Flow-based Distributionally Robust Optimization

Chen Xu [1] Jonghyeok Lee [1] Xiuyuan Cheng [2] Yao Xie [1]

[1]Georgia Institute of Technology [2]Duke University

Introduction

Flow-based models are continuous-time models that gradually transform base distributions into desired target distributions. Such models have been popular in the field of normalizing flow, where the target distribution is pre-specified as the standard multivariate Gaussian. In this context, flow-based models allow accurate density estimation of and efficient sampling from the base distribution.

However, in many problems, our objective is not to model the current data distribution but to find a density transport map that fits other goals. In this sense, we need to *extrapolate* the data distribution along certain directions.

In this work, our particular consideration is the problem of distributionally robust optimization (DRO) [2], which aims to find robust optimal solutions that minimize certain risk over the worst-case distribution within a pre-specified uncertainty set. An essential aspect of solving DRO is to **find the worst-case distribution**, which is an infinite-dimensional optimization problem that is particularly challenging in high dimension for general risk functions.

Problem setup

Consider a real-valued *risk function* $\mathcal{R}$ taking as inputs a d-dimensional distribution P with a finite second moment and a measurable test function ϕ. Specifically, we assume there is a pre-specified loss function ℓ so that $\mathcal{R}(P,\phi)=\mathbb{E}_{X\sim P}[\ell(X,\phi)]$. We are interested in solving the following distributionally robust optimization (DRO) problem:

$$\min_{\phi\in\Phi}\max_{P\in\mathcal{P}_r}\ \mathcal{R}(P,\phi). \tag{1}$$

In (1), Φ denotes the constraint set for ϕ and $\mathcal{P}_r$ is the Wasserstein-2 (W_2) ball around the data distribution P_X. Namely, we let

$$\mathcal{P}_r=\{P:W_2^2(P,P_X)\le r^2\}, \tag{2}$$

where by the Monge formulation, the W_2 distance $W_2^2(P_2,P_1)$ between two d-dimensional distributions P_2,P_1 with finite second moments is written as

$$W_2^2(P_2,P_1)=\min_{T:T_\#P_1=P_2}\mathbb{E}_{X\sim P_1}\|T(X)-X\|_2^2. \tag{3}$$

In (3), $T:\mathbb{R}^d\to\mathbb{R}^d$ denotes the transport map and $T_\#$ denotes the push-forward operation by T, where $(T_\#P)(A)=P(T^{-1}(A))$ for a measureable set A.

Our approach

We focus on solving for the worst-case distribution within $\mathcal{P}_r$ (i.e., the maximizer of the inner problem of (1)), assuming ϕ is given. Note that for each $r>0$, there is an $h>0$ dependent on r, such that we have the following equivalent unconstrained problem:

$$\min_{P\in\mathcal{P}}-\mathcal{R}(P,\phi)+\frac{1}{2h}W_2^2(P,P_X). \tag{4}$$

Under (3), it is also equivalent to solving the optimal transport map $T:\mathbb{R}^d\to\mathbb{R}^d$:

$$\min_{T:\mathbb{R}^d\to\mathbb{R}^d}\mathbb{E}_{X\sim P_X}\left[-\ell(T(X),\phi)+\frac{1}{2h}\|T(X)-X\|_2^2\right]. \tag{5}$$

To yield a tractable and meaningful solution of T, we parametrize it as the solution map of a NeuralODE model [1]. Specifically, given a neural network $f(x(t),t;\psi)$ with trainable parameters ψ, define the solution map T_s^t over interval $[s,t)$ as

$$T_s^t(x;\psi)=x+\int_s^t f(x(s'),s';\psi)ds'. \tag{6}$$

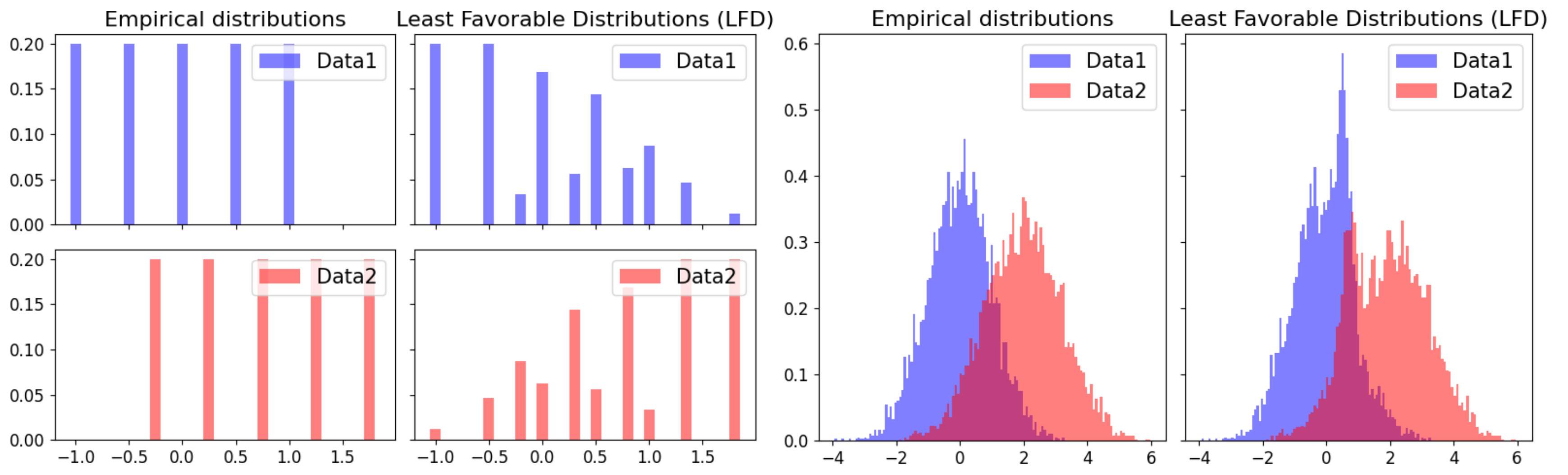
Using (6), the problem of finding T in (5) thus reduces to training ψ in the following problem:

$$\min_{\psi}\mathbb{E}_{X\sim P_X}\left[-\ell(T_0^1(X;\psi),\phi)+\frac{1}{2h}\|T_0^1(X;\psi)-X\|_2^2\right]. \tag{7}$$

We have also developed a step-wise training algorithm using (7), which is inspired by the JKO approach [5].

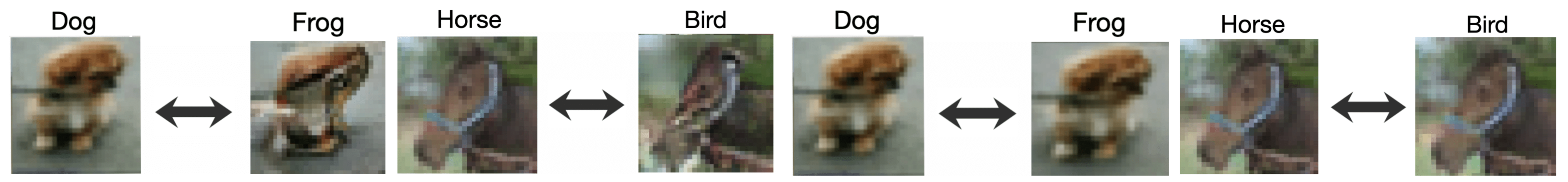
Main contributions

1. Re-formulate the problem of finding worst-case distributions in DRO into its Wasserstein proximal form and subsequently reduce the infinite-dimensional problem into **solving for a transport map.**
2. Parametrize the transport map by **continuous-time flow** neural networks and develop an efficient block-wise training algorithm to train the flow parameters.
3. Demonstrate the **effectiveness** of FlowDRO on various high-dimensional problems from adversarial attack and differential privacy.



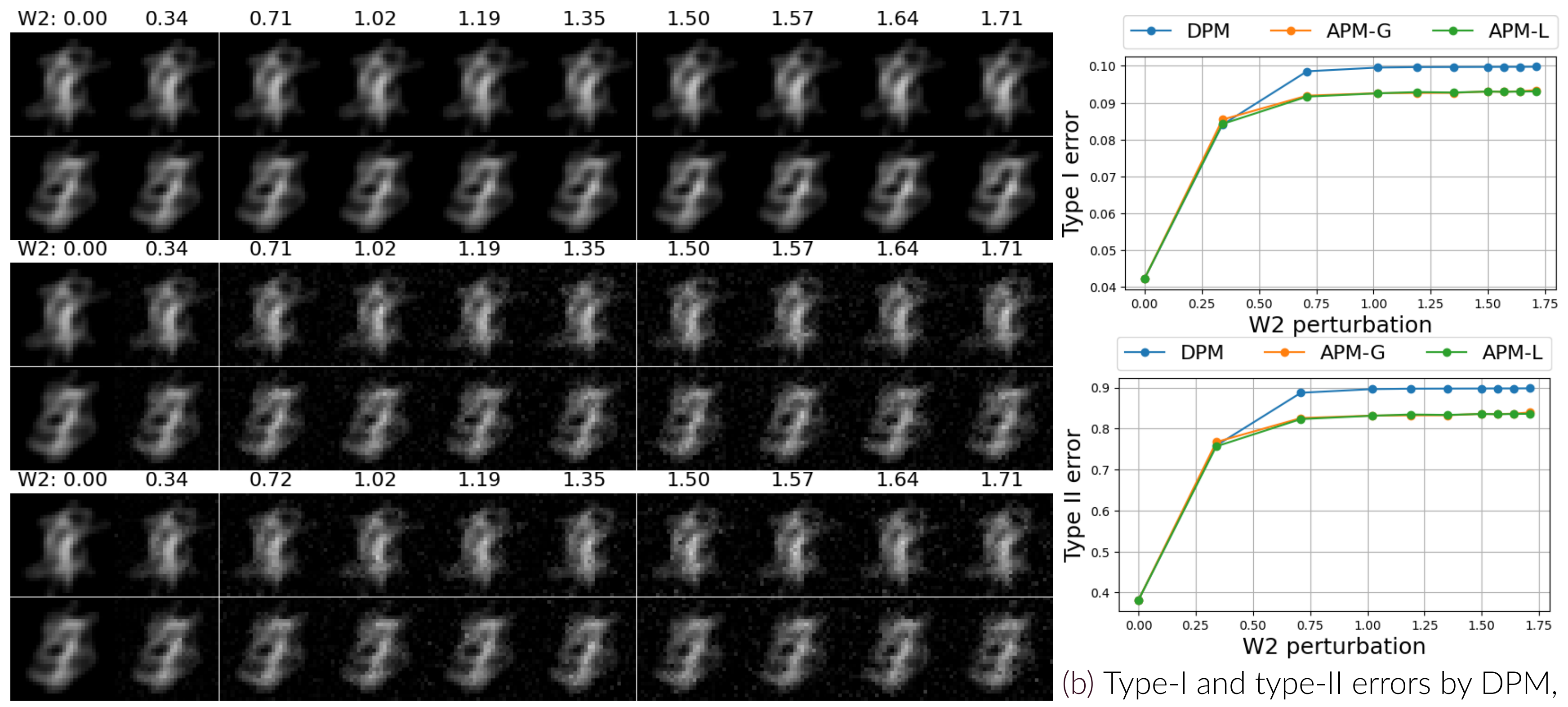
(a) WDRO on training samples (b) Proposed `FlowDRO` on test samples

Figure 1. Comparison of WDRO and `FlowDRO` on the 1D example following [4, Figure 1]. **Left of (a) and (b):** Empirical distributions of two sets of *training* (shown in (a)) and *test* (shown in (b)) samples from $\mathcal{N}(0,1)$ and $\mathcal{N}(2,1.2)$. **Right of (a) and (b):** Least-favorable distributions (LFD) found by WDRO and `FlowDRO`, where LFDs are within the W_2 ball (2) with radius $\varepsilon=0.1$. Note that WDRO solves a convex problem to obtain the LFD by moving the probability mass on *discrete training samples*. In particular, WDRO does not generalizes to a test sample unless it coincides exactly with some of the training samples. In comparison, `FlowDRO` yields a one-to-one continuous-time transport map that can be directly applied to both training and test samples, meanwhile leading to a continuous LFD.



(a) Attacked by `FlowDRO` (b) Attacked by PGD-ℓ_2

Figure 2. Adversarial samples found by `FlowDRO` and by PGD-ℓ_2. Captions show prediction by the pre-trained classifier ϕ on input images X_{test} (before attack) and $\tilde{X}_{test}$ (after attack).

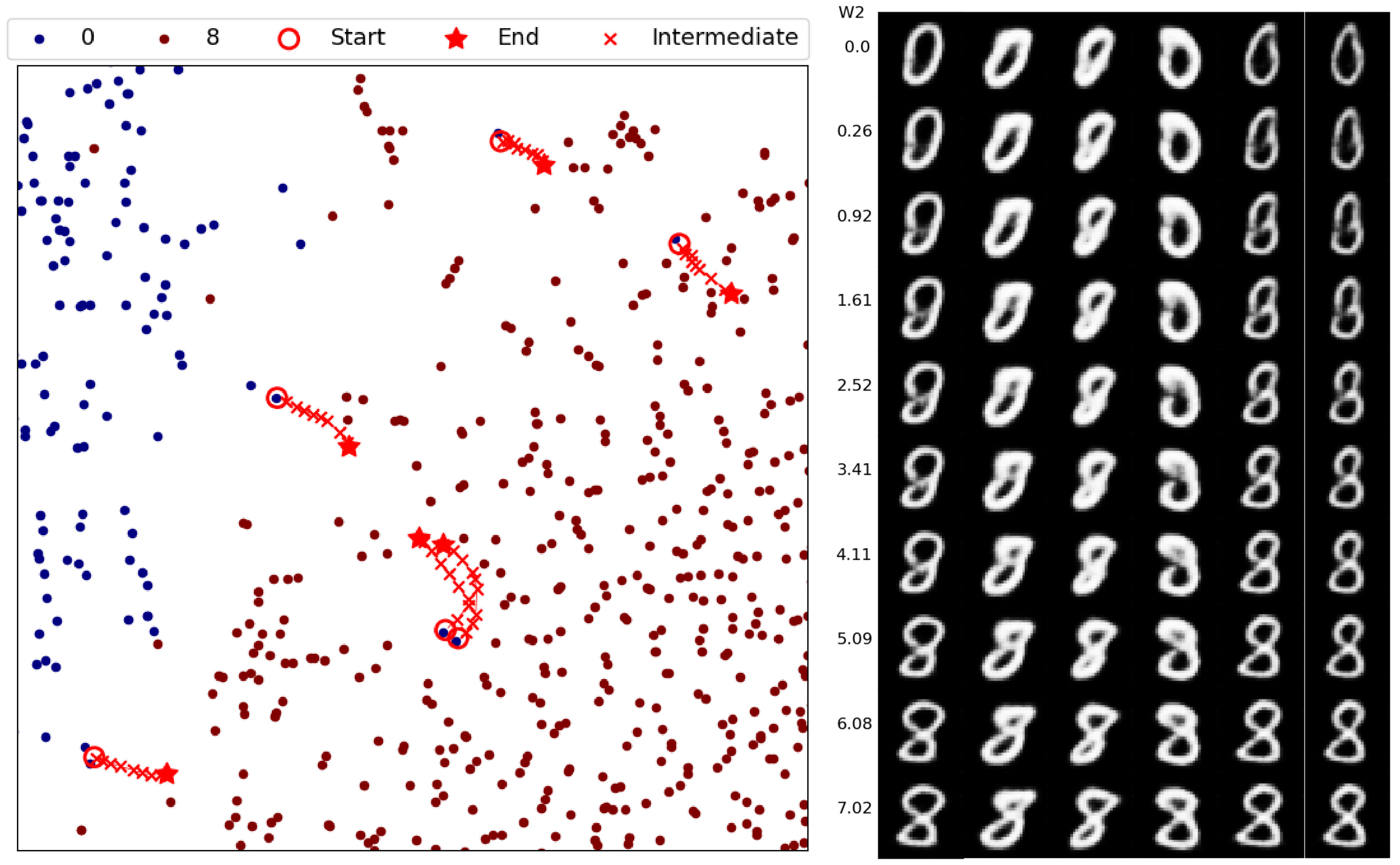


(a) DPM (top), APM-G (middle), and APM-L (bottom) (b) Type-I and type-II errors by DPM, APM-G, and APM-L.

Figure 3. Differential privacy example of one-neighborhood digit missing. Figures (a) visualize privacy-protected queries $M_r(q(S))$ by DPM, APM-G, and APM-L within different Wasserstein-2 balls with radius r around the distribution of $q(S)$. Figure (b) examines the type-I and type-II errors over different r. We control the value of r by different DP mechanisms to be identical for a fair comparison.

Numerical results

We showcase the proposed `FlowDRO` on adversarial attack of MNIST and CIFAR10 and differential privacy on membership inference attack. We also show preliminary results on solving the min-max problem (1) to get more robust decision functions ϕ (see Figure 5).



(a) Digit deformation trajectories under 2D T-SNE embedding (b) Gradual change of X_{test} in pixel space

Figure 4. `FlowDRO` perturbation of MNIST digits over blocks and time integration. Figure (a) visualizes the perturbation trajectories from digits 0 to 8 under 2D T-SNE embedding. Figure (b) shows the trajectory in pixel space, along with the corresponding W_2 perturbation between original and perturbed images over time.

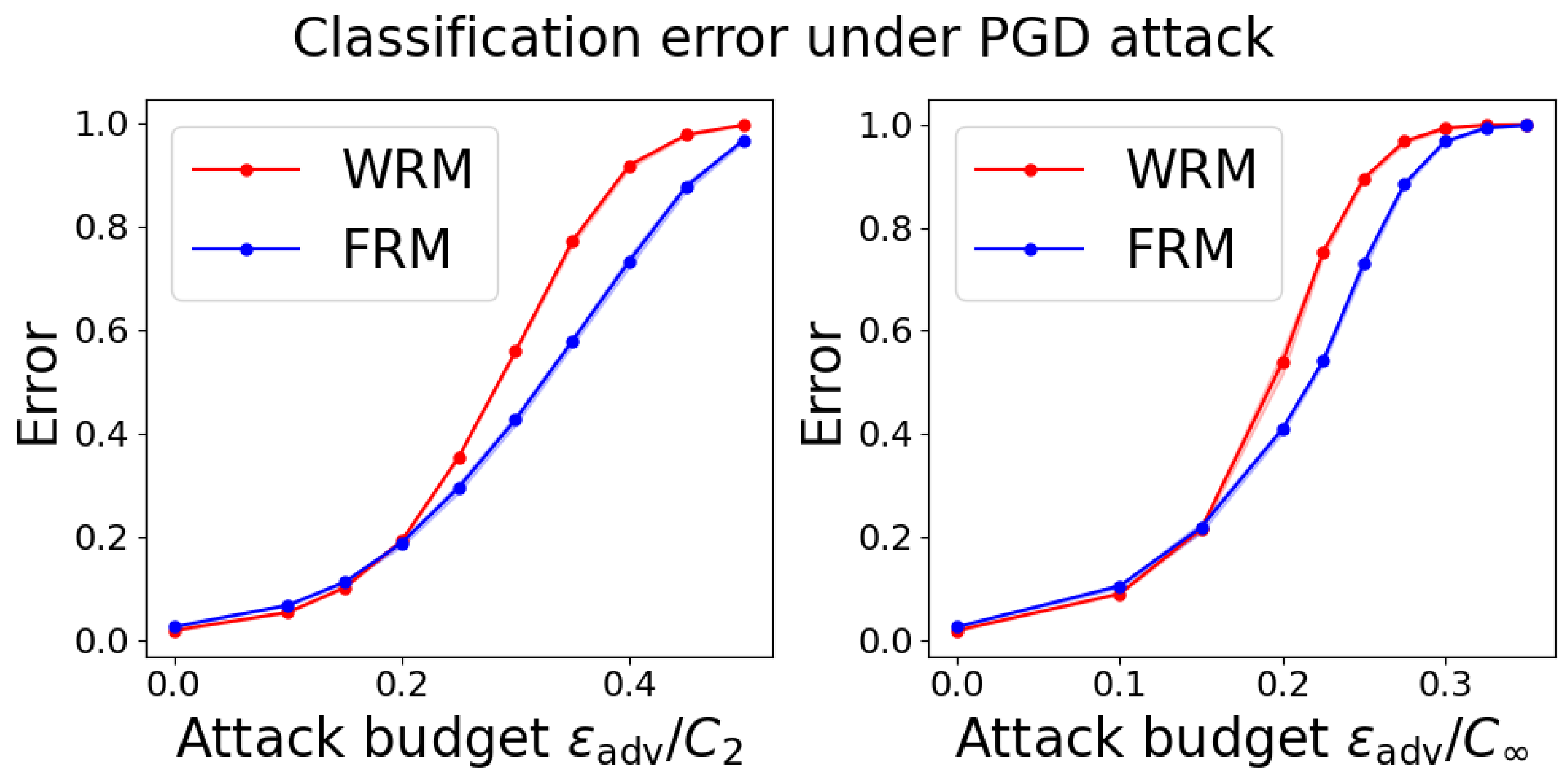


Figure 5. Performance of trained robust classifier on MNIST, using FRM (ours leveraging `FlowDRO`) vs. WRM [3].

References

[1] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[2] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[3] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[4] Liyan Xie, Rui Gao, and Yao Xie. Robust hypothesis testing with wasserstein uncertainty sets. *arXiv preprint arXiv:2105.14348*, 2021.

[5] Chen Xu, Xiuyuan Cheng, and Yao Xie. Normalizing flow neural networks by JKO scheme. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.