Solving Models of Economic Dynamics with Ridgeless Kernel Regressions

Mahdi Ebrahimi Kahou[1], Jesse Perla[2], and Geoff Pleiss[3,4]

[1]Department of Economics, Bowdoin College.
[2]Vancouver School of Economics, University of British Columbia.
[3]Department of Statistics, University of British Columbia.
[4]Vector Institute.

October 28, 2025

Abstract

This paper proposes 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 shooting methods enforce the asymptotic boundary conditions by targeting a known steady state—which is numerically unstable, hard to tune, and unable to address cases with steady-state multiplicity. Instead, our approach solves the underdetermined problem without imposing the asymptotic boundary condition, using regularization to select the unique solution fulfilling transversality among admissible trajectories. In particular, ridgeless kernel methods recover this path by selecting the minimum norm solution, coinciding with the non-explosive trajectory. We provide theoretical guarantees showing that kernel solutions satisfy asymptotic boundary conditions without imposing them directly, and we establish a consistency result ensuring convergence within the solution concept of differential-algebraic equations. Finally, we illustrate the method in canonical models and demonstrate its ability to handle problems with multiple steady states.

Keywords: Kernel Methods; Economic Dynamics; Machine Learning; Growth; Inductive Bias.
JEL codes: E13, C45, C63.

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: https://github.com/HighDimensionalEconLab/kernel_econ_alignment.

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, typically arising from transversality conditions in the embedded optimal control problem. Unlike shooting algorithms that aim toward a known steady state, we show that kernel methods can satisfy these asymptotic conditions by selecting the least explosive trajectory among all candidate solutions—and without even calculating the steady-state, let alone imposing it as a condition. This yields robust and computationally efficient algorithms, even under steady-state multiplicity and hysteresis, more stable than parametric approaches such as deep learning, and often easier to tune than classic shooting methods.

Beyond capturing forward-looking behavior, asymptotic boundary conditions are essential for well-posedness in the sense of Hadamard (Hadamard, 1902; Hadamard and Hadamard, 1932)—ensuring 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 regularization to select among the resulting trajectories.[1] For a large class of models, the unique valid solution is also the only non-explosive one; all others fail transversality. By choosing an appropriate function norm and penalizing it, kernel methods recover this non-explosive solution—thereby satisfying the asymptotic boundary conditions.

Kernel Methods. Kernel methods are a natural fit for this task since they provide a formal 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 solutions that perfectly satisfy the differential equation at some finite number of points.

A "kernel machine" is a non-parametric approximation where a function is represented as a weighted sum of the distances to all existing data (i.e., a kernel).[2] For goals ranging from empirical risk-minimization to solving a system of function equations, kernels and RKHS map an infinite dimensional problem in a function space to a finite dimensional problem in the weights of the kernel machine for all data.

[1] See Tikhonov (1963) and Willoughby (1979) 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).

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

Related Work. Our approach is connected with both traditional methods for solving linearized systems and to a recent literature that uses 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 inspiration from this broader approach: algorithms that select non-explosive roots lie on the solution manifold and automatically satisfy transversality conditions.

The use of ML methods is becoming increasingly popular for solving and estimating economic models. Applications span a wide range, including wealth inequality Han et al. (2022), financial frictions Fernández-Villaverde et al. (2023), the heterogeneous impacts of climate change Barnett et al. (2023), portfolio choice problems Azinovic and Žemlička (2023), heterogeneous agent New Keynesian models Kase et al. (2022), human capital accumulation in the labor market Jungerman (2023), and labor market dynamics in search and matching environments Payne et al. (2024). These models often rely on neural networks Azinovic et al. (2022); Maliar et al. (2021); Ebrahimi Kahou et al. (2021, 2024) and even some work with Gaussian Processes Scheidegger and Bilionis (2019).

Recent work in optimal control explores how to achieve stable solutions using ML-based methods Nakamura-Zimmerer et al. (2022b,a); Chen (2023); Chang et al. (2019). More directly within economics, Ebrahimi Kahou et al. (2024) discusses the intuitive connection between the inductive bias of deep neural networks 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 on 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:

- **Inductive bias alignment:** theoretical and empirical evidence that the minimum-norm implicit bias of kernel methods aligns with many asymptotic boundary conditions found in economic problems;

- **Learning the right set of steady states:** evidence that kernel machines identify the steady states of dynamical systems—the ones corresponding to the optimal solution—which leads to highly accurate generalization outside the training data, even without enforcing asymptotic boundary conditions;

- **Consistency of ML estimates:** guarantees that the approximation error of our kernel

methods can be bounded, with the method converging to the true minimum norm solution (and thus solutions satisfying asymptotic boundary conditions) as training data increases; and

- **Robustness and speed:** demonstrations that kernel machines can be competitive in both speed and robustness with traditional methods for modeling economic systems, even on small-scale problems.

Structure. The remainder of this paper is structured as follows. Section 2 describes the class of economic models and provides assumptions on primitives which lead to explosive solutions violating transversality. Section 3 describes how kernel methods map to our class of problems, and discusses cases where a min norm solution is sufficient to fulfill transversality. Theorem 3 provides a consistency result showing how kernel methods converge to the min norm solution, aligning with the solution concept of the DAE itself. Results from our core applications and future directions are presented in Section 4, while Section 5 concludes the paper.

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.[3]

Problem class. In these models, the first-order necessary conditions of the underlying decision problems can be stacked into a system of ordinary differential and algebraic equations (DAEs). The variables can be partitioned 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).[4] Canonical examples of such equations arise from applying Pontryagin's Maximum Principle to the present-value Hamiltonian of a dynamic optimization problem, as in Acemoglu (2008). In that case, $\boldsymbol{\mu}(t)$ represents the present-

[3] An inherent characteristic of discounted infinite-horizon optimal control problems, whether deterministic or stochastic, are 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)).

[4] 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). While this paper analyzes the convergence for DAEs, the computational methods can be used for systems augmenting inequality constraints in addition to Equation (3) and differential inclusions in Equations (1) and (2).

value Lagrange multipliers associated with the state variables $\boldsymbol{x}(t)$. The dynamical system with primitives $\boldsymbol{F}, \boldsymbol{G} : \mathbb{R}^M \times \mathbb{R}^M \times \mathbb{R}^P \to \mathbb{R}^M$ and $\boldsymbol{H} : \mathbb{R}^M \times \mathbb{R}^M \times \mathbb{R}^P \to \mathbb{R}^P$ with a discount rate $r > 0$ is

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

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

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

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

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

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

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

This system is an autonomous, semi-explicit differential-algebraic equation (DAE) 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. Fitting to our notation define: capital $x(t)$, consumption $y(t)$, flow utility $\log(y)$, present-value co-state variable $\mu(t)$, discount rate $r > 0$, depreciation rate $0 < \delta < 1$, and a monotonically increasing and strictly concave production function $f(x)$ where $f(0) = 0$. Writing the standard equations in our notation with $x, \mu, y \in \mathbb{R}^1$,[5]

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

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

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

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

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

[5]The derivation follows the standard planning problem maximizes the lifetime discounted utility of consumption: $\int_0^\infty e^{-rt} u\left(c(t)\right) dt$, with $u(c) = \log(c)$, subject to the law of motion for capital: $\dot{k}(t) = f\left(k(t)\right) - \delta k(t) - c(t)$, where the production function is $f(k) = k^a$ with $0 < a < 1$, and the depreciation rate satisfies $0 < \delta < 1$. In this case, the present-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 mapped to our notation (see Acemoglu (2008)).

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

Key necessary conditions for the DAE to be *well-posed* are that for the given $\boldsymbol{x}_0$, there exists a unique $\boldsymbol{\mu}(0)$ fulfilling the asymptotic boundary condition Equation (5). Assumption 1 provides further assumptions to ensure that there exists a unique $\boldsymbol{y}(0)$ fulfilling Equation (3) given $\boldsymbol{x}(0)$ and $\boldsymbol{\mu}(0)$.

Assumption 1 (Conditions for Well-posedness and Regularity)**.** *Assume that*

- $\boldsymbol{F}$, $\boldsymbol{G}$, *and* $\boldsymbol{H}$ *in Equations* (1) *to* (4) *are Lipschitz with respect to* $\|\cdot\|_\infty$*;*
- $\boldsymbol{F}$ *and* $\boldsymbol{G}$ *have Lipschitz first derivatives and* $\boldsymbol{H}$ *has Lipschitz first and second derivatives;*
- *The Jacobian of* $\boldsymbol{H}$ *with respect to* $\boldsymbol{y}$ *is nonsingular along the relevant trajectories:* $\det\big(\nabla_{\boldsymbol{y}}\boldsymbol{H}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\big) \neq 0$*, and its inverse is Lipschitz continuous with a Lipschitz first derivative in the domain* $t \in [0,T]$.

By the implicit function theorem the last condition in Assumption 1 implies a unique, locally Lipschitz, map $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x},\boldsymbol{\mu})$ fulfilling $\boldsymbol{H}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y}) = \boldsymbol{0}$—so no initial/boundary condition for $\boldsymbol{y}(0)$ is required. This ensures that the system is a semi-explicit DAE with index 1.[6]

The conditions in Assumption 1 ensure uniqueness given an initial condition—eliminating sources of multiplicity that may arise due to the jump variables, $\boldsymbol{y}(t)$.[7]

In addition, we add another standard assumptions on problem formulation to ensure that both the state and co-state variables in a solution are bounded and strictly positive.[8]

Assumption 2 (Bounded Solutions)**.** *Assume that:*

- *For any given* $\boldsymbol{x}_0$*, there is a unique solution,* $\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t)$ *to Equations* (1) *to* (5)*.*
- *There exist bounds such that,* $\boldsymbol{0} < \underline{\boldsymbol{x}} < \boldsymbol{x}(t) < \bar{\boldsymbol{x}} < \infty, \boldsymbol{0} < \underline{\boldsymbol{\mu}} < \boldsymbol{\mu}(t) < \bar{\boldsymbol{\mu}} < \infty$*, and* $0 < \underline{\boldsymbol{y}} < \boldsymbol{y}(t) < \bar{\boldsymbol{y}} < \infty$ *for all* $t > 0$ *on the solution path for a given* $\boldsymbol{x}_0$*.*

[6] Algorithms that reduce the index of a semi-explicit DAE to an ODE, such as the Pantelides algorithm, may augment $\boldsymbol{x}(t)$ and $\boldsymbol{\mu}(t)$ to eliminate the algebraic equation (cf. Equation (3)). While this procedure is often carried out by hand in macroeconomics, we will instead work directly with the more natural DAE formulation.

[7] For example, some models have multiple $(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t))$ fulfill Equations (1) to (5) for a given $\boldsymbol{x}_0$. These often come out of coordination failures in monetary economics and multiplicity in rational expectations equilibria. While these models with multiplicity are not considered in our paper, there may be hysteresis and multiplicity of steady-states—i.e., $\lim_{t\to\infty}\boldsymbol{x}(t)$ may depend on $\boldsymbol{x}_0$. For example, these methods could solve transition paths and steady-states in dynamic models of trade, such as Ravikumar et al. (2019), where steady-state depends on the current account initial conditions.

[8] See Arrow and Kurz (1970, p. 51) for when these assumptions hold. In practice, it is usually sufficient to use a present-value rather than current-value Hamiltonian, and solve a de-trended model in cases with growth.

2.2 Shooting Methods and Explosive Non-Solutions

Before we explore the kernel-based methods in Section 3 and solve the neoclassical growth model in Section 4.1, we will analyze shooting methods to provide intuition on our solution algorithm.

The central computational challenge is that applying the boundary-value Equation (5) numerically is not directly feasible, as it is asymptotic. Shooting methods, and related algorithms solving for an implicit equation with a finite-horizon BVP Methods, use the insight that if we could replace the asymptotic boundary condition in Equation (5) with the right $\boldsymbol{\mu}(0) = \boldsymbol{\mu}_0$, then Assumptions 1 and 2 ensure the solution is a well-posed initial value problem (IVP)—routinely solved for DAEs with millions of equations using software such as Hindmarsh et al. (2005).

First, we need to contrast two types of functions fulfilling the ODEs: a *solution* trajectory fulfills the full set of well-posed equations Equations (1) to (5); and a *non-solution* trajectory fulfills the ill-posed system Equations (1) to (4), but fails the asymptotic boundary condition Equation (5). Assumptions 1 and 2 ensure that we have a unique bounded solution, but there are usually a continuum of non-solutions which can each be associated with a different $\boldsymbol{\mu}(0)$. The solution is indexed by its initial condition $\boldsymbol{\mu}_0$, a single point inside an M-dimensional set of non-solutions.

Shooting methods. In many applications it is straightforward to compute a steady state $\boldsymbol{x}(\infty)$, $\boldsymbol{\mu}(\infty)$, $\boldsymbol{y}(\infty)$, which is bounded by Assumption 2. A shooting algorithm takes an initial co-state $\tilde{\boldsymbol{\mu}}_0$, integrates the dynamics to a large T using a standard DAE/ODE IVP solver, and evaluates

$$\Psi(\tilde{\boldsymbol{\mu}}_0; T) \equiv \begin{bmatrix} \boldsymbol{x}(T; \tilde{\boldsymbol{\mu}}_0) - \boldsymbol{x}(\infty) \\ \boldsymbol{\mu}(T; \tilde{\boldsymbol{\mu}}_0) - \boldsymbol{\mu}(\infty) \\ \boldsymbol{y}(T; \tilde{\boldsymbol{\mu}}_0) - \boldsymbol{y}(\infty) \end{bmatrix}^\top .$$

One then finds $\tilde{\boldsymbol{\mu}}_0$ such that $\Psi(\tilde{\boldsymbol{\mu}}_0; T) \approx \mathbf{0}$ using a root–finding method (e.g., bisection, Newton). The transversality condition in Equation (5) is implicitly used in calculating the steady-state, but it is always present. The stability of this procedure is governed by the Jacobian $\nabla\Psi(\tilde{\boldsymbol{\mu}}_0; T)$.[9]

Challenges with shooting methods. Even when the steady state is easily calculated, in practice the algorithm is unstable and highly sensitive to both the initial guess $\tilde{\boldsymbol{\mu}}_0$ and the chosen horizon T.[10]

[9] Related approaches convert the DAE into a finite-horizon boundary-value problem (BVP) by discretizing time (e.g., via finite differences) and solving a nonlinear system that enforces the initial conditions, artificial terminal conditions, and the dynamics. Although often preferable to shooting, the resulting Jacobian has similar conditioning because the system effectively embeds the same $\Psi(\cdot; T)$ structure.

[10] In many important applications in growth, trade, international economics, and spatial economics, calculating the steady state itself is difficult. Moreover, there may be initial-condition dependence and multiple steady states. In such cases, shooting methods require first solving for all candidate steady states $x(\infty)$ given x_0, analyzing fixed-point stability via the Hessians of Equations (1) to (3), and then partitioning $\mathbb{R}^M$ into basins of attraction. This process is infeasible outside of the simplest, low-dimensional settings.

If T is too small, the algorithm may converge but produces a biased approximation because it forces trajectories to approach the steady state too quickly. If T is too large, then non-solution trajectories—those failing Equation (5)—separate rapidly from the true solution. In that case the Jacobian $\nabla\Psi(\tilde{\boldsymbol{\mu}}_0; T)$ becomes ill-conditioned: $\Psi(\tilde{\boldsymbol{\mu}}_0; T)$ is close to zero in a small neighborhood around the true $\boldsymbol{\mu}_0$ and diverges sharply outside it. This lack of smoothness renders root-finding highly sensitive in higher dimensions. Thus one faces a tradeoff: a small T yields stability but bias, while a large T is unbiased but numerically unstable.

This sensitivity is inherent to optimal control problems due to their saddle-path structure: all trajectories except those along the unique solution manifold diverge. For this problem class the situation is even sharper: Theorem 1 shows that under our assumptions, any non-solution trajectory diverges at least as fast as the exponential discounting rate r appearing in the transversality condition Equation (5). Consequently, even small deviations in the initial condition $\tilde{\boldsymbol{\mu}}_0$ accumulate exponentially over time, directly leading to instability in computing $\nabla\Psi(\cdot; T)$.[11]

Theorem 1. *[Divergence Rate of Non-Solutions] Let $\tilde{\boldsymbol{\mu}}_0$ be the initial condition associated with a non-solution, and let $\boldsymbol{y}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}})$ denote the solution of Equation (3) given Assumption 1, i.e., $\boldsymbol{H}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \boldsymbol{y}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}})) = \mathbf{0}$. Suppose there exist points $\tilde{\boldsymbol{x}}^* \in \mathbb{R}^M$ and $\tilde{\boldsymbol{y}}^* \in \mathbb{R}^P$ such that*

$$\lim_{\tilde{\boldsymbol{\mu}}\to\infty} \boldsymbol{y}(\tilde{\boldsymbol{x}}^*, \tilde{\boldsymbol{\mu}}) = \tilde{\boldsymbol{y}}^*, \quad \lim_{\tilde{\boldsymbol{\mu}}\to\infty} \boldsymbol{F}(\tilde{\boldsymbol{x}}^*, \tilde{\boldsymbol{\mu}}, \boldsymbol{y}(\tilde{\boldsymbol{x}}^*, \tilde{\boldsymbol{\mu}})) = \mathbf{0}, \quad \lim_{\tilde{\boldsymbol{\mu}}\to\infty} \boldsymbol{G}(\tilde{\boldsymbol{x}}^*, \tilde{\boldsymbol{\mu}}, \boldsymbol{y}(\tilde{\boldsymbol{x}}^*, \tilde{\boldsymbol{\mu}})) \leq \mathbf{0}.$$

Then

$$\lim_{t\to\infty} \frac{\dot{\tilde{\boldsymbol{\mu}}}^{(m)}(t)}{\tilde{\boldsymbol{\mu}}^{(m)}(t)} \geq r, \qquad \text{for some } m = 1, \ldots, M.$$

Furthermore, if $\lim_{\boldsymbol{\mu}\to\infty} \boldsymbol{G}(\tilde{\boldsymbol{x}}^*, \boldsymbol{\mu}, \boldsymbol{y}(\tilde{\boldsymbol{x}}^*, \boldsymbol{\mu}))^{(m)} < 0$ *some m, then*

$$\lim_{t\to\infty} \frac{\dot{\tilde{\boldsymbol{\mu}}}^{(m)}(t)}{\tilde{\boldsymbol{\mu}}^{(m)}(t)} > r,$$

Proof. Rewrite Equation (2) componentwise and take limits

$$\lim_{t\to\infty} \frac{\dot{\boldsymbol{\mu}}^{(m)}(t)}{\boldsymbol{\mu}^{(m)}(t)} = \lim_{t\to\infty} \left(r - \boldsymbol{G}(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t))^{(m)} \right) = \lim_{\boldsymbol{\mu}\to\infty} r - \boldsymbol{G}(\tilde{\boldsymbol{x}}^*, \boldsymbol{\mu}, \tilde{\boldsymbol{y}}^*)^{(m)}.$$

Note that if $\lim_{\boldsymbol{\mu}\to\infty} \boldsymbol{G}(\tilde{\boldsymbol{x}}^*, \boldsymbol{\mu}, \tilde{\boldsymbol{y}}^*)^{(m)} < 0$ for some m, then $\lim_{t\to\infty} \frac{\dot{\tilde{\boldsymbol{\mu}}}^{(m)}(t)}{\tilde{\boldsymbol{\mu}}^{(m)}(t)} > r$

□

[11] Formally, the condition number of $\nabla\Psi(\cdot; T)$, the ratio of the largest to smallest singular values, will diverge with T. It is worth noting that if the focus is only on transition dynamics, the explicit calculation of the steady state is not strictly required beyond providing a target for the shooting method. While $\Psi(\cdot; T)$ uses the steady state for guidance, its main role is to steer the algorithm away from explosive non-solutions—and it can be discarded once a valid solution is found. We will exploit this observation in our kernel-based methods.

This result confirms the intuition for why shooting methods are difficult to use in practice. Not only do the non-solutions diverge, they do so exponentially, at least as fast as the transversality condition in Equation (5). Consequently, a small error in the initial condition at time zero, $\tilde{\boldsymbol{\mu}}_0$, accumulates exponentially over time—which leads to the numerical instability of calculating $\nabla\Psi(\cdot;T)$ in shooting methods. .

Verification for neoclassical growth. The condition in Theorem 1 is easily verified in the neoclassical growth model. If $\tilde{\mu}(t) \to \infty$, then by Equation (8) we have $\tilde{y}^* = \lim_{\tilde{\mu}\to\infty} \tilde{\mu}^{-1} = 0$. Use this result with Equation (6) to find that $\tilde{x}^* = \lim_{t\to\infty} \tilde{x}(t)$, must satisfies $f(\tilde{x}^*) = \delta\, \tilde{x}^*$. Rearrange using $f(0) = 0$ to get $\frac{f(\tilde{x}^*)-f(0)}{\tilde{x}^*} = \delta$. By strict concavity $f'(\tilde{x}^*) < \frac{f(\tilde{x}^*)-f(0)}{\tilde{x}^*} = \delta$. Finally, Equation (7) gives us $G(\tilde{x}^*, \tilde{\mu}(t), \tilde{y}^*) = f'(\tilde{x}^*) - \delta < 0$. The rate of divergence, by Theorem 1, is strictly faster than r.

The results of Theorem 1 only partially characterizes the set of non-solutions—concentrating on the primary case which makes shooting methods difficult and our own methods successful. Other cases which pose no challenge for shooting methods—such as a divergent $\hat{\boldsymbol{x}}(t)$ with stationary $\hat{\boldsymbol{\mu}}(t)$ tend to be problem dependent and rely on Assumption 2 (see standard saddle-path analysis in Acemoglu (2008)). For example, in the growth example a divergent $\tilde{x}(t)$ leads negative $\tilde{y}(t)$ to which contradict Assumption 2—and can be eliminated by adding in extra bounds to the optimization problem.

The key takeaway from these results is that shooting methods, and similar approaches imposing the steady-state at a terminal condition, are inherently sensitive to the choice of T since all non-solutions explode exponentially leading to instability when evaluating the system pointwise at the terminal condition.

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 Theorem 1 shows that the pointwise evaluation of transversality at a large T is sensitive, it hints that more global approaches to contrast solutions from non-solutions might be effective. In particular, many function norms and semi-norms can provide a numerically stable alternative. If a non-solution diverges, the norm of the corresponding state or co-state variables will eventually exceed that of the optimal solution for a large enough T, thereby violating Equation (5). Therefore, any algorithm that solves Equations (1) to (4) while controlling a norm of the state and co-state variables over a finite time horizon can get arbitrarily close to the optimal solution.

While we could use this insight for different types of function approximation and norms (e.g.,

Ebrahimi Kahou et al. (2024) empirically shows similar results arising from inductive bias in deep learning), kernel methods have the advantage that the norms are precisely defined in the Reproducing Kernel Hilbert Space (RKHS) and are numerically stable even for large T.

Sobolev norm solutions. Before discussing function approximation in the RKHS space 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.

Differential equations of this sort typically align with a space defined by the Sobolev norm of the solution's derivative. In particular, we consider the Sobolev-$2,2$ space of functions, which is defined as the space of functions $w(\cdot)$ where the function, as well as its first and second (weak) derivatives are square integrable over the domain $[0,T]$:

$$\mathcal{W}^{2,2}([0,T]) = \{w(\cdot) : w(\cdot), \dot{w}(\cdot), \ddot{w}(\cdot) \in L^2([0,T])\}.$$

For any $w(\cdot) \in \mathcal{W}^{2,2}([0,T])$, we have that $\dot{w}(\cdot)$ lives in the Sobolev-$1,2$ space-the space of square-integrable functions with square-integrable first (weak) derivatives. The Sobolev-$1,2$ norm of $\dot{w}(\cdot)$, also defined as the Sobolev-$2,2$ semi-norm of $w(\cdot)$, can be viewed as a measure of complexity of the differential equation solution $w(\cdot)$:

$$|w(\cdot)|_{\mathcal{W}^{2,2}([0,T])} := \|\dot{w}(\cdot)\|_{\mathcal{W}^{1,2}([0,T])} := \left(\int_0^T \left(|\dot{w}(t)|^2 + |\ddot{w}(t)|^2\right) dt\right)^{1/2}.$$

In other words: solutions (and non-solution on a finite domain), and their derivatives belong to a Sobolev space with a well-defined semi-norm.[12] These are reasonable assumptions given the connections between infinite-horizon optimal solutions in economic growth models and Sobolev spaces; see Van et al. (2007) and Chichilnisky (1977).

Back to our setup: intuitively, solutions to Equations (1) to (4) either converge and fulfill Equation (5) or diverge, for example with $\dot{\tilde{\boldsymbol{\mu}}}(t) > 0$. This would lead to solutions having a lower semi-norm. As T increases, the solution and non-solutions increasingly separate since Theorem 1 shows $\frac{\dot{\tilde{\boldsymbol{\mu}}}^{(m)}(t)}{\tilde{\boldsymbol{\mu}}^{(m)}(t)} \geq r$, for some $m = 1, \cdots, M$. Therefore, given the bounding in Assumption 2 to find the solution that fulfills transversality, it is sufficient to choose among all non-solutions those where $\|\boldsymbol{x}\|_{\mathcal{W}^{1,2}([0,T])} + \|\boldsymbol{\mu}\|_{\mathcal{W}^{1,2}([0,T])}$ is minimized.

[12] We note that the semi-norm $|w(\cdot)|_{\mathcal{W}^{2,2}}$ is a more natural measure of complexity than the norm $\|w(\cdot)\|_{\mathcal{W}^{2,2}}$. Consider for example a problem where $w(T)$ converges to some steady state $w(\infty)$ as $T \to \infty$. Convergence to a steady state implies that $\dot{w}(T) \to 0$ as $T \to \infty$, and so it is possible for $\|\dot{w}(\cdot)\|_{L^2([0,T])}$ to be bounded. However, if $w(\infty) > 0$, then $\|w(\cdot)\|_{\mathcal{W}^{1,2}([0,T])}$ will diverge as $T \to \infty$. In other words, norms of $w(\cdot)$ are sensitive to the scaling and location of the steady state, while norms of $\dot{w}(\cdot)$ are not.

3 Method

In this section we first define our approximation class and present an algorithm for solving an underdetermined DAEs, Equations (1) to (4), using ridgeless kernel regression. Next, Theorem 2 shows that a minimum-norm solution is a sufficient condition to satisfy the pointwise transversality requirement—ensuring that the unique solution to the underdetermined kernel regression is the one fulfilling transversality. Finally, we establish in Theorem 3 that ridgeless kernel regression is consistent and therefore asymptotically enforces the minimum norm condition and, by Theorem 2, 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—an exemplar-based approximation in which functions are expressed as combinations of their values at the data, weighted by a function that 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," while guaranteeing that it has the minimum function norm amongst all possible solutions satisfying this condition. As a non-parametric method, the number of parameters defining our approximation grows with the data. Given a kernel function $k(\cdot, \cdot)$, our approximations take the form:

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

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.[13] 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.[14]

In this paper, we use a Matérn kernel for $k(\cdot, \cdot)$, with smoothness ν and lengthscale ℓ (see Definition 1 in appendix Appendix B). The choice of the Matérn kernel family is driven by theoretical

[13] 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.

[14] In particular, since most kernels will have $\lim_{t\to\infty} k(t, t_j) = 0$ for any t_j, approximating the derivatives leads to $\lim_{t\to\infty} \hat{\boldsymbol{\mu}}(t) = \hat{\boldsymbol{\mu}}_0$ for any fixed $\mathcal{D}$. Alternatively, if the function itself was approximated then, for any fixed $\mathcal{D}$, $\lim_{t\to\infty} \hat{\boldsymbol{\mu}}(t) = 0$. That said, given that our goal is to solve for only short- and medium-run behavior, and our consistency results in Theorem 3 only apply on $[0, T]$, approximating the function itself often works in practice.

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 $\hat{\boldsymbol{x}}(t)$ and $\hat{\boldsymbol{\mu}}(t)$ in Equation (11), $\|\hat{\boldsymbol{x}}^{(m)}\|^2_{\mathcal{H}}$ and $\|\hat{\boldsymbol{\mu}}^{(m)}\|^2_{\mathcal{H}}$, are constructed as follows:

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

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

The subscript $\mathcal{H}$ denotes the Reproducing Kernel Hilbert Space (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 Smola and Schölkopf (1998, Chapters 2 & 3). 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 objective in this function space such as $\min_{h\in\mathcal{H}}\{\sum_{i=1}^{N}\Phi(h(t_i))+\lambda||h||^2_{\mathcal{H}}\}$ for some loss $\Phi(\cdot)$ and regularization term $\lambda>0$. The Representer Theorems show this has a unique solution of 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; Smola and Schölkopf, 1998), with norm $\|h\|^2_{\mathcal{H}}=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\,\alpha_j\,k(t_i,t_j)$. Furthermore, standard results (e.g., Liang and Rakhlin (2020) and Hastie et al. (2022)) show that in the limit as $\lambda\to 0$ this is equivalent to directly minimizes that norm subject to interpolating the data, i.e., $\min_{h\in\mathcal{H}}||h||^2_{\mathcal{H}}$ s.t. $\Phi(h(t_i))=0$ for all $t_i\in\mathcal{D}$.

Mapping to 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 to solve,

$$\begin{aligned}
&\min_{\hat{\boldsymbol{x}},\hat{\boldsymbol{\mu}},\hat{\boldsymbol{y}}} \left(\sum_{m=1}^{M} \|\hat{\hat{\boldsymbol{x}}}^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^{M} \|\hat{\hat{\boldsymbol{\mu}}}^{(m)}\|_{\mathcal{H}}^2\right) \\
&\text{s.t. } \hat{\hat{\boldsymbol{x}}}(t_i) = \boldsymbol{F}(\hat{\boldsymbol{x}}(t_i), \hat{\boldsymbol{\mu}}(t_i), \hat{\boldsymbol{y}}(t_i)), \quad \text{for all } t_i \in \mathcal{D} \\
&\quad \hat{\hat{\boldsymbol{\mu}}}(t_i) = r\hat{\boldsymbol{\mu}}(t_i) - \hat{\boldsymbol{\mu}}(t_i) \odot \boldsymbol{G}(\hat{\boldsymbol{x}}(t_i), \hat{\boldsymbol{\mu}}(t_i), \hat{\boldsymbol{y}}(t_i)), \quad \text{for all } t_i \in \mathcal{D} \\
&\quad \boldsymbol{0} = \boldsymbol{H}(\hat{\boldsymbol{x}}(t_i), \hat{\boldsymbol{\mu}}(t_i), \hat{\boldsymbol{y}}(t_i)), \quad \text{for all } t_i \in \mathcal{D}.
\end{aligned} \tag{13}$$

The power of kernel methods is that the Representer Theorems shows that this infinite-dimensional non-parametric problem in a function space becomes tractable and finite-dimensional in $\boldsymbol{\alpha}_j^x$, $\boldsymbol{\alpha}_j^\mu$, $\boldsymbol{\alpha}_j^y$, $\hat{\boldsymbol{\mu}}_0$, and $\hat{\boldsymbol{y}}_0$ since solutions can be expressed as Equations (11) and (12).[15] When the scale is such that constrained optimizers are insufficient, one can instead solve the optimization problem for a fixed, small regularization penalty using the more standard ERM formulation. For details, see appendix Appendix A.

Intuitively, minimizing Equation (13) finds the function with the minimum RKHS norm among all possible interpolating solutions. Next we need to show why this minimum norm solution will be the one which fulfill transversality.

3.2 Transversality

While the intuition that a minimum function norm solution is sufficient to enforce transversality is intuitive, we need to proceed cautiously to relate the norms of the DAE solution concept to that 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 defined over compact domains, it is well established that the Matérn RKHS with $\nu = (P - 1/2)$ and a specific value of ℓ is exactly equal to the $\mathcal{W}^{P,2}([0,T])$ norm for all $P \geq 1$ (see appendix Appendix B). Thus, modelling the derivatives with the $\nu = 1/2$ Matérn kernel (with appropriate lengthscale) produces minimum $\mathcal{W}^{1,2}([0,T])$ derivatives (and thus minimum $\mathcal{W}^{2,2}([0,T])$-seminorm solutions). We are assuming more curvature than is strictly necessary, given that the solution concept of the DAE only requires a first derivative. However, targeting the $\mathcal{W}^{2,2}([0,T])$-semi-norm provides the necessary control to ensure transversality, as we

[15] The primary tradeoffs 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 require evaluating $k(t, t_i)$ for all data. Computational methods can solve these challenges approximately up to millions of observables, Gardner et al. (2018); Wang et al. (2019), but as with all non-parametric methods it eventually becomes a limitation. 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.

will now show.

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

Theorem 2 (Bounding with Norms)**.** *Let f and g be elements of $\mathcal{W}^{2,2}([0,T])$. Then, for some $D_1 < \infty$,*

$$\sup_t |f(t)g(t)| \leq D_1 \left(\|f\|_{\mathcal{H}}^2 + \|g\|_{\mathcal{H}}^2 \right), \tag{14}$$

and from the Sobolev embedding theorem, there exists some $D_2 < \infty$ such that,

$$\sup_t |f(t)g(t)| \leq D_2 \left(\left\|\dot{f}\right\|_{\mathcal{H}}^2 + \|\dot{g}\|_{\mathcal{H}}^2 \right), \tag{15}$$

where $\mathcal{H}$ is the Matérn RKHS with $\nu = \frac{1}{2}$

Proof. See appendix Appendix C. □

This theorem connects the norm of the derivatives to the maximum value the product of two functions can obtain. Intuitively, this result shows why controlling the RKHS norms of the sum of the derivatives might rule out unbounded and explosive functions. Functions that diverges pointwise on $[0,T]$ must have an RKHS norm that becomes unbounded as T grows.

To reiterate, this shows that a minimum norm of the approximated derivatives is sufficient to control the pointwise transversality condition and aligns with the underlying Sobolev norm of the solution concept of the DAE. While we have assumed more smoothness than is strictly necessary, e.g., $\mathcal{W}^{2,2}([0,T])$ instead of $\mathcal{W}^{1,2}([0,T])$, this is not a significant limitation in practice as we demonstrate in Section 4.[16]

3.3 Consistency

While applying Theorem 2 shows that the min norm solution of the DAE will be the one which fulfills 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

[16] However, this also shows where these methods are likely to break down. For example, in cases which are not twice-differentiable almost everywhere, or where the level of the function rather than just its derivatives are essential for selecting the trajectory fulfilling Equation (5).

demonstrate that our approximation is a *consistent estimator* of the true minimum norm solution, meaning that as $N \to \infty$ the empirical solutions $\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{y}}_N$ converge to Equations (1) to (5).

Theorem 3 (Consistency)**.** *Given some $0 < K < \infty$, let $\mathbb{S}$ be the set of functions $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{y})$ that satisfy Equations (1) to (4) and Assumption 1 with $\boldsymbol{x}(0) = \boldsymbol{x}_0$ and $\|\boldsymbol{\mu}(0)\|_\infty, \|\boldsymbol{y}(0)\|_\infty \leq K$. Then the minimum norm solution*

$$(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*) = \inf_{(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\in\mathbb{S}} \sum_{m=1}^{M} \|\dot{x}^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^{M} \|\dot{\mu}^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^{P} \|\dot{y}^{(m)}\|_{\mathcal{H}}^2$$

exists and has bounded $\mathcal{H}$ norm. Moreover, if $t \in \mathcal{D}$ are drawn uniformly i.i.d. from $[0, T]$ then the solutions $\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{y}}_N$ from Equation (13) with the Matérn-1/2 kernel satisfies Equations (1) to (4) almost everywhere in the limit as $N \to \infty$ and

$$\begin{aligned} &\lim_{N\to\infty} \sum_{m=1}^{M} \|\hat{\dot{x}}_N^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^{M} \|\hat{\dot{\mu}}_N^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^{P} \|\hat{\dot{y}}_N^{(m)}\|_{\mathcal{H}}^2 \\ \overset{\text{a.s.}}{=}\ &\inf_{(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\in\mathbb{S}} \sum_{m=1}^{M} \|\dot{x}^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^{M} \|\dot{\mu}^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^{P} \|\dot{y}^{(m)}\|_{\mathcal{H}}^2. \end{aligned}$$

Proof. See appendix Appendix D. □

Though this proof borrows techniques from the statistical learning literature on kernel methods (e.g., Wainwright, 2019), it is non-trivial for several reasons. First, we need to show that solutions to the DAE exist in the RKHS induced by the Matérn kernel, and that a minimum norm solution exists. Second, we need to prove that uniform convergence of derivatives implies uniform convergence of the DAE solution functions themselves. Given these results, we can be confident that with enough data and a large enough T, the min semi-norm solution exists, will be consistently approximated by the empirical solutions $\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{y}}_N$ fulfilling Equations (1) to (4), and that it will be sufficient to ensure the transversality condition Equation (5) holds.

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—and summarize additional experiments in Section 4.4.

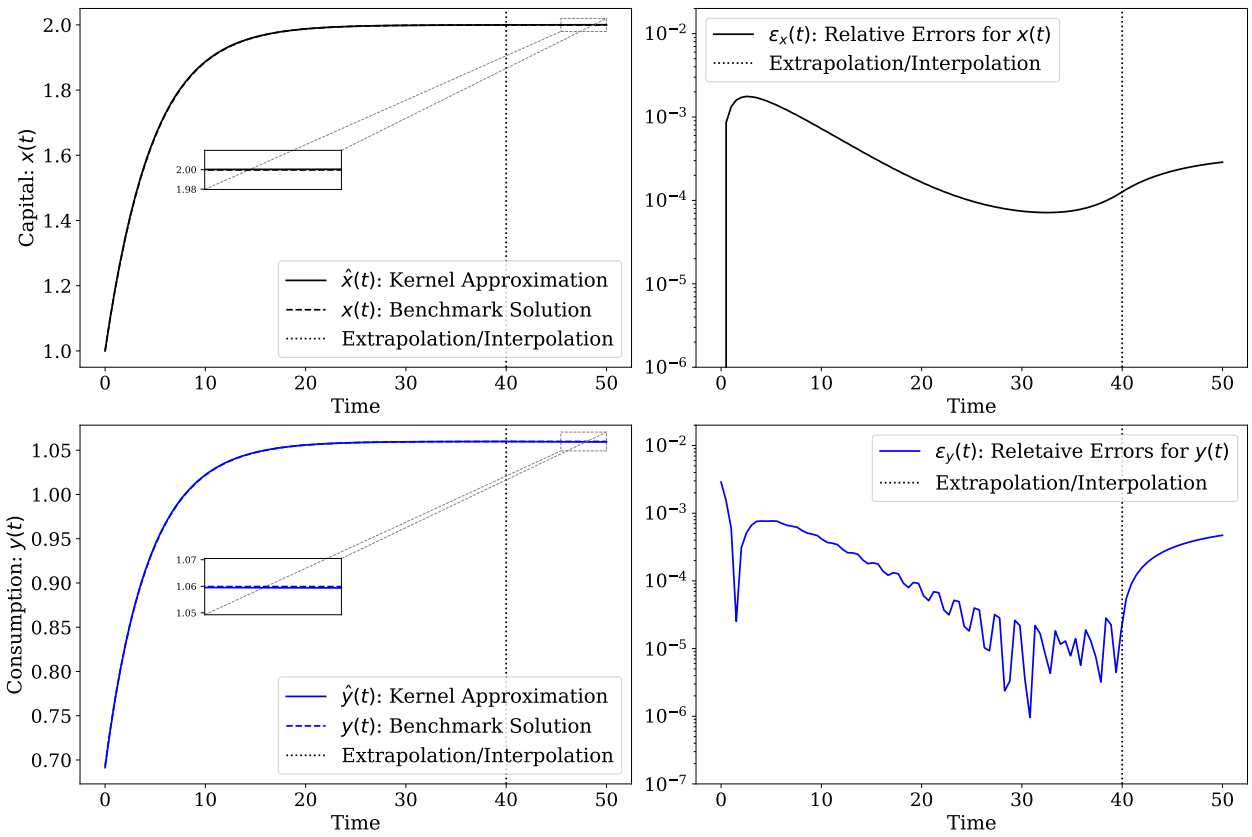

Figure 1: Solution of the neoclassical growth model (Equations (6) to (9)) without imposing the transversality condition (Equation (10)). The vertical dotted lines marks the end of the training set. The top panels show the results for capital, $x(t)$ and associated relative error, and the bottom two show results for consumption, $y(t)$. The relative error, are low, on the order of 0.01% even when extrapolating beyond $\mathcal{D}$.

In all cases, we use a Matérn kernel with $\nu = \frac{1}{2}$, $\ell = 10$, and $\sigma = 1$ (see Definition 1) and use $\mathcal{D} := \{0, 1, 2, \ldots, 40\}$. 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. In our baseline model, all cases are solved using open-source constrained optimizers 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). Our baseline parameters are $x_0 = 1.0, \delta = 0.1, r = 0.11, f(x) = x^a$ and $a = \frac{1}{3}$.

Results. Figure 1 shows the consumption and capital relative to a benchmark.[17] 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 since we approximate the derivatives rather than the functions themselves and Matérn kernels are zero-reverting.

Sufficiency of the minimum norm solution. As discussed in Section 2.2, as long as $f(x)$ is monotone and strictly concave, $f(0) = 0$, and $\delta > 0$, then this model fulfills Theorem 1. That is, non-solutions of $\tilde{\mu}(t)$ diverge asymptotically at a rate greater than r. Since these are the key failures of transversality, the only modification to the optimization problem Equation (13) required in practice was to impose the bound $\hat{y}_0 > 0$, which prevented all other violations of Assumption 2.[18]

Robustness. While Theorem 3 demonstrate consistency, it is helpful to check our methods' sensitivity to hyperparameters and features of $\mathcal{D}$: Section 1.1 of the Supplemental Appendix shows that our methods perform well with a much sparser and irregular $\mathcal{D}$; section 1.2 of Supplemental Appendix indicates low sensitivity to different kernel hyperparameters; and section 1.3 of Supplemental Appendix demonstrates that the approximation 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—an important consideration for problems where the convergence rate is unknown and the appropriate choice of T is not clear a priori.

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 $b_2 = 2.5$. The derivative of the production function, $f'(x)$, exhibits a discontinuity at $\bar{x} = \left(\frac{b_2}{b_1 - 1}\right)^{\frac{1}{a}}$. As a result this problem has two steady states, but a unique transition path for any given initial condition.

This poses a significant challenge for classical algorithms, such as shooting and 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 $\boldsymbol{x}_0$. 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.

[17] In this case we solved this as a mixed initial-boundary value problem using the analytically calculated boundary condition as a boundary condition. As discussed in Section 2.

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

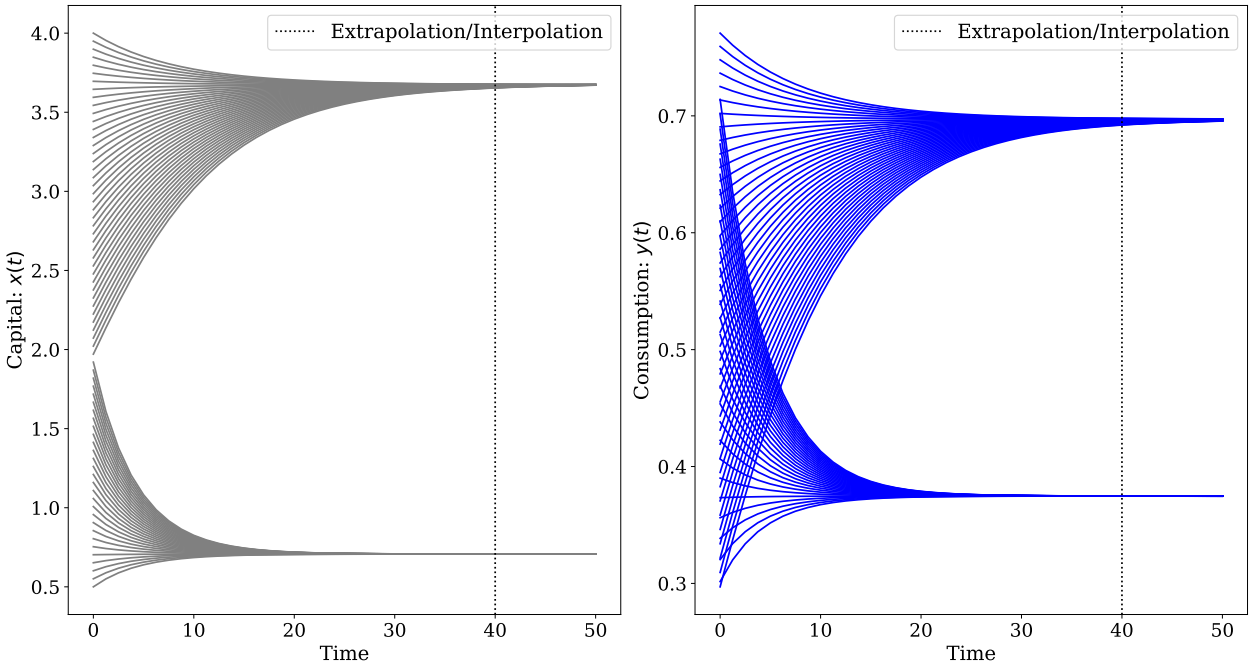

Figure 2: Solution of the neoclassical growth model (Equations (6) to (9)) with multiple steady states due to the concave-convex production function in Section 4.2. The left panel shows the solution trajectories for capital, $x(t)$, for 70 different initial conditions $x_0 \in [0.5, 4]$. The right panel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

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

While Theorem 1 does not hold globally, it is true local to both steady-states—eliminating the key non-solutions. However, 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 correctly corresponds with 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 consequence of the minimum norm solution. Consider the two possible trajectories 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; for instance, see Ljungqvist and Sargent (2018). In this context, the transversality condition is also referred to as the "no-bubble" condition.[19]

Model. The model values a stream of dividends, where a "bubble" is defined as a price path whose dynamics cannot be explained by the dividends process.

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

$$\dot{x}(t) = c + gx(t), \tag{16}$$

$$\dot{\mu}(t) = r\mu(t) - x(t) := r\mu(t) - \mu(t)\frac{x(t)}{\mu(t)}, \tag{17}$$

$$0 = \lim_{t\to\infty} e^{-rt}\mu(t)x(t), \tag{18}$$

where c and g are constants governing the dividend process, and $r > 0$ denotes the discount rate of a risk-neutral investor. Equation (18) is the "no-bubble" condition. The set of solutions to Equations (16) and (17), without imposing Equation (18), can be found analytically:

$$\mu(t) = \textstyle\int_0^\infty e^{-r\tau}x(t+\tau)\mathrm{d}\tau = \mu_f(t) + \zeta e^{rt}, \tag{19}$$

$$\mu_f(t) := \tfrac{c}{r-g} + \left(x_0 - \tfrac{c}{r-g}\right)e^{(r-g)t}, \tag{20}$$

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 (18)) is imposed, this problem is well-posed, with a unique solution of $\mu(t) = \mu_f(t)$ (i.e., $\zeta = 0$).

Selecting the "no-bubble" solution. A key advantage of this example is that Equation (19) characterizes the full set of deterministic non-solutions to Equations (16) and (17) which do not impose Equation (18). For a given function norm, apply the triangle inequality to the set of solutions from Equation (19) to yield $\|\mu_f\|_{\mathcal{W}} \leq \|\mu\|_{\mathcal{W}} \leq \|\mu_f\|_{\mathcal{W}} + \zeta\|e^{rt}\|_{\mathcal{W}}$. Differentiating yields $\|\dot{\mu}_f\|_{\mathcal{W}} \leq \|\dot{\mu}\|_{\mathcal{W}} \leq \|\dot{\mu}_f\|_{\mathcal{W}} + \zeta r\|e^{rt}\|_{\mathcal{W}}$ where $\mathcal{W}$ is a norm such as $\mathcal{W}^{1,2}([0,T])$. Finally, note that the this semi-norm is minimized when $\zeta = 0$.

To verify the divergence rate in Theorem 1, note from Equation (17) that if $\boldsymbol{x}(t)$ converges to a finite steady state while $\boldsymbol{\mu}(t)$ diverges, then $\lim_{t\to\infty} \boldsymbol{G}(\boldsymbol{x}(t), \boldsymbol{\mu}(t)) = \lim_{t\to\infty} \frac{\boldsymbol{x}(t)}{\boldsymbol{\mu}(t)} = 0$, so $\dot{\boldsymbol{\mu}}(t)/\boldsymbol{\mu}(t) \to r$ and $\boldsymbol{\mu}(t)$ grows asymptotically at rate r. However, note that the speed of separating solutions from non-solutions is slower than our previous example, where the non-solutions

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

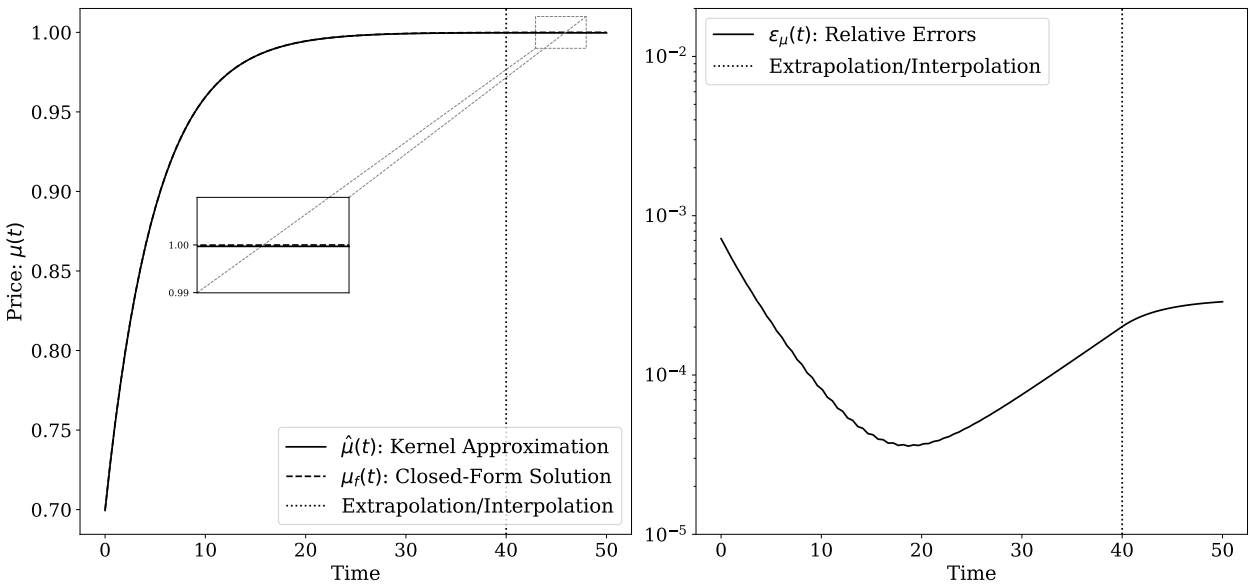

Figure 3: Solution of the linear asset pricing model (Equations (16) and (17)) without imposing the "no-bubble" condition (Equation (18)). The vertical dotted lines marks the end of the training set.

asymptotically diverged at rates strictly faster than r.

Results. Our baseline parameters are $x_0 = 1.0, c = 0.02, g = -0.2$, and $r = 0.1$. Figure 3 shows the results of the ridgeless kernel machine alongside the fundamental price from Equation (20). Even without imposing the long-run "no-bubble" condition in Equation (18), the kernel method recovers the "fundamental" price with high accuracy. As before, the minimum norm still selects the correct solution and as T grows the norms of non-solutions will grow quickly due to Theorem 1.

4.4 Other Examples

To show that our method can handle problems of increasing dimensionality, we test it on a standard model of human capital and economic growth in section 2.1 of the Supplemental Appendix. In our formulation of the model, there are two jump variables and three co-state variables. With classical methods such as shooting, finding the optimal solution requires a five-dimensional search. By comparison, our algorithm computes the solution in under a second. To demonstrate the applicability of our algorithm beyond macro and finance, in section 2.2 of the Supplemental Appendix, we test it on the optimal advertising model, a classic framework in the marketing literature.

4.5 Future Directions

There are several natural directions for future applications and extensions. One is extending kernel methods to handle inequality and complementarity constraints, which would make them applicable

to models such as lifecycle consumption and macroeconomic models with financial frictions. Another direction is adapting kernel methods to stochastic and recursive settings, where conditions like transversality must hold globally and involve expectations over random trajectories. Finally, it is important to study the performance of kernel methods in very high-dimensional settings, such as those found in the trade and spatial economics literature.

The complexity of kernel methods depends on sample size rather than the dimensionality of the state space. If a kernel captures similarity in the state space, only a modest number of samples may suffice to solve dynamic economic problems, making these methods well-suited for both very high-dimensional and recursive stochastic models.

5 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.

Appendix A Ridge Regression Formulation

The optimization problem in Equation (13) is equivalent to the limit of ridge regression, as follows:

$$\lim_{\lambda \to 0} \left\{ \min_{\hat{\dot{\boldsymbol{x}}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{y}}} \left\{ \sum_{t_i \in \mathcal{D}} \left[\left\| \hat{\dot{\boldsymbol{x}}}(t_i) - \boldsymbol{F}(\hat{\boldsymbol{x}}(t_i), \hat{\boldsymbol{\mu}}(t_i), \hat{\boldsymbol{y}}(t_i)) \right\|_2^2 \right.\right.\right.$$
$$\left. + \left\| \hat{\boldsymbol{\mu}}(t_i) - r\, \hat{\boldsymbol{\mu}}(t_i) + \hat{\boldsymbol{\mu}}(t_i) \odot \boldsymbol{G}(\hat{\boldsymbol{x}}(t_i), \hat{\boldsymbol{\mu}}(t_i), \hat{\boldsymbol{y}}(t_i)) \right\|_2^2 + \| \boldsymbol{H}(\hat{\boldsymbol{x}}(t_i), \hat{\boldsymbol{\mu}}(t_i), \hat{\boldsymbol{y}}(t_i)) \|_2^2 \right]$$
$$\left. \left. + \| \hat{\boldsymbol{x}}(0) - \boldsymbol{x}_0 \|_2^2 + \lambda \Big(\sum_{m=1}^{M} \| \hat{\dot{\boldsymbol{x}}}^{(m)} \|_{\mathcal{H}}^2 + \sum_{m=1}^{M} \| \hat{\dot{\boldsymbol{\mu}}}^{(m)} \|_{\mathcal{H}}^2 \Big) \right\} \right\},$$

The formulation in Equation (13) is referred to as *ridgeless* due to the vanishing ridge penalty term as $\lambda \to 0$. In practice, rather than taking the limit, one can simply choose a small fixed λ. For the applications presented in this paper, we solved the problem for λ in the range 10^{-4} to 10^{-6}. The results are omitted, as they were nearly identical to the solution of the optimization problem in Equation (13).

Appendix B Connection Between Sobolev-$P, 2$ Spaces and 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 ν and lengthscale ℓ. For the purposes of this paper, we will define $k_{\nu,\ell}$ as:

$$k_{\nu,\ell}(t,t') = \kappa(|t-t'|), \qquad \widehat{\kappa}(\omega) := \left(1+\ell^2\omega^2\right)^{-\nu-1/2}, \tag{21}$$

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

Note that Equation (21) corresponds to the standard Matérn kernel definition (Rasmussen and Williams, 2006) after appropriately scaling the inputs and outputs. $\nabla^{(r)} k_{\nu,\ell}(\cdot,\cdot)$ will denote the r^{th} derivative of $k_{\nu,\ell}$ with respect to its its first argument.

Given some interval $[0,T] \subseteq \mathbb{R}$, $\mathcal{H}^{\nu,\ell}([0,T])$ denotes the RKHS of $[0,T] \to \mathbb{R}$ functions where the reproducing kernel is equal to $k_{\nu,\ell}$. $\mathcal{W}^{P,2}([0,T])$ denotes to the Sobolev-$P, 2$ space of $[0,T] \to \mathbb{R}$ functions. Whenever possible we will drop the superscripts for $\mathcal{H}^{\nu,\ell}([0,T])$.

Theorem 4 (Equivalence of $\mathcal{W}^{P,2}$ and $\mathcal{H}^{\nu,\ell}$)**.** *For any positive integer P, any $\ell > 0$, and any $\nu = P - 1/2$ there exists positive constants $C_1(\nu,\ell)$ and $C_2(\nu,\ell)$ so that*

$$C_1(\nu,\ell)\|w\|_{\mathcal{W}^{P,2}([0,T])} \leq \|w\|_{\mathcal{H}^{\nu,\ell}([0,T])} \leq C_2(\nu,\ell)\|w\|^{\mathcal{W}^{P,2}([0,T])}.$$

for all $w \in \mathcal{W}^{P,2}$. In other words, the Sobolev norm $\|\cdot\|_{\mathcal{W}^{P,2}([0,T])}$ and the RKHS norm $\|\cdot\|_{\mathcal{H}^{\nu,\ell}([0,T])}$ are equivalent, and thus $\mathcal{W}^{P,2}([0,T]) = \mathcal{H}^{\nu,\ell}([0,T])$.

Proof. We begin by first establishing an equivalence between $\mathcal{W}^{P,2}(\mathbb{R})$ and $\mathcal{H}^{\ell,\nu}(\mathbb{R})$. In the Fourier domain, the Sobolev norm for $\mathbb{R} \to \mathbb{R}$ functions is given by

$$\|w\|_{\mathcal{W}^{P,2}(\mathbb{R})} = \left\| \widehat{w}(\cdot)\left(1+(\cdot)^2\right)^{P/2} \right\|_{L_2(\mathbb{R})},$$

and, for any RKHS $\mathcal{H}(\mathbb{R})$ with stationary reproducing kernels, the RKHS norm for $\mathbb{R} \to \mathbb{R}$ functions is given by

$$\|w\|_{\mathcal{H}(\mathbb{R})} = \left\| \widehat{w}(\cdot)\left(\widehat{\kappa}(\cdot)\right)^{-1/2} \right\|_{L_2(\mathbb{R})}$$

The Matérn kernel, as defined in Equation (21), has a Fourier transform that decays at at a rate of $(1+|\cdot|^2)^{-P}$ and thus $\|\cdot\|_{\mathcal{H}^{\nu,\ell}(\mathbb{R})}$ is bounded above and below by a constant multiple of $\|\cdot\|_{\mathcal{W}^{P,2}(\mathbb{R})}$ where the constant only depends on ℓ. This argument can be generalized to prove an equivalence between $\mathcal{W}^{P,2}([0,T])$ and $\mathcal{H}^{\ell,\nu}([0,T])$, as the domain $[0,T]$ trivially has a Lipschitz boundary. See

(Wendland, 2004, Corollary 10.48) for details. □

Corollary 5 (Equality of $\mathcal{W}^{P,2}$ and $\mathcal{H}^{\nu,\ell}$ with specific ν, ℓ values)**.** *For any P, there exists a value of ℓ so that, for all $w \in \mathcal{W}^{P,2}([0,T])$:*

$$\|w\|_{\mathcal{W}^{P,2}([0,T])} = \|w\|_{\mathcal{H}^{P-1/2,\ell}([0,T])}.$$

Appendix C Proof of Theorem 2

Theorem 2 (Restated). Let f and g be elements of $\mathcal{W}^{2,2}([0,T])$. Then, for some $D_1 < \infty$,

$$\sup_t |f(t)g(t)| \leq D_1 \left(\|f\|_{\mathcal{H}}^2 + \|g\|_{\mathcal{H}}^2 \right), \tag{22}$$

and from the Sobolev embedding theorem, there exists some $D_2 < \infty$ such that,

$$\sup_t |f(t)g(t)| \leq D_2 \left(\left\|\dot{f}\right\|_{\mathcal{H}}^2 + \|\dot{g}\|_{\mathcal{H}}^2 \right), \tag{23}$$

where $\mathcal{H}$ is the Matérn RKHS with $\nu = \frac{1}{2}$

Proof. Let $k(\cdot,\cdot)$ be the kernel associated with the RKHS $\mathcal{H}$ where $\sup_t k(t,t) \leq K < \infty$. For $f, g \in \mathcal{H}$ by the reproducing property,

$$f(t) = \langle f, k(t,\cdot)\rangle_{\mathcal{H}} \quad \text{and} \quad g(t) = \langle g, k(t,\cdot)\rangle_{\mathcal{H}} \tag{24}$$

Then substitute with Equation (24), use the Cauchy-Schwarz inequality, and then the kernel bound to get

$$\begin{aligned} |f(t)g(t)| &\leq |f(t)||g(t)| \leq ||f||_{\mathcal{H}}||k(t,\cdot)||_{\mathcal{H}^M}\, ||g||_{\mathcal{H}}||k(t,\cdot)||_{\mathcal{H}} \\ &\leq K||f||_{\mathcal{H}}||g||_{\mathcal{H}} \end{aligned}$$

Then, by the Arithmetic-Geometric Mean Inequality,

$$\leq \frac{K}{2}\left(||f||_{\mathcal{H}}^2 + ||g||_{\mathcal{H}}^2\right)$$

By Theorem 4, there exists a constant K' such that

$$\frac{K}{2}\left(||f||_{\mathcal{H}}^2 + ||g||_{\mathcal{H}}^2\right) \leq K'\left(||f||_{\mathcal{W}^{1,2}([0,T])}^2 + ||g||_{\mathcal{W}^{1,2}([0,T])}^2\right)$$

From the Sobolev embedding theorem

$$K'\left(||f||_{\mathcal{W}^{1,2}([0,T])}^2 + ||g||_{\mathcal{W}^{1,2}([0,T])}^2\right) \leq K'\left(||\dot{f}||_{\mathcal{W}^{1,2}([0,T])}^2 + ||\dot{g}||_{\mathcal{W}^{1,2}([0,T])}^2\right)$$

Finally, from Theorem 4, there exists some D such that

$$\leq D\left(\left\|\dot{f}\right\|_{\mathcal{H}}^2 + \|\dot{g}\|_{\mathcal{H}}^2\right).$$

□

Appendix D Proof of Theorem 3

Theorem 3 (Restated). Given some $0 < K < \infty$, let $\mathbb{S}$ be the set of functions $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{y})$ that satisfy Equations (1) to (4) and Assumption 1 with $\boldsymbol{x}(0) = \boldsymbol{x}_0$ and $\|\boldsymbol{\mu}(0)\|_\infty, \|\boldsymbol{y}(0)\|_\infty \leq K$. Then the minimum norm solution

$$(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*) = \inf_{(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\in\mathbb{S}} \sum_{m=1}^M \|\dot{x}^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^M \|\dot{\mu}^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^P \|\dot{y}^{(m)}\|_{\mathcal{H}}^2$$

exists and has bounded $\mathcal{H}$ norm. Moreover, if $t \in \mathcal{D}$ are drawn uniformly i.i.d. from $[0, T]$ then the solutions $\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{y}}_N$ from Equation (13) with the Matérn-1/2 kernel satisfies Equations (1) to (4) almost everywhere in the limit as $N \to \infty$ and

$$\begin{aligned} &\lim_{N\to\infty} \sum_{m=1}^M \|\hat{\dot{x}}_N^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^M \|\hat{\dot{\mu}}_N^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^P \|\hat{\dot{y}}_N^{(m)}\|_{\mathcal{H}}^2 \\ \overset{\text{a.s.}}{=}\ &\inf_{(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\in\mathbb{S}} \sum_{m=1}^M \|\dot{x}^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^M \|\dot{\mu}^{(m)}\|_{\mathcal{H}}^2 + \sum_{m=1}^P \|\dot{y}^{(m)}\|_{\mathcal{H}}^2. \end{aligned}$$

Throughout this proof we will drop the domain $[0, T]$ from the $\mathcal{W}$ for brevity. For notational simplicity we redefine $\dot{\boldsymbol{\mu}}(t) = r\boldsymbol{\mu}(t) - \boldsymbol{\mu}(t) \odot \boldsymbol{G}(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t)) \equiv \boldsymbol{B}(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t))$. It is trivial to see that $\boldsymbol{B}$ is Lipschitz with a Lipschitz first derivative if $\boldsymbol{G}$ is.

We break up this proof into a series of lemmas.

Lemma 6. *For every $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{y}) \in \mathbb{S}$, the functions $x^{(1)}, \ldots, x^{(M)}$, $\mu^{(1)}, \ldots, \mu^{(M)}$, and $y^{(1)}, \ldots, y^{(P)}$, are all elements of $\mathcal{W}^{2,2}$.*

Proof. First, we note that Assumption 1 allows us to rewrite the DAE as an ODE. Specifically, Equation (4) can be rewritten an equation of $\dot{\boldsymbol{y}}(t)$ by differentiating both sides with respect to t

$$\begin{aligned}\mathbf{0} &= \nabla_{\boldsymbol{x}}\boldsymbol{H}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\dot{\boldsymbol{x}}(t) + \nabla_{\boldsymbol{\mu}}\boldsymbol{H}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\dot{\boldsymbol{\mu}}(t) + \nabla_{\boldsymbol{y}}\boldsymbol{H}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\dot{\boldsymbol{y}}(t) \\ &= \nabla_{\boldsymbol{x}}\boldsymbol{H}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\boldsymbol{F}(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t)) + \nabla_{\boldsymbol{\mu}}\boldsymbol{H}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\boldsymbol{B}(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t)) \\ &\quad + \nabla_{\boldsymbol{y}}\boldsymbol{H}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\dot{\boldsymbol{y}}(t).\end{aligned}$$

By assumption, $\nabla_{\boldsymbol{y}}\boldsymbol{H}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})$ is non-singular on all relevant trajectories, so we have that

$$\begin{aligned}\dot{\boldsymbol{y}}(t) &= \nabla_{\boldsymbol{y}}\boldsymbol{H}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})^{-1}\Big(\nabla_{\boldsymbol{x}}\boldsymbol{H}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\boldsymbol{F}(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t)) \\ &\qquad + \nabla_{\boldsymbol{\mu}}\boldsymbol{H}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\boldsymbol{B}(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t))\Big) \\ &=: \boldsymbol{C}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y}).\end{aligned}$$

By Lipschitz continuity of $\boldsymbol{F}$, $\boldsymbol{B}$, the derivative of $\boldsymbol{H}$, the inverse of $\nabla_{\boldsymbol{y}}\boldsymbol{H}$, and all their respective derivatives, we have that $\boldsymbol{C}$ is Lipschitz continuous with a Lipschitz first derivative over the interval $[0,T]$.

We thus can rewrite the DAE as the following ODE system:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t)), \quad \dot{\boldsymbol{\mu}}(t) = \boldsymbol{B}(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t)), \quad \dot{\boldsymbol{y}}(t) = \boldsymbol{C}(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t)),$$

where $\boldsymbol{F},\boldsymbol{B},\boldsymbol{C}$ are all Lipschitz continuous with Lipschitz first derivatives. By Lipschitz continuity of $\boldsymbol{F}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and the Cauchy-Lipschitz-Picard theorem (e.g. Brezis and Brézis, 2011, Thm. 7.3), we have that every $(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y}) \in \mathbb{S}$ are continuous and continuously differentiable over $[0,\infty)$. This fact implies that $\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y}$ and $\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}}$ are bounded over $[0,T]$. Now consider $\ddot{\boldsymbol{x}}$, $\ddot{\boldsymbol{\mu}}$, $\ddot{\boldsymbol{y}}$

$$\begin{aligned}\ddot{\boldsymbol{x}}(t) &= \dot{\boldsymbol{F}}(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t)) \\ \ddot{\boldsymbol{\mu}}(t) &= \dot{\boldsymbol{B}}(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t)) \\ \ddot{\boldsymbol{y}}(t) &= \dot{\boldsymbol{C}}(\boldsymbol{x}(t),\boldsymbol{\mu}(t),\boldsymbol{y}(t)).\end{aligned}$$

Because $\boldsymbol{F}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ admit Lipschitz first derivatives, we have that $\ddot{\boldsymbol{x}}$, $\ddot{\boldsymbol{\mu}}$, and $\ddot{\boldsymbol{y}}$ are also bounded over $[0,T]$. Since bounded functions are trivially square integrable, we have that

- $\|x^{(1)}\|_{\mathcal{W}^{2,2}}, \ldots, \|x^{(M)}\|_{\mathcal{W}^{2,2}} < \infty$,
- $\|\mu^{(1)}\|_{\mathcal{W}^{2,2}}, \ldots, \|\mu^{(M)}\|_{\mathcal{W}^{2,2}} < \infty$, and
- $\|y^{(1)}\|_{\mathcal{W}^{2,2}}, \ldots, \|y^{(P)}\|_{\mathcal{W}^{2,2}} < \infty$.

□

Lemma 7. *Let M' be the total dimension of the DAE (i.e., $M' = 2M + P$), and define $\mathcal{W}^{2,2}_M$ as the Hilbert space of functions $(\boldsymbol{x}(\cdot), \boldsymbol{\mu}(\cdot), \boldsymbol{y}(\cdot)) : [0,T] \to \mathbb{R}^{M'}$ equipped with the norm:*

$$\|(\boldsymbol{x}(\cdot), \boldsymbol{\mu}(\cdot), \boldsymbol{y}(\cdot))\|^2_{\mathcal{W}^{2,2}_{M'}} = \textstyle\sum_{m=1}^M \|x^{(m)}\|^2_{\mathcal{W}^{2,2}} + \sum_{m=1}^M \|\mu^{(m)}\|^2_{\mathcal{W}^{2,2}} + \sum_{m=1}^P \|y^{(m)}\|^2_{\mathcal{W}^{2,2}}.$$

Then the set $\mathbb{S}$ is closed in $\mathcal{W}^{2,2}_{M'}$.

Proof. Consider a Cauchy sequence $(\boldsymbol{x}_n, \boldsymbol{\mu}_n, \boldsymbol{y}_n) \in \mathbb{S}$. By Lemma 6, we have that $x_n^{(m)} \in \mathcal{W}^{2,2}$ for all $m \in [1, M]$. By completeness of $\mathcal{W}^{2,2}$, we have that $x_n^{(m)} \to x^{(m)}$ for some $x^{(m)} \in \mathcal{W}^{2,2}$; i.e. for every ϵ, there exists some n' such that $\|y^{(m)}_{n'} - y^{(m)}\|_{\mathcal{W}^{2,2}} < \epsilon$.

$$\begin{aligned}
\sup_{t\in[0,T]} \|x^{(m)}_{n'}(t) - x^{(m)}(t)\|_\infty &= \sup_{t\in[0,T]} \left\langle k(t,\cdot), x^{(m)}_{n'}(t) - x^{(m)}(t) \right\rangle_{\mathcal{W}^{2,2}} \\
&\le \sup_{t\in[0,T]} \|k(t,\cdot)\|_{\mathcal{W}^{2,2}} \|x^{(m)}_{n'}(t) - x^{(m)}(t)\|_{\mathcal{W}^{2,2}} \\
&\le C\|x^{(m)}_{n'}(t) - x^{(m)}(t)\|_{\mathcal{W}^{2,2}} < C\epsilon,
\end{aligned}$$

where $k(\cdot,\cdot)$ is the reproducing kernel associated with $\mathcal{W}^{2,2}$ and C is some universal constant. The penultimate inequality comes from fact that $\mathcal{W}^{2,2}$ is equivalent to a Matérn RKHS which has a bounded-everywhere reproducing kernel. Thus, $x^{(m)}_{n'}$ converges uniformly to $x^{(m)}$, and so

$$\begin{aligned}
\dot{\boldsymbol{x}}(t) = \lim_{n\to\infty} \dot{\boldsymbol{x}}_n(t) &= \lim_{n\to\infty} \boldsymbol{F}(\boldsymbol{x}_n(t), \boldsymbol{\mu}_n(t), \boldsymbol{y}_n(t)) \\
&= \boldsymbol{F}(\lim_{n\to\infty} \boldsymbol{x}_n(t), \lim_{n\to\infty} \boldsymbol{\mu}_n(t), \lim_{n\to\infty} \boldsymbol{y}_n(t)) = \boldsymbol{F}(\boldsymbol{x}(t), \boldsymbol{\mu}(t), \boldsymbol{y}(t))
\end{aligned}$$

where the penultimate equality comes from the continuity of $\boldsymbol{F}$. Analogous results hold for $\boldsymbol{\mu}_n$ and $\boldsymbol{y}_n$. Moreover, by uniform convergence we have that $\|\boldsymbol{\mu}(0) - \boldsymbol{\mu}_n(0)\|_\infty$ and $\|\boldsymbol{y}(0) - \boldsymbol{y}_n(0)\|_\infty$ are arbitrarily small and thus $\|\boldsymbol{\mu}(0)\|_\infty, \|\boldsymbol{y}(0)\|_\infty < K$. Therefore $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{y}) \in \mathbb{S}$. □

Lemma 8. *Denote M' and $\mathcal{W}^{2,2}_{M'}$ as in Lemma 7. Let $\mathcal{H}_{M'}$ be the Hilbert space of functions $(\dot{\boldsymbol{x}}(\cdot), \dot{\boldsymbol{\mu}}(\cdot), \dot{\boldsymbol{y}}(\cdot)) : [0,T] \to \mathbb{R}^{M'}$ equipped with the norm*

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

where $\mathcal{H}$ is the RKHS associated with the Matérn-1/2 kernel with some lengthscale $0 < \ell < \infty$. Let

$$B := \inf_{(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\in\mathbb{S}} \|(\dot{\boldsymbol{x}}, \dot{\boldsymbol{\mu}}, \dot{\boldsymbol{y}})\|_{\mathcal{H}_{M'}}$$

Then there exists some $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*) \in \mathbb{S}$ *that achieves this infimum.*

Proof. By Theorem 4, $\mathcal{W}^{2,2}$ is equivalent to the Matérn-1/2 RKHS $\mathcal{H}$. Note that the elements of $\dot{\boldsymbol{w}}(\cdot)$ for any $\boldsymbol{w}(\cdot) \in \mathcal{W}^{2,2}_{M'}$ are $\mathcal{W}^{1,2}$ functions, and thus

$$\boldsymbol{w}(\cdot) \in \mathcal{W}^{2,2}_{M'} \implies \dot{\boldsymbol{w}}(\cdot) \in \mathcal{H}_{M'}.$$

Define the operator $D : \mathcal{W}^{2,2}_{M'} \to \mathcal{H}_{M'}$ as $D(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{y}) = (\dot{\boldsymbol{x}}, \dot{\boldsymbol{\mu}}, \dot{\boldsymbol{y}})$, where here $\dot{\boldsymbol{x}}, \dot{\boldsymbol{\mu}}, \dot{\boldsymbol{y}}$ denote weak derivatives. Note that D is a surjective and bounded linear operator between two Hilbert spaces, and that the nullspace of D (i.e. the set of constant functions) is a closed set. Since $\mathbb{S} \subset \mathcal{W}^{2,2}_{M'}$ is a closed subset of a Hilbert space (Lemma 7), $\dot{\mathbb{S}} := D(\mathbb{S}) \subset \mathcal{W}^{2,2}_{M'}$ is also closed subset of a Hilbert space (see e.g. Brezis and Brézis, 2011, Exercise 2.10). By the existence portion of the Hilbert projection theorem, there exists a (potentially non-unique) minimum $\mathcal{H}_{M'}$-norm element of $\dot{\mathbb{S}}$ (i.e. there exists some $(\dot{\boldsymbol{x}}^*, \dot{\boldsymbol{\mu}}^*, \dot{\boldsymbol{y}}^*) \in \dot{\mathbb{S}}$ such that $\|(\dot{\boldsymbol{x}}^*, \dot{\boldsymbol{\mu}}^*, \dot{\boldsymbol{y}}^*)\|_{\mathcal{H}_{M'}} = \inf_{(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y})\in\dot{\mathbb{S}}} \|(\dot{\boldsymbol{x}}, \dot{\boldsymbol{\mu}}, \dot{\boldsymbol{y}})\|_{\mathcal{H}_{M'}}$. We conclude the proof by setting $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$ to be some element in $\mathbb{S}$ such that $(\dot{\boldsymbol{x}}^*, \dot{\boldsymbol{\mu}}^*, \dot{\boldsymbol{y}}^*) = D(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$. □

Lemma 9. *For any* $0 < C < \infty$*, define the sets*

$$\mathbb{F}^C := \{w \in \mathcal{H} \ : \ \|w\|_{\mathcal{H}} \leq C\}$$
$$\textstyle\int \mathbb{F}^C := \{\int_0^{(\cdot)} w(\tau)d\tau \ : \ w \in \mathbb{F}\}.$$

Denoting $\widehat{\mathcal{R}}_N$ *as the empirical Rademacher complexity for some dataset* $t_1, \ldots, t_N \in [0, T]$*, we have that*

$$\widehat{\mathcal{R}}_N(\mathbb{F}^C) \lesssim CN^{-1/2}, \qquad \widehat{\mathcal{R}}_N(\textstyle\int \mathbb{F}^C) \lesssim TCN^{-1/2}.$$

Proof. The Rademacher complexity $\widehat{\mathcal{R}}_N(\mathbb{F}^C) \lesssim CN^{-1/2}$ follows a standard result for reproducing kernel Hilbert spaces, using the fact that $\mathcal{H}$ (the Matérn-1/2 RKHS) has a bounded-everywhere reproducing kernel. Bounding the Rademacher complexity of $\int \mathbb{F}^C$ mirrors the standard proof of

the $\widehat{\mathcal{R}}_N(\mathbb{F}^C)$ bound:

$$
\begin{aligned}
\widehat{\mathcal{R}}(\textstyle\int \mathbb{F}^C) &:= \mathbb{E}_{\epsilon_i}\left[\sup_{w\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N}\epsilon_i \int_0^{t_i} w(\tau)d\tau\right] && (\epsilon_i \overset{\text{i.i.d.}}{\sim} \text{Rad}) \\
&= \mathbb{E}_{\epsilon_i}\left[\sup_{w\in\mathcal{H}} \left\langle \frac{1}{N}\sum_{i=1}^{N}\epsilon_i \int_0^{t_i} k(\tau,\cdot)d\tau,\ w(\cdot)\right\rangle_{\mathcal{H}}\right] \\
&\le \mathbb{E}_{\epsilon_i}\left[\left\|\frac{C}{N}\sum_{i=1}^{N}\epsilon_i \int_0^{t_i} k(\tau,\cdot)d\tau\right\|_{\mathcal{H}}\right] && \text{(Cauchy-Schwarz inequality)} \\
&\le \sqrt{\mathbb{E}_{\epsilon_i}\left[\left\|\frac{C}{N}\sum_{i=1}^{N}\epsilon_i \int_0^{t_i} k(\tau,\cdot)d\tau\right\|_{\mathcal{H}}^2\right]} && \text{(Jensen inequality)} \\
&= \sqrt{\mathbb{E}_{\epsilon_i}\left[\frac{C^2}{N^2}\sum_{i,j=1}^{N}\epsilon_i\epsilon_j \left\langle \int_0^{t_i} k(\tau,\cdot)d\tau,\ \int_0^{t_j} k(\tau,\cdot)d\tau\right\rangle_{\mathcal{H}}\right]} \\
&= \sqrt{\frac{C^2}{N^2}\sum_{i}^{N}\left\|\int_0^{t_i} k(\tau,\cdot)d\tau\right\|_{\mathcal{H}}^2} && (\epsilon_i \text{ are uncorrelated}) \\
&\le \sqrt{\frac{C^2}{N^2}\sum_{i}^{N}\left(\int_0^{t_i} \|k(\tau,\cdot)\|_{\mathcal{H}}\,d\tau\right)^2} && \text{(triangle inequality)} \\
&\le \sqrt{\frac{C^2}{N^2}\sum_{i}^{N}\left(\int_0^{T} \sup_{t\in[0,T]}\|k(t,\cdot)\|_{\mathcal{H}}\,d\tau\right)^2} \\
&= TCN^{-1/2}\sup_{t\in[0,T]}\|k(t,\cdot)\|_{\mathcal{H}}\,.
\end{aligned}
$$

Recognizing that $k(t,\cdot)$ is a bounded-everywhere reproducing kernel completes the proof. □

Lemma 10. *Define $\mathcal{H}_{M'}$ as in Lemma 8. For any $0 < C < \infty$, define the sets*

$$
\begin{aligned}
\mathbb{F}^C_{M'} &:= \{(\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}}) \in \mathcal{H}_{M'} \ :\ \|(\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}})\|_{\mathcal{H}_D} \le C\} \\
\textstyle\int \mathbb{F}^C_{M'} &:= \left\{\textstyle\int_0^{(\cdot)}(\dot{\boldsymbol{x}}(\tau),\dot{\boldsymbol{\mu}}(\tau),\dot{\boldsymbol{y}}(\tau))d\tau \ :\ (\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}}) \in \mathbb{F}\right\}
\end{aligned}
$$

Then $\widehat{\mathcal{R}}_N(\mathbb{F}^C_{M'}) \lesssim CMN^{-1/2}$ and $\widehat{\mathcal{R}}_N(\int \mathbb{F}^C_{M'}) \lesssim TMCN^{-1/2}$.

Proof. The proof follows a standard summation argument for Rademacher complexity:

$$
\begin{aligned}
&\widehat{\mathcal{R}}(\mathbb{F}^C_{M'}) \\
:=&\mathbb{E}\left[\sup_{(\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}})\in\mathcal{H}_{M'}} \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j_x=1}^{M}\epsilon_{ij_x}\dot{x}^{(j_x)}(t_i)+\sum_{j_y=1}^{M}\epsilon_{ij_y}\dot{y}^{(j_y)}(t_i)+\sum_{j_z=1}^{M}\epsilon_{ij_z}\dot{z}^{(j_z)}(t_i)\right)\right] \\
&\qquad\qquad (\epsilon_{ij_x},\epsilon_{ij_y},\epsilon_{ij_z} \overset{\text{i.i.d.}}{\sim} \text{Rad}) \\
\leq&\sum_{j_x=1}^{M}\mathbb{E}\left[\sup_{\dot{x}^{(j_x)}\in\mathcal{H}}\frac{1}{N}\sum_{i=1}^{N}\epsilon_{ij_x}\dot{x}^{(j_x)}(t_i)\right]+\sum_{j_y=1}^{M}\mathbb{E}\left[\sup_{\dot{y}^{(j_y)}\in\mathcal{H}}\frac{1}{N}\sum_{i=1}^{N}\epsilon_{ij_y}\dot{y}^{(j_y)}(t_i)\right] \\
&+\sum_{j_z=1}^{P}\mathbb{E}\left[\sup_{\dot{z}^{(j_z)}\in\mathcal{H}}\frac{1}{N}\sum_{i=1}^{N}\epsilon_{ij_z}\dot{z}^{(j_z)}(t_i)\right] \\
\lesssim& CMN^{-1/2}, \qquad \text{(Lemma 9)}
\end{aligned}
$$

where the last inequality comes from the fact that $\|(\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}})\|_{\mathcal{H}_{M'}} \leq C$ implies that $\|\dot{\boldsymbol{x}}^{(m)}\|_{\mathcal{H}}$, $\|\dot{\boldsymbol{\mu}}^{(m)}\|_{\mathcal{H}}$, $\|\dot{\boldsymbol{y}}^{(m)}\|_{\mathcal{H}} \leq C$ for all m. An analogous proof holds for $\widehat{\mathcal{R}}(\int \mathbb{F}^C_{M'})$. □

Lemma 11. *Denote $\mathcal{H}_{M'}$ as in Lemma 7. For any $(\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}}) \in \mathcal{W}^{2,2}_M$, $\hat{\boldsymbol{\mu}}_0 \in \mathbb{R}^M$, $\hat{\boldsymbol{y}}_0 \in \mathbb{R}^P$, define the differential equation error function*

$$
\begin{aligned}
e_{\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}},\hat{\boldsymbol{\mu}}_0,\hat{\boldsymbol{y}}_0}(\cdot) &:= \left\| \begin{bmatrix} \dot{\boldsymbol{x}}(\cdot) - \boldsymbol{F}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y}) \\ \dot{\boldsymbol{\mu}}(\cdot) - \boldsymbol{B}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y}) \\ \boldsymbol{H}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y}) \end{bmatrix} \right\|_\infty \\
\boldsymbol{x}(t) := \boldsymbol{x}_0 + \int_0^t \dot{\boldsymbol{x}}(\tau)d\tau, &\quad \boldsymbol{\mu}(t) := \hat{\boldsymbol{\mu}}_0 + \int_0^t \dot{\boldsymbol{\mu}}(\tau)d\tau, \quad \boldsymbol{y}(t) := \hat{\boldsymbol{y}}_0 + \int_0^t \dot{\boldsymbol{y}}(\tau)d\tau,
\end{aligned} \tag{25}
$$

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

$$
\left\{ e_{\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}},\hat{\boldsymbol{\mu}}_0,\hat{\boldsymbol{y}}_0}(\cdot) \;:\; \|(\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}})\|_{\mathcal{W}^{2,2}_M} \leq C, \|\hat{\boldsymbol{\mu}}_0\|_\infty \leq K, \|\hat{\boldsymbol{y}}_0\|_\infty \leq K \right\}.
$$

Then every error function in $\mathbb{G}^{C,K}$ is bounded by some constant $\tilde{C}$ and $\widehat{\mathcal{R}}(\mathbb{G}^{C,K}) \lesssim CN^{-1/2}$.

Proof. Note that each error function in $\mathbb{G}^{C,K}$ is a Lipschitz function ($\|\cdot\|_\infty$), each of which is applied to the summation of two sub-functions:

1. a $(\dot{\boldsymbol{x}}(\cdot),\dot{\boldsymbol{\mu}}(\cdot),\dot{\boldsymbol{y}}(\cdot)) \in \mathcal{H}_{M'}$ with norm less than C (i.e. an element of $\mathbb{F}^C_{M'}$, as defined in Lemma 10), and

2. lipschitz functions ($\boldsymbol{F}$, $\boldsymbol{B}$, $\boldsymbol{H}$) applied to the integral of a vector-valued RKHS function $(\dot{\boldsymbol{x}},\dot{\boldsymbol{\mu}},\dot{\boldsymbol{y}})$ with norm less than C (i.e. a Lipschitz function applied to an element of $\int \mathbb{F}^C_{M'}$, as

defined in Lemma 10).

Boundedness of the error functions falls from the fact $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{y})$ are bounded, $\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{y}}_0$ are bounded, and $\boldsymbol{F}$, $\boldsymbol{B}$, $\boldsymbol{H}$ are continuous (and thus bounded over $[0, T]$). The Rademacher complexity falls from standard Lipschitz and summation rules: $\widehat{\mathcal{R}}(\mathbb{G}^{C,K}) \lesssim \widehat{\mathcal{R}}(\mathbb{F}^C_{M'}) + \widehat{\mathcal{R}}(\int \mathbb{F}^C_{M'}) \lesssim CMN^{-1/2}$. □

Now we are ready to prove Theorem 3.

Proof of Theorem 3. We begin by noting that

$$\begin{aligned} B &:= \inf_{(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y}\in\mathbb{S}} \textstyle\sum_{m=1}^{M} \|\hat{x}^{(m)}\|^2_{\mathcal{H}} + \sum_{m=1}^{M} \|\hat{\mu}^{(m)}\|^2_{\mathcal{H}} + \sum_{m=1}^{P} \|\hat{y}^{(m)}\|^2_{\mathcal{H}} \\ &= \inf_{(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{y}\in\mathbb{S}} \|(\dot{\boldsymbol{x}}, \dot{\boldsymbol{\mu}}, \dot{\boldsymbol{y}})\|_{\mathcal{H}_{M'}} < \infty \end{aligned}$$

is implied by Lemma 6. Let $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$ be some element in $\mathbb{S}$ that achieves this infimum (the existence of which is guaranteed by Lemma 8). We know that $\|(\dot{\hat{\boldsymbol{x}}}_N, \dot{\hat{\boldsymbol{\mu}}}_N, \dot{\hat{\boldsymbol{y}}}_N)\|_{\mathcal{H}_{M'}} \leq \|(\dot{\boldsymbol{x}}^*, \dot{\boldsymbol{\mu}}^*, \dot{\boldsymbol{y}}^*)\|_{\mathcal{H}_{M'}} = B$—since $(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{y}^*)$ satisfies the constraints for Equation (13)—and thus $(\dot{\hat{\boldsymbol{x}}}_N, \dot{\hat{\boldsymbol{\mu}}}_N, \dot{\hat{\boldsymbol{y}}}_N) \in \mathbb{F}^B_{M'}$ (as defined by Lemma 10).

Defining $e_{\dot{\hat{\boldsymbol{x}}}_N, \dot{\hat{\boldsymbol{\mu}}}_N, \dot{\hat{\boldsymbol{y}}}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{y}}_0}(\cdot)$ as in Lemma 11, we have that $e_{\dot{\hat{\boldsymbol{x}}}_N, \dot{\hat{\boldsymbol{\mu}}}_N, \dot{\hat{\boldsymbol{y}}}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{y}}_0}(t_i) = 0$ for each t_i in $\mathcal{D}$. Applying a standard uniform large law argument (e.g. Wainwright, 2019, Thm. 4.2) we have that, for any $\delta > 0$,

$$\begin{aligned} \int_0^T e_{\dot{\hat{\boldsymbol{x}}}_N, \dot{\hat{\boldsymbol{\mu}}}_N, \dot{\hat{\boldsymbol{y}}}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{y}}_0}(\tau) d\tau &= \left| \frac{1}{N} \sum_{i=1}^{N} e_{\dot{\hat{\boldsymbol{x}}}_N, \dot{\hat{\boldsymbol{\mu}}}_N, \dot{\hat{\boldsymbol{y}}}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{y}}_0}(t_i) - \int_0^T e_{\dot{\hat{\boldsymbol{x}}}_N, \dot{\hat{\boldsymbol{\mu}}}_N, \dot{\hat{\boldsymbol{y}}}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{y}}_0}(\tau) d\tau \right| \\ &\leq 2\mathbb{E}_{t_i}\left[\widehat{\mathcal{R}}(\mathbb{G}^{B,K})\right] + \delta \end{aligned}$$

with probability $1 - 2\exp(-\frac{N\delta^2}{8\tilde{C}^2})$ (where $\tilde{C}$ is the constant defined in Lemma 11). Since $\widehat{\mathcal{R}}(\mathbb{G}^{B,K}) \lesssim BMN^{-1/2}$ (Lemma 11), we have that $\int_0^T e_{\dot{\hat{\boldsymbol{x}}}_N, \dot{\hat{\boldsymbol{\mu}}}_N, \dot{\hat{\boldsymbol{y}}}_N, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{y}}_0}(\tau) d\tau \xrightarrow{\text{a.s.}} 0$, which implies that $\lim_{N\to\infty}(\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{\mu}}_N, \hat{\boldsymbol{y}}_N)$ satisfies the differential equation almost everywhere.

Define $(\dot{\boldsymbol{x}}, \dot{\boldsymbol{\mu}}, \dot{\boldsymbol{y}})$ as the continuously-differentiable representative of the function $\lim_{N\to\infty}(\dot{\hat{\boldsymbol{x}}}_N, \dot{\hat{\boldsymbol{\mu}}}_N, \dot{\hat{\boldsymbol{y}}}_N)$ ((see e.g. Brezis and Brézis, 2011, Thm. 8.2)). Since $(\dot{\boldsymbol{x}}, \dot{\boldsymbol{\mu}}, \dot{\boldsymbol{y}})$ is continuously-differentiable and satisfies the differential equation everywhere, it must also be an element of $\mathbb{S}$. All together, this implies that $\|(\dot{\boldsymbol{x}}, \dot{\boldsymbol{\mu}}, \dot{\boldsymbol{y}})\|_{\mathcal{H}_{M'}} \geq B$, and so $\|(\dot{\boldsymbol{x}}, \dot{\boldsymbol{\mu}}, \dot{\boldsymbol{y}})\|_{\mathcal{H}_{M'}} = \|\lim_{N\to\infty}(\dot{\hat{\boldsymbol{x}}}_N, \dot{\hat{\boldsymbol{\mu}}}_N, \dot{\hat{\boldsymbol{y}}}_N)\|_{\mathcal{H}_{M'}} = B$. □

References

Acemoglu, Daron (2008), *Introduction to modern economic growth*. Princeton university press.

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

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

Azinovic, Marlon, Luca Gaegauf, and Simon Scheidegger (2022), "Deep equilibrium nets." *International Economic Review*, n/a (n/a), URL https://onlinelibrary.wiley.com/doi/abs/10.1111/iere.12575.

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

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

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.

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.

Blanchard, Olivier J and Mark W Watson (1982), "Bubbles, rational expectations and financial markets."

Blanchard, Olivier Jean and Charles M Kahn (1980), "The solution of linear difference models under rational expectations." *Econometrica: Journal of the Econometric Society*, 1305–1311.

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

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.

Chen, Guoyuan (2023), "Deep neural network approximations for the stable manifolds of the hamilton-jacobi equations."

Chichilnisky, Graciela (1977), "Nonlinear functional analysis and optimal economic growth." *Journal of Mathematical Analysis and Applications*, 61 (2), 504–520, URL https://www.sciencedirect.com/science/article/pii/0022247X77901342.

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.

Ebrahimi Kahou, Mahdi, Jesús Fernández-Villaverde, Sebastian Gomez, Jesse Perla, and Jan Rosa (2024), "Spooky boundaries at a distance: Exploring transversality and stability with deep learning." *Working Paper*.

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.

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.

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.

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

Hadamard, Jacques and Jacqueline Hadamard (1932), "Le problème de cauchy et les équations aux dérivées partielles linéaires hyperboliques: leçons professées à l'université yale."

Han, Jiequn, Yucheng Yang, and Weinan E (2022), "Deepham: A global solution method for heterogeneous agent models with aggregate shocks."

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.

Hindmarsh, Alan C., Peter N. Brown, Keith E. Grant, Steven L. Lee, Radu Serban, Dan E. Shumaker, and Carol S. Woodward (2005), "SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers." *ACM Transactions on Mathematical Software*, 31 (3), 363–396.

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

Kase, Hanno, Leonardo Melosi, and Matthias Rottner (2022), *Estimating nonlinear heterogeneous agents models with neural networks*. Centre for Economic Policy Research.

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.

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

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

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

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

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

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.

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.

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

Rasmussen, Carl Edward and Christopher Williams (2006), *Gaussian processes for machine learning*, volume 1. MIT Press.

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.

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.

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

Sethi, Suresh P (1973), "Optimal control of the vidale-wolfe advertising model." *Operations research*, 21 (4), 998–1013.

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

Smola, Alexander J and Bernhard Schölkopf (1998), *Learning with kernels*, volume 4. Citeseer.

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.

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.

Vidale, M. L. and H. B. Wolfe (1957), "An operations-research study of sales response to advertising." *Operations Research*, 5 (3), 370–381.

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

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.

Weber, Thomas A (2006), "An infinite-horizon maximum principle with bounds on the adjoint variable." *Journal of Economic Dynamics and Control*, 30 (2), 229–241.

Wendland, Holger (2004), "Scattered data approximation."

Willoughby, Ralph A. (1979), "Solutions of ill-posed problems (a. n. tikhonov and v. y. arsenin)." *SIAM Review*, 21 (2), 266–267, URL https://doi.org/10.1137/1021044.

Supplement to "Solving Models of Economic Dynamics with Ridgeless Kernel Regressions"

Mahdi Ebrahimi Kahou[1], Jesse Perla[2], Geoff Pleiss[3,4]

[1]Department of Economics, Bowdoin College.

[2]Vancouver School of Economics, University of British Columbia.

[3]Department of Statistics, University of British Columbia.

[4]Vector Institute.

This supplement contains the details of the robustness checks on our algorithm and provides additional applications alongside those presented in the paper.

1 Robustness

This section provides robustness checks and an exploration of sample efficiency for the neoclassical growth model.

1.1 Sparse training data and data efficiency

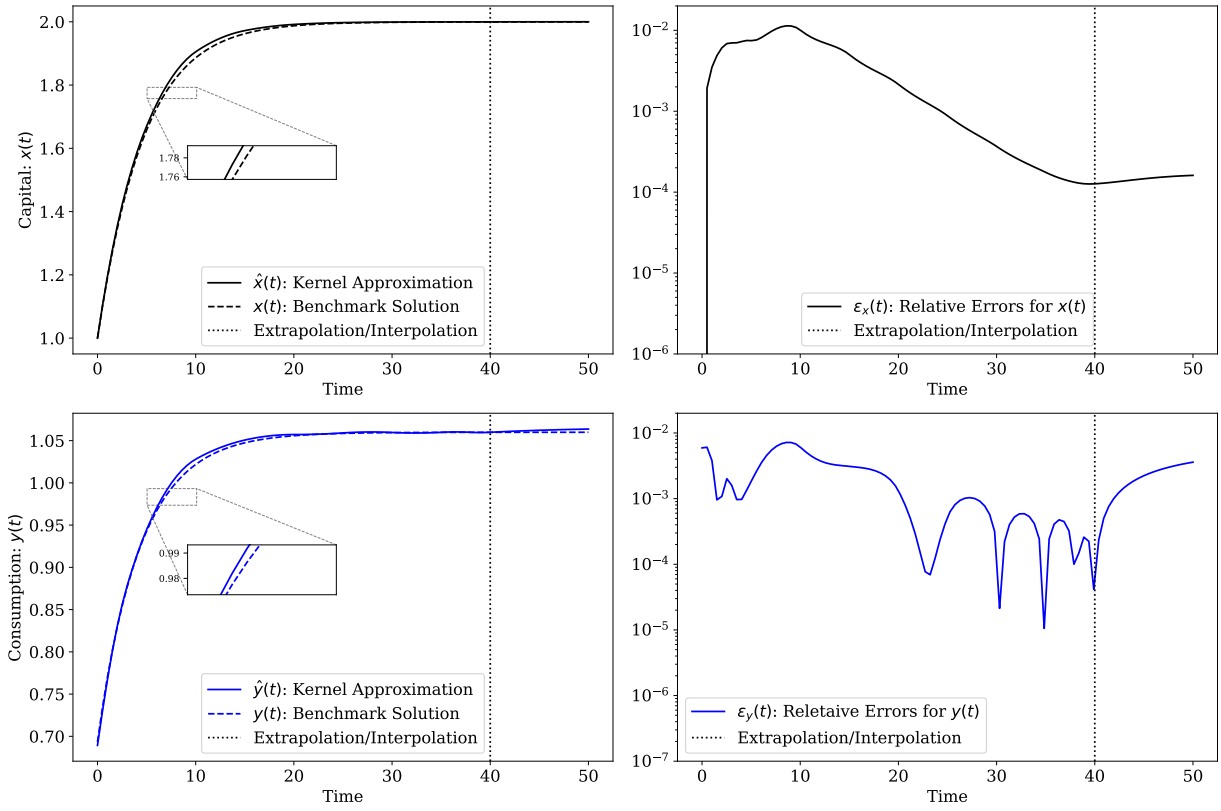

Figure 4: Solution of the neoclassical growth model (Equations (6) to (8) of the main paper) without imposing the transversality condition (Equation (10) of the main paper), for sparse training data $\mathcal{D} := \{0, 1, 3, 5, 10, 15, 20, 25, 30, 35, 38, 40\}$. The top panels show the results for capital, $x(t)$ and associated relative errors, and the bottom two show the results for consumption, $y(t)$. Accurate solutions can be obtained even with a very sparse $\mathcal{D}$.

Since the number of parameters in the optimization grows linearly with the number of grid points, the cardinality of the training set, $\mathcal{D}$, becomes an impediment in higher dimensions with many state and jump variables. Therefore, obtaining accurate approximate solutions with sparse training data is crucial.

Figure 4 shows the result of the neoclassical growth model (i.e., Equations (6) to (8) of the

main paper) solved with $\mathcal{D} := \{0, 1, 3, 5, 10, 15, 20, 25, 30, 35, 38, 40\}$. The top-left panel shows the approximate and benchmark capital paths, denoted by $\hat{x}(t)$ and $x(t)$, respectively. The top-right panel shows the relative errors between the approximate and benchmark solutions for capital, denoted by $\varepsilon_x(t)$. The bottom-left panel shows the approximate and benchmark consumption paths, denoted by $\hat{y}(t)$ and $y(t)$, respectively. The bottom-right panel shows the relative errors between the approximate and benchmark solutions for consumption, denoted by $\varepsilon_y(t)$.

These results show that one can obtain very accurate approximate solution, even with a very sparse training data.

1.2 Robustness to the choice of the kernel and kernel parameters

Table 1 shows the result of the approximate solution of the neoclassical growth model, described in Equations (6) to (8) of the main paper, for different Matérn kernels and kernel parameters.

The first three rows report the performance of the approximate solutions using three different kernels. We present the maximum and minimum absolute values of the relative errors for both the capital path, $\hat{x}(t)$, and the consumption path, $\hat{y}(t)$. The first row shows the baseline solution using the Matérn kernel with $\nu = \frac{1}{2}$. The second and third rows present results for the Matérn kernels with $\nu = \frac{3}{2}$ and $\nu = \frac{5}{2}$, respectively.

ν	ℓ	Max of Rel. Error: $\hat{x}(t)$	Max of Rel. Error: $\hat{y}(t)$	Min of Rel. Error: $\hat{x}(t)$	Min of Rel. Error: $\hat{y}(t)$
1/2	10	1.8e-03	2.9e-03	7.1e-05	9.6e-07
3/2	10	5.9e-04	3.0e-02	1.5e-05	3.4e-06
5/2	10	1.4e-04	2.4e-02	2.7e-05	6.0e-08
1/2	2	3.1e-03	2.8e-03	1.3e-07	5.5e-07
1/2	20	1.9e-03	8.2e-02	7.6e-05	2.9e-05

Table 1: The robustness of the approximate solutions of the neoclassical growth model (i.e., Equations (6) to (8) of the main paper) is tested using different Matérn kernels, $\nu = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}$, and length scales $\ell = 2, 10, 20$.

The last two rows show the performance of the approximate solutions for two different *length scales*, $\ell = 2$, and $\ell = 20$.

Throughout these experiments, we achieve highly accurate approximate solutions. Therefore, the results demonstrate insensitivity to the selection of Matérn kernels and the length scales.

1.3 Smaller time horizons: accurate short-run dynamics

One might suspect that achieving an accurate optimal solution, which does not violate the transversality condition, is only possible if one uses a large time horizon in the training data. For instance, we use $\mathcal{D} = \{0, 1, 2, \ldots, 30\}$ to obtain the results depicted in Figure 1. In this experiment, we establish that we can still achieve accurate short-run dynamics by using a smaller time horizon.

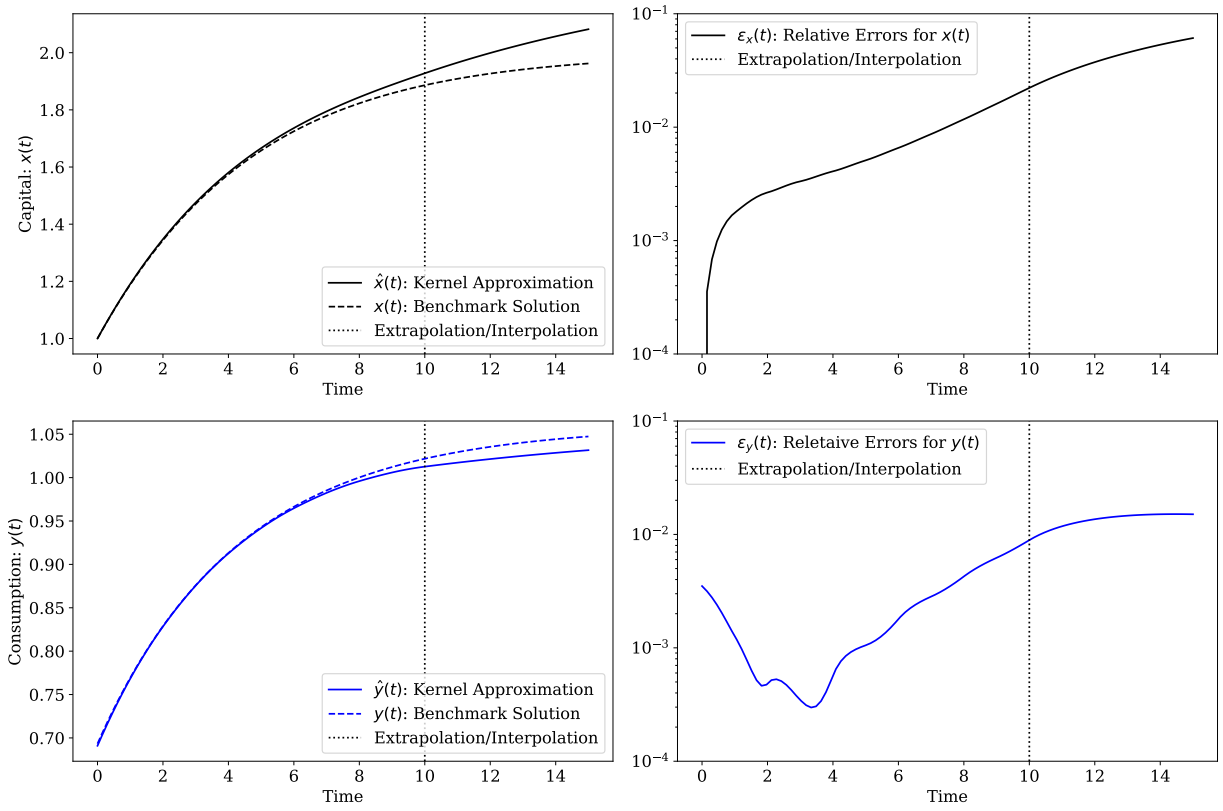

Figure 5: Solution of the neoclassical growth model (Equations (6) to (8) of the main paper) without imposing the transversality condition (Equation (10) of the main paper), for training data with a smaller time horizon $\mathcal{D} := \{0, \cdots, 10\}$. The top panels show the results for capital, $x(t)$ and associated relative errors, and the bottom two show the results for consumption, $y(t)$. Accurate short-run dynamics obtained even a smaller time horizon.

Figure 5 shows the approximate solutions for the neoclassical growth model (i.e., Equations (6) to (8) of the main paper) for training data with smaller time horizon, defined as $\mathcal{D} := \{0, 1, 2, \cdots, 10\}$. The top-left panel shows the approximate and benchmark capital paths, denoted by $\hat{x}(t)$ and $x(t)$, respectively. The top-right panel shows the relative errors between the approximate and benchmark solutions for capital, denoted by $\varepsilon_x(t)$. The bottom-left panel shows the approximate and bench-

mark consumption paths, denoted by $\hat{y}(t)$ and $y(t)$, respectively. The bottom-right panel shows the relative errors between the approximate and benchmark solutions for consumption, denoted by $\varepsilon_y(t)$.

2 More Applications

In this section, we discuss additional applications. In appendix 2.1, we solve a model that incorporates the time evolution of human capital and its interaction with physical capital in economic growth. The model includes seven variables: two state variables along with their co-state variables, and three jump variables. In appendix 2.2, we solve an optimal advertising model based on the work of Sethi (1973).

2.1 Human capital and growth

In this example, we solve the neoclassical growth model with human and physical capital as illustrated in Acemoglu (2008). The optimal paths for the state variables $\boldsymbol{x}(t) := [x_k(t), x_h(t)]$, co-state variables $\boldsymbol{\mu}(t) := [\mu_k(t), \mu_h(t)]$, and jump variables $\boldsymbol{y}(t) := [y_c(t), y_k(t), y_h(t)]$ solve

$$\dot{x}_k(t) = y_k(t) - \delta_k x_k(t), \tag{S1}$$

$$\dot{x}_h(t) = y_h(t) - \delta_h x_h(t), \tag{S2}$$

$$\dot{\mu}_k(t) = r\mu_k(t) - \mu_k(t)\big[f_1(x_k(t),\, x_h(t)) - \delta_k\big], \tag{S3}$$

$$\dot{\mu}_h(t) = r\mu_h(t) - \mu_h(t)\big[f_2(x_k(t),\, x_h(t)) - \delta_h\big], \tag{S4}$$

$$0 = \mu_k(t) y_c(t) - 1, \tag{S5}$$

$$0 = \mu_k(t) - \mu_h(t), \tag{S6}$$

$$0 = f(x_k(t),\, x_h(t)) - y_c(t) - y_k(t) - y_h(t), \tag{S7}$$

with two transversality conditions

$$0 = \lim_{t\to\infty} e^{-rt} x_k(t)\mu_k(t), \tag{S8}$$

$$0 = \lim_{t\to\infty} e^{-rt} x_h(t)\mu_h(t), \tag{S9}$$

for given initial conditions $x_k(0) = x_{k_0}$, $x_h(0) = x_{h_0}$.

The production function is defined as $f\left(x_k(t), x_h(t)\right) = x_k(t)^{a_k} x_h(t)^{a_h}$. Here, $f_1\left(\cdot,\cdot\right)$ is the

derivative with respect to the first input and $f_2(\cdot,\cdot)$ is the derivative with respect to the second input. The two constants in the production function, a_k and a_h, are positive numbers, such that $a_k + a_h < 1$. Additionally, $\delta_k > 0$, $\delta_h > 0$, and $r > 0$.

Human capital is denoted by $x_h(t)$, physical capital by $x_k(t)$, consumption by $y_c(t)$, investment in human capital by $y_h(t)$, and investment in physical capital by $y_k(t)$. Here $\mu_k(t)$ and $\mu_h(t)$ are the co-state variables.

This problem is more challenging than the neoclassical growth model introduced in Section 4 of the main paper because it has twice the number of state and co-state variables, and adds additional algebraic equations, Equations (S6) and (S7). Furthermore, the dynamics become coupled through no-arbitrage conditions between investment in physical and human capital.

For an arbitrary initial condition, this formulation leads to a time-zero discontinuity where $x_k(0+\epsilon)$ and $x_h(0+\epsilon)$ jump to the solution manifold. We choose $x_k(0)$ solve for a $x_h(0)$ consistent the no-arbitrage $f_h(k(\epsilon), h(\epsilon)) - \delta_h = f_k(k(\epsilon), h(\epsilon)) - \delta_k$. The dynamics show that even in cases with more challenging coupling, kernel methods can find consistent solutions without imposing the transversality conditions.

In this experiment we, use $\delta_k = 0.1$, $\delta_h = 0.05$, $\alpha_k = \frac{1}{3}$, $\alpha_h = \frac{1}{4}$, $r = 0.11$, $x_{k_0} = 1.5$, and $x_{h_0} = 1.37$ as the numerical values for the economic parameters. We use $\mathcal{D} = \{0, 1, \cdots, 80\}$ as the training data. In order to stabilize the solution given the coupling of the differential equations, we also found it necessary to add to the objective $\lambda_p \times (||i_k||_{\mathcal{H}} + ||i_h||_{\mathcal{H}} + ||c||_{\mathcal{H}})$ with $\lambda_p = 5 \times 10^{-3}$.

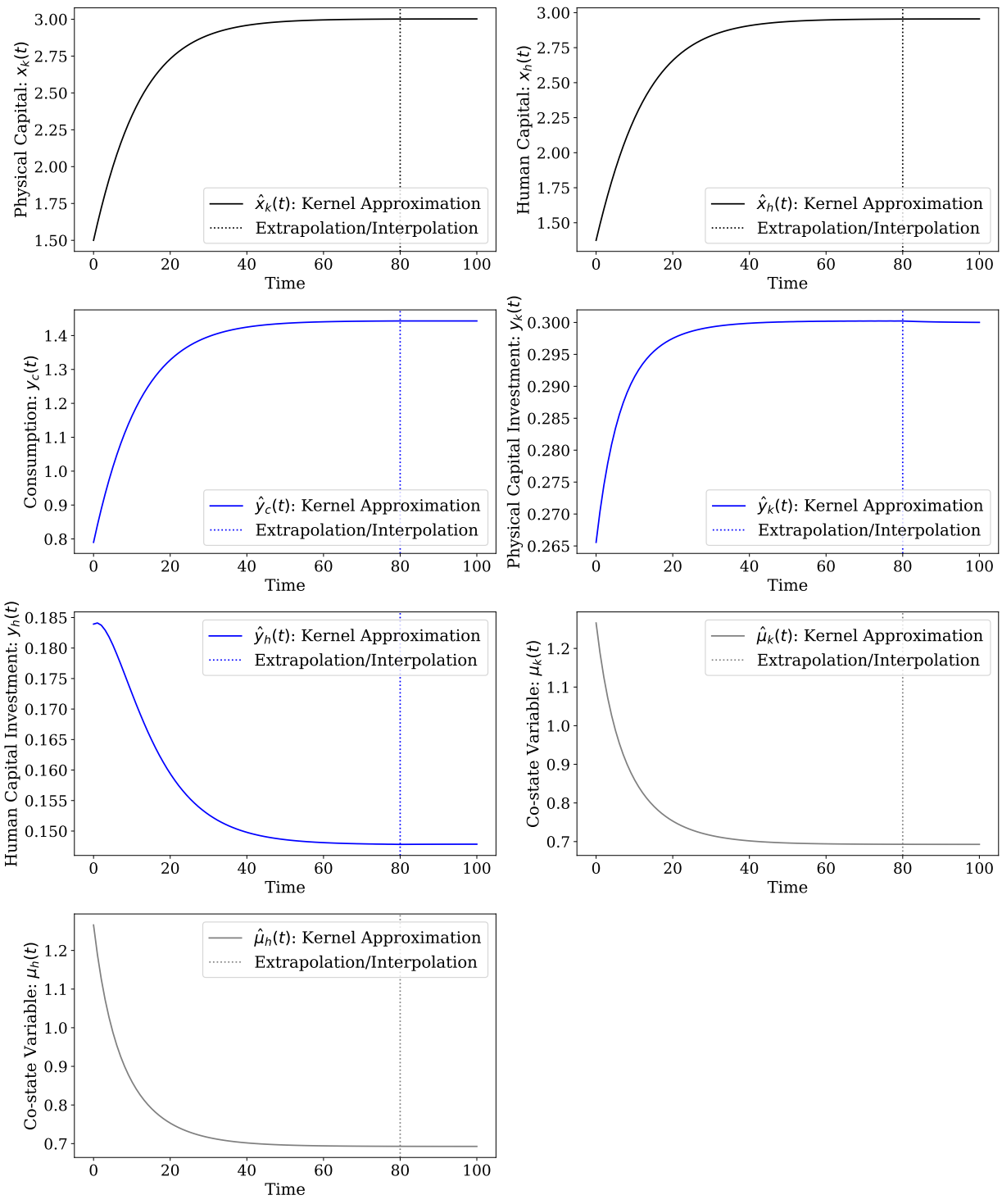

Figure 6: Solution of the growth model with human capital (see Equations (S1) to (S7)) without imposing the transversality conditions. A non-explosive approximate solution is obtained; as shown, all paths converge to the correct steady states.

Figure 6 shows the approximate paths of physical capital ($\hat{x}_k(t)$), human capital ($\hat{x}_h(t)$), consumption ($\hat{y}_c(t)$), investment in physical capital ($\hat{y}_k(t)$), and investment in human capital ($\hat{y}_h(t)$), along with the co-state variables ($\hat{\mu}_k(t)$ and $\hat{\mu}_h(t)$). The vertical dashed line shows the boundary between the interpolation and extrapolation regions.

The correct set of steady states. How do we know the approximate solutions converge to the correct set of steady states? A set of solutions that violates the transversality conditions is characterized by paths such that $\lim_{t\to\infty} e^{-rt}\mu_k(t)x_k(t) \neq 0$ and $\lim_{t\to\infty} e^{-rt}\mu_h(t)x_h(t) \neq 0$. Since the approximate solutions shown in Figure 6 do not exhibit this behavior, we can be confident that they converge to the correct set of steady states.

2.2 Optimal Advertising

In this example we solve an optimal advertising model based on the classical model of expenditure on advertising introduced in (Vidale and Wolfe, 1957). The optimal paths $\boldsymbol{x}(t)$, $\boldsymbol{y}(t)$, and $\boldsymbol{\mu}(t)$, solve

$$\dot{x}(t) = [1 - x(t)]\, y(t) - \beta x(t), \tag{S10}$$

$$\dot{\mu}(t) = r\mu(t) - \gamma + \beta\mu(t) + \mu(t)y(t) \tag{S11}$$

$$0 = y(t)^{\frac{1-\kappa}{\kappa}} - \kappa\mu(t)\,[1 - x(t)] \tag{S12}$$

for a given initial condition $\boldsymbol{x}(0) = \boldsymbol{x}_0$, and a transversality condition

$$\lim_{t\to\infty} e^{-rt}x(t)\mu(t) = 0, \tag{S13}$$

x represents the market share of the company, y is a variable corresponding to advertising expenditure, and μ is the co-state variable.

The parameter κ is a constant between 0 and 1, β is strictly positive, r is the discount rate, the constant γ is defined as $\gamma := \frac{\beta+r}{c}$, and c is the cost of advertising. See (Weber, 2006; Sethi, 1973) for a detailed treatment of this problem.

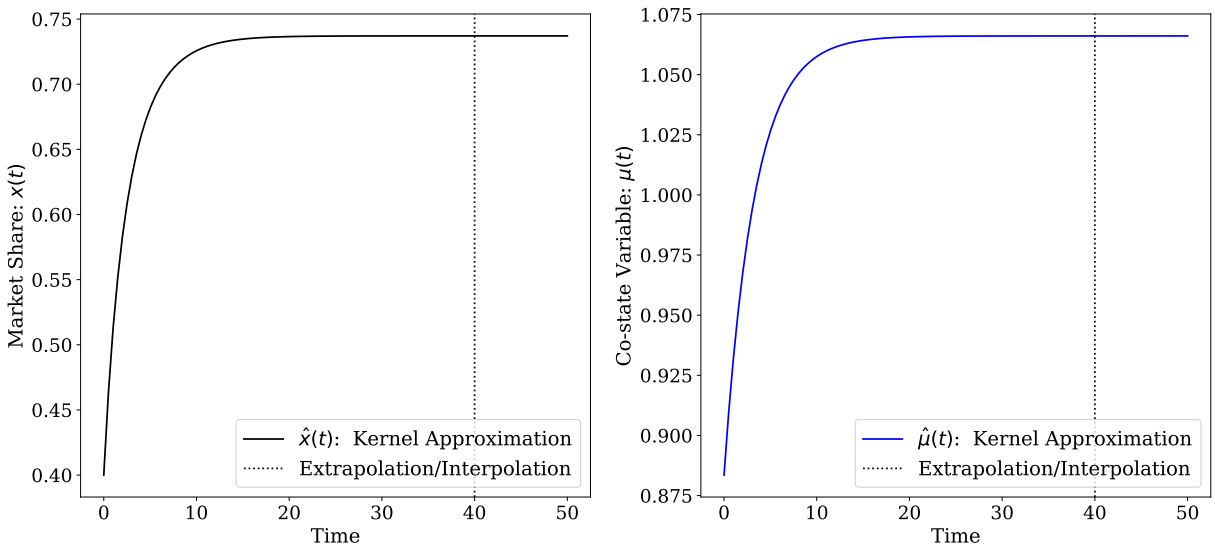

Figure 7: Solution of the optimal advertising model (i.e., Equations (S10) to (S12)) without imposing the transversality condition.

In this example we use $x_0 = 0.4$, $r = 0.11$, $c = 0.5$, $\beta = 0.05$, $\kappa = 0.5$, and $\gamma = \frac{\beta+r}{c} = 0.32$ as the numerical values for the parameters. We use $\mathcal{D} = \{0, 1, \cdots, 40\}$ as the training data.

Results. Figure 7 shows the approximate market share, denoted by $\hat{x}(t)$, and the approximate co-state variable, denoted by $\hat{\mu}(t)$, obtained using a Matérn kernel with $\nu = \frac{1}{2}$. The vertical dashed line indicates the boundary between the interpolation and extrapolation regions. Despite not applying the transversality condition (Equation (S13)), the kernel approximation accurately recovers the optimal solution.

The correct set of steady states. How do we know the approximate solution recovers the optimal solution? As shown in Figure 7, the approximate state, and co-state variable approach a finite number. Therefore, $\lim_{t\to\infty} e^{-rt}\hat{\mu}(t)\hat{x}(t) = 0$.