NBER WORKING PAPER SERIES

SPOOKY BOUNDARIES AT A DISTANCE: INDUCTIVE BIAS, DYNAMIC MODELS, AND BEHAVIORAL MACRO

Mahdi E. Kahou
Jesús Fernández-Villaverde
Sebastian Gomez-Cardona
Jesse Perla
Jan Rosa

Working Paper 32850
http://www.nber.org/papers/w32850

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2024

Spooky Boundaries at a Distance: Inductive Bias, Dynamic Models, and Behavioral Macro
Mahdi E. Kahou, Jesús Fernández-Villaverde, Sebastian Gomez-Cardona, Jesse Perla, and Jan Rosa
NBER Working Paper No. 32850
August 2024
JEL No. C0,E0

ABSTRACT

In the long run, we are all dead. Nonetheless, when studying the short-run dynamics of economic models, it is crucial to consider boundary conditions that govern long-run, forward-looking behavior, such as transversality conditions. We demonstrate that machine learning (ML) can automatically satisfy these conditions due to its inherent inductive bias toward finding flat solutions to functional equations. This characteristic enables ML algorithms to solve for transition dynamics, ensuring that long-run boundary conditions are approximately met. ML can even select the correct equilibria in cases of steady-state multiplicity. Additionally, the inductive bias provides a foundation for modeling forward-looking behavioral agents with self-consistent expectations.

Mahdi E. Kahou
Bowdoin College
mekahou@alumni.ubc.ca

Jesús Fernández-Villaverde
Department of Economics University of Pennsylvania
The Ronald O. Perelman Center for Political Science and Economics
133 South 36th Street Suite 150
Philadelphia, PA 19104
and CEPR
and also NBER
jesusfv@econ.upenn.edu

Sebastian Gomez-Cardona
Morningstar
sebastiangomez87@gmail.com

Jesse Perla
Vancouver School of Economics
University of British Columbia
jesse.perla@ubc.ca

Jan Rosa
University of British Columbia
Vancouver School of Economics
jan.rosa1993@gmail.com

1 Introduction

> But this *long run* is a misleading guide to current affairs. *In the long run* we are all dead, *J.M. Keynes, A Tract on Monetary Reform (1923), p. 65.*

> Newton believed that his laws of motion do not guarantee the stability of the solar system: he concluded that God periodically intervenes to keep things going smoothly, *E. Sober, Ockham's Razors: A User's Manual (2015), p. 60. Refers to Newton's Letter to Richard Bentley dated February 11, 1693*

Steady states play a paradoxical role in dynamic economic models. On one hand, they are only reached asymptotically ("long after we are dead," as Keynes would put it), and the model's short-run behavior often differs significantly from them. On the other hand, steady states are crucial for solving these models. Researchers commonly impose long-run assumptions based on economic reasoning, such as transversality conditions, to ensure consistent short-run dynamics.

However, imposing these conditions often complicates solving models, particularly those with many dimensions. Generally, these conditions require solving the model over a broad range of possible values of the state variables. For example, recursive formulations necessitate accurate solutions for arbitrary values of the state variables, even though the solution may only be relevant from a single initial condition.

This paper presents two key contributions. First, we demonstrate that it is possible to meet long-run boundary conditions without strictly enforcing them as a constraint on the model's dynamics. Specifically, we show how machine learning (ML) methods enable us to conduct short-run simulations while still satisfying boundary conditions, even in scenarios with multiple steady states, due to the inductive bias inherent in ML algorithms. Inductive bias, a concept widely discussed in the philosophy of science and ML, reflects a preference for simplicity and parsimony (akin to Ockham's razor) when fitting a general model with limited observations.

Inductive bias is also closely related to the double descent phenomenon, which describes the ability of highly parameterized models —such as deep neural networks with thousands, millions, or even billions of parameters— to escape the classical bias-variance tradeoff and achieve minimal errors in fitting and forecasting (see Belkin et al., 2019, and Belkin, 2021).[1] We document how

[1]While it might seem counterintuitive that a model with billions of parameters could be considered "simple," counting parameters is not the correct measure of parsimony in a function space (the space in which dynamic model

neural networks, with four orders of magnitude more parameters than grid points (the "data" in this context), can solve economic models, producing parsimonious solutions with minimal errors that adhere to long-run constraints. This contribution justifies the accuracy of ML methods, including deep learning, in solving high-dimensional models (e.g., large spatial models used to analyze climate change as in Cruz and Rossi-Hansberg, 2023), even those with multiple steady states and hysteresis.

The intuition behind why the inductive bias delivers the results is straightforward. Economists' assumptions, such as the transversality condition, reject explosive trajectories of the model's state (or co-state) variables because they violate the boundary conditions that discipline agents' forward-looking forecasts. If we measure these explosive trajectories as functions, their functional (semi-)norms are large.[2] In contrast, trajectories that satisfy the long-run boundary conditions have small functional (semi-)norms. Inductive bias reflects a preference for solutions with small functional (semi-)norms (in whatever context they appear). Thus, ML tends to select solutions that satisfy long-run conditions, even without being explicitly programmed to search for them.

Our second contribution is to argue how inductive bias can serve as a micro-foundation for modeling forward-looking behavioral agents. Instead of relying on ad hoc behavioral rules, we propose equipping agents with a general ML model, such as a richly parameterized deep neural network, and allowing them to learn from a few observations. Inductive bias ensures that agents will learn a solution that (i) is easy to compute, (ii) exhibits minimal errors, and (iii) satisfies the necessary long-run constraints. Moreover, by periodically retraining the neural network, long-run errors are minimized. As Sober's quote at the beginning of this paper suggests, periodic intervention —here, retraining— ensures that long-run behavior remains stable and smooth. Remarkably, this periodic retraining leads to approximately time-consistent policies.

In this way, we may enable economists to construct behavioral approximations of forward-looking agents while limiting "additional 'free parameters,' unrestricted by theory," which was a central goal of the rational expectations revolution (Lucas and Sargent, 1981, and Sargent, 2024). Thus, our work builds on the tradition of Sargent (1993, p.16) and Evans and Honkapohja (2001, p.69-70), who connected perfect-foresight models, transversality conditions, stability, and

solutions reside). Instead, complexity is assessed within the function space of the solutions themselves, using tools such as Kolmogorov complexity, the Vapnik-Chervonenkis dimension, Rademacher complexity, and other related concepts depending on the context.

[2]In a deterministic model, the solutions themselves are the functions. In stochastic or recursive models, the function (semi-)norm applies to the policy itself, which generates the divergent trajectories.

bounded rationality. By having agents solving a problem with a bias toward min-norm solutions, our results are akin to the sparsity models of Gabaix (2014) and Gabaix (2023) and the new classes of attention cost functions in Caplin et al. (2022).[3]

After formally introducing inductive bias, we apply it to understanding transversality in two canonical models: the linear asset pricing model and the neoclassical growth model. Building on the established knowledge of these models, we demonstrate how ML methods implicitly (or explicitly) regularize solutions to achieve min-norms that satisfy boundary conditions without directly calculating long-run behavior. Moreover, the solutions exhibit excellent short-run accuracy.

In the case of the neoclassical growth model, we also highlight the connection between our results and the classical turnpike theorems (see McKenzie, 1976, and Marimón, 1989). These theorems establish that models with long —but finite— horizons share almost identical short- to medium-run dynamics with infinite-horizon models. ML methods seem to "understand" turnpike theorems.

We deliberately chose two simple models to support our case. Attempting to illustrate our arguments with a large model (e.g., a mid-sized New Keynesian model) would be a fool's errand. The many moving parts of a complex larger model would obscure the intuition and make it difficult to establish benchmarks for evaluating our solution.

Indeed, both applications transparently underscore that long-run boundary conditions —such as transversality and no-bubble conditions— arise from economic assumptions essential for model consistency.[4] These are not merely technical conditions; they are intrinsic to the economics of rational expectations. While these long-run boundary conditions that discipline forward expectations may not always be explicitly stated, they are implicitly present when solving infinite-horizon control problems or using recursive methods.

In linear models, boundary conditions are often articulated in terms of stability. For example, in Blanchard and Kahn (1980) and Klein (2000), these conditions are satisfied by selecting the unique non-explosive solution through spectral methods. Checking the eigenvalues for a linear,

[3] As an example, Gabaix, 2014, p. 1694, discusses the role of axioms in his framework and their connection to the compressed sensing literature, where underdetermined problems are solved using a min-norm assumption (e.g., the ℓ_1 norm, nuclear norms, etc.).

[4] The classic literature on rational expectations carefully articulates these assumptions in economic terms. For example, Sargent and Wallace (1973) introduce an asymptotic boundary condition that "the money supply [is] not expected to increase too swiftly." Similarly, Blanchard and Kahn (1980) emphasize the need to "rule out exponential growth of the expectation." Knife-edge stability is also a common feature of monetary models that assume perfect foresight or rational expectations, as discussed in Obstfeld and Rogoff (1983).

time-invariant policy provides a sufficient condition to eliminate unstable trajectories that violate transversality. In global methods, boundary conditions are frequently applied implicitly (e.g., during steady-state calculations) or may appear to be bypassed, such as in collocation on a compact space. However, they remain a necessary condition for optimality (e.g., see Ekeland and Scheinkman, 1986, and Kamihigashi, 2005).[5]

In summary, we lay the theoretical groundwork for using deep learning to find equilibria in dynamic models. Examples of this burgeoning literature include Ebrahimi Kahou et al. (2021), Maliar et al. (2021), Azinovic et al. (2022), Han et al. (2022), Kase et al. (2022), Barnett et al. (2023), Fernández-Villaverde et al. (2023), Jungerman (2023), Duarte et al. (2024), and Payne et al. (2024). Interestingly, these studies do not directly impose transversality conditions, and none explicitly address the issue. Alternatively, Ebrahimi Kahou et al. (2024) demonstrate how to solve a min-norm problem using kernel methods, which provably satisfy transversality conditions.

We also connect our work to the recent literature on behavioral macroeconomics, such as Gabaix (2020), Caplin et al. (2022), and Gabaix (2023). The key difference between that literature and our approach is that while parametric methods like rational inattention identify the unique solution to a well-posed parametric problem, we document a bias toward particular solutions in optimization problems with multiple solutions. Finally, our exploration of how ML can construct self-consistent expectations echoes the ideas found in Bianchi et al. (2022).

The remainder of the paper is organized as follows. Section 2 introduces the concept of inductive bias and its role in ML. Section 3 presents our first canonical model, the linear asset pricing model. Section 4 describes our second canonical model, the neoclassical growth model. Section 5 concludes. A technical appendix provides additional results.

2 Inductive Bias in ML

Before we can interpret inductive bias in the context of dynamic, forward-looking models, we need to understand it in the context of solving functional equations. Let $\mathcal{X}$ be a space and write the economic model as a set of functional equations $\ell(x, f) = 0$ for all $x \in \mathcal{X}$ where $f : \mathcal{X} \to \mathcal{R}$. For example, in a growth model, the $\ell(x, f)$ might be a combination of the Euler equation residual at a particular capital level x and the resource constraint where f is the investment policy.

[5] Appendix B explains why these conditions are traditionally overlooked in low-dimensional problems solved globally and why this approach becomes inadequate in high-dimensional contexts.

A typical solution method in economics parameterizes a policy (or value) function with $f_\theta \in \mathcal{H}(\Theta)$, where $\mathcal{H}$ is a hypothesis class of function approximations. In our case, we will deal with high-dimensional parameterizations (e.g., neural networks with many parameters $\theta \in \Theta$).

A simple algorithm, which generalizes classic approaches such as collocation, is to find a f_θ that interpolates the equations of the economic model at a finite set of points. We can either choose a grid or sample N points $\mathcal{D} \subset \mathcal{X}$ and minimize the empirical risk (ERM):

$$\min_{\theta \in \Theta} \left\{ \frac{1}{N} \sum_{x \in \mathcal{D}} ||\ell(x, f_\theta)||_2^2 \right\}. \tag{1}$$

Solving for this minimum only requires gradients of a scalar loss.[6] This helps ensure that gradient-based optimization algorithms solving problem (1) do not scale exponentially with $|\theta|$ due to the curse of dimensionality. In fact, optimization algorithms where $|\theta| \gg N$ are typically faster and more reliable than $|\theta| \approx N$ in this class of problems.

Assuming that $\mathcal{H}(\Theta)$ is flexible and highly overparameterized ($|\theta| \gg N$), optimization methods can reliably find a f_θ such that we get interpolation, i.e., $\ell(x, f_\theta) \approx 0$ for all $x \in \mathcal{D}$. However, given the over-parameterization, there are many $\theta \in \Theta$ that could interpolate the data.

Thus, the key question is to characterize toward which solution the ML approximations converge in practice. Rather than just choosing a random interpolating function —which might have an arbitrary degree of overfitting— ML algorithms converge in practice toward the "simplest" interpolating solutions. This is called the inductive bias in the ML literature. The classic interpretation of inductive bias is Ockham's razor, i.e., the simplest solution should be the most likely. However, what does "simple" mean in this context?

Inductive bias and the min-norm solution. One framework for how to define "simplest" is as the min-norm interpolating solution (Belkin, 2021). More formally, when solving ERM with a highly overparameterized approximation, we can (always loosely, and sometimes formally) think

[6] Alternatively, we can solve the underdetermined nonlinear system of equations $\ell(x, f_\theta) = 0$ for all $x \in \mathcal{D}$. However, the Jacobian is enormous and dense, so it is usually more efficient to solve problem (1). Efficiently calculating gradients with automatic differentiation is the core feature of ML libraries such as Pytorch, Tensorflow, and JAX.

of the argmin of problem (1) as equivalent to solving:

$$f_\theta^* \equiv \min_{f_\theta \in \mathcal{H}(\Theta)} \|f_\theta\|_\psi \tag{2}$$

$$\text{s.t. } \ell(x, f_\theta) = 0, \text{ for all } x \in \mathcal{D}, \tag{3}$$

where ψ is some function semi-norm. That is, it finds the simplest function, measured by ψ, that interpolates the data according to condition (3).[7]

In general, we will not be able to characterize ψ directly, as it depends on a combination of $\mathcal{H}(\Theta), \ell(\cdot,\cdot)$, and the optimization method used to solve problem (1). While this is an interpretation of the argmin of problem (1) rather than a direct optimization problem itself, equations (2) and (3) provide a useful intuition we will use throughout this paper.

Inductive bias and the double descent phenomenon. As we mentioned before, we will be dealing with highly parameterized neural networks $\ell(x, f_\theta)$. Classic statistics intuition suggests that having excess parameters makes it more likely to overfit, leading to large $||\ell(x, f_\theta)||_2^2$ outside of $\mathcal{D}$ and excess sensitivity to the details of $\mathcal{H}$ and to initial conditions for the optimization algorithm. Surprisingly, this is not what one finds in practice.

The argmins of problem (1) with neural networks are often nearly the same function for a significant portion of $\mathcal{X} \setminus \mathcal{D}$. To be more precise, for two different initial conditions of $\theta \in \Theta$, the minimization will converge to two different θ_1 and θ_2, but where the functions themselves are in an equivalence class. That is, $f_{\theta_1}(x) \approx f_{\theta_2}(x)$ for a surprisingly large region of $x \in \mathcal{X} \setminus \mathcal{D}$. As long as Θ is large, the exact features of $\mathcal{H}(\Theta)$ become less important, and different approximation architectures behave similarly.

The general theory for why ML methods select min-norm solutions is, thus, connected to the literature on double descent and generalization theory.[8] Double descent is the phenomenon where, counterintuitively, the seemingly inescapable bias-variance tradeoff —where adding too

[7] The precise min-norm of the inductive bias is provable in some cases, such as in overparameterized linear regression and ridgeless kernel regression (Hastie et al., 2022) and kernel methods (Ebrahimi Kahou et al., 2024), and holds for some limiting cases of neural networks (Belkin, 2021, and Ma and Ying, 2021).

[8] The ML literature is still exploring the double descent phenomenon and generalization and their connection to inductive bias. The prevailing view is that double descent arises from a combination of the optimization algorithms used, the intrinsic complexity of the data, and the geometry of the loss function within the parameter space. Since flat minima occupy a disproportionately large volume in high-dimensional spaces, the bias toward these minima is mostly independent of the optimization method employed. See Smith et al. (2021), Chiang et al. (2022), and Zhang et al. (2021). Spiess et al. (2023) provide recent examples of double descent in causal inference econometrics.

many parameters leads to overfitting— seems to disappear if one adds vastly more parameters (even trillions in some large language models!). That is, the cure to overfitting is to keep adding parameters, not eliminating them. While ML practitioners have known of the double descent phenomenon since the late 1980s (Vallet et al., 1989), researchers in computer science and statistics are still exploring its theoretical foundations.

Inductive bias and long-run expectations. For this paper, we take the min-norm interpretation as given and wait for the computer scientists to make progress on characterizing it precisely. But we can take this interpretation as given since it has been robustly confirmed across many algorithms and hypothesis classes $\mathcal{H}$. We know that, with sufficient parameters, ML algorithms are biased to choose the flattest solution among all functions that interpolate the data.[9]

Returning to our primary objective of solving dynamic models and modeling behavioral agents, these flat interpolating solutions correspond to the unique, non-explosive solutions that satisfy long-run boundary conditions. In particular, we will demonstrate in our examples that by solving problem (1) without explicitly imposing the long-run boundary conditions, the min-norm solutions, as interpreted by problem (2), will approximately satisfy these boundary conditions in many cases, even when $\mathcal{D}$ contains no points near a steady state. From a behavioral economics perspective, the ML solution that is consistent with both the laws of motion and Ockham's razor is the one that meets our constraints on the long-run expectations of forward-looking agents.

Simulating ergodic distributions and steady states. Given that the solutions we just discussed only approximately fulfill the long-run conditions, they will only be approximately stable. This is in contrast to linear rational expectations equilibria that are stable by construction. Given that our goal is to show that one can focus on short-run dynamics without directly imposing long-run dynamics, this is not an inherent limitation of ML methods. We build on a long tradition in applied mathematics by assuming that agents may consider a data-generated process for forecasts motivated by Ockham's razor even if it slowly diverges in the long run, as long as they periodically can re-optimize to refine the errors. As we mentioned in the introduction, this is a useful framework for thinking about how behavioral agents solve their optimization problems in practice, i.e., choosing parsimony over stability.

[9] A function is flat when all of its derivatives exist and are zero. Hence, an inductive bias toward flat functions would penalize those derivatives. In that sense, an inductive bias is a form of interpretable regularization.

However, these errors mean that we need to be careful in calculating the steady state (or ergodic distribution) when we do not have provable stability and convergence. Small deviations from stability accumulate in the long run, leading to numerically unstable calculations of the steady state and ergodic distributions, which could feed back into errors in the short-term dynamics, especially if instability biases the $\mathcal{D}$ in problem (1).[10] Nonetheless, in many applications (e.g., the responses to climate change in spatial models), knowing the steady state and ergodic distributions is irrelevant, and ML methods work without any problem.

Examples. To illustrate the theory and its implications, the remainder of this paper examines two classic examples with well-established baselines: (i) a linear asset pricing model, which formally demonstrates the sufficiency of this condition and clarifies its connection to function norms, and (ii) the neoclassical growth model, the canonical forward-looking model with saddle-path dynamics, which further elucidates links to turnpike theory. Although we illustrate the results using deterministic sequence space models, the same principles apply to recursive and stochastic versions.[11] One of the most surprising findings is that this approach remains effective even in the presence of steady-state multiplicity and hysteresis, as seen in the neoclassical growth model with a concave-convex production function. The algorithm succeeds despite being unaware of the existence of multiple steady states or the appropriate domain of attraction of each of them.

Although intentionally low-dimensional, these examples offer valuable insights into how economists can leverage deep learning to solve high-dimensional models while satisfying transversality conditions. Later, we will revisit the broader argument that these conditions stem from behavioral assumptions and propose that inductive bias might lead to economically justifiable —and computable— approaches to equilibrium selection and approximately time-consistent policies for behavioral macroeconomics.

3 Linear Asset Pricing

We start showing our results with the basic model of risk-neutral asset pricing, which has traditionally served a pedagogical role in exploring long-run boundary conditions for models with

[10] Without ensuring stability, simulations may (almost surely) diverge. Furthermore, adding too many points in the fitting process close to the (possibly misspecified) steady state or ergodic distribution would distort the relative error compared to the short-run dynamics of interest.

[11] See Appendix B for the formulation of the neoclassical growth model in a state-space representation.

forward-looking agents (e.g., Ljungqvist and Sargent, 2018). Linearity will help us to illustrate the connection between multiplicity, asymptotic boundary conditions, and norms in function spaces.[12]

3.1 Model

The risk-neutral price, $p(t)$, of a claim to an exogenous stream of dividends, $y(t)$, is given by the recursive equation:

$$p(t) = y(t) + \beta p(t+1), \tag{4}$$

where $\beta \in (0,1)$ is the discount factor and $t = \{0, 1, \cdots, \infty\}$. The interpretation is standard but worth repeating: the price of a claim to the asset today is the period payoff plus the forecast of the discounted price of the asset tomorrow.[13] While the model is cast in discrete time, the notation uses the continuous extension $p : \mathbb{R} \to \mathbb{R}$, careful only to evaluate (4) for discrete t.[14]

Forward-looking behavior and multiplicity. This recursive structure of equation (4) captures the inherent forward-looking nature of the asset pricing problem. The price at time $p(t)$ cannot be known without forecasts of $p(t+1)$ and inductively $p(t+2), \ldots p(t+\infty)$, showing the connection between long-run forecasts and asymptotic boundary conditions.

Given the simple linear structure, one can characterize the set of solutions to (4) as:

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

where $\zeta \geq 0$ and $p_f(t) \equiv \sum_{\tau=0}^{\infty} \beta^{\tau} y(t+\tau)$ represents the price based on fundamentals (i.e., the discounted present value of dividends). The bubble, $p(t) - p_f(t) = \zeta \beta^{-t}$, is explosive for all $\zeta > 0$ since $\beta < 1$. From the perspective of the risk-neutral agent, this multiplicity captures that there are many internally consistent, fully rational forecasts of future prices that fulfill the problem's

[12] See Blanchard (1979) and Brunnermeier (2017) for details. Beyond its pedagogical value, linear asset pricing serves as the foundational framework for studying fiat currency, hyperinflation, and other pure bubbles in economics (e.g., Flood and Garber, 1980, and Brunnermeier, 2017). Bianchi and Nicolò (2021) present a general approach to deal with the problem of indeterminacy in linear rational expectations models.

[13] The requirement on the $y(t)$ primitive is that it does not grow faster than β^{-1} (i.e., $\lim_{t\to\infty} |y(t+1)/y(t)| < \beta^{-1}$) to ensure that present discounted values are well defined. For example, dividends could follow a recursive state-space model, as in Ljungqvist and Sargent (2018), with bounds on the spectral radius of the evolution equation.

[14] ML methods tend to work better in continuous time, as discussed in Fernández-Villaverde et al. (2023) and Ebrahimi Kahou et al. (2024). We stay in discrete time for simplicity.

recursive structure, each mapping to a different-sized bubble at time zero (i.e., $p(0) - p_f(0) = \zeta$).

Asymptotic boundary conditions. An economic interpretation of the multiplicity of solutions to equation (4) is that there are many possible paths where the agent rationally forecasts a bubble growing asymptotically, but only one where the agent's long-run forecasts do not diverge relative to discounting. This leads to the no-bubble condition:

$$0 = \lim_{t \to \infty} \beta^t p(t). \tag{6}$$

Economically, this boundary condition is motivated by stability. Agents who believe the boundary condition (6) reject asset price forecasts, which can perpetually grow faster than the discount rate, leading to constraints on their long-run expectations.[15] With this condition, the system (4) and (6) is now a well-posed problem with the unique solution, $p(t) = p_f(t)$.

This simple model illustrates a central computational challenge in economics. With forward-looking agents, the dynamics of short-term forecasts (e.g., $p(t)$ for $t \ll t_N \equiv \max \mathcal{D}$) can only be made by imposing an asymptotic condition (6). Otherwise, any $p(0) \geq p_f(0)$ is a possible equilibrium price. Because of this, even if we have no intrinsic interest in forecasting long-run dynamics, we nevertheless must consider the entire sequences of $p(t)$ for $\{t = 0, \cdots, \infty\}$, subject to the distant boundary condition (6) to have an internally consistent model of the short run.

Below, we will demonstrate how we may be able to fulfill these long-run, economically important boundary conditions without actually solving for $\lim_{T \to \infty} \beta^T p(T)$ directly.

Stability and function norms. The no-bubble condition is a special case of a broader class of transversality conditions, which often arise out of conditions to ensure stability in optimal control, as we will see in Section 4. In optimal control, a policy is stable if it will not diverge, and transversality conditions ensure that repeated applications of a policy rule fulfill that condition.

Intuitively, the key to our methods is that the solution that fulfills the transversality condition is the stable one, which we see in our case (5). Among all of the possible solutions, the asymptotic boundary condition (6) selects the least explosive one.

[15] See Tirole (1982) and Brunnermeier (2017) for a discussion of how bubble terms violate the transversality condition and its connection to expectations. More broadly, many models directly connect long-run boundary conditions and explosive price paths. For instance, in the monetary models of Brock (1974), Brock (1975), and Obstfeld and Rogoff (1983), the solutions that violate the transversality condition correspond to either explosive paths of asset prices or the relative price of capital.

To formalize a comparison of different functions that solve (4), consider a function semi-norm, i.e., $||p||_\psi$. An important example is a Sobolev norm defined on $t \in [0, T]$ where $||p||^2_{W^{1,2}} \equiv \int_0^T |p'(t)|^2 dt$. In particular, functions with larger derivatives, $p'(t)$, will have larger norms, and explosive functions have exploding norms relative to flat ones since $\lim_{T\to\infty} |p'(T)| = \infty$.[16]

Coming back to our specific problem, explosive solutions will have larger norms than those based entirely on fundamentals. Take the general solution that characterizes solutions without applying a boundary condition in equation (5), recall that $\zeta \geq 0$, and apply the triangle-inequality to compare its norm to the $p_f(\cdot)$ that uniquely solves the problem with a no-bubble condition:

$$\|p\|_\psi \equiv \|p_f + \zeta\beta^t\|_\psi \leq \|p_f\|_\psi + \zeta \|\beta^t\|_\psi. \tag{7}$$

The norm is minimized when $\zeta = 0$, and in that case $\|p_f\|_\psi = \|p\|_\psi$. To see that these are the same function, up to an equivalence class, compare the solutions: $\|p - p_f\|_\psi \equiv \|p_f + \zeta\beta^t - p_f\|_\psi = \zeta\|\beta^t\|_\psi$. Hence, $\zeta = 0$ implies $\|p - p_f\|_\psi = 0$.[17]

We introduced the $W^{1,2}$ norm for illustrative purposes and because it is representative of norms in ML that are biased toward flatness by penalizing the gradients. Since we only used the triangle inequality in the proof, these results hold for any semi-norm ψ. Therefore, we can have some confidence that these methods are not sensitive to the choice of norm ψ in practice.

3.2 Min-Norm Solution

The previous section showed us that the unique solution to the model with the asymptotic boundary condition is the smallest (and hence flattest, given relevant norms that penalize gradients) solution to the dynamic, forward-looking equation (4). Coming back to ML, its inductive bias toward flatter functions fortuitously aligns with both the behavioral assumptions on forward-looking agents and the long-run boundary conditions those assumptions imply. In a divine coincidence, "small" solution functions are both the preferred choice of ML methods and economic assumptions on long-run expectations of agents.

[16] Focusing on a closed interval with a finite T for the norms to exist is innocuous here, and can be relaxed by adjusting our definition of norms. For instance, Van et al. (2007) use exponential discounting of functions in the definition of norms.

[17] If we have a semi-norm, rather than a norm, then it only provides an equivalence class, $\|p_f - p\|_\psi = 0$ rather than ensuring p and p_f are identical pointwise. While this is not an issue in this particular example, it can be relevant when considering two functions in an equivalence class that fulfill the model equations, and long-run boundary conditions could have economically relevant differences. In that case, one needs to adapt the approach.

Next, we examine how well this observation works in practice at selecting the unique, non-explosive solution. To implement the simple algorithm with neural networks, first define an approximation $p_\theta \in \mathcal{H}(\Theta)$ for a highly parameterized neural network. Next, for $\mathcal{X} = [0, \infty)$ choose $\mathcal{D} \equiv \{t_1, \cdots, t_N\} \subset \mathcal{X}$ and minimize (4) numerically:

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{t \in \mathcal{D}} \left[p_\theta(t) - y(t) - \beta p_\theta(t+1) \right]^2, \tag{8}$$

where our baseline example generates dividends using $y(t+1) = c + (1+g)y(t)$, for $y(0) = y_0, c > 0$ and $g \geq -1$. This is just a particular case of problems (1) and (3).

Given the connection between problems (1) and (3), we interpret the solution to problem (8) according to:

$$\begin{aligned} &\min_{p_\theta \in \mathcal{H}(\Theta)} \|p_\theta\|_\psi \\ &\text{s.t. } p_\theta(t) = y(t) + \beta p_\theta(t+1), \text{ for all } t \in \mathcal{D}. \end{aligned} \tag{9}$$

Outside of some specific cases, we do not know the precise ψ that the inductive bias of the ML algorithm leads to, but this interpretation nevertheless provides us with intuition. In fact, we know from equation (7) that the results are not sensitive to the particular semi-norm.[18]

Parameterization. The parameter values of the model are set at $\beta = 0.9$, $c = 0.01$, and $y_0 = 0.08$ and we consider $g = -0.1$ and $g = 0.02$. The first g leads to prices reaching a steady state, while the second leads to prices growing indefinitely, but given that $g < \beta^{-1} - 1$, there is a well-defined price on a balanced growth path (BGP). We do not provide g to the algorithm or calculate a BGP but give our function approximation the ability to scale a neural network output exponentially. More concretely, when $g > 0$, our $\mathcal{H}$ is chosen to be $p_\theta(t) = \exp(\phi t)\text{NN}(t; \theta_{\text{NN}})$, where $\theta \equiv \{\phi, \theta_{\text{NN}}\}, \phi \in \mathbb{R}$, and $\text{NN}(t; \cdot)$ is a neural network to be defined.[19]

Results. For our numerical solutions, we fit problem (8) using a choice of hypothesis space $\mathcal{H}(\Theta)$, where $p_\theta(t) = \text{NN}(t; \theta)$ (or $p_\theta(t) = \exp(\phi t)\text{NN}(t; \theta_{\text{NN}})$). Neural networks serve as a broad class of

[18] See Blanc et al. (2020), Damian et al. (2021), and Ma and Ying (2021) for more on characterizing the approximate function norms of the inductive bias. In some limiting cases, it can be proven to be $W^{1,2}$.

[19] Insights from problem structure will often achieve better generalization (e.g., Ebrahimi Kahou et al., 2021, encode symmetry in the design of $\mathcal{H}(\Theta)$ to improve the generalization of the approximation).

approximators, encompassing many approximations familiar to economists, such as splines and orthogonal polynomials. We employ neural networks because they showcase how our arguments function with a maximally flexible functional form, but inductive bias is present in all ML models.

Our baseline example uses $p(t;\theta) \equiv f(W_1 \cdot \sigma(W_2 \cdot \sigma(W_3 \cdot \sigma(W_4 t + b_4) + b_3) + b_2) + b_1)$ where $f(\cdot) = \log(1 + \exp(\cdot))$ to ensure positivity and $\sigma(\cdot) = \tanh(\cdot)$ is the hyperbolic tangent function applied pointwise. The components of $\theta \in \Theta$ are $W_4 \in \mathbb{R}^{128\times 1}, W_3 \in \mathbb{R}^{128\times 128}, W_2 \in \mathbb{R}^{128\times 128}, W_1 \in \mathbb{R}^{1\times 128}, b_4 \in \mathbb{R}, b_3 \in \mathbb{R}^{128}, b_2 \in \mathbb{R}^{128}$, and $b_1 \in \mathbb{R}$. In ML terminology, this would be referred to as a multilevel perceptron (MLP) with four hidden layers, 128 nodes for the hidden layers, tanh as the activation function, and a final layer of *Sofplus*. To give a sense of the degree of overfitting, we use 30 grid points with 50K parameters, i.e., overparameterized by about four orders of magnitude.[20]

For the grid, we use $\mathcal{D} = 0, 1, 2, \ldots, 29$, where the key performance metric is to compare the solution relative to $p_f(x)$ for $x \in \mathcal{D}$. While our focus is on the short term, we also plot an extrapolation region for $t > 30$ to assess how closely the steady state is forecasted and to gauge stability. Recall that when $g = 0.02$, there is a BGP rather than a steady state. In this case, we check whether the model accurately learns g to enforce the no-bubble condition for the short run (i.e., whether $\log(1 + g) \approx \phi$ given the $p_\theta(t) = \exp(\phi t)\mathrm{NN}(t;\theta_{\mathrm{NN}})$ approximation).

Previously, we asserted that, in practice, inductive bias would yield the same interpolating function up to an equivalence class. Since this statement may be sensitive to accumulated numerical errors, we rerun the optimizer from different initial conditions for θ and report the median, 10th, and 90th percentiles. The primary metric is the relative error, $\varepsilon_p(t) \equiv (p_\theta(t) - p_f(t))/p_f(t)$where our goal is to ensure that the inductive bias results in low errors in the short to medium run.[21]

Figure 1 plots our results. The dotted line represents the closed-form baseline, the solid line depicts the median, and the shaded region illustrates the 10th to 90th percentiles. The top left panel compares the solutions for $g = -0.1$, while the top right panel shows the corresponding relative errors. The bottom left panel contrasts the solutions for $g = 0.02$, and the bottom right panel plots the relative errors.

The inductive bias toward the min-norm solution results in highly accurate approximations in the short run. The approximations perform well even in the extrapolation region for $t > 30$, despite

[20] The results are not especially sensitive to the design of $\mathcal{H}$ as long as the problem dimensionality is high enough and multiple layers are used.

[21] For these examples, we use the Limited Memory Broyden-Fletcher-Goldfarb-Shannon (L-BFGS) optimizer due to its robustness and speed, and calculate the gradients of the objective function (8) using all of $\mathcal{D}$ (i.e., full batch).

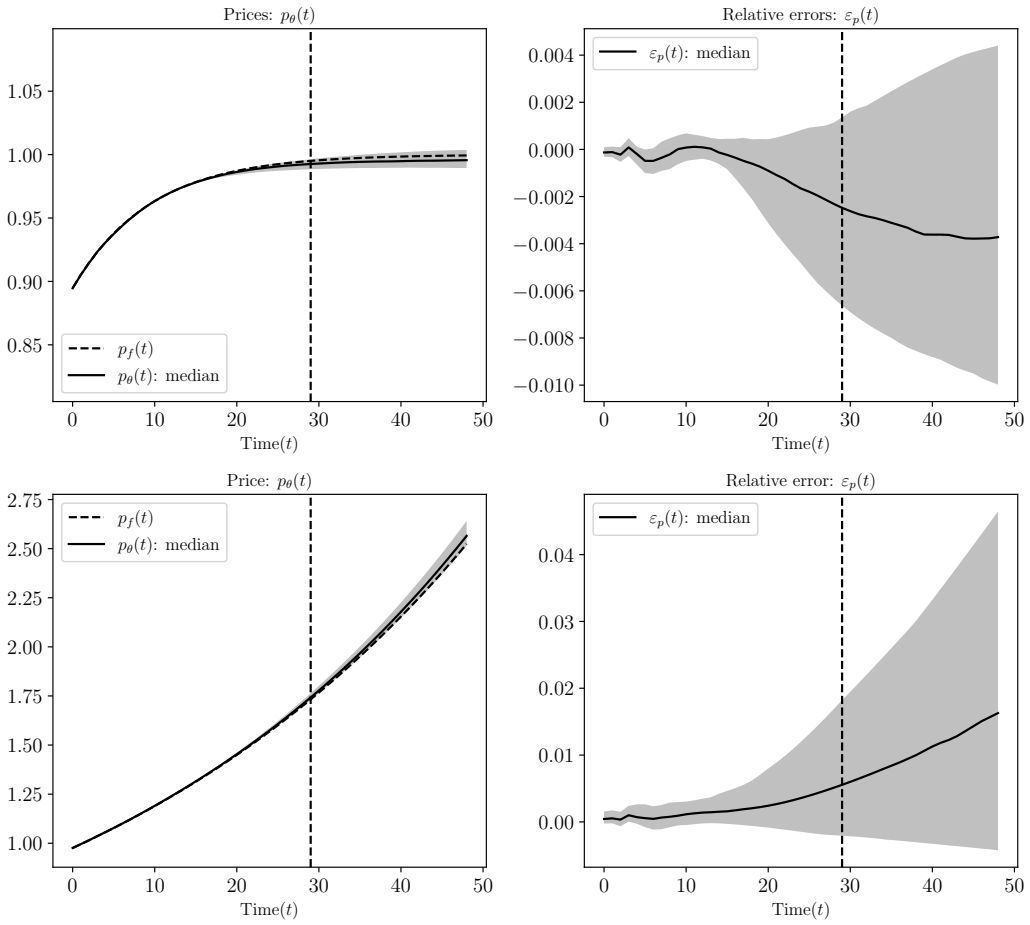

Figure 1: Solutions to problem (8) for an ensemble of 100 initial conditions for θ.

this not being our primary objective. The difference between $p_\theta(t)$ and $p_f(t)$ is imperceptible in the short run, as seen in the two left panels. For the stationary case with $g = -0.1$, the relative errors up to $t = 10$ are within numerical precision on average and well within 0.1% even at the 90th percentile of experiments. Even in the non-stationary case, up to $t = 10$, the solutions are close to numerical precision, and the 90th percentile of errors remains around 0.5%.

Although the dispersion of solutions in the extrapolation region of the BGP is small, it is still present: the median error is approximately 0.5%, with the 90th percentile errors reaching nearly 2% at $t = 30$, as shown in the bottom right panel. Nevertheless, this outcome reinforces the effectiveness of our methods, as it demonstrates that an imperfect characterization of the long run does not result in substantial errors in the short run, even without knowledge of g.

We can also apply the min-norm interpretation to $\mathcal{H}$. By leveraging our economic understanding of the problem's structure, we allowed $\mathcal{H}$ to rescale itself using $p_\theta(t) = \exp(\phi t)\mathrm{NN}(t; \theta_{\mathrm{NN}})$,

where $\phi \in \mathbb{R}$ is learned. In this context, our neural network is inclined to select the ϕ that minimizes the norm of the $\text{NN}(t;\theta)$ function. This result reveals a strong bias toward the closed-form solution $\phi = \log(1+g)$, as it minimizes the explosiveness of $\text{NN}(t;\theta_{\text{NN}})$ when rescaled by its learned ϕ. Other solutions with different ϕ values require more explosive $\text{NN}(t;\theta_{\text{NN}})$ to compensate, leading to higher norms. This highlights that deep learning solutions are not merely black boxes; successful outcomes depend on integrating economic insights into the design of $\mathcal{H}$.[22]

In summary, our results indicate that over-parameterized neural networks learn the steady state and there is strong evidence of the bias toward flat functions. The approximate functions remain flat even in regions where data were not provided during the optimization process.

Behavioral interpretation We can consider our solution method as a behavioral approximation of agents with an inductive bias. That is, agents would train their neural networks with just a few observations (30) and, by applying Ockham's razor, consider that asset pricing trajectories that lead to more stable solutions are the most likely. While the solution does not exactly choose a stable path, as we can see with the slight divergences in the long run in the top panels of Figure 1, it gets the short-term correct to numerical precision. In other words, a behavioral agent does rather well in terms of pricing assets.

Even more surprisingly, the policies are approximately time-consistent. If we re-optimized at $t = 5$, for example, the agent would similarly choose a path with low error in the short run that is (slightly) unstable in the long run. Inductively, with a sufficiently fast re-optimization of the sequential problem, the inductive bias would lead the solution to stay close to the steady state.[23]

4 The Neoclassical Growth Model

The neoclassical growth model serves as a classic example of the importance of transversality conditions in ruling out sub-optimal paths. As in our previous example, we will analyze these conditions with an emphasis on the behavioral foundations of the boundary conditions and their connection to inductive bias. To push the argument further, in Section 4.3, we examine a model

[22] In Appendix A.4, we present a case where the functional form is misspecified.

[23] This is analogous to model predictive control (MPC), a control-theory method commonly used in industry that solves finite-horizon control problems periodically, in contrast to the infinite-horizon open-loop control that finds a single time-invariant policy that is never updated. Under reasonable regularity conditions, the MPC solutions are approximately time-consistent and converge to the open-loop infinite-horizon solution (Mayne et al., 2000).

inspired by Skiba (1978) that features multiple steady states. Despite the added complexity — stemming from the nonlinear nature of the growth model and the need to satisfy both initial values and boundary conditions— the saddle-path characteristics of these problems lead the inductive bias to select even more stable solutions than those discussed in Section 3.

4.1 Model

We follow the standard treatment of the neoclassical growth model in Acemoglu (2008) and Ljungqvist and Sargent (2018) with a log utility. Then, we can jump straight to the dynamic system of equations that characterize the optimal path:

$$k(t+1) = z(t)^{1-\alpha} f\left(k(t)\right) + (1-\delta)k(t) - c(t), \tag{10}$$

$$c(t+1) = \beta c(t)\left[z(t+1)^{1-\alpha} f'\big(k(t+1)\big) + 1 - \delta\right], \tag{11}$$

$$0 = \lim_{t\to\infty} \beta^t c(t)^{-1} k(t+1), \tag{12}$$

given $\beta \in (0,1), \delta \in (0,1)$, and an initial condition $k(0) = k_0$. Equation (10) is the law of motion derived from the resource constraint, equation (11) is a forward-looking Euler equation, and equation (12) is a forward-looking transversality condition that prevents capital from accumulating too fast relative to the marginal utility of consumption, $c(t)^{-1}$.

The total factor productivity (TFP) process, $z(t)$, follows $z(t+1) = (1+g)z(t)$ given a growth rate $0 \leq g < 1/\beta - 1$ and initial condition $z(0) = z_0$. Our baseline production function is $f(k) = k^\alpha$ for $\alpha \in (0,1)$, which has a unique steady state (or BGP) and transition dynamics toward it. Section 4.3 will move toward a concave-convex production.

Forward-looking behavior and saddle-path stability. Given only the initial condition k_0 and equations (10) and (11), the system has multiple steady states with associated transition dynamics unless the transversality condition is imposed. In the simple case of $z(t) = 1$ for all t, there are two possible equilibria: one with $k_{\max}(t)$, where the limit approaches the global maxima of output at $f'(k^*_{\max}) = \delta k^*_{\max}$ at that point $c^*_{\max} = 0$, and another, $k(t)$ and $c(t)$, with interior c^* and k^*, where the economy converges to the saddle-path with positive consumption.[24]

[24] There is a trivial third steady state (which we ignore) with $k = 0$ that is not an attracting basin, cannot be reached unless $k_0 = 0$, and is unstable for any perturbation of k_0. The case where $g > 0$ for the TFP process is similar but requires rescaling by the geometric growth rate.

As in our previous example, while the role of the transversality condition (12) is to ensure the self-consistency of forecasts, it also eliminates multiplicity numerically. The transversality condition eliminates the $\lim_{t\to\infty} k_{\max}(t) = k^*_{\max}$ as a possible steady state because, if $\lim_{t\to\infty} c_{\max}(t) = 0$, then the marginal utility of consumption goes to infinity, which makes the transversality condition (12) impossible to fulfill given that capital accumulation is limited by the resource constraint. While the Euler equation (11) could be consistent with trajectories that have maximal output and no consumption in the limit, a transversality-violating trajectory would require the agent to forecast an (asymptotic) infinite marginal utility of consumption.

Stability and function norms. Consider the case of numerically solving the undetermined system of equations (10) and (11) subject to $k(0) = k_0$, but without imposing the transversality condition (12). As discussed above, there are two possible capital and consumption trajectories, but only one of them fulfills transversality.

To provide intuition for why our methods are successful despite the possible multiplicity of solutions, we need to explain why the trajectories that lead to no consumption and maximal output have a higher norm than the saddle-path solution. The key reason for this, in this formulation of the problem, is that $k^*_{\max} \gg k^*$. In our parameterization, the output maximizing capital level is approximately 10-20 times larger. Consequently, since the output maximizing trajectory, $k_{\max}(t)$, and the saddle-path optimal trajectory, $k(t)$, grow from the same $k(0) = k_0$ initial condition, the trajectories of the output maximizing capital will be steeper. Hence, we would expect that $||k_{\max}||_\psi > ||k||_\psi$ for norms that penalize gradients, as does $\psi = W^{1,2}$.

In other words, among all the solutions of (10) and (11), the optimal one has the smallest norm because of the saddle-path nature of this problem. All sub-optimal solutions are (locally) explosive. In contrast, the optimal solution is non-explosive.[25]

4.2 Min-Norm Solution

As in the case of linear asset pricing, we solve the model without applying the long-run boundary condition and rely on the inductive bias of the ML algorithms to select the min-norm solution.

[25] One can reformulate this problem in terms of the co-state variable (i.e., the marginal utility) to see this with an even more stark result. In this formulation, the sub-optimal paths are globally explosive. See Appendix A.5 for a detailed discussion, especially the right panel of Figure A.5.

First, define an approximation $k_\theta \in \mathcal{H}(\Theta)$ for a highly parameterized neural network. Next, for $\mathcal{X} = [0, \infty)$ choose $\mathcal{D} \equiv \{t_1, \cdots, t_N\} \subset \mathcal{X}$ and minimize (11) subject to $k(0) = k_0$:

$$\min_{\theta \in \Theta} \left[\frac{1}{N} \sum_{t \in \mathcal{D}} \left(\frac{c(t+1; k_\theta)}{c(t; k_\theta)} - \beta \big[z(t+1)^{1-\alpha} f'\big(k_\theta(t+1)\big) + (1-\delta) \big] \right)^2 + \big(k_\theta(0) - k_0\big)^2 \right], \tag{13}$$

where $z(t) = z(0)(1+g)^t$, and consumption is defined as a function of k_θ through the feasibility constraint:

$$c(t; k_\theta) = z(t)^{1-\alpha} f(k_\theta(t)) + (1-\delta) k_\theta(t) - k_\theta(t+1).$$

As before, we will use (2) to interpret solutions to problem (13) as the interpolating solution that minimizes some norm, $||k_\theta||_\psi$.

Results. The baseline parameters are $f(k) \equiv k^\alpha$, $\beta = 0.9, \alpha = 0.33, \delta = 0.1, k_0 = 0.4, z(0) = 1$, and $g = 0$. We choose $\mathcal{D} = \{0, 1, 2, \ldots, 29\}$ and minimize equation (13) with a choice of $\mathcal{H}(\Theta)$. The baseline example uses $k(t; \theta) = \text{NN}(t; \theta)$ for a neural network with four hidden layers, 128 nodes for the hidden layers, tanh as the activation function, and a final layer of *Softplus*.[26]

We solve the ERM using L-BFGS and all of $\mathcal{D}$. The relative errors are $\varepsilon_k(t) \equiv \frac{k_\theta(t) - k(t)}{k(t)}$ and $\varepsilon_c(t) \equiv \frac{c_\theta(t) - c(t)}{c(t)}$. Our goal is to ensure that inductive bias leads to low errors in the short to medium run, even if there may be a small degree of instability in the long run.

Figure 2 shows our results. The left panel plots the median of the approximate solutions for capital and consumption, $k_\theta(t)$ and $c_\theta(t)$, compared to a benchmark (dashed lines) solved with value function iteration. Despite not being provided the transversality condition (12), the approximation always provides the correct dynamics. The right panel shows the median, and the shaded region the 10th to 90th percentiles for the errors relative to the baseline, $\varepsilon_k(t)$ and $\varepsilon_c(t)$.

The errors are close to numerical precision in the short to medium run. In the long-run extrapolation region, where $T > 30$, the relative error grows, albeit slowly, and with a median error of less than 0.1% for $k_\theta(t)$. This confirms that inductive bias will (1) choose the correct trajectories consistent with the transversality condition; (2) small extrapolation errors, in the long run, do not feed back to large errors in the short run; and (3) the solution is almost but not quite, stable. The last point suggests that while this has fairly accurate extrapolation, we should

[26] In the case with $g > 0$, we let the approximation choose to scale a neural network exponentially with $k_\theta(t) = \exp(\phi t)\text{NN}(t, \theta_{\text{NN}})$ where $\theta \equiv \{\phi, \theta_{\text{NN}}\}$. We will not provide the approximation with the g and let it decide whether it wants to normalize the solution

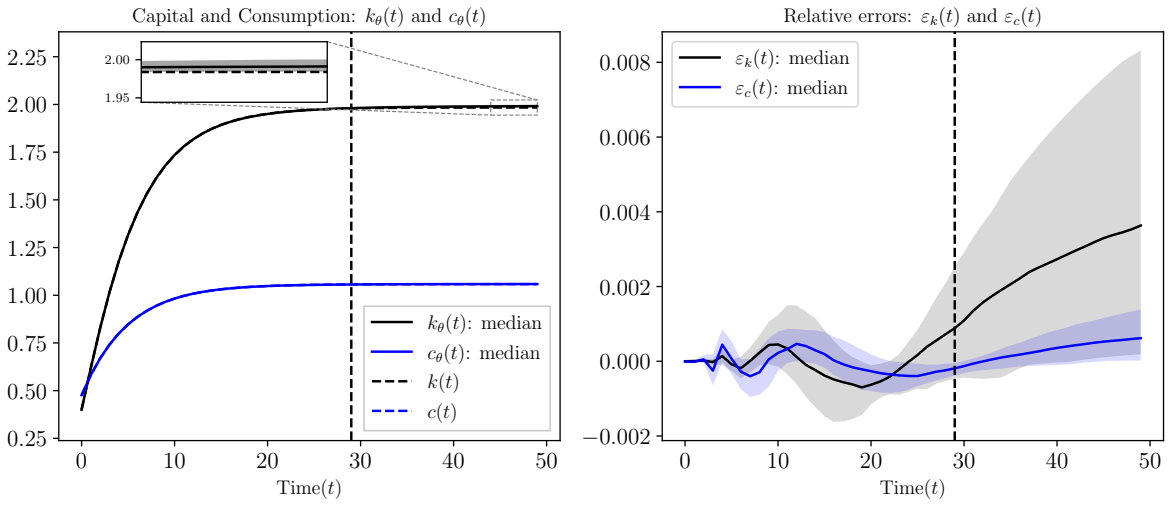

Figure 2: Ensemble of 100 initial conditions for θ solving (13) with $g = 0$ and $k_0 = 0.4$.

be cautious in using ML methods for the simulation of long-run and ergodic distributions. Almost stable is not always good enough for simulating a steady-state or ergodic distribution.

Results with a BGP. Next, we solve the same model with $g = 0.02$ and show that the inductive bias still leads to the correct solution. As before, when we solve a version with a BGP, we will neither manually detrend nor provide g to the approximation.

Figure 3 plots the results. The left panels show the median of the approximate capital and consumption paths, $k_\theta(t)$ and $c_\theta(t)$, with the baseline solutions as dashed lines. The right panels show the median, and the shaded region the 10th to 90th percentiles for the errors relative to the baseline, $\varepsilon_k(t)$ and $\varepsilon_c(t)$.

The short- to medium-run errors are extremely small, and even the extrapolation region has roughly a median error of 0.1% for consumption at $t = 50$ despite the exponential growth and the extrapolation. The inductive bias has led the approximation to choose a min-norm solution and a rescaling despite not being given the growth rate, the transversality condition, or the BGP. The intuition is the same as that for the asset pricing example with growth. If $k(t)$ is approximated by a $\exp(\phi t)\text{NN}(t; \theta_{\text{NN}})$, the ϕ that leads to the smallest norm for the $\text{NN}(t; \theta_{\text{NN}})$ is $\phi \approx \log(1 + g)$. All other ϕ lead to explosive $\text{NN}(t, \theta_{\text{NN}})$ functions as $t \to \infty$. See Appendix A.3 for more details.

State-space formulation. Appendix B shows that the transversality condition (12) is also required in a recursive state-space formulation. In that case, the inductive bias toward min-norm

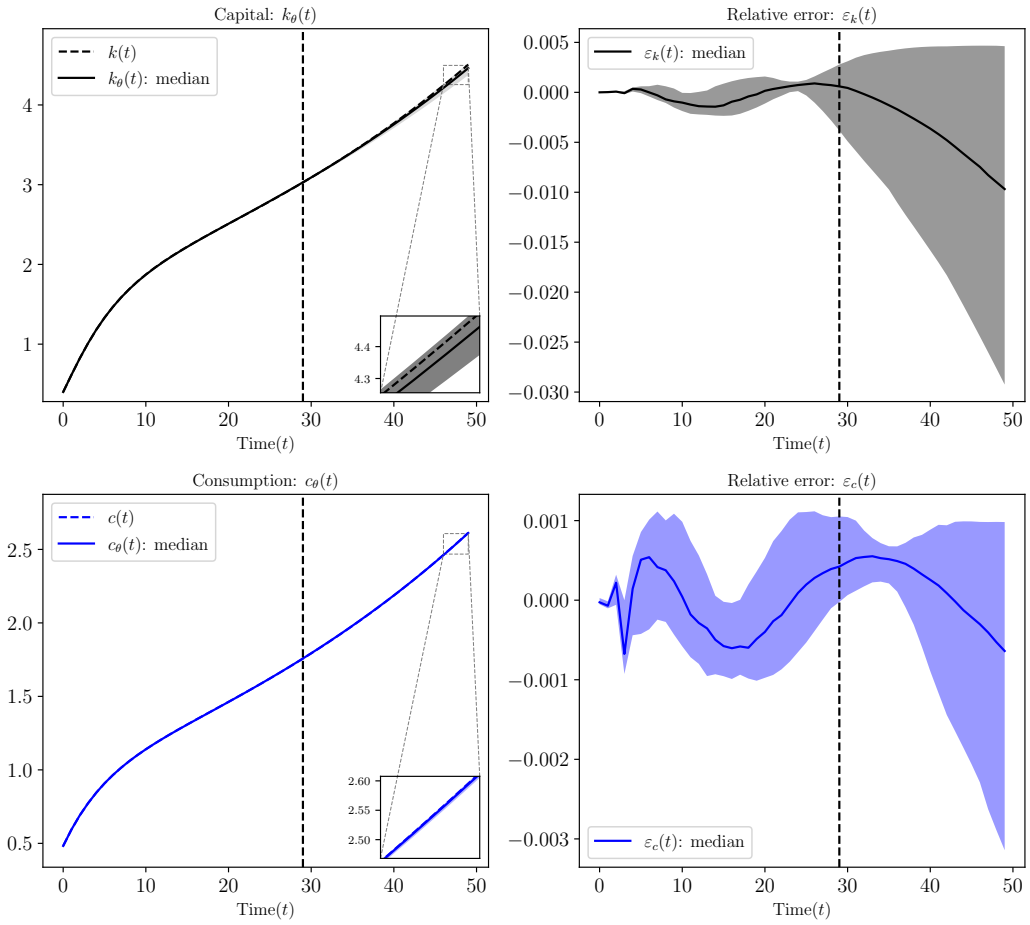

Figure 3: Ensemble of 100 initial conditions for θ solving (13) with $g = 0.02$ and $k_0 = 0.4$.

solutions applies to the policy function of the state space. To sketch out the logic: without applying the transversality condition, multiple solutions are fulfilling the Euler equation and resource constraints for an investment policy $k'(\cdot)$ such that $k_{t+1} = k'(k_t)$. As in the sequential case, we would need to show that the inductive bias toward flat min-norm solutions chooses the correct policy (i.e., that the min-norm solution is the correct one).

To see this, use the Sobelev norm $||k'||_{W^{1,2}} \equiv \int_{k\in\mathcal{X}} ||\nabla_k k'(k)||^2 \mathrm{d}k$ on a compact subset of the domain. Policies that, on average, exhibit larger gradients (i.e., are less flat) will then have larger norms. Hence, higher values of $||k'||_{W^{1,2}}$ will result in greater variations in the underlying state when these policies are applied to iterate trajectories. Specifically, in scenarios where $k_{t+1} = k'(k_t)$, larger norms of policy functions will cause k_t to explode relative to policies with smaller gradients.[27]

[27]In principle, higher norms of $||k'||_{W^{1,2}}$ could lead to a more volatile k_t, but our model does not exhibit cyclical behavior. The inductive bias here is sensitive to the formulation of the problem. For example, if one approximates the $c(\cdot)$ function instead of $k'(\cdot)$, the min-norm solution may not be sufficiently separated from the other solutions that violate transversality. See our more complete discussion in Appendix A.5.

Robustness. Appendix A.1 shows that we achieve nearly as accurate short- to medium-run forecasts with a sparser and irregular grid, even with as few as nine points, using the same number of parameters for $\mathcal{H}(\Theta)$.

Another practical concern is whether it is necessary, as it is with shooting and similar approaches, to pick $t_N \equiv \max\{\mathcal{D}\}$ close to the steady state but not so large that the solutions diverge due to numerical instability. Appendix A.2 demonstrates that the methods are still accurate when fit with $\mathcal{D} = \{0, 1, \cdots, 9\}$, well below the point of convergence to the steady state. The short-run errors are a fraction of a percent and only reach a median relative error of roughly 0.25% at the extrapolation threshold.

Appendix B.3 provides an example of where these methods are unsuccessful, which is illustrative of the importance of choosing the right problem formulation. Recall that the key to the success of inductive bias was that $k^*_{\max} \gg k^*$, which helped ensure that $||k_{\max}||_\psi \gg ||k||_\psi$. Hence, it was relatively easy for optimization algorithms to use inductive bias to regularize and choose $k(t)$ rather than the $k_{\max}(t)$ that fulfills the Euler equation but not transversality. Instead, if we approximate using $c(t)$, $c^*_{\max} = 0$ and $c^* > 0$ may not be cleanly separated (and hence $||c||_\psi$ will be harder to disentangle from the $||c_{\max}||_\psi$ through regularization). The consequence is that solving the formulation with $c(t)$ is more likely to find the incorrect solutions that fulfill the Euler equations but not transversality.[28]

Finally, Appendix A.4 fits a misspecified function form of a BGP version and shows that it still achieves a good approximation even when given an incorrect functional form.

Behavioral interpretation and turnpikes. The interpretation of inductive bias from the perspective of behavioral macro is that the agents choosing the $k(t)$ policy are biased toward the one that is less explosive.[29] A more specific interpretation of the neoclassical growth model is that the inductive bias is toward solutions on the turnpike.

Turnpike theorems (see McKenzie, 1976, and Marimón, 1989) show that models with long

[28] See Appendix B.3. The key to these methods is to scale and approximate the system so that failures in transversality result in an explosive norm, which ML algorithms can easily correct through inductive bias. In this case, writing the model in terms of $k(t)$ or the marginal utility of consumption, $u'(c(t)) = 1/c(t)$, would work well. The most useful guidance comes from the transversality condition $\lim_{t\to\infty} \beta^t \lambda(t) k(t) = 0$. This suggests that formulating the problem using the co-state variable, $\lambda(t)$, is often the most reliable approach.

[29] This statement is loose because the trajectories are only locally explosive since the resource constraint limits the growth of capital in the long run, albeit at a high level. Other formulations, such as solving for the marginal utility $u'(c) = c^{-1}$ or co-state variables, would have explosive trajectories in the limit.

—but finite— horizons have almost the same short- to medium-run dynamics as infinite-horizon models. Even if trajectories have local transition dynamics at $t = 0$ and the terminal point T, for a large enough time frame, it is optimal to remain close to the time-invariant steady state except close to 0 and T. The implication is that in order to solve finite-horizon models, one can find the turnpike to use as a boundary condition and solve for the transition dynamics from $t = 0$.

Relating this to our paper, inductive bias favors transition dynamics that tend toward the turnpike without explicitly characterizing it, given that the turnpike trajectory is the unique path that does not diverge. The tradeoff we face is that, despite having very accurate local dynamics from $t = 0$, ML methods may not precisely identify the turnpike. A policy at $t = 0$ might be insufficiently stable under commitment. If the agent never re-evaluated her policy, she would fail to stay on the turnpike. However, since our focus is on short-run dynamics, the imperfect time consistency of long-run policies is not a significant concern. From a behavioral perspective, periodically recalculating these solutions to refine local dynamics would ensure stability.

4.3 Multiple Steady-States and Hysteresis

When there are multiple steady states, each with its domain of attraction, how would inductive bias move us toward the correct steady state for a given initial condition?[30]

To show how ML methods work in practice, we solve the same model as before but replace the production function with $f(k) \equiv \max\{k^\alpha, b_1 k^\alpha - b_2\}$ for $b_1 > 1$ and $b_2 > 0$ as inspired by Skiba (1978). In that case, there are two sets of steady states, denoted (k_1^*, c_1^*) and (k_2^*, c_2^*), with different domains of attraction. As before, without applying transversality, there will be a $(k_{\max}^*, 0)$ that would solve the problem with vanishing consumption, but inductive bias eliminates it.

Let $a = 0.5, g = 0, b_1 = 3$, and $b_2 = 2.5$. This parameterization leads to steady states $k_1^* = 2.75$ and $k_2^* = 4$. The model is solved, as before, by choosing a $\mathcal{D}$ and minimizing function (13), where the only change relative to the previous method is the new $f(\cdot)$ and $f'(\cdot)$. In particular, we do not provide the algorithm with any hints that there are multiple steady states.[31]

Figure 4 plots the results of this experiment for various initial conditions crossing between the

[30] To understand the challenge, consider how this problem is solved using classical methods: (1) solve the spectral problem with the Jacobian to find potential fixed points; (2) use the second-order conditions to determine which points are attracting basins; (3) identify the domains of attraction to divide the state space; and (4) for a given initial condition, select the appropriate steady state as the boundary condition.

[31] Unlike the previous examples, we solve the ERM problem with the Adam optimizer, which is slower than L-BFGS but introduces more inductive bias.

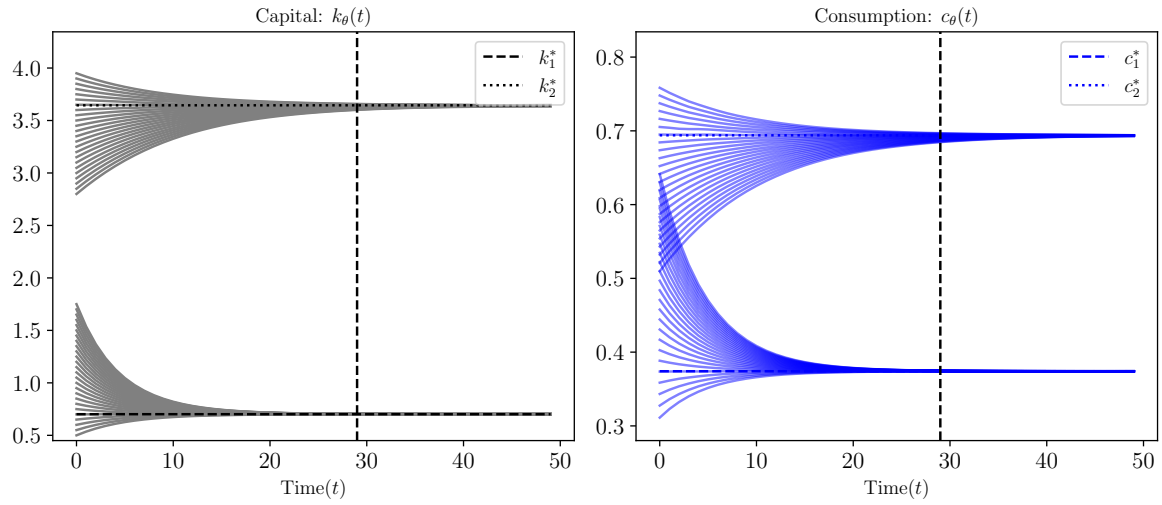

Figure 4: Solutions to (10) and (11) with a convex-concave production function.

two domains of attraction. The left panel shows the capital paths for various initial conditions for capital, while the right panel shows the corresponding consumption paths. For clarity, we do not split the trajectories close to the boundary between the two regions (which occurs at the unstable root) since this approach is not intended to sharply characterize basins of attraction.[32]

The inductive bias chooses solutions converging to the correct set of steady states even in the presence of steady-state multiplicity. ML accurately generates transition dynamics from each initial condition toward the appropriate domain of attraction.

To understand why ML chooses the correct steady states, notice that there are two forces at play. Inductive bias plays a role in both of them. First, the discontinuity in the marginal product of capital would introduce a discontinuity in the Euler equation (11) were trajectories to pass between the two regions. Discontinuities in the Euler equation lead to large changes in $k(t)$, which the inductive bias avoids. Second, any trajectories moving toward the wrong steady state would require steeper transitions since they first need to move closer to the domain of attraction. Regardless of the source of higher gradients, an inductive bias toward flat solutions rejects those trajectories. The success of this experiment opens up the possibility of solving complex economic models with significant hysteresis and multiple steady states, as we see in many spatial models.

[32]In practice, it seems to get close to the boundaries between the regions but eventually makes mistakes. We urge caution in cases where the domains of attraction are very close. In that case, traditional methods are required.

5 Conclusion

This paper presents a theoretical framework and two applications to explore how ML methods can solve short-term transition dynamics while maintaining consistency with forward-looking expectations. The central insight is that transversality conditions —essential boundary conditions that ensure the consistency of forward-looking expectations— can be approximately satisfied through an inductive bias toward flat and simple interpolating solutions, in line with Ockham's razor. Beyond its role in this paper, inductive bias is also fundamental to the double descent phenomenon in ML. It underpins the success of deep learning across a wide range of problems in economics and other fields.

While we demonstrated these methods using simple, interpretable, deterministic problems with well-established theory and reference implementations, the success of this approach suggests that ML may help solve high-dimensional problems that are otherwise infeasible due to the curse of dimensionality. However, the findings also caution against hastily adopting ML methods to solve models without reference implementations. First, satisfying transversality conditions can be sensitive to problem formulation. Second, nearly stable solutions may still fall short in numerically simulating ergodic distributions and steady states without periodic policy refinement. However, in many situations, this is not a concern since the steady state or ergodic distributions are not of interest in themselves.

In our core results, we emphasized that long-run boundary conditions are behavioral assumptions shared by both agents and economists to ensure self-consistent dynamics. Specifically, we solve for policy functions from the agents' perspective that satisfy Euler equations and other intertemporal conditions, given a well-specified process for future dynamics. The agents' inductive bias leads them to select solutions consistent with stable long-run expectations, ultimately resulting in a self-consistent rational expectations equilibrium. This suggests a future direction where ML models could serve as the foundation for behavioral macroeconomics, with the evolution of the underlying environment and state space being both dynamic and learned.

References

Acemoglu, D. (2008): *Introduction to Modern Economic Growth*, Princeton University Press.

Azinovic, M., L. Gaegauf, and S. Scheidegger (2022): "Deep Equilibrium Nets," *International Economic Review*, 63, 1471–1525.

Barnett, M., W. Brock, L. P. Hansen, R. Hu, and J. Huang (2023): "A deep learning analysis of climate change, innovation, and uncertainty," Papers 2310.13200, arXiv.org.

Belkin, M. (2021): "Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation," *Acta Numerica*, 30, 203–248.

Belkin, M., D. Hsu, S. Ma, and S. Mandal (2019): "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences of the United States of America*, 116, 15849–15854.

Bianchi, F., S. C. Ludvigson, and S. Ma (2022): "Belief distortions and macroeconomic fluctuations," *American Economic Review*, 112, 2269–2315.

Bianchi, F. and G. Nicolò (2021): "A generalized approach to indeterminacy in linear rational expectations models," *Quantitative Economics*, 12, 843–868.

Blanc, G., N. Gupta, G. Valiant, and P. Valiant (2020): "Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process," *Proceedings of Machine Learning Research vol*, 125, 1–31.

Blanchard, O. J. (1979): "Speculative bubbles, crashes and rational expectations," *Economics Letters*, 3, 387–389.

Blanchard, O. J. and C. M. Kahn (1980): "The solution of linear difference models under rational expectations," *Econometrica*, 48, 1305–1311.

Brock, W. A. (1974): "Money and growth: The case of long run perfect foresight," *International Economic Review*, 750–777.

——— (1975): "A simple perfect foresight monetary model," *Journal of Monetary Economics*, 1, 133–150.

Brunnermeier, M. K. (2017): "Bubbles," in *The New Palgrave Dictionary of Economics*, Palgrave Macmillan, 1–8.

Caplin, A., M. Dean, and J. Leahy (2022): "Rationally inattentive behavior: Characterizing and Generalizing Shannon entropy," *Journal of Political Economy*, 130, 1676–1715.

Chiang, P.-y., R. Ni, D. Y. Miller, A. Bansal, J. Geiping, M. Goldblum, and T. Goldstein (2022): "Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent," in *The Eleventh International Conference on Learning Representations*.

Cruz, J.-L. and E. Rossi-Hansberg (2023): "The economic geography of global warming," *Review of Economic Studies*, 91, 899–939.

Damian, A., T. Ma, and J. D. Lee (2021): "Label noise SGD provably prefers flat global minimizers," *Advances in Neural Information Processing Systems*, 34, 27449–27461.

Duarte, V., D. Duarte, and D. Silva (2024): "Machine learning for continuous-time finance," Working paper, CESifo.

Ebrahimi Kahou, M., J. Fernández-Villaverde, J. Perla, and A. Sood (2021): "Exploiting symmetry in high-dimensional dynamic programming," Working Paper 28981, National Bureau of Economic Research.

Ebrahimi Kahou, M., J. Yu, J. Perla, and G. Pleiss (2024): "How inductive bias in machine learning aligns with optimality in economic dynamics," Tech. Rep. 2406.01898, arXiv.org.

Ekeland, I. and J. A. Scheinkman (1986): "Transversality conditions for some infinite horizon discrete time optimization problems," *Mathematics of Operations Research*, 11, 216–229.

Evans, G. W. and S. Honkapohja (2001): *Learning and Expectations in Macroeconomics*, Princeton University Press.

Fernández-Villaverde, J., S. Hurtado, and G. Nuño (2023): "Financial frictions and the wealth distribution," *Econometrica*, 91, 869–901.

Flood, R. P. and P. M. Garber (1980): "Market fundamentals versus price-level bubbles: The first tests," *Journal of Political Economy*, 88, 745–770.

Gabaix, X. (2014): "A sparsity-based model of bounded rationality," *Quarterly Journal of Economics*, 129, 1661–1710.

——— (2020): "A behavioral New Keynesian model," *American Economic Review*, 110, 2271–2327.

——— (2023): "Behavioral macroeconomics via sparse dynamic programming," *Journal of the European Economic Association*, 21, 2327–2376.

Han, J., Y. Yang, and W. E (2022): "DeepHAM: A global solution method for heterogeneous agent models with aggregate shocks," Tech. Rep. 2112.14377, arXiv.org.

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2022): "Surprises in high-dimensional ridgeless least squares interpolation," *Annals of Statistics*, 50, 949.

Jungerman, W. (2023): "Dynamic Monopsony and Human Capital," Working Paper.

Kamihigashi, T. (2005): "Necessity of the transversality condition for stochastic models with bounded or CRRA utility," *Journal of Economic Dynamics and Control*, 29, 1313–1329.

Kase, H., L. Melosi, and M. Rottner (2022): "Estimating nonlinear heterogeneous agents models with neural networks," Discussion Paper 17391, CEPR.

Klein, P. (2000): "Using the generalized Schur form to solve a multivariate linear rational expectations model," *Journal of Economic Dynamics and Control*, 24, 1405–1423.

Ljungqvist, L. and T. J. Sargent (2018): *Recursive Macroeconomic Theory*, MIT Press, 4 ed.

Lucas, R. and T. J. Sargent (1981): *Rational Expectations and Econometric Practice*, vol. 2, University of Minnesota Press.

Ma, C. and L. Ying (2021): "On linear stability of SGD and input-smoothness of neural networks," *Advances in Neural Information Processing Systems*, 34, 16805–16817.

Maliar, L., S. Maliar, and P. Winant (2021): "Deep learning for solving dynamic economic models." *Journal of Monetary Economics*, 122, 76–101.

Marimón, R. (1989): "Stochastic turnpike property and stationary equilibrium," *Journal of Economic Theory*, 47, 282–306.

Mayne, D. Q., J. B. Rawlings, C. V. Rao, and P. O. Scokaert (2000): "Constrained model predictive control: Stability and optimality," *Automatica*, 36, 789–814.

McKenzie, L. W. (1976): "Turnpike theory," *Econometrica*, 841–865.

Obstfeld, M. and K. Rogoff (1983): "Speculative hyperinflations in maximizing models: Can we rule them out?" *Journal of Political Economy*, 91, 675–687.

Payne, J., A. Revei, and Y. Yang (2024): "Deep learning for search and matching models," Tech. Rep. 4768566, SSRN.

Sargent, T. J. (1993): *Bounded Rationality in Macroeconomics*, Oxford University Press.

——— (2024): "Macroeconomics after Lucas," Working Paper.

Sargent, T. J. and N. Wallace (1973): "The stability of models of money and growth with perfect foresight," *Econometrica*, 1043–1048.

Skiba, A. K. (1978): "Optimal growth with a convex-concave production function," *Econometrica*, 527–539.

Smith, S. L., B. Dherin, D. Barrett, and S. De (2021): "On the origin of implicit regularization in stochastic gradient descent," in *International Conference on Learning Representations*.

Spiess, J., G. Imbens, and A. Venugopal (2023): "Double and single descent in causal inference with an application to high-dimensional synthetic control," Working Paper 31802, National Bureau of Economic Research.

Tirole, J. (1982): "On the possibility of speculation under rational expectations," *Econometrica*, 1163–1181.

Vallet, F., J.-G. Cailton, and P. Refregier (1989): "Linear and Nonlinear Extension of the Pseudo-Inverse Solution for Learning Boolean Functions," *Europhysics Letters*, 9, 315.

Van, C. L., R. Boucekkine, and C. Saglam (2007): "Optimal control in infinite horizon problems: a Sobolev space approach," *Economic Theory*, 32, 497–509.

Zhang, C., S. Bengio, M. Hardt, B. Recht, and O. Vinyals (2021): "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, 64, 107–115.

Spooky Boundaries at a Distance: Technical Appendix

Mahdi Ebrahimi Kahou[1] Jesús Fernández-Villaverde[2]
Sebastián Gómez-Cardona[3] Jesse Perla[4] Jan Rosa[4]

August 12, 2024

Appendix A Robustness

This appendix contains additional robustness results for the neoclassical growth model of Section 4 in the main text.

A.1 Sparse Grids

In our baseline example, we choose $\mathcal{D} \equiv \{0, \dots 29\}$ and minimize equation (13) to find a $k_\theta(t)$ where $|\theta| \approx 40,000$. Alternatively, we use a sparser set of grid points and interpolate when $t \notin \mathcal{D}$. In particular, consider a grid with more points close to the area with high curvature and fewer closer to the steady state, $\mathcal{D}^{\text{Sparse 1}} \equiv \{0, 1, 2, 4, 6, 8, 12, 16, 20, 24, 29\}$, and another grid with fewer points spread evenly over the domain, $\mathcal{D}^{\text{Sparse 2}} \equiv \{0, 1, 4, 8, 12, 16, 20, 24, 29\}$.

Figure A.1 shows the results of these two experiments for an ensemble of 100 initial conditions. The left panel compares the benchmark solution, $k(t)$, relative to the $k_\theta(t)$ for $\mathcal{D}^{\text{Sparse 1}}$ and $\mathcal{D}^{\text{Sparse 2}}$. The right panel compares the benchmark $c(t)$ against the corresponding $c_\theta(t)$. In both cases, the shaded areas show the 10th and the 90th percentiles.

The distribution of the relative error of $k_\theta(t)$ is small, even in the extrapolation region. In the case of $c_\theta(t)$, the error is so small that the 10th and 90th percentile ranges are not visible. This experiment establishes that we can achieve very accurate solutions with sparse grids, even though

[1]Bowdoin College, [2]University of Pennsylvania, [3]Morningstar, and [4]University of British Columbia, Vancouver School of Economics.

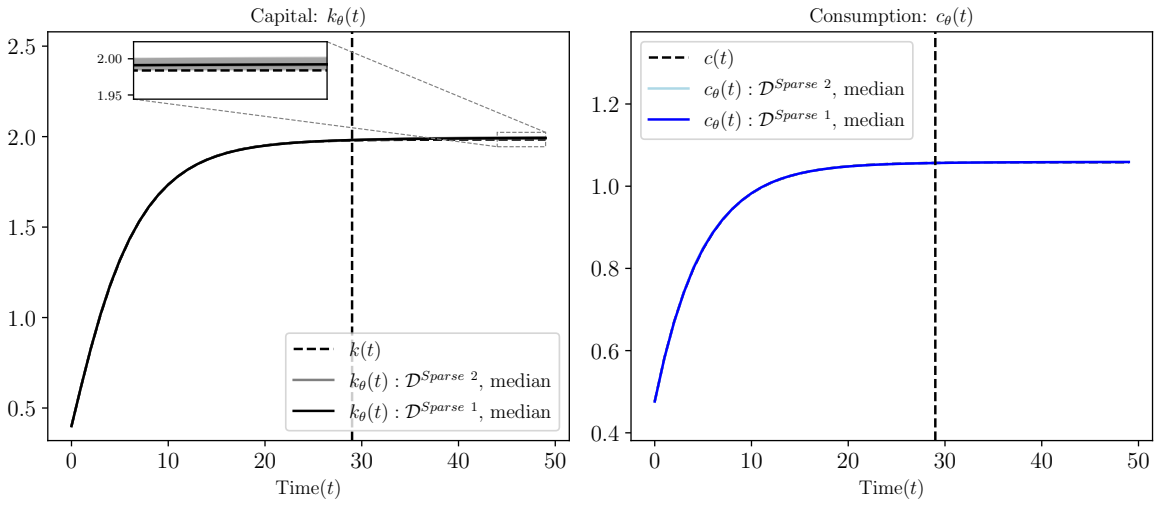

Figure A.1: Solutions to equation (13) with $\mathcal{D}^{\text{Sparse 1}}$ and $\mathcal{D}^{\text{Sparse 2}}$.

the problem remains overparameterized by around four orders of magnitude. ML algorithms do not intrinsically require a large amount of data as long as they have a strong inductive bias.

A.2 Solving on a Short Horizon

A challenge in solving for transition dynamics of models with classic algorithms, such as shooting methods, is the difficulty in choosing the T at which point the solution is close to a steady state. If T is too small, we move toward the steady state too quickly. If T is too large, numerical instabilities can accumulate as the solution iterates forward. Choosing the value of T is an art and requires a good prior on the speed of convergence for a particular model.

To test whether this concern holds with our methods, we solve our model by minimizing equation (13) with the same $\mathcal{H}(\Theta)$, but choose $\mathcal{D} \equiv \{0, 1, \cdots, 9\}$. Not only are there few grid points, but the $t_N = 9$ is far below the point of convergence to the steady state.

Figure A.2 shows the results of this experiment for an ensemble of 100 initial conditions. The left panel shows the median of the approximate capital paths, denoted by $k_\theta(t)$ and the benchmark solution. The right panel shows the median of the approximate consumption paths, denoted by $c_\theta(t)$ and the benchmark solution. The shaded areas represent the 10th and 90th percentiles.

The conclusion is that for the short- to medium-run dynamics, the solutions are very accurate, and the lack of grid points close to the steady state does not feed back to large errors in the short run (as it would with a shooting method). The extrapolation errors are larger than in the baseline

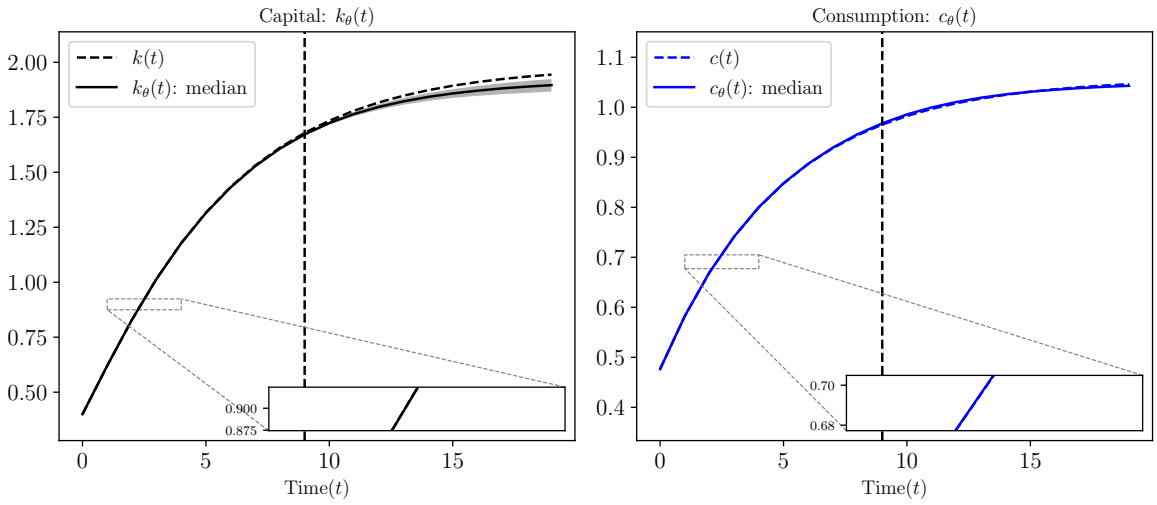

Figure A.2: Solutions to equation (13) with $\mathcal{D} \equiv \{0, 1, \cdots, 9\}$.

case, but getting the long run right was not the goal of the exercise. As discussed, extrapolating and simulating to the steady state is dangerous in general because these solutions are not provably stable. This experiment suggests that the ML methods relying on the inductive bias are not very sensitive to choosing data close to the steady state as long as they are not used to extrapolate too far out of the sample.

A.3 Learning the Scaling Factor

When designing the $\mathcal{H}(\Theta)$ with a BGP, we added in a learnable rescaling: $k_\theta(t) = \exp(\phi t)\text{NN}(t, \theta_{\text{NN}})$, where $\theta \equiv \{\phi, \theta_{\text{NN}}\}$. Given a $\mathcal{D}$ with a large maximum value t_N, the min-norm solution for $\text{NN}(t; \theta_{\text{NN}})$ is achieved by setting $\phi = \log(1 + g)$—at which point $\text{NN}(t; \theta_{\text{NN}})$ could be non-explosive. However, if t_N is relatively small, then we would not expect the approximation to exactly choose the $\phi = \log(1 + g)$ case. A smaller ϕ might yield a lower norm $\text{NN}(t; \theta_{\text{NN}})$ for interpolating a particular $\mathcal{D}$. How well, then, does the algorithm learn g?

Taking the results of Figure 3, which generated solutions using 100 initial conditions, Figure A.3 plots a histogram of the approximated ϕ and compares them to the true growth rate, $g = 0.02$. The results show that the min-norm is biased toward smaller growth rates, as we might expect. However, the solutions in Figure 3 are still extremely accurate. The variations in ϕ within Figure A.3 have compensated changes to $\text{NN}(t; \theta_{\text{NN}})$. A very accurate approximation of the growth rate is not necessary to achieve accurate short- and medium-run dynamics.

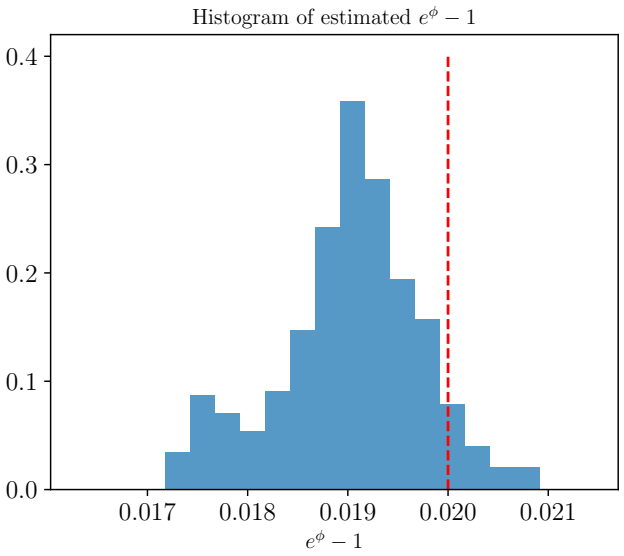

Figure A.3: The distribution of the learned $e^{\phi} - 1$ for the ensemble of 100 seeds used in Figure 3; $g = 0.02$, shown as the dashed line.

A.4 Learning a Misspecified $\mathcal{H}(\Theta)$

In Figure 3, we used economic insights to choose a $\mathcal{H}(\Theta)$ that included a term for exponential growth. Is it still helpful to suggest problem structure when designing $\mathcal{H}(\Theta)$ if the suggestion is misspecified?

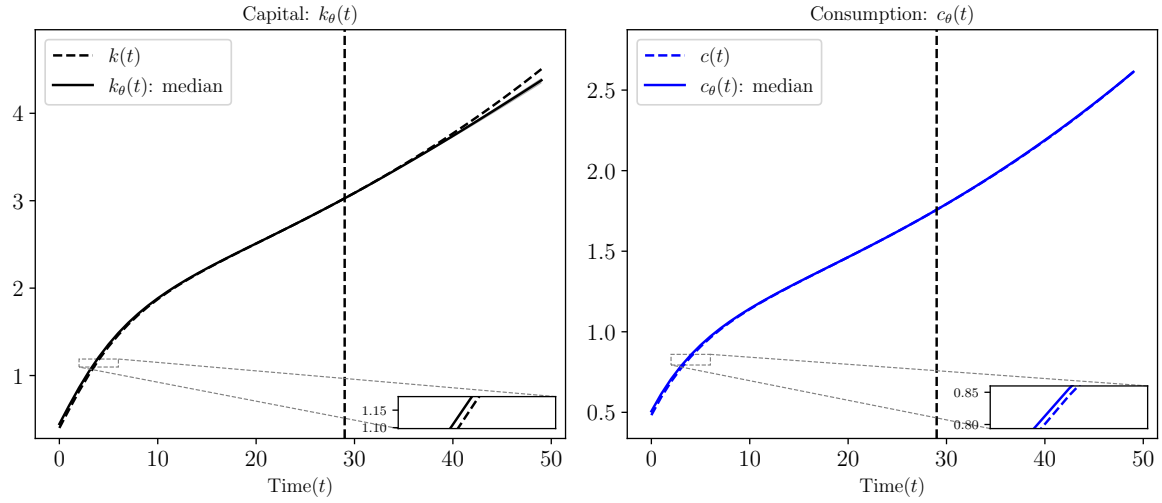

Figure A.4: Solutions to problem (13) with the misspecified $k_\theta(t) = t \cdot \mathrm{NN}(t;\theta) + k_0$ and $g = 0.02$.

To analyze this case, we solve a version where the scaling is assumed to be linear rather than exponential. In particular, $k_\theta(t) = t \cdot \mathrm{NN}(t;\theta) + k_0$. The linear scaling allows some degree of

growth but as $t_N \to \infty$, the NN$(t;\theta)$ would still need to have an infinite norm in order to capture the true dynamics of the BGP.

Figure A.4 displays the solutions to problem (13) with this specification for 100 initial conditions. The left panel shows the benchmark and the median of the solution for capital, while the right panel does the same for consumption. Although the 10th and 90th percentiles are included, they are so close to each other that they remain indistinguishable even after zooming in.

Compared to the well-specified case of Figure 3, the long-run extrapolation slowly diverges (and would continue to do so for any finite t_N), but this does not cause any issues for the short- and medium-run dynamics.

A.5 Function Norms and the Transversality Condition

Section 4 characterized the set of functions fulfilling the Euler equation and resource constraints as (i) $k_{\max}(t), c_{\max}(t)$, with steady states $k^*_{\max}$ such that $f'(k^*_{\max}) = \delta$ and $c^*_{\max} = 0$; and (ii) $k(t), c(t)$ with interior steady states k^* and c^*. The transversality condition (12) eliminated the first solution to prevent the marginal utility of consumption, $u'(c) = c^{-1}$, from becoming infinite.

When relying on the inductive bias of the function norms in lieu of the transversality condition, we must argue that $||k_{\max}||_\psi > ||k||_\psi$ for a large class of norms, ψ. To see this, Figure A.5 plots the two solutions to the under-determined system. The blue curves show a set of capital, consumption, and marginal utility paths, denoted respectively by $k_{\max}(t)$, $c_{\max}(t)$, and $u'(c_{\max}(t))$, that violate the transversality condition. The black curves show the optimal paths that satisfy the transversality condition and that eventually converge to k^*, c^*. Focusing on the left panel, we see that the path of the $k_{\max}(t)$ function has much steeper changes than that of $k(t)$. Therefore, for a large class of norms and semi-norms, which penalize either the average level or gradients, we have $\|k_{\max}\|_\psi > \|k\|_\psi$.

The middle and right panels of Figure A.5 also provide intuition on why these methods can be fragile to the right formulation. While $\|k_{\max}\|_\psi > \|k\|_\psi$ for a large norm given the big spread between k^* and $k^*_{\max}$, this is not the case for $c(t)$. If a norm penalized the gradients (e.g., $\int_0^T |c'(t)|dt$), then the norms of $||c_{\max}||_\psi$ and $||c||_\psi$ would be similar. If the level enters the norm, it may even bias the solution toward the wrong answer (i.e., where $c^*_{\max} = 0$). The right panel shows the other extreme, where using the marginal utility makes an even starker difference between the two solutions. The general advice, true for both the sequential formulation and the state-

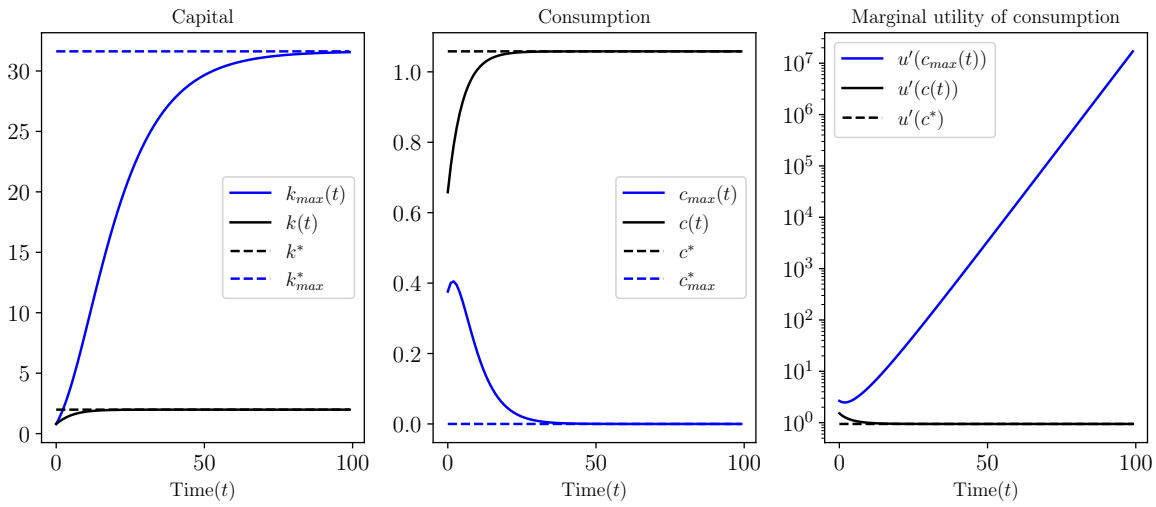

Figure A.5: Comparison between the optimal solution and those violating the transversality condition.

space version, is that it is best to approximate functions that are most explosive if they violate transversality. Co-state variables are the best; state variables often work well, but jump variables are often bounded in a way that makes min-norm solutions harder to disentangle. We see that a similar issue holds in Appendix B.3 for the recursive formulation.

Appendix B State-Space Formulation

This appendix describes the recursive state-space formulation of the neoclassical growth model, in contrast to the sequence-space baseline of Section 4. Inductive bias will serve a similar role in providing a sufficiency condition for transversality, but it will involve norms of the policy functions rather than the trajectories themselves.

B.1 Model

For the state-space $(k, z) \in \mathbb{R}^2_+$, equations (10) and (11) become:

$$u'(c) = u'(c')\beta\left[z'^{1-\alpha} f'(k') + 1 - \delta\right] \tag{B.1}$$

$$k' = z^{1-\alpha} f(k) + (1-\delta)k - c \tag{B.2}$$

where k', c', and z' are the next period capital, consumption, and TFP, respectively, and $u(c) = \log c$. All model primitives and parameters remain the same as in the baseline. The transversality condition (12) must hold for all initial conditions in the state-space formulation:

$$0 = \lim_{T\to\infty} \beta^T u'\left(c_T(k_0, z_0)\right) k_{T+1}(k_0, z_0) \quad \text{for all } (k_0, z_0) \in \mathbb{R}^2_+. \tag{B.3}$$

In this notation, $k_{T+1}(k_0, z_0)$ requires iterating the $k'(\cdot, \cdot)$ policy and $z' = (1+g)z$ law of motion $T+1$ times from (k_0, z_0). Consumption, $c_T(k_0, z_0)$, is found by first iterating to find (k_T, z_T) and then using equation (B.2) to calculate $c_T = z_T^{1-\alpha} f(k_T) + (1-\delta)k_T - k'(k_T, z_T)$.

Transversality with classic methods. The iteration of the policy $k'(\cdot, \cdot)$ in equation (B.3) links stability and transversality. If $k'(\cdot, \cdot)$ was explosive —e.g., $|\nabla_k k'(k, z)| > 1$ for k and z above some threshold— capital would explode until it asymptotically approached the capital maximizing the BGP (or $k^*_{\max}$ if $g = 0$) via equation (B.1). This, in turn, would lead to an infinite marginal utility of consumption in equation (B.3), violating transversality.

In practice, classical methods do not apply the transversality condition as a limit and instead enforce it indirectly in several ways:

- For sequence-space methods, a steady state is found (perhaps after detrending the BGP), which is then used as a terminal boundary condition with shooting methods. Those approaches implicitly use the transversality condition when solving for the correct steady state.

- Linear rational expectations models and LQ control, such as those in Blanchard and Kahn (1980) and Klein (2000), select the non-explosive root via spectral methods.

- With global solution methods, such as projection and collocation, the transversality is implicitly fulfilled by restricting the domain for the state space. For example, in the growth model, we might approximate with Chebyshev polynomials on a compact hypercube on $[k_{\min}, \bar{k}] \times [z_{\min}, \bar{z}]$. If we chose $\bar{k} < k^*_{\max}$ and $k_{\min} < k^*$, then policy functions violating transversality are rejected since they cannot fulfill the Euler equation before hitting corners. Alternatively, by bounding $c \geq c_{\min} > 0$, algorithms implicitly reject functions that fail transversality by bounding the marginal utility of consumption, $u'(c) \leq u'(c_{\min}) < \infty$.

In low dimensions, where we have a strong prior on the relevant regions of the state space, economists can artfully tinker to ensure that a compact hypercube is placed at the appropriate location and does not contain the solutions violating transversality. Moreover, by plotting the dynamics of the model, we can see when simulations diverge (see Fernández-Villaverde et al., 2016, p.10).[1]

However, this process is not feasible in high dimensions since we cannot constrain ourselves to a compact hypercube and may not have a good prior on the location of a steady state. Even evaluating whether transversality conditions are fulfilled for a given policy is computationally infeasible since it requires iterating the policy function for all initial conditions.

Notice here the connection to the issue of stability in ML methods. Simple forward iterations can accumulate numerical errors and be numerically unstable when the solution is only "approximately" stable. This phenomenon appears even in small models.

B.2 Min-Norm Solution

We approximate the capital policy, $k'_\theta(\cdot,\cdot) \in \mathcal{H}(\Theta)$, using a highly parameterized neural network. Choose $\mathcal{D} \subset \mathbb{R}^2_+$ with N points and minimize the equivalent of equation (13):

$$\min_{\theta\in\Theta} \frac{1}{N} \sum_{(k,z)\in\mathcal{D}} \left[\frac{u'\big(c(k,z;k'_\theta)\big)}{u'\big(c(k'_\theta(k,z),(1+g)z;k'_\theta)\big)} - \beta\big[(1+g)zf'\big(k'_\theta(k,z)\big)+1-\delta\big] \right]^2. \tag{B.4}$$

Consumption is defined through the feasibility constraint for a given policy for capital $k'_\theta(\cdot,\cdot)$:

$$c\big(k,z;k'_\theta\big) \equiv f(k) + (1-\delta)k - k'_\theta(k,z). \tag{B.5}$$

Following the interpretation of ERM as a minimum norm solution, we can think of solutions to equation (B.4) as finding:

$$\min_{k'_\theta\in\mathcal{H}(\Theta)} ||k'_\theta||_\psi \tag{B.6}$$

$$\text{s.t. } \frac{u'\big(c(k,z;k'_\theta)\big)}{u'\big(c(k'_\theta(k,z),(1+g)z;k'_\theta)\big)} = \beta\big[(1+g)zf'\big(k'_\theta(k,z)\big)+1-\delta\big], \quad \text{for all } (k,z)\in\mathcal{D}. \tag{B.7}$$

[1] This is part of the appeal of perturbative solutions, which are provably stable even in high dimensions (if properly pruned).

The norm in problem (B.6) typically depends on gradients due to its bias toward flat solutions. For example, it might have properties similar to those of a Sobolev norm $||k'_\theta||^2_{W^{1,2}} \equiv \int ||\nabla k'_\theta(k,z)||_2^2 dF(k,z)$ for some measure F over the state space, or on a compact subset of the domain.

To informally argue why this bias would choose the non-explosive solution, consider iterating the policy function $k_{t+1} = k'_\theta(k_t, z_t)$. A bias toward solutions with smaller gradients with $|\nabla_k k'_\theta(k,z)| < 1$ for large k will lead to policies that have smaller changes in capital, $k_{t+1} - k_t$. If a steady state exists, it will reach the $k_t \approx k'_\theta(k_t, z_t)$ fixed point. Iterating forward with the policy, the bias leads to trajectories that fulfill the transversality condition (B.3). In Appendix B.3. we demonstrate this by plotting the $k'_\theta(\cdot,\cdot)$ for the trajectories that fulfill the Euler equation with and without transversality.

Results. We solve the minimization problem (B.4) for $\beta = 0.9$, $\alpha = 0.33$, $\delta = 0.1$, $g = 0$, $z_0 = 1$, and $k_0 = 0.4$. In our baseline case, $\mathcal{D}$ is a uniform grid of 16 points between $k_1 = 0.8$ and $k_{N_k} = 2.5$. When $g \neq 0$, we can use a grid $\mathcal{D} \equiv \{k_1, \cdots, k_{N_k}\} \times \{z_1, \cdots, z_{N_z}\}$ of $N = N_z \times N_k$ total points, but the methods could use sampled or simulated points in the state-space. The design of $\mathcal{H}(\Theta)$ is a neural network $\text{NN}(k, z; \theta)$ identical to the sequential version of the model, except that it takes two inputs (k, z) rather than the univariate t. As before, we solve with the L-BFGS optimization algorithm, which is fast and requires little tuning.

Figure B.1 shows the median of solutions for capital (top row) and consumption (bottom row) for an ensemble of 100 initial conditions. The consumption path $\tilde{c}(t)$ is calculated with equation (B.5) given the trajectory of the state space. The benchmark solutions, $k(t)$ and $c(t)$, are obtained using value function iteration. The left panels show the median of the approximate capital, $\hat{k}(t)$, and consumption, $\hat{c}(t)$, paths, along with the benchmark solutions (i.e., $\hat{k}(t)$ and $\hat{c}(t)$ are the results of iterating the solution from a particular initial condition). The right panels show the median of the relative errors for capital, $\varepsilon_k(t) \equiv (\hat{k}(t) - k(t))/k(t)$, and consumption, $\varepsilon_c(t) \equiv (\hat{c}(t) - c(t))/c(t)$. The shaded regions show the 10th and 90th percentiles. The gray region in the top-left panel shows the interpolation region, defined as the convex hull of $\mathcal{D}$. The dashed parts of the curves show the median of the relative errors in the extrapolation region. The shaded regions show the 10th and 90th percentiles of the solutions for the 100 random seeds for optimization of θ.

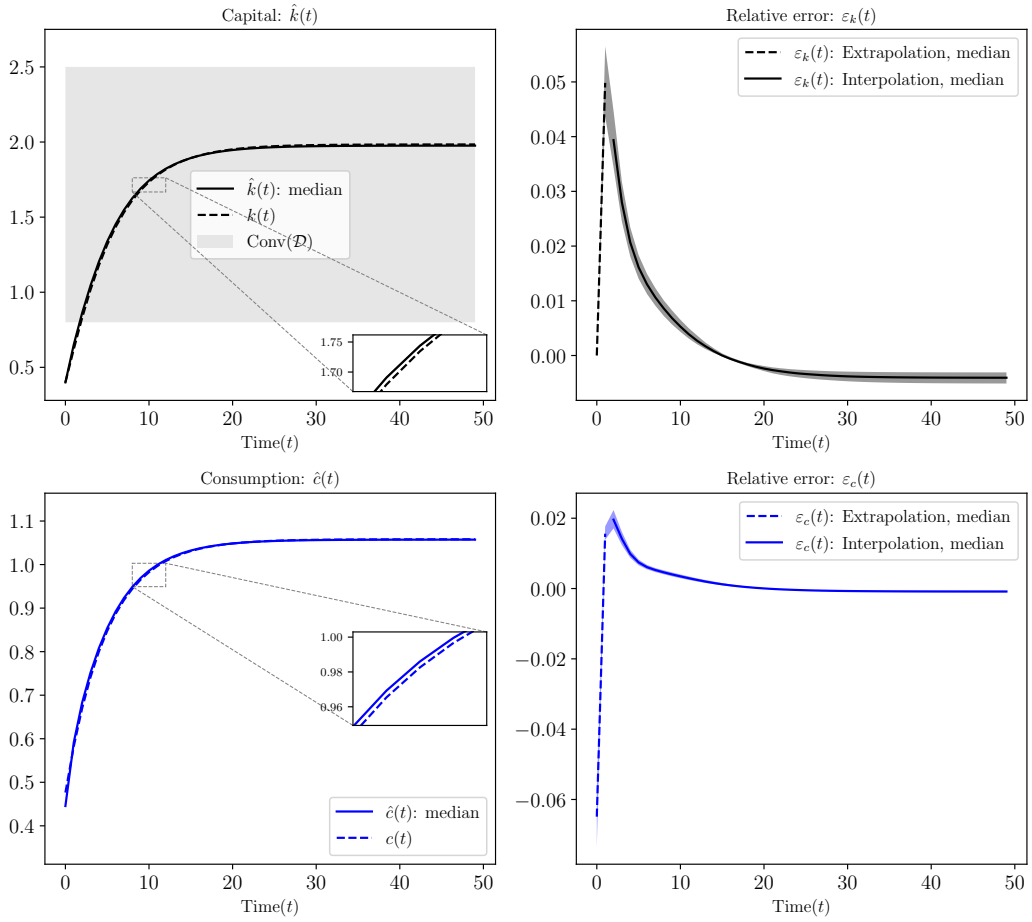

Figure B.1: Solutions obtained by solving problem (B.4) for $g = 0$.

The results show that the inductive bias rules out solutions that violate the transversality conditions in all cases and achieves a good approximation despite only using 16 data points. Even when k_0 is outside the minimum value of $\mathcal{D}$, the errors are small. An inductive bias leads to good generalization behavior even outside of the convex hull of Conv($\mathcal{D}$).

BGP. Since we know that the solution will be homothetic when $g = 0.02$, we now design $\mathcal{H}(\Theta)$ as $k'_\theta(k, z) \equiv z \cdot \text{NN}(k/z, z; \theta)$. We set $\mathcal{D}$ as the cartesian product of 16 points in $[0.8, 3.5]$ for capital with 8 points in $[0.8, 1.8]$. As before, using a small $\mathcal{D}$ highlights the strength of the inductive bias. This implementation minimizes the problem (B.4) with different $\mathcal{H}(\Theta)$ and $\mathcal{D}$ for 100 seeds on the initial condition for the optimizer.[2]

Figure B.2 shows the results for a simulated trajectory from $k_0 = 0.4$ and $z_0 = 1$ and compares

[2]In the exactly homothetic case, we could further simplify this to a univariate NN($k/z; \theta$), but we leave in the z parameter as a check for cases that are almost homothetic and as a further check that the inductive bias avoids overfitting.

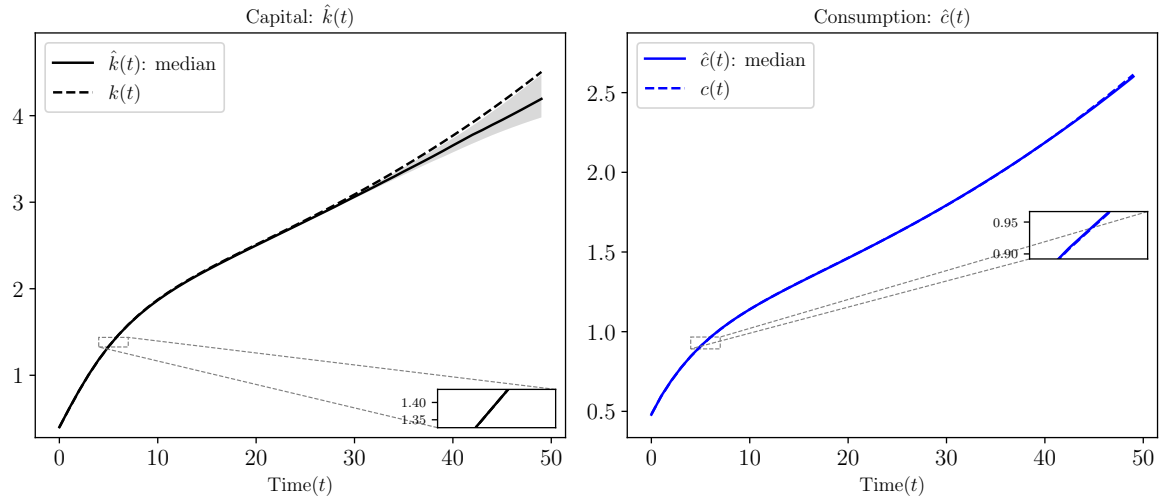

Figure B.2: Solutions to problem B.4 for $g = 0.02$.

the dynamics given the benchmark solution. The left panel shows the median of the approximate capital path, denoted by $\hat{k}(t)$. The right panel shows the median of the approximate consumption path, denoted by $\hat{c}(t)$. The shaded regions show the 10th and 90th percentiles.

The results indicate that, even in the case of growing TFP, the solution is very accurate in the short run, and the differences relative to the benchmark are difficult to see even after zooming in on the graph. The long-run extrapolation is less accurate than in the benchmark (where we could manually rescale due to homotheticity). In other words, we can obtain very accurate short- and medium-run solutions, even though the initial condition for capital lies outside the interpolation region.

B.3 Failures of Euler Residuals Minimization

Appendix A.5 discussed the importance of choosing the right formulation of the problem to ensure that the inductive bias toward min-norm solutions would select the solution that fulfills transversality. This issue is often even more stark in state-space formulations of the problem. Understanding this phenomenon is especially important before we move toward high-dimensional problems in macroeconomics, where failures of transversality are less obvious.

We demonstrate this problem by comparing an equivalent formulation of the neoclassical growth model where we approximate $c_\theta(k, z)$ to our previous results in Figures B.1 and B.2. The inductive bias toward min-norm solutions will consistently choose the wrong solution that violates

transversality.

Let $z = 1$ and $g = 0$ for simplicity, approximate $c_\theta(k) \in \mathcal{H}(\Theta)$ with a neural network, and implicitly define the investment choice as $k'\big(k; c_\theta\big) \equiv f(k) + (1-\delta)k - c_\theta(k)$. The equivalent to the ERM objective function (B.4) becomes:

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{k \in \mathcal{D}} \underbrace{\left[\frac{u'\big(c_\theta(k)\big)}{u'\Big(c_\theta\big(k'(k; c_\theta)\big)\Big)} - \beta \left[f'\big(k'(k; c_\theta)\big) + 1 - \delta \right] \right]}_{\equiv \varepsilon_E^c(k)}^2 . \tag{B.8}$$

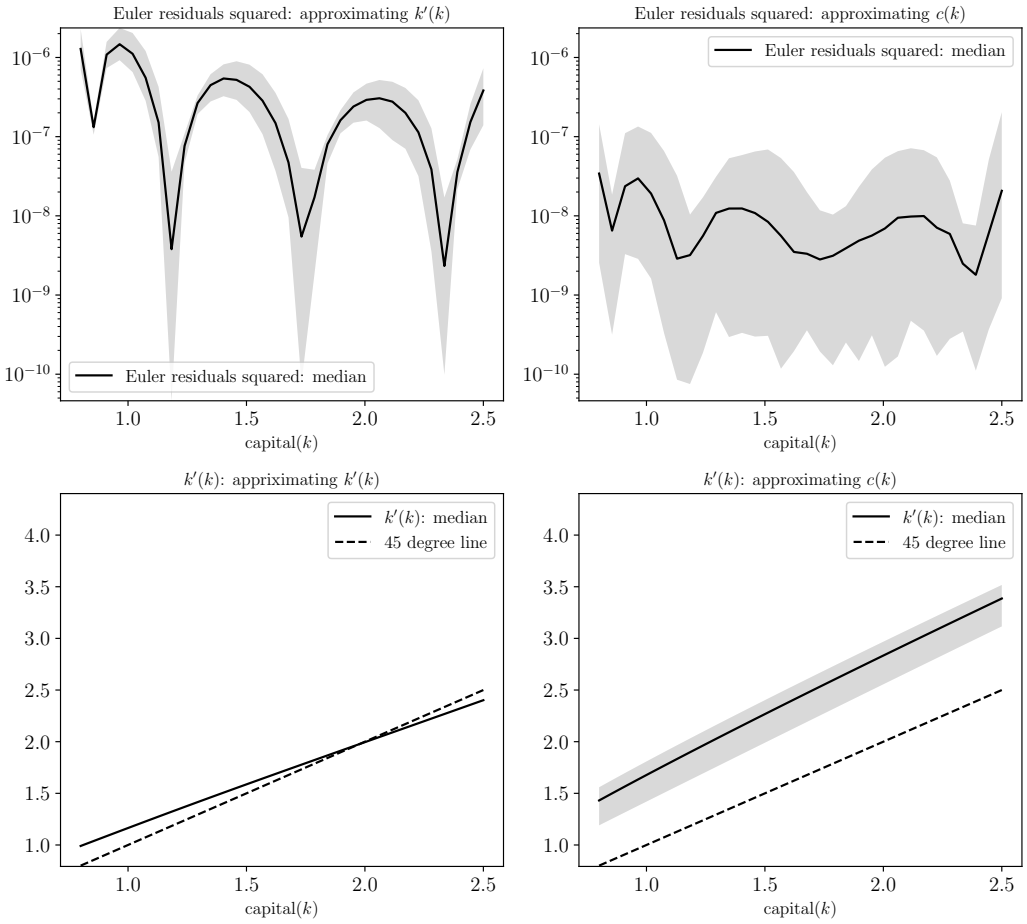

Figure B.3: Comparison between approximating the policy function for capital $k'(k)$ vs. the consumption function $c(k)$ with a deep neural network.

Figure B.3 shows the comparison between approximating the policy function for capital $k'_\theta(\cdot)$ vs. approximating the consumption function $c(\cdot)$ with a deep neural network.[3] The left panels

[3]Primitives and parameters are identical to our baseline case. Given the parameters, the steady-state solution fulfilling transversality is $k^* \approx 2.0$.

show the results using the baseline k'_θ approximation, as in Figure B.1, but plots the Euler error in the top panel and the policy function $k'_\theta(k)$ in the bottom panel. The $k'(k)$ in the bottom panel crosses the 45-degree line around $k^* \approx 2.0$, which is the closed-form steady state. The top right panel instead plots the square Euler error when approximation the c_θ function, while the bottom right panel plots the implied $k'(k; c_\theta)$ policy from $k'\big(k; c_\theta\big) \equiv f(k) + (1-\delta)k - c_\theta(k)$. The solid curves show the medians, and the shaded regions show the 10th and 90th percentiles over 100 different seeds.

Approximating the consumption functions with a neural network leads to solutions that violate the transversality condition. Given the c_θ approximation, the squared Euler residual error, $\varepsilon^c_E(k)$, is defined in equation (B.8) and when approximating with k'_θ, an equivalent definition of $\varepsilon^k_E(k)$ exists from equation (B.4). The Euler errors in both cases are very small and close to numerical precision, so the optimizer has a solution that interpolates the Euler equation and implicitly fulfills the resource constraint on $\mathcal{D}$. If anything, the Euler errors are smaller for the c_θ approximation. However, the bottom right panel does not have the $k'(k)$ intersecting the 45-degree line. It has chosen a c_θ such that $\nabla_k k'(k; c_\theta) > 1$ for all k. This leads to explosive $\tilde{k}(t)$ trajectories and fails the transversality condition in all cases.

The reason why the inductive bias works in the wrong direction in this formulation can be seen if we return to the middle panel of Figure A.5. The consumption trajectory that violates transversality converges to 0 and would have a smaller norm for many ψ that penalizes the level of the function. Even without penalizing the level, the slope of the solution fulfilling transversality is not systematically smaller in absolute value.

To conclude, low Euler (or value-function) errors are insufficient to ensure that an ML algorithm has successfully solved the problem, and inductive bias with the wrong problem formulation might systematically choose the policy that violates transversality. The broad advice is to ensure that the problem is formulated in a way that violations of transversality lead to explosive behavior (e.g., diverging states or formulating in terms of the marginal utility or co-state variables).

References

Blanchard, O. J. and C. M. Kahn (1980): "The solution of linear difference models under rational expectations," *Econometrica*, 48, 1305–1311.

Fernández-Villaverde, J., J. Rubio-Ramírez, and F. Schorfheide (2016): "Solution and estimation methods for DSGE models," in *Handbook of Macroeconomics*, ed. by J. B. Taylor and H. Uhlig, Elsevier, vol. 2, 527–724.

Klein, P. (2000): "Using the generalized Schur form to solve a multivariate linear rational expectations model," *Journal of Economic Dynamics and Control*, 24, 1405–1423.