Multi-Objective Optimization via Wasserstein-Fisher-Rao Gradient Flow

Yinuo Ren[1], Tesi Xiao[2], Tanmay Gangwani[2], Anshuka Rangi[2], Holakou Rahmanian[2], Lexing Ying[1,3], Subhajit Sanyal[2]

[1]Institute for Computational and Mathematical Engineering (ICME), Stanford University
[2]Amazon [3]Department of Mathematics, Stanford University

Multi-objective optimization

Consider optimizing over m possibly conflicting objective functions simultaneously:

$$\min_{\mathbf{x}\in\mathcal{D}} \mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_m(\mathbf{x}))$$

- **Pareto Optimality:** $\mathbf{x}^* \in \mathcal{D}$ s.t. $\nexists \mathbf{x}' \in \mathcal{D}$, $\begin{cases} f_i(\mathbf{x}') \le f_i(\mathbf{x}^*), & \forall i \in [m] \\ f_j(\mathbf{x}') < f_j(\mathbf{x}^*), & \exists j \in [m] \end{cases}$
- **Locally Pareto Optimal:** $\mathbf{x}^*$ is Pareto optimal in a neighborhood of $\mathbf{x}^*$
- **Pareto Front** $\mathcal{P}$: Set of all Pareto optimal solutions

Our goal is to find a set of *diversified* solutions $\hat{\mathcal{P}}$ that profiles the Pareto front $\mathcal{P}$.

Challenge: Previous methods struggle to deal with Pareto fronts with *complicated geometry*, that are *non-convex*, *non-smooth*, or even *discontinuous*, without any prior knowledge.

Wasserstein-Fisher-Rao Gradient Flow

Name	Metric	Gradient Flow
Wasserstein	$\inf\left\{\int_0^1\int \|\mathbf{v}_t\|^2 \mathrm{d}\rho_t \mathrm{d}t \middle\vert \partial_t\rho_t = -\nabla\cdot(\rho_t\mathbf{v}_t)\right\}$	$\partial_t\rho_t = \nabla\cdot(\rho_t\nabla\delta_\rho\mathcal{E}[\rho_t])$
Fisher-Rao	$\inf\left\{\int_0^1\int \widetilde{\beta}_t^2 \mathrm{d}\rho_t \mathrm{d}t \middle\vert \partial_t\rho_t = \rho_t\widetilde{\beta}_t\right\}$	$\partial_t\rho_t = -\rho_t\widetilde{\delta_\rho\mathcal{E}[\rho_t]}$
Wasserstein-Fisher-Rao	$\inf\left\{\int_0^1\int \left(\|\mathbf{v}_t\|^2 + \widetilde{\beta}_t^2\right) \mathrm{d}\rho_t \mathrm{d}t \middle\vert \partial_t\rho_t = -\nabla\cdot(\rho_t\mathbf{v}_t) + \rho_t\widetilde{\beta}_t\right\}$	$\partial_t\rho_t = \nabla\cdot(\rho_t\nabla\delta_\rho\mathcal{E}[\rho_t]) - \rho_t\widetilde{\delta_\rho\mathcal{E}[\rho_t]}$

where $\widetilde{\cdot} = \cdot - \mathbb{E}_{\rho_t}[\cdot]$, $\delta_\rho\mathcal{E}[\rho]$ is the Fréchet derivative of $\mathcal{E}[\rho]$

We perform the Wasserstein-Fisher-Rao gradient flow that evolve a probability distribution ρ_t over $\mathcal{D}$ to minimize a functional $\mathcal{E}[\rho_t]$ which should be designed such that its minimizers satisfies:

- **Global Pareto Optimality:** ρ^* should not cover those only *locally* Pareto optimal
- **Diversity:** ρ^* should be *close to* and *span the entirety* of $\mathcal{P}$

Methodology

Let $\mathcal{E}[\rho] = \alpha_1\mathcal{F}_1[\rho] + \alpha_2\mathcal{F}_2[\rho] + \beta\mathcal{G}[\rho] - \gamma\mathcal{H}[\rho]$, where each term is defined as follows:

- **Objective Functions:** ensure *local Pareto optimality*

$$\mathcal{F}_1[\rho] = \int_{\mathcal{D}} \|\mathbf{g}^\dagger(\mathbf{x})\|^2\rho(\mathrm{d}\mathbf{x}), \text{ where } \mathbf{g}^\dagger(\mathbf{x}) = \operatorname*{argmin}_{\|\mathbf{g}\|\le 1}\min_{i\in[m]} -\mathbf{g}^\top\nabla f_i(\mathbf{x})$$

Small $\|\mathbf{g}^\dagger\|$ indicates misalignment among the objective function, *i.e.* $\mathbf{x}$ is close to local Pareto optimality [2];

- **Dominance Potential:** promote *global Pareto optimality*

$$\mathcal{F}_2[\rho] = \int_{\mathcal{D}}\int_{\mathcal{P}} D(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y}))\mu_{\mathcal{P}}(\mathrm{d}\mathbf{y})\rho(\mathrm{d}\mathbf{x}),$$

where the asymmetric kernel $D(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y})) = \prod_{i=1}^m \max\{0, f_i(\mathbf{x}) - f_i(\mathbf{y})\}$ and is non-zero if and only if $\mathbf{x}$ is dominated by $\mathbf{y}$.

- **Entropy:** encourage *diversity* $-\mathcal{H}[\rho] = \int_{\mathcal{D}} \rho(\mathbf{x})\log\rho(\mathbf{x})\mathrm{d}\mathbf{x}$
- **Repulsive Potential:** encourage *diversity*

$$\mathcal{G}[\rho] = \frac{1}{2}\int_{\mathcal{D}\times\mathcal{D}} \rho(\mathrm{d}\mathbf{x})R(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y}))\rho(\mathrm{d}\mathbf{y}),$$

where the repulsive kernel $R(\mathbf{x}, \mathbf{y}) = \frac{1}{\|\mathbf{x}-\mathbf{y}\|}$ or $\exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}\right)$.

Theoretical Analysis

Theorem 1. *The following decay of the functional $\mathcal{E}[\rho_t]$ holds:*

$$\partial_t\mathcal{E}[\rho_t] = -\int_{\mathcal{D}} \rho_t\left\|\nabla\delta_\rho\mathcal{E}[\rho_t]\right\|^2 + \rho_t\widetilde{\delta_\rho\mathcal{E}[\rho_t]}^2\mathrm{d}\mathbf{x} \le 0.$$

Furthermore, if $\beta\wedge\gamma > 0$, the density ρ_t converges to the unique minimizer ρ^* of $\mathcal{E}[\rho]$, as $t\to\infty$.

Theorem 2. *Assume $\inf_{\mathbf{x}\in\mathcal{D}}\rho_0(\mathbf{x})/\rho^*(\mathbf{x}) \ge e^{-M}$ with $\beta = 0$, the following exponential convergence holds:*

$$\mathrm{KL}(\rho_t\|\rho^*) \le Me^{-\gamma t} + e^{-\gamma t + Me^{-\gamma t}}\mathrm{KL}(\rho_0\|\rho^*).$$

Algorithm

We adopt *interacting particle method*, discretize ρ_t by $\rho_t \approx \frac{1}{N}\sum_{k=1}^n \delta(\mathbf{x} - \mathbf{x}_k)$, and approximate the Wasserstein-Fisher-Rao gradient flow by the *splitting scheme* [3] that alternatively updates the following:

- **Overdamped Langevin Dynamics** (*Transportation*):

$$\partial_t\rho_t = \nabla\cdot\left(\rho_t\nabla\left(\delta_\rho\mathcal{F} + \delta_\rho\mathcal{G}[\rho_t]\right)\right) + \gamma\Delta\rho_t,$$

as a *Fokker-Planck equation*, corresponds to the following Langevin dynamics:

$$\mathrm{d}\mathbf{x}_t = -\nabla\left(\delta_\rho\mathcal{F} + \delta_\rho\mathcal{G}[\rho_t]\right)\mathrm{d}t + \sqrt{2\gamma}\mathrm{d}\mathbf{w}_t,$$

which can be discretized into

$$\mathbf{x}_k^{(\ell+1/2)} = \mathbf{x}_k^{(\ell)} - \frac{\tau}{2}\nabla\left(\delta_\rho\mathcal{F} + \delta_\rho\mathcal{G}[\rho_t]\right) + \sqrt{\gamma\tau}\varepsilon_k^{(\ell)}$$

- **Birth-Death Dynamic** (*Teleportation*): $\partial_t\log\rho_t = -\widetilde{\delta_\rho\mathcal{E}[\rho_t]} := -\Lambda_t$, where

$$\Lambda_{(\ell+1/2)\tau} \approx \delta_\rho\mathcal{E}[\rho_t](\mathbf{x}_k^{(\ell+1/2)}) - \frac{1}{N}\sum_{k'=1}^N \delta_\rho\mathcal{E}[\rho_t](\mathbf{x}_{k'}^{(\ell+1/2)}).$$

To update $\mathbf{x}_k^{(\ell+1/2)}$ to $\mathbf{x}_k^{(\ell+1)}$, depending on $\operatorname{sgn}\Lambda_{(\ell+1/2)\tau}$, one would

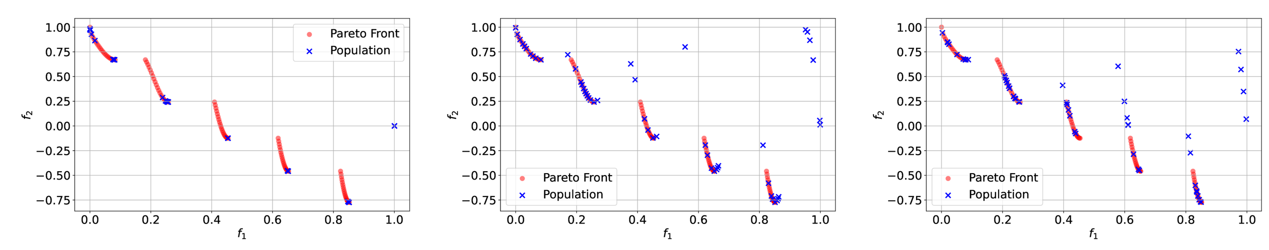
- Either duplicate it w/prob $\exp\left(-\Lambda_{(\ell+1/2)\tau}\tau/2\right) - 1$ and remove one random
- Or remove it w/prob $1 - \exp\left(-\Lambda_{(\ell+1/2)\tau}\tau/2\right)$ and duplicate one random

Main Methodological Takeaways

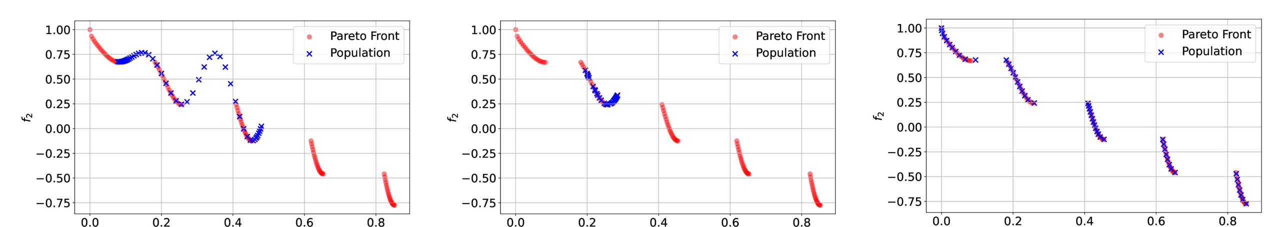
- *Transportation*: Langevin dynamics move the particles towards the Pareto front while keeping each other apart
- *Teleportation*: Birth-death dynamics eliminate the particles that are only locally Pareto optimal, ensuring *global Pareto optimality* even on *challenging* tasks with *complicated* Pareto fronts

Experiment Results

- ZDT3 Problem [6]:

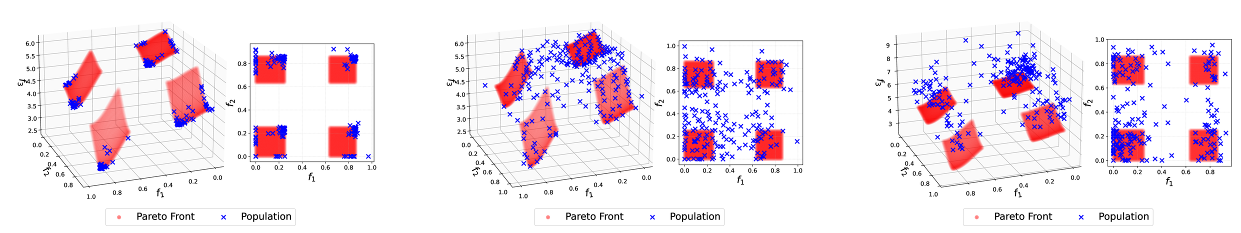


(a) Weighted Sum (b) MOO-SVGD (c) MOO-LD

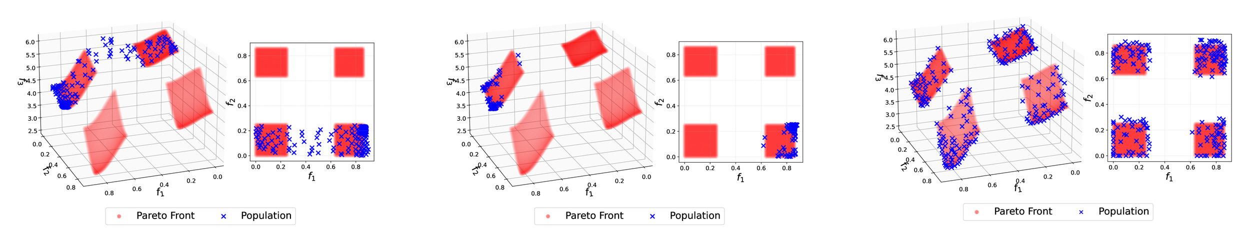


(d) COSMOS (e) GMOOAR-HV (f) Particle-WFR

- DTLZ7 Problem [1]:



(a) Weighted Sum (b) MOO-SVGD (c) MOO-LD

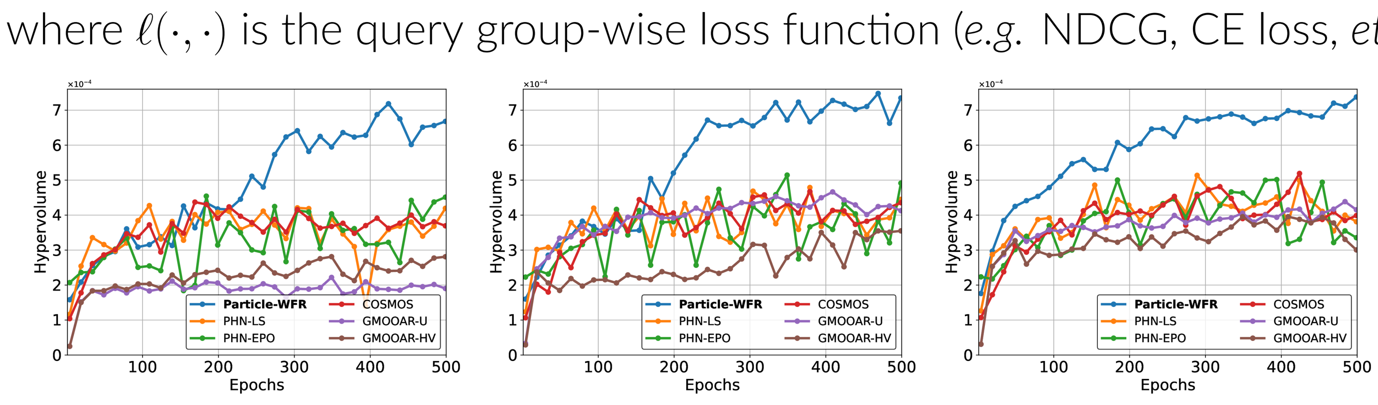


(d) COSMOS (e) GMOOAR-HV (f) Particle-WFR

- MSLR-WEB10K Dataset [5]: a *Learning-To-Rank (LTR)* [4] dataset:
  - Query groups: $\Psi = \{\Psi^{(p)}\}_{p=1}^{|\Psi|}$, $|\Psi| = 10^4$
  - Items : $|\Psi^{(p)}| = n^{(p)}$, and $\forall j \in [n^{(p)}]$, an item is characterized by a feature vector $\mathbf{x}_j^{(p)} \in \mathbb{R}^{d_f}$, and 6 associated relevance labels $y_j^{(p),i}$, $i \in [6]$
  - Feasible region $\mathcal{D}$: the space of *3-layer Multi-Layer Perceptrons (MLPs)*, parametrized by θ
  - Objective functions: loss functions corresponding to each label $\{y_j^{(p),i}\}_{j=1}^{n^{(p)}}$

$$\mathcal{L}_i(\theta;\Psi) = \frac{1}{|\Psi|}\sum_{p=1}^{|\Psi|}\ell\left(\{f_\theta(\mathbf{x}_j^{(p)})\}_{j=1}^{n^{(p)}}; \{y_j^{(p),i}\}_{j=1}^{n^{(p)}}\right),$$

where $\ell(\cdot,\cdot)$ is the query group-wise loss function (*e.g.* NDCG, CE loss, *etc.*)



(a) $N = 8$ (b) $N = 12$ (c) $N = 16$

where *hypervolume* is the volume of the dominated region of $\hat{\mathcal{P}}$ w.r.t. a reference point $\boldsymbol{r}$, higher is better

References

[1] K. Deb, L. Thiele, M. Laumanns, and E. Zitzler. Scalable multi-objective optimization test problems. In *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)*, volume 1, pages 825–830. IEEE, 2002.

[2] J.-A. Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012.

[3] T. O. Gallouët and L. Monsaingeon. A jko splitting scheme for kantorovich–fisher–rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130, 2017.

[4] T.-Y. Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

[5] T. Qin and T.-Y. Liu. Introducing letor 4.0 datasets. *arXiv preprint arXiv:1306.2597*, 2013.

[6] E. Zitzler, K. Deb, and L. Thiele. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary computation*, 8(2):173–195, 2000.

Check out our paper for more details!

{yinuoren, lexing}@stanford.edu, {tesixiao, gangwt, anshuka, holakou, subhajs}@amazon.com