Flowing Datasets with Wasserstein over Wasserstein Gradient Flows

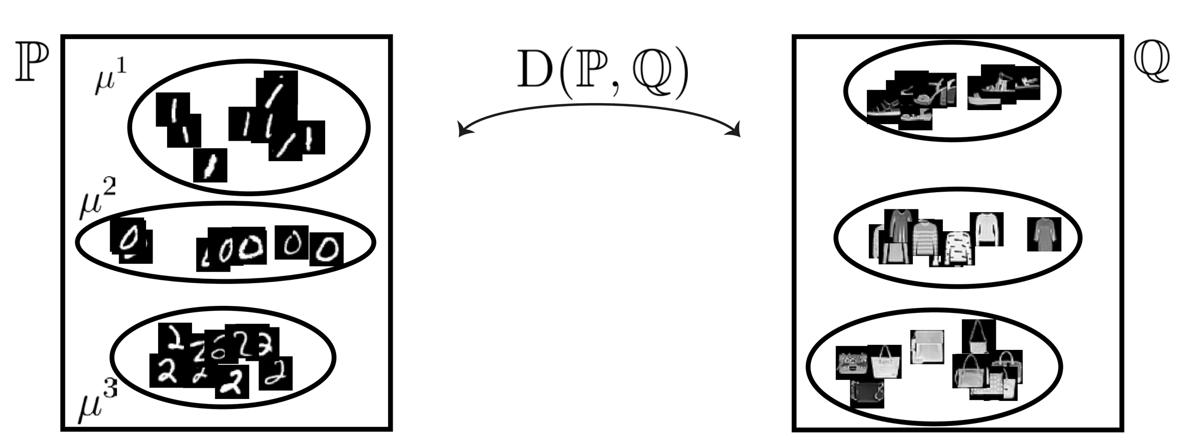
Clément Bonet*[1], Christophe Vauthier*[2], Anna Korba[1]

[1]ENSAE, CREST, Institut Polytechnique de Paris; [2]Université Paris-Saclay, Laboratoire de Mathématique d'Orsay

Contributions

Goal: move labeled dataset in a coherent way

- Labeled datasets modeled as $\mathbb{P}=\frac{1}{C}\sum_{c=1}^C \delta_{\mu^{c,n}} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ where $\mu^{c,n}=\frac{1}{n}\sum_{i=1}^n \delta_{x_i^c}$
- Endow $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ with OT distance WoW
- Minimize $\mathbb{F}:\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)\to\mathbb{R}$ using WoW gradient flows
- Application on image datasets



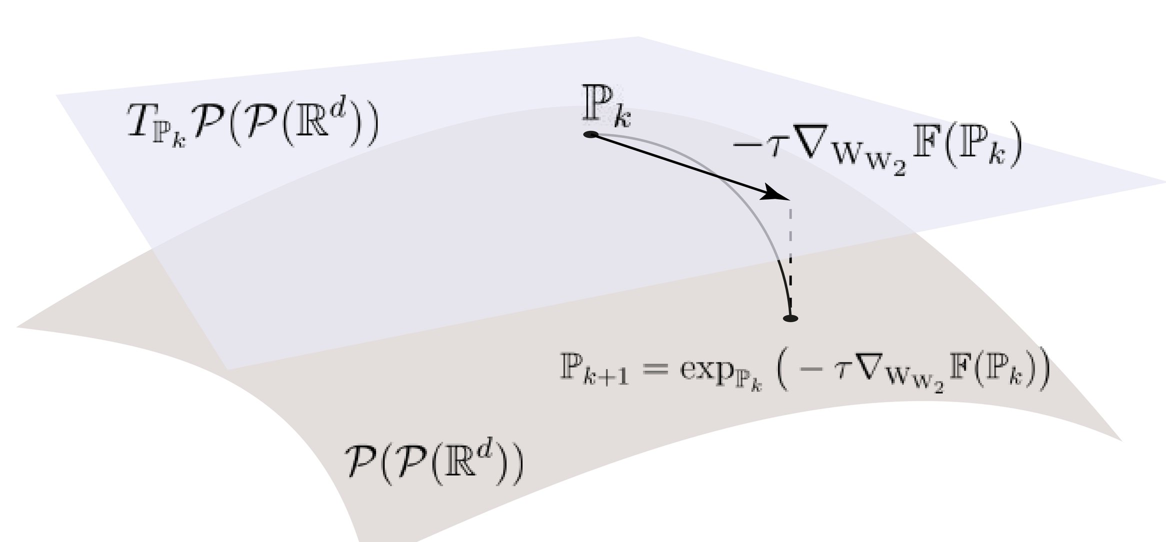
Wasserstein over Wasserstein Space

WoW distance: Let $\mathbb{P},\mathbb{Q}\in\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$,

$$W_{W_2}(\mathbb{P},\mathbb{Q})^2=\inf_{\Gamma\in\Pi(\mathbb{P},\mathbb{Q})}\int W_2^2(\mu,\nu)\ \mathrm{d}\Gamma(\mu,\nu)$$

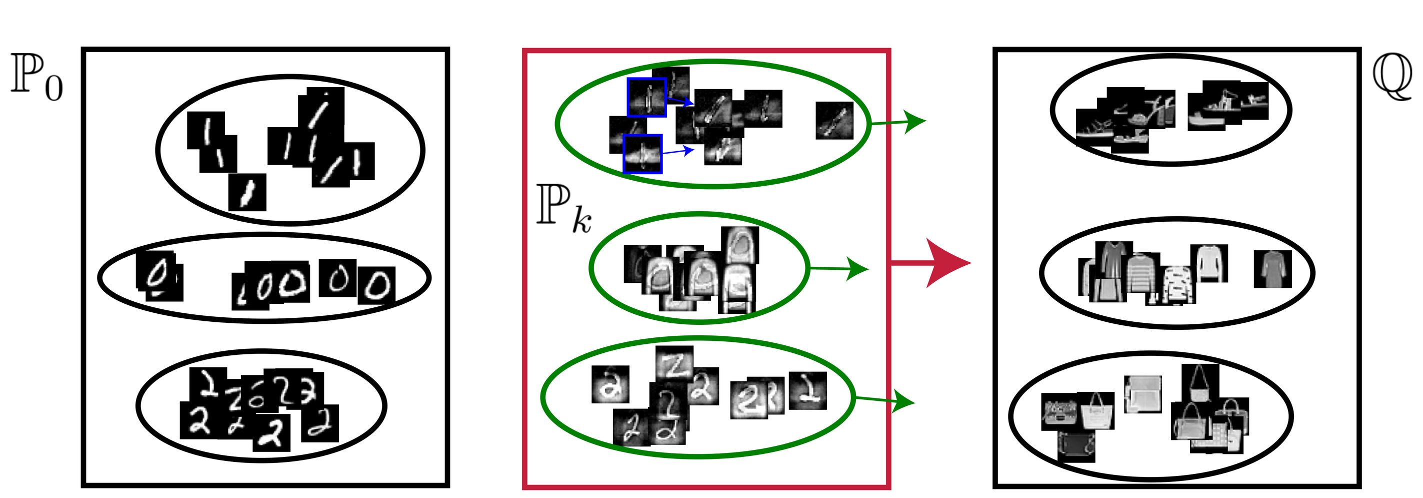
→ Riemannian structure

WoW Gradient Descent: $\mathbb{P}_{k+1}=\exp_{\mathbb{P}_k}\big(-\tau\nabla_{W_{W_2}}\mathbb{F}(\mathbb{P}_k)\big)$



In practice: For $\mathcal{M}=\mathbb{R}^d$, $\mathbb{P}_k=\frac{1}{C}\sum_{c=1}^C\delta_{\mu_k^{c,n}}$:

$$\forall k\geq 0,\ x_{i,k+1}^c=x_{i,k}^c-\tau\nabla_{W_{W_2}}\mathbb{F}(\mathbb{P}_k)(\mu_k^{c,n})(x_{i,k}^c)$$



Minimization of the MMD on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$

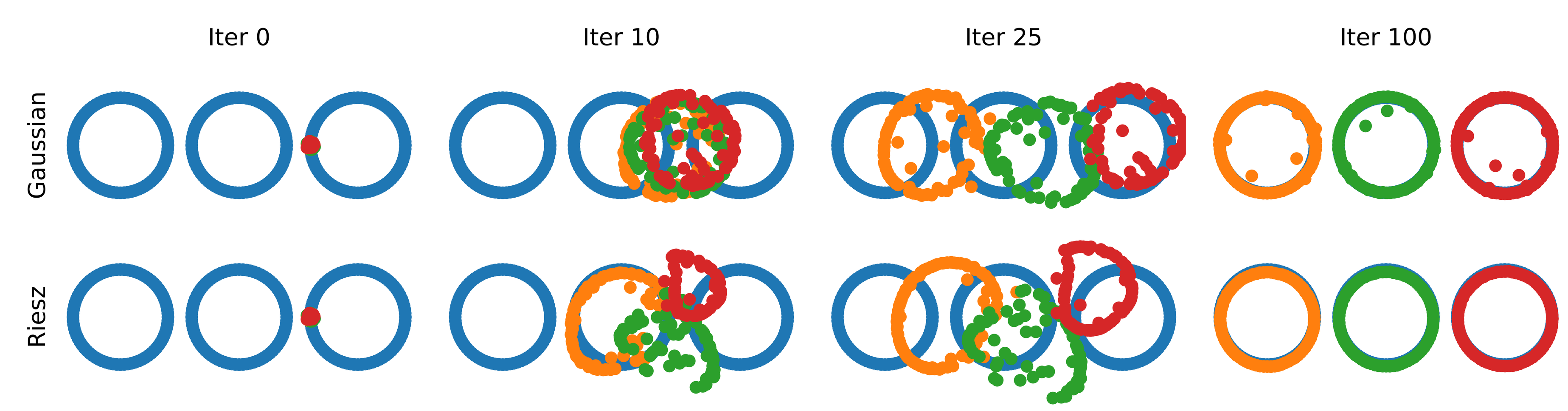
Goal: minimize $\mathbb{F}(\mathbb{P})=\frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P},\mathbb{Q})=\mathbb{V}(\mathbb{P})+\mathbb{W}(\mathbb{P})+\mathrm{cst}$, with $\mathbb{V}(\mathbb{P})=\int\mathcal{V}(\mu)\mathrm{d}\mathbb{P}(\mu)$, $\mathcal{V}(\mu)=-\int K(\mu,\nu)\mathrm{d}\mathbb{Q}(\nu)$, $\mathbb{W}(\mathbb{P})=\frac{1}{2}\iint K(\mu,\nu)\mathrm{d}\mathbb{P}(\mu)\mathrm{d}\mathbb{P}(\nu)$

SW distance: $\mathrm{SW}_2^2(\mu,\nu)=\int_{S^{d-1}}W_2^2(P^\theta_\#\mu,P^\theta_\#\nu)\mathrm{d}\sigma(\theta)$, $P^\theta(x)=\langle x,\theta\rangle$

Kernel: $K(\mu,\nu)=e^{-\frac{1}{2h}\mathrm{SW}_2^2(\mu,\nu)}$ (Gaussian) or $K(\mu,\nu)=-\mathrm{SW}_2(\mu,\nu)$ (Riesz)

Computational complexity: $O(C^2Ln\log n)$

$\nabla_{W_{W_2}}\mathbb{F}(\mathbb{P})(\mu^{c,n})(x_i^c)=nC\nabla_{i,c}F(\mathbf{x})$ for $\mathbf{x}=(x_i^c)_{i,c}$: obtained in closed-form or by auto-differentiation of $F(\mathbf{x}):=\mathbb{F}(\mathbb{P})$



Wasserstein over Wasserstein Gradients

For $(x,v)\in T\mathcal{M}$, define $\pi^{\mathcal{M}}\big((x,v)\big)=x$.

Couplings. For any $\gamma\in\mathcal{P}_2(T\mathcal{M})$, let $\phi^{\mathcal{M}}(\gamma)=\pi^{\mathcal{M}}_\#\gamma$, $\phi^{\exp}(\gamma)=\exp_\#\gamma$.

$$\exp_{\mathbb{P}}^{-1}(\mathbb{Q}):=\Big\{\tilde{\Gamma}\in\mathcal{P}_2\big(\mathcal{P}_2(T\mathcal{M})\big),\ \phi^{\mathcal{M}}_\#\tilde{\Gamma}=\mathbb{P},\ \phi^{\exp}_\#\tilde{\Gamma}=\mathbb{Q},\ \iint\|v\|_x^2\mathrm{d}\gamma(x,v)\mathrm{d}\tilde{\Gamma}(\gamma)=W_{W_2}^2(\mathbb{P},\mathbb{Q})\Big\}.$$

WoW Gradient

Let $\mathbb{F}:\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)\to\mathbb{R}$. $\mathbb{F}$ is WoW differentiable at $\mathbb{P}$ if there exists $\nabla_{W_{W_2}}\mathbb{F}(\mathbb{P}):\mathcal{P}_2(\mathcal{M})\to T\mathcal{P}_2(\mathcal{M})$ s.t. for any $\mathbb{Q}\in\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$, $\tilde{\Gamma}\in\exp_{\mathbb{P}}^{-1}(\mathbb{Q})$,

$$\mathbb{F}(\mathbb{Q})=\mathbb{F}(\mathbb{P})+\iint\langle\nabla_{W_{W_2}}\mathbb{F}(\mathbb{P})(\pi^{\mathcal{M}}_\#\gamma)(x),v\rangle_x\ \mathrm{d}\gamma(x,v)\mathrm{d}\tilde{\Gamma}(\gamma)+o\big(W_{W_2}(\mathbb{P},\mathbb{Q})\big).$$

Potentials: $\mathbb{V}(\mathbb{P})=\int\mathcal{F}(\mu)\mathrm{d}\mathbb{P}(\mu)$, $\nabla_{W_{W_2}}\mathbb{V}(\mathbb{P})=\nabla_{W_2}\mathcal{F}$

Interactions: $\mathbb{W}(\mathbb{P})=\iint\mathcal{W}(\mu,\nu)\mathrm{d}\mathbb{P}(\mu)\mathrm{d}\mathbb{P}(\nu)$,

$$\nabla_{W_{W_2}}\mathbb{W}(\mathbb{P})(\mu)=\int\big(\nabla_{W_{2,1}}\mathcal{W}(\mu,\nu)+\nabla_{W_{2,2}}\mathcal{W}(\mu,\nu)\big)\ \mathrm{d}\mathbb{P}(\nu)$$

For $K_\nu(\mu)=K(\mu,\nu)=e^{-\frac{1}{2h}\mathrm{SW}_2^2(\mu,\nu)}$, $\nabla_{W_{W_2}}\mathbb{F}(\mathbb{P})(\mu)=\int\nabla_{W_2}K_\nu(\mu)\ \mathrm{d}(\mathbb{P}-\mathbb{Q})(\nu)$, $\nabla_{W_2}K_\nu(\mu)=-\frac{1}{h}e^{-\frac{1}{2h}\mathrm{SW}_2^2(\mu,\nu)}\int_{S^{d-1}}\psi'_\theta(\langle x,\theta\rangle)\theta\ \mathrm{d}\sigma(\theta)$, $\psi'_\theta(u)=u-F^{-1}_{P^\theta_\#\nu}\big(F_{P^\theta_\#\mu}(u)\big)$.

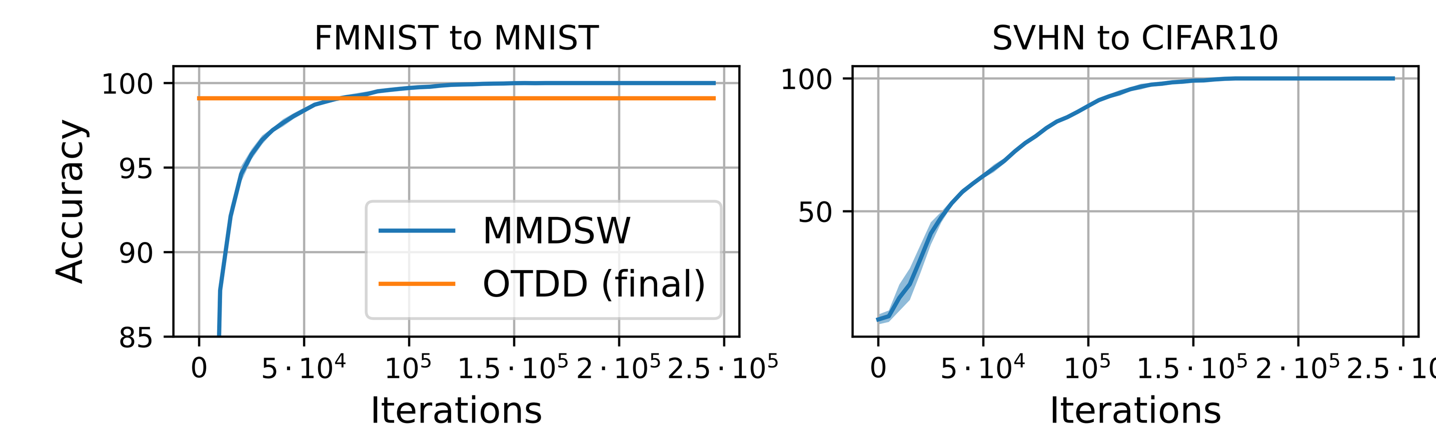
Tangent space: $T_{\mathbb{P}}\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)=\overline{\{\nabla_{W_2}\varphi,\ \varphi\in\mathrm{Cyl}\big(\mathcal{P}_2(\mathcal{M})\big)\}}$

Properties: There is at most one element in $\partial\mathbb{F}(\mathbb{P})\cap T_{\mathbb{P}}\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$. If $\xi\in\partial\mathbb{F}(\mathbb{P})\cap T\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$, then ξ is a strong differential of $\mathbb{F}$ at $\mathbb{P}$ (*i.e.* the Taylor expansion holds for any $\tilde{\Gamma}\in\mathcal{P}_2\big(\mathcal{P}_2(T\mathcal{M})\big)$ s.t. $\phi^{\mathcal{M}}_\#\tilde{\Gamma}=\mathbb{P}$).

Domain Adaptation

Minimize $\mathbb{F}(\mathbb{P})=\frac{1}{2}\mathrm{MMD}_K(\mathbb{P},\mathbb{Q})$ starting from $\mathbb{P}_0$ (FMNIST or SVHN) towards $\mathbb{Q}$ (MNIST or CIFAR10).

- Pretrain a classifier on $\mathbb{Q}$
- Monitor accuracy of the classifier along the flow



Applications

Dataset distillation. Synthesize a dataset $\mathbb{Q}=\frac{1}{n}\sum_{c=1}^C\delta_{\nu^{c,n}}$ (n big) with a dataset $\mathbb{P}=\frac{1}{C}\sum_{c=1}^C\delta_{\mu^{c,k}}$ (k small).

→ $\min_{\mathbb{P}}\ \mathbb{E}_{\theta,\omega}[\mathrm{MMD}_K^2(\phi^{\theta,\omega}_\#\mathbb{P},\phi^{\theta,\omega}_\#\mathbb{Q})]$ with $\phi^{\theta,\omega}(\mu)=\psi^\theta_\#\mathcal{A}^\omega_\#\mu$, $\mathcal{A}^\omega$ a random augmentation, ψ^θ a randomly initialized neural network.

Evaluation: train a classifier on the new synthetic dataset $\mathbb{P}$

Transfer learning (k-shot learning). Augment a dataset $\mathbb{Q}=\frac{1}{C}\sum_{c=1}^C\delta_{\nu^{c,k}}$ (k small) adding flowed samples starting from $\mathbb{P}_0=\frac{1}{C}\sum_{c=1}^C\delta_{\mu^{c,n}}$ a known bigger dataset.

→ $\min_{\mathbb{P}}\ \mathrm{MMD}_K^2(\mathbb{P},\mathbb{Q})$ starting from $\mathbb{P}=\mathbb{P}_0$

Evaluation: train a classifier on the augmented dataset $\hat{\mathbb{Q}}$

Dataset distillation

Dataset	k	$\psi^\theta=\mathcal{A}^\omega=\mathrm{Id}$		Baselines	
		DM	MMDSW	Random	Full data
MNIST	1	$61.1_{\pm6.5}$	**66.5**$_{\pm5.5}$	$55.8_{\pm2.0}$	99.4
	10	$88.2_{\pm2.8}$	**93.2**$_{\pm0.7}$	$92.2_{\pm1.1}$	
	50	$95.9_{\pm0.9}$	$97.0_{\pm0.2}$	$97.6_{\pm0.2}$	
FMNIST	1	$54.4_{\pm3.2}$	**60.0**$_{\pm4.1}$	$49.0_{\pm7.5}$	92.4
	10	$74.6_{\pm1.0}$	**76.7**$_{\pm1.0}$	$75.3_{\pm0.7}$	
	50	$81.3_{\pm0.5}$	**84.2**$_{\pm0.1}$	$83.2_{\pm0.2}$	

Transfer learning

Dataset	k	Train on $\mathbb{Q}$	MMDSW	OTDD	(Hua et al., 2023)
M to F	1	$26.0_{\pm5.3}$	**40.5**$_{\pm4.7}$	$30.5_{\pm4.2}$	$36.4_{\pm3.3}$
	5	$38.5_{\pm6.7}$	$61.5_{\pm4.6}$	$59.7_{\pm1.8}$	**62.7**$_{\pm1.1}$
	10	$53.9_{\pm7.9}$	$65.4_{\pm1.5}$	$64.0_{\pm1.4}$	**66.2**$_{\pm1.0}$
	100	$71.1_{\pm1.5}$	**74.7**$_{\pm0.8}$	-	$73.5_{\pm0.7}$
M to K	1	$18.4_{\pm3.1}$	**20.9**$_{\pm2.0}$	$18.8_{\pm2.1}$	$19.4_{\pm1.9}$
	5	$25.9_{\pm4.0}$	$37.4_{\pm2.2}$	$31.3_{\pm1.4}$	**39.0**$_{\pm1.0}$
	10	$30.9_{\pm4.6}$	**44.7**$_{\pm1.8}$	$34.1_{\pm0.9}$	$44.1_{\pm1.2}$
	100	$60.1_{\pm1.1}$	**66.8**$_{\pm0.8}$	$66.3_{\pm0.9}$	$62.4_{\pm1.2}$