Statistical and Geometrical properties of the regularized Kernel Kullback Leibler divergence

Clémentine Chazal, Anna Korba, Francis Bach

What is sampling in Maching Learning ?

It consists in approaching an unknown target probability distribution $q \in \mathscr{P}(\mathbb{R}^d)$ and to sample form it, i.e. generate $x_1, \ldots, x_n \sim q$.

What is sampling in Maching Learning ?

It consists in approaching an unknown target probability distribution $q \in \mathscr{P}(\mathbb{R}^d)$ and to sample form it, i.e. generate $x_1, \ldots, x_n \sim q$.

<u>Applications</u> :

- **Bayesian inference** : the density of q is known up to a normalization constant : $q = \frac{\tilde{q}}{Z}$.
- **<u>Generative modelling</u>** : Samples from q are known : $y_1, \ldots, y_m \sim q$.

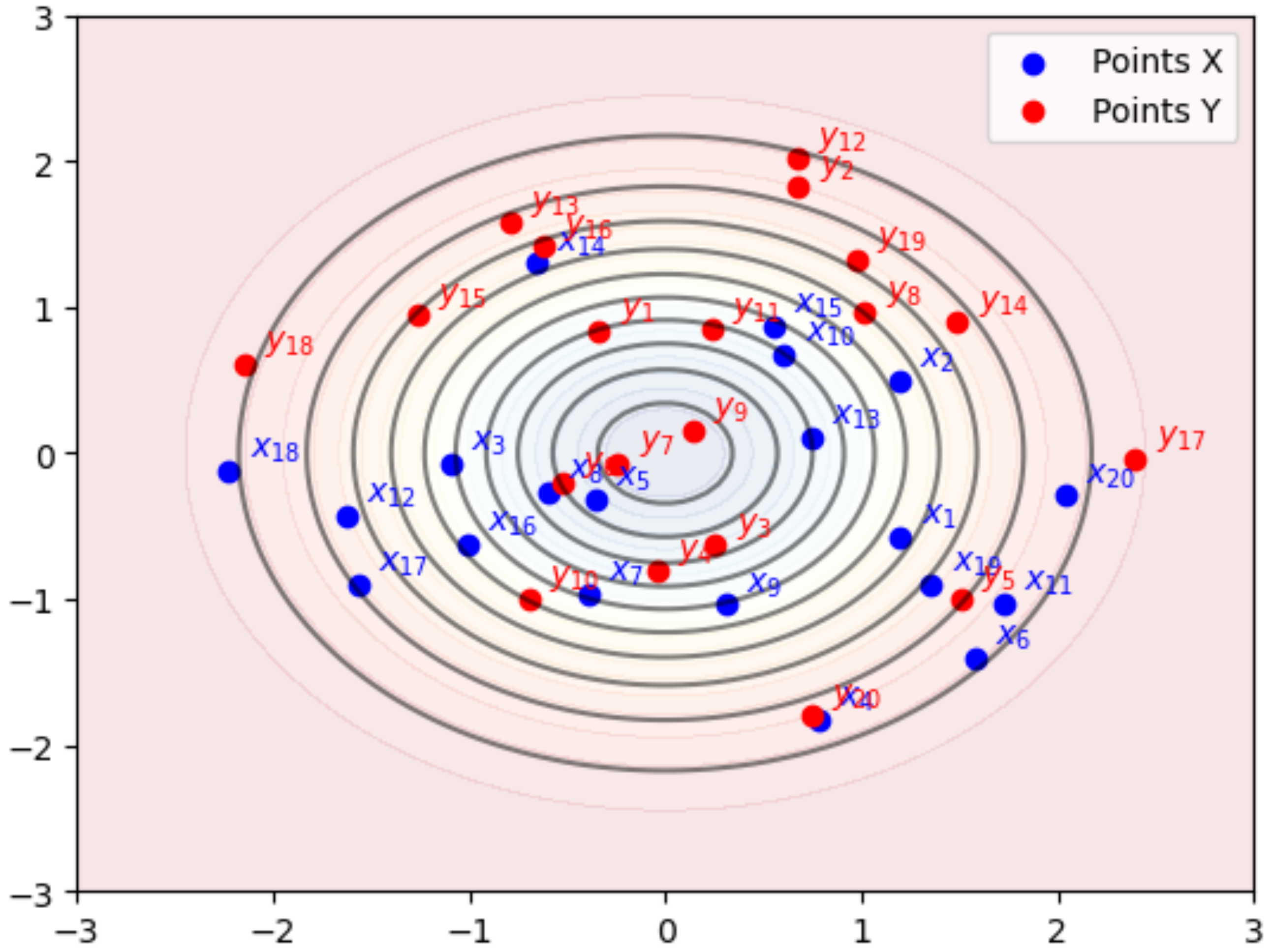
What is sampling in Maching Learning ?

It consists in approaching an unknown target probability distribution $q \in \mathscr{P}(\mathbb{R}^d)$ and to sample form it, i.e. generate $x_1, \ldots, x_n \sim q$.

<u>Applications</u> :

- **Bayesian inference** : the density of q is known up to a normalization constant : $q = \dfrac{\tilde{q}}{Z}$.

- <u>**Generative modelling**</u> : Samples from q are known : $y_1, \ldots, y_m \sim q$.



We want $\hat{p}_0 = \dfrac{1}{n} \sum_{i=1}^{n} \delta_{x^{(i)}}$ to be close to q and so close to $\hat{q} = \dfrac{1}{m} \sum_{j=1}^{m} \delta_{y_j}$.

Sampling as an optimization problem over $\mathscr{P}(\mathbb{R}^d)$

Let D be a **distance** or a **divergence** in $\mathscr{P}(\mathbb{R}^d)$:

- $\forall p, q \in \mathscr{P}(\mathbb{R}^d), \quad D(p\,||\,q) \geqslant 0.$
- $D(p\,||\,q) = 0 \Leftrightarrow p = q.$

Sampling as an optimization problem over $\mathscr{P}(\mathbb{R}^d)$

Let D be a **distance** or a **divergence** in $\mathscr{P}(\mathbb{R}^d)$:

- $\forall p, q \in \mathscr{P}(\mathbb{R}^d), \quad D(p\,||\,q) \geqslant 0.$
- $D(p\,||\,q) = 0 \Leftrightarrow p = q.$

Sampling can be formulated as a minimization problem,

$$\min_{p \in \mathscr{P}(\mathbb{R}^d)} \mathscr{F}(p)$$

where $\mathscr{F}(p) = D(p\,||\,q)$ for a fixed target q.

Sampling as an optimization problem over $\mathscr{P}(\mathbb{R}^d)$

Let D be a **distance** or a **divergence** in $\mathscr{P}(\mathbb{R}^d)$:

- $\forall p, q \in \mathscr{P}(\mathbb{R}^d), \quad D(p\,||\,q) \geqslant 0.$
- $D(p\,||\,q) = 0 \Leftrightarrow p = q.$

Sampling can be formulated as a minimization problem,

$$\min_{p \in \mathscr{P}(\mathbb{R}^d)} \mathscr{F}(p)$$

where $\mathscr{F}(p) = D(p\,||\,q)$ for a fixed target q.

Questions :

- How to solve optimization in $\mathscr{P}(\mathbb{R}^d)$?
- Choice of the divergence D : Regularized KKL

Solving an optimization problem over $\mathscr{P}(\mathbb{R}^d)$

How to solve the minimization $\min_{p \in \mathscr{P}(\mathbb{R}^d)} \mathscr{F}(p)$?

Solving an optimization problem over $\mathscr{P}(\mathbb{R}^d)$

How to solve the minimization $\min_{p \in \mathscr{P}(\mathbb{R}^d)} \mathscr{F}(p)$? ➡ Gradient descent **?**

Solving an optimization problem over $\mathscr{P}(\mathbb{R}^d)$

How to solve the minimization $\min_{p \in \mathscr{P}(\mathbb{R}^d)} \mathscr{F}(p)$? → Gradient descent **?** → Impossible in $\mathscr{P}(\mathbb{R}^d)$ **!**

Solving an optimization problem over $\mathscr{P}(\mathbb{R}^d)$

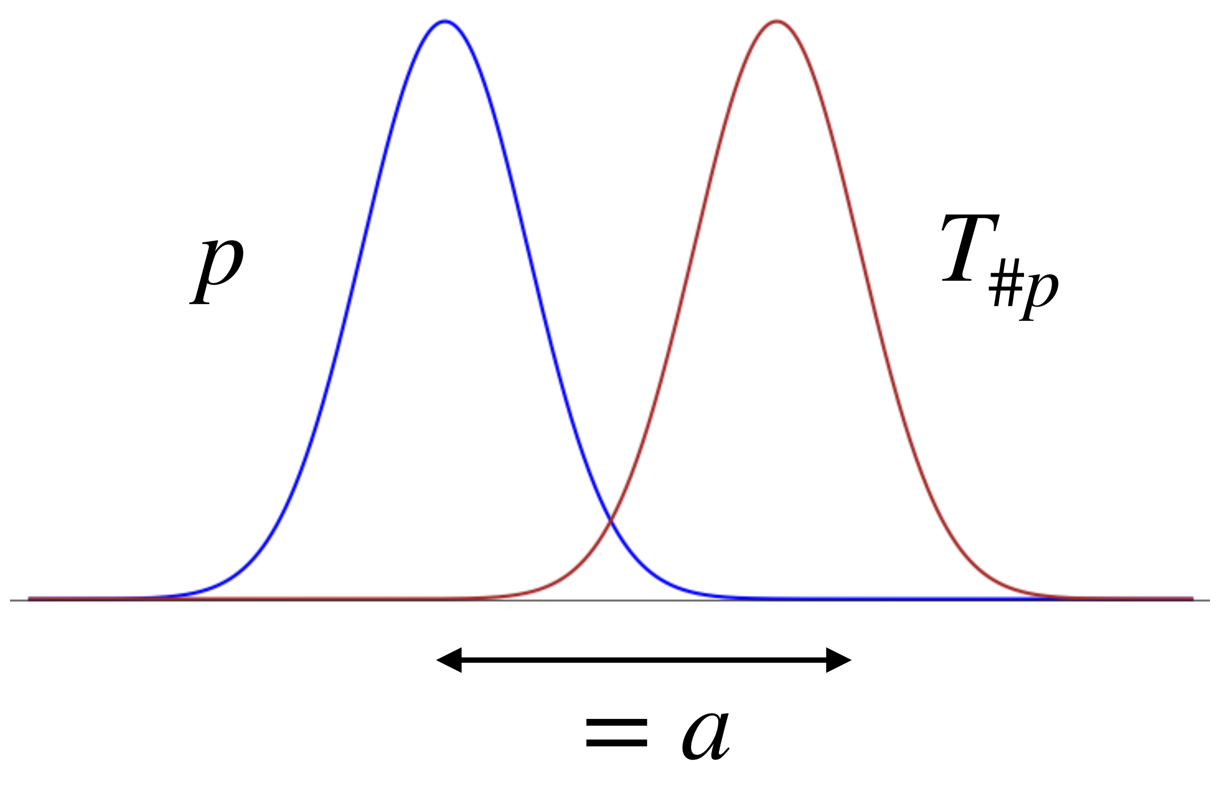
How to solve the minimization $\min_{p \in \mathscr{P}(\mathbb{R}^d)} \mathscr{F}(p)$? → Gradient descent **?** → Impossible in $\mathscr{P}(\mathbb{R}^d)$ **!**

Reminder

Let $p \in \mathscr{P}(\mathbb{R}^d)$ and $T : \mathbb{R}^d \to \mathbb{R}^d$, the **push forward** distribution of p by T is

$$T_{\#p}(A) = p(T^{-1}(A)), \forall A \subset \mathbb{R}^d.$$

<u>Example :</u> $T(x) = x + a$



Solving an optimization problem over $\mathscr{P}(\mathbb{R}^d)$

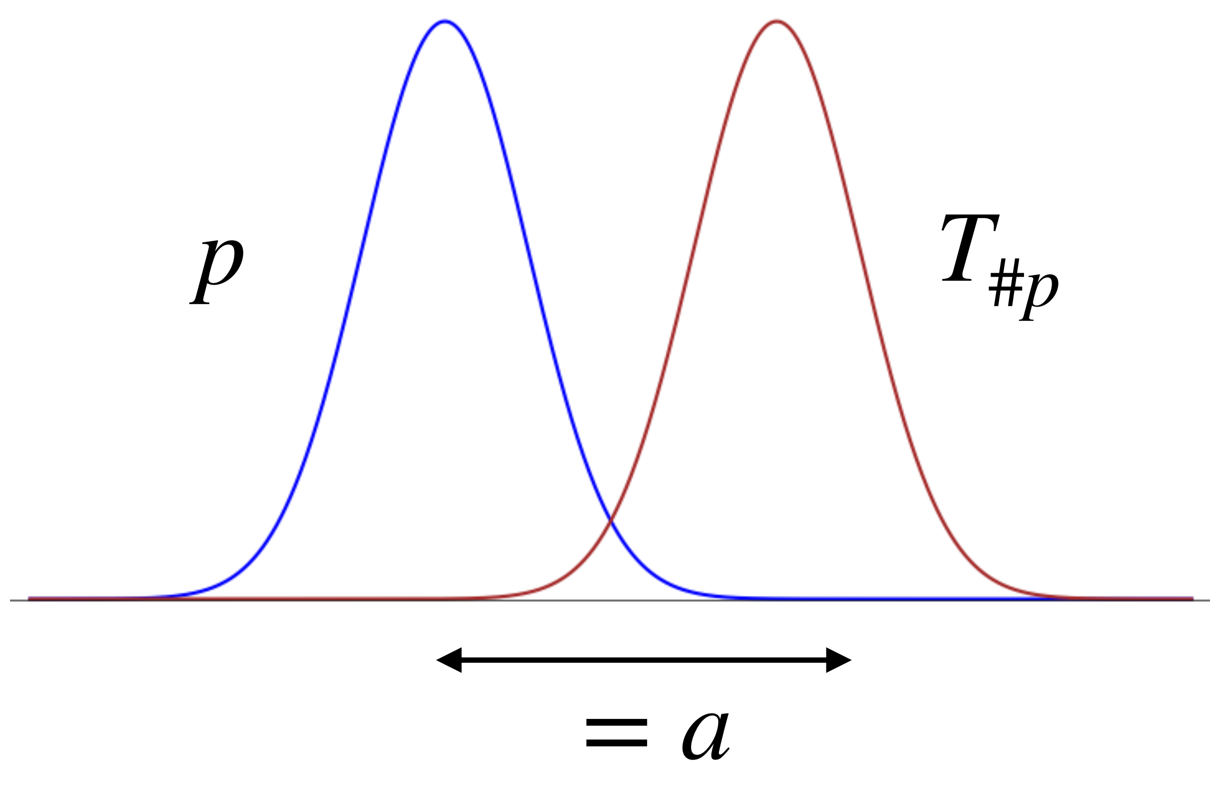
How to solve the minimization $\min_{p \in \mathscr{P}(\mathbb{R}^d)} \mathscr{F}(p)$? ➡ Gradient descent **?** ➡ Impossible in $\mathscr{P}(\mathbb{R}^d)$ **!**

Reminder

Let $p \in \mathscr{P}(\mathbb{R}^d)$ and $T : \mathbb{R}^d \to \mathbb{R}^d$, the **push forward** distribution of p by T is

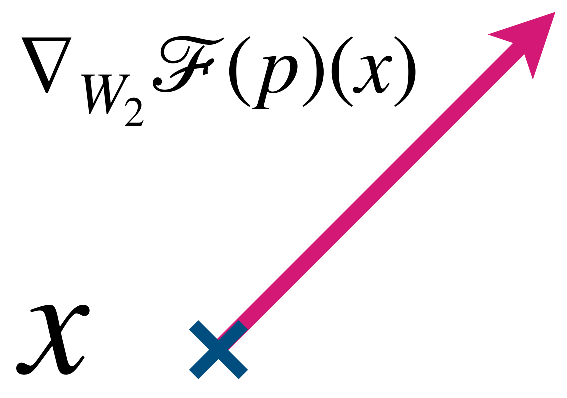
$$T_{\#p}(A) = p(T^{-1}(A)), \ \forall A \subset \mathbb{R}^d.$$

<u>Example :</u> $T(x) = x + a$



Idea to solve the minimization : Instead of doing Gradient descent on p we do it on the particules in $\mathbb{R}^d$ constituting the masse of p.

Wasserstein Gradient

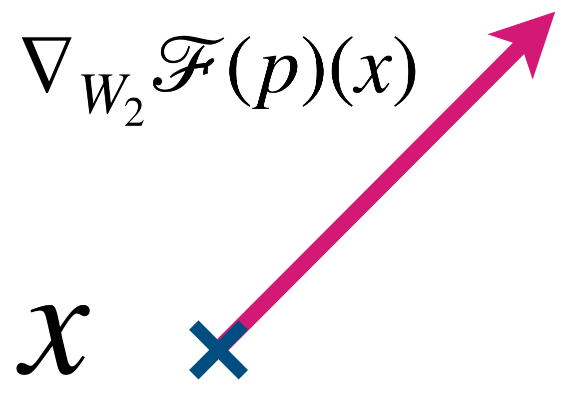


Definition: If for all $h : \mathbb{R}^d \to \mathbb{R}^d$, $\varepsilon > 0$,

$$\mathscr{F}((I_d + \varepsilon h)_{\#}p) = \mathscr{F}(p) + \varepsilon \langle \nabla_{W_2}\mathscr{F}(p), h \rangle_p + o(\varepsilon)$$

handles, then $\nabla_{W_2}\mathscr{F}(p) : \mathbb{R}^d \to \mathbb{R}^d$ is the **Wasserstein gradient** of $\mathscr{F}$. This is a **vector field** .

Wasserstein Gradient



Definition: If for all $h : \mathbb{R}^d \to \mathbb{R}^d$, $\varepsilon > 0$,

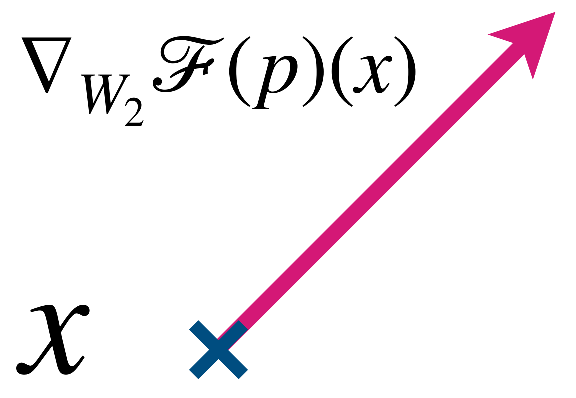
$$\mathscr{F}((I_d + \varepsilon h)_\# p) = \mathscr{F}(p) + \varepsilon \langle \nabla_{W_2}\mathscr{F}(p), h \rangle_p + o(\varepsilon)$$

handles, then $\nabla_{W_2}\mathscr{F}(p) : \mathbb{R}^d \to \mathbb{R}^d$ is the **Wasserstein gradient** of $\mathscr{F}$. This is a **vector field** .

Wasserstein Gradient Descent : $t = 1,\ldots,T$, step size γ, we do

$$p_{t+1} = (I_d - \gamma \nabla_{W_2}\mathscr{F}(p_t))_{\# p_t}$$

Wasserstein Gradient



Definition: If for all $h : \mathbb{R}^d \to \mathbb{R}^d$, $\varepsilon > 0$,

$$\mathscr{F}((I_d + \varepsilon h)_{\#}p) = \mathscr{F}(p) + \varepsilon\langle \nabla_{W_2}\mathscr{F}(p), h\rangle_p + o(\varepsilon)$$

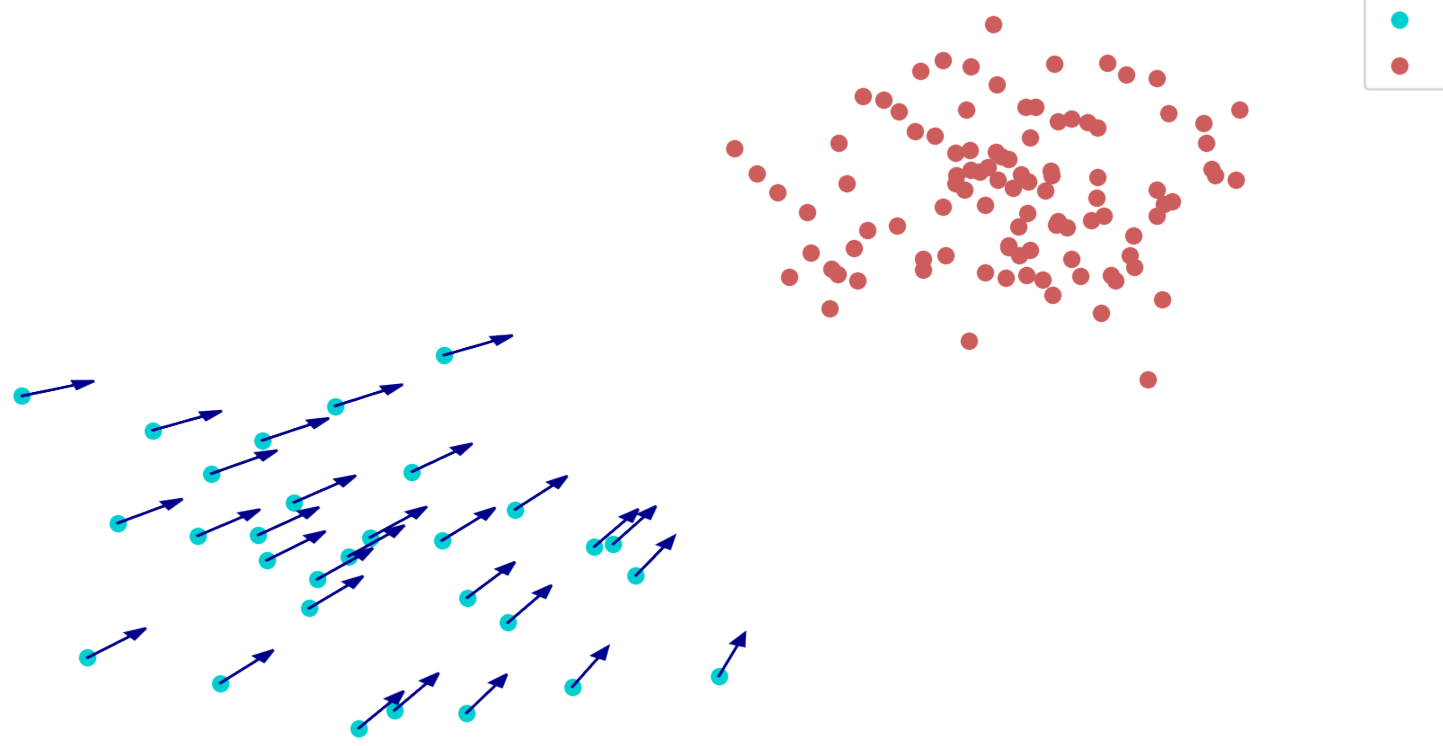
handles, then $\nabla_{W_2}\mathscr{F}(p) : \mathbb{R}^d \to \mathbb{R}^d$ is the **Wasserstein gradient** of $\mathscr{F}$. This is a **vector field** .

Wasserstein Gradient Descent : $t = 1,...,T$, step size γ, we do

$$p_{t+1} = (I_d + \gamma\nabla_{W_2}\mathscr{F}(p_t))_{\#p_t}$$

Discretized WGD: Let $x_t^{(1)}, \ldots, x_t^{(n)} \in \mathbb{R}^d$ and $\hat{p}_t = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_t^{(i)}}$,

$$x_{t+1}^{(i)} = x_t^{(i)} - \gamma\nabla_{W_2}\mathscr{F}(\hat{p}_t)(x_t^{(i)}),\ \forall i = 1,..,n$$



Choice of the divergence D

Let come back to the principal problem

$$\min_{p \in \mathscr{P}(\mathbb{R}^d)} D(p \,||\, q).$$

The choice of the divergence D is critical.

Examples:

- Maximum Mean Discrepancy
- Kullback Leibler divergence $\mathrm{KL}(p \,||\, q) = \int \log \frac{p}{q} dp$

- The choice of D is based on several factors: its geometry, the facility with which its Wasserstein gradient can be calculated, the possibility of evaluating it on all types of probabilities…
- We chose to use the **Regularized Kernel Kullback Leibler divergence.**

Reminders on kernel methods

Let $\mathcal{X} \subset \mathbb{R}^d$, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a symetric positive kernel, i.e.

- $\forall x_1, \ldots, x_n \in \mathcal{X}, \quad \forall a_1, \ldots, a_n \in \mathbb{R}, \sum_i \sum_j a_i a_j k(x_i, x_j) \geqslant 0$
- $\forall x, y \in \mathcal{X}, \quad k(x, y) = k(y, x)$

Reminders on kernel methods

Let $\mathscr{X} \subset \mathbb{R}^d$, $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ a symetric positive kernel, i.e.

- $\forall x_1, \ldots, x_n \in \mathscr{X}, \quad \forall a_1, \ldots, a_n \in \mathbb{R}, \sum_i \sum_j a_i a_j k(x_i, x_j) \geqslant 0$

- $\forall x, y \in \mathscr{X}, \quad k(x, y) = k(y, x).$

<u>**Theorem**</u> **:** There exists a unique Hilbert space $\mathscr{H} \subset \{f : \mathscr{X} \to \mathbb{R}\}$ s.t.

- $\forall x \in \mathscr{X}, \quad k(\,.\,, x) \in \mathscr{H}$

- $\forall f \in \mathscr{H}, \quad f(x) = \langle f, k(\,.\,, x) \rangle_{\mathscr{H}}$

$\mathscr{H}$ is the RKHS (Reproducing kernel Hilbert space) associated with k, and k is the unique reproducing kernel of $\mathscr{H}$.

Kernel Kullback Leibler divergence (KKL)

Let $\mathscr{H}$ be an RKHS on $\mathbb{R}^d$ with kernel k. Let $p \in \mathscr{P}(\mathbb{R}^d)$, the covariance operator of p, $\Sigma_p : \mathscr{H} \mapsto \mathscr{H}$, is

$$\Sigma_p = \int k(\,.\,, x)k(\,.\,, x)^* dp(x).$$

* *Bach, F. (2022). Information theory with kernel methods. IEEE Transactions on Information Theory, 69(2), 752-775.*

Kernel Kullback Leibler divergence (KKL)

Let $\mathscr{H}$ be an RKHS on $\mathbb{R}^d$ with kernel k. Let $p \in \mathscr{P}(\mathbb{R}^d)$, the covariance operator of p, $\Sigma_p : \mathscr{H} \mapsto \mathscr{H}$, is

$$\Sigma_p = \int k(\,.\,,x)k(\,.\,,x)^* dp(x).$$

Using $p \mapsto \Sigma_p$, $q \mapsto \Sigma_q$, the KKL divergence is defined as

$$\mathrm{KKL}(p\,||\,q) = \mathrm{Tr}\ \Sigma_p(\log \Sigma_p - \log \Sigma_q)$$

for $p \ll q \in \mathscr{P}(\mathbb{R}^d)$ and is equal to $+\infty$ if $p \not\ll q$.

* *Bach, F. (2022). Information theory with kernel methods. IEEE Transactions on Information Theory, 69(2), 752-775.*

Kernel Kullback Leibler divergence (KKL)

Let $\mathscr{H}$ be an RKHS on $\mathbb{R}^d$ with kernel k. Let $p \in \mathscr{P}(\mathbb{R}^d)$, the covariance operator of p, $\Sigma_p : \mathscr{H} \mapsto \mathscr{H}$, is

$$\Sigma_p = \int k(\,.\,,x)k(\,.\,,x)^* dp(x).$$

Using $p \mapsto \Sigma_p$, $q \mapsto \Sigma_q$, the KKL divergence is defined as

$$\text{KKL}(p\,||\,q) = \text{Tr } \Sigma_p(\log \Sigma_p - \log \Sigma_q)$$

for $p \ll q \in \mathscr{P}(\mathbb{R}^d)$ and is equal to $+\infty$ if $p \not\ll q$.

Theorem [Bach 2022] : If k^2 is universal and $k(x,x) = 1\, \forall x \in \mathbb{R}^d$, then $\text{KKL}(p\,||\,q) = 0 \Leftrightarrow p = q$.

* *Bach, F. (2022). Information theory with kernel methods. IEEE Transactions on Information Theory, 69(2), 752-775.*

Kernel Kullback Leibler divergence (KKL)

Let $\mathcal{H}$ be an RKHS on $\mathbb{R}^d$ with kernel k. Let $p \in \mathcal{P}(\mathbb{R}^d)$, the covariance operator of p, $\Sigma_p : \mathcal{H} \mapsto \mathcal{H}$, is

$$\Sigma_p = \int k(\,.\,,x)k(\,.\,,x)^* dp(x).$$

Using $p \mapsto \Sigma_p$, $q \mapsto \Sigma_q$, the KKL divergence is defined as

$$\mathrm{KKL}(p||q) = \mathrm{Tr}\ \Sigma_p(\log \Sigma_p - \log \Sigma_q)$$

for $p \ll q \in \mathcal{P}(\mathbb{R}^d)$ and is equal to $+\infty$ if $p \not\ll q$.

Theorem [Bach 2022] : If k^2 is universal and $k(x,x) = 1\, \forall x \in \mathbb{R}^d$, then $\mathrm{KKL}(p||q) = 0 \Leftrightarrow p = q$.

Regularized KKL : For $\alpha \in [0,1]$, for any $p, q \in \mathcal{P}(\mathbb{R}^d)$

$$\mathrm{KKL}_\alpha(p||q) = \mathrm{KKL}(p||(1-\alpha)q + \alpha p)$$

* *Bach, F. (2022). Information theory with kernel methods. IEEE Transactions on Information Theory, 69(2), 752-775.*

Close form expression of the regularized KKL

Let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ and $\hat{q} = \frac{1}{m}\sum_{j=1}^{m} \delta_{y_j}$. Define $K_p = (k(x_i, x_j))_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$, $K_q = (k(y_i, y_j))_{i,j=1}^{m} \in \mathbb{R}^{m\times m}$ and $K_{pq} = (k(x_i, y_j))_{i,j=1}^{n,m} \in \mathbb{R}^{n\times m}$. Then, for any $\alpha \in]0,1[$,

$$\mathrm{KKL}_\alpha(\hat{p}\,||\,\hat{q}) = \mathrm{Tr}\left(\frac{1}{n}K_p \log \frac{1}{n}K_p\right) - \mathrm{Tr}\left(I_\alpha K \log(K)\right),$$

Where $I_\alpha = \begin{pmatrix} \frac{1}{\alpha}I & 0 \\ 0 & 0 \end{pmatrix}$ and $K = \begin{pmatrix} \frac{\alpha}{n}K_p & \sqrt{\frac{\alpha(1-\alpha)}{nm}}K_{pq} \\ \sqrt{\frac{\alpha(1-\alpha)}{nm}}K_{qp} & \frac{1-\alpha}{m}K_q \end{pmatrix}$.

- There is also a closed form for the Wasserstein Gradient on empirical measures !

Properties of the KKL

- **Convergence in α :** Let $p, q \in \mathscr{P}(\mathbb{R})$. Assume $\dfrac{dp}{dq} \leqslant \dfrac{1}{\mu}$. Then,

  $$|\,\mathrm{KKL}_\alpha(p\,||\,q) - \mathrm{KKL}(p\,||\,q)\,| \leqslant \left(\alpha\left(1+\frac{1}{\mu}\right) + \frac{\alpha^2}{1-\alpha}\left(1+\frac{1}{\mu^2}\right)\right)|\,\mathrm{Tr}\left(\Sigma_p \log \Sigma_q\right)|$$

Properties of the KKL

- **Convergence in α :** Let $p, q \in \mathscr{P}(\mathbb{R})$. Assume $\dfrac{dp}{dq} \leqslant \dfrac{1}{\mu}$. Then,

$$|\,\mathrm{KKL}_\alpha(p\,||\,q) - \mathrm{KKL}(p\,||\,q)\,| \leqslant \left(\alpha\left(1+\frac{1}{\mu}\right) + \frac{\alpha^2}{1-\alpha}\left(1+\frac{1}{\mu^2}\right)\right)|\,\mathrm{Tr}\left(\Sigma_p \log \Sigma_q\right)|$$

- **Convergence in n:**

$$\mathbb{E}\,|\,\mathrm{KKL}_\alpha(\hat{p}\,||\,\hat{q}) - \mathrm{KKL}_\alpha(p\,||\,q)\,| \leqslant \frac{35}{\sqrt{m \wedge n}}\frac{1}{\alpha\mu}(2\sqrt{c} + \log n) + \frac{1}{m \wedge n}\left(1 + \frac{1}{\mu} + \frac{c(24\log n)^2}{\alpha\mu^2}(1 + \frac{n}{m \wedge m})\right).$$

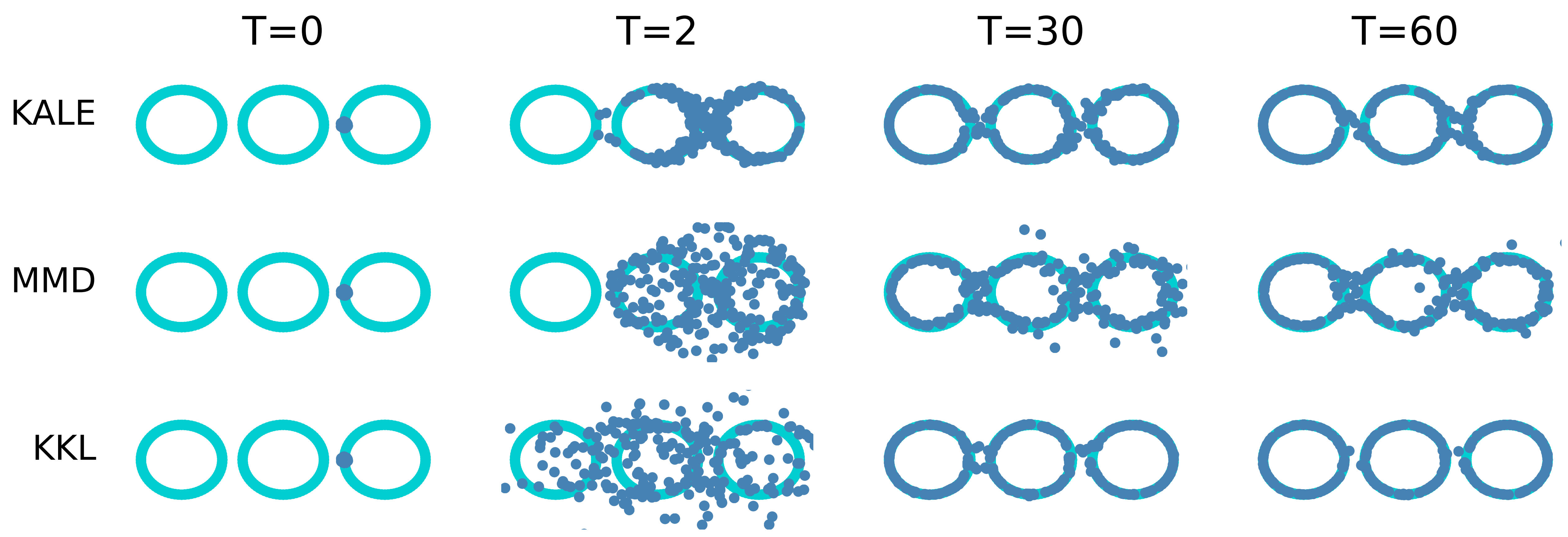
Properties of the KKL

- **Convergence in α :** Let $p, q \in \mathscr{P}(\mathbb{R})$. Assume $\frac{dp}{dq} \leqslant \frac{1}{\mu}$. Then,
$$|\,\mathrm{KKL}_\alpha(p\,||\,q) - \mathrm{KKL}(p\,||\,q)\,| \leqslant \left(\alpha\left(1+\frac{1}{\mu}\right)+\frac{\alpha^2}{1-\alpha}\left(1+\frac{1}{\mu^2}\right)\right)|\,\mathrm{Tr}\left(\Sigma_p \log \Sigma_q\right)|$$

- **Convergence in n:**
$$\mathbb{E}\,|\,\mathrm{KKL}_\alpha(\hat{p}\,||\,\hat{q}) - \mathrm{KKL}_\alpha(p\,||\,q)\,| \leqslant \frac{35}{\sqrt{m \wedge n}}\frac{1}{\alpha\mu}(2\sqrt{c}+\log n)+\frac{1}{m \wedge n}\left(1+\frac{1}{\mu}+\frac{c(24\log n)^2}{\alpha\mu^2}(1+\frac{n}{m \wedge m})\right).$$

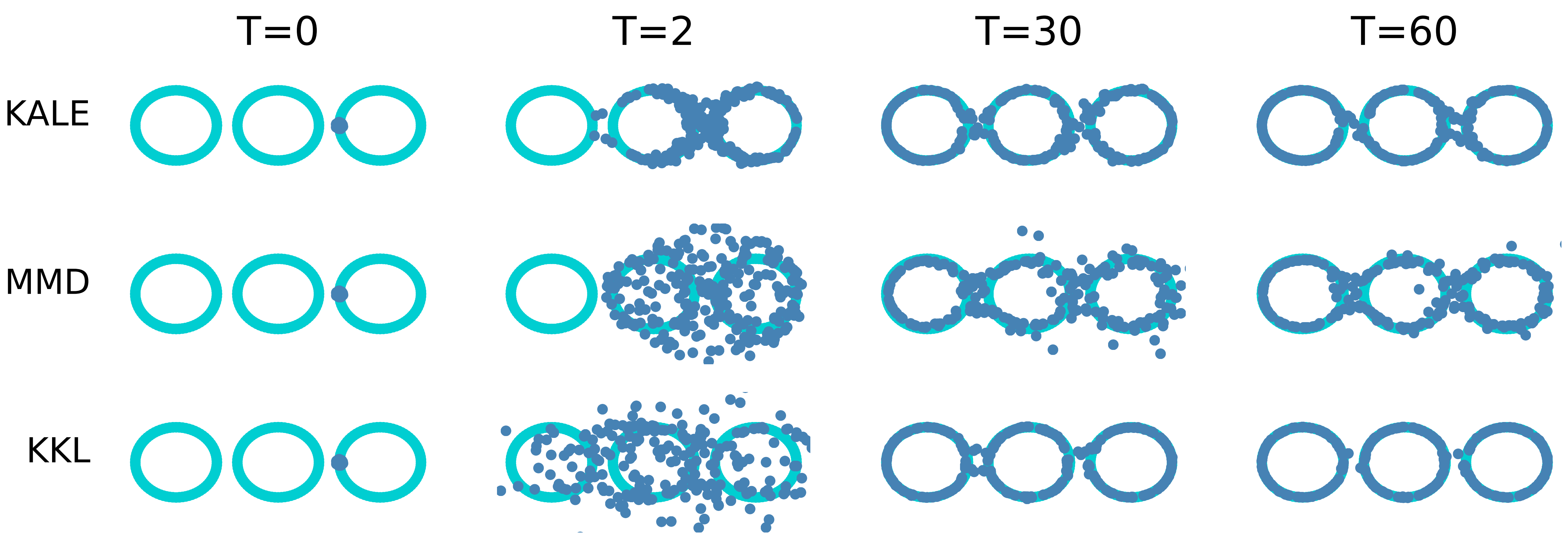
- $\alpha \to \mathrm{KKL}_\alpha(p\,||\,q)$ is decreasing

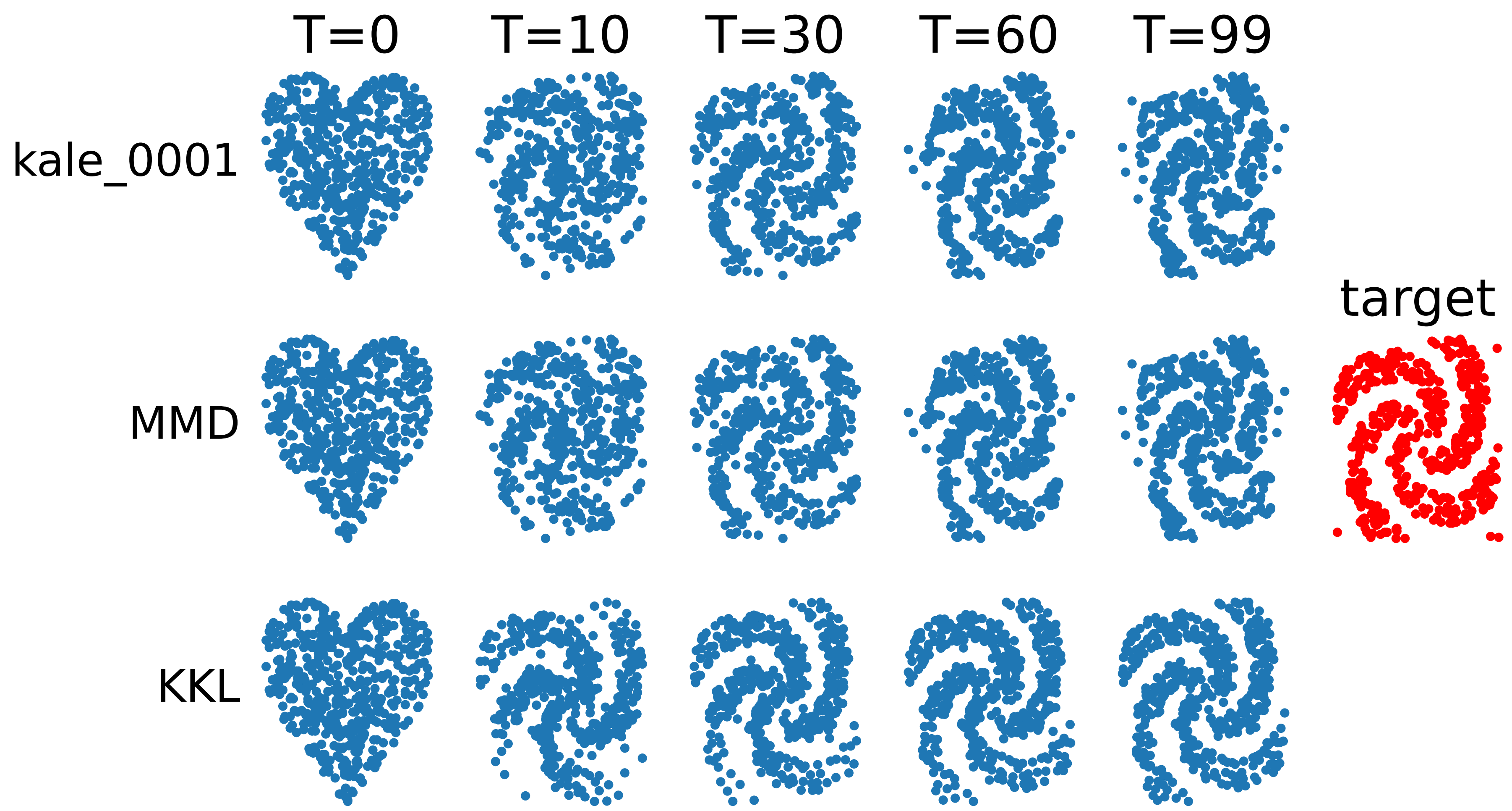
Sampling experiments [Glaser 2021]*



* Glaser, P., Arbel, M., & Gretton, A. (2021). KALE flow: A relaxed KL gradient flow for probabilities with disjoint support. *Advances in Neural Information Processing Systems*, *34*, 8018-8031.

Sampling experiments [Glaser 2021]*





* Glaser, P., Arbel, M., & Gretton, A. (2021). KALE flow: A relaxed KL gradient flow for probabilities with disjoint support. *Advances in Neural Information Processing Systems*, *34*, 8018-8031.