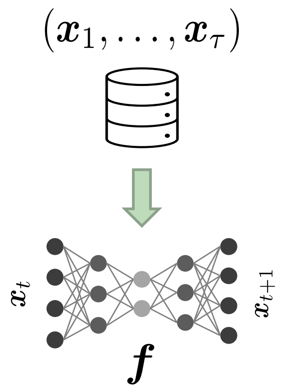
Learning Multiple Nonlinear Dynamical Systems with Side Information

Naoya Takeishi (HES-SO/RIKEN), Yoshinobu Kawahara (KyushuU/RIKEN)

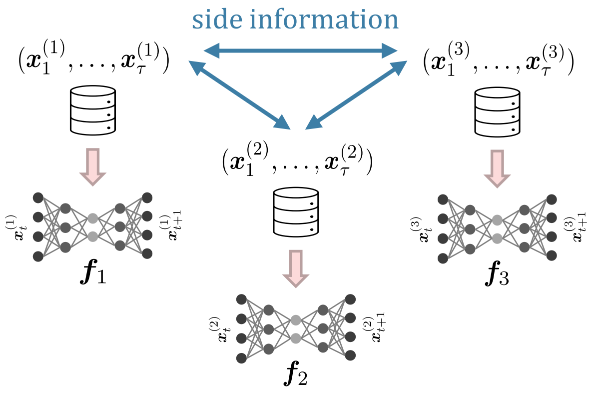
The 59th IEEE CDC, December 2020

Problem Setting

- Machine learning of dynamical system
  - Given a dataset (i.e., a sequence of states) $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\tau), \boldsymbol{x} \in \mathbb{R}^d$,
  - learn $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^d$ s.t. $\boldsymbol{x}_{t+1} \approx \boldsymbol{f}(\boldsymbol{x}_t)$
  - e.g., linear models, kernel machines, neural nets, ...



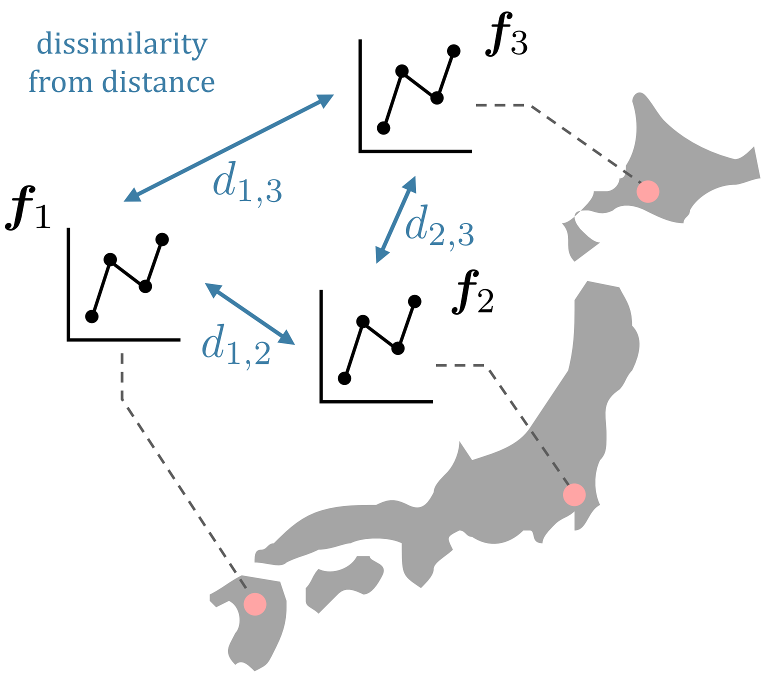
- Machine learning of multiple dynamical systems
  - Given multiple datasets $(\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{\tau_1}^{(1)}), \ldots, (\boldsymbol{x}_1^{(n)}, \ldots, \boldsymbol{x}_{\tau_n}^{(n)})$,
  - learn $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$ s.t. $\boldsymbol{x}_{t+1}^{(1)} \approx \boldsymbol{f}_1(\boldsymbol{x}_t^{(1)}), \ldots, \boldsymbol{x}_{t+1}^{(n)} \approx \boldsymbol{f}_n(\boldsymbol{x}_t^{(n)})$
  - Side information on relation between datasets/dynamics is often available and helpful for learning

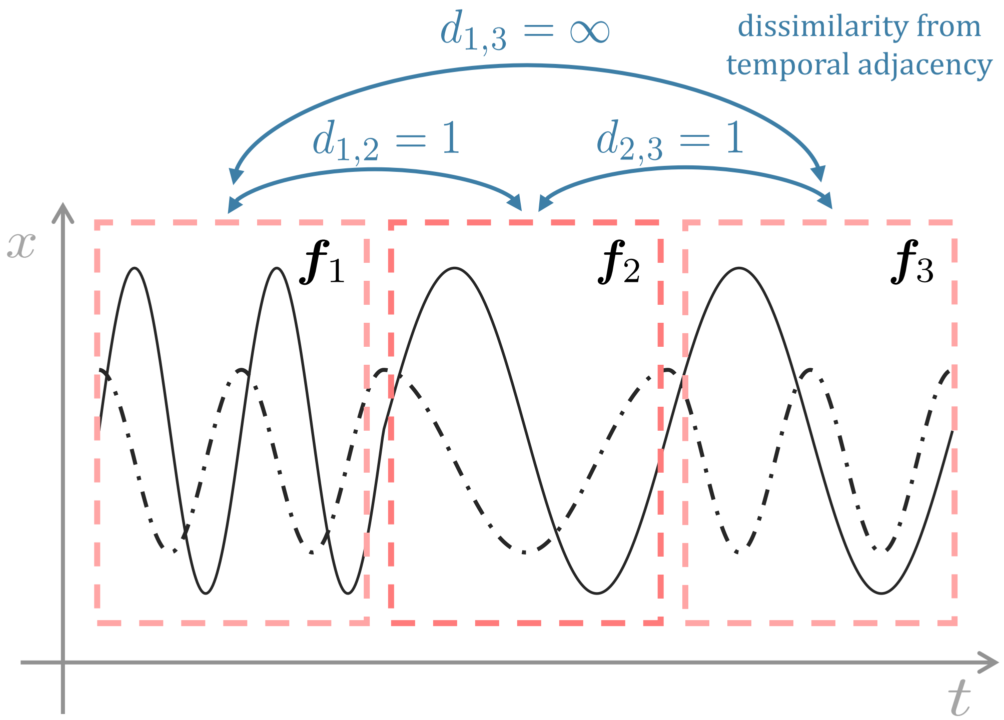


Examples of Side Information

When learning dynamics from **geometrically distributed sensors**, distance between sensors may be informative for similarity of dynamics, e.g., closer sensors measure similar dynamics.

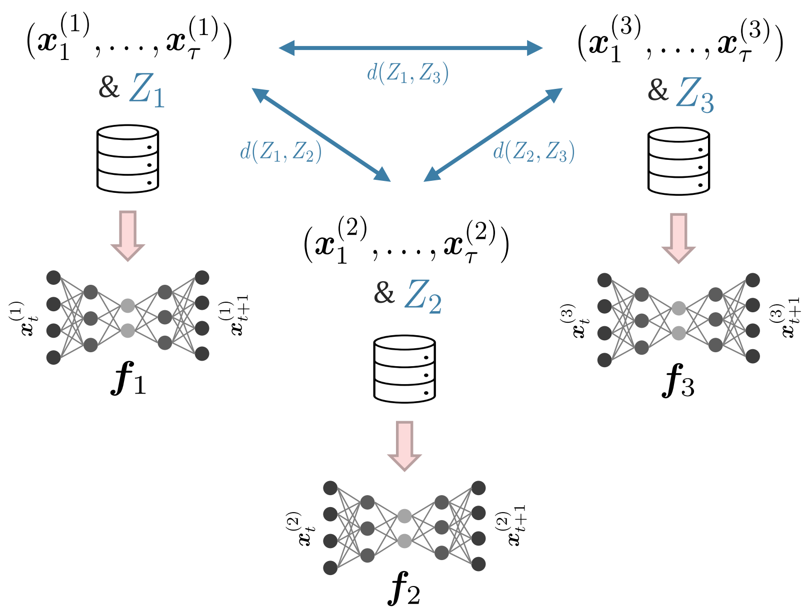
Learning **time-varying dynamical systems** can also be handled as a special case, where we may use t as side information, e.g., dynamics in adjacent time periods are similar.





Problem Setting Again

- Input:
  - n sets of measurements (state sequences) $(\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{\tau_1}^{(1)}), \ldots, (\boldsymbol{x}_1^{(n)}, \ldots, \boldsymbol{x}_{\tau_n}^{(n)})$
  - n sets of side information $Z_1, \ldots, Z_n$
    - ✓ along with some dissimilarity measure $d_{\mathcal{Z}}(Z_i, Z_j)$
    - ✓ Z may be location of sensors, timestamps, ..., additional measurements, text description, ...
- Output: dynamics models $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$ such that $\boldsymbol{x}_{t+1}^{(1)} \approx \boldsymbol{f}_1(\boldsymbol{x}_t^{(1)}), \ldots, \boldsymbol{x}_{t+1}^{(n)} \approx \boldsymbol{f}_n(\boldsymbol{x}_t^{(n)})$



- Core idea: Regularization with side information (next slide)
  - if Z_i and Z_j are similar (i.e., $d_{\mathcal{Z}}(Z_i, Z_j)$ is small), then $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ should also be similar

Regularization with Side Information

- Regularization with side information
  - if Z_i and Z_j are similar (i.e., $d_{\mathcal{Z}}(Z_i, Z_j)$ is small),
    then $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ should also be similar (i.e., some dissimilarity $d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j)$ is small)

- Formulation as multi-task learning for dynamical systems

$$\underset{\boldsymbol{f}_1, \dots, \boldsymbol{f}_n}{\text{minimize}} \quad \underbrace{\frac{1}{n} \sum_{i=1}^{n} L_i(\boldsymbol{f}_i; \{\boldsymbol{x}_1^{(i)}, \dots, \boldsymbol{x}_{\tau_i}^{(i)}\})}_{\substack{\text{original objective function to learn } \boldsymbol{f}_1, \dots, \boldsymbol{f}_n \\ \text{e.g., squared loss } L_i = \sum_{t=1}^{\tau_i - 1} \|\boldsymbol{f}_i(\boldsymbol{x}_t^{(i)}) - \boldsymbol{x}_{t+1}^{(i)}\|_2^2}} + \underbrace{\lambda \cdot \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j)}{d_{\mathcal{Z}}(Z_i, Z_j)}}_{\text{regularization term to make } d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j) \propto d_{\mathcal{Z}}(Z_i, Z_j)}$$

- What $d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j)$ should be? (next slide)

Dissimilarity Measure of Dynamical Systems

$$\underset{\boldsymbol{f}_1,\dots,\boldsymbol{f}_n}{\text{minimize}} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} L_i(\boldsymbol{f}_i;\, \{\boldsymbol{x}_1^{(i)},\dots,\boldsymbol{x}_{\tau_i}^{(i)}\})}_{\substack{\text{original objective function to learn } \boldsymbol{f}_1,\dots,\boldsymbol{f}_n \\ \text{e.g., squared loss } L_i = \sum_{t=1}^{\tau_i - 1}\|\boldsymbol{f}_i(\boldsymbol{x}_t^{(i)}) - \boldsymbol{x}_{t+1}^{(i)}\|_2^2}} + \lambda \cdot \underbrace{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\frac{d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j)}{d_{\mathcal{Z}}(Z_i, Z_j)}}_{\text{regularization term to make } d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j) \propto d_{\mathcal{Z}}(Z_i, Z_j)}$$

- $d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j)$ should measure the dissimilarity between $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$
- We use an operator-theoretic metric [Ishikawa+ 18]

K_i, K_j: Perron–Frobenius operators in RKHS

$$d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j) = \sqrt{1 - \frac{k_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j)^2}{k_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_i)k_{\mathrm{DS}}(\boldsymbol{f}_j, \boldsymbol{f}_j)}}, \quad k_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j)^2 = \mathrm{tr}\left(\bigwedge^m \sum_{t=0}^{\tau} (K_j^t)^* K_i^t\right)$$

- because it is applicable to nonlinear $\boldsymbol{f}$ & agnostic of parametric form of $\boldsymbol{f}$

Dissimilarity Measure of Dynamical Systems

cont'd

$$d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j) = \sqrt{1 - \frac{k_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j)^2}{k_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_i) k_{\mathrm{DS}}(\boldsymbol{f}_j, \boldsymbol{f}_j)}}, \quad k_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j)^2 = \mathrm{tr}\left(\bigwedge^m \sum_{t=0}^{\tau} (K_j^t)^* K_i^t\right)$$

- Let $\mathcal{H}$ be an RKHS equipped with a kernel function $k_{\mathcal{H}}(\cdot,\cdot)$.
- Let $\phi(\boldsymbol{x}) = k_{\mathcal{H}}(\boldsymbol{x}, \cdot)$ be the corresponding feature map.
- For a nonlinear dynamical system $\boldsymbol{x}_{t+1} = \boldsymbol{f}(\boldsymbol{x}_t)$, consider linear operator K:

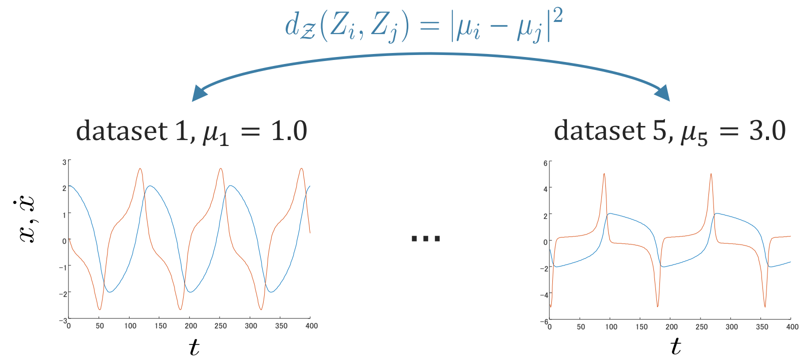
$$K\phi(\boldsymbol{x}) = \phi(\boldsymbol{f}(\boldsymbol{x})),$$

  which is called Perron–Frobenius operator in RKHS corresponding to $\boldsymbol{f}$.
- K can be **estimated from trajectory** $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\tau)$ with the kernel $k_{\mathcal{H}}(\cdot,\cdot)$.

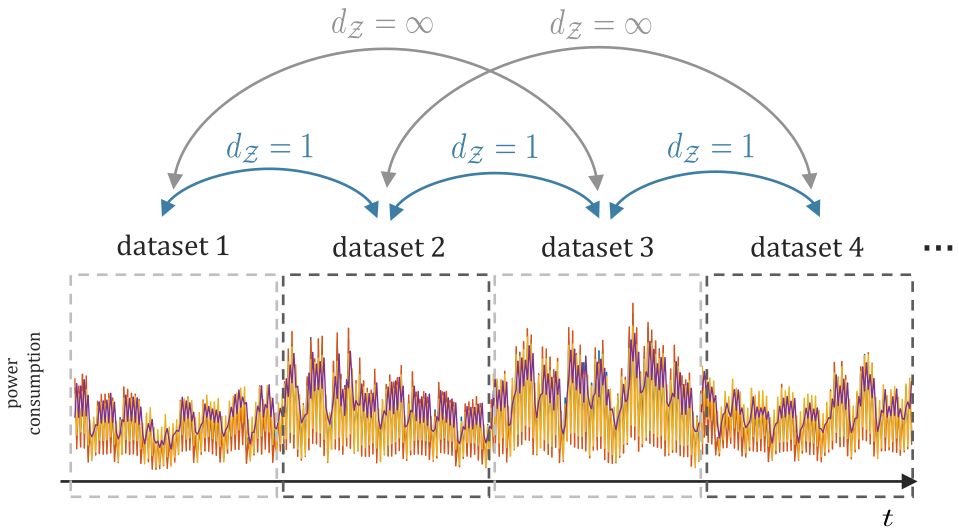
$d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j)$ is (almost everywhere) differentiable wrt. the parameters of $\boldsymbol{f}_i$ & $\boldsymbol{f}_j$

Experiment: Datasets

- Synthetic datasets
  - ➢ Van der Pol oscillator (→)
    - ✓ i-th dataset is generated via $\ddot{x} - \mu_i(1 - x^2)\dot{x} + x = 0$
    - ✓ Side information: $d_{\mathcal{Z}}(Z_i, Z_j) = |\mu_i - \mu_j|^2$
  - ➢ Rössler system



- Real-world datasets
  - ➢ Solar power production time-series
    - ✓ Side information: Z_i = location of i-th plant
  - ➢ Power consumption time-series (→)
    - ✓ Side information: $d_{\mathcal{Z}}(Z_i, Z_j) = 1$ if $j = i + 1$



Experiment: Baseline Methods

1. No multi-task regularization

2. Multi-task w/ Parameter L2 dist. $d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j) = \sum_k |\theta_{i,k} - \theta_{j,k}|^2$

3. Multi-task w/ Parameter L1 dist. $d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j) = \sum_k |\theta_{i,k} - \theta_{j,k}|$

$\theta_{i,k}$: k-th parameter of $\boldsymbol{f}_i$

cf. Proposed method (= multi-task w/ dynamics dist.)

$$\underset{\boldsymbol{f}_1,\dots,\boldsymbol{f}_n}{\text{minimize}} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} L_i(\boldsymbol{f}_i;\, \{\boldsymbol{x}_1^{(i)}, \dots, \boldsymbol{x}_{\tau_i}^{(i)}\})}_{\text{original objective function to learn } \boldsymbol{f}_1,\dots,\boldsymbol{f}_n} + \underbrace{\lambda \cdot \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j)}{d_{\mathcal{Z}}(Z_i, Z_j)}}_{\text{regularization term to make } d_{\mathrm{DS}}(\boldsymbol{f}_i, \boldsymbol{f}_j) \propto d_{\mathcal{Z}}(Z_i, Z_j)}$$

e.g., squared loss $L_i = \sum_{t=1}^{\tau_i - 1} \|\boldsymbol{f}_i(\boldsymbol{x}_t^{(i)}) - \boldsymbol{x}_{t+1}^{(i)}\|_2^2$

Experiment: Results

Test one-step prediction errors (smaller is better)

	Baseline 1 No regularization	Baseline 2 Parameter L2 dist.	Baseline 3 Parameter L1 dist.	Proposed method (dynamics dist.)
VdP	$2.936\,(.56) \times 10^{-1}$	$2.402\,(.82) \times 10^{-1}$	$2.659\,(.51) \times 10^{-1}$	$\mathbf{2.272}\,(.71) \times 10^{-1}$
Rössler	$1.358\,(.05) \times 10^{-2}$	$1.359\,(.05) \times 10^{-2}$	$1.358\,(.05) \times 10^{-2}$	$\underline{\mathbf{1.319}}\,(.05) \times 10^{-2}$
Solar	$9.231\,(.01) \times 10^{-4}$	$\mathbf{9.229}\,(.01) \times 10^{-4}$	$\mathbf{9.226}\,(.01) \times 10^{-4}$	$\underline{\mathbf{9.101}}\,(.03) \times 10^{-4}$
Demand	$1.486\,(.03) \times 10^{-3}$	$1.487\,(.03) \times 10^{-3}$	$1.488\,(.03) \times 10^{-3}$	$\underline{\mathbf{1.439}}\,(.03) \times 10^{-3}$

Significant difference, **bold**: from (A) / underline: from (B), by paired t-test at $p < 10^{-3}$.

- Proposed regularization achieves better prediction
  - Probably, L2/L1 distance of parameters cannot capture dynamics' dissimilarity

Summary

$$\underset{\boldsymbol{f}_1,\dots,\boldsymbol{f}_n}{\text{minimize}} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} L_i(\boldsymbol{f}_i;\ \{\boldsymbol{x}_1^{(i)},\dots,\boldsymbol{x}_{\tau_i}^{(i)}\})} + \lambda \cdot \underbrace{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n} \frac{d_{\mathrm{DS}}(\boldsymbol{f}_i,\boldsymbol{f}_j)}{d_{\mathcal{Z}}(Z_i,Z_j)}}$$

original objective function to learn $\boldsymbol{f}_1,\dots,\boldsymbol{f}_n$
e.g., squared loss $L_i = \sum_{t=1}^{\tau_i-1}\|\boldsymbol{f}_i(\boldsymbol{x}_t^{(i)}) - \boldsymbol{x}_{t+1}^{(i)}\|_2^2$

regularization term to make $d_{\mathrm{DS}}(\boldsymbol{f}_i,\boldsymbol{f}_j) \propto d_{\mathcal{Z}}(Z_i,Z_j)$

- **Multi-task learning** formulation for dynamical systems
  - ➢ to utilize side information
- We adopt the operator-theoretic metric [Ishikawa+ 18] for measuring dissimilarity between dynamical systems,
  - ➢ which can be estimated from trajectories during training
  - ➢ good for utilizing side information