Teaching Economics to the Machines

Hui Chen

MIT Sloan and NBER

joint with Yuhan Cheng, Yanchu Liu, and Ke Tang

August 25, 2023

Outline

“The End of Theory

Chris Anderson, Editor in Chief of Wired, 2008

“All models are wrong, but some are useful.” So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don’t have to settle for wrong models. Indeed, they don’t have to settle for models at all.

How to teach a robot to play pool?

How should we teach a robot to play pool?

- Data-driven approach:
  - ↪ x: shot angle, position, force ...
  - ↪ y: ball direction, speed, spin ...
  - ↪ However, labeled data are often expensive to collect.

How should we teach a robot to play pool?

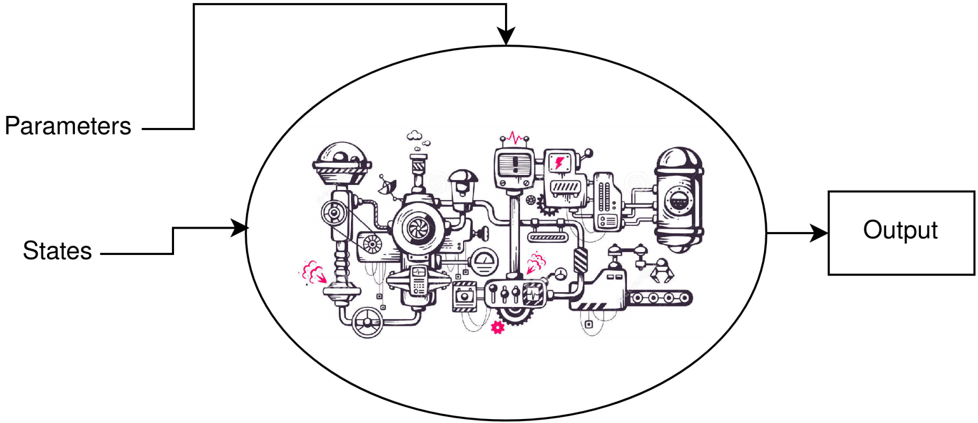
- Data-driven approach:
  - $\hookrightarrow$ x: shot angle, position, force ...
  - $\hookrightarrow$ y: ball direction, speed, spin ...
  - $\hookrightarrow$ However, labeled data are often expensive to collect.
- Theory-driven approach:
  - $\hookrightarrow$ Model of elastic collisions
  - $\hookrightarrow$ Conservation of linear momentum and kinetic energy $\Rightarrow$ complete predictions for ball movements
  - $\hookrightarrow$ Model could be misspecified (inelastic collision, imperfect surface ...)

How should we teach a robot to play pool?

- Data-driven approach:
  - ↪ x: shot angle, position, force ...
  - ↪ y: ball direction, speed, spin ...
  - ↪ However, labeled data are often expensive to collect.
- Theory-driven approach:
  - ↪ Model of elastic collisions
  - ↪ Conservation of linear momentum and kinetic energy $\Rightarrow$ complete predictions for ball movements
  - ↪ Model could be misspecified (inelastic collision, imperfect surface ...)
- How to combine the knowledge from theory and data?
  - ↪ Our answer: transfer learning framework
  - ↪ First learn from theory, then from real data.

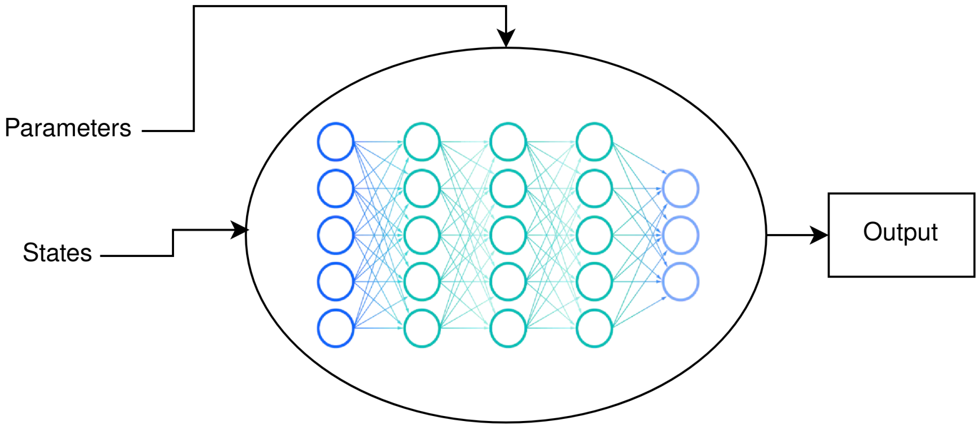
Outline

Deep surrogate: “look-up table” in the age of AI



$$f\,\underbrace{(\,s_t|\,\theta)}_{\mathbf{x}_t} = y_t$$

Deep surrogate: “look-up table” in the age of AI



$$\phi(\mathbf{x}_t|\theta_{NN}) = y_t$$

Learning from theory

Chen, Didisheim and Scheidegger (2022)

- Consider a set of economic restrictions represented by $y = F(s|\theta)$. Using the augmented state $x \equiv (s, \theta)$ and define
$$y = f(x) \equiv F(s|\theta)$$
- By imposing a hierarchical prior on θ, we can derive a model-implied joint probability distribution of x and y, $\mathbb{P}_{(\mathscr{X},\mathscr{Y})}$.
- Generate a training set $S = ((x_i, y_i))_{i=1}^{m}$ of iid samples drawn from $\mathscr{X} \times \mathscr{Y}$; for a given a loss function $\mathscr{L}$, find a function from the set $\mathscr{H}$ of deep neural networks to minimize the empirical loss of approximation:

$$\hat{f}_{\mathscr{H},S} = \arg\min \left\{ \frac{1}{m} \sum_{i=1}^{m} \mathscr{L}(f(x_i), y_i) : f \in \mathscr{H} \right\}$$

Learning from theory

Chen, Didisheim and Scheidegger (2022)

- What does “learning from theory” mean? Inherit the structural restrictions.

Learning from theory

Chen, Didisheim and Scheidegger (2022)

- What does "learning from theory" mean? Inherit the structural restrictions.
- The advantages of deep surrogates:
  - ↪ Knowledge of true DGP: Generate as much training data as desired with (essentially) no error.
  - ↪ Expressivity: Universal approximation theorem for shallow and deep networks (Hornik, Stinchcombe, and White 1989, Hanin and Sellke 2017, Lu et al. 2017).
  - ↪ Curse of dimensionality: With suitable target function and activation function, can train accurate surrogate with sample sizes that grow polynomially (vs. exponentially) in the dimensionality of the model (Berner et al. 2020).

Performance gains

Example: Option pricing model under double-exponential jump-diffusion with stochastic volatility

- SPX: ~ 4000 option prices on a typical day
- Horse race:
  - ↪ FFT implementation vs. deep surrogate
  - ↪ Single core vs. GPU

	FFT	Deep Surrogate	Deep Surrogate + GPU
pricing, 1-day	10s		
estimation, 1-day	180s		
estimation, 1-year	125h		

Performance gains

Example: Option pricing model under double-exponential jump-diffusion with stochastic volatility

- SPX: ~ 4000 option prices on a typical day
- Horse race:
  - ↪ FFT implementation vs. deep surrogate
  - ↪ Single core vs. GPU

	FFT	Deep Surrogate	Deep Surrogate + GPU
pricing, 1-day	10s	0.6s	
estimation, 1-day	180s	3.2s per iter	
estimation, 1-year	125h	2.2h	

Performance gains

Example: Option pricing model under double-exponential jump-diffusion with stochastic volatility

- SPX: ~ 4000 option prices on a typical day
- Horse race:
  - ↪ FFT implementation vs. deep surrogate
  - ↪ Single core vs. GPU

	FFT	Deep Surrogate	Deep Surrogate + GPU
pricing, 1-day	10s	0.6s	0.06s
estimation, 1-day	180s	3.2s per iter	0.3s
estimation, 1-year	125h	2.2h	0.2h

Outline

Teaching economics to the machines

- The neural networks trained on the simulated data from a theoretical model inherit the structural restrictions from the theory.
- Valid restrictions can help regulate learning from real data.
  - $\hookrightarrow$ Bias-variance tradeoff
- The idea: Do not strictly impose the theoretical restrictions. Let them guide/regulate the training of the ML model on real data.
- Potential benefits: Variance reduction, faster training on real data, less demand for real data ...

Two related approaches

- Structural estimation: Impose equilibrium conditions when estimating structural parameters from the data.
  - ↪ Rational expectations econometrics (Saracoglu and Sargent 1978, Hansen and Sargent 1980)

- Bayesian approach: Deriving informative priors from an economic model to estimate VARs for macro variables.
  - ↪ Random walk (Doan, Litterman, and Sims 1984)
  - ↪ DSGE models (Ingram and Whiteman 1994, Del Negro and Schorfheide 2004)

Two related approaches

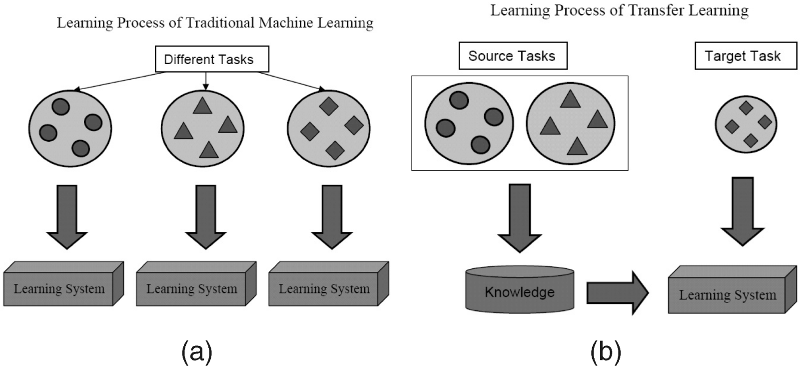
- Structural estimation: Impose equilibrium conditions when estimating structural parameters from the data.
  - Rational expectations econometrics (Saracoglu and Sargent 1978, Hansen and Sargent 1980)
- Bayesian approach: Deriving informative priors from an economic model to estimate VARs for macro variables.
  - Random walk (Doan, Litterman, and Sims 1984)
  - DSGE models (Ingram and Whiteman 1994, Del Negro and Schorfheide 2004)
- Del Negro and Schorfheide (2004): DSGE-VAR
  - simulate time-series data from a DSGE model (with a hierarchical prior on the DSGE model parameters);
  - fit a VAR to the simulated data to form a prior for the VAR parameters;
  - derive the posterior using real data.

Two related approaches

- Structural estimation: Impose equilibrium conditions when estimating structural parameters from the data.
  - ↪ Rational expectations econometrics (Saracoglu and Sargent 1978, Hansen and Sargent 1980)
- Bayesian approach: Deriving informative priors from an economic model to estimate VARs for macro variables.
  - ↪ Random walk (Doan, Litterman, and Sims 1984)
  - ↪ DSGE models (Ingram and Whiteman 1994, Del Negro and Schorfheide 2004)
- Del Negro and Schorfheide (2004): DSGE-VAR
  - ↪ simulate time-series data from a DSGE model (with a hierarchical prior on the DSGE model parameters);
  - ↪ fit a VAR to the simulated data to form a prior for the VAR parameters;
  - ↪ derive the posterior using real data.
- Our contribution: “Generalizing” the above approach to nonlinear models via transfer learning. But it is not a posterior (an average of the simulated and real data).

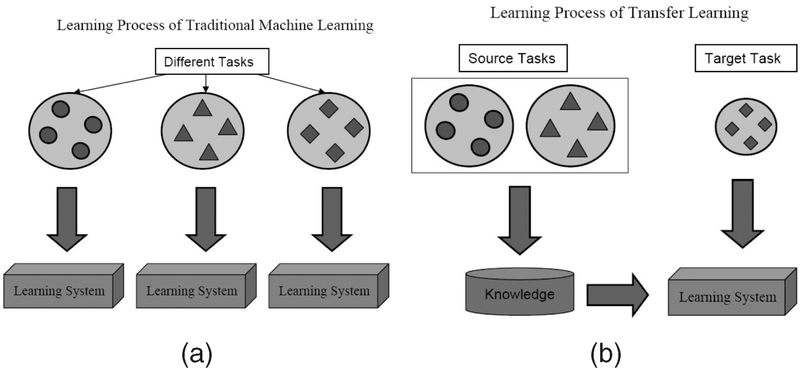
Transfer learning

- Transfer Learning: Using pre-trained models on large datasets (source domain) to accelerate learning on a smaller dataset (source domain) for a specific task.
  - ↪ computer vision: pre-train CNNs on ImageNet ⇒ medical image analysis
  - ↪ natural language processing: pre-trained LLMs ⇒ sentiment analysis



Transfer learning

- Transfer Learning: Using pre-trained models on large datasets (source domain) to accelerate learning on a smaller dataset (source domain) for a specific task.
  - ↪ computer vision: pre-train CNNs on ImageNet ⇒ medical image analysis
  - ↪ natural language processing: pre-trained LLMs ⇒ sentiment analysis
- In our setting: Source domain = economic model; target domain = actual data.



Source domain

- Denote a feed-forward neural network with L layers as $\Phi(\sigma_1, \cdots, \sigma_L; W_1, \cdots, W_L)$.
  - $\hookrightarrow$ σ_i: activation function; W_i: weights
- Draw training sample $\big((x_i, f(x_i))\big)_{i=1}^m$ from the model-implied joint distribution, $\mathbb{P}_{(\mathscr{X},\mathscr{Y})}$.
- In the source domain, we train the neural networks to learn from data generated by the theoretical model:

$$\widehat{W_1}, \widehat{W_2}, \dots, \widehat{W_L} = \arg\min \lambda_0 \mathscr{L}_0 + \sum_j \lambda_j \mathscr{L}_j$$

- In the example for option pricing, use weighted MAE for option prices and Greeks in the loss function.

$$\mathscr{L}_0 = \sum_{i=1}^N w_i |\Phi(x_i) - f(x_i)|$$

$$\mathscr{L}_j = \sum_{i=1}^N w_{ij} \left| \frac{\partial \Phi(x_i)}{\partial x_{ij}} - \frac{\partial f(x_i)}{\partial x_{ij}} \right|$$

Target domain

- Training in the target domain on the real data $\big((x_i, y_i)\big)_{i=1}^{n}$. It follows a fine-tuning method with lower learning rate and smaller number of epochs.
- We inherit the same network architecture from the source domain and use the weights from the source domain as the starting point,

$$\widetilde{W}_1, \cdots, \widetilde{W}_L = \arg\min \sum_{i=1}^{N} w_i \left| \Phi(\sigma_1, \cdots, \sigma_L; W_1, ..., W_L)(x_i) - y_i \right|$$

with initial values $\widehat{W}_1, \widehat{W}_2, \cdots, \widehat{W}_L$.

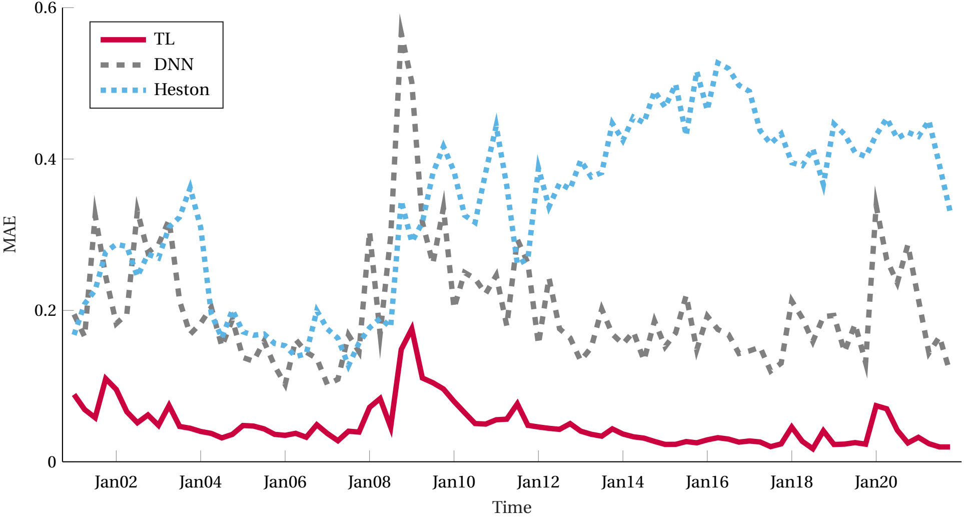
- Alternative approaches:
  - ↪ replacing part of the network from the source domain with new layers;
  - ↪ fixing early part of the network (frozen layers)

Outline

Example application

- A horse race between theory-based transfer learning (TL) and standard deep neural networks (DNN) for option pricing.
  - ↪ Use the Black-Scholes model in the source domain.
  - ↪ Augment the set of features with past returns, bid-ask spreads, put-call ratios ... (total of $d = 16$ features)
- Architecture: 18-layer ResNet, 22 neurons in each hidden layer, LeakyRelu activation
- At the end of each quarter, use data from the past 9 months to train the models; use the models to "predict" option prices in the next quarter.
  - ↪ Mimicking a setting with limited labeled data in the target domain.
- Other candidate models: Heston model, multi-target DNN

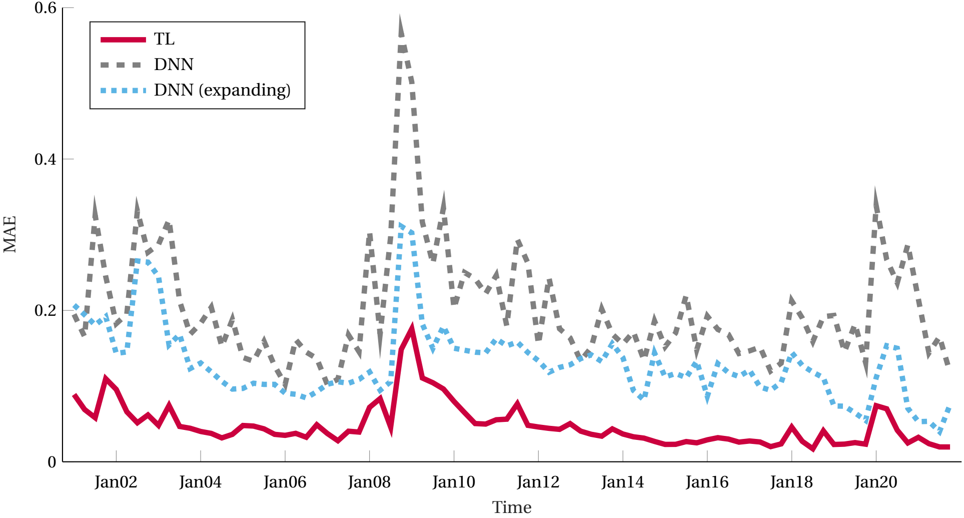
TL outperforms DNN and Heston model



Performance attribution

- Regressing the relative pricing errors between TL and DL models on a variety of attributes.
- Transfer learning has a bigger advantage relative to deep learning:
  - ↪ when the market is more volatile (IV);
  - ↪ when inputs are "unusual" (measured by the Mahalanobis distance);
  - ↪ for options with lower bid-ask spreads;
  - ↪ for options with short (< 7 days) or long (> 90 days) time to maturity;
  - ↪ for ATM and ITM options.

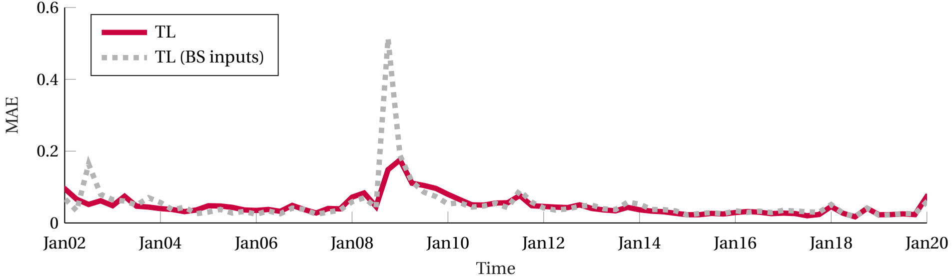
Having more data does help with DNN

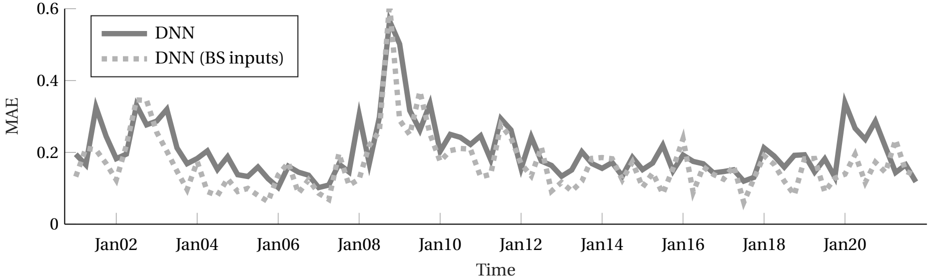


What is source of TL out-performance?

- The theory helps identify relevant inputs?
  - ⟶ Compare to DNN with the same inputs

Theory does more than identifying inputs

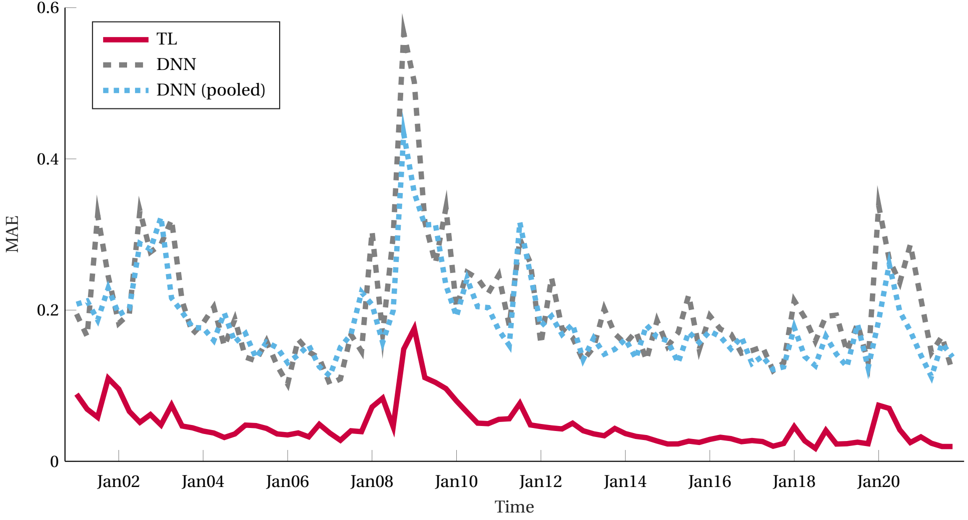




What is source of TL out-performance?

- The theory helps identify relevant inputs?
  - ↪ Compare to DNN with the same inputs
- More data?
  - ↪ Pool real data with simulated data
  - ↪ Also consistent with the Bayesian interpretation

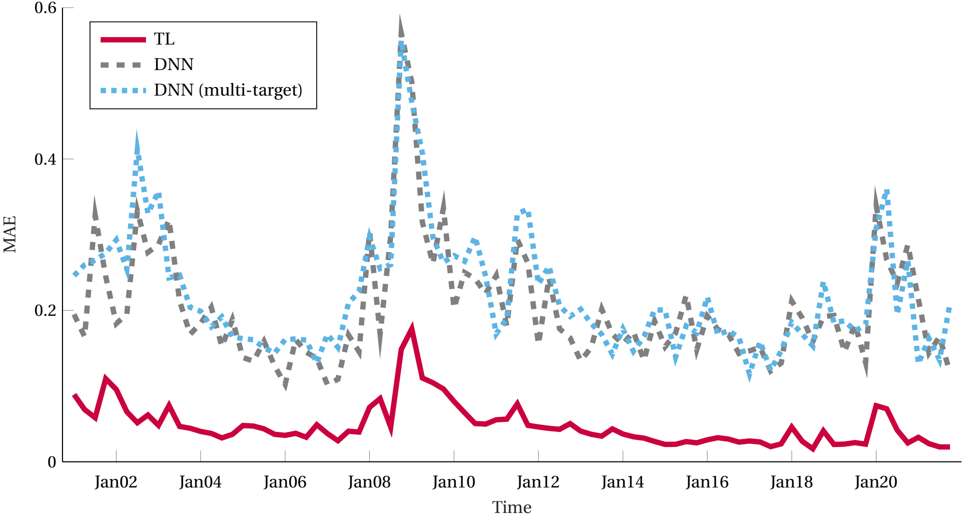
TL vs. pooling simulated and real data



What is source of TL out-performance?

- The theory helps identify relevant inputs?
  - ↪ Compare to DNN with the same inputs
- More data?
  - ↪ Pool real data with simulated data
  - ↪ Also consistent with the Bayesian interpretation
- Regularization by penalizing against deviations from the model?
  - ↪ Multi-target: Make the loss dependent on distance from data and model

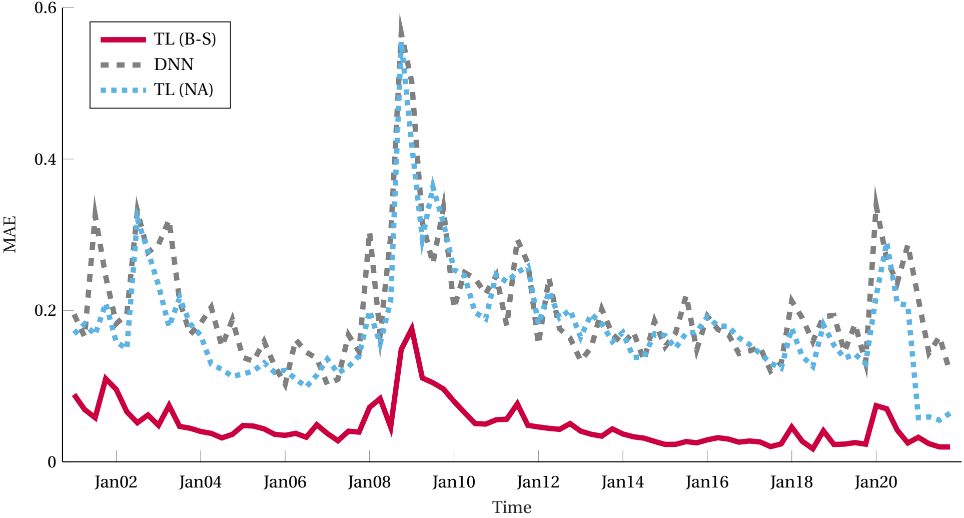
TL vs. mixing the empirical and model targets



What is source of TL out-performance?

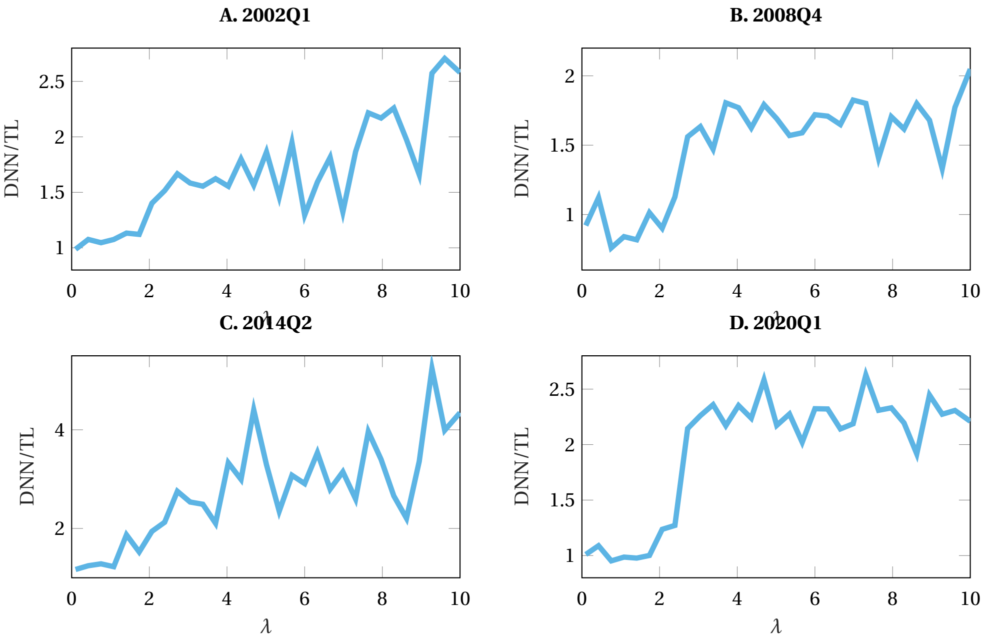
- The theory helps identify relevant inputs?
  - ↪ Compare to DNN with the same inputs
- More data?
  - ↪ Pool real data with simulated data
  - ↪ Also consistent with the Bayesian interpretation
- Regularization by penalizing against deviations from the model?
  - ↪ Multi-target: Make the loss dependent on distance from data and model
- Would any restrictions work?
  - ↪ Use no-arbitrage restrictions in the source domain

Using no-arbitrage restrictions in the source domain



Relative sample size in source domain

- When deriving informative priors from a DSGE model, the relative sample size λ of simulated data from the model effectively controls the prior precision (Del Negro and Schorfheide 2004).
- They document an inverse U-shaped relation between forecasting accuracy and λ.
- Intuition: Too little and too much weight on the DSGE model can both hurt performance.
- In our transfer learning setting, bigger sample size in the source domain improves the approximation accuracy for the theoretical model.
- However, it does not mean the weights inherited from the source domain will have a bigger influence of the final weights.


A. 2002Q1
DNN/TL
2.5
2
1.5
1
0
2
4
6
8
10
λ
B. 2008Q4
DNN/TL
2
1.5
1
0
2
4
6
8
10
λ
C. 2014Q2
DNN/TL
4
2
0
2
4
6
8
10
λ
D. 2020Q1
DNN/TL
2.5
2
1.5
1
0
2
4
6
8
10
λ

Outline

Conclusion

- Guiding ML with economic theory through transfer learning.
  - ↪ Faster training with smaller sample size.
  - ↪ Dealing with structural breaks.
  - ↪ Providing us with hints on how to improve the theories (e.g., through feature importance analysis on inputs outside of theory).
- It opens the door for a variety of applications that could benefit from the theory-data hybrid approach.

Conclusion

- Guiding ML with economic theory through transfer learning.
  - ↪ Faster training with smaller sample size.
  - ↪ Dealing with structural breaks.
  - ↪ Providing us with hints on how to improve the theories (e.g., through feature importance analysis on inputs outside of theory).

- It opens the door for a variety of applications that could benefit from the theory-data hybrid approach.

Theories are dead? Long live our theories!