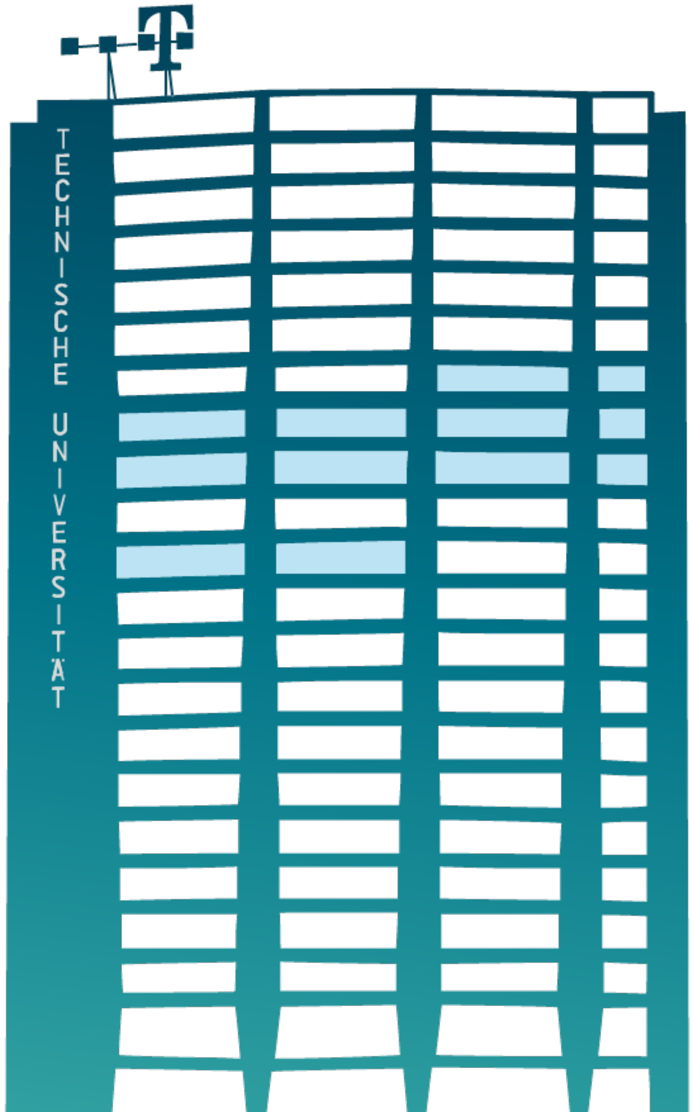
DAI-Labor
TU Berlin



The Fréchet Mean in Dynamic Time Warping Spaces for Data Imputation

David Schultz

26.02.2021

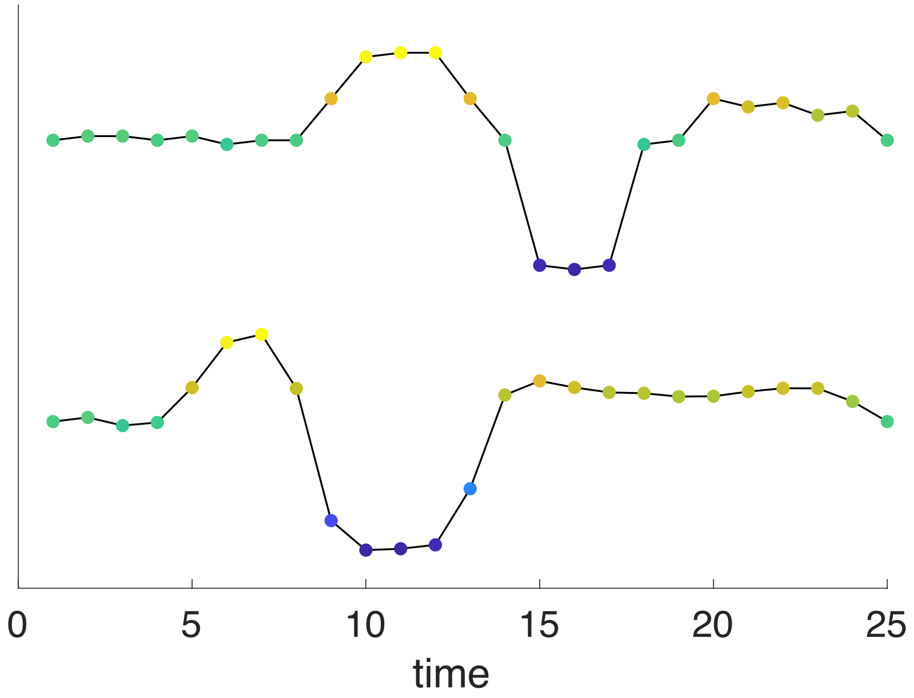
Outline

- Motivation
- Dynamic Time Warping (DTW)
- Fréchet Mean under DTW
- Nonsmooth Analysis
- Data Imputation

Motivation

- **Problem: How to average time series such that features are preserved?**

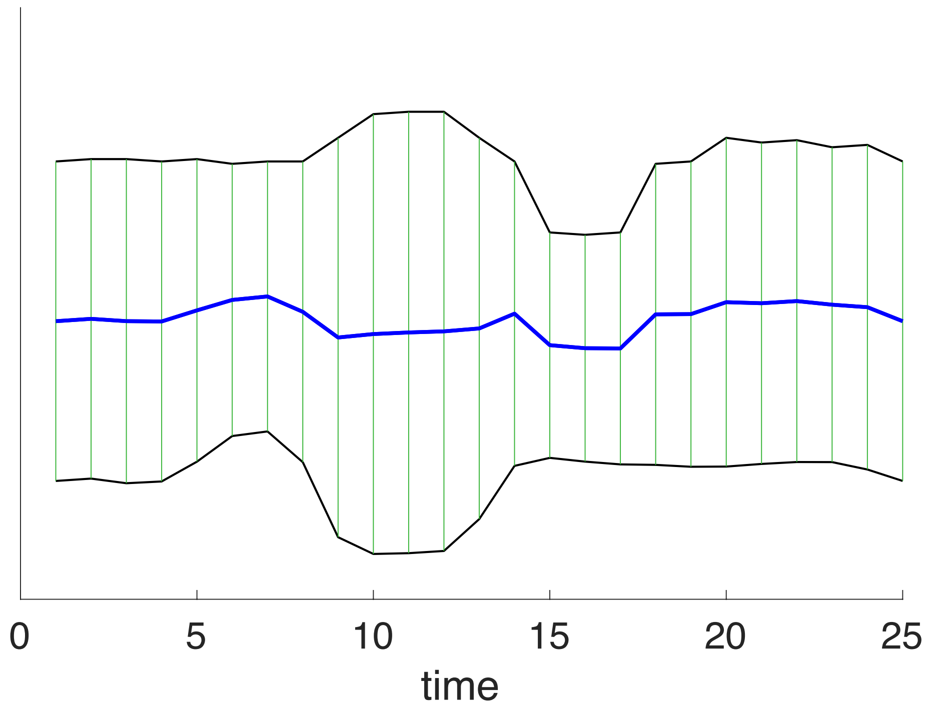
- Features
  - General shape of time series
  - Valleys
  - Mountains



- Assumptions
  - Varying speed (deformations in time)
  - Varying length

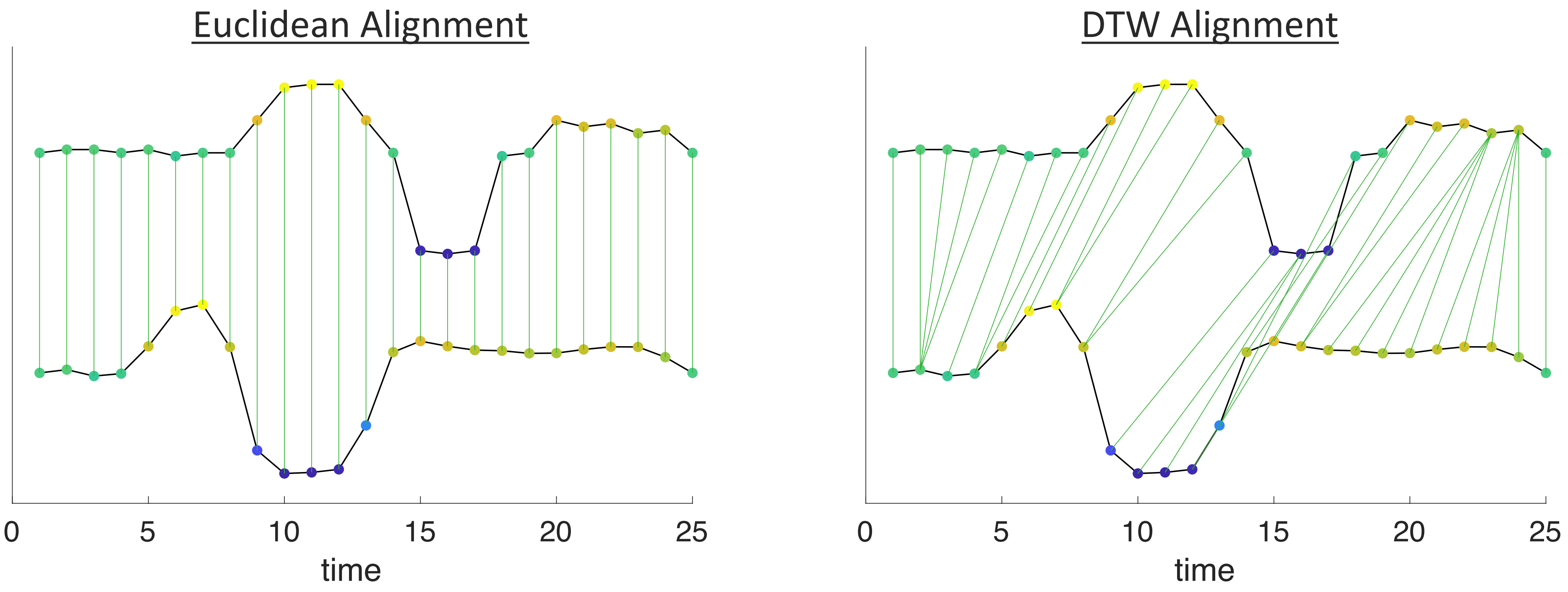
Motivation

- The **arithmetic mean / Euclidean mean** may not be appropriate, because it averages pointwise and does not account for temporal variations



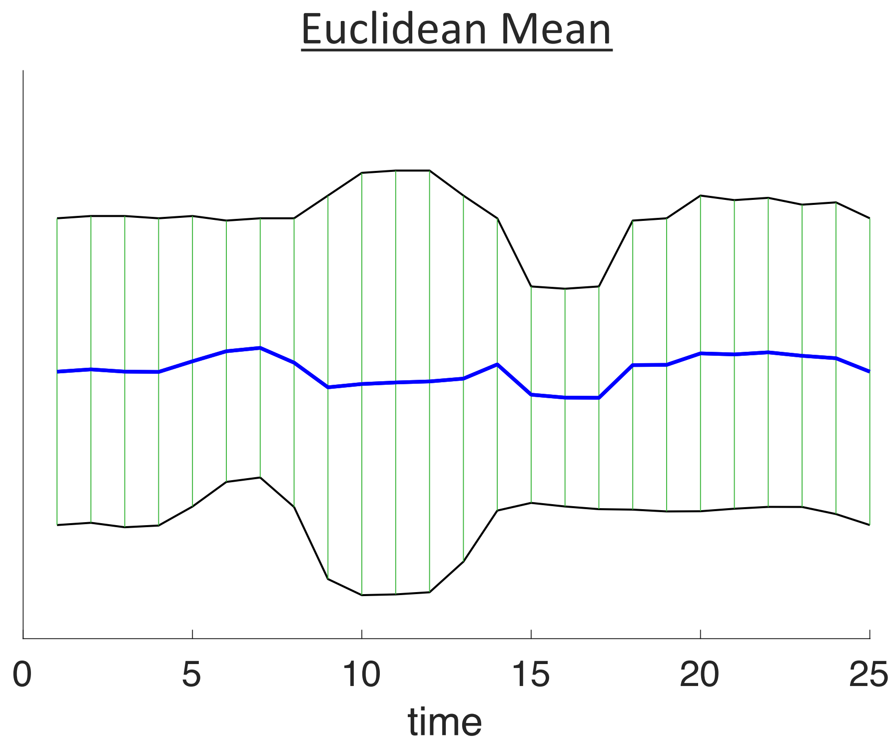
Motivation

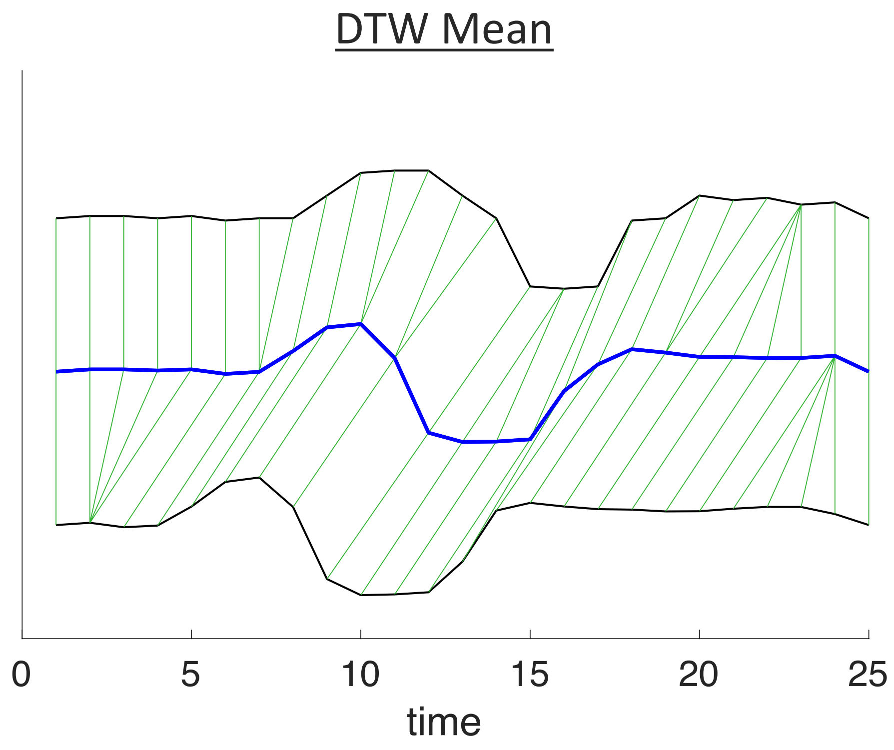
- One standard approach to cope with temporal variations is **Dynamic Time Warping (DTW)**, which tries to find more natural alignments.



Motivation

- **Idea: Average under Dynamic Time Warping**





Dynamic Time Warping

- A **time series** is a sequence $x = (x_1, \ldots, x_n)$ of elements $x_i \in \mathbb{R}$.

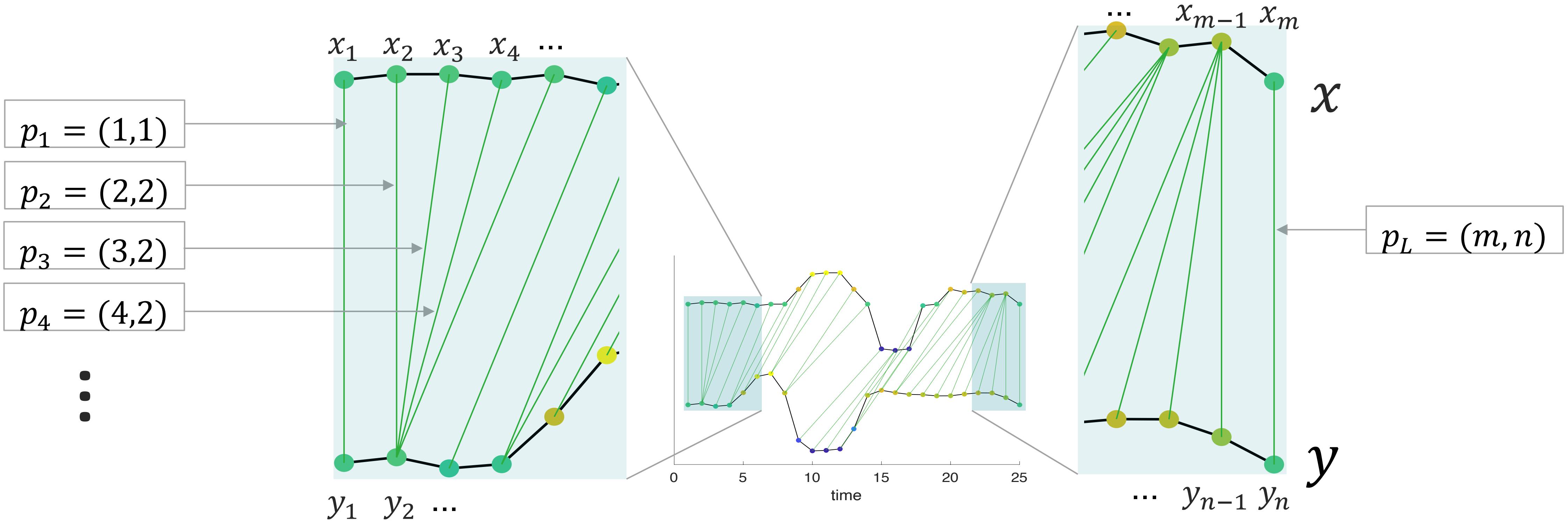
- **Notation**
  - $\mathcal{T}_n = \mathbb{R}^n$ Set of all time series of length n
  - $\mathcal{T} = \bigcup_{n \in \mathbb{N}} \mathcal{T}_n$ Set of all time series of finite length

Dynamic Time Warping

- Denote $[n] = \{1, \dots, n\}$
- A **warping path** of order $m \times n$ is a sequence $p = (p_1, \dots, p_L)$ of index pairs $p_l = (i_l, j_l) \in [m]\times[n]$ such that
  - $p_1 = (1,1)$ and $p_L = (m, n)$ (Boundary condition)
  - $p_{l+1} - p_l \in \{(1,0), (0,1), (1,1)\}$ (Step condition)
- We denote the set of all warping paths of order $m \times n$ by $\mathcal{P}_{m,n}$

Dynamic Time Warping

- A **warping path** $p = (p_1, \dots, p_L)$ defines an alignment between two time series $x = (x_1, \dots, x_m) \in \mathcal{T}_m$ and $y = (y_1, \dots, y_n) \in \mathcal{T}_n$



Dynamic Time Warping

- The **cost** of aligning time series x and y along warping path p is defined as

$$C_p(x, y) = \sqrt{\sum_{(i,j)\in p} \left|x_i - y_j\right|^2}$$

- The (non-metric!) **DTW distance** between two time series x and y is defined by

$$\mathrm{dtw}(x, y) = \min_{p\in\mathcal{P}_{m,n}} C_p(x, y)$$

- Finding an optimal warping path takes $\mathcal{O}(mn)$ time using a dynamic program [1].

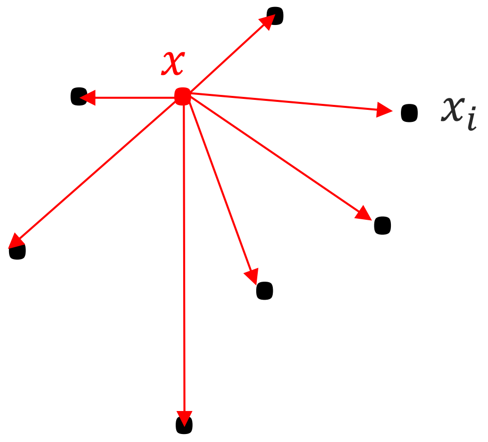
Fréchet Mean under DTW

- **Motivation**
  - The DTW spaces $(\mathcal{T}_m, \mathrm{dtw})$ and $(\mathcal{T}, \mathrm{dtw})$ do not possess a meaningful addition operator which is consistent with the dtw distance => Arithmetic mean is undefined

→ How can we **generalize** the concept of arithmetic mean to DTW spaces?

Fréchet Mean under DTW

Motivation: Fréchet Mean in Euclidean Spaces

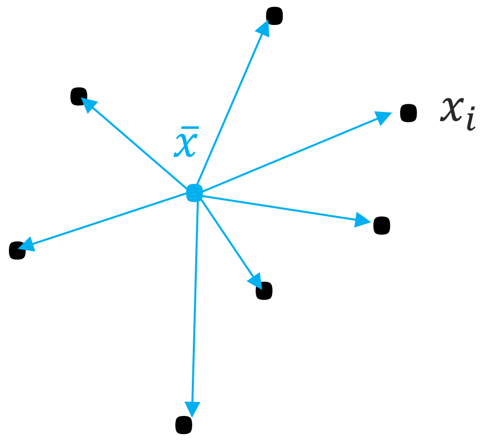


- $X = (x_1, \dots, x_N)$ i.i.d. sample of elements $x_i \in \mathbb{R}^n$ from distribution P

- **Fréchet function** (for Euclidean distance)
  - $F(x) = \frac{1}{N} \sum_{i=1}^{N} \|x - x_i\|_2^2$

Fréchet Mean under DTW

Motivation: Fréchet Mean in Euclidean Spaces



- $X = (x_1, \dots, x_N)$ i.i.d. sample of elements $x_i \in \mathbb{R}^n$ from distribution P

- **Fréchet function** (for Euclidean distance)
  - $F(x) = \frac{1}{N} \sum_{i=1}^{N} \|x - x_i\|_2^2$

- **Result:** The **arithmetic mean** $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ is the unique minimizer of the Fréchet function. The value $F(\bar{x})$ is the **sample variance of** X.

Fréchet Mean under DTW

Fréchet Mean in DTW Spaces

- $X = (x_1, \ldots, x_N)$ i.i.d. sample of time series $x_i \in \mathcal{T}$ from distribution P

- **Fréchet function**
  - $F: \mathcal{T} \to \mathbb{R}, \qquad F(x) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{dtw}(x, x_i)^2$

- **Definition.** A **Sample Fréchet Mean** is any minimizer x^* of the Fréchet function. The value $\mathrm{F}(x^*)$ is the **sample variance** of X.

Fréchet Mean under DTW

In summary, we need to solve the following optimization problem

- **DTW Mean problem**

$$\min_{x \in \mathcal{T}} F(x) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{dtw}(x, x_i)^2$$

Fréchet Mean under DTW

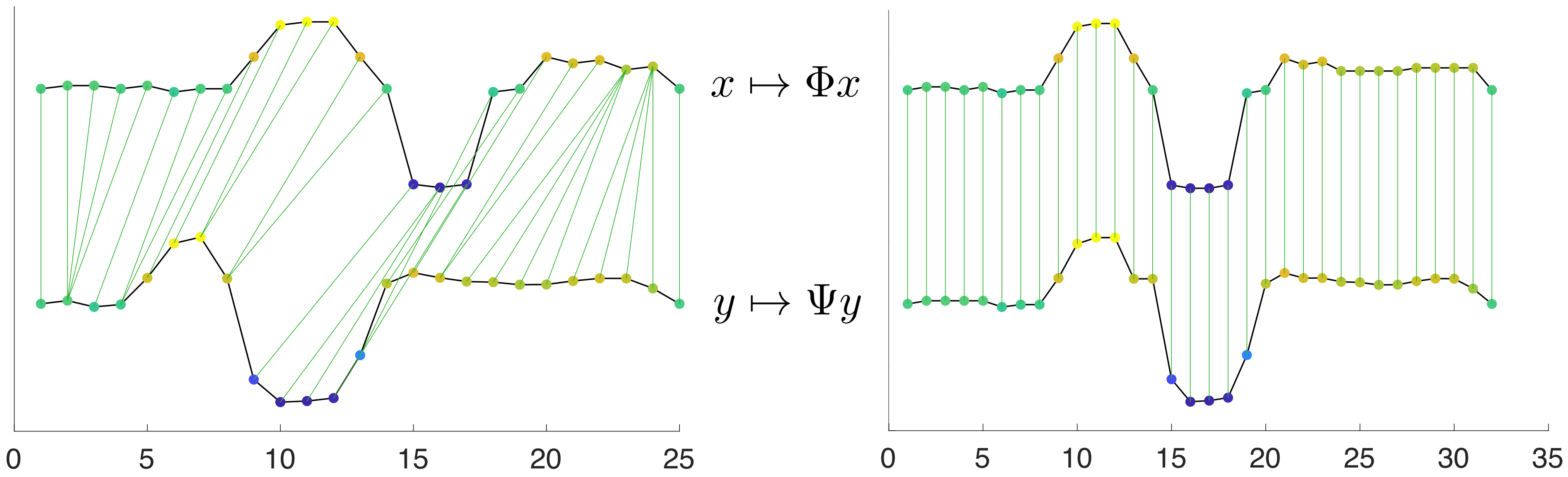
Properties

- **Existence:** A solution exists [4]
- **Uniqueness:** Fréchet means are generally not unique [4]
- **Complexity:** The optimization problem NP-hard [5]
- **Exact Solutions:** An essentially optimal-time ($\mathcal{O}(n^{2N}2^N nN)$) algorithm for finding exact solutions was proposed in [6]. ($n =$ max time series length, $N =$ sample size)
- **Heuristic Solutions:** Generalized gradient methods [2]
- **Analysis:** Nonsmooth Analysis was provided for the restricted DTW Mean problem in [2]
- **Consistency:** The sample mean of the restricted DTW Mean problem is a strongly consistent estimator of the population mean.

Nonsmooth Analysis for the restricted DTW Mean Problem

- **Proposition [2]**: Let $x \in \mathcal{T}_m$ and $y \in \mathcal{T}_n$ be two time series and $p \in \mathcal{P}_{m,n}$ be a warping path of length L. There exist embeddings $\Phi: \mathcal{T}_m \to \mathbb{R}^L$ and $\Psi: \mathcal{T}_n \to \mathbb{R}^L$ such that

$$C_p(x, y) = \|\Phi x - \Psi y\|_2$$



Nonsmooth Analysis for the restricted DTW Mean Problem

- Using the lemma, we can comfortably compute gradients of the squared cost [2]

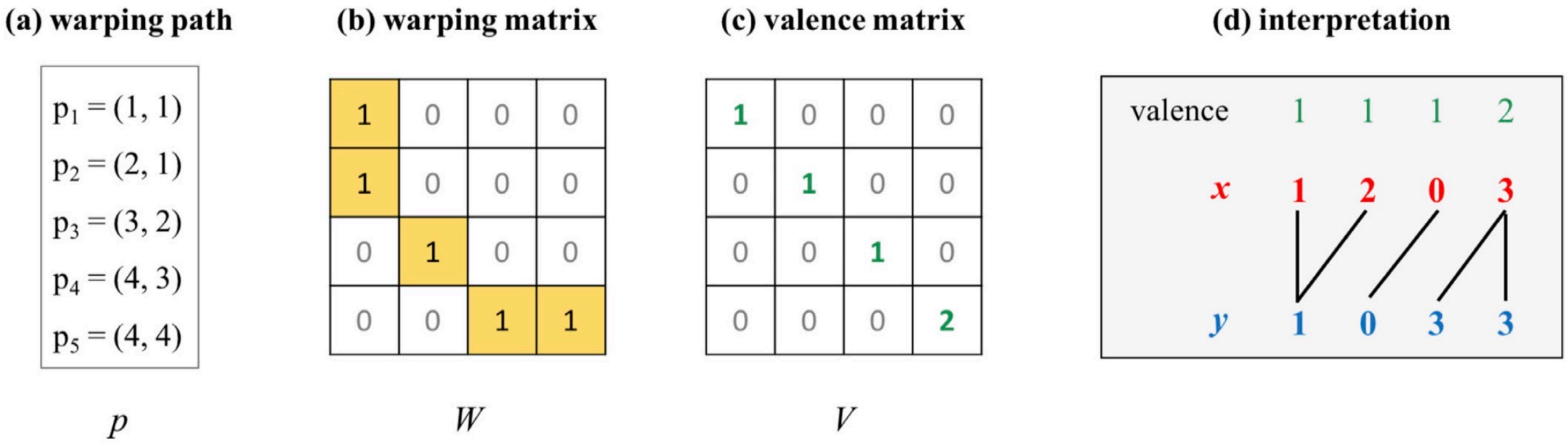
$$\nabla_x C_p^2(x, y) = \nabla_x \|\Phi x - \Psi y\|_2^2 = \Phi^T \Phi x - \Phi^T \Psi y = Vx - Wy$$

- ... and determine the unique minimum of the cost for a given warping path

$$x = V^{-1} W y$$

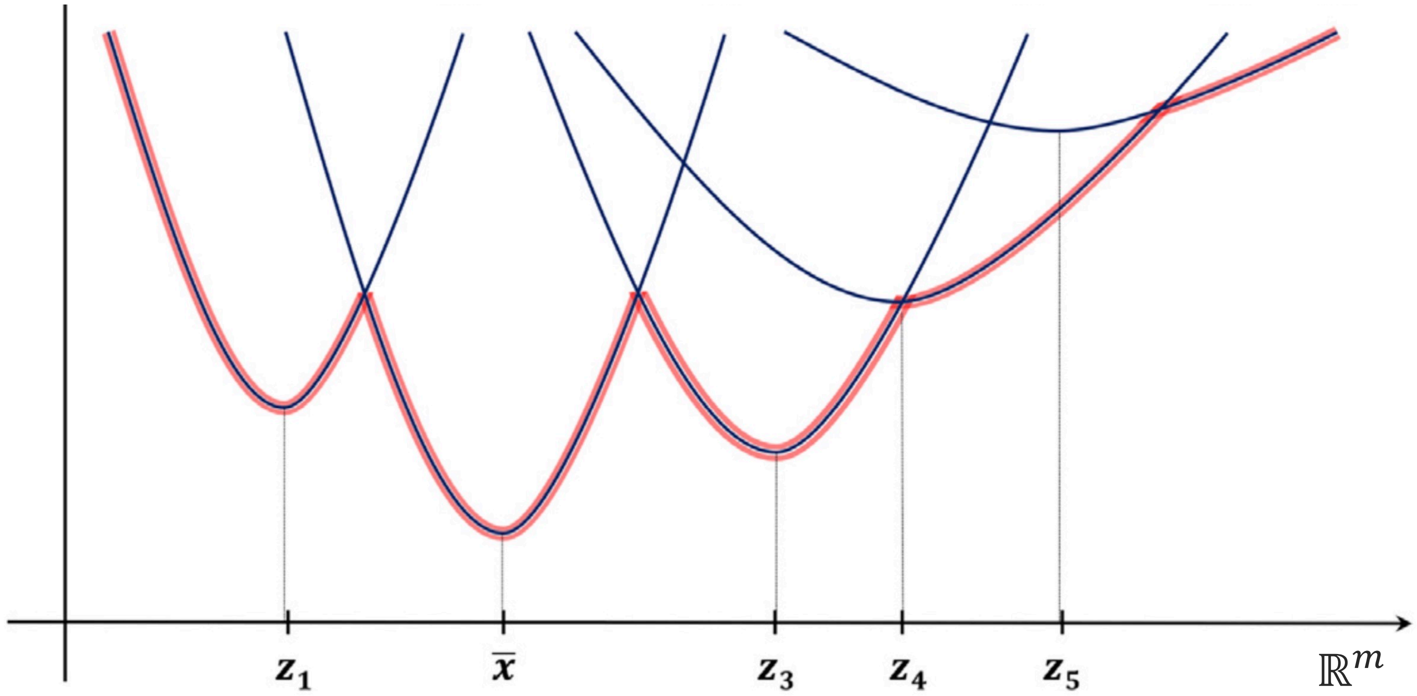
- V is the (diagonal) **valence matrix**. It counts for each diagonal entry $V_{i,i}$ how many elements from time series y are aligned to x_i by warping path p.
- W is the **warping matrix** which is often used to illustrate a warping path.

Nonsmooth Analysis for the restricted DTW Mean Problem



Nonsmooth Analysis for the restricted DTW Mean Problem

- A decomposition reveals that the Fréchet function F_m is the pointwise minimum of differentiable, convex **component functions** [2].



DAI-Labor

Nonsmooth Analysis for the restricted DTW Mean Problem

- At differential points, the gradient of F_m can be obtained by computing the gradient of the active component function.
- At nondifferential points, there exists a concept of generalized gradient ([7]).
- For gradient descent based optimizers, a gradient of one of the active component functions can be used.

Nonsmooth Analysis for the restricted DTW Mean Problem

- **Necessary conditions of optimality [2].** If z is a local minimizer of F_m, there exist warping paths $p_1, \dots, p_N$ inducing valence matrices $V_1, \dots, V_N$ and warping matrices $W_1, \dots, W_N$ such that
  - $F_m(z) = \sum_i C_{p_i}(z, x_i)^2$ (i.e. warping paths are optimal)
  - $z = (\sum_i V_i)^{-1}(\sum_i W_i x_i)$

- Hence, any sample mean is of the form

$$z = \left(\sum_i V_i\right)^{-1}\left(\sum_i W_i x_i\right)$$

Nonsmooth Analysis for the restricted DTW Mean Problem

- DTW Mean

$$z = \left(\sum_i V_i\right)^{-1} \left(\sum_i W_i x_i\right)$$

Each z_j is the arithmetic mean of all elements that are aligned to z_j

Normalization

Sum of (transformed) samples

- Euclidean Mean

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Nonsmooth Analysis for the restricted DTW Mean Problem

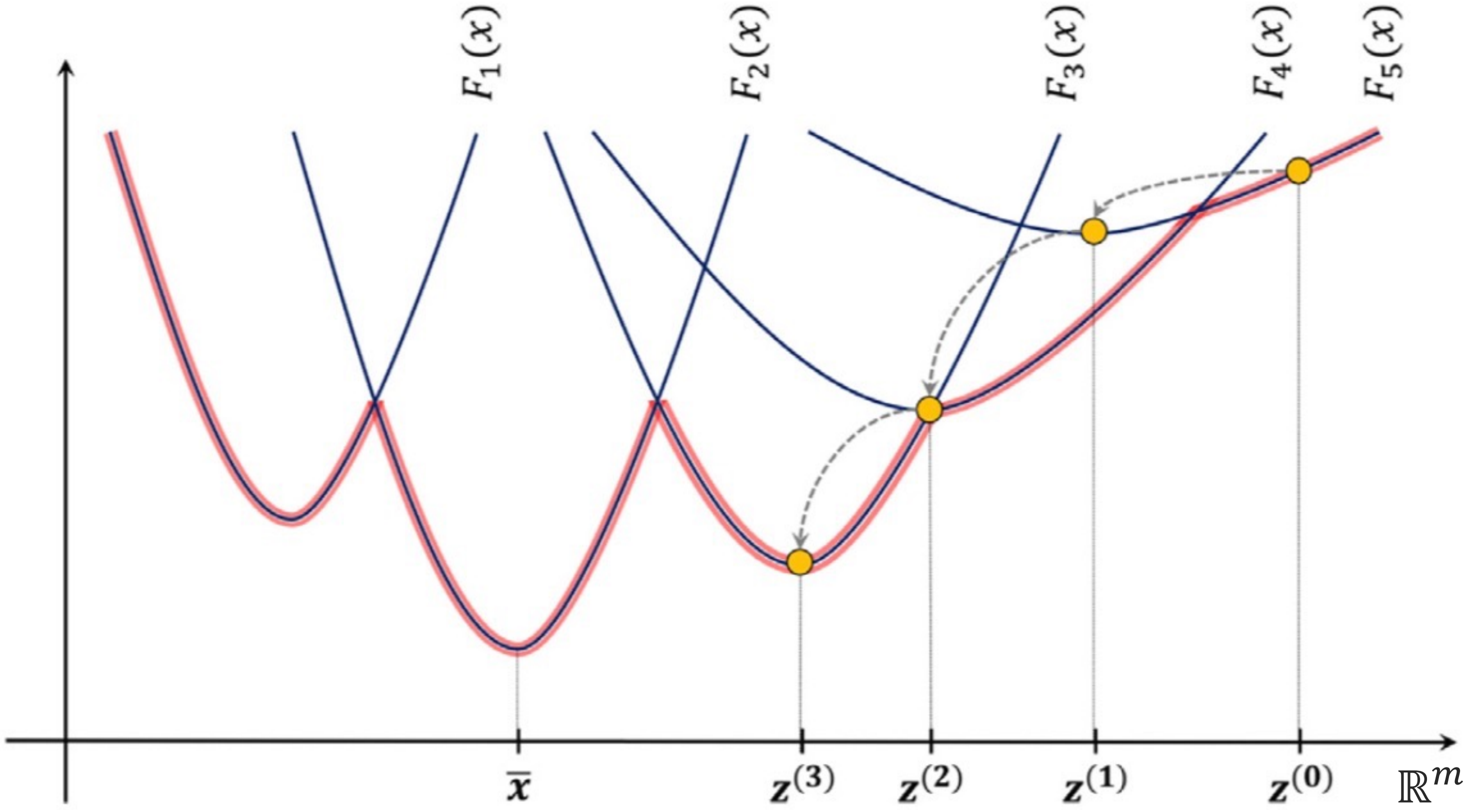
- We know how a mean looks like, the most difficult problem is to find the correct warping paths

- Sample Mean algorithms
  - Exact Dynamic Program [6]
  - Stochastic Subgradient Methods [2]
  - Majorize-Minimize Algorithm, known as DTW Barycenter Averaging (DBA) [2,7]

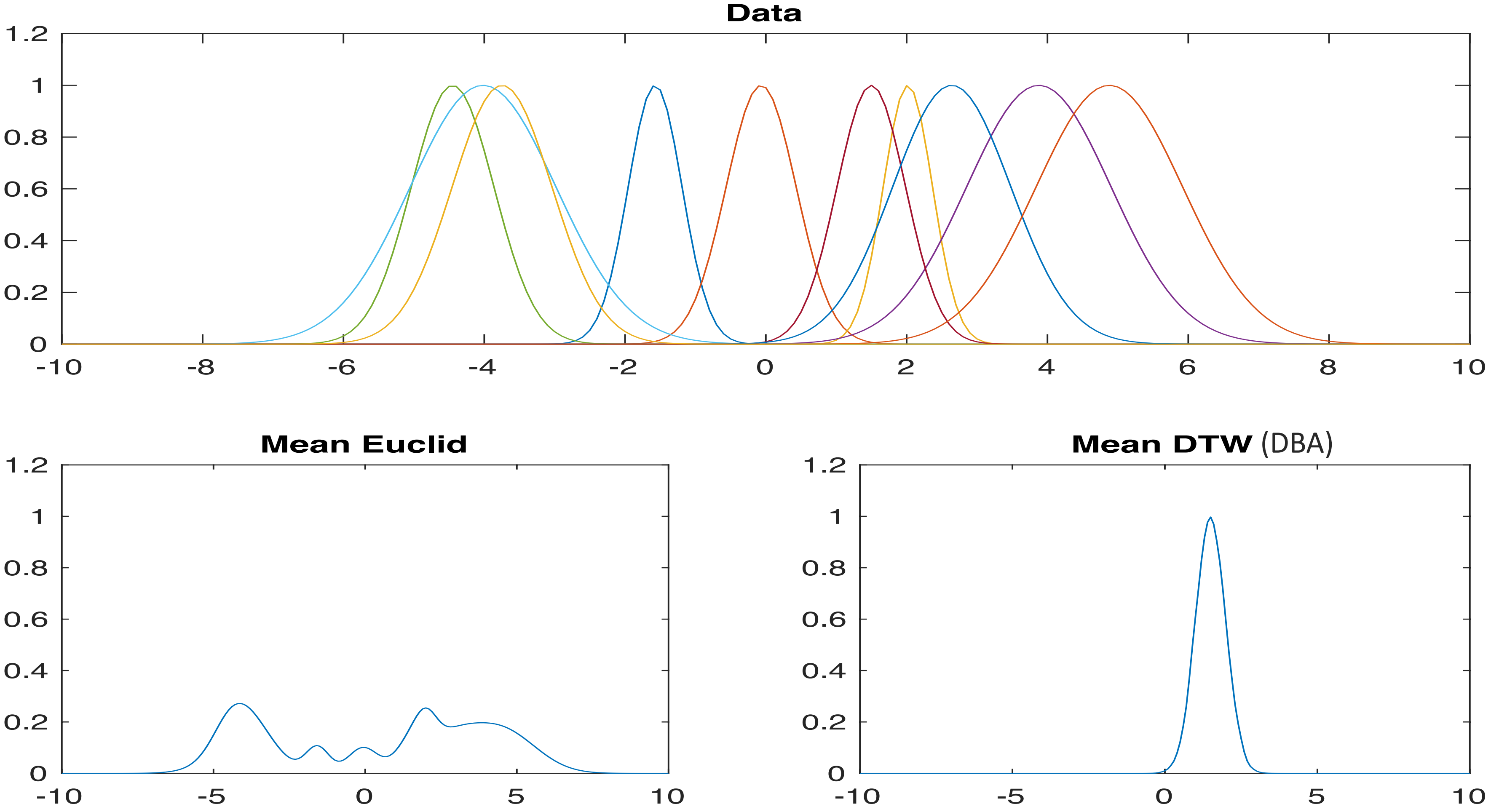
- Many more (possibly global) optimization techniques are applicable.

Nonsmooth Analysis for the restricted DTW Mean Problem

Majorize-Minimize Algorithm (DBA – DTW Barycenter Averaging)



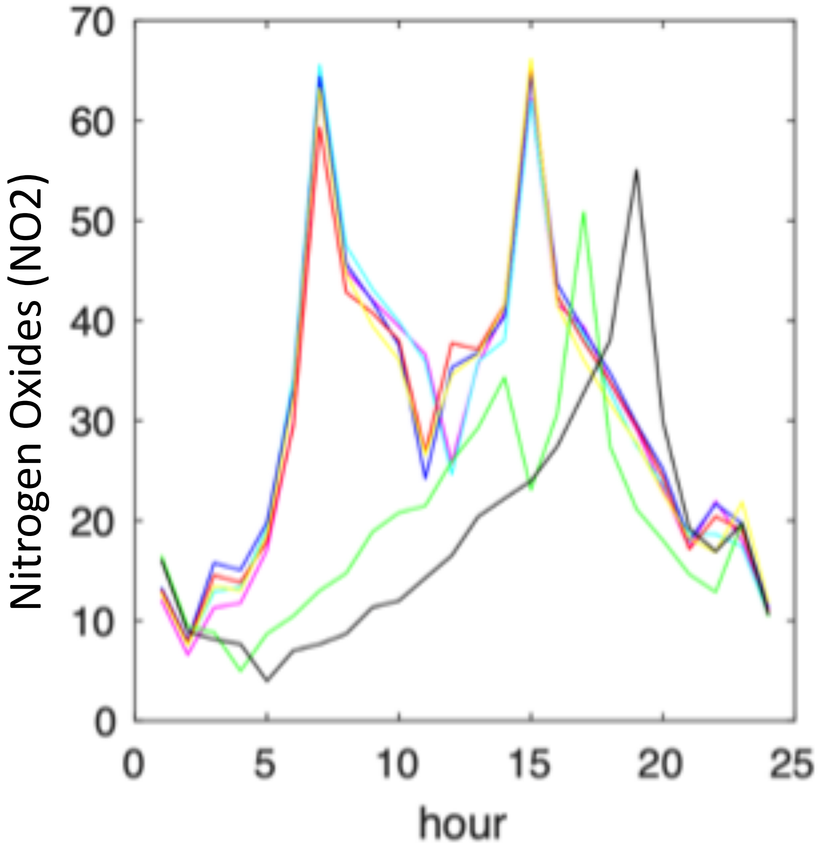
Example



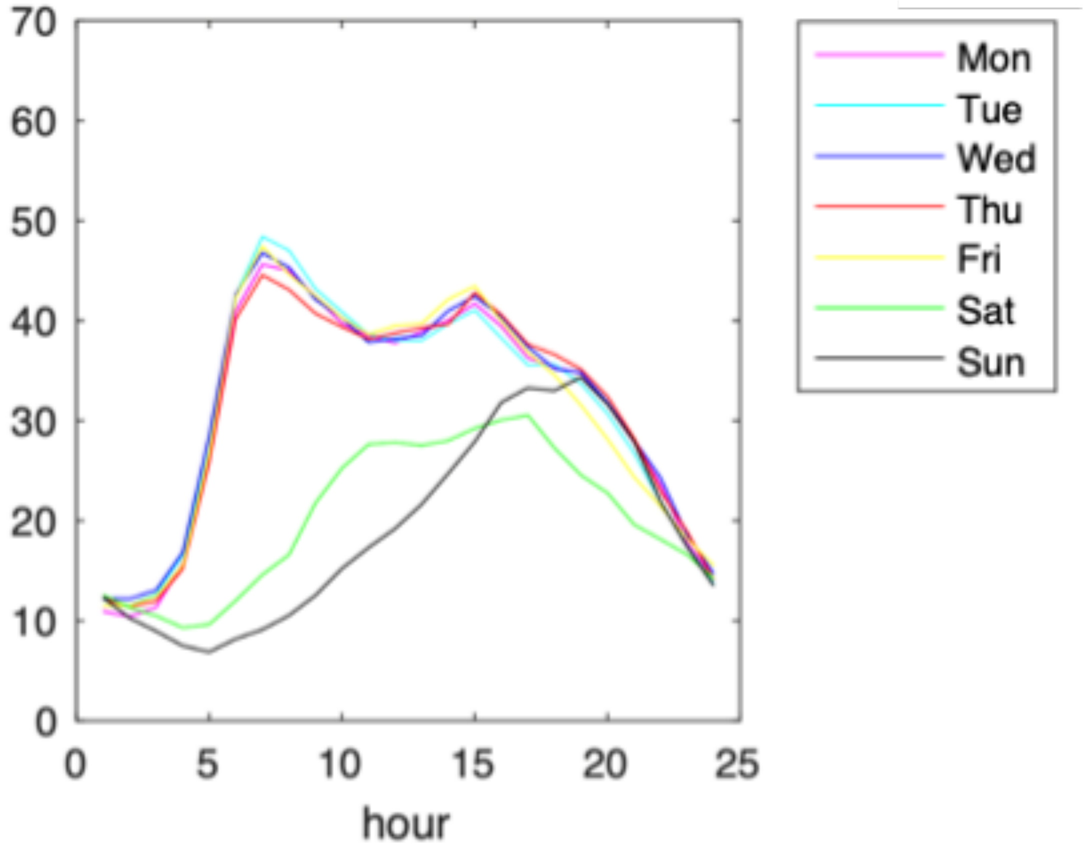
Data
1.2
1
0.8
0.6
0.4
0.2
0
-10
-8
-6
-4
-2
0
2
4
6
8
10
Mean Euclid
1.2
1
0.8
0.6
0.4
0.2
0
-10
-5
0
5
10
Mean DTW (DBA)
1.2
1
0.8
0.6
0.4
0.2
0
-10
-5
0
5
10

Example: Emmission profiles

DTW Mean



Euclidean Mean



Data Imputation with Fréchet Means

Let $x = (x_1, \ldots, x_n)$ be fixed and $y = (y_1, \ldots, y_m)$ with missing values at indices $\mathcal{M} \subseteq \{1, \ldots, m\}$. We want to solve the imputation problem

$$\min_{\{y_j | j \in \mathcal{M}\}} \operatorname{dtw}(x, y) . \tag{8.3}$$

- Intuitively, x is a time series from which we copy the missing values for y after aligning x and y using DTW

- x could be
  - a time series from the dataset
  - a Fréchet Mean (e.g. from a k-Means algorithm on the dataset)

Milestones

- Goals for MS1
  - Intro (Motivation, Problem, Goal)
  - Literature (related work and approaches we want to consider)
  - Our approach
  - Planned experimental setup
  - Time Plan / Work packages
  - Current status (some implementations should be done)

Milestones

- Goals for MS2
  - Implementations (nearly) finished
  - Preliminary experimental results
  - Remaining tasks

- Goals for final Presentation
  - Final results
  - Discussion
  - Conclusions

Suggested next steps

- Start with Literature research and project plan
- Define workpackages and assign tasks to team members

- Communication tools
  - Matrix, hosted by TU (https://chat.tu-berlin.de)
  - Slack
  - ISIS Group Forum

- Github TUBIT
  - Please invite me

References

- [1] Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, *26*(1), 43-49.
- **[2] Schultz, D., & Jain, B. (2018). Nonsmooth analysis and subgradient methods for averaging in dynamic time warping spaces. *Pattern Recognition*, *74*, 340-358.**
- [3] Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. Annales de l'institut Henri Poincaré, 215–310, 1948.
- [4] Jain, B. J., & Schultz, D. (2016). On the existence of a sample mean in dynamic time warping spaces. *arXiv preprint arXiv:1610.04460*.
- [5] Bulteau, L., Froese, V., & Niedermeier, R. (2018). Hardness of consensus problems for circular strings and time series averaging. *arXiv preprint arXiv:1804.02854*.
- [6] Brill, M., Fluschnik, T., Froese, V., Jain, B., Niedermeier, R., & Schultz, D. (2019). Exact mean computation in dynamic time warping spaces. *Data Mining and Knowledge Discovery*, *33*(1), 252-291.
- **[7] Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, *44*(3), 678-693.**