Soft-DTW: A Differentiable Loss Function for Time Series

Marco Cuturi

université PARIS-SACLAY

ENSAE ParisTech — École nationale de la statistique et de l'administration économique

université PARIS-SACLAY

Mathieu Blondel

NTT

From: Plume App

Ground truth (reality)

From: Plume App

Ground truth (reality)

How wrong was this prediction?

This depends on the loss function used to train the algorithm.

From: Plume App

- In this talk we propose to use the celebrated **Dynamic Time Warping** discrepancy as a loss.

- Loss functions should be **differentiable**. We show that an **appropriate smoothing** , **soft-DTW**, helps

- We apply this to **several problems**:
  - Computation of **barycenters**,
  - Clustering of time series,
  - Learning with structured (time series) output

From: Plume App

0. The DTW Geometry

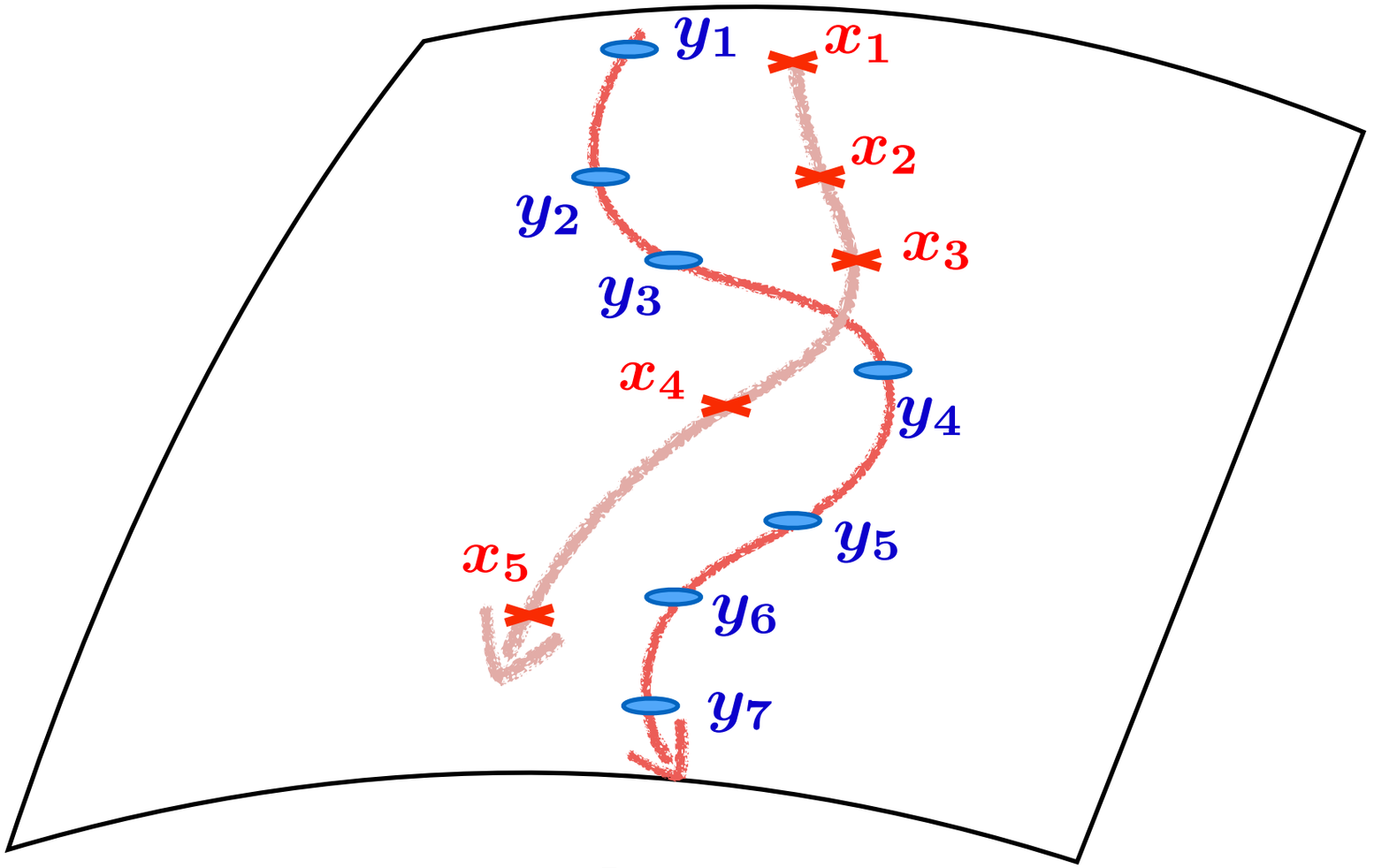
1. Soft-DTW

2. Soft-DTW as a Loss Function

Dynamic Time Warping

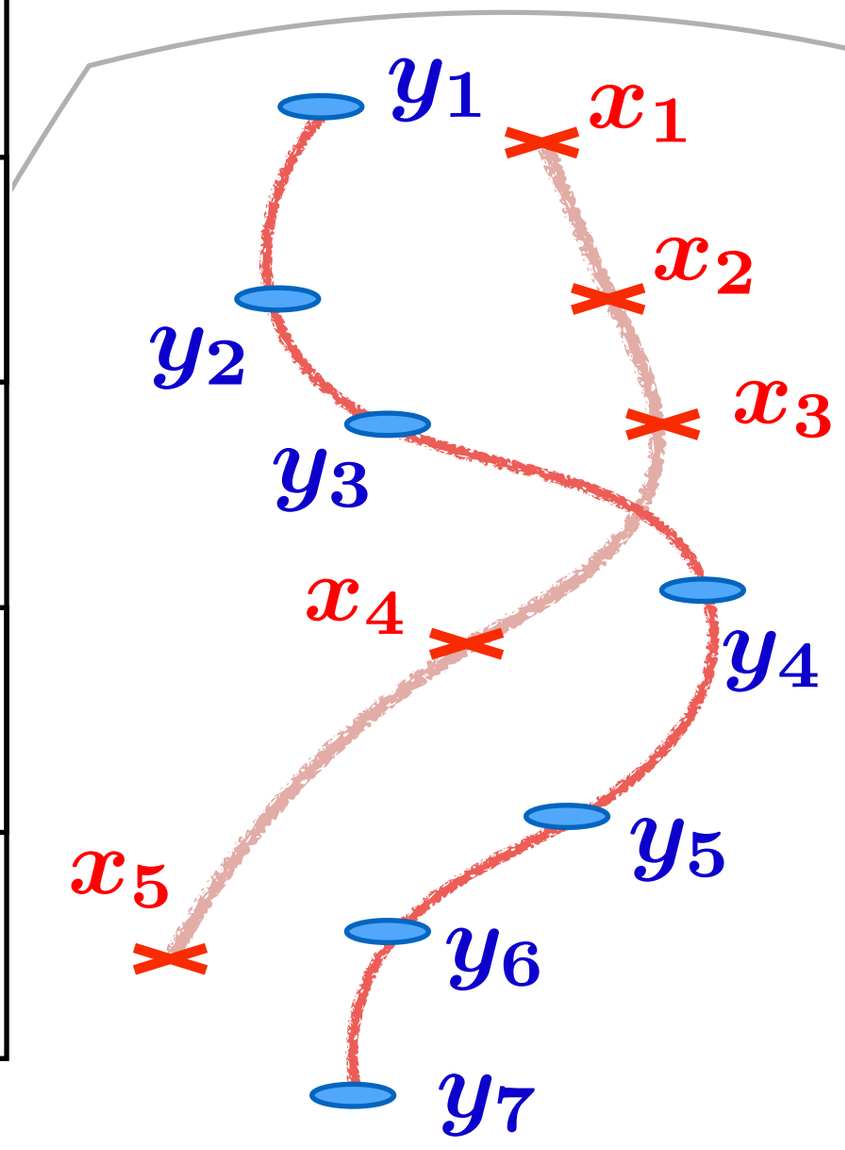
A discrepancy function between two **time series of observations** supported **on a metric space.**

(Ω, δ)

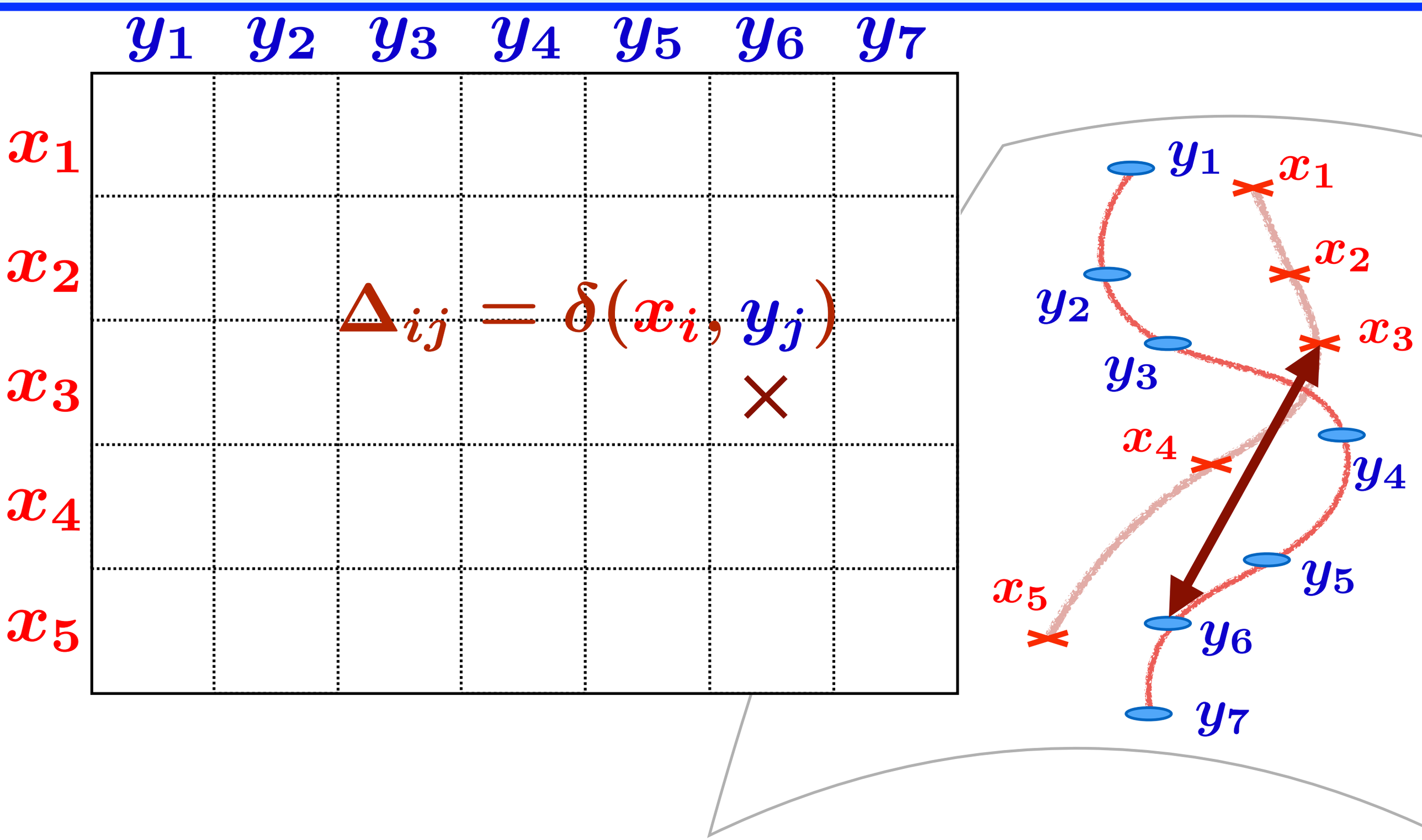


Pairwise Distance Matrix

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1							
x_2							
x_3							
x_4							
x_5							

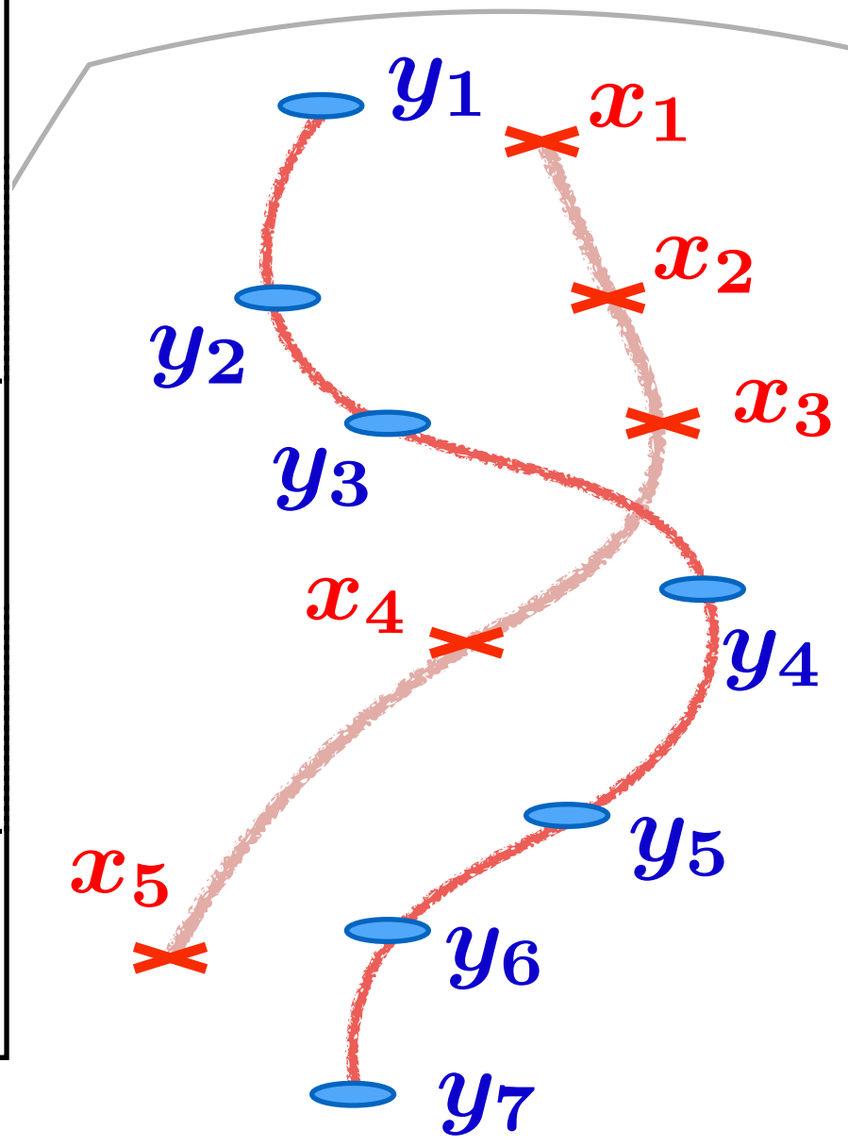


Pairwise Distance Matrix



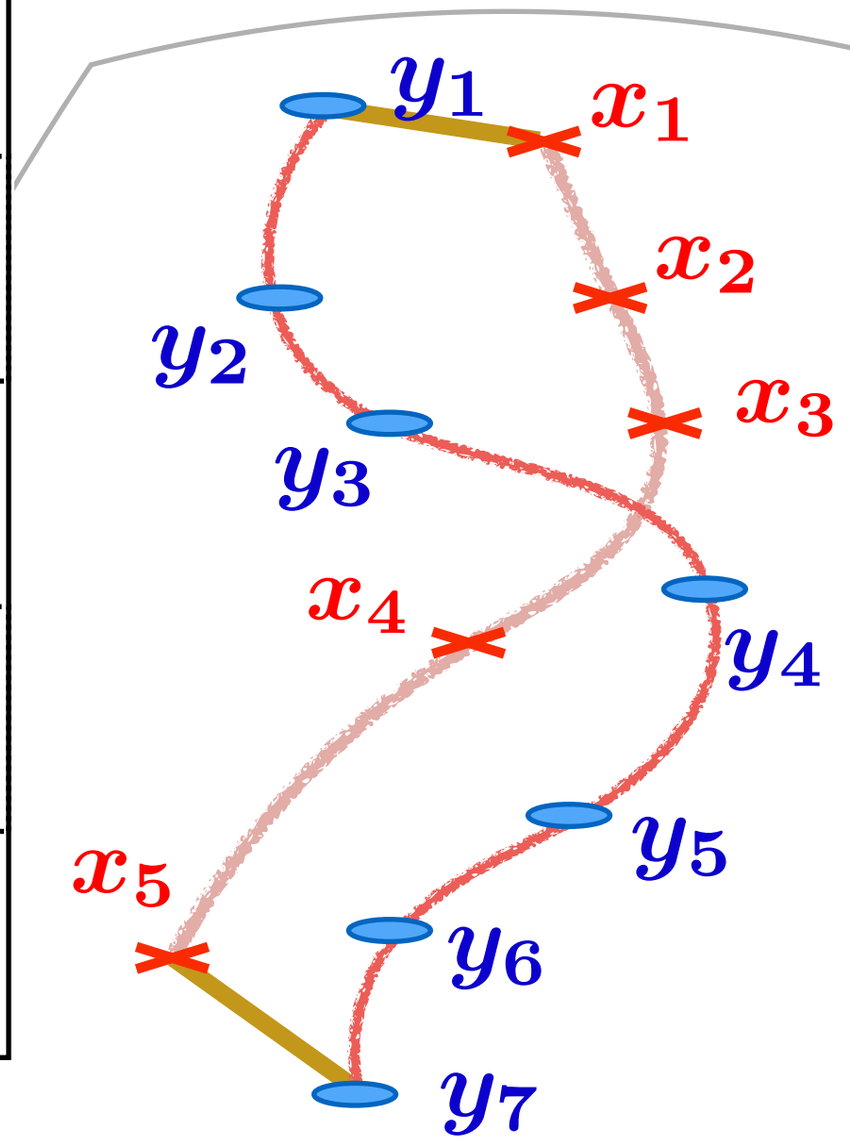
Pairwise Distance Matrix

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1	Δ_{11}	Δ_{12}	Δ_{13}	Δ_{14}	Δ_{15}	Δ_{16}	Δ_{17}
x_2	Δ_{21}	Δ_{22}	Δ_{23}	Δ_{24}	Δ_{25}	Δ_{26}	Δ_{27}
x_3	Δ_{31}	Δ_{32}	Δ_{33}	Δ_{34}	Δ_{35}	Δ_{36}	Δ_{37}
x_4	Δ_{41}	Δ_{42}	Δ_{43}	Δ_{44}	Δ_{45}	Δ_{46}	Δ_{47}
x_5	Δ_{51}	Δ_{52}	Δ_{53}	Δ_{54}	Δ_{55}	Δ_{56}	Δ_{57}

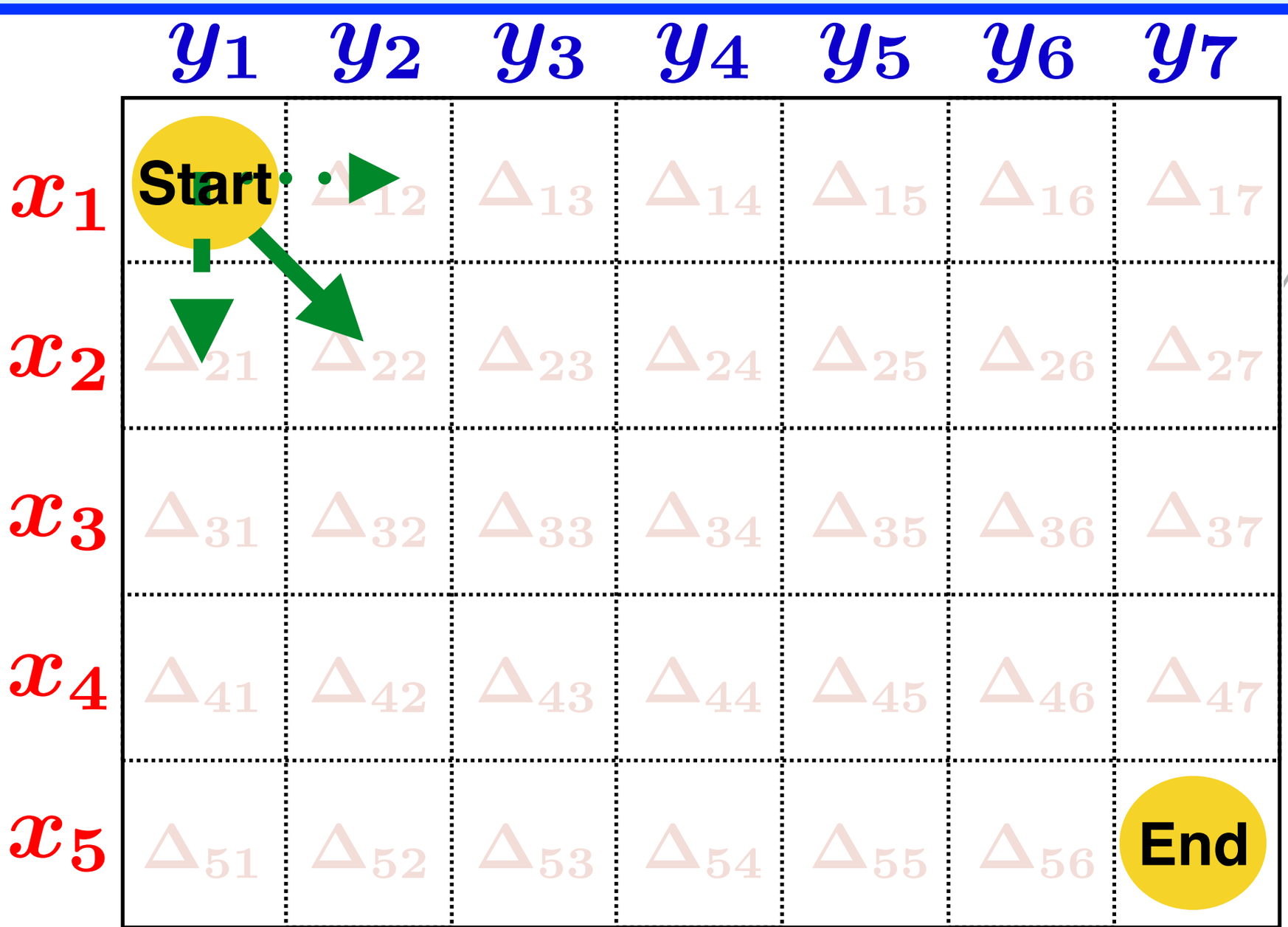


Alignment Path

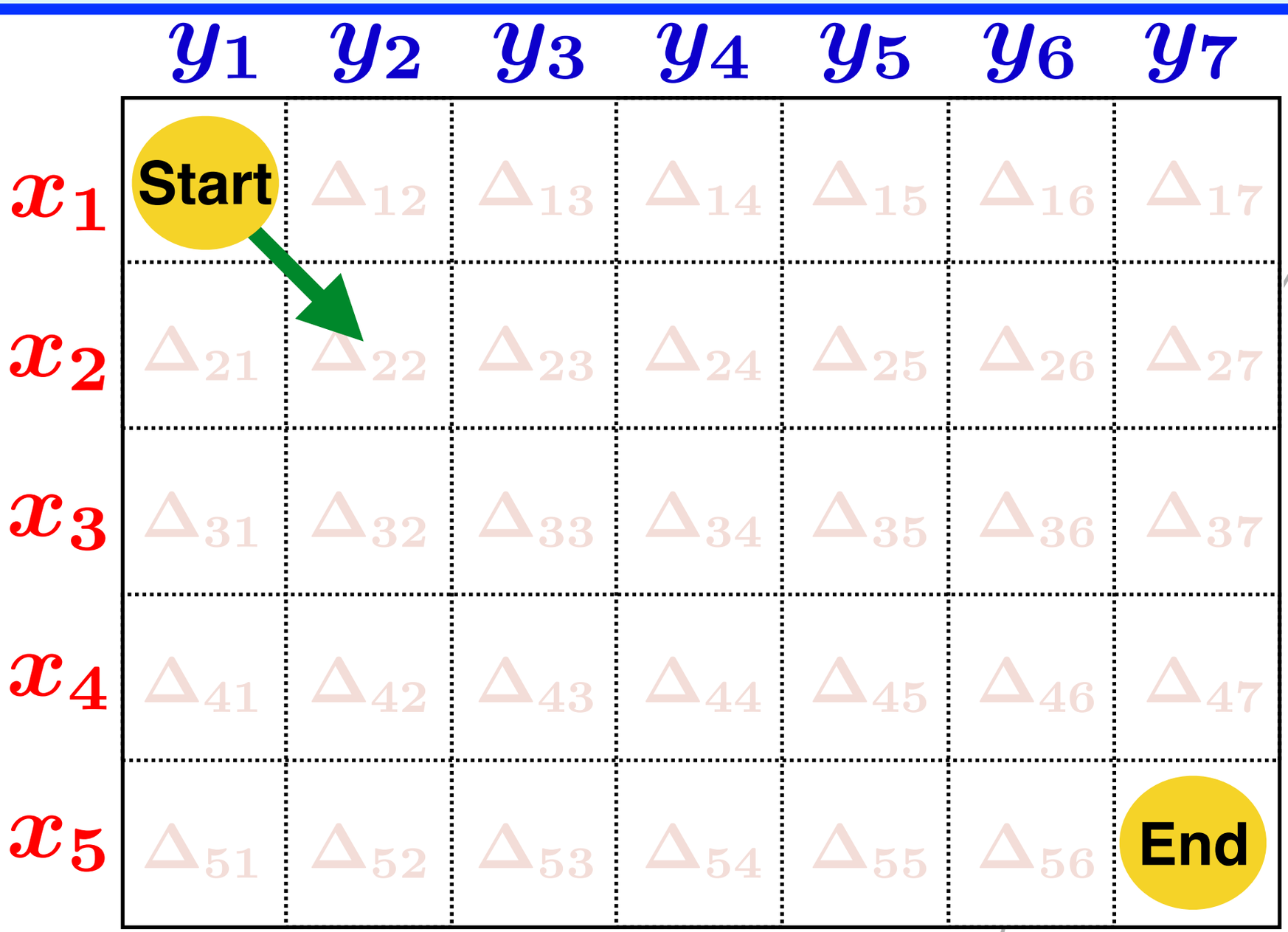
	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1	Start	Δ_{12}	Δ_{13}	Δ_{14}	Δ_{15}	Δ_{16}	Δ_{17}
x_2	Δ_{21}	Δ_{22}	Δ_{23}	Δ_{24}	Δ_{25}	Δ_{26}	Δ_{27}
x_3	Δ_{31}	Δ_{32}	Δ_{33}	Δ_{34}	Δ_{35}	Δ_{36}	Δ_{37}
x_4	Δ_{41}	Δ_{42}	Δ_{43}	Δ_{44}	Δ_{45}	Δ_{46}	Δ_{47}
x_5	Δ_{51}	Δ_{52}	Δ_{53}	Δ_{54}	Δ_{55}	Δ_{56}	End

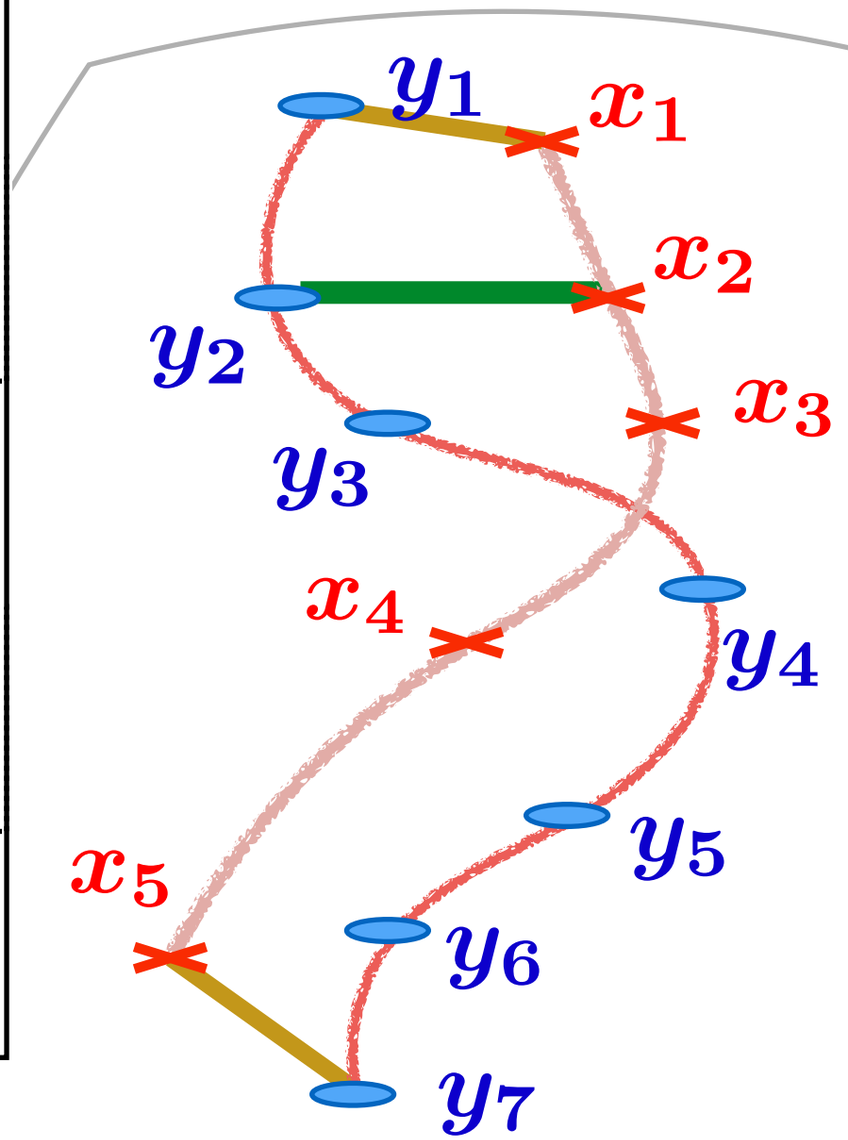


Alignment Path

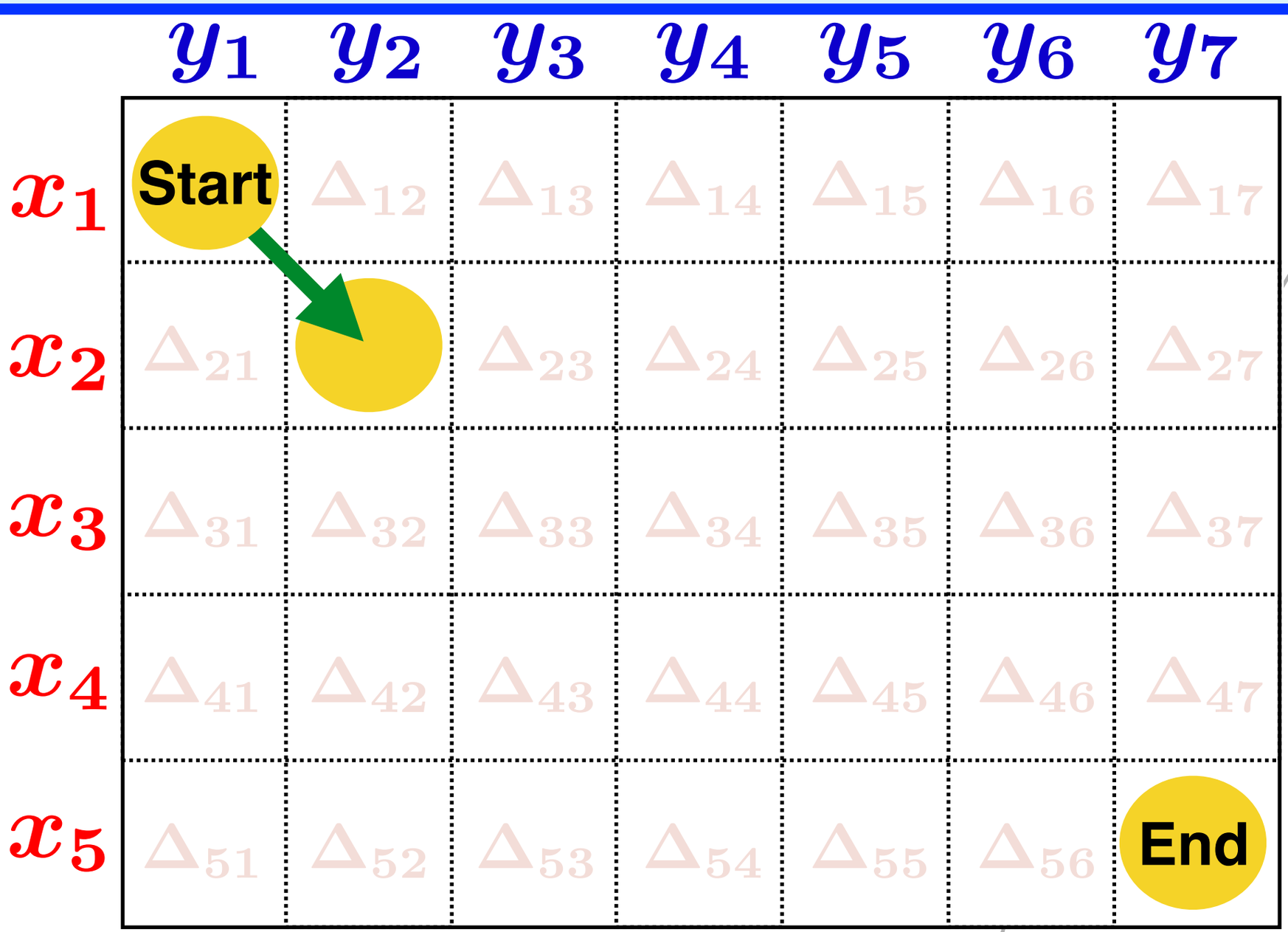


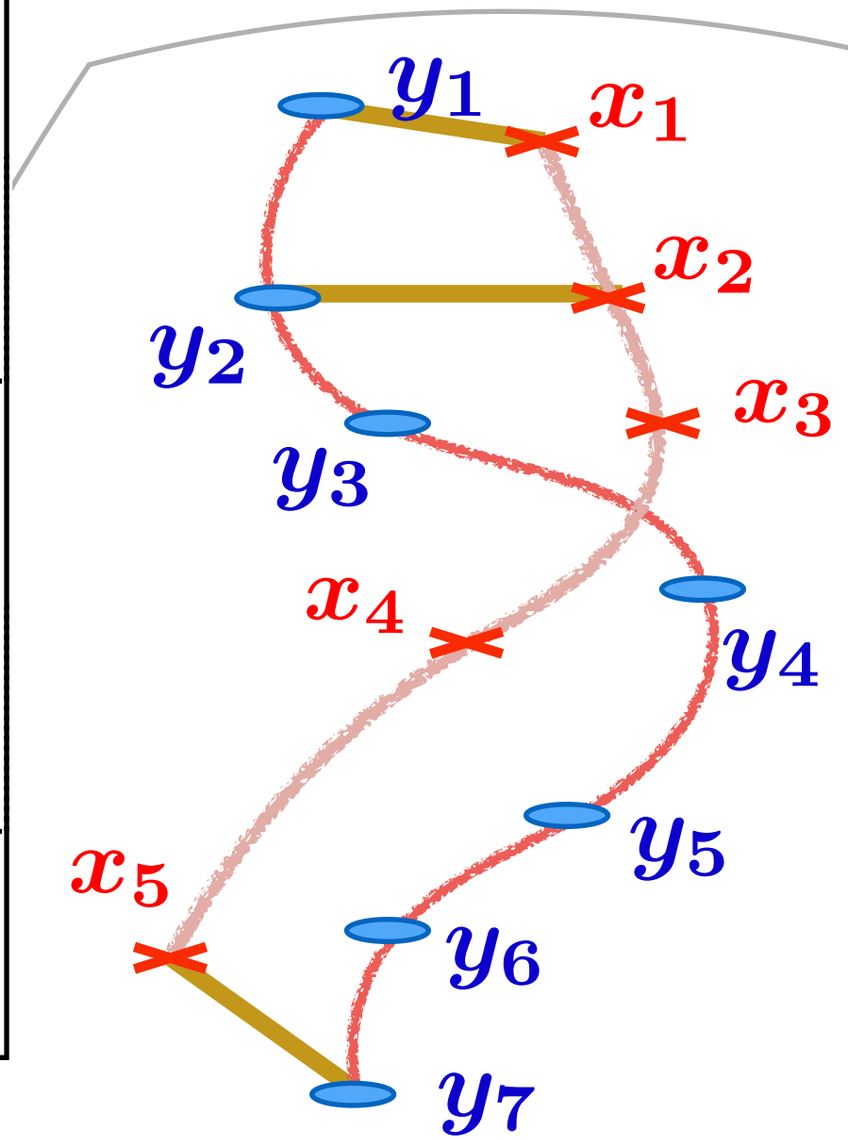
Alignment Path



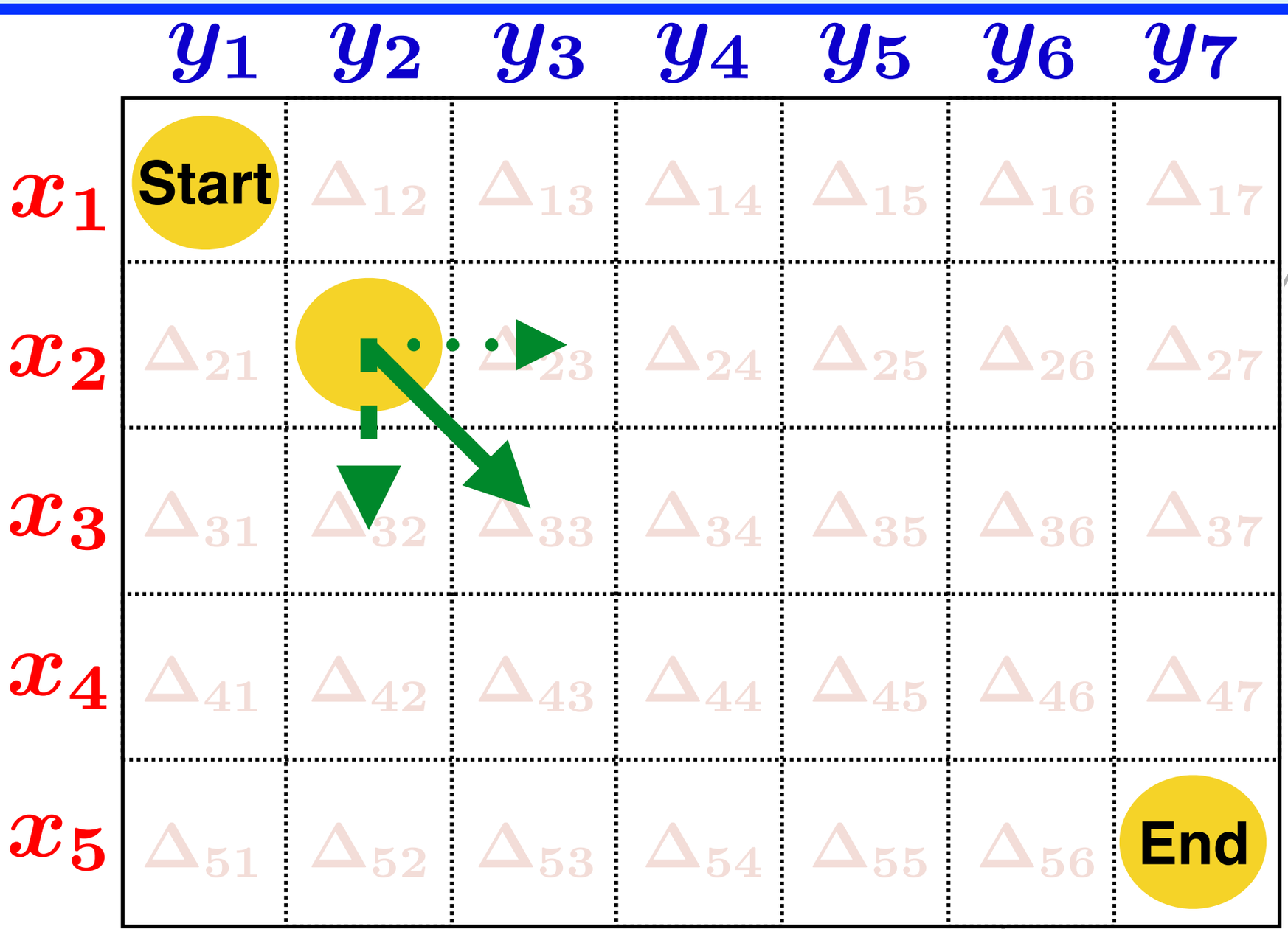


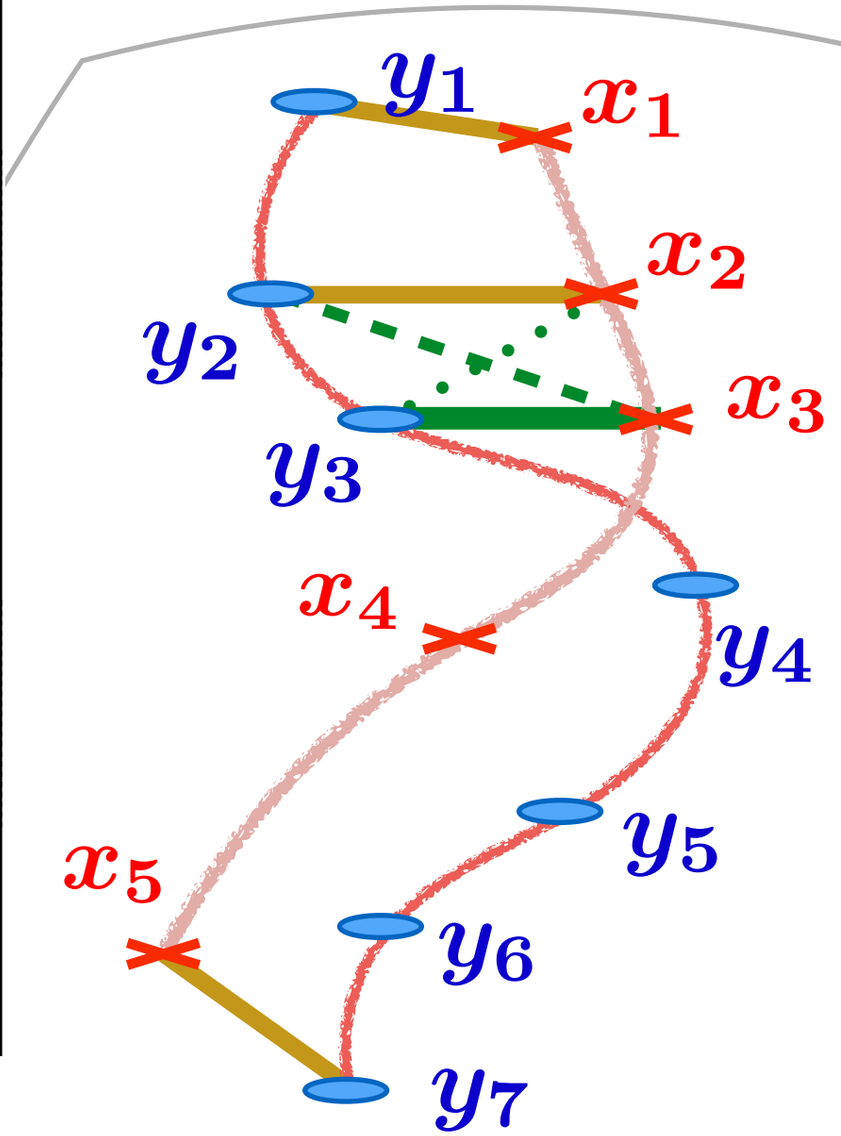
Alignment Path





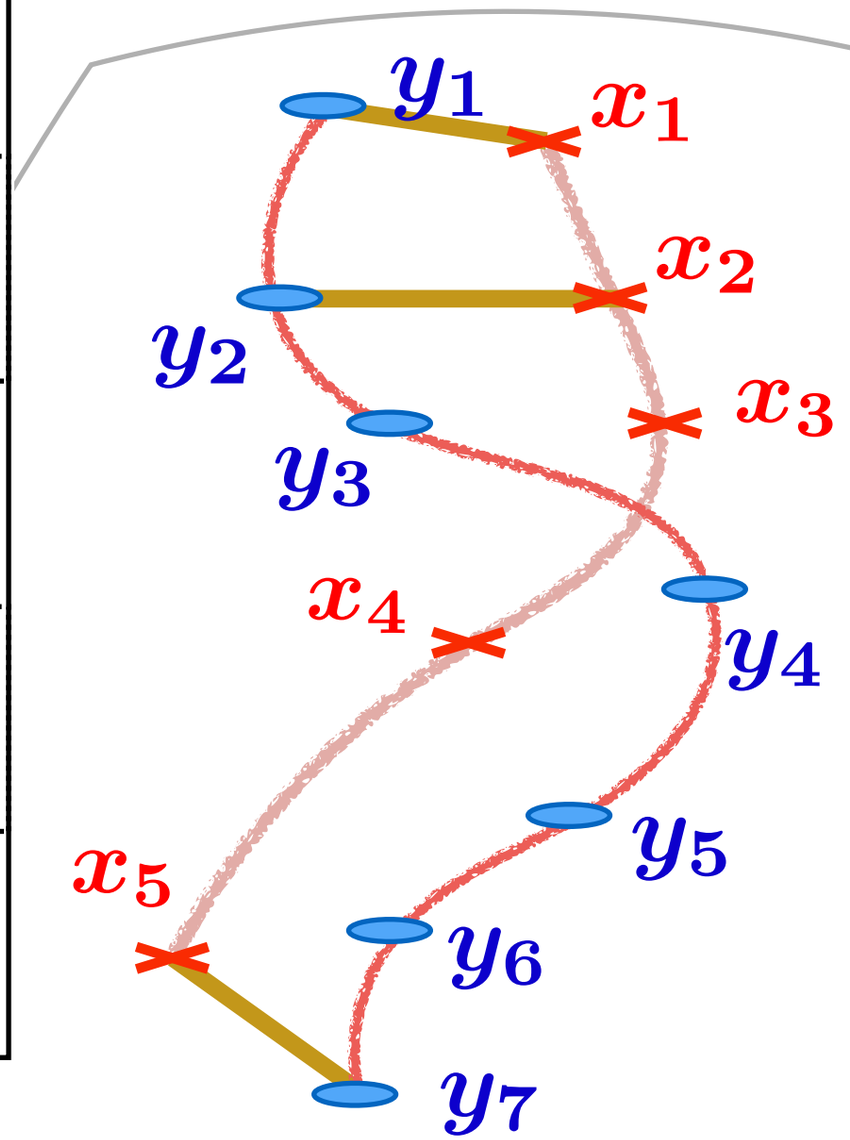
Alignment Path





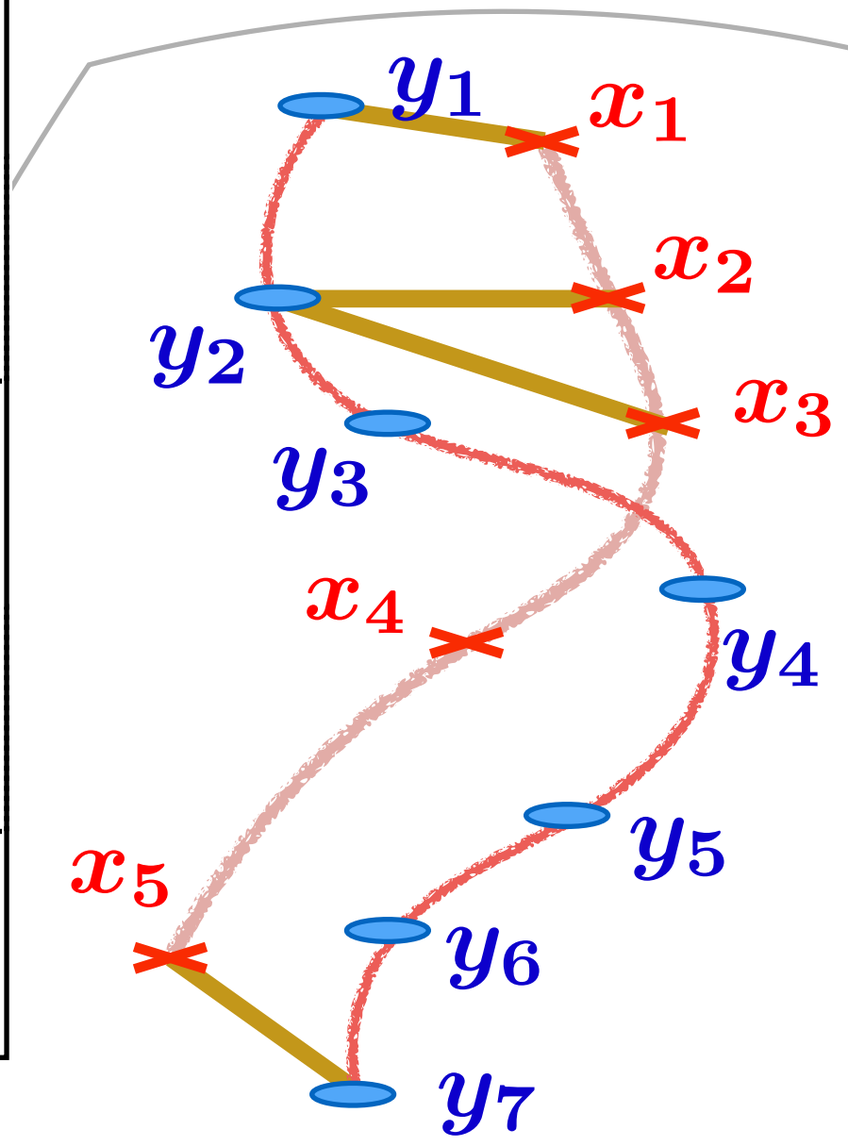
Alignment Path

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1	Start	Δ_{12}	Δ_{13}	Δ_{14}	Δ_{15}	Δ_{16}	Δ_{17}
x_2	Δ_{21}		Δ_{23}	Δ_{24}	Δ_{25}	Δ_{26}	Δ_{27}
x_3	Δ_{31}	Δ_{32}	Δ_{33}	Δ_{34}	Δ_{35}	Δ_{36}	Δ_{37}
x_4	Δ_{41}	Δ_{42}	Δ_{43}	Δ_{44}	Δ_{45}	Δ_{46}	Δ_{47}
x_5	Δ_{51}	Δ_{52}	Δ_{53}	Δ_{54}	Δ_{55}	Δ_{56}	End

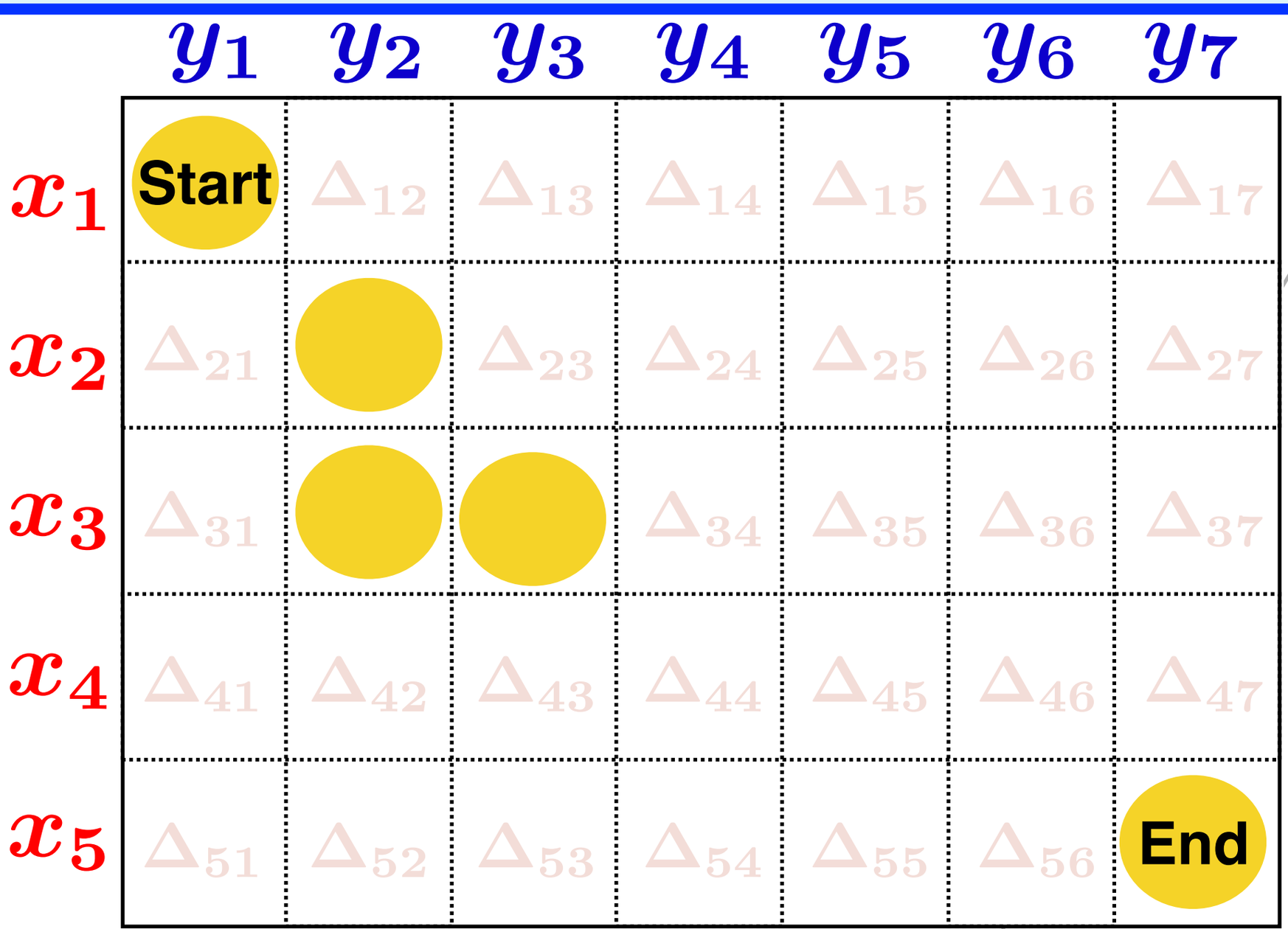


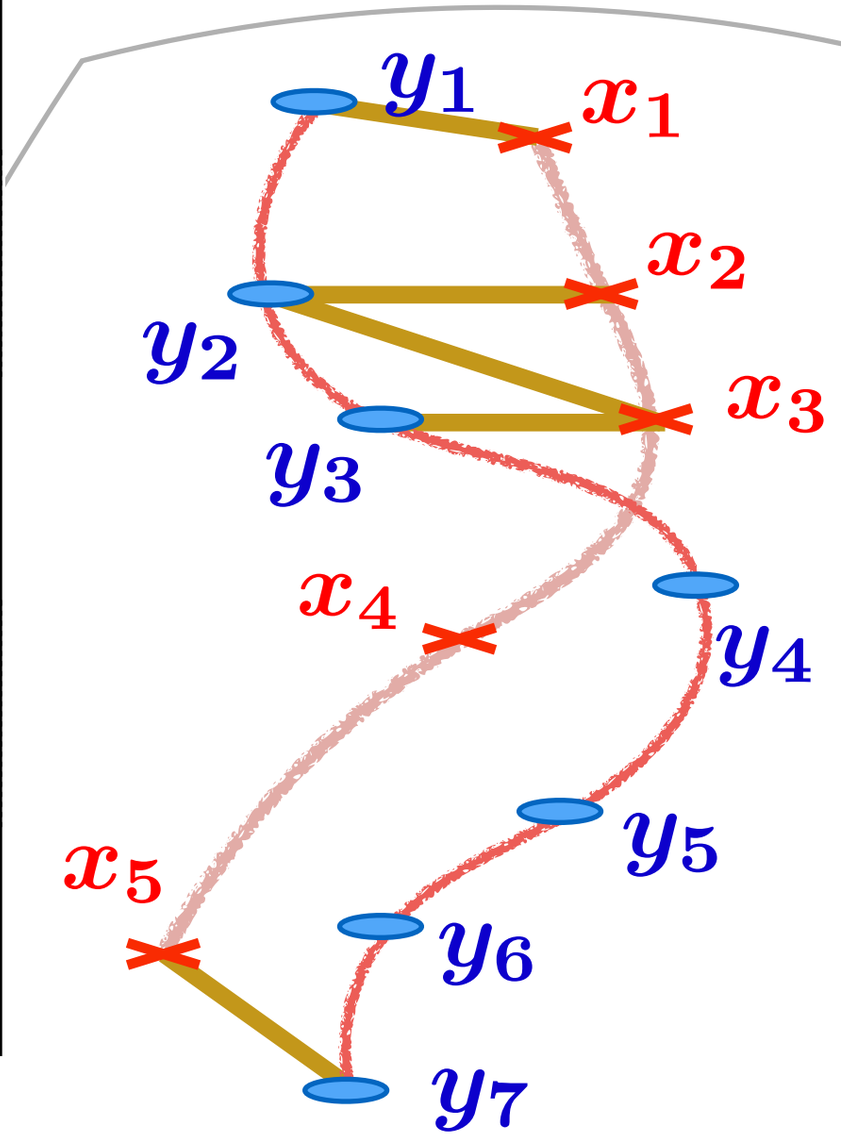
Alignment Path

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1	Start	Δ_{12}	Δ_{13}	Δ_{14}	Δ_{15}	Δ_{16}	Δ_{17}
x_2	Δ_{21}		Δ_{23}	Δ_{24}	Δ_{25}	Δ_{26}	Δ_{27}
x_3	Δ_{31}		Δ_{33}	Δ_{34}	Δ_{35}	Δ_{36}	Δ_{37}
x_4	Δ_{41}	Δ_{42}	Δ_{43}	Δ_{44}	Δ_{45}	Δ_{46}	Δ_{47}
x_5	Δ_{51}	Δ_{52}	Δ_{53}	Δ_{54}	Δ_{55}	Δ_{56}	End

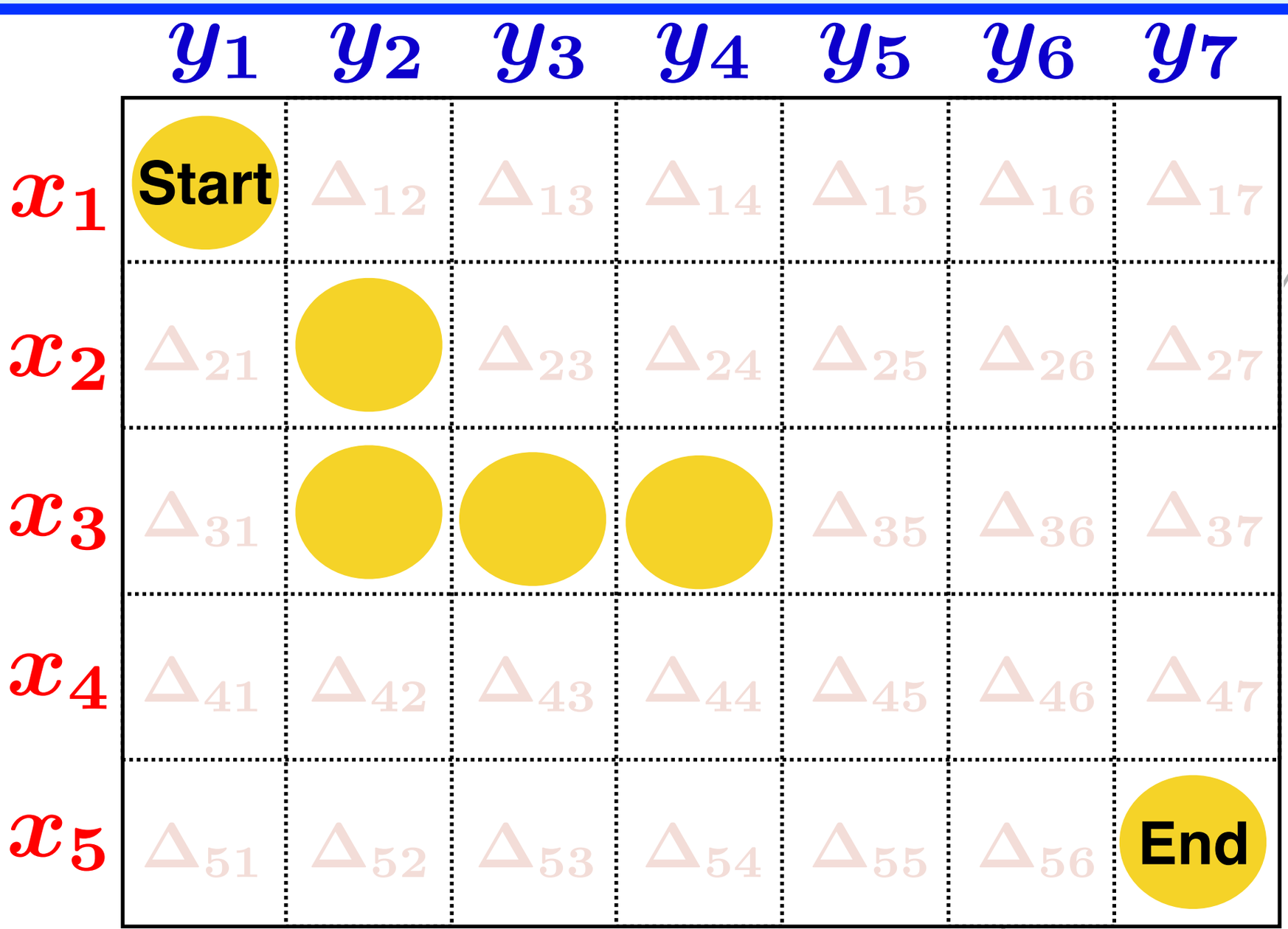


Alignment Path

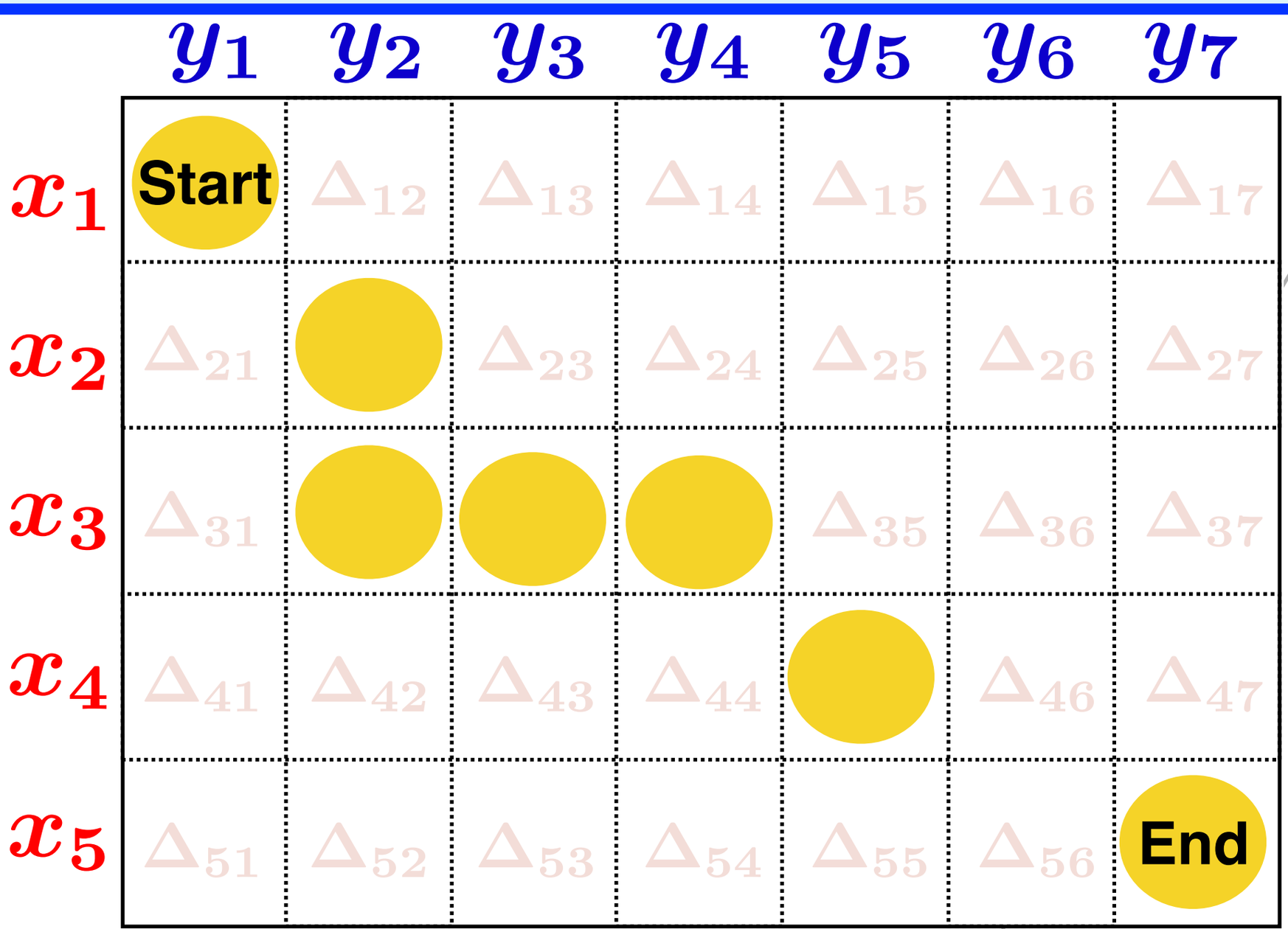


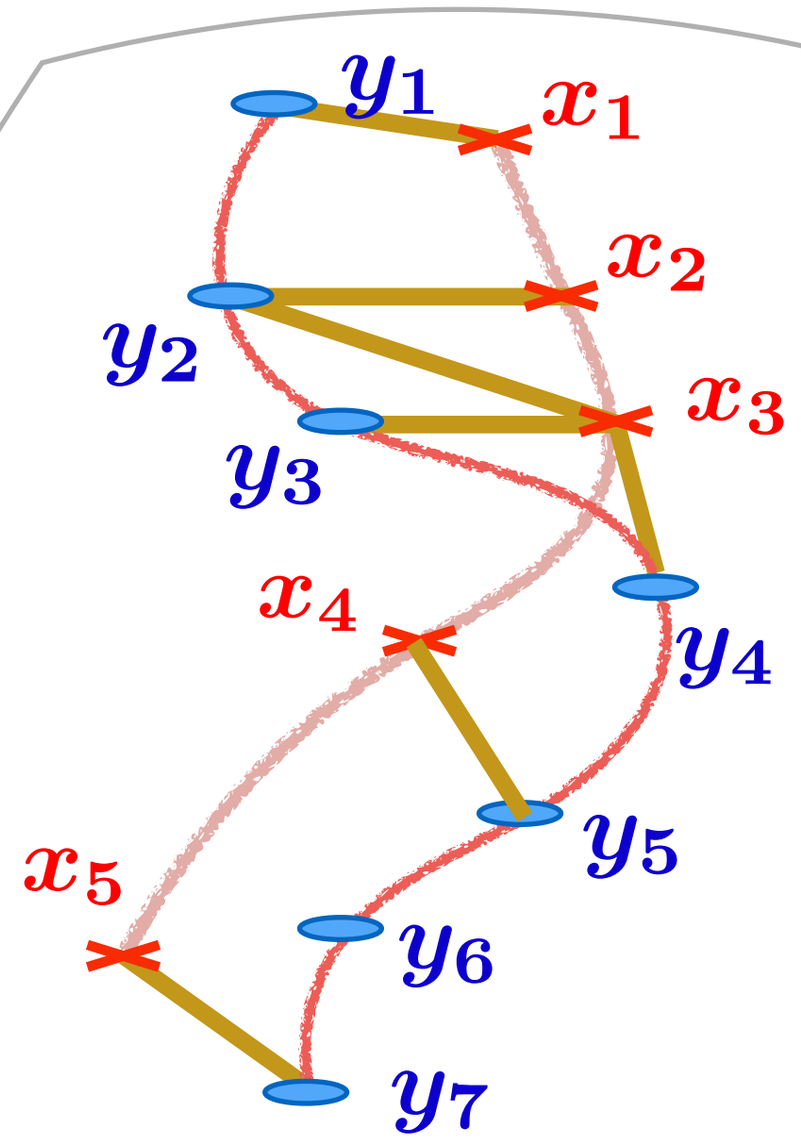


Alignment Path

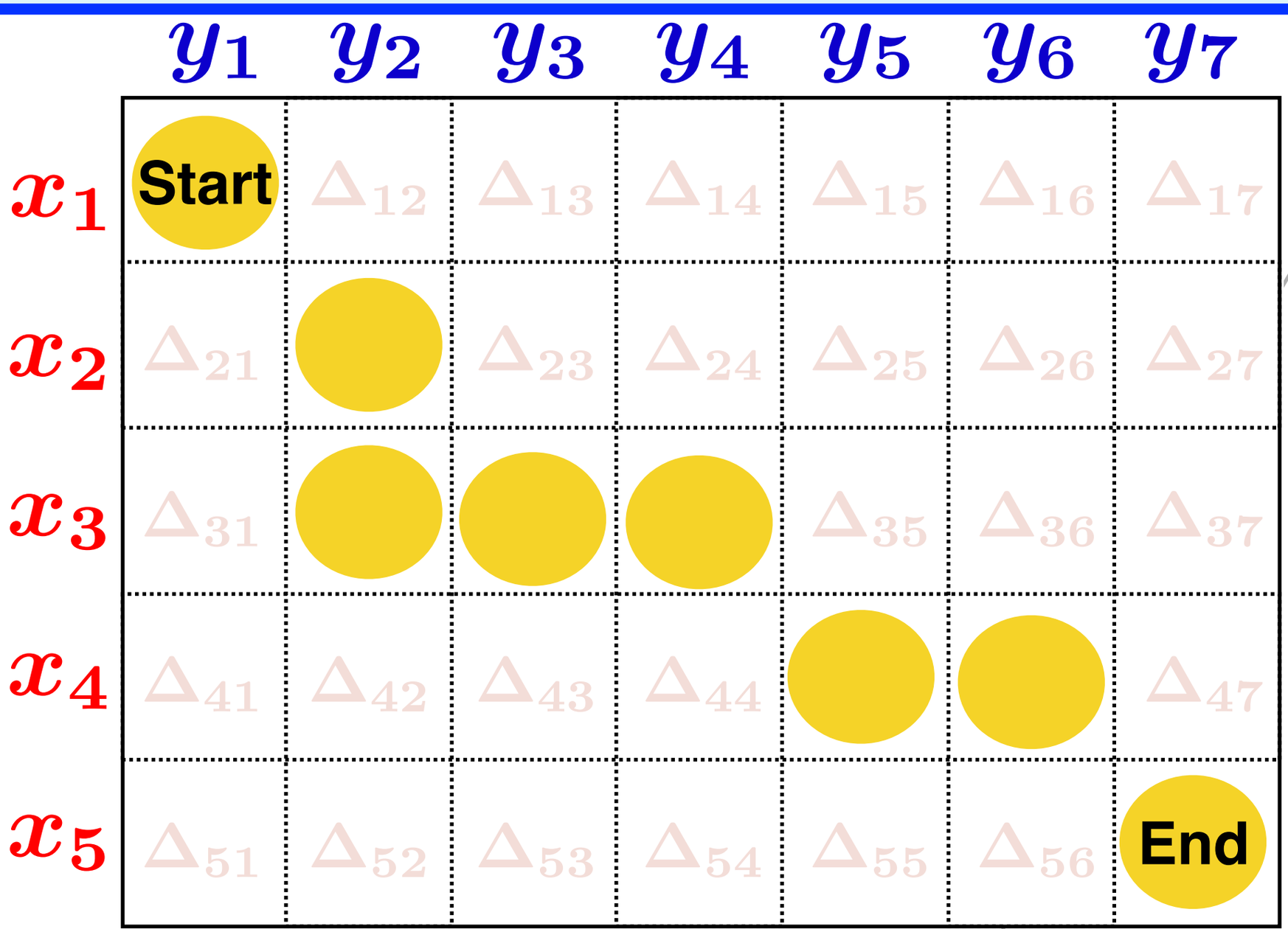


Alignment Path

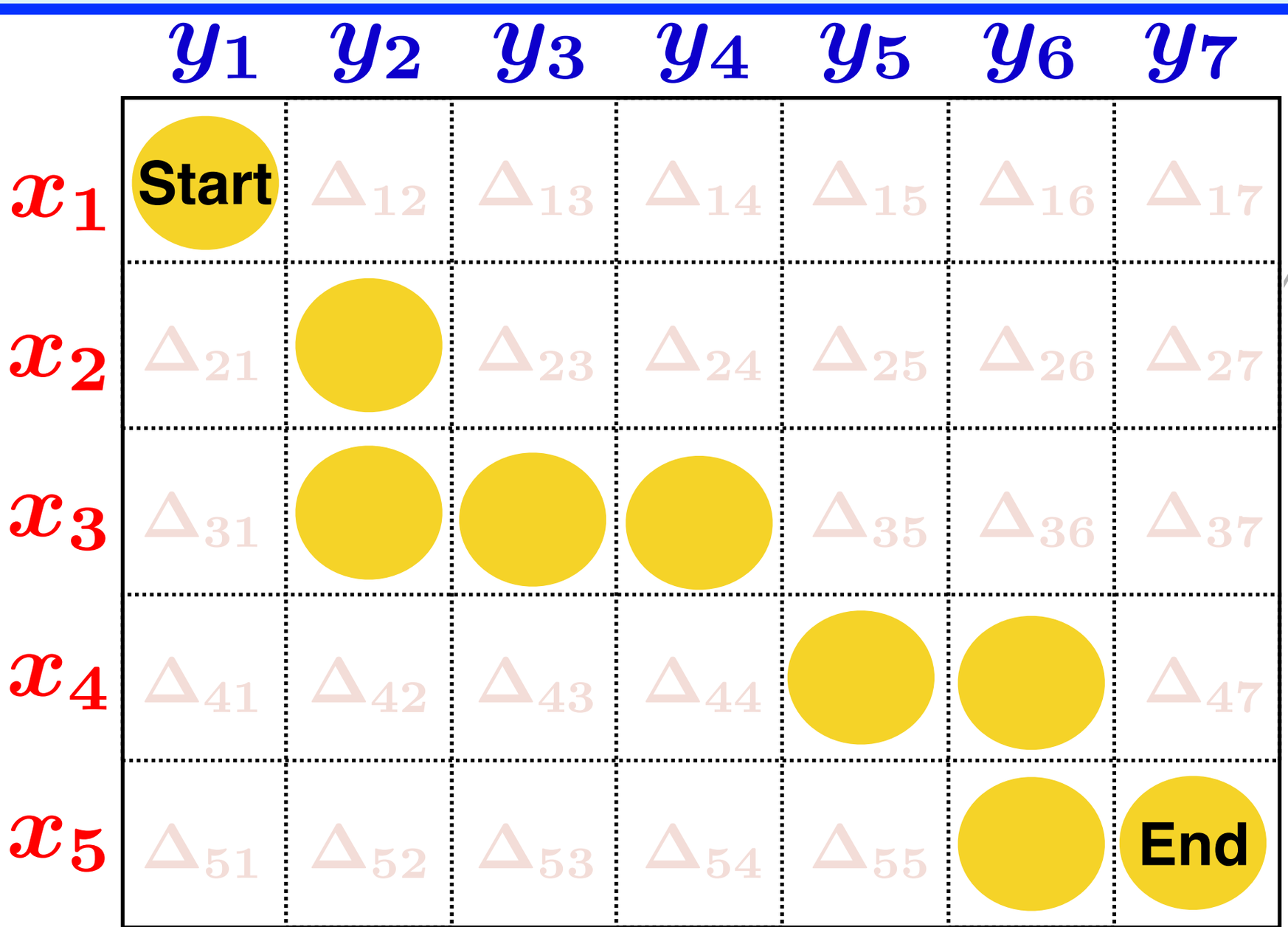




Alignment Path

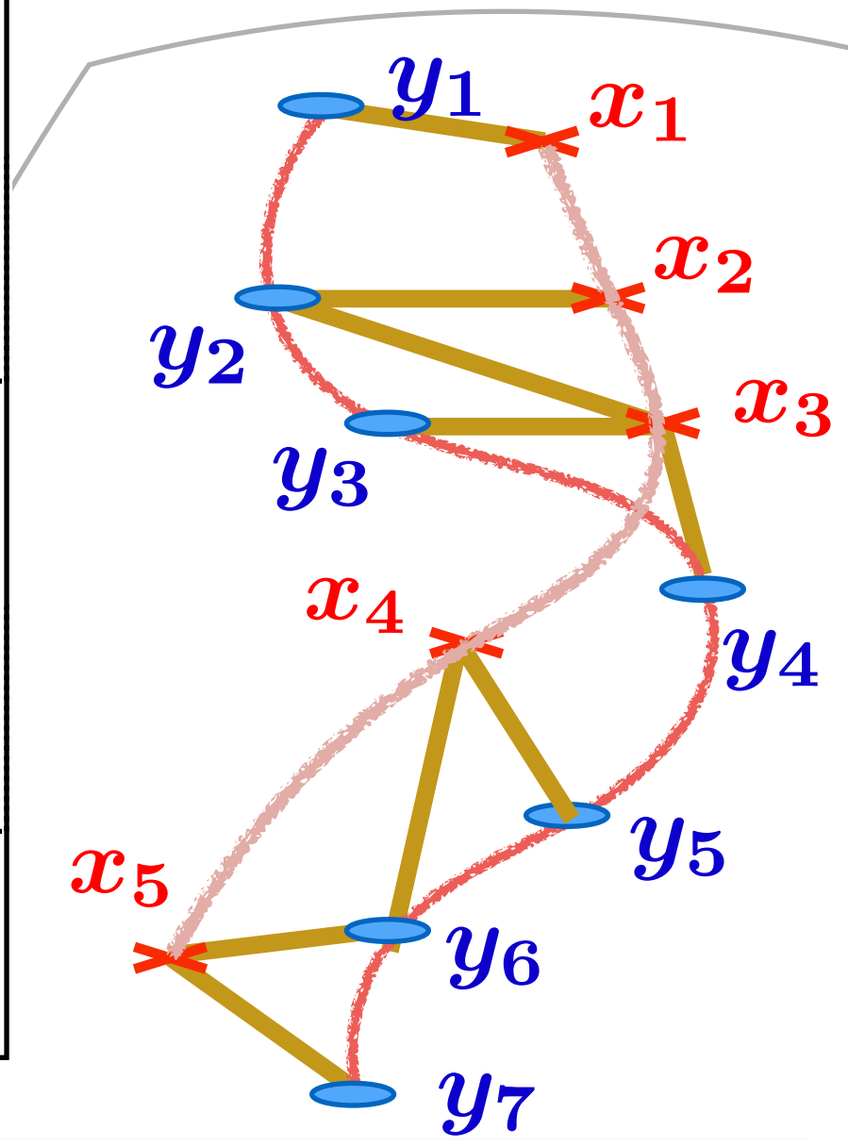


Alignment Path



Path Cost

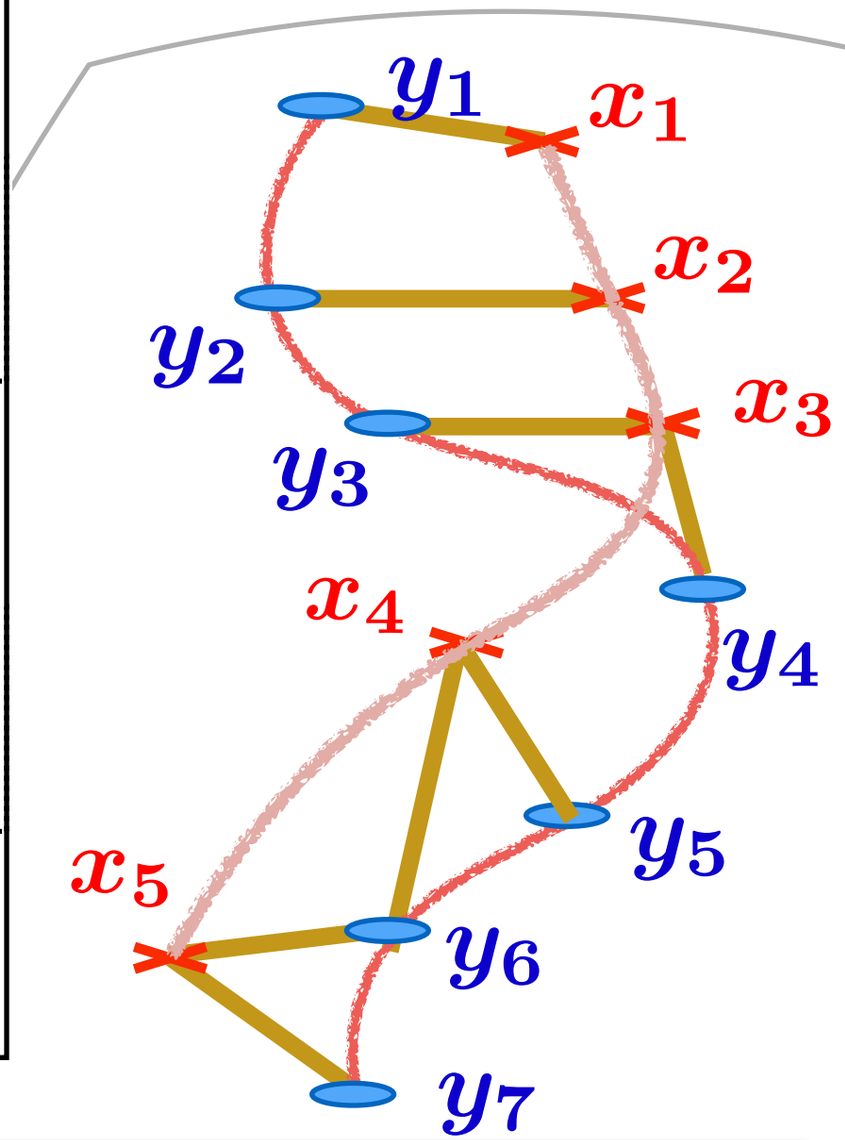
	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1	Δ_{11}	Δ_{12}	Δ_{13}	Δ_{14}	Δ_{15}	Δ_{16}	Δ_{17}
x_2	Δ_{21}	Δ_{22}	Δ_{23}	Δ_{24}	Δ_{25}	Δ_{26}	Δ_{27}
x_3	Δ_{31}	Δ_{32}	Δ_{33}	Δ_{34}	Δ_{35}	Δ_{36}	Δ_{37}
x_4	Δ_{41}	Δ_{42}	Δ_{43}	Δ_{44}	Δ_{45}	Δ_{46}	Δ_{47}
x_5	Δ_{51}	Δ_{52}	Δ_{53}	Δ_{54}	Δ_{55}	Δ_{56}	Δ_{57}



$$\text{Cost} = \Delta_{11} + \Delta_{22} + \Delta_{32} + \Delta_{33} + \Delta_{34} + \Delta_{46} + \Delta_{56} + \Delta_{57}$$

Path Cost

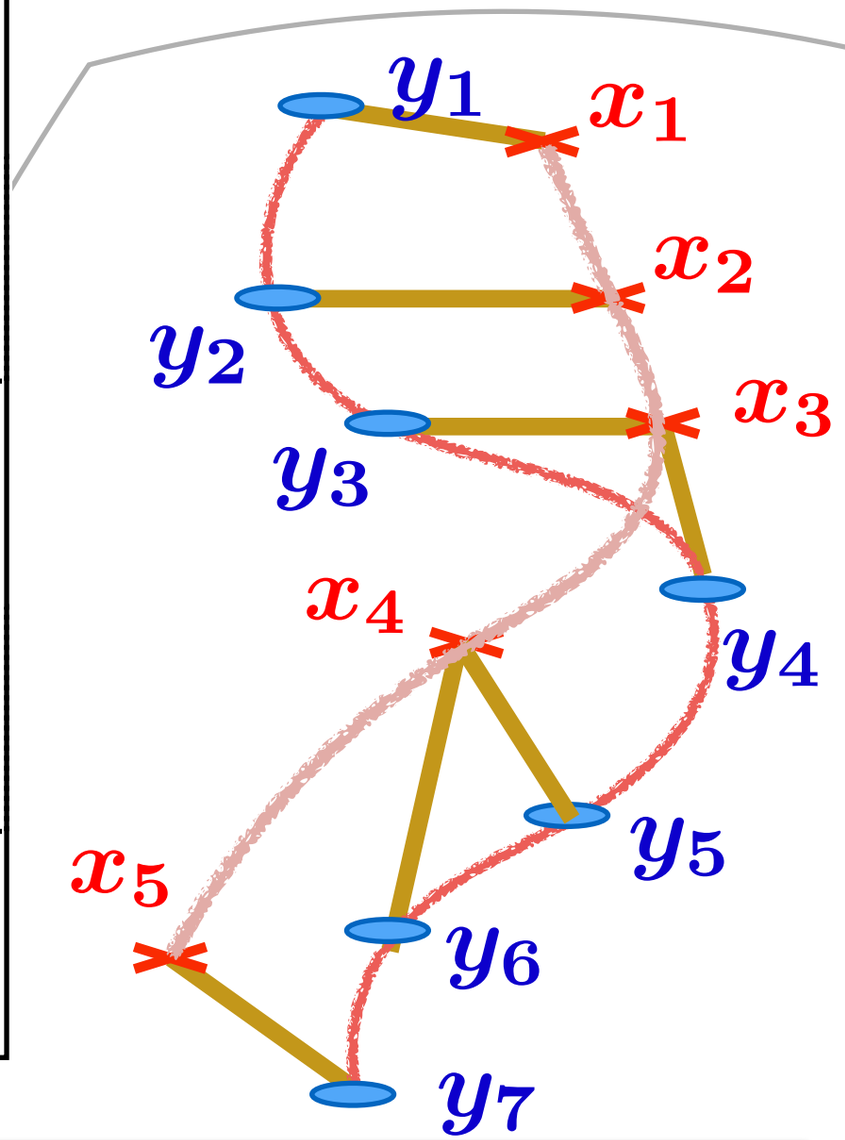
	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1	Δ_{11}	Δ_{12}	Δ_{13}	Δ_{14}	Δ_{15}	Δ_{16}	Δ_{17}
x_2	Δ_{21}	Δ_{22}	Δ_{23}	Δ_{24}	Δ_{25}	Δ_{26}	Δ_{27}
x_3	Δ_{31}	Δ_{32}	Δ_{33}	Δ_{34}	Δ_{35}	Δ_{36}	Δ_{37}
x_4	Δ_{41}	Δ_{42}	Δ_{43}	Δ_{44}	Δ_{45}	Δ_{46}	Δ_{47}
x_5	Δ_{51}	Δ_{52}	Δ_{53}	Δ_{54}	Δ_{55}	Δ_{56}	Δ_{57}



$$\text{Cost} = \Delta_{11} + \Delta_{22} \quad + \quad \Delta_{33} + \Delta_{34} + \Delta_{46} + \Delta_{56} + \Delta_{57}$$

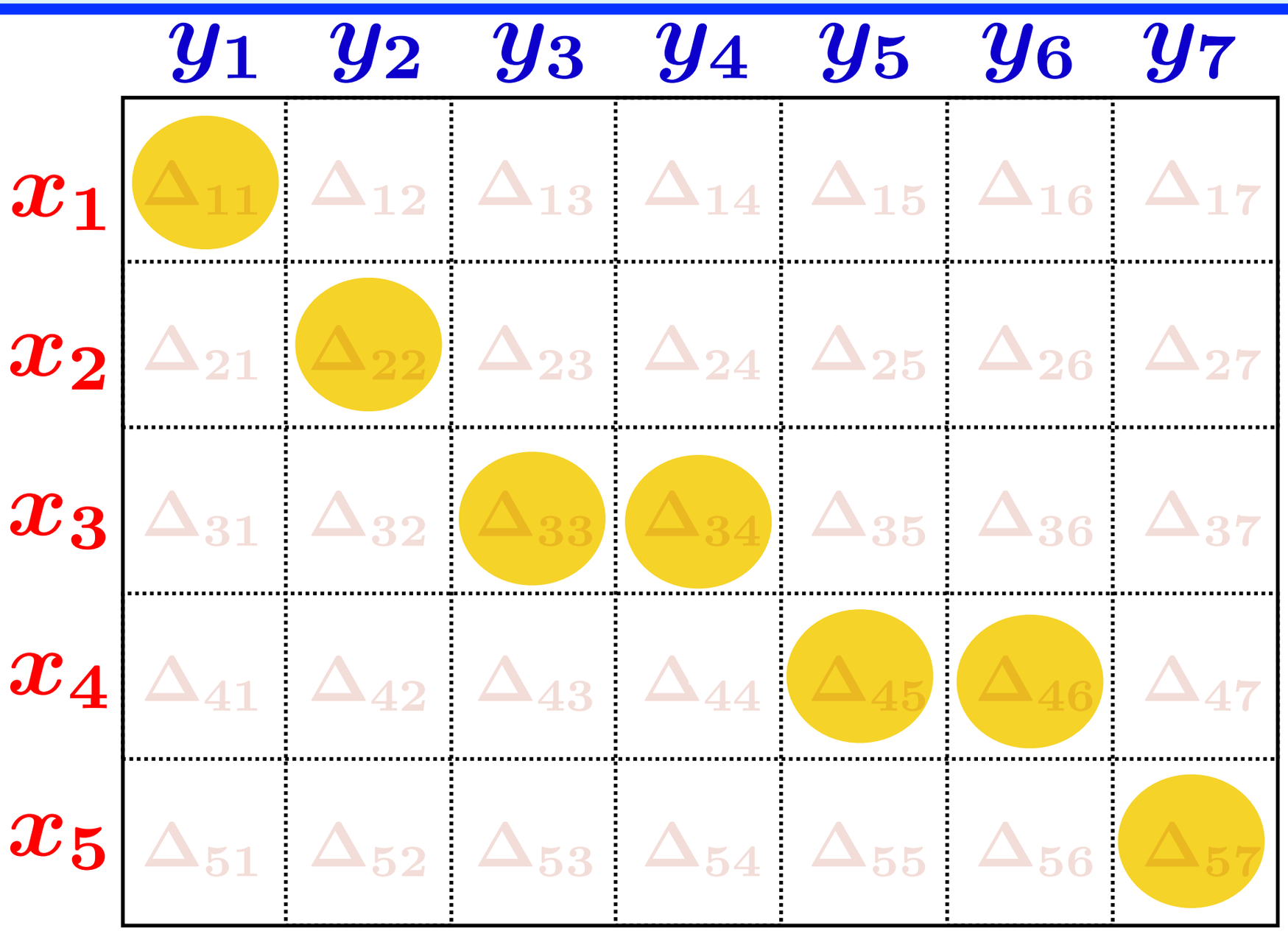
Path Cost

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1	Δ_{11}	Δ_{12}	Δ_{13}	Δ_{14}	Δ_{15}	Δ_{16}	Δ_{17}
x_2	Δ_{21}	Δ_{22}	Δ_{23}	Δ_{24}	Δ_{25}	Δ_{26}	Δ_{27}
x_3	Δ_{31}	Δ_{32}	Δ_{33}	Δ_{34}	Δ_{35}	Δ_{36}	Δ_{37}
x_4	Δ_{41}	Δ_{42}	Δ_{43}	Δ_{44}	Δ_{45}	Δ_{46}	Δ_{47}
x_5	Δ_{51}	Δ_{52}	Δ_{53}	Δ_{54}	Δ_{55}	Δ_{56}	Δ_{57}



$$\text{Cost} = \Delta_{11} + \Delta_{22} \quad + \quad \Delta_{33} + \Delta_{34} + \Delta_{46} \quad + \quad \Delta_{57}$$

Path Cost


y_1 y_2 y_3 y_4 y_5 y_6 y_7
x_1
x_2
x_3
x_4
x_5
Δ_{11} Δ_{12} Δ_{13} Δ_{14} Δ_{15} Δ_{16} Δ_{17}
Δ_{21} Δ_{22} Δ_{23} Δ_{24} Δ_{25} Δ_{26} Δ_{27}
Δ_{31} Δ_{32} Δ_{33} Δ_{34} Δ_{35} Δ_{36} Δ_{37}
Δ_{41} Δ_{42} Δ_{43} Δ_{44} Δ_{45} Δ_{46} Δ_{47}
Δ_{51} Δ_{52} Δ_{53} Δ_{54} Δ_{55} Δ_{56} Δ_{57}

Path Cost

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1	**1**	**0**	**0**	**0**	**0**	**0**	**0**
x_2	**0**	**1**	**0**	**0**	**0**	**0**	**0**
x_3	**0**	**0**	**1**	**1**	**0**	**0**	**0**
x_4	**0**	**0**	**0**	**0**	**1**	**1**	**0**
x_5	**0**	**0**	**0**	**0**	**0**	**0**	**1**

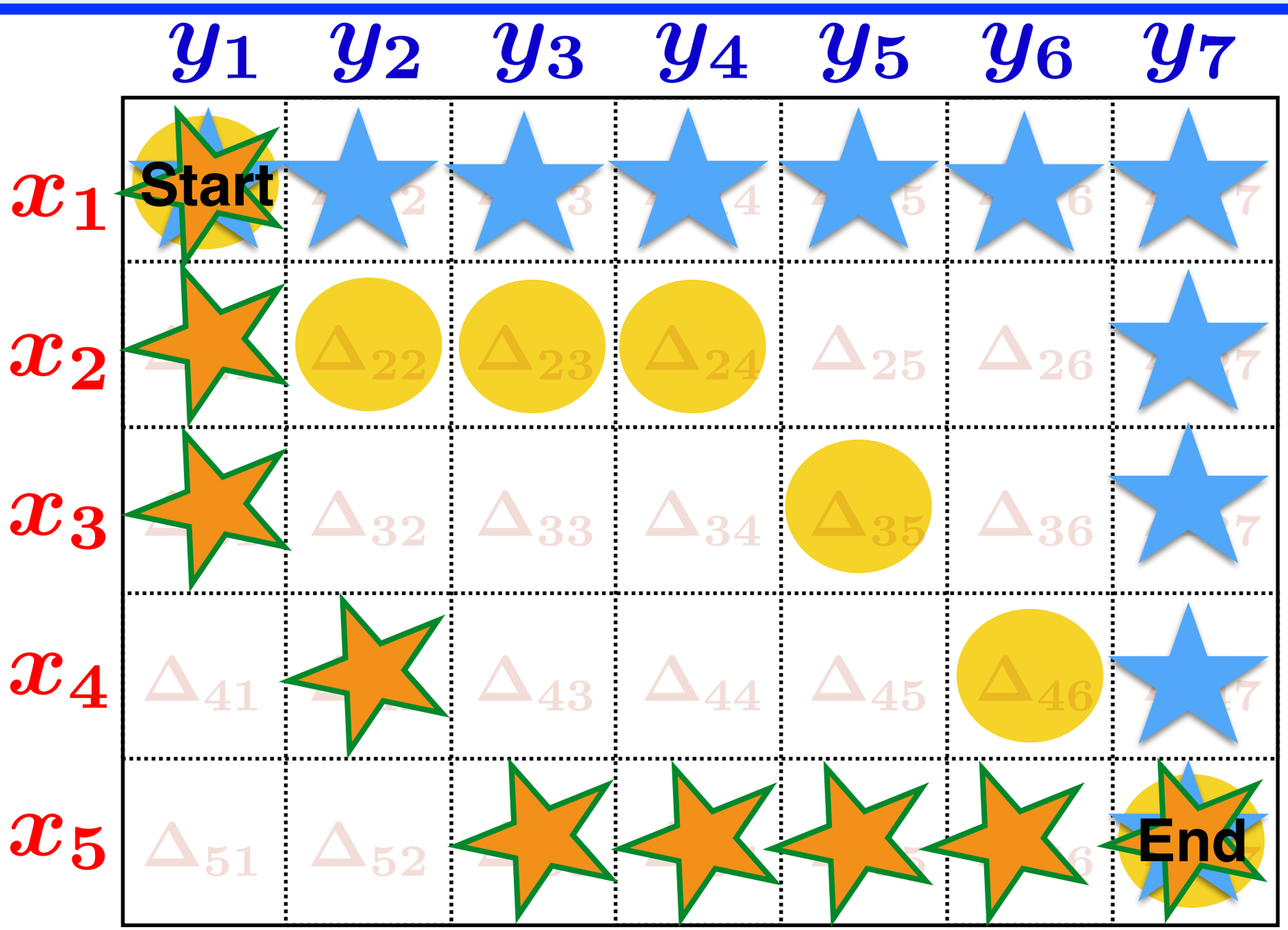
$= A$

$$\text{Cost} = \langle A, \Delta \rangle, A \in \{0,1\}^{n \times m}$$

Min Cost Alignment Matrix?

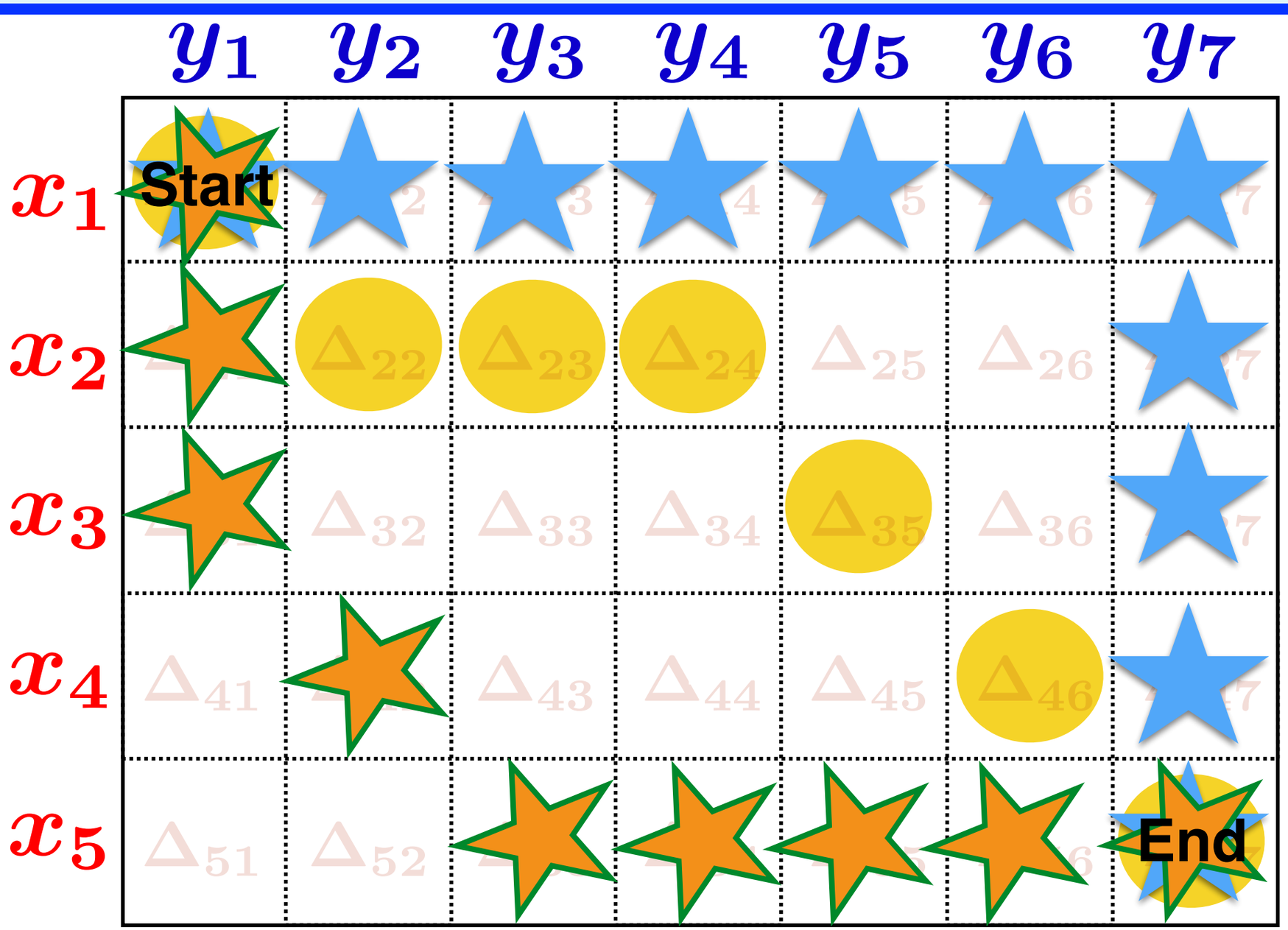
	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1	Start	Δ_{12}	Δ_{13}	Δ_{14}	Δ_{15}	Δ_{16}	Δ_{17}
x_2	Δ_{21}	Δ_{22}	Δ_{23}	Δ_{24}	Δ_{25}	Δ_{26}	Δ_{27}
x_3	Δ_{31}	Δ_{32}	Δ_{33}	Δ_{34}	Δ_{35}	Δ_{36}	Δ_{37}
x_4	Δ_{41}	Δ_{42}	Δ_{43}	Δ_{44}	Δ_{45}	Δ_{46}	Δ_{47}
x_5	Δ_{51}	Δ_{52}	Δ_{53}	Δ_{54}	Δ_{55}	Δ_{56}	End

Min Cost Alignment Matrix?



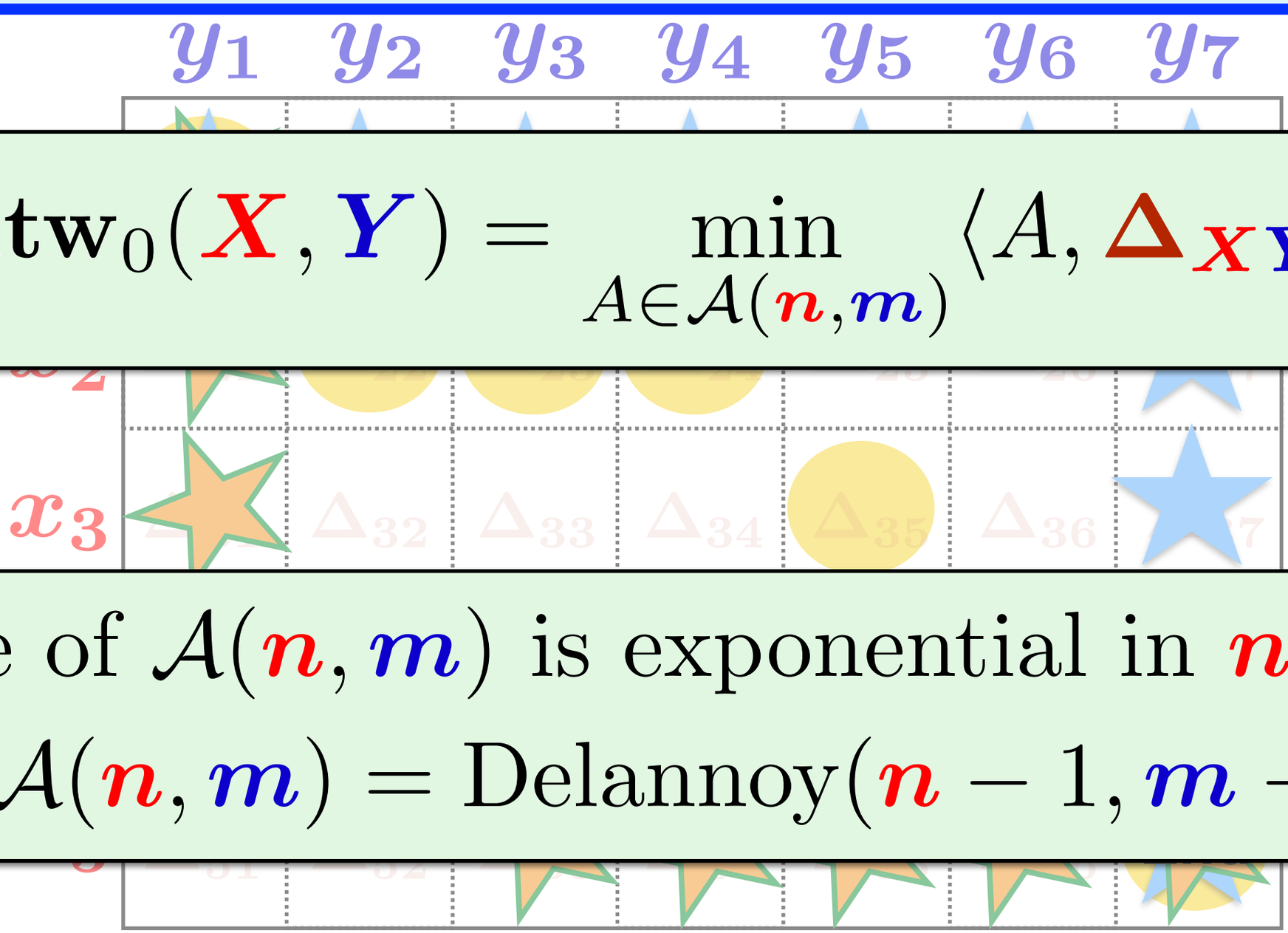
Set of all valid path matrices: $\mathcal{A}(\boldsymbol{n}, \boldsymbol{m}) \subset \{0,1\}^{\boldsymbol{n} \times \boldsymbol{m}}$

Best Alignment Matrix



Set of all valid path matrices: $\mathcal{A}(\boldsymbol{n}, \boldsymbol{m}) \subset \{0,1\}^{\boldsymbol{n} \times \boldsymbol{m}}$

Best Alignment Matrix



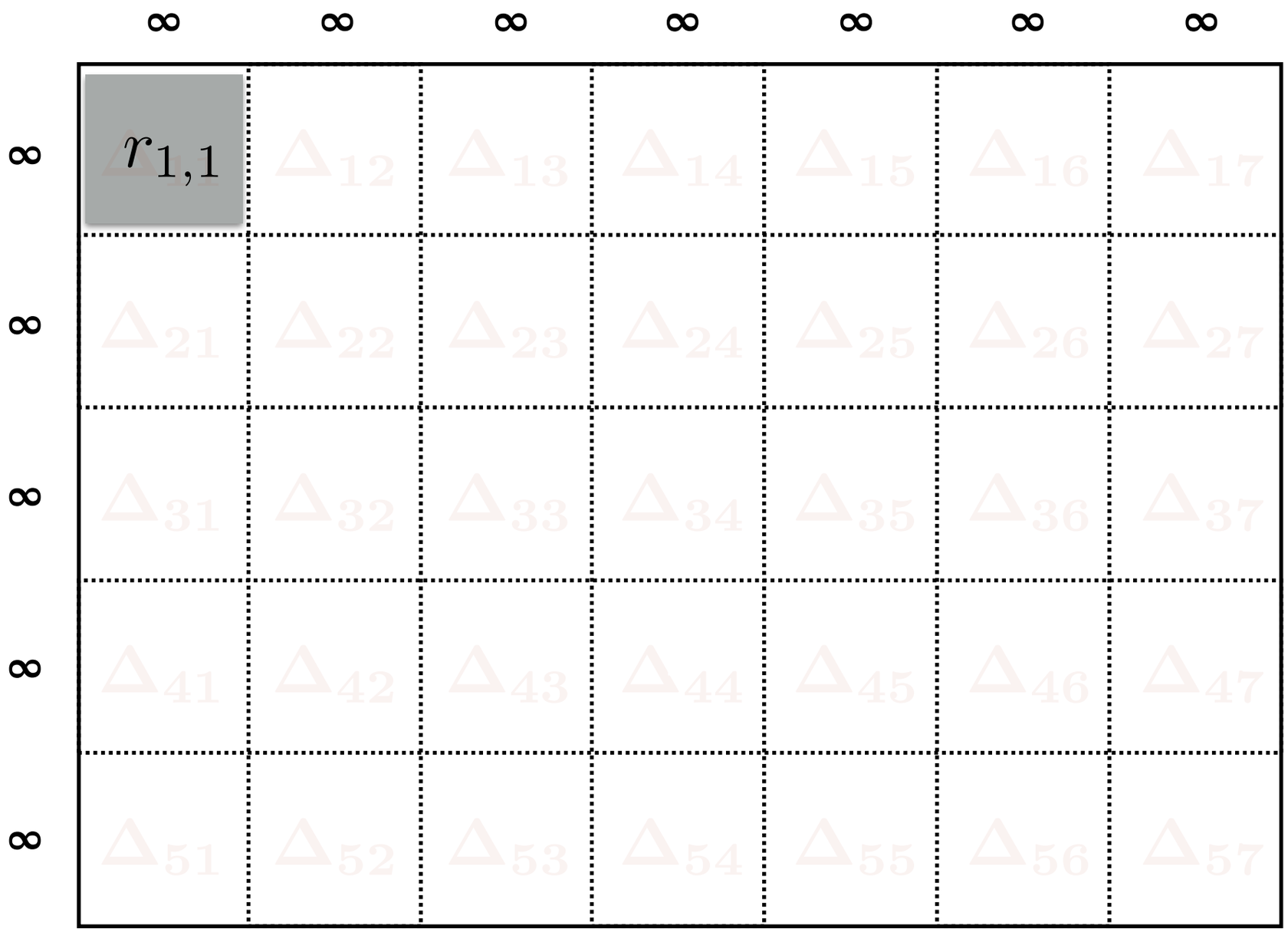
$$\mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \min_{A \in \mathcal{A}(\boldsymbol{n}, \boldsymbol{m})} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

Size of $\mathcal{A}(\boldsymbol{n}, \boldsymbol{m})$ is exponential in $\boldsymbol{n}, \boldsymbol{m}$.

$$\#\mathcal{A}(\boldsymbol{n}, \boldsymbol{m}) = \text{Delannoy}(\boldsymbol{n} - 1, \boldsymbol{m} - 1)$$

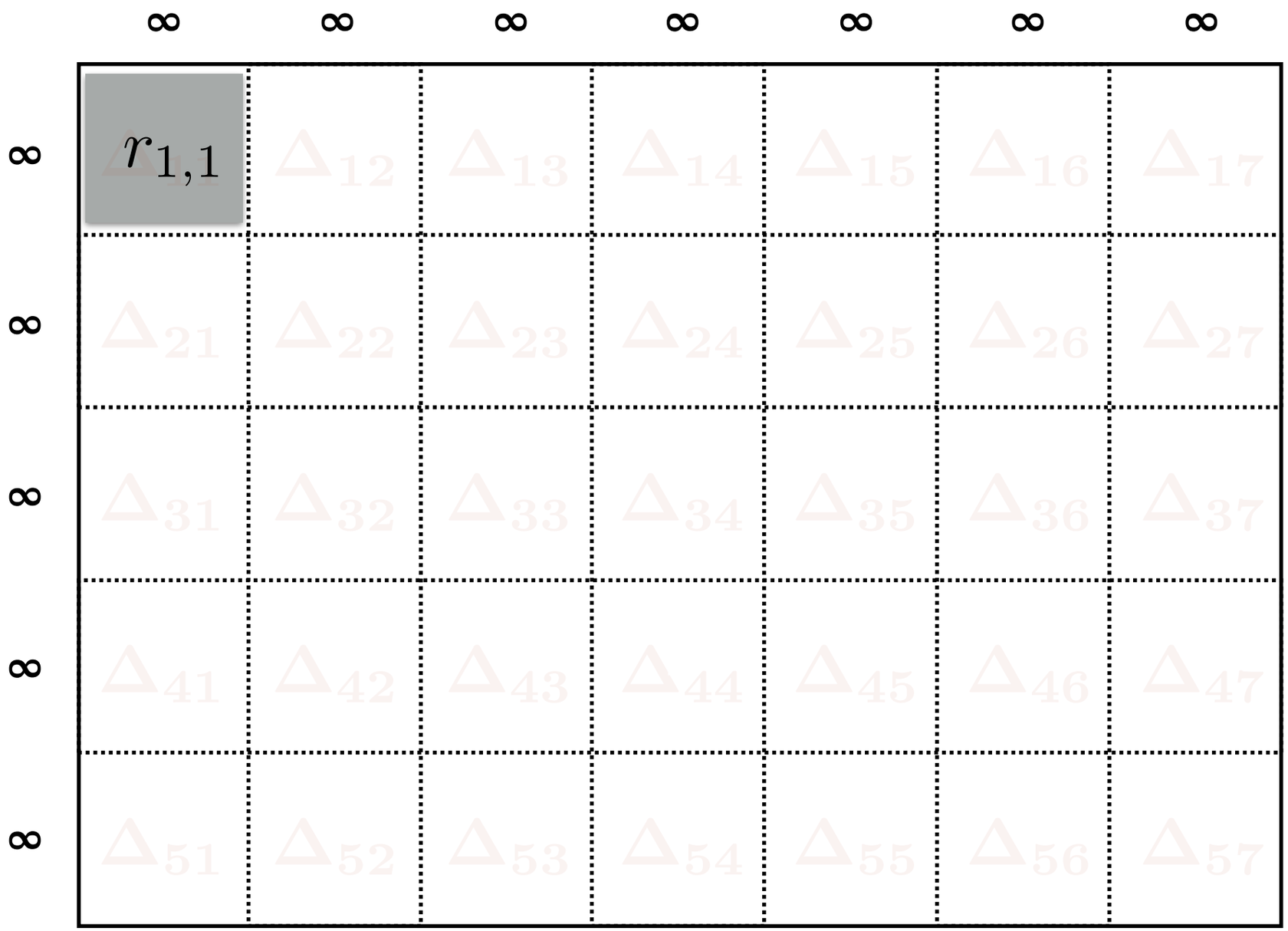
Set of all valid path matrices: $\mathcal{A}(\boldsymbol{n}, \boldsymbol{m}) \subset \{0, 1\}^{\boldsymbol{n} \times \boldsymbol{m}}$

Best Path: Bellman Recursion



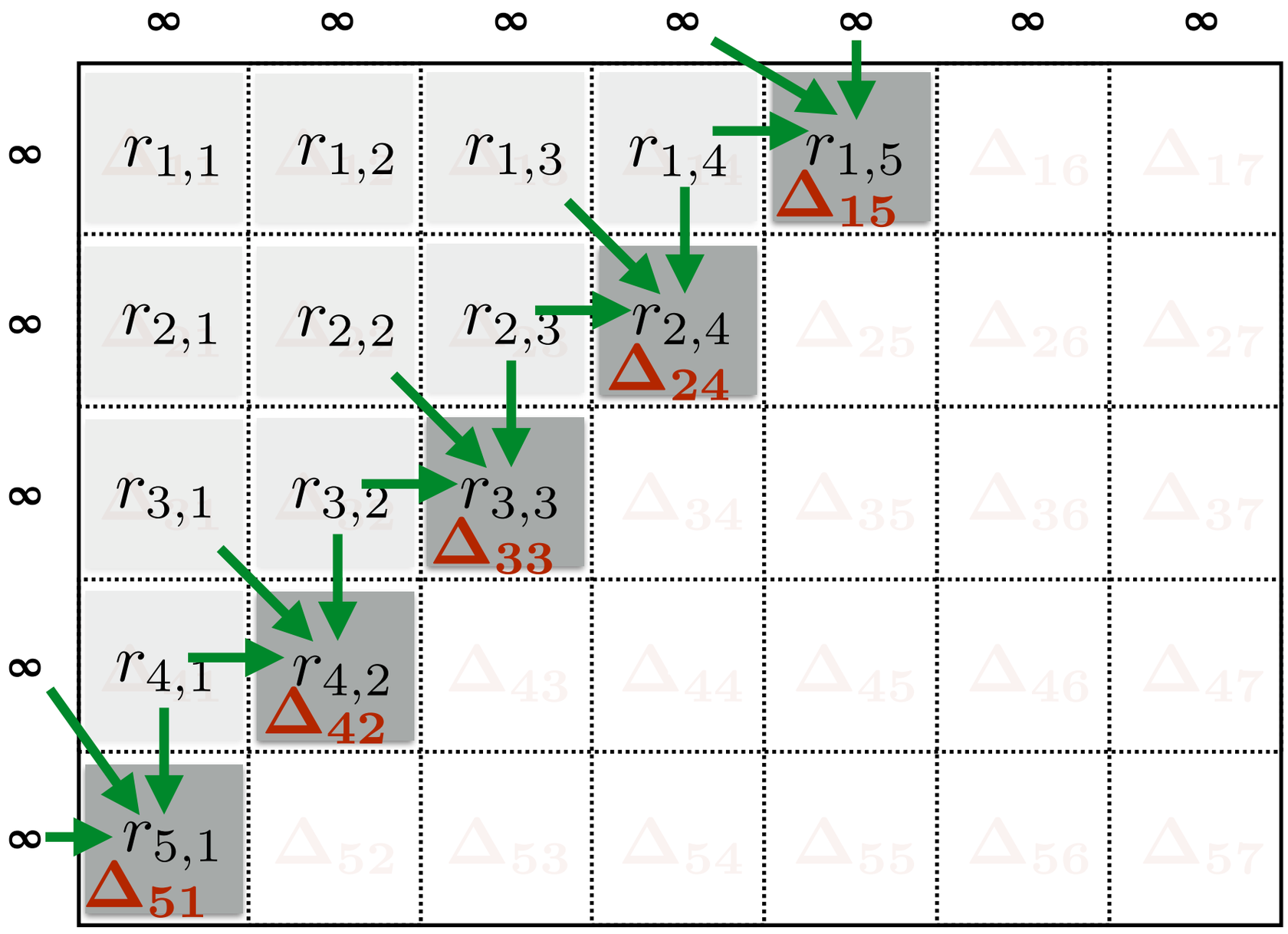
$$r_{1,1} = \Delta_{11} \qquad r_{0,j} = r_{i,0} = 0$$

Best Path: Bellman Recursion



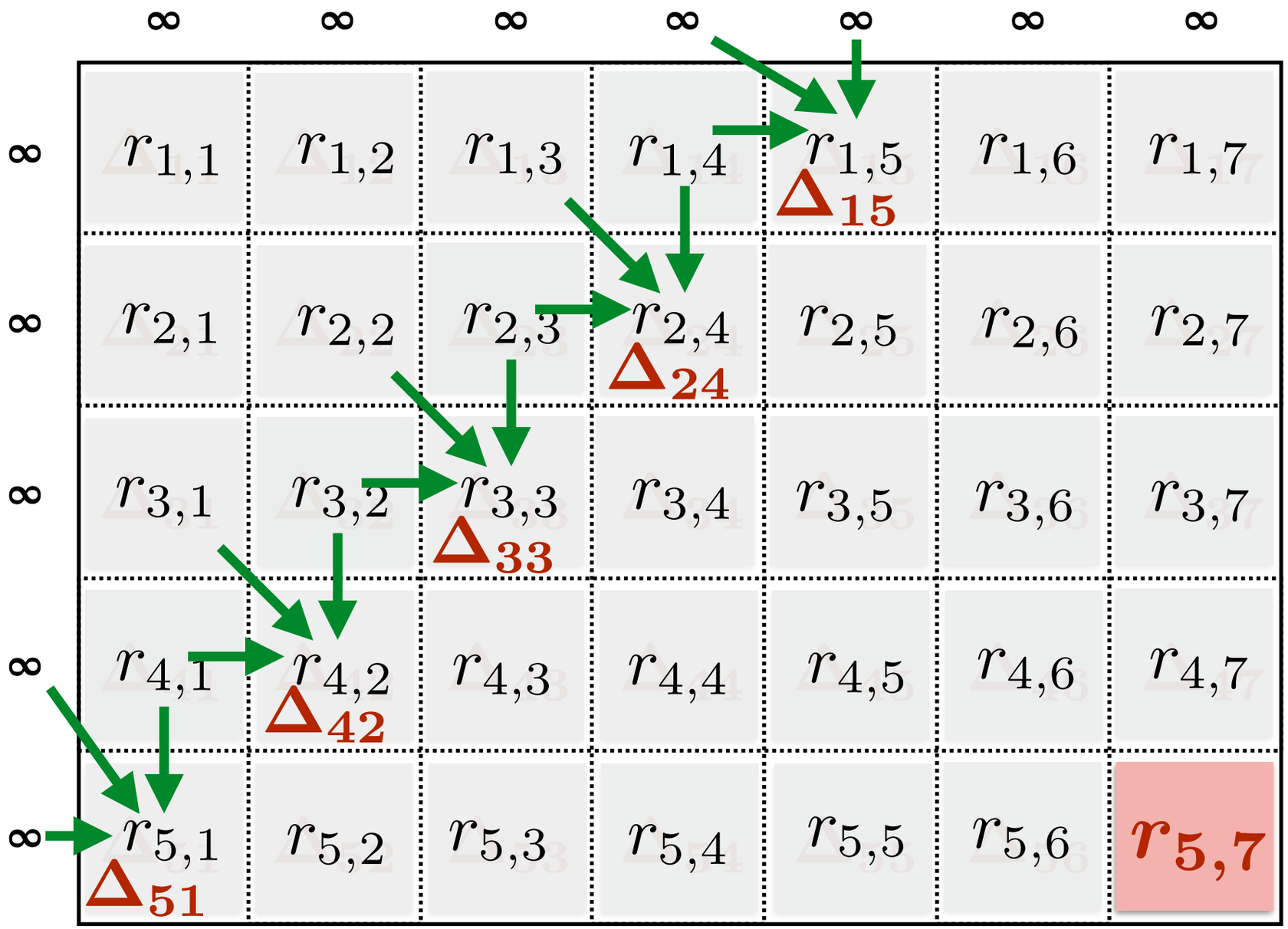
$$r_{i,j} = \min\left(r_{i-1,j-1}\,,\, r_{i-1,j}\,,\, r_{i,j-1}\right) + \boldsymbol{\Delta_{i,j}}$$

Best Path: Bellman Recursion



$$r_{i,j} = \min(r_{i-1,j-1}\,,\, r_{i-1,j}\,,\, r_{i,j-1}) + \boldsymbol{\Delta_{i,j}}$$

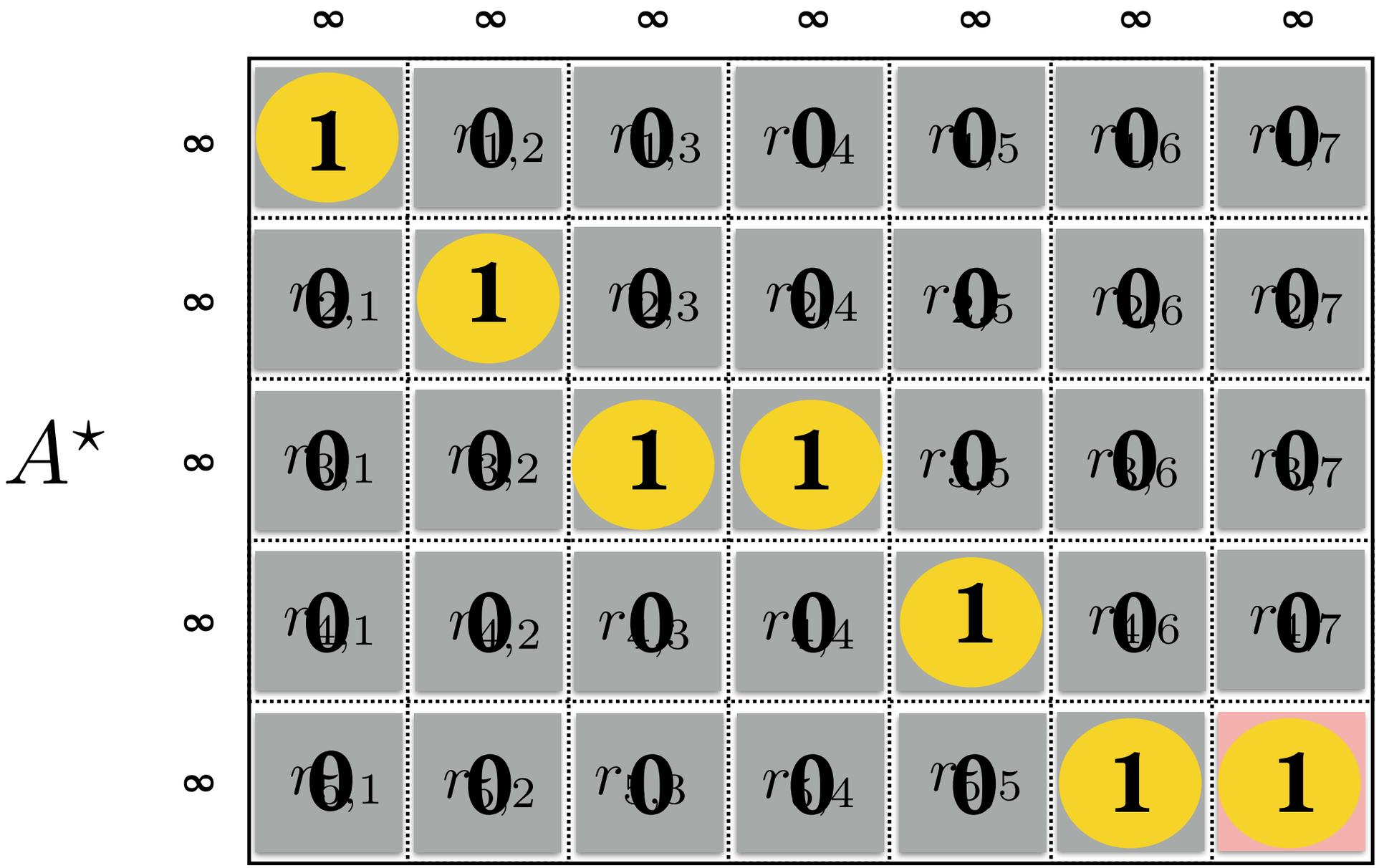
Best Path: Bellman Recursion



$$\mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{r_{n,m}}$$

Optimal Path

$A^\star$



$$\mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{r}_{\boldsymbol{n},\boldsymbol{m}} = \langle A^\star, \boldsymbol{\Delta} \rangle$$

Optimal Path

$A^\star$

	∞	∞	∞	∞	∞	∞	∞
∞	**1**	**0**	**0**	**0**	**0**	**0**	**0**
∞	**0**	**1**	**0**	**0**	**0**	**0**	**0**
∞	**0**	**0**	**1**	**1**	**0**	**0**	**0**
∞	**0**	**0**	**0**	**0**	**1**	**0**	**0**
∞	**0**	**0**	**0**	**0**	**0**	**1**	**1**

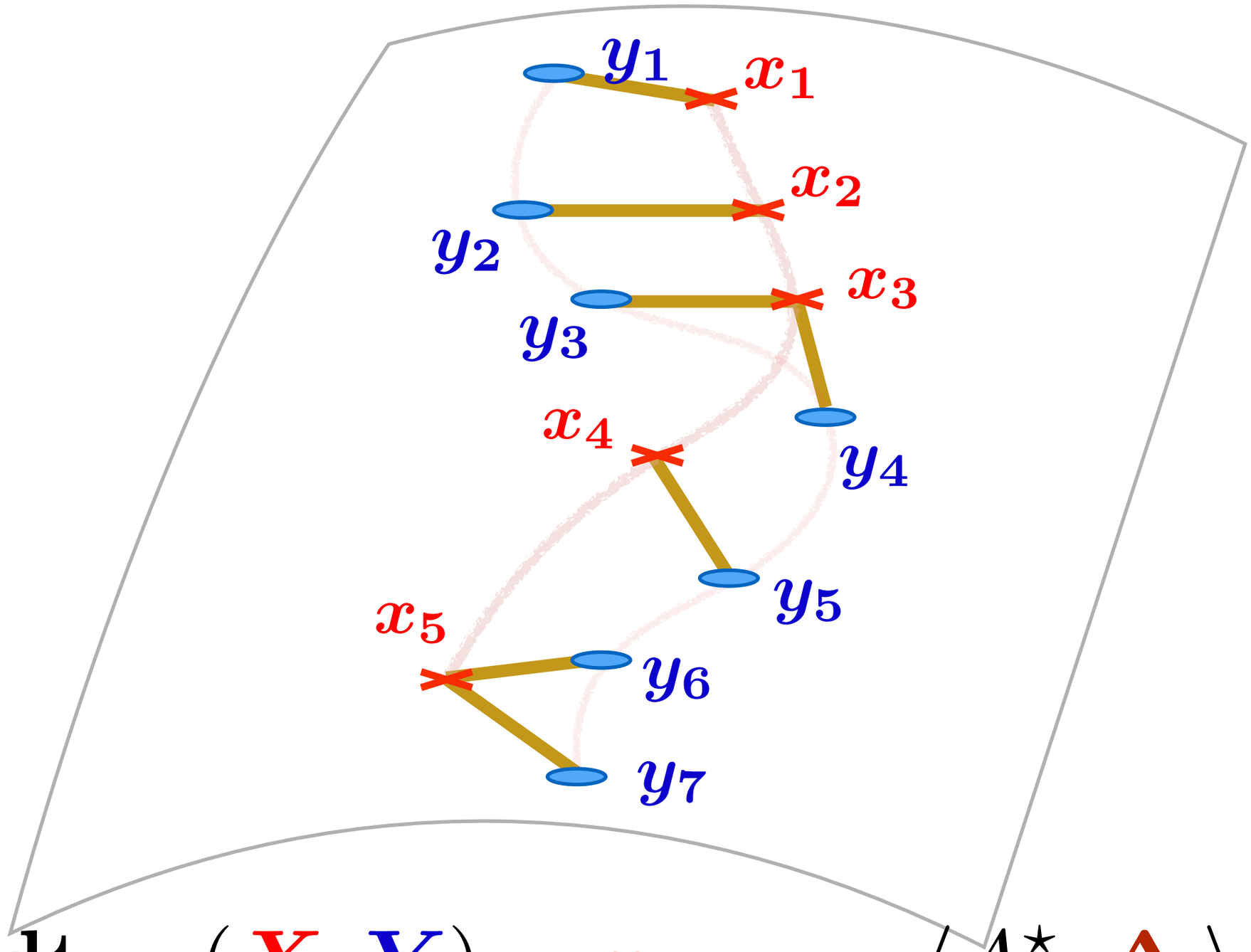
$$\mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{r}_{\boldsymbol{n},\boldsymbol{m}} = \langle A^\star, \boldsymbol{\Delta} \rangle$$

0. The DTW Geometry

1. Soft-DTW

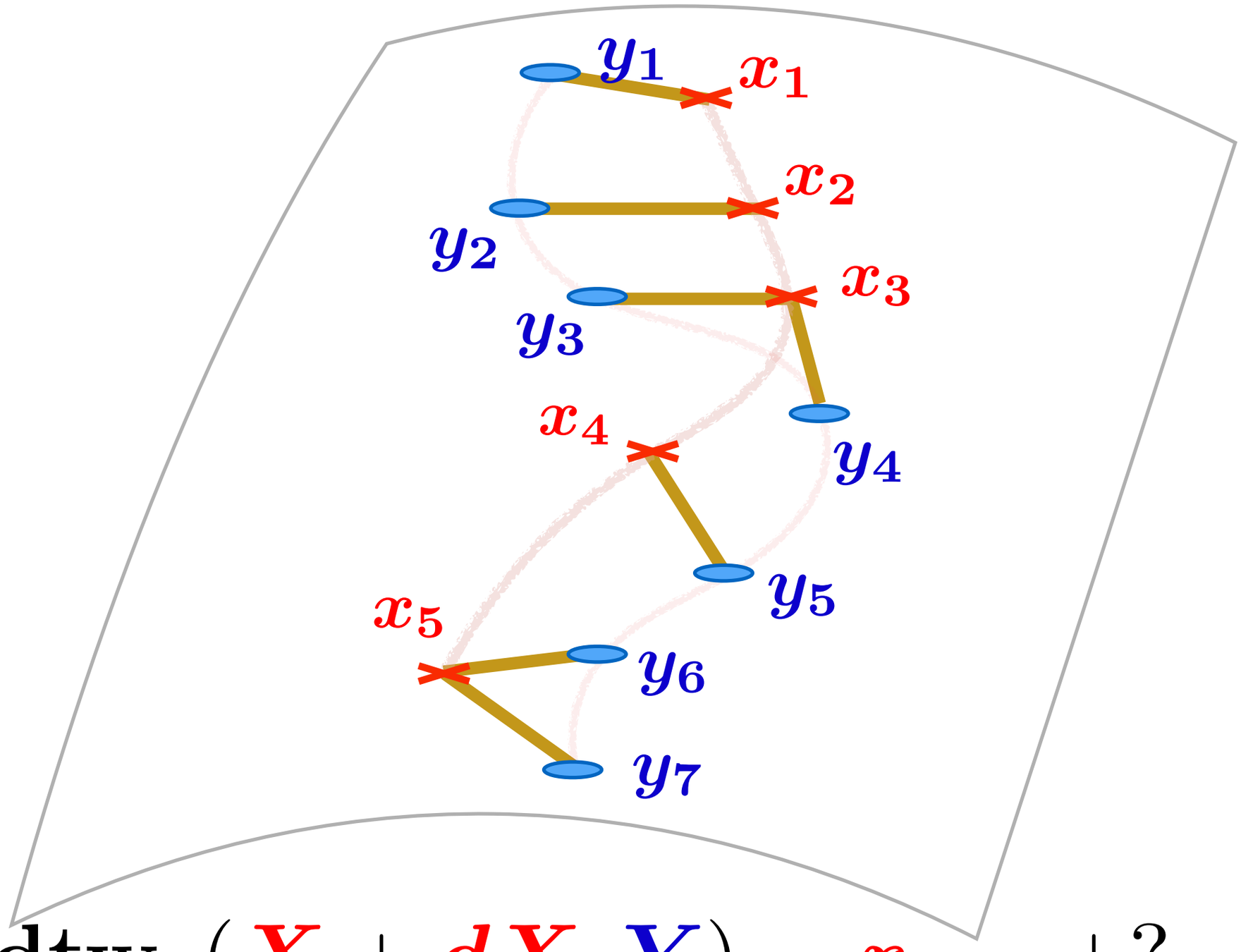
2. Soft-DTW as a Loss Function

DTW as a Loss: Differentiability?



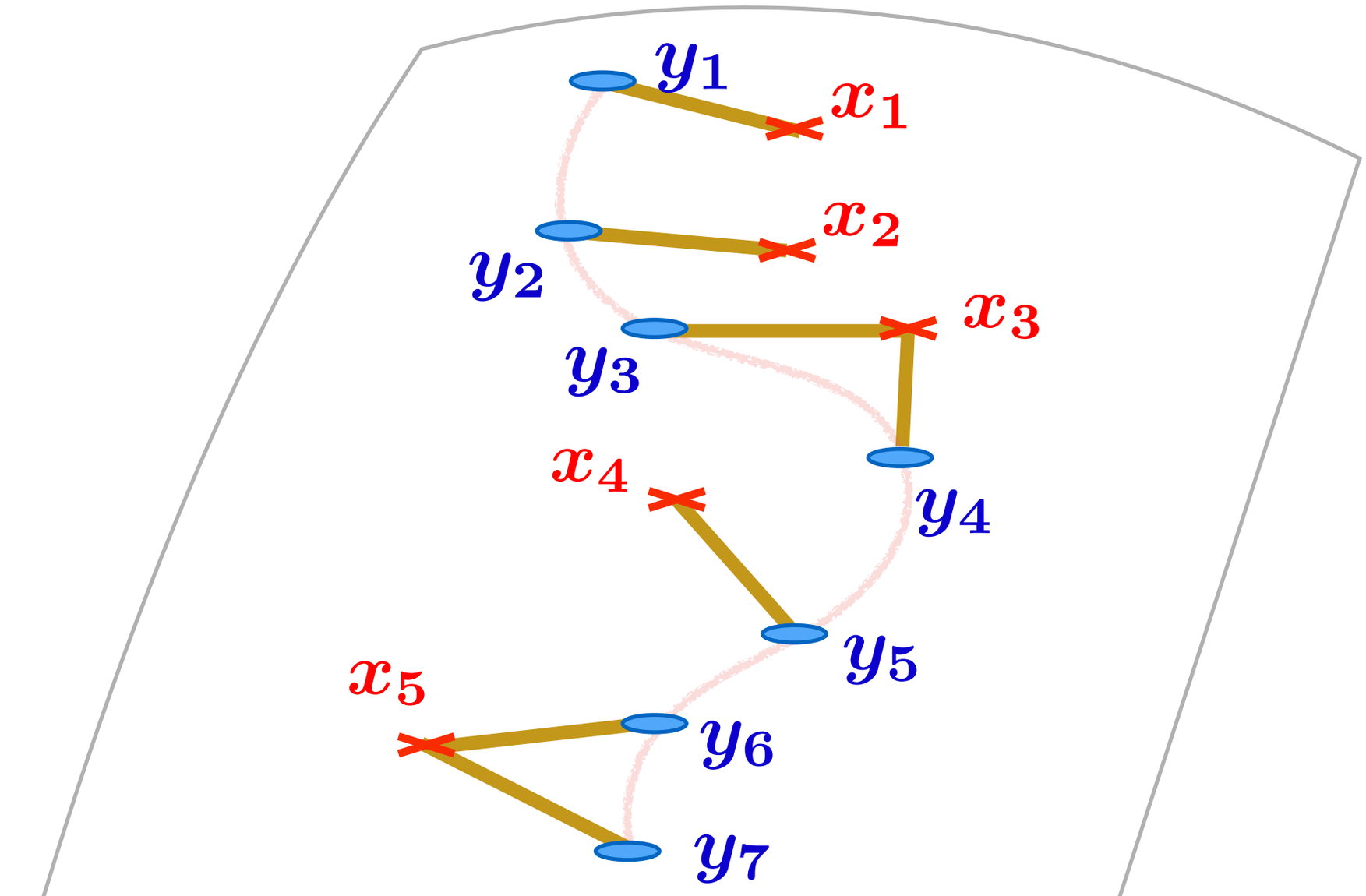
$$\mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{r}_{n,m} = \langle A^\star, \boldsymbol{\Delta} \rangle$$

DTW as a Loss: Differentiability?



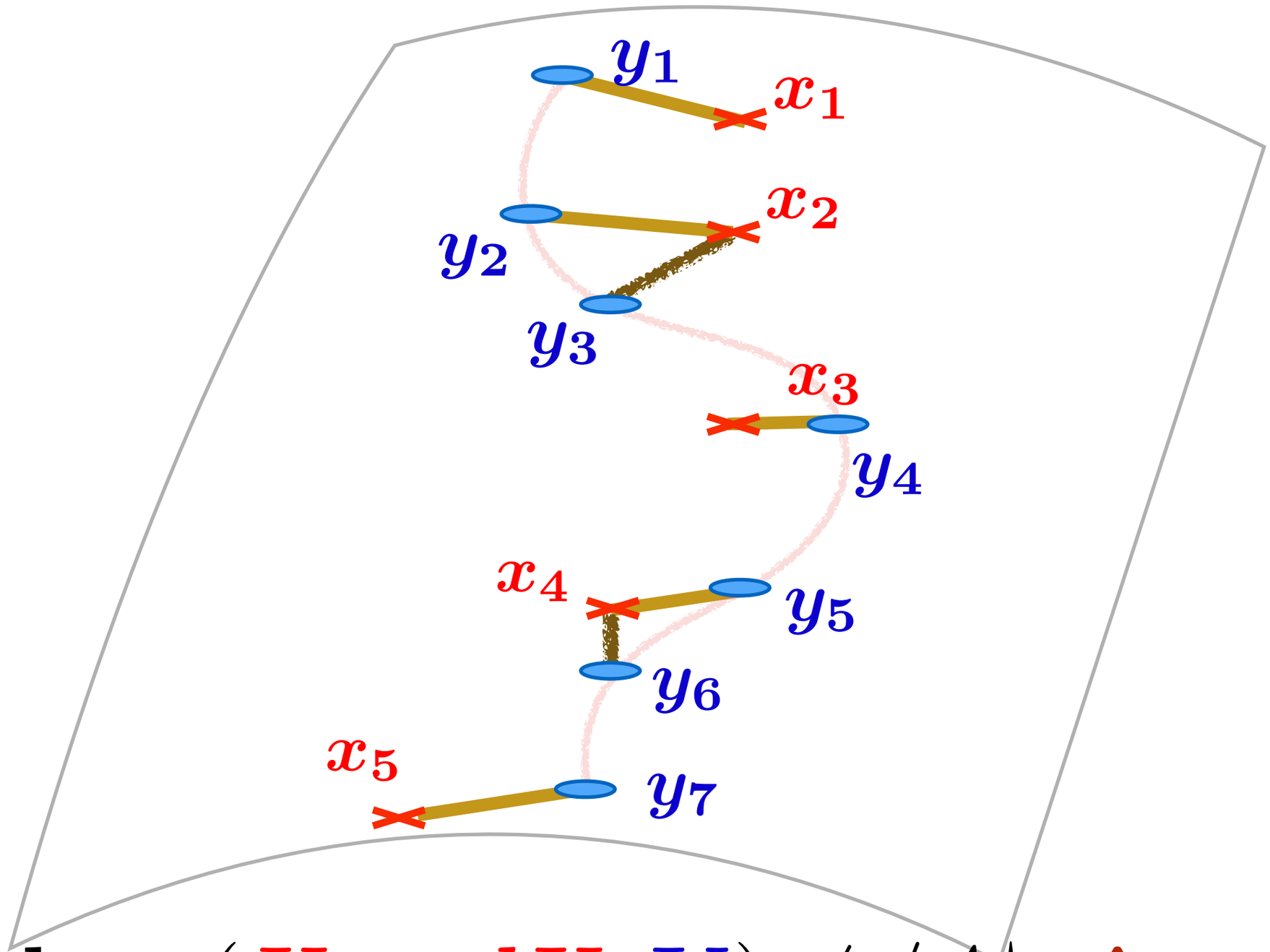
$$\mathbf{dtw}_0(\boldsymbol{X} + \boldsymbol{dX}, \boldsymbol{Y}) = \boldsymbol{r_{n,m}} + ?$$

DTW as a Loss: Differentiability?



$$\mathbf{dtw}_0(\boldsymbol{X} + \boldsymbol{dX}, \boldsymbol{Y}) = \langle A^\star, \boldsymbol{\Delta}_{\boldsymbol{X}+\boldsymbol{dX}\,\boldsymbol{Y}} \rangle$$

DTW as a Loss: Differentiability?



$$\mathbf{dtw}_0(\boldsymbol{X} + \boldsymbol{dX}, \boldsymbol{Y}) \neq \langle A^\star, \boldsymbol{\Delta}_{\boldsymbol{X}+\boldsymbol{dX}\,\boldsymbol{Y}} \rangle$$

DTW as a Loss: Differentiability?

$$\mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \min_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

$$\frac{\partial\, \mathbf{dtw}_0(\boldsymbol{X},\boldsymbol{Y})}{\partial \boldsymbol{X}} = \frac{\partial\, \mathbf{dtw}_0(\boldsymbol{X},\boldsymbol{Y})}{\partial \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}}} \frac{\partial \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}}}{\partial \boldsymbol{X}}$$

$$\mathbb{R}^{\boldsymbol{n}\times\boldsymbol{m}} \to \mathbb{R} \qquad \mathbb{R}^{|\Omega|\times \boldsymbol{n}} \to \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{m}}$$

DTW as a Loss: Differentiability?

$$\mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \min_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

$$\left(\frac{\partial\, \mathbf{dtw}_0(\boldsymbol{X},\boldsymbol{Y})}{\partial \boldsymbol{X}} \right)^{T} = \left(\frac{\partial\, \mathbf{dtw}_0(\boldsymbol{X},\boldsymbol{Y})}{\partial \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}}} \right)^{T} \left(\frac{\partial \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}}}{\partial \boldsymbol{X}} \right)^{T}$$

$$\mathbb{R}^{\boldsymbol{n}\times\boldsymbol{m}} \to \mathbb{R} \qquad \mathbb{R}^{|\Omega|\times\boldsymbol{n}} \to \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{m}}$$

DTW as a Loss: Differentiability?

$$\mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \min_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

$$\nabla_{\boldsymbol{X}} \mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \left(\frac{\partial \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}}}{\partial \boldsymbol{X}}\right)^T$$

$$\nabla_{\boldsymbol{\Delta}} \min_{\mathcal{A}(\boldsymbol{n},\boldsymbol{m})} \langle \cdot, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle = A^\star$$

$$\mathbb{R}^{\boldsymbol{n}\times\boldsymbol{m}} \longrightarrow \mathbb{R}^{|\Omega|\times\boldsymbol{n}}$$

iff optimal solution is unique

DTW as a Loss: Differentiability?

$$\mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \min_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

- $\mathbf{dtw}_0$ is piecewise **linear** w.r.t $\boldsymbol{\Delta}$
- if $\boldsymbol{\Delta}_{ij} = \boldsymbol{\delta}(\boldsymbol{x}_i, \boldsymbol{y}_j) = \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2$, $\mathbf{dtw}_0$ is piecewise **quadratic** w.r.t. $\boldsymbol{X}$.

DTW as a Loss: Differentiability?

$$\mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \min_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

- If $A^\star$ is **unique**,

$$\nabla_X \mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \left(\frac{\partial \boldsymbol{\Delta}(\boldsymbol{X}, \boldsymbol{Y})}{\partial \boldsymbol{X}} \right)^T A^\star$$

- $\mathbf{dtw}_0$ has a discontinuous gradient.

Any way to fix this?

$$\mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \min_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

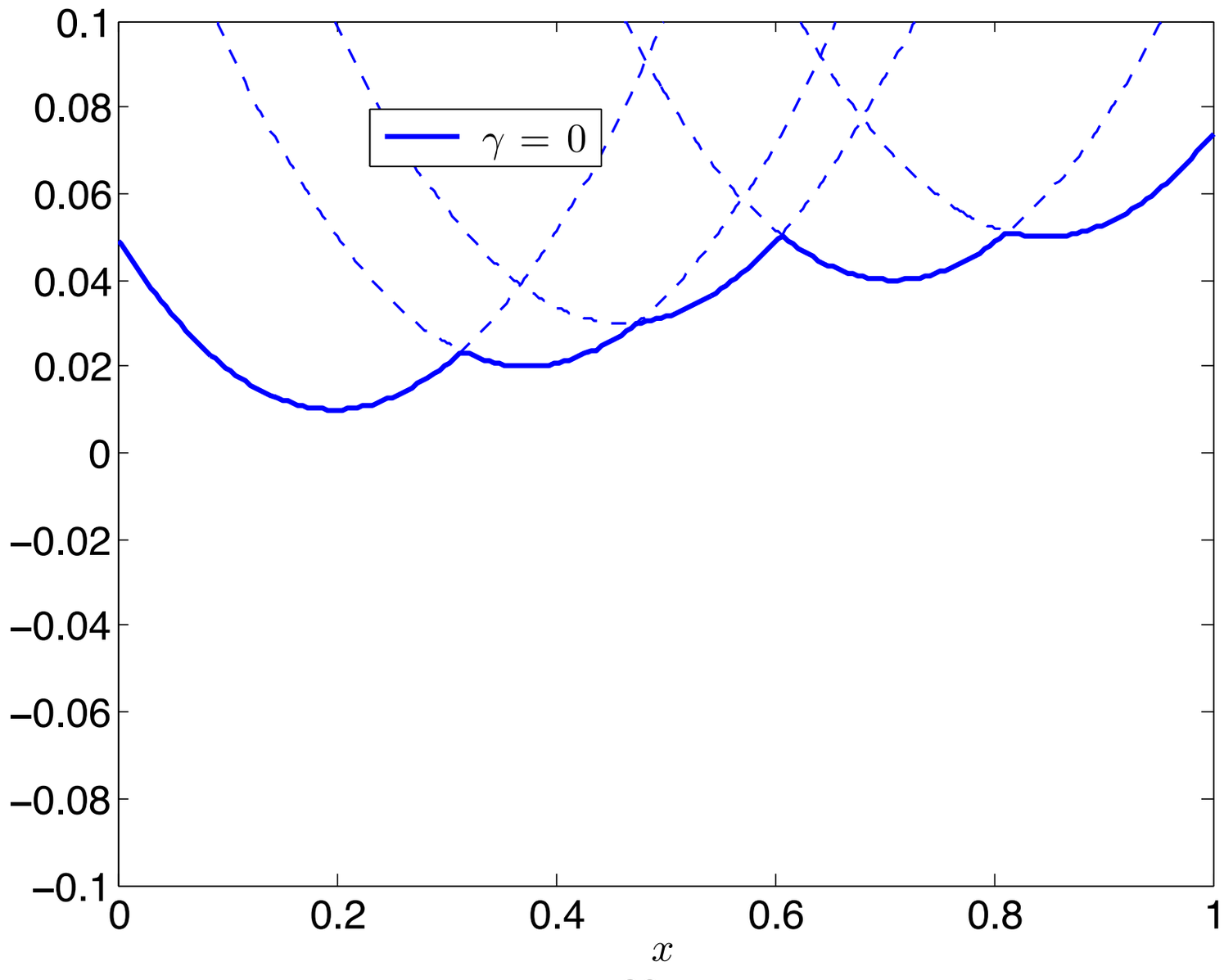
Problem: non-differentiability of min operator over finite family of values.

Fix

$$\min^{\gamma}(u_1, \ldots, u_n) = \begin{cases} \min_{i \leq n} u_i, & \gamma = 0, \\ -\gamma \log \sum_{i=1}^{n} e^{-u_i/\gamma}, & \gamma > 0. \end{cases}$$

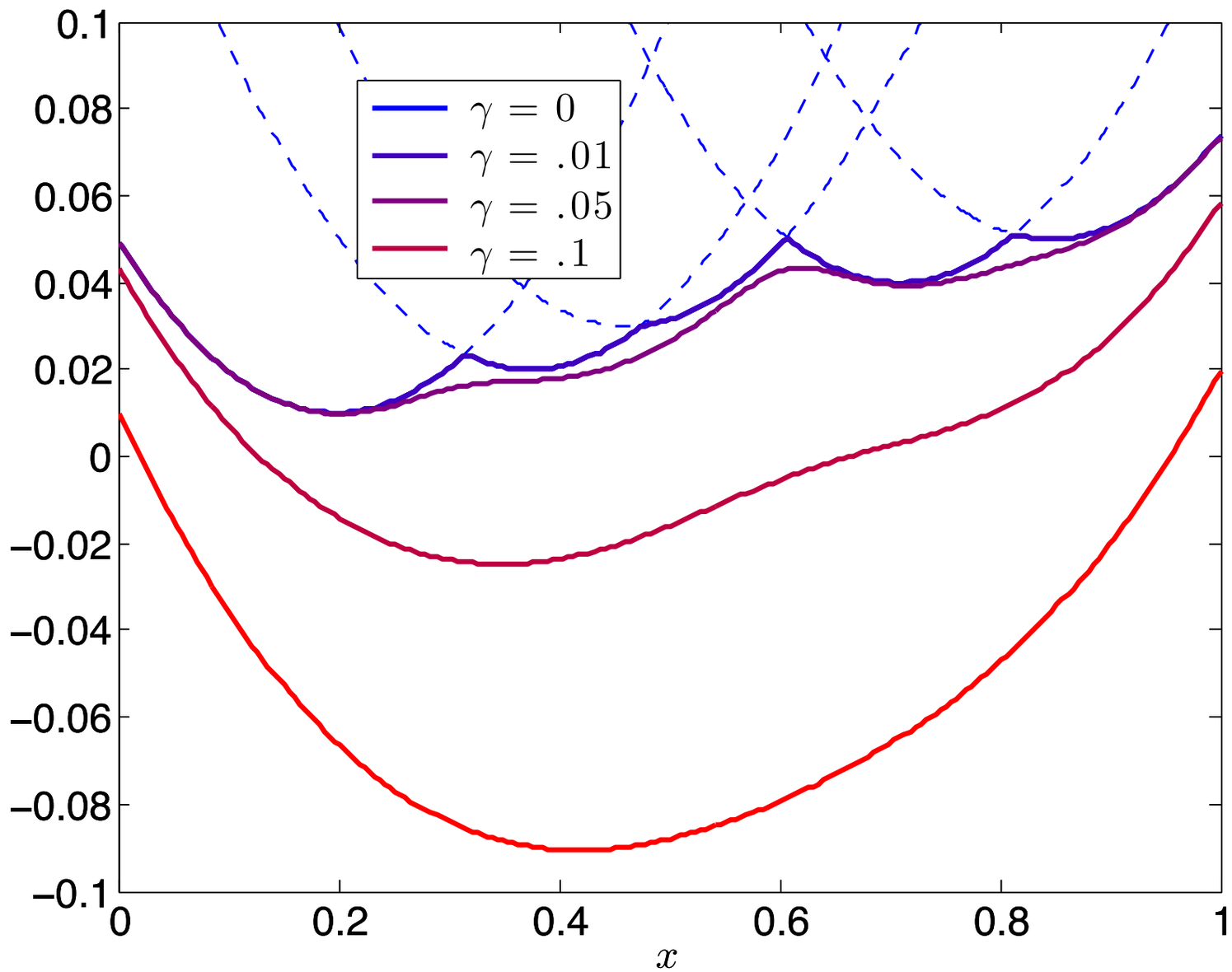
Example softmin of quadratic functions

$$f(\boldsymbol{x}) = \min_{i=1,\ldots,s}{}^{\gamma} \; a_i \boldsymbol{x}^2 + b_i \boldsymbol{x} + c_i$$



Example softmin of quadratic functions

$$f(\boldsymbol{x}) = \min_{i=1,\ldots,s}{}^{\gamma} \; a_i \boldsymbol{x}^2 + b_i \boldsymbol{x} + c_i$$



Soft-DTW

$$\mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \min_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

Fix: Replace min by $\min^\gamma, \gamma > 0$

$$\mathbf{dtw}_\gamma(\boldsymbol{X}, \boldsymbol{Y}) = \min^\gamma_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

$$\mathbf{dtw}_\gamma(\boldsymbol{X}, \boldsymbol{Y}) = -\gamma \log \sum_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} e^{-\frac{\langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle}{\gamma}}$$

Soft-DTW

$$\mathbf{dtw}_0(\boldsymbol{X},\boldsymbol{Y}) = \min_{A\in\mathcal{A}(\boldsymbol{n},\boldsymbol{m})} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

Fix: Replace min by $\min^\gamma, \gamma > 0$

$$\mathbf{dtw}_{\boldsymbol{\gamma}}(\boldsymbol{X},\boldsymbol{Y}) = \min\nolimits^{\boldsymbol{\gamma}}_{A\in\mathcal{A}(\boldsymbol{n},\boldsymbol{m})} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

$$\mathbf{dtw}_{\boldsymbol{\gamma}}(\boldsymbol{X},\boldsymbol{Y}) = -\boldsymbol{\gamma}\log \sum_{A\in\mathcal{A}(\boldsymbol{n},\boldsymbol{m})} e^{-\frac{\langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle}{\boldsymbol{\gamma}}}$$

k_{GA}

See CVBM'07

Recursive Computation

$$\mathbf{dtw}_0(\boldsymbol{X}, \boldsymbol{Y}) = \min_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

$$r_{i,j} = \min(r_{i-1,j-1}\,,\, r_{i-1,j}\,,\, r_{i,j-1}) + \boldsymbol{\Delta}_{\boldsymbol{i},\boldsymbol{j}}$$

$$\mathbf{dtw}_{\gamma}(\boldsymbol{X}, \boldsymbol{Y}) = \min^{\gamma}_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

$$r_{i,j} = \min^{\gamma}(r_{i-1,j-1}\,,\, r_{i-1,j}\,,\, r_{i,j-1}) + \boldsymbol{\Delta}_{\boldsymbol{i},\boldsymbol{j}}$$

CVBM'07 reformulated

DTW as a Loss: Differentiability?

$$\mathbf{dtw}_{\gamma}(\boldsymbol{X}, \boldsymbol{Y}) = \min_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})}^{\gamma} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

$$\nabla_{\boldsymbol{X}} \mathbf{dtw}_{\gamma}(\boldsymbol{X}, \boldsymbol{Y}) = \left(\frac{\partial \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}}}{\partial \boldsymbol{X}} \right)^T$$

$$\nabla_{\boldsymbol{\Delta}} - \gamma \log \sum_{\mathcal{A}(\boldsymbol{n},\boldsymbol{m})} e^{-\langle \cdot, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle / \gamma}$$

$$\mathbb{R}^{\boldsymbol{n} \times \boldsymbol{m}} \longrightarrow \mathbb{R}^{|\Omega| \times \boldsymbol{n}}$$

$$\mathbb{E}_{\gamma}[A] := \frac{\sum_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} A e^{-\frac{\langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle}{\gamma}}}{\sum_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} e^{-\frac{\langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle}{\gamma}}}$$

DTW as a Loss: Differentiability?

$$\mathbf{dtw}_{\gamma}(\boldsymbol{X}, \boldsymbol{Y}) = \min_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})}^{\gamma} \langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

$$\nabla_{\boldsymbol{X}} \mathbf{dtw}_{\gamma}(\boldsymbol{X}, \boldsymbol{Y}) = \left(\frac{\partial \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}}}{\partial \boldsymbol{X}} \right)^{T}$$

$$\nabla_{\boldsymbol{\Delta}} - \gamma \log \sum_{\mathcal{A}(\boldsymbol{n},\boldsymbol{m})} e^{-\langle \cdot, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle / \gamma}$$

$$\mathbb{R}^{\boldsymbol{n} \times \boldsymbol{m}} \rightarrow \mathbb{R}^{|\Omega| \times \boldsymbol{n}}$$

$$\mathbb{E}_{\gamma}[A] := \frac{\sum_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} A e^{-\frac{\langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle}{\gamma}}}{\sum_{A \in \mathcal{A}(\boldsymbol{n},\boldsymbol{m})} e^{-\frac{\langle A, \boldsymbol{\Delta}_{\boldsymbol{X}\boldsymbol{Y}} \rangle}{\gamma}}}$$

Expectation of Path under Gibbs distribution

Recursive Computation (Backward)

$$r_{i,j} = \min{}^{\gamma}\left(r_{i-1,j-1}\,,\, r_{i-1,j}\,,\, r_{i,j-1}\right) + \boldsymbol{\Delta_{i,j}}$$

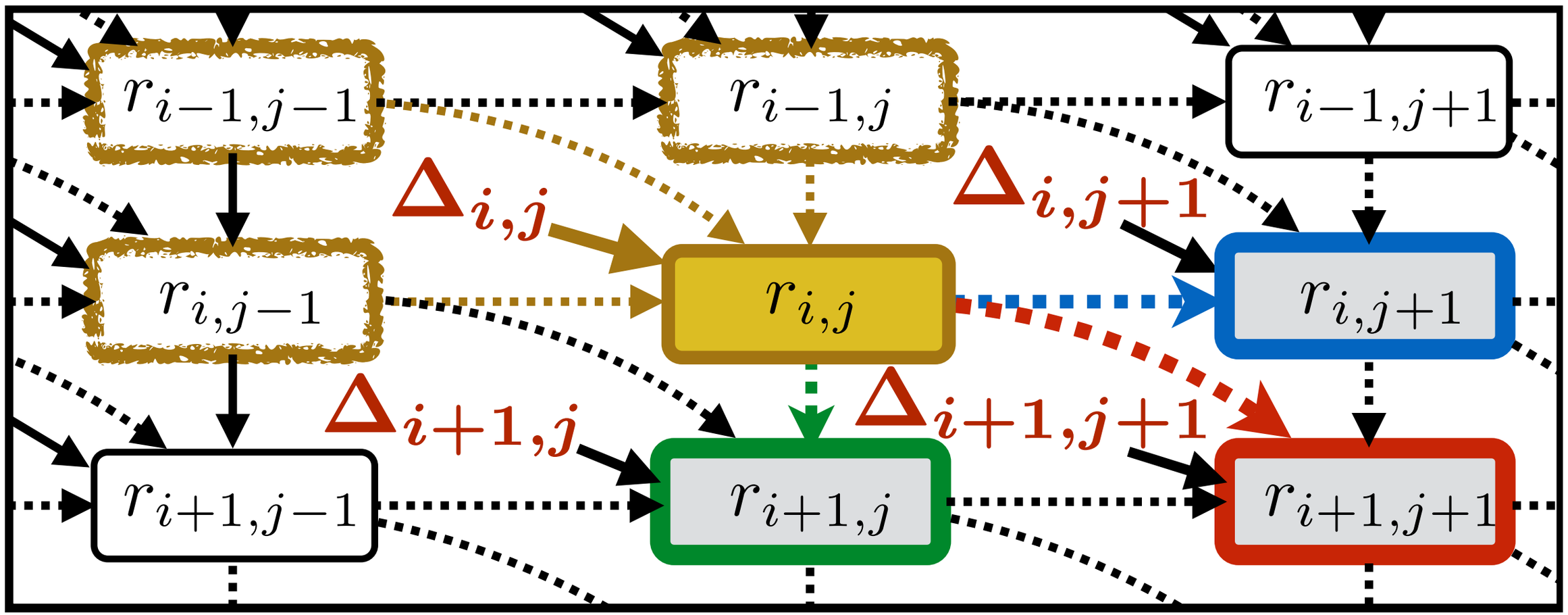
$$\frac{\partial\, \mathbf{dtw}_\gamma(\boldsymbol{X},\boldsymbol{Y})}{\partial \boldsymbol{X}} = \frac{\partial\, \mathbf{dtw}_\gamma(\boldsymbol{X},\boldsymbol{Y})}{\partial \boldsymbol{\Delta}} \frac{\partial \boldsymbol{\Delta}}{\partial \boldsymbol{X}}$$

$$\frac{\partial\, \mathbf{dtw}_\gamma(\boldsymbol{X},\boldsymbol{Y})}{\partial \boldsymbol{\Delta}} = \left[\frac{\partial \boldsymbol{r_{n,m}}}{\partial \boldsymbol{\Delta_{ij}}}\right]_{ij} = \left[\frac{\partial \boldsymbol{r_{n,m}}}{\partial r_{i,j}} \frac{\partial r_{i,j}}{\partial \boldsymbol{\Delta_{ij}}}\right]_{ij} = \left[\frac{\partial \boldsymbol{r_{n,m}}}{\partial r_{i,j}}\right]_{ij}$$

$$e_{i,j} = \frac{\partial \boldsymbol{r_{n,m}}}{\partial r_{i,j}}$$

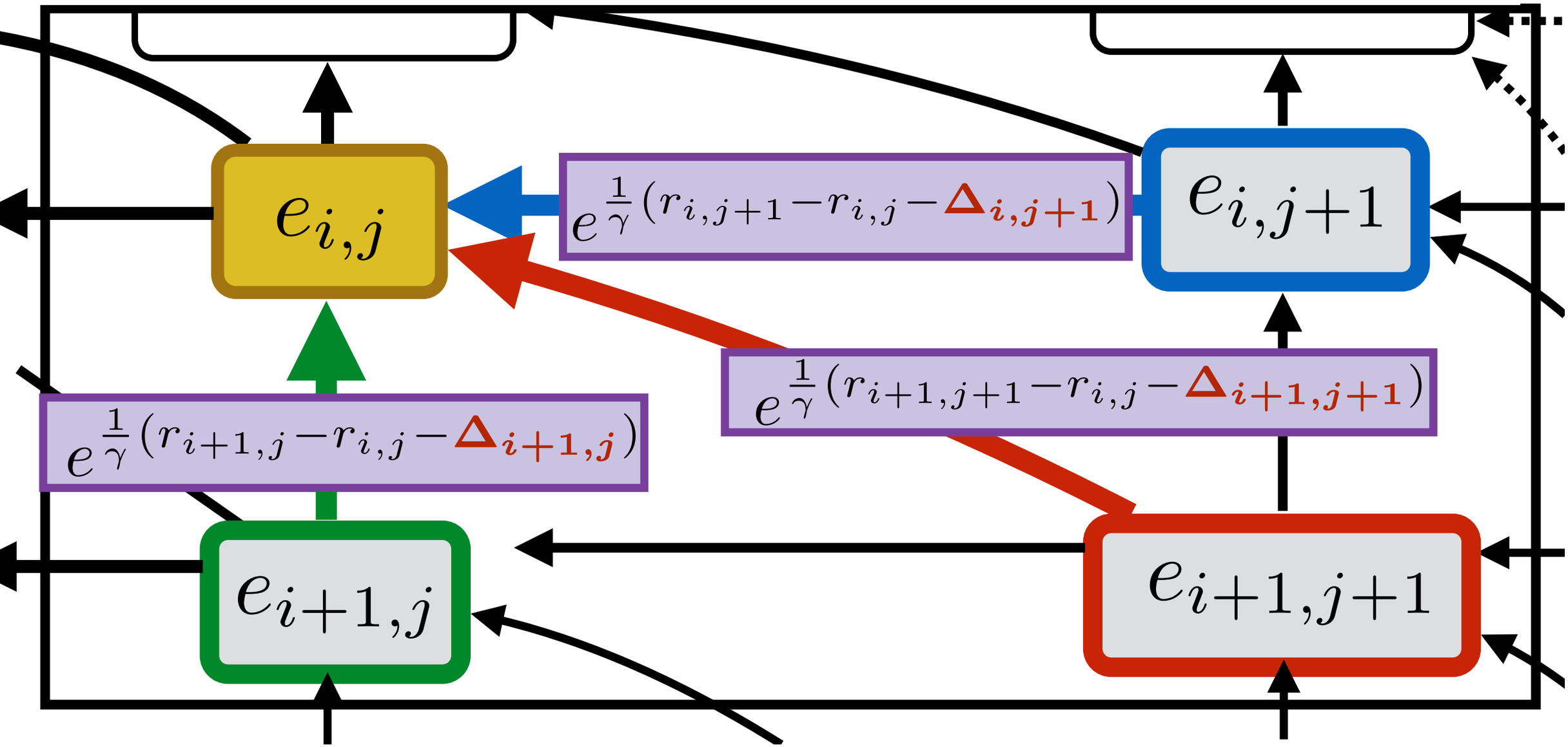
Computation Graph: Forward

Bellman's recursion has the following computational graph



Backward Pass

with a few simplifications, the backward pass boils down to the following updates.



Backward Recurrence

$$a = e^{\frac{1}{\gamma}(r_{i+1,j} - r_{i,j} - \boldsymbol{\Delta_{i+1,j}})}$$

$$b = e^{\frac{1}{\gamma}(r_{i,j+1} - r_{i,j} - \boldsymbol{\Delta_{i,j+1}})}$$

$$c = e^{\frac{1}{\gamma}(r_{i+1,j+1} - r_{i,j} - \boldsymbol{\Delta_{i+1,j+1}})}$$

$$e_{i,j} = e_{i+1,j} \cdot a + e_{i,j+1} \cdot b + e_{i+1,j+1} \cdot c$$

$$\nabla_X \mathbf{dtw}_\gamma(\boldsymbol{X}, \boldsymbol{Y}) = \left(\frac{\partial \boldsymbol{\Delta}(\boldsymbol{X}, \boldsymbol{Y})}{\partial \boldsymbol{X}} \right)^T E$$

Backward Recurrence

$$a = e^{\frac{1}{\gamma}(r_{i+1,j} - r_{i,j} - \boldsymbol{\Delta_{i+1,j}})}$$

$$b = e^{\frac{1}{\gamma}(r_{i,j+1} - r_{i,j} - \boldsymbol{\Delta_{i,j+1}})}$$

$$c = e^{\frac{1}{\gamma}(r_{i+1,j+1} - r_{i,j} - \boldsymbol{\Delta_{i+1,j+1}})}$$

$$e_{i,j} = e_{i+1,j} \cdot a + e_{i,j+1} \cdot b + e_{i+1,j+1} \cdot c$$

$$\nabla_X \mathbf{dtw}_\gamma(\boldsymbol{X}, \boldsymbol{Y}) = \left(\frac{\partial \boldsymbol{\Delta}(\boldsymbol{X}, \boldsymbol{Y})}{\partial \boldsymbol{X}} \right)^T E$$

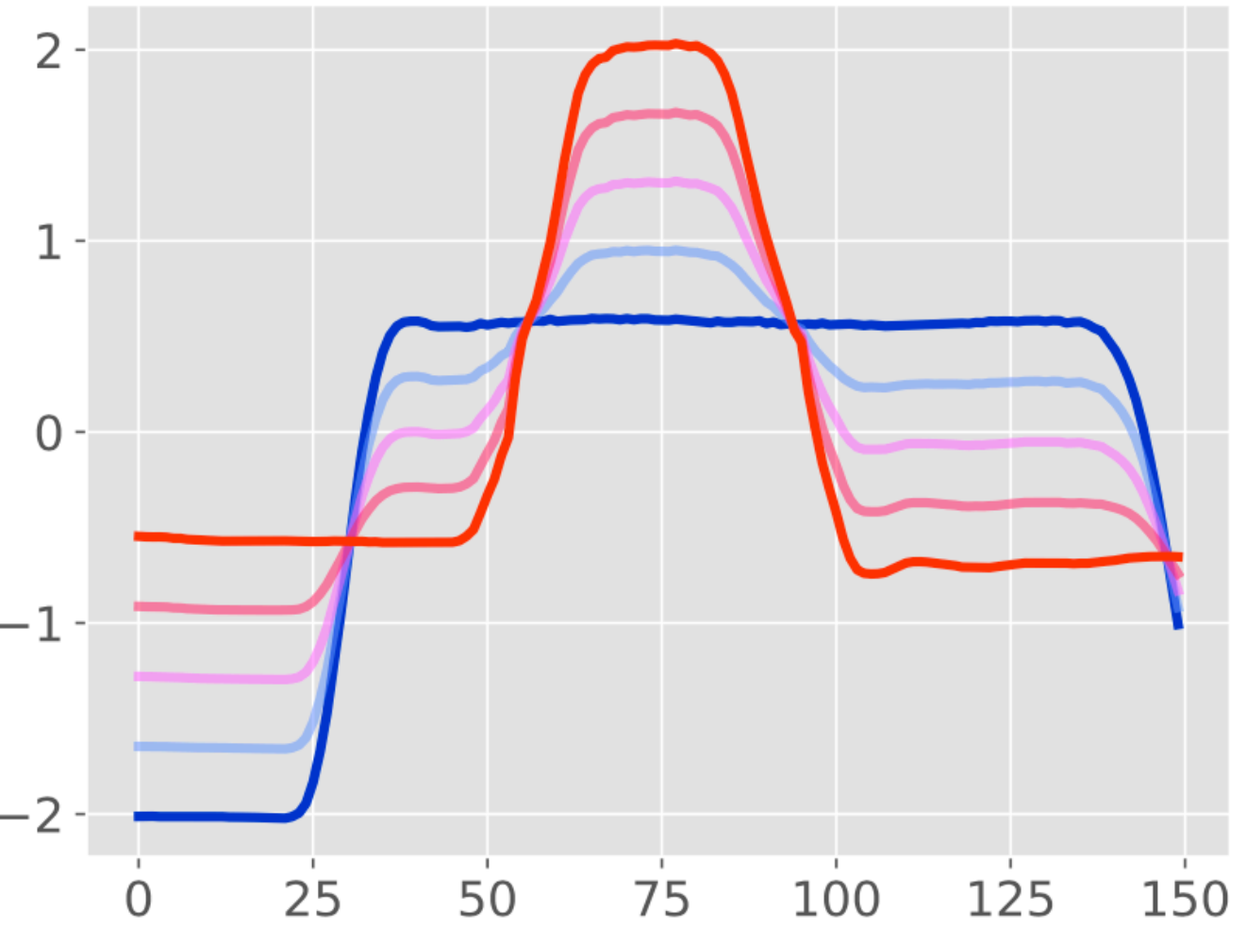
0. The DTW Geometry

1. Soft-DTW

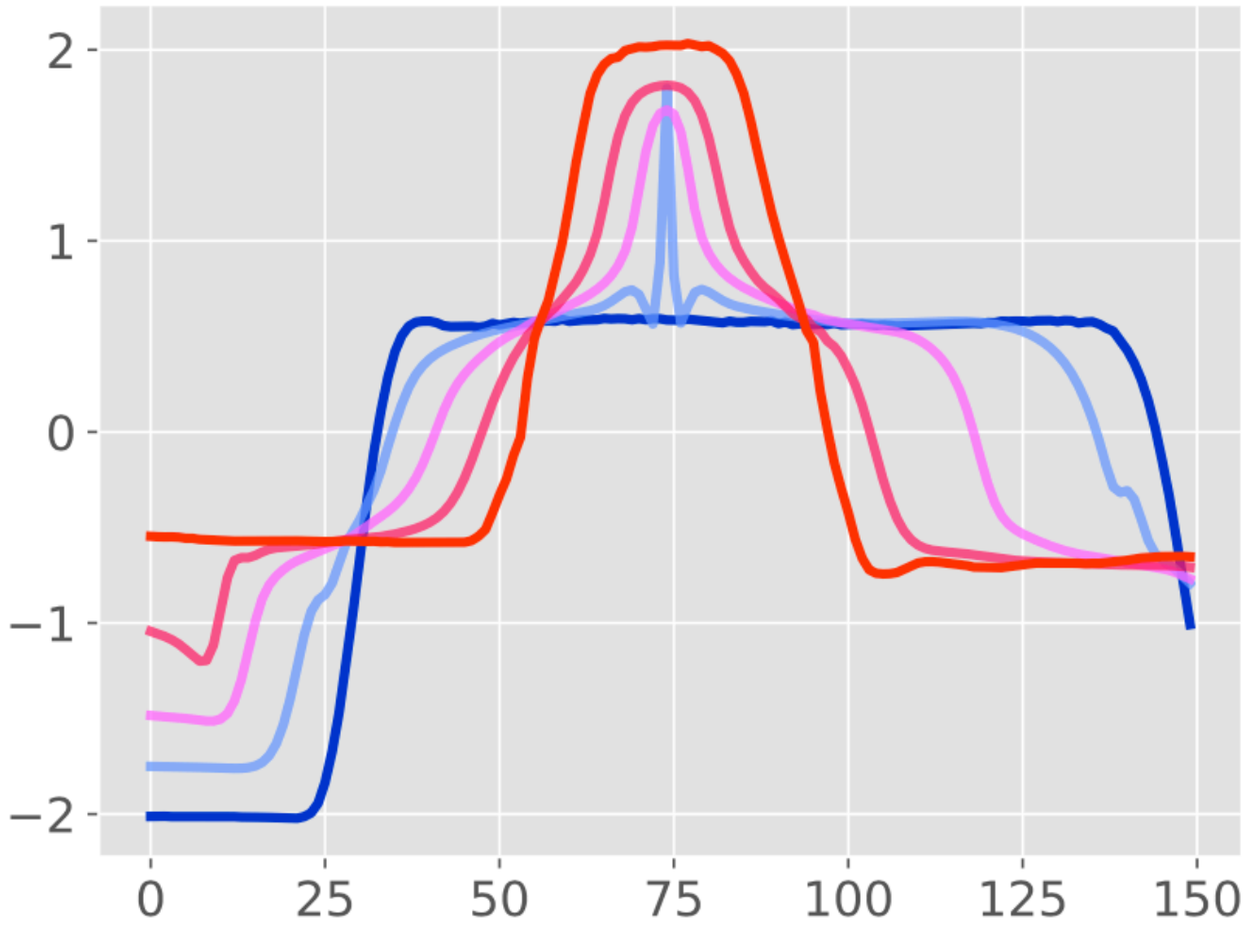
2. Soft-DTW as a Loss Function

Interpolation Between 2 Time Series

$$\min_{\mathbf{X}} \left[\lambda \, \mathbf{dtw}_\gamma(\mathbf{X}, \boldsymbol{Y_1}) + (1-\lambda)\, \mathbf{dtw}_\gamma(\mathbf{X}, \boldsymbol{Y_2})\right]$$



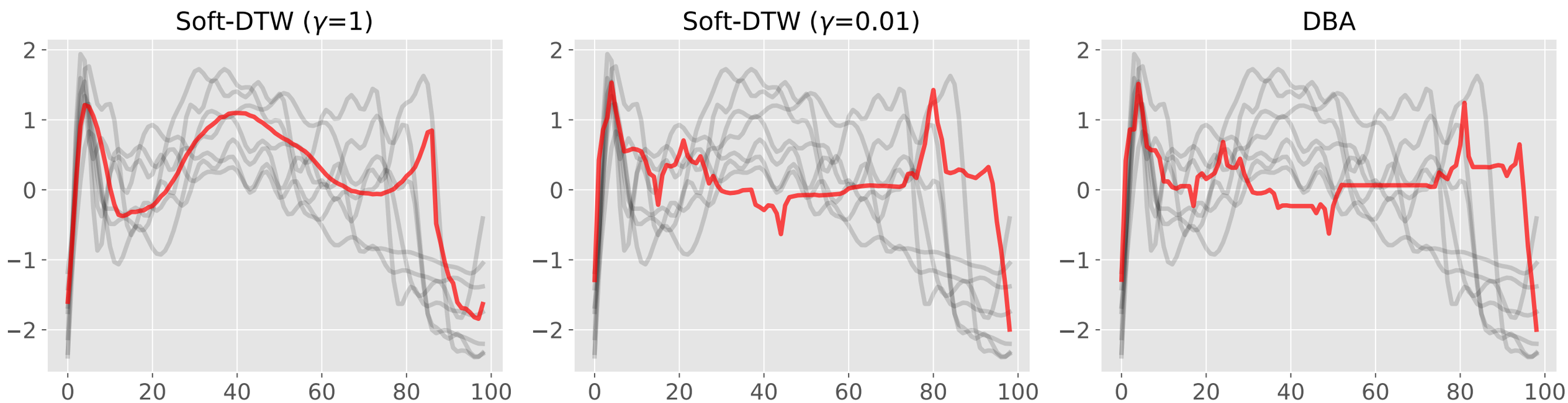
Euclidean loss



Soft-DTW loss ($\gamma = 1$)

sDTW Barycenter

$$\min_{\boldsymbol{X}} \sum_{j=1} \frac{\lambda_j}{m_j} \mathbf{dtw}_\gamma(\boldsymbol{X}, \boldsymbol{Y_j})$$



[DBA] Petitjean et al., A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44 (3):678–693, 2011.

sDTW Barycenter

$$\min_{\boldsymbol{X}} \sum_{j=1} \frac{\lambda_j}{m_j} \mathbf{dtw}_\gamma(\boldsymbol{X}, \boldsymbol{Y_j})$$

Table 1. Percentage of the datasets on which the proposed soft-DTW barycenter is achieving lower DTW loss (Equation (4) with $\gamma = 0$) than competing methods.

	Random initialization	Euclidean mean initialization
Comparison with DBA		
$\gamma = 1$	40.51%	3.80%
$\gamma = 0.1$	93.67%	46.83%
$\gamma = 0.01$	100%	79.75%
$\gamma = 0.001$	97.47%	89.87%
Comparison with subgradient method		
$\gamma = 1$	96.20%	35.44%
$\gamma = 0.1$	97.47%	72.15%
$\gamma = 0.01$	97.47%	92.41%
$\gamma = 0.001$	97.47%	97.47%

sDTW Barycenter

use $\gamma > 0$ to compute argmin,
evaluate energy with $\mathbf{dtw}_0$.

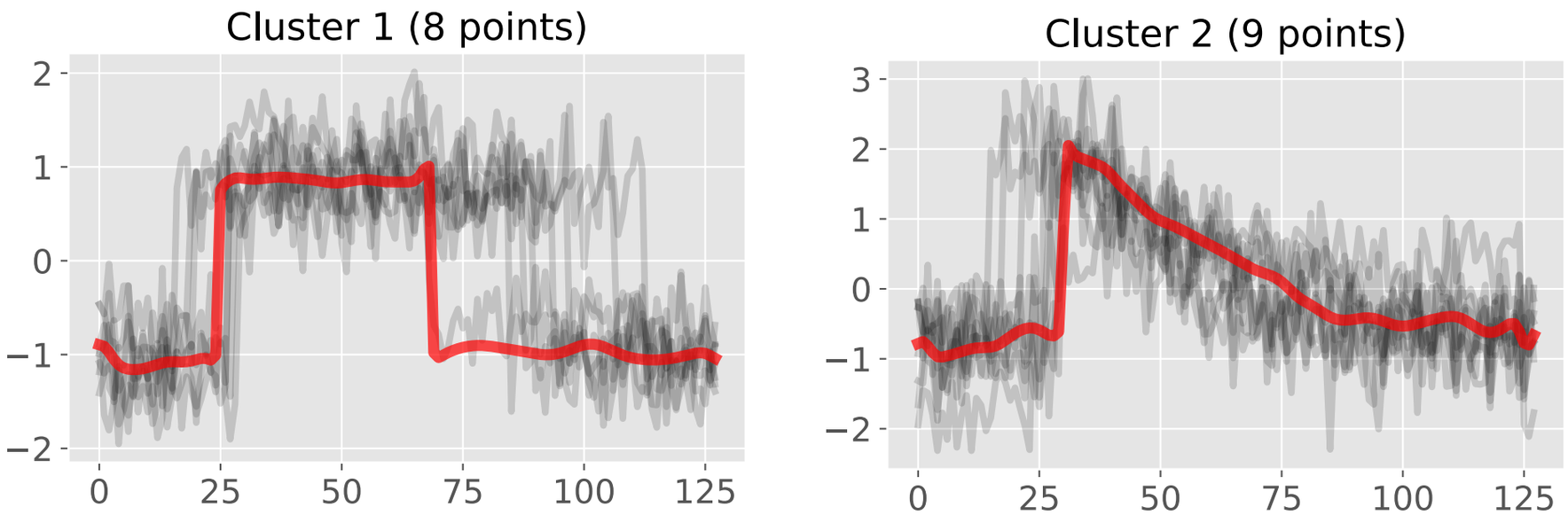
$\boldsymbol{Y}_j$)

…rcentage of the datasets on which the proposed soft-DTW barycenter is achieving lower DTW loss (Equation (4) with $\gamma = 0$) than competing methods.

	Random initialization	Euclidean mean initialization
Comparison with DBA		
$\gamma = 1$	40.51%	3.80%
$\gamma = 0.1$	93.67%	46.83%
$\gamma = 0.01$	100%	79.75%
$\gamma = 0.001$	97.47%	89.87%
Comparison with subgradient method		
$\gamma = 1$	96.20%	35.44%
$\gamma = 0.1$	97.47%	72.15%
$\gamma = 0.01$	97.47%	92.41%
$\gamma = 0.001$	97.47%	97.47%

sDTW Clustering

$$\min_{\boldsymbol{X_1},\ldots,\boldsymbol{X_k}} \sum_{j=1}^{N} \min_{\boldsymbol{i}=1,\ldots,k} \mathbf{dtw}_\gamma(\boldsymbol{X_i}, \boldsymbol{Y_j})$$



sDTW Clustering

$$\min_{\boldsymbol{X_1},\ldots,\boldsymbol{X_k}} \sum_{j=1}^{N} \min_{\boldsymbol{i}=1,\ldots,k} \mathbf{dtw}_\gamma(\boldsymbol{X_i}, \boldsymbol{Y_j})$$

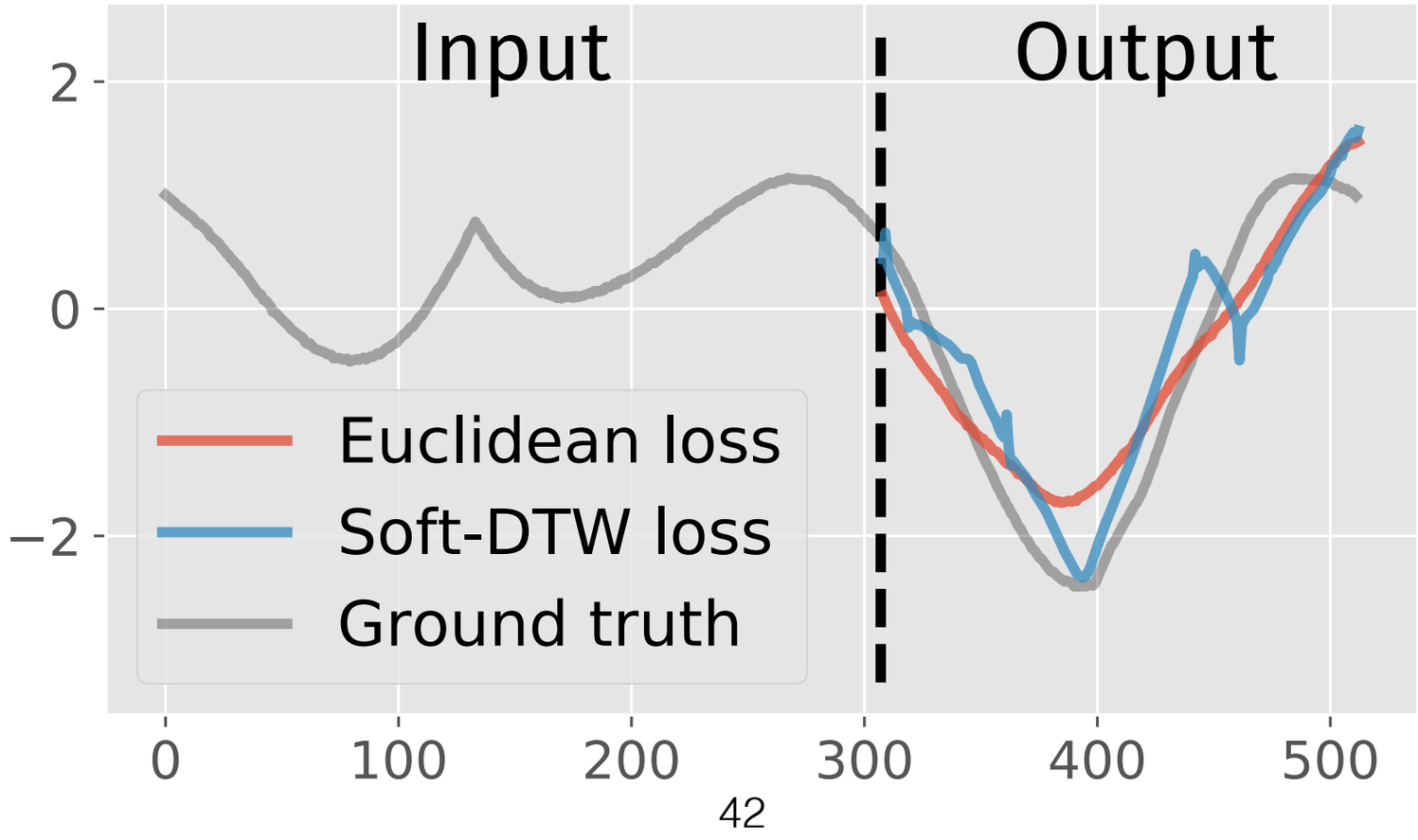
Table 2. Percentage of the datasets on which the proposed soft-DTW based k-**means** is achieving lower DTW loss (Equation (5) with $\gamma = 0$) than competing methods.

	Random initialization	Euclidean mean initialization
Comparison with DBA		
$\gamma = 1$	15.78%	29.31%
$\gamma = 0.1$	24.56%	24.13%
$\gamma = 0.01$	59.64%	55.17%
$\gamma = 0.001$	77.19%	68.97%
Comparison with subgradient method		
$\gamma = 1$	42.10%	46.44%
$\gamma = 0.1$	57.89%	50%
$\gamma = 0.01$	76.43%	65.52%
$\gamma = 0.001$	96.49%	84.48%

sDTW Prediction Loss

$$f_{\boldsymbol{\theta}} : \mathcal{I} \to \mathbb{R}^{|\Omega| \times n}, (x_i, \boldsymbol{Y_i})_i \in \mathcal{I} \times \mathbb{R}^{|\Omega| \times m_i}$$

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \frac{1}{m_i} \mathbf{dtw}_\gamma(f_{\boldsymbol{\theta}}(x_i), \boldsymbol{Y_i})$$



sDTW Prediction Loss

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \frac{1}{m_i} \mathbf{dtw}_\gamma\left(f_{\boldsymbol{\theta}}(x_i), \boldsymbol{Y_i}\right)$$

Table 3. Averaged rank obtained by a multi-layer perceptron (MLP) under Euclidean and soft-DTW losses. Euclidean initialization means that we initialize the MLP trained with soft-DTW loss by the solution of the MLP trained with Euclidean loss.

Training loss	Random initialization	Euclidean initialization
When evaluating with DTW loss		
Euclidean	3.46	4.21
soft-DTW ($\gamma = 1$)	3.55	3.96
soft-DTW ($\gamma = 0.1$)	3.33	3.42
soft-DTW ($\gamma = 0.01$)	2.79	2.12
soft-DTW ($\gamma = 0.001$)	**1.87**	**1.29**
When evaluating with Euclidean loss		
Euclidean	**1.05**	**1.70**
soft-DTW ($\gamma = 1$)	2.41	2.99
soft-DTW ($\gamma = 0.1$)	3.42	3.38
soft-DTW ($\gamma = 0.01$)	4.13	3.64
soft-DTW ($\gamma = 0.001$)	3.99	3.29

Poster: #96

- **Dynamic Time Warping** is a natural and flexible discrepancy to compare time series, yet it is **non-differentiable.**

- **Soft-DTW** is a differentiable approximation, with better convexity properties.

- Using **soft-DTW** quantity results in better minima, even when measured with the original DTW

- Code available on

https://github.com/mblondel/soft-dtw