TEACHER-POLICY VALUE IDENTIFICATION FROM CONTROL DEMONSTRATIONS WITH TEACHER-RELATIVE PERFORMANCE GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

SDRE controllers are useful teachers for high-dimensional nonlinear systems, but their algebraic value surrogate need not coincide with the infinite-horizon cost-to-go realized by the resulting policy. We propose a control-supervised approach to learn this teacher-policy value from held-input transitions and control labels, without value labels or statewise drift evaluations during training. A single neural value function, trained by one-step policy-value consistency and control supervision, yields both an estimate of the teacher-policy value and a gradient-defined controller that avoids per-state Riccati solves at deployment. Since the teacher need not be greedy with respect to its own value, the training objective balances value identification and control imitation. We prove that a uniform one-step residual bound controls the teacher-value identification error up to discretization error. Through a completion-of-the-square identity, value consistency and control supervision bound the HJB defect. For bounded finite-energy trajectories, the deployed cost plus accumulated control mismatch is bounded by the teacher cost plus residual/discretization errors, without optimal references or pointwise imitation assumptions. LQR anchoring ensures local exponential stability for all network parameters, independent of training accuracy. Numerical experiments on 40D Cucker–Smale, 51D Allen–Cahn, and 64D Burgers confirm stable closed-loop performance and rollout-value recovery while replacing online Riccati solves with amortized neural feedback.

1 INTRODUCTION

Optimal feedback control of nonlinear systems must stabilize the state while optimizing a performance objective over a region of operation, and it must respond rapidly to states throughout that region. For linear time-invariant systems with infinite-horizon quadratic costs, the linear quadratic regulator (LQR) provides an optimal stabilizing feedback law through an algebraic Riccati equation. For general nonlinear systems, dynamic programming characterizes such feedback through the Hamilton–Jacobi–Bellman (HJB) equation, but direct state-space discretization is prohibitive in high dimension. State-dependent Riccati equation (SDRE) control offers a practical alternative by expressing the drift in semilinear form and solving an algebraic Riccati equation at the current state (Banks et al., 2007; Çimen, 2008; Saluzzi, 2026). Related work learns state-dependent factorizations of nonlinear dynamics from data and uses them to construct SDRE controllers (Richards et al., 2023). Repeated Riccati solves can, however, dominate the online cost of large-scale control, which has motivated neural surrogates for fast feedback evaluation (da Costa et al., 2021; Albi et al., 2022; Samadi Khoshkho et al., 2023).

Following the learning-from-demonstration literature, we call the controller that produces the demonstrations the teacher. Direct action regression addresses the online bottleneck but does not explicitly identify the cost-to-go realized by the teacher. A scalar value model can estimate the cost-to-go and generate feedback through its gradient. Existing value-based SDRE methods fit an algebraic value surrogate constructed from state-dependent Riccati solutions, together with SDRE-derived derivative targets (Albi et al., 2022), with related kernel and tensor approximations studied by Oster et al. (2025). Such SDRE-based supervision has also been used to initialize HJB residual minimization (Borovykh et al., 2022). However, the pointwise Riccati construction does not generally ensure that this algebraic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

surrogate coincides with the cumulative cost realized by the continuously updated SDRE teacher. Fitting the surrogate therefore does not by itself identify the teacher's closed-loop performance.

Other approaches to optimal value learning use trajectory supervision, dynamics models, or policy iteration. Characteristic and Pontryagin-based methods generate value and gradient labels by solving two-point boundary value problems (Nakamura-Zimmerer et al., 2021b; Azmi et al., 2021), model-based neural approaches enforce an HJB residual using the system dynamics (Lutter et al., 2020; Schiassi et al., 2024; Fotiadis & Vamvoudakis, 2023), and physics-informed reinforcement learning uses integral Bellman residuals to evaluate policies within a policy-iteration framework (Wang & Wu, 2024). The overall objective of these approaches is to approximate optimal values and feedback policies, whereas fixed-policy evaluation concerns the cost-to-go realized by a prescribed controller.

Estimating the cost-to-go of a fixed controller from transition data is a policy-evaluation problem in reinforcement learning (Li, 2023). Temporal-difference methods learn from differences between successive predictions (Sutton, 1988), and their continuous-time and integral counterparts evaluate policies through differential or finite-time Bellman relations (Doya, 2000; Vrabie & Lewis, 2009; Lee & Sutton, 2021). Offline policy evaluation estimates the values of fixed policies from logged transitions under coverage and approximation conditions (Perdomo et al., 2023). Demonstration-based methods include interactive imitation learning (Ross et al., 2011) and combinations of temporal-difference learning and action supervision (Hester et al., 2018), while offline policy improvement must control errors from evaluating actions outside the data distribution (Fujimoto et al., 2019; Kumar et al., 2020).

In this work, we learn a single scalar neural model whose output estimates the realized cost-to-go of a fixed, possibly suboptimal SDRE teacher under continuous feedback, while its gradient defines a deployable feedback controller. Training uses teacher control labels and short state transitions, without value labels or pointwise evaluations of the system drift. A one-step approximation to the teacher's Bellman relation yields a value-consistency loss from the observed transitions and known running cost, which we combine with control supervision. These two objectives need not be aligned: even exact recovery of the teacher value does not necessarily reproduce the teacher's actions through its gradient. We therefore use a weighted objective to balance value identification and control imitation. Small empirical losses, however, do not guarantee closed-loop stability; we therefore enforce local stability through the value parameterization.

In neural feedback design, QRnet augments the linear-quadratic regulator (LQR) with a neural network to represent a scalar value function for nonlinear systems (Nakamura-Zimmerer et al., 2021a). This architecture, however, does not by itself guarantee stability of the resulting feedback controller. Subsequent stability-oriented designs parameterize the value gradient (costate) or the control policy directly: Nakamura-Zimmerer et al. (2022a) demonstrate improved closed-loop stability, while Nakamura-Zimmerer et al. (2022b) establish a local-stability guarantee. These stability-oriented architectures do not explicitly represent a scalar value function that directly estimates the realized cost-to-go. Inspired by these works, we propose an LQR-anchored neural value architecture that is trained to recover the SDRE teacher's realized value while guaranteeing local stability by construction. The equilibrium continuous-time algebraic Riccati equation (CARE) solution fixes the local quadratic LQR term, and the neural correction contributes only higher-order terms near the origin. Consequently, the gradient-defined feedback retains the stabilizing LQR linearization for every fixed choice of network parameters, yielding local exponential stability independently of training accuracy. At deployment, a value-gradient evaluation replaces the per-state Riccati solve.

Our contributions are fourfold. **(1)** We formulate teacher-policy value identification and propose a single scalar model that combines approximate policy-value consistency with control supervision, without value labels or drift evaluations during training. **(2)** We establish a regional error bound for teacher-policy value identification in terms of a uniform one-step residual bound and the held-input discretization error. **(3)** We derive Bellman–control compatibility and a regional bound on the sum of deployed cost and accumulated control mismatch relative to the teacher value, without optimal references or pointwise imitation assumptions. **(4)** We propose a new LQR-anchored value-function parameterization that guarantees local exponential stability of the resulting closed-loop system independently of training accuracy. Experiments on 40-dimensional Cucker–Smale, 51-dimensional Allen–Cahn, and 64-dimensional Burgers systems evaluate closed-loop performance and quantify the deployment speedup from replacing per-state Riccati solves with neural feedback.

Section 2 introduces the nonlinear regulation problem and distinguishes the SDRE teacher's realized policy value from its algebraic surrogate. Section 3 presents the LQR-anchored value parameterization and the control-supervised learning objective. Section 4 establishes value-identification and teacher-relative performance bounds, together with local exponential stability. Section 5 evaluates value recovery, closed-loop performance, and feedback evaluation time on three high-dimensional systems. Section 6 summarizes the findings and outlines future directions. The appendices contain supporting results, proofs, and implementation details.

2 Problem formulation and the SDRE teacher

Notation. $\|x\|$ is the Euclidean norm of x, $A \succ B$ ($A \succeq B$) means that $A - B$ is positive (semi-)definite, and $\|v\|_R^2 := v^\top R v$ is the R-weighted squared norm for $R \succ 0$. ∇ and ∇^2 denote the gradient and Hessian with respect to the state, and $\dot{z}$ denotes the time derivative of z. Finally, Ω denotes an open state domain containing the origin, Π the set of admissible policies on Ω, and $\mathcal{K} \subset \Omega$ the compact region on which the regional guarantees are stated.

2.1 Nonlinear optimal regulation

We consider, for $t \geq 0$, the general nonlinear dynamics with affine control input

$$\dot{x}(t) = f(x(t)) + B(x(t))u(t), \qquad x(0) = x_0, \tag{1}$$

where $x(t) \in \Omega \subseteq \mathbb{R}^n$ is the state, and $u(t) \in \mathbb{R}^m$ is the control input. The drift $f : \Omega \to \mathbb{R}^n$ and the input matrix $B : \Omega \to \mathbb{R}^{n \times m}$ are assumed to be smooth with $f(0) = 0$.

Let $Q = Q^\top \in \mathbb{R}^{n \times n}$ and $R = R^\top \in \mathbb{R}^{m \times m}$ be the state and control weighting matrices, respectively, with $Q \succeq 0$ and $R \succ 0$. The running cost ℓ and the corresponding infinite-horizon cost-to-go from initial state x_0 are

$$\ell(x, u) := x^\top Q x + u^\top R u, \qquad J(u; x_0) := \int_0^\infty \ell(x(t), u(t))\, \mathrm{d}t. \tag{2}$$

For a feedback policy $\pi : \Omega \to \mathbb{R}^m$, we write $J(\pi; x_0)$ for the cost obtained with $u(t) = \pi(x(t))$.

Definition 2.1 (Admissible policy)**.** *A feedback policy $\pi : \Omega \to \mathbb{R}^m$ is called admissible on Ω, written $\pi \in \Pi$, if (i) it is continuous with $\pi(0) = 0$; (ii) the origin is Lyapunov stable under the closed-loop dynamics, and for every $x_0 \in \Omega$, the closed-loop trajectory exists uniquely for all $t \geq 0$, remains in Ω, and converges to the origin; and (iii) $J(\pi; x_0) < \infty$ for every $x_0 \in \Omega$.*

For $\pi \in \Pi$, its policy value is $V^\pi(x) := J(\pi; x)$. For any continuously differentiable scalar function $V(x)$, define the Hamiltonian H evaluated pointwise at (x, u) as

$$H(x, u, V) := \nabla V(x)^\top \big(f(x) + B(x)u\big) + \ell(x, u). \tag{3}$$

If V^π is continuously differentiable, the generalized Hamilton–Jacobi–Bellman equation (GHJB) reads $H(x, \pi(x), V^\pi) = 0$. This identity does not imply that $\pi(x)$ minimizes $H(x, u, V^\pi)$ over $u \in \mathbb{R}^m$.

The optimal value is $V^*(x) := \inf_{\pi \in \Pi} J(\pi; x)$. When V^* is continuously differentiable and an admissible optimal policy exists, V^* satisfies the HJB equation

$$H\big(x, u^*, V^*\big) = 0, \tag{4}$$

with the optimal policy given by

$$u^*(x) = \arg\min_{u \in \mathbb{R}^m} H(x, u, V^*) = -\frac{1}{2} R^{-1} B(x)^\top \nabla V^*(x). \tag{5}$$

2.2 The SDRE teacher and its policy value

Assume that the drift admits a state-dependent coefficient factorization $f(x) = A(x)x$, where $A : \Omega \to \mathbb{R}^{n \times n}$ is a chosen smooth matrix-valued function. This factorization need not be unique, and it gives the semilinear state dynamics:

$$\dot{x} = A(x)x + B(x)u. \tag{6}$$

For each $x \in \Omega$, the state-dependent Riccati equation (SDRE) method computes a symmetric stabilizing solution $P(x) \in \mathbb{R}^{n\times n}$ of

$$A(x)^\top P(x) + P(x)A(x) - P(x)B(x)R^{-1}B(x)^\top P(x) + Q = 0. \tag{7}$$

The stabilizing solution is assumed to exist and to vary smoothly on Ω, and we assume that $P_0 := P(0) \succ 0$ at the equilibrium $x = 0$. The induced SDRE feedback is

$$u_S(x) := -R^{-1}B(x)^\top P(x)x. \tag{8}$$

We assume that the SDRE feedback u_S is admissible on Ω in the sense of Definition 2.1. Define the SDRE algebraic value surrogate $V_{\mathrm{alg}} : \Omega \to \mathbb{R}$ by

$$V_{\mathrm{alg}}(x) := x^\top P(x)x, \qquad x \in \Omega. \tag{9}$$

Let $x_S(t; x)$ denote the trajectory initialized at $x_S(0; x) = x$ under the continuously updated SDRE feedback u_S. Its realized policy value is

$$V_S(x) := V^{u_S}(x) = J(u_S; x) = \int_0^\infty \ell\big(x_S(t; x), u_S(x_S(t; x))\big)\,\mathrm{d}t. \tag{10}$$

Since $f(0) = 0$ and $u_S(0) = 0$, the origin is a zero-cost equilibrium and $V_S(0) = 0$. Along the SDRE closed loop $x(t) = x_S(t; x_0)$, the algebraic surrogate satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}V_{\mathrm{alg}}(x(t)) + \ell(x(t), u_S(x(t))) = x(t)^\top \dot{P}(x(t))x(t). \tag{11}$$

The additional term on the right-hand side is generally nonzero. By equation (10), however, $\frac{\mathrm{d}}{\mathrm{d}t}V_S(x(t)) + \ell(x(t), u_S(x(t))) = 0$. When V_S is continuously differentiable, this identity is equivalent to the GHJB relation $H(x, u_S(x), V_S) = 0$, with H defined in equation (3). Thus, $V_S \neq V_{\mathrm{alg}}$ in general.

Table 2 in the Appendix lists the main symbols together with the information available to the learner.

3 Control-supervised Bellman learning algorithm

3.1 LQR-anchored value-function parameterization

Let $V_\theta : \Omega \to \mathbb{R}$ be a differentiable value model with trainable parameters θ and $u_\theta : \Omega \to \mathbb{R}^m$ the feedback it induces (equations (12) and (14) below). We seek a model such that (i) $V_\theta \approx V_S$, (ii) $u_\theta \approx u_S$, and (iii) u_θ locally stabilizes the origin. Requirement (i) is addressed by a one-step policy-evaluation residual; (ii) uses the control labels to constrain the control-relevant gradient $B(x)^\top \nabla V_\theta(x)$; and (iii) is enforced by the value parameterization.

To preserve the LQR quadratic term at the origin, we parameterize the scalar value function as

$$V_\theta(x) := x^\top P_0 x \cdot \exp(g_\theta(x) - g_\theta(0)), \tag{12}$$

where $g_\theta : \Omega \to \mathbb{R}$ is a twice continuously differentiable scalar neural network. The exponential factor is positive and equals one at the origin, so the model matches the LQR value through second order at the equilibrium while allowing nonlinear corrections away from it. In particular, for every θ it satisfies

$$V_\theta(0) = 0, \qquad \nabla V_\theta(0) = 0, \qquad \nabla^2 V_\theta(0) = 2P_0. \tag{13}$$

The induced feedback is greedy with respect to V_θ:

$$u_\theta(x) := \operatorname*{arg\,min}_{u\in\mathbb{R}^m} H(x, u, V_\theta) = -\frac{1}{2}R^{-1}B(x)^\top \nabla V_\theta(x). \tag{14}$$

The identities in equation (13) imply $u_\theta(0) = 0$ and $\partial u_\theta/\partial x(0) = -R^{-1}B(0)^\top P_0$, so the feedback retains the stabilizing LQR linearization for every fixed θ, yielding local exponential stability independently of training accuracy (Theorem 4.4).

3.2 TEACHER DATA GENERATION AND TRAINING PROCEDURE

Fix a sampling interval $h > 0$. For each initial state x, let $z(s; x)$ solve the nonlinear dynamics with the teacher control value $u_S(x)$ held fixed:

$$\dot{z}(s; x) = f(z(s; x)) + B(z(s; x))u_S(x), \qquad z(0; x) = x, \qquad F_h(x) := z(h; x). \tag{15}$$

The state-dependent dynamics are evaluated at $z(s; x)$ throughout the interval. At the next sampling time, the teacher control is recomputed at the successor state. Thus $F_h^k(x)$ is the state at time kh under this sampled implementation, and generally differs from $x_S(kh; x)$ under continuously updated teacher feedback. For N sampled states $\{x_i\}_{i=1}^N \subset \Omega$, let $u_i := u_S(x_i)$ denote the SDRE teacher label and $x_i^+ := F_h(x_i)$ the successor state. The sampled states and h are chosen so that the corresponding trajectories exist and remain in Ω over $[0, h]$. The demonstration data are

$$\mathcal{D}_h := \left\{(x_i, u_i, x_i^+)\right\}_{i=1}^N. \tag{16}$$

Numerical integration approximates F_h in equation (15); the analysis treats F_h as the exact held-input transition. Since Q and R are known, the left-point approximation of the one-step teacher cost is $h\ell(x_i, u_i)$, so the data provide the approximate policy-evaluation relation

$$V_S(x_i) - V_S(x_i^+) \approx h\ell(x_i, u_i), \tag{17}$$

with an error due to left-point quadrature and holding the teacher control value $u_S(x_i)$ fixed over the sampling interval. Section 3.3 defines the corresponding one-step residual, and Section 4 quantifies the resulting discretization error.

During training, the learner observes $\mathcal{D}_h$, can evaluate $B(x)$, and knows Q, R, h, and $P_0 = P(0)$, but does not evaluate $f(x)$, $A(x)$, or the SDRE solution $P(x)$. Training uses mini-batch stochastic optimization: at each step we evaluate V_θ at the sampled states and successors, obtain u_θ from the value gradient, and update θ with a stochastic-gradient estimate of equation (21); Algorithm 1 in Appendix B records the complete procedure.

3.3 TEACHER-VALUE CONSISTENCY AND CONTROL SUPERVISION

Based on the approximate teacher-policy evaluation relation (17), we define the one-step Bellman residual on the same per-unit-time scale as the running cost ℓ:

$$r_h[V](x) = \frac{V(x) - V(F_h(x))}{h} - \ell(x, u_S(x)). \tag{18}$$

The associated value-based Bellman loss

$$\mathcal{L}_V(\theta) := \frac{1}{N}\sum_{i=1}^N \big(r_h[V_\theta](x_i)\big)^2, \qquad r_h[V_\theta](x_i) = \frac{V_\theta(x_i) - V_\theta(x_i^+)}{h} - \ell(x_i, u_i), \tag{19}$$

penalizes one-step policy-evaluation inconsistency on the teacher demonstrations, with V_S as the value-learning target.

The SDRE teacher need not be greedy with respect to its own policy value V_S. Consequently, even if $V_\theta = V_S$, the induced feedback u_θ defined by equation (14) need not coincide with u_S. To promote control imitation, we use the teacher labels to constrain the control-relevant gradient $B(x)^\top \nabla V_\theta(x)$. The empirical control-supervision loss is

$$\mathcal{L}_u(\theta) := \frac{1}{N}\sum_{i=1}^N \|u_\theta(x_i) - u_i\|_2^2. \tag{20}$$

The total loss is

$$\mathcal{L}(\theta) = \eta_b \mathcal{L}_V(\theta) + \eta_u \mathcal{L}_u(\theta), \tag{21}$$

where $\eta_b \geq 0$ and $\eta_u \geq 0$ are adjustable weights that balance teacher-value identification and control imitation, with default $\eta_b = \eta_u = 1$.

The value-recovery and performance guarantees in Section 4 are stated under uniform residual bounds and regional trajectory conditions. Small empirical losses alone do not establish these conditions.

4 THEORY: VALUE IDENTIFICATION, BELLMAN–CONTROL COMPATIBILITY, AND PERFORMANCE GUARANTEES

In this section, we first establish value-recovery (Theorem 4.1) and Bellman–control compatibility bounds (Theorem 4.2), then combine them to bound the learned policy's cost relative to the SDRE teacher (Theorem 4.3). We also establish local exponential stability for each fixed θ through LQR anchoring, independently of training accuracy (Theorem 4.4). Proofs and supporting results are provided in Appendix A.

4.1 IDENTIFICATION OF THE TEACHER-POLICY VALUE

Theorem 4.1. *Let F_h be defined by equation* (15)*. Let $\mathcal{K} \subset \Omega$ be a compact region containing the origin in its interior, and suppose that V_S is continuous on $\mathcal{K}$. Assume that there exist constants $h_0, \Gamma, C_S > 0$ such that, for every $0 < h \le h_0$,*

$$F_h(\mathcal{K}) \subseteq \mathcal{K}, \tag{22}$$

$$h \sum_{k=0}^{\infty} \left\| F_h^k(x) \right\|^2 \le \Gamma \|x\|^2, \qquad x \in \mathcal{K}, \tag{23}$$

$$|r_h[V_S](x)| \le C_S h \|x\|^2, \qquad x \in \mathcal{K}. \tag{24}$$

For any fixed $h \in (0, h_0]$, let V_θ be given by equation (12)*. If, for some $\varepsilon_V \ge 0$,*

$$|r_h[V_\theta](x)| \le \varepsilon_V \|x\|^2, \qquad x \in \mathcal{K}, \tag{25}$$

then

$$|V_\theta(x) - V_S(x)| \le \Gamma(\varepsilon_V + C_S h) \|x\|^2, \qquad x \in \mathcal{K}. \tag{26}$$

Remark 4.1. Under the local regularity assumptions of Lemma A.1 in Appendix A.1, including $V_S \in C^2$ near the origin, equations (22)–(24) hold on a small neighborhood of the origin. Here, for a larger compact region $\mathcal{K} \subset \Omega$, we assume that these conditions hold. The bound in equation (26) guarantees the teacher-value recovery up to the learned residual and an $O(h)$ discretization error, provided the residual bounds are uniform in $\mathcal{K}$.

4.2 BELLMAN–CONTROL COMPATIBILITY AND TEACHER-RELATIVE PERFORMANCE

Theorem 4.2. *Let $\mathcal{K} \subset \Omega$ be a compact region containing the origin in its interior, and suppose that equation* (22) *holds for the map (15) for every $0 < h \le h_0$. Assume that f, B, and u_S are C^1 on a neighborhood of $\mathcal{K}$, and fix θ such that g_θ is C^2 on that neighborhood. After possibly reducing h_0, there exists a constant $C > 0$, which may depend on $\mathcal{K}$ and the fixed network parameters θ, but is independent of $x \in \mathcal{K}$ and $0 < h \le h_0$, such that*

$$\Big| \underbrace{\ell(x, u_\theta(x)) + \nabla V_\theta(x)^\top \big(f(x) + B(x)u_\theta(x)\big)}_{\text{HJB defect}} + \underbrace{r_h[V_\theta](x)}_{\text{Bellman residual}} + \underbrace{\|u_\theta(x) - u_S(x)\|_R^2}_{\text{Control mismatch}} \Big| \le Ch \|x\|^2. \tag{27}$$

Remark 4.2. It follows from equation (27) that

$$|H(x, u_\theta(x), V_\theta)| \le |r_h[V_\theta](x)| + \|u_\theta(x) - u_S(x)\|_R^2 + Ch \|x\|^2, \qquad x \in \mathcal{K}. \tag{28}$$

Thus the value residual and control mismatch jointly bound the HJB defect, up to a discretization term. Theorem 4.3 below converts the compatibility relation into a teacher-relative cost bound.

Theorem 4.3. *Let $\mathcal{K}$, h, and V_θ satisfy the hypotheses of Theorems 4.1 and 4.2. Let C be the constant in equation* (27)*, and let $x_{u_\theta}(t; x)$ denote the trajectory of equation* (1) *under $u(t) = u_\theta(x(t))$, initialized at x. Fix an initial state $x \in \mathcal{K}$. Suppose that this trajectory is well defined for all $t \ge 0$, remains in $\mathcal{K}$, and satisfies, for some $\Gamma_\theta > 0$,*

$$\int_0^\infty \|x_{u_\theta}(t; x)\|^2 \, \mathrm{d}t \le \Gamma_\theta \|x\|^2. \tag{29}$$

Then $x_{u_\theta}(t;x) \to 0$ *as* $t \to \infty$, $J(u_\theta;x) < \infty$, *and*

$$J(u_\theta;x) + \int_0^\infty \|u_\theta(x_{u_\theta}(t;x)) - u_S(x_{u_\theta}(t;x))\|_R^2 \, \mathrm{d}t \\ \leq V_S(x) + \left[\Gamma(\varepsilon_V + C_S h) + \Gamma_\theta(\varepsilon_V + Ch)\right] \|x\|^2 . \tag{30}$$

In particular,

$$J(u_\theta;x) - V_S(x) \leq \left[\Gamma(\varepsilon_V + C_S h) + \Gamma_\theta(\varepsilon_V + Ch)\right] \|x\|^2 . \tag{31}$$

Remark 4.3. The bound in equation (30) controls both the deployed cost and the accumulated R-weighted control mismatch, without assuming that the control imitation error is small. The learned trajectories must still satisfy the state-energy bound (29), which does not follow from a small empirical Bellman loss alone.

If the trajectory assumptions hold for every initial state in $\mathcal{K}$, then u_θ is admissible on $\mathcal{K}$. Indeed, regularity and the anchored parameterization ensure continuity, $u_\theta(0) = 0$, and uniqueness of trajectories; Theorem 4.3 gives convergence and finite cost, while Theorem 4.4 provides Lyapunov stability. Proposition A.2 provides sufficient conditions for these trajectory assumptions on a value sublevel set.

4.3 Stability by construction

Theorem 4.4. *Assume that* $f(0) = 0$ *and that* f *and* B *are* C^1 *near the origin. Let* $P_0 \succ 0$ *be the stabilizing equilibrium CARE solution of Section 2, and define the corresponding Hurwitz matrix*

$$A_c := \frac{\partial f}{\partial x}(0) - B(0)R^{-1}B(0)^\top P_0.$$

For any fixed θ *such that* g_θ *is* C^2 *near the origin, let* V_θ *and* u_θ *be defined by equation* (12) *and equation* (14)*. Then*

$$u_\theta(x) = -R^{-1}B(0)^\top P_0 x + \mathcal{O}(\|x\|^2), \qquad x \to 0, \tag{32}$$

and

$$\frac{\partial(f + Bu_\theta)}{\partial x}(0) = A_c. \tag{33}$$

Hence the origin of equation (1) *under the feedback* u_θ *is locally exponentially stable, independently of training accuracy.*

Remark 4.4. Under additional local assumptions, we can show that the realized-cost gaps $V_S(x) - V^*(x)$ and $J(u_\theta;x) - V^*(x)$ are both $\mathcal{O}(\|x\|^4)$ as $x \to 0$, see Corollary A.4 in Appendix A.8.

5 Numerical experiments

5.1 Experimental setup

We evaluate teacher-value identification, closed-loop performance, and feedback evaluation time on the 40D Cucker–Smale system, the 51D Allen–Cahn equation, and a 64D Burgers-type equation, with control dimensions $m = 20, 1, 2$, respectively. Cucker–Smale tests regulation of 20 interacting agents through their accelerations. Allen–Cahn tests stabilization of a reaction–diffusion equation through a single localized actuator, using 51 finite-difference grid points. Burgers combines nonlinear convection, diffusion, and a destabilizing reaction, with two localized actuators and 64 interior Chebyshev nodes. Model equations and data generation are given in Appendix C.1.

All value networks use equation (12). We compare our method ($\eta_b = \eta_u = 1$) with Control-only ($\eta_b = 0$, $\eta_u = 1$), u_{Jac}-QRnet (Nakamura-Zimmerer et al., 2022b), local LQR, and online SDRE. For Table 1, each learned method uses the same 10,000 teacher transitions and three network initialization seeds. Each network is tested on 15 initial states obtained by normalizing three independently sampled initial states and multiplying each by five scale factors $s \in \{1, 1.25, 1.5, 1.75, 2\}$, with the normalization procedure defined in Appendix C.3. The first factor places the test states at the largest observed training initial-state norm; the remaining four place them beyond that range. Training setting details are given in Appendix C.2.

Table 1: Performance and CPU feedback evaluation time with 10,000 training transitions across five test scales. For learned methods, E_{AUC} and cost ratios report the mean and standard deviation of three seed-wise averages over 15 initial states; deterministic baselines use the same states. The LQR value candidate is $x^\top P_0 x$. Cost ratios exclude terminal corrections. CPU time is measured per feedback evaluation at a fixed state, excluding state integration. We report the median over repeated calls, averaged over training seeds for learned methods.

System	Method	$E_{\mathrm{AUC}} \downarrow$	$J_T(\pi)/J_T(u_S)$	Time (ms)↓
Cucker–Smale ($n = 40$)	Ours	$\mathbf{0.043 \pm 0.011}$	1.0078 ± 0.0008	0.5626
	Control-only	0.056 ± 0.028	1.0089 ± 0.0009	0.5621
	u_{Jac}-QRnet	–	1.0043 ± 0.0001	0.1574
	LQR	0.067	1.0104	0.0022
	SDRE	–	1.0000	8.1160
Allen–Cahn ($n = 51$)	Ours	$\mathbf{0.125 \pm 0.011}$	0.9939 ± 0.0021	0.3531
	Control-only	0.231 ± 0.057	0.9944 ± 0.0007	0.3558
	u_{Jac}-QRnet	–	1.0040 ± 0.0085	0.0959
	LQR	1.034	1.3376	0.0017
	SDRE	–	1.0000	8.7593
Burgers ($n = 64$)	Ours	$\mathbf{0.228 \pm 0.035}$	0.9899 ± 0.0088	0.3405
	Control-only	0.277 ± 0.025	1.0396 ± 0.0723	0.3396
	u_{Jac}-QRnet	–	1.0054 ± 0.0103	0.0907
	LQR	1.164	1.0721	0.0019
	SDRE	–	1.0000	16.3579

5.2 Value recovery and closed-loop performance

Along each online-SDRE trajectory x_S, we measure value recovery by

$$E_{\mathrm{AUC}} = \int_0^T \left| V_\theta(x_S(t)) - \widehat{V}_S(x_S(t)) \right| \mathrm{d}t \Big/ \int_0^T \left| \widehat{V}_S(x_S(t)) \right| \mathrm{d}t,$$

where $\widehat{V}_S(x_S(t))$ is the remaining SDRE rollout cost from t to T with the local-LQR terminal estimate $x_S(T)^\top P_0 x_S(T)$.

Averaged over the five test scales, our method achieves the lowest mean E_{AUC} among the evaluated value models on all three problems (Table 1), with the largest improvement over Control-only on Allen–Cahn. The finite-horizon cost ratio $J_T(\pi)/J_T(u_S)$ compares the deployed controller's accumulated running cost with that of the teacher. Our method achieves mean ratios of 1.0078, 0.9939, and 0.9899 on Cucker–Smale, Allen–Cahn, and Burgers, respectively. The ratios below one on Allen–Cahn and Burgers indicate small empirical improvements in average relative cost over the teacher. Our controller requires approximately 0.34–0.56 ms per CPU feedback evaluation, corresponding to a 14.4–$48.0\times$ speedup over online SDRE by replacing per-state Riccati solves with neural feedback.

Appendices C.4.1 and C.4.2 examine the effects of training-set size and initial-condition scale, respectively, on both E_{AUC} and the cost ratio. With the networks fixed, Appendix C.5 further shows that the sampled median absolute compatibility remainder decreases as h is reduced.

5.3 Effect of Bellman supervision

We examine how Bellman supervision affects teacher-value recovery and the resulting feedback on Allen–Cahn. Figure 1 shows that our method, Control-only, and u_{Jac}-QRnet produce nearly identical control signals, while the value curve from our method is closer to the teacher rollout reference than that from Control-only. Figure 2 compares three representative Bellman weights, $\eta_b \in \{10^{-3}, 1, 10^3\}$, with $\eta_u = 1$. For the illustrated initial condition, $\eta_b = 1$ gives value estimates close to the teacher rollout cost-to-go and a control signal that nearly coincides with SDRE. Reducing η_b to 10^{-3} preserves this close control agreement but produces more pronounced overestimation of the teacher cost-to-go at early times. Increasing η_b to 10^3 also gives value estimates close to the

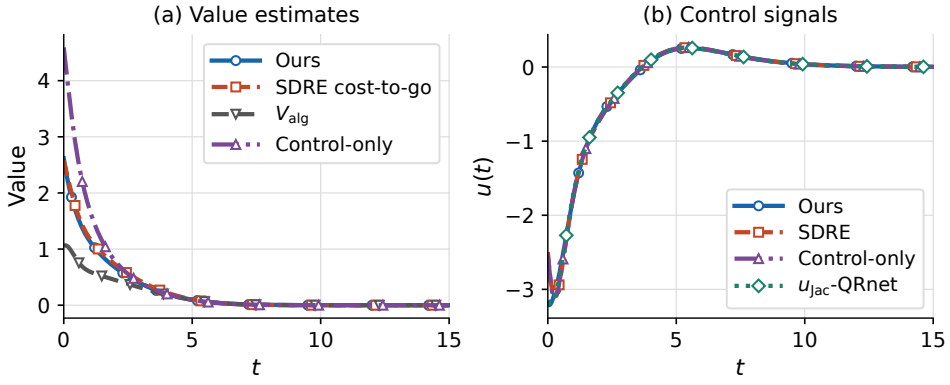


Figure 1: Value estimates (a) and control signals (b) for Allen–Cahn with $x_0(\xi) = 1 + \xi(1-\xi)$. Each network value is evaluated along its own closed-loop trajectory. SDRE cost-to-go is the teacher's remaining cost to $T = 15$; V_{alg} is evaluated along the teacher trajectory.

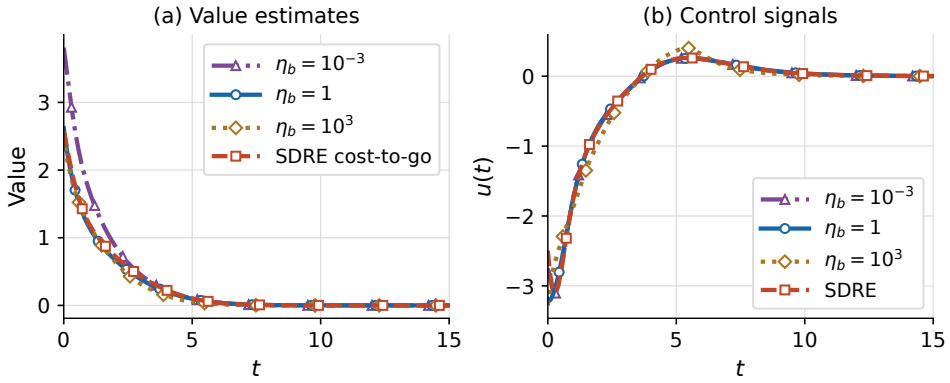


Figure 2: Effect of the Bellman-loss weight η_b on Allen–Cahn value estimates (a) and control signals (b), with $\eta_u = 1$. The initial state and trajectory conventions are as in Figure 1.

teacher reference, while introducing small transient deviations from the SDRE control signal. In this example, $\eta_b = \eta_u = 1$ provides a practical balance between teacher-value recovery and control imitation.

6 Conclusion

We propose a control-supervised method that uses a single scalar network to learn an SDRE teacher's realized cost-to-go and generate a gradient-defined controller from held-input transitions and control labels, without value labels or drift evaluations during training. The method balances value identification and control imitation, which can conflict because the teacher need not be greedy with respect to its own value. We establish bounds on value-identification error and the HJB defect, together with a regional teacher-relative performance bound that includes accumulated control mismatch. Finally, the proposed LQR-anchored value-function parameterization guarantees local exponential stability of the resulting closed-loop system, independently of training accuracy. Experiments on three high-dimensional systems show improved value recovery over control-only training and closed-loop costs close to SDRE, with 14.4–$48.0\times$ speedup in CPU feedback evaluation by replacing online Riccati solves.

Future work will investigate adaptive weighting of value consistency and control supervision across the state space. Another direction is to extend the framework to continuous-control reinforcement learning by coupling policy-value identification with iterative policy improvement.

Reproducibility statement

Appendix A provides the proofs and supporting estimates. Appendices B and C describe the training procedure, benchmark systems, evaluation protocol, and additional numerical results.

Data and code availability

The generated datasets and source code are intended for release in anonymized supplementary material. Repository metadata and author-identifying paths will be removed before submission.

Ethics statement

The study uses simulated dynamical systems and does not involve human participants, animals, personal data, or other sensitive data.

Author contributions, funding, and competing interests

A CRediT contribution statement and funding information will be added after double-blind review. The authors should confirm the final competing-interest declaration before submission.

AI use statement

Generative AI tools assisted with mathematical reasoning, code development, experimental analysis, and manuscript preparation. The authors made all final scientific decisions and take full responsibility for the final work.

References

Giacomo Albi, Sara Bicego, and Dante Kalise. Gradient-augmented supervised learning of optimal feedback laws using state-dependent Riccati equations. *IEEE Control Systems Letters*, 6:836–841, 2022. doi: 10.1109/LCSYS.2021.3086697.

Behzad Azmi, Dante Kalise, and Karl Kunisch. Optimal feedback law recovery by gradient-augmented sparse polynomial regression. *Journal of Machine Learning Research*, 22(48):1–32, 2021. URL https://www.jmlr.org/papers/v22/20-755.html.

H. T. Banks, B. M. Lewis, and H. T. Tran. Nonlinear feedback controllers and compensators: A state-dependent Riccati equation approach. *Computational Optimization and Applications*, 37(2): 177–218, 2007. doi: 10.1007/s10589-007-9015-2.

Anastasia Borovykh, Dante Kalise, Alexis Laignelet, and Panos Parpas. Data-driven initialization of deep learning solvers for Hamilton–Jacobi–Bellman PDEs. *IFAC-PapersOnLine*, 55(30):168–173, 2022. doi: 10.1016/j.ifacol.2022.11.047.

Tayfun Çimen. State-dependent riccati equation (SDRE) control: a survey. *IFAC Proceedings Volumes*, 41(2):3761–3775, 2008. doi: 10.3182/20080706-5-KR-1001.00635.

Rômulo Fernandes da Costa, Osamu Saotome, Elvira Rafikova, and Renato Machado. Fast real-time SDRE controllers using neural networks. *ISA Transactions*, 118:133–143, 2021. doi: 10.1016/j.isatra.2021.02.019.

Kenji Doya. Reinforcement learning in continuous time and space. *Neural Computation*, 12(1): 219–245, 2000. doi: 10.1162/089976600300015961.

Filippos Fotiadis and Kyriakos G Vamvoudakis. A physics-informed neural networks framework to solve the infinite-horizon optimal control problem. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 6014–6019. IEEE, 2023. doi: 10.1109/CDC49753.2023.10383404.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062. PMLR, 2019. URL https://proceedings.mlr.press/v97/fujimoto19a.html.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, Gabriel Dulac-Arnold, John Agapiou, Joel Z. Leibo, and Audrunas Gruslys. Deep Q-learning from demonstrations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):3223–3230, 2018. doi: 10.1609/aaai.v32i1.11757.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020.

Jaeyoung Lee and Richard S. Sutton. Policy iterations for reinforcement learning problems in continuous time and space—fundamental theory and methods. *Automatica*, 126:109421, 2021. doi: 10.1016/j.automatica.2020.109421.

Shengbo Eben Li. *Reinforcement learning for sequential decision and optimal control*. Springer, 2023. doi: 10.1007/978-981-19-7784-8.

Michael Lutter, Boris Belousov, Kim Listmann, Debora Clever, and Jan Peters. HJB optimal feedback control with deep differential value functions and action constraints. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 640–650. PMLR, 2020. URL https://proceedings.mlr.press/v100/lutter20a.html.

Tenavi Nakamura-Zimmerer, Qi Gong, and Wei Kang. QRnet: Optimal regulator design with LQR-augmented neural networks. *IEEE Control Systems Letters*, 5(4):1303–1308, 2021a. doi: 10.1109/LCSYS.2020.3034415.

Tenavi Nakamura-Zimmerer, Qi Gong, and Wei Kang. Adaptive deep learning for high-dimensional Hamilton–Jacobi–Bellman equations. *SIAM Journal on Scientific Computing*, 43(2):A1221–A1247, 2021b. doi: 10.1137/19M1288802.

Tenavi Nakamura-Zimmerer, Qi Gong, and Wei Kang. Neural network optimal feedback control with enhanced closed loop stability. In *2022 American Control Conference (ACC)*, pp. 2373–2378, 2022a. doi: 10.23919/ACC53348.2022.9867619.

Tenavi Nakamura-Zimmerer, Qi Gong, and Wei Kang. Neural network optimal feedback control with guaranteed local stability. *IEEE Open Journal of Control Systems*, 1:210–222, 2022b. doi: 10.1109/OJCSYS.2022.3205863.

Mathias Oster, Luca Saluzzi, and Tizian Wenzel. A comparison study of supervised learning techniques for the approximation of high dimensional functions and feedback control. *Dynamic Games and Applications*, 15(2):454–480, 2025. doi: 10.1007/s13235-024-00610-6.

Juan C. Perdomo, Akshay Krishnamurthy, Peter Bartlett, and Sham Kakade. A complete characterization of linear estimators for offline policy evaluation. *Journal of Machine Learning Research*, 24 (284):1–50, 2023. URL https://www.jmlr.org/papers/v24/22-0341.html.

Spencer M. Richards, Jean-Jacques Slotine, Navid Azizan, and Marco Pavone. Learning control-oriented dynamical structure from data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29051–29062. PMLR, 2023. URL https://proceedings.mlr.press/v202/richards23a.html.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Luca Saluzzi. The state-dependent riccati equation in nonlinear optimal control: Analysis and numerical approximation. *Journal of Optimization Theory and Applications*, 209(1):34, 2026. doi: 10.1007/s10957-026-02980-w.

Mohammadamin Samadi Khoshkho, Zahra Samadikhoshkho, and Michael G. Lipsett. Distilled neural state-dependent Riccati equation feedback controller for dynamic control of a cable-driven continuum robot. *International Journal of Advanced Robotic Systems*, 20(3):17298806231174737, 2023. doi: 10.1177/17298806231174737.

Enrico Schiassi, Andrea D'Ambrosio, and Roberto Furfaro. Bellman neural networks for the class of optimal control problems with integral quadratic cost. *IEEE Transactions on Artificial Intelligence*, 5(3):1016–1025, 2024. doi: 10.1109/TAI.2022.3206735.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3 (1):9–44, 1988. doi: 10.1007/BF00115009.

Draguna Vrabie and Frank L. Lewis. Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Networks*, 22(3):237–246, 2009. doi: 10.1016/j.neunet.2009.03.008.

Yujia Wang and Zhe Wu. Physics-informed reinforcement learning for optimal control of nonlinear systems. *AIChE Journal*, 70(10):e18542, 2024. doi: 10.1002/aic.18542.

Table 2: Main notation and information available to the learner.

Symbol	Meaning	Learner access
x, u	State and control input	observed
f, B	Drift and input matrix	f: no; B: yes
Q, R	State and control weighting matrices	known
A, P	SDRE factorization and Riccati solution	no
P_0	Equilibrium CARE solution $P(0)$	known
u_S	SDRE teacher feedback	control labels u_i only
V_{alg}	Algebraic SDRE surrogate $x^\top P(x)x$	not used as a label
V_S	Realized SDRE policy value $J(u_S; x)$	unknown target
V^*, u^*	Optimal value and feedback	unknown
F_h	Nonlinear one-step map with fixed control value $u_S(x)$	successors x_i^+
V_θ, u_θ	Learned value and induced feedback (θ: parameters)	learned outputs

A Proofs and supporting theoretical results

Appendix A.1 collects common held-input map estimates and local teacher properties. Appendices A.2–A.5 prove the main theorems. Appendix A.6 gives a regional energy certificate, and Appendices A.7–A.8 establish optimal-reference bounds and their local fourth-order refinement.

Common notation. Throughout this appendix, we retain the notation introduced in the main text and use the following additional notation. For the dynamics (1), define the linearization matrices at the equilibrium $(x, u) = (0, 0)$ by

$$A_0 := \frac{\partial f}{\partial x}(0), \qquad B_0 := B(0).$$

We also write

$$a_S(x) := f(x) + B(x)u_S(x), \qquad b_x(y) := f(y) + B(y)u_S(x), \qquad x, y \in \Omega.$$

Here a_S is the vector field under continuously updated teacher feedback, whereas b_x is the vector field with the control value $u_S(x)$ held fixed. In particular,

$$b_x(x) = a_S(x).$$

A.1 Preliminary estimates

We first establish uniform short-time displacement estimates and a first-order expansion for the held-input map for subsequent use. Let $\mathcal{K} \subset \Omega$ be compact with the origin in its interior. Assume that f, B, and u_S are C^1 on an open neighborhood U of $\mathcal{K}$, with $f(0) = u_S(0) = 0$. Differentiability near the origin and compactness away from it give $\|a_S(x)\| \le C_a \|x\|$ on $\mathcal{K}$. Choose $\rho > 0$ such that the closed ρ-neighborhood $\mathcal{K}^+$ of $\mathcal{K}$ is a compact subset of U. The vector fields b_x and their derivatives with respect to y are uniformly bounded for $x \in \mathcal{K}$ and $y \in \mathcal{K}^+$. Local existence and these bounds give $h_0 > 0$, independent of x, such that $z(t; x)$ in equation (15) exists for $0 \le t \le h_0$ and stays within distance ρ of x. Thus the segments from x to $z(t; x)$ also lie in $\mathcal{K}^+$, and, for a uniform constant $L > 0$,

$$\|b_x(z(t; x)) - b_x(x)\| \le L \|z(t; x) - x\| .$$

Writing $F_t(x) := z(t; x)$ for this single held-input interval, the integral equation gives

$$F_t(x) - x = ta_S(x) + \int_0^t \big[b_x(F_s(x)) - b_x(x)\big] \, \mathrm{d}s.$$

Grönwall's inequality gives a constant $C_F > 0$, independent of x and t, such that

$$\|F_t(x) - x\| \le C_F t \|x\| , \qquad x \in \mathcal{K}, \quad 0 \le t \le h_0. \tag{34}$$

Consequently, the triangle inequality yields $\|F_t(x)\| \le (1+C_F h_0)\,\|x\|$ on the same range. Moreover, the integral equation, together with the Lipschitz and displacement bounds, yields the first-order expansion

$$\begin{aligned} F_h(x) &= x + h a_S(x) + \xi_h(x), \\ \|\xi_h(x)\| &\le C_\xi h^2\,\|x\|, \end{aligned} \tag{35}$$

for $x \in \mathcal{K}$ and $0 < h \le h_0$, where $C_\xi := LC_F/2$. All constants are independent of x and h in these ranges.

Lemma A.1. *Suppose that f, B, P, and V_S are C^2 near the origin, and that $P_0 = P(0) \succ 0$ is the stabilizing solution of the equilibrium CARE introduced in Section 2. Let F_h be the held-input map (15). Then there exist a compact neighborhood $\mathcal{N} \subset \Omega$ of the origin and constants $h_0, M, \lambda, C_S > 0$ such that, for every $x \in \mathcal{N}$ and $0 < h \le h_0$, the continuous teacher trajectory and all iterates of the map F_h remain in $\mathcal{N}$ and satisfy*

$$\|x_S(t;x)\| \le Me^{-\lambda t}\,\|x\|, \qquad \left\|F_h^k(x)\right\| \le Me^{-\lambda kh}\,\|x\|, \quad t \ge 0,\ k \ge 0. \tag{36}$$

Consequently,

$$\int_0^\infty \|x_S(t;x)\|^2\,\mathrm{d}t \le \frac{M^2}{2\lambda}\,\|x\|^2, \qquad x \in \mathcal{N}. \tag{37}$$

Moreover,

$$|r_h[V_S](x)| \le C_S h\,\|x\|^2, \qquad x \in \mathcal{N}, \quad 0 < h \le h_0. \tag{38}$$

Proof of Lemma A.1. Equation (8) gives

$$u_S(0) = 0, \qquad \frac{\partial u_S}{\partial x}(0) = -R^{-1}B_0^\top P_0.$$

We first consider the continuous teacher dynamics. The C^2 assumptions and the preceding identities imply

$$a_S(x) = A_c x + \mathcal{O}(\|x\|^2)$$

locally. Since P_0 is the stabilizing CARE solution, A_c is Hurwitz. Hence the continuous-time Lyapunov equation

$$A_c^\top H + HA_c = -I$$

has a unique symmetric positive-definite solution H. Define the quadratic function $W(x) := x^\top Hx$. Since $\nabla W(x) = 2Hx$, its derivative along the continuous teacher dynamics satisfies

$$\begin{aligned} \nabla W(x)^\top a_S(x) &= 2x^\top H\big(A_c x + \mathcal{O}(\|x\|^2)\big) \\ &= x^\top (A_c^\top H + HA_c)x + \mathcal{O}(\|x\|^3) \\ &= -\|x\|^2 + \mathcal{O}(\|x\|^3). \end{aligned}$$

We next consider the held-input map (15). Applying equation (34) and equation (35) on a sufficiently small closed ball about the origin gives

$$\begin{aligned} W(F_h(x)) - W(x) &= h\nabla W(x)^\top a_S(x) + \nabla W(x)^\top \xi_h(x) + (F_h(x)-x)^\top H(F_h(x)-x) \\ &= h\nabla W(x)^\top a_S(x) + \mathcal{O}(h^2\,\|x\|^2) \\ &= -h\,\|x\|^2 + \mathcal{O}(h\,\|x\|^3 + h^2\,\|x\|^2). \end{aligned}$$

The remainder bounds are uniform in x on this ball and in $0 < h \le h_0$.

Choose $r > 0$ and reduce $h_0 > 0$ sufficiently so that

$$\nabla W(x)^\top a_S(x) \le -\frac{1}{2}\,\|x\|^2, \tag{39}$$

$$W(F_h(x)) - W(x) \le -\frac{h}{2}\,\|x\|^2 \tag{40}$$

for $\|x\| \le r$ and $0 < h \le h_0$. Choose $\alpha > 0$ such that $\mathcal{N} := \{x : W(x) \le \alpha\}$ is contained in this ball. The first inequality makes $\mathcal{N}$ positively invariant under the continuous teacher dynamics.

Compactness and local Lipschitz continuity of a_S ensure existence for all $t \geq 0$. The second inequality gives $F_h(\mathcal{N}) \subseteq \mathcal{N}$, so all iterates of F_h are well defined and remain in $\mathcal{N}$.

The quadratic function W satisfies

$$\lambda_{\min}(H) \|z\|^2 \leq W(z) \leq \lambda_{\max}(H) \|z\|^2 .$$

Set $\gamma := 1/(2\lambda_{\max}(H))$ and reduce h_0 if necessary so that $\gamma h_0 < 1$. The preceding decay inequalities and the upper eigenvalue bound imply

$$\nabla W(x)^\top a_S(x) \leq -\gamma W(x), \qquad W(F_h(x)) \leq (1-\gamma h) W(x),$$

for $x \in \mathcal{N}$ and $0 < h \leq h_0$. Scalar comparison and iteration give

$$W(x_S(t;x)) \leq e^{-\gamma t} W(x),$$
$$W(F_h^k(x)) \leq (1-\gamma h)^k W(x) \leq e^{-\gamma k h} W(x),$$

for $x \in \mathcal{N}$, $t \geq 0$, $k \geq 0$, and $0 < h \leq h_0$. By the same eigenvalue bounds, both estimates in equation (36) follow with

$$M := \sqrt{\frac{\lambda_{\max}(H)}{\lambda_{\min}(H)}}, \qquad \lambda := \frac{\gamma}{2}.$$

Integrating the squared continuous-time estimate yields equation (37).

It remains to prove equation (38). The teacher-policy GHJB relation established in Section 2 gives

$$\nabla V_S(x)^\top a_S(x) = -\ell(x, u_S(x)).$$

The sublevel set $\mathcal{N}$ can be chosen within the neighborhood where V_S is C^2. Since $V_S \geq 0$ and $V_S(0) = 0$, we have $\nabla V_S(0) = 0$. Compactness of $\mathcal{N}$ gives

$$L_V := \sup_{y \in \mathcal{N}} \left\| \nabla^2 V_S(y) \right\| < \infty.$$

Convexity of $\mathcal{N}$ then implies $\|\nabla V_S(x)\| \leq L_V \|x\|$ for $x \in \mathcal{N}$.

Moreover, $F_h(\mathcal{N}) \subseteq \mathcal{N}$, so the segment joining x to $F_h(x)$ lies in $\mathcal{N}$. Taylor's theorem, together with equation (34) and equation (35), therefore yields

$$\begin{aligned} \left| V_S(F_h(x)) - V_S(x) - h \nabla V_S(x)^\top a_S(x) \right| &\leq \|\nabla V_S(x)\| \, \|\xi_h(x)\| + \frac{L_V}{2} \|F_h(x) - x\|^2 \\ &\leq L_V \left(C_\xi + \frac{C_F^2}{2} \right) h^2 \|x\|^2 , \end{aligned}$$

uniformly for $x \in \mathcal{N}$ and $0 < h \leq h_0$. Using the GHJB relation above and equation (18), division by h proves equation (38) with any positive constant $C_S \geq L_V(C_\xi + C_F^2/2)$. □

Remark A.1. Lemma A.1 verifies equations (22)–(24) on $\mathcal{N}$. Invariance and one-step consistency follow directly from the lemma. For W and H constructed in the proof, summing the discrete Lyapunov decrease inequality (40) gives

$$h \sum_{k=0}^{\infty} \left\| F_h^k(x) \right\|^2 \leq 2W(x) \leq 2\lambda_{\max}(H) \|x\|^2 ,$$

uniformly for $x \in \mathcal{N}$ and $0 < h \leq h_0$. Thus equation (23) holds with $\Gamma = 2\lambda_{\max}(H)$. On a larger region $\mathcal{K}$, these regional conditions must be verified separately.

A.2 Proof of Theorem 4.1

Proof. Fix $h \in (0, h_0]$ and $x \in \mathcal{K}$. Let $e := V_\theta - V_S$ and $x_k := F_h^k(x)$. By equation (22), $x_k \in \mathcal{K}$ for every $k \geq 0$. Subtracting the two residuals in equation (18) and summing along the iterates gives, for every integer $N \geq 1$,

$$e(x) - e(x_N) = h \sum_{k=0}^{N-1} \big(r_h[V_\theta](x_k) - r_h[V_S](x_k) \big).$$

The residual bounds (24) and (25), together with equation (23), imply

$$|e(x)| \le |e(x_N)| + h(\varepsilon_V + C_S h) \sum_{k=0}^{N-1} \|x_k\|^2 \le |e(x_N)| + \Gamma(\varepsilon_V + C_S h) \|x\|^2 .$$

Since $h > 0$ is fixed, equation (23) gives $\sum_{k=0}^{\infty} \|x_k\|^2 < \infty$, and hence $x_N \to 0$. Continuity of V_θ and V_S, together with $V_\theta(0) = V_S(0) = 0$, implies $e(x_N) \to 0$. Letting $N \to \infty$ in the preceding inequality proves equation (26). □

A.3 Proof of Theorem 4.2

Proof. Under the hypotheses of the theorem, the uniform estimates (34) and (35) apply on $\mathcal{K}$ for sufficiently small h_0.

By equation (13), $\nabla V_\theta(0) = 0$. The C^2 regularity of V_θ, together with differentiability near the origin and compactness away from it, gives $\|\nabla V_\theta(x)\| = \mathcal{O}(\|x\|)$ uniformly on $\mathcal{K}$. After reducing h_0 if necessary, the segments joining x to $F_h(x)$ lie in a fixed compact subset of the neighborhood where V_θ is C^2. The Hessian is therefore uniformly bounded on these segments. Taylor's theorem and the held-input estimates give

$$\begin{aligned} V_\theta(F_h(x)) - V_\theta(x) &= \nabla V_\theta(x)^\top (F_h(x) - x) + \mathcal{O}(\|F_h(x) - x\|^2) \\ &= h \nabla V_\theta(x)^\top a_S(x) + \mathcal{O}(h^2 \|x\|^2), \end{aligned}$$

uniformly for $x \in \mathcal{K}$ and $0 < h \le h_0$. Dividing by h and using equation (18) and equation (3) yields

$$|r_h[V_\theta](x) + H(x, u_S(x), V_\theta)| \le Ch \|x\|^2 , \qquad x \in \mathcal{K},$$

where $C > 0$ may depend on $\mathcal{K}$ and the fixed network parameters, but is independent of x and h.

By equation (14), $B(x)^\top \nabla V_\theta(x) = -2R u_\theta(x)$. Completing the square in the control variable gives

$$H(x, u_S(x), V_\theta) = H(x, u_\theta(x), V_\theta) + \|u_\theta(x) - u_S(x)\|_R^2 .$$

Substituting this identity into the preceding residual estimate proves equation (27). The triangle inequality then gives equation (28). □

A.4 Proof of Theorem 4.3

Proof. Fix an initial state x satisfying the trajectory hypotheses of the theorem, and write $x_\theta(t) := x_{u_\theta}(t; x)$. The regularity assumptions of Theorem 4.2, together with equation (12) and equation (14), imply that the learned closed-loop vector field $f + Bu_\theta$ is C^1 on a neighborhood of $\mathcal{K}$. Since $\mathcal{K}$ is compact and $x_\theta(t) \in \mathcal{K}$ for all $t \ge 0$,

$$\sup_{t \ge 0} \|\dot{x}_\theta(t)\| \le \sup_{y \in \mathcal{K}} \|f(y) + B(y) u_\theta(y)\| < \infty.$$

Hence x_θ is uniformly continuous on $[0, \infty)$. Together with equation (29), this implies $x_\theta(t) \to 0$. Indeed, otherwise uniform continuity would give infinitely many disjoint intervals of a common positive length on which $\|x_\theta(t)\|$ stays bounded away from zero, contradicting the finite state energy.

By equation (27), equation (25), and the chain rule,

$$\frac{\mathrm{d}}{\mathrm{d}t} V_\theta(x_\theta(t)) + \ell(x_\theta(t), u_\theta(x_\theta(t))) + \|u_\theta(x_\theta(t)) - u_S(x_\theta(t))\|_R^2 \le (\varepsilon_V + Ch) \|x_\theta(t)\|^2 .$$

For any $T > 0$, integration, the nonnegativity of V_θ from equation (12), and equation (29) give

$$\begin{aligned} &\int_0^T \ell(x_\theta(t), u_\theta(x_\theta(t)))\, \mathrm{d}t + \int_0^T \|u_\theta(x_\theta(t)) - u_S(x_\theta(t))\|_R^2 \, \mathrm{d}t \\ \le\, &V_\theta(x) - V_\theta(x_\theta(T)) + (\varepsilon_V + Ch) \int_0^T \|x_\theta(t)\|^2 \, \mathrm{d}t \\ \le\, &V_\theta(x) + \Gamma_\theta(\varepsilon_V + Ch) \|x\|^2 . \end{aligned}$$

The two integrals on the left are nonnegative, nondecreasing in T, and uniformly bounded. They therefore have finite limits as $T \to \infty$, giving $J(u_\theta; x) < \infty$ and finite accumulated control mismatch. Combining the resulting bound with equation (26) proves equation (30). Dropping the nonnegative control-mismatch term yields equation (31). □

A.5 Proof of Theorem 4.4

Proof. Fix θ. Since $g_\theta \in C^2$, we have $e^{g_\theta(x)-g_\theta(0)} = 1 + \mathcal{O}(\|x\|)$, while ∇g_θ is locally bounded near the origin. Differentiating equation (12) yields

$$\nabla V_\theta(x) = e^{g_\theta(x)-g_\theta(0)} \left[2P_0 x + (x^\top P_0 x)\nabla g_\theta(x)\right] = 2P_0 x + \mathcal{O}(\|x\|^2).$$

Together with $B(x) = B_0 + \mathcal{O}(\|x\|)$, substitution into equation (14) proves equation (32).

Since B is C^1 and V_θ is C^2, u_θ is C^1 near the origin. Moreover, $f(0) = u_\theta(0) = 0$, so the origin is a closed-loop equilibrium. The product rule and equation (32) give

$$\frac{\partial(f + Bu_\theta)}{\partial x}(0) = A_0 + B_0 \frac{\partial u_\theta}{\partial x}(0) = A_0 - B_0 R^{-1} B_0^\top P_0 = A_c,$$

which proves equation (33). The closed-loop vector field is C^1, and A_c is Hurwitz and independent of θ. Lyapunov's indirect method therefore yields local exponential stability for every fixed θ, independently of training accuracy. □

A.6 A regional energy estimate

The following proposition provides sufficient conditions for the trajectory assumptions in Remark 4.3 on a value sublevel set.

Proposition A.2. *Assume the hypotheses of Theorem 4.2 and the residual bound (25) on $\mathcal{K}$. Let C be the constant in equation* (27)*. For some $\rho > 0$, suppose that*

$$\mathcal{K}_\rho := \{x \in \mathcal{K} : V_\theta(x) \le \rho\} \subset \operatorname{int}(\mathcal{K}).$$

If, for some $q > \varepsilon_V + Ch$,

$$\ell(x, u_\theta(x)) + \|u_\theta(x) - u_S(x)\|_R^2 \ge q\,\|x\|^2, \qquad x \in \mathcal{K}_\rho, \tag{41}$$

then, for every $x \in \mathcal{K}_\rho$, the learned trajectory exists uniquely for all $t \ge 0$, remains in $\mathcal{K}_\rho$, and satisfies

$$\int_0^\infty \|x_{u_\theta}(t;x)\|^2 \,\mathrm{d}t \le \frac{\overline{c}_\theta}{q - \varepsilon_V - Ch}\,\|x\|^2, \qquad x \in \mathcal{K}_\rho, \tag{42}$$

where $\overline{c}_\theta > 0$ is any constant satisfying $V_\theta(x) \le \overline{c}_\theta \|x\|^2$ on $\mathcal{K}_\rho$.

Proof. The regularity assumptions imply that $f + Bu_\theta$ is C^1 near $\mathcal{K}$, so local solutions exist uniquely. Combining equation (27), equation (25), and equation (41) gives

$$\frac{\mathrm{d}}{\mathrm{d}t} V_\theta(x_{u_\theta}(t;x)) \le -(q - \varepsilon_V - Ch)\,\|x_{u_\theta}(t;x)\|^2$$

as long as the trajectory lies in $\mathcal{K}_\rho$.

Continuity of V_θ and compactness of $\mathcal{K}$ imply that $\mathcal{K}_\rho$ is compact. Since $\mathcal{K}_\rho \subset \operatorname{int}(\mathcal{K})$, its boundary lies in $\{V_\theta = \rho\}$. The anchored parameterization gives $V_\theta(0) = 0 < \rho$, so this boundary excludes the origin and the derivative is strictly negative there. Thus $\mathcal{K}_\rho$ is positively invariant. Compactness then permits continuation of each solution for all $t \ge 0$.

The exponential factor in equation (12) is bounded on $\mathcal{K}_\rho$, so the stated quadratic upper bound holds for some $\overline{c}_\theta > 0$. Integrating the differential inequality and using $V_\theta \ge 0$ gives

$$(q - \varepsilon_V - Ch)\int_0^T \|x_{u_\theta}(t;x)\|^2 \,\mathrm{d}t \le V_\theta(x) - V_\theta(x_{u_\theta}(T;x)) \le \overline{c}_\theta\,\|x\|^2.$$

Letting $T \to \infty$ proves equation (42). Thus equation (29) holds for every initial state $x \in \mathcal{K}_\rho$, with $\Gamma_\theta = \overline{c}_\theta/(q - \varepsilon_V - Ch)$. □

Remark A.2. When $Q \succ 0$, condition (41) holds with $q = \lambda_{\min}(Q)$. For this choice, the additional requirement $q > \varepsilon_V + Ch$ becomes $\lambda_{\min}(Q) > \varepsilon_V + Ch$. If the hypotheses of Theorem 4.1 also hold on $\mathcal{K}$, then Theorem 4.3 applies to every initial state in $\mathcal{K}_\rho$. This certificate is sufficient and may be conservative.

A.7 COMPARISON WITH THE OPTIMAL COST

The next result adds an optimal reference. It is not needed for the teacher-relative guarantee in Section 4, but shows how teacher quality and imitation error transfer to optimal value and policy performance.

Theorem A.3. *Let $\mathcal{K}$, h, and V_θ satisfy the hypotheses of Theorem 4.1, with the constants appearing there. Suppose that V^* is C^1 on a neighborhood of $\mathcal{K}$, satisfies equation* (4) *on $\mathcal{K}$, and induces u^* through equation* (5)*. Assume that the closed-loop vector fields under u_S and u_θ are locally Lipschitz near $\mathcal{K}$. Suppose that, for every $x \in \mathcal{K}$, the trajectories $x_S(t;x)$ and $x_{u_\theta}(t;x)$ are well defined for all $t \geq 0$ and remain in $\mathcal{K}$. Assume that equation* (29) *holds for every $x \in \mathcal{K}$ with a common constant $\Gamma_\theta > 0$, and that, for some $\Gamma_S > 0$,*

$$\int_0^\infty \|x_S(t;x)\|^2 \, \mathrm{d}t \leq \Gamma_S \|x\|^2, \qquad x \in \mathcal{K}. \tag{43}$$

Finally, suppose that

$$\|u_S(x) - u^*(x)\|_R \leq \varepsilon_S \|x\|, \qquad \|u_\theta(x) - u_S(x)\|_R \leq \varepsilon_u \|x\|, \qquad x \in \mathcal{K}. \tag{44}$$

Then, for every $x \in \mathcal{K}$,

$$|V_\theta(x) - V^*(x)| \leq \left[\Gamma(\varepsilon_V + C_S h) + \Gamma_S \varepsilon_S^2\right] \|x\|^2, \tag{45}$$

$$0 \leq J(u_\theta; x) - V^*(x) \leq \Gamma_\theta(\varepsilon_S + \varepsilon_u)^2 \|x\|^2. \tag{46}$$

Proof of Theorem A.3. By equation (5), $B(x)^\top \nabla V^*(x) = -2Ru^*(x)$. Completing the square and using equation (4) gives, for every $v \in \mathbb{R}^m$,

$$H(x, v, V^*) = H(x, u^*(x), V^*) + \|v - u^*(x)\|_R^2 = \|v - u^*(x)\|_R^2.$$

Fix $x \in \mathcal{K}$. For $\pi \in \{u_S, u_\theta\}$, let $x_\pi(t;x)$ denote the closed-loop trajectory initialized at x. Integrating this identity along the trajectory and applying the chain rule yields

$$\begin{aligned} &\int_0^T \ell(x_\pi(t;x), \pi(x_\pi(t;x)))\, \mathrm{d}t + V^*(x_\pi(T;x)) - V^*(x) \\ &= \int_0^T \|\pi(x_\pi(t;x)) - u^*(x_\pi(t;x))\|_R^2 \, \mathrm{d}t. \end{aligned} \tag{47}$$

The closed-loop vector fields are bounded on the compact set $\mathcal{K}$, so both trajectories are uniformly continuous. Combining the energy bounds (43) and (29) with the convergence argument used in the proof of Theorem 4.3, we obtain $x_\pi(t;x) \to 0$. Hence $V^*(x_\pi(T;x)) \to V^*(0) = 0$.

Moreover, by equation (44) and the triangle inequality, $\|u_\theta(x) - u^*(x)\|_R \leq (\varepsilon_S + \varepsilon_u)\|x\|$ for all $x \in \mathcal{K}$. Together with the teacher mismatch bound and the energy estimates, this makes the squared policy mismatch integrable for both policies. Letting $T \to \infty$ in equation (47) therefore proves $J(\pi;x) < \infty$ and

$$J(\pi;x) - V^*(x) = \int_0^\infty \|\pi(x_\pi(t;x)) - u^*(x_\pi(t;x))\|_R^2 \, \mathrm{d}t. \tag{48}$$

Taking $\pi = u_S$, using $J(u_S;x) = V_S(x)$, and applying equation (44) and equation (43) gives

$$0 \leq V_S(x) - V^*(x) \leq \Gamma_S \varepsilon_S^2 \|x\|^2.$$

Combining this estimate with equation (26) and the triangle inequality proves equation (45). For $\pi = u_\theta$, the same identity, the preceding mismatch bound, and equation (29) prove equation (46).

□

Remark A.3. The teacher-energy bound (43) holds locally under the hypotheses of Lemma A.1, by equation (37). On a larger $\mathcal{K}$, it remains a regional assumption. Under its hypotheses, Proposition A.2 supplies the learned-trajectory conditions for initial states in $\mathcal{K}_\rho$. The corresponding teacher-trajectory conditions must still be verified before applying the performance bounds to those initial states.

In equation (44), ε_u measures control imitation error, whereas ε_S measures the teacher's deviation from u^* and requires additional analysis or access to the optimal reference. These upper bounds do not exclude a learned policy outperforming the teacher.

A.8 LOCAL COST ESTIMATES

The exact performance identity (48), combined with local exponential stability and LQR matching, gives the following refinement under the local hypotheses stated below.

Corollary A.4. *Assume the local hypotheses of Lemma A.1 and Theorem 4.4. Suppose, in addition, that V^* is C^3 near the origin, satisfies equation* (4) *and equation* (5) *there, and obeys the LQR-jet matching condition*

$$\nabla^2 V^*(0) = 2P_0. \tag{49}$$

Then, as $x \to 0$,

$$u_S(x) - u^*(x) = \mathcal{O}(\|x\|^2), \qquad u_\theta(x) - u^*(x) = \mathcal{O}(\|x\|^2), \tag{50}$$

and

$$V_S(x) - V^*(x) = \mathcal{O}(\|x\|^4), \qquad J(u_\theta; x) - V^*(x) = \mathcal{O}(\|x\|^4). \tag{51}$$

Proof of Corollary A.4. Fix θ. Since $V^*(0) = 0$ is a minimum and V^* is differentiable, $\nabla V^*(0) = 0$. The C^3 regularity, equation (49), the regularity of B, and equation (5) therefore give

$$\begin{aligned} \nabla V^*(x) &= 2P_0 x + \mathcal{O}(\|x\|^2), \\ u^*(x) &= -R^{-1} B_0^\top P_0 x + \mathcal{O}(\|x\|^2). \end{aligned}$$

The same first-order feedback expansion holds for u_S by equation (8) and the regularity of B and P, and for u_θ by equation (32). Subtracting these expansions proves equation (50).

Lemma A.1 and Theorem 4.4 give local exponential stability under both policies. Choose the initial neighborhood small enough that both trajectories remain where the HJB identity and the quadratic policy-error estimates hold. The finite-horizon identity (47) then applies for $\pi \in \{u_S, u_\theta\}$. Exponential decay gives $V^*(x_\pi(T; x)) \to 0$ and makes the squared policy mismatch integrable. Taking $T \to \infty$ therefore establishes equation (48) under these local assumptions.

Consequently, equation (50) and equation (48) yield

$$0 \le J(\pi; x) - V^*(x) = \mathcal{O}\left(\int_0^\infty \|x_\pi(t; x)\|^4 \, \mathrm{d}t \right) = \mathcal{O}(\|x\|^4), \qquad \pi \in \{u_S, u_\theta\},$$

where the last estimate follows by integrating the fourth power of the exponential decay bound. Since $J(u_S; x) = V_S(x)$, this proves equation (51).

□

Remark A.4. The fourth-order estimates in equation (51) concern the realized policy costs and rely on the additional matching condition (49). They do not imply $V_\theta(x) - V^*(x) = \mathcal{O}(\|x\|^4)$. The neighborhood and the implied constants may depend on the fixed network parameters θ.

B ALGORITHM DETAILS

Algorithm 1: Control-supervised Bellman value learning

Input: Demonstrations $\mathcal{D}_h$, B, Q, R, P_0, step h, and weights η_b, η_u.
Initialize the locally LQR-matched value model V_θ by equation (12);
for *each optimization step* **do**
  sample a mini-batch $\{(x_i, u_i, x_i^+)\}$ from $\mathcal{D}_h$;
  compute $V_\theta(x_i)$, $V_\theta(x_i^+)$, and $\nabla V_\theta(x_i)$;
  compute $u_\theta(x_i)$ by equation (14);
  compute $r_h[V_\theta](x_i)$ using x_i^+ and u_i;
  update θ with a stochastic-gradient optimizer using the mini-batch estimate of equation (21);
Output: Neural network parameter θ.

Euclidean versus R-weighted control loss. The control loss (20) uses the squared Euclidean norm, while the theoretical bounds use $\|u_\theta - u_S\|_R^2$. Since $R \succ 0$, their sample averages satisfy

$$\lambda_{\min}(R)\mathcal{L}_u(\theta) \le \frac{1}{N}\sum_{i=1}^{N} \|u_\theta(x_i) - u_i\|_R^2 \le \lambda_{\max}(R)\mathcal{L}_u(\theta).$$

For the three benchmarks, $R = I_{20}/20$, $R = 0.1$, and $R = 0.5I_2$, so the R-weighted sample average is $\mathcal{L}_u/20$, $0.1\mathcal{L}_u$, and $0.5\mathcal{L}_u$, respectively.

C Experimental details and additional results

C.1 Benchmark systems

All three benchmarks have a state-independent input matrix $B(x) \equiv B$ and use the running cost ℓ defined in equation (2). To generate each teacher transition, the control is held fixed at its left-endpoint value $u_S(x)$ over an interval of length h, while the nonlinear state dynamics are integrated using RK45 for Cucker–Smale and Allen–Cahn and BDF for Burgers.

For Cucker–Smale, we consider $N_a = 20$ agents with positions $y \in \mathbb{R}^{N_a}$, velocities $v \in \mathbb{R}^{N_a}$, and state $x = [y^\top, v^\top]^\top \in \mathbb{R}^{40}$. The dynamics are

$$\begin{bmatrix}\dot{y}\\ \dot{v}\end{bmatrix} = \begin{bmatrix}0_{N_a} & I_{N_a}\\ 0_{N_a} & L(y)\end{bmatrix}\begin{bmatrix}y\\ v\end{bmatrix} + \begin{bmatrix}0_{N_a}\\ I_{N_a}\end{bmatrix} u, \tag{52}$$

where 0_{N_a} is the $N_a \times N_a$ zero matrix and

$$[L(y)]_{ij} = \begin{cases} -\dfrac{1}{N_a}\displaystyle\sum_{k\neq i}\psi(y_i, y_k), & i = j,\\ \dfrac{1}{N_a}\psi(y_i, y_j), & i \neq j,\end{cases} \qquad \psi(a,b) = \frac{1}{1+|a-b|^2}.$$

The cost matrices are $Q = I_{40}/20$ and $R = I_{20}/20$. Training initial states are sampled uniformly from $(-1, 1)^{40}$, with $h = 0.2$.

For Allen–Cahn, we consider the PDE with Neumann boundary conditions (Albi et al., 2022),

$$\partial_t x(\xi, t) = 0.1\partial_{\xi\xi} x + x(1 - x^2) + \chi_{[0.6,0.9]}(\xi)u(t), \qquad \xi \in [0, 1],$$

where $u(t) \in \mathbb{R}$ and $\chi_{[0.6,0.9]}$ is the indicator function of the controlled interval. Finite differences on a uniform grid with $n = 51$ points give

$$\dot{x} = A(x)x + Bu, \qquad A(x) = 0.1D_{\xi\xi} + I_{51} - \mathrm{diag}(x \odot x), \tag{53}$$

where $D_{\xi\xi}$ is the discrete Laplacian incorporating the Neumann boundary conditions, $\odot$ denotes componentwise multiplication, and $B_i = \chi_{[0.6,0.9]}(\xi_i)$. We use $Q = 0.02I_{51}$ and $R = 0.1$. Training initial fields are sampled as

$$x_0(\xi) = a_{\mathrm{amp}} \sum_{k=1}^{5} a_k \cos(2\pi k\xi + \phi_k) + b,$$

with $a_k \sim \mathcal{U}(-1/k, 1/k)$, $\phi_k \sim \mathcal{U}(0, 2\pi)$, $a_{\mathrm{amp}} \sim \mathcal{U}(0.5, 2)$, and $b \sim \mathcal{U}(-1, 1)$. The held-input transition interval is $h = 0.01$.

For Burgers, Chebyshev pseudospectral discretization of the unstable Burgers-type PDE in Nakamura-Zimmerer et al. (2021a) gives

$$\dot{x} = -\frac{1}{2}D(x \odot x) + \nu D_{\xi\xi}x + \alpha \odot x \odot e^{-\beta x} + Bu, \tag{54}$$

with $x \in \mathbb{R}^{64}$, $u \in \mathbb{R}^2$, $\xi \in [-1, 1]$, and homogeneous Dirichlet boundary conditions. If $\widetilde{D}$ is the full 66-node Chebyshev differentiation matrix, then D and $D_{\xi\xi}$ are the interior blocks of $\widetilde{D}$ and $\widetilde{D}^2$, respectively; in general $D_{\xi\xi} \neq D^2$. The exponential acts componentwise. We use $\nu = 0.02$, $\beta = 0.1$,

$R = 0.5I_2$, and Q equal to the diagonal matrix of interior Clenshaw–Curtis quadrature weights. The reaction and actuator profiles are

$$\alpha(\xi) = 25(0.2^2 - \xi^2)\chi_{[-0.2,0.2]}(\xi),$$
$$b_1(\xi) = -25(\xi + 0.8)(\xi + 0.4)\chi_{[-0.8,-0.4]}(\xi),$$
$$b_2(\xi) = -25(\xi - 0.4)(\xi - 0.8)\chi_{[0.4,0.8]}(\xi).$$

Here $\alpha_i = \alpha(\xi_i)$ and $B_{ij} = b_j(\xi_i)$. The SDRE factorization is $A(x) = -\frac{1}{2}D\,\mathrm{diag}(x) + \nu D_{\xi\xi} + \mathrm{diag}(\alpha \odot e^{-\beta x})$. Training initial fields are $x_0(\xi) = \sum_{k=1}^{4} a_k \sin(k\pi\xi)$, where $a_k \sim \mathcal{U}(-1/k, 1/k)$. The held-input transition interval is $h = 0.01$.

C.2 Training data and network settings

Training-set construction. For each system, the training examples are one-step teacher transitions (x_i, u_i, x_i^+) as defined in equation (16), where $u_i = u_S(x_i)$ and $x_i^+ = F_h(x_i)$. For Cucker–Smale, each training trajectory has length $T_{\mathrm{train}} = 10$ with $h = 0.2$ and contributes 50 transitions; 200 trajectories therefore yield 10,000 training transitions. For Allen–Cahn and Burgers, each training trajectory has length $T_{\mathrm{train}} = 2$ with $h = 0.01$ and contributes 200 transitions; 50 trajectories per system therefore yield 10,000 training transitions. The 2,000- and 5,000-transition training sets are constructed in the same way, with the trajectory length and sampling interval unchanged and only the number of training trajectories varied.

Network architecture. Our method and Control-only use the scalar value parameterization (12) and the feedback (14). The correction g_θ has three hidden layers of width 256, $\tanh$ activations, and a linear scalar output. For numerical stability, the exponent $g_\theta(x) - g_\theta(0)$ is clipped to $[-12, 12]$ for Cucker–Smale and $[-10, 10]$ for Allen–Cahn and Burgers. Since $g_\theta(0) - g_\theta(0) = 0$ and g_θ is continuous, clipping is inactive on a neighborhood of the origin. It therefore leaves the local LQR matching and the local stability guarantee unchanged. Our method uses $\eta_b = \eta_u = 1$; Control-only uses $\eta_b = 0$, $\eta_u = 1$.

The u_{Jac}-QRnet (Nakamura-Zimmerer et al., 2022b) baseline directly represents the feedback as

$$u_{\mathrm{QR}}(x) = -K_0 x + \phi_\vartheta(x) - \phi_\vartheta(0) - \mathrm{D}\phi_\vartheta(0)x, \qquad K_0 = R^{-1}B(0)^\top P_0.$$

Here ϕ_ϑ has three hidden layers of width 256 with $\tanh$ activations and a linear output in $\mathbb{R}^m$; $\mathrm{D}\phi_\vartheta(0)$ denotes its Jacobian at the origin. Thus $u_{\mathrm{QR}}(0) = 0$ and $\mathrm{D}u_{\mathrm{QR}}(0) = -K_0$. This network is trained on SDRE control labels and has no scalar value output. The LQR baseline is $u_{\mathrm{LQR}}(x) = -K_0 x$. States and controls are used in their original coordinates, without input normalization or feature standardization. All networks are trained with Adam. For Cucker–Smale, the initial learning rate is 2×10^{-4}, the batch size is 128, and the maximum number of epochs is 4000. The learning rate is multiplied by 0.8 every 100 epochs, with early-stopping patience 400. For Allen–Cahn and Burgers, the initial learning rate is 2×10^{-3}, the batch size is 1280, and the maximum number of epochs is 3000. Cosine annealing reduces the learning rate to 10^{-6}, with early-stopping patience 300. Training was performed on an NVIDIA GeForce RTX 5060 Laptop GPU.

C.3 Evaluation protocol

For each problem, we sample three raw initial states independently from the training initial-state distribution. Let

$$E_{\max} = \max_{z \in \mathcal{X}_{0,\mathrm{train}}} \|z\|_Q, \qquad \|x\|_Q = \sqrt{x^\top Q x},$$

where $\mathcal{X}_{0,\mathrm{train}}$ denotes the set of training initial states. Each sampled state is rescaled as

$$\widehat{x}_0 = \frac{E_{\max}}{\|x_0\|_Q} x_0,$$

so that $\|\widehat{x}_0\|_Q = E_{\max}$ while preserving its direction. For each of the three directions, we then construct five test initial states $s\widehat{x}_0$ with

$$s \in \{1, 1.25, 1.5, 1.75, 2\}.$$

Thus, $s = 1$ places the test state at the largest observed training initial-state norm, whereas the remaining four scales place it beyond that range. The normalization is used only to construct the test initial states. The resulting 15 states are shared by all methods and network seeds, giving 45 rollouts for each learned method and 15 for each deterministic baseline, online SDRE and LQR. The horizons are $T = 30, 15, 15$ and the control-update intervals are $h = 0.2, 0.01, 0.01$ for Cucker–Smale, Allen–Cahn, and Burgers, respectively. Teacher rollouts are extended when the relative terminal correction exceeds 10^{-3}, with at most two extensions. These rollout values are used only for evaluation.

For a controller π and initial state x_0, we define the reported finite-horizon cost as

$$J_T(\pi; x_0) := \int_0^T \ell\big(x_\pi(t; x_0), u_\pi(t)\big)\, \mathrm{d}t,$$

where $u_\pi(t)$ is the applied sample-and-hold control.

The reported cost ratios exclude terminal corrections and are computed over all test rollouts. Therefore, a ratio $J_T(\pi; x_0)/J_T(u_S; x_0) < 1$ indicates that π accumulates less running cost than the online SDRE teacher over the simulated horizon for that initial state.

For value error and cost ratio, we first average over the 15 initial states for each network seed, then report the mean and standard deviation over the three seeds.

CPU latency is measured on an AMD Ryzen 9 8945HX processor at batch size one, with GPU acceleration disabled, using 1000 warm-up calls and 10,000 timed calls. We report the median per seed, averaged across seeds for learned methods.

C.4 Additional empirical results

Figure 3 illustrates convergence toward the origin under the learned controller in the three benchmark systems. Figure 4 supplements the Allen–Cahn example in Figure 1 with Cucker–Smale and Burgers value and control curves. Since the Cucker–Smale control is 20-dimensional, we plot only the time evolution of its ℓ^2 norm. For the illustrated Cucker–Smale example, our method and Control-only yield very similar value estimates. For Burgers, our method agrees more closely with the teacher's finite-horizon rollout cost-to-go than Control-only, although a discrepancy remains. A broader comparison of value recovery is provided in Table 1.

C.4.1 Sensitivity to the number of training transitions

Table 3 examines the sensitivity of our method to the number of training transitions. The network architecture, loss weights, training procedure, network seeds, validation and test data were kept fixed.

Increasing the training-set size from 2,000 to 5,000 transitions reduces the mean E_{AUC} for all three problems. With 10,000 transitions, E_{AUC} decreases further for Cucker–Smale, while Allen–Cahn and Burgers show modest increases relative to their 5,000-transition results but remain substantially below their 2,000-transition values. Thus, larger training sets generally improve value recovery over the tested range, although the improvement is not monotone for every problem. In contrast, the closed-loop cost ratio remains close to one across all training-set sizes. These results suggest that training-set size has a more pronounced effect on value recovery than on closed-loop control performance. We use 10,000 transitions as a common training budget for the final comparisons.

C.4.2 Generalization across initial-condition scales

Table 4 reports results at three representative scales, $s \in \{1, 1.5, 2\}$, selected from the five scales used in the evaluation. As defined in Appendix C.3, $s = 1$ places the test initial states at the largest observed training initial-state Q-norm, whereas $s = 1.5$ and $s = 2$ place them beyond that range. For each scale, each of the three trained networks is evaluated on the same three initial-condition directions, and its results are averaged over these directions to obtain one value per network seed. The table reports the mean and standard deviation of the resulting three seed-level averages.

Across the three reported scales, the mean cost ratio deviates from one by at most 1.8%. This indicates that our method generalizes well in closed-loop control performance across initial-condition scales, including scales beyond the observed training range. In contrast, value recovery shows a different

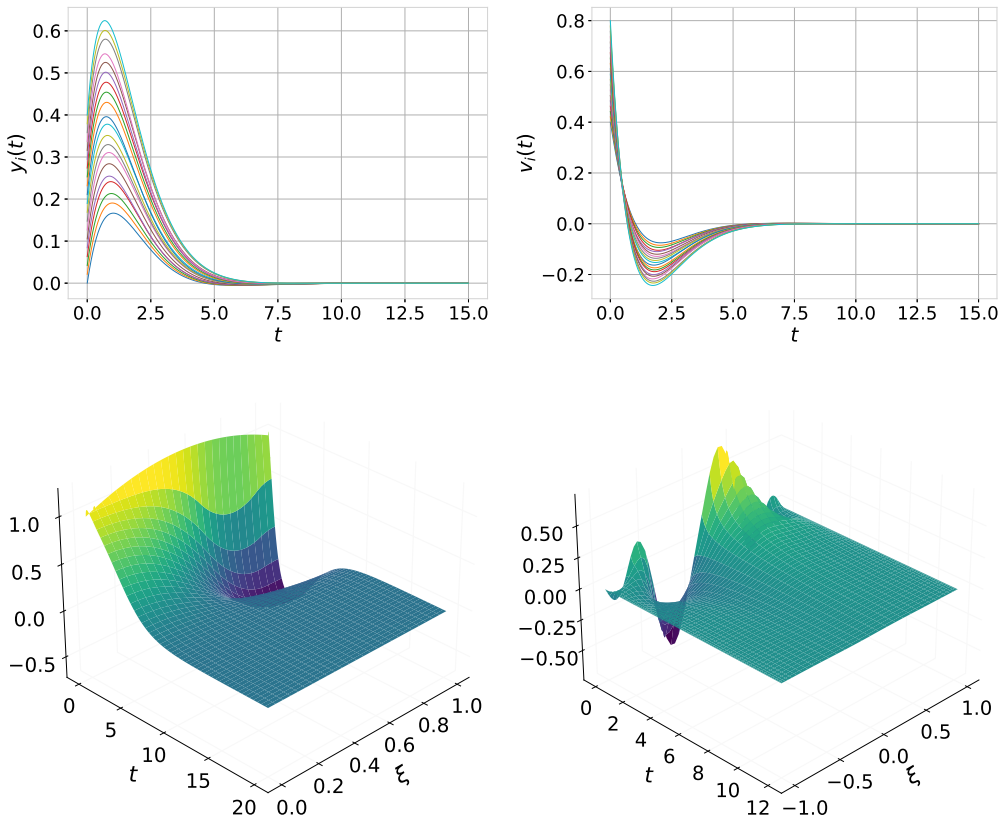


Figure 3: Representative closed-loop states under u_θ. Top row: Cucker–Smale positions and velocities. Bottom row, from left to right: Allen–Cahn and Burgers.

Table 3: Sample-size sensitivity of our method. Each entry is the mean and standard deviation of the three seed-wise averages over the same 15 test initial conditions at normalized scales $s \in \{1, 1.25, 1.5, 1.75, 2\}$.

System	Training transitions	$E_{\mathrm{AUC}} \downarrow$	$J_T(\pi)/J_T(u_S)$
Cucker–Smale	2,000	0.152 ± 0.006	1.0253 ± 0.0017
	5,000	0.110 ± 0.016	1.0133 ± 0.0022
	10,000	0.043 ± 0.011	1.0078 ± 0.0008
Allen–Cahn	2,000	0.235 ± 0.040	0.9918 ± 0.0051
	5,000	0.116 ± 0.007	0.9975 ± 0.0028
	10,000	0.125 ± 0.011	0.9939 ± 0.0021
Burgers	2,000	0.328 ± 0.060	1.0014 ± 0.0074
	5,000	0.206 ± 0.033	0.9871 ± 0.0080
	10,000	0.228 ± 0.035	0.9899 ± 0.0088

dependence on scale: it remains nearly unchanged for Cucker–Smale but becomes less accurate for Allen–Cahn and Burgers as the initial-condition scale increases, as reflected by the increase in E_{AUC}.

C.5 Step-size diagnostics for Bellman–control compatibility

Recall that, for a step size h, the one-step Bellman residual defined in equation (18) is

$$r_h[V_\theta](x) = \frac{V_\theta(x) - V_\theta(F_h(x))}{h} - \ell(x, u_S(x)). \tag{55}$$

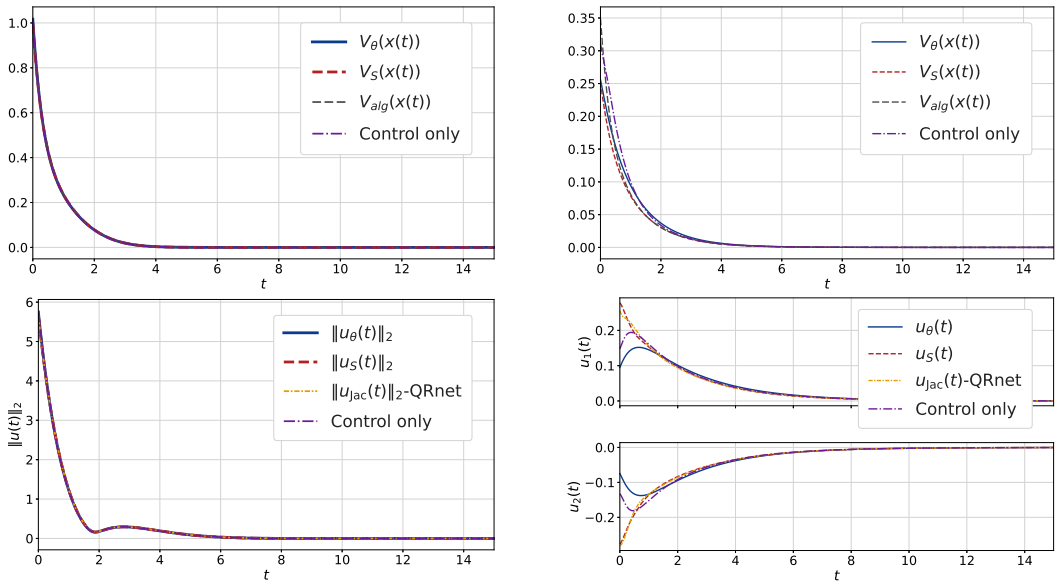


Figure 4: Value estimates and control signals for representative Cucker–Smale (left) and Burgers (right) trajectories. The top panels compare scalar outputs with the SDRE realized cost-to-go V_S and the algebraic surrogate V_{alg}. The bottom panels compare control signals with Control-only, SDRE and u_{Jac}-QRnet.

Table 4: Scale-dependent closed-loop cost and value-recovery error for our method trained with 10,000 transitions. Results are reported as the mean ± standard deviation over three network seeds.

System	Metric	Scale s		
		1	1.5	2
Cucker–Smale	$J_T(\pi)/J_T(u_S)$	1.0059 ± 0.0010	1.0073 ± 0.0006	1.0106 ± 0.0011
	E_{AUC}	0.040 ± 0.009	0.044 ± 0.011	0.043 ± 0.013
Allen–Cahn	$J_T(\pi)/J_T(u_S)$	0.9884 ± 0.0012	0.9939 ± 0.0021	0.9993 ± 0.0033
	E_{AUC}	0.093 ± 0.008	0.131 ± 0.013	0.147 ± 0.010
Burgers	$J_T(\pi)/J_T(u_S)$	0.9820 ± 0.0036	0.9902 ± 0.0088	0.9973 ± 0.0147
	E_{AUC}	0.131 ± 0.038	0.222 ± 0.037	0.338 ± 0.030

We denote the Bellman–control compatibility remainder appearing in equation (27) by

$$\delta_h(x) = H(x, u_\theta(x), V_\theta) + r_h[V_\theta](x) + \|u_\theta(x) - u_S(x)\|_R^2. \tag{56}$$

We examine how these two quantities vary with h. The trained networks and 50 independently sampled states are held fixed, while the held-input teacher successor $F_h(x)$ from equation (15) is recomputed at each value of h.

Table 5 shows that the median $|\delta_h|$ decreases as h is reduced in all three problems. This trend is consistent with the $\mathcal{O}(h\|x\|^2)$ bound established in Theorem 4.2. By contrast, the median Bellman residual remains at approximately the same scale for Allen–Cahn and Burgers, although it decreases for Cucker–Smale. This behavior is consistent with its continuous-time limit. Indeed, for a fixed smooth value model, equation (18) gives

$$r_h[V_\theta](x) = -H(x, u_S(x), V_\theta) + \mathcal{O}(h), \qquad \text{as } h \to 0.$$

Thus, reducing h need not eliminate the Bellman residual of a fixed learned value function.

Table 5: Step-size diagnostics for our method. Each entry is the median over 50 sampled states and three network seeds. The networks remain fixed while $F_h(x)$ is recomputed at each h.

System	h	Median $\lvert r_h[V_\theta]\rvert$	Median $\lvert\delta_h\rvert$
Cucker–Smale	0.20000	3.003×10^{-4}	3.035×10^{-4}
	0.10000	1.515×10^{-4}	1.608×10^{-4}
	0.05000	7.542×10^{-5}	8.279×10^{-5}
	0.02500	3.867×10^{-5}	4.049×10^{-5}
Allen–Cahn	0.01000	6.234×10^{-3}	8.314×10^{-4}
	0.00500	6.351×10^{-3}	5.668×10^{-4}
	0.00250	6.463×10^{-3}	3.981×10^{-4}
	0.00125	6.946×10^{-3}	2.155×10^{-4}
Burgers	0.01000	5.243×10^{-3}	4.016×10^{-4}
	0.00500	5.165×10^{-3}	1.998×10^{-4}
	0.00250	5.236×10^{-3}	1.251×10^{-4}
	0.00125	5.268×10^{-3}	6.393×10^{-5}