Spooky Boundaries at a Distance: Exploring Transversality and Stationarity with Deep Learning

Mahdi Ebrahimi Kahou[1] Jesús Fernández-Villaverde[2] Sebastián Gómez-Cardona[1] Jesse Perla[1] Jan Rosa[1]

October 23, 2022

[1]University of British Columbia, Vancouver School of Economics

[2]University of Pennsylvania

Motivation

Motivation

- Dynamic models usually require **economic conditions** eliminating explosive solutions (e.g., transversality or no-bubble).
    - These are variations of "boundary conditions" in ODEs and PDEs on **forward-looking** behavior.
    - Deterministic, stochastic, sequential, recursive formulations all require conditions in some form.
- These forward-looking boundary conditions are the key limitation on increasing dimensionality:
    - Otherwise, in sequential setups, we can easily solve high-dimensional initial value problems.
    - In recursive models accurate solutions are required for arbitrary values of the state variables.
- **Question:** Can we avoid precisely calculating steady-state, BGP, and stationary distribution, which are never reached, and still have accurate short/medium-run dynamics disciplined by these boundary conditions?

Contribution

- Show that **deep learning** solutions to many dynamic forward-looking models automatically fulfill the long-run boundary conditions we need (transversality and no-bubble).
  - We show how to design the approximation using economic insight.
- Solve classic models with known solutions (asset pricing and neoclassical growth) and show excellent short/medium term dynamics –even when **non-stationary** or with **steady state multiplicity**.
- Suggests these methods may solve high-dimensional problems while avoiding the key computational limitation.
  - We have to understand low-dimensional problems first.
- **Intuition**: DL has an "implicit bias" toward smooth and simple functions. Explosive solutions are not smooth.

But first, what is a deep learning solution and the implicit bias?

Background: Deep learning for functional equations

Models as functional equations

Equilibrium conditions in economics can be written as functional equations:

- Take some function(s) $\psi \in \Psi$ where $\psi : X \to Y$ (e.g., optimal policy and consumption function in neoclassical growth model).
- Domain X could be state (e.g., capital) or time if sequential.
- The "model" is $\ell : \Psi \times X \to \mathcal{R}$ (e.g., Euler residuals and feasibility condition).
- The solution is the root of the model (residuals operator), i.e., $0 \in \mathcal{R}$, at each $x \in X$ (e.g., optimal policy is the root of the Euler over the space of capital).

Then a **solution** is an $\psi^* \in \Psi$ where $\ell(\psi^*, x) = 0$ for all $x \in X$.

Example: one formulation of neoclassical growth

An Example of a recursive case:

- Domain: $x = \begin{bmatrix} k \end{bmatrix}$ and $X = \mathbb{R}_+$.
- Solve for the optimal policy $k'(\cdot)$ and consumption function $c(\cdot)$: So $\psi : \mathbb{R} \to \mathbb{R}^2$ and $Y = \mathbb{R}^2_+$.
- Residuals are the Euler equation and feasibility condition, so $\mathcal{R} = \mathbb{R}^2$:

$$\ell(\underbrace{\begin{bmatrix} k'(\cdot) & c(\cdot) \end{bmatrix}}_{\equiv \psi}, \underbrace{k}_{\equiv x}) = \underbrace{\begin{bmatrix} u'(c(k)) - \beta u'(c(k'(k)))\left(f'(k'(k)) + 1 - \delta\right) \\ f(k) - c(k) - k'(k) + (1-\delta)k \end{bmatrix}}_{\text{model}}$$

- Finally, $\psi^* = [k'(\cdot), c(\cdot)]$ is a solution if it has zero residuals on domain X.

Classical solution method for functional equations

1. **Pick** finite set of N points $\hat{X} \subset X$ (e.g., a grid).
2. **Choose** approximation $\hat{\psi}(\cdot;\theta) \in \mathcal{H}(\Theta)$ with coefficients $\Theta \subseteq \mathbb{R}^M$ (e.g., Chebyshev polynomials).
3. **Fit** with nonlinear least-squares

$$\min_{\theta \in \Theta} \sum_{x \in \hat{X}} \ell(\hat{\psi}(\cdot;\theta), x)^2$$

   If $\theta \in \Theta$ is such that $\ell(\hat{\psi}(\cdot;\theta), x) = 0$ for all $x \in \hat{X}$ we say it **interpolates** $\hat{X}$.

4. The goal is to have good **generalization**:
   - The approximate function is close to the solution outside of $\hat{X}$.
   - That is $\hat{\psi}(x;\theta) \approx \psi^*(x)$ for $x \notin \hat{X}$.

A deep learning approach

- **Deep neural networks** are **highly-overparameterized** functions designed for good generalization.
  - Number of coefficients much larger than the grid points ($M \gg N$).
- Example: one layer neural network, $\hat{\psi} : \mathbb{R}^Q \to \mathbb{R}$:

$$\hat{\psi}(x;\theta) = W_2 \cdot \sigma\left(W_1 \cdot x + b_1\right) + b_2$$

- $W_1 \in \mathbb{R}^{P\times Q}$, $b_1 \in \mathbb{R}^{P\times 1}$, $W_2 \in \mathbb{R}^{1\times P}$, and $b_2 \in \mathbb{R}$.
- $\sigma(\cdot)$ is a nonlinear function applied element-wise (e.g., $\max\{\cdot, 0\}$).
- $\Theta \equiv \{b_1, W_1, b_2, W_2\}$ are the coefficients, in this example $M = PQ + P + P + 1$.
- Making it "deeper" by adding another "layer":

$$\hat{\psi}(x;\theta) \equiv W_3 \cdot \sigma(W_2 \cdot \sigma(W_1 \cdot x + b_1) + b_2) + b_3.$$

- Architecture of the neural networks can be flexibly informed by the economic insight and theory. However, not crucial for this paper.

Deep learning optimizes in a space of functions: which $\hat{\psi}$?

- Since $M \gg N$, it is possible for $\hat{\psi}$ to interpolate and the objective value will be ≈ 0.
- Since $M \gg N$ there are many solutions (e.g., θ_1 and θ_2),
  - Agree on the grid points: $\hat{\psi}(x;\theta_1) \approx \hat{\psi}(x;\theta_2)$ for $x \in \hat{X}$.
- Since individual θ are irrelevant it is helpful to think of optimization directly within $\mathcal{H}$

$$\boxed{\min_{\hat{\psi}\in\mathcal{H}} \sum_{x\in\hat{X}} \ell(\hat{\psi}, x)^2}$$

But which $\hat{\psi}$?

Deep learning and interpolation

- For M large enough, optimizers **tend to** converge to **unique** smooth and simple $\hat{\psi}$ (w.r.t to some norm $\|\cdot\|_S$). Unique both in $\hat{X}$ and X. There is a bias toward a specific class functions.
- **How to interpret:** interpolating solutions for some functional norm $\|\cdot\|_S$

$$\min_{\hat{\psi}\in\mathcal{H}} ||\hat{\psi}||_S$$
$$\text{s.t. } \ell(\hat{\psi}, x) = 0, \quad \text{for } x \in \hat{X}$$

  - CS and literature refers to this as the **inductive bias** or **implicit bias**: optimization process is biased toward particular $\hat{\psi}$
  - Small values of $\|\cdot\|_S$ corresponds to flat solutions with small gradients.
  - Characterizing $\|\cdot\|_S$ (e.g., Sobolev) is an active research area in CS at the heart of deep learning theory.

Examples

Deep learning and interpolation in practice

Reminder: in practice we solve

$$\boxed{\min_{\theta \in \Theta} \sum_{x \in \hat{X}} \ell\left(\hat{\psi}(\cdot;\theta), x\right)^2}$$

- The smooth interpolation is imposed **implicitly** through the optimization process.
- No explicit norm minimization or penalization is required.

In this paper: we describe how the $\min_{\hat{\psi} \in \mathcal{H}} ||\hat{\psi}||_S$ solutions are also the ones which automatically fulfill transversality and no-bubble conditions.

- They are disciplined by long-run boundary conditions. Therefore, we can obtain accurate short/medium-run dynamics.

Outline

To explore how we can have accurate short-run dynamics, we show deep learning solutions to

1. Classic linear-asset pricing model.
2. Sequential formulation of the neoclassical growth model.
3. Sequential neoclassical growth model with multiple steady states.
4. Recursive formulation of the neoclassical growth model.
5. Non-stationarity, such as balanced growth path.

Linear asset pricing

Sequential formulation

- Dividends, $y(t)$, y_0 as given, and follows the process:

$$y(t+1) = c + (1+g)y(t)$$

- Writing as a linear state-space model with $x(t+1) = Ax(t)$ and $y(t) = Gx(t)$ and

$$x(t) \equiv \begin{bmatrix} 1 & y(t) \end{bmatrix}^\top, A \equiv \begin{bmatrix} 1 & 0 \\ c & 1+g \end{bmatrix}, G \equiv \begin{bmatrix} 0 & 1 \end{bmatrix}$$

- "Fundamental" price given $x(t)$ is PDV with $\beta \in (0,1)$ and $\beta(1+g) < 1$

$$\boxed{p_f(t) \equiv \sum_{j=0}^{\infty} \beta^j y(t+j) = G(I - \beta A)^{-1} x(t).}$$

Recursive formulation

With standard transformation, all solutions $p_f(t)$ fulfill the recursive equations

$$p(t) = Gx(t) + \beta p(t+1) \tag{1}$$

$$x(t+1) = Ax(t) \tag{2}$$

$$0 = \lim_{T\to\infty} \beta^T p(T) \tag{3}$$

$$x_0 \text{ given} \tag{4}$$

That is, a system of two difference equations with one boundary and one initial condition.

- The boundary condition (3) is an **assumption** necessary for the problem to be well-posed and have a unique solution.
- It ensures that $p(t) = p_f(t)$ by imposing long-run boundary condition.
- But without this assumption there can be "bubbles" with $p(t) \neq p_f(t)$, only fulfilling (1) and (2).
- Intuition: system of $\{p(t), x(t)\}$ difference equations requires total of two boundaries or initial values to have a unique solution.

Solutions without no-bubble condition

Without the no-bubble condition:

- Solutions in this deterministic asset pricing model are of the form:

$$\boxed{p(t) = p_f(t) + \zeta\ \beta^{-t}.} \tag{5}$$

- For any $\zeta \geq 0$. The initial condition $x(0)$ determines $p_f(t)$.
- There are infinitely many solutions.
- The no-bubble condition chooses $\zeta = 0$.

Interpolation problem: without no-bubble condition

- A set of points in time $\hat{X} = \{t_1, \ldots, t_{\max}\}$.
- A family of over-parameterized functions $p(\cdot;\theta) \in \mathcal{H}(\Theta)$.
- Generate $x(t)$ using the law of motion and $x(0)$, equation (2).

  In practice we minimize the residuals of the recursive form for the price:

$$\boxed{\min_{\theta\in\Theta} \frac{1}{|\hat{X}|} \sum_{t\in\hat{X}} \left[p(t;\theta) - Gx(t) - \beta p(t+1;\theta)\right]^2} \tag{6}$$

- This minimization **does not contain** no-bubble condition. It has infinitely many minima.
- Does the implicit bias of over-parameterized interpolation weed out the bubbles? **Yes**.
- **Intuition**: bubble solutions are explosive, i.e., big functions with big derivatives.

  Let's analyze this more rigorously.

Interpolation formulation: min-norm mental model

The min-norm **mental model** can be written as:

$$
\begin{aligned}
\min_{p \in \mathcal{H}} \quad & \|p\|_S && (7)\\
\text{s.t.} \quad & p(t) - Gx(t) - \beta p(t+1) = 0 \quad \text{for } t \in \hat{X} && (8)\\
& 0 = \lim_{T \to \infty} \beta^T p(T) && (9)
\end{aligned}
$$

Where $x(t)$ for $t \in \hat{X}$ is defined by $x(0)$ initial condition and recurrence $x(t+1) = Ax(t)$ in (2)

- The minimization of norm $\|p\|_S$ has "inductive bias" towards particular solutions for $t \in [0, \infty] \setminus \hat{X}$.

Is the no-bubble condition still necessary?

- To analyze, drop the no-bubble condition and examine the class of solutions.
- In this case, we know the interpolating solutions to (8) without imposing (9)

$$p(t) = p_f(t) + \zeta\beta^{-t} \tag{10}$$

- Applying the triangle inequality

$$\boxed{\|p_f\|_S \leq \|p\|_S \leq \|p_f\|_S + \zeta\ \|\beta^{-t}\|_S} \tag{11}$$

- Relative to classic methods the "deep learning" problem now has a new objective, minimizing $\|p\|_S$.
  - That is, $p(t) = p_f(t)$, the solution fulfills the no-bubble condition, and (9) is satisfied at the optima.
- The new objective of minimizing the norm, makes the no-bubble condition **redundant**.

Min-norm norm formulation: redundancy of no-bubble condition

Given the no-bubble condition is automatically fulfilled, could solve the following given some $\mathcal{H}$ and compare to $p_f(t)$

$$
\begin{aligned}
\min_{p\in\mathcal{H}} \quad & \|p\|_S && (12)\\
\text{s.t.} \quad & p(t) - Gx(t) - \beta p(t+1) = 0 \quad \text{for } t \in \hat{X} && (13)
\end{aligned}
$$

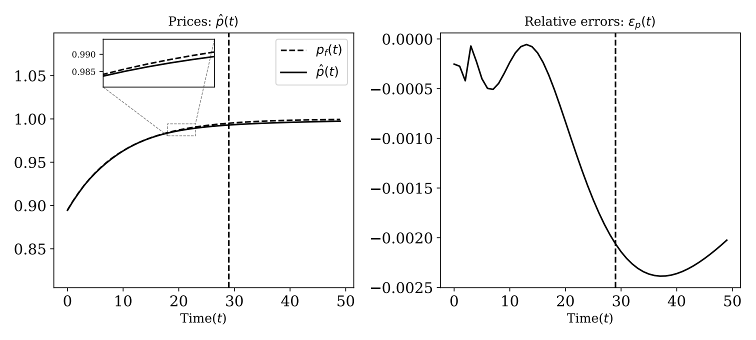
A reminder: in practice, given the $\hat{X}$, we directly implement this as $p(\cdot;\theta) \in \mathcal{H}(\Theta)$ and fit with

$$
\min_{\theta\in\Theta} \frac{1}{|\hat{X}|} \sum_{t\in\hat{X}} \left[p(t;\theta) - Gx(t) - \beta p(t+1;\theta)\right]^2 \tag{14}
$$

Since law of motion is deterministic, given $x(0)$ we generate $x(t)$ with $x(t+1) = Ax(t)$ for $t \in \hat{X}$

- The $\hat{X}$ does not need to be contiguous and $|\hat{X}|$ may be relatively small.
- Most important: no steady state calculated, nor large $T \in \hat{X}$ required.

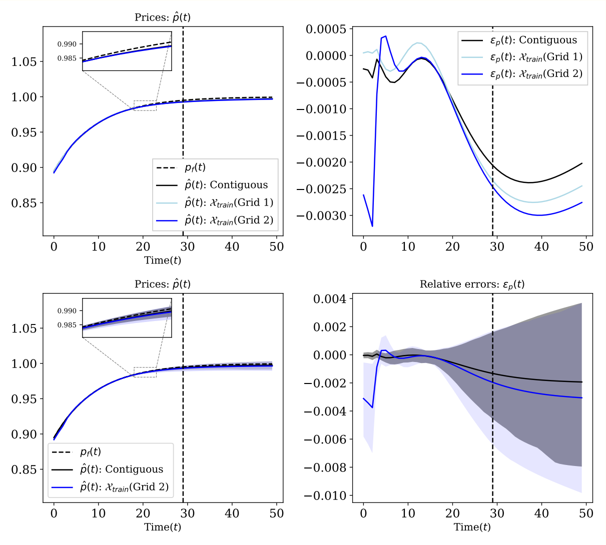
Results



1. **Pick** $\hat{X} = \{0, 1, 2, ..., 29\}$ and $t > 29$ is "extrapolation" where $c = 0.01$, $g = -0.1$, and $y_0 = 0.8$.
2. **Choose** $p(t;\theta) = NN(t;\theta)$ where "NN" has 4 hidden layers of 128 nodes. $|\Theta| = 49.9K$ coefficients.
3. **Fit** using L-BFGS and PyTorch in just a **few seconds**. Could use Adam/SGD/etc.
4. Low generalization errors, even without imposing no-bubble condition.

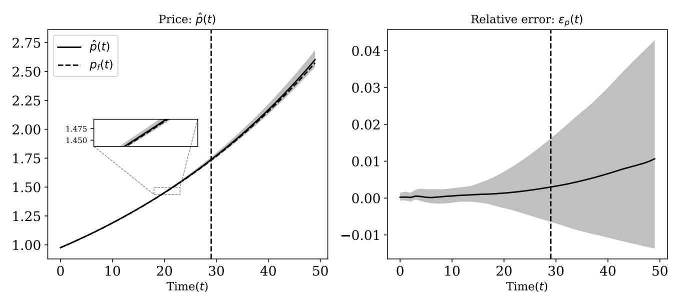
Relative errors define as $\varepsilon_p(t) \equiv \frac{\hat{p}(t)-p(t)}{p(t)}$.

Contiguous vs. sparse grid



- **Pick** $\hat{X}(\text{Grid 1}) = \{0, 1, 2, 4, 6, 8, 12, 16, 20, 24, 29\}$ and $\hat{X}(\text{Grid 2}) = \{0, 1, 4, 8, 12, 18, 24, 29\}$.
- Contrary to popular belief, can use **less grid points** relative to alternatives.
- The solutions are very close (with different seeds)
  - Hypothesis verified, the solutions agree on the seen and unseen grid points.

Growing dividends



- **Pick** same $\hat{X}$ but now $c = 0.0$, $g = 0.02$.
- **Choose** $p(t;\theta) = e^{\phi t} NN(t;\theta_1)$ where $\theta \equiv \{\phi, \theta_1\} \in \Theta$ are the coefficients.
  - Here we used economic intuition of problem to design $\mathcal{H}(\Theta)$ to generalize better.
- Non-stationary but can figure out the growth.
- Bonus: learns the growth rate: $\phi \approx \ln(1+g)$ and even extrapolates well! Growth rate

Neoclassical growth in sequence space

Sequential formulation

$$
\begin{aligned}
\max_{\{c(t),k(t+1)\}_{t=0}^{\infty}} \quad & \sum_{t=0}^{\infty} \beta^t u\left(c(t)\right) && (15)\\
\text{s.t.} \quad & k(t+1) = z(t)^{1-\alpha} f\left(k(t)\right) + (1-\delta)k(t) - c(t) && (16)\\
& z(t+1) = (1+g)z(t) && (17)\\
& k(t) \geq 0 && (18)\\
& 0 = \lim_{T\to\infty} \beta^T u'\left(c(T)\right) k(T+1) && (19)\\
& k_0, z_0 \text{ given} && (20)
\end{aligned}
$$

- Preferences: $u(c) = \frac{c^{1-\sigma}-1}{1-\sigma}$, $\sigma > 0$, $\lim_{c\to 0} u'(c) = \infty$, and $\beta \in (0,1)$.
- Cobb-Douglas production function: $f(k) = k^{\alpha}$, $\alpha \in (0,1)$ before scaling by TFP z_t.
- Skip standard steps... Euler equation: $u'(c(t)) = \beta u'(c(t+1))\big[z(t+1)^{1-\alpha} f'(k(t+1)) + 1 - \delta\big]$.

Interpolation problem: without transversality condition

- A set of points in time $\hat{X} = \{t_1, \ldots, t_{\max}\}$.
- A family of over-parameterized functions $k(\cdot;\theta) \in \mathcal{H}(\Theta)$.
- Generate $z(t)$ using the law of motion and $z(0)$, equations (17).
- Use the feasibility condition and define $c(t;k) \equiv z(t)^{1-\alpha} f\left(k(t)\right) + (1-\delta)k(t) - k(t+1)$.

In practice we minimize the Euler and initial conditions residuals:

$$\min_{\theta\in\Theta}\left(\frac{1}{|\hat{X}|}\sum_{t\in\hat{X}}\lambda_1\left[\underbrace{\frac{u'\left(c(t;k(\cdot,\theta))\right)}{u'\left(c(t+1;k(\cdot;\theta))\right)} - \beta\left[z(t+1)^{1-\alpha}f'(k(t+1;\theta)) + 1 - \delta\right]}_{\text{Euler residuals}}\right]^2 + \lambda_2\left[\underbrace{k(0;\theta) - k_0}_{\text{Initial condition residuals}}\right]^2\right)$$

- λ_1 and λ_2 positive weights.

Interpolation problem: without transversality condition

- This minimization **does not contain** the transversality condition.
  - Without the transversality condition it has infinitely many minima.

- **No explicit** norm minimization.

- Does the implicit bias weed out the solutions that violate the transversality condition? **Yes**.

- **Intuition**: The solutions that violate the transversality condition are big functions with big derivatives.

  Let's analyze this more rigorously.

Interpolation formulation: min-norm mental model

$$\min_{k \in \mathcal{H}} \quad \|k\|_S \tag{21}$$

$$\text{s.t.} \quad u'(c(t;k)) = \beta u'(c(t+1;k))\left[z(t+1)^{1-\alpha} f'(k(t+1)) + 1 - \delta\right] \quad \text{for } t \in \hat{X} \tag{22}$$

$$k(0) = k_0 \tag{23}$$

$$0 = \lim_{T \to \infty} \beta^T u'(c(T;k))k(T+1) \tag{24}$$

$$\boxed{c(t;k) \equiv z(t)^{1-\alpha} f\left(k(t)\right) + (1-\delta)k(t) - k(t+1)} \tag{25}$$

Where $z(t)$ for $t \in \hat{X}$ is defined by $z(0)$ initial condition and recurrence $z(t+1) = (1+g)z(t)$.

Is the transversality condition still necessary? Case of $g = 0$, $z = 1$

Sketch of the proof:

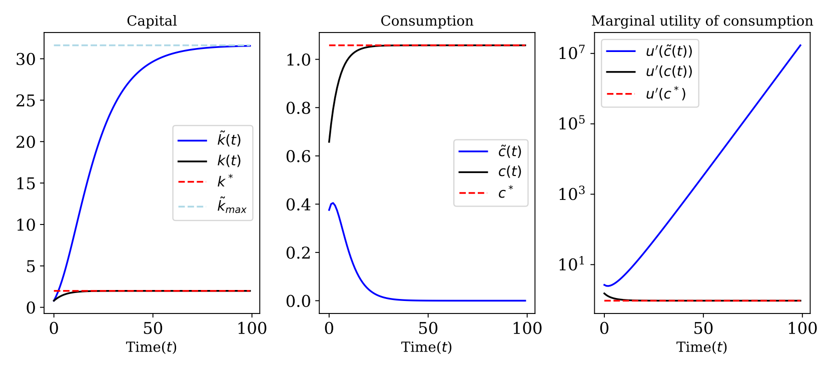
- Let $\{k(t), c(t)\}$ be the sequence of optimal solution.
- Let $\{\tilde{k}(t), \tilde{c}(t)\}$ be a sequence of solution that satisfy all the equations **except** transversality condition (24).

1. $\tilde{c}(t)$ approaches zero.
2. $\tilde{k}(t)$ approaches $\tilde{k}_{\text{max}} \equiv \delta^{\frac{1}{\alpha-1}}$, and $k(t)$ approaches $k^* \equiv \left(\frac{\beta^{-1}+\delta-1}{\alpha}\right)^{\frac{1}{\alpha-1}}$.
3. Both $\tilde{k}(t)$ and $k(t)$ are monotone. $\tilde{k}_{\text{max}} \gg k^*$. Therefore,

$$0 \leq \|k\|_S \leq \|\tilde{k}\|_S.$$

Is the transversality condition still necessary? Case of $g = 0$, $z = 1$

Example: the violation of the transversality condition.



- The solution that violate the transversality are associated with **"big"** capital path.
- The new objective of minimizing the norm, makes the transversality condition **redundant**.

Min-norm formulation: redundancy of transversality condition

Given the transversality condition is automatically fulfilled, one could solve

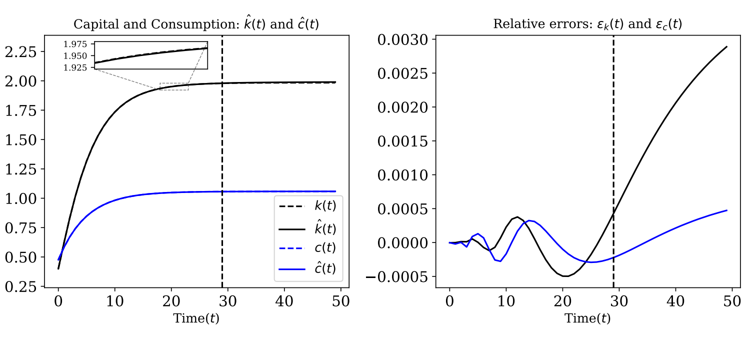
$$\begin{aligned} \min_{k\in\mathcal{H}} \quad & \|k\|_S \\ \text{s.t.} \quad & u'(c(t;k)) = \beta u'(c(t+1;k))\left[z(t+1)^{1-\alpha} f'(k(t+1)) + 1 - \delta\right] \quad \text{for } t \in \hat{X} \\ & k(0) = k_0 \end{aligned}$$

Reminder: in practice we solve

$$\min_{\theta\in\Theta} \left(\frac{1}{|\hat{X}|} \sum_{t\in\hat{X}} \lambda_1 \left[\frac{u'\left(c(t;k(\cdot,\theta))\right)}{u'\left(c(t+1;k(\cdot;\theta))\right)} - \beta\left[z(t+1)^{1-\alpha} f'(k(t+1;\theta)) + 1 - \delta\right] \right]^2 + \lambda_2 \left[\underbrace{k(0;\theta) - k_0}_{\text{Initial condition residuals}} \right]^2 \right)$$

- $|\hat{X}|$ may be relatively small, no steady state calculated, nor large $T \in \hat{X}$ required. ▸▸ Sparse Grids

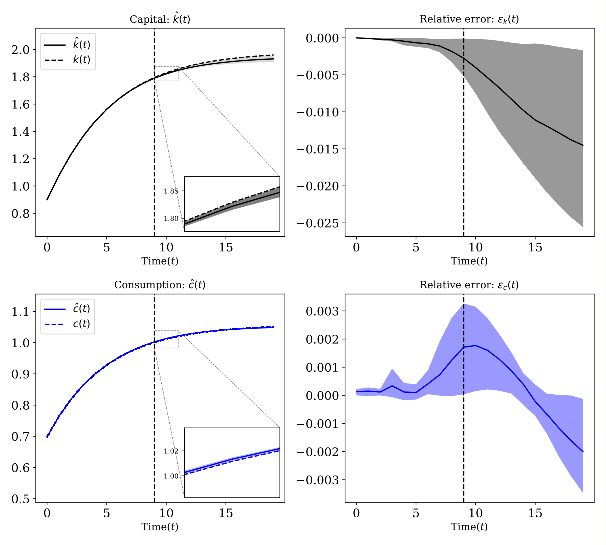
Results



1. **Pick** $\hat{X} = \{0, 1, ..., 30\}$ and $t > 30$ is "extrapolation" $\alpha = \frac{1}{3}$, $\sigma = 1$, $\beta = 0.9$, $g = 0.0$, and $k_0 = 0.4$
2. **Choose** $k(t; \theta) = NN(t; \theta)$ where "NN" has 4 hidden layers of 128 nodes. $|\Theta| = 49.9K$ coefficients.
3. **Fit** using L-BFGS in just a **few seconds**. Comparing with value function iteration solution.
4. Low generalization errors, even without imposing the transversality condition. Small k_0.

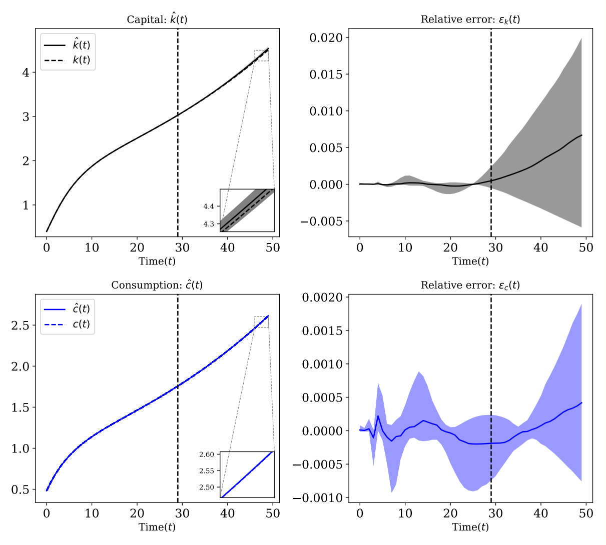
Relative errors defined as $\varepsilon_c(t) \equiv \frac{\hat{c}(t) - c(t)}{c(t)}$, $\varepsilon_k(t) \equiv \frac{\hat{k}(t) - k(t)}{k(t)}$.

Far from the steady state



- **Pick** $\hat{X} = \{0, 1, \ldots, 9\}$
- No large $T \in \hat{X}$ is required.
    - Even for medium time horizons the solutions do not violate TVC.
    - Long-run errors do not impair the accuracy of short run dynamics.
- Generalization errors are small.

Growing TFP



- **Pick** same $\hat{X}$ but now $g = 0.02$.
- **Choose** $k(t;\theta) = e^{\phi t} NN(t;\theta_{NN})$ where $\theta \equiv \{\phi, \theta_{NN}\} \in \Theta$ is the coefficient vector
    - Here we used economic intuition of problem to design the $\mathcal{H}(\Theta)$ to generalize better.
- Non-stationary but can figure out the BGP.
- Learns the growth rate: $\phi \approx \ln(1+g)$
- Economic insight leads to great extrapolation!
- It works very well even in the presence of misspecifcation.

Linear growth

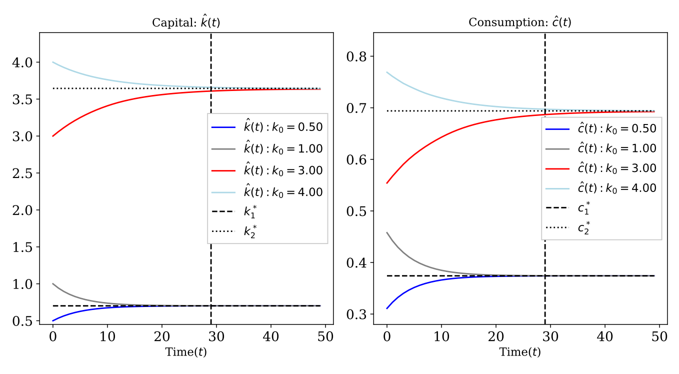
The neoclassical growth model with multiple steady states

Sequential formulation

$$
\begin{aligned}
\max_{\{c_t, k_{t+1}\}_{t=0}^{\infty}} \quad & \sum_{t=0}^{\infty} \beta^t u(c_t) \\
\text{s.t.} \quad & k_{t+1} = f(k_t) + (1-\delta)k_t - c_t \\
& k_t \geq 0 \\
& 0 = \lim_{T\to\infty} \beta^T u'(c_T) k_{T+1} \\
& k_0 \text{ given.}
\end{aligned}
$$

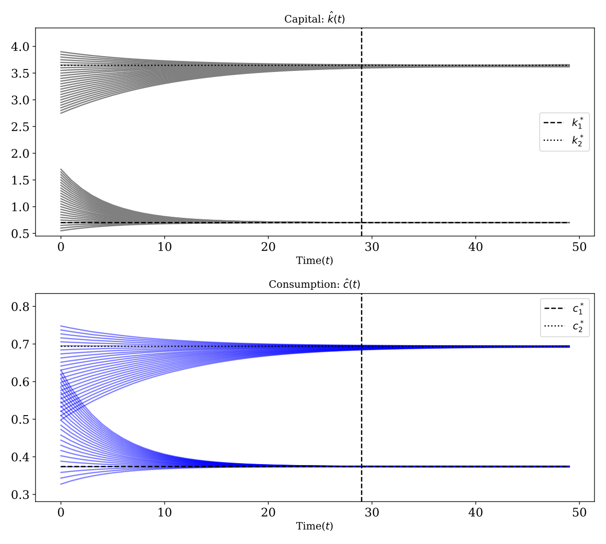
1. Preferences: $u(c) = \frac{c^{1-\sigma}-1}{1-\sigma}$, $\sigma > 0$, $\lim_{c\to 0} u'(c) = \infty$, and $\beta \in (0,1)$.
2. **"Butterfly production function"**: $f(k) = a \max\{k^{\alpha}, b_1 k^{\alpha} - b_2\}$, $\alpha \in (0,1)$:
    - There is a kink in the production function at $k^* \equiv \left(\frac{b_2}{b_1 - 1}\right)^{\frac{1}{\alpha}}$.
    - This problem has **two** steady states, k_1^* and k_2^* and their corresponding consumption levels c_1^* and c_2^*.

Results



1. **Pick** $\hat{X} = \{0, \ldots, 30\}$, $\alpha = \frac{1}{3}$, $\sigma = 1$, $\beta = 0.9$, $g = 0.0$, $a = 0.5$, $b_1 = 3$, $b_2 = 2.5$ and $k_0 \in \{0.5, 1.0, 3.0, 4.0\}$
2. **Choose** $k(t; \theta) = NN(t; \theta)$ where "NN" has 4 hidden layers of 128 nodes. $|\Theta| = 49.9K$ coefficients.
3. **Fit** using Adam optimizer.

Results: different initial conditions



- Different initial conditions in $k_0 \in [0.5, 1.75] \cup [2.75, 4]$.
- In the vicinity of k_1^* and k_2^* the paths converge to the right steady-states.
  - The implicit bias picks up the right path.
- Low generalization errors, even without imposing the transversality condition.

Details

Recursive version of the neoclassical growth model here

Recursive formulation (with a possible BGP)

Skipping the Bellman formulation and going to the first order conditions in the state space , i.e., (k, z)

$$
\begin{aligned}
&u'(c(k,z)) = \beta u'(c(k'(k,z),z'))\left[z'^{1-\alpha} f'(k'(k,z)) + 1 - \delta\right] \\
&k'(k,z) = z^{1-\alpha} f(k) + (1-\delta)k - c(k,z) \\
&z' = (1+g)z \\
&k' \geq 0 \\
&0 = \lim_{T\to\infty} \beta^T u'(c_T) k_{T+1} \quad \forall (k_0, z_0) \in X
\end{aligned}
$$

- Preferences: $u(c) = \frac{c^{1-\sigma}-1}{1-\sigma}$, $\sigma > 0$, $\lim_{c\to 0} u'(c) = \infty$, and $\beta \in (0,1)$.
- Cobb-Douglas production function: $f(k) = k^\alpha$, $\alpha \in (0,1)$ before scaling by TFP z.

Interpolation problem: without transversality condition

- A set of points $\hat{X} = \{k_1, \ldots, k_{N_k}\} \times \{z_1, \ldots, z_{N_z}\}$.
- A family of over-parameterized functions $k'(\cdot, \cdot; \theta) \in \mathcal{H}(\Theta)$.
- Use the feasibility condition and define $c(k, z; k') \equiv z^{1-\alpha} f(k) + (1-\delta)k - k'(k, z)$.

In practice we minimize the Euler residuals:

$$\min_{\theta \in \Theta} \frac{1}{|\hat{X}|} \sum_{(k,z) \in \hat{X}} \left[\underbrace{\frac{u'\Big(c\big(k, z; k'(.;\theta)\big)\Big)}{u'\Big(c\big(k'(k, z; \theta), (1+g)z; k'(.;\theta)\big)\Big)} - \beta \left[((1+g)z)^{1-\alpha} f'\left(k'(k, z; \theta)\right) + 1 - \delta \right]}_{\text{Euler residual}} \right]^2$$

Interpolation problem: without the transversality condition

- This minimization **does not contain** the transversality condition.
  - Without the transversality condition it has more than one minima.
- **No explicit** norm minimization.
- Does the implicit bias weed out the solutions that violate the transversality condition? **Yes**
- **Intuition**: The solutions that violate the transversality condition are "bigger" than those don not violate it.

  Let's analyze this more rigorously.

Interpolation formulation: min-norm mental model

$$\min_{k' \in \mathcal{H}} \quad \|k'\|_S \tag{26}$$

$$\text{s.t.} \quad u'\Big(c\big(k, z; k'\big)\Big) = \beta u'\Big(c\big(k'(k, z), (1+g)z; k'\big)\Big) \times$$

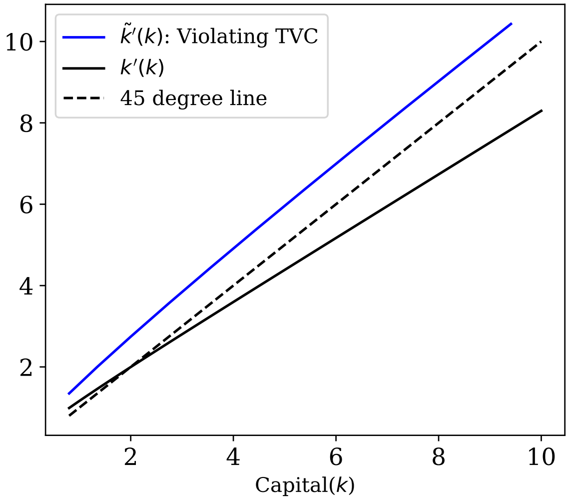
$$\big[((1+g)z)^{1-\alpha} f'(k'(k, z)) + 1 - \delta\big] \quad \text{for } (k, z) \in \hat{X} \tag{27}$$

$$0 = \lim_{T \to \infty} \beta^T u'(c(T))k(T+1) \quad \text{for all } (k_0, z_0) \in X \tag{28}$$

where

$$c(k, z; k') \equiv z^{1-\alpha} f(k) + (1-\delta)k - k'(k, z)$$

Is the transversality condition necessary? Case of g = 0, z = 1



Min-norm formulation: redundancy of transversality condition

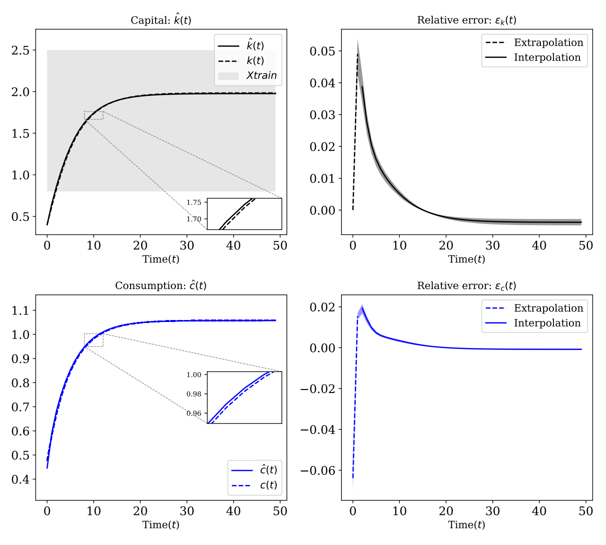
We can drop the transversality condition:

$$
\begin{aligned}
\min_{k' \in \mathcal{H}} \quad & \|k'\|_S \\
\text{s.t.} \quad & u'\Big(c\big(k, z; k'\big)\Big) = \beta u'\Big(c\big(k'(k,z), (1+g)z; k'\big)\Big) \times \\
& \big[((1+g)z)^{1-\alpha} f'(k'(k,z)) + 1 - \delta\big] \quad \text{for } (k,z) \in \hat{X}
\end{aligned}
$$

In practice, given $\hat{X}$, we directly implement this as $k'(\cdot, \cdot; \theta) \in \mathcal{H}(\Theta)$ and fit with

$$
\min_{\theta \in \Theta} \frac{1}{|\hat{X}|} \sum_{(k,z) \in \hat{X}} \left[\frac{u'\Big(c\big(k, z; k'(.;\theta)\big)\Big)}{u'\Big(c\big(k'(k,z;\theta), (1+g)z; k'(.;\theta)\big)\Big)} - \beta \left[((1+g)z)^{1-\alpha} f'\left(k'(k,z;\theta)\right) + 1 - \delta\right] \right]^2
$$

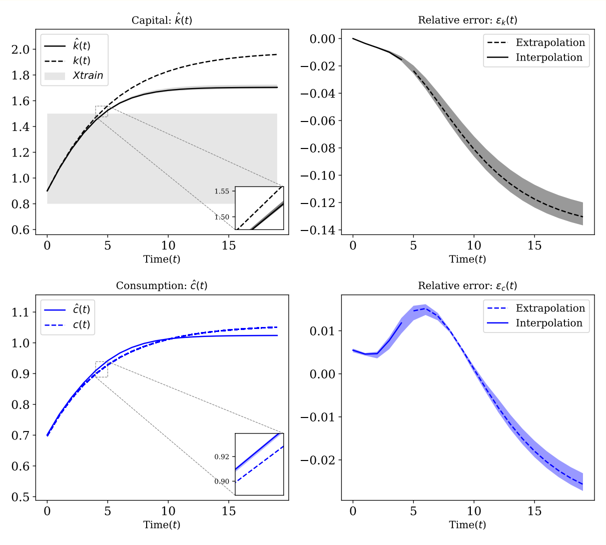
Results: one initial condition



- **Pick** $\hat{X} = [0.8, 2.5] \times \{1\}$ and $k_0 = 0.4 \notin \hat{X}$ is "extrapolation" $\alpha = \frac{1}{3}$, $\sigma = 1$, $\beta = 0.9$.
- **Choose** $k'(k, z; \theta) = NN(k, z; \theta)$ where "NN" has 4 hidden layers of 128 nodes. $|\Theta| = 49.9K$ coefficients.
- **Fit** using L-BFGS and PyTorch in just a few seconds.
- Low generalization errors, even without imposing transversality condition.

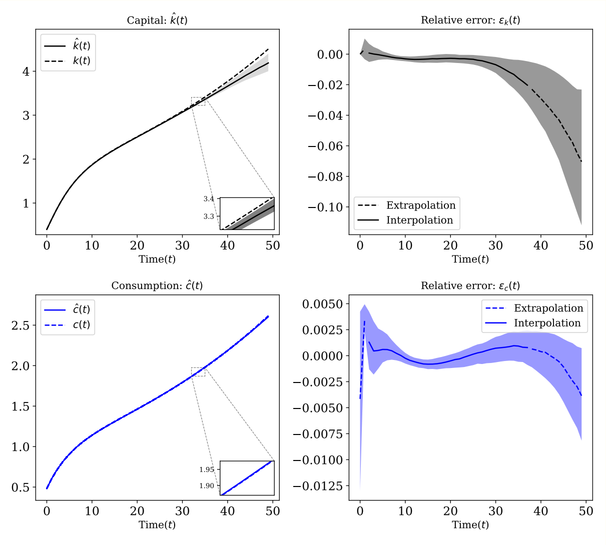
For all $k \in X$

Far from the steady state



- **Pick** $\hat{X} = [0.8, 1.5]$, $k^* \notin [0.8, 1.5]$.
- A local grid around the k_0 is enough.
  - Accurate solutions in the interpolation region.
- Generalization errors are not bad.

Growing TFP



- **Pick** $\hat{X} = [0.8, 3.5] \times [0.8, 1.8]$ but now $g = 0.02$.
- **Choose** $k'(k, z; \theta) = zNN(k, \frac{k}{z}; \theta)$.
  - Here we used economic intuition to design the $\mathcal{H}(\Theta)$.
- Relative errors are very small inside the grid.
- Small generalization errors.

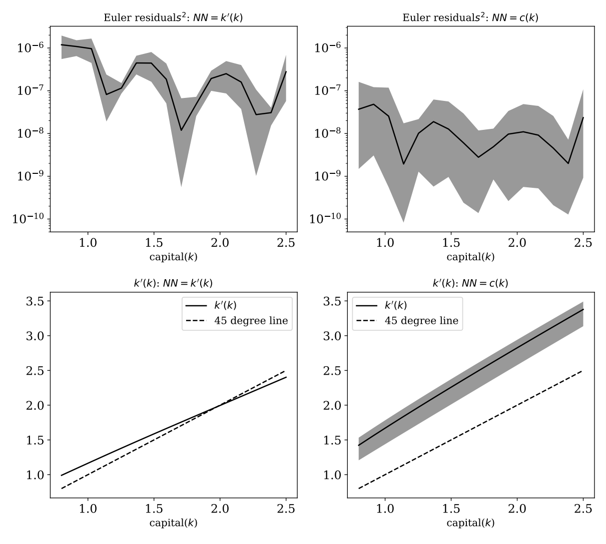
Are Euler and Bellman residuals enough?

Euler residuals are not enough

- We picked a grid $\hat{X}$ and approximated $k'(k)$ with an over-parameterized function.
  - The approximate solutions do not violate the transversality condition.
- What happens if we approximate the consumption functions $c(k)$ with an over-parameterized function.
  - We get an interpolating solution, i.e, very small Euler residuals.
  - However, the solutions **violate** the transversality condition.

Intuition: consumption functions with low derivatives leads to optimal policies for capital with big derivatives.

Small Euler residuals can be misleading



- Left panels: approximating $k'(z)$ with a deep neural network.
  - The solutions do not violate the TVC.
  - $k'(k)$ intersects with 45° line at $k^* \approx 2$.
- Right panels: approximating $c(k)$ with a deep neural network.
  - The solutions **violate** the TVC.
  - $k'(k)$ intersects with 45° line at $\tilde{k}_{\max} \approx 30$.
  - Euler residuals are systematically lower.

Conclusion

Conclusion

- Solving functional equations with deep learning is an extension of collocation/interpolation methods.
- With **massive over-parameterization**, optimizers tend to choose those interpolating functions which are not explosive and with smaller gradients (i.e., **inductive bias**).
- Over-parameterized solutions **automatically** fulfill **forward-looking** boundary conditions:
  - Shedding light on the convergence of deep learning based solutions in dynamic problems in macroeconomics.
- If we solve models with deep-learning without (directly) imposing long-run boundary conditions,
  - Short/medium-run errors are small, and long-run errors after **"we are all dead"** are even manageable.
  - Long-run errors do not affect transition dynamics even in the presence of **non-stationarity** and **steady-state multiplicity**.
  - Gives hope for solving high-dimensional models still disciplined by forward-looking economic assumptions.

Appendix

Sobolev semi-norms ⏩ back

Let ψ_1 and ψ_2 be two differentiable function from a compact space $\mathcal{X}$ in $\mathbb{R}$ to $\mathbb{R}$ such that

$$\int_{\mathcal{X}} \left| \frac{d\psi_1}{ds} \right|^2 ds > \int_{\mathcal{X}} \left| \frac{d\psi_2}{ds} \right|^2 ds \tag{30}$$

then

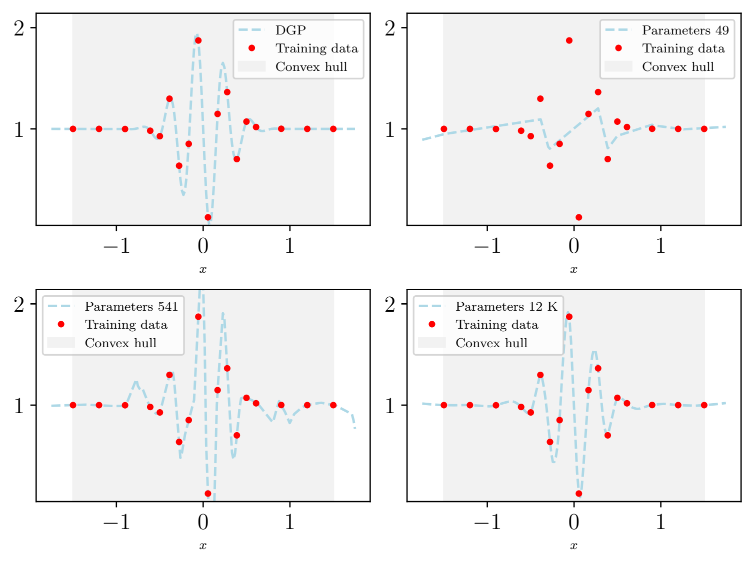
$$\|\psi_1\|_S > \|\psi_2\|_S. \tag{31}$$

Moreover, since $\|\cdot\|_S$ is a semi-norm, it satisfies the triangle inequality

$$\|\psi_1 + \psi_2\|_S \leq \|\psi_1\|_S + \|\psi_2\|_S. \tag{32}$$

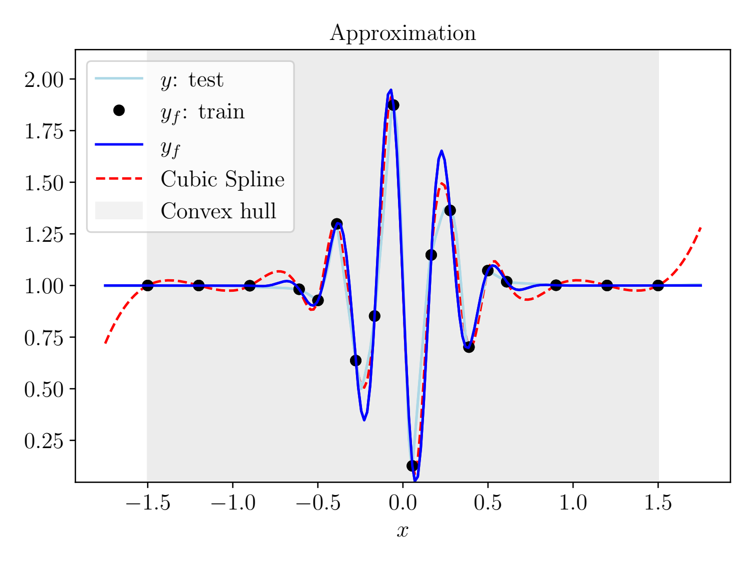
Recently shown the optimizers penalize Sobolev semi-norms: Ma, C., Ying, L. (2021)

Smooth interpolation



Smooth interpolation: Comparison with cubic splines

back



Smooth interpolation: A simple dynamical system

Consider the following system

$$K_{t+1} = \eta K_t.$$

This system have the following solutions

$$K(t) = K_0 \eta^t.$$

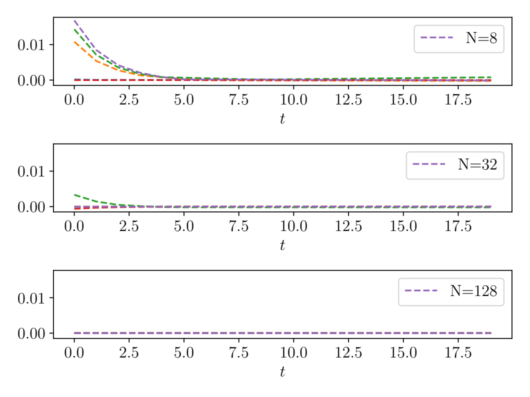
- Without specifying the initial condition, K_0, this is an ill-defined problem, i.e., there are infinity many solutions.
- The solution to:

$$\begin{aligned} \min_{K \in \mathcal{H}} \quad & \|K\|_S \\ \text{s.t.} \quad & K(t+1) - \eta K(t) = 0 \quad \text{for } t = t_1, \ldots, t_N \end{aligned}$$

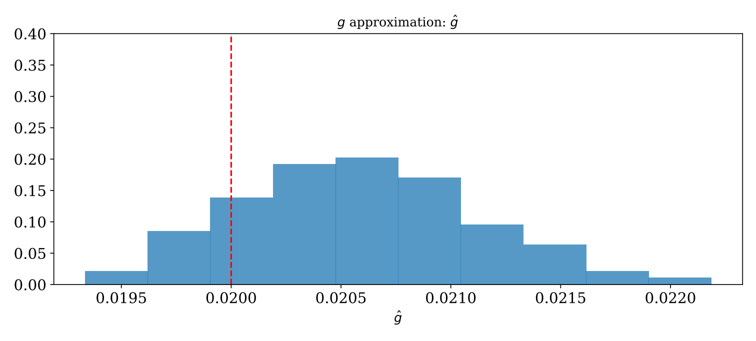
  is $K(t) = 0$.

Smooth interpolation: A simple dynamical system results

Three layers deep neural network, for $N = 8$, 32, and 128. Each trajectory corresponds to different random initialization of the optimization procedure (seed).



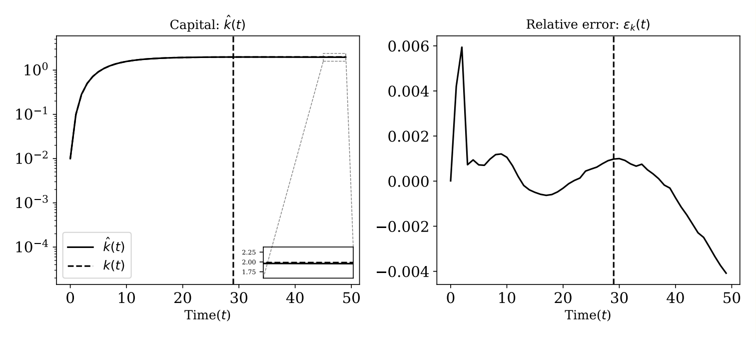
Learning the growth rate



$$\hat{g} \equiv e^{\hat{\phi}} - 1.$$

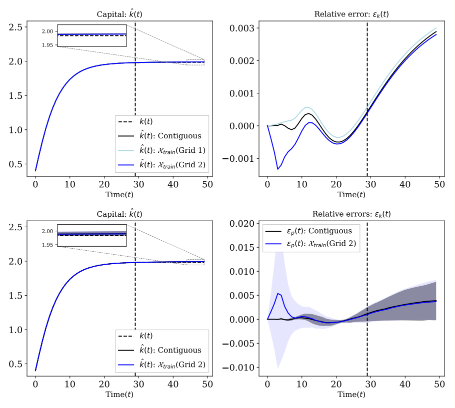
The histogram for approximate growth rate over 100 seeds. back

Learning the growth rate



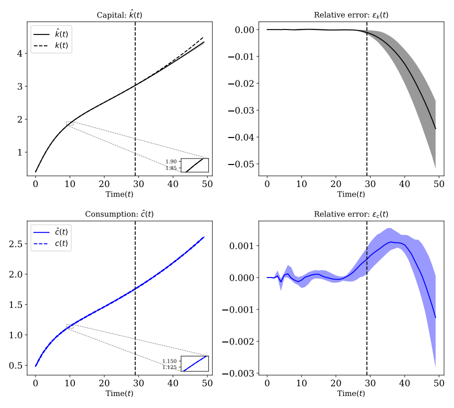
back

Contiguous vs. dense grid



- $\hat{X}(\text{Grid 1}) = \{0, 1, 2, 4, 6, 8, 12, 16, 20, 24, 29\}$, $\hat{X}(\text{Grid 2}) = \{0, 1, 4, 8, 12, 18, 24, 29\}$.
- Contiguous grid : $\hat{X} = \{0, 1, 2, ..., 29\}$. back

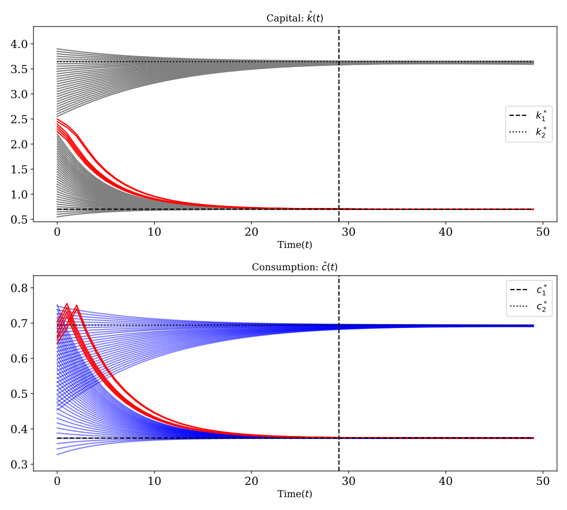
Misspecification of growth



$$k(t;\theta) = tNN(t;\theta) + \phi$$

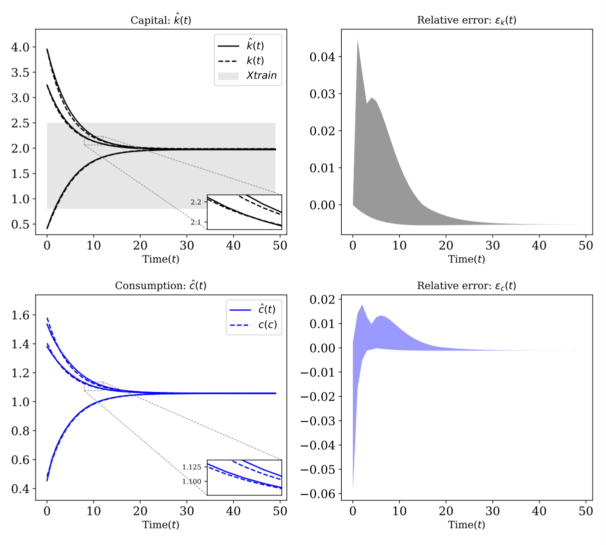
back

Neoclassical growth with multiple steady-states: where things fail



back

Results: initial conditions over the state space



- The solution has to satisfy the transversality condition for all points in X
  - $\lim_{T\to\infty} \beta^T u'(c(T))k(T+1) = 0 \quad \forall\ k_0 \in X$
- Left: Three different initial condition for capital, two of them outside X.
- Shaded regions: error range in capital and consumption for 70 different initial condition in $[0.5, 4.0]$.

back