References

[1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011. 3

[2] David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020. 2, 3, 4

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005. 1

[4] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017. 2

[5] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014. 3, 6

[6] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020. 2

[7] R Flamary, N Courty, D Tuia, and A Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 1:1–40, 2016. 2

[8] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019. 1, 2, 6, 4, 5

[9] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. *Advances in neural information processing systems*, 31, 2018. 2

[10] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas J Spanos, and Dawn Song. Efficient task-specific data valuation for nearest neighbor algorithms. *arXiv preprint arXiv:1908.08619*, 2019. 2

[11] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR, 2019. 1, 2

[12] Kevin Fu Jiang, Weixin Liang, James Zou, and Yongchan Kwon. Opendataval: a unified benchmark for data valuation. *arXiv preprint arXiv:2306.10577*, 2023. 3, 7, 4

[13] Hoang Anh Just, Feiyang Kang, Jiachen T Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. Lava: Data valuation without pre-specified learning algorithms. *arXiv preprint arXiv:2305.00054*, 2023. 1, 2, 4, 5, 6, 8

[14] Feiyang Kang, Hoang Anh Just, Anit Kumar Sahu, and Ruoxi Jia. Performance scaling via optimal transport: Enabling data selection from partially revealed sources. *arXiv preprint arXiv:2307.02460*, 2023. 2

[15] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017. 1

[16] Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. *arXiv preprint arXiv:2110.14049*, 2021. 2

[17] Yongchan Kwon and James Zou. Data-oob: out-of-bag estimate as a simple and efficient data value. In *International Conference on Machine Learning*, pages 18135–18152. PMLR, 2023. 3

[18] Yongchan Kwon, Manuel A Rivas, and James Zou. Efficient computation and analysis of distributional shapley values. In *International Conference on Artificial Intelligence and Statistics*, pages 793–801. PMLR, 2021. 2

[19] Jinkun Lin, Anqi Zhang, Mathias Lécuyer, Jinyang Li, Aurojit Panda, and Siddhartha Sen. Measuring the effect of training data on deep learning predictions via randomized experiments. In *International Conference on Machine Learning*, pages 13468–13504. PMLR, 2022. 2

[20] Zelei Liu, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui. Gtg-shapley: Efficient and accurate participant contribution evaluation in federated learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–21, 2022. 1, 2, 6, 4, 5

[21] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 8

[22] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019. 4

[23] Eduardo Fernandes Montesuma and Fred Maurice Ngole Mboula. Wasserstein barycenter for multi-source domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16785–16793, 2021. 2, 4

[24] Brendan Pass. Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 49(6):1771–1790, 2015. 3

[25] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 2, 1

[26] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020. 2

[27] Alain Rakotomamonjy, Rémi Flamary, Gilles Gasso, M El Alaya, Maxime Berar, and Nicolas Courty. Optimal transport for conditional domain matching and label shift. *Machine Learning*, pages 1–20, 2022. 2

[28] Alain Rakotomamonjy, Kimia Nadjahi, and Liva Ralaivola. Federated wasserstein distance. *arXiv preprint arXiv:2310.01973*, 2023. 2, 3, 4, 5, 7

[29] Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedo-*

nia, September 18–22, 2017, Proceedings, Part II 10, pages 737–753. Springer, 2017. 2

[30] Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on artificial intelligence and statistics*, pages 849–858. PMLR, 2019. 2

[31] Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, and Ali Jadbabaie. Robust federated learning: The case of affine distribution shifts. *Advances in Neural Information Processing Systems*, 33:21554–21565, 2020. 4

[32] Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low. Collaborative machine learning with incentive-aware model rewards. In *International conference on machine learning*, pages 8927–8936. PMLR, 2020. 2

[33] Rachael Hwee Ling Sim, Xinyi Xu, and Bryan Kian Hsiang Low. Data valuation in machine learning:"ingredients", strategies, and open challenges. In *Proc. IJCAI*, pages 5607–5614, 2022. 1

[34] Tianshu Song, Yongxin Tong, and Shuyue Wei. Profit allocation for federated learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2577–2586. IEEE, 2019. 1, 6, 4, 5

[35] Tiffany Tuor, Shiqiang Wang, Bong Jun Ko, Changchang Liu, and Kin K Leung. Overcoming noisy and irrelevant data in federated learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5020–5027. IEEE, 2021. 2

[36] Rosanna Turrisi, Rémi Flamary, Alain Rakotomamonjy, and Massimiliano Pontil. Multi-source domain adaptation via weighted joint distributions optimal transport. In *Uncertainty in Artificial Intelligence*, pages 1970–1980. PMLR, 2022. 2

[37] Cédric Villani. *Topics in optimal transportation*. American Mathematical Soc., 2021. 2, 3

[38] Cédric Villani et al. *Optimal transport: old and new*. Springer, 2009. 2

[39] Jiachen T Wang and Ruoxi Jia. Data banzhaf: A robust data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 6388–6421. PMLR, 2023. 2

[40] Shengsheng Wang, Bilin Wang, Zhe Zhang, Ali Asghar Heidari, and Huiling Chen. Class-aware sample reweighting optimal transport for multi-source domain adaptation. *Neurocomputing*, 523:213–223, 2023. 2

[41] Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A principled approach to data valuation for federated learning. *Federated Learning: Privacy and Incentive*, pages 153–167, 2020. 6, 5

[42] Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. *Advances in Neural Information Processing Systems*, 34:16104–16117, 2021. 2

[43] Xinyi Xu, Zhaoxuan Wu, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Validation free and replication robust volume-based data valuation. *Advances in Neural Information Processing Systems*, 34:10837–10848, 2021. 2, 8

[44] Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement learning. In *International Conference on Machine Learning*, pages 10842–10851. PMLR, 2020. 3

Data Valuation and Detections in Federated Learning

Supplementary Material

7. Technical Concepts Definition

To formulate equation 3, we utilized the mathematical property of Wasserstein Distance as below.

Property 1 *(Triangle Inequality of Wasserstein Distance) For any $p \geq 1, P, Q, \gamma \in \mathcal{P}_p(X)$, $\mathcal{W}_p$ is a metric on $\mathcal{P}_p(X)$, as such it satisfies the triangle inequality as [25]*

$$\mathcal{W}_p(P, Q) \leq \mathcal{W}_p(P, \gamma) + \mathcal{W}_p(\gamma, Q), \tag{13}$$

in order to attain equality, *geodesics* and *Interpolating point* are defined as structuring tools of metric spaces.

Definition 2 *(Geodesics [3]) Let $(\mathcal{X}, d)$ be a metric space. A constant speed geodesic $x : [0,1] \to \mathcal{X}$ between $x_0, x_1 \in \mathcal{X}$ is a continuous curve such that $\forall a, b \in [0,1], d\big(x(a), x(b)\big) = |a - b| \cdot d(x_0, x_1)$.*

Definition 3 *(Interpolating point [3]) Any point x_t from a constant speed geodestic $(x(t))_{t\in[0,1]}$ is an interpolating point and verifies $d(x_0, x_1) = d(x_0, x_t) + d(x_t, x_1)$.*

The above definitions and properties are used to define the interpolating measure of the Wasserstein distance:

Definition 4 *(Wasserstein Geodesics, Interpolating measure [3, 15]) Let $P, Q \in \mathcal{P}_p(X)$ with $X \subseteq \mathbb{R}^d$ compact, convex and equipped with $\mathcal{W}_p$. Let $\pi^\star \in \Pi(P, Q)$ be an optimal transport plan between two distributions P and Q. For $t \in [0,1]$, let $\gamma_t = (\pi_t)_{\#}\pi^\star$ where $\pi_t(x, y) = (1-t)x + ty$, i.e. γ_t is the push-forward measure of $\pi^\star$ under the map π_t. Then, the curve $\bar{\mu} = (\gamma_t)_{t\in[0,1]}$ is a constant speed geodesic, also called a Wasserstein geodesics between P and Q.*

8. Proof for Theorem 1

Since the $Q^{(k)}$ is the wasserstein barycenter for all interpolating measures $\gamma_i^{(k)}$ where $i \in [1, N]$, we have

$$A^{(k)} = \sum_i^N [\mathcal{W}_p(P_i, \gamma_i^{(k)}) + \mathcal{W}_p(Q^{(k)}, \gamma_i^{(k)})] \tag{14}$$

$$= \sum_i^N \mathcal{W}_p(P_i, \gamma_i^{(k)}) + \sum_i^N \mathcal{W}_p(Q^{(k)}, \gamma_i^{(k)}) \tag{15}$$

$$\leq \sum_i^N \mathcal{W}_p(P_i, \gamma_i^{(k)}) + \sum_i^N \mathcal{W}_p(Q^{(k-1)}, \gamma_i^{(k)}) \tag{16}$$

Define the interpolating measure between $\gamma_i^{(k)}$ and $Q^{(k-1)}$ as $\eta_{Q_i}^{(k)'}$, we have $\mathcal{W}_p(Q^{(k-1)}, \gamma_i^{(k)}) = \mathcal{W}_p(Q^{(k-1)}, \eta_{Q_i}^{(k)'}) + \mathcal{W}_p(\eta_{Q_i}^{(k)'}, \gamma_i^{(k)})$.
Based on Algorithm 1, $\eta_{P_i}^{(k+1)}$ is the interpolating measure for P_i and $\gamma_i^{(k)}$. Thus, we can derive that,

$$\begin{aligned}
&\mathcal{W}_p(P_i, \eta_{P_i}^{(k+1)}) + \mathcal{W}_p(\eta_{P_i}^{(k+1)}, \gamma_i^{(k)}) \\
&\leq \mathcal{W}_p(P_i, \eta_{P_i}^{(k)}) + \mathcal{W}_p(\eta_{P_i}^{(k)}, \gamma_i^{(k)}) \qquad (17) \\
&\mathcal{W}_p(Q^{(k-1)}, \eta_{Q_i}^{(k)'}) + \mathcal{W}_p(\eta_{Q_i}^{(k)'}, \gamma_i^{(k)}) \\
&\leq \mathcal{W}_p(Q^{(k-1)}, \eta_{Q_i}^{(k)}) + \mathcal{W}_p(\eta_{Q_i}^{(k)}, \gamma_i^{(k)}) \qquad (18)
\end{aligned}$$

These two inequalities lead to

$$\begin{aligned}
&\mathcal{W}_p(P_i, \eta_{P_i}^{(k+1)}) + \mathcal{W}_p(\eta_{P_i}^{(k+1)}, \gamma_i^{(k)}) \\
&+ \mathcal{W}_p(Q^{(k-1)}, \eta_{Q_i}^{(k)'}) + \mathcal{W}_p(\eta_{Q_i}^{(k)'}, \gamma_i^{(k)}) \\
&\leq \mathcal{W}_p(P_i, \eta_{P_i}^{(k)}) + \mathcal{W}_p(\eta_{P_i}^{(k)}, \gamma_i^{(k)}) \qquad (19) \\
&+ \mathcal{W}_p(Q^{(k-1)}, \eta_{Q_i}^{(k)}) + \mathcal{W}_p(\eta_{Q_i}^{(k)}, \gamma_i^{(k)}) \qquad (20)
\end{aligned}$$

Simultaneously, the $\gamma_i^{(k)}$ is the interpolating measure for $\eta_{P_i}^{(k)}$ and $\eta_{Q_i}^{(k)}$. So we have

$$\begin{aligned}
&\mathcal{W}_p(\eta_{P_i}^{(k)}, \gamma_i^{(k)}) + \mathcal{W}_p(\eta_{Q_i}^{(k)}, \gamma_i^{(k)}) \\
&\leq \mathcal{W}_p(\eta_{P_i}^{(k)}, \gamma_i^{(k-1)}) + \mathcal{W}_p(\eta_{Q_i}^{(k)}, \gamma_i^{(k-1)}) \qquad (21)
\end{aligned}$$

and

$$\begin{aligned}
&\mathcal{W}_p(P_i, \eta_{P_i}^{(k+1)}) + \mathcal{W}_p(\eta_{P_i}^{(k+1)}, \gamma_i^{(k)}) \\
&+ \mathcal{W}_p(Q^{(k-1)}, \eta_{Q_i}^{(k)'}) + \mathcal{W}_p(\eta_{Q_i}^{(k)'}, \gamma_i^{(k)}) \\
&\leq \mathcal{W}_p(P_i, \eta_{P_i}^{(k)}) + \mathcal{W}_p(Q^{(k-1)}, \eta_{Q_i}^{(k)}) \\
&+ \mathcal{W}_p(\eta_{P_i}^{(k)}, \gamma_i^{(k-1)}) + \mathcal{W}_p(\eta_{Q_i}^{(k)}, \gamma_i^{(k-1)}) \qquad (22) \\
&= \mathcal{W}_p(P_i, \gamma_i^{(k-1)}) + \mathcal{W}_p(Q^{(k-1)}, \gamma_i^{(k-1)}) \qquad (23)
\end{aligned}$$

Hence, we can now derive that

$$A^{(k)} \leq \sum_{i}^{N} \mathcal{W}_p(P_i, \gamma_i^{(k)}) + \sum_{i}^{N} \mathcal{W}_p(Q^{(k-1)}, \gamma_i^{(k)}) \quad (24)$$

$$= \sum_{i}^{N} [\mathcal{W}_p(P_i, \eta_{P_i}^{(k+1)}) + \mathcal{W}_p(\eta_{P_i}^{(k+1)}, \gamma_i^{(k)})]$$

$$+ \sum_{i}^{N} [\mathcal{W}_p(Q^{(k-1)}, \eta_{Q_i}^{(k)'}) + \mathcal{W}_p(\eta_{Q_i}^{(k)'}, \gamma_i^{(k)})] \quad (25)$$

$$= \sum_{i}^{N} [\mathcal{W}_p(P_i, \eta_{P_i}^{(k+1)}) + \mathcal{W}_p(\eta_{P_i}^{(k+1)}, \gamma_i^{(k)})$$

$$+ \mathcal{W}_p(Q^{(k-1)}, \eta_{Q_i}^{(k)'}) + \mathcal{W}_p(\eta_{Q_i}^{(k)'}, \gamma_i^{(k)})] \quad (26)$$

$$\leq \sum_{i}^{N} [\mathcal{W}_p(P_i, \gamma_i^{(k-1)}) + \mathcal{W}_p(Q^{(k-1)}, \gamma_i^{(k-1)})] = A^{(k-1)} \quad (27)$$

Thus, the sequence $A_{(k)}$ is non-increasing. By the triangle inequality, we have for any $k \in \mathbb{N}$,

$$\sum_{i}^{N} \mathcal{W}_p(P_i, Q) \leq A^{(k)} \quad (28)$$

Using the monotone convergence theorem, since $A^{(k)}$ is non-increasing and bounded sequence below, then it converges to its infimum.

9. Proof for Theorem 2

In this section, we give a detailed proof for Theorem 2, which is a restatement of the proof for Theorem 1 in [13]. First, denote joint distribution of random data-label pairs $(x, f_t(x))_{x\sim P_i(x)}$ and $(x, f_v(x))_{x\sim Q(x)}$ as $P_i^{f_t}$ and Q^{f_v} respectively, which are the same notation as P_i and Q but made with explicit dependence on f_t and f_v for clarity. The distributions of $(f_t(x))_{x\sim P_i(x)}$ and $(f_v(x))_{x\sim Q(x)}$ as $P_{i_{f_t}}$ and Q_{f_v} respectively. Besides, we define conditional distributions $P_i(x|y) := \frac{P_i(x)I[f_t(x)=y]}{\int P_i(x)I[f_t(x)=y]dx}$ and $Q(x|y) := \frac{Q(x)I[f_v(x)=y]}{\int Q(x)I[f_v(x)=y]dx}$. Also, we denote $\pi \in \Pi(P_i, Q)$ as a coupling between a pair of distributions P_i, Q and $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as distance metric function. Generally, the p-Wasserstein distance with respect to cost function $\mathcal{C}$ is defined as $\mathcal{W}_p(P_i, Q) := \inf_{\pi\in\Pi(P_i,Q)} \mathbb{E}_{(x,y)\sim\pi}[\mathcal{C}(x, y)]$.

To prove Theorem 2, the concept of probabilistic *cross*-Lipschitzness is needed, and it is assumed that two labeling functions should produce consistent labels with high probability on two close instances.

Definition 5 *(Probabilistic Cross-Lipschitzness). Two labeling functions $f_t : \mathcal{X} \to \{0, 1\}^V$ and $f_v : \mathcal{X} \to \{0, 1\}^V$ are (ϵ, δ)-probabilistic cross-Lipschitz w.r.t. a joint distribution π over $\mathcal{X} \times \mathcal{X}$ if for all $\epsilon > 0$:*

$$P_{(x_1,x_2)\sim\pi}[||f_t(x_1) - f_v(x_2)|| > \epsilon d(x_1, x_2)] \leq \delta \quad (29)$$

Given labeling functions f_t, f_v and a coupling π, we can bound the probability of finding pairs of training and validation instances labeled differently in a $(1/\epsilon)$-ball with respect to π.

Let $\pi^*_{x,y}$ be the coupling between $P_i^{f_t}$ and Q^{f_v} such that

$$\pi^*_{x,y} := \arg_{\pi\in\Pi(P_i^{f_t}, Q^{f_v})} \inf \mathbb{E}_{((x_i,y_i),(x_q,y_q))\sim\pi}[\mathcal{C}((x_i, y_i), (x_q, y_q))]. \quad (30)$$

We define two couplings π^* and $\widetilde{\pi}^*$ between $P_i(x), Q(x)$ as follows:

$$\pi^*(x_i, x_q) := \int_{\mathcal{Y}} \int_{\mathcal{Y}} \pi^*_{x,y}((x_i, y_i), (x_q, y_q))dy_i dy_q. \quad (31)$$

For $\widetilde{\pi}^*$, we first need to define a coupling between $P_{i_{f_t}}$ and Q_{f_v}:

$$\pi^*_y(y_i, y_q) := \int_{\mathcal{X}} \int_{\mathcal{X}} \pi^*_{x,y}((x_i, y_i), (x_q, y_q))dx_i dx_q \quad (32)$$

and another coupling between $P_i^{f_t}, Q^{f_v}$:

$$\widetilde{\pi}^*_{x,y}(x_i, y_i), (x_q, y_q)) := \pi^*_y(y_i, y_q)P_i(x_i|y_i)Q(x_q|y_q). \quad (33)$$

Finally, $\widetilde{\pi}^*$ is constructed as follows:

$$\widetilde{\pi}^*(x_i, x_q) := \int_{\mathcal{Y}} \int_{\mathcal{Y}} \pi^*_y(y_i, y_q)P_i(x_i|y_i)Q(x_q|y_q)dy_i dy_q. \quad (34)$$

Next, we are going to prove Theorem 2. The lefthand side of inequality in the theorem can be written as

$$\mathbb{E}_{x\sim Q(x)}[\mathcal{L}(f_v(x), f(x))]$$
$$= \mathbb{E}_{x\sim Q(x)}[\mathcal{L}(f_v(x), f(x))]$$
$$- \mathbb{E}_{x\sim P_i(x)}[\mathcal{L}(f_t(x), f(x))] + \mathbb{E}_{x\sim P_i(x)}[\mathcal{L}(f_t(x), f(x))] \quad (35)$$

$$\leq \mathbb{E}_{x\sim P_i(x)}[\mathcal{L}(f_t(x), f(x))]$$
$$+ |\mathbb{E}_{x\sim Q(x)}[\mathcal{L}(f_v(x), f(x))] - \mathbb{E}_{x\sim P_i(x)}[\mathcal{L}(f_t(x), f(x))]| \quad (36)$$

We bound $|\mathbb{E}_{x\sim Q(x)}[\mathcal{L}(f_v(x), f(x))] -$

$\mathbb{E}_{x\sim P_i(x)}[\mathcal{L}(f_t(x), f(x))]|$ as follows:

$$
\begin{aligned}
&|\mathbb{E}_{x\sim Q(x)}[\mathcal{L}(f_v(x), f(x))] - \mathbb{E}_{x\sim P_i(x)}[\mathcal{L}(f_t(x), f(x))]| \\
&= |\int_{\mathcal{X}^2} [\mathcal{L}(f_v(x_q), f(x_q)) - \mathcal{L}(f_t(x_i), f(x_i))] d\pi^*(x_i, x_q)| \qquad (37)\\
&= |\int_{\mathcal{X}^2} [\mathcal{L}(f_v(x_q), f(x_q)) - \mathcal{L}(f_v(x_q), f(x_i)) \\
&+ \mathcal{L}(f_v(x_q), f(x_i)) - \mathcal{L}(f_t(x_i), f(x_i))] d\pi^*(x_i, x_q)| \qquad (38)\\
&\leq \underbrace{\int_{\mathcal{X}^2} |\mathcal{L}(f_v(x_q), f(x_q)) - \mathcal{L}(f_v(x_q), f(x_i))| d\pi^*(x_i, x_q)}_{U_1} \\
&+ \underbrace{\int_{\mathcal{X}^2} |\mathcal{L}(f_v(x_q), f(x_i)) - \mathcal{L}(f_t(x_i), f(x_i))| d\pi^*(x_i, x_q)}_{U_2} \qquad (39)
\end{aligned}
$$

where the last inequality is due to triangle inequality. Now, we bound U_1 and U_2 separately. For U_1, we have

$$
\begin{aligned}
U_1 &\leq k \int_{\mathcal{X}^2} ||f(x_q) - f(x_i)|| d\pi^*(x_i, x_q) \qquad (40)\\
&\leq k\epsilon \int_{\mathcal{X}^2} \mathrm{d}(x_i, x_q) d\pi^*(x_i, x_q), \qquad (41)
\end{aligned}
$$

where both inequalities are due to Lipschitzness of $\mathcal{L}$ and f. Recall that $\pi_y^*(y_i, y_q) := \int_{\mathcal{X}} \int_{\mathcal{X}} \pi_{x,y}^*((x_i, y_i), (x_q, y_q)) dx_i dx_q$ and $\widetilde{\pi}_{x,y}^*(x_i, y_i), (x_q, y_q)) := \pi_y^*(y_i, y_q) P_i(x_i|y_i) Q(x_q|y_q)$. And for U_2, we can derive that

$$
\begin{aligned}
U_2 &\leq k \int_{\mathcal{Y}^2} \int_{\mathcal{X}^2} ||y_q - y_i|| d\pi_{x,y}^*((x_i, y_i), (x_q, y_q)) \qquad (42)\\
&= k \int_{\mathcal{Y}^2} ||y_q - y_i|| d\pi_y^*(y_i, y_q) \qquad (43)\\
&= k \int_{\mathcal{X}^2} \int_{\mathcal{Y}^2} ||y_q - y_i|| d\widetilde{\pi}_{x,y}^*((x_i, y_i), (x_q, y_q)) \qquad (44)\\
&= k \int_{\mathcal{Y}^2} \int_{\mathcal{X}^2} ||f_v(x_q) - f_t(x_i)|| d\widetilde{\pi}_{x,y}^*((x_i, y_i), (x_q, y_q)), \qquad (45)
\end{aligned}
$$

where the last step holds since if $y_i \neq f_t(x_i)$ or $y_q \neq f_v(x_q)$, then $\widetilde{\pi}_{x,y}^*((x_i, y_i), (x_q, y_q)) = 0$. Define the region $\mathcal{A} = (x_i, y_i) : ||f_v(x_q) - f_t(x_i)|| < \epsilon_{tv} \mathrm{d}(x_i, x_q)$, then

$$
\begin{aligned}
U_2 &\leq k \int_{\mathcal{Y}^2} \int_{\mathcal{X}^2} ||f_v(x_q) - f_t(x_i)|| d\widetilde{\pi}_{x,y}^*((x_i, y_i), (x_q, y_q)) \qquad (46)\\
&= k \int_{\mathcal{Y}^2} \int_{\mathcal{X}^2 \setminus \mathcal{A}} ||f_v(x_q) - f_t(x_i)|| d\widetilde{\pi}_{x,y}^*((x_i, y_i), (x_q, y_q)) \\
&+ k \int_{\mathcal{Y}^2} \int_{\mathcal{A}} ||f_v(x_q) - f_t(x_i)|| d\widetilde{\pi}_{x,y}^*((x_i, y_i), (x_q, y_q)) \qquad (47)\\
&\leq k \int_{\mathcal{Y}^2} \int_{\mathcal{X}^2 \setminus \mathcal{A}} 2V d\widetilde{\pi}_{x,y}^*((x_i, y_i), (x_q, y_q)) \\
&+ k \int_{\mathcal{Y}^2} \int_{\mathcal{A}} ||f_v(x_q) - f_t(x_i)|| d\widetilde{\pi}_{x,y}^*((x_i, y_i), (x_q, y_q)). \qquad (48)
\end{aligned}
$$

Define $\widetilde{f}_t(x_i) = f_t(x_i)$ and $\widetilde{f}_v(x_q) = f_v(x_q)$ if $(x_i, x_q) \in \mathcal{A}$, and $\widetilde{f}_t(x_i) = \widetilde{f}_v(x_q) = 0$ otherwise (note that $||\widetilde{f}_v(x_q) - \widetilde{f}_t(x_i)|| < \epsilon_{tv} \mathrm{d}(x_i, x_q)$ for all $(x_i, x_q) \in \mathcal{X}^2$), then we can bound the second term as follows:

$$
\begin{aligned}
&k \int_{\mathcal{Y}^2} \int_{\mathcal{A}} ||f_v(x_q) - f_t(x_i)|| d\widetilde{\pi}_{x,y}^*((x_i, y_i), (x_q, y_q)) \qquad (49)\\
&\leq k \int_{\mathcal{Y}^2} d\pi_y^*(y_i, y_q) \\
&\int_{\mathcal{A}} ||f_v(x_q) - f_t(x_i)|| dP_i(x_i|y_i) dQ(x_q|y_q) \qquad (50)\\
&= k \int_{\mathcal{Y}^2} d\pi_y^*(y_i, y_q) \\
&\int_{\mathcal{X}}^2 ||\widetilde{f}_v(x_q) - \widetilde{f}_t(x_i)|| dP_i(x_i|y_i) dQ(x_q|y_q) \qquad (51)\\
&= k \int_{\mathcal{Y}^2} d\pi_y^*(y_i, y_q) \\
&\int_{\mathcal{X}}^2 ||\mathbb{E}_{x_q\sim Q(\cdot|y_q)}[\widetilde{f}_v(x_q)] - \mathbb{E}_{x_i\sim P_i(\cdot|y_i)}[\widetilde{f}_t(x_i)]|| \qquad (52)\\
&\leq k\epsilon_{tv} \int_{\mathcal{Y}^2} d\pi_y^*(y_i, y_q) \mathcal{W}_p(P_i(\cdot|y_i), Q(\cdot|y_q)). \qquad (53)
\end{aligned}
$$

The last inequality is a consequence of the duality form of the Kantorovich-rubinstein theorem [37]. Combining all

parts, we have

$$U_1 + U_2 \leq k\epsilon \int_{\mathcal{X}^2} \mathrm{d}(x_i, x_q) d\pi^*(x_i, x_q)$$
$$+ k \int_{\mathcal{Y}^2} \int_{\mathcal{X}^2 \setminus \mathcal{A}} 2V d\widetilde{\pi}^*_{x,y}((x_i, y_i), (x_q, y_q))$$
$$+ k\epsilon_{tv} \int_{\mathcal{Y}^2} d\pi^*_y(y_i, y_q) \mathcal{W}_d(P_i(\cdot|y_i), Q(\cdot|y_q)) \quad (54)$$
$$\leq k\epsilon \int_{\mathcal{X}^2} \mathrm{d}(x_i, x_q) d\pi^*(x_i, x_q) + 2kV\delta_{tv}$$
$$+ k\epsilon_{tv} \int_{\mathcal{Y}^2} d\pi^*_y(y_i, y_q) \mathcal{W}_p(P_i(\cdot|y_i), Q(\cdot|y_q)) \quad (55)$$
$$= 2kV\delta_{tv} + k \int_{(\mathcal{X}\times\mathcal{Y})^2} [\epsilon d(x_i, x_q)$$
$$+ \epsilon_{tv} \mathcal{W}_p(P_i(\cdot|y_i), Q(\cdot|y_q))] d\pi^*_{x,y}((x_i, y_i), (x_q, y_q)) \quad (56)$$
$$\leq 2kV\delta_{tv} + k \int_{(\mathcal{X}\times\mathcal{Y})^2} [\epsilon d(x_i, x_q) +$$
$$c\epsilon \mathcal{W}_p(P_i(\cdot|y_i), Q(\cdot|y_q))] d\pi^*_{x,y}((x_i, y_i), (x_q, y_q)) \quad (57)$$
$$= k\epsilon \mathbb{E}_{\pi^*_{x,y}}[\mathcal{C}((x_i, y_i), (x_q, y_q))] + 2kV\delta_{tv} \quad (58)$$
$$= k\epsilon \mathcal{W}_p(P_i^{f_t}, Q^{f_v}) + 2kV\delta_{tv}. \quad (59)$$

Thus the inequality in theorem 2 has been proved. For a more detailed discussion, please refer to [13].

10. Experiments Details

10.1. Client Evaluation

We follow the same settings in [20] and [34] and divide the combination of distribution and size for client data into five different cases.

(1) Same Distribution and Same Size: All five clients possess the same number of images for each class;

(2) Different Distributions and Same Size: Each participant has the same number of samples. However, the Participant 1 dataset contains 80% for two classes. The other clients evenly divide the remaining 20% of the samples. Similar procedures are applied to the rest;

(3) Same Distribution and Different Sizes: Randomly sample from the entire training set according to pre-defined ratios to form the local dataset for each participant, while ensuring that there are the same number of images for each class in each participant. The ratios for client 1-5 are: 10%,15%, 20%, 25% and 30%;

(4) Noisy Labels and Same Size: Adopt the dataset from case (1), and flip the labels of a pre-defined percentage of samples in each participant's local dataset. The ratios for client 1-5 are: 0%,5%, 10%, 15% and 20%;

(5) Noisy Features and Same Size: Adopt the dataset from case (1), and add different percentages of Gaussian noise into the input images. The ratios for client 1-5 are: 0%,5%, 10%, 15% and 20%.

10.2. Implementation Details

The code implementation is developed based on Pytorch. We also developed our code and conduct comparisons based on the reference from the following sources:

- Opendataval [12]
- LAVA [13]
- WBTransport [23]
- GTG-Shapley [20]

10.3. Experimental Baselines

Original Shapley The Shapley value is a concept from cooperative game theory that has been applied to machine learning to attribute a value to each feature (or client) in a predictive model. It provides a way to fairly distribute the "contribution" of each player in a cooperative game. In the context of federated learning, it can be assumed that the prediction of a model in the server is the outcome of a game, and each client is a player contributing to that prediction. The Shapley value assigns a value to each client based on its marginal contribution to the model's prediction across all possible combinations of data for different clients.

Formally, a coalitional game is defined as: There is a set N (of n players) and a function v that maps subsets of players to the real numbers: $2^N \to \mathbb{R}$, with $v(\emptyset) = 0$, where $\emptyset$ denotes the empty set. Then a general formula to calculate the Shapley value is provided as follows:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)) \quad (60)$$

where n is the total number of players and the sum extends over all subsets S of N not containing player i.

MR/OR [34] These two metrics are based on the contribution index, which is a concept proposed by the paper to replace the original Shapley value. Since direct computing of the contribution index can be time-consuming, two gradient-based methods are provided to reduce the time. The first one reconstructs models by updating the initial global model in federated learning with the gradients in different rounds. Then it calculates the contribution index by the performance of these reconstructed models. The second method calculates the contribution index in each round by updating the global model in the previous round with the gradients in the current round. Contribution indexes of multiple rounds are then added together with elaborated weights to get the final result.

TMC-Shapley [8] The Truncated Monte Carlo Shapley (TMC-Shapley) is a data evaluation metric to quantify the value of each training datum to the predictor performance.

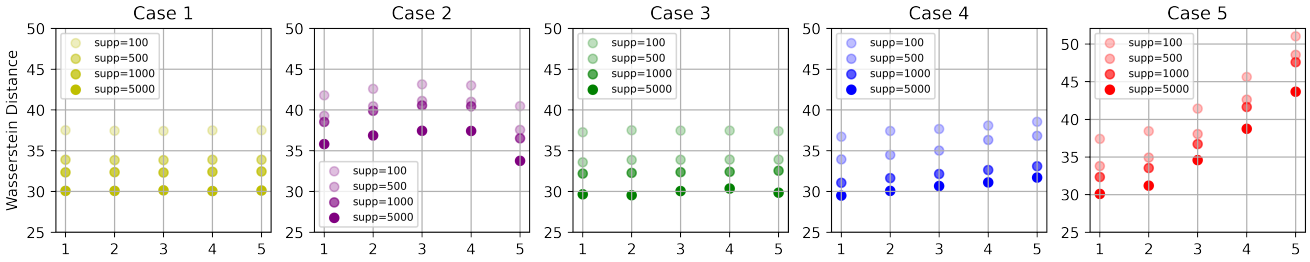

Figure 6. Wassestein Distances Under Different Support for **Q** (x-axis is the client number)

The Monte Carlo and gradient-based methods are developed to efficiently estimate the value in practical settings. However, this metric only considers the context of supervised machine learning, and the privacy demand is neglected in the paper.

FedShapley [41] This metric is proposed as a variant of the Shapley value amenable to federated learning. The key idea of the modification is to characterize the aggregate value of the set of clients in the same round through the model performance change caused by the addition of their data and then use the Shapley value to distribute the value of the set to each client. Compared with the canonical SV, it can be calculated without incurring extra communication costs and is capable of capturing the effect of participation order on data value.

GTG-Shapley [20] The Guided Truncation Gradient Shapley (GTG-Shapley) approach is a modification of the original Shapley value to address the challenge of significant computation costs in practice. It reconstructs federated learning models from gradient updates for Shapley value calculation instead of repeatedly training with different combinations of participants. A guided Monte Carlo sampling technique is introduced into the algorithm, enhancing the efficiency of calculation and reducing the computation costs.

11. Discussions

11.1. Hyperparameter Analysis

Our observations have shown that epoch K exerts minimal influence on the approximated distance. Conversely, with an increasing quantity of supports of the interpolating measure S, the approximated distance progressively approaches the exact distance. Consequently, in the context of evaluating relative contributions, a choice of few epochs and supports can effectively approximate the relative distance, leading to a reduction in computational complexity.

11.2. Consistent evaluation time

We find the truncation techniques in [8, 20] depend on the test performance, when the performance with a certain subset of clients is above the pre-specified threshold, contributions of remaining clients are assigned to 0 without additional evaluations. Therefore, the evaluation time varies with different truncation times and in the worst case the truncation will be conducted at the last round, making the complexity approaching $\mathcal{O}(2^N)$. In addition, the gradients in [34] with noisy data make MR and OR approaches have larger elapsed time than other cases. However, FedBary is robust and the elapsed time will not be affected by data characteristics.

Data	#Noisy	#Removed	acc.before	acc.after
CIFAR10	500	494	0.67	0.73
Fashion	1000	580	0.56	0.64

Table 4. Accuracy before/after removing detected noisy samples

11.3. Client detection is better than server detection

The detection accuracy is 100% in the client side with $\nabla\mathcal{W}_p(P_i, \eta_Q)$ and 76% in the server side with $\nabla\mathcal{W}_p(\eta_Q, Q)$. We conjecture this result is due to the gradient towards the P_i is more informative and straightforward. For the mislabeled data detection, our approach could only detect 45% of noisy data. However, it is worth noting that when accessing data, the mislabeled detection accuracy is only 47%, and the bottom plots show two approaches are almost overlapping, which shows Fedbary does not sacrifice much performance on detections on the benchmark of accessing data.

11.4. Boost FL Performance

11.5. Future Explorations

Noisy Label Detection The current approach, which relies on an augmented matrix based on a Gaussian approximation for the conditional distribution, demonstrates relatively poor performance in differentiating clean subsets from mislabeled ones compared to the case of noisy features. To address this issue, a potential future direction is to implement exact calculations for the Wasserstein distance by utilizing an interpolating measure and an appropriate embedding approach, to design a filtering approach for the mislabeled case.

Accurate Server Detection Detecting noisy data from the server side is crucial for defending against potential attacks from untrusted clients. The current approach, primarily driven by client-side analysis, excels in detection due to the client's access to their own data and the ability to measure gradients with respect to the interpolating measure such as γ_i or η_{Q_i} shared from the server side. However, the detection ability from the server side is limited since it cannot access local client data, and using η_{P_i} from the local client or γ_i is less effective in our explorations. Strengthening the server-side detection capabilities is of paramount importance in the context of the security application.