Deep Spectral Ranking

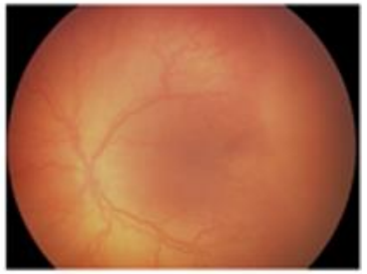
Presenter: İlkay Yıldız, yildizi@ece.neu.edu

İlkay Yıldız[1], Jennifer Dy[1], Deniz Erdoğmuş[1], Susan Ostmo[2], J. Peter Campbell[2], Michael F. Chiang[3], Stratis Ioannidis[1]

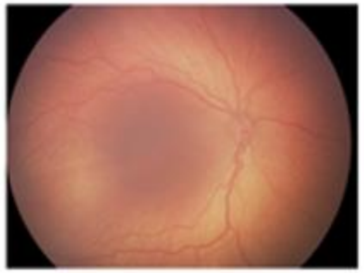
1. Electrical and Computer Engineering Dept., Northeastern University, Boston, MA, USA
2. Casey Eye Institute, Oregon Health and Science University, Portland, OR, USA
3. National Eye Institute, National Institutes of Health, Bethesda, MD, USA

Motivation

- Ranking observations are classic in many domains.



>



More diseased *image*

>

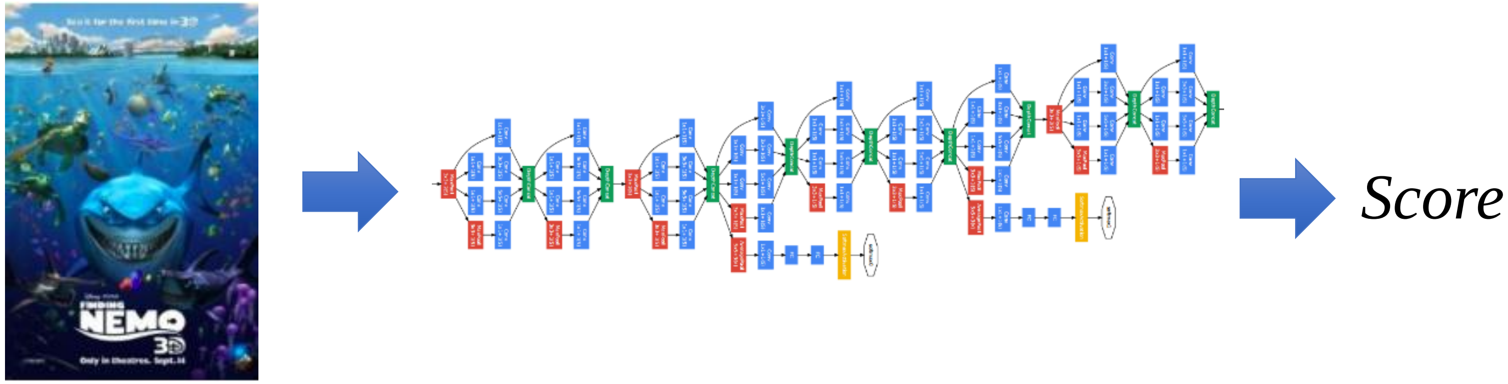
>

Better *Movie*

- **<u>Inference:</u>** Plackett-Luce Model
  - Each sample has a positive *score*.
  - Probability that a sample is ranked higher is proportional to this score.

Motivation

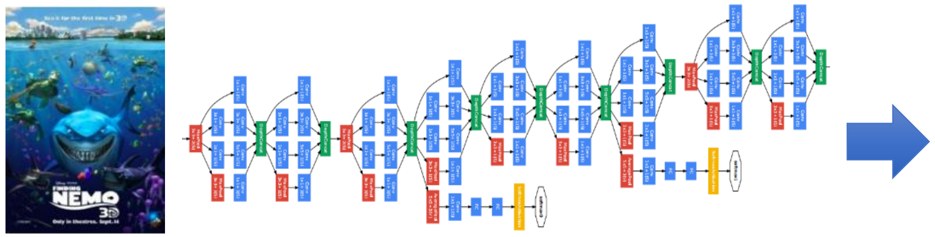
- **<u>Regression</u>**
  - **<u>Shallow:</u>** Newton's method *[Tian et al., 2019]*
  - **<u>DNN:</u>** Siamese network attains 0.92 AUC from 80 samples! *[Yıldız et al., 2019]*



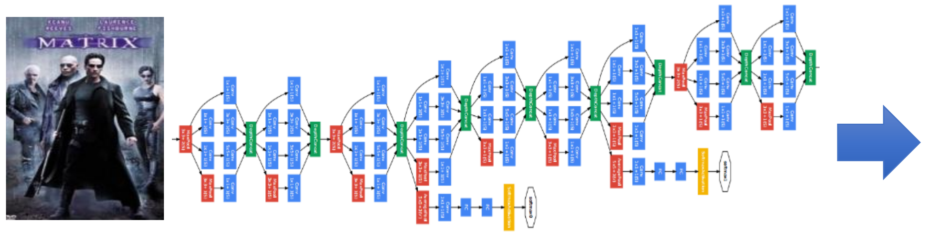
[1] Tian, P., Guo, Y., ..., Chiang, M. F., Dy, J., Erdogmus, D., and Ioannidis, S. (2019). A severity score for Retinopathy of Prematurity. SIGKDD.
[2] Yıldız, I., Tian, P., Dy, J., Erdogmus, D., Brown, J., ..., Chiang, M. F., and Ioannidis, S. (2019). Classification and comparison via neural networks. Neural Networks.

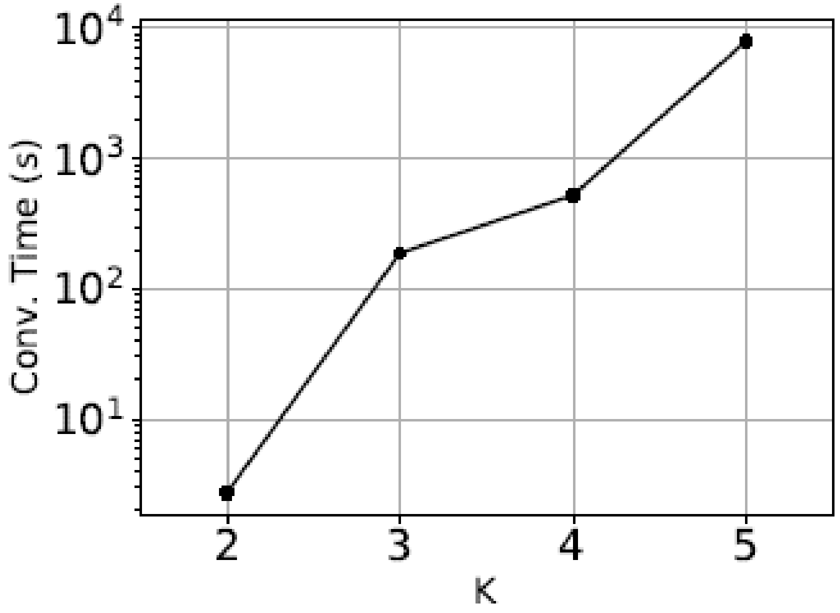
Challenges

- Traditional DNN method: *siamese architecture* with K identical base networks
- Large –and potentially variable– memory footprint



K samples ⇔ K networks





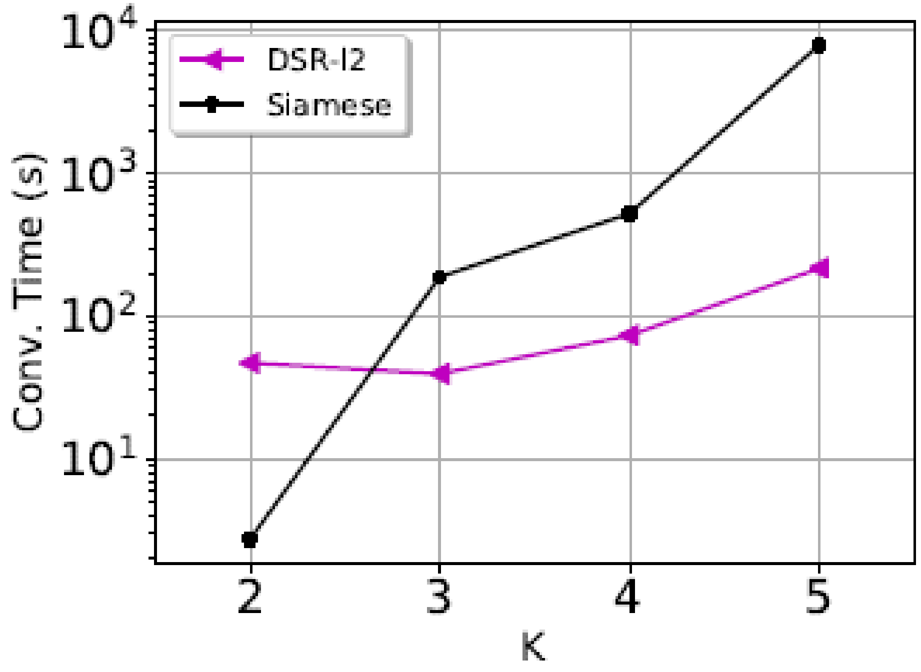
- An SGD epoch over $\binom{n}{K} = O(n^K)$ observations
  - *Exponential* training time

Challenges

- A *spectral* algorithm for shallow regression

[Yıldız et al., 2020]

- Solve via *ADMM* → Scores form the stationary distribution of a *Markov Chain*.



- Does not generalize well to deep models!
  - *Scores* form a *distribution* ↔ ℓ_2 penalty leads to vanishing gradients and stationary points!

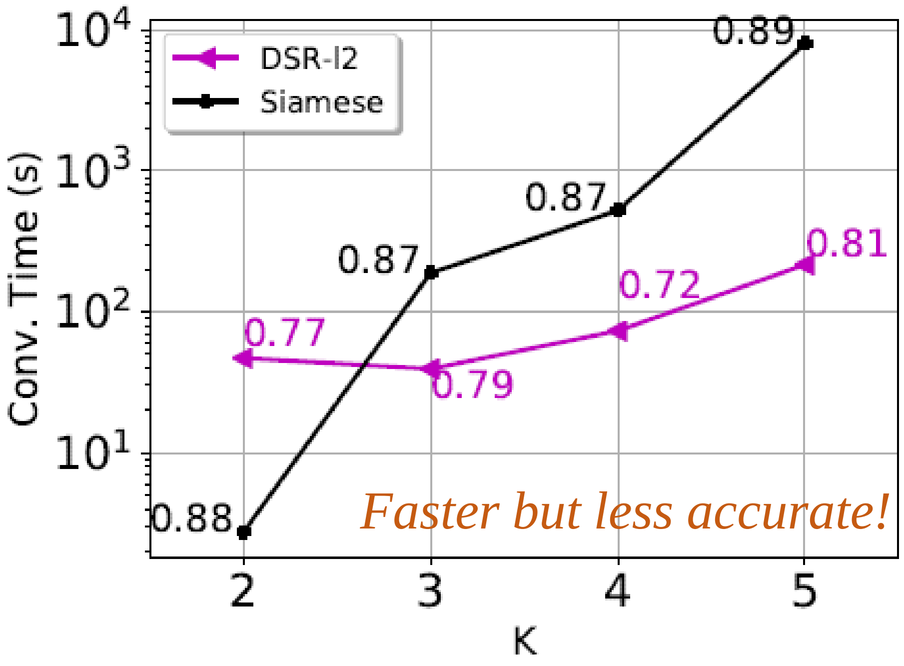
[3] Yıldız, I., Dy, J., Erdogmus, D., ..., Chiang, M. F., and Ioannidis, S. (2020). Fast and accurate ranking regression. AISTATS.

Challenges

- A *spectral* algorithm for shallow regression

[Yıldız et al., 2020]

  - Solve via *ADMM* → Scores form the stationary distribution of a *Markov Chain*.



- Does not generalize well to deep models!
  - *Scores* form a *distribution* ↔ ℓ_2 penalty leads to vanishing gradients and stationary points!

[3] Yıldız, I., Dy, J., Erdogmus, D., ..., Chiang, M. F., and Ioannidis, S. (2020). Fast and accurate ranking regression. AISTATS.

Contributions

- Bridge the gap between *DNN* models and *spectral* algorithms for ranking regression
- Replace ℓ_2-penalty of ADMM with *KL divergence*: still amenable to a spectral method!
- Significantly outperform ℓ_2-penalty ADMM *and* siamese network

Problem Formulation

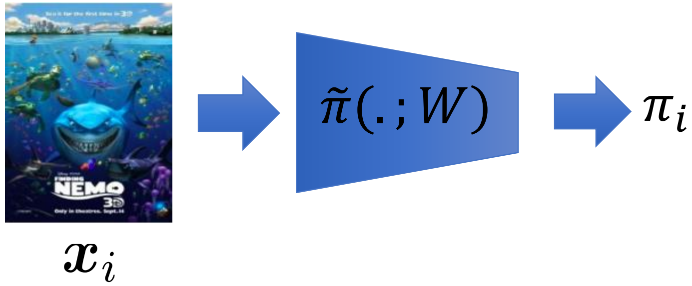
Plackett-Luce Model

- For each sample $i \in \mathcal{N}$, there exists a score $\pi_i \in \mathbb{R}_+$.
- Given query set of alternatives: $A_\ell \subseteq \mathcal{N}$
- m independent choice observations: $c_\ell \in A_\ell,\ \ell \in \mathcal{M}$

$$\boxed{\mathbf{P}(c_\ell \mid A_\ell, \boldsymbol{\pi}) = \pi_{c_\ell} / \sum_{j \in A_\ell} \pi_j = \pi_\ell / \sum_{j \in A_\ell} \pi_j}$$

Maximum Likelihood Estimation (MLE) for Ranking Regression

$$
\begin{aligned}
\text{Minimize}_{\boldsymbol{\pi}, \boldsymbol{W}} \quad & \mathcal{L}(\mathcal{D} \mid \boldsymbol{\pi}) \equiv \sum_{\ell=1}^{m} \left(\log \sum_{j \in A_\ell} \pi_j - \log \pi_\ell \right) \\
\text{subject to:} \quad & \boldsymbol{\pi} = \tilde{\boldsymbol{\pi}}(\boldsymbol{X}; \boldsymbol{W}) = [\tilde{\pi}_i = \tilde{\pi}(\boldsymbol{x}_i; \boldsymbol{W})]_{i \in \mathcal{N}} \\
& \boldsymbol{\pi} \geq \mathbf{0},
\end{aligned}
$$



- Traditional siamese architecture has base network: $\tilde{\pi}(\cdot\,; W)$.

Alternating Direction Method of Multipliers (ADMM) with generalized penalty

$$\begin{aligned} \text{Minimize}_{\boldsymbol{\pi}, \boldsymbol{W}} \quad & \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\pi}) \equiv \sum_{\ell=1}^{m} \left(\log \sum_{j \in A_\ell} \pi_j - \log \pi_\ell \right) \\ \text{subject to:} \quad & \boldsymbol{\pi} = \tilde{\boldsymbol{\pi}}(\boldsymbol{X}; \boldsymbol{W}), \quad \boldsymbol{\pi} \geq \mathbf{0}, \end{aligned}$$

- **ADMM:** Decouple optimization of scores and parameters

$$L_\rho(\boldsymbol{\pi}, \boldsymbol{W}, \boldsymbol{y}) = \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\pi}) + \boldsymbol{y}^\top (\boldsymbol{\pi} - \tilde{\boldsymbol{\pi}}(\boldsymbol{X}; \boldsymbol{W})) + \rho \cdot D_p(\boldsymbol{\pi} || \tilde{\boldsymbol{\pi}}(\boldsymbol{X}; \boldsymbol{W}))$$

$$\boldsymbol{\pi}^{k+1} = \underset{\boldsymbol{\pi} \in \mathbb{R}_+^n}{\arg\min} L_\rho(\boldsymbol{\pi}, \boldsymbol{W}^k, \boldsymbol{y}^k)$$

Efficient spectral approach over the exponential ranking data!

$$\boldsymbol{W}^{k+1} = \underset{\boldsymbol{W} \in \mathbb{R}^{d'}}{\arg\min} \; \rho D_p\left(\boldsymbol{\pi}^{k+1} || \tilde{\boldsymbol{\pi}}(\boldsymbol{X}; \boldsymbol{W})\right) - \boldsymbol{y}^{k\top} \tilde{\boldsymbol{\pi}}(\boldsymbol{X}; \boldsymbol{W})$$

Linear in number of samples!

$$\boldsymbol{y}^{k+1} = \boldsymbol{y}^k + \rho(\boldsymbol{\pi}^{k+1} - \tilde{\boldsymbol{\pi}}(\boldsymbol{X}; \boldsymbol{W}^{k+1}))$$

Dual variable update

Scores are still amenable to a spectral solution!

$$\frac{\partial L_\rho(\boldsymbol{\pi}, W^k, y^k)}{\partial \pi_i} = 0 \ \forall i \iff \sum_{j \neq i} \pi_j \lambda_{ji}(\boldsymbol{\pi}) - \sum_{j \neq i} \pi_i \lambda_{ij}(\boldsymbol{\pi}) = \pi_i \sigma_i, \qquad (15)$$

Theorem 4.2. *Eq. (15) are the balance equations of a continuous-time MC with transition rates:*

$$\mu_{ji}(\boldsymbol{\pi}) = \begin{cases} \lambda_{ji}(\boldsymbol{\pi}) + \frac{2\pi_i\sigma_i\sigma_j}{\sum_{t\in\mathcal{N}_-}\pi_t\sigma_t - \sum_{t\in\mathcal{N}_+}\pi_t\sigma_t} \\ \qquad \text{if } j \in \mathcal{N}_+ \text{ and } i \in \mathcal{N}_- \\ \lambda_{ji}(\boldsymbol{\pi}) \quad \text{otherwise,} \end{cases} \qquad (16)$$

Stationary scores are also the stationary distribution of the continuous time MC.

$$\sigma_i(\boldsymbol{\pi}) = \rho \frac{\partial D_p(\boldsymbol{\pi} || \tilde{\boldsymbol{\pi}}^k)}{\partial \pi_i} + y_i^k$$

ILSRX: $\boxed{\pi^{l+1} = \mathrm{ssd}\left(M(\pi^l)\right)}$

ADMM + KL proximal penalty

l_2 proximal penalty: $D_p(\boldsymbol{\pi}||\tilde{\boldsymbol{\pi}}) = \| \tilde{\boldsymbol{\pi}} - \boldsymbol{\pi} \|_2^2$

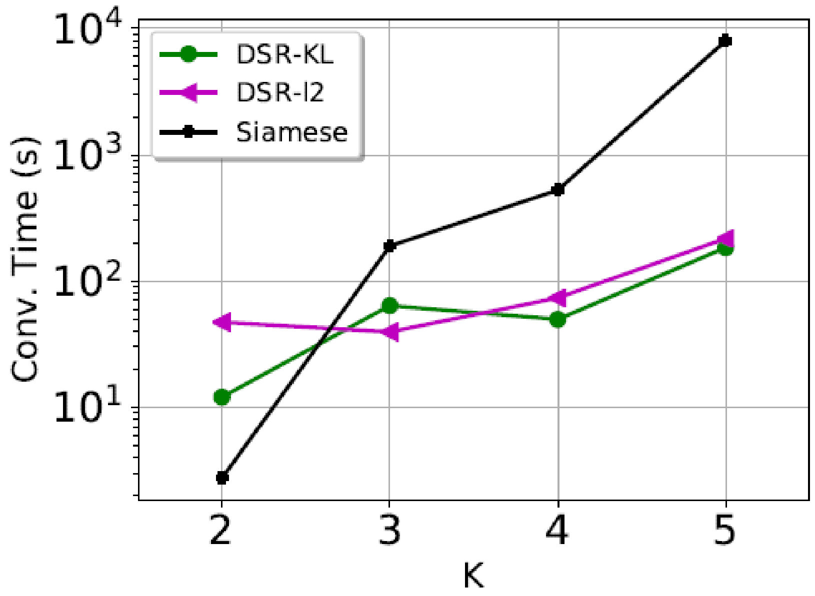
$$\boldsymbol{W}^{k+1} = \underset{\boldsymbol{W}\in\mathbb{R}^{d'}}{\arg\min} \ \| \tilde{\boldsymbol{\pi}}(\boldsymbol{X};\boldsymbol{W}) - (\boldsymbol{\pi}^{k+1} + \tfrac{1}{\rho}\boldsymbol{y}^k) \|_2^2$$

KL proximal penalty: $D_p(\boldsymbol{\pi}||\tilde{\boldsymbol{\pi}}) = \sum_{i=1}^n \pi_i \log \frac{\pi_i}{\tilde{\pi}_i}$

$$\boldsymbol{W}^{k+1} = \underset{\boldsymbol{W}\in\mathbb{R}^{d'}}{\arg\min} \ \sum_{i=1}^n (-\tfrac{y_i^k}{\rho}\tilde{\pi}_i - \pi_i^{k+1} \log \tilde{\pi}_i)$$

- ℓ_2 loss: Prone to vanishing gradients and reaching stationary points.

- Naturally suited as $\boldsymbol{\pi} = [\pi_i]_{i\in[n]} \in \mathbb{R}_+^n$ is a distribution.
- Max-entropy loss: better fitting and convergence



ADMM + KL penalty

l_2 proximal penalty: $D_p(\boldsymbol{\pi}||\tilde{\boldsymbol{\pi}}) = \|\tilde{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2^2$

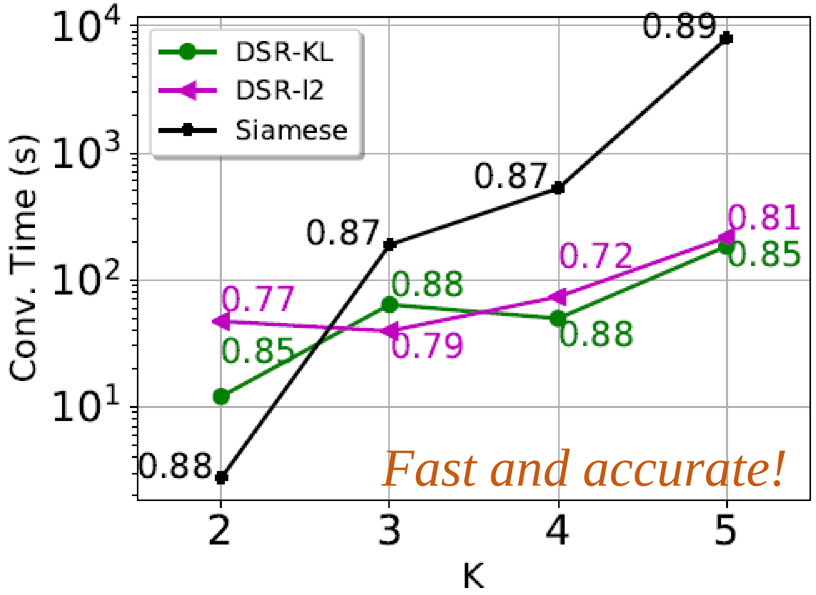
$$\boldsymbol{W}^{k+1} = \underset{\boldsymbol{W}\in\mathbb{R}^{d'}}{\arg\min} \ \|\tilde{\boldsymbol{\pi}}(\boldsymbol{X};\boldsymbol{W}) - (\boldsymbol{\pi}^{k+1} + \tfrac{1}{\rho}\boldsymbol{y}^k)\|_2^2$$

- ℓ_2 loss: Prone to vanishing gradients and reaching stationary points.

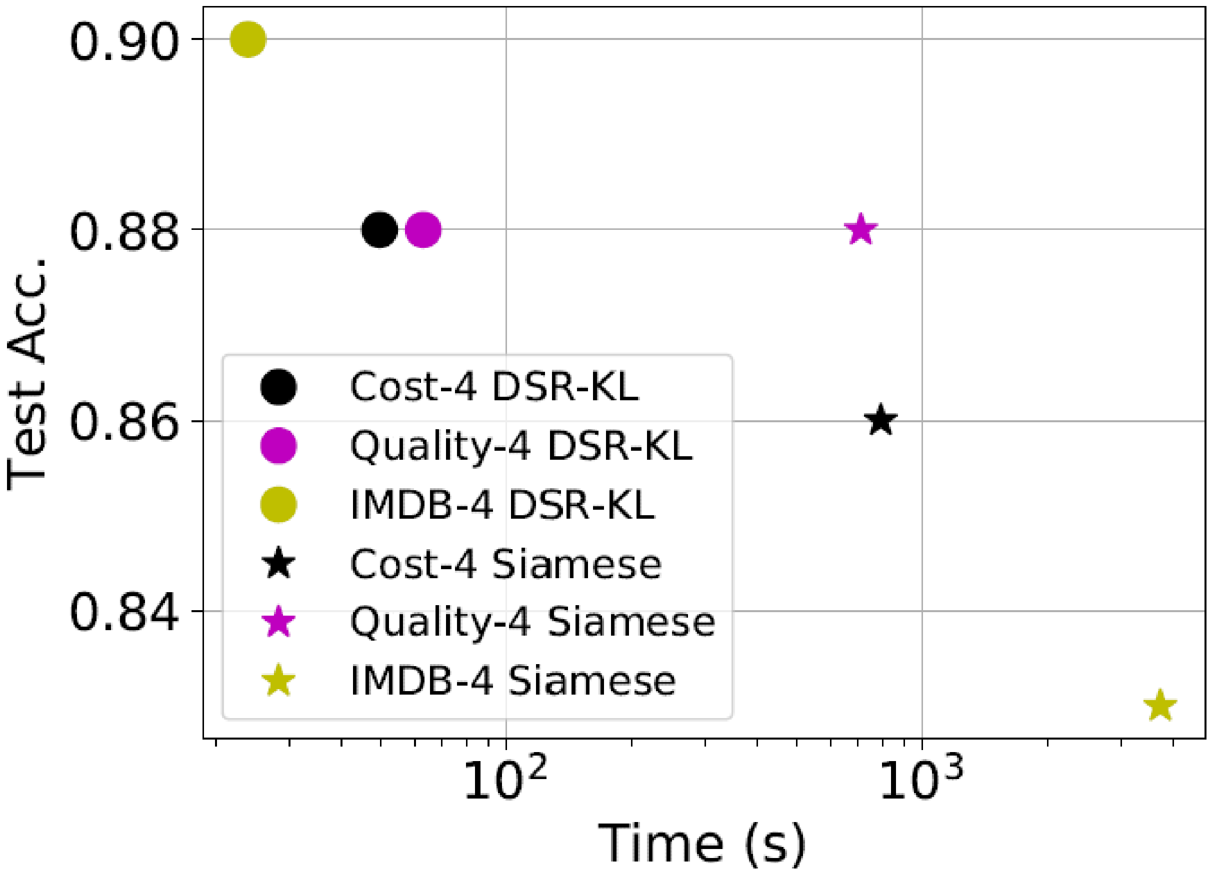
KL proximal penalty: $D_p(\boldsymbol{\pi}||\tilde{\boldsymbol{\pi}}) = \sum_{i=1}^n \pi_i \log \frac{\pi_i}{\tilde{\pi}_i}$

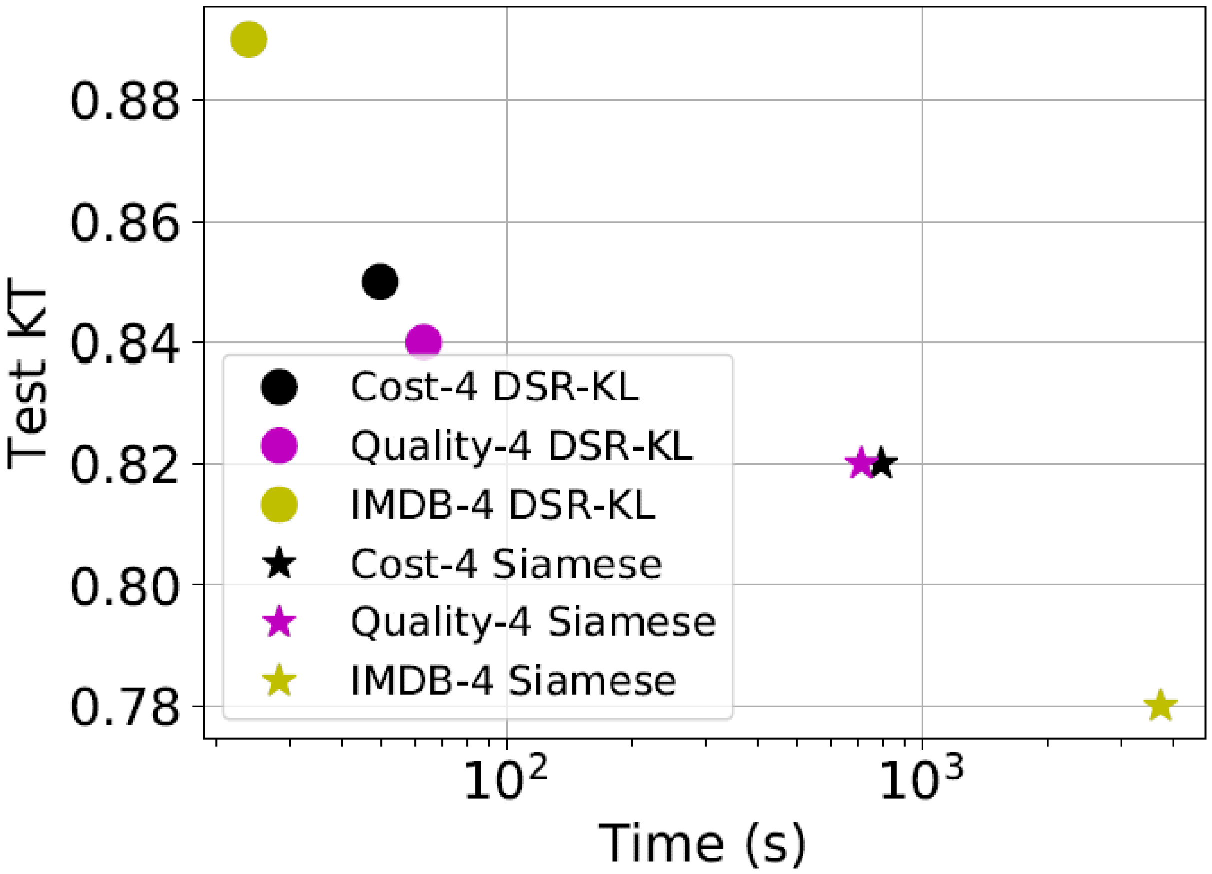
$$\boldsymbol{W}^{k+1} = \underset{\boldsymbol{W}\in\mathbb{R}^{d'}}{\arg\min} \ \sum_{i=1}^n \left(-\frac{y_i^k}{\rho}\tilde{\pi}_i - \pi_i^{k+1}\log\tilde{\pi}_i\right)$$

- Naturally suited as $\boldsymbol{\pi} = [\pi_i]_{i\in[n]} \in \mathbb{R}_+^n$ is a distribution.
- Max-entropy loss: better fitting and convergence



Faster Training AND Better Predictions!





THANK YOU!

Presenter: İlkay Yıldız, yildizi@ece.neu.edu

Supported by NIH (R01EY019474), NSF (SCH-1622542 at MGH, SCH-1622536 at Northeastern, SCH-1622679 at OHSU), Facebook Statistics Research Award, and by unrestricted departmental funding from Research to Prevent Blindness (OHSU)