Additional Material for "Solving Models of Economic Dynamics with Ridgeless Kernel Regressions"

MAHDI EBRAHIMI KAHOU
Department of Economics, Bowdoin College

JESSE PERLA
Vancouver School of Economics, University of British Columbia

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

This document collects additional material that is not part of the paper or its supplement.

1. Human Capital and Growth

In this example, we solve the neoclassical growth model with human and physical capital following 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

$$\dot{x}_k(t) = y_k(t) - \delta_k x_k(t), \tag{E1}$$

$$\dot{x}_h(t) = y_h(t) - \delta_h x_h(t), \tag{E2}$$

$$\dot{\mu}_k(t) = r\mu_k(t) - \mu_k(t)\big[f_1(x_k(t), x_h(t)) - \delta_k\big], \tag{E3}$$

$$\dot{\mu}_h(t) = r\mu_h(t) - \mu_h(t)\big[f_2(x_k(t), x_h(t)) - \delta_h\big], \tag{E4}$$

$$0 = \mu_k(t) y_c(t) - 1, \tag{E5}$$

$$0 = \mu_k(t) - \mu_h(t), \tag{E6}$$

$$0 = f(x_k(t), x_h(t)) - y_c(t) - y_k(t) - y_h(t), \tag{E7}$$

with two transversality conditions

$$0 = \lim_{t\to\infty} e^{-rt} x_k(t)\mu_k(t), \tag{E8}$$

$$0 = \lim_{t\to\infty} e^{-rt} x_h(t)\mu_h(t), \tag{E9}$$

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

Production is $f(x_k, x_h) = x_k^{a_k} x_h^{a_h}$ with $a_k, a_h > 0$, $a_k + a_h < 1$, and $\delta_k, \delta_h, r > 0$, and f_1 and f_2 are its partial derivatives.

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 $y_k(t)$.

This problem is more challenging than the neoclassical growth model of Main Paper Section 4.1 because it has twice the number of state and co-state variables, and adds the algebraic equations Equations (E6) and (E7). Furthermore, the dynamics become coupled through no-arbitrage conditions between investment in physical and human capital.

The algebraic equations restrict the admissible initial stocks: differentiating Equation (E6) along a solution and substituting Equations (E3) and (E4) yields the no-arbitrage condition $f_1 - \delta_k = f_2 - \delta_h$, which must already hold at $t = 0$. With finite controls the states cannot jump, so arbitrary pairs $(x_k(0), x_h(0))$ admit no smooth solution. We therefore choose $x_k(0)$ and solve for the consistent $x_h(0)$ from $f_2(x_k(0), x_h(0)) - \delta_h = f_1(x_k(0), x_h(0)) - \delta_k$.

In this experiment we use $\delta_k = 0.1$, $\delta_h = 0.05$, $a_k = \frac{1}{3}$, $a_h = \frac{1}{4}$, $r = 0.11$, $x_{k,0} = 1.5$, and the consistent $x_{h,0} \approx 1.37$. We use $\mathcal{D} = \{0, 1, \ldots, 80\}$ as the training data.

As in the multicountry model of Main Paper Section 4.5, $\nabla_{\boldsymbol{y}} \boldsymbol{H}$ is singular (Supplemental Appendix Section 1): Equation (E6) contains no jump variable, and only the sum

$y_k(t) + y_h(t)$ enters Equation (E7), so the split of investment between physical and human capital is not determined by the algebraic block pointwise. Differentiating the no-arbitrage condition once more gives a linear restriction on (y_k, y_h) that, together with Equation (E7), pins down the split along any exact solution. The collocation residuals do not impose these derived conditions. The minimum-norm solution recovers the consistent split on its own, and on the test grid the solved paths satisfy $f_1 - \delta_k = f_2 - \delta_h$ to within 9.8×10^{-8}.

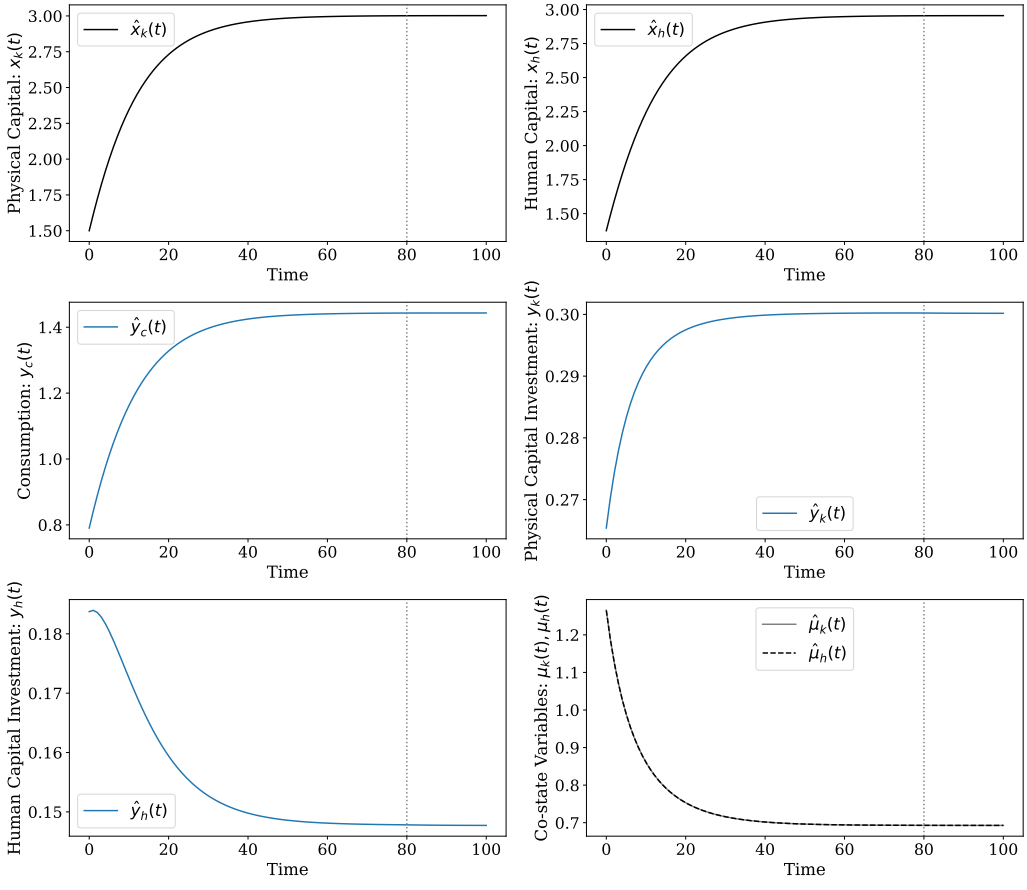


FIGURE 1. Solution of the growth model with human capital (see Equations (E1) to (E7)) without imposing the transversality conditions. The paths converge to the steady state.

Figure 1 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 dotted line shows the boundary between the interpolation and extrapolation regions. Even with this coupling, the kernel method finds a consistent solution without imposing the transversality conditions.

Transversality. The approximate state and co-state paths in Figure 1 settle at finite values, so the discounted products in Equations (E8) and (E9) vanish since $r > 0$.

2. Optimal Advertising

We solve an optimal advertising model based on Vidale and Wolfe (1957). The optimal paths $x(t)$, $y(t)$, and $\mu(t)$ solve

$$\dot{x}(t) = [1 - x(t)]\, y(t) - \beta x(t), \tag{E10}$$

$$\dot{\mu}(t) = r\mu(t) - \gamma + \beta\mu(t) + \mu(t)y(t), \tag{E11}$$

$$0 = y(t)^{\frac{1-\kappa}{\kappa}} - \kappa\mu(t)\,[1 - x(t)] \tag{E12}$$

for a given initial condition $x(0) = x_0$ and the transversality condition

$$\lim_{t\to\infty} e^{-rt} x(t)\mu(t) = 0. \tag{E13}$$

Here x is the market share of the company, y is the transformed advertising control, and μ is the co-state variable. Equations (E10) to (E12) are the necessary conditions of the problem

$$\max_{y(\cdot)} \int_0^\infty e^{-rt} \left[\gamma\, x(t) - y(t)^{1/\kappa}\right] \mathrm{d}t$$

subject to Equation (E10). The current-value Hamiltonian $\gamma x - y^{1/\kappa} + \mu\,[(1-x)y - \beta x]$ delivers Equation (E11) as its co-state equation and Equation (E12) as the first-order condition for y.

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

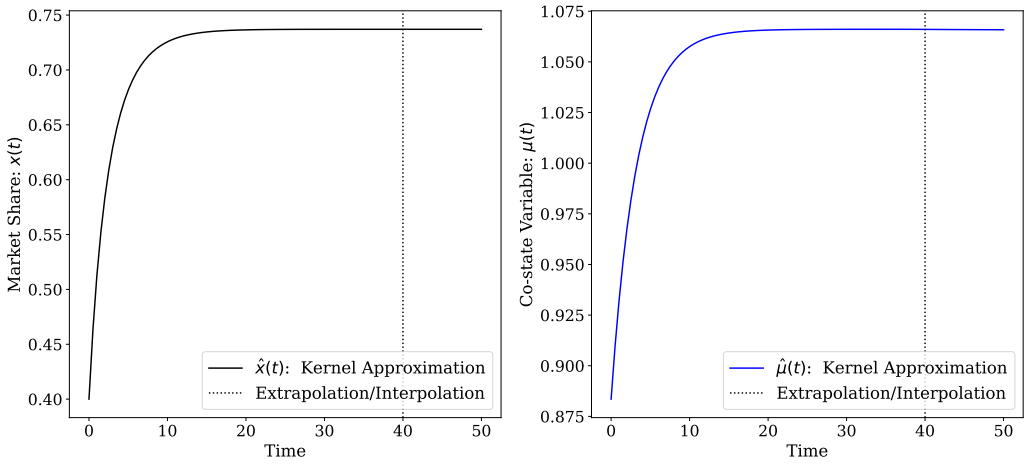


FIGURE 2. Solution of the optimal advertising model (Equations (E10) to (E12)) without imposing the transversality condition in Equation (E13).

In this example we use $x_0 = 0.4$, $r = 0.11$, $c = 0.5$, $\beta = 0.05$, $\kappa = 0.5$, and $\gamma = \frac{\beta+r}{c} = 0.32$. We use $\mathcal{D} = \{0, 1, \ldots, 40\}$ as the training data with lengthscale $\ell = 15$.

Results. Figure 2 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 dotted line indicates the boundary between the interpolation and extrapolation regions. Despite not imposing the transversality condition in Equation (E13), the kernel approximation converges to the steady state.

Transversality. The approximate state and co-state paths in Figure 2 settle at finite values, consistent with Equation (E13) since $r > 0$.

3. SEQUENTIAL ROBUSTNESS EXPERIMENTS

These experiments complement Supplemental Appendix Sections 2, 4 and 6.1.

3.1 *Equilibrium selection*

In the concave–convex model of Main Paper Section 4.2, initial conditions just above the kink $\bar{x}$ admit a candidate path toward each steady state, both satisfying the differential-algebraic equation. Table 1 solves both candidates for four such x_0 and reports their squared reproducing kernel Hilbert space (RKHS) norms, the welfare of each computed path, and the benchmark value. At every x_0 the high-destination candidate has the smaller norm in all three components, while the low-destination candidate has the higher welfare. The minimum-norm selection ranks candidates by norm, not by welfare, and in the vicinity of the kink the two rankings differ.

TABLE 1. Candidate paths toward the high and low steady states of the concave–convex model from four initial conditions: squared RKHS norms of $\dot{x}$, $\dot{y}$, and $\dot{\mu}$, welfare of the computed path, benchmark value, and the maximum absolute collocation residual of the re-solved low candidate (the high candidate is the solution of Main Paper Figure 2).

	$x_0 = 1.97$		$x_0 = 2.02$		$x_0 = 2.07$		$x_0 = 2.12$	
	High	Low	High	Low	High	Low	High	Low
$\|\dot{x}\|_K^2$	0.0194	0.0718	0.0188	0.0776	0.0183	0.0842	0.0177	0.0915
$\|\dot{y}\|_K^2$	0.00111	0.0945	0.00108	0.0941	0.00104	0.0937	0.00101	0.0934
$\|\dot{\mu}\|_K^2$	0.151	0.485	0.125	0.452	0.104	0.421	0.0862	0.394
Welfare	-6.8871	-6.5805	-6.7207	-6.5052	-6.5607	-6.4314	-6.4065	-6.3590
Benchmark V	-6.8848	-6.5897	-6.7185	-6.5189	-6.5587	-6.4466	-6.4048	-6.3727
Max residual	–	4.2e-08	–	4.2e-08	–	4.3e-08	–	4.4e-08

3.2 *Classical shooting and the horizon*

Supplemental Appendix Section 6.1 argues that the terminal condition of shooting becomes exponentially sensitive to the initial guess as the horizon grows. The neoclassical growth model in Main Paper Section 4.1 is especially well suited to shooting methods since the steady state is known in closed form and there is a single unknown co-state, so shooting reduces to a scalar root-find for c_0, whose location researchers can pin down with strong priors. This section shows that despite the favorable conditions, two standard variations on shooting methods can have significant issues with robustness and sensitivity:

- Bisection first finds a bracket whose endpoints diverge in opposite directions and then repeatedly halves it. It uses only the sign of the terminal miss, the gap between capital at the horizon and its steady-state value, never its value. Therefore, it rests on the same fact as the minimum-norm selection of Main Paper Theorem 1, that non-solutions diverge. However, it is sensitive to the horizon on both sides, since a short T biases the root c_0 and a long T amplifies even machine-precision error in c_0 until the terminal condition cannot be met. Even with a well-tuned T, the deeper limitation is that bisection scales exponentially with the number of co-states. The terminal miss has one sign per co-state, so bracketing the root needs a hyper-rectangle whose number of corners grows exponentially in the dimension.

- Root-finding with Newton or nonlinear least squares uses gradient information instead, which differentiable integrators such as Perla (2026) supply through the solution itself. The advantage is that root-finding can, in principle, grow only linearly with the number of co-states, since automatic differentiation delivers the Jacobian at one extra pass per co-state. But it is more sensitive to T than bisection, since numerical errors accumulate along divergent non-solutions, and it can become increasingly difficult even to find initial conditions where the objective and its Jacobian are numerically stable.

In contrast, the kernel solution of Main Paper Equation (13) controls the norm of the solution directly rather than through a terminal condition and is far less sensitive to the horizon.[1]

Setup. Using the baseline model and parameterization in Main Paper Section 4.1, we eliminate the co-state through $\mu = c^{-\gamma}$ and solve the system $\dot{k} = f(k) - \delta k - c$ and $\dot{c} = \frac{c}{\gamma}\big(f'(k) - r - \delta\big)$, with $k(0) = k_0$ and the unknown $c(0) = c_0$.

Under this parameterization the steady state is $k_{ss} = 3.603$ and $c_{ss} = 1.173$. All examples show transitions from our benchmark $k_0 = 1.0$, where the reference solution has $c_0^* = 0.6718$.

The eigenvalues of the Jacobian at the steady state of this system are the standard tool for analyzing stability and the saddle-path dynamics.[2] For our baseline parameterization, this leads to an "unstable" eigenvalue of $\lambda_u = 0.1467$ and a "stable" eigenvalue of $\lambda_s = -0.1049$. Standard analysis, see Acemoglu (2008), shows that a generic path without transversality imposed escapes the steady state at the rate of the unstable eigenvalue (i.e., $\propto \exp(\lambda_u t)$). Hence, in this case the solutions (i.e., those fulfilling the system of equations as well as transversality) and non-solutions (i.e., those fulfilling the system of equations but not transversality) diverge at approximately the rate λ_u.[3]

In our experiments, to give these methods the best possible chance of success, we analyze their robustness near the known solution c_0^*, drawing guesses uniformly in $c_0^*(1 \pm \rho)$ for different radii $\rho > 0$. For the reference solution $k^*(t)$ and $c^*(t)$, we define the absolute relative errors as $\varepsilon_k(t) := \left|\frac{k(t) - k^*(t)}{k^*(t)}\right|$ and $\varepsilon_c(t) := \left|\frac{c(t) - c^*(t)}{c^*(t)}\right|$.

Bisection. First, we consider an efficient and robust version of the bisection method. The goal is to find a $\underline{\hat{c}}_0$ and $\overline{\hat{c}}_0$ such that $k(T; \underline{\hat{c}}_0) - k_{ss} > 0$ and $k(T; \overline{\hat{c}}_0) - k_{ss} < 0$, so that the root of $k(T; c_0) - k_{ss}$ falls in $[\underline{\hat{c}}_0, \overline{\hat{c}}_0]$. Then the algorithm finds the sign at the midpoint of this region and recursively updates the bracketing interval.

Since it is often difficult to find a bracket, as arbitrary points $\underline{\hat{c}}_0$ and $\overline{\hat{c}}_0$ often have the same sign or are numerically unstable, a standard approach is to start from a point where the calculation is numerically stable and then expand multiplicatively until a bracket is found. In our experiment, we apply this algorithm starting from initial guesses drawn uniformly in $c_0^*(1 \pm \rho)$.[4]

[1] For the kernel method, the large T solutions are much more stable, but even a small T can be relatively accurate. See Supplemental Appendix Section 2.3, which solves this model accurately from a training window of 20 years.

[2] In this standard case, the Jacobian is $\begin{bmatrix} f'(k_{ss}) - \delta & -1 \\ c_{ss} f''(k_{ss})/\gamma & 0 \end{bmatrix}$, where $f'(k_{ss}) - \delta = r$. See Acemoglu (2008).

[3] Given this divergence, and since numerical stability is at the heart of the challenges with shooting methods, we solve the ordinary differential equation (ODE) with an adaptive fifth-order Runge–Kutta solver at relative tolerance 10^{-12} and absolute tolerance 10^{-14}, so that through $T = 100$ the integration error sits below the floating-point error. When solving with Newton's method, this differential equation solver has a well-established sensitivity theory. We use forward-mode automatic differentiation for the best possible numerical accuracy (see Perla (2026) for the standard algorithms and their sensitivity analysis).

[4] Specifically, from the guess $\hat{c}_0$ the bracket expands toward the root in steps that start at a quarter of $\rho\,\hat{c}_0$ and grow by a factor of 1.5 until the sign of $k(T) - k_{ss}$ flips. A path that runs into $k = 0$ before T counts as a guess that is too high. The expansion and the subsequent bisection share a budget of 60 integrations, and the returned c_0 is the end of the final bracket with the smaller terminal miss.

The results of the bisection algorithm for different T with $\rho = 0.3$ are given in Table 2. It shows the relative errors of the initial and terminal conditions as well as the mean relative errors of the transitions. Summarizing the results:

- The method is relatively fast in this one-dimensional problem. However, as discussed above, the cost grows exponentially with the dimension of the co-state.
- When it converges and for T large enough to avoid the bias, the algorithm targets the terminal capital fairly accurately, but from $T = 200$ on it does not converge for any of the random draws of initial conditions.
- The mean relative errors of both $k(t)$ and $c(t)$ are poor for the smaller T (except near T where the algorithm mechanically pins the terminal capital). The method reaches a high degree of accuracy only for a relatively small range of T that is not so high that the path diverges and not so low that the finite horizon biases the root. Finding this T is a central challenge in using shooting methods.

TABLE 2. Bisection shooting by horizon T, from 20 random guesses per horizon within the relative radius 0.3 of c_0^*. Convergence is defined as $|k(T) - k_{ss}|/k_{ss} \leq 10^{-6}$ along the path from the returned c_0. Runtime is the median wall-clock time of the search. The remaining columns are over all 20 searches, which all found a bracket: the median relative terminal miss, the median $\varepsilon_c(0)$ of the returned c_0, and the means over searches of the mean $\varepsilon_k(t)$ and $\varepsilon_c(t)$ at 100 equally spaced dates on $[0, 110]$ and of the mean $\varepsilon_k(t)$ on $[0, T]$.

T	Converged	Runtime (s)	$\lvert k(T) - k_{ss}\rvert/k_{ss}$	$\varepsilon_c(0)$	Mean $\varepsilon_k(t)$, $t \in [0, 110]$	Mean $\varepsilon_c(t)$, $t \in [0, 110]$	Mean $\varepsilon_k(t)$, $t \in [0, T]$
10	20/20	0.0048	1.2e-16	4.5e-02	4.7e+00	7.0e-01	1.6e-01
20	20/20	0.011	1.8e-15	2.3e-03	3.3e+00	5.2e-01	3.3e-02
40	20/20	0.02	3.1e-14	1.2e-05	1.2e+00	2.2e-01	2.1e-03
60	20/20	0.03	2.0e-12	7.7e-08	1.0e-01	2.8e-02	1.8e-04
80	20/20	0.041	1.5e-11	5.0e-10	1.0e-03	3.2e-04	1.7e-05
100	20/20	0.049	1.1e-09	2.8e-12	6.5e-06	2.1e-06	1.6e-06
150	20/20	0.068	6.8e-07	4.9e-13	1.1e-10	4.0e-11	2.8e-08
200	0/20	0.091	8.7e-04	4.9e-13	1.1e-10	4.0e-11	3.4e-05
300	0/20	0.11	7.0e+00	4.9e-13	7.8e-10	2.5e-10	9.9e-01

Newton. Second, we consider Newton's method on the terminal condition $g(c_0) := k(T; c_0) - k_{ss}$. Newton's method for systems of equations updates iterate i as $c_0^{i+1} = c_0^i - \nabla g(c_0^i)^{-1} g(c_0^i)$, and with a one-dimensional co-state the inverse Jacobian is simply $g'(c_0)^{-1}$. We evaluate $g(c_0)$ with the ODE solver described above and its derivative by forward-mode automatic differentiation through the integrator.[5] Since there is no bracket, success depends on the quality of the initial guess, so we draw guesses uniformly in $c_0^*(1 \pm \rho)$ for three radii, $\rho \in \{10^{-8}, 10^{-4}, 10^{-1}\}$.

[5] We solve with a maximum of 50 iterations, but successful solutions typically converge to numerical precision in far fewer steps. An iterate whose path leaves the positive orthant before T ends the run as a failure, since g and its derivative are then undefined. While there are many variations on this method to solve for the root such as damping the Newton's update, the numerical instability of divergent non-solutions are inherently challenging.

The results for each ρ and T are given in Table 3. The accuracy columns match Table 2 but are computed over the converged draws only, since a Newton run that fails returns no c_0. Summarizing the results:

- Where it converges at an appropriate T, Newton is fast and precise. It takes a handful of steps and about a millisecond, and it finds the same root as bisection at each T, so the accuracy columns coincide with those of Table 2.
- Convergence depends on the horizon and the guess together. Since non-solutions diverge from the initial deviation from the true c_0^* at roughly the rate of the unstable eigenvalue, the ρ needed for numerical stability shrinks exponentially in T. For $\rho = 0.1$, the algorithm converges only at horizons with heavy bias (i.e., $T \leq 20$).
- The usable horizons are even narrower than for bisection, and Newton reaches them only from a guess that must be increasingly close to the correct value as T increases. Hence, while in principle the gradient lets Newton scale to higher-dimensional co-states far better than bisection, the need for a good initial guess remains and becomes a significant limitation, since it must be met for each co-state and is highly sensitive to the horizon.

TABLE 3. Newton shooting by radius ρ and horizon T, from 20 random guesses drawn uniformly in $c_0^*(1 \pm \rho)$. Convergence is defined as the run reaching $|k(T) - k_{ss}|/k_{ss} \leq 10^{-6}$ before its path leaves the positive orthant or the step budget is spent. Runtime is the median wall-clock time over all 20 runs, converged or not. The remaining columns are as in Table 2 but over the converged runs only. A dash marks a cell with no converged run.

ρ	T	Converged	Runtime (s)	$\lvert k(T) - k_{ss}\rvert/k_{ss}$	$\varepsilon_c(0)$	Mean $\varepsilon_k(t), t \in [0, 110]$	Mean $\varepsilon_c(t), t \in [0, 110]$	Mean $\varepsilon_k(t), t \in [0, T]$
10^{-8}	10	20/20	0.00051	1.1e-07	4.5e-02	4.7e+00	7.0e-01	1.6e-01
	20	20/20	0.0007	1.1e-10	2.3e-03	3.3e+00	5.2e-01	3.3e-02
	40	20/20	0.00074	3.6e-09	1.2e-05	1.2e+00	2.2e-01	2.1e-03
	60	20/20	0.00086	1.2e-11	7.7e-08	1.0e-01	2.8e-02	1.8e-04
	80	20/20	0.00096	7.3e-09	5.0e-10	1.0e-03	3.2e-04	1.6e-05
	100	20/20	0.0011	5.4e-09	2.8e-12	6.5e-06	2.1e-06	1.6e-06
	150	0/20	0.0075	–	–	–	–	–
10^{-4}	10	20/20	0.0005	1.1e-07	4.5e-02	4.7e+00	7.0e-01	1.6e-01
	20	20/20	0.00069	1.1e-10	2.3e-03	3.3e+00	5.2e-01	3.3e-02
	40	20/20	0.00085	8.2e-11	1.2e-05	1.2e+00	2.2e-01	2.1e-03
	60	10/20	0.0071	4.7e-08	7.7e-08	1.0e-01	2.8e-02	1.8e-04
	80	0/20	0.0075	–	–	–	–	–
	100	0/20	0.0073	–	–	–	–	–
	150	0/20	0.0072	–	–	–	–	–
10^{-1}	10	15/20	0.00054	4.9e-09	4.5e-02	4.7e+00	7.0e-01	1.6e-01
	20	6/20	0.0072	9.8e-08	2.3e-03	3.3e+00	5.2e-01	3.3e-02
	40	0/20	0.0073	–	–	–	–	–
	60	0/20	0.0072	–	–	–	–	–
	80	0/20	0.0074	–	–	–	–	–
	100	0/20	0.0076	–	–	–	–	–
	150	0/20	0.0075	–	–	–	–	–

3.3 *Robustness to kernel smoothness*

Table 4 reruns the multicountry growth model with an endogenous world interest rate of Main Paper Section 4.5 with a Matérn $\nu = \frac{3}{2}$ kernel in place of the baseline $\nu = \frac{1}{2}$, at otherwise identical settings. The identical settings include the number of training points,

the ridge weight, and the stopping tolerance on the scaled residual norm, the three settings rows of Table 4. Under a common γ the world rate, capital, and q_i are invariant to the initial bond distribution, so the largest difference in each across the three bond configurations, the bond-invariance gap, is a measure of how far a solution is from that result. The smoother kernel lowers the validation residuals by about a factor of four, leaves the bond-invariance gaps at the same order, and solves in similar time.

TABLE 4. The bond-configuration experiment of Main Paper Section 4.5 under the baseline Matérn $\nu = 1/2$ kernel and a Matérn $\nu = 3/2$ kernel at identical settings. Stopping atol is the tolerance on the scaled residual norm at which the optimization stops. The first three bond-invariance gaps are the largest differences across the three initial bond configurations in the world rate, capital, and q_i on the test window $[0, 50]$. The fourth is the largest of these differences, over all three variables, on the full training horizon $[0, 160]$. Each residual row reports the largest value across the three configurations.

	Matérn $\nu = 1/2$	Matérn $\nu = 3/2$
Train points	192	192
Stopping atol	1.0e-07	1.0e-07
Ridge	3.0e-13	3.0e-13
Bond-invariance gap: $r^\star$ (test window)	1.2e-04	6.6e-05
Bond-invariance gap: $x_{1,i}$ (test window)	5.1e-07	5.7e-07
Bond-invariance gap: q_i (test window)	1.9e-05	1.7e-05
Bond-invariance gap (train horizon)	2.9e-03	1.9e-03
Max train residual	3.1e-06	2.7e-06
Mean validation residual	2.6e-05	6.2e-06
Max validation residual	2.4e-04	5.7e-05
Baseline solve time (s)	0.77	0.87

4. KERNEL INTERPOLATION OF A KNOWN FUNCTION

This section separates the kernel approximation from the solution of a model. A known function is interpolated in the representation of Main Paper Equation (11), with no differential equation, to show why the fitted path converges to a constant, how the length-scale fails when it is too small or too large, and where the choice between the Matérn and Gaussian kernels matters.

Setup. The function is the logistic curve $x(t) = \frac{x_{ss}}{1+\left(\frac{x_{ss}}{x_0}-1\right)e^{-rt}}$ with $x_0 = 0.1$, $x_{ss} = 1$, and $r = 0.15$, which converges to its true asymptote x_{ss}. It is observed at the 11 dates $t = 0, 4, \ldots, 40$, where it has reached 0.978, and fitted as $\hat{x}(t) = x_0 + \sum_j \alpha_j \tilde{K}(t, t_j)$ with $\tilde{K}(t, t_j) = \int_0^t k(s, t_j)\, ds$, the coefficients being those of minimum RKHS norm $\boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha}$ among all that match the observations. Since each kernel is integrable, the fitted derivative $\sum_j \alpha_j k(t, t_j)$ vanishes far from the dates and the level converges to the learned asymptote $\hat{x}(\infty) := x_0 + \sum_j \alpha_j \tilde{K}(\infty, t_j)$. Figure 3 shows the fit on $[0, 80]$ for the Matérn-$1/2$ and Gaussian kernels at $\ell = 0.25, 5, 100$. Each panel reports the mean absolute error

on the training window and on the extrapolation window, the learned asymptote $\hat{x}(\infty)$, and the condition number of the Gram matrix $\boldsymbol{K}$. Figure 4 shows the kernels centered at $t = 20$, with their integrals. The date 20 is chosen for illustration only.

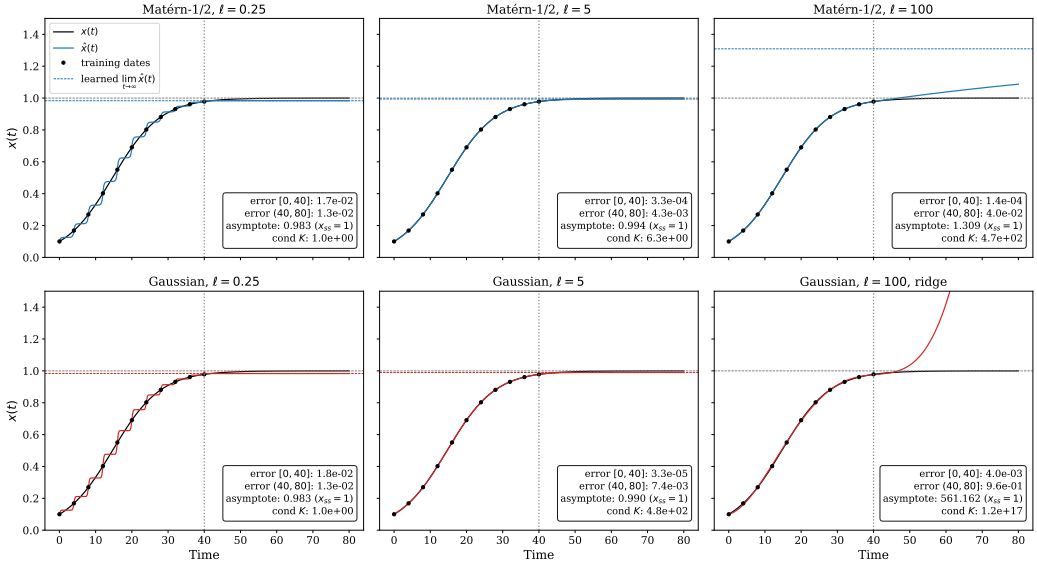


FIGURE 3. Interpolation of the logistic curve at 11 dates with the Matérn-1/2 kernel (top) and the Gaussian kernel (bottom) at three lengthscales. The dashed colored line is the learned asymptote $\hat{x}(\infty)$ and the dashed grey line is the true asymptote x_{ss}. Each panel reports the mean absolute error on the training window and on the extrapolation window, the asymptote, and the computed condition number of the Gram matrix $K(t_i, t_j)$. The Gaussian at $\ell = 100$ is solved with the ridge $\lambda\boldsymbol{\alpha}^\top(\boldsymbol{K} + \epsilon\boldsymbol{I})\boldsymbol{\alpha}$, $\lambda = \epsilon = 10^{-6}$, because its Gram matrix is numerically singular. Every other panel is ridgeless and matches the observations exactly, up to floating-point error.

Length scale about the time scale of the function. At $\ell = 5$ both kernels match the observations exactly, up to floating-point error, track the curve between them accurately, and converge close to the true asymptote $x_{ss} = 1$.

Too small. At $\ell = 0.25$, a sixteenth of the spacing, each kernel is a spike that does not reach its neighbors, so the fitted derivative is zero between dates and the level is a staircase. Each observation is matched at the midpoint of a jump, and the minimum norm spreads each increment over the two adjacent jumps, so each plateau lies between its neighboring observations, close to their average, and the asymptote is the last observation plus half of the last jump. The fit carries no information between dates, and the two kernels give very similar staircases.

Too large. At $\ell = 100$, more than twice the window, the fit is accurate on the window and the difficulty is in the extrapolation. The kernel keeps the slope it has at the last date for a time of order ℓ, long after the function itself has flattened, so the asymptote overshoots x_{ss}.

The Gaussian at $\ell = 100$ fails for a different reason, and the cause is the kernel itself, not numerical precision. Kernels this wide, centered at different dates, are nearly collinear as functions on the training interval, as shown in Figure 4. The Gaussian is far worse than the Matérn-$1/2$ here, because each Matérn-$1/2$ kernel has a cusp at its date, which keeps the kernels distinguishable. This shows in the condition number of the Gram matrix, which is 1.2×10^{17} for the Gaussian, large enough that the matrix is numerically singular, against 4.7×10^{2} for the Matérn-$1/2$.

Fitting a transition of width $1/r$ with kernels of this width and this level of collinearity needs coefficients of enormous size and opposite sign. Across the training interval they very nearly cancel and the fit is accurate, but the cancellation does not survive past the last date, and the extrapolation runs away. That is a property of the minimum-norm interpolant itself. As ℓ grows past the window, interpolation with the Gaussian kernel approaches polynomial interpolation through the dates and extrapolates as a polynomial does.

Because the Gram matrix is numerically singular, this panel alone is solved with a small ridge, $\lambda\boldsymbol{\alpha}^{\top}(\boldsymbol{K} + \epsilon\boldsymbol{I})\boldsymbol{\alpha}$ with $\lambda = \epsilon = 10^{-6}$. Every other panel is ridgeless. With the ridge regularization the fit matches the observations to within a percent but still departs from the curve right after the last date. The ridge sidesteps the singular Gram matrix by solving a regularized least-squares problem instead of exact interpolation, and that selects a different function, but the true curve rises over a span of about $1/r$ and then stays at x_{ss}, far faster variation than a kernel of width $\ell = 100$ represents with small norm, so the ridge cannot recover it and the extrapolation still runs away.

This is the tradeoff noted in Main Paper Section 2: the method replaces the horizon sensitivity of shooting with the conditioning of a Gram matrix, a standard problem with standard remedies, the lengthscale first among them, which a poorly chosen kernel makes severe. Supplemental Appendix Table 1 reports the same pattern on the solved model, where every Matérn solve and the Gaussian at $\ell = 2$ are accepted and the Gaussian at $\ell = 10$ and 20 is rejected on conditioning.

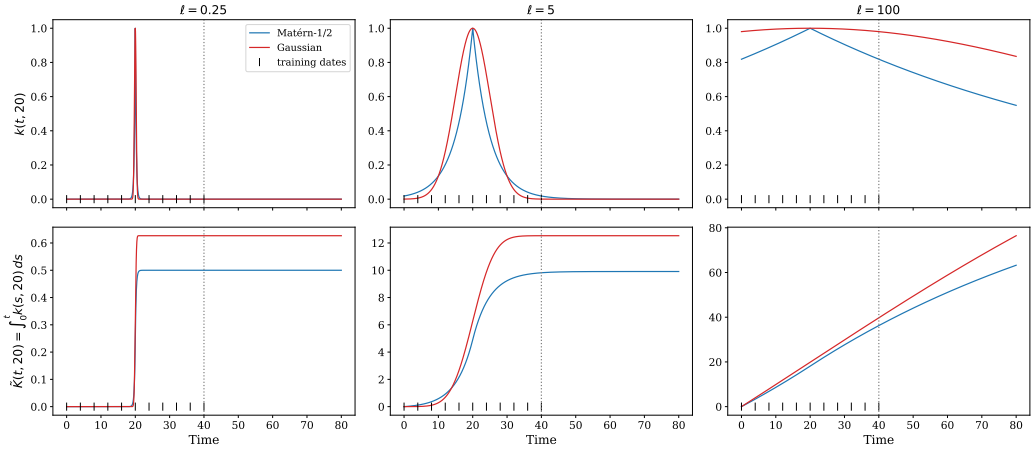


FIGURE 4. The two kernels centered at the date $t = 20$ (top) and their integrals from zero (bottom) at the three lengthscales of Figure 3. Ticks mark the dates.

What this shows about extrapolation. Convergence to an asymptote is by construction for any integrable kernel. In this experiment the asymptote is close to x_{ss} when the lengthscale is large enough for the kernels to track the function between the dates, but not so large that they reach far beyond the window. When the lengthscale is much smaller or much larger, the asymptote is inaccurate. In the solved models of the paper the correct asymptote is not supplied by the kernel but selected by the model equations together with the minimum-norm criterion. By Main Paper Theorem 1, among the trajectories of the model the solution has finite derivative seminorm and every other trajectory, including every one that approaches the wrong level, has infinite seminorm.

5. The Ridge Weight and the Number of Collocation Dates

Main Paper Theorem 3 is about the ridgeless interpolant of Main Paper Equation (13) as the number of dates $N \to \infty$, while the sequential kernel solves minimize the ridge objective of Main Paper Section A with a small weight. This section solves the neoclassical growth model of Main Paper Section 4.1, whose solution is known from the boundary-value benchmark, for a set of ridge weights and numbers of dates and reports how the solution depends on each.

Setup. The objective as implemented is

$$\min_{\boldsymbol{\alpha}^x,\boldsymbol{\alpha}^\mu,\boldsymbol{\alpha}^y,\hat{y}_0,\hat{\mu}_0} \frac{1}{3N}\sum_{t_i\in\mathcal{D}}\left[\text{resource}(t_i)^2+\text{Euler}(t_i)^2+\text{shadow}(t_i)^2\right]+\lambda\sum_{j\in\{x,\mu,y\}}\boldsymbol{\alpha}_j^\top\boldsymbol{K}\boldsymbol{\alpha}_j, \tag{E14}$$

with the three rows the residuals of Main Paper Equations (6) and (7) and of the constant relative risk aversion (CRRA) form of Main Paper Equation (8) stated in Main Paper Section 4.1 at the dates, $\boldsymbol{K}_{ij} = k(t_i, t_j)$ the Gram matrix of the Matérn-1/2 kernel with $\ell = 10$, and the two free initial values $\hat{\mu}_0$ and $\hat{y}_0$ unpenalized. The residual term

is the mean over the $3N$ rows, so λ here is $3N$ times the weight in the unnormalized display of Main Paper Section A. λ is held fixed within each solve. The training window is $T = 100$, as in the paper, and the dates are $t_1 = 0$, where the initial condition is pinned, together with $N - 1$ dates drawn i.i.d. uniformly on $[0, T]$. The ridge weights are $\lambda \in \{10^{-2}, 10^{-4}, \ldots, 10^{-14}\}$ and the numbers of dates $N \in \{11, 21, 51, 101, 201\}$, with 7 independent draws of the dates per N, each solved at every λ from the same initialization with every kernel coefficient zero. Each solve stops on the gradient of Equation (E14) at $\max(10^{-4}\lambda, 10^{-13})$. No acceptance check is applied, and every solve, converged or not, enters the medians. The computed solution at $\lambda = 10^{-10}$ serves as the reference, called the interpolant below. Every number reported in this section, in Table 5, Figure 5, and the text, is the median over the seven draws. The mean relative error of capital is over 100 equally spaced dates on $[0, 110]$ against the benchmark, as in Main Paper Section 4.1.

For fixed N, Equation (E14) is the quadratic-penalty form of the constrained problem Main Paper Equation (13). As $\lambda \to 0$, the global minimizers of Equation (E14) approach the minimum-norm interpolants that solve Main Paper Equation (13). The proof of Main Paper Theorem 3 uses two properties of the solution of Main Paper Equation (13), that its squared RKHS norm is at most B^2, the squared RKHS norm of the true solution, and that the empirical mean of the DAE error at the dates (training points) is zero. Since the true solution is feasible for the ridge version in Main Paper Section A, its minimizer has squared norm at most B^2 and empirical mean error at most $\sqrt{\lambda}\, B$, so the same argument gives consistency along any sequence $\lambda_N \to 0$.

TABLE 5. Median over seven draws of the dates of the mean relative error of $\hat{x}(t)$ against the benchmark, by ridge weight λ (rows) and number of dates N (columns). A count in parentheses is the number of the seven solves that met the gradient tolerance within 300 steps where it is below seven. The medians include every solve.

λ	$N = 11$	$N = 21$	$N = 51$	$N = 101$	$N = 201$
10^{-2}	8.6e-02	7.4e-02	6.8e-02	6.2e-02	6.0e-02
10^{-4}	2.3e-02	1.0e-02	1.6e-03	1.7e-03	1.9e-03
10^{-6}	3.1e-02 (6)	1.7e-02	2.8e-03	8.7e-04 (6)	3.7e-04
10^{-8}	3.1e-02 (6)	1.7e-02	2.8e-03 (6)	9.0e-04	4.0e-04
10^{-10}	3.1e-02 (6)	1.7e-02	2.8e-03	9.3e-04	4.0e-04
10^{-12}	3.1e-02 (6)	1.7e-02	2.8e-03	9.4e-04	4.0e-04
10^{-14}	3.1e-02 (6)	1.7e-02	2.9e-03	9.5e-04	4.0e-04

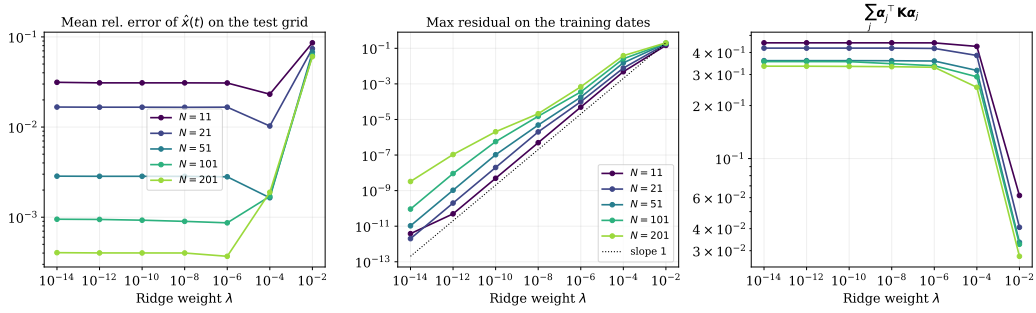


FIGURE 5. Medians over the seven draws against λ, one line per N: the mean relative error of $\hat{x}(t)$ (left), the largest absolute residual of Equation (E14) at the training dates (middle, with a slope-one reference), and the squared RKHS norm $\sum_j \boldsymbol{\alpha}_j^\top \boldsymbol{K}\boldsymbol{\alpha}_j$ of the solution (right).

Results. In the middle panel of Figure 5 the residual at the training dates falls in proportion to λ, while the mean relative error of capital in the left panel and the RKHS norm in the right panel stop moving once $\lambda \leq 10^{-6}$. In line with Main Paper Theorem 3, the mean relative error of capital falls with N across the columns of Table 5, for instance at $\lambda = 10^{-10}$ from 0.031 at $N = 11$ to 4×10^{-4} at $N = 201$. It is important to note that a larger λ introduces bias rather than failure. In Table 5, the error at $\lambda = 10^{-4}$ is below the error at $\lambda = 10^{-10}$ for $N \leq 51$, where the ridge acts as regularization, and above it for $N \geq 101$, where the bias of a fixed λ no longer shrinks with N. The left panel of Figure 5 shows the same dip and rise at $\lambda = 10^{-4}$. This is expected, since the results of Main Paper Section A concern the limit $\lambda \to 0$. Seven of the 245 solves missed the tolerance, marked in Table 5, and they do not move the medians.

5.1 *The ridge weight and the kernel*

The ridge weight, λ, matters more for smoother kernels. Table 6 reports solutions of the model on the 101 equally spaced dates of Main Paper Section 4.1 with the Matérn kernels at $\nu = 1/2, 3/2, 5/2$ and $\ell = 10$ and the Gaussian kernel at $\ell = 2$, the accepted kernels of Supplemental Appendix Table 1. As there, the Gaussian kernel is denoted by $\nu = \infty$, its limit in the Matérn family. Each kernel is solved at $\lambda \in \{10^{-8}, 10^{-10}, 10^{-12}, 10^{-14}\}$, stopping on the gradient of Equation (E14) at $\max(10^{-4}\lambda, 10^{-13})$ or after 300 steps. The table also reports a constrained solve of Main Paper Equation (13) itself, with the collocation equations as equality constraints and the RKHS norm as the objective, by an interior-point method. The $\lambda = 10^{-10}$ rows are the solves of Supplemental Appendix Table 1. In eight of the 16 ridge cases the optimizer reaches the limit of 300 steps before the gradient falls below the tolerance, ending with a gradient of at most 2.1×10^{-11}. These are the cases with the two smoother kernels, whose Gram matrices are too ill-conditioned to reach the gradient of 10^{-13} of the stopping rule. It is worth noting that they are not failures, as their residuals and errors in Table 6 show.

TABLE 6. Ridge solutions at four weights and the constrained minimum-norm solution for each kernel of Supplemental Appendix Table 1, on the 101 equally spaced dates. The columns are the largest absolute residual of Equation (E14) at the training dates, the squared RKHS norm, and the maximum relative errors of $\hat{x}(t)$ and $\hat{y}(t)$ on $[0, 110]$ against the benchmark.

Kernel	λ	Max training residual	$\sum_j \boldsymbol{\alpha}_j^\top \boldsymbol{K}\boldsymbol{\alpha}_j$	Max of Rel. Error: $\hat{x}(t)$	Max of Rel. Error: $\hat{y}(t)$
$\nu = 1/2, \ell = 10$	10^{-8}	3.3e-06	0.3305	1.5e-03	3.1e-03
	10^{-10}	3.3e-08	0.3305	1.5e-03	3.1e-03
	10^{-12}	3.3e-10	0.3305	1.5e-03	3.1e-03
	10^{-14}	3.3e-12	0.3305	1.5e-03	3.1e-03
	constrained	3.1e-10	0.3305	1.5e-03	3.1e-03
$\nu = 3/2, \ell = 10$	10^{-8}	3.5e-05	1.1513	2.3e-04	9.1e-04
	10^{-10}	3.5e-07	1.1535	2.3e-04	1.1e-03
	10^{-12}	3.5e-09	1.1535	2.3e-04	1.2e-03
	10^{-14}	3.5e-11	1.1535	2.3e-04	1.2e-03
	constrained	4.4e-13	1.1535	2.3e-04	1.2e-03
$\nu = 5/2, \ell = 10$	10^{-8}	4.2e-04	4.0291	2.1e-05	4.4e-04
	10^{-10}	4.7e-06	4.2913	3.6e-05	4.1e-04
	10^{-12}	4.7e-08	4.2977	4.0e-04	3.4e-03
	10^{-14}	4.7e-10	4.2979	4.2e-04	3.6e-03
	constrained	5.0e-12	4.2979	4.2e-04	3.6e-03
$\nu = \infty, \ell = 2$	10^{-8}	1.2e-04	0.6022	1.4e-04	2.8e-04
	10^{-10}	4.0e-06	0.6802	9.2e-05	2.3e-04
	10^{-12}	7.9e-08	0.6972	8.8e-04	1.2e-02
	10^{-14}	9.0e-10	0.6991	1.2e-03	1.7e-02
	constrained	9.5e-11	0.6992	1.5e-03	2.1e-02

Results. For the three Matérn kernels ($\nu = 1/2, 3/2, 5/2$) the ridge and constrained solves agree as $\lambda \to 0$, with the maximum relative errors of $\hat{x}(t)$ and $\hat{y}(t)$ converging to the same values. The ridge weight does little for the Matérn-$1/2$ kernel and almost as little for Matérn-$3/2$, and a real dependence on it appears only for the smoother kernels, Matérn-$5/2$ and the Gaussian. This tracks the conditioning of the Gram matrix, whose condition number rises with smoothness from 380 for Matérn-$1/2$ to 1.7×10^8 for the Gaussian. With a nearly singular Gram matrix the ridgeless limit is therefore delicate. The explicit ridge, and any solver device that changes how exactly the collocation equations are enforced, can move the approximation error at large t by an order of magnitude. Even so, the constrained Gaussian solution is still accurate to within a few percent, with a maximum consumption error of 2.1×10^{-2}. The Matérn-$1/2$ kernel of the paper does not have this problem at these settings.

6. SENSITIVITY ANALYSIS OF THE MULTICOUNTRY ECONOMY

6.1 *The Two-Country World Economy*

The implicit-differentiation machinery of Supplemental Appendix Section 7 applies unchanged to the general-equilibrium model of Main Paper Section 4.5, where the world interest rate is an endogenous algebraic path. For the $N = 2$ baseline we differentiate the world rate, each country's consumption, and each country's capital at $T \in \{0, 10, 30, 80\}$, from impact through the late transition, with capital still below the steady state at $T = 30$ ($x_{1,1} = 3.76$, $x_{1,2} = 3.72$) and nearly at it by $T = 80$ ($x_{1,1} = x_{1,2} = 4.68$, against the steady

state $x_1^\star = 4.786$), with respect to the following initial conditions and structural parameters: the initial capital stocks $k_{i,0} = x_{1,i}(0)$, a bond redistribution τ that moves the initial positions to $(b_{1,0} + \tau,\, b_{2,0} - \tau)$ and, with $L_1 = L_2$, stays on the clearing set $\sum_i L_i b_{i,0} = 0$, the curvatures γ_i separately, and (a, δ, r, η) together with country technology scales z_i (production $z_i f\left(x_{1,i}\right)$, with $z_i = 1$ in the baseline). Table 7 reports the $T = 30$ and $T = 80$ slices.

Three theoretical predictions organize the table. First, by the aggregation result of Supplemental Appendix Section 4.3, under the common curvature $\gamma_i = \gamma$ a pure bond redistribution moves neither the world rate nor any capital stock: the τ row of the $r^\star$ and $x_{1,i}$ columns consists of numerical zeros, at most 6.62×10^{-8} at the two tabulated dates. What τ does move is consumption, by equal and opposite amounts in the two countries: $\partial y_{1,1}(T)/\partial\tau = -\partial y_{1,2}(T)/\partial\tau \approx 0.0902$ at $T = 30$, growing to 0.0987 at $T = 80$ along the Euler equation while $r^\star(t) > r$.

Second, the long-run capital stock depends only on $(a, \delta, r, \eta, z_i)$, so the capital sensitivities to the initial conditions and the curvatures γ_i must die out as T grows. Between $T = 30$ and $T = 80$ the $k_{i,0}$ rows of the capital columns shrink by roughly a factor of ten, and the largest γ_i entry, $\partial x_{1,1}(T)/\partial\gamma_1$, moves from -0.168 to -0.0447, while the $(a, \delta, r, \eta, z_i)$ rows remain large because those parameters move $x_1^\star$ itself. The surviving entries again approach the differentiated steady state. At $T = 80$, $\partial x_{1,i}(T)/\partial z_i \approx 6.96$ against the closed-form $\partial x_1^\star/\partial z_i = x_1^\star/\left[(1-a)\, z_i\right] = 7.14$, with the remaining gap reflecting the transition at that horizon.

Third, the rate column $r^\star(T)$ shows the world price responding to anything that moves world savings or productivity, and every row except τ is nonzero at $T = 30$. Every rate sensitivity except $\partial r^\star/\partial r$ decays toward zero by $T = 80$ as $r^\star(t) \to r$. In the long run the interest rate is pinned by discounting alone, and $\partial r^\star(80)/\partial r \approx 0.98$.

Table 8 repeats the computation through the neural network (NN) solution. Across all four saved dates, the largest absolute gap between the two methods is 0.0548, about 0.1% of the derivative's magnitude of 55.

6.2 *The Ten-Country World Economy*

Table 9 reports world aggregates of an $N = 10$ economy at the four saved dates: the world rate, the world saving rate (world investment over world output), the dispersion of net foreign asset (NFA) positions (a population-weighted root mean square of the b_i, divided by steady-state output), and the population-weighted Gini coefficient of output. The middle block gives their semi-elasticities, $100\,\partial \ln(\cdot)/\partial\gamma$, along the common-curvature direction $(1, \ldots, 1)$ (each γ_i is a coefficient of relative risk aversion). At $T = 10$, multiplying the rate semi-elasticity by the rate level and dividing by 100 recovers the plain derivative $\partial r^\star(10)/\partial\gamma$, which Section 6.3 reports again for $N = 40$. Table 10 repeats the computation through the NN solution. The two agree closely, and the only visible differences, in the rate semi-elasticity at $T = 0$ and the output Gini at $T = 80$, are both immaterial, the first about 2% and the second on a Gini level below 10^{-5}.

TABLE 7. Sensitivities of the two-country world economy at $T \in \{30, 80\}$, differentiated through the kernel solution. The entries are the derivatives of the world rate $r^\star(T)$, the consumptions $y_{1,i}(T)$, and the capital stocks $x_{1,i}(T)$ with respect to each initial condition and structural parameter. τ is the clearing-preserving initial bond transfer from country 2 to country 1. z_i scales country i's technology.

			$T = 30$		
θ	$r^\star(T)$	$y_{1,1}(T)$	$y_{1,2}(T)$	$x_{1,1}(T)$	$x_{1,2}(T)$
$k_{1,0}$	-0.00208	0.0891	-0.0414	0.204	0.152
$k_{2,0}$	-0.00253	-0.0706	0.129	0.194	0.238
τ	-3.86×10^{-9}	0.0902	-0.0902	-2.83×10^{-9}	-2.78×10^{-9}
γ_1	0.00282	-0.0599	0.0297	-0.168	-0.167
γ_2	0.00262	0.0345	-0.0625	-0.156	-0.156
a	0.0816	2.98	2.44	18.4	18.3
δ	-0.15	-4.65	-3.78	-26.0	-24.7
r	0.873	-1.31	-2.32	-46.9	-46.8
η	-0.00707	0.45	0.607	5.92	6.5
z_1	0.00362	1.71	0.0702	5.22	-0.263
z_2	0.0044	0.12	1.64	-0.335	5.16
			$T = 80$		
θ	$r^\star(T)$	$y_{1,1}(T)$	$y_{1,2}(T)$	$x_{1,1}(T)$	$x_{1,2}(T)$
$k_{1,0}$	-1.45×10^{-4}	0.0727	-0.0683	0.0187	0.0182
$k_{2,0}$	-1.75×10^{-4}	-0.107	0.113	0.0222	0.0226
τ	5.7×10^{-8}	0.0987	-0.0987	-6.62×10^{-8}	-5.69×10^{-8}
γ_1	4.37×10^{-4}	-0.084	0.075	-0.0447	-0.0446
γ_2	4.07×10^{-4}	0.0801	-0.0886	-0.0415	-0.0415
a	0.011	4.5	3.82	30.7	30.7
δ	-0.0339	-7.97	-6.82	-55.0	-55.0
r	0.98	-3.49	-4.44	-76.4	-76.4
η	-0.00438	0.224	0.415	5.06	5.07
z_1	2.5×10^{-4}	1.92	0.117	6.96	-0.0317
z_2	3.02×10^{-4}	0.184	1.85	-0.0386	6.95

6.3 *The Forty-Country Sensitivity*

At $N = 40$ the preconditioner of Supplemental Appendix Section 4.4 carries the same derivative through the conjugate-gradient solver. The $N = 40$ problem is re-solved from its solution with the ridge weight set to 3×10^{-13} and the preconditioner built once there, with build damping 10^{-12}. The tangent system, with the operator of Supplemental Ap-

TABLE 8. Sensitivities of the two-country world economy at $T \in \{30, 80\}$, differentiated through the NN solution. The kernel counterpart is Table 7.

			$T = 30$		
θ	$r^\star(T)$	$y_{1,1}(T)$	$y_{1,2}(T)$	$x_{1,1}(T)$	$x_{1,2}(T)$
$k_{1,0}$	-0.00208	0.089	-0.0414	0.204	0.152
$k_{2,0}$	-0.00253	-0.0707	0.129	0.194	0.238
τ	1.84×10^{-10}	0.0902	-0.0902	-2.33×10^{-8}	-2.33×10^{-8}
γ_1	0.00282	-0.0599	0.0298	-0.168	-0.167
γ_2	0.00262	0.0345	-0.0625	-0.156	-0.156
a	0.0816	2.98	2.44	18.4	18.3
δ	-0.15	-4.65	-3.78	-26.0	-24.6
r	0.873	-1.31	-2.32	-46.9	-46.8
η	-0.00707	0.45	0.607	5.92	6.5
z_1	0.00363	1.71	0.0702	5.22	-0.263
z_2	0.0044	0.12	1.64	-0.336	5.16
			$T = 80$		
θ	$r^\star(T)$	$y_{1,1}(T)$	$y_{1,2}(T)$	$x_{1,1}(T)$	$x_{1,2}(T)$
$k_{1,0}$	-1.51×10^{-4}	0.0727	-0.0683	0.0188	0.0182
$k_{2,0}$	-1.83×10^{-4}	-0.107	0.113	0.0222	0.0227
τ	-2.85×10^{-9}	0.0987	-0.0987	1.09×10^{-7}	1.09×10^{-7}
γ_1	4.38×10^{-4}	-0.084	0.075	-0.0442	-0.0442
γ_2	4.07×10^{-4}	0.0802	-0.0885	-0.0411	-0.0411
a	0.0111	4.5	3.82	30.7	30.7
δ	-0.034	-7.98	-6.82	-55.1	-55.1
r	0.98	-3.49	-4.44	-76.4	-76.4
η	-0.00433	0.223	0.414	5.05	5.06
z_1	2.63×10^{-4}	1.92	0.117	6.96	-0.0317
z_2	3.19×10^{-4}	0.184	1.85	-0.0386	6.95

pendix Equation (S11) at $\varepsilon = 0$ and right-hand side $-\tilde{\boldsymbol{J}}^\top \partial \boldsymbol{r}/\partial\gamma$, is solved by conjugate gradient (CG) to a relative tolerance of 10^{-8} within 400 iterations using that preconditioner. The derivative of $r^\star(10)$ with respect to a common increase in risk aversion is 0.0123 at $N = 40$, in line with 0.0109 at $N = 2$ and 0.0117 at $N = 10$. Re-solving the model and computing the tangent together take 4.1 seconds, against 16.6 seconds to solve the model afresh. Including compilation these are 4.59 and 16.9 seconds.

TABLE 9. World aggregates of the ten-country economy: levels, semi-elasticities along the common-curvature direction, and the redistribution-neutrality residual, differentiated through the kernel solution.

	$T=0$	$T=10$	$T=30$	$T=80$
Level				
World interest rate $r^\star$	0.131	0.0773	0.0514	0.0417
World saving rate X/Y	0.255	0.219	0.192	0.178
NFA dispersion	0.333	0.645	0.944	1.1
Output Gini	0.0421	0.00679	5.88×10^{-4}	4.4×10^{-6}
Common risk aversion: $100\,\partial\ln(\cdot)/\partial\gamma$ (% per unit γ)				
World interest rate $r^\star$	8.02	15.08	11.15	2.11
World saving rate X/Y	−21.52	−10.41	−2.32	0.15
NFA dispersion	0	0.16	7.44	14.70
Output Gini	0	10.42	5.49	−5.98
Redistribution neutrality				
$\max_i \lvert\partial r^\star/\partial b_{i,0}\rvert$	5×10^{-6}	3.9×10^{-8}	1.2×10^{-8}	1.4×10^{-7}

TABLE 10. World aggregates of the ten-country economy: levels, semi-elasticities along the common-curvature direction, and the redistribution-neutrality residual, differentiated through the NN solution.

	$T=0$	$T=10$	$T=30$	$T=80$
Level				
World interest rate $r^\star$	0.132	0.0773	0.0514	0.0417
World saving rate X/Y	0.255	0.219	0.192	0.178
NFA dispersion	0.333	0.644	0.944	1.1
Output Gini	0.0421	0.00679	5.88×10^{-4}	5.1×10^{-6}
Common risk aversion: $100\,\partial\ln(\cdot)/\partial\gamma$ (% per unit γ)				
World interest rate $r^\star$	8.22	15.08	11.15	2.11
World saving rate X/Y	−21.53	−10.41	−2.34	0.15
NFA dispersion	0	0.15	7.43	14.73
Output Gini	0	10.41	5.49	−4.98
Redistribution neutrality				
$\max_i \lvert\partial r^\star/\partial b_{i,0}\rvert$	1.6×10^{-9}	5.9×10^{-11}	1.1×10^{-10}	4.2×10^{-10}

7. Recursive Variants

This section reports two further solutions of the state-space models of Supplemental Appendix Section 5, with the methods, calibration, and diagnostics of Supplemental Appendix Section 5.3: the deterministic growth model with trend growth in the law of motion of productivity, and an NN solution of the deterministic small open economy. Table 11 collects the results in the format of Supplemental Appendix Table 3; NN entries are medians over five seeds, except that the debt and asset variants of the open economy use one seed.

7.1 *Trend growth in levels*

Starting from the deterministic growth model of Supplemental Appendix Section 5.1, let total factor productivity grow at the trend rate $(1-a)g$ for $g \geq 0$,

$$\mathrm{d}z(t) = (1-a)\, g\, \mathrm{d}t, \qquad \text{so} \qquad e^{z(t)} f(k) = e^{z_0} e^{(1-a)gt} k^a, \tag{E15}$$

which is the Hicks-neutral index consistent with balanced growth of per-capita quantities at rate g: the labor-augmenting index $e^{z/(1-a)}$ grows at g. Nothing else changes. The primitives are Supplemental Appendix Equation (S20) with $\boldsymbol{F} = \big((1-a)g,\ e^z f(k) - \delta k - c\big)$ and $\boldsymbol{B} = \mathbf{0}$, and the residual is Supplemental Appendix Equation (S25) with $\Gamma \equiv 0$. The system is solved in levels. The state is still (z, k), the policy class and change of variables are those of Supplemental Appendix Section 5.3 (with ray slope $s_\zeta = 1.51$), and the trend enters through the law of motion Equation (E15) and nowhere else. With $g > 0$ capital and consumption grow without bound along the solution and the fitted policy is still a function of the two states. The benchmark path used for comparison is computed on the detrended system as a boundary value problem and rescaled to levels. The trend and the mean reversion of the stochastic model are not combined, since a drift $(1-a)g - \kappa z$ would mean-revert rather than trend.

The experiment sets $g = 0.02$ and solves from five initial conditions, drawing z_0 and $\log k_0$ with variance 0.01 around $(0, 0)$, with test paths from ten draws with variance 0.0025. The kernel's $\mathcal{D}$ is the five benchmark paths sampled at $t = 0, 1, \ldots, 60$, and the NN simulates the five trajectories with the Runge–Kutta step Supplemental Appendix Equation (S47). The errors are of the same order as the $g = 0$ rows of Supplemental Appendix Table 3, and the diagnostic Supplemental Appendix Equation (S49) is smaller, since the discounted product in Supplemental Appendix Equation (S22) falls at the rate $r + (\gamma - 1)g > r$ along a balanced growth path.

7.2 *The deterministic small open economy*

Under the restriction $\sigma_z = \sigma_r = 0$ of Supplemental Appendix Section 5.2, a NN solves for one trajectory from (k_0, b_0), along which $b(t)$ is a function of $k(t)$, so the policies are functions of capital alone, investment $\iota(k)$ and the current account $CA(k)$, with the change of variables of Supplemental Appendix Section 5.2. Its residuals are Supplemental Appendix Equation (S41) with $\Gamma \equiv 0$ and, in place of Supplemental Appendix Equation (S40), the second row of Supplemental Appendix Equation (S35) rescaled to levels

of consumption,

$$\dot{c} - \frac{r^\star - r}{\gamma} c = \frac{c^{\gamma+1}}{\gamma} R_2, \qquad \dot{c} := \nabla c \cdot \boldsymbol{F}, \tag{E16}$$

which again differs from R_2 by a positive factor. The trajectory is one Runge–Kutta path of 65 states on $[0, 40]$, trained in a single epoch. Because $r^\star = r$, neither residual involves b. Consumption is constant along the path, and the fitted policies are the same for every b_0, with the asset path shifted by b_0 and consumption by $r^\star b_0$. A fit from any one position therefore covers the three initial positions of Main Paper Section 4.4, $b_0 \in \{-0.84, 0, 0.84\}$, and separate solutions at each reproduce that section's benchmark. The continuum of steady states appears through the asset and consumption paths. The test entries in Table 11 are worst-case values over the path. The test residual is the maximum residual along the diagnostic path on $[0, 160]$, and the path error is the maximum benchmark error on $[0, 50]$, over the relative errors in capital, q, and consumption and the asset error relative to steady-state output. The transversality diagnostics are $\log\left[e^{-rT} c(T)^{-\gamma} q(T) k(T)\right]$ and $\log\left[e^{-rT} c(T)^{-\gamma} |b(T)|\right]$ on the single path at $T = 160$, reported for $b_0 = 0$ in the table and equal to -4.85 and -4.7 for $b_0 = -0.84$ and -4.97 and -5.08 for $b_0 = 0.84$.

TABLE 11. Recursive variants: kernel and NN policies, in the format of Supplemental Appendix Table 3. For the growth model the diagnostic is Supplemental Appendix Equation (S49). For the open economy the two diagnostics are $\log\left[e^{-rT} c(T)^{-\gamma} q(T) k(T)\right]$ and $\log\left[e^{-rT} c(T)^{-\gamma} |b(T)|\right]$ at $T = 160$. For the open economy the test residual is the maximum along the diagnostic path and the path error the maximum benchmark error over capital, q, consumption, and scaled assets.

Model	Solver	Size	Simulated $\mathcal{D}$	Test $\lvert\mathcal{R}\rvert$	Path error	Transversality	Time
Growth model with trend, $g = 0.02$							
Deterministic	Kernel	305 coeffs	5 traj $\times$ 61, $\Delta t = 1$	5.34×10^{-6}	3.55×10^{-5}	-11.98	0.181
Deterministic	NN	337 params	5 traj $\times$ 61, $\Delta t = 1$	1.49×10^{-7}	2.25×10^{-6}	-11.79	1.46
Deterministic small open economy (diagnostics: capital, assets)							
Deterministic	NN	338 params	1 traj $\times$ 65, $\Delta t = 0.625$	4.22×10^{-7}	3.42×10^{-4}	$-4.91, -4.88$	0.149

References

Acemoglu, Daron (2008), *Introduction to modern economic growth*. Princeton university press. [2, 7]

Perla, Jesse (2026), "tinydiffeq: Differentiable ODE/SDE/DAE solvers for JAX." https://github.com/HighDimensionalEconLab/tinydiffeq. Version 2.6.1. [6, 7]

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

Vidale, M. L. and H. B. Wolfe (1957), "An operations-research study of sales response to advertising." *Operations Research*, 5 (3), 370–381. [4]

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. [4]