e_1(P_1),P_3\right) + \mathcal{W}_s\left(P_3,e_2(P_2)\right)\right)^s \\
=&\left(\mathcal{WRT}_s\left(P_1,P_3\right) + \mathcal{WRT}_s\left(P_3,P_2\right)\right)^s \\
=&\left(\frac{2\mathcal{WRT}_s\left(P_1,P_3\right) + 2\mathcal{WRT}_s\left(P_3,P_2\right)}{2}\right)^s \\
\le&\frac{2^s\mathcal{WRT}_s^s\left(P_1,P_3\right) + 2^s\mathcal{WRT}_s^s\left(P_3,P_2\right)}{2} \qquad \text{(30a)} \\
=&2^{s-1}\mathcal{WRT}_s^s\left(P_1,P_3\right) + 2^{s-1}\mathcal{WRT}_s^s\left(P_3,P_2\right),
\end{aligned}$$

where (30a) comes from the property of Jensen's inequality for convex function; that is, $f(x) = x^s, x \in [0,+\infty)$ is a convex function when $s \in [1,+\infty)$. ■

Lemma 16 (Lemma 6) *Let $P_1, P_2, P_3 \subset \mathbb{R}^d$ be three nonnegative weighted nonempty finite point sets, where each point set has the total weights equal to 1. Then for any real numbers*

$\epsilon > 0,\ s \in [2, +\infty) \cup \{1\}$,

$$|\mathcal{WRT}_s^s(P_1, P_2) - \mathcal{WRT}_s^s(P_1, P_3)| \le \left(1 + \frac{1}{\epsilon}\right)^{s-1} \mathcal{WRT}_s^s(P_2, P_3) + \left(\frac{2^{s-1} + (2\epsilon)^{s-1}}{2} - 1\right) \mathcal{WRT}_s^s(P_1, P_2). \tag{31}$$

Proof W.l.o.g., we assume that the induced rigid transformations of $\mathcal{WRT}_s^s(P_1, P_3)$, $\mathcal{WRT}_s^s(P_3, P_2)$ are $e_1, e_2 \in SE(d)$, respectively. That is,

$$\begin{aligned} \mathcal{WRT}_s^s(P_1, P_3) &= \mathcal{W}_s^s(e_1(P_1), P_3), \\ \mathcal{WRT}_s^s(P_2, P_3) &= \mathcal{W}_s^s(e_2(P_2), P_3). \end{aligned} \tag{32}$$

First, we consider the case $\mathcal{WRT}_s^s(P_1, P_2) \ge \mathcal{WRT}_s^s(P_1, P_3)$. Then we have

$$\begin{aligned} &|\mathcal{WRT}_s^s(P_1, P_2) - \mathcal{WRT}_s^s(P_1, P_3)| \\ =&\mathcal{WRT}_s^s(P_1, P_2) - \mathcal{WRT}_s^s(P_1, P_3) \\ \le&\mathcal{W}_s^s\left(e_1(P_1), e_2(P_2)\right) - \mathcal{WRT}_s^s(P_1, P_3) \\ \le&\left(\mathcal{W}_s\left(e_1(P_1), P_3\right) + \mathcal{W}_s\left(P_3, e_2(P_2)\right)\right)^s - \mathcal{WRT}_s^s(P_1, P_3) \\ =&\left(\mathcal{WRT}_s(P_1, P_3) + \mathcal{WRT}_s(P_3, P_2)\right)^s - \mathcal{WRT}_s^s(P_1, P_3) \\ \le&\left(1 + \frac{1}{\epsilon}\right)^{s-1} \mathcal{WRT}_s^s(P_2, P_3) + (1+\epsilon)^{s-1} \mathcal{WRT}_s^s(P_1, P_3) - \mathcal{WRT}_s^s(P_1, P_3) \qquad (33a) \\ \le&\left(1 + \frac{1}{\epsilon}\right)^{s-1} \mathcal{WRT}_s^s(P_2, P_3) + \left(\frac{2^{s-1} + (2\epsilon)^{s-1}}{2}\right) \mathcal{WRT}_s^s(P_1, P_3) - \mathcal{WRT}_s^s(P_1, P_3) \qquad (33b) \\ =&\left(1 + \frac{1}{\epsilon}\right)^{s-1} \mathcal{WRT}_s^s(P_2, P_3) + \left(\frac{2^{s-1} + (2\epsilon)^{s-1}}{2} - 1\right) \mathcal{WRT}_s^s(P_1, P_3) \\ \le&\left(1 + \frac{1}{\epsilon}\right)^{s-1} \mathcal{WRT}_s^s(P_2, P_3) + \left(\frac{2^{s-1} + (2\epsilon)^{s-1}}{2} - 1\right) \mathcal{WRT}_s^s(P_1, P_2), \qquad (33c) \end{aligned}$$

where (33a) comes from Lemma 6, and (33b) comes from the property of Jensen's inequality for convex function; that is, $f(x) = x^{s-1}, x \in [0, +\infty)$ is a convex function when $s \in [2, +\infty) \cup \{1\}$. For this case, we assume $\mathcal{WRT}_s^s(P_1, P_2) \ge \mathcal{WRT}_s^s(P_1, P_3)$, thus (33c) holds.

For the other case $\mathcal{WRT}_s^s(P_1, P_2) < \mathcal{WRT}_s^s(P_1, P_3)$, we directly have

$$\begin{aligned} &|\mathcal{WRT}_s^s(P_1, P_2) - \mathcal{WRT}_s^s(P_1, P_3)| \\ =&\mathcal{WRT}_s^s(P_1, P_3) - \mathcal{WRT}_s^s(P_1, P_2) \\ \le&\left(1 + \frac{1}{\epsilon}\right)^{s-1} \mathcal{WRT}_s^s(P_2, P_3) + \left(\frac{2^{s-1} + (2\epsilon)^{s-1}}{2} - 1\right) \mathcal{WRT}_s^s(P_1, P_2) \end{aligned} \tag{34}$$

by exchanging the roles of P_2 and P_3 in Eq.(33). ■

Remark 17 *For both "static" cases and "dynamic" cases, the conclusions for $1^{st}, 2^{nd}$ are special cases of s^{th} norm, when $s = 1, s = 2$, respectively.*

6. Experiments

We run DWB on

- synthetic datasets: Gaussian dataset, von Mises–Fisher dataset, and
- real-world datasets: MNIST, Human Connectome Project (HCP).

We run DWB-RT on

- synthetic dataset: von Mises–Fisher dataset, and
- real-world datasets: ModelNet40, MNIST, Human Connectome Project (HCP).

All of the experimental results were obtained on a Linux workstation with 3.80GHz Intel(R) Core(TM) i7-10700K CPU and 64GB Memory; the algorithms are implemented in Python3.8.

Gaussian dataset: To construct an instance of ensemble clustering, we generate a synthetic dataset of 2000 points randomly sampled from $k = 50$ Gaussian distributions from $\mathbb{R}^{100}$. We run k-means clustering 1000 times, where each time has a different random initialization for k center points, to generate 1000 different clustering solutions. According to the model introduced by (Ding et al. (2016)), each instance is a geometric prototype problem with 1000 different 50-point sets in $\mathbb{R}^{2000}$. It can also be seen as a special case of DWB problem with the same weight. That is, for static case, it is a DWB problem with $m = 1000, k = 50, d = 100$.

ModelNet40: ModelNet40 (Wu et al. (2015)) is a comprehensive clean collection of $3D$ CAD models. We choose 5 kinds of objects: airplane, bed, bookshelf, chair and vase. And for each object, 500 CAD models were selected as our original input dataset. First, we convert these CAD models into point clouds. And then each point cloud was grouped into $k = 300$ clusters; each cluster was represented by its cluster center; the weight of each center is proportional to the total number of points of the cluster. Then, only considering dynamic case, we can get DWB-RT problems with $m = 500, k = 300, d = 3$.

MNIST: MNIST(LeCun et al. (1998)) is a popular handwritten benchmark with digits from 0 to 9. For each digit, we select 3000 28×28 grayscale images. First, for each image, its 28×28 pixels were represented by 60 weighted $2D$ points via k-means clustering (Pedregosa et al. (2011); Lloyd (1982)): group the pixels into 60 clusters and represent each cluster with its cluster center; the weight of each center is proportional to the total pixel value of the cluster. Then, for static and dynamic cases, we can get DWB and DWB-RT problems with $m = 3000, k = 60, d = 2$, respectively.

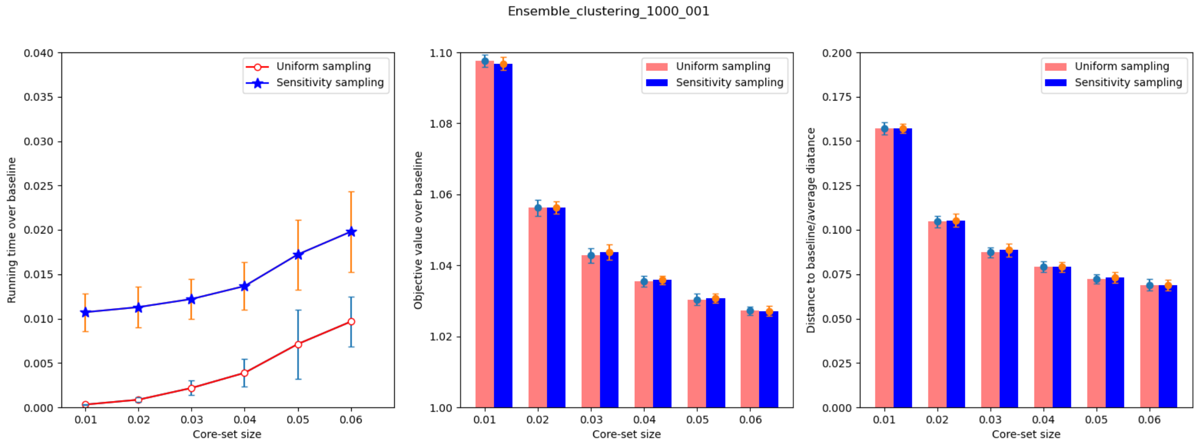
von Mises–Fisher dataset: The von Mises–Fisher distribution (Fisher (1953)) is a probability distribution on the $(p-1)$-sphere in $\mathbb{R}^p$. Here, we generate a synthetic dataset of 1000 Mises–Fisher distributions on the 2-sphere in $\mathbb{R}^3$ randomly, where the support size of each distribution is 100. Then, for static and dynamic cases, we aim at computing the DWB and DWB-RT of the 1000 Mises–Fisher distributions with $m = 1000, k = 100, d = 3$, respectively.

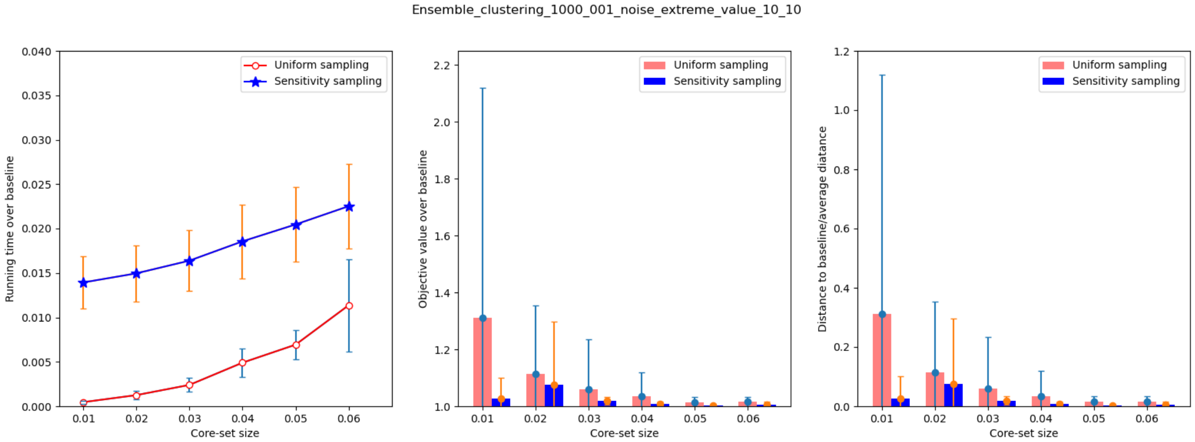
Human Connectome Project (HCP): Human Connectome Project (HCP) (Van Essen et al. (2013)) is a dataset of high-quality neuroimaging data in over 1100 healthy young adults, aged 22–35. We select one subject randomly, and took 500 3D brain images of this subject. Then, for static and dynamic cases, we aim at computing DWB and DWB-RT of these 500 3D brain images with $m = 500, k = 300, d = 3$, respectively.

For all these applications, we construct the core-set using both uniform sampling and sensitivity-based sampling(Ding and Liu (2018)) method. In our algorithms, the DWD is computed by method in (Bonneel et al. (2011)), and the DWB is calculated by algorithms in (Cuturi and Doucet (2014); Álvarez-Esteban et al. (2016)). For dynamic cases, the DWD-RT is computed by method in (Cohen and Guibasm (1999)), and the DWB-RT is calculated by algorithms in (Ding and Xu (2014)).

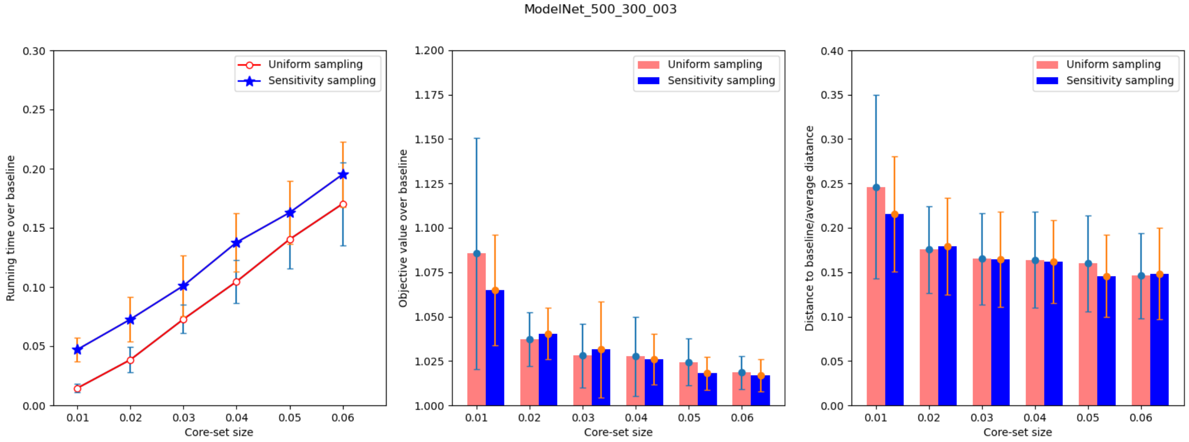
For each application, we consider three criteria: running time, objective value, and difference to ground truth. For all three applications, we use the DWB/DWB-RT of the original input dataset as the ground truth; then we compute its matching cost to the DWB/DWB-RT of core-set, denoted by x, as well as the average matching cost over the original input dataset to the ground truth, denoted by Ave; finally, we obtain the ratio x/Ave. In general, the lower the ratio x/Ave, the closer the obtained discrete Wasserstein barycenter to the ground truth.

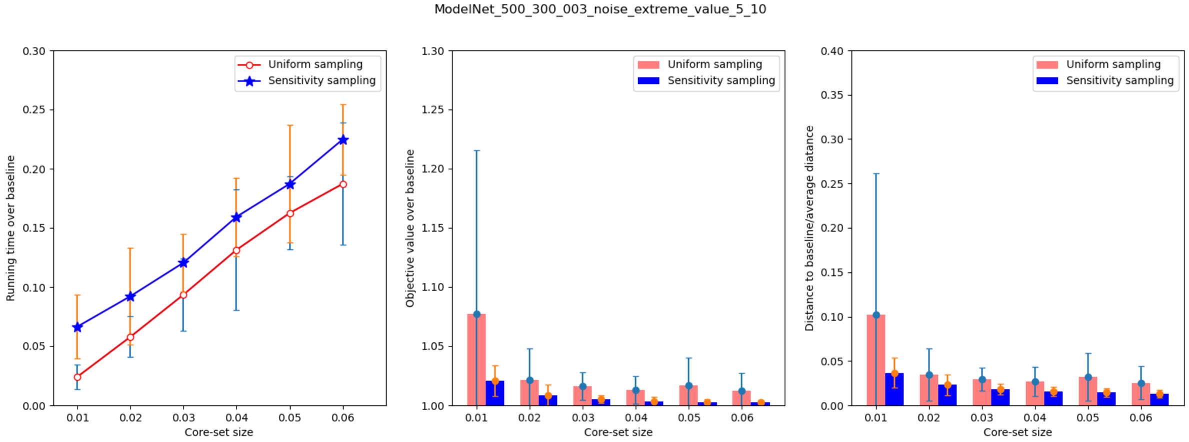
For experiments on Gaussian database, we vary the core-set size from 1% to 6% of the input size.



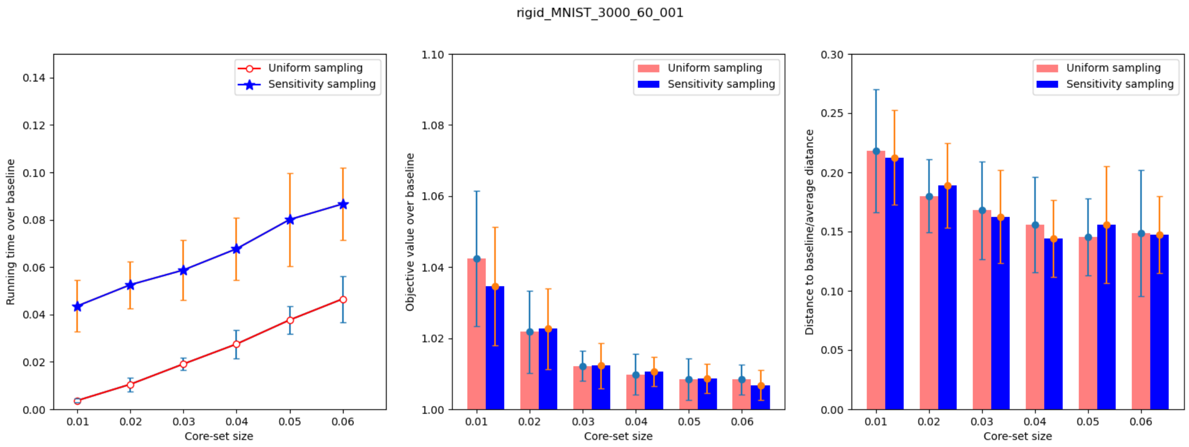


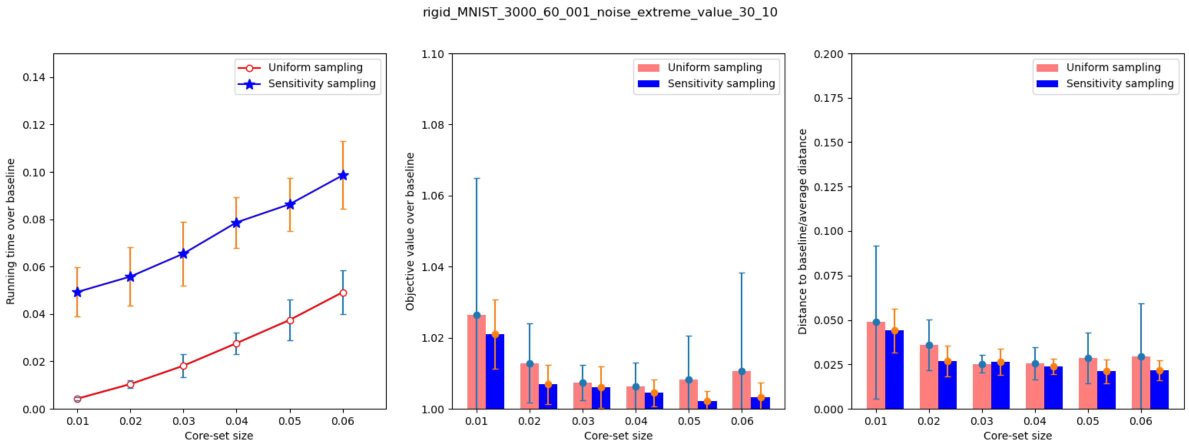
For experiments on ModelNet40, we vary the core-set size from 3% to 18% of the input size.

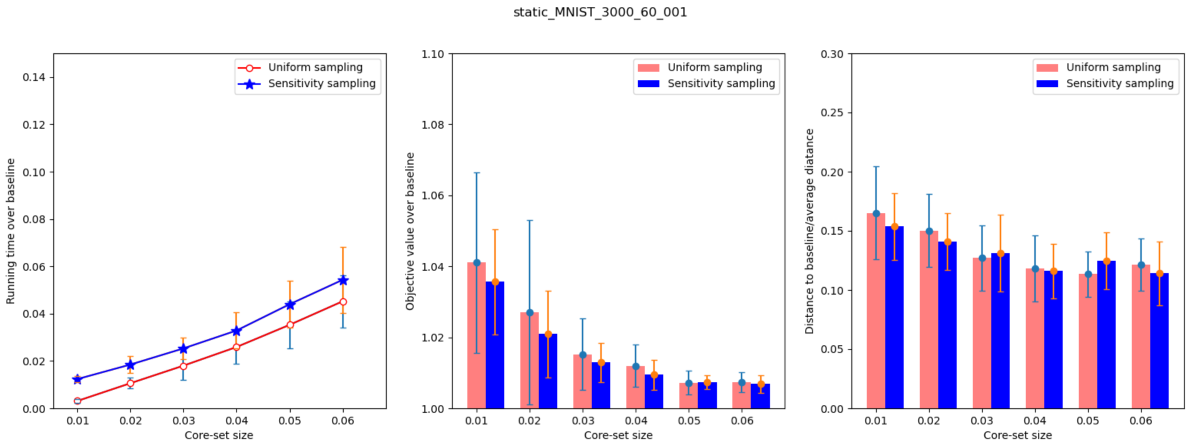


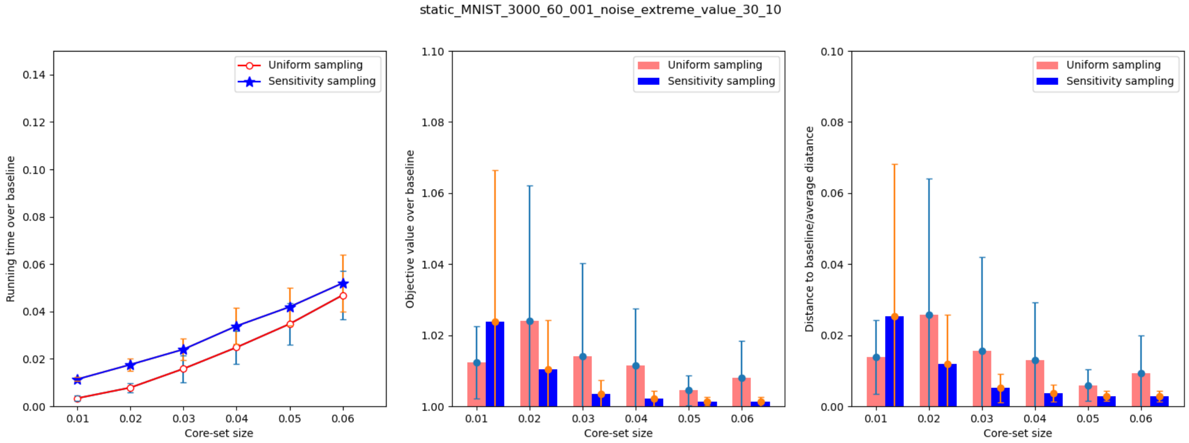


For experiments on MNIST, we vary the core-set size from 1% to 6% of the input size.

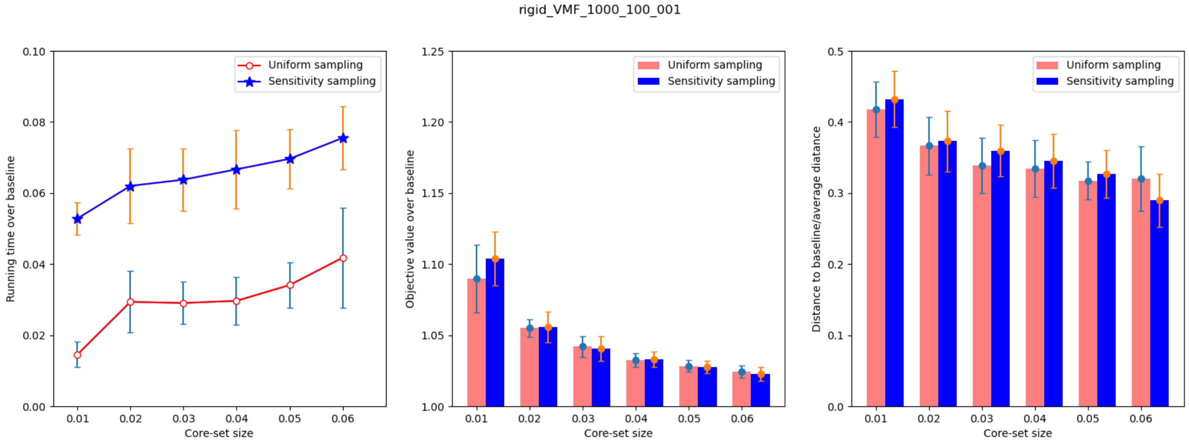

rigid_MNIST_3000_60_001
Uniform sampling
Sensitivity sampling
Running time over baseline
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Objective value over baseline
1.00
1.02
1.04
1.06
1.08
1.10
Distance to baseline/average diatance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.01
0.02
0.03
0.04
0.05
0.06
Core-set size

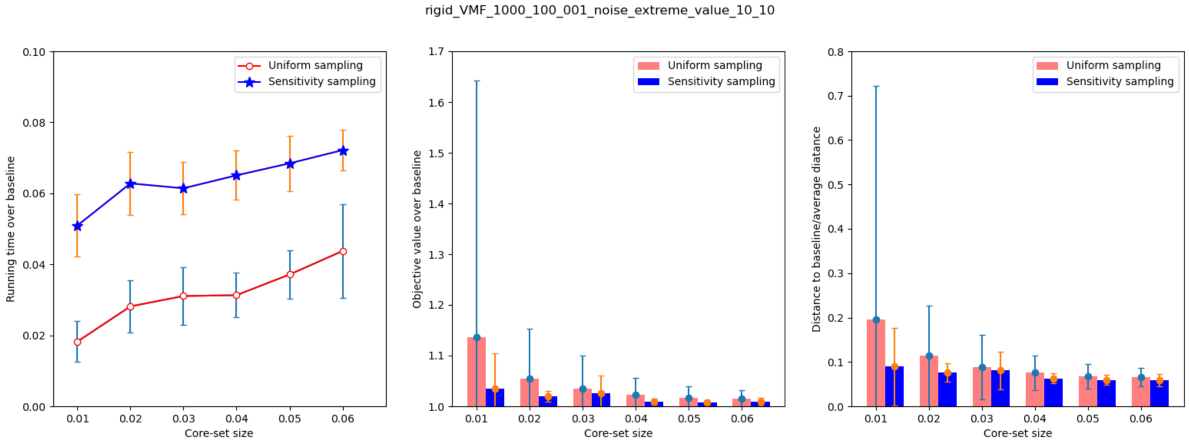

rigid_MNIST_3000_60_001_noise_extreme_value_30_10
Uniform sampling
Sensitivity sampling
Running time over baseline
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Objective value over baseline
1.00
1.02
1.04
1.06
1.08
1.10
Distance to baseline/average diatance
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
0.01
0.02
0.03
0.04
0.05
0.06
Core-set size

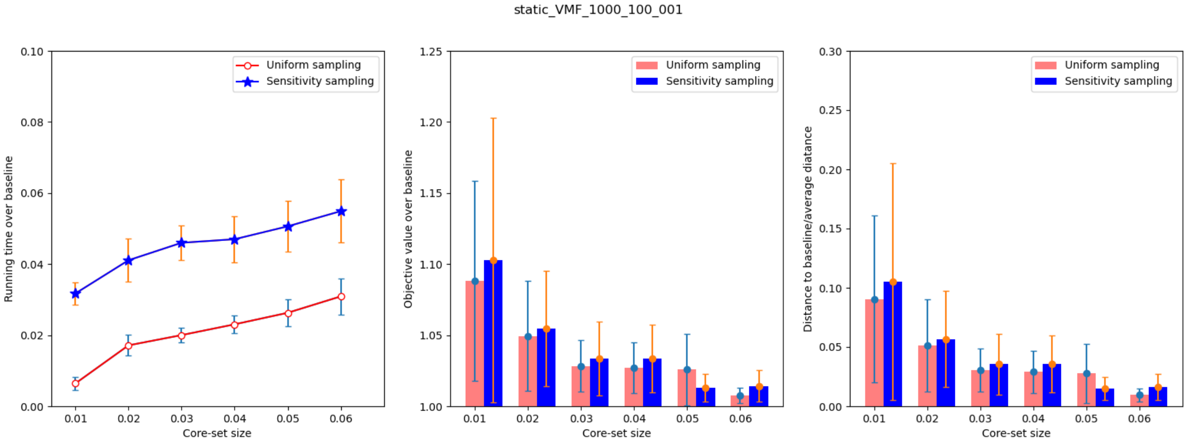

static_MNIST_3000_60_001
Uniform sampling
Sensitivity sampling
Running time over baseline
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Objective value over baseline
1.00
1.02
1.04
1.06
1.08
1.10
Distance to baseline/average diatance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.01
0.02
0.03
0.04
0.05
0.06
Core-set size

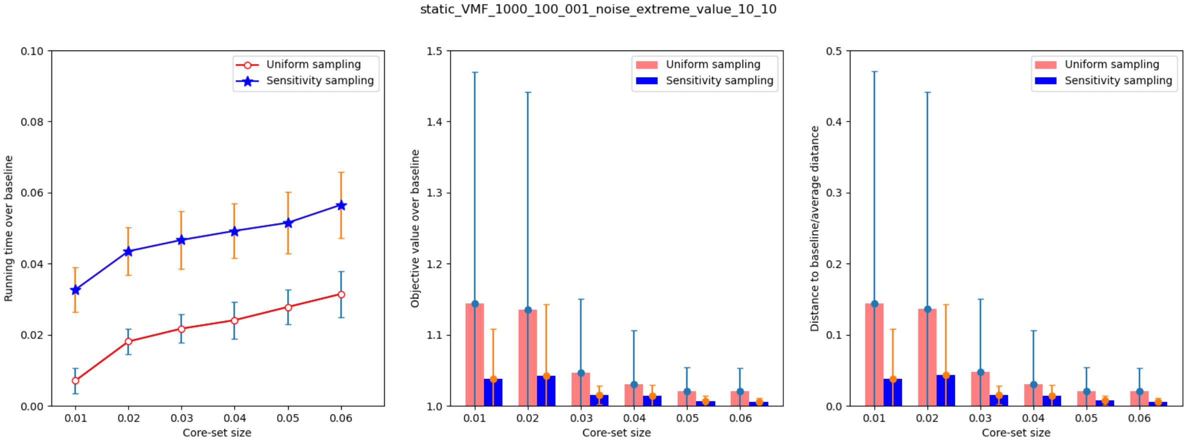


For experiments on Mises–Fisher distribution, we vary the core-set size from 1% to 6% of the input size.

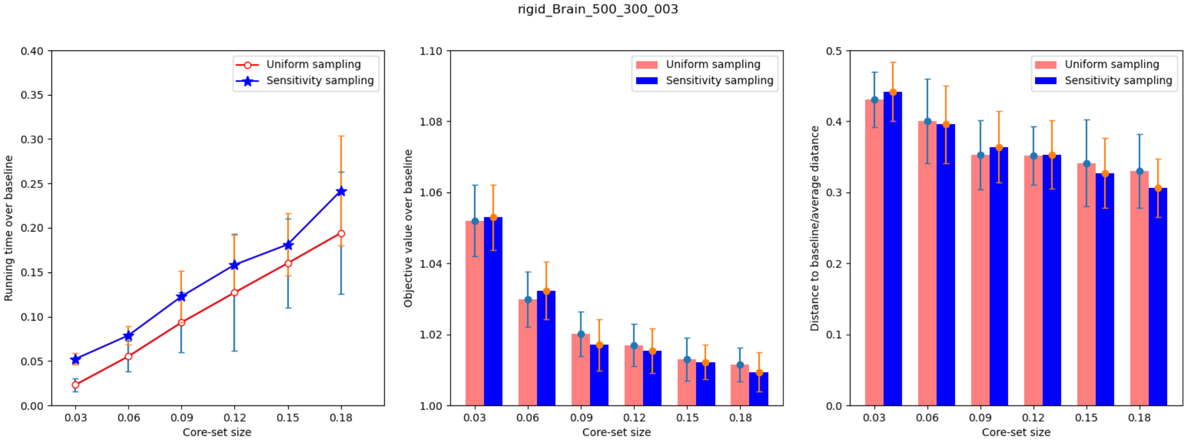


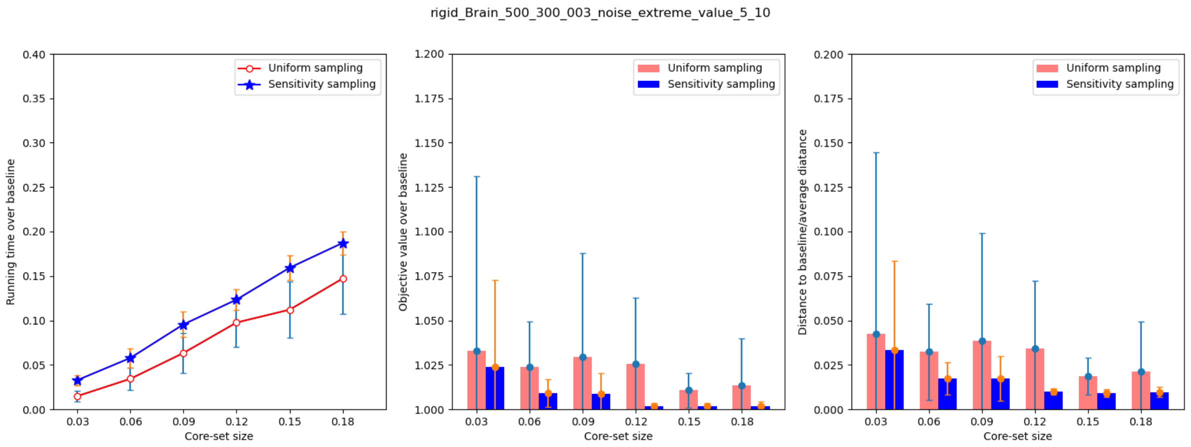


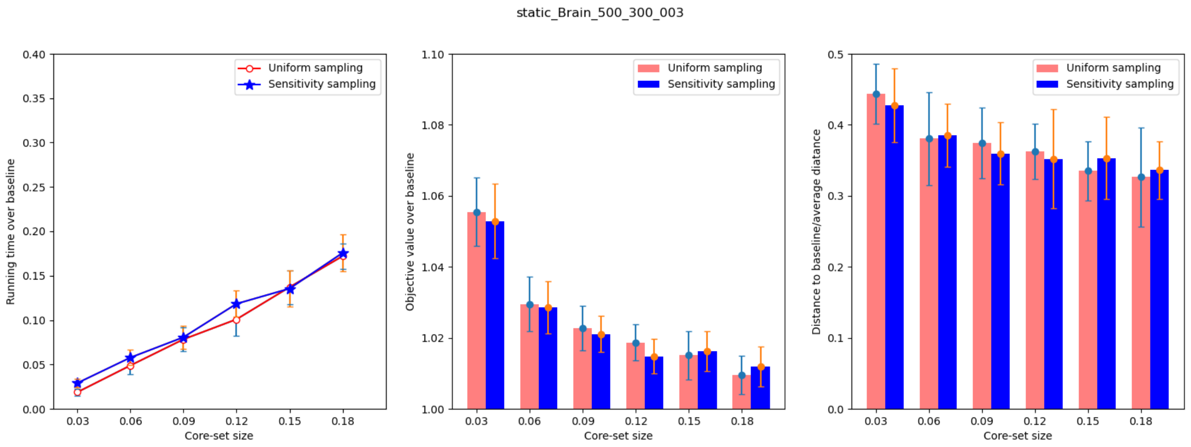


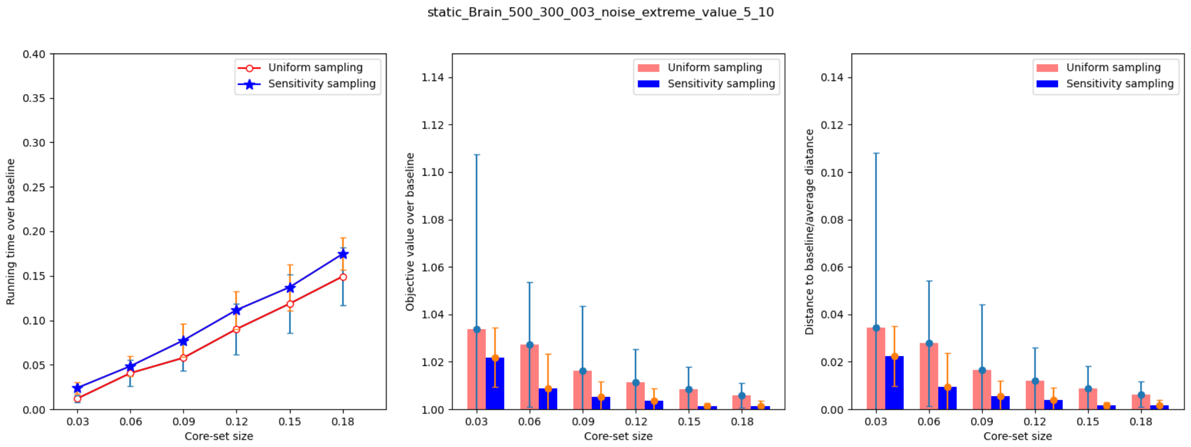


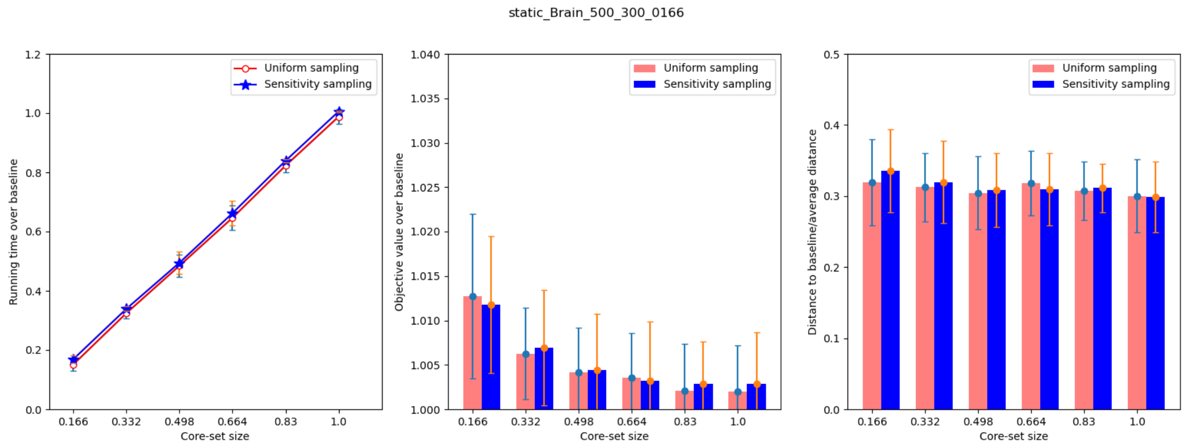
For experiments on HCP, first, for both static and rigid cases, we vary the core-set size from 3% to 18% of the input size. Then, for static case, we vary the core-set size from 16.6% to 100% of the input size.



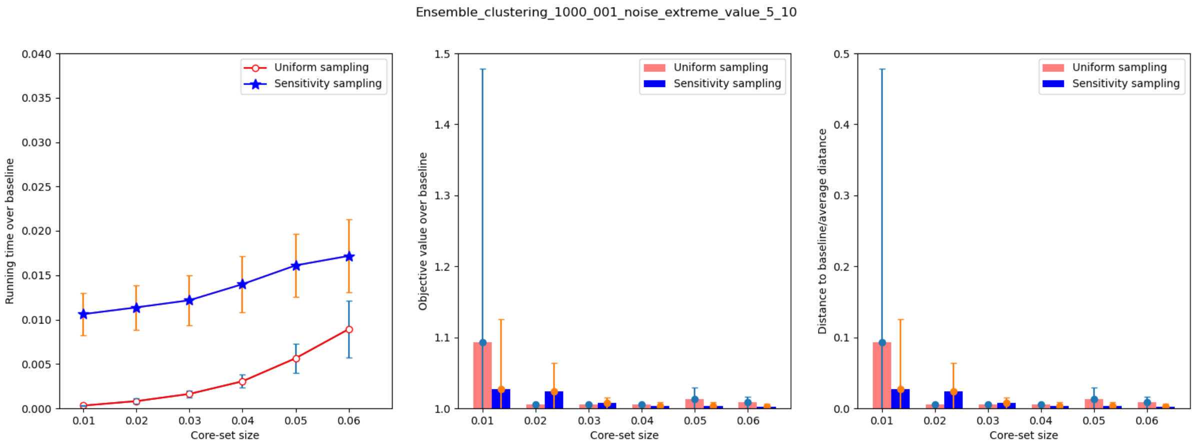

rigid_Brain_500_300_003_noise_extreme_value_5_10


static_Brain_500_300_003

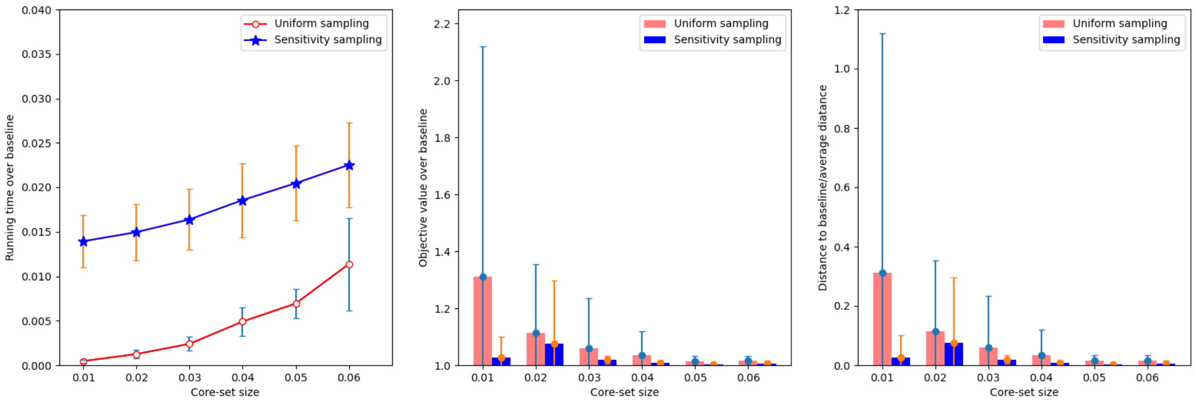

static_Brain_500_300_003_noise_extreme_value_5_10



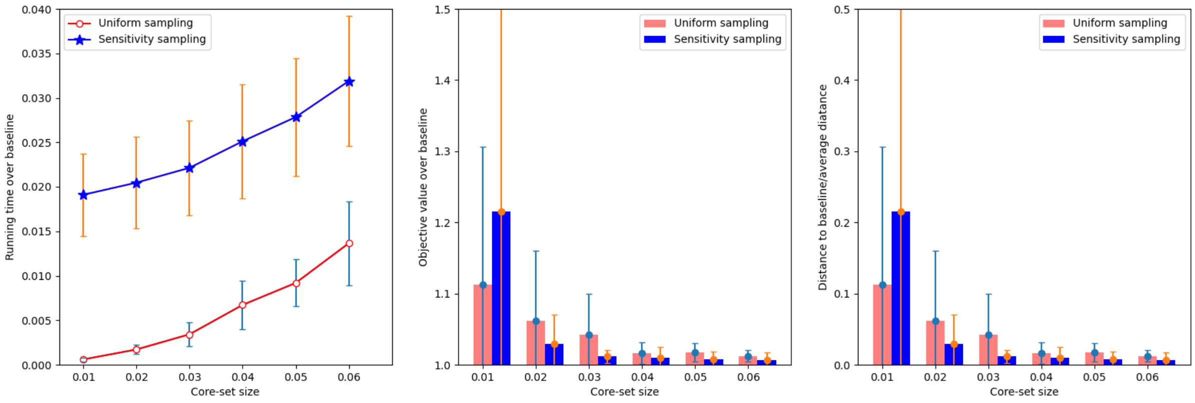
Next, we add 0.5%, 1.0%, 1.5%, 2.0%, 2.5%, 3.0% noise for ensemble clustering.



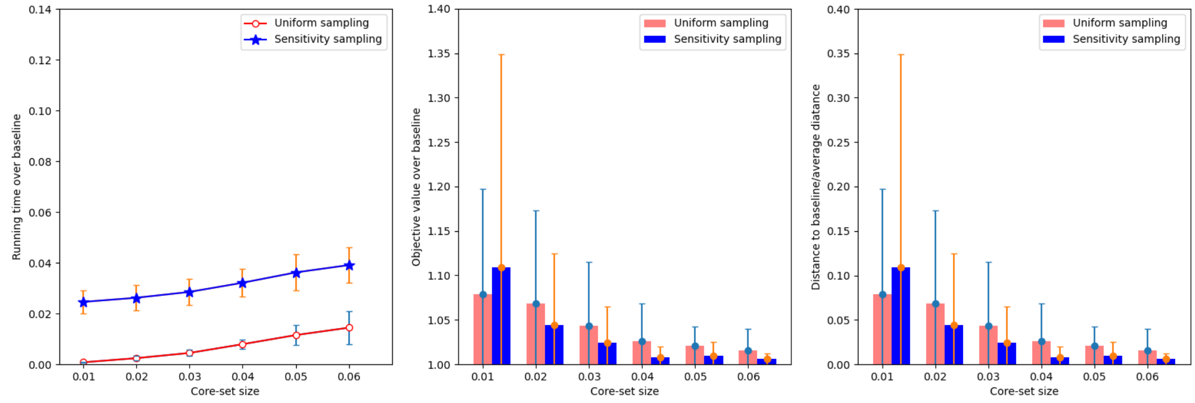
Ensemble_clustering_1000_001_noise_extreme_value_10_10



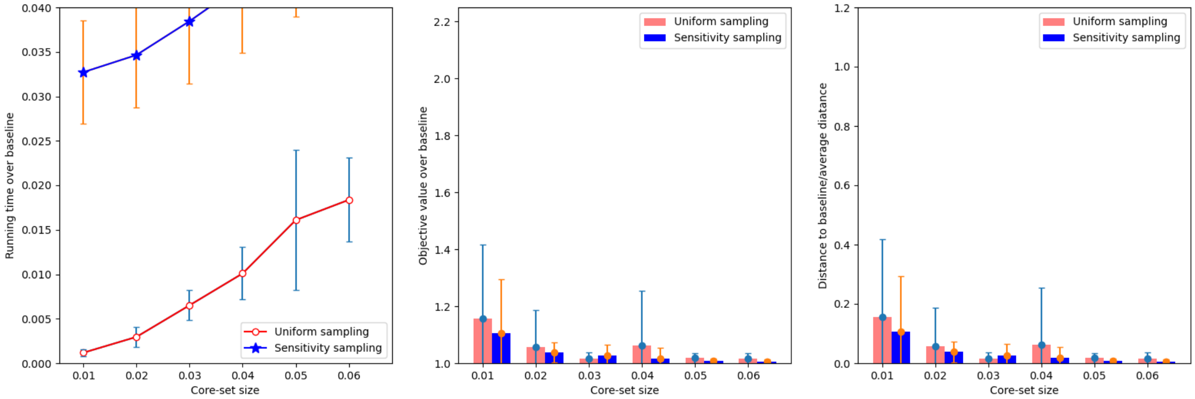
Ensemble_clustering_1000_001_noise_extreme_value_15_10



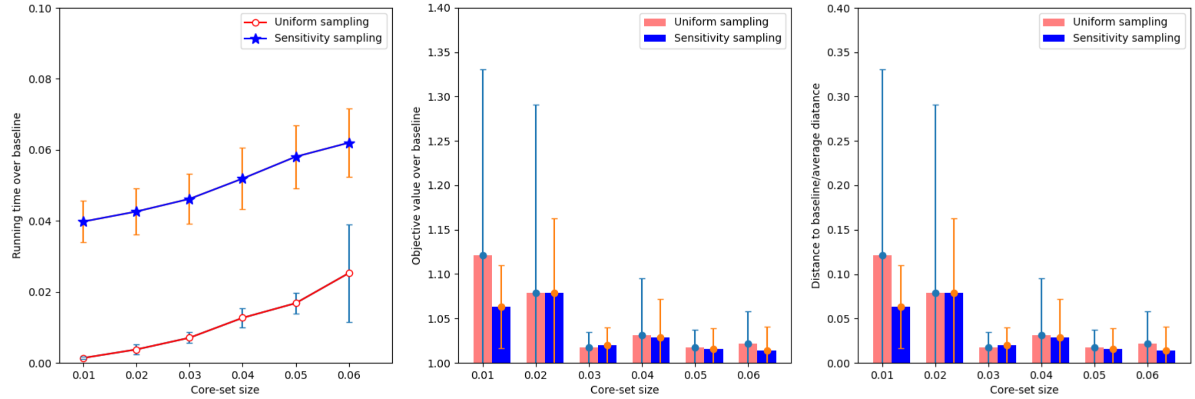
Ensemble_clustering_1000_001_noise_extreme_value_20_10



Ensemble_clustering_1000_001_noise_extreme_value_25_10



Ensemble_clustering_1000_001_noise_extreme_value_30_10



Results For each application, we run 20 trials and record the average results.

7. References

MNIST(LeCun et al. (1998))
Python scikit-learn (Pedregosa et al. (2011))
k-means clustering (Lloyd (1982))
ot.emd (Bonneel et al. (2011))
ot.lp.free-support-barycenter (Cuturi and Doucet (2014); Álvarez-Esteban et al. (2016))
python pot (Flamary et al. (2021))
ModelNet40 (Wu et al. (2015))
Mises–Fisher distribution (Fisher (1953))
HCP (Van Essen et al. (2013))

acknowledgment ”Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH

Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University."

References

Pedro C Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–12, 2011.

Scott Cohen and L Guibasm. The earth mover's distance under transformation sets. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1076–1083. IEEE, 1999.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

Hu Ding and Manni Liu. On geometric prototype and applications. *arXiv preprint arXiv:1804.09655*, 2018.

Hu Ding and Jinhui Xu. Finding median point-set using earth mover's distance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

Hu Ding, Lu Su, and Jinhui Xu. Towards distributed ensemble clustering for networked sensing systems: a novel geometric approach. In *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pages 1–10, 2016.

Ronald Aylmer Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305, 1953.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL `http://jmlr.org/papers/v22/20-451.html`.

Michael Langberg and Leonard J Schulman. Universal ε-approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 598–607. SIAM, 2010.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12: 2825–2830, 2011.

Y Rubner, C Tomasi, and L Guibas. The earth mover's distance as a metric for image. Technical report, Technical Report STAN-CS-TN-98-86, Computer Science Department, Stanford, 1998.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Kasturi Varadarajan and Xin Xiao. A near-linear algorithm for projective clustering integer points. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1329–1342. SIAM, 2012.

Wikipedia contributors. Rigid transformation — Wikipedia, the free encyclopedia, 2021. URL `https://en.wikipedia.org/w/index.php?title=Rigid_transformation&oldid=1039406690`. [Online; accessed 19-September-2021].

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.