Multiple Wasserstein Gradient Descent for Multi-objective Distributional Optimization

Dai Hai Nguyen[1], Hiroshi Mamitsuka[2], Atsuyoshi Nakamura[1]

[1]Hokkaido University

[2]Kyoto University

Multiple-**O**bjective **D**istributional **O**ptimization (MODO)

$$\min_{q \in \mathcal{P}_2(\mathcal{X})} \mathbf{F}(q) \triangleq [F_1(q), F_2(q), \ldots, F_K(q)]$$

- $\mathcal{X} \subseteq \mathbb{R}^d$: a compact domain, $\mathcal{P}_2(\mathcal{X})$: space of probability distributions on $\mathcal{X}$
- $K \geq 2$: number of objectives
- Each $F_k(q)$ is defined over the space $\mathcal{P}_2(\mathcal{X})$

Multiple-**O**bjective **D**istributional **O**ptimization (MODO)

$$\min_{q \in \mathcal{P}_2(\mathcal{X})} \mathbf{F}(q) \triangleq [F_1(q), F_2(q), \ldots, F_K(q)]$$

- $\mathcal{X} \subseteq \mathbb{R}^d$: a compact domain, $\mathcal{P}_2(\mathcal{X})$: space of probability distributions on $\mathcal{X}$
- $K \geq 2$: number of objectives
- Each $F_k(q)$ is defined over the space $\mathcal{P}_2(\mathcal{X})$
- MODO can be regarded as multi-objective optimization (MOO) on the space of probability distributions, where MOO is:

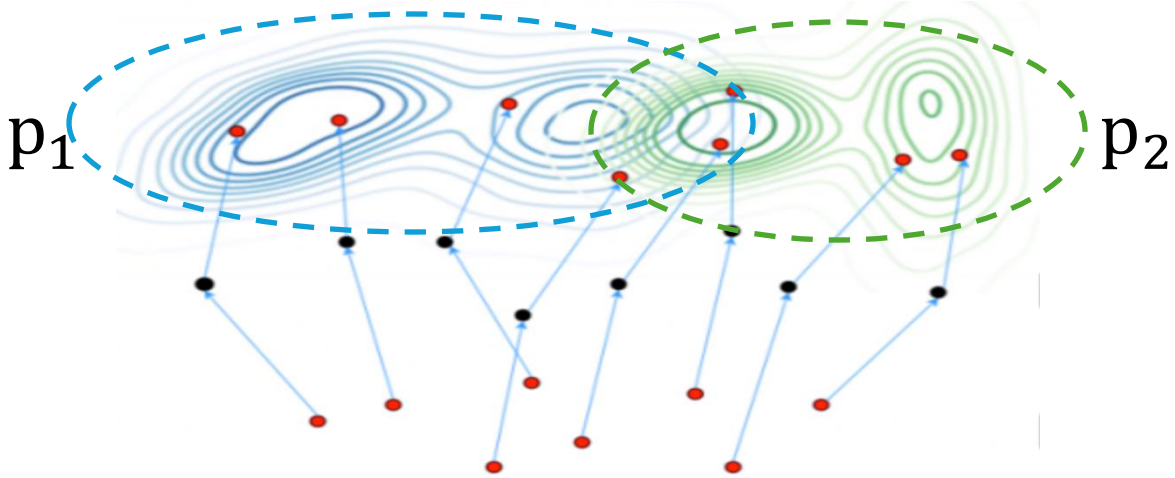
$$\min_{\mathbf{x} \in \mathcal{X}} \mathbf{f}(\mathbf{x}) \triangleq [f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_K(\mathbf{x})]$$

Multiple-Objective Distributional Optimization (MODO)

- Example: **Multi-Target Sampling**

Given a set of target distributions $\{\mathrm{p}_1(\mathbf{x}), \mathrm{p}_2(\mathbf{x}), \ldots, \mathrm{p}_K(\mathbf{x})\}$, with parameters $\mathbf{x}$, we aim to find the optimal distribution q^* that minimize:

$$\min_{\mathrm{q}\in\mathcal{P}_2(\mathcal{X})} [\mathbf{KL}(\mathrm{q}, \mathrm{p}_1), \mathbf{KL}(\mathrm{q}, \mathrm{p}_2), \ldots, \mathbf{KL}(\mathrm{q}, \mathrm{p}_K)]$$

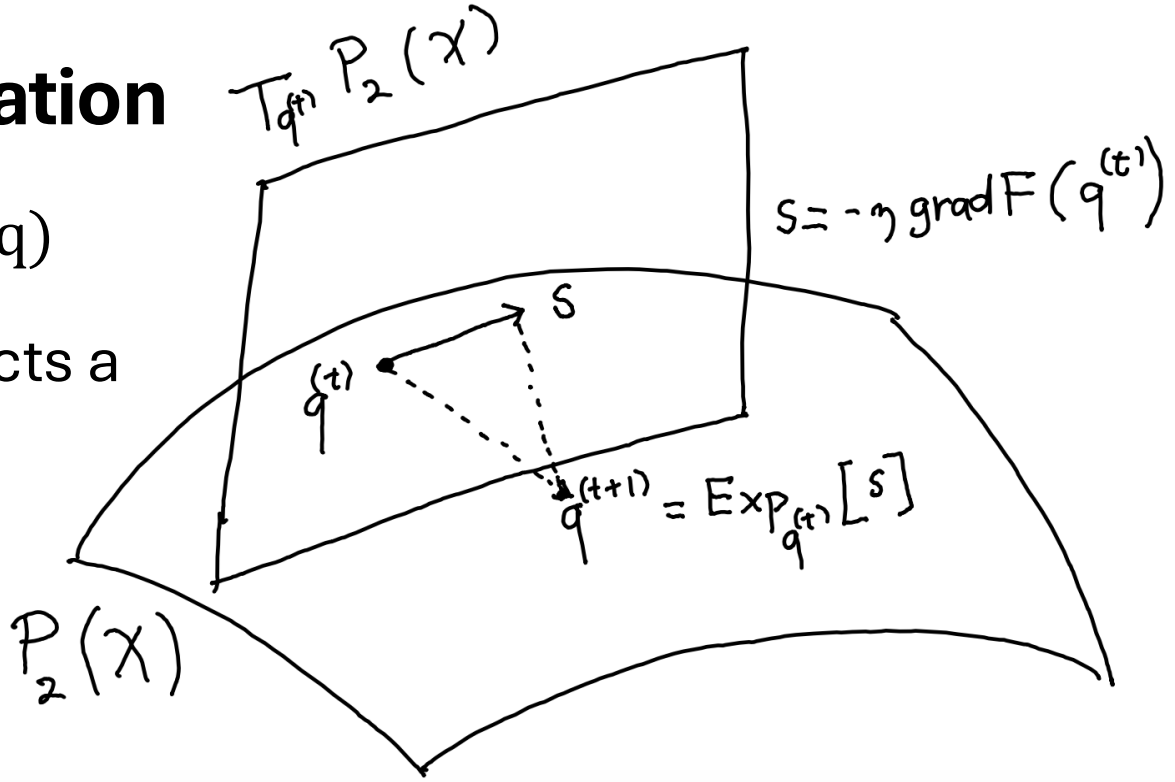


based on picture in [Phan et al, NeurIPS 2022]

Related work: Distributional Optimization

Optimization over distribution space: $\min_{q \in \mathcal{P}_2(\mathcal{X})} F(q)$

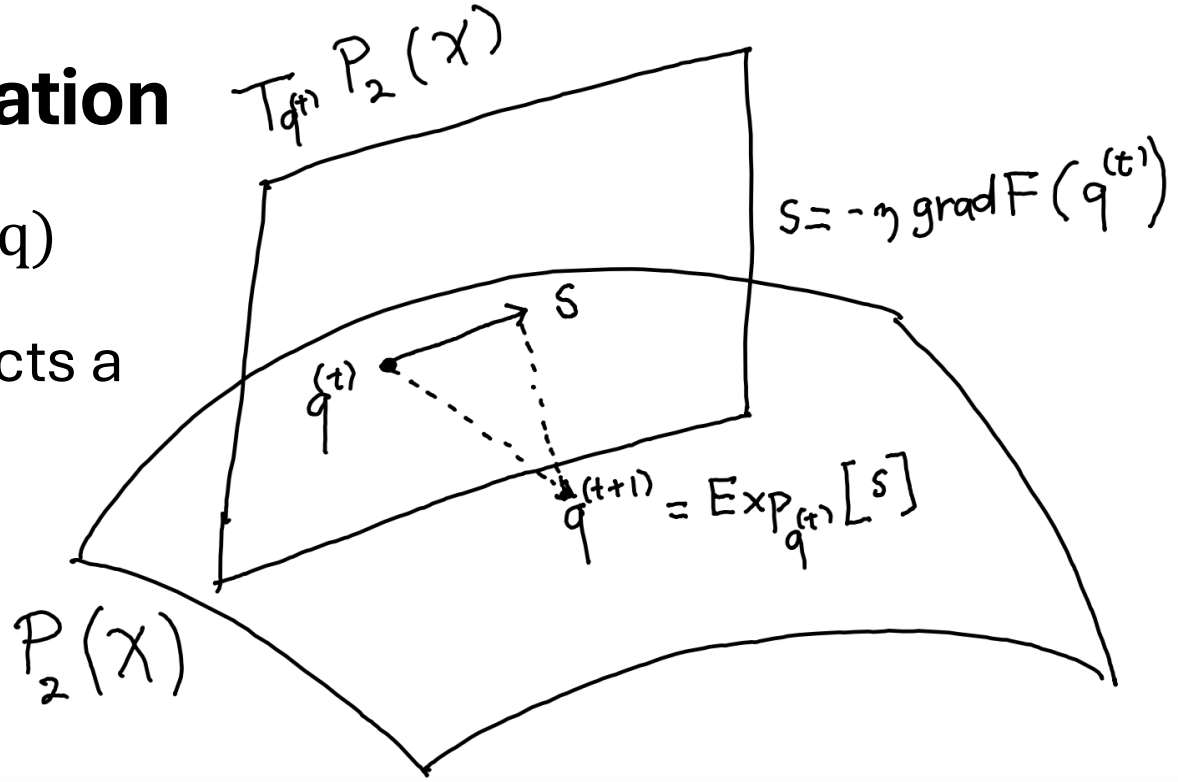
- Wasserstein gradient descent (**WGD**) constructs a sequence of $\{q^{(t)}\}$ to decrease F
- $\mathcal{T}_{q^{(t)}}\mathcal{P}_2(\mathcal{X})$: tangent space at $q^{(t)}$
- $s \in \mathcal{T}_{q^{(t)}}\mathcal{P}_2(\mathcal{X})$: tangent vector at $q^{(t)}$



Related work: Distributional Optimization

Optimization over distribution space: $\min_{q \in \mathcal{P}_2(\mathcal{X})} F(q)$

- Wasserstein gradient descent (**WGD**) constructs a sequence of $\{q^{(t)}\}$ to decrease F
- $\mathcal{T}_{q^{(t)}}\mathcal{P}_2(\mathcal{X})$: tangent space at $q^{(t)}$
- $s \in \mathcal{T}_{q^{(t)}}\mathcal{P}_2(\mathcal{X})$: tangent vector at $q^{(t)}$



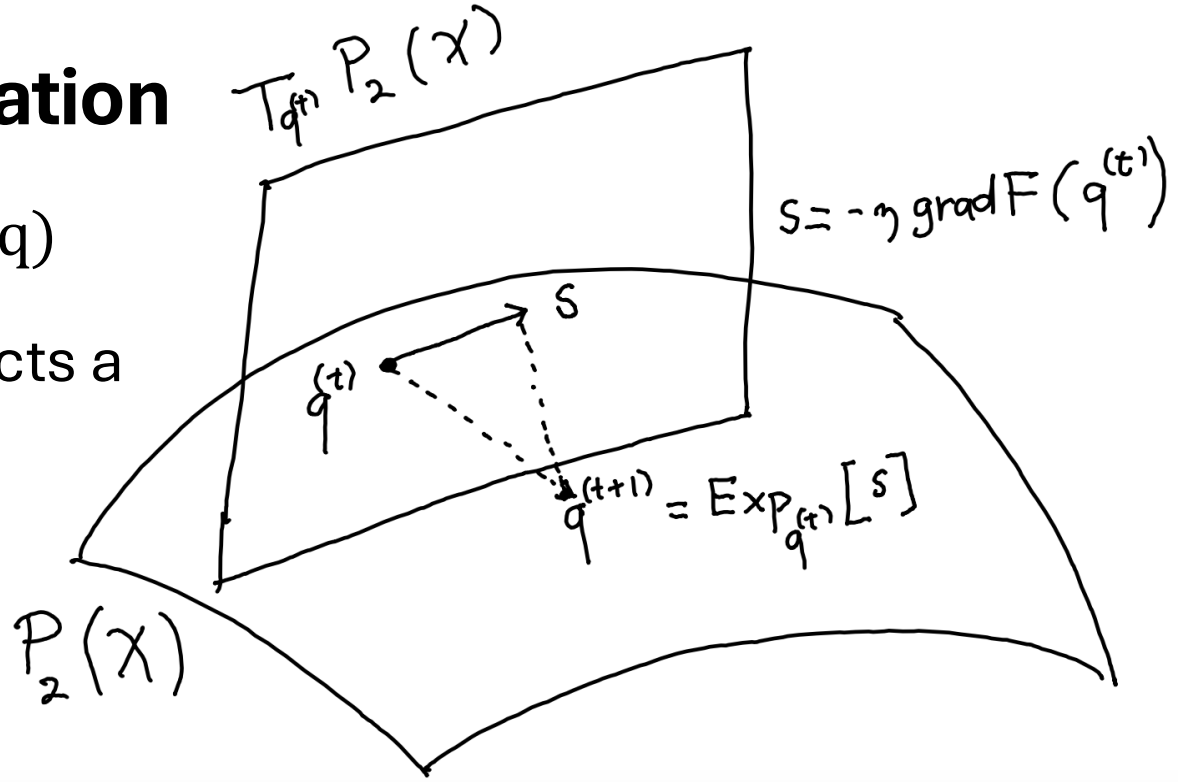
- Let $\gamma: [0,1] \to \mathcal{P}_2(\mathcal{X})$ be a curve satisfying $\gamma(0) = q^{(t)}$ and $\gamma'(0) = s$, **first variation** $\boldsymbol{\delta F(q^{(t)})}$ is defined as

$$\lim_{h \to 0} \frac{1}{h}[F(\gamma(h)) - F(q^{(t)})] = \int \delta F(q^{(t)})(\mathbf{x}) s(\mathbf{x}) d\mathbf{x}$$

Related work: Distributional Optimization

Optimization over distribution space: $\min_{q \in \mathcal{P}_2(\mathcal{X})} F(q)$

- Wasserstein gradient descent (**WGD**) constructs a sequence of $\{q^{(t)}\}$ to decrease F
- $\mathcal{T}_{q^{(t)}}\mathcal{P}_2(\mathcal{X})$: tangent space at $q^{(t)}$
- $s \in \mathcal{T}_{q^{(t)}}\mathcal{P}_2(\mathcal{X})$: tangent vector at $q^{(t)}$



- Let $\gamma: [0,1] \to \mathcal{P}_2(\mathcal{X})$ be a curve satisfying $\gamma(0) = q^{(t)}$ and $\gamma'(0) = s$, **first variation** $\boldsymbol{\delta F(q^{(t)})}$ is defined as

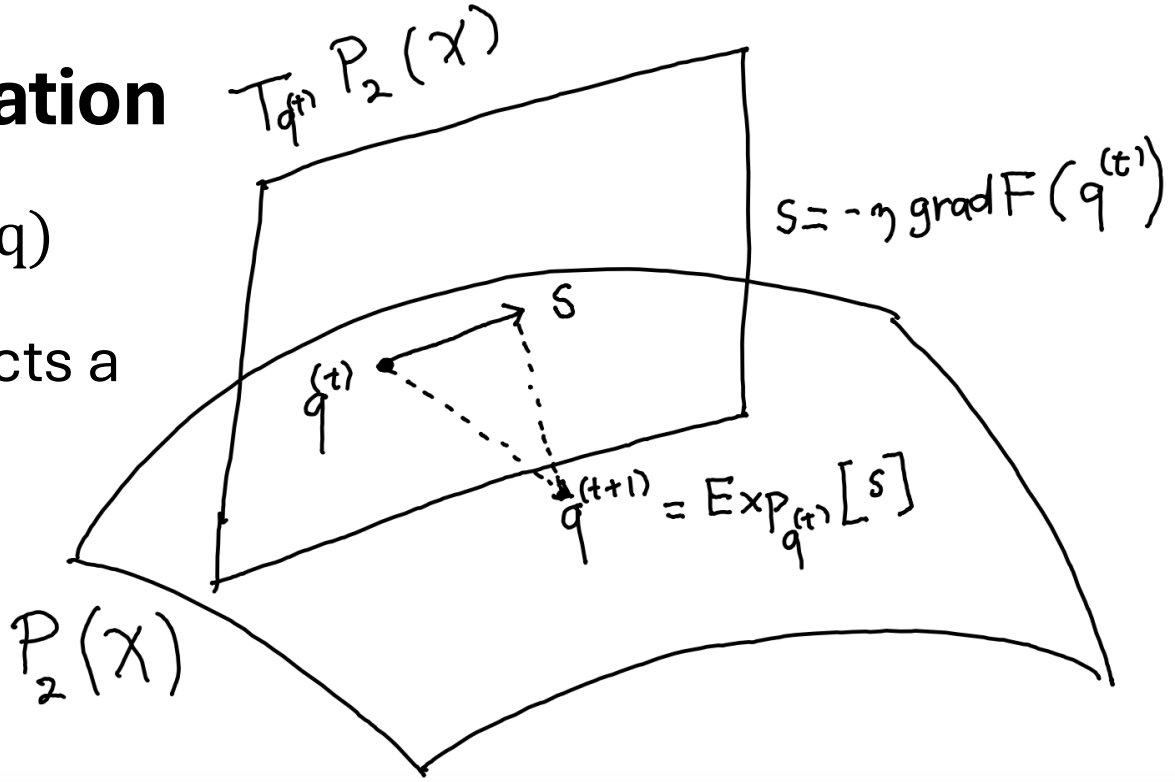
$$\lim_{h \to 0} \frac{1}{h}[F(\gamma(h)) - F(q^{(t)})] = \int \delta F(q^{(t)})(\mathbf{x}) s(\mathbf{x}) d\mathbf{x}$$

- **Continuity equation**: $s(\mathbf{x}) + \mathbf{div}\left(q^{(t)}(\mathbf{x})\mathbf{v}(\mathbf{x})\right) = 0, \mathbf{v}(x) = d\mathbf{x}/dt$

Related work: Distributional Optimization

Optimization over distribution space: $\min_{q \in \mathcal{P}_2(\mathcal{X})} F(q)$

- Wasserstein gradient descent (**WGD**) constructs a sequence of $\{q^{(t)}\}$ to decrease F
- $\mathcal{T}_{q^{(t)}}\mathcal{P}_2(\mathcal{X})$: tangent space at $q^{(t)}$
- $s \in \mathcal{T}_{q^{(t)}}\mathcal{P}_2(\mathcal{X})$: tangent vector at $q^{(t)}$



- Let $\gamma: [0,1] \to \mathcal{P}_2(\mathcal{X})$ be a curve satisfying $\gamma(0) = q^{(t)}$ and $\gamma'(0) = s$, **first variation** $\boldsymbol{\delta F(q^{(t)})}$ is defined as

$$\lim_{h \to 0} \frac{1}{h}[F(\gamma(h)) - F(q^{(t)})] = \int \delta F(q^{(t)})(\mathbf{x}) s(\mathbf{x}) d\mathbf{x}$$

- **Continuity equation**: $s(\mathbf{x}) + \mathbf{div}\left(q^{(t)}(\mathbf{x})\mathbf{v}(\mathbf{x})\right) = 0, \mathbf{v}(x) = d\mathbf{x}/dt$
- **Wasserstein gradient**: $\operatorname{grad} F(q^{(t)})(\mathbf{x}) + \mathbf{div}\left(q^{(t)}(\mathbf{x}) \nabla \delta F(q^{(t)})(\mathbf{x})\right) = 0$

Finding Pereto Stationary Distributions for MODO

- Optimization over distribution space: $\min_{q \in \mathcal{P}_2(\mathcal{X})} \mathbf{F}(q) \triangleq [F_1(q), F_2(q), \dots, F_K(q)]$
- **Pareto optimality.** Consider $p, q \in \mathcal{P}_2(\mathcal{X})$, say q dominates p if $F_k(q) \geq F_k(p)$ for $k = 1,2, \dots, K$, and $\mathbf{F}(p) \neq \mathbf{F}(q)$. $q \in \mathcal{P}_2(\mathcal{X})$ is Pareto optimal if for all $p \in \mathcal{P}_2(\mathcal{X})$, q is not dominated by p.
- **Pareto Stationarity.** $q \in \mathcal{P}_2(\mathcal{X})$ is Pareto stationary if

$$\min_{\mathbf{w} \in \mathcal{W}} \langle \operatorname{grad} \mathbf{F}(q)\mathbf{w}, \operatorname{grad} \mathbf{F}(q)\mathbf{w} \rangle_q = 0$$

Further, $q \in \mathcal{P}_2(\mathcal{X})$ is ϵ-Pareto stationary if

$$\min_{\mathbf{w} \in \mathcal{W}} \langle \operatorname{grad} \mathbf{F}(q)\mathbf{w}, \operatorname{grad} \mathbf{F}(q)\mathbf{w} \rangle_q \leq \epsilon^2$$

where $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^K | \mathbf{w} \succcurlyeq 0, \mathbf{w}^\intercal 1 = 1\}$, $\operatorname{grad} \mathbf{F}(q) = [\operatorname{grad} F_1(q), \dots, \operatorname{grad} F_K(q)]$, and $\operatorname{grad} \mathbf{F}(q)\mathbf{w} = \sum_{k=1}^K w_k \operatorname{grad} F_k(q)$

Multiple Wasserstein Gradient Descent (MWGraD) for MODO

Update Distributions via Velocities

- Goal: Constructs a sequence of $\{q^{(t)}\}_{t=0}^{T}$, starting from a simple $q^{(0)}$ (e.g. Gaussian), minimizing all objectives simultaneously

- Particle update:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \alpha \mathbf{v}^{(t)}(\mathbf{x}^{(t)})$$

where $\mathbf{x}^{(t)} \sim q^{(t)}$, $\mathbf{v}^{(t)}$ is the velocity field, and α is stepsize.

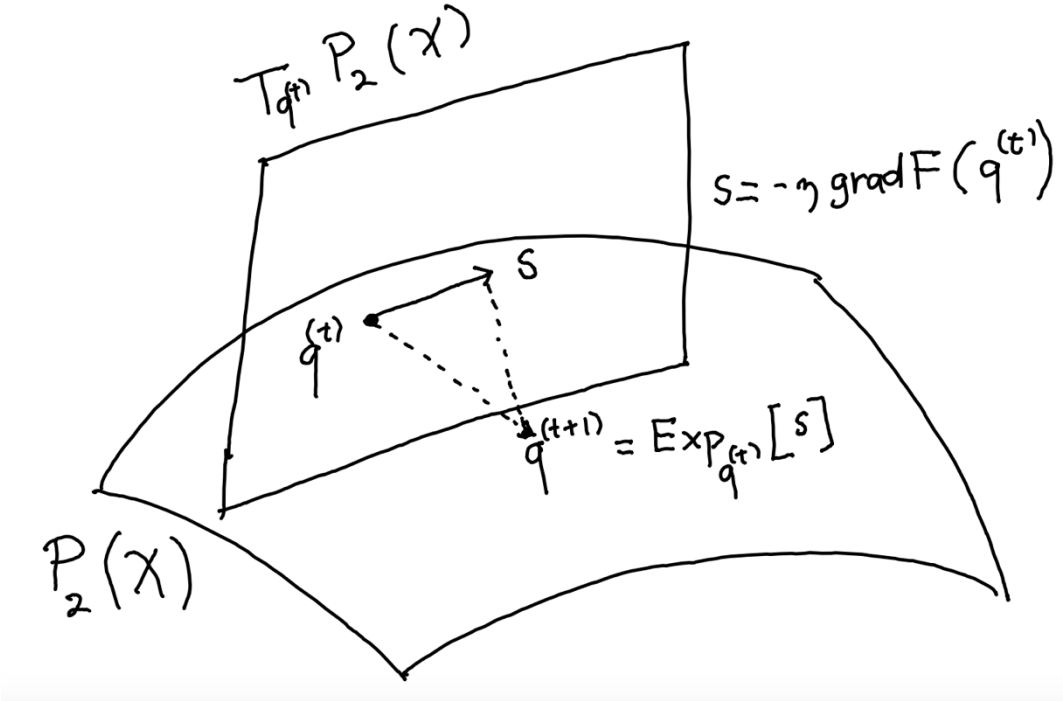
- Tangent vector $s^{(t)}$ associated with $\mathbf{v}^{(t)}$:

$$s^{(t)}\left(\mathbf{x}^{(t)}\right) = \mathbf{div}\left(q^{(t)}\left(\mathbf{x}^{(t)}\right)\mathbf{v}^{(t)}\left(\mathbf{x}^{(t)}\right)\right) \in \mathcal{T}_{q^{(t)}}\mathcal{P}_2(\mathcal{X})$$

Multiple **W**asserstein **Gra**dient **D**escent **(MWGraD)** for MODO

Velocity-based reformulation of MODO

- Let $\gamma: [0,1] \to \mathcal{P}_2(\mathcal{X})$ be a curve satisfying $\gamma(0) = \mathrm{q}^{(\mathrm{t})}$ and $\gamma'(0) = s$



- Goal: find $\mathrm{s}^{(\mathrm{t})}$ that maximize the minimum decrease across all objectives

$$\max_{\mathrm{s}} \min_{\mathrm{k}\in[\mathrm{K}]} \frac{1}{h}\left(\mathrm{F_k}\left(\mathrm{q}^{(\mathrm{t})}\right) - \mathrm{F_k}\left(\gamma(\mathrm{h})\right)\right)$$

Multiple **W**asserstein **Gra**dient **D**escent **(MWGraD)** for MODO

Reformulation of MODO

- Aim: find $s^{(t)}$ that maximize the minimum decrease across all objectives

$$\max_{s} \min_{k\in[K]} \frac{1}{h}\left(F_k\left(q^{(t)}\right) - F_k\left(\gamma(h)\right)\right)$$

First variation:

$$\lim_{h\to 0} \frac{1}{h}\left[F_k\left(\gamma(h)\right) - F_k\left(q^{(t)}\right)\right] = \int \delta F_k\left(q^{(t)}\right)(\mathbf{x})\, s(\mathbf{x})\, d\mathbf{x}$$

Continuity equation:

$$s(\mathbf{x}) = \mathbf{div}\left(q^{(t)}(\mathbf{x})\mathbf{v}(\mathbf{x})\right)$$

$$\max_{\mathbf{v}\in\boldsymbol{\mathcal{V}}} \min_{k\in[K]} \left\{ \int \left\langle \nabla \delta F_k\left(q^{(t)}\right)(\mathbf{x}), \mathbf{v}(\mathbf{x})\right\rangle \mathbf{q}^{(t)}(\mathbf{x})\, d\mathbf{x} \right\}$$

Multiple Wasserstein Gradient Descent (MWGraD) for MODO

Reformulation of MODO

- Aim: find $s^{(t)}$ that maximize the minimum decrease across all objectives

$$\max_{s} \min_{k\in[K]} \frac{1}{h}\left(F_k\left(q^{(t)}\right) - F_k\left(\gamma(h)\right)\right)$$

First variation:

$$\lim_{h\to 0} \frac{1}{h}\left[F_k\left(\gamma(h)\right) - F_k\left(q^{(t)}\right)\right] = \int \delta F_k\left(q^{(t)}\right)(\mathbf{x})\, s(\mathbf{x})\, d\mathbf{x}$$

Continuity equation:

$$s(\mathbf{x}) = \mathbf{div}\left(q^{(t)}(\mathbf{x})\mathbf{v}(\mathbf{x})\right)$$

$$\max_{\mathbf{v}\in\boldsymbol{\mathcal{V}}} \min_{k\in[K]} \left\{ \int \left\langle \nabla \delta F_k\left(q^{(t)}\right)(\mathbf{x}), \mathbf{v}(\mathbf{x})\right\rangle \mathbf{q}^{(t)}(\mathbf{x})\, d\mathbf{x} - \underbrace{\frac{1}{2}\int \langle \mathbf{v}(\mathbf{x}), \mathbf{v}(\mathbf{x})\rangle \mathbf{q}^{(t)}(\mathbf{x})\, d\mathbf{x}}_{\text{Regularization term}} \right\}$$

Multiple Wasserstein Gradient Descent (MWGraD) for MODO

Reformulation of MODO: at each iteration t, find $\mathbf{v}^{(t)}$ that optimize:

$$\max_{\mathbf{v}\in\boldsymbol{\mathcal{V}}} \min_{k\in[K]} \left\{ \int \left\langle \nabla \delta F_k\left(q^{(t)}\right)(\mathbf{x}), \mathbf{v}(\mathbf{x}) \right\rangle \mathbf{q}^{(t)}(\mathbf{x}) d\mathbf{x} - \frac{1}{2} \int \langle \mathbf{v}(\mathbf{x}), \mathbf{v}(\mathbf{x}) \rangle \mathbf{q}^{(t)}(\mathbf{x}) d\mathbf{x} \right\}$$

Theorem 1. The solution $\mathbf{v}^{(t)}$ is given by:

$$\mathbf{v}^{(t)}(\mathbf{x}) = \mathbf{V}^{(t)}(\mathbf{x})\mathbf{w}^* = \underbrace{\sum_{k=1}^{K} w_k^* \mathbf{v}_k^{(t)}(\mathbf{x})}_{\text{Convex combination of velocities}}, \text{ for } \forall \mathbf{x}$$

where $\mathbf{v}_k^{(t)}(\mathbf{x}) = \nabla \delta F_k\left(q^{(t)}\right)(\mathbf{x})$, for $k \in [K]$, $\mathbf{V}^{(t)}(\mathbf{x}) = \left[\mathbf{v}_1^{(t)}(\mathbf{x}), \mathbf{v}_2^{(t)}(\mathbf{x}), \ldots, \mathbf{v}_K^{(t)}(\mathbf{x})\right]$, and

$$\mathbf{w}^* = \underset{w\in\mathcal{W}}{\operatorname{argmin}} \underbrace{\frac{1}{2} \int \left\| \mathbf{V}^{(t)}(\mathbf{x})\mathbf{w} \right\|_2^2 \mathbf{q}^{(t)}(\mathbf{x}) d\mathbf{x}}_{\text{Squared norm of combined velocity}}$$

Multiple **W**asserstein **Gra**dient **D**escent **(MWGraD)** for MODO

- **Practical implementation**

1) Particles approximation: use $\left\{\mathbf{x}_i^{(t)}\right\}_{i=1}^{m}$ to approximate $q^{(t)}$, and update via:

$$\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} - \alpha \mathbf{v}^{(t)}(\mathbf{x}_i^{(t)})$$

2) Velocity Estimation: $\mathbf{v}^{(t)}(\mathbf{x}) = \mathbf{V}^{(t)}(\mathbf{x})\mathbf{w}^* = \sum_{k=1}^{K} w_k^* \nabla \delta F_k\left(q^{(t)}\right)(\mathbf{x})$,

Estimate $\delta F_k\left(q^{(t)}\right)$ using: **SVGD** [Liu et al., 2016], **Blob** method [Carrillo et al., 2019], **Neural network** (or variational transport) [Liu et al 2021]

3) Min-norm oracle (Approximate): $\mathbf{w}^* = \underset{w \in \mathcal{W}}{\operatorname{argmin}} \frac{1}{2} \int \left\| \mathbf{V}^{(t)}(\mathbf{x})\mathbf{w} \right\|_2^2 \mathbf{q}^{(t)}(\mathbf{x}) d\mathbf{x}$

Instead of solving exactly, take a projected gradient descent step:

$$\mathbf{w}^{(t+1)} = \Pi_{\mathcal{W}}\left[\mathbf{w}^{(t)} - \beta \Delta \mathbf{w}^{(t)}\right],$$

where $\Delta \mathbf{w}^{(t)}$ is the gradient of min-norm oracle w.r.t. $\mathbf{w}$, β is the step size.

Multiple Wasserstein Gradient Descent (MWGraD) for MODO

Assumption 1 (Wasserstein gradient error) The deviation between the true velocity $\mathbf{v}_k^{(t)} = \nabla \delta F_k\left(q^{(t)}\right)$ and the approximation $\tilde{\mathbf{v}}_k^{(t)}$ is upper bounded.

$$\mathbb{E}_{q^{(t)}} \left\| \tilde{\mathbf{v}}_k^{(t)} - \mathbf{v}_k^{(t)} \right\|_2^2 \leq \sigma^2$$

Assumption 2 (Geodesic smoothness) F_k is geodesically ℓ_k-smooth with respect to the 2-Wasserstein distance, for $k \in [K]$, in the sense that: for $p, q \in \mathcal{P}_2(\mathcal{X})$,

$$F_k(q) \leq F_k(p) + \left\langle \mathrm{grad} F_k(p), \mathrm{Exp}_p^{-1}(q) \right\rangle_p + \frac{\ell_k}{2} \mathcal{W}_2^2(p, q)$$

Theorem 2. Let Assumptions 1 and 2 hold, and $\epsilon > 0$ be a small constant. Set $\alpha \leq \mathcal{O}(\epsilon^2)$, $\beta \leq \mathcal{O}(\epsilon^2)$, $T \geq \max\left\{\Theta\left(\frac{1}{\alpha\epsilon^2}\right), \Theta\left(\frac{1}{\beta\epsilon^2}\right)\right\}$, we then have that

$$\min_{0 \leq t \leq T-1} \left\langle \mathrm{grad}\, \mathbf{F}\left(q^{(t)}\right)\mathbf{w}^{(t)}, \mathrm{grad}\, \mathbf{F}\left(q^{(t)}\right)\mathbf{w}^{(t)} \right\rangle_{q^{(t)}} \leq \frac{1}{T} \sum_{t=0}^{T} \left\langle \mathrm{grad}\, \mathbf{F}\left(q^{(t)}\right)\mathbf{w}^{(t)}, \mathrm{grad}\, \mathbf{F}\left(q^{(t)}\right)\mathbf{w}^{(t)} \right\rangle_{q^{(t)}}$$

$$\leq \epsilon^2 + 3\sigma^2$$

convergence with exact computation of velocities

approximation error of velocities

Experiments: Synthetic dataset

Objective: $\min_{q \in \mathcal{P}_2(\mathcal{X})} [\mathbf{KL}(q, p_1), \mathbf{KL}(q, p_2), \ldots, \mathbf{KL}(q, p_K)]$ (**Multi-target sampling**)

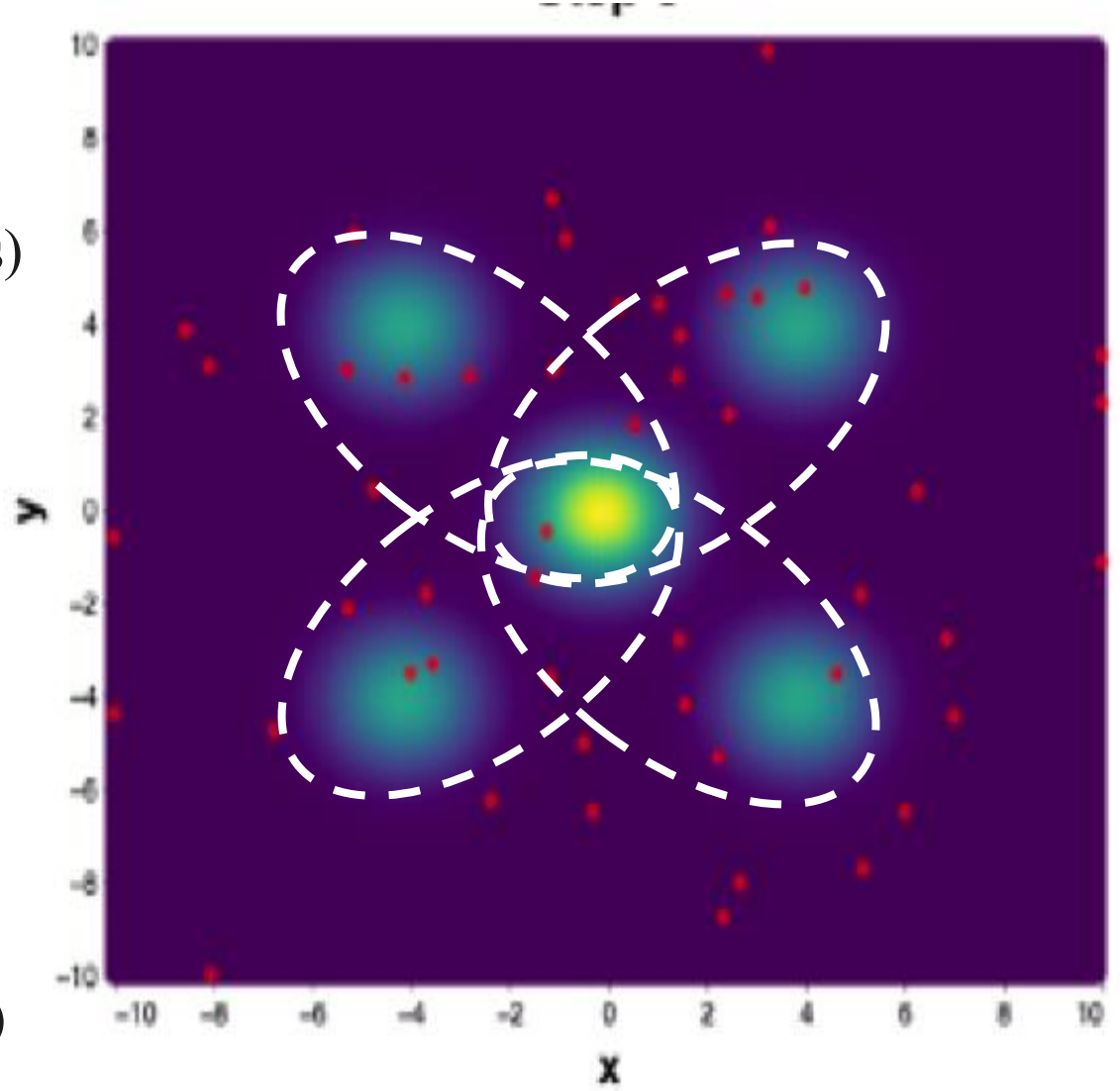
Setup:

- K= 4 target distributions (each a mixture of 2 Gaussians)
- Targets share a high-density region
- Approximate q using 50 particles initialized from a standard normal

Methods compared:

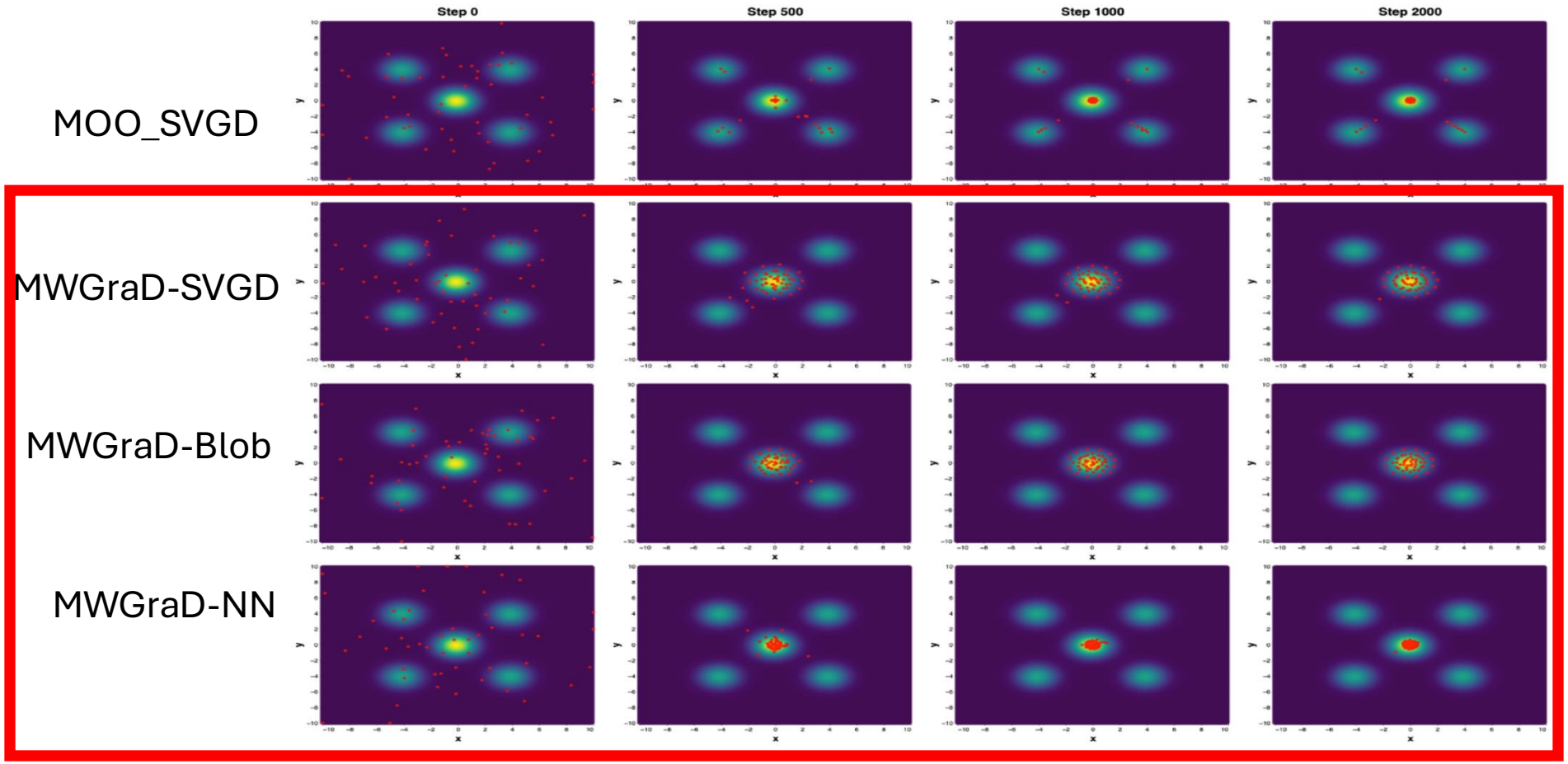
- **MOO-SVGD** [Liu et al., 2021]
- **MWGraD Variants** (*ours*):

  - MWGraD with SVGD (**MWGraD-SVGD**)
  - MWGraD with Blob method (**MWGraD-Blob**)
  - MWGraD with neural networks (**MWGraD-NN**)



Experiments: Synthetic dataset

Objective: $\min_{q \in \mathcal{P}_2(\mathcal{X})} [\mathbf{KL}(q, p_1), \mathbf{KL}(q, p_2), \ldots, \mathbf{KL}(q, p_K)]$



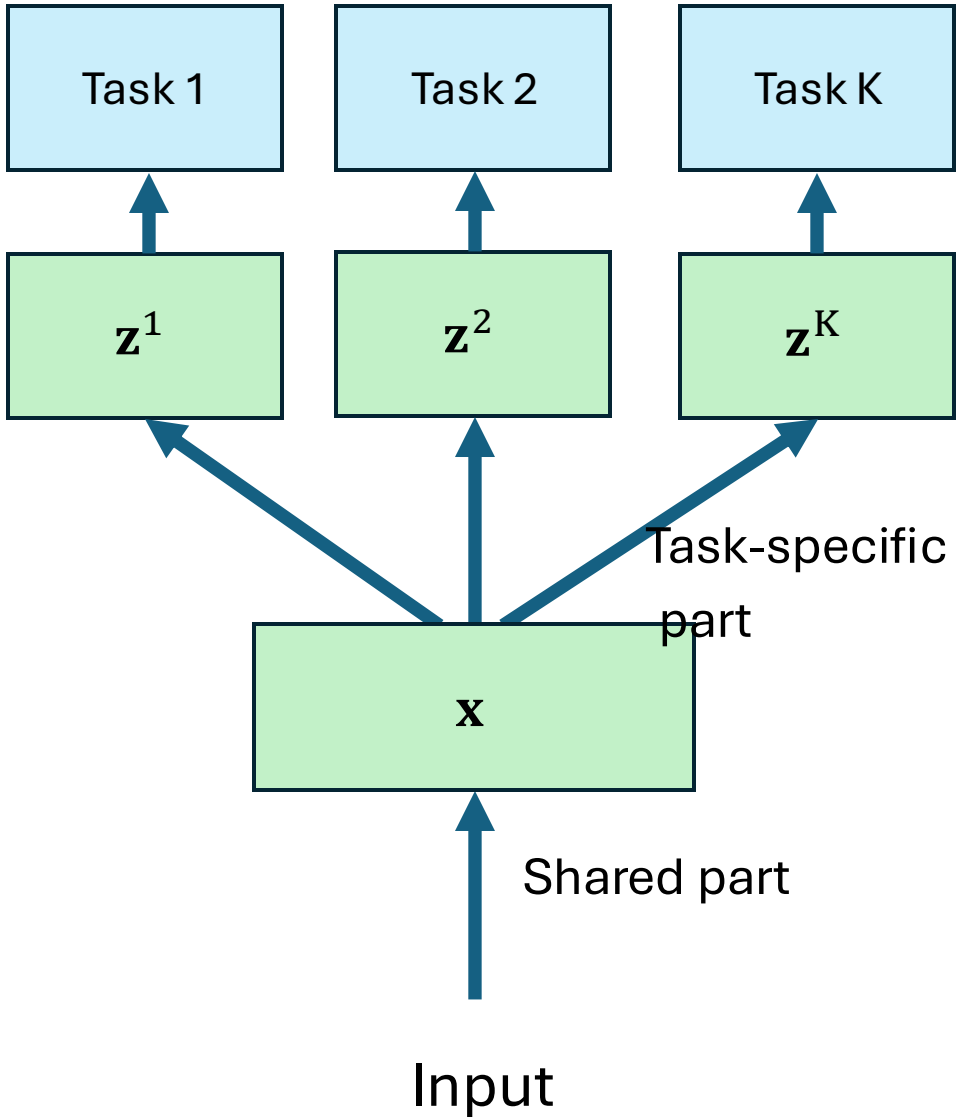
Ours

MWGraD tends to cover the joint high density of objective functions

Experiments: Multi-task learning

Bayesian multi-task learning:

- Assume K prediction tasks and a given dataset D
- Parameters for each task k: shared part: $\mathbf{x}$, task-k-specific part: $\mathbf{z}^k$
- Aim: estimate/sample from $\mathrm{p}(\mathbf{x}, \mathbf{z}^k \,|\mathrm{D})$ for each task k



Procedure: repeat (1) and (2)

(1) Fix $\mathbf{x}$, sample from $\mathrm{p}(\mathbf{z}^k \,|\, \mathbf{x}, \mathrm{D})$ for each task k

(by **SVGD** [Liu et al., 2016])

(2) Fix $\mathbf{z}^1, \mathbf{z}^2, .., \mathbf{z}^K$, perform multi-target sampling from multiple posteriors

$$\mathrm{p}(\mathbf{x}|\ \mathbf{z}^1, \mathrm{D}), \mathrm{p}(\mathbf{x} \,|\, \mathbf{z}^2, \mathrm{D}), \ldots, \mathrm{p}(\mathbf{x} \,|\mathbf{z}^K, \mathrm{D})$$

(by **MWGraD** variants)

Experiments: Multi-task learning

- **Datasets**: Multi-Fashion-MNIST, Multi-MNIST and Multi-Fashion. Each of them consists of 120,000 training and 20,000 testing images from MNIST and FashionMNIST by overlaying an image on top of another.
- Number of tasks: K = 2, Number of particles: m= 5. Run 2000 updates of particles

Ours

Datasets	Tasks	MGDA	MOO-SVGD	MT-SGD	MWGraD -SVGD	MWGraD -Blob	MWGraD -NN
Multi-Fashon+MNIST	#1	94.4±0.6	94.8±0.4	96.2±0.3	95.7±0.4	**96.7**±0.5	95.9±0.4
	#2	85.5±0.5	85.6±0.2	87.8±0.6	88.9±0.6	**92.5**±0.4	88.2±0.3
Multi-MNIST	#1	93.4±0.4	93.1±0.3	94.4±0.5	94.5±0.4	**97.6**±0.2	**97.7**±0.5
	#2	91.8±0.6	91.2±0.2	92.9±0.5	93.2±0.6	**96.7**±0.5	95.5±0.4
Multi-Fashion	#1	84.1±0.8	83.8±0.8	84.9±0.6	85.1±0.7	86.8±0.3	**87.2**±0.4
	#2	83.3±0.4	83.1±0.3	84.6±0.5	84.3±0.4	**87.2**±0.5	85.3±0.6

Table 1: Experimental results on Multi-Fashion+MNIST, Multi-MNIST, and Multi-Fashion. We report the ensemble accuracy (higher is better) averaged over three independent runs with different initializations.

Conclusions

- Addressed the MODO problem: simultaneously minimizing multiple functions of probability distributions

- Introduced MWGraD, an iterative particle-based algorithm for solving MODO

- Provided theoretical analyses and presented experimental results on both synthetic and real-world datasets, demonstrating the effectiveness of MWGraD in identifying the joint high-density regions of objective functions

References

[1] Nguyen et al., Multiple Wasserstein gradient descent for multi-objective distributional optimization, UAI 2025.

[2] Nguyen et al., Moreau-Yoshida variational transport: a general framework for solving regularized distributional optimization problems, *Machine learning 2024*.

[3] Nguyen et al., Mirror variational transport: a Particle-based algorithm for distributional optimization on constrained domains, *Machine learning 2023*.

[4] Liu et al., Profiling pareto-front with multi-objective Stein variational gradient descent, *NeurIPS 2021*.

[5] Liu et al., Infinite-dimensional optimization for zero-sum games via variational transport, *ICML 2021*.

[6] Liu et al., Stein variational gradient descent: a general purpose Bayesian inference algorithm, *NeurIPS 2016*.

Thank you for listening

Visit our poster **ID:128**

Appendix : Multiple gradient descent algorithm (MGDA)

- **MOO**

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbf{f}(\mathbf{x}) \triangleq [\mathrm{f}_1(\mathbf{x}), \mathrm{f}_2(\mathbf{x}), \dots, \mathrm{f}_\mathrm{K}(\mathbf{x})]$$

- **Pareto Optimality**. For two solution $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, say $\mathbf{y}$ dominates $\mathbf{x}$ iff $\mathrm{f}_\mathrm{k}(\mathbf{y}) \geq \mathrm{f}_\mathrm{k}(\mathbf{x})$ for $\mathrm{k} = 1,2, \dots, \mathrm{K}$, and $\mathbf{f}(\mathbf{y}) \neq \mathbf{f}(\mathbf{x})$. $\mathbf{x}$ is Pareto optimal if no other solution dominates it.

- **Pareto stationary solution**. $\mathbf{x}^* \in \mathcal{X}$ is Pareto stationary iff there exists $\mathbf{w} \in \mathcal{W}$ such that

$$\nabla\mathbf{f}(\mathbf{x}^*)\mathbf{w} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{w}_\mathbf{k} \nabla\mathrm{f}_\mathrm{k}(\mathbf{x}^*) = 0,$$

where

$$\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^\mathrm{K} | \mathbf{w} \succcurlyeq 0, \mathbf{w}^\intercal 1 = 1\}, \nabla\mathbf{f}(\mathbf{x}) = [\nabla\mathrm{f}_1(\mathbf{x}), \nabla\mathrm{f}_2(\mathbf{x}), \dots, \nabla\mathrm{f}_\mathrm{K}(\mathbf{x})].$$

Appendix: Multiple gradient descent algorithm (MGDA)

- is iterative and gradient-based
- $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \alpha \mathbf{d}^{(t)}$ ($\mathbf{d}^{(t)}$ is search/update direction at t-th iteration).
- **Key idea**: find $\mathbf{d}^{(t)}$ that maximizes the minimum decrease across all the objectives by solving:

(Primal)
$$\mathbf{d}^{(t)} = \underset{\mathbf{d} \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \max_{k \in [K]} \left\langle \nabla f_k\left(\mathbf{x}^{(t)}\right), \mathbf{d} \right\rangle + \frac{1}{2} \|\mathbf{d}\|_2^2 \right\}$$

(Dual)
$$\mathbf{w}^{(t)} = \underset{\mathbf{w} \in \mathcal{W}}{\operatorname{argmin}} \left\| \sum_{k=1}^{K} w_k \nabla f_k\left(\mathbf{x}^{(t)}\right) \right\|_2^2$$

Then, $\mathbf{d}^{(t)} = \sum_{k=1}^{K} \mathbf{w}_k^{(t)} \nabla f_k\left(\mathbf{x}^{(t)}\right) = \nabla \mathbf{f}\left(\mathbf{x}^{(t)}\right) \mathbf{w}^{(t)}$