Stabilizing Training with Soft Dynamic Time Warping: A Case Study for Pitch Class Estimation with Weakly Aligned Targets

Johannes Zeitler, Simon Deniffel, Michael Krause, and Meinard Müller

Summary

Training instabilities when training on weakly aligned targets with soft dynamic time warping (SDTW) loss

Analyze the cause of instabilities

Propose stabilizing strategies

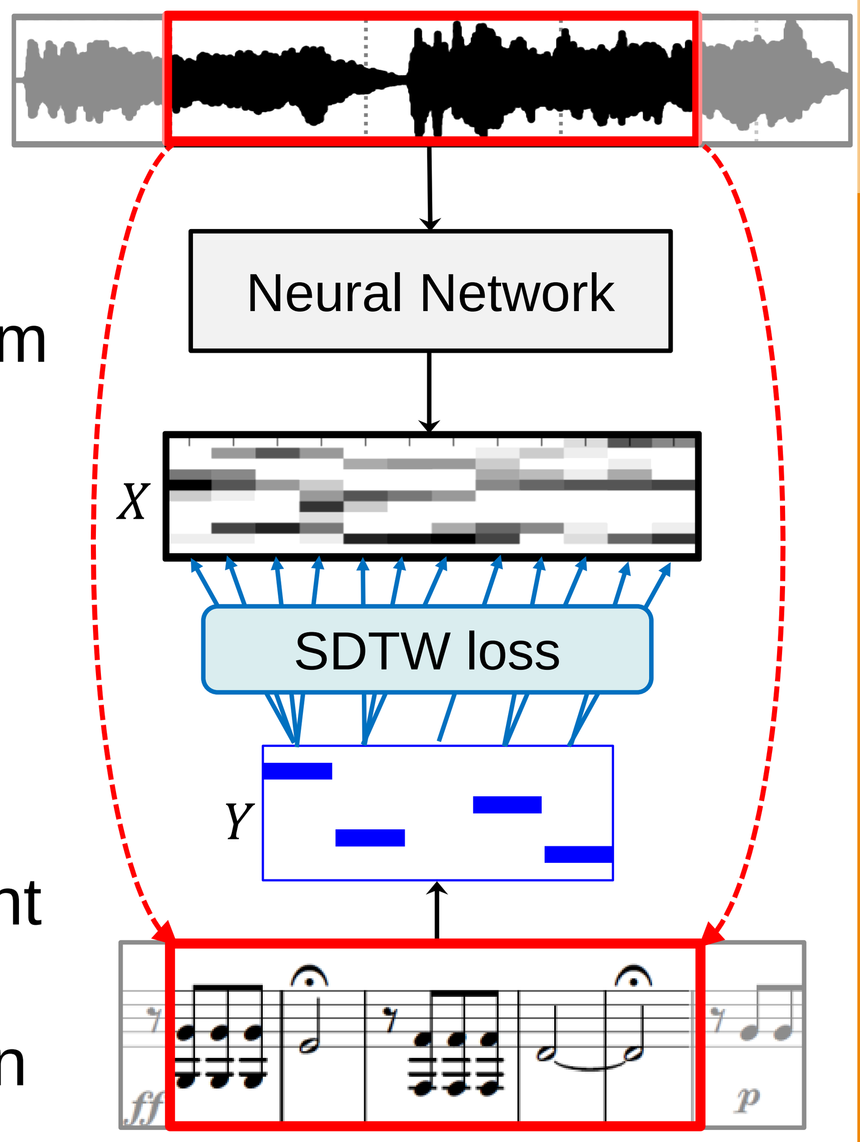
Training with Weakly Aligned Targets

Only annotate start & end of audio segments

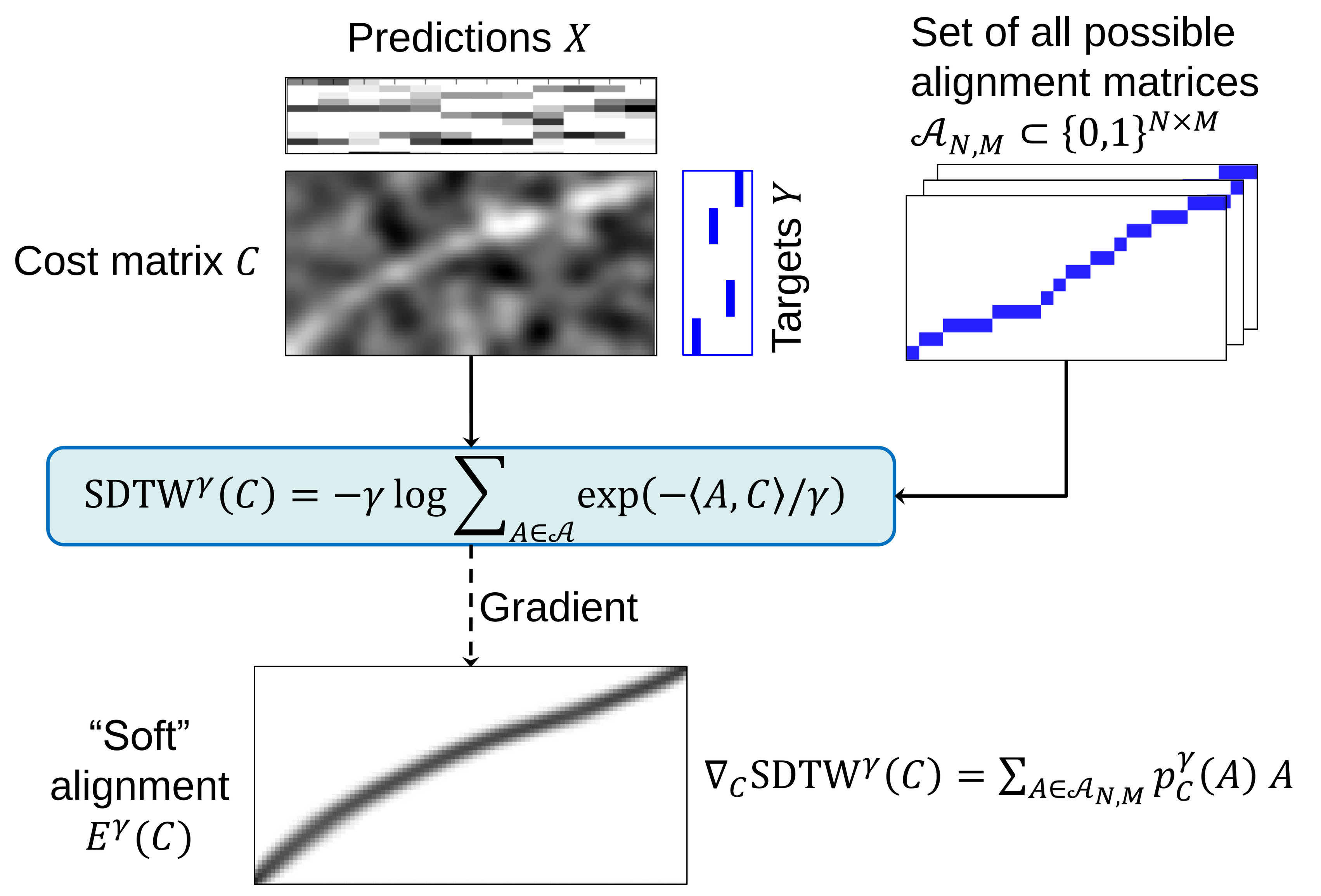
Retrieve note events between start & end from musical score

Weak targets Y provide information about note event order, but not duration

Use SDTW for alignment of predictions X and targets Y in loss function



Soft Dynamic Time Warping (SDTW) [1]



Experiments & Evaluation

CNN-based pitch class estimation from classical music recordings

Prediction sequence length 500

Average weak target sequence length ≈ 25

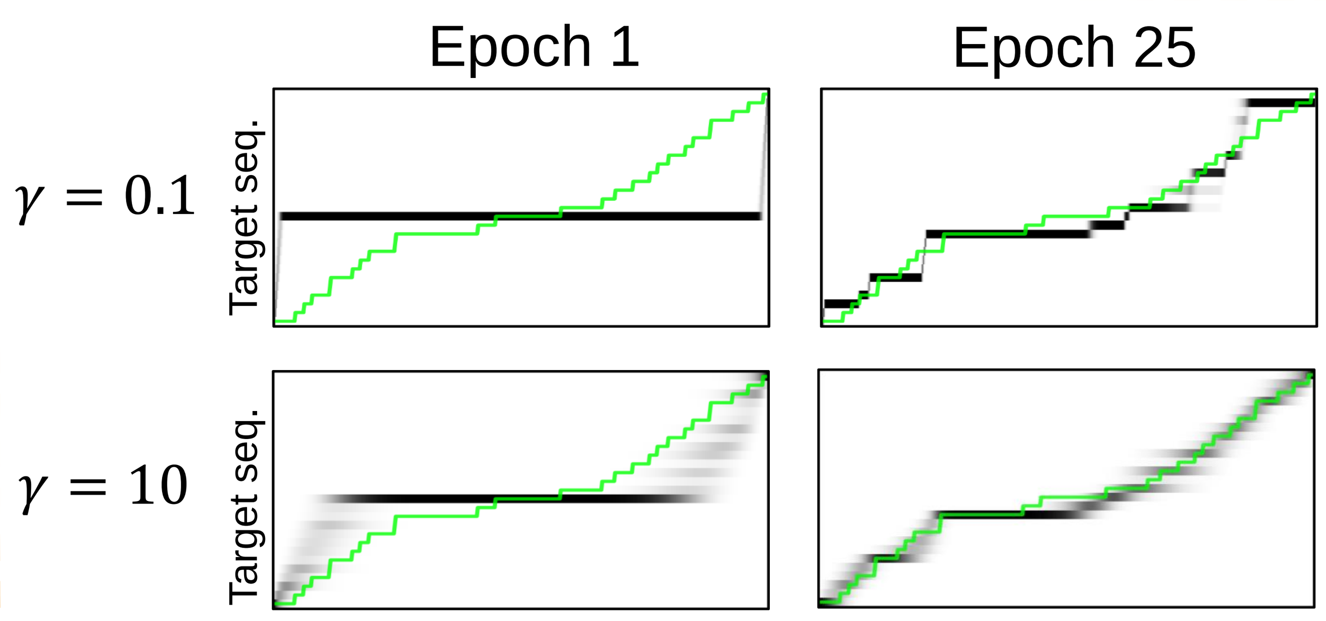
Inspect soft alignments $E^{\gamma}(C)$

	F-measure
Baseline: strongly aligned targets	$\mathbf{0.82 \pm 0.07}$

Standard SDTW & Stabilizing Strategies

Standard SDTW

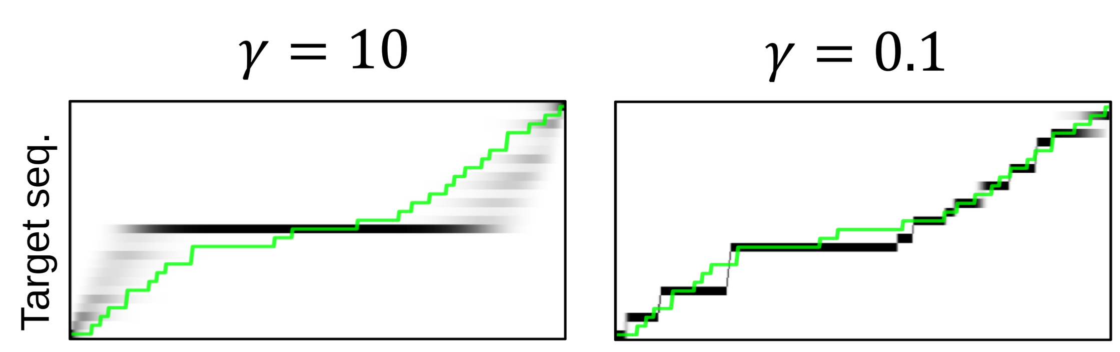
- Poor prediction accuracy in the early training stages
- Noisy cost matrix C
- Collapsed or blurry soft alignments $E^{\gamma}(C)$
- Erroneous gradient updates, training diverges



F-measure: $\gamma = 0.1$: 0.56 ± 0.37; $\gamma = 10$: 0.57 ± 0.37

Hyperparameter scheduling

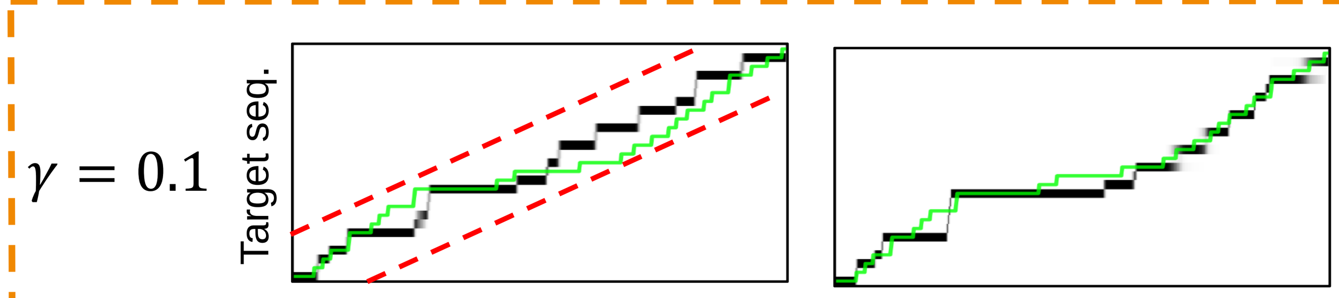
- Increase initial softmin temperature γ
- Blurry alignments partially overlap with the reference
- Reduce γ at a later training stage for sharp alignments



F-measure: 0.80 ± 0.04

Diagonal prior

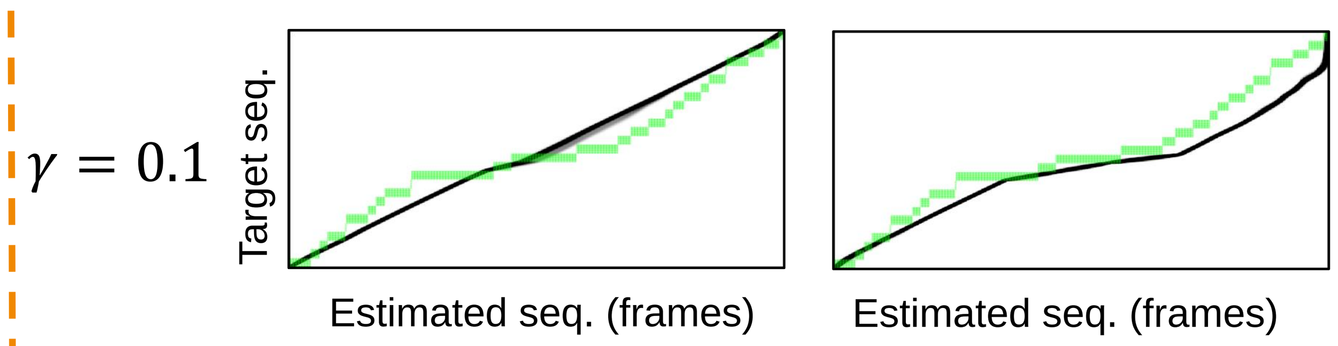
- Penalize off-diagonal elements in the computation of the soft alignment matrix



F-measure: <u>0.81 ± 0.02</u>

Sequence unfolding

- Uniformly stretch target sequence to length of predicted sequence
- Diagonal step is cheaper than "around corner"



F-measure: 0.53 ± 0.04

[1] Marco Cuturi and Mathieu Blondel, "Soft-DTW: a differentiable loss function for time-series," ICML 2017