# TREET: Transfer Entropy Estimation via Transformers

Omer Luxembourg, Dor Tsur, Prof. Haim Permuter

## Introduction and Background

- **Transfer entropy (TE)**

$$\mathsf{TE}_{X\to Y}(k,l) := \mathsf{I}\left(X^l; Y_l | Y_{l-k}^{l-1}\right)$$

- **Applications** span neuroscience, embodied cognitive systems, social networks, and finance
- **Estimation challenges:** traditional methods suffer from TE estimation for long sequences.

## Contributions

- **TREET** – new estimator for TE with transformers.
- **Various Applications** – TE estimation benchmark, capacity estimation, feature analysis, and probability density estimator.

## Derivation

- Donsker-Vardahan based variational formula,

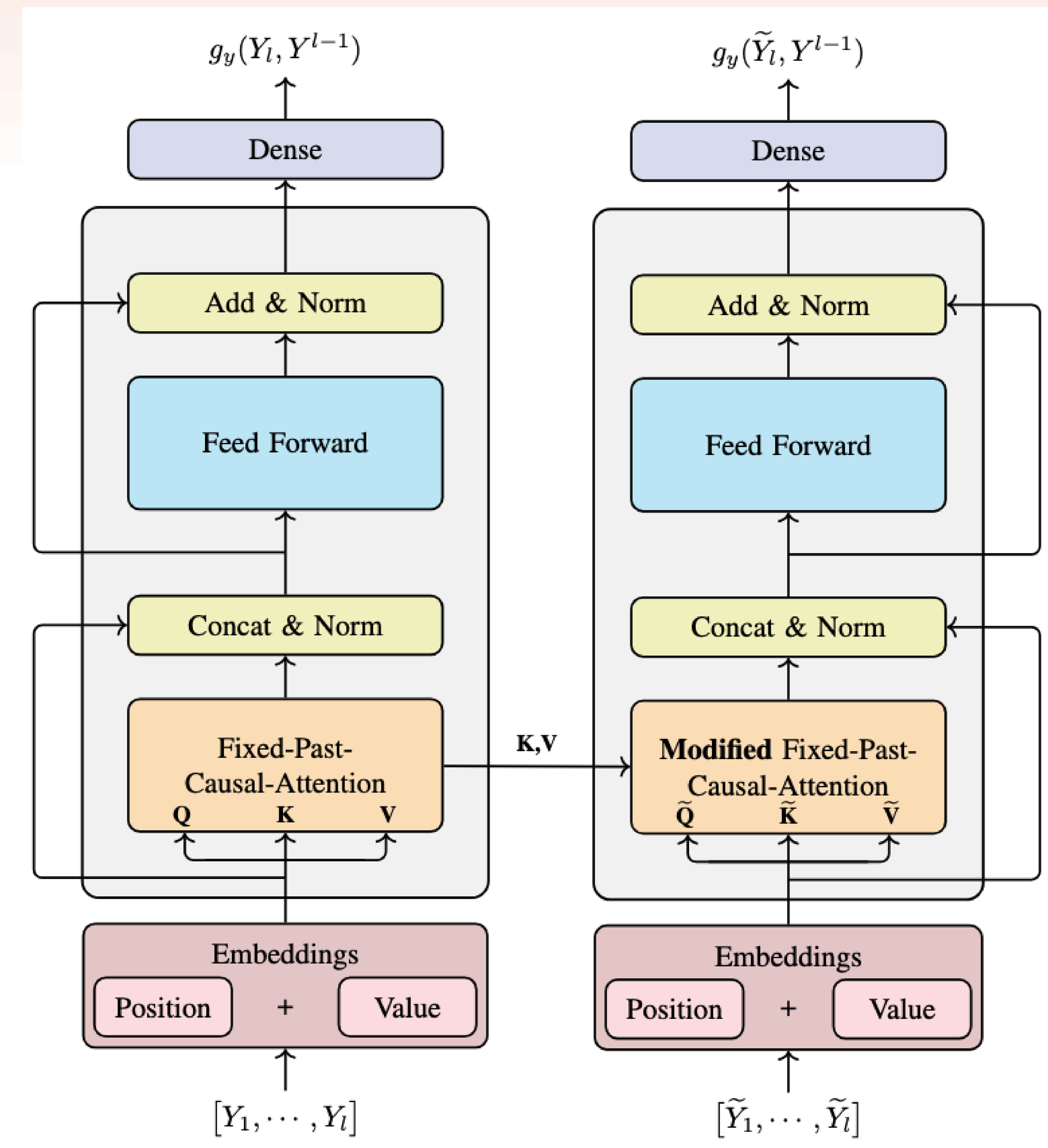
$$\mathsf{D}_{\mathsf{KL}}(P\|Q) = \sup_{f:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[f] - \log\left(\mathbb{E}_Q[e^f]\right)$$

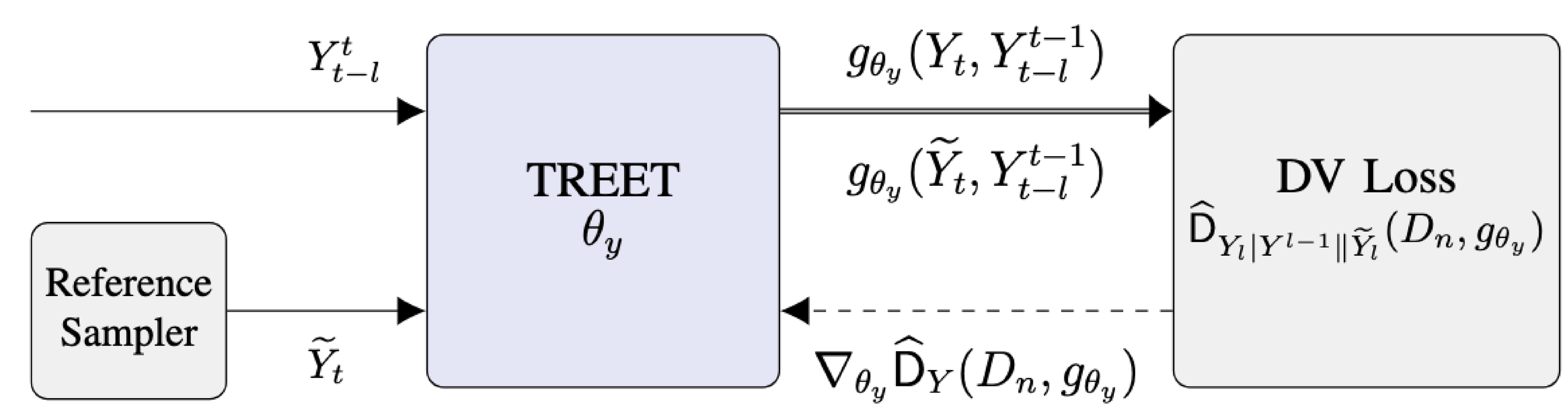
- Unsupervised learning based on sequential data,

$$\widehat{\mathsf{TE}}_{X\to Y}(D_n; l) =$$

$$\sup_{g_{xy}\in\mathcal{G}^{XY}_{\mathrm{ctf}}} \widehat{\mathsf{D}}_{Y_l|Y^{l-1}X^l\|\widetilde{Y}_l}(D_n, g_{xy}) - \sup_{g_y\in\mathcal{G}^{Y}_{\mathrm{ctf}}} \widehat{\mathsf{D}}_{Y_l|Y^{l-1}\|\widetilde{Y}_l}(D_n, g_y)$$

## Architecture and Algorithm




## Results and Applications

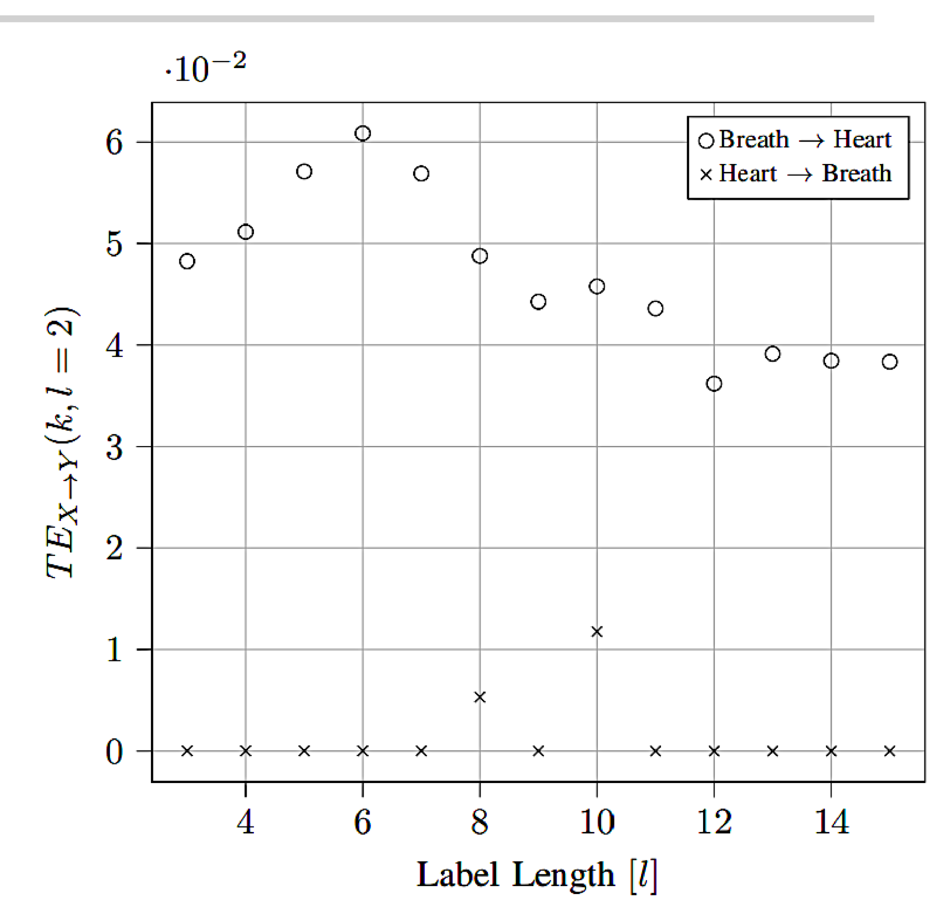
**TE Estimation Benchmark** – TREET achieves stable and consistent results for all memory length $l$

| $\lambda$ | | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|---|
| Model, $l$ | Ground Truth | 0.829 | 0.811 | 0.699 | 0.415 | 0.132 | 0.019 | 0.001 |
| TREET | 1 | 0.812 | 0.792 | 0.667 | 0.395 | 0.126 | 0.016 | 0.0 |
| | 2 | 0.826 | 0.796 | 0.681 | 0.392 | 0.117 | 0.014 | 0 |
| | 4 | 0.825 | 0.798 | 0.682 | 0.393 | 0.115 | 0.013 | 0 |
| | 7 | 0.82 | 0.799 | 0.679 | 0.405 | 0.121 | 0.015 | 0.001 |
| | 9 | 0.815 | 0.802 | 0.686 | 0.403 | 0.126 | 0.017 | 0.001 |
| | 19 | 0.824 | 0.805 | 0.69 | 0.405 | 0.128 | 0.017 | 0.001 |
| | 49 | 0.829 | 0.811 | 0.694 | 0.409 | 0.128 | 0.018 | 0.001 |
| | 99 | 0.829 | 0.81 | 0.693 | 0.41 | 0.115 | 0.017 | 0.001 |
| TENE | 1 | 0.823 | 0.807 | 0.696 | 0.416 | 0.126 | 0.014 | 0 |
| | 2 | 0.814 | 0.782 | 0.688 | 0.39 | 0.115 | 0.013 | 0 |
| | 4 | 0.43 | 0.76 | 0.602 | 0.382 | 0.119 | 0.013 | -0.001 |
| | 7 | >10 | >10 | 0.698 | 0.354 | 0.09 | 0.013 | 0.0 |
| | 9 | >10 | >10 | >10 | 0.359 | 0.091 | 0.014 | 0.0 |
| | 19 | >10 | >10 | >10 | 1.796 | 0.038 | 0.021 | 0 |
| | 49 | <-10.0 | <-10.0 | <-10.0 | >10 | 0.027 | 0.01 | -0.002 |
| | 99 | >10 | >10 | >10 | >10 | 0.102 | -0.002 | -0.002 |
| Copnet | 1 | 0.835 | 0.81 | 0.688 | 0.397 | 0.111 | 0.0 | -0.022 |
| | 2 | 0.819 | 0.786 | 0.676 | 0.377 | 0.106 | -0.004 | -0.017 |
| | 4 | 0.9 | 0.864 | 0.747 | 0.469 | 0.193 | 0.087 | 0.079 |
| | 7 | 1.297 | 1.268 | 1.135 | 0.903 | 0.649 | 0.571 | 0.575 |
| | 9 | 1.703 | 1.679 | 1.558 | 1.357 | 1.106 | 1.054 | 1.053 |
| | 19 | 4.862 | 4.834 | 4.75 | 4.574 | 4.431 | 4.428 | 4.432 |
| | 49 | >10 | >10 | >10 | >10 | >10 | >10 | >10 |
| | 99 | >10 | >10 | >10 | >10 | >10 | >10 | >10 |

**TE Estimation and Optimization** - channel capacity estimation of Gaussian moving average (GMA) with delay 100 – with sufficient context $l$, TREET is the best estimator.

| $l$ | TREET | | DINE | |
|---|---|---|---|---|
| | Estimated capacity [nat] | Absolute error (%) | Estimated capacity [nat] | Absolute error (%) |
| 60 | 0.19 | 53 | **0.30** | **25** |
| 70 | 0.29 | 29 | **0.35** | **15** |
| 80 | 0.28 | 31 | **0.34** | **17** |
| 90 | 0.29 | 28 | **0.30** | **26** |
| 100 | **0.37** | **9** | 0.36 | 12 |
| 110 | **0.38** | **7** | 0.33 | 20 |
| 120 | **0.36** | **11** | 0.33 | 18 |
| 130 | **0.36** | **11** | 0.34 | 17 |
| 140 | **0.35** | **14** | 0.26 | 35 |

**Feature Analysis with TREET** – Sleep Apnea patients dataset feature analysis. TREET can identify the process that transfer most of the information. For Apnea patients, breathing is affecting the heart rate in contrast to healthy patients.




**Continuous Time Probability Density Estimation with TREET** – hidden Markov model (HMM) without (1) and with(2) state delay. As out-of-the-box estimator, TE is consistent probability estimator and robust for memory length interaction.

| Model | Gaussian | | Uniform | |
|---|---|---|---|---|
| | KLD | TV | KLD | TV |
| Kalman | 1.076 | 0.467 | 1.304 | 0.408 |
| KDE | 1.135 | 0.608 | 0.847 | 0.417 |
| MDN | 1.098 | 0.632 | 0.667 | 0.460 |
| DINE | **0.795** | <u>0.525</u> | **0.475** | **0.291** |
| TREET | <u>0.797</u> | **0.524** | <u>0.482</u> | <u>0.296</u> |

| $k$ | TREET | | DINE | | MDN | | Kalman | |
|---|---|---|---|---|---|---|---|---|
| | KL | TV | KL | TV | KL | TV | KL | TV |
| 2 | <u>0.933</u> | <u>0.569</u> | 0.934 | 0.570 | 1.460 | 0.708 | **0.636** | **0.405** |
| 5 | <u>1.036</u> | 0.601 | **0.875** | **0.556** | 1.454 | 0.707 | 1.284 | <u>0.599</u> |
| 10 | **0.984** | **0.585** | 1.299 | 0.662 | 1.450 | 0.706 | <u>1.239</u> | <u>0.603</u> |
| 15 | **0.992** | **0.587** | 1.441 | 0.704 | 1.448 | 0.707 | <u>1.232</u> | <u>0.602</u> |
| 25 | **0.993** | **0.587** | 1.451 | 0.706 | 1.443 | 0.704 | <u>1.196</u> | <u>0.598</u> |
| 50 | **0.993** | **0.589** | 1.452 | 0.707 | 1.453 | 0.707 | <u>1.193</u> | <u>0.597</u> |