# Solving Models of Economic Dynamics with Ridgeless Kernel Regressions

MAHDI EBRAHIMI KAHOU
Department of Economics, Bowdoin College

JESSE PERLA
Vancouver School of Economics, University of British Columbia

GEOFF PLEISS
Department of Statistics, University of British Columbia and Vector Institute

We propose a ridgeless kernel method for solving infinite-horizon, deterministic, continuous-time models in economic dynamics, formulated as systems of differential-algebraic equations with asymptotic boundary conditions (e.g., transversality). Traditional methods enforce the asymptotic boundary conditions by targeting a known steady state, an approach that is numerically unstable, hard to tune, and hard to apply with multiple steady states or hysteresis. Instead, our approach solves the problem without imposing the transversality conditions, using regularization to select the unique solution fulfilling transversality. Ridgeless kernel methods recover this path by selecting the minimum-norm solution, which coincides with the non-explosive trajectory. We provide theoretical guarantees showing that kernel solutions satisfy the asymptotic boundary conditions without imposing them directly, and we prove that they are consistent estimators of the true solution. Finally, we illustrate the method in low- and high-dimensional models, including multiple steady states and hysteresis, and compare it with deep learning.

KEYWORDS. Kernel Methods, Economic Dynamics, Machine Learning, Growth, Inductive Bias.

JEL CLASSIFICATION. E13, C45, C63.

## 1. INTRODUCTION

This paper proposes using ridgeless kernel regression to solve a broad class of infinite-horizon, deterministic, continuous-time models in economic dynamics. The main computational challenge in these models is satisfying asymptotic boundary conditions, typ-

Mahdi Ebrahimi Kahou : m.ebrahimikahou@bowdoin.edu
Jesse Perla: jesse.perla@ubc.ca
Geoff Pleiss: geoff.pleiss@stat.ubc.ca

[1]We thank Misha Belkin, Amy Greenwald, William Jungerman, Lilia Maliar, Christian Matthes, Hikaru Saijo, Felipe Schwartzman, and James Yu. We are especially grateful to Rachel Childers for comments and discussions. We gratefully acknowledge support from a SSHRC Insight Grant number 435-2023-0119. Replication materials are provided at https://github.com/HighDimensionalEconLab/kernels.

ically arising from transversality conditions in the embedded optimal control problem. We show that, unlike shooting algorithms that aim toward a known steady state, kernel methods can satisfy these asymptotic conditions by selecting the least explosive among all candidate trajectories, without even calculating the steady state, let alone imposing it as a condition. This yields robust and computationally efficient algorithms that are often more stable than classic methods such as shooting or solving implicit boundary-value problems, even under steady-state multiplicity and hysteresis. Furthermore, even when it is less precise than neural networks (NNs), the kernel method provides a formal framework for characterizing solutions to dynamic problems in a function space, which can be useful for both theory and econometrics.

Beyond capturing forward-looking behavior, asymptotic boundary conditions are essential for well-posedness in the sense of Hadamard (1902, 1932), which requires existence, uniqueness, and continuous dependence on initial conditions. Without them, the system may admit a continuum of trajectories satisfying both the initial conditions and the differential-algebraic equations (DAEs), thereby violating well-posedness.

The central idea of this paper is to treat the problem as underdetermined: impose initial conditions and laws of motion, but not the asymptotic boundary conditions, and then use implicit or explicit regularization to select among the resulting trajectories.[2] For the class of problems we focus on, the unique valid solution is the only non-explosive one, and all other trajectories fail transversality. Therefore, the job of regularization is to rule out the explosive ones.

By choosing an appropriate function norm and penalizing it, kernel methods recover this non-explosive solution, and thus satisfy the asymptotic boundary conditions in a way amenable to formal characterization.

*Kernel Methods.* Kernel methods are a natural fit for this task since they provide a framework for defining and penalizing function norms through the theory of the Reproducing Kernel Hilbert Space (RKHS) induced by the choice of the kernel. In particular, we focus on ridgeless kernel methods (Belkin et al., 2019), which select the minimum-norm solution among all functions that perfectly satisfy the DAE at a finite set of points, the training data.

A "kernel machine" is a non-parametric approximation in which a function is represented as a weighted sum of a kernel centered at each data point.[3]

*Related Work.* Our approach is connected both to traditional methods for solving linearized systems and to a recent literature that uses machine learning (ML) to solve nonlinear optimal control problems.

Perturbation solutions to models of economic dynamics use classic stability analysis from linear-quadratic (LQ) control to find solutions that satisfy asymptotic boundary conditions (e.g., Blanchard and Kahn, 1980). These methods exploit the linearity of time-invariant policies in LQ control, ensuring stability by ruling out explosive roots inconsistent with transversality and selecting the unique stabilizing solution. Our paper draws

[2] See Tikhonov (1963) and Tikhonov and Arsenin (1977) for classic treatments of regularization in ill-posed problems. Our approach is also related to implicit and explicit regularization in deep learning (Ebrahimi Kahou et al., 2024).

[3] See Rasmussen and Williams (2006) and Murphy (2022) for more details on kernel methods, Reproducing Kernel Hilbert Spaces, and the Representer Theorems.

inspiration from this broader approach: algorithms that select non-explosive roots yield solutions that lie on the solution manifold and automatically satisfy transversality conditions.

The use of machine learning methods is becoming increasingly popular for solving and estimating economic models. Applications span a wide range, including wealth inequality (Han et al., 2026), financial frictions (Fernández-Villaverde et al., 2023), the heterogeneous impacts of climate change (Barnett et al., 2023), portfolio choice problems (Azinovic and Žemlička, 2023), consumption-saving with multiple durable goods (Druedahl and Røpke, 2026), heterogeneous agent New Keynesian models (Kase et al., 2022), human capital accumulation in the labor market (Jungerman, 2023, Adenbaum et al., 2024), and labor market dynamics in search and matching environments (Payne et al., 2024). These models often rely on NNs (Azinovic et al., 2022, Maliar et al., 2021, Ebrahimi Kahou et al., 2021, 2024, Azinovic-Yang and Žemlička, 2025), but sometimes include kernel methods in the form of Gaussian processes (Scheidegger and Bilionis, 2019). Outside economics, kernel and Gaussian-process methods are used to solve nonlinear partial differential equations (Chen et al., 2021), but without the asymptotic boundary conditions that are central here.

Recent work in optimal control explores how to achieve stable solutions using ML-based methods (Nakamura-Zimmerer et al., 2022b,a, Chen, 2024, Chang et al., 2019). More directly within economics, Ebrahimi Kahou et al. (2024) discusses the intuitive connection between the inductive bias of ML, i.e., which solution an algorithm prefers among those that fit the data equally well, and turnpikes (McKenzie, 1976) in dynamic economic models, but does not provide a formal theory or consider kernel methods. Our paper contributes to this literature, providing a formal argument for why the inductive bias of ML algorithms promotes stability in infinite-horizon control when using particular kernel machines.

*Contributions.* Core results of our paper include:

- **Minimum norm and transversality:** theoretical and empirical evidence that minimum-norm kernel methods select the solution satisfying many asymptotic boundary conditions found in economic problems;
- **Learning the right set of steady states:** evidence that kernel machines converge to the economically relevant steady states of dynamical systems without computing them, namely those corresponding to cleanly characterized solutions with an economic interpretation, even with multiple steady states (a growth model with a concave-convex production function) or hysteresis (a small open economy). This leads to fairly accurate generalization outside the training data, even without enforcing asymptotic boundary conditions;
- **Consistency of ML estimates:** a guarantee that our kernel estimates converge to the true solution, and thus to one satisfying the asymptotic boundary conditions, as the training data increase;
- **Robustness and speed:** demonstrations that kernel machines can be competitive in both speed and robustness with traditional and deep-learning methods for modeling economic systems on small and large problems; and

- **Scalability:** a multicountry growth model with an endogenous world interest rate, solved for up to 100 countries in about a minute.

*Structure.* The remainder of this paper is structured as follows. Section 2 describes the class of economic models and our assumptions, and shows that every trajectory violating transversality is explosive. Section 3 describes how kernel methods map to our class of problems and shows that paths whose derivatives have finite RKHS norm satisfy transversality. Section 3.3 shows that kernel solutions converge to the true solution of the DAE as the number of training points grows. Section 4 presents the core applications: the neoclassical growth model, a growth model with multiple steady states, linear asset pricing, a small open economy, and a multicountry growth model with an endogenous world interest rate. Section 5 describes the state-space formulation used for the recursive and stochastic examples, and Section 6 discusses future directions, including extending the theory to that formulation.[4] Section 7 concludes with the scope of the theory and the tradeoffs of kernel methods relative to other machine learning approaches such as NNs.

## 2. Setup

We focus on an important class of dynamic problems in economics and finance: deterministic, continuous-time systems that arise from discounted infinite-horizon optimal control problems, together with algebraic constraints that encode conditions such as instantaneous market clearing.[5]

*Problem class.* In these models, the first-order necessary conditions of the underlying decision problems can be stacked into a system of differential-algebraic equations (DAEs). Partition the variables into three vectors: *state variables* $\boldsymbol{x}(t) \in \mathbb{R}^M_{++}$, with an initial condition $\boldsymbol{x}_0$; *co-state variables* $\boldsymbol{\mu}(t) \in \mathbb{R}^M_{++}$, with accompanying asymptotic boundary conditions that typically arise from the embedded control problem; and *jump variables* $\boldsymbol{y}(t) \in \mathbb{R}^P$, which relate the state and co-state variables (e.g., co-state $=$ the marginal utility of consumption) and/or impose intratemporal constraints (e.g., market clearing conditions).[6] Canonical examples of such equations arise from applying Pontryagin's Maximum Principle to the current-value Hamiltonian of a dynamic optimization problem, as in Acemoglu (2008). In that case, $\boldsymbol{\mu}(t)$ represents the current-value

[4] We detail the implementation of these formulations in Supplemental Appendix Section 5 (Ebrahimi Kahou et al., 2026) and provide replication examples online at https://github.com/HighDimensionalEconLab/kernels.

[5] An inherent characteristic of discounted infinite-horizon optimal control problems, whether deterministic or stochastic, is the presence of asymptotic boundary conditions such as transversality conditions. These conditions can be formulated sequentially or recursively in a state space, and in continuous- or discrete-time (see discussions of necessary conditions in Michel (1982), Benveniste and Scheinkman (1982), Van et al. (2007)).

[6] When present, jump variables constrain the solution manifold and are not matched with boundary or initial values. The connection between the number of *jump variables* and stability local to a steady state is discussed in Blanchard and Kahn (1980).

Lagrange multipliers (co-state variables) associated with the state variables $\boldsymbol{x}(t)$.[7] Let $\mathbb{V} \subseteq \mathbb{R}^M_{++} \times \mathbb{R}^M_{++} \times \mathbb{R}^P$ be the domain of the variables. The dynamical system with primitives $\boldsymbol{F}, \boldsymbol{G} : \mathbb{V} \to \mathbb{R}^M$ and $\boldsymbol{H} : \mathbb{V} \to \mathbb{R}^P$ with a discount rate $r > 0$ is

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t)), \tag{1}$$

$$\dot{\boldsymbol{\mu}}(t) = r\boldsymbol{\mu}(t) - \boldsymbol{\mu}(t) \odot \boldsymbol{G}(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t)), \tag{2}$$

$$\boldsymbol{0} = \boldsymbol{H}(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t)), \tag{3}$$

subject to $M$ initial values, and $M$ asymptotic boundary conditions (i.e., transversality conditions)

$$\boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{4}$$

$$\boldsymbol{0} = \lim_{t \to \infty} e^{-rt} \boldsymbol{x}(t) \odot \boldsymbol{\mu}(t) \tag{5}$$

The symbol $\odot$ denotes element-wise multiplication between vectors of the same dimension.

This system is an autonomous differential-algebraic equation with a mix of initial conditions, Equation (4), and asymptotic boundary conditions, Equation (5).

*Example: Neoclassical Growth Model.* The canonical example of this class in macroeconomics is the neoclassical growth model. To fit our notation, define: capital $x(t) \in \mathbb{R}_{++}$, consumption $y(t) \in \mathbb{R}_{++}$, current-value co-state variable $\mu(t) \in \mathbb{R}_{++}$ with flow utility $\log(y)$, discount rate $r > 0$, depreciation rate $0 < \delta < 1$, and a twice continuously differentiable, strictly increasing, strictly concave production function $f(x)$ satisfying the Inada conditions.[8]

$$\dot{x}(t) = f\left(x(t)\right) - \delta x(t) - y(t) := F(x(t), \mu(t), y(t)) \tag{6}$$

$$\dot{\mu}(t) = r\mu(t) - \mu(t) \underbrace{\left[f'\left(x(t)\right) - \delta\right]}_{:= G(x(t), \mu(t), y(t))} \tag{7}$$

$$0 = \mu(t)y(t) - 1 := H(x(t), \mu(t), y(t)) \tag{8}$$

$$x(0) = x_0 \tag{9}$$

$$0 = \lim_{t \to \infty} e^{-rt} \mu(t) x(t) \tag{10}$$

[7] The proofs do not use non-negativity of the jump variables or strict positivity of $\boldsymbol{x}$ and $\boldsymbol{\mu}$, which holds in our baseline examples in Sections 4.1 to 4.3. They therefore apply unchanged when components of $\boldsymbol{x}(t)$ are naturally signed, such as the net foreign asset positions in Sections 4.4 and 4.5.

[8] The derivation follows the standard planning problem that maximizes the lifetime discounted utility of consumption: $\int_0^\infty e^{-rt} u\left(c(t)\right) dt$, subject to the law of motion for capital: $\dot{k}(t) = f\left(k(t)\right) - \delta k(t) - c(t)$, where the depreciation rate satisfies $0 < \delta < 1$. In this case, the current-value Hamiltonian is $u\left(c(t)\right) + \mu(t)\left[f\left(k(t)\right) - \delta k(t) - c(t)\right]$. The DAE follows from applying Pontryagin's Maximum Principle along with a standard transversality condition, and is mapped to our notation (see Acemoglu (2008)).

This example fulfills the assumptions of Section 2.1 below, as verified in Section 4.1. Next we will analyze the key conditions required for our algorithm, and demonstrate the intuition with this running example.

### 2.1 *Assumptions for Unique and Bounded Solutions*

Next we make the regularity and uniqueness assumptions required for the DAE to be *well-posed*, with the norms required for our method and consistency proof. First, Assumption 1 provides the regularity conditions on the domain.

ASSUMPTION 1 (Regularity). *Write* $\|v\|_\infty := \max_i |v_i|$ *for a vector* $v$. *Assume that* $\boldsymbol{F}$, $\boldsymbol{G}$, *and* $\boldsymbol{H}$ *in Equations* (1) *to* (3) *are locally Lipschitz on* $\mathbb{V}$ *with respect to* $\|\cdot\|_\infty$.[9]

In addition, we add standard assumptions on the problem formulation to ensure that a unique solution exists and that its state and co-state variables are bounded and strictly positive.[10]

ASSUMPTION 2 (Unique Bounded Solution). *Fix* $\boldsymbol{x}_0$. *Call* $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{y})$ *a* trajectory *if* $\boldsymbol{x}, \boldsymbol{\mu} : [0,\infty) \to \mathbb{R}^M$ *are continuously differentiable,* $\boldsymbol{y} : [0,\infty) \to \mathbb{R}^P$ *is continuous,* $(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t)) \in \mathbb{V}$ *for all* $t \geq 0$, *and Equations* (1) *to* (4) *hold for all* $t \geq 0$. *Assume:*

(i) *Exactly one trajectory,* $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$, *satisfies Equation* (5).

(ii) $\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*$ *are twice continuously differentiable and*

$$\int_0^\infty \|\dot{\boldsymbol{x}}^*(t)\|_2^2 + \|\dot{\boldsymbol{\mu}}^*(t)\|_2^2 + \|\dot{\boldsymbol{y}}^*(t)\|_2^2 + \|\ddot{\boldsymbol{x}}^*(t)\|_2^2 + \|\ddot{\boldsymbol{\mu}}^*(t)\|_2^2 + \|\ddot{\boldsymbol{y}}^*(t)\|_2^2 \,\mathrm{d}t < \infty.$$

(iii) *There exists a box*

$$\mathbb{B} := [\underline{\boldsymbol{x}}, \bar{\boldsymbol{x}}] \times [\underline{\boldsymbol{\mu}}, \bar{\boldsymbol{\mu}}] \times [\underline{\boldsymbol{y}}, \bar{\boldsymbol{y}}] \subset \mathbb{V}$$

*that contains* $(\boldsymbol{x}^*(t), \boldsymbol{\mu}^*(t), \boldsymbol{y}^*(t))$ *for all* $t \geq 0$.

The first condition, Assumption 2(i), assumes uniqueness rather than deriving it from the primitives.[11] The part specific to DAEs is whether a unique $\boldsymbol{y}$ fulfills the algebraic Equation (3), since the jump variables have neither initial nor boundary values. A continuously differentiable $\boldsymbol{H}$ with invertible $\nabla_{\boldsymbol{y}} \boldsymbol{H}$ on $\mathbb{V}$ lets us locally substitute $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{\mu})$ into Equations (1) and (2), which is sufficient but stronger than necessary.

[9] Following Evans and Gariepy (1992, Section 3.1.1, p. 80): for any $A \subseteq \mathbb{R}^n$, a map $f : A \to \mathbb{R}^m$ is *locally Lipschitz* if its restriction to every compact $K \subseteq A$ is Lipschitz, with a constant that may depend on $K$. Here $A = \mathbb{V}$, compact means closed and bounded in $\mathbb{R}^n$, and the choice of norm only rescales the constant.

[10] See Arrow and Kurz (1970, p. 51) for when these assumptions hold. In practice, it is usually sufficient to use a current-value rather than present-value Hamiltonian, since the present-value multipliers $e^{-rt}\boldsymbol{\mu}(t)$ decay to zero and violate the lower bounds. In cases with growth, either solve a de-trended model or carry the trend in the approximation class as in Supplemental Appendix Section 2.4.

[11] Multiple steady states do not by themselves conflict with Assumption 2: with hysteresis, as in Section 4.4, the limiting state depends on $\boldsymbol{x}_0$ but the trajectory is unique for each initial condition (Supplemental Appendix Section 1.1).

See Brenan et al. (1996) and Supplemental Appendix Section 1 for sufficient conditions on the algebraic equations.

The remaining two conditions describe the solution itself. Condition (ii) requires its first and second derivatives to be square-integrable. In our applications the solution converges to an interior steady state with exponentially decaying derivatives, which is more than enough for this to hold. Condition (iii) then requires the solution to stay in a compact box $\mathbb{B}$ inside $\mathbb{V}$: the state and co-state variables are bounded from above and strictly away from zero, while the jump variables need only be bounded. Unlike (i), both restrict only the unique solution. Trajectories that fail Equation (5) are unrestricted, and Theorem 1 shows that none of them has finite derivative energy, the time integral of its squared state and co-state time derivatives.

### 2.2 *Shooting Methods and Explosive Non-Solutions*

The central computational challenge is that the boundary condition Equation (5) is asymptotic, so it cannot be imposed directly. If we knew the right $\boldsymbol{\mu}(0) = \boldsymbol{\mu}_0$, and the algebraic Equation (3) determined $\boldsymbol{y}$ along the path, the remaining system Equations (1) to (4) would be an initial value problem.

First, we need to contrast two types of functions fulfilling the DAE. A *solution* trajectory fulfills the full set of well-posed equations Equations (1) to (5), while a *non-solution* trajectory fulfills the underdetermined system Equations (1) to (4) but fails the asymptotic boundary condition Equation (5). Assumption 2 ensures that we have a unique bounded solution, but there is usually a continuum of non-solutions which, when the algebraic Equation (3) determines $\boldsymbol{y}$ along the path, can each be associated with a different $\boldsymbol{\mu}(0)$. The solution is then the single value of $\boldsymbol{\mu}(0)$, among this $M$-dimensional family of candidates, for which transversality holds.

One classical approach is a shooting method, which guesses $\tilde{\boldsymbol{\mu}}_0$, integrates Equations (1) to (3) to a large horizon $T$, and adjusts the guess until $\boldsymbol{x}(T)$ is close to a known steady state, which approximates Equation (5) when the solution converges there. The difficulty is that a guess at any fixed distance from $\boldsymbol{\mu}_0$ either leaves the domain $\mathbb{V}$ in finite time, so that it does not define a trajectory on $[0, \infty)$, or produces a non-solution which, since $e^{-rt}\boldsymbol{x}(t) \odot \boldsymbol{\mu}(t)$ fails to vanish by Assumption 2(i), does not decay relative to $e^{rt}$ and typically becomes exponentially large as the horizon increases. Near the finite-horizon root a fixed terminal tolerance is then typically met only by guesses in a neighborhood that shrinks exponentially with $T$, while a small $T$ forces the path toward the steady state too early and biases the approximation. Shooting and related finite-horizon boundary-value methods work well with a good initial guess in low dimensions, but the tuning becomes harder with the dimension, and the long-run target must be known, which complicates models with hysteresis or multiple steady states (Supplemental Appendix Section 6.1).

### 2.3 *Key Insight*

The key insight into our methods is that we can use the rapid divergence separating the unique solution from the non-solutions to our advantage.

While pointwise evaluation of transversality at a large $T$ is sensitive, more global approaches to distinguish solutions from non-solutions are effective. In particular, function norms and seminorms provide an alternative that is far more numerically stable for large $T$, since they integrate over the horizon rather than match the path at its endpoint. Theorem 1 below shows that every non-solution has infinite derivative energy $\int_0^\infty \|\dot{\boldsymbol{x}}(t)\|_2^2 + \|\dot{\boldsymbol{\mu}}(t)\|_2^2\, dt$ while the solution's is finite, so for each non-solution the energy over $[0, T]$ eventually exceeds the solution's. Any norm whose square is bounded below by a multiple of this energy and which is finite on the solution, such as the Sobolev and RKHS norms of the derivatives below, is therefore infinite for every non-solution and finite for the solution. Minimizing such a norm subject to Equations (1) to (4) therefore selects the solution, which Theorem 3 makes precise for the sampled version of the problem.

While we could use this insight for different types of function approximation and norms (e.g., Ebrahimi Kahou et al. (2024) empirically show similar results arising from inductive bias in deep learning), kernel methods have the advantage that the norms are precisely defined in the RKHS and remain stable as $T$ grows.[12]

*Sobolev norm solutions.* Before discussing function approximation in the RKHS and determining whether it aligns with the solution concept of the DAE, we can consider the natural function space in which solutions to our problem class reside. This provides us with a precise way to compare degrees of divergence that is not pointwise.

To measure how fast a trajectory moves over $[0, \infty)$, we use the Sobolev-$1,2$ norm of its derivative:

$$\|\dot{w}(\cdot)\|_{\mathcal{W}^{1,2}} := \left( \int_0^\infty \left( |\dot{w}(t)|^2 + |\ddot{w}(t)|^2 \right) dt \right)^{1/2}.$$

This is a seminorm of $w(\cdot)$ because it ignores the level of $w$ and penalizes only its first and second derivatives.

In other words, Assumption 2(ii) says that the solution has a finite $\|\dot{w}(\cdot)\|_{\mathcal{W}^{1,2}}$ over $[0, \infty)$.[13]

Back to our setup: Theorem 1 below shows that every trajectory satisfying Equations (1) to (4) but not Equation (5) has infinite derivative energy $\int_0^\infty \|\dot{\boldsymbol{x}}(t)\|_2^2 + \|\dot{\boldsymbol{\mu}}(t)\|_2^2\, dt$, while the solution's is finite by Assumption 2(ii). This energy is bounded by $\|\dot{\boldsymbol{x}}\|_{\mathcal{W}^{1,2}}^2 + \|\dot{\boldsymbol{\mu}}\|_{\mathcal{W}^{1,2}}^2$, so the solution is the unique trajectory with finite derivative seminorm, and over a long enough horizon $[0, T]$ its seminorm is smaller than that of any given non-solution. Given Assumption 2, the trajectory fulfilling transversality is therefore the one of finite, and hence minimal, derivative seminorm.

[12] The numerical difficulty of kernel methods is instead the conditioning of the Gram matrix, which does not grow with $T$ and is handled by standard numerical methods: an appropriate choice of kernel bandwidth (to which the results are insensitive, Supplemental Appendix Section 2), the ridge continuation of Section A, and the preconditioning of Supplemental Appendix Section 4.4 at large $N$.

[13] We note that the seminorm $\|\dot{w}(\cdot)\|_{\mathcal{W}^{1,2}}$ is a more natural measure of complexity than the full norm $\|w(\cdot)\|_{\mathcal{W}^{2,2}}$. Consider for example a problem where $w(T)$ converges to some nonzero steady state $w(\infty)$ as $T \to \infty$. Such a function is not square-integrable over $[0, \infty)$, so its full norm diverges while the seminorm can stay finite.

THEOREM 1 (Divergence of Non-Solutions). *For a DAE satisfying Assumptions 1 and 2, define* $\mathbb{S}$ *as the set of trajectories in the sense of Assumption 2, which satisfy Equations (1) to (4) but not necessarily Equation (5), and define* $\mathbb{M}$ *as the set of* $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{y})$ *with* $\boldsymbol{x}, \boldsymbol{\mu}$ *continuously differentiable and finite derivative energy* $\int_0^\infty \|\dot{\boldsymbol{x}}(t)\|_2^2 + \|\dot{\boldsymbol{\mu}}(t)\|_2^2 \, dt < \infty$. *Then* $\mathbb{S} \cap \mathbb{M}$ *contains exactly one element* $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$*, and it is the unique trajectory that satisfies Equation (5).*

PROOF. See Section C. □

## 3. METHOD

In this section we first define our approximation class and present an algorithm for solving an underdetermined DAE, Equations (1) to (4), using ridgeless kernel regression. After introducing kernel methods, Theorem 2 will show that a finite-norm solution is a sufficient condition to satisfy the pointwise transversality requirement, so that any solution to the underdetermined kernel regression fulfills transversality. Finally, we establish in Theorem 3 that the ridgeless kernel regression problem is consistent and therefore asymptotically solves the full system Equations (1) to (5). Although intuitive, the result is nontrivial because it requires demonstrating alignment and mathematical consistency between the RKHS of the approximate solution, its derivatives, and the solution concept of the DAE itself.

### 3.1 *Kernel Method*

Our algorithm models $\boldsymbol{x}(t)$, $\boldsymbol{\mu}(t)$, and $\boldsymbol{y}(t)$, together with their derivatives, using kernel machines, in which functions are expressed as weighted combinations of a kernel function centered at each data point, where the kernel encodes a notion of distance (Murphy, 2022). More specifically, our approximation ensures that the DAE is satisfied on some finite set of (possibly irregular) points $\mathcal{D} := \{t_1, \ldots, t_N\}$, which we refer to as the "training data," and selects, among the paths in the approximation class satisfying this condition, one of minimum function norm. Because the method is non-parametric, the number of parameters defining our approximation grows with the data. Given a kernel function $k(\cdot, \cdot)$, our approximations take the form:

$$\begin{aligned} \dot{\hat{\boldsymbol{x}}}(t) &= \textstyle\sum_{j=1}^N \boldsymbol{\alpha}_j^x k(t, t_j), & \dot{\hat{\boldsymbol{\mu}}}(t) &= \textstyle\sum_{j=1}^N \boldsymbol{\alpha}_j^\mu k(t, t_j), & \dot{\hat{\boldsymbol{y}}}(t) &= \textstyle\sum_{j=1}^N \boldsymbol{\alpha}_j^y k(t, t_j), \\ \hat{\boldsymbol{x}}(t) &= \boldsymbol{x}_0 + \textstyle\int_0^t \dot{\hat{\boldsymbol{x}}}(\tau) d\tau, & \hat{\boldsymbol{\mu}}(t) &= \hat{\boldsymbol{\mu}}_0 + \textstyle\int_0^t \dot{\hat{\boldsymbol{\mu}}}(\tau) d\tau, & \hat{\boldsymbol{y}}(t) &= \hat{\boldsymbol{y}}_0 + \textstyle\int_0^t \dot{\hat{\boldsymbol{y}}}(\tau) d\tau, \end{aligned} \tag{11}$$

where $\boldsymbol{\alpha}_j^x$, $\boldsymbol{\alpha}_j^\mu$, $\boldsymbol{\alpha}_j^y$, $\hat{\boldsymbol{\mu}}_0$, and $\hat{\boldsymbol{y}}_0$ are parameters fit to fulfill our DAE.[14] We approximate the time derivative and integrate to obtain the function values. While it is possible to approximate the function itself, the benefits of this formulation are that the initial condition $\hat{\boldsymbol{x}}(0) = \boldsymbol{x}_0$ is directly enforced, and extrapolation performance is significantly improved.[15]

[14] For a given kernel, the integration in Equation (11) can be done in closed form or numerically, i.e. $\int_0^t k(\tau, t_j) \mathrm{d}\tau$ for each $t_j$.

[15] For any fixed $\mathcal{D}$, $\lim_{t\to\infty} k(t, t_j) = 0$ implies $\lim_{t\to\infty} \dot{\hat{\boldsymbol{\mu}}}(t) = \lim_{t\to\infty} \dot{\hat{\boldsymbol{x}}}(t) = \lim_{t\to\infty} \dot{\hat{\boldsymbol{y}}}(t) = 0$. For an integrable kernel, the level therefore converges to a finite constant learned from the fitted

In this paper, we use a Matérn kernel for $k(\cdot,\cdot)$, with smoothness $\nu = \frac{1}{2}$ and length-scale $\ell$ (see Definition 1 in Section B). The choice of the Matérn kernel family is driven by theoretical considerations to formally align with the solution concept of the DAE, but many other kernels (e.g., Gaussian kernels) have similar performance in practice, even if they impose more smoothness than is strictly necessary.

*Function Norms.* Central to these methods is that the kernel function $k(\cdot,\cdot)$ has an associated function space, its Reproducing Kernel Hilbert Space $\mathcal{H}$, which provides an inner product and an associated function norm. Moreover, the norm of functions in this space can be calculated as a quadratic form of its coefficients. In particular, for the approximated $\hat{\boldsymbol{x}}(t)$, $\hat{\boldsymbol{\mu}}(t)$, and $\hat{\boldsymbol{y}}(t)$ in Equation (11), the squared norms are computed as follows:

$$\|\hat{\boldsymbol{x}}^{(m)}\|_{\mathcal{H}}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i^{x^{(m)}}\, \alpha_j^{x^{(m)}}\, k(t_i,t_j), \qquad \|\hat{\boldsymbol{\mu}}^{(m)}\|_{\mathcal{H}}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i^{\mu^{(m)}}\, \alpha_j^{\mu^{(m)}}\, k(t_i,t_j),$$
$$\|\hat{\boldsymbol{y}}^{(p)}\|_{\mathcal{H}}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i^{y^{(p)}}\, \alpha_j^{y^{(p)}}\, k(t_i,t_j), \tag{12}$$

where superscript $(m)$ or $(p)$ denotes the coefficients corresponding to the $m$-th state or co-state variable or the $p$-th jump variable. For instance, $\alpha_i^{x^{(m)}}$ is the $i$-th element of the learnable coefficients for the $m$-th state variable.

The subscript $\mathcal{H}$ denotes the RKHS associated with the kernel $k(\cdot,\cdot)$. The RKHS norm $\|\cdot\|_{\mathcal{H}}$ measures the complexity of a function in a way determined by $k$ and is used to control the smoothness of the approximating functions. For more details, see Schölkopf and Smola (2002, Chapters 2 & 4). In the case of Matérn kernels, as will be discussed later, this norm has a concrete smoothness interpretation: it directly controls the magnitude of the function's derivatives up to a certain order.

*Ridgeless Kernel Regression.* Kernel methods are tailored to solve empirical risk minimization (ERM) style objectives in this function space, such as $\min_{h\in\mathcal{H}}\{\sum_{i=1}^{N}\Phi(h(t_i)) + \lambda||h||_{\mathcal{H}}^2\}$ for some loss $\Phi(\cdot)$ and regularization parameter $\lambda > 0$. The Representer Theorem shows that every minimizer has the form $h(t) = \sum_{j=1}^{N}\alpha_j k(t,t_j)$ for some $\alpha_j \in \mathbb{R}$ (Schölkopf et al., 2001, Schölkopf and Smola, 2002), with norm $\|h\|_{\mathcal{H}}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\,\alpha_j\,k(t_i,t_j)$. Furthermore, for a continuous nonnegative loss that admits zero-loss interpolants, as $\lambda \to 0$ the minimizers approach (along some subsequence) those of the minimum-norm interpolation problem $\min_{h\in\mathcal{H}}||h||_{\mathcal{H}}^2$ s.t. $\Phi(h(t_i)) = 0$ for all $t_i \in \mathcal{D}$ (see Liang and Rakhlin, 2020, Hastie et al., 2022, for least squares); Section A proves this for our problem. In our problem the constraints also involve integrals of the derivatives and the whole path, so Equation (11) is a finite-dimensional approximation class rather than the representer form of the minimizer.

---

coefficients. In particular, the Matérn-1/2 kernel $k(t,t') = \sigma^2 e^{-|t-t'|/\ell}$ gives $\lim_{t\to\infty}\hat{\boldsymbol{\mu}}(t) = \hat{\boldsymbol{\mu}}_0 + \sigma^2\ell\sum_{j=1}^{N}\boldsymbol{\alpha}_j^{\mu}\left(2 - e^{-t_j/\ell}\right)$, which is generally not $\hat{\boldsymbol{\mu}}_0$. Alternatively, if the function itself were approximated without an intercept, then $\lim_{t\to\infty}\hat{\boldsymbol{\mu}}(t) = 0$. That said, given that our goal is to solve for only short- and medium-run behavior, approximating the function itself often works in practice.

In our problem, the loss will interpolate the system of equations in Equations (1) to (4) and minimize the sum of RKHS norms of the approximating functions. The unknowns are the derivatives $\hat{\dot{\boldsymbol{x}}}, \hat{\dot{\boldsymbol{\mu}}}, \hat{\dot{\boldsymbol{y}}}$, each component of which lies in $\mathcal{H}$, and the initial values $\hat{\boldsymbol{\mu}}_0 \in \mathbb{R}^M$, $\hat{\boldsymbol{y}}_0 \in \mathbb{R}^P$, with $\hat{\boldsymbol{x}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{y}}$ recovered by integration as in Equation (11):

$$\min_{\hat{\dot{\boldsymbol{x}}}, \hat{\dot{\boldsymbol{\mu}}}, \hat{\dot{\boldsymbol{y}}}, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{y}}_0} \left( \sum_{m=1}^{M} \|\hat{\dot{\boldsymbol{x}}}^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^{M} \|\hat{\dot{\boldsymbol{\mu}}}^{(m)}\|_{\mathcal{H}}^2 + \sum_{p=1}^{P} \|\hat{\dot{\boldsymbol{y}}}^{(p)}\|_{\mathcal{H}}^2 \right) \tag{13}$$

$$\begin{aligned}
\text{s.t. } & \hat{\dot{\boldsymbol{x}}}(t_i) = \boldsymbol{F}(\hat{\boldsymbol{x}}(t_i), \hat{\boldsymbol{\mu}}(t_i), \hat{\boldsymbol{y}}(t_i)), \quad \text{for all } t_i \in \mathcal{D} \\
& \hat{\dot{\boldsymbol{\mu}}}(t_i) = r\hat{\boldsymbol{\mu}}(t_i) - \hat{\boldsymbol{\mu}}(t_i) \odot \boldsymbol{G}(\hat{\boldsymbol{x}}(t_i), \hat{\boldsymbol{\mu}}(t_i), \hat{\boldsymbol{y}}(t_i)), \quad \text{for all } t_i \in \mathcal{D} \\
& \boldsymbol{0} = \boldsymbol{H}(\hat{\boldsymbol{x}}(t_i), \hat{\boldsymbol{\mu}}(t_i), \hat{\boldsymbol{y}}(t_i)), \quad \text{for all } t_i \in \mathcal{D} \\
& (\hat{\boldsymbol{x}}(t), \hat{\boldsymbol{\mu}}(t), \hat{\boldsymbol{y}}(t)) \in \mathbb{B} \quad \text{for all } t \geq 0, \qquad (\star)
\end{aligned}$$

where $\mathbb{B} \subset \mathbb{V}$ is the box of Assumption 2(iii), which contains the solution. The constraint $(\star)$ is a technical device for Theorem 3: it keeps every feasible path in the domain of $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{H}$, where Assumption 1 applies, and it is the only place where Assumption 2(iii) enters. Our implementations do not impose it. If a global minimizer of the problem without $(\star)$, over the same function class, lies in $\mathbb{B}$, then it also solves the problem with $(\star)$ imposed. We formally prove in Section E (Lemma 7) that a solution to Equation (13) exists for any $\mathcal{D} = \{t_1, \ldots, t_N\}$.

*Implementation.* We parameterize the derivatives by Equation (11), a finite-dimensional subspace of $\mathcal{H}$ with norms given by Equation (12), which restricts Equation (13) to a finite-dimensional problem in the coefficients and initial values. The algorithm does not solve the constrained problem Equation (13) directly. It minimizes the ridge objective, the residuals plus a small multiple $\lambda$ of the squared norms, with continuation of $\lambda$ toward zero where needed, as described in Section A. The constrained problem is the object of the theory, and the ridge problem is what we compute. Proposition 1 in Section A gives the connection between the two.[16] As a robustness check, for selected models we also solve the constrained problem in Equation (13) directly with a general nonlinear solver.[17]

Intuitively, minimizing Equation (13) finds the path with the minimum RKHS norm among all interpolating paths in $\mathbb{B}$. Next we need to show why this minimum-norm solution will be the one which fulfills transversality.

[16] The primary computational tradeoff for kernel methods is that they require pairwise evaluation across all of the data (i.e., $k(t_i, t_j)$ for all $i, j$), and an out-of-data function evaluation requires evaluating $k(t, t_i)$ for all data. Computational methods can solve these challenges in practice for up to millions of observations (Gardner et al., 2018, Wang et al., 2019), but as with all non-parametric methods it eventually becomes a limitation. This scaling limitation is generally less pressing in economic research than in production ML systems in industry. Another consideration is that while in our current applications it is essential, the strong dependence of solutions on the RKHS norms may or may not be a benefit if the solution concept of the underlying problem is poorly aligned with the RKHS induced by the kernel.

[17] We use https://github.com/cvanaret/Uno, which is a second-order interior-point method. However, this direct approach forms and factorizes the exact Hessian of the Lagrangian, which scales poorly with $N$.

### 3.2 *Transversality*

While it is intuitive that a minimum-norm solution is sufficient to enforce transversality, we need to proceed cautiously to relate the norms of the DAE solution concept to those of the RKHS. While many different kernels, and associated norms, could be used in practice, we will emphasize Matérn kernels due to their formal connection to Sobolev spaces.

*Matérn kernels.* For functions on $[0,\infty)$, the RKHS of the Matérn kernel with $\nu = 1/2$ is the Sobolev space $\mathcal{W}^{1,2}$, with norms that are equivalent up to constants depending only on the lengthscale $\ell$ (Theorem 4 in Section B). Thus, modelling the derivatives with the $\nu = 1/2$ Matérn kernel places them in $\mathcal{W}^{1,2}$, and the RKHS norm in Equation (13) controls the $\mathcal{W}^{1,2}$ norm of the derivatives. We are assuming more curvature than is strictly necessary, given that the solution concept of the DAE only requires a first derivative. However, controlling the $\mathcal{W}^{1,2}$ norm of the derivatives is sufficient to ensure transversality, as we will now show.

Given that Equation (5) is expressed in terms of the product $\boldsymbol{x}(t) \odot \boldsymbol{\mu}(t)$, while Equation (13) and the Representer Theorems are written in terms of function norms, we first need to show how penalized norms are sufficient to control $\boldsymbol{x}(t) \odot \boldsymbol{\mu}(t)$ globally, and consequently bound the pointwise transversality condition.

THEOREM 2 (Bounding with Norms). *Let $\mathcal{H}$ be the Matérn RKHS with $\nu = \frac{1}{2}$ and lengthscale $\ell > 0$ on $[0,\infty)$, let $\dot{x}^{(m)}, \dot{\mu}^{(m)} \in \mathcal{H}$ for every $m = 1,\dots,M$, and let $\boldsymbol{x}(t) = \boldsymbol{x}(0) + \int_0^t \dot{\boldsymbol{x}}(\tau)d\tau$ and $\boldsymbol{\mu}(t) = \boldsymbol{\mu}(0) + \int_0^t \dot{\boldsymbol{\mu}}(\tau)d\tau$ for finite initial values $\boldsymbol{x}(0), \boldsymbol{\mu}(0) \in \mathbb{R}^M$. Then*

$$\lim_{t\to\infty} e^{-rt}\, \boldsymbol{x}(t) \odot \boldsymbol{\mu}(t) \;=\; \mathbf{0};$$

*that is, $(\boldsymbol{x}, \boldsymbol{\mu})$ satisfies the asymptotic boundary condition Equation (5).*

PROOF. See Section D. □

This theorem applies in particular to the approximations in Equation (11), and shows that finite RKHS norms of the derivatives are enough to deliver the asymptotic boundary condition Equation (5). Intuitively, controlling the sum of the RKHS norms of the derivatives rules out exponentially explosive paths.

To reiterate, a finite norm of the approximated derivatives is sufficient for transversality and aligns with the underlying Sobolev norm of the solution concept of the DAE. While we have assumed more smoothness than is strictly necessary, this is not a significant limitation in practice as we demonstrate in Section 4.[18]

[18] However, this also shows where these methods might break down or where consistency proofs along the lines of Theorem 3 are more challenging. For example, in cases where the solution lacks the smoothness required by Assumption 2(ii), or where the level of the function rather than just its derivatives is essential for selecting the trajectory fulfilling Equation (5).

### 3.3 *Consistency*

While Theorems 1 and 2 show that finite derivative norms single out the trajectory fulfilling Equation (5), we need to show that our kernel solution minimizing Equation (13) approximates the true solution despite only satisfying the DAE on a finite number of points. To that end, we demonstrate that our approximation is a *consistent estimator* of the true solution.

THEOREM 3 (Consistency). *Suppose Assumptions 1 and 2 hold, and let* $(\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{y}}_N)$ *be any solution of Equation (13), with* $\mathcal{H}$ *the Matérn RKHS with* $\nu = \frac{1}{2}$ *and lengthscale* $\ell > 0$ *on* $[0, \infty)$*, for the points* $\mathcal{D} = \{t_1, \ldots, t_N\}$*, drawn i.i.d. from a distribution* $\pi$ *on* $[0, \infty)$ *with a strictly positive density and finite second moment. Then*

$$\lim_{N\to\infty}(\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{y}}_N) \overset{\text{a.s.}}{=} (\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*),$$

*where* $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$ *is the unique trajectory satisfying Equation (5) (Assumption 2), and the convergence is uniform on compact subsets of* $[0, \infty)$.

PROOF. See Section E. □

Theorem 3 is stated for a minimizer over the whole RKHS and for collocation points drawn i.i.d. from $\pi$, while our implementations restrict the derivatives to the finite-dimensional class Equation (11) and use fixed time grids.

Though this proof borrows techniques from the statistical learning literature on kernel methods (e.g., Wainwright, 2019), it is nontrivial for several reasons. First, we need to show that a solution to Equation (13) exists and that the true solution satisfies its constraints. Second, we need to prove that uniform convergence of derivatives on compact sets implies the same for the DAE solution functions themselves. Given these results, we can be confident that with enough data, the empirical solutions $\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{y}}_N$ will (approximately) fulfill Equations (1) to (5), with eventual convergence to the true solution satisfying the transversality condition.

## 4. RESULTS

We solve two standard baselines in dynamic economics: the neoclassical growth model in Section 4.1, and a model of risk-neutral asset pricing in Section 4.3. These problems are chosen because they are standard examples in textbooks (e.g., Ljungqvist and Sargent, 2018), they admit reference solutions from classical methods, and they have established results regarding the set of solutions to the ill-posed versions without the asymptotic boundary conditions.

In Section 4.2, we present a case with multiple steady states, a challenging setting where our methods are particularly useful. In Section 4.4, we solve a small open economy model with a continuum of steady states indexed by initial conditions. In Section 4.5, we close the model with $N$ countries trading a single bond, so that the world interest rate is determined endogenously.

In all cases, we use a Matérn kernel with $\nu = \frac{1}{2}$ and $\sigma = 1$ (see Definition 1), with lengthscale $\ell = 10$ and $\mathcal{D} := \{0, 1, 2, \ldots, 40\}$ unless noted otherwise in the example. Where possible we compare the relative errors of a kernel solution to a benchmark obtained with classic methods, e.g., $\varepsilon_{\boldsymbol{w}(t)} := \left|\frac{\hat{\boldsymbol{w}}(t) - \boldsymbol{w}(t)}{\boldsymbol{w}(t)}\right|$ with corresponding definitions for the other variables. All baseline cases are solved with the solvers of Section 3 and execute in less than a second.

### 4.1 *Neoclassical Growth Model*

In this section, we solve the neoclassical growth model (also known as the Ramsey–Cass–Koopmans model) summarized by Equations (6) to (10), generalized from log to constant relative risk aversion (CRRA) utility: the flow utility is $\frac{y^{1-\gamma}-1}{1-\gamma}$, so the algebraic condition Equation (8) becomes $0 = \mu(t)y(t)^{\gamma} - 1$, with $\gamma = 1$ recovering the log case. Our baseline parameters map the annual calibration of Aruoba et al. (2006): $a = \frac{1}{3}$, $\gamma = 2$ (their $\tau$), $r = -4\log(0.9896) \approx 0.042$ from their quarterly discount factor, $\delta = 0.1$ (the annual analogue of the quarterly $0.025$ in King and Rebelo, 1999), $f(x) = x^a$, and $x_0 = 1.0$. Because the lower discount rate slows convergence to the steady state, this example trains on $\mathcal{D} = \{0, 1, \ldots, 100\}$ and tests on $[0, 110]$.

*Assumption verification.* The primitives are continuously differentiable on the positive orthant, and the saddle path to the unique interior steady state is the unique trajectory satisfying Equation (10), with exponentially decaying derivatives (Acemoglu, 2008), so Assumptions 1 and 2 hold.

*Results.* Figure 1 shows the consumption and capital relative to a benchmark.[19] The kernel approximation recovers the optimal solution almost perfectly, fulfilling Equation (10) despite not being provided the steady state as a boundary condition.

The vertical dotted lines mark the end of the training set. While not our primary goal, the solution method extrapolates accurately, allowing it to learn the steady state. As discussed in Section 3, this is possible because we approximate the derivatives rather than the functions themselves: the Matérn derivative basis decays to zero, so the extrapolated level flattens toward a finite constant learned from the DAE.

For this model, any trajectory that violates transversality has a divergent derivative seminorm (Theorem 1), while the optimal solution's seminorm is finite. Because this divergence is the key failure of transversality, no modification to the optimization problem Equation (13) was required in practice.[20]

*Robustness.* While Theorem 3 demonstrates consistency, it is helpful to check our methods' sensitivity to hyperparameters and features of $\mathcal{D}$. Supplemental Appendix Section 2.1 shows that our methods perform well with a much sparser and irregular $\mathcal{D}$, and Supplemental Appendix Section 2.2 indicates low sensitivity to different kernel hyperparameters. Supplemental Appendix Section 2.3 demonstrates that the approximation

[19] The benchmark solves the model as a mixed initial-boundary value problem, using the analytically calculated steady state as a boundary condition, as discussed in Section 2.

[20] In general, as problems get larger, it is helpful to add non-negativity or box-bounding constraints from Assumption 2 to the optimization problem minimizing Equation (13). While these will often be nonbinding in the optimal solution, they can help optimizers converge.

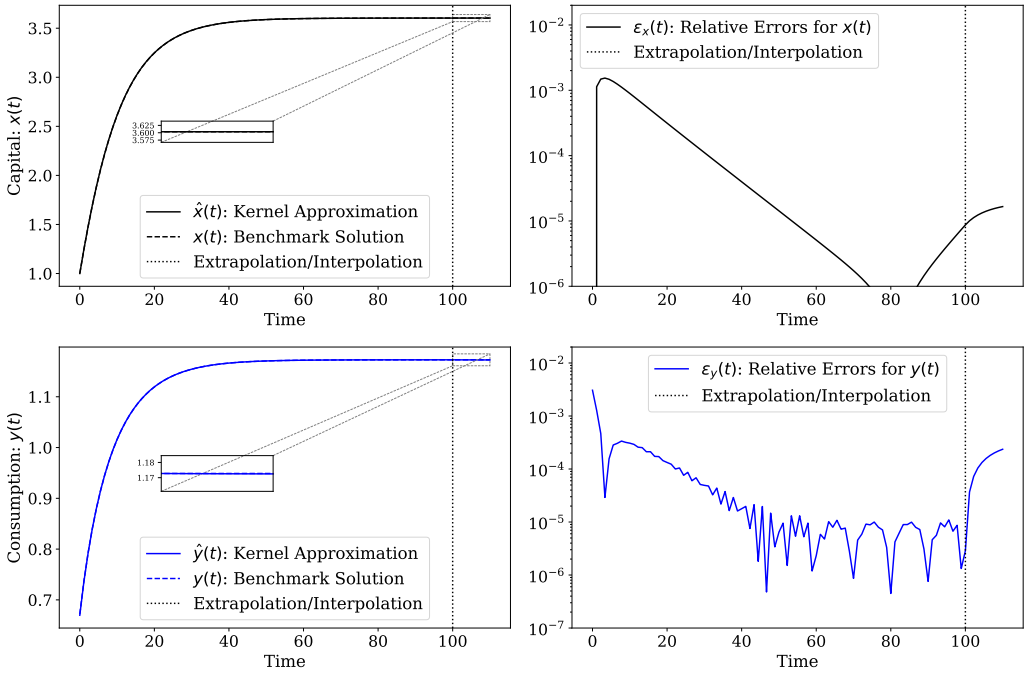



FIGURE 1. Solution of the neoclassical growth model (Equations (6) to (9)) without imposing the transversality condition (Equation (10)). The vertical dotted lines mark the end of the training set. The top panels show the results for capital, $x(t)$, and its relative error, and the bottom two show results for consumption, $y(t)$. The relative errors are low, on the order of $0.02\%$ on average and below $0.31\%$ everywhere, even when extrapolating beyond $\mathcal{D}$. Median computation time: $0.136$ seconds.

remains effective in the short to medium term, even if $\mathcal{D}$ does not contain large time values, which correspond to the solution getting very close to the steady state. This is an important consideration for problems where the convergence rate is unknown and the appropriate choice of $T$ is not clear a priori. Finally, Supplemental Appendix Section 2.4 shows how to solve models on a balanced growth path in levels, without detrending, by rescaling the approximation class with a learned balanced-growth parameter. The algorithm estimates the growth factor effectively, since the correct growth rate leads to the smallest RKHS norm of the detrended function.

### 4.2 *Neoclassical Growth Model with Multiple Steady States*

We now turn to a more complex version of the neoclassical growth model where $f(x) := A\max\{x^a, b_1 x^a - b_2\}$, as in Azariadis and Drazen (1990), Skiba (1978) where $A = 0.5, b_1 = 3.0$, and $b_2 = 2.5$. This example keeps log utility ($\gamma = 1$) with $a = \frac{1}{3}$, $\delta = 0.1$, and $r = 0.11$. The derivative of the production function, $f'(x)$, has a jump discontinuity at the kink $\bar{x} = \left(\frac{b_2}{b_1 - 1}\right)^{\frac{1}{a}}$, where we evaluate Equation (7) with the derivative of the attaining branch. As a result, this problem has two steady states, one on each smooth branch. Away from

the Skiba indifference point separating their basins, where the planner is indifferent between two optimal paths, there is a unique optimal transition path for any given initial condition.

*Assumption verification.* Assumption 1 fails at the kink, and Assumption 2(i) fails because the same $x_0$ can have a candidate path toward each steady state (Supplemental Appendix Section 1.1). Both hold locally around each steady state, and we show that the method works in practice.

This poses a significant challenge for classical algorithms, such as shooting and boundary value problem (BVP) methods, because the algorithms rely on analytic characterization of the steady-state value to impose the correct boundary value for a particular initial condition $x_0$. In practice, the set of steady states and the partitioning of the initial conditions into domains of attraction are not known a priori and are rarely computed outside of simple problems.

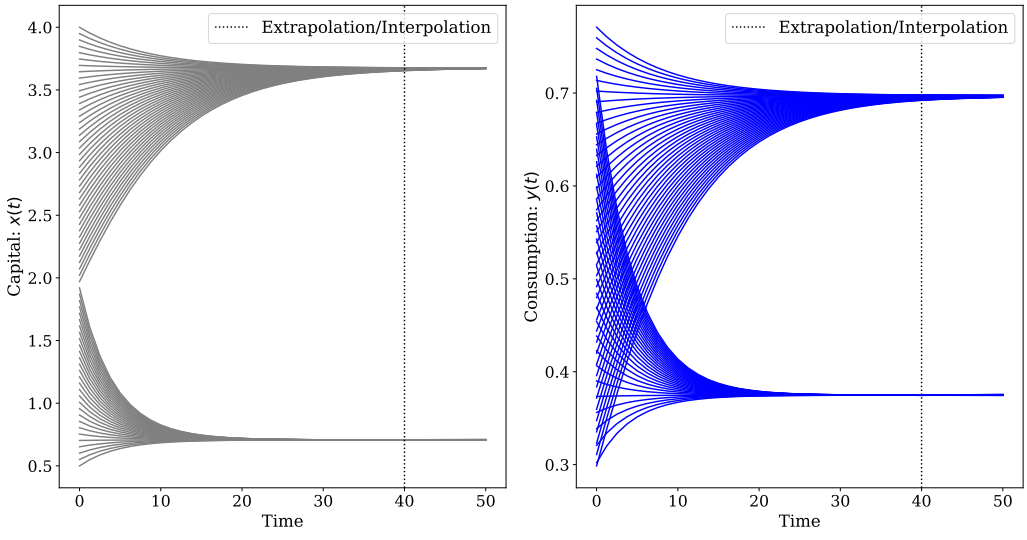



FIGURE 2. Solution of the neoclassical growth model (Equations (6) to (9)) with multiple steady states due to the concave-convex production function in Section 4.2. The left panel shows the solution trajectories for capital, $x(t)$, for 70 different initial conditions $x_0 \in [0.5, 4]$. The right panel shows the solution trajectories for consumption, $y(t)$. Median computation time per initial condition: 0.0021 seconds.

*Results.* Figure 2 shows the results using Matérn kernels for 70 different initial conditions across the basins of attraction for the two steady states.

In this case there is an additional concern that an approximate solution might jump into a different basin of attraction and go to the wrong steady state. While possible, we found that all of the examples converge to the "correct" steady state, i.e., the steady state that corresponds to the supplied initial condition $x_0$, despite not being provided the set of steady states or their basins of attraction.

Intuitively, this behavior is also a consequence of the minimum-norm solution. Consider the two possible trajectories from $x_0$ to each of the two steady states: the trajectory with smaller gradients and less steep dynamics will have a smaller norm.

### 4.3 *Linear Asset Pricing*

Models of asset pricing and rational bubbles are relatively simple and often admit closed-form solutions. These models have traditionally served a pedagogical role in exploring transversality conditions (see, e.g., Ljungqvist and Sargent, 2018). In this context, the transversality condition is also referred to as the "no-bubble" condition.[21]

*Model.* The model values a stream of dividends, where a "bubble" is defined as the residual after calculating the present-discounted value of the dividends.

Let $x(t) \in \mathbb{R}_{++}$ be the flow payoffs from a claim to an asset, and $\mu(t) \in \mathbb{R}_{++}$ be the price of a claim to that asset. For simplicity, we assume that the flow payoffs, $x(t)$, follow a deterministic linear process. For a given $x_0$, the key equations are

$$\dot{x}(t) = c + gx(t), \tag{14}$$

$$\dot{\mu}(t) = r\mu(t) - x(t) = r\mu(t) - \mu(t)\underbrace{\frac{x(t)}{\mu(t)}}_{:=G(x(t),\mu(t))}, \tag{15}$$

$$0 = \lim_{t\to\infty} e^{-rt}\mu(t)x(t), \tag{16}$$

where $c$ and $g$ are constants governing the dividend process, and $r > 0$ denotes the discount rate of a risk-neutral investor. Equation (16) instantiates the asymptotic boundary condition Equation (5) for this model. The standard "no-bubble" condition is $\lim_{t\to\infty} e^{-rt}\mu(t) = 0$. Because the dividend converges to $-c/g > 0$ in our parameterization, the two conditions select the same solution, and we refer to Equation (16) as the no-bubble condition.[22]

The set of trajectories satisfying Equations (14) and (15), without imposing Equation (16), can be found analytically for $r > g$ and $g \neq 0$:

$$\mu(t) = \mu_f(t) + \zeta e^{rt}, \tag{17}$$

$$\mu_f(t) := \int_0^\infty e^{-r\tau}x(t+\tau)\mathrm{d}\tau = -\tfrac{c}{rg} + \frac{1}{r-g}\left(x_0 + \tfrac{c}{g}\right)e^{gt}, \tag{18}$$

where $\mu_f(t)$ is interpreted as the "fundamental" price of the asset, and $\zeta \geq 0$ is indeterminate. However, when the "no-bubble" condition (i.e., Equation (16)) is imposed, this problem is well-posed, with a unique solution of $\mu(t) = \mu_f(t)$ (i.e., $\zeta = 0$).

*Assumption verification.* The primitives are smooth on the positive orthant, and by Equation (17) $\mu_f$ is the unique trajectory satisfying Equation (16), with exponentially decaying derivatives since $g < 0$, so Assumptions 1 and 2 hold.

[21] For asset pricing models, see Blanchard and Watson (1982), Diba and Grossman (1988), which characterize the set of solutions not fulfilling transversality and connect them to economic bubbles.

[22] The product form is not equivalent in general: if $x(t) \to 0$ (e.g., $c = 0$ and $g < 0$), every bubble $\zeta e^{rt}$ satisfies Equation (16) while violating $\lim_{t\to\infty} e^{-rt}\mu(t) = 0$.

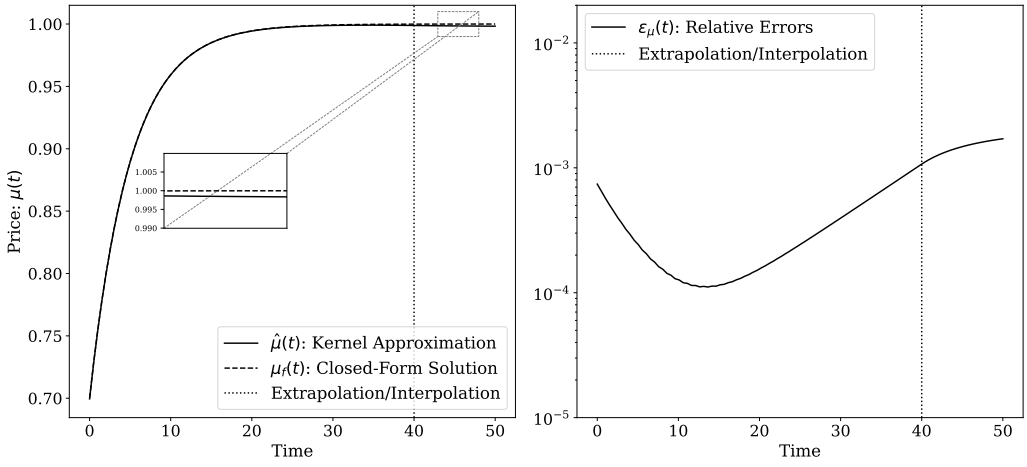



FIGURE 3. Solution of the linear asset pricing model (Equations (14) and (15)) without imposing the "no-bubble" condition (Equation (16)). The vertical dotted lines mark the end of the training set. Median computation time: 0.0599 seconds.

*Selecting the "no-bubble" solution.* This is an example where the result of Theorem 1 can be seen clearly. Equation (17) gives, in closed form, every trajectory that fails Equation (16), indexed by $\zeta > 0$ through a bubble $\zeta e^{rt}$. Its derivative $\zeta r e^{rt}$ has $\mathcal{W}^{1,2}([0,T])$ norm growing like $e^{rT}$, so the seminorm diverges for any bubble.

*Results.* Our baseline parameters are $x_0 = 0.01, c = 0.02, g = -0.2$, and $r = 0.1$. Figure 3 shows the results of the ridgeless kernel machine alongside the fundamental price from Equation (18). Even without imposing the long-run "no-bubble" condition in Equation (16), the kernel method recovers the "fundamental" price with high accuracy. As before, minimizing the norm still selects the correct solution, and as $T$ grows the norms of non-solutions will grow due to Theorem 1.

### 4.4 *Small Open Economy*

Building on the growth model, we next solve a continuous-time small open economy in which capital accumulation can be financed through an internationally traded bond, adapting the capital-installation technology of Ravikumar et al. (2019) to continuous time. The closest continuous-time antecedent is Turnovsky and Sen (1991). With a fixed world interest rate, the model deliberately lacks an isolated steady state: the long-run net foreign asset position inherits its initial condition (Schmitt-Grohé and Uribe, 2003). Rather than closing the model with a stationarity-inducing device (e.g., a debt-elastic interest premium), we retain the continuum of steady states, so no terminal state can be computed from primitives before solving for the transition path.

*Model.* In the notation of Equations (1) to (5) there are $M = 2$ states and co-states and $P = 2$ jump variables. The states are capital, $x_1(t) \in \mathbb{R}_{++}$, and net foreign assets, $x_2(t) \in \mathbb{R}$, which earn a fixed world interest rate $r^\star$ and may take either sign (negative values are external debt). The co-states are the marginal value of installed capital,

$\mu_1(t) \in \mathbb{R}_{++}$, and the marginal value of wealth, $\mu_2(t) \in \mathbb{R}_{++}$. The jump variables are consumption, $y_1(t) \in \mathbb{R}_{++}$, and investment expenditure, $y_2(t) \in \mathbb{R}_{++}$. The household has CRRA flow utility $\frac{y_1^{1-\gamma}-1}{1-\gamma}$ with discount rate $r$, production is $f(x_1) = x_1^a$, and investment is installed through the concave technology $\chi y_2^{\eta} x_1^{1-\eta}$ with $\eta \in (0,1)$ and $\chi := \delta^{1-\eta}$, normalized so that steady-state investment expenditure equals replacement, $y_2^\star = \delta x_1^\star$.[23] Written in our notation, the DAE is

$$\dot{x}_1(t) = \chi\, y_2(t)^{\eta}\, x_1(t)^{1-\eta} - \delta x_1(t) := F_1(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t)) \tag{19}$$

$$\dot{x}_2(t) = r^\star x_2(t) + f\left(x_1(t)\right) - y_1(t) - y_2(t) := F_2(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t)) \tag{20}$$

$$\dot{\mu}_1(t) = r\mu_1(t) - \mu_1(t) \underbrace{\left[(1-\eta)\,\chi \left(\frac{y_2(t)}{x_1(t)}\right)^{\eta} - \delta + \frac{\mu_2(t)}{\mu_1(t)} f'\left(x_1(t)\right)\right]}_{:= G_1(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t))} \tag{21}$$

$$\dot{\mu}_2(t) = r\mu_2(t) - \mu_2(t) \underbrace{r^\star}_{:= G_2(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t))} \tag{22}$$

$$0 = \mu_2(t)\, y_1(t)^{\gamma} - 1 := H_1(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t)) \tag{23}$$

$$0 = \eta\, \chi\, \mu_1(t) \left(\frac{y_2(t)}{x_1(t)}\right)^{\eta-1} - \mu_2(t) := H_2(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t)) \tag{24}$$

$$x_1(0) = k_0, \qquad x_2(0) = b_0 \tag{25}$$

$$\mathbf{0} = \lim_{t\to\infty} e^{-rt}\, \boldsymbol{x}(t) \odot \boldsymbol{\mu}(t) \tag{26}$$

Equations (19) to (22) are differential and Equations (23) and (24) are algebraic. The ratio $q(t) := \mu_1(t)/\mu_2(t)$ is the shadow value of installed capital in consumption units (Tobin's $q$), and Equation (24) makes the investment rate an algebraic function of it, $y_2(t)/x_1(t) = (\eta\chi q(t))^{\frac{1}{1-\eta}}$. The asset component of Equation (26), $\lim_{t\to\infty} e^{-rt}\mu_2(t)\, x_2(t) = 0$, combines the no-Ponzi and transversality restrictions on foreign borrowing.

*Assumption verification.* The primitives are continuously differentiable for positive capital, co-states, and jump variables, and the continuum of steady states below is hysteresis rather than multiplicity of the trajectory (Supplemental Appendix Section 1.1), so Assumptions 1 and 2 hold.[24]

We solve an equivalent reduced system of Equations (19) to (24). Inverting Equation (24) for the investment rate eliminates that row, and the co-states enter the remaining equations only through $q(t)$ and $\mu_2(t)$, leaving four differential equations and the single algebraic condition Equation (23) in the paths $(x_1, x_2, q, \mu_2, y_1)$. All five paths are

[23] The derivation follows the household problem $\max \int_0^\infty e^{-rt} \frac{y_1(t)^{1-\gamma}-1}{1-\gamma}\, \mathrm{d}t$ subject to Equations (19) and (20), with Hamiltonian $\frac{y_1^{1-\gamma}-1}{1-\gamma} + \mu_2\left[r^\star x_2 + f(x_1) - y_1 - y_2\right] + \mu_1\left[\chi y_2^{\eta} x_1^{1-\eta} - \delta x_1\right]$. The DAE follows from applying Pontryagin's Maximum Principle, as in the neoclassical growth model of Section 2. The installation technology is the continuous-time analogue of the one in Ravikumar et al. (2019).

[24] Net foreign assets can be negative, but they take a constant sign for all $t > T_0$, so the change of variables noted in Section 2 applies and our results carry over.

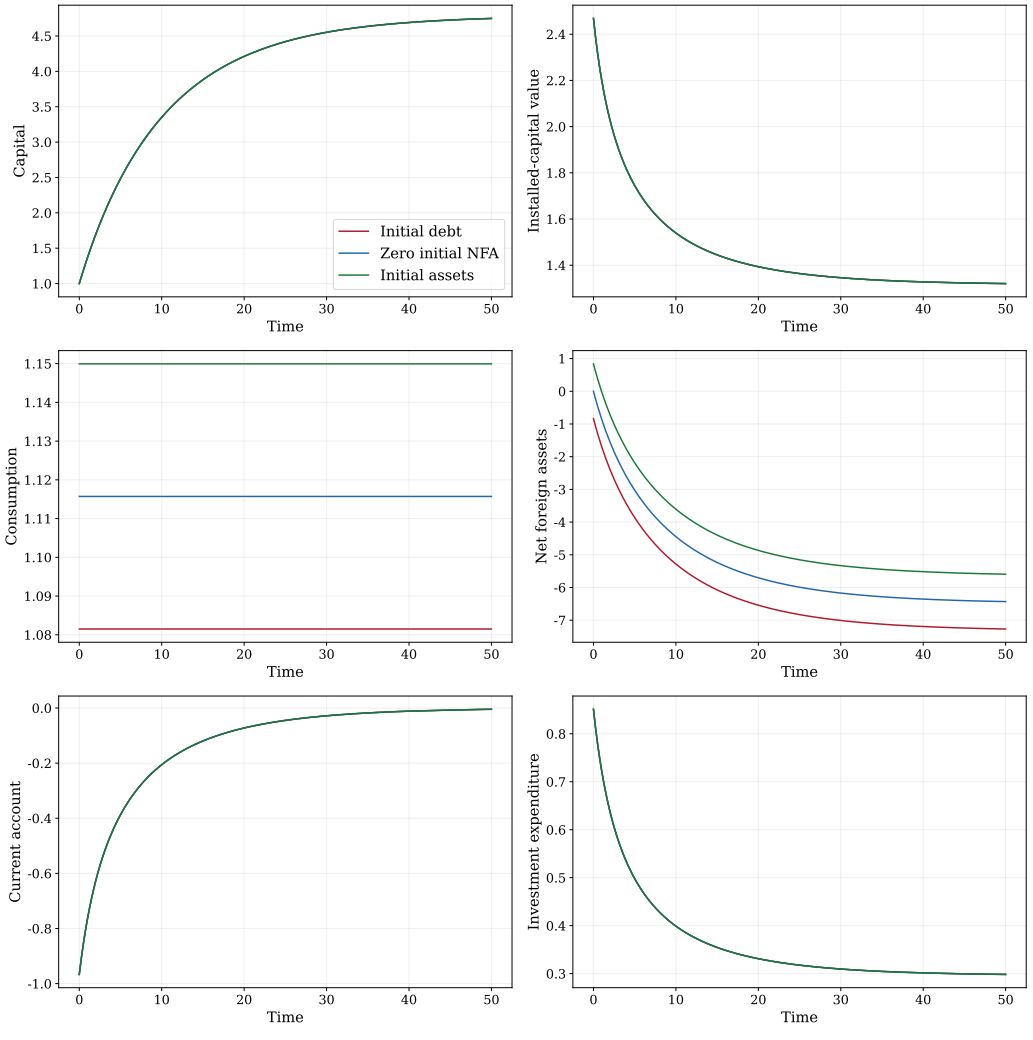



FIGURE 4. Solution of the small open economy model (Equations (19) to (25)) without imposing the transversality conditions (Equation (26)), for three initial net foreign asset (NFA) positions $b_0$. Capital, the installed-capital value $q(t)$, investment, and the current account coincide across the three cases. Consumption is flat at a different permanent level for each $b_0$. The net foreign asset paths shift in parallel toward different long-run positions, each a point on the continuum of steady states characterized by Supplemental Appendix Equation (S2). Median computation time for the baseline: 0.161 seconds.

represented in levels. Positivity of $x_1$, $q$, $\mu_2$, and $y_1$ is not imposed and holds ex post at the solution.

Our baseline sets $r^\star = r$, so steady-state capital is unique while the long-run asset position is indexed by its initial condition: a continuum of steady states, with flat consumption along every transition. Supplemental Appendix Section 3 derives the steady states and the exact comparative statics in $b_0$ that the results below check.

*Results.* Our baseline parameters follow the annual calibration of Ravikumar et al. (2019): $a = 0.33$, $\gamma = 2$, $\eta = 0.76$, $r = -\log(0.96) \approx 0.041$, $\delta = -\log(0.94) \approx 0.062$, and $r^\star = r$, with $k_0 = 1$ and initial net foreign assets $b_0 \in \{-0.84,\, 0,\, 0.84\} \approx \{-\frac{1}{2}f(x_1^\star),\, 0,\, \frac{1}{2}f(x_1^\star)\}$. These are initial debt, a zero initial position, and initial assets.[25] Figure 4 shows the kernel solutions for the three initial positions. The paths recover an independent benchmark with a maximum error of $1.22 \times 10^{-3}$: relative errors of $2.38 \times 10^{-4}$ for capital, $2.77 \times 10^{-4}$ for $q$, and $2.53 \times 10^{-5}$ for consumption, and $1.22 \times 10^{-3}$ for net foreign assets measured relative to steady-state output, since $x_2(t)$ changes sign and has no meaningful relative error.[26] The solutions display the exact comparative statics of the model derived in Supplemental Appendix Section 3. Capital, $q(t)$, and investment are identical across the three cases to $1.9 \times 10^{-13}$, and the current account to $2.02 \times 10^{-8}$. Consumption differs across them by $r^\star \Delta b_0$ to $4.03 \times 10^{-9}$. The net foreign asset paths are parallel, $\Delta x_2(t) = \Delta b_0$ to $4.17 \times 10^{-7}$, converging to permanently different long-run positions. Flat consumption is a sharp check on the correctness of the solution. With $r^\star = r$, Equations (22) and (23) require perfect consumption smoothing, so any curvature in $y_1(t)$ would signal an error. It is flat as an outcome rather than a restriction. $y_1$ is an unrestricted path in the approximation class, and the minimum-norm solution drives its fitted RKHS norm to $3.8 \times 10^{-14}$, with $\max_t |y_1(t) - y_1(0)| = 1.72 \times 10^{-8}$. No transversality, no-Ponzi, steady-state, or other terminal condition was imposed. The minimum-norm solution selects the path fulfilling Equation (26) separately for each initial condition, even though the terminal asset position differs in every case.

### 4.5 *Multicountry Growth with an Endogenous World Interest Rate*

Our final example expands the small open economy of Section 4.4 to a multi-country growth model with interest rates determined in equilibrium, and transition dynamics generated by heterogeneity in both the capital and bond initial positions. Consider $N$ countries with fixed populations $L_i$ that trade a single bond, so the world interest rate becomes an endogenous algebraic path rather than a fixed parameter.[27] Supplemental Appendix Section 4 derives the model from the household problem and states the reduced system that is solved, while this section states the equations and reports results.

*Model.* Each country $i = 1, \ldots, N$ has the same technology and discount rate as in Section 4.4, with output scaled by a country productivity $z_i > 0$, equal to one in the baseline, flow utility with country-specific CRRA coefficient $\gamma_i$, and per-worker states, co-states, and jump variables $\left(x_{1,i}, x_{2,i}, \mu_{1,i}, \mu_{2,i}, y_{1,i}, y_{2,i}\right) \in \mathbb{R}_{++} \times \mathbb{R} \times \mathbb{R}^4_{++}$. Relative

[25] This example uses a longer horizon than the baseline examples: with the Matérn kernel $\nu = \frac{1}{2}$ and $\sigma = 1$, $\mathcal{D}$ contains 192 points on $[0, 160]$, concentrated at early dates (177 equally spaced on $[0, 40]$ and 15 equally spaced on $(40, 160]$), with lengthscale $\ell = 20$, tested on 201 points in $[0, 50]$.

[26] The benchmark solves a joint boundary value problem for $(x_1, q, x_2)$ with initial consumption $y_1(0)$ as an unknown parameter, imposing $x_1(0) = k_0$, $x_2(0) = b_0$, a log-linearized stable-manifold terminal condition for $q$, and a finite-horizon version of the intertemporal budget constraint Supplemental Appendix Equation (S3) with a tail correction along that stable arm.

[27] The closest continuous-time version of this is Alvarez (2017), which imposes continuously balanced trade. Here countries borrow and lend, which preserves the initial-asset history dependence of the small open economy, and Ravikumar et al. (2019) remains the target calibration.

to the small open-economy model, the fixed rate $r^\star$ is replaced by a world interest rate $r^\star(t)$, which is an additional jump variable. In the notation of Equations (1) to (5) there are $M = 2N$ states and co-states and $P = 2N + 1$ jump variables. For each $i$,

$$\dot{x}_{1,i}(t) = \chi\, y_{2,i}(t)^{\eta}\, x_{1,i}(t)^{1-\eta} - \delta x_{1,i}(t) \tag{27}$$

$$\dot{x}_{2,i}(t) = r^\star(t)\, x_{2,i}(t) + z_i f\left(x_{1,i}(t)\right) - y_{1,i}(t) - y_{2,i}(t) \tag{28}$$

$$\dot{\mu}_{1,i}(t) = r\mu_{1,i}(t) - \mu_{1,i}(t)\left[(1-\eta)\,\chi\left(\frac{y_{2,i}(t)}{x_{1,i}(t)}\right)^{\eta} - \delta + \frac{\mu_{2,i}(t)}{\mu_{1,i}(t)}\, z_i f'\left(x_{1,i}(t)\right)\right] \tag{29}$$

$$\dot{\mu}_{2,i}(t) = r\mu_{2,i}(t) - \mu_{2,i}(t)\, r^\star(t) \tag{30}$$

$$0 = \mu_{2,i}(t)\, y_{1,i}(t)^{\gamma_i} - 1 \tag{31}$$

$$0 = \eta\,\chi\,\mu_{1,i}(t)\left(\frac{y_{2,i}(t)}{x_{1,i}(t)}\right)^{\eta-1} - \mu_{2,i}(t) \tag{32}$$

and the world bond market clears,

$$0 = \sum_{i=1}^{N} L_i\, x_{2,i}(t), \tag{33}$$

which closes the algebraic block. The initial conditions and asymptotic boundary conditions are

$$x_{1,i}(0) = k_{i,0}, \qquad x_{2,i}(0) = b_{i,0} \text{ with } \sum_{i=1}^{N} L_i\, b_{i,0} = 0, \tag{34}$$

$$\mathbf{0} = \lim_{t\to\infty} e^{-rt}\, \boldsymbol{x}(t) \odot \boldsymbol{\mu}(t). \tag{35}$$

Supplemental Appendix Section 4.1 derives these equations from the household problem and market clearing.[28]

*Assumption verification.* The primitives are continuously differentiable for positive capital, co-states, and jump variables, so Assumption 1 holds. Unlike the previous examples, no algebraic equation determines the world interest rate $r^\star(t)$ at an instant. See Supplemental Appendix Section 1 for uniqueness in this case and Supplemental Appendix Section 4.2 for the model reduction that resolves it. The continuum of steady states is hysteresis as in Section 4.4, and we show that the method works in practice.

As in Section 4.4, we solve an equivalent reduced system of these equations in the paths $(x_{1,i}, x_{2,i}, q_i, \mu_{2,i}, y_{1,i}, r^\star)$, where $q_i := \mu_{1,i}/\mu_{2,i}$ is Tobin's $q$, the value of installed capital in consumption units (see Supplemental Appendix Section 4.2). A steady state requires $r^\star = r$ from Equation (30), so every country has the same steady-state capital,

[28] Summing the budget constraints Equation (28) weighted by $L_i$ gives $\dot{B}(t) = r^\star(t)B(t) + S(t)$ for the aggregate bond position $B := \sum_i L_i x_{2,i}$ and the world excess supply of goods $S := \sum_i L_i \left[z_i f\left(x_{1,i}\right) - y_{1,i} - y_{2,i}\right]$. Given $B(0) = 0$, bond-market clearing and goods-market clearing are equivalent, and only one is an independent restriction: this is Walras's law for this economy (Supplemental Appendix Section 4.2).

Supplemental Appendix Equation (S1), while the long-run cross-country asset distribution is restricted only by clearing: the continuum of steady states in consumption and bond holdings is similar to that of the small open economy, except with dimension $N-1$ (Supplemental Appendix Section 4.3).

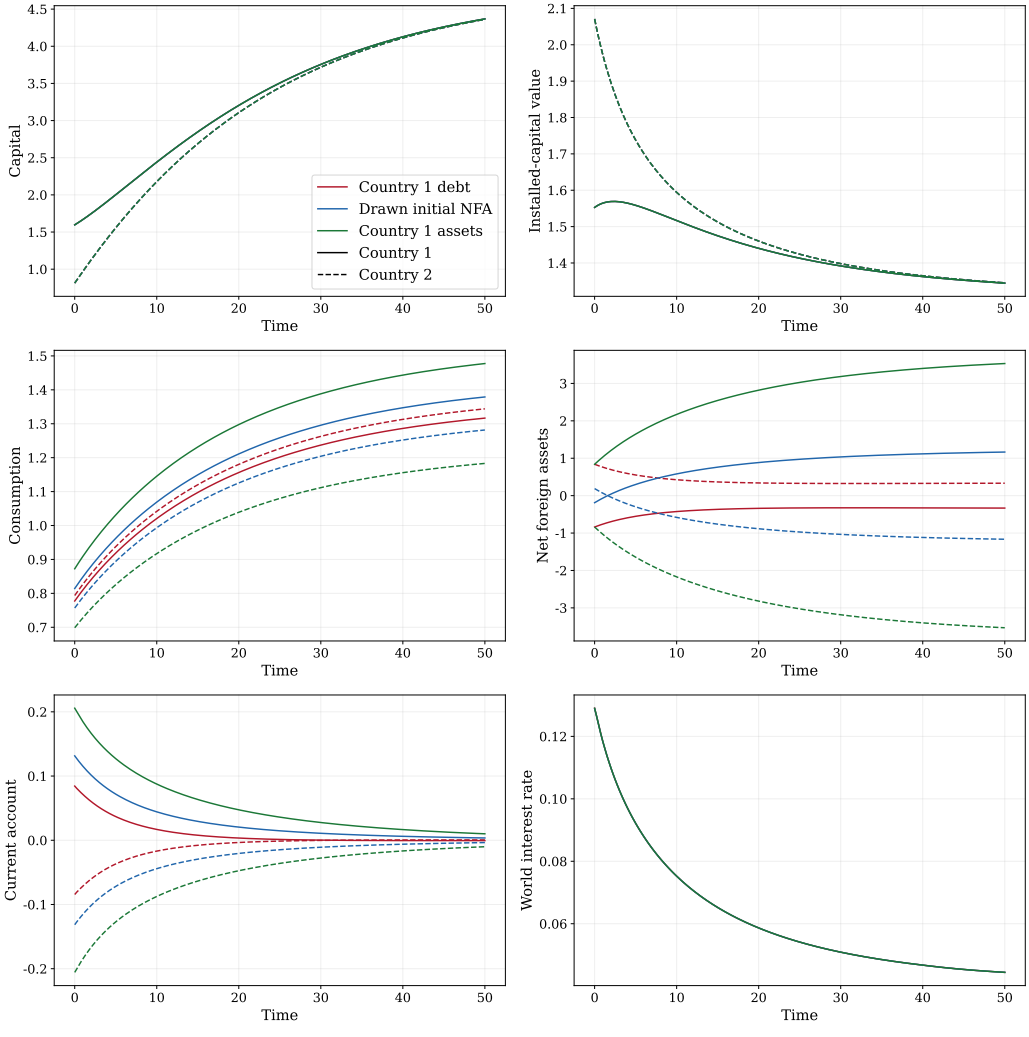



FIGURE 5. Solution of the two-country world economy (Equations (27) to (34)) without imposing the transversality conditions (Equation (35)), for three configurations of opposite initial net foreign assets (solid: country 1, dashed: country 2). The world interest rate is high while the capital-scarce world accumulates and converges to $r$ from above. Net foreign asset paths are mirror images by market clearing and converge to permanently different long-run distributions across the three configurations. Median computation time for the baseline: $0.768$ seconds.

*Results.* The baseline has $N = 2$ countries with $L_1 = L_2 = 1$, a common curvature $\gamma_i = \gamma = 2$, and the calibration of Section 4.4, differing only in their initial conditions.

Initial capital stocks are drawn once from a log-normal distribution around the initial condition of Section 4.4, $\log k_{i,0} \sim \mathcal{N}\left(0, 0.25^2\right)$, and are shared by all three configurations below. The draw is $(k_{1,0}, k_{2,0}) \approx (1.60, 0.81)$. The initial bond positions vary across configurations, with $(b_{1,0}, b_{2,0})$ approximately $(-0.84, 0.84)$, $(-0.19, 0.19)$, or $(0.84, -0.84)$. By the aggregation result of Supplemental Appendix Section 4.3, the bond configuration is immaterial for the world rate and the capital paths under the common $\gamma$, though not for consumption or the asset paths.[29]

Figure 5 shows the kernel solutions. Across the three configurations, training residuals are below $3.12 \times 10^{-6}$ and bond-market clearing holds to $1.79 \times 10^{-6}$ over the full training horizon. The world interest rate starts near $13\%$ and declines to $r$ as both countries accumulate capital toward the common $x_1^\star$, while the three initial bond configurations produce permanently different long-run creditor positions. This is the $N$-country analogue of the history dependence of the small open economy.

By the aggregation result of Supplemental Appendix Section 4.3, under a common $\gamma$ the paths of the world rate, capital, and $q_i$ are invariant to the initial bond distribution. On the plotted interval $[0, 50]$, the largest cross-configuration gaps are $1.16 \times 10^{-4}$ for the rate, $5.08 \times 10^{-7}$ for capital, and $1.94 \times 10^{-5}$ for $q$.

*Scaling in the number of countries.* To show how the method scales, Figure 6 reports computation times as $N$ grows, drawing populations, productivities, and initial conditions at each $N$ and comparing two linear solvers for the same problem and solver (Perla, 2026): a dense factorization, and an iterative method using the conjugate gradient (CG) linear solver with a specialized preconditioner designed to take advantage of the economic structure of the problem (see Supplemental Appendix Section 4.4 for details). The problem scales well. At $N = 100$, the reduced system has $5N + 1 = 501$ equations and the discretized system on the 192 collocation points has $96{,}192$ residual equations in $96{,}493$ unknowns. The dense factorization takes $47.6$ seconds at $N = 10$, while the preconditioned CG solver takes $3.62$ seconds at $N = 10$ and $55.7$ seconds at $N = 100$, with cost growing roughly linearly in $N$. Every solution satisfies the same maximum training-residual tolerance of $10^{-5}$.

[29] The training grid has 192 points on $[0, 160]$, concentrated at early dates (177 equally spaced on $[0, 40]$ and 15 equally spaced on $(40, 160]$), and we use a Matérn kernel with $\nu = \frac{1}{2}$ and lengthscale $\ell = 20$.

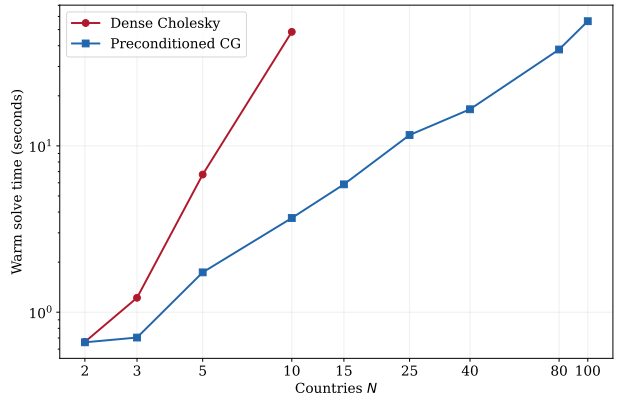



FIGURE 6. Computation time of the multicountry model against the number of countries $N$ (log–log, double precision), using a Matérn-$1/2$ kernel with $\ell = 20$ and drawing country populations, productivities, and initial conditions at every width. Each point is the median of two warm solves, timing the solver call alone. The dense solver is reported through $N = 10$, and the preconditioned CG solver of Supplemental Appendix Section 4.4 through $N = 100$, with a median computation time of $55.7$ seconds at that size. Both solvers use the same settings and return accepted solutions wherever both are reported.

## 5. STATE SPACE FORMULATIONS

In this section we present a recursive and stochastic state-space formulation of this class of models. Supplemental Appendix Section 5 works through the examples, and Section 6 suggests future research expanding the theory in Section 3 to this class of models.

### 5.1 *Setup*

*The mathematical problem.* Let $\boldsymbol{x}(t) \in \mathbb{R}^M$ be the state and $\boldsymbol{W}(t)$ a standard $M$-dimensional Brownian motion on a filtered probability space. The first $M_z \geq 0$ components of $\boldsymbol{x}$, denoted $\boldsymbol{x}_z$, are exogenous. The remaining $M - M_z$, denoted $\boldsymbol{x}_e$, are endogenous. The unknowns are now functions of the state rather than of time: a co-state function $\boldsymbol{\mu} : \Omega \to \mathbb{R}^{M-M_z}$, one per endogenous state, and a policy $\boldsymbol{y} : \Omega \to \mathbb{R}^P$ for the jump variables, defined on a region $\Omega \subseteq \mathbb{R}^M$ of the state space.[30] Given these functions, the stochastic state-space formulation of the problem, written as a system of stochastic differential equations (SDEs), is

$$\mathrm{d}\boldsymbol{x}(t) = \boldsymbol{F}\big(\boldsymbol{x}(t), \boldsymbol{\mu}(\boldsymbol{x}(t)), \boldsymbol{y}(\boldsymbol{x}(t))\big)\,\mathrm{d}t + \boldsymbol{B}\big(\boldsymbol{x}(t), \boldsymbol{\mu}(\boldsymbol{x}(t)), \boldsymbol{y}(\boldsymbol{x}(t))\big)\,\mathrm{d}\boldsymbol{W}(t), \tag{36}$$

$$\mathrm{d}\boldsymbol{\mu}(\boldsymbol{x}(t)) = \Big[r\boldsymbol{\mu}(\boldsymbol{x}(t)) - \boldsymbol{\mu}(\boldsymbol{x}(t)) \odot \boldsymbol{G}\big(\boldsymbol{x}(t), \boldsymbol{\mu}(\boldsymbol{x}(t)), \boldsymbol{y}(\boldsymbol{x}(t))\big)\Big]\mathrm{d}t$$

[30] This informal treatment omits the regularity conditions that a consistency proof would require, and the full set is not known. They include sufficient smoothness of $\boldsymbol{\mu}$ to evaluate Equation (42) pointwise when $\boldsymbol{B} \neq \boldsymbol{0}$, and conditions beyond Assumption 1 for a unique global strong solution that remains in $\Omega$ with sufficient integrability for the expectation in Equation (40). Identifying them is a research agenda of its own (see Section 6).

$$+ \nabla \boldsymbol{\mu}(\boldsymbol{x}(t))\, \boldsymbol{B}\big(\boldsymbol{x}(t), \boldsymbol{\mu}(\boldsymbol{x}(t)), \boldsymbol{y}(\boldsymbol{x}(t))\big)\, \mathrm{d}\boldsymbol{W}(t), \tag{37}$$

$$\mathbf{0} = \boldsymbol{H}\big(\boldsymbol{x}(t), \boldsymbol{\mu}(\boldsymbol{x}(t)), \boldsymbol{y}(\boldsymbol{x}(t))\big), \tag{38}$$

$$\boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{39}$$

$$\mathbf{0} = \lim_{t\to\infty} \mathbb{E}\big[e^{-rt}\, \boldsymbol{x}_e(t) \odot \boldsymbol{\mu}(\boldsymbol{x}(t)) \,\big|\, \boldsymbol{x}_0\big] \qquad \text{for every } \boldsymbol{x}_0 \in \Omega, \tag{40}$$

where $\boldsymbol{F}$, $\boldsymbol{G}$, and $\boldsymbol{H}$ are those of Equations (1) to (3) and the transversality condition Equation (40) is now an expectation over the Brownian motion $\boldsymbol{W}(t)$. The only additional primitive is the volatility matrix $\boldsymbol{B} : \mathbb{R}^M \times \mathbb{R}^{M-M_z} \times \mathbb{R}^P \to \mathbb{R}^{M\times M}$, and $\nabla\boldsymbol{\mu}$ is the Jacobian of $\boldsymbol{\mu}$. As in Section 2, only the state carries an initial condition, and Equation (40) enforces an asymptotic boundary condition on $\boldsymbol{\mu}(\boldsymbol{x}(t))$.

*Residual formulation.* Let $\mathcal{L}$ be the infinitesimal generator of the diffusion Equation (36), which for a twice continuously differentiable test function $\varphi : \Omega \to \mathbb{R}$ is

$$\begin{aligned}\mathcal{L}\varphi(\boldsymbol{x}) := \nabla\varphi(\boldsymbol{x}) \cdot \boldsymbol{F}\big(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{y}(\boldsymbol{x})\big) \\ + \tfrac{1}{2}\operatorname{tr}\Big(\boldsymbol{B}\big(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{y}(\boldsymbol{x})\big)\, \boldsymbol{B}\big(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{y}(\boldsymbol{x})\big)^{\top} \nabla^2\varphi(\boldsymbol{x})\Big),\end{aligned} \tag{41}$$

and which acts componentwise on a vector-valued test function. Equations (37) and (38) hold along every trajectory from every $\boldsymbol{x}_0 \in \Omega$ if, for every $\boldsymbol{x} \in \Omega$,[31]

$$r\boldsymbol{\mu}(\boldsymbol{x}) = \boldsymbol{\mu}(\boldsymbol{x}) \odot \boldsymbol{G}\big(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{y}(\boldsymbol{x})\big) + \mathcal{L}\boldsymbol{\mu}(\boldsymbol{x}), \tag{42}$$

$$\mathbf{0} = \boldsymbol{H}\big(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{y}(\boldsymbol{x})\big), \tag{43}$$

where the transversality condition Equation (40) is still required for each $\boldsymbol{x}_0$. Stacking, define

$$\boldsymbol{R}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{y}) := \begin{bmatrix} r\boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{x}) \odot \boldsymbol{G}\big(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{y}(\boldsymbol{x})\big) - \mathcal{L}\boldsymbol{\mu}(\boldsymbol{x}) \\ \boldsymbol{H}\big(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{y}(\boldsymbol{x})\big) \end{bmatrix}. \tag{44}$$

Global solution methods bound this uniformly, $\min_{\boldsymbol{\mu},\boldsymbol{y}} \sup_{\boldsymbol{x}\in\Omega} \big\|\boldsymbol{R}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{y})\big\|_2^2$ subject to Equation (40), and ideal algorithms drive this error to zero as their resolution, such as the number of grid points, grows.

However, outside of linear models or low-dimensional and compact $\Omega$, finding a global solution of this sort is typically intractable, regardless of the approximation class used for $\boldsymbol{\mu}$ and $\boldsymbol{y}$. Instead, most ML algorithms (e.g., kernels, NNs, or even spline regressions) find a statistical solution: rather than bounding the error uniformly on the whole domain, they achieve a small error on the important regions of the state space as measured by some distribution over the states.

## 5.2 *Method*

*Risk minimization.* The distribution over states is a modeling primitive for these types of statistical formulations. It may be externally given, for example as empirical samples

[31] See Øksendal (2003, Theorem 7.3.3) for the generator and Achdou et al. (2022) for the derivation of such conditions from the Hamilton–Jacobi–Bellman equation.

as in standard supervised learning, or generated adaptively from the solution itself, by simulating the law of motion of the state under the unknown policy functions from an initial condition.

For an external distribution $\rho$ on $\Omega$, the problem is risk minimization in the sense of statistical decision theory (Murphy, 2022, section 5.4),

$$\min_{\boldsymbol{\mu},\boldsymbol{y}} \mathbb{E}_{\boldsymbol{x}\sim\rho}\left[\left\|\boldsymbol{R}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{y})\right\|_2^2\right] \quad \text{s.t.} \quad \text{Equation (40)}, \tag{45}$$

where $\boldsymbol{\mu}, \boldsymbol{y}$ lie in a function space of sufficient regularity and measurability for the problem to be well defined.

Alternatively, the distribution of relevant states may be induced by the solution itself. For the adaptive case, let $\mathbb{W}_T$ be Wiener measure on $[0,T]$, the law of a standard Brownian motion path (Øksendal, 2003, Chapter 2), and let $\boldsymbol{x}_0 \sim \rho_0$, which may be a point mass. One version of this problem is

$$\begin{aligned} \min_{\boldsymbol{\mu},\boldsymbol{y}} \quad & \mathbb{E}_{\boldsymbol{x}_0\sim\rho_0,\,\boldsymbol{W}\sim\mathbb{W}_T}\left[\tfrac{1}{T}\int_0^T \left\|\boldsymbol{R}(\boldsymbol{x}(t);\boldsymbol{\mu},\boldsymbol{y})\right\|_2^2 \mathrm{d}t\right] \\ \text{s.t.} \quad & \text{Equations (36), (39) and (40)}, \end{aligned} \tag{46}$$

so that $\boldsymbol{x}(t)$ is the trajectory from $\boldsymbol{x}_0$ under $(\boldsymbol{\mu},\boldsymbol{y})$ themselves given the Brownian path $\boldsymbol{W}$, and the only distributional primitives are $\rho_0$ and $T$.[32] Crucially, the expectation is over random variables that are independent of the unknown functions.[33] In deterministic state-space formulations, the expectation is over $\rho_0$ alone, which can even be a point mass, in which case the trajectory solving Equations (36) and (39) is deterministic.

*Empirical risk for fixed distributions.* The empirical counterpart to Equation (45) is empirical risk minimization (Murphy, 2022, section 5.4): for an external $\rho$, take samples $\mathcal{D} := \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$ from $\rho$ and solve

$$\min_{\hat{\boldsymbol{\mu}},\hat{\boldsymbol{y}}} \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{R}(\boldsymbol{x}_i;\hat{\boldsymbol{\mu}},\hat{\boldsymbol{y}})\right\|_2^2. \tag{47}$$

The inductive bias plays the same role as in Equation (13): among the functions that attain zero empirical risk, select the interpolant of minimum seminorm,

$$\min_{\hat{\boldsymbol{\mu}},\hat{\boldsymbol{y}}} \left\{\sum_{m=1}^{M-M_z}\|\hat{\boldsymbol{\mu}}^{(m)}\|_\psi^2 + \sum_{m=1}^{P}\|\hat{\boldsymbol{y}}^{(m)}\|_\psi^2\right\} \quad \text{s.t.} \quad \boldsymbol{R}(\boldsymbol{x};\hat{\boldsymbol{\mu}},\hat{\boldsymbol{y}}) = \boldsymbol{0} \quad \text{for all } \boldsymbol{x}\in\mathcal{D}, \tag{48}$$

where $\|\cdot\|_\psi$ is a seminorm on functions on $\Omega$ that measures the smoothness of the unknown functions across states rather than in time.

[32] To emphasize short-run accuracy, use a discount factor $e^{-\tilde{r}t}$ inside the integral, as in Wang et al. (2020).

[33] Alternatively, let $\rho(\cdot;\boldsymbol{\mu},\boldsymbol{y})$ be the time average over $[0,T]$ of the law of $\boldsymbol{x}(t)$. Then the objective is $\mathbb{E}_{\boldsymbol{x}\sim\rho(\cdot;\boldsymbol{\mu},\boldsymbol{y})}\left[\|\boldsymbol{R}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{y})\|_2^2\right]$, i.e. Equation (45) with a distribution that depends on the unknowns. The challenge is that an empirical version of that form must differentiate $\rho(\cdot;\boldsymbol{\mu},\boldsymbol{y})$ with respect to the parameters of the unknown functions, or else it is a misspecified counterpart to Equation (45) and ignores how the unknown functions move the states (Murphy, 2023, section 6.3).

For kernel methods, $\|\cdot\|_\psi$ is the RKHS norm $\|\cdot\|_{\mathcal{H}}$ of a kernel $k : \Omega \times \Omega \to \mathbb{R}$. We approximate each function as a weighted sum of the kernel at the sampled states, so Equation (48) parallels the problem of Section 3.1 with states in place of times. The norm typically penalizes derivatives with respect to $\boldsymbol{x}$. The finite-dimensional problem has $N$ coefficients per function regardless of the dimension of $\Omega$, which enters only through the scalar-valued kernel capturing the degree of similarity between two states (Rasmussen and Williams, 2006).

Although the size of the empirical problem is independent of the dimension of $\Omega$, whether a modest $N$ suffices as the dimension grows depends on the particular problem and on the choice of kernel, which encodes a "representation" of the state (Section 6). In some cases NNs are preferred for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{y}}$ because they learn a representation (e.g., Achdou et al., 2026), even though their parameter count grows with the dimension of $\Omega$. A further tradeoff is that an NN solving Equation (47) typically has no fully characterizable seminorm, or only a loose and approximate one.[34]

As in the sequential case of Section 3, we solve this problem without enforcing the transversality condition (i.e., we rely on the inductive bias in Equation (48) to fulfill Equation (40) with enough samples). We show this works well in practice in Supplemental Appendix Section 5.3 for both kernel and deep learning solutions, and outline the open question of proving conditions for it in Section 6.

*Sample average approximation for adaptive distributions.* Next consider Equation (46), where the states are generated by the dynamics themselves. Since their distribution depends on the unknowns, we express it through fixed exogenous randomness and replace the expectation by what is called a *sample average approximation* (SAA) in stochastic optimization (Kleywegt et al., 2002), used for optimal control by Phelps et al. (2016) and for mean field games by Ruthotto et al. (2020).

The constraints in Equation (46) are the SDE Equations (36) and (39) with the approximants $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{y}}$ in its drift and volatility. As before, Equation (40) is not imposed. Evaluating the objective means solving that SDE forward from $\boldsymbol{x}_0$ and evaluating $\boldsymbol{R}$ along the resulting trajectory $\hat{\boldsymbol{x}}(t)$. Since those states depend on $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{y}}$, minimizing requires the derivative of $\hat{\boldsymbol{x}}(t)$ with respect to their parameters, whether kernel coefficients or NN weights.[35] The algorithm then depends on the discretization: our deterministic implementations use a fixed-step fourth-order Runge–Kutta method, which is accurate with few time steps, and our stochastic ones Euler–Maruyama (Supplemental Appendix Section 5.3).

For the SAA, draw a finite collection of initial conditions and noise realizations and hold them fixed while optimizing over $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{y}})$. Let $t_i = (i-1)\Delta t$ for $i = 1, \ldots, N$ with $\Delta t := T/(N-1)$, so that $t_1 = 0$ and $t_N = T$. For trajectories $s = 1, \ldots, S$, draw an initial

[34] Nevertheless, Equation (48) provides a standard interpretation to understand the inductive bias for these types of interpolation problems (see Belkin, 2021, Ebrahimi Kahou et al., 2024).

[35] Differentiating ordinary differential equation (ODE) and SDE solutions with respect to parameters in their coefficients is well developed (Chen et al., 2018, Li et al., 2020). Childers et al. (2022) differentiate through the solution of discrete-time state-space models for estimation, and Ebrahimi Kahou et al. (2024), Han et al. (2026), and Yang et al. (2025) do the same for simulated stochastic difference equations when training NN policies in discrete time.

condition $\boldsymbol{x}_0^{(s)} \sim \rho_0$ and represent the Brownian path through its increments

$$\Delta \boldsymbol{W}_i^{(s)} := \sqrt{\Delta t}\, \boldsymbol{Z}_i^{(s)}, \qquad \boldsymbol{Z}_i^{(s)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M), \tag{49}$$

for $i = 1, \ldots, N-1$, independently across $s$ and $i$ and of the initial conditions. The Euler–Maruyama discretization of Equations (36) and (39) starts from $\hat{\boldsymbol{x}}_1^{(s)} = \boldsymbol{x}_0^{(s)}$, since $t_1 = 0$, and sets

$$\hat{\boldsymbol{x}}_{i+1}^{(s)} = \hat{\boldsymbol{x}}_i^{(s)} + \Delta t\, \boldsymbol{F}\big(\hat{\boldsymbol{x}}_i^{(s)}, \hat{\boldsymbol{\mu}}(\hat{\boldsymbol{x}}_i^{(s)}), \hat{\boldsymbol{y}}(\hat{\boldsymbol{x}}_i^{(s)})\big) + \boldsymbol{B}\big(\hat{\boldsymbol{x}}_i^{(s)}, \hat{\boldsymbol{\mu}}(\hat{\boldsymbol{x}}_i^{(s)}), \hat{\boldsymbol{y}}(\hat{\boldsymbol{x}}_i^{(s)})\big)\, \Delta \boldsymbol{W}_i^{(s)}, \tag{50}$$

for $i = 1, \ldots, N-1$ and $s = 1, \ldots, S$. Using the resulting $SN$ states, the SAA problem is

$$\begin{aligned} \min_{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{y}}} \quad & \frac{1}{SN} \sum_{s=1}^{S} \sum_{i=1}^{N} \big\| \boldsymbol{R}\big(\hat{\boldsymbol{x}}_i^{(s)}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{y}}\big) \big\|_2^2 \\ \text{s.t.} \quad & \text{Equation (50)} \quad \text{for } i = 1, \ldots, N-1 \text{ and } s = 1, \ldots, S. \end{aligned} \tag{51}$$

Given the draws, every $\hat{\boldsymbol{x}}_i^{(s)}$ is a differentiable function of the parameters of $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{y}})$ through Equation (50), and the gradient of the objective passes through it (Murphy, 2023, section 6.3). When $\boldsymbol{B} = \mathbf{0}$ the increments drop out and any deterministic integrator can replace Equation (50). Our implementations renew the draws every epoch, which avoids overfitting to one realization of the noise without requiring a large $S$ in any single optimization step. Supplemental Appendix Section 5.3 describes the details.

*Repeated risk minimization.* An advantage of SAA is that, for fixed $T$ and under standard conditions, minimizers of Equation (51) converge to minimizers of Equation (46) as $S$ grows (Kleywegt et al., 2002). An alternative is repeated risk minimization (RRM), which alternates between solving Equation (47) on states generated by the current policy and regenerating those states under the updated policy, so that no derivative passes through the dynamics. RRM converges in some cases, but in general it may cycle or stop at a fixed point that does not minimize the policy-dependent risk Equation (46) (Perdomo et al., 2020). When a large $T$ is used to approximate an invariant distribution, a separate and compounding concern is the sensitivity of that distribution to policy-induced or numerical perturbations of the transition law (Mitrophanov, 2005).

### 5.3 *Examples*

The examples in Supplemental Appendix Section 5.3 contrast the two approaches with the same residual Equation (44). Our kernel implementations use Equation (48), with the states of $\mathcal{D}$ taken from the benchmark path in the deterministic case and from a previously solved NN policy in the stochastic ones. Our NN implementations use Equation (51).

*The stochastic growth model.* The stochastic growth model of Supplemental Appendix Section 5.1 adds to the model of Section 4.1 a productivity state $z$ following an Ornstein–Uhlenbeck process, so $\boldsymbol{x} = (\boldsymbol{x}_z, \boldsymbol{x}_e) = (z, k)$, with consumption $y = c$ as the

jump variable, output $e^z f(k)$, and constant relative risk aversion $\gamma$.[36] The state follows

$$\mathrm{d}z(t) = -\kappa\, z(t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_1(t), \tag{52}$$

$$\mathrm{d}k(t) = \big[e^{z(t)} f\big(k(t)\big) - \delta k(t) - c\big(z(t), k(t)\big)\big]\mathrm{d}t, \tag{53}$$

so $\boldsymbol{B} = \mathrm{diag}(\sigma, 0)$ and the generator Equation (41) is

$$\mathcal{L}\varphi(z,k) = -\kappa z\,\partial_z\varphi + \big[e^z f(k) - \delta k - c(z,k)\big]\partial_k\varphi + \tfrac{\sigma^2}{2}\,\partial_{zz}\varphi. \tag{54}$$

With $G = e^z f'(k) - \delta$ and $H = \mu c^\gamma - 1$, the residual Equation (44) is

$$\boldsymbol{R}(z,k;\mu,c) = \begin{bmatrix} r\mu - \mu\big[e^z f'(k) - \delta\big] - \mathcal{L}\mu \\ \mu\, c^\gamma - 1 \end{bmatrix}, \tag{55}$$

and the transversality condition Equation (40) is

$$\lim_{t\to\infty} \mathbb{E}\big[e^{-rt}\, k(t)\, \mu(z(t),k(t)) \,\big|\, z_0, k_0\big] = 0.$$

Setting $\sigma = 0$ and $z_0 = 0$ gives the growth model of Section 4.1 with its unknowns as functions of $k$, which we solve the same way with results in Supplemental Appendix Section 5.3.

*Other examples.* We also solve a stochastic version of the open economy of Section 4.4, with productivity and the world interest rate both stochastic, four states, and two co-states (Supplemental Appendix Section 5.2).

## 6. Future Directions

We see several directions for extending the methods and theory of this paper.

- *Theory for state-space and stochastic formulations.* The minimum-norm selection of Section 5 works in practice (Supplemental Appendix Section 5.3) but is a conjecture. The norm is over states rather than time, so the $\mathcal{W}^{1,2}$ control of the paths' time derivatives that drives Theorem 2 is absent, and Equation (5) becomes a statement in expectation over random trajectories. A theory would also have to settle the existence and measurability of the objects in Equations (36) to (40) and discuss alignment with richer solution concepts such as viscosity solutions (see Achdou et al., 2022), before the inductive bias of a kernel or an NN could be characterized. While the theoretical approach in Section 3 might provide some foundations, we expect this to require a new framework.

- *Tradeoffs in using kernels versus NNs.* For the sequential models, Supplemental Appendix Section 6 compares the two solvers: well-tuned NNs are typically more accurate, while kernel solutions are often faster and require less tuning. In the state-space formulation, Section 5.2 lays out the tradeoffs. A kernel fixes what it means

[36] Supplemental Appendix Section 5.1 derives these from the Hamilton–Jacobi–Bellman equation and describes the reductions used in the implementation.

for two states to be similar and has one coefficient per data point, regardless of the dimension of the state. An NN in effect chooses its own change of variables for the state, at a parameter count that grows with the dimension of the state. Finally, only the kernel has an explicit norm, while the norm an NN implicitly minimizes is unknown in general. Which approach is preferable as the dimension of the state grows remains an open question.

- *Kernels as a lens on other methods.* Wide NNs behave as kernel regressions with the neural tangent kernel (Jacot et al., 2018), and ridgeless kernel regression is a standard approach for understanding interpolating estimators in deep learning (Belkin, 2021). In practice, Ebrahimi Kahou et al. (2024) document that deep learning selects the non-explosive path without a terminal condition; Theorems 1 to 3 establish that mechanism for minimum-norm interpolation in an RKHS. Since the kernel description holds only in the "lazy" regime, where the parameters stay near their initialization and the NN is effectively linear in them (Chizat et al., 2019), characterizing the inductive bias of the NNs used in practice, and whether it aligns with the Sobolev seminorms of Section 2.3, is the natural next step.

- *Adaptive state-space methods for kernels.* Only our NN implementations use the adaptive problem Equation (51); the kernel implementations fix their states in advance. Adaptation is harder for kernels because the approximation is determined by both the weights and the reference states, so moving the states moves the function class itself. There is a long tradition in Bayesian optimization of choosing the points of a kernel method adaptively (e.g., Maus et al., 2024), and adapting those tools to select the states at which the residual is enforced is open.

- *Non-parametric structural estimation.* A kernel solution comes with an explicit RKHS norm and the guarantees of Section 3, an advantage over NNs, which provide no comparable guarantees or bounds on the function space. That makes kernels a candidate for structural estimation of macroeconomic models whose solutions are approximated non-parametrically. For example, reading the kernel solution as a Gaussian process (Rasmussen and Williams, 2006) could provide uncertainty quantification along with the point estimate, and an estimator could build on the consistency result of Theorem 3. The solvers are also differentiable through the solution itself, which Supplemental Appendix Section 7 uses to compute the sensitivity of the solved paths to every structural parameter and initial condition at once and to check them against closed forms.

- *Constraints and variational inequalities.* Inequality and complementarity constraints would bring in lifecycle consumption with borrowing limits, irreversible investment, occasionally binding financial frictions, and constrained differential games, whose equilibrium conditions are differential variational inequalities rather than DAEs. The kernel method should extend once the complementarity conditions are smoothed or reformulated, since Equation (13) is already a constrained optimization problem.

## 7. Conclusion

This paper introduces ridgeless kernel methods for solving deterministic, infinite-horizon, continuous-time dynamic models formulated as DAEs. For a broad class of models, we show that selecting the minimum-norm solution guarantees transversality, addressing the main computational obstacle of traditional approaches. Kernel methods are natural for this task, since their function norms are explicit, allowing us to establish sufficiency results linking kernel solutions to the DAE solution concept.

The theory covers deterministic, sequential models with interior solutions. Section 5 applies the same minimum-norm construction to stochastic models in state space, with kernels and with NNs, where it selects the transversal solution in practice but without the guarantees of Section 3. Relative to other methods there are tradeoffs in speed, accuracy, and generalization. A well-tuned kernel method is often faster and easier to tune than a well-tuned NN, but less accurate (Supplemental Appendix Section 6). What it provides is a precise and interpretable notion of generalization through its RKHS norm, which underlies the theory here and opens the door to econometric analysis.

## Appendix A: Ridge Regression Formulation

For $\lambda > 0$, the ridge version of Equation (13) replaces the interpolation constraints by a penalty on the squared residuals, keeping the unknowns, the integration in Equation (11), and the constraint $(\star)$:

$$\begin{aligned}\min_{\hat{\boldsymbol{x}},\hat{\boldsymbol{\mu}},\hat{\boldsymbol{y}},\hat{\boldsymbol{\mu}}_0,\hat{\boldsymbol{y}}_0} \sum_{t_i\in\mathcal{D}} \Big[ &\Big\|\dot{\hat{\boldsymbol{x}}}(t_i) - \boldsymbol{F}(\hat{\boldsymbol{x}}(t_i),\hat{\boldsymbol{\mu}}(t_i),\hat{\boldsymbol{y}}(t_i))\Big\|_2^2 \\ &+ \Big\|\dot{\hat{\boldsymbol{\mu}}}(t_i) - r\,\hat{\boldsymbol{\mu}}(t_i) + \hat{\boldsymbol{\mu}}(t_i)\odot \boldsymbol{G}(\hat{\boldsymbol{x}}(t_i),\hat{\boldsymbol{\mu}}(t_i),\hat{\boldsymbol{y}}(t_i))\Big\|_2^2 \\ &+ \|\boldsymbol{H}(\hat{\boldsymbol{x}}(t_i),\hat{\boldsymbol{\mu}}(t_i),\hat{\boldsymbol{y}}(t_i))\|_2^2\Big] \\ &+ \lambda\Big(\sum_{m=1}^{M}\|\hat{\boldsymbol{x}}^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^{M}\|\hat{\boldsymbol{\mu}}^{(m)}\|_{\mathcal{H}}^2 + \sum_{p=1}^{P}\|\hat{\boldsymbol{y}}^{(p)}\|_{\mathcal{H}}^2\Big) \\ \text{s.t.}\quad &(\hat{\boldsymbol{x}}(t),\hat{\boldsymbol{\mu}}(t),\hat{\boldsymbol{y}}(t))\in\mathbb{B}\quad \text{for all } t\geq 0.\end{aligned} \tag{56}$$

Write $R_N$ for the sum of squared residuals and $\mathcal{L}$ for the sum of squared RKHS norms, so that the objective is $R_N + \lambda\mathcal{L}$, and let $B_N^2$ be the optimal value of Equation (13). The formulation in Equation (13) is referred to as *ridgeless* because it is the $\lambda \to 0$ limit of Equation (56) in the following sense.

Proposition 1. *Under Assumptions 1 and 2, Equation (56) has a solution for every $\lambda > 0$ and every $\mathcal{D}$. Every solution satisfies $\mathcal{L} \leq B_N^2$ and $R_N \leq \lambda B_N^2$. If $\lambda_k \to 0$ and $h_k$ solves Equation (56) with $\lambda = \lambda_k$, then a subsequence of $h_k$ converges, weakly in $\mathcal{H}$ for the derivatives and uniformly on compact sets for the paths, to a solution of Equation (13).*

Proof. Let $h_N^*$ solve Equation (13), which exists by Lemma 7. It is feasible for Equation (56) with $R_N(h_N^*) = 0$, so the infimum of Equation (56) is at most $\lambda B_N^2$, and any

minimizing sequence $h_n$ eventually satisfies $\lambda\mathcal{L}(h_n) \le R_N(h_n) + \lambda\mathcal{L}(h_n) \le \lambda B_N^2 + 1$. Its derivatives therefore lie in the class $\mathbb{F}_{M'}^C$ of Lemma 9, with $C^2 = B_N^2 + 1/\lambda$, and its paths lie in $\mathbb{B}$, so Lemma 12 gives a subsequence converging to some $h_\infty$ whose path lies in $\mathbb{B}$, with $\mathcal{L}(h_\infty) \le \liminf \mathcal{L}(h_n)$ by weak lower semicontinuity. The derivatives and paths converge pointwise, and $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{H}$ are continuous on $\mathbb{B}$, so $R_N(h_n) \to R_N(h_\infty)$. Hence $h_\infty$ attains the infimum.

For any solution $h_\lambda$, comparison with $h_N^*$ gives $R_N(h_\lambda) + \lambda\mathcal{L}(h_\lambda) \le \lambda B_N^2$, and both terms are nonnegative.

Finally, let $h_k$ solve Equation (56) with $\lambda_k \to 0$. Then $\mathcal{L}(h_k) \le B_N^2$ and the paths lie in $\mathbb{B}$, so Lemma 12 with $C = B_N$ gives a subsequence converging to some $h_0$ whose path lies in $\mathbb{B}$, with $\mathcal{L}(h_0) \le B_N^2$. As above $R_N(h_k) \to R_N(h_0)$ along the subsequence, and $R_N(h_k) \le \lambda_k B_N^2 \to 0$, so $R_N(h_0) = 0$: $h_0$ satisfies the interpolation constraints and $(\star)$. It is feasible for Equation (13) with objective at most $B_N^2$, hence a solution. □

In practice the limit need not be taken: a small fixed $\lambda$ approximates the ridgeless solution in the sense of Proposition 1, and the optimizer can instead anneal $\lambda$ toward zero during the iterations. All of our kernel and NN implementations use the `nlls-gram` Levenberg–Marquardt solver (Perla, 2026), which provides this continuation and, for the kernel method, carries the RKHS norm as its metric.

## Appendix B: Connection Between Sobolev Spaces and the Matérn RKHS

Definition 1 (Matern Kernel). Let $k_{\nu,\ell}(\cdot,\cdot) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ denote the Matérn covariance function with smoothness $\nu$ and lengthscale $\ell$. For the purposes of this paper, we will define $k_{\nu,\ell}$ as:

$$k_{\nu,\ell}(t,t') = \varphi(|t-t'|), \qquad \widehat{\varphi}(\omega) := \left(1+\ell^2\omega^2\right)^{-\nu-1/2}, \tag{57}$$

where $\widehat{(\cdot)}$ corresponds to the Fourier transform.

Note that Equation (57) corresponds to the standard Matérn kernel definition (Rasmussen and Williams, 2006) after appropriately scaling the inputs and outputs. For $\nu = 1/2$ the kernel is exponential in the time domain, $\varphi(s) = a_\ell e^{-s/\ell}$ for $s \ge 0$, where the amplitude $a_\ell > 0$ depends only on $\ell$ and on the normalization of the Fourier transform (for $\widehat{\varphi}(\omega) = \int_{\mathbb{R}} \varphi(|s|)e^{-i\omega s}ds$, $a_\ell = 1/(2\ell)$). We will use $\mathcal{H}^{\nu,\ell}$ to denote the RKHS of $[0,\infty) \to \mathbb{R}$ functions whose reproducing kernel is the restriction of $k_{\nu,\ell}$ to $[0,\infty)\times[0,\infty)$, and $\mathcal{W}^{1,2}$ for the Sobolev space of $[0,\infty) \to \mathbb{R}$ functions that are square integrable with a square-integrable weak derivative, with norm $\|w\|_{\mathcal{W}^{1,2}}^2 := \|w\|_{L^2}^2 + \|\dot{w}\|_{L^2}^2$ (Brezis and Brézis, 2011, Ch. 8). Whenever possible we will drop the superscripts for $\mathcal{H}^{\nu,\ell}$.

Theorem 4 (Equivalence of $\mathcal{W}^{1,2}$ and $\mathcal{H}^{1/2,\ell}$). *For any $\ell > 0$, $\mathcal{H}^{1/2,\ell} = \mathcal{W}^{1,2}$ as sets, identifying each Sobolev function with its continuous representative, and for every $w \in \mathcal{W}^{1,2}$*

$$\|w\|_{\mathcal{H}^{1/2,\ell}}^2 = \frac{1}{2a_\ell\ell}\left(\|w\|_{L^2}^2 + \ell^2\|\dot{w}\|_{L^2}^2 + \ell\, w(0)^2\right). \tag{58}$$

*Consequently*

$$C_1(\ell)\|w\|_{\mathcal{W}^{1,2}} \le \|w\|_{\mathcal{H}^{1/2,\ell}} \le C_2(\ell)\|w\|_{\mathcal{W}^{1,2}},$$
$$C_1(\ell)^2 = \frac{\min(1,\ell^2)}{2a_\ell\ell}, \qquad C_2(\ell)^2 = \frac{\max(1,\ell^2)+\ell}{2a_\ell\ell},$$

*so the two norms are equivalent.*

PROOF. *The whole line.* Let $\mathcal{H}(\mathbb{R})$ be the RKHS of $k_{1/2,\ell}$ on $\mathbb{R}$, and equip $\mathcal{W}^{1,2}(\mathbb{R})$ with the inner product $\langle u,v\rangle := \frac{1}{2a_\ell\ell}\int_{\mathbb{R}} uv + \ell^2\dot{u}\dot{v}\,dt$, which is equivalent to the standard one, so the space is complete. For fixed $s\in\mathbb{R}$, $k_s := k_{1/2,\ell}(s,\cdot) = a_\ell e^{-|\cdot-s|/\ell}$ lies in $\mathcal{W}^{1,2}(\mathbb{R})$ with weak derivative $\dot{k}_s(t) = -(a_\ell/\ell)\operatorname{sgn}(t-s)\,e^{-|t-s|/\ell}$. Let $v\in\mathcal{W}^{1,2}(\mathbb{R})$. Integrating by parts on $(-\infty,s)$ and on $(s,\infty)$, where $\ddot{k}_s = k_s/\ell^2$, and using that $v$ and $k_s$ are continuous and vanish at $\pm\infty$,

$$\ell^2\int_{\mathbb{R}} \dot{k}_s\dot{v}\,dt \;=\; \ell^2\big(\dot{k}_s(s^-)-\dot{k}_s(s^+)\big)v(s) - \int_{\mathbb{R}} k_s v\,dt \;=\; 2a_\ell\ell\,v(s) - \int_{\mathbb{R}} k_s v\,dt,$$

so $\langle k_s, v\rangle = v(s)$. Thus $k_{1/2,\ell}$ is reproducing for this inner product, and since a reproducing kernel determines its Hilbert space uniquely (Aronszajn, 1950), $\mathcal{H}(\mathbb{R}) = \mathcal{W}^{1,2}(\mathbb{R})$ with $\|u\|^2_{\mathcal{H}(\mathbb{R})} = \frac{1}{2a_\ell\ell}\int_{\mathbb{R}} u^2 + \ell^2\dot{u}^2\,dt$.

*The half line.* Let $\mathcal{N}$ be the subspace of $u\in\mathcal{H}(\mathbb{R})$ vanishing on $[0,\infty)$, which is closed because point evaluations are continuous. Restriction to $[0,\infty)$ is injective on $\mathcal{N}^\perp$, and for $w$ in its image write $u_w\in\mathcal{N}^\perp$ for the unique preimage, which is the extension of $w$ of minimal $\mathcal{H}(\mathbb{R})$-norm. The image with $\|w\|_+ := \|u_w\|_{\mathcal{H}(\mathbb{R})}$ is a Hilbert space, and for $s\ge 0$ and $n\in\mathcal{N}$, $\langle k_s, n\rangle = n(s) = 0$, so $k_s\in\mathcal{N}^\perp$ and $\langle k_s|_{[0,\infty)}, w\rangle_+ = \langle k_s, u_w\rangle = u_w(s) = w(s)$. Hence this space is the RKHS of the restricted kernel (Aronszajn, 1950): $\mathcal{H}^{1/2,\ell}$ consists of the restrictions of $\mathcal{W}^{1,2}(\mathbb{R})$ functions, which is all of $\mathcal{W}^{1,2}$ (extend by even reflection), with $\|w\|_{\mathcal{H}^{1/2,\ell}} = \min\{\|u\|_{\mathcal{H}(\mathbb{R})} : u|_{[0,\infty)} = w\}$. To compute the minimum, let $u$ be any extension. On $(-\infty,0]$, $(u-\ell\dot{u})^2\ge 0$ gives $u^2+\ell^2\dot{u}^2 \ge 2\ell u\dot{u} = \ell\frac{d}{dt}(u^2)$, with equality if and only if $u = \ell\dot{u}$, so

$$\int_{-\infty}^{0} u^2+\ell^2\dot{u}^2\,dt \;\ge\; \ell\,u(0)^2 \;=\; \ell\,w(0)^2,$$

with equality for $u(t) = w(0)e^{t/\ell}$. Adding the integral over $[0,\infty)$ and dividing by $2a_\ell\ell$ gives Equation (58).

*Equivalence.* The lower bound follows from $\|w\|^2_{L^2} + \ell^2\|\dot{w}\|^2_{L^2} \ge \min(1,\ell^2)\|w\|^2_{\mathcal{W}^{1,2}}$ and $w(0)^2\ge 0$. For the upper bound, since $w(t)\to 0$ as $t\to\infty$ for $w\in\mathcal{W}^{1,2}$ (Brezis and Brézis, 2011, Ch. 8), $w(0)^2 = -\int_0^\infty \frac{d}{dt}w^2\,dt = -2\int_0^\infty w\dot{w}\,dt \le \|w\|^2_{L^2} + \|\dot{w}\|^2_{L^2}$. □

The Matérn kernels with $\nu = P - 1/2$ for integers $P\ge 2$ are related to the Sobolev spaces $\mathcal{W}^{P,2}$ in the same way (Wendland, 2004, Ch. 10), but only $\nu = 1/2$ is used in our proofs.

## APPENDIX C: PROOF OF THEOREM 1

THEOREM 1 (RESTATED). For a DAE satisfying Assumptions 1 and 2, define $\mathbb{S}$ as the set of trajectories in the sense of Assumption 2, which satisfy Equations (1) to (4) but not necessarily Equation (5), and define $\mathbb{M}$ as the set of $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{y})$ with $\boldsymbol{x}, \boldsymbol{\mu}$ continuously differentiable and finite derivative energy $\int_0^\infty \|\dot{\boldsymbol{x}}(t)\|_2^2 + \|\dot{\boldsymbol{\mu}}(t)\|_2^2 \, dt < \infty$. Then $\mathbb{S} \cap \mathbb{M}$ contains exactly one element $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$, and it is the unique trajectory that satisfies Equation (5).

PROOF. *The solution lies in* $\mathbb{S} \cap \mathbb{M}$. By Assumption 2(i), $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$ is a trajectory, so it lies in $\mathbb{S}$; by Assumption 2(ii), the integral defining $\mathbb{M}$ is bounded by the finite integral in that assumption, so it lies in $\mathbb{M}$.

*Every element of* $\mathbb{S} \cap \mathbb{M}$ *satisfies Equation* (5). Let $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{y}) \in \mathbb{S} \cap \mathbb{M}$ and set $A := \big(\int_0^\infty \|\dot{\boldsymbol{x}}(t)\|_2^2 + \|\dot{\boldsymbol{\mu}}(t)\|_2^2 \, \mathrm{d}t\big)^{1/2} < \infty$. Each component $w$ of $\boldsymbol{x}$ or $\boldsymbol{\mu}$ is continuously differentiable by the definition of a trajectory, and $\dot{w}^2 \le \|\dot{\boldsymbol{x}}\|_2^2 + \|\dot{\boldsymbol{\mu}}\|_2^2$ pointwise, so $\int_0^t \dot{w}^2 \, \mathrm{d}\tau \le A^2$ for every $t \ge 0$. The fundamental theorem of calculus and the Cauchy–Schwarz inequality against the constant function $1$ then give

$$|w(t)| \;\le\; |w(0)| + \int_0^t |\dot{w}(\tau)| \, \mathrm{d}\tau \;\le\; |w(0)| + \sqrt{t}\Big(\int_0^t \dot{w}(\tau)^2 \, \mathrm{d}\tau\Big)^{1/2} \;\le\; |w(0)| + \sqrt{t}\, A.$$

Hence, with $W_0 := \max_m \max\big(|x^{(m)}(0)|, |\mu^{(m)}(0)|\big)$, both $|x^{(m)}(t)|$ and $|\mu^{(m)}(t)|$ are at most $W_0 + \sqrt{t}\, A$ for each $m$, and multiplying,

$$|x^{(m)}(t)\, \mu^{(m)}(t)| \;\le\; \big(W_0 + \sqrt{t}\, A\big)^2 \;\le\; 2W_0^2 + 2A^2 t \qquad \text{for all } t \ge 0,$$

using $(a+b)^2 \le 2a^2 + 2b^2$. Finite derivative energy therefore does not prevent $x^{(m)}\mu^{(m)}$ from growing, but it caps the growth at a linear rate in $t$. The transversality condition Equation (5) weights this product by $e^{-rt}$, so

$$\big|e^{-rt} x^{(m)}(t)\, \mu^{(m)}(t)\big| \;\le\; e^{-rt}\big(2W_0^2 + 2A^2 t\big) \;\longrightarrow\; 0 \qquad \text{as } t \to \infty,$$

because $r > 0$ gives $e^{-rt} \to 0$ and $te^{-rt} \to 0$.[37] Since this holds for every $m$, $\lim_{t\to\infty} e^{-rt} \boldsymbol{x}(t) \odot \boldsymbol{\mu}(t) = \mathbf{0}$, which is Equation (5).

By Assumption 2(i) the only trajectory satisfying Equation (5) is $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$, so $\mathbb{S} \cap \mathbb{M} = \{(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)\}$, and this element is the unique trajectory satisfying Equation (5). □

## APPENDIX D: PROOF OF THEOREM 2

THEOREM 2 (RESTATED). Let $\mathcal{H}$ be the Matérn RKHS with $\nu = \frac{1}{2}$ and lengthscale $\ell > 0$ on $[0, \infty)$, let $\dot{x}^{(m)}, \dot{\mu}^{(m)} \in \mathcal{H}$ for every $m = 1, \dots, M$, and let $\boldsymbol{x}(t) = \boldsymbol{x}(0) + \int_0^t \dot{\boldsymbol{x}}(\tau) d\tau$ and $\boldsymbol{\mu}(t) = \boldsymbol{\mu}(0) + \int_0^t \dot{\boldsymbol{\mu}}(\tau) d\tau$ for finite initial values $\boldsymbol{x}(0), \boldsymbol{\mu}(0) \in \mathbb{R}^M$. Then

$$\lim_{t\to\infty} e^{-rt}\, \boldsymbol{x}(t) \odot \boldsymbol{\mu}(t) \;=\; \mathbf{0};$$

that is, $(\boldsymbol{x}, \boldsymbol{\mu})$ satisfies the asymptotic boundary condition Equation (5).

[37] Discounting is essential here. With $r = 0$, finite derivative energy no longer implies Equation (5): $x^{(m)}(t) = \mu^{(m)}(t) = (1+t)^{1/3}$ has a square-integrable derivative, while $x^{(m)}(t)\, \mu^{(m)}(t) = (1+t)^{2/3} \to \infty$.

PROOF. By Theorem 4, $\mathcal{H} = \mathcal{W}^{1,2}$ with $\|w\|_{\mathcal{W}^{1,2}} \le C_1^{-1}\|w\|_{\mathcal{H}}$ for every $w \in \mathcal{H}$, so for each $m = 1, \dots, M$

$$\|\dot{x}^{(m)}\|_{L^2([0,\infty))} \le \|\dot{x}^{(m)}\|_{\mathcal{W}^{1,2}} \le C_1^{-1}\|\dot{x}^{(m)}\|_{\mathcal{H}} < \infty.$$

Since $x^{(m)}(t) = x^{(m)}(0) + \int_0^t \dot{x}^{(m)}(\tau)d\tau$, the Cauchy-Schwarz inequality gives

$$\begin{aligned}\left|x^{(m)}(t)\right| &\le \left|x^{(m)}(0)\right| + \int_0^t \left|\dot{x}^{(m)}(\tau)\right| d\tau \\ &\le \underbrace{\left|x^{(m)}(0)\right|}_{:=a_m} + \sqrt{t}\,\underbrace{\left\|\dot{x}^{(m)}\right\|_{L^2([0,\infty))}}_{:=b_m},\end{aligned}$$

with $0 \le a_m, b_m < \infty$. An identical argument will show that $|\mu^{(m)}(t)| \le c_m + d_m\sqrt{t}$ for finite $c_m, d_m$. Hence, since $r > 0$,

$$e^{-rt}\left|x^{(m)}(t)\mu^{(m)}(t)\right| \le e^{-rt}\left(a_m + b_m\sqrt{t}\right)\left(c_m + d_m\sqrt{t}\right) = \mathcal{O}\big(t\,e^{-rt}\big) \longrightarrow 0.$$

□

## APPENDIX E: PROOF OF THEOREM 3

THEOREM 3 (RESTATED). Suppose Assumptions 1 and 2 hold, and let $(\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{y}}_N)$ be any solution of Equation (13), with $\mathcal{H}$ the Matérn RKHS with $\nu = \frac{1}{2}$ and lengthscale $\ell > 0$ on $[0,\infty)$, for the points $\mathcal{D} = \{t_1, \dots, t_N\}$, drawn i.i.d. from a distribution $\pi$ on $[0,\infty)$ with a strictly positive density and finite second moment. Then

$$\lim_{N\to\infty}(\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{y}}_N) \overset{\text{a.s.}}{=} (\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*),$$

where $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$ is the unique trajectory satisfying Equation (5) (Assumption 2), and the convergence is uniform on compact subsets of $[0,\infty)$.

Throughout this appendix, $\mathcal{H}$ denotes the Matérn-$1/2$ RKHS of functions on $[0,\infty)$ and $\kappa := \sup_{t\in[0,\infty)} \sqrt{k(t,t)} < \infty$ denotes the (finite) uniform bound on its reproducing kernel. We will repeatedly use the fact that $\mathcal{H} = \mathcal{W}^{1,2}$ as sets, with $C_1\|w\|_{\mathcal{W}^{1,2}} \le \|w\|_{\mathcal{H}} \le C_2\|w\|_{\mathcal{W}^{1,2}}$ for constants $0 < C_1 \le C_2 < \infty$ depending only on the lengthscale (Theorem 4).

Two consequences are used often enough to record here: for any $w \in \mathcal{H}$,

$$\sup_{t\in[0,\infty)} |w(t)| \le \kappa\|w\|_{\mathcal{H}}, \quad |w(t) - w(s)| \le \|\dot{w}\|_{L^2}\sqrt{|t-s|} \le C_1^{-1}\|w\|_{\mathcal{H}}\sqrt{|t-s|}, \tag{59}$$

the first by the reproducing property and Cauchy-Schwarz (see e.g. Wainwright, 2019, Ch. 12), the second by Cauchy-Schwarz applied to $w(t) - w(s) = \int_s^t \dot{w}(\tau)d\tau$ and $\|\dot{w}\|_{L^2} \le \|w\|_{\mathcal{W}^{1,2}} \le C_1^{-1}\|w\|_{\mathcal{H}}$. In particular, every $w \in \mathcal{H}$ is continuous, and bounded subsets of $\mathcal{H}$ are uniformly bounded and equicontinuous.

We will also need the representer of the integration functional. For $t \geq 0$, set

$$g_t := \int_0^t k(\tau, \cdot)\, d\tau \in \mathcal{H}, \quad \text{so that} \quad \langle g_t, w \rangle_{\mathcal{H}} = \int_0^t w(\tau)\, d\tau \quad \text{for all } w \in \mathcal{H}, \tag{60}$$

where the integral is well-defined because $\tau \mapsto k(\tau, \cdot)$ is continuous and bounded in $\mathcal{H}$. The triangle inequality gives $\|g_t\|_{\mathcal{H}} \leq \kappa t$, and thus both $w \mapsto w(t)$ and $w \mapsto \int_0^t w$ are bounded linear functionals on $\mathcal{H}$.

Throughout this proof, $\mathbb{B}$ denotes the box in $(\star)$, the compact subset of $\mathbb{V}$ from Assumption 2(iii), and $K := \max_{(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\in\mathbb{B}} \max(\|\boldsymbol{x}\|_\infty, \|\boldsymbol{\mu}\|_\infty, \|\boldsymbol{y}\|_\infty)$. By Assumption 1, $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{H}$ are Lipschitz on $\mathbb{B}$ with respect to $\|\cdot\|_\infty$. The minimization in Equation (13) is over all $\dot{x}^{(m)}, \dot{\mu}^{(m)}, \dot{y}^{(p)} \in \mathcal{H}$ and $\boldsymbol{\mu}_0 \in \mathbb{R}^M$, $\boldsymbol{y}_0 \in \mathbb{R}^P$, with $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{y})$ recovered by integration as in Equation (11).

We break up this proof into a series of lemmas. The first lemma uses Assumption 2 to characterize the derivative norms of DAE solutions.

LEMMA 5. *$(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$, the true solution to the DAE in Equations (1) to (5), satisfies*

$$B^2 := \textstyle\sum_{m=1}^M \|\dot{x}^{*(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^M \|\dot{\mu}^{*(m)}\|_{\mathcal{H}}^2 + \sum_{p=1}^P \|\dot{y}^{*(p)}\|_{\mathcal{H}}^2 < \infty.$$

PROOF. Define

$$C_0 = \int_0^\infty \|\dot{\boldsymbol{x}}^*(t)\|_2^2 + \|\dot{\boldsymbol{\mu}}^*(t)\|_2^2 + \|\dot{\boldsymbol{y}}^*(t)\|_2^2 + \|\ddot{\boldsymbol{x}}^*(t)\|_2^2 + \|\ddot{\boldsymbol{\mu}}^*(t)\|_2^2 + \|\ddot{\boldsymbol{y}}^*(t)\|_2^2 \, \mathrm{d}t,$$

which is finite by Assumption 2(ii). Since $\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*$ are twice continuously differentiable, each component of $\dot{\boldsymbol{x}}^*, \dot{\boldsymbol{\mu}}^*, \dot{\boldsymbol{y}}^*$ lies in $\mathcal{W}^{1,2}$ and, by definition of the Sobolev norm,

$$C_0 = \textstyle\sum_{m=1}^M \|\dot{x}^{*(m)}\|_{\mathcal{W}^{1,2}}^2 + \sum_{m=1}^M \|\dot{\mu}^{*(m)}\|_{\mathcal{W}^{1,2}}^2 + \sum_{p=1}^P \|\dot{y}^{*(p)}\|_{\mathcal{W}^{1,2}}^2.$$

By Theorem 4, $B^2 \leq C_2^2 C_0 < \infty$. □

The next two lemmas address the validity of the optimization problem in Equation (13): a solution to this empirical risk minimization problem always exists, and the true solution is a constraint-satisfying potential candidate.

LEMMA 6. *For all sets $\mathcal{D} = \{t_1, \ldots, t_N\}$, the true DAE solution $\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*$ satisfies the constraints in Equation (13).*

PROOF. The DAE solution $\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*$ satisfies Equations (1) to (3) everywhere on $[0, \infty)$, and thus it satisfies the first three constraints in Equation (13) at every $t_i \in \mathcal{D}$. The constraint $(\star)$ is Assumption 2(iii). □

LEMMA 7. *For all sets $\mathcal{D} = \{t_1, \ldots, t_N\}$, a solution to Equation (13) exists.*

PROOF. Write the objective of Equation (13) as

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{y}) = \textstyle\sum_{m=1}^M \left( \|\dot{x}^{(m)}\|_{\mathcal{H}}^2 + \|\dot{\mu}^{(m)}\|_{\mathcal{H}}^2 \right) + \sum_{p=1}^P \|\dot{y}^{(p)}\|_{\mathcal{H}}^2$$

A candidate solution is defined by the derivatives $\{\dot{x}^{(m)}, \dot{\mu}^{(m)}\}_{m=1}^{M}, \{\dot{y}^{(p)}\}_{p=1}^{P}$ as well as initial conditions $\boldsymbol{\mu}_0, \boldsymbol{y}_0$, from which $\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{y}$ are recovered through the integration in Equation (11). We begin by noting that the feasible set is nonempty and the objective attains a finite value on it, since $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$ is feasible (Lemma 6) with objective value $\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*) = B^2 < \infty$ (Lemma 5).

Let $(\boldsymbol{x}_n, \boldsymbol{\mu}_n, \boldsymbol{y}_n)$ be a minimizing sequence in the feasible set with $\mathcal{L}(\boldsymbol{x}_n, \boldsymbol{\mu}_n, \boldsymbol{y}_n) \le B^2$ for all $n$. The constraint $(\star)$ at $t = 0$ forces $(\boldsymbol{\mu}_{0,n}, \boldsymbol{y}_{0,n}) = (\boldsymbol{\mu}_n(0), \boldsymbol{y}_n(0)) \in [\underline{\boldsymbol{\mu}}, \bar{\boldsymbol{\mu}}] \times [\underline{\boldsymbol{y}}, \bar{\boldsymbol{y}}]$, a compact subset of $\mathbb{R}^{M+P}$. Moreover, closed balls in a Hilbert space are weakly sequentially compact. Thus, after passing to a subsequence, which we do not relabel,

$$(\dot{x}_n^{(m)}, \dot{\mu}_n^{(m)}, \dot{y}_n^{(p)}) \overset{\text{weakly}}{\longrightarrow} (\dot{x}_\infty^{(m)}, \dot{\mu}_\infty^{(m)}, \dot{y}_\infty^{(p)}) \in \mathcal{H}, \qquad (\boldsymbol{\mu}_{0,n}, \boldsymbol{y}_{0,n}) \to (\boldsymbol{\mu}_{0,\infty}, \boldsymbol{y}_{0,\infty}).$$

We next claim the limit also lives in the feasible set. Weak convergence means $\langle h, \dot{x}_n^{(m)} \rangle \to \langle h, \dot{x}_\infty^{(m)} \rangle$ for every $h \in \mathcal{H}$ (and likewise for $\dot{\mu}^{(m)}, \dot{y}^{(p)}$). Taking $h$ to be a coordinate copy of $k(t, \cdot)$ gives $\dot{\boldsymbol{x}}_n(t) \to \dot{\boldsymbol{x}}_\infty(t)$, $\dot{\boldsymbol{\mu}}_n(t) \to \dot{\boldsymbol{\mu}}_\infty(t)$, and $\dot{\boldsymbol{y}}_n(t) \to \dot{\boldsymbol{y}}_\infty(t)$ for every $t \in [0, \infty)$. Moreover, taking $h$ to be a coordinate copy of $g_t$ and using Equation (60) gives $\int_0^t \dot{\boldsymbol{x}}_n \to \int_0^t \dot{\boldsymbol{x}}_\infty$, and hence $\boldsymbol{x}_n(t) \to \boldsymbol{x}_\infty(t)$, $\boldsymbol{\mu}_n(t) \to \boldsymbol{\mu}_\infty(t)$, $\boldsymbol{y}_n(t) \to \boldsymbol{y}_\infty(t)$, again for every $t \in [0, \infty)$. Each of the first three constraints of Equation (13) is a continuous function of the finitely many values $\dot{\boldsymbol{x}}(t_i), \dot{\boldsymbol{\mu}}(t_i), \boldsymbol{x}(t_i), \boldsymbol{\mu}(t_i), \boldsymbol{y}(t_i)$ (by continuity of $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{H}$), so each of the constraints passes to the limit. Regarding the constraint $(\star)$: for each fixed $t$, $(\boldsymbol{x}_n(t), \boldsymbol{\mu}_n(t), \boldsymbol{y}_n(t)) \in \mathbb{B}$ converges to $(\boldsymbol{x}_\infty(t), \boldsymbol{\mu}_\infty(t), \boldsymbol{y}_\infty(t))$, which therefore lies in $\mathbb{B}$ because $\mathbb{B}$ is closed. Altogether, we have that the limit lives in the feasible set.

Finally, the norms of Hilbert spaces are weakly lower semicontinuous. Because the objective function $\mathcal{L}$ is the sum of squared $\mathcal{H}$ norms, we have that

$$\mathcal{L}(\boldsymbol{x}_\infty, \boldsymbol{\mu}_\infty, \boldsymbol{y}_\infty) \le \underbrace{\liminf_n \mathcal{L}(\boldsymbol{x}_n, \boldsymbol{\mu}_n, \boldsymbol{y}_n)}_{=:I} \le B^2.$$

The limit is feasible with objective value at most $I$, and $I$ is trivially the infimum because the sequence is minimizing, so the infimum is attained. □

Because the proof of Theorem 3 establishes consistency via a uniform law of large numbers, the next lemmas establish the Rademacher complexity over the kernel-estimated DAE solutions, as well as the Rademacher complexity on an *error function* that measures how much the kernel-estimated DAE solution violates Equations (1) to (3).

LEMMA 8. *Let* $0 \le C < \infty$. *Define the function classes*

$$\mathbb{F}^C := \{w \in \mathcal{H} \,:\, \|w\|_{\mathcal{H}} \le C\}$$
$$\textstyle\int \mathbb{F}^C := \left\{t \mapsto \int_0^t w(\tau) d\tau \,:\, w \in \mathbb{F}^C\right\}.$$

*Letting* $\pi$ *be a distribution with support over* $[0, \infty)$ *and finite second moment, and denoting* $\widehat{\mathcal{R}}_N$ *as the empirical Rademacher complexity for a dataset* $t_1, \ldots, t_N \overset{\text{i.i.d.}}{\sim} \pi$, *we have that*

$$\widehat{\mathcal{R}}_N(\mathbb{F}^C) \le C\kappa N^{-1/2}, \qquad \mathbb{E}_{t_i}\left[\widehat{\mathcal{R}}_N(\textstyle\int \mathbb{F}^C)\right] \le C\kappa\sqrt{\mathbb{E}_{t\sim\pi}[t^2]}N^{-1/2}.$$

PROOF. The bound $\widehat{\mathcal{R}}_N(\mathbb{F}^C) \le C\kappa N^{-1/2}$ is a standard result for reproducing kernel Hilbert spaces with a bounded kernel; it follows from the computation below with $k(t_i,\cdot)$ in place of $g_{t_i}$, using $\|k(t_i,\cdot)\|_{\mathcal{H}} = \sqrt{k(t_i,t_i)} \le \kappa$. For $\int \mathbb{F}^C$ we use the integral representer $g_t$ of Equation (60):

$$
\begin{aligned}
\widehat{\mathcal{R}}_N(\textstyle\int \mathbb{F}^C) &:= \mathbb{E}_{\epsilon_i}\left[\sup_{w\in\mathbb{F}^C} \frac{1}{N}\sum_{i=1}^N \epsilon_i \int_0^{t_i} w(\tau)d\tau\right] && (\epsilon_i \overset{\text{i.i.d.}}{\sim} \text{Rad}) \\
&= \mathbb{E}_{\epsilon_i}\left[\sup_{w\in\mathbb{F}^C} \left\langle \frac{1}{N}\sum_{i=1}^N \epsilon_i g_{t_i},\, w\right\rangle_{\mathcal{H}}\right] \\
&\le \mathbb{E}_{\epsilon_i}\left[\left\|\frac{C}{N}\sum_{i=1}^N \epsilon_i g_{t_i}\right\|_{\mathcal{H}}\right] && \text{(Cauchy-Schwarz inequality)} \\
&\le \sqrt{\mathbb{E}_{\epsilon_i}\left[\left\|\frac{C}{N}\sum_{i=1}^N \epsilon_i g_{t_i}\right\|_{\mathcal{H}}^2\right]} && \text{(Jensen inequality)} \\
&= \sqrt{\frac{C^2}{N^2}\sum_{i,j}\mathbb{E}_{\epsilon_i}\left[\epsilon_i\epsilon_j\right]\left\langle g_{t_i},\, g_{t_j}\right\rangle_{\mathcal{H}}} = \sqrt{\frac{C^2}{N^2}\sum_{i=1}^N \|g_{t_i}\|_{\mathcal{H}}^2} && (\epsilon_i \text{ are uncorrelated}) \\
&\le \sqrt{\frac{C^2}{N^2}\sum_{i=1}^N \left(\int_0^{t_i} \|k(\tau,\cdot)\|_{\mathcal{H}}\, d\tau\right)^2} && \text{(triangle inequality)} \\
&\le \sqrt{\frac{C^2\kappa^2}{N^2}\sum_{i=1}^N t_i^2} = \frac{C\kappa}{N^{1/2}}\sqrt{\frac{1}{N}\sum_{i=1}^N t_i^2},
\end{aligned}
$$

where the last inequality uses $\|k(\tau,\cdot)\|_{\mathcal{H}} = \sqrt{k(\tau,\tau)} \le \kappa$. Taking expectations over $t_i \overset{\text{i.i.d.}}{\sim} \pi$ and applying Jensen's inequality gives $\mathbb{E}_{t_i}[\widehat{\mathcal{R}}_N(\int \mathbb{F}^C)] \le C\kappa\sqrt{\mathbb{E}_{t\sim\pi}[t^2]}\, N^{-1/2}$, which is finite by the second-moment hypothesis. □

LEMMA 9. *Let $\pi$ and $C$ be as in Lemma 8, and write $M' := 2M + P$ for the total dimension of the DAE. Define*

$$
\mathbb{F}^C_{M'} := \left\{(\dot{\boldsymbol{x}}, \dot{\boldsymbol{\mu}}, \dot{\boldsymbol{y}}) \,:\, \textstyle\sum_{m=1}^M \left(\|\dot{x}^{(m)}\|_{\mathcal{H}}^2 + \|\dot{\mu}^{(m)}\|_{\mathcal{H}}^2\right) + \sum_{p=1}^P \|\dot{y}^{(p)}\|_{\mathcal{H}}^2 \le C^2\right\}
$$

$$
\textstyle\int \mathbb{F}^C_{M'} := \left\{t \mapsto \int_0^t (\dot{\boldsymbol{x}}(\tau), \dot{\boldsymbol{\mu}}(\tau), \dot{\boldsymbol{y}}(\tau))\, d\tau \,:\, (\dot{\boldsymbol{x}}, \dot{\boldsymbol{\mu}}, \dot{\boldsymbol{y}}) \in \mathbb{F}^C_{M'}\right\}
$$

*Then $\widehat{\mathcal{R}}_N(\mathbb{F}^C_{M'}) \le C\kappa M' N^{-1/2}$ and $\mathbb{E}_{t_i}\left[\widehat{\mathcal{R}}_N(\int \mathbb{F}^C_{M'})\right] \le C\kappa M'\sqrt{\mathbb{E}_{t\sim\pi}[t^2]}\, N^{-1/2}$.*

PROOF. Every component of an element of $\mathbb{F}^C_{M'}$ has $\mathcal{H}$-norm at most $C$, i.e. lies in $\mathbb{F}^C$ as defined in Lemma 8. The claim then follows from the standard summation rule for Rademacher complexity:

$$
\widehat{\mathcal{R}}_N(\mathbb{F}^C_{M'})
$$

$$:=\mathbb{E}\left[\sup_{(\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}})\in\mathbb{F}^C_{M'}}\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{M}\epsilon^x_{ij}\dot{x}^{(j)}(t_i)+\sum_{j=1}^{M}\epsilon^\mu_{ij}\dot{\mu}^{(j)}(t_i)+\sum_{j=1}^{P}\epsilon^y_{ij}\dot{y}^{(j)}(t_i)\right)\right]$$

$$(\epsilon^x_{ij},\epsilon^\mu_{ij},\epsilon^y_{ij}\overset{\text{i.i.d.}}{\sim}\text{Rad})$$

$$\leq\underbrace{\sum_{j=1}^{M}\mathbb{E}\left[\sup_{\dot{x}^{(j)}\in\mathbb{F}^C}\frac{1}{N}\sum_{i=1}^{N}\epsilon^x_{ij}\dot{x}^{(j)}(t_i)\right]}_{\leq MC\kappa N^{-1/2}}+\underbrace{\sum_{j=1}^{M}\mathbb{E}\left[\sup_{\dot{\mu}^{(j)}\in\mathbb{F}^C}\frac{1}{N}\sum_{i=1}^{N}\epsilon^\mu_{ij}\dot{\mu}^{(j)}(t_i)\right]}_{\leq MC\kappa N^{-1/2}}$$

$$+\underbrace{\sum_{j=1}^{P}\mathbb{E}\left[\sup_{\dot{y}^{(j)}\in\mathbb{F}^C}\frac{1}{N}\sum_{i=1}^{N}\epsilon^y_{ij}\dot{y}^{(j)}(t_i)\right]}_{\leq PC\kappa N^{-1/2}}$$

$$\leq C\kappa M' N^{-1/2}, \qquad \text{(Lemma 8)}$$

where the inequality relaxes the joint supremum over $\mathbb{F}^C_{M'}$ to a sum of componentwise suprema over $\mathbb{F}^C$. An identical argument using the second bound of Lemma 8 gives the claim for $\int\mathbb{F}^C_{M'}$. □

LEMMA 10. *For $\boldsymbol{\mu}_0\in\mathbb{R}^M$, $\boldsymbol{y}_0\in\mathbb{R}^P$, and $\{\dot{x}^{(m)},\dot{\mu}^{(m)}\}_{m=1}^M,\{\dot{y}^{(p)}\}_{p=1}^P\in\mathcal{H}$, set*

$$\boldsymbol{x}(t):=\boldsymbol{x}_0+\int_0^t\dot{\boldsymbol{x}}(\tau)d\tau,\quad \boldsymbol{\mu}(t):=\boldsymbol{\mu}_0+\int_0^t\dot{\boldsymbol{\mu}}(\tau)d\tau,\quad \boldsymbol{y}(t):=\boldsymbol{y}_0+\int_0^t\dot{\boldsymbol{y}}(\tau)d\tau,$$

*and, whenever $(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t))\in\mathbb{V}$ for all $t\geq 0$, define the DAE error function*

$$e_{\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}},\boldsymbol{\mu}_0,\boldsymbol{y}_0}(t):=\left\|\begin{bmatrix}\dot{\boldsymbol{x}}(t)-\boldsymbol{F}(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t))\\ \dot{\boldsymbol{\mu}}(t)-(r\boldsymbol{\mu}(t)-\boldsymbol{\mu}(t)\odot\boldsymbol{G}(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t)))\\ \boldsymbol{H}(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t))\end{bmatrix}\right\|_\infty. \tag{61}$$

*For any $0\leq C<\infty$, define $\mathbb{G}^C$ as the set of error functions*

$$\mathbb{G}^C:=\left\{e_{\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}},\boldsymbol{\mu}_0,\boldsymbol{y}_0}(\cdot)\ :\ (\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}})\in\mathbb{F}^C_{M'},\ (\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t))\in\mathbb{B}\text{ for all }t\geq 0\right\}.$$

*Then every error function in $\mathbb{G}^C$ is bounded by some constant $\widetilde{C}$ (depending only on $C$, $\kappa$, $\mathbb{B}$, $r$, $\boldsymbol{F}$, $\boldsymbol{G}$, and $\boldsymbol{H}$) and*

$$\mathbb{E}_{t_i}\left[\widehat{\mathcal{R}}_N(\mathbb{G}^C)\right]\leq\sqrt{2}LM'\left(C\kappa\left(1+\sqrt{\mathbb{E}_{t\sim\pi}[t^2]}\right)+K\right)N^{-1/2},$$

*where $L$ is a finite constant that depends only on $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{H}$, $r$, and $\mathbb{B}$.*

PROOF. By Assumption 1, $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{H}$ are Lipschitz with respect to $\|\cdot\|_\infty$ on the compact set $\mathbb{B}$, and hence bounded there. Therefore on $\mathbb{R}^M\times\mathbb{R}^M\times\mathbb{B}$ the function

$$(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{d},\boldsymbol{e})\longmapsto\left\|\begin{bmatrix}\boldsymbol{a}-\boldsymbol{F}(\boldsymbol{c},\boldsymbol{d},\boldsymbol{e})\\ \boldsymbol{b}-(r\boldsymbol{d}-\boldsymbol{d}\odot\boldsymbol{G}(\boldsymbol{c},\boldsymbol{d},\boldsymbol{e}))\\ \boldsymbol{H}(\boldsymbol{c},\boldsymbol{d},\boldsymbol{e})\end{bmatrix}\right\|_\infty$$

is Lipschitz continuous with respect to $\|\cdot\|_\infty$ in $(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{d},\boldsymbol{e})$: the map $(\boldsymbol{c},\boldsymbol{d},\boldsymbol{e}) \mapsto \boldsymbol{d} \odot \boldsymbol{G}(\boldsymbol{c},\boldsymbol{d},\boldsymbol{e})$ is Lipschitz on $\mathbb{B}$ because $\boldsymbol{d}$ is bounded there and $\boldsymbol{G}$ is bounded and Lipschitz. Let $L$ be its Lipschitz constant with respect to $\|\cdot\|_\infty$, which depends only on $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{H}$, $r$, and $\mathbb{B}$. Being real-valued, the function extends to all of $\mathbb{R}^M \times \mathbb{R}^M \times \mathbb{R}^M \times \mathbb{R}^M \times \mathbb{R}^P$ with the same Lipschitz constant (McShane extension). Since $\|u-v\|_\infty \le \|u-v\|_2$, the extension is also $L$-Lipschitz with respect to the Euclidean norm, which is the hypothesis of the vector-valued contraction inequality used below.

For boundedness: $(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t)) \in \mathbb{B}$ by the design of $\mathbb{G}^C$, and each component of $\dot{\boldsymbol{x}}$ and $\dot{\boldsymbol{\mu}}$ satisfies $\sup_t |\dot{x}^{(m)}(t)| \le \kappa C$ by Equation (59). Therefore $e_{\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}},\boldsymbol{\mu}_0,\boldsymbol{y}_0}(\cdot)$ is a Lipschitz function applied to inputs lying in a fixed bounded set, and is thus bounded by a constant $\widetilde{C}$ depending only on $C$, $\kappa$, $\mathbb{B}$, $r$, $\boldsymbol{F}$, $\boldsymbol{G}$, and $\boldsymbol{H}$.

For the Rademacher complexity, note that each $e \in \mathbb{G}^C$ is the composition of the above $L$-Lipschitz map with the vector of functions $(\dot{\boldsymbol{x}}(\cdot),\dot{\boldsymbol{\mu}}(\cdot),\boldsymbol{x}(\cdot),\boldsymbol{\mu}(\cdot),\boldsymbol{y}(\cdot))$, whose components range over $\mathbb{F}^C_{M'}$ and $\int \mathbb{F}^C_{M'}$ translated by the constants $\boldsymbol{x}_0,\boldsymbol{\mu}_0,\boldsymbol{y}_0$, where $\|\boldsymbol{x}_0\|_\infty, \|\boldsymbol{\mu}_0\|_\infty, \|\boldsymbol{y}_0\|_\infty \le K$ by $(\star)$ at $t=0$. The class of constant functions bounded by $K$ in each of the $M'$ coordinates has empirical Rademacher complexity at most $KM'N^{-1/2}$, since $\mathbb{E}_\epsilon|N^{-1}\sum_i \epsilon_i| \le N^{-1/2}$. By the vector-valued contraction inequality of Maurer (2016), which contributes the factor $\sqrt{2}$, and the summation rule,

$$\begin{aligned}\mathbb{E}_{t_i}\left[\widehat{\mathcal{R}}_N(\mathbb{G}^C)\right] &\le \sqrt{2}L\left(\mathbb{E}_{t_i}\left[\widehat{\mathcal{R}}_N(\mathbb{F}^C_{M'})\right] + \mathbb{E}_{t_i}\left[\widehat{\mathcal{R}}_N(\int \mathbb{F}^C_{M'})\right] + KM'N^{-1/2}\right)\\ &\le \sqrt{2}LM'\left(C\kappa\left(1+\sqrt{\mathbb{E}_{t\sim\pi}[t^2]}\right)+K\right)N^{-1/2},\end{aligned}$$

using Lemma 9 for the first two terms. □

The next two lemmas will help to establish that, if the "error function" in Lemma 10 converges to zero, then the kernel-estimated DAE solutions will converge to true DAE solutions.

LEMMA 11. *Let $\pi$ be a measure on $[0,\infty)$ with strictly positive density with respect to Lebesgue measure. Let $\{\dot{x}^{(m)},\dot{\mu}^{(m)}\}_{m=1}^M, \{\dot{y}^{(p)}\}_{p=1}^P \in \mathcal{H}$, define $\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y}$ by integration as in Lemma 10 with $\boldsymbol{x}(0)=\boldsymbol{x}_0$, and suppose $(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t)) \in \mathbb{B}$ for all $t \ge 0$. If $(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})$ satisfies Equations (1) to (3) $\pi$-almost everywhere, then $\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y} \in C^1([0,\infty))$ and Equations (1) to (4) hold* everywhere *on $[0,\infty)$.*

PROOF. Elements of $\mathcal{H}$ are continuous (Equation (59)), so by the fundamental theorem of calculus $\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y}$ are $C^1$ with derivatives $\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}}$, and Equation (4) holds by construction. Since the path lies in $\mathbb{B}$ and $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{H}$ are continuous on $\mathbb{B}$ (Assumption 1), the error function $e := e_{\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}},\boldsymbol{\mu}(0),\boldsymbol{y}(0)}$ of Equation (61) is continuous on $[0,\infty)$. By hypothesis $e = 0$ outside a set $\mathbb{D}$ with $\pi(\mathbb{D}) = 0$. Since $\pi$ has a strictly positive density with respect to Lebesgue measure, no nonempty open interval is $\pi$-null; thus $\mathbb{D}$ has empty interior, or equivalently $[0,\infty)\setminus\mathbb{D}$ is dense. A continuous function vanishing on a dense set vanishes everywhere, so Equations (1) to (3) hold on all of $[0,\infty)$. □

LEMMA 12. *Let $0 \le C < \infty$. For each $n$, let $(\dot{\boldsymbol{x}}_n, \dot{\boldsymbol{\mu}}_n, \dot{\boldsymbol{y}}_n) \in \mathbb{F}^C_{M'}$ (defined as in Lemma 9), let $\boldsymbol{\mu}_{0,n} \in \mathbb{R}^M$, $\boldsymbol{y}_{0,n} \in \mathbb{R}^P$, define $\boldsymbol{x}_n, \boldsymbol{\mu}_n, \boldsymbol{y}_n$ from these by integration as in Lemma 10, and suppose $(\boldsymbol{x}_n(t), \boldsymbol{\mu}_n(t), \boldsymbol{y}_n(t)) \in \mathbb{B}$ for all $t \ge 0$. Then there is a subsequence $n_k$ and a limit $(\dot{\boldsymbol{x}}_\infty, \dot{\boldsymbol{\mu}}_\infty, \dot{\boldsymbol{y}}_\infty, \boldsymbol{\mu}_{0,\infty}, \boldsymbol{y}_{0,\infty})$, with $\boldsymbol{x}_\infty, \boldsymbol{\mu}_\infty, \boldsymbol{y}_\infty$ defined from it by the same integration, such that*

1. $(\dot{x}^{(m)}_{n_k}, \dot{\mu}^{(m)}_{n_k}, \dot{y}^{(p)}_{n_k}) \overset{\text{weakly}}{\longrightarrow} (\dot{x}^{(m)}_\infty, \dot{\mu}^{(m)}_\infty, \dot{y}^{(p)}_\infty) \in \mathcal{H}$ *and* $(\boldsymbol{\mu}_{0,n_k}, \boldsymbol{y}_{0,n_k}) \to (\boldsymbol{\mu}_{0,\infty}, \boldsymbol{y}_{0,\infty})$,
2. $\dot{\boldsymbol{x}}_{n_k} \to \dot{\boldsymbol{x}}_\infty$, $\dot{\boldsymbol{\mu}}_{n_k} \to \dot{\boldsymbol{\mu}}_\infty$, $\dot{\boldsymbol{y}}_{n_k} \to \dot{\boldsymbol{y}}_\infty$ *and* $\boldsymbol{x}_{n_k} \to \boldsymbol{x}_\infty$, $\boldsymbol{\mu}_{n_k} \to \boldsymbol{\mu}_\infty$, $\boldsymbol{y}_{n_k} \to \boldsymbol{y}_\infty$ *uniformly on compact subsets of* $[0, \infty)$,
3. $\|\dot{x}^{(m)}_\infty\|_{\mathcal{H}} \le \liminf_k \|\dot{x}^{(m)}_{n_k}\|_{\mathcal{H}}$ *for every* $m$, *and likewise for* $\dot{\boldsymbol{\mu}}_\infty, \dot{\boldsymbol{y}}_\infty$, *and*
4. $(\boldsymbol{x}_\infty(t), \boldsymbol{\mu}_\infty(t), \boldsymbol{y}_\infty(t)) \in \mathbb{B}$ *for all* $t \ge 0$.

PROOF. Each component of $(\dot{\boldsymbol{x}}_n, \dot{\boldsymbol{\mu}}_n, \dot{\boldsymbol{y}}_n)$ lies in the ball $\{w \in \mathcal{H} : \|w\|_{\mathcal{H}} \le C\}$. Closed balls in a Hilbert space are weakly sequentially compact, and $(\boldsymbol{\mu}_{0,n}, \boldsymbol{y}_{0,n}) = (\boldsymbol{\mu}_n(0), \boldsymbol{y}_n(0))$ lies in the compact set $[\underline{\mu}, \bar{\mu}] \times [\underline{y}, \bar{y}]$; passing to a subsequence gives Claim 1. Claim 3 follows from the weak lower semicontinuity of the Hilbert norm.

For Claim 2, Equation (59) shows that all functions in the radius $C$ RKHS ball are uniformly bounded (by $\kappa C$) and uniformly equicontinuous on $[0, \infty)$. By the Arzelà-Ascoli theorem and a diagonal argument over $[0, j]$, $j \in \mathbb{N}$, a further subsequence converges uniformly on compact sets. The uniform limit must agree with the weak limit, since weak convergence in $\mathcal{H}$ implies pointwise convergence ($w_{n_k}(t) = \langle w_{n_k}, k(t, \cdot)\rangle_{\mathcal{H}} \to \langle w_\infty, k(t, \cdot)\rangle_{\mathcal{H}} = w_\infty(t)$). Uniform convergence of $\boldsymbol{x}_{n_k}, \boldsymbol{\mu}_{n_k}, \boldsymbol{y}_{n_k}$ on $[0, T]$ then follows from

$$\begin{aligned}\sup_{t\in[0,T]} \|\boldsymbol{x}_{n_k}(t) - \boldsymbol{x}_\infty(t)\|_\infty &\le \int_0^T \|\dot{\boldsymbol{x}}_{n_k}(\tau) - \dot{\boldsymbol{x}}_\infty(\tau)\|_\infty \, d\tau \\ &\le T \sup_{\tau\in[0,T]} \|\dot{\boldsymbol{x}}_{n_k}(\tau) - \dot{\boldsymbol{x}}_\infty(\tau)\|_\infty \quad \to 0,\end{aligned}$$

and analogously for $\boldsymbol{\mu}_{n_k}$ and $\boldsymbol{y}_{n_k}$, whose initial values also converge.

Claim 4 holds because $\mathbb{B}$ is closed and $(\boldsymbol{x}_{n_k}(t), \boldsymbol{\mu}_{n_k}(t), \boldsymbol{y}_{n_k}(t)) \in \mathbb{B}$ converges to $(\boldsymbol{x}_\infty(t), \boldsymbol{\mu}_\infty(t), \boldsymbol{y}_\infty(t))$ for each $t$. □

Now we are ready to prove Theorem 3.

PROOF OF THEOREM 3. A solution to Equation (13) exists (Lemma 7) and the true solution is feasible (Lemma 6).

Let $B^2$ be the (deterministic and fixed) objective value of the true solution, as defined in Lemma 5, and let

$$B_N^2 = \textstyle\sum_{m=1}^M \|\dot{\hat{x}}^{(m)}_N\|^2_{\mathcal{H}} + \sum_{m=1}^M \|\dot{\hat{\mu}}^{(m)}_N\|^2_{\mathcal{H}} + \sum_{p=1}^P \|\dot{\hat{y}}^{(p)}_N\|^2_{\mathcal{H}}$$

be the objective value attained by the empirical solution for some $\mathcal{D} = \{t_1, \ldots, t_N\}$. Because the true solution is feasible for every $\mathcal{D}$ and $(\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{y}}_N)$ minimizes the objective,

$$B_N \le B < \infty, \qquad (\dot{\hat{\boldsymbol{x}}}_N, \dot{\hat{\boldsymbol{\mu}}}_N, \dot{\hat{\boldsymbol{y}}}_N) \in \mathbb{F}^B_{M'},$$

where $\mathbb{F}^B_{M'}$ is defined as in Lemma 9. This comparison is the only place where the minimality of the empirical solution is used; the remainder of the proof uses only its feasibility and the bound $B_N \le B$.

Define $e_N(\cdot) := e_{\dot{\hat{\boldsymbol{x}}}_N, \dot{\hat{\boldsymbol{\mu}}}_N, \dot{\hat{\boldsymbol{y}}}_N, \hat{\boldsymbol{\mu}}_N(0), \hat{\boldsymbol{y}}_N(0)}(\cdot)$ and $\mathbb{G}^B$ as in Lemma 10. By feasibility, $e_N \in \mathbb{G}^B$ and $e_N(t_i) = 0$ for every $t_i \in \mathcal{D}$. Since $e_N \ge 0$, a uniform law of large numbers (Wainwright, 2019, Thm. 4.10) gives, for any $\delta > 0$,

$$\begin{aligned}\int_0^\infty e_N(\tau)\pi(d\tau) &= \int_0^\infty e_N(\tau)\pi(d\tau) - \frac{1}{N}\sum_{i=1}^N \underbrace{e_N(t_i)}_{=0} \\ &\le \sup_{e\in\mathbb{G}^B}\left|\int_0^\infty e(\tau)\pi(d\tau) - \frac{1}{N}\sum_{i=1}^N e(t_i)\right| \\ &\le 4\mathbb{E}_{t_i\sim\pi}\left[\widehat{\mathcal{R}}_N(\mathbb{G}^B)\right] + \delta\end{aligned}$$

with probability at least $1 - \exp(-\frac{N\delta^2}{2\widetilde{C}^2})$, where $\widetilde{C}$ is the constant from Lemma 10.[38][39] These tail probabilities are summable in $N$, so by the Borel-Cantelli lemma $\limsup_N \sup_{e\in\mathbb{G}^B} |\int_0^\infty \frac{1}{N}\sum_{i=1}^N e(t_i)| \le \delta$ almost surely for each fixed $\delta$, using $\mathbb{E}_{t_i\sim\pi}[\widehat{\mathcal{R}}_N(\mathbb{G}^B)] \to 0$ (Lemma 10). Applying this along a countable sequence $\delta_k \downarrow 0$ shows that the supremum tends to zero almost surely. Since it bounds $\int_0^\infty e_N\, d\pi$ for every choice of minimizer of Equation (13), this yields

$$\int_0^\infty e_N(\tau)\pi(d\tau) \xrightarrow{\text{a.s.}} 0. \tag{62}$$

Now work on the almost-sure event of Equation (62), and let $(\hat{\boldsymbol{x}}_{N_k}, \hat{\boldsymbol{\mu}}_{N_k}, \hat{\boldsymbol{y}}_{N_k})$ be an arbitrary subsequence. By Lemma 12 with $C = B$ there is a further subsequence $(\hat{\boldsymbol{x}}_{N_{k'}}, \hat{\boldsymbol{\mu}}_{N_{k'}}, \hat{\boldsymbol{y}}_{N_{k'}})$ along which $(\dot{\hat{x}}^{(m)}_{N_{k'}}, \dot{\hat{\mu}}^{(m)}_{N_{k'}}, \dot{\hat{y}}^{(p)}_{N_{k'}})$ converge weakly in $\mathcal{H}$ and $(\hat{\boldsymbol{x}}_{N_{k'}}, \hat{\boldsymbol{\mu}}_{N_{k'}}, \hat{\boldsymbol{y}}_{N_{k'}})$ converges uniformly on compact sets to some limit $(\boldsymbol{x}_\infty, \boldsymbol{\mu}_\infty, \boldsymbol{y}_\infty)$, with

$$\begin{gathered}(\boldsymbol{x}_\infty(t), \boldsymbol{\mu}_\infty(t), \boldsymbol{y}_\infty(t)) \in \mathbb{B} \quad \text{for all } t \ge 0, \\ \textstyle\sum_{m=1}^M \|\dot{x}^{(m)}_\infty\|^2_{\mathcal{H}} + \sum_{m=1}^M \|\dot{\mu}^{(m)}_\infty\|^2_{\mathcal{H}} + \sum_{p=1}^P \|\dot{y}^{(p)}_\infty\|^2_{\mathcal{H}} \;\le\; B^2 \;<\; \infty,\end{gathered} \tag{63}$$

the latter because a sum of limits inferior is at most the limit inferior of the sum. By continuity of $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{H}$ on $\mathbb{B}$ (Assumption 1) and the pointwise convergence of

[38] The supremum over $\mathbb{G}^B$ is measurable: the class is parameterized by a subset of the separable space $\mathcal{H}^{M'} \times \mathbb{R}^{M+P}$, and for fixed $t_1, \dots, t_N$ the values $e(t_i)$ and $\int e\, d\pi$ depend continuously on the parameters (by Equation (59) and dominated convergence), so the supremum may be taken over a countable dense subset.

[39] Wainwright (2019) defines the Rademacher complexity with an absolute value inside the supremum and obtains the constant 2. The error function of the true solution is identically zero and lies in $\mathbb{G}^B$, so $\sup_{e\in\mathbb{G}^B} \sum_i \epsilon_i e(t_i)$ and $\sup_{e\in\mathbb{G}^B} \sum_i (-\epsilon_i) e(t_i)$ are both nonnegative and $\sup_{e\in\mathbb{G}^B} |\sum_i \epsilon_i e(t_i)|$ is at most their sum. Since $-\epsilon$ has the same law as $\epsilon$, the complexity with the absolute value is at most twice $\widehat{\mathcal{R}}_N(\mathbb{G}^B)$, which gives the constant 4.

the derivatives and paths (Lemma 12), $e_{N_{k'}}(t) \to e_\infty(t)$ for every $t$, where $e_\infty(\cdot) = e_{\dot{\boldsymbol{x}}_\infty, \dot{\boldsymbol{\mu}}_\infty, \dot{\boldsymbol{y}}_\infty, \boldsymbol{\mu}_{0,\infty}, \boldsymbol{y}_{0,\infty}}(\cdot)$. Since $e_{N_{k'}} \geq 0$, Fatou's lemma gives us

$$\int_0^\infty e_\infty(\tau)\pi(d\tau) \;\leq\; \liminf_{N_{k'}\to\infty} \int_0^\infty e_{N_{k'}}(\tau)\pi(d\tau) \;=\; 0.$$

Therefore $e_\infty(\cdot) = 0$ $\pi$-almost everywhere, i.e. $(\boldsymbol{x}_\infty, \boldsymbol{\mu}_\infty, \boldsymbol{y}_\infty)$ satisfies Equations (1) to (3) $\pi$-almost everywhere, and Lemma 11 shows that $\boldsymbol{x}_\infty, \boldsymbol{\mu}_\infty, \boldsymbol{y}_\infty$ are $C^1$ and satisfy Equations (1) to (4) *everywhere* on $[0,\infty)$.

By Equation (63), $\dot{x}_\infty^{(m)}, \dot{\mu}_\infty^{(m)} \in \mathcal{H}$ for every $m$, and the initial values $\boldsymbol{x}_\infty(0) = \boldsymbol{x}_0$ and $\boldsymbol{\mu}_\infty(0) \in [\underline{\boldsymbol{\mu}}, \bar{\boldsymbol{\mu}}]$ are finite, so Theorem 2 yields Equation (5). Being $C^1$, $\mathbb{B}$-valued, and a solution of Equations (1) to (4) everywhere, $(\boldsymbol{x}_\infty, \boldsymbol{\mu}_\infty, \boldsymbol{y}_\infty)$ is a trajectory in the sense of Assumption 2 that satisfies Equation (5), so by Assumption 2(i) it equals $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$.

It remains to upgrade this subsequential statement into a claim about the full sequence. Suppose for the sake of contradiction that the full sequence does not converge to $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$ uniformly on compact sets. Writing

$$d_T(N) \;:=\; \sup_{t\in[0,T]} \Big( \|\hat{\boldsymbol{x}}_N(t) - \boldsymbol{x}^*(t)\|_\infty + \|\hat{\boldsymbol{\mu}}_N(t) - \boldsymbol{\mu}^*(t)\|_\infty + \|\hat{\boldsymbol{y}}_N(t) - \boldsymbol{y}^*(t)\|_\infty \Big)$$

for the worst-case approximation error over the window $[0,T]$, this supposition implies that there exists some horizon $T < \infty$, some $\varepsilon > 0$, and indices $N_1 < N_2 < \cdots$ such that $d_T(N_k) \geq \varepsilon$ for every $k$. However, the argument above applies to the subsequence $(\hat{\boldsymbol{x}}_{N_k}, \hat{\boldsymbol{\mu}}_{N_k}, \hat{\boldsymbol{y}}_{N_k})$, yielding a further subsequence $(\hat{\boldsymbol{x}}_{N_{k'}}, \hat{\boldsymbol{\mu}}_{N_{k'}}, \hat{\boldsymbol{y}}_{N_{k'}})$ that converges uniformly on compact sets to $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$. Therefore, $d_T(N_{k'}) \to 0$, and so there are (infinitely many) terms in $N_{k'}$ (and thus infinitely many terms in $N_k$) where the worst-case approximation error is less than $\varepsilon$; a contradiction. Taken together, no such $T$ and $\varepsilon$ exist; hence the full sequence $(\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{y}}_N)$ converges almost surely (uniformly on compact subsets) to $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$. □

## References

Acemoglu, Daron (2008), *Introduction to modern economic growth*. Princeton university press. [4, 5, 14]

Achdou, Yves, Johannes Brumm, and Lukas Frank (2026), "Mastering stochastic OLG models in continuous time." *arXiv preprint arXiv:2608.11134*. [28]

Achdou, Yves, Jiequn Han, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll (2022), "Income and wealth distribution in macroeconomics: A continuous-time approach." *The Review of Economic Studies*, 89 (1), 45–86. [26, 30]

Adenbaum, Jacob, Fil Babalievsky, and William Jungerman (2024), "Learning on the job." URL https://jacobadenbaum.github.io/files/AdenbaumBabalievskyJungerman2024.pdf. [3]

Alvarez, Fernando (2017), "Capital accumulation and international trade." *Journal of Monetary Economics*, 91, 1–18. [21]

Aronszajn, Nachman (1950), "Theory of reproducing kernels." *Transactions of the American Mathematical Society*, 68 (3), 337–404. [34]

Arrow, Kenneth J and Mordecai Kurz (1970), *Public investment, the rate of return, and optimal fiscal policy*. The Johns Hopkins Press. [6]

Aruoba, S. Borağan, Jesús Fernández-Villaverde, and Juan F. Rubio-Ramírez (2006), "Comparing solution methods for dynamic equilibrium economies." *Journal of Economic Dynamics and Control*, 30 (12), 2477–2508. [14]

Azariadis, Costas and Allan Drazen (1990), "Threshold externalities in economic development." *The quarterly journal of economics*, 105 (2), 501–526. [15]

Azinovic, Marlon, Luca Gaegauf, and Simon Scheidegger (2022), "Deep equilibrium nets." *International Economic Review*, 63 (4), 1471–1525. [3]

Azinovic, Marlon and Jan Žemlička (2023), "Economics-inspired neural networks with stabilizing homotopies." *arXiv preprint arXiv:2303.14802*. [3]

Azinovic-Yang, Marlon and Jan Žemlička (2025), "Deep learning in the sequence space." *arXiv preprint arXiv:2509.13623*. [3]

Barnett, Michael, William Brock, Lars Peter Hansen, Ruimeng Hu, and Joseph Huang (2023), "A deep learning analysis of climate change, innovation, and uncertainty." [3]

Belkin, Mikhail (2021), "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation." *Acta Numerica*, 30, 203–248. [28, 31]

Belkin, Mikhail, Daniel Hsu, Siyuan Ma, and Soumik 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 (32), 15849–15854. [2]

Benveniste, L.M and J.A Scheinkman (1982), "Duality theory for dynamic optimization models of economics: The continuous time case." *Journal of Economic Theory*, 27 (1), 1–19, URL https://www.sciencedirect.com/science/article/pii/0022053182900126. [4]

Blanchard, Olivier J and Mark W Watson (1982), "Bubbles, rational expectations and financial markets." In *Crises in the Economic and Financial Structure* (Paul Wachtel, ed.), 295–316, D.C. Heath and Company, Lexington, MA. NBER Working Paper 945. [17]

Blanchard, Olivier Jean and Charles M. Kahn (1980), "The solution of linear difference models under rational expectations." *Econometrica*, 48 (5), 1305–1311. [2, 4]

Brenan, Kathryn E, Stephen L Campbell, and Linda R Petzold (1996), *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, volume 14 of *Classics in Applied Mathematics*. SIAM, Philadelphia. [7]

Brezis, Haim and Haim Brézis (2011), *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer. [33, 34]

Chang, Ya-Chien, Nima Roohi, and Sicun Gao (2019), "Neural lyapunov control." In *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), volume 32, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2019/file/2647c1dba23bc0e0f9cdf75339e120d2-Paper.pdf. [3]

Chen, Guoyuan (2024), "Deep neural network approximations for the stable manifolds of the Hamilton–Jacobi–Bellman equations." *IEEE Transactions on Automatic Control*, 69 (10), 7239–7246. [3]

Chen, Ricky TQ, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud (2018), "Neural ordinary differential equations." *Advances in neural information processing systems*, 31, 6571–6583. [28]

Chen, Yifan, Bamdad Hosseini, Houman Owhadi, and Andrew M. Stuart (2021), "Solving and learning nonlinear PDEs with Gaussian processes." *Journal of Computational Physics*, 447, 110668. [3]

Childers, David, Jesús Fernández-Villaverde, Jesse Perla, Christopher Rackauckas, and Peifan Wu (2022), "Differentiable state-space models and Hamiltonian Monte Carlo estimation." Working Paper 30573, National Bureau of Economic Research. [28]

Chizat, Lénaïc, Edouard Oyallon, and Francis Bach (2019), "On lazy training in differentiable programming." In *Advances in Neural Information Processing Systems*, volume 32, Curran Associates, Inc. [31]

Diba, Behzad T. and Herschel I. Grossman (1988), "The theory of rational bubbles in stock prices." *The Economic Journal*, 98 (392), 746–754, URL http://www.jstor.org/stable/2233912. [17]

Druedahl, Jeppe and Jacob Røpke (2026), "Deep learning algorithms for solving convex finite-horizon models." Working paper, University of Copenhagen. [3]

Ebrahimi Kahou, Mahdi, Jesús Fernández-Villaverde, Sebastián Gómez-Cardona, Jesse Perla, and Jan Rosa (2024), "Spooky boundaries at a distance: Inductive bias, dynamic models, and behavioral macro." Working Paper 32850, National Bureau of Economic Research. [2, 3, 8, 28, 31]

Ebrahimi Kahou, Mahdi, Jesús Fernández-Villaverde, Jesse Perla, and Arnav Sood (2021), "Exploiting symmetry in high-dimensional dynamic programming." Working Paper 28981, National Bureau of Economic Research, URL http://www.nber.org/papers/w28981. [3]

Ebrahimi Kahou, Mahdi, Jesse Perla, and Geoff Pleiss (2026), "Supplement to 'Solving Models of Economic Dynamics with Ridgeless Kernel Regressions'." *Quantitative Economics Supplemental Material*. [4]

Evans, Lawrence C. and Ronald F. Gariepy (1992), *Measure Theory and Fine Properties of Functions*. Studies in Advanced Mathematics, CRC Press, Boca Raton, FL. [6]

Fernández-Villaverde, Jesús, Samuel Hurtado, and Galo Nuño (2023), "Financial frictions and the wealth distribution." *Econometrica*, 91 (3), 869–901, URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA18180. [3]

Gardner, Jacob, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson (2018), "Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration." In *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), volume 31, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2018/file/27e8e17134dd7083b050476733207ea1-Paper.pdf. [11]

Hadamard, Jacques (1902), "Sur les problèmes aux dérivées partielles et leur signification physique." *Princeton university bulletin*, 49–52. [2]

Hadamard, Jacques (1932), *Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques*. Hermann et Cie, Paris. [2]

Han, Jiequn, Yucheng Yang, and Weinan E (2026), "DeepHAM: A global solution method for heterogeneous agent models with aggregate shocks." *Quantitative Economics*, 17 (2), 297–341. [3, 28]

Hastie, Trevor, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani (2022), "Surprises in high-dimensional ridgeless least squares interpolation." *Annals of statistics*, 50 (2), 949. [10]

Jacot, Arthur, Franck Gabriel, and Clément Hongler (2018), "Neural tangent kernel: Convergence and generalization in neural networks." In *Advances in Neural Information Processing Systems*, volume 31, Curran Associates, Inc. [31]

Jungerman, William (2023), "Dynamic monopsony and human capital." Working Paper. [3]

Kase, Hanno, Leonardo Melosi, and Matthias Rottner (2022), "Estimating nonlinear heterogeneous agent models with neural networks." Discussion Paper 17391, Centre for Economic Policy Research. [3]

King, Robert G. and Sergio T. Rebelo (1999), "Resuscitating real business cycles." In *Handbook of Macroeconomics*, volume 1, 927–1007, Elsevier. [14]

Kleywegt, Anton J., Alexander Shapiro, and Tito Homem-de Mello (2002), "The sample average approximation method for stochastic discrete optimization." *SIAM Journal on Optimization*, 12 (2), 479–502. [28, 29]

Li, Xuechen, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud (2020), "Scalable gradients for stochastic differential equations." In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 3870–3882. [28]

Liang, Tengyuan and Alexander Rakhlin (2020), "Just interpolate: Kernel "ridgeless" regression can generalize." *The Annals of Statistics*, 48 (3), pp. 1329–1347, URL https://www.jstor.org/stable/26931513. [10]

Ljungqvist, L. and T.J. Sargent (2018), *Recursive Macroeconomic Theory, fourth edition*. MIT Press, URL https://books.google.ca/books?id=IG1qDwAAQBAJ. [13, 17]

Maliar, Lilia, Serguei Maliar, and Pablo Winant (2021), "Deep learning for solving dynamic economic models." *Journal of Monetary Economics*, 122, 76–101. [3]

Maurer, Andreas (2016), "A vector-contraction inequality for rademacher complexities." In *International Conference on Algorithmic Learning Theory*, 3–17. [41]

Maus, Natalie, Kyurae Kim, Geoff Pleiss, David Eriksson, John P. Cunningham, and Jacob R. Gardner (2024), "Approximation-aware Bayesian optimization." In *Advances in Neural Information Processing Systems*, volume 37. [31]

McKenzie, Lionel W. (1976), "Turnpike theory." *Econometrica*, 44 (5), 841–865, URL http://www.jstor.org/stable/1911532. [3]

Michel, Phillippe (1982), "On the transversality condition in infinite horizon optimal problems." *Econometrica: Journal of the Econometric Society*, 50 (4), 975–985. [4]

Mitrophanov, A. Yu. (2005), "Sensitivity and convergence of uniformly ergodic Markov chains." *Journal of Applied Probability*, 42 (4), 1003–1014. [29]

Murphy, Kevin P. (2022), *Probabilistic Machine Learning: An Introduction*. Adaptive Computation and Machine Learning, MIT Press, URL https://github.com/probml/pml-book. [2, 9, 27]

Murphy, Kevin P. (2023), *Probabilistic Machine Learning: Advanced Topics*. Adaptive Computation and Machine Learning, MIT Press, URL https://github.com/probml/pml-book. [27, 29]

Nakamura-Zimmerer, Tenavi, Qi Gong, and Wei Kang (2022a), "Neural network optimal feedback control with enhanced closed loop stability." In *2022 American Control Conference (ACC)*, 2373–2378, IEEE. [3]

Nakamura-Zimmerer, Tenavi, Qi Gong, and Wei Kang (2022b), "Neural network optimal feedback control with guaranteed local stability." *IEEE Open Journal of Control Systems*, 1, 210–222. [3]

Øksendal, Bernt (2003), *Stochastic Differential Equations: An Introduction with Applications*, 6 edition. Springer, Berlin. [26, 27]

Payne, Jonathan, Adam Rebei, and Yucheng Yang (2024), "Deep learning for search and matching models." URL https://dx.doi.org/10.2139/ssrn.4768566. [3]

Perdomo, Juan C., Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt (2020), "Performative prediction." In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 7599–7609. [29]

Perla, Jesse (2026), "nlls_gram: Metric-aware Levenberg–Marquardt nonlinear least squares for JAX." https://github.com/HighDimensionalEconLab/nlls_gram. Version 2.10.0. [24, 33]

Phelps, Chris, Johannes O. Royset, and Qi Gong (2016), "Optimal control of uncertain systems using sample average approximations." *SIAM Journal on Control and Optimization*, 54 (1), 1–29. [28]

Rasmussen, Carl Edward and Christopher Williams (2006), *Gaussian processes for machine learning*, volume 1. MIT Press. [2, 28, 31, 33]

Ravikumar, B., Ana Maria Santacreu, and Michael Sposi (2019), "Capital accumulation and dynamic gains from trade." *Journal of International Economics*, 119, 93–110, URL https://www.sciencedirect.com/science/article/pii/S002219961930042X. [18, 19, 21]

Ruthotto, Lars, Stanley J. Osher, Wuchen Li, Levon Nurbekyan, and Samy Wu Fung (2020), "A machine learning framework for solving high-dimensional mean field game and mean field control problems." *Proceedings of the National Academy of Sciences*, 117 (17), 9183–9193. [28]

Scheidegger, Simon and Ilias Bilionis (2019), "Machine learning for high-dimensional dynamic stochastic economies." *Journal of Computational Science*, 33, 68–82, URL https://www.sciencedirect.com/science/article/pii/S1877750318306161. [3]

Schmitt-Grohé, Stephanie and Martín Uribe (2003), "Closing small open economy models." *Journal of International Economics*, 61 (1), 163–185. [18]

Schölkopf, Bernhard, Ralf Herbrich, and Alex J Smola (2001), "A generalized representer theorem." In *COLT*, 416–426. [10]

Schölkopf, Bernhard and Alexander J Smola (2002), *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA. [10]

Skiba, A. K. (1978), "Optimal growth with a convex-concave production function." *Econometrica*, 46, 527–539, URL http://www.jstor.org/stable/1914229. [15]

Tikhonov, Andrei Nikolaevich (1963), "On the solution of ill-posed problems and the method of regularization." In *Doklady akademii nauk*, volume 151, 501–504, Russian Academy of Sciences. [2]

Tikhonov, Andrey N. and Vasiliy Y. Arsenin (1977), *Solutions of Ill-Posed Problems*. V. H. Winston & Sons, Washington, D.C. [2]

Turnovsky, Stephen J and Partha Sen (1991), "Fiscal policy, capital accumulation, and debt in an open economy." *Oxford Economic Papers*, 43 (1), 1–24. [18]

Van, Cuong Le, Raouf Boucekkine, and Cagri Saglam (2007), "Optimal control in infinite horizon problems: a sobolev space approach." *Economic Theory*, 32 (3), 497–509. [4]

Wainwright, Martin J (2019), *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press. [13, 36, 43]

Wang, Haoran, Thaleia Zariphopoulou, and Xun Yu Zhou (2020), "Reinforcement learning in continuous time and space: A stochastic control approach." *Journal of Machine Learning Research*, 21 (198), 1–34. [27]

Wang, Ke, Geoff Pleiss, Jacob Gardner, Stephen Tyree, Kilian Q Weinberger, and Andrew Gordon Wilson (2019), "Exact gaussian processes on a million data points." In *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), volume 32, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2019/file/01ce84968c6969bdd5d51c5eeaa3946a-Paper.pdf. [11]

Wendland, Holger (2004), *Scattered Data Approximation*. Number 17 in Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press. [34]

Yang, Yucheng, Chiyuan Wang, Andreas Schaab, and Benjamin Moll (2025), "Structural reinforcement learning for heterogeneous agent macroeconomics." *arXiv preprint arXiv:2512.18892*. [28]

Co-editor [Name Surname; will be inserted later] handled this manuscript.